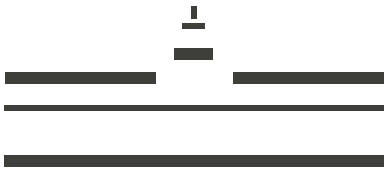


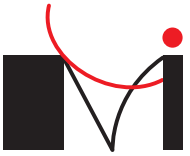
Westfälische
Wilhelms-Universität
Münster

› Bayesian Inversion in Biomedical Imaging

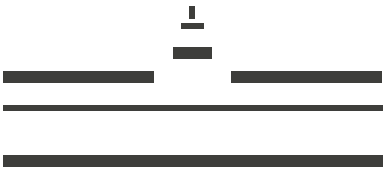
Felix Lucka

- 2014 -

wissen.leben
WWU Münster



Applied
Mathematics
Münster



Westfälische
Wilhelms-Universität
Münster

Fach: Mathematik

Bayesian Inversion in Biomedical Imaging

Inaugural Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften

– Dr. rer. nat. –

im Fachbereich Mathematik und Informatik

der Mathematisch-Naturwissenschaftlichen Fakultät

der Westfälischen Wilhelms-Universität Münster

eingereicht von

Felix Lucka

aus Hannover

– 2014 –

Dekan:	Prof. Dr. Martin Stein
Erster Gutachter:	Prof. Dr. Martin Burger (Universität Münster)
Zweiter Gutachter:	Prof. Dr. Samuli Siltanen (University of Helsinki, Finland)
Dritter Gutachter:	Priv.-Doz. Dr. Carsten H. Wolters (Universität Münster)
Tag der mündlichen Prüfung:	23.01.2015
Tag der Promotion:	23.01.2015

Abstract

Biomedical imaging techniques allow to assess the structure or function of living organisms in a non-invasive way. In recent years, they established various important new diagnostic and therapeutic approaches in clinical applications and became the key tool in many scientific studies. One prominent example is the understanding of function and pathology of the human brain on the macroscopic level.

Besides innovations in the instrumentation, the development of new and improved methods for processing and analysis of the measured data has become a vital field of research. Building on traditional signal processing, this area nowadays also comprises mathematical modeling, numerical simulation and inverse problems. The latter describes the reconstruction of quantities of interest from measured data and a given generative model. Unfortunately, most inverse problems are *ill-posed*, which means that a robust and reliable reconstruction is not possible unless additional *a-priori* information on the quantity of interest is incorporated into the solution method. *Bayesian inversion* is a mathematical methodology to formulate and employ a-priori information in computational schemes to solve the inverse problem.

This thesis develops a recent overview on Bayesian inversion and exemplifies the presented concepts and algorithms in various numerical studies including challenging biomedical imaging applications with experimental data. A particular focus is on using *sparsity* as a-priori information within the Bayesian framework. The back-and-forth between developments in theory, algorithms and their translation into real imaging applications was the guiding motivation behind the work for the thesis.

Key contributions

- Conceptual aspects of Bayesian inversion: The comparison of Bayesian inversion to other approaches to solve the inverse problem such as variational regularization and compressed sensing is an important aspect of this thesis. This entails the comparison between different estimation methods in Bayesian inference. For this, a unique combination of various computed examples and new, ground-breaking theoretical results are presented. Another focus is on the different ways in which sparsity is incorporated into Bayesian inversion and its implications.

- Development of fast algorithms for Bayesian inversion: Markov chain Monte Carlo (MCMC) methods are required to solve various computational tasks in Bayesian inversion. As common MCMC schemes are not applicable in high-dimensional settings, it was not possible to compute and evaluate several Bayesian inversion techniques in these scenarios so far. In this thesis, fast MCMC methods are developed that allow to perform such computations even in very high-dimensional problems. The results not only initiated the theoretical developments mentioned in the previous point but also challenge common beliefs about the applicability of Bayesian inversion in high-dimensional scenarios in general.

- Application of Bayesian inversion to experimental data: Reconstructing the brain-activity-related ion currents by measuring the induced electromagnetic fields outside the skull (EEG/MEG source reconstruction) constitutes a challenging, severely ill-posed inverse problem. In addition, its solution requires the usage of sophisticated preprocessing, modeling and simulation techniques. *Hierarchical* Bayesian inversion as one way to incorporate sparsity in the Bayesian framework is examined for EEG/MEG source reconstruction of experimental data. As a second application, a computed tomography (CT) scenario is examined. Bayesian inversion relying on the MCMC techniques mentioned in the previous point is developed for *total variation* and *Besov space* priors and applied to analyze a specific experimental data set.

Keywords: Bayesian inference; biomedical imaging; inverse problems; sparsity; compressed sensing; image deblurring; computed tomography; electroencephalography; magnetoencephalography; source reconstruction; ℓ_1; total variation; Besov space; Cauchy prior; Student's t prior; hierarchical Bayesian modeling; Bayesian estimation; maximum a-posteriori; conditional mean; Bayes cost; exact recovery conditions; source condition; Bregman distance; posterior sampling; MCMC; slice sampler; ADMM; simulated annealing; ordered overrelaxation; discretization invariance; noise modeling; realistic head modeling; finite element method

Acknowledgements

I want to thank everyone who made this thesis possible, especially:

- Martin Burger and Carsten H. Wolters for their supervision and support for traveling, conferences, summer schools, publishing and so on, over all the years. Most of all for giving me the freedom to pursue my own ideas and projects.
- Samuli Siltanen for reviewing this thesis.
- Ümit Aydin, Johannes Vorwerk, Bastian Pietras and Pia Heins for proofreading and help with printing this thesis.
- Everyone who helped me coping with the greatest horror an applied mathematician can possibly face: Processing experimental data! Thanks to Ümit Aydin, Benjamin Lanfer, Johannes Vorwerk, Andreas Wollbrink, Arno Janssen, Sumientra Rampersad and Seok Lew for help with the EEG/MEG data and Esa Niemi and Samuli Siltanen for help with the CT data.
- All other colleagues in Münster for helpful and/or silly discussions and a great time. Especially Sebastian Westerheide, Markus Knappitsch, Ralf Engbers, Frank Wübbeling, Christoph Brune, Carolin Gietz, Claudia Giesbert, Michi Möller, Martin Benning, Christian Himpe, Rene Milk, Andreas Buhr, Christian Engwer, Hendrik Dirks, Lena Frerking, Eva Brinkmann, Sven Wagner, Andreas Nüßing, Markus Junghöfer, Max Bruchmann and Christian Dobel.
- Andrea Bertozzi, Stanley Osher and their groups for inviting me to the UCLA and Mark Cohen and his group for discussions.
- Carola Schönlieb and her group for inviting me to Cambridge.
- More Finns: Sampsa Pursiainen, Ville Kolehmainen and Tapio Helin for support and discussions.
- My parents for their support at all times.
- Meinen Großeltern für ihre Unterstützung.

Studienstiftung
des deutschen Volkes

I am grateful for receiving a scholarship from the German National Academic Foundation (Studienstiftung des deutschen Volkes) including financial and ideational support.

In addition to the scholarship, this thesis profited from other sources of founding:

- I am grateful that the experimental data was made available to me:
  - The EEG/MEG data was recorded by Ümit Aydin and Andreas Wollbrink at the Institute for Biomagnetism and Biosignalanalysis (University of Münster) funded by the German Research Foundation (DFG), project WO1425/2-1.
  - The CT data was recorded by Aki Kallonen at the Tomography Laboratory of Keijo Hämäläinen (University of Helsinki Department of Physics). The mathematical measurement models were created and data preprocessed by Esa Niemi and Samuli Siltanen (University of Helsinki Department of Mathematics and Statistics), funded by the Finnish Centre of Excellence in Inverse Problems Research 2012–2017 (Academy of Finland decision number 250215).
- Several research visits and conference travels were funded by the German Research Foundation (DFG) through the trilateral Chinese-Finnish-German research project "Sparsity-constrained inversion with tomographic applications". Thanks to Samuli Siltanen, Matti Lassas, Peter Maass, Martin Burger, Jianguo Huang, Xiaoqun Zhang and all others involved in this project.
- Several conference travels were funded by the German Research Foundation (DFG) through project WO1425/2-1.

Contents

List of Figures

Notation and Abbreviations

As most of the notation and abbreviations will be introduced in the corresponding chapters, this listing serves as a reference for later look-up.

General notation

A_i	i-th column of a matrix A
A_i^T	i-th row of a matrix A
$u_{(i,j)}$	(i,j)-th pixel of u interpreted as a 2D pixel image
$A_{[i]}/u_{[i]}$	i-th bloc of matrix columns of A/ vector entries of u: 3.5.5
$u_{-i}/u_{[-i]}$	all components of u except the i-th one/bloc.
$Au \overset{ls}{=} f$	the linear system is solved in a least-squares sense.
$\mathbb{1}_{\mathcal{C}}(u)$	indicator function on the set $\mathcal{C}$
$\mathrm{unif}(a,b)$	continuous uniform distribution on $[a,b]$
I_n	identity matrix in n dimensions
$u \sim \ldots$	u follows a ... distribution or u is distributed like ...
$\mathcal{N}(\mu,\Sigma)$	multivariate normal probability distribution with mean $\mu \in \mathbb{R}^n$ and symmetric, positive semi-definite covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$

Important notation with a fixed meaning within the whole thesis

(in order of appearance)

$u^{\dagger,\infty} \in \mathcal{U}$	unknown quantity to recover: 1.2, (1.1)
$\mathcal{U}$	function space over domain Ω: 1.2
Ω	domain of $u^{\dagger,\infty}$: 1.2
$\mathcal{A} : \mathcal{U} \to \mathcal{X}$	infinite dimensional forward operator: 1.2, (1.1)
$f^{\dagger,\infty} \in \mathcal{X}$	infinite dimensional, noise-free data: 1.2, (1.1)

$m \in \mathbb{N}$	dimension of the measurements: 1.2
$f^\dagger \in \mathbb{R}^m$	finite dimensional projection of $f^{\dagger,\infty}$: 1.2
$\varepsilon \in \mathbb{R}^m$	noise variable: 1.2, (1.2)
$Noi : \mathbb{R}^m \to \mathbb{R}^m$	noise function: 1.2, (1.2)
$f \in \mathbb{R}^m$	finite dimensional, noisy data: 1.2, , (1.2)
$n \in \mathbb{N}$	dimension of the discretized unknowns: 1.2
$u^\dagger \in \mathbb{R}^n$	discretization of $u^{\dagger,\infty}$: 1.2, (1.3)
$A : \mathbb{R}^n \to \mathbb{R}^m$	discretization of $\mathcal{A}$: 1.2, (1.3)
$\mathcal{J}(u)$	regularization functional / prior energy: 1.2, (3.12), 3.2.3, (3.10)
$p_{like}(f\|u)$	likelihood probability distribution: 1.2, 3.1
$D^T \in \mathbb{R}^{h\times n}$	prior representation operator: 1.3, 3.2.4
$\|u\|_0 := \mathrm{card}(\{i\|u_i \neq 0\})$	number of non-zero elements of u ("ℓ_0 norm"): 1.3, 3.5
$M \in \mathbb{N}$	natural number describing the measurement dimension m: 2
$N \in \mathbb{N}$	natural number describing the discretization dimension n: 2
$r \in \mathbb{R}^3$	spatial vector: 2.4
$p_{prior}(u)$	a-priori probability distribution: 3.2
$p_{post}(u\|f)$	a-posteriori probability distribution: 3, (3.1)
$\hat{u}_{\mathrm{MAP}}$	maximum a-posteriori estimate: 3, (3.2)
$\hat{u}_{\mathrm{CM}}$	conditional mean estimate: 3, (3.3)
Σ_ε	covariance matrix of the Gaussian noise model: 3.1, (3.6)
$\gamma \in \mathbb{R}^h$	hyperparameter used in HBM: 3.3
$p_{hyper}(\gamma)$	hyperprior: 3.3
$D^p_{\mathcal{J}}(u, v)$	Bregman distance: A.1
K_0, K	burn-in and run length of MCMC: Algorithm 4.1
$\mathrm{erfc}(y)$	complementary error function $:= \frac{2}{\sqrt{\pi}} \int_y^\infty e^{-t^2}\,\mathrm{d}t$: 4.1.8
$\mathrm{erfcinv}(z)$	inverse complementary error function: 4.1.8
N_O	Number of samples in OOR: : 4.3.1

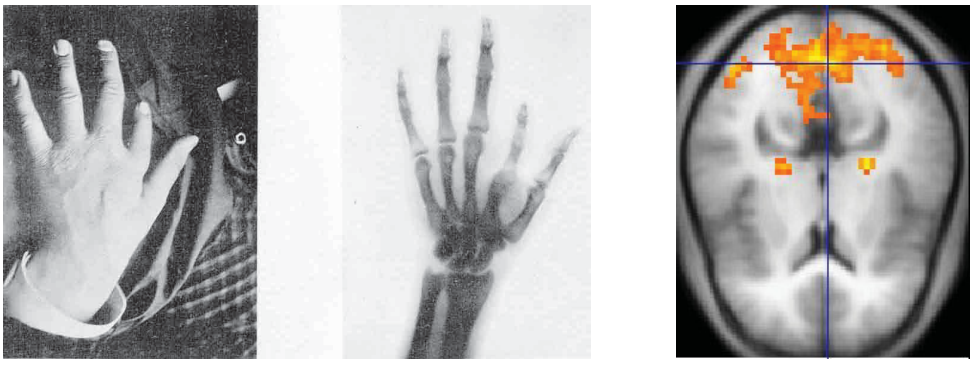
Frequently Used Abbreviations

acf	autocorrelation function: 4.1.6
ADMM	alternating direction method of multipliers: 4.2.2
AEP/AEF	auditory evoked potentials/fields: 2.4.2; 5.4.5
cdf	cumulative distribution function: 4.1.2
CM	conditional mean (estimate): 3; 3.4.3; 5; 6.1
CS	compressed sensing: 1.3; 3.5; 5.4.6
CStd	conditional standard deviation (estimate): 3.4.1; 5.3
CT	computed tomography: 1.1; 2.3; 5.3
DLE	dipole localization error: A.7
EEG	electroencephalography: 1.1; 2.4; 5.4
EMEG	electromagnetoencephalography: 1.1; 2.4; 5.4
EMD	earth mover's distance: A.7
EP/EF	evoked potentials/fields 2.4.2; 5.4.5
FEM	finite element method: 2.4.1; 5.4
GM	gray matter: 5.4.1
HBM	hierarchical Bayesian model/modeling: 3.3
i.i.d.	independent and identically distributed
MAP	maximum a-posteriori (estimate): 3; 3.4.3; 3.5; 5; 6.1
MCMC	Markov chain Monte Carlo: 4.1.3
MEG	magnetoencephalography: 1.1; 2.4; 5.4
MH	Metropolis-Hastings: 4.1.4
MNE	minimum norm estimate: 5.4.4
MRI	magnetic resonance imaging: 1.1; 5.4.1
NM	near mean (estimate): 4.2.5
OOR	Ordered overrelaxation: 4.3.1
SA	simulated annealing: 4.2.4
SC	single component (density): 4.1.5; 4.1.7
sLORETA	standardized low resolution brain electromagnetic tomography: 5.4.4
SRWMH	symmetric random-walk Metropolis Hastings: 4.1.4
SSC	strong source condition: 3.5.4
SSEP/SSEF	somatosensory evoked potentials/fields: 2.4.2; 5.4.5
SSR	sub-sampling rate: 4.1.3.
TV	total variation: 1.3; 3.2.4 5.2.2
WM	white matter: 5.4.1
WMNE	weighted minimum norm estimate: 5.4.4

1 Introduction

1.1. Biomedical Imaging

Background *Biomedical imaging* techniques try to map structure or function of living organisms to image-like data formats in a non-invasive way. Commonly, "non-invasive" in this context is understood in the sense that no solid instrumentation is introduced into the organism. Biomedical imaging techniques established various important new diagnostic and therapeutic approaches in clinical applications and became the key tool in many scientific studies. One prominent example is the understanding of function and pathology of the human brain on the macroscopic level: Neuroimaging allows



(a) (b)

Figure 1.1.: (a) Wilhelm Röntgen discovered X-rays in 1885. The photography shows an X-ray examination of a deformed hand published shortly after in *Nouvelle iconographie de la Salpétrière* (Tome 9; plaque XXI). (b) Brain regions of activation are visualized by a statistical extraction of the BOLD signal (yellow-red scale) on the background of an averaged structural T1-MRI image (grey scale). Source for both images: Wikimedia Commons.

neuroscientists to observe far more direct correlates of brain function than previously used measures such as behavioral experiments or lesion studies.

A classical example of an imaging technique is given by *radiography*, i.e., the use of X-rays to expose the anatomical structure of the body's tissues: Density and composition of the different tissues determine how much of a beam of X-rays transmitted through the body is absorbed; see Figure 1.1a. Functional imaging techniques often target correlates of the process they want to expose. A prominent example is the use of *functional magnetic resonance imaging* (*fMRI*) to image brain activity: Active neurons consume oxygen at a higher rate than inactive ones, causing the body to adjust its blood flow quickly (*hemodynamic response*). The difference in magnetic susceptibility of oxygen-rich and oxygen-poor blood (*BOLD*-contrast) can be detected using *magnetic resonance imaging* (*MRI*) technology; see Figure 1.1b. In this example, the BOLD signal is used as an indirect marker of the targeted neuronal activity.

The discovery of suitable, non-destructive interactions of biological tissues and processes with physical quantities such that the interaction exposes desired, measurable information provides the physical basis for imaging techniques. However, in contrast to a single X-ray projection (Figure 1.1a), the measured information often does not directly allow for human interpretation. Therefore, the development of modern imaging techniques crucially relied on the increasing availability and power of computers in the 1960's. *Computed tomography* (*CT*) refers to techniques that assemble single projection measurements to a tomographic image by the use of computational algorithms. A prominent example is the use of X-ray projections from multiple directions to produce 3D anatomical images (*X-ray CT*; see Figure 1.2). In general, all modern imaging techniques rely on computational algorithms to process and decode the measured information to reconstruct the quantities of interest in an image-like data format. The development of new and improved methods for these tasks has become a vital field of research. Building on traditional signal processing, this area nowadays also comprises mathematical modeling, numerical simulation and inverse problems. The latter describes the reconstruction of the quantities of interest from the measured data and a given generative model. Unfortunately, most inverse problems are mathematically *ill-posed*, which means that a robust and reliable reconstruction is not possible unless additional *a-priori* information on the quantity of interest is incorporated into the solution method. The next section will introduce different *inversion methods* that formulate and employ a-priori information in computational schemes to solve the inverse problem. This thesis is particularly devoted to the methodology of *Bayesian inversion*.

EEG and MEG Besides X-ray CT, the main imaging techniques examined in this thesis are *electroencephalography* (*EEG*) and *magnetoencephalography* (*MEG*): Neuronal

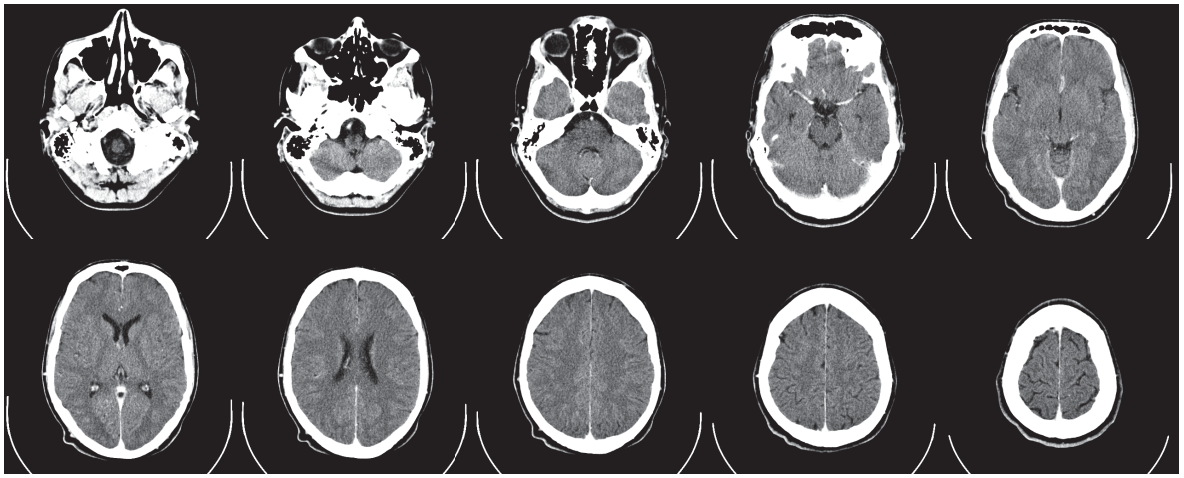


Figure 1.2.: Transversal slices of an X-ray CT of a human head. Source: Wikimedia Commons.

activity is accompanied by the flow of ionic currents. This current flow induces electromagnetic fields which propagate through the head's tissues and can be measured on the outside. Traditionally, EEG measurements were directly used for diagnostic or scientific purposes, much like *electrocardiography* (*ECG*) is used for the diagnosis of heart pathology. For these reasons, EEG and MEG are often not considered as imaging techniques in the original sense. However, a computational reconstruction of the neuronal activity to produce 3D images that align with other functional imaging techniques such as fMRI is possible (*EEG/MEG source reconstruction*). In the past decades, strong efforts to facilitate the use of EEG and MEG for this purpose were undertaken (cf. Michel and Murray 2012) and this thesis will present parts of the corresponding methodology.

Recent trends A single imaging modality is typically limited in two important aspects: First, it can only resolve certain temporal and spatial scales, and second, only a specific type of information is delivered (e.g., only anatomical *or* functional information). *Multimodal integration* tries to overcome these limitations by either fusing information from different measurements or by developing simultaneous measurement devices (see Figure 1.3). In this thesis, we will examine the results of combined EEG-MEG source analysis. *Imaging from coupled physics* or *hybrid imaging* follows a different approach: Instead of simply measuring two distinct imaging modalities at the same time, one imaging modality is used to probe the quantity of interest while the other is used to measure this interaction.

Traditional techniques were designed to deliver images to be read and interpreted by trained radiologists or scientists. This is a *qualitative* usage of the image information. Recently, there is an increasing demand for a *quantitative* usage of image data: Numer-

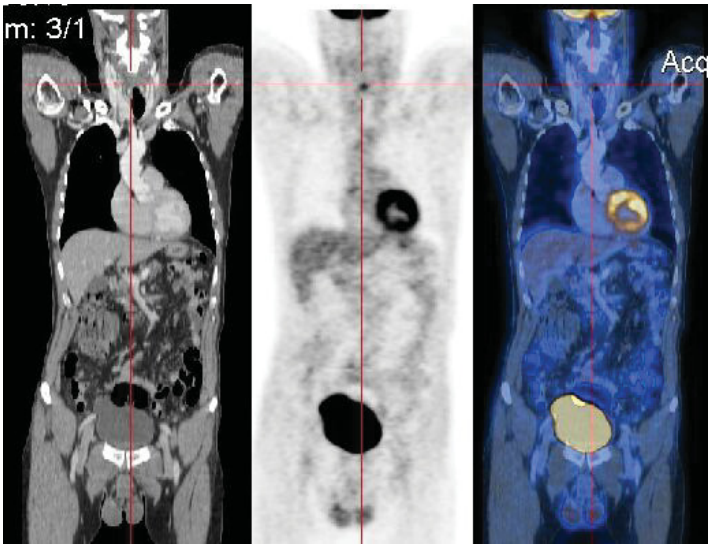


Figure 1.3.: Multimodal imaging by simultaneous acquisition of CT (left, gray scale) and *positron emission tomography* (*PET*, middle, inverted gray scale). The right image shows an overlay produced by a color-coded image fusion of the single modalities. Source: Wikimedia Commons.

ous images are subject to subsequent, automatized image processing procedures and the objective, quantitative results are used to statistically test a scientific hypothesis. An example from neuroimaging is given by *dynamic causal modeling* (*DCM*), a methodology to test hypotheses about the modulation of brain networks on the basis of the reconstructed spatio-temporal brain activity from groups of subjects. Another example is the automated acquisition and analysis of cell microscopy images to measure the statistics of dynamical cell processes, e.g., to assess the dynamics of neurotransmitters in synaptic transmission. On the basis of such procedures, hypotheses about the effects of diseases or drugs on these dynamics can be tested.

An apparent feature of the above examples is the focus on understanding (and potentially also modeling) dynamical processes. While this seems a natural aim for functional imaging, techniques traditionally used for imaging static, anatomical structures such as CT or MRI are also increasingly used and further developed to yield temporal information. This is especially important for applications where a coupling between anatomy and function is of interest, e.g., in cardiac imaging.

In the applications sketched above, two potential methodical challenges arise:

1. The shift towards subsequent quantitative, statistical analysis of the reconstructed images would benefit from a quantification of the uncertainties of the reconstruction procedure. Bayesian inversion techniques can deliver such measures.

2. The amount of data acquired and processed in these applications may require compression techniques to be integrated into the reconstruction procedure. We

will discuss *compressed sensing* techniques on several occasions throughout this theses.

1.2. Inverse Problems

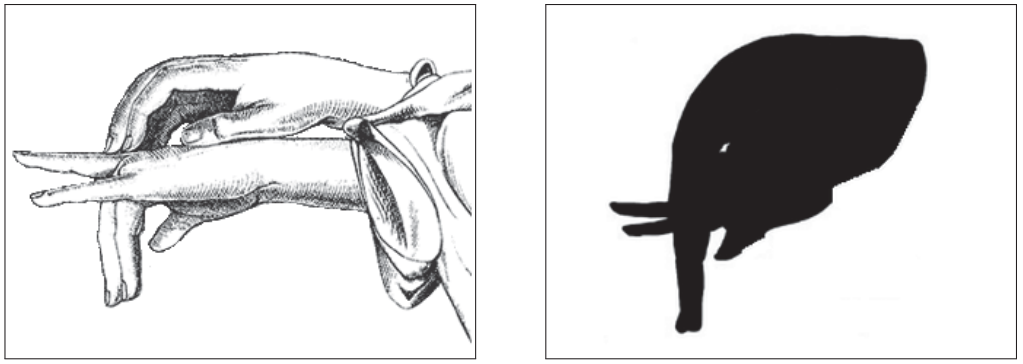
In this section, we will formalize the problem of image reconstruction that arises in all biomedical imaging applications we consider.

Setting The unknown image $u^{\dagger,\infty}$ is represented by an element of an infinite dimensional function space $\mathcal{U}$ over some domain Ω. From now on, "image" is to be understood in this abstract sense rather than with a correspondence to image-like data. The physical process by which the imaging modality generates measurable data is described by a *forward operator* $\mathcal{A} : \mathcal{U} \to \mathcal{X}$ which maps $u^{\dagger,\infty}$ to the infinite dimensional, noise-free data $f^{\dagger,\infty}$:

$$f^{\dagger,\infty} = \mathcal{A}(u^{\dagger,\infty}) \tag{1.1}$$

Most often, $\mathcal{A}$ is given as the solution operator of an underlying PDE model. We will examine concrete examples in Chapter 2. In practice, only a finite dimensional projection $f^{\dagger} \in \mathbb{R}^m$, $f^{\dagger} = Pf^{\dagger,\infty}$ can be measured, which is often also corrupted by noise:

$$f = Noi(f^{\dagger}, \varepsilon) = Noi(P\mathcal{A}(u^{\dagger,\infty}), \varepsilon) \tag{1.2}$$



(a) $u^{\dagger,\infty}$ **(b)** $\mathcal{A}(u^{\dagger,\infty})$

Figure 1.4.: An illustration of a typical non-uniqueness of inverse problems: (a) $u^{\dagger,\infty}$ is a 3D object and (b) $\mathcal{A}(u^{\dagger,\infty})$ is its 2D shadow given a light source and a wall. While it is simple to compute the shadow (forward problem), recovering $u^{\dagger,\infty}$ from the 2D projection is under-determined. This situation is stereotypical for many biomedical imaging applications which try to reconstruct quantities from lower-dimensional projections. Images were taken from *Hand Shadows To Be Thrown Upon The Wall* (Henry Bursill, 1895, available through www.gutenberg.net) and modified.

Here, $\varepsilon \in R^m$ denotes a noise variable, which can be modeled as deterministic or stochastic and $Noi : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ is a deterministic noise function.
While the discretization and perturbation of $f^{\dagger,\infty}$ is a natural consequence of the measurement process, discretization or other restrictions imposed on $u^{\dagger,\infty}$ are an intentional choice we have to make. As such, its consequences with regard to further assumptions on $u^{\dagger,\infty}$ and the concrete imaging application have to be considered. We will return to this issue in forthcoming sections and assume for now that a *computational model*,

$$f = Noi(A(u^\dagger), \varepsilon), \tag{1.3}$$

was chosen, where $u^\dagger \in \mathbb{R}^n$ represents a discretization of $u^{\dagger,\infty}$ and $A : \mathbb{R}^n \to \mathbb{R}^m$ subsumes discretization and projection of $\mathcal{A}$. A specific but often encountered scenario is to assume that A is linear and the noise function is a simple addition:

$$f = Au^\dagger + \varepsilon \tag{1.4}$$

Inverse Problems Solving any of the forward equations (1.1)-(1.4) for $u^{\dagger,\infty}$ or $u^\dagger$ constitutes an *inverse problem*: Generally speaking, we want to reconstruct the cause that led to an observed result. As this is also the basis of our everyday rational decisions, it seems like a simple task at first glance. However, the reversal of a (physical) causal relationship necessarily bears a potential ambiguity as many causes can lead to the same result. In addition, it suffers from an unavoidable loss of information or entropy increase due to energy dispersal. The mathematical field of inverse problems formalizes and examines scenarios, where these difficulties become severe, and simple approaches to invert (1.1)-(1.4) fail.
It turns out that inverse problems are typically *ill-posed* in the sense of Hadamard (Hadamard 1923):

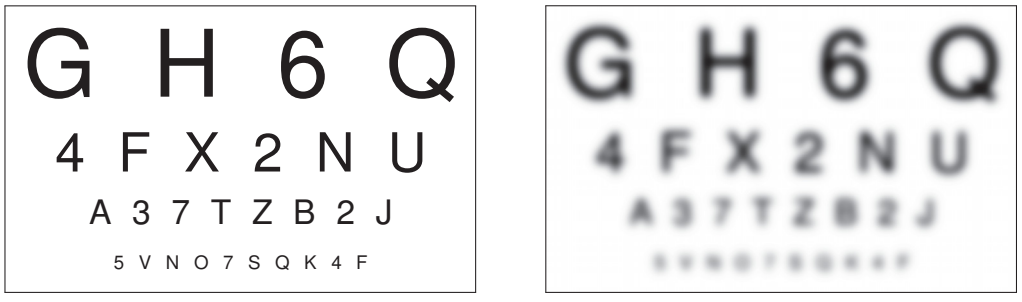
- The potential ambiguity in the cause-effect reversal can manifest in the non-uniqueness of the solution to (1.1)-(1.4): $\mathcal{A}$, respectively A is not invertible; see Figure 1.4.

- The abstract "loss of information" due to dispersal or dissipation manifests in the properties of $\mathcal{A}$. Typically, $\mathcal{A}$ is a compact operator whose singular values decay very fast. This leads to an increased loss (or compression) of information in the higher singular functions; see Figure 1.5. Therefore, a recovery of this information becomes difficult and unstable. In mathematical terms, the inverse of $\mathcal{A}$ restricted to its co-kernel is unbounded. As a result, the solution of (1.1) is not a continuous function of $f^{\dagger,\infty}$. These properties are inherited by A, which is ill-conditioned.

- While $f^{\dagger,\infty}$ may be in the range of $\mathcal{A}$ (whereby (1.1) would have a solution), $f = Noi(P\mathcal{A}(u^{\dagger,\infty}), \varepsilon)$ might not be in the range of $\mathcal{A}$. Even if so, the spectral properties of typical noise functions are very distinct from those of compact operators. For instance, in the situation of (1.4), ε typically also adds components to the singular vectors of A with very small singular values. As a consequence, the contribution of the noise dominates the signal in these singular functions. In simple terms, something is added to the signal, which should not be there anymore. Using a simple, straight forward inversion, these singular functions would be strongly amplified. Hence, the solution is dominated by the inversion of the noise.

In the following, we will sketch some of the main inversion frameworks developed in the field of inverse problems.

Regularization Theory The first approach relies on techniques and concepts developed for analyzing, solving or simulating the underlying PDEs. In particular, concepts from functional analysis such as spectral theory, the theory of weak solutions, Sobolev spaces and variational calculus are used to analyze $\mathcal{A}$ with respect to $\mathcal{U}$ and $\mathcal{X}$ to identify the structure of the ill-posedness of inverting (1.1). Typical questions to be answered concern the singular system of $\mathcal{A}$, existence and uniqueness of solutions and stability with respect to deterministic perturbations $f^{\dagger,\infty} + \delta\varepsilon$. Subsequently, the ill-posed problem is approximated by a well-posed one in a reasonable, controlled manner (*regularization*). *Variational regularization* is a particular regularization strategy, which is of importance to this thesis. The idea is to define the regularized solution as a minimizer of a suitable



(a) $u^{\dagger,\infty}$ **(b)** $\mathcal{A}(u^{\dagger,\infty})$

Figure 1.5.: An illustration of the effects of ill-condition: A simple Gaussian blurring (see Section 2.2) has a different impact on the image information on different spatial scales. While it does not prevent the visual system from correctly identifying the large characters, it renders the recognition of the small ones impossible.

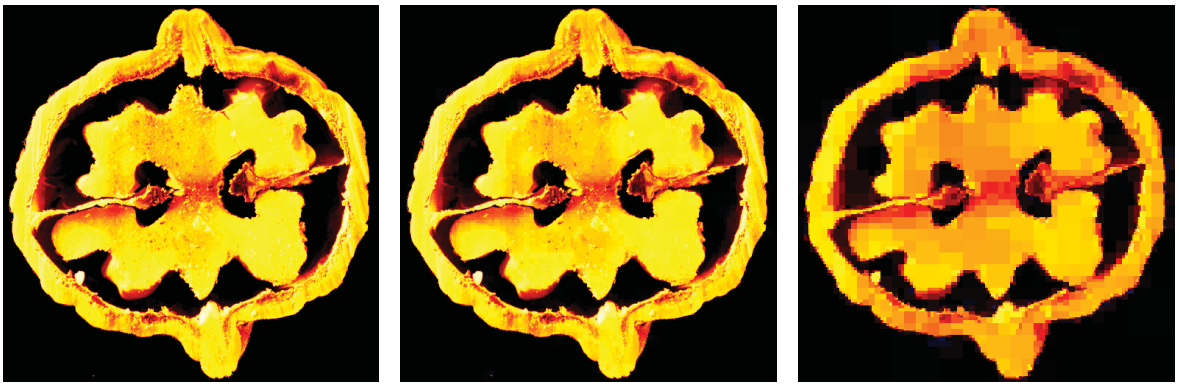
functional. Applied to the discrete forward equation (1.2), we obtain:

$$u_\lambda = \underset{u \in \mathcal{W} \subset \mathcal{U}}{\operatorname{argmin}} \left\{ \mathcal{H}_f(u) + \lambda \mathcal{J}(u) \right\}. \tag{1.5}$$

Here, $\mathcal{H}_f : \mathcal{W} \to \mathbb{R}$ measures the misfit between measured and predicted data (*data fidelity term*), usually in a distance suitable for the noise function in (1.2). The *regularization functional*, $\mathcal{J} : \mathcal{W} \to \mathbb{R}$ has to render the minimization problem (1.5) well-posed by ensuring existence, uniqueness and stability of u_λ. This can be analyzed by methods from variational calculus. In addition, $\mathcal{J}(u)$ can be used to penalize unwanted features of u_λ, thereby encoding *a-priori* knowledge about the solution. Especially if $\mathcal{H}_f(u)$ is a convex functional, using a $\mathcal{J}(u)$ which is also convex is a popular and well studied choice. Concepts from *convex analysis* can be used to examine the properties of u_λ by the optimality condition of (1.5). Questions of interest may be the rate of convergence in the noiseless limit or whether and how an exact recovery of certain features of $u^{\dagger,\infty}$ is possible. Typical for this approach is to formulate the inversion in a function space setting. A technical difficulty arises when stochastic noise models are considered.

Statistical Inference An approach starting from a stochastic noise model is to treat the inverse problem as a special instance of a statistical inference problem. For example, (1.4) can be seen as a classical linear regression problem. The unknowns of interest u become a possible *model of reality* belonging to a *class of models* $\mathcal{U}$. Forward modeling, discretization and noise contamination are summarized by a *forward mapping* $u \mapsto p_{like}(f|u)$, which links u to a *likelihood probability distribution* for f. The likelihood distribution is determined by (1.3). We will return to this issue in Section 3.1. If the forward mapping is one-to-one, the statistical model is *identifiable*. This is closely related to the uniqueness of solutions examined in regularization theory. Any function $\hat{u} : \mathbb{R}^m \to \mathcal{U}$ (e.g., given by (1.5)) which maps given data to an estimate of $u^\dagger$ is an *estimator*. As it relies on the random realization of the data, it is also a random variable. *Statistical decision theory* was developed to classify and validate estimators. Many concepts are closely related to concepts of regularization theory. One example is the *consistency* of estimators: Loosely speaking, an estimator for $u^\dagger$ is consistent, if it converges to $u^\dagger$ if the data gets "better". This is related to the general definition of regularization strategies which demand inverse methods to be continuous and to converge in the noiseless limit.

The statistical approach is very well suited to treat the stochastic nature of the noise. However, many concepts are less suitable to treat other features of typical inverse problems scenarios such as the ill-posedness or the inherently infinite dimensional nature



(a) (b) (c)

Figure 1.6.: (a) Photograph of a cross-section of a Walnut (see Section 2.3.3) and its Haar wavelet reconstruction after keeping only the (b) 10% or (c) 1% largest coefficients.

of the problem.

Bayesian Inference As this thesis will emphasize on the Bayesian approach to inverse problems, an extensive introduction will be given in Chapter 3. Loosely speaking, in the Bayesian approach the ill-posedness of the inversion is understood in the sense that f alone does not contain enough suitable information to determine $u^\dagger$. The remaining uncertainty can only be removed by incorporating additional a-priori information. Probability distributions are used to encode this and all other information available.

1.3. Sparsity

All inverse approaches rely on incorporating a-priori information on $u^{\dagger,\infty}$ in some way. *Sparsity* is the assumption that the most distinctive features of $u^{\dagger,\infty}$ can be approximated using only a few elements of a suitable representation system, i.e., a basis, frame or dictionary. We will speak of such $u^{\dagger,\infty}$ as being *compressible*. In Figure 1.6, an example for the compressibility of photographic images is presented. A straightforward realization of using sparsity as a-priori information to solve (1.3) would be to minimize the number of non-zero coefficients of u in a representation system encoded by $D^T \in \mathbb{R}^{h\times n}$:

$$\hat{u}_0 := \operatorname{argmin} |D^T u|_0, \quad s.\,t. \quad Au = f, \tag{1.6}$$

where $|v|_0 := \operatorname{card}(\{i|v_i \neq 0\})$ counts the number of non-zero elements of a vector. The non-convex optimization problem (1.6) is of combinatorial complexity. A common way to cope with that is to replace $|D^T u|_0$ by a convex surrogate, e.g., the ℓ_1-norm:

$$\hat{u}_1 := \operatorname{argmin} \left\| D^T u \right\|_1, \quad s.\,t. \quad Au = f, \tag{1.7}$$

Considering the ill-posed nature of (1.3), relaxing the strict equality constraint $Au = f$ is preferable. A possible way is to use a variational reformulation of the equality constraint and to minimize a weighted sum of sparsity and equality constraints:

$$\hat{u}_\lambda = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|f - A\,u\|_2^2 + \lambda \left\| D^T u \right\|_1 \right\} \tag{1.8}$$

This is obviously an instance of (1.5), and using variational regularization is the most common way to formulate sparsity constraints for inverse problems. One popular example is given by *total variation* (TV) inversion (Rudin et al. 1992), which imposes sparsity constraints on the gradient of $u^{\dagger,\infty}$ and will be examined in the upcoming chapters. Using sparsity constraints in Bayesian inversion is far less elaborate up to now and will be a central topic of this thesis.

Compressed Sensing Solving inverse problems using sparsity constraints is closely related to a recently developed signal processing framework called *compressed sensing* (Candes et al. 2006, Donoho 2006, Foucart and Rauhut 2013). In an abstract sense, inverse problems can also be viewed as a specific instance of signal transmission problems: The unknown signal $u^{\dagger,\infty}$ is *sampled* or *encoded* by $P\mathcal{A}$ into $f^\dagger$. Signal processing theory provides conditions under which a suitable recovery or *decoding* of $u^{\dagger,\infty}$ or $u^\dagger$ from $f^\dagger$ or f is possible. A classical example is the Nyquist-Shannon sampling theorem which guarantees the perfect recovery of frequency-limited signals if a certain sampling rate (*Nyquist rate*) is used. Compressed sensing extends these results for signals that are also *sparse* or *compressible*: Conditions were developed under which a non-linear decoding by (1.7) perfectly recovers a sparse, linearly encoded $u^{\dagger,\infty}$ or $u^\dagger$ using a sampling rate lower than the Nyquist rate (*undersampling*). Many applications of compressed sensing aim at designing sensing schemes which are optimal in the sense that the same reconstruction quality is achieved with as few data acquired as possible. This is also desirable in many biomedical imaging applications. For instance, the exposure to the ionizing X-radiation in CT is harmful for living tissue. Therefore, designing scanning procedures that provide the same diagnostic information while using less radiation is advantageous. In dynamic imaging modalities , higher temporal resolutions may be achieved using reduced spatial sensing schemes. However, in other inverse problems scenarios like the spatial inversion in EEG/MEG, the situation is contrary to the typical compressed sensing application: The sensing scheme used is fixed and already insufficient for a satisfactory recovery of the solution. Only incorporating additional information such as sparsity can compensate for this. On the other hand, the temporal inversion of EEG/MEG is a different issue since the temporal mapping is nearly one-to-one even at high sampling rates. Many inverse problems exhibit a similar split of its spatial and

temporal characteristics. In this case, complementing inverse methods in one domain with concepts of compressed sensing in the other domain can be advantageous.

1.4. Organization

This thesis is organized as follows: The next chapter introduces the imaging applications examined in this thesis. Chapter 3 develops the conceptual principles of the Bayesian approach to inverse problems while Chapter 4 presents the computational methods required to apply it in practice. Extensive numerical studies to illustrate and examine all the aspects discussed by then will be carried out in Chapter 5, which will also include the application of Bayesian inversion to challenging experimental data scenarios. Chapter 6 revisits and reflects the theoretical considerations of Chapter 3 in the light of the computational results and develops new theoretical ideas. Finally, Chapter 7 draws conclusions and highlights topics and directions for future work.
To enhance its readability, the presentation of the main text is kept intentionally concise. Several chapters or sections end with a "Notes and Comments" subsection, which supplements the main text with more detailed references and considerations about advanced topics. In addition, the appendix contains supplementary material and technical details.
Parts of the contents of this thesis were already published in journal articles, or presented in other forms. A detailed record of this is given in Section A.9.

1.5. Notes and Comments

Treating inverse problems by regularization theory, statistical or Bayesian inference expresses a different conceptual perspective on the same problem. All three have their own advantages and shortcomings and lead to more or less intuitive formulations in particular scenarios. However, while seemingly different in nature, they often lead to similar reconstruction methods. As a consequence, there is a growing interest to overcome the traditional boundaries of the separate fields. Section 3.5.6 will point to some important contributions in this direction.
Detailed references to Bayesian inference and compressed sensing will be given in Sections 3.7 and 3.5.6, respectively. Extensive introductions to regularization theory are given by Engl et al. (1996), Kirsch (1996), Schuster et al. (2012). Vogel (2002) complements these by a detailed introduction to the computational solution of inverse problems. Good introductions to the statistical approach to inverse problems can be found in Evans and Stark (2002), O'Sullivan (1986). A very recent, general and

application oriented introduction to linear and nonlinear inverse problems is given by Mueller and Siltanen (2012).

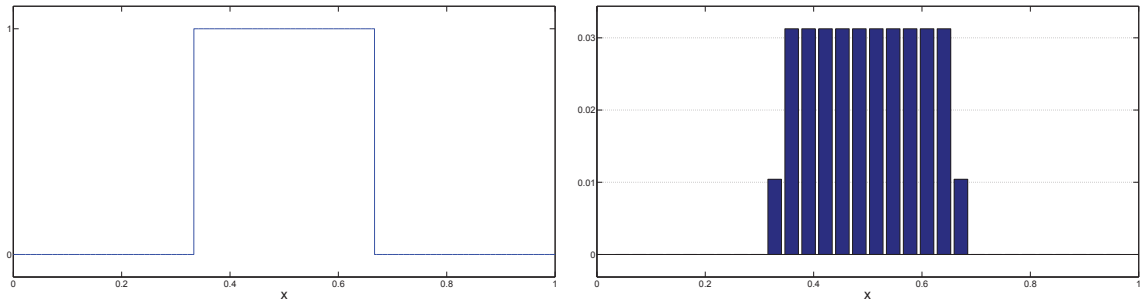
2

Computational Scenarios

In this chapter, we will introduce the computational scenarios used in this thesis. The image deblurring scenarios in Section 2.2 will mainly be used to illustrate the theoretical and computational aspects presented in Chapters 3 and 4. The more complex CT and EEG/MEG scenarios introduced in Sections 2.3 and 2.4 will also provide the basis for real data analysis in Chapter 5.

2.1. Inverse Crimes

Working with artificial inverse problems scenarios to examine certain theoretical or computational aspects comprises a potential pitfall: The scenario used to simulate the data and the one assumed for the inversion are usually more coherent than in real-world applications. In the extreme case, they even coincide: Inverse model and reality are identified. The inverse results obtained in such situations are usually of a better quality than those obtained in real-world scenarios. As a consequence, they may give an overly optimistic impression about the performance of a particular inverse method. We will speak of "inverse crimes" when referring to this difficulty of designing and interpreting computational studies with simulated data. While it is not possible to commit no inverse crimes at all when working with simulated data, it is easy to avoid some rather obvious ones. For instance, simulated data should never be generated using the same discretization (1.3) used in the inversion. Further information and illustration of this phenomena can be found in Siltanen (2009), Chapter 3, Kaipio and Somersalo (2007) and Colton and Kress (1992).



(a) Unknown function $u^{\dagger,\infty}$. **(b)** Clean measurement $f^{\dagger}$ for $m = 30$.

Figure 2.1.: "Boxcar" scenario.

2.2. Image Deblurring

Image deblurring problems are simple, intuitive and illustrative examples of inverse problems. Still, they already exhibit the elementary features of ill-posedness. We will examine two artificial scenarios, one in 1D and one in 2D.

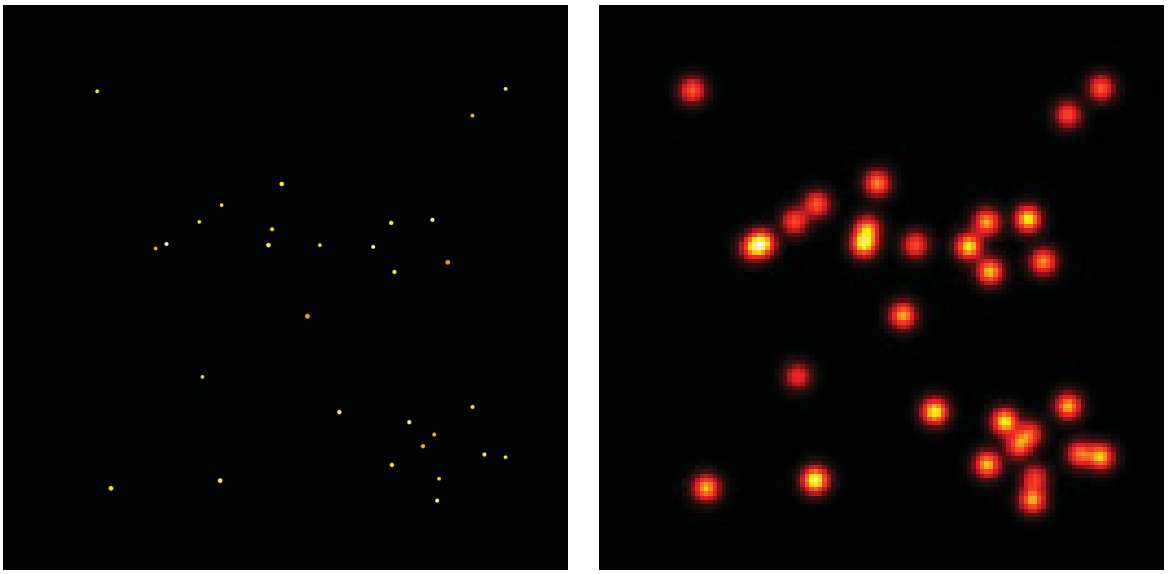
2.2.1. Boxcar Reconstruction in 1D

The first scenario mimics a measurement made by a *charge coupled device* (*CCD*) used in digital cameras or medical imaging devices. These devices integrate the amount of light illuminating a pixel over a certain period of time. It is adopted from Lassas and Siltanen (2004) which inspired some of our computational studies.
Using our previous notation, we model the unknown, continuous light intensity by a positive function $u^{\dagger,\infty} \in \mathcal{U}$, $u^{\dagger,\infty} : [0,1] \to \mathbb{R}^+$, where $\mathcal{U}$ is a suitable space of positive functions on $\Omega = [0,1]$. $\mathcal{A}$ is simply the identity on $\mathcal{U}$, while P models the integration into the m pixels of the CCD. The pixel are constructed as the inner subintervals of an equidistant division of $[0,1]$ into 2^M subintervals. As a result, the j-th of the $m = 2^M - 2$ pixel is represented by the interval $[\frac{j}{2^M}, \frac{j+1}{2^M}]$. The measurement $f_j^{\dagger}$ is then given by:

$$f_j^{\dagger} = \int_{\frac{j}{2^M}}^{\frac{j+1}{2^M}} u^{\dagger,\infty}(x)\,\mathrm{d}x \tag{2.1}$$

For discretizing $u^{\dagger,\infty}$, we choose the grid $x_i^n = \frac{i}{2^N}$, $i = 1,\dots,n$, with $n = 2^N - 1$ and $N > M$. The discretization of the forward mapping implied by (2.1) in terms of the $m \times n$ matrix A can then be implemented by the trapezoidal quadrature rule. The j^{th} row of A is given by

$$A_j^T := [\underbrace{0,0,\dots,0}_{j\,2^{(N-M)}-1}, \tfrac{1}{2}\delta_h, \underbrace{\delta_h,\delta_h,\dots,\delta_h}_{2^{(N-M)}-1}, \tfrac{1}{2}\delta_h, 0,0,\dots,0], \tag{2.2}$$



(a) Unknown function $u^{\dagger,\infty}$. **(b)** Clean measurement $f^\dagger$ for $m = 129 \times 129$.

Figure 2.2.: "Spots" scenario

where $\delta_h := \frac{1}{n+1}$ defines the grid size (we will always use B_i to denote the i-th column of a matrix B and B_i^T to denote its i-th row). The unknown function $u^{\dagger,\infty}$ we actually use is the indicator function (also called *boxcar* function) on $[\frac{1}{3}, \frac{2}{3}]$; see Figure 2.1a. In Figure 2.1b, the corresponding $f^\dagger$ is shown. It was computed directly by (2.1).
This first scenario will simply be called "Boxcar" from now on. It is a simplification of the task to reconstruct a spatially distributed intensity image that is known to consist of piecewise homogeneous parts with sharp edges. One example is given by the recovery of the body's organs and their boundaries from X-ray CT measurements; see Sections 2.3 and 5.3.

2.2.2. Point Source Reconstruction in 2D

As a second scenario, we examine the convolution of a 2D intensity function $u^{\dagger,\infty} : [0,1]^2 \to \mathbb{R}^+$ with a Gaussian kernel with standard deviation $\sigma_{\mathcal{A}} = 0.015$:

$$f^{\dagger,\infty} = g * u, \qquad \text{with} \qquad g(x,y) = \frac{1}{2\pi\,\sigma_{\mathcal{A}}^2} \exp\left(\frac{-(x^2+y^2)}{2\,\sigma_{\mathcal{A}}^2}\right) \tag{2.3}$$

The measurement is, again, given by a subsequent integration of $f^{\dagger,\infty}$ into measurement pixel. Here, we subdivide the image into $m = (2^N + 1) \times (2^N + 1)$ equidistant pixel. The unknowns will be reconstructed on the same pixel grid using Neumann boundary conditions; hence, $n = (2^N - 1) \times (2^N - 1)$. In this scenario, A will not be computed explicitly, but direct, *matrix-free* implementations of all computational operations

involving it will be used, see Section A.2. $f^\dagger$ will be computed using the same routines, but to avoid an obvious inverse crime, the grid used for these computations will be 4 times finer.
The concrete $u^{\dagger,\infty}$ we reconstruct is shown in Figure 2.2a. It consists of 30 point sources, i.e., circular spots of constant intensity. Their locations, radii and intensities were generated using a simple stochastic model. Figure 2.2b shows $f^\dagger$ for $m = 129 \times 129$. We will call this scenario "Spots" from now on.

2.3. Computed Tomography

In general, the propagation of electromagnetic radiation in biological tissues is a complex process. While *Maxwell's equations* give a valid mathematical description for it on the microscopic scale, quantities of interest for medical imaging applications can be described by the *radiative transport equation* (*RTE*). The RTE is a PDE for the spatial distribution of the steady-state spectral intensity $I(\lambda)$ of the radiation. In absence of internal sources, it includes *absorption* and *scattering* processes. For the almost monochromatic, high-energy X-radiation, scattering can be neglected and the RTE reduces to a simple ODE along the ray l (parameterized by $t \in [0, T]$):

$$\frac{d}{dt} I(t) = -u^{\dagger,\infty}(t)\, I(t) \tag{2.4}$$

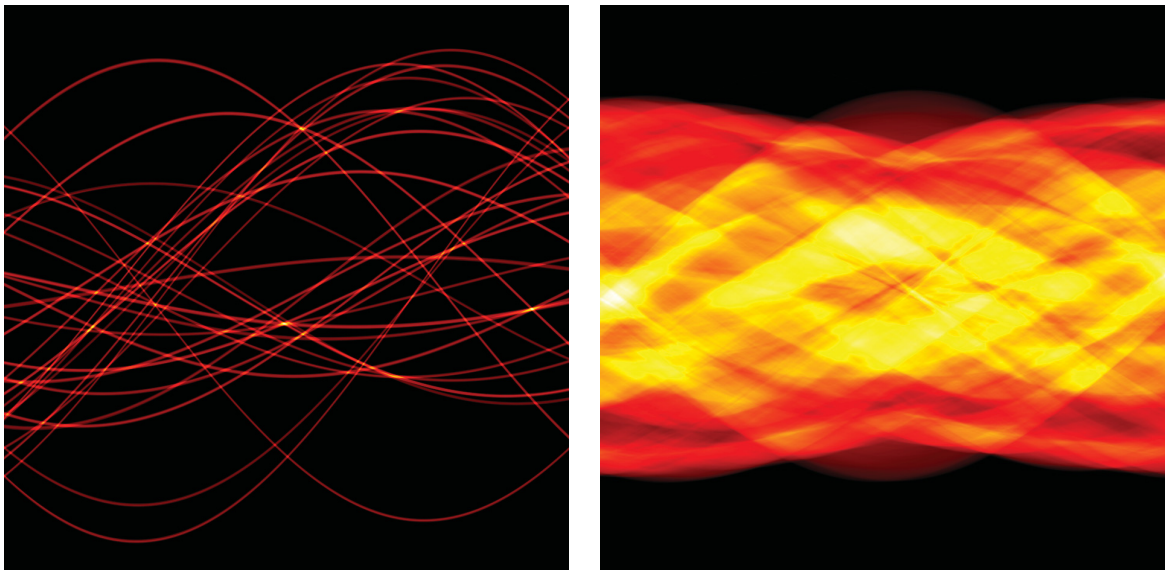
Here, $u^{\dagger,\infty}$ corresponds to the *mass absorption coefficient*, which is assumed to be proportional to the tissue density. If we assume that the X-ray source with intensity I_0 is placed at $t = 0$ we can compute the intensity I_l measured by a detector at $t = T$ as:

$$I_l = I_0 \exp\left(- \int_0^T u^{\dagger,\infty}(t)\, \mathrm{d}t \right) \tag{2.5}$$

This formula, known as *Beer's law*, can be rearranged to

$$f_l^{\dagger,\infty} := \log\left(\frac{I_0}{I_l} \right) = \int_0^T u^{\dagger,\infty}(t)\, \mathrm{d}t = \int_l u^{\dagger,\infty}(t)\, \mathrm{d}l(t). \tag{2.6}$$

For given measurements along a set of lines $\mathcal{L}_*$, the inverse problem is to recover $u^{\dagger,\infty}(x, y)$ from its integrals along these lines. This is a problem of *integral geometry* formulated and treated in the work of Johann Radon in 1917 (Radon 1917, 1986).



(a) (b)

Figure 2.3.: Radon transform $\mathcal{R}[u^{\dagger,\infty}]$ for (a) $u^{\dagger,\infty}$ used in the "Spots" scenario (see Section 2.2.2 and Figure 2.2a) and (b) for Figure 1.6a interpreted as a 2D function.

2.3.1. The Radon Transform

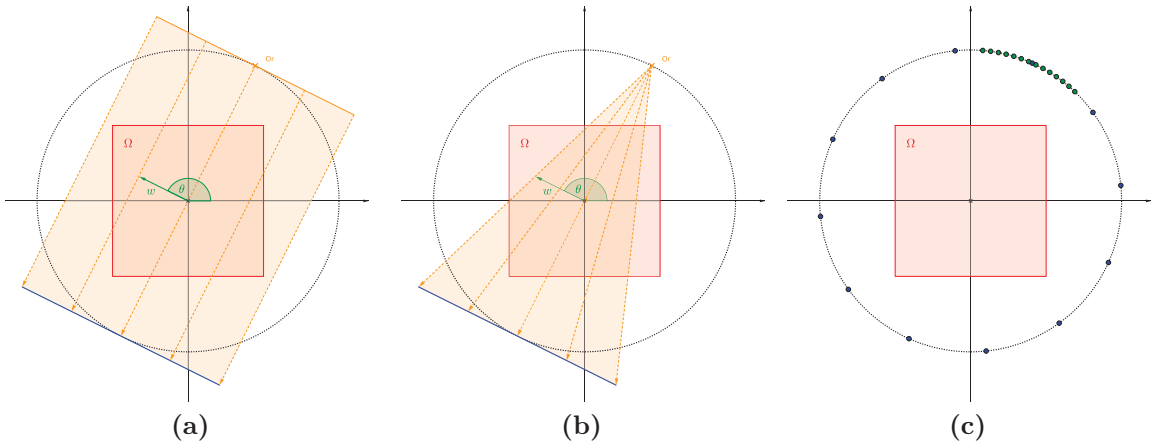
Radon introduced an invertible integral transform for piecewise continuous, compactly supported functions of two variables based on line-integrals like (2.6). Any line $l \subset \mathbb{R}^2$ can be described by the angle θ of its normal vector w and its (signed) distance s to the origin:

$$l(\theta, s) = \{(x(t), y(t)) = (t \sin\theta + s\cos\theta, -t\cos\theta + s\sin\theta) \,|\, t \in \mathbb{R}\}$$

The space $\mathcal{L}$ of all lines in $\mathbb{R}^2$ can now be parameterized by $\theta \in [0, \pi)$ and $s \in \mathbb{R}$. The *Radon transform* $\mathcal{R}[u]$ of a function $u : \mathbb{R}^2 \to \mathbb{R}$ defines a function on $\mathcal{L}$ by:

$$\begin{aligned}\mathcal{R}[u](\theta, s) &= \int_{l(\theta,s)} u(x(t), y(t))\,\mathrm{d}l(t) \\ &= \int_{-\infty}^{\infty} u(t\sin\theta + s\cos\theta, -t\cos\theta + s\sin\theta)\,\mathrm{d}l(t)\end{aligned} \tag{2.7}$$

We will denote the transform for a fixed angle θ by $\mathcal{R}_\theta[u](s) := \mathcal{R}[u](\theta, s)$. Figure 2.3 shows two exemplary Radon transforms. The Radon transform of a Dirac delta distribution is a distribution supported on the graph of a sine wave. Therefore, visualizations of Radon transforms look like compositions of multiple sine waves and are usually called *sinograms*. This connection to trigonometric functions is formalized by the *Fourier slice*



(a) (b) (c)

Figure 2.4.: (a) Parallel beam geometry. (b) Fan beam geometry. (c) Different angle distribution schemes: Sparse (blue dots) and limited (green dots) angle scanning.

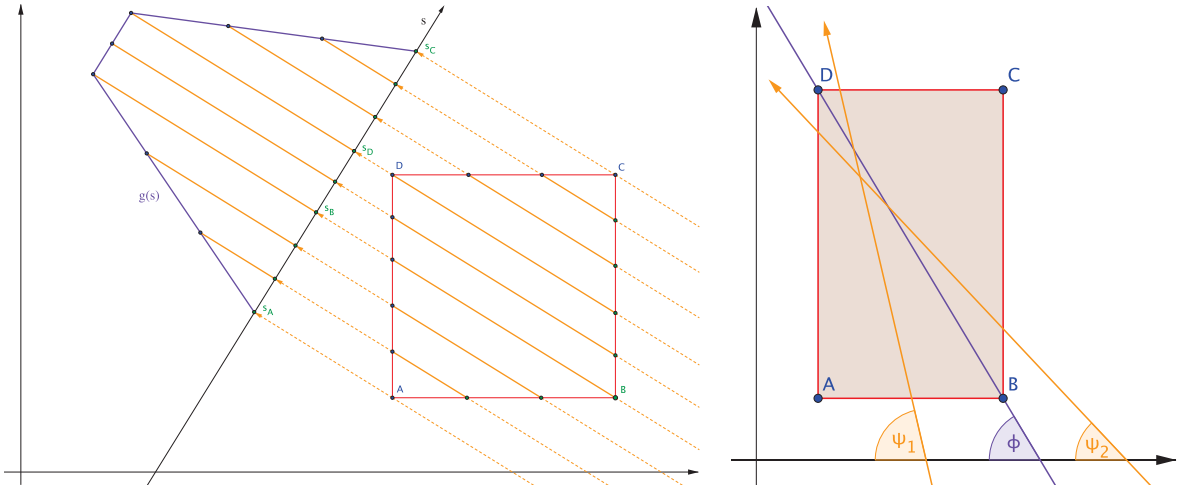
theorem, which states that for piecewise continuous, compactly supported u:

$$\mathcal{F}[\mathcal{R}_\theta[u]](\nu) = \mathcal{F}[u](\nu w) \tag{2.8}$$

This means that the Radon transform along a certain angle fully determines one slice of the 2D Fourier transform of u. Therefore, $\mathcal{R}[u]$ fully determines $\mathcal{F}[u]$ and, thereby, u, and the complete Radon transform is invertible. Based on these relations, direct analytical as well as approximative, numerically stable inversion formulas to analyze CT data can be derived (*filtered back projections*, see NATTERER 1986). These approaches often require a sufficiently dense sampling of the Radon transform. In the next section, we will develop a formulation similar to those of the other scenarios which will also be applicable if only a sparse sampling of the Radon transform is used.

2.3.2. Computational Model for Computed Tomography

We assume that $u^{\dagger,\infty}$ is supported in $\Omega = [0,1]^2$. $\mathcal{A}$ is given as the Radon transform restricted to a subset $\mathcal{L}_* \subset \mathcal{L}$. We will define $\mathcal{L}_*$ by the subset of *measurement angles* $\theta_i \in [0, 2\pi)$, $i = 1, \dots, m_\theta$ and the *beam geometry* used. The latter describes the spatial radiation pattern for a fixed angle. We will consider *parallel-beam* and *fan-beam* geometries. In parallel-beam geometry all beams transverse the target parallel to each other; see Figure 2.4a. This directly corresponds to $\mathcal{R}_{\theta_i}[u](s)$, is easy to implement and intuitive to examine. In the fan beam geometry, a point source and a detector placed opposite to each other are rotated around the origin. This leads to an angular spread of the lines from the source to the detector; see Figure 2.4b. The projection for a given measurement angle θ_i does not directly correspond to $\mathcal{R}_{\theta_i}[u](s)$ but is spread out in angular direction. Furthermore, changing source and detector, i.e., taking measurements



(a) (b)

Figure 2.5.: (a) Geometrical drawing to illustrate the derivation of $g(s)$ for a rectangle. (b) Geometrical drawing to illustrate formula (2.11).

for angles θ_i and $\theta_i + \pi$ does not lead to the same results anymore. This beam geometry corresponds to the configuration used to collect the experimental data used in this thesis.

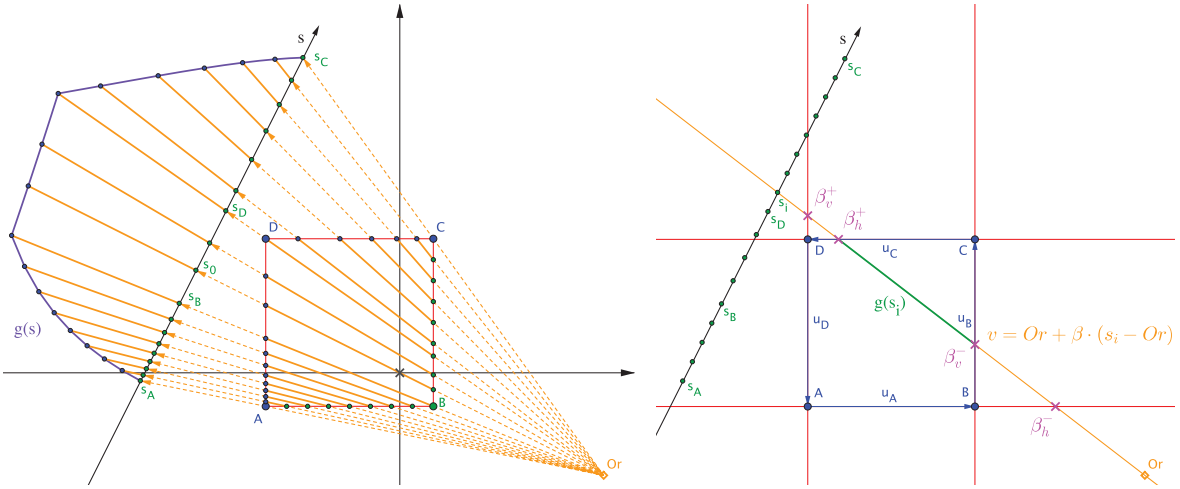
Choosing a dense, equidistant distribution of the measurement angles $\theta_i \in [0, 2\pi)$, $i = 1, \ldots, m_\theta$ leads to the best reconstruction quality. However, there are situations where this is not possible or desirable:

- In *limited angle tomography*, the measurement setup restricts the range of θ to $[\theta_{min}, \theta_{max}]$. This occurs, e.g., in *mammography*, *dental radiology*, *intraoperative* or *rotational aniography* or *electron tomography*. The inversion of limited angle data is *severely ill-posed*. Only specific features of the solution can be recovered and several artifacts may appear. See Frikel (2013) for a recent overview on the theoretical implications of this scanning setup.

- One focus of research in CT is to reduce the radiation dose delivered to the patient (Yu et al. 2009) while another is reconstruct spatio-temporal images (*4D-CT*). For both aims, a reduction of the number of measurement angles is an often discussed option (*sparse angle tomography*). Thereby, the loss of spatial resolution and quality of the images has to be tolerable and outweighed by the potential benefits.

Figure 2.4c illustrates both angle distribution schemes.

The transmitted intensity $\mathcal{A}u^{\dagger,\infty}$ is not measured directly. As in the image deblurring scenarios examined in the previous section, we rather measure its integral over a sensor pixel, which is, again, modeled by the operator P: We assume that we have m_s



(a) (b)

Figure 2.6.: (a) Geometrical drawing to derive $g(s)$ of a rectangle in fan-beam geometry. (b) Geometrical drawing to illustrate the computation of $g(s_i)$: First, the ray (orange line) parameterized as $Or + \beta(s_i - Or)$ is crossed with the horiontal and vertical extensions of the rectangle edges (red lines), resulting in the beta coordinates of the crossings (pink crosses), $\beta_h^-, \beta_h^+, \beta_v^-$ and β_v^+ (if the ray is too parallel to these lines, these values have to be corrected). Then, $g(s_i)$ (visualized by the green line segment) is given by the length of $[\beta_h^-, \beta_h^+] \cap [\beta_v^-, \beta_v^+]$ normalized by the length of $(s_i - Or)$.

measurement pixel of equal size δ_s. In total, this leads to $m = m_s \cdot m_\theta$ measurements. The unknowns u will, again, be discretized using a pixel basis with $n = 2^N \times 2^N$. To compute A for this basis (and for certain wavelets in Section 5) we need to compute $f^\dagger = P\mathcal{A}v$ for v being the indicator function of a rectangle. For this, let A, B, C, D denote the corners of the rectangle, ω its width and ϱ its height. For a fixed measurement angle θ in parallel- or fan-beam geometry, let $g(s)$ denote the transmitted intensity. We first consider parallel beams for which we can deduce $g(s)$ from simple geometric considerations: $g(s)$ will be zero until the beam hits the first corner of the rectangle. After that, it grows linearly as the beam enters at one side of the rectangle and leaves at an adjacent side. When it hits the second corner, it becomes constant as the beam enters at one side and leaves on the opposite side. After hitting the third corner, it starts to decrease to zero again, with the same slope as in the beginning. Figure 2.5a shows a sketch of the situation. As a consequence, $g(s)$ is a simple, piecewise linear function, which we can easily parameterize by the s-values of the beams hitting the corners, s_A, s_B, s_C and s_D, and its value ν in the constant part. The s-values can be computed by solving the beam parametrization

$$A = \begin{pmatrix} x_A \\ y_A \end{pmatrix} = \begin{pmatrix} t_A \sin\theta + s_A \cos\theta \\ -t_A \cos\theta + s_A \sin\theta \end{pmatrix} \tag{2.9}$$

for s_A:

$$s_A = x_A \cos\theta + y_A \sin\theta \tag{2.10}$$

The longest intersection of a beam with the rectangle determines ν: Define ϕ as the short angle of the rectangle's diagonal to the x-axis, and ψ be the corresponding short angle of the beam to the x-axis. If ψ is smaller than ϕ, the beam enters and leaves at the right and left side; otherwise, it enters and leaves at bottom and top (see Figure 2.5b). A simple computation yields:

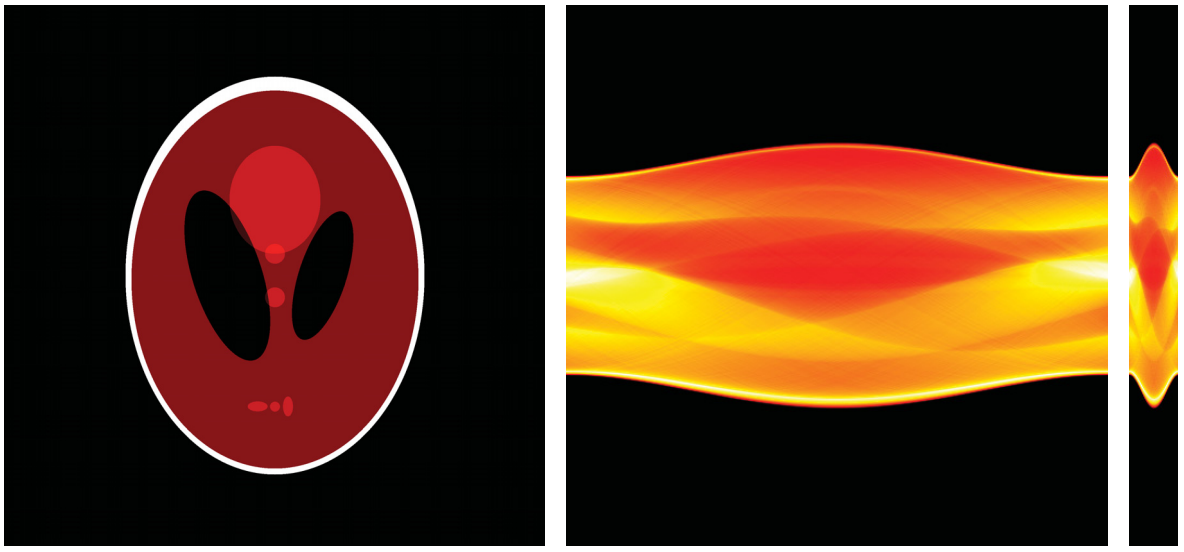
$$\nu = \begin{cases} \dfrac{\omega}{\sin\theta} & \text{if} \quad \psi < \phi \\ \dfrac{\varrho}{\cos\theta} & \text{otherwise} \end{cases}, \qquad \phi = \arctan\frac{\varrho}{\omega}, \quad \psi = \theta - \frac{\pi}{2} \tag{2.11}$$

The integral of the parameterized, piecewise linear function $g(s)$ over the pixels of a given sensor grid can then be computed explicitly.
For the fan-beam geometry, the situation is more complicated and no simple parametrization of $g(s)$ is available; see Figure 2.6a. Therefore, we will approximate $g(s)$ by a piecewise linear function $\bar{g}(s)$, which can then also be integrated over the sensor pixels explicitly. We construct the grid on which $\bar{g}(s)$ is defined starting from $S = \{s_A, s_B, s_C, s_D\}$. Ordering the four elements in S yields three intervals. Each of these intervals is then divided into equidistant sub-intervals such that no sub-interval exceeds a chosen length δ_{fan}. Then, for each s_i of the resulting grid S, we compute the intersection of a ray from the source to the sensor point corresponding to s_i with the rectangle to determine $g(s_i)$. The details of our approach are explained in Figure 2.6b. These basic operations can be implemented in a fast and robust way. The step size δ_{fan} determines the scale on which a linear approximation of $g(s)$ is acceptable. This corresponds to approximating the fan-beams that fall into this sensor interval by parallel beams. For a fixed δ_{fan} and target, the error of this approximation depends on the distance d between source and target. The parallel-beam geometry can be seen as an approximation of the fan-beam geometry in the limit of $d \to \infty$.

2.3.3. Computational Scenarios

Phantom Reconstruction in Parallel-Beam Geometry

The first computational scenario we consider is an artificial data scenario in parallel-beam geometry. The unknown function $u^{\dagger,\infty}$ is a slightly scaled version of the *Shepp-Logan phantom* (Shepp and Logan 1974), a toy model of the human head defined by 10 ellipses; see Figure 2.7a. Figure 2.7b shows the sinogram for a dense angular spacing, while 2.7c shows the sinogram for the sparse angle scenario we will examine: $m_s = 500$,



(a) (b) (c)

Figure 2.7.: "Phantom-CT" scenario. (a) Unknown function $u^{\dagger,\infty}$. (b) Clean measurement $f^\dagger$ for $m_s = m_\theta = 500$ (c) Clean measurement $f^\dagger$ for $m_s = 500$, $m_\theta = 45$.

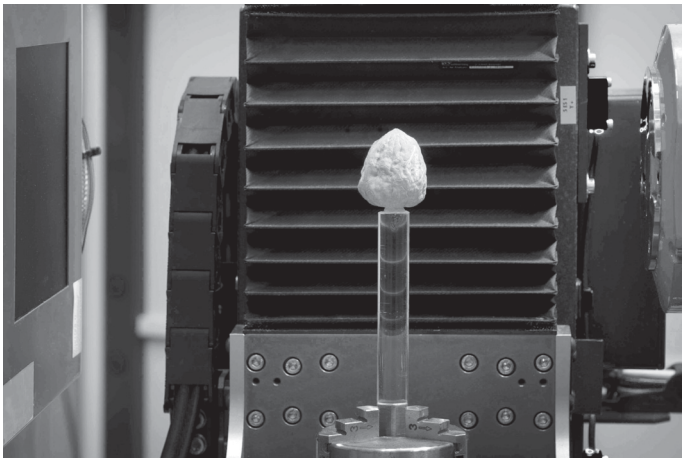
$m_\theta = 45$. We will refer to this scenario as "Phantom-CT".

Walnut Reconstruction in Fan-Beam Geometry

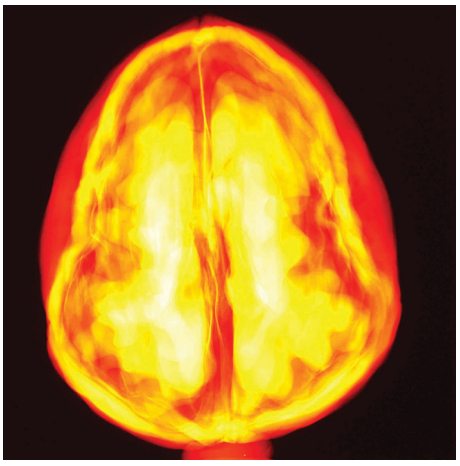
The second scenario is a real data scenario modeled using the fan-beam geometry. The data set was also analyzed in Hämäläinen et al. (2014, 2013). The target is a walnut, chosen for its resemblance to a human brain. The apparent advantage of this setting is that an unlimited number of measurements can be taken without radiation concerns. Furthermore, photographs of cross-sections of other walnuts can easily be taken and examined to generate a-priori information used by the inverse method. Figure 1.6a shows such a photograph.
The data were recorded using a fixed setup of an X-ray source facing a planar detector. The walnut was placed on a rotatable bar in-between, see Figure 2.8a. Figure 2.8b shows one projection of the whole walnut. In this thesis, we will only consider the 2D reconstruction of the central slice of the walnut from the corresponding horizontal line of measurement pixels. The sinogram data is shown in Figure 2.8c. We will refer to this scenario as "Walnut-CT".

2.3.4. Notes and Comments

Buzug (2008) is a general reference for CT, a mathematical one is given by Natterer (1986).



(a)



(b)

(c)

Figure 2.8.: (a) Photograph of the measurement setup; Image courtesy of Samuli Siltanen. (b) A single projection of the walnut. (c) Full sinogram using an angular spacing of 3°. The senor resolution has been reduced by 4, i.e., every 4 subsequent pixels of the 2296 original pixels were added up.

Radiography (see Section 1.1) played a pioneering role for biomedical imaging partly because the RTE can be reduced to a simple form of the RTE for the high-energy X-rays. Optical tomography working with lower-energy radiation has to consider more sophisticated approximations of the RTE, e.g., scattering needs to be accounted for. In return, such imaging techniques can be used for different applications, especially for the examination of soft tissues. See Arridge (1999), Arridge and Schotland (2009) for an introduction and overview.

In most CT scanners, source and sensor array rotate around the patient. For obtaining high sampling rates, a high rotation frequency needs to be realized. As a result, such scanners are heavy and immobile. Using multiple fixed sources and detectors that collect data at a high temporal rate is discussed as a potential alternative to this scanning

(a) (b)

Figure 2.9.: (a) EEG electrode cap. (b) MEG device. Source for both images: Wikimedia Commons.

paradigm, also for realizing 4D-CT. Such a setup naturally leads to a sparse angle tomography problem. See Niemi et al. (2013) for a further discussion.

2.4. EEG/MEG Source Reconstruction

A detailed introduction into EEG/MEG source reconstruction was given in Section 1 in Lucka (2011), including the neurophysiological generators, the mathematical modeling of the forward problem and an overview on the different inverse approaches developed in this field. To integrate EEG/MEG source reconstruction into the concept of this thesis, we will only summarize the most important aspects here. On several other occasions, more detailed explanations of certain sub-topics will be given.

The phenomenon of electromagnetic fields generated by living organisms is called *bioelectromagnetism*. In the neuronal tissue of the brain, bio-chemical activity causes ion current flows that induce electromagnetic fields. Maxwell's equations and the *material equations* provide the accurate physical description of these fields, but similar to CT, only suitable simplifications of these equations lead to tractable computational models. We will consider the following forward equations for the electric potential Φ, caused by a *primary current density* $u^{\dagger,\infty} : \Omega \to \mathbb{R}^3$:

$$\nabla \cdot (\sigma \nabla \Phi) = \nabla \cdot u^{\dagger,\infty} \qquad \text{in } \Omega \tag{2.12}$$

$$\nu \cdot (\sigma \nabla \Phi) = 0 \qquad \text{on } \partial\Omega \text{ (\textit{no-penetration condition})} \tag{2.13}$$

$$\int_{\partial\Omega} \Phi \, \mathrm{d}S = 0 \qquad (\textit{fix ground potential}), \tag{2.14}$$

where $\sigma(r)$ is the *electric conductivity* of the tissues. Given Φ, the corresponding magnetic field $\mathbf{B}$ can be computed by the *Biot-Savart law*:

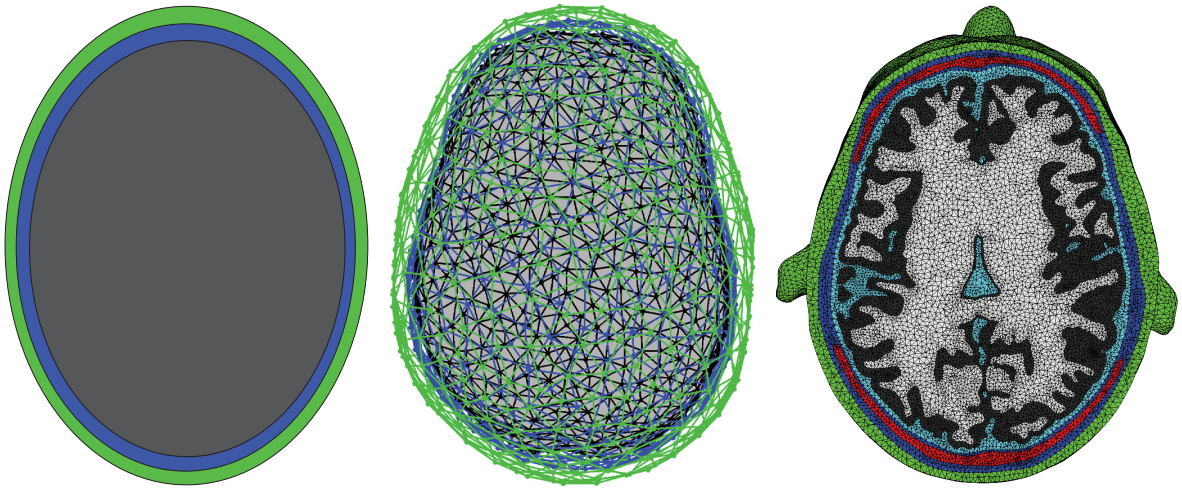
$$\mathbf{B}(r') = \frac{\mu_0}{4\pi} \int_\Omega \left(u^{\dagger,\infty}(r) - \sigma(r) \cdot \nabla\Phi(r)\right) \times \frac{r' - r}{\|r' - r\|_2^3} \, \mathrm{d}r \qquad \text{for } r' \in \mathbb{R}^3 \backslash \bar{\Omega}, \tag{2.15}$$

where μ_0 is the *magnetic constant* and ν the normal of $\partial\Omega$. Following our notation, $\mathcal{A}u^{\dagger,\infty} = f^{\dagger,\infty} = (\Phi|_{\partial\Omega}, \mathbf{B})$. Probing the fields with P to generate $f^\dagger$ describes the concrete measurement setup. We might perform EEG or MEG recordings alone or simultaneously (which we will refer to as *EMEG*). For EEG, we model the electrode measurement of the electric potential (see Figure 2.9a) by a simple point evaluation of Φ at the electrode position. For probing the extremely weak magnetic field caused by neuronal activity, we will use a specific type of magnetometers called *SQUIDs*. SQUIDs measure the *magnetic flux* through a coil by a quantum effect. Here, the magnetic flux is the surface integral of the normal component of $\mathbf{B}$ over the coil area. *Gradiometers* combine two of such coils on top of each other to measure the spatial derivative of the magnetic flux in normal direction of the coils. The elementary, physical, measurement sensors can furthermore be combined (*sensor montage*) to form measurement *channels*. Montages can be designed to reduce certain noise or artifact signals or to enhance the contrast of a certain brain activity. They can already be implemented in an analog way (as the combination of two magnetometers into a gradiometer) or realized after the digitization of the physical measurement channels. We will return to this issue in Section 5.4.5.

2.4.1. Computational Model for EEG/MEG Source Reconstruction

Developing a computational model to simulate (2.12) and (2.15) requires considering three related difficulties.

Head modeling First, the dielectric properties of the different head tissues (called *compartments*) have to be modeled to define $\sigma(r)$. Depending on the desired degree of realism, building such a *volume conductor model* of the head can be a sophisticated task. An early, but still commonly used, approach is to approximate the compartment boundaries by closed surfaces with a simple analytical form such as spheres or ellipsoids (see Figure 2.10a). For every compartment, a constant conductivity is assumed. More sophisticated models replace the parametric surfaces by realistically shaped ones, usually discretely defined by *triangulations* (see Figure 2.10b). A triangulation of the complete



(a) (b) (c)

Figure 2.10.: Different approaches to volume conductor modeling: (a) The geometry is approximated by an ellipse. (b) The geometry is defined by triangulated surfaces. (c) The complete volume is triangulated.

volume instead of the surfaces allows to define $\sigma(r)$ individually for every tetrahedron (see Figure 2.10c). This offers the possibility to incorporate local conductivity models: In general, $\sigma(r)$ describes the mobility of electric charge carriers in a medium. Isotropic media, i.e., with no directional structure, can be described by a scalar value. For highly directed tissues, such as the fibrous white matter, more sophisticated models are required. Equation (2.12) allows for a tensor representation of $\sigma(r)$, which we will refer to as *anisotropic conductivity*.

With increasing flexibility for realistic, anisotropic, individual head modeling, the model construction also needs an increasing amount of precise, individual anatomical imaging information and more sophisticated image analysis methods.

Source Modeling A second difficulty is to find a mathematical model for the primary current density $u^{\dagger,\infty}$. Together with the choice of σ, it determines which numerical methods we can use to solve (2.12). Commonly, the microscopic current flow is modeled by a mathematical current dipole with a suitable dipole moment (see Brazier 1949, deMunck et al. 1988). An ensemble of such microscopic dipoles can be approximated by a single, macroscopic, *equivalent current dipole* $q_{dip}\,\delta(r - r_{dip})$. Most numerical methods developed for EEG/MEG solve (2.12) and (2.15) for such a dipole source term. An arbitrary u can then be discretized by a finite number of unit-strength dipoles:

$$u^{\dagger,\infty} \approx \sum_{i}^{n} u_i^{\dagger}\, q_i \delta(r - r_i), \qquad \|q_i\|_2 = 1, \tag{2.16}$$

where $u^\dagger \in \mathbb{R}^n$ represents the amplitudes of the dipoles. By the linearity of (2.12) and (2.15) the measurements generated by $u^{\dagger,\infty}$ can then be approximated by

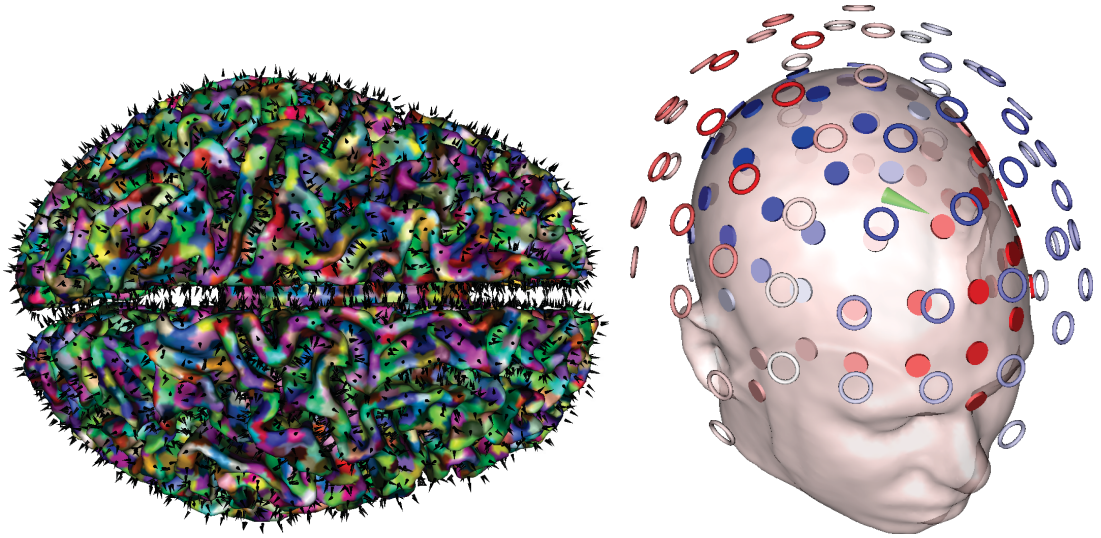
$$f^\dagger = P\mathcal{A}u^{\dagger,\infty} \approx \sum_i^n P\mathcal{A}\, u_i^\dagger\, q_i\delta(r - r_i) = \sum_i^n u_i^\dagger\, P\mathcal{A}q_i\delta(r - r_i) =: Au^\dagger. \tag{2.17}$$

Source reconstruction relying on the *lead-field* matrix A to be used in the inverse problems approaches introduced in Section 1.2 is called *current density reconstruction* (*CDR*). The set of dipoles $\{(q_i, r_i) \,|\, i = 1, \ldots, n\}$ used for the discretization define the *source space*. Physiological as well as computational constraints have to be considered in its construction; see Figure 2.11a for an example.

Numerical Solver Dependent on the type of volume conductor and source model used, different numerical approaches to solve (2.12) and (2.15) are available. For simple conductor geometries, such as multiple concentric spheres, explicit asymptotic formulas for dipole sources can be derived. For volume conductors defined by realistically shaped surfaces, *boundary element* (*BE*) methods were developed. BE methods reformulate (2.12) as an integral equation on the compartment boundaries and compute a discrete solution on the given surface triangulation. *Finite element* (*FE*) methods have to be used if the volume conductor is given as a triangulation of the complete volume. They rely on the weak formulation of (2.12) and approximate its solution by a Galerkin approach using local basis functions. Using a singular source model such as the current dipole (we have $\delta(r) \in H^{-3/2-\varepsilon}(\Omega)\ \forall \varepsilon > 0$) is a potential difficulty for both the FE and the BE approach (Vorwerk et al. 2012). Different solutions have been proposed: The *Venant direct method* (Buchner et al. 1997), the *partial integration direct method* (Schimpf et al. 2002) and the *subtraction approach* (Wolters et al. 2007).

2.4.2. Computational Scenarios

All EMEG studies in this thesis rely on anatomical MRI data and EEG/MEG recordings of a healthy, 25-year-old, male subject. In Section 5.4.1, we will describe how to construct head models with a varying degree of realism based on the MR images. The most realistic head model (which is depicted in Figure 2.10c) will differentiate between ten tissue compartments of which some will be modeled anisotropic. The construction of corresponding source spaces for different purposes will be described in Section 5.4.3. The head models and source spaces will be used in the computational studies with simulated as well as with real EEG/MEG recordings.

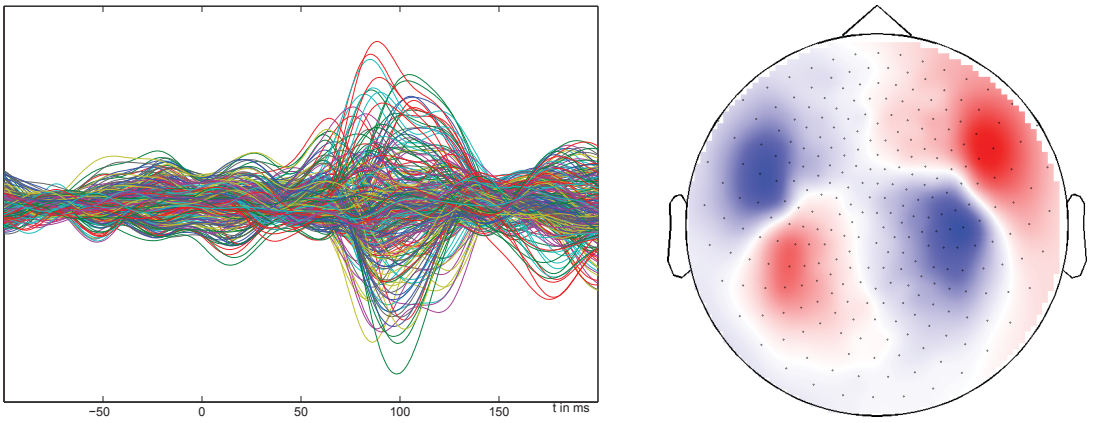


(a) (b)

Figure 2.11.: (a) Illustration of a source space: The dipoles (black cones) are located in the gray matter compartment and oriented along the normal direction of the interface between gray and white matter. This surface is shown with a parceling representing the assignment of its surface triangles to the nearest dipole (cf. Section 5.4.3). (b) Sensor configuration used in the "simEMEG" scenario consisting of 63 electrodes (disks) and 63 magnetometers (rings) colored by the fields generated by a single dipolar source in the brain (green cone).

EEG vs. MEG Scenario

For the first scenario the most realistic head model is used. The artificial sensor configuration employed was designed to examine the differences between EEG and MEG based source reconstruction: From 134 points corresponding to a regular sampling of the skin surface of the head model, $m = 63$ locations corresponding to realistic electrode positions were selected to create the EEG sensors (cf. Figure 8 in Lucka et al. 2012). To create a matching MEG sensor configuration, these positions were shifted by 3 cm away from the surface in normal direction. This corresponds to the average sensor-to-surface distance we encountered in real MEG sensor configurations. At each of the new positions, a magnetometer directed towards the center of the head model is placed. Figure 2.11b shows the sensor configurations and simulated fields for a single dipole source. For the source space locations, a volumetric division of the gray matter based on a 3D-lattice with 6 mm spacing is used. On each of the resulting 1336 locations, 3 dipoles in x, y and z direction are placed, leading to $n = 3 \times 1336$ unknowns to recover. We will use this setup to reconstruct real source configurations $u^{\dagger,\infty}$ consisting of one to three single dipoles located uniformly at random in the gray matter and refer to this scenario as "simEMEG".

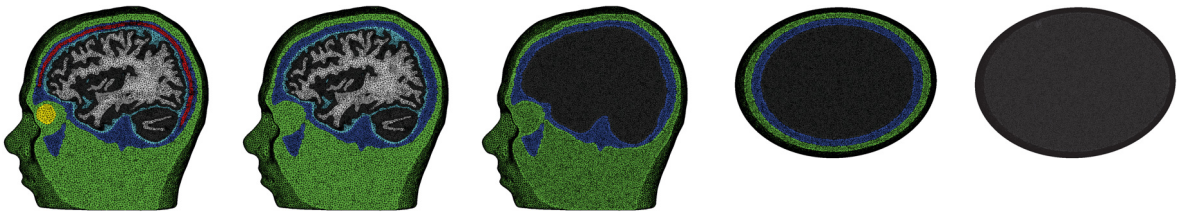


(a) AEF (MEG) butterfly **(b)** AEF (MEG) topography

Figure 2.12.: Auditory evoked fields, averaged over 89 trials. (a) *Butterfly plot*: Each graph corresponds to the trail-averaged temporal evolution of one channel. (b) *Topography plot* at $t = 92$ ms after stimulus onset: The channel positions are mapped to a disk representing the view from above onto the head surface. The measurement data is then interpolated and color-coded to visualize the spatial field distribution.

Evoked Potentials and Fields

Evoked potentials/fields (*EP/EF*) describe the EEG/MEG recordings of the brain's response to a specific external stimulus. An important class of EPs/EFs are *sensory evoked potentials/fields* (*SEP/SEF*) which follow a sensory stimulation. In this thesis, we will investigate *auditory evoked potentials/fields* (*AEP/AEF*) following the presentation of a tone and *somatosensory evoked potentials/fields* (*SSEP/SSEF*) elicited by the stimulation (tactile or electric) of a sensory nerve in the periphery. Single SEP/SEF signals also reflect all other ongoing brain processes aside the stimulus related activity. In addition, they are contaminated by internal and external noise and nuisance sources. To reduce this interference by non-stimulus related signals, an averaging strategy is employed: The same SEP/SEF are recorded multiple times (*trials*) and the signals are averaged. As the SEP/SEF are time-locked to the stimulus onset whereas the interference signals are not (apart from those caused by the stimulus generation), the interference signals will average out in the long run. Figure 2.12 shows AEF signals for a simulation by a 350 Hz tone, averaged over 89 trials. In addition to averaging, several pre-processing techniques were employed; details will be described in Section 5.4.5. We will invert the fields at $t = 92$ ms, using the same head model and source space as in the previous scenario but realistic sensor configurations which are registered to the head model.



(a) (b) (c) (d) (e)

Figure 2.13.: "Head model cascade" scenario. From left to right, the degree of realism is decreasing. A detailed description will be given in Section 5.4.6.

Head Model Cascade

From the most realistic head model, a cascade of less realistic head models is derived (see Figure 2.13). Source spaces of varying spatial density n that align for all head models are constructed and realistic EEG and MEG sensor configurations, similar to those in the real data scenario, are employed. The aim of this setting will be to examine the interplay between realistic head modeling and sparse inverse methods by testing if the corresponding lead-field matrices fulfill certain recovery conditions; cf. Section 1.3. The details of the model generation process and the computations will be given in Sections 5.4.2 and 5.4.6.

2.4.3. Notes and Comments

As mentioned at the beginning of this Section, Lucka (2011) contains a more detailed introduction into several of the topics discussed here, including a lot of references. In addition, some points will be discussed in forthcoming chapters. We will therefore limit the discussion here to some complementary topics.

The theoretical aspects of solving the inverse problem of EEG/MEG are less developed compared to CT. A particular reason is the lack of an analytical framework that provides similar insights into the structure of forward and inverse problem of EEG/MEG as integral geometry does for the Radon transform. In particular, there is no equivalent of the Fourier slice theorem with all its consequences. Dassios and Fokas made important contributions to this topic, see Dassios and Fokas (2013) for a recent overview on their work.

Forward modeling and computation approaches are a constant matter of debate in EEG/MEG source reconstruction. Several aspects are to consider: The *modeling error* describes the error in forward computation caused by replacing the true conductivity $\sigma^{\dagger,\infty}$ by the volume conductor model σ. It has to be compared to the *numerical error* of the computational approach used to solve (2.12) for a fixed σ. Both numerical and modeling error strongly rely on the concrete source configuration $u^{\dagger,\infty}$ used. However,

when using state-of-the-art numerical methods the overall error is usually dominated by the modeling error. See Vorwerk (2011), Vorwerk et al. (2014, 2012) for a recent overview on this topic.

Reducing the modeling error requires a lot of practical efforts like acquiring additional MRI scans or performing sophisticated segmentation procedures. Especially in studies with a large number of subjects, a trade-off between accuracy and modeling effort has to be made.

Modeling the measurement of the surface electrodes by a point evaluation introduces a modeling error that can be avoided if more detailed *complete electrode models* (*CEM*) are used. In Pursiainen, Lucka and Wolters (2012) a detailed comparison between both models is conducted. A realistic modeling of the measurement electrodes is especially important if the surface covered by electrodes is large, like encountered with high density electrode caps on (premature) infants.

Using the mathematical model of a dipole to describe the primary currents is advantageous if used in combination with simple surfaces or BE methods. Therefore, the first forward computation approaches used in practice all relied on this source model (see Vorwerk 2011, for an overview), thereby establishing it as a *de facto* standard in the field. The first FE approaches for EEG/MEG tried to reproduce the results of the former methods and therefore tried to model the source terms as single dipoles as well. As mentioned above, the singular model is a potential difficulty for FE approaches. However, a mathematical model is not an end in itself. It is an approximation with the explicit intention that it leads to a tractable problem. If it fails to do so, there is no real need to stick to it. Instead of developing FE methods that can cope with singular source models, one could therefore also employ less singular source models like $u^{\dagger,\infty} \in H(\mathrm{div}, \Omega, ; \mathbb{R}^3)$ (see, e.g., Calvetti et al. 2009, Pursiainen et al. 2012, Tanzer et al. 2005).

Trail averaging as a recording paradigm has certain limitations. It is based on the assumption that the external stimulus always provokes the same spatio-temporal response of the brain. In this perception, the brain is a passive, linear signal processing machine. In particular, it assumes that there is no dynamic internal state of the brain with which the stimulus interacts. Inter-individual variability of the brain response cannot be explained or studied with this model. In recent years, there is a shift towards studying the brain's internal states (*microstates*) and their functional role (Michel and Murray 2012). This is one motivation for performing *single-trail* analysis of EEG/MEG data. Other areas where such an analysis is of interest include the spontaneous, un-evoked generation of pathological activity such as inter-ictal spikes in epilepsy (see Aydin et al. 2014, for a case study and further references) and *Brain-Computer Interface* (*BCI*) applications (Nicolas-Alonso and Gomez-Gil 2012).

3 The Bayesian Approach to Inverse Problems

In this chapter, we will introduce the basic principles of Bayesian inference applied to inverse problems and imaging. First, different stochastic noise models for (1.2) will be discussed in Section 3.1. Thereby, f will become a random variable governed by the *likelihood distribution* $p_{like}(f|u)$, and the inverse problem a problem of statistical inference (cf. Section 1.2). Due to the ill-posedness of inverse problems, standard statistical inference based on $p_{like}(f|u)$ alone is not suited to obtain satisfactory results. Bayesian inference strategies extend the standard framework to cope with these problems: The main idea is to rethink the concept of probability in order to incorporate a-priori information on the solution. The core assumption is that probability and information are dual to each other. Therefore, all variables are naturally modeled as random variables, not only f but also $u^{\dagger,\infty}$ and $u^{\dagger}$ (and potentially also $\mathcal{A}$ and A). However, this randomness introduced should not be confused with real physical properties of the objects in question. It rather reflects our lack of information about them. Encoding the available a-priori information into a probability distribution $p_{prior}(u)$ (the *a-priori* or *prior probability distribution*) is called *Bayesian modeling* and will be discussed in Sections 3.2 and 3.3. In an abstract sense, solving the inverse problem now amounts to combining all sources of information about $u^{\dagger,\infty}$: The information before the measurement (encoded in the prior) with the information gained by performing the measurement (encoded in the likelihood). This information can also be represented by a probability distribution $p_{post}(u|f)$, called

a-posteriori or *posterior probability distribution*, which can be computed by *Bayes' rule*:

$$p_{post}(u|f) = \frac{p_{like}(f|u)\; p_{prior}(u)}{p(f)} \tag{3.1}$$

Bayesian estimation is the process of extracting the information of interest from the posterior and will be discussed in Section 3.4. Two particular examples thereof are the popular point estimators for $u^\dagger$, the *maximum a-posteriori estimate* (*MAP*),

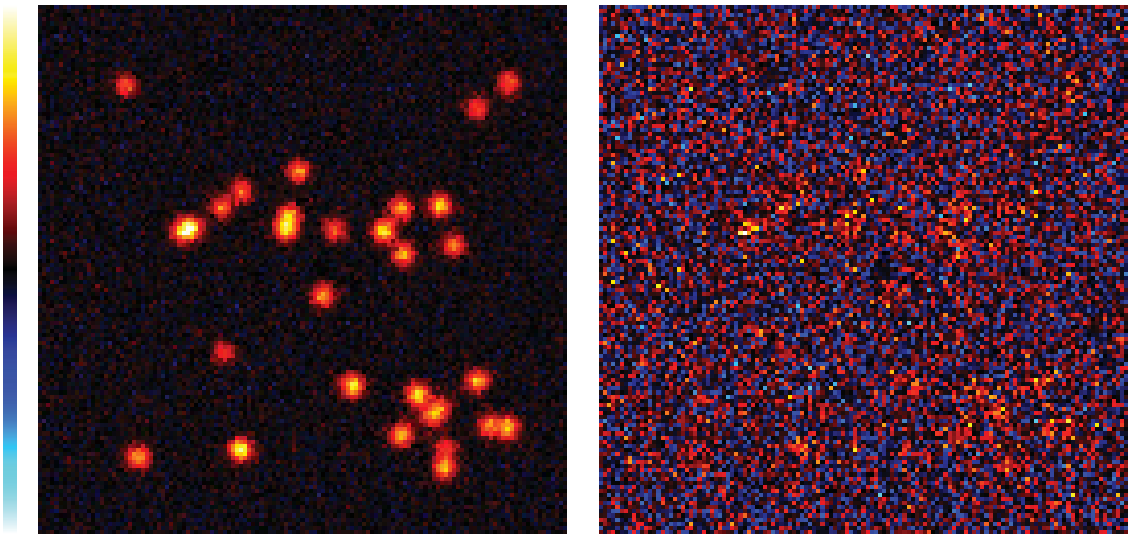
$$\hat{u}_{\mathrm{MAP}} := \operatorname*{argmax}_{u\in\mathbb{R}^n} \{\; p_{post}(u|f)\} \,, \tag{3.2}$$

and the *conditional mean estimate* (*CM*),

$$\hat{u}_{\mathrm{CM}} := \mathbb{E}\,[u|f] = \int u\; p_{post}(u|f)\,\mathrm{d}u. \tag{3.3}$$

The general analysis of such estimators is the topic of *Bayesian decision theory*, which will be introduced as well.
Sparsity as introduced in Section 1.3, is a specific type of a-priori information. We will encounter several ways to encode it into prior distributions. For some sparse priors, $\hat{u}_{\mathrm{MAP}}$ will allow for a more detailed examination by concepts from compressed sensing theory (cf. Section 1.3). This will be discussed in the last section of this chapter. The next chapter, *Bayesian Computation* will then discuss the practical aspects of Bayesian inference, for instance, how to compute $\hat{u}_{\mathrm{MAP}}$ and $\hat{u}_{\mathrm{CM}}$.

Notation and Remark: The augmentation of the classical, deterministic inverse problems setting into a fully stochastic one requires some technical terms and concepts. However, the aim of this chapter is to provide an intuitive, gentle introduction rather than a solid mathematical one in the sense of probability theory. For this sake, some terms will be used somewhat loosely. We will speak of distributions and densities instead of probability distributions and probability densities in the following and will only differentiate between the random variable X and its concrete realization $X = x$ and between the terms probability, probability distributions and probability density where it is necessary. The multivariate normal distribution will simply be called "Gaussian distribution", and random variables distributed according to this distribution will be called "Gaussian random variables". Probability densities will often only be described up to their normalization factor, e.g., as $p(t) \propto \exp(-t^2)$. This lack of mathematical precision can be justified in our setting: All random variables defined in this chapter are finite dimensional random variables on $\mathbb{R}^n$ and either have a probability distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$ or are a



(a) (b) $\sigma = 0.05 \cdot \left\|f^\dagger\right\|_\infty$ (c) $\sigma = 0.5 \cdot \left\|f^\dagger\right\|_\infty$

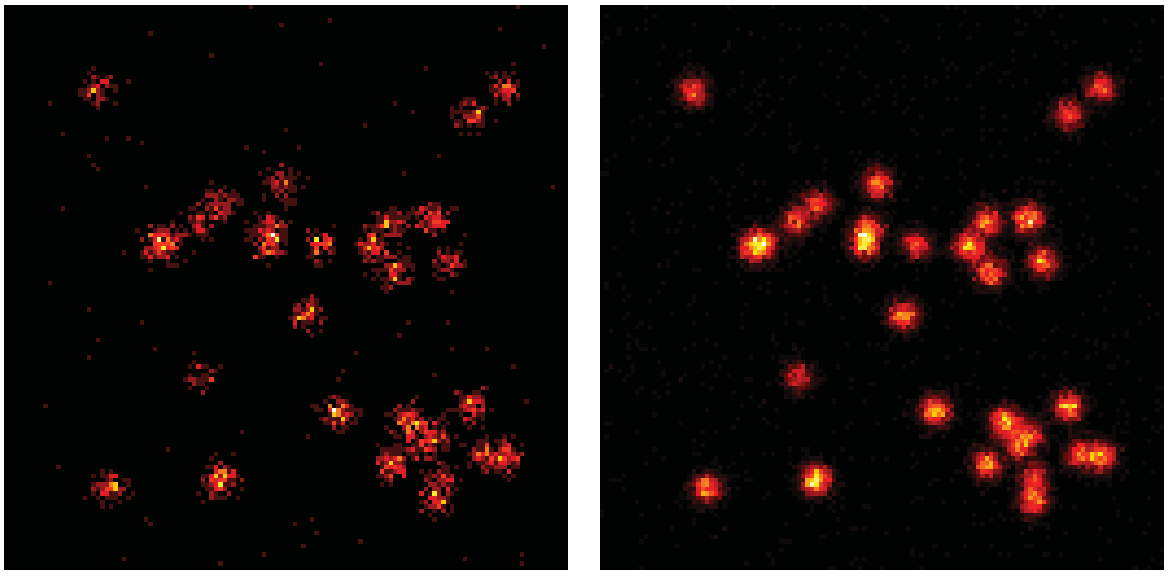
Figure 3.1.: Noisy measurement data f for $f^\dagger$ from Figure 2.2b and additive i.i.d. Gaussian noise model.

Dirac measure. All these measures are Radon measures. Since a regular version of the conditional probability densities exists (Ambrosio et al. 2008, Klenke 2008), problems potentially arising with conditional probability densities are not relevant.

3.1. Stochastic Noise Modeling

In (1.2) or (1.3), the noise variable ε summarizes all features of the inverse problem that are poorly known but are, unlike $u^{\dagger,\infty}$ or $u^\dagger$, not of central interest. As such, it can describe:

- Errors from the use of simplified forward models. For instance, our models for CT (2.4) and EMEG (2.12) were derived from more complicated models by making a couple of simplifications.
- Discretization errors of the forward models.
- Poorly known parameters of the forward models. In EMEG, even when assuming that (2.12) in combination with a volume conductor model of σ can be considered a sufficiently accurate model, the difficulty of determining the bulk conductivities of the tissue compartments remains a source of uncertainty.
- Noise generated by the measurement device. This can comprise sensor-specific noise, amplification noise, electric circuit noise, or digitization errors.
- Background activity $v \in \mathcal{U}$ distinct from $u^{\dagger,\infty}$. In EMEG, this amounts to residual



(a) Short counting time **(b)** Long counting time

Figure 3.2.: Noisy measurement data f for $f^\dagger$ from Figure 2.2b and Poisson noise model.

brain activity that is not related to the recording paradigm (cf. Section 2.4.2).

- Recorded signals from other sources. In CT, this can amount to photons that arrive at the sensor but are not in any way related to the X-rays used for scanning. In EMEG, a large number of such signals are known. Examples include bioelectromagnetic signals originating from other processes in the body, such as the heart-activity, or from external electromagnetic fields, for instance those originating from the power supply of the measurement site.

These features can be deterministic or inherently random, static or dynamic in the course of the measurement. Several techniques can be employed to assess and mitigate the contribution of the different factors. We will discuss some of them when examining the real data scenarios in Chapter 5. For now, we assume that this has been done and that ε contains the residual contributions that we have to take into account. The noise function $Noi(f^\dagger, \varepsilon)$ models the way in which ε interacts with the clean data $f^\dagger$. However, apart from this, ε may also directly depend on $u^\dagger$ in some way (usually on $A(u^\dagger)$). Noise function and the distribution of ε, $p_{noise}(\varepsilon)$ determine the *stochastic noise model*. This model determines the likelihood distribution, which describes the probability that a known u leads to an observed f(cf. Section 1.2). The most simple model is the *independent additive noise model*:

$$f = Noi(A(u), \varepsilon) = A(u) + \varepsilon, \tag{3.4}$$

where ε is mutually independent from u. The corresponding likelihood is given as

$$p_{like}(f|u) = p_{noise}\left(f - A(u)\right) \tag{3.5}$$

A common example is given by assuming that ε follows a multivariate normal distribution with zero mean and a covariance matrix Σ_ε^{-1}:

$$p_{like}(f|u) = (2\pi)^{-\frac{m}{2}}\ |\Sigma_\varepsilon|^{-\frac{1}{2}}\ \exp\left(-\frac{1}{2}\|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}}\right), \tag{3.6}$$

where $\|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}} := (f - A(u))^T \Sigma_\varepsilon^{-1}(f - A(u))$. The special case $\Sigma_\varepsilon^{-1} \propto I_m$ is called *i.i.d.* (*independent and identically distributed*) noise model. Due to its *alpha-stability* and the *central limit theorem*, assuming a Gaussian distribution is often a reasonable model for a macroscopic quantity, which is actually the sum of an ensemble of microscopic random variables with bounded variability and spatially decaying correlations. Especially in a real data scenario this seems to be a suitable noise approximation: The multitude of potential noise contributions described above suggests that deriving an explicit model based on physical considerations may be hopeless. In addition, using such a complex noise model will limit the number of computational techniques we can use for the practical inversion. A Gaussian model can already be estimated from first and second order statistics. This can, of course, also be seen as a disadvantage: A Gaussian model cannot account for anything but these statistics. In Sections 5.4.5 and 5.3.2, we perform Gaussian modeling for the real data scenarios we consider and examine the sensitivity of inverse methods to noise modeling errors.

In many imaging applications, the measurements consist of counting discrete events. These events have a certain probability to occur that is related to $f^\dagger$. In principle, each event can be modeled by a *Bernoulli variable* and their sum, which is the measurement, by a *binomial distribution*. However, the discrete, combinatorial character of the binomial distribution is not easy to handle and continuous approximations would be advantageous. For sensors that count a large number of events during measurement time ($f_i^\dagger$ is relatively large) the *de Moivre-Laplace theorem* (Klenke 2008) assures that after re-scaling of the event counts, using a Gaussian model is a good limiting approximation. However, for sensors where this is not true, the binomial distribution can be approximated with the limiting distribution in the case of rare events, the *Poisson distribution*. One possible Poisson noise model assumes that $f_i \sim \mathrm{Pois}(A(u^\dagger) + \eta)$, where $\eta \geqslant 0$ models the counts of events that are not related to $u^\dagger$, for instance, background

radiation. The likelihood is then given as

$$p_{like}(f|u) = \prod_{i=1}^{m} \frac{(A(u)+\eta)_i^{f_i}}{f_i!} \exp\left(-(A(u)+\eta)_i\right) \overset{u}{\propto} \exp\left(-|A(u)+\eta|_1 + f^T \log(A(u)+\eta)\right) \quad (3.7)$$

In contrast to the Gaussian limit, signal and noise are *not* independent in this noise model. Figures 3.1 and 3.2 illustrate the impact of low and high noise levels in both noise models.

In both artificial image deblurring scenarios, i.i.d. Gaussian noise will be employed: $\Sigma_\varepsilon = \sigma^2 I_m$. For the "Boxcar" scenario (cf. Section 2.2.1), $\sigma = 0.001$ will be used, corresponding to a *relative noise level* of $\sigma / \left\|f^\dagger\right\|_\infty = 0.032$. For the "Spots" (cf. Section 2.2.2) we will fix the relative noise level to 0.1 leading to $\sigma = 0.1\left\|f^\dagger\right\|_\infty$. For the "Phantom-CT" scenario, we will fix the relative noise level to 0.01. The reason for choosing a relative noise level in these two scenarios is that m, the dimension of the measurement space, will vary.

In the statistical approach, the ill-posedness of the inverse problem manifests in the properties of the likelihood. The ill-conditioning leads to a wide spread of the likelihood: If two very distinct u and v lead to very similar $A(u)$, $A(v)$ then $p_{like}(u|f)$ and $p_{like}(v|f)$ will also be very similar. As a result, the likelihood cannot be concentrated on particular regions of $\mathbb{R}^n$ but is rather flat and uninformative. If the problem is even underdetermined, the likelihood cannot even be normalized. Standard statistical inference based only on $p_{like}(f|u)$ is therefore doomed to fail: Computing the *maximum likelihood estimate* (*ML*) $\hat{u}_{\mathrm{ML}}$ would correspond to maximizing $p_{like}(f|u)$. For the Gaussian noise model (3.6), this would lead to

$$\hat{u}_{\mathrm{ML}} = \operatorname*{argmax}_{u\in\mathbb{R}^n} \left\{ \exp\left(-\frac{1}{2}\|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}} \right) \right\} = \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ \|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}} \right\} \quad (3.8)$$

This means that every (weighted) least-squares solution to $f = A(u)$ is an ML estimator. This is not a major improvement: The ML estimator inherits all the ill-posedness of the inverse problem, in particular its instability. Therefore, statistical approach to inverse problems examines the properties of alternative estimators. For instance, *penalized maximum likelihood estimation* defines estimators using variational regularization schemes such as (1.5). As discussed in the introduction to this chapter, the Bayesian approach tries to compensate the deficits of the likelihood by introducing the prior.

Notes and Comments

References to statistical inference for inverse problems were given in the first Chapter. We chose to define the noise model for the finite dimensional projections $f^\dagger = P\mathcal{A}(u^{\dagger,\infty})$ only. Dealing with infinite dimensional noise models (or the limit of finite dimensional models) turns out to be quite involved. Section 1.4. in KEKKONEN et al. (2014) gives a recent literature overview on this topic. Most of these works address the mathematical difficulties of infinite dimensional models. However, the physical validity of such models is often problematic. Consider the photon-count based measurement of CCD pixels: For a fixed measurement time, a fixed number of photons arrives at the sensor array. The limit $m \to \infty$ means that the size of the pixels goes to zero. Thereby, the average number of photons arriving in a pixel also goes to zero. Neither Gauss nor Poisson distribution are valid approximations of the binomial distribution in this limit. New mathematical models are required to accurately model such measurement situations (see HELIN et al. 2010, for an example of such an approach). Practically, the limit of small pixel sizes in CCD devices also amplifies a number of internal noise contributions and changes the overall signal-to-noise ratio.

3.2. Bayesian Modeling

3.2.1. General Concepts

The central step in Bayesian inference is the construction of priors, called Bayesian modeling. In this step, we encode all a-priori information on u. The main difficulty is that a-priori information is often of *qualitative* nature, whereas a prior represents *quantitative* information. For instance, consider $u^{\dagger,\infty}$ in the point source reconstruction scenario (Figure 2.2a). Our qualitative a-priori information could be that the solution consists of a few, small, circular bright objects. To turn this into quantitative information, we have to specify what we mean by "few", "small" and "bright" and find a suitable probability density that would actually produce images similar to $u^{\dagger,\infty}$. In general, many kinds of a-priori information can be available:

- A certain representation that captures the most distinct features of u; maybe a basis, frame or dictionary (cf. Section 1.3). Commonly, geometrical features are chosen, e.g., spatial smoothness. In spatio-temporal scenarios, this can also include known features of the temporal evolution.

- Information on the statistics of u with respect to the above representation; possibly represented by parameters describing the distribution. Examples include *moments*, *location*, *shape* or *scale* parameter, *decay characteristic* of the tails, *multi-modality*

(in the sense of probability distributions).

- "Natural" constraints. Especially in biomedical imaging, the objects to recover are restricted in some way. For instance, the neuronal activity in EMEG cannot lead to arbitrarily high currents. Related to this is the a-priori definition of *regions of interest* (*ROIs*) to restrict Ω.
- Multimodal integration (cf. Section 1.1): Information from another imaging modality is incorporated to enhance the recovery of u. Examples include the utilization of fMRI activations to enhance EMEG source localization. This *asymmetric data fusion* operating on the level of the prior is not to be confused with *symmetric data fusion* or *joint reconstruction*, which operates on the level of the likelihood, e.g., combined EEG/MEG source reconstruction.
- Empirical information extracted from the solutions of previously inverted data. Preferably, these solutions should come from similar scenarios with a better data situation. One example is to construct a prior for limited-angle CT reconstructions from solutions of full-angle CT scans of the same body part.

In this section, we will discuss how to account for the first three points. Apart from encoding all available information the prior has to counteract the ill-posedness of the inverse problem. For this, it has to be sufficiently *informative* or *tight*, especially in regions where the likelihood is flat and uninformative. In addition, it should lead to a posterior tractable in practical computations.

3.2.2. Incorporating Hard Constraints

The easiest part is to account for *hard constraints*, such as restricting the support of u to a ROI or to restrict $u(x)$ to a physical or physiological plausible interval, for instance, to enforce non-negativity. A prior model $\tilde{p}_{prior}(u)$ defined on $\mathbb{R}^n$ can be restricted to the *feasible set* $\mathcal{C} \subset \mathbb{R}^n$ of all u admissible to the constraints by amending it with an indicator function on $\mathcal{C}$:

$$p_{prior}(u) \propto \tilde{p}_{prior}(u) \cdot \mathbb{1}_{\mathcal{C}}(u) = \begin{cases} \tilde{p}_{prior}(u) & \text{if} \quad u \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

3.2.3. Gibbs Priors

Originating from statistical physics, *Gibbs distributions* often provide an accurate description of the statistics of high-dimensional quantities u:

$$p_{prior}(u) \propto \exp\left(-\lambda \mathcal{J}(u)\right) \tag{3.10}$$

Here, $\mathcal{J}(u)$ is a functional measuring an *energy* of the configurational state of u. In the context of inverse problems, especially in imaging, "energy" is to be understood in an abstract sense, often rather related to *entropy* in the sense of information theory. In the special case of $\mathcal{J}(u)$ being *convex*, the prior is called *log-concave*.
Using a Gaussian likelihood (3.6) with a Gibbs prior, the posterior is given by:

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}} - \lambda\mathcal{J}(u)\right) \tag{3.11}$$

Computing $\hat{u}_{\text{MAP}}$ for this posterior (cf. (3.2)) can be done by minimizing the negative logarithm of the posterior:

$$\hat{u}_{\text{MAP}} = \operatorname*{argmin}_{u\in\mathbb{R}^n}\left\{\frac{1}{2}\|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}} + \lambda\mathcal{J}(u)\right\} \tag{3.12}$$

Comparing this expression with (1.5) and (1.8) reveals the close connection that the MAP estimate establishes between Bayesian inference and variational regularization.

3.2.4. Gibbs Priors Based on ℓ_p^q-Norms

A typical construction of Gibbs energies for Bayesian inversion relies on ℓ_p vector norms. Let's assume that the characteristic features of u can be extracted by a mapping $D^T : \mathbb{R}^n \mapsto \mathcal{Y}$ and an a-priori estimate of the mean of u, μ_u is known. Then, a canonical energy construction scheme would be given by

$$\mathcal{J}(u) = dist\left(D^T(u), D^T(\mu_u)\right)^q, \tag{3.13}$$

where $dist : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ is a quasi-metric which grows sub-exponentially fast. In this thesis, we will only consider the case where D^T is a linear mapping to $\mathbb{R}^h$ and $\mu_u = 0$. On $\mathbb{R}^h$, defining $dist(D^T u, 0)$ by vector norms is then a convenient choice which leads to

$$p_{prior}(u) \propto \exp\left(-\lambda\left\|D^T u\right\|_p^q\right) = \exp\left(-\lambda\left(\sum_i^h |D_i^T u|^p\right)^{q/p}\right), \tag{3.14}$$

where $D_i \in \mathbb{R}^h$ is, again, the i-th column of D. We will call this construction ℓ_p^q-*prior*, the special case of $q = p$ simply ℓ_p *prior*. Besides D, which we assume to be normalized in some way, the three scalar parameters λ, p and q control different aspects of the multivariate distribution: λ controls the *scale* of $p_{prior}(u)$, while p and q determine the *shape* of it. p controls the geometry of the level-sets of $p_{prior}(u)$ in $\mathbb{R}^n$ while q determines the radial profile of it, i.e., the 1D distribution conditioned along a certain direction $v \in \mathbb{R}^n$. Figure 3.3 illustrates different choices of λ, p and q. Note that (3.14) also

defines a prior in the case of $0 < p < 1$, although $\|\cdot\|_p$ does not define a proper norm anymore. In this case, the level-sets are not convex anymore (and the prior is, therefore, not log-concave). In the case of $0 < q < 1$, the prior is not log-concave either. Both cases lead to practical challenges for many computational techniques.
Often, the components of u have certain sub-structures that we want to respect when formulating a prior. In EMEG for example, we want to reconstruct a vector field based on a source space $\{q_i, r_i\}$ (cf. Section 2.4.1). The source space can be chosen in such a way that the first three components of u correspond to three orthogonal dipoles $\{q_1^x, q_1^y, q_1^z\}$ placed at the same location. Often, one does not have any prior knowledge on how to choose this local coordinate system and want our prior to be invariant with respect to rotations of it. In addition, we may have a-priori knowledge about the spatial distribution of amplitude of the currents, but not about their direction. Using (3.14) with $p \neq 2$ would not fulfill these requirements. We first have to group all components of u that belong to one of the N different source locations, compute the amplitude of the resulting current and formulate our prior information in terms of the amplitudes. If we just want to impose a simple ℓ_p^q prior with $D = I_N$ on the amplitudes, we can just re-arrange the components of u to a matrix $U \in \mathbb{R}^{N\times 3}$ and defining the prior by the matrix norm

$$\|C\|_{r,p} := \left(\sum_i^N \left(\sum_j^l |C_{ij}|^r \right)^{p/r} \right)^{1/p}, \qquad \text{for} \quad C \in \mathbb{R}^{N\times l} \tag{3.15}$$

as

$$p_{prior}(u) \propto \exp\left(-\lambda \|U\|_{2,p}^q\right) = \exp\left(-\lambda \left(\sum_i^N \left(u_{i,x}^2 + u_{i,y}^2 + u_{i,z}^2\right)^{p/2} \right)^{q/p} \right) \tag{3.16}$$

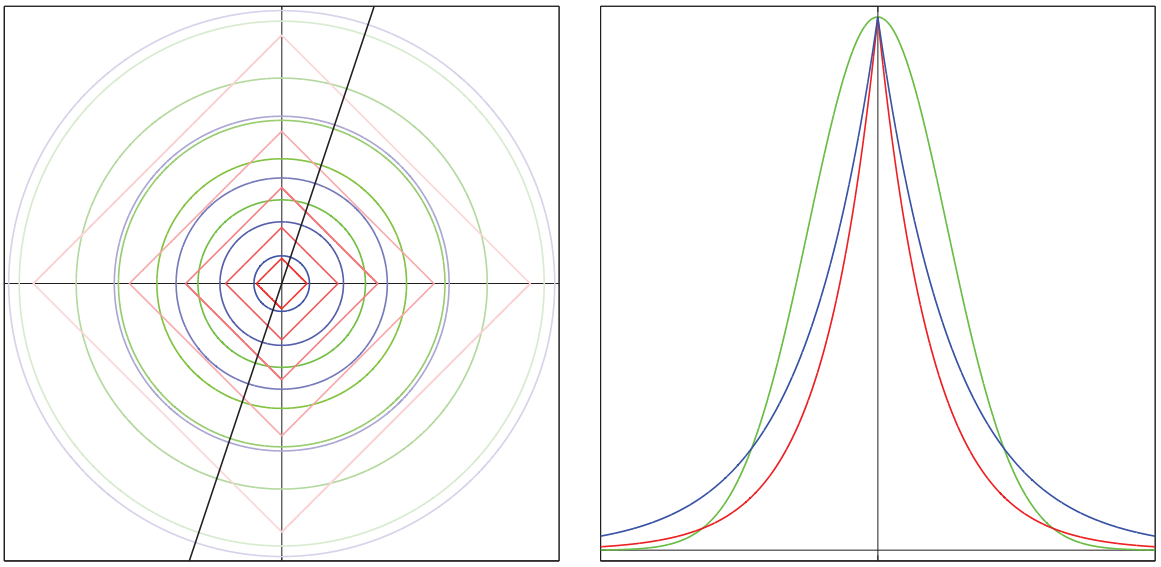
We will call such constructions ℓ_p^q-*block priors* as one formally has to order the components of u into blocks.

The Gaussian Case

A special case is given by the ℓ_2 prior which corresponds to a Gaussian distribution

$$\exp\left(-\lambda \left\| D^T u \right\|_2^2\right) = \exp\left(-\frac{1}{2} u^T (2\lambda D D^T) u\right) = \exp\left(-\frac{1}{2} u^T C^{-1} u\right), \tag{3.17}$$

with covariance matrix $C = (2\lambda D D^T)^{-1}$. By the same reasoning as for the noise modeling in Section 3.1, Gaussian distributions can be considered as the most fundamental priors in Bayesian modeling. They can be used to model various different kinds of



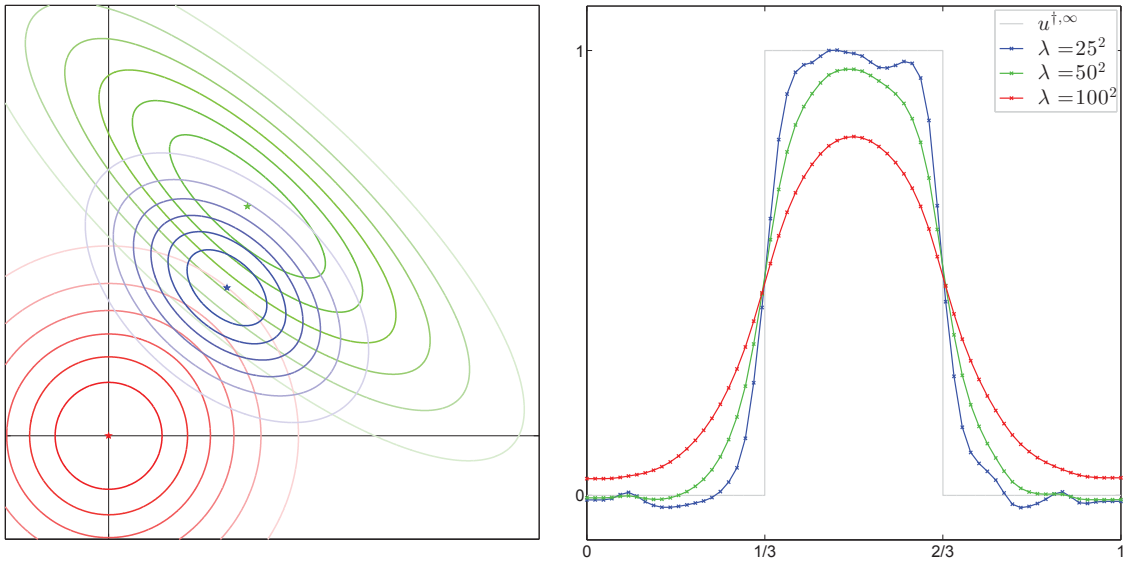
(a) (b)

Figure 3.3.: (a) Level sets of different (unnormalized) ℓ_p^q priors for the values $\{1/6, 2/6, \ldots, 1\}$. Green lines: $p = q = 2$, $\lambda = 2$. Red lines: $p = q = 1$, $\lambda = 2$. Blue lines: $p = 2$, $q = 1$, $\lambda = 1.82$. (b) Radial profiles thereof along the black line in (a).

a-priori information. Section 3.4. in Kaipio and Somersalo (2005) contains a detailed introduction into Gaussian modeling and an illustrative collection of examples. Besides these capacities, using Gaussians for both likelihood and prior distribution facilitates the computational inference considerably: The resulting posterior is also a Gaussian (see Figure 3.4a). Its mean and covariance can be computed explicitly. Furthermore, $\hat{u}_{\text{MAP}}$ and $\hat{u}_{\text{CM}}$ coincide. Despite these advantages, Bayesian inference methods with Gaussian priors also suffer from severe drawbacks. For instance, consider recovering the Boxcar function (cf. Figure 2.1) with a Gaussian prior with D^T being the forward difference operator:

$$D_i = e_{i+1} - e_i, \qquad \Longrightarrow D_i^T u = u_{i+1} - u_i, \qquad i = 1, \ldots, n-1 \tag{3.18}$$

D is a discretization of the first spatial derivative with Neumann boundary conditions. This choice puts a prior on the increments $\xi_i := u_{i+1} - u_i$ (*increment prior*) and aims at recovering functions that can rather be characterized by their jumps than by their values, which seems to be appropriate in this scenario. Figure 3.4b shows $\hat{u}_{\text{MAP}}(\lambda)$ for three values of λ. The estimates are either too noisy or too smooth. The original function $u^{\dagger,\infty}$ has a sparse jump-set: $D^T u$ is sparse for any given n. In general, using Gaussian priors, sparse solutions cannot be recovered. In addition, Bayesian inference with Gaussian priors may suffer from other drawbacks, such as systematic errors. In



(a) (b)

Figure 3.4.: (a) Illustration of Bayesian inference with a Gaussian prior: Level sets of likelihood (green), prior (red) and resulting posterior (blue). The markers indicate the corresponding maxima. (b) MAP estimates for the "Boxcar" scenario ($n = 63$) using a Gaussian prior with different λ and D being the forward difference operator.

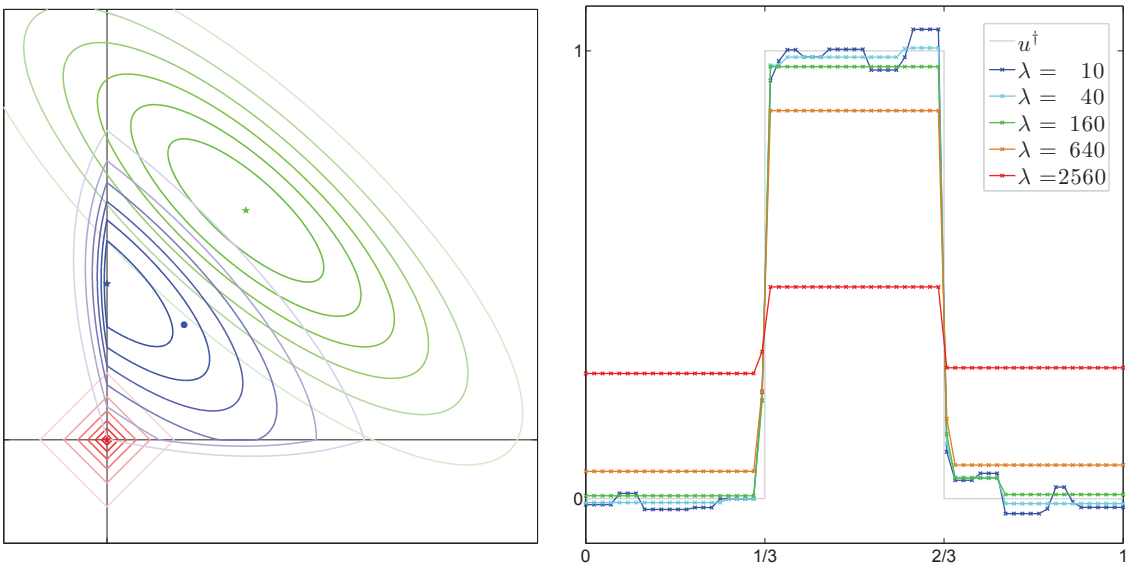
Lucka (2011), the phenomenon of *depth-bias* in EEG/MEG source reconstruction was examined: Using ℓ_2 priors, deep-lying sources are not reconstructed in the correct depth, but always on the surface of the source space.
For these reasons, a lot of recent research, including the work for this thesis, was devoted to examine non-Gaussian prior models.

ℓ_1 priors

If we compare the MAP estimate for ℓ_p^q priors,

$$\hat{u}_{\mathrm{MAP}} = \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2} \|f - A(u)\|^2_{\Sigma_\varepsilon^{-1}} + \lambda \left\| D^T u \right\|_p^q \right\}, \tag{3.19}$$

with (1.8), and recall that solutions to (1.8) were sparse, we recognize that using an ℓ_1 prior leads to a sparse MAP estimate. Comparing Figure 3.5a to 3.4a, one can observe that due to the shape of the ℓ_1 prior, the MAP estimate (blue star) lies on the vertical coordinate axis, which means that the component in horizontal direction is zero. Comparing Figure 3.5b with Figure 3.4b, we see that using an ℓ_1 increment prior we can obtain reconstructions with a sparse jump set, i.e., they are neither smooth nor noisy. Hence, ℓ_1 priors seem like promising candidates for sparse Bayesian inference and will be examined in more detail in this thesis.



(a) (b)

Figure 3.5.: (a) Illustration of Bayesian inference with an ℓ_1 prior: Level sets of likelihood (green), prior (red) and resulting posterior (blue). The star markers indicate the corresponding maxima, the dot marker the CM estimate of the posterior. (b) MAP estimates for the "Boxcar" scenario ($n = 63$) using an ℓ_1 prior with different λ and D being the forward difference operator.

Total Variation Priors

Total variation (TV) deblurring techniques (Burger and Osher 2013, Rudin et al. 1992) try to solve (3.12) for $\mathcal{J}(u)$ being a discrete version of the *total-variation* seminorm:

$$\mathrm{TV}(u^\infty) := \sup_{\substack{\varphi \in C_0^\infty(\Omega;\mathbb{R}^s) \\ \|\varphi\|_\infty \leqslant 1}} \int_\Omega u^\infty \operatorname{div} \varphi \, \mathrm{d}x, \tag{3.20}$$

which is defined for functions $u^\infty \in L^p(\Omega)$, $\Omega \subset \mathbb{R}^s$. Restricted to the Sobolev space $W^{1,1}$, $TV(u^\infty)$ becomes

$$\mathrm{TV}(u^\infty) = \int_\Omega \|\nabla u^\infty\|_2 \mathrm{d}x, \qquad u^\infty \in W^{1,1}. \tag{3.21}$$

Total variation imaging is a prominent example of *edge-preserving* image reconstruction techniques that are used in scenarios where the exact reconstruction of feature edges is of superior importance to, e.g., the contrast of these features.
Using a discretized version $TV_{dis}(u)$ of $\mathrm{TV}(u^\infty)$ to define a Gibbs prior as $p_{prior}(u) \propto \exp(-\lambda TV_{dis}(u))$ has become popular in Bayesian inversion as well (*TV priors*). The ℓ_1 increment prior used in the last paragraph is one possible realization of a TV prior in 1D. In 2D and higher dimensions, a direct implementation of (3.21) leads to *isotropic*

TV techniques: The ℓ_2-norm of the gradient is implemented correctly and apart from discretization errors, the prior is invariant with respect to rotations of the coordinate system. For instance, for a 2D image where we can index the components of u as $u_{(i,j)}$ with $i = 1, \ldots, N$, $j = 1, \ldots, N$, $n = N^2$, we can use forward differences in both spatial directions to define

$$p_{iTV}(u) \propto \exp\left(-\lambda \sum_{(i,j)}^{n} \sqrt{(u_{(i+1,j)} - u_{(i,j)})^2 + (u_{(i,j+1)} - u_{(i,j)})^2}\right) \tag{3.22}$$

This corresponds to an ℓ_1-block prior (cf. (3.16)): $p_{iTV}(u) \propto \exp\left(-\lambda \|Gu\|_{21}\right)$, where each row of G contains the discrete spatial derivatives, $G^T_{(i,j)} = (u_{(i+1,j)} - u_{(i,j)}, u_{(i,j+1)} - u_{(i,j)})$. As in EMEG, we use a block prior on the amplitudes of a vector field (the gradient field) in order to obtain rotation invariance. In contrast, *anisotropic TV* techniques, derived from replacing (3.20) by an equivalent semi-norm (Berkels et al. 2006), can be implemented using a conventional ℓ_1 prior. In the most simple case, (3.21) will be replaced by

$$a\mathrm{TV}(u^{\dagger,\infty}) := \int_\Omega \left|\partial_x u^{\dagger,\infty}\right| + \left|\partial_y u^{\dagger,\infty}\right| \, \mathrm{d}x, \tag{3.23}$$

which can be used to define

$$p_{aTV}(u) \propto \exp\left(-\lambda \sum_{(i,j)}^{n} \left|u_{(i+1,j)} - u_{(i,j)}\right| + \left|u_{(i,j+1)} - u_{(i,j)}\right|\right) = \exp\left(-\lambda \left\| \begin{bmatrix} D_x^T \\ D_y^T \end{bmatrix} u \right\|_1\right). \tag{3.24}$$

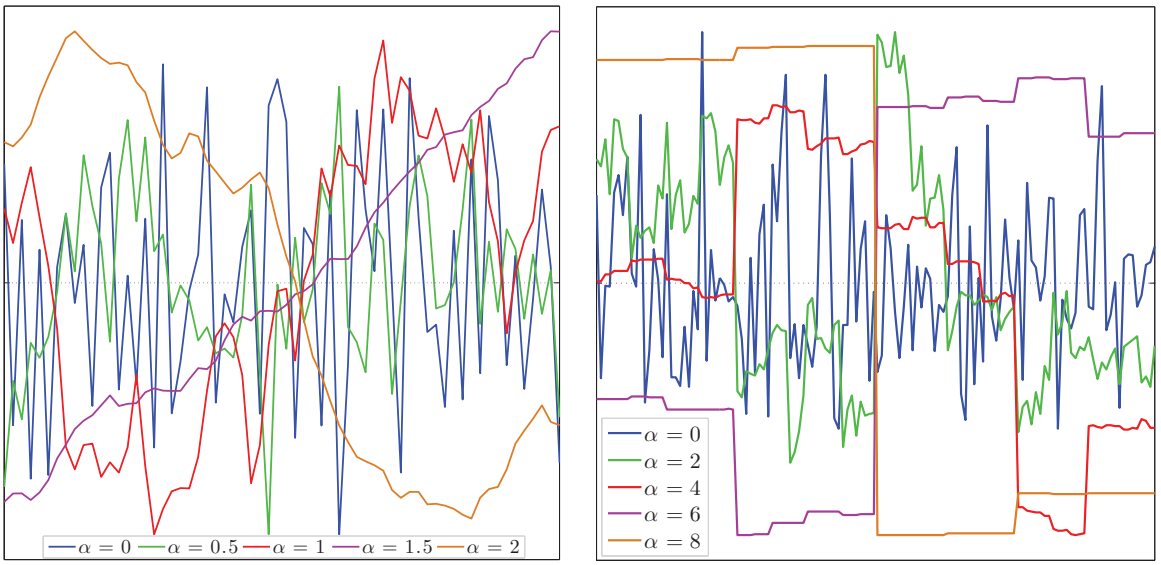
Here, D_x^T and D_y^T implement the partial derivative in x and y direction. This prior favors image edges aligned to the coordinate axes, but given a sufficiently high spatial resolution, the differences between isotropic and anisotropic models are negligible.
Although TV priors are commonly defined as ℓ_p^q priors with $p = q = 1$ (as above), we will also examine their extension to other values of p and q.

Orthonormal Bases, Wavelets and Besov Space Priors

A special class of ℓ_p priors can be derived by choosing $D^T = WV^T$, where V is an orthonormal basis $\{v_1, \ldots, v_n\}$ and W is a diagonal matrix of positive weights $\{w_1, \ldots, w_n\}$:

$$p_{prior}(u) \propto \prod_{i}^{n} \exp\left(-\lambda w_i^p |\langle v_i, u\rangle|^p\right) \tag{3.25}$$

We consider bases V that were constructed by discretizing the first n basis functions of an orthonormal basis of a suitable function space on Ω. The relative size of the coefficients $|\langle v_i, u\rangle|$ for growing i often encodes certain regularity features of u. Take for



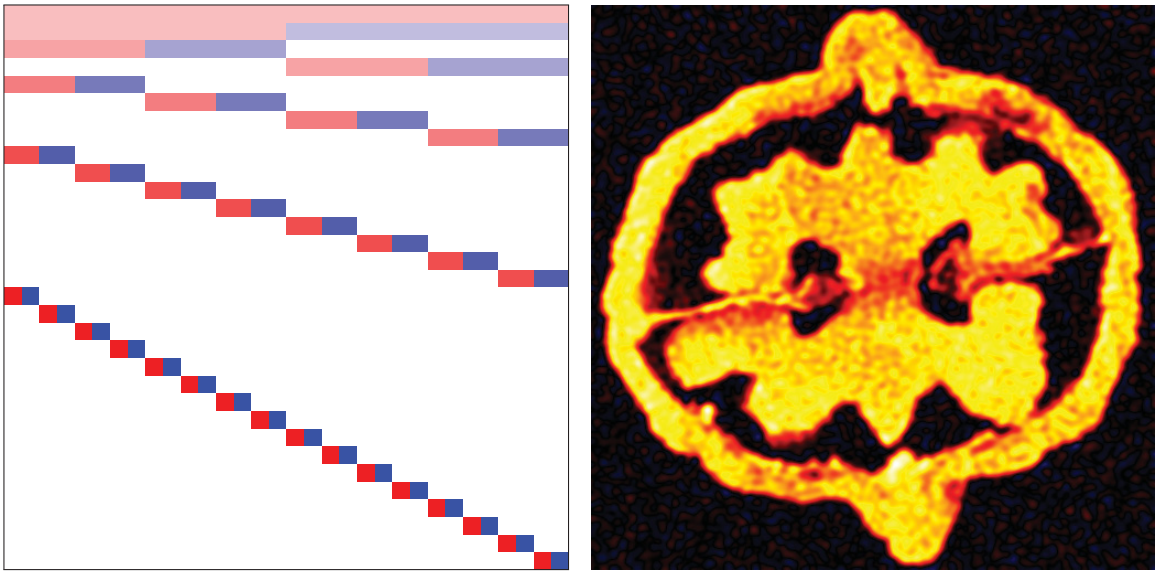
(a) (b)

Figure 3.6.: Random draws from (3.25) for $p = 1$ and (a) a discrete cosine wave basis ($n = 64$) with $w_i = i^\alpha$ and (b) a Haar wavelet basis ($n = 128$) with w_i being the scale weights raised to the power of α

instance V to be composed of discrete cosine waves (with increasing frequency). For smooth functions u, the coefficients will be large for some small frequencies and decay quickly for large ones. Non-smooth functions u that consist of many oscillations on small spatial scales will have a lot of large high frequency coefficients. Using a prior model like (3.25), we can adjust the relative variance of the different frequencies by w_i to choose whether smooth or non-smooth functions u should a-priori be more or less likely. An illustration is given in Figure 3.6.

A very promising class of orthonormal bases for imaging applications is given by wavelets that form a *multiresolution analysis*. These bases consist of wavelets generated from dyadic dilations and translations of certain generating functions that are piecewise continuous and compactly supported (see Figures 3.7a and A.3). As a result, u is decomposed into different scales and locations. The compact localization in space is an advantage of wavelet analysis over Fourier analysis, which is not suited to represent localized small scale variations or discontinuities of u in a compact way. Especially for the analysis of "natural" images or signals, wavelet-based multiresolution approaches can yield superior results (cf. Figures 1.6c and 3.7b).

As for the discrete cosine prior, using these bases allows to assign different variances to image features on different spatial scales which induces a certain type of spatial regularity. A specific, scale-dependent choice of w_i yields the *Besov space priors* introduced in Lassas et al. (2009). They have some appealing properties for Bayesian inversion



(a) (b)

Figure 3.7.: (a) Visualization of the Haarwavelet basis in 1D: Each row of the image represents a basis function v_i. The color scale ranges from blue (negative) to red (positive), cf. Figures 2.12b and 2.11b. One can clearly observe the dilation and translation. (b) Reconstruction of Figure 1.6a after keeping only the 1% largest *discrete cosine transformation* (*dct*) coefficients.

which will be examined in Section 5.2.6.

Notes and Comments

The definition of multivariate priors by ℓ_p^q norms is inspired by variational regularization approaches using corresponding energies (cf. (3.19)), and the infinite dimensional Hilbert/Banach space setting behind the continuous models. As it also facilitates practical computations, similar approaches are used in related fields where high dimensional settings dominate, for instance in statistical (machine) learning. In traditional Bayesian statistics, one would deduce priors from concepts such as maximum entropy or by characteristic functions. This leads to the use of *exponential families* as prior models, which also possess a structure that facilitates practical computations.

For these reasons, the priors commonly used in traditional Bayesian statistics and in Bayesian inverse problems differ. In particular, the ℓ_1 prior can be considered a product of 1D Laplace distributions but is not a multivariate extension of the 1D Laplace distribution as often confused in Bayesian inversion literature. One possibility to define such an extension is discussed in Eltoft et al. (2006)). Another issue is the incorporation of covariance information by D. An explicit relation to the prior covariance is only given in the Gaussian case where $C = (2\lambda DD^T)^{-1}$. We will further

discuss this issue in Section 6.1.

Bayesian modeling is one of the big advantages of Bayesian inference, but not really exploited here. We only described general constructions to encode scaling and general relationships between coordinates. Sophisticated models include contextual, geometrical, physiological or empirical information and draw from the rich field of (stochastic) mathematical modeling. On the other hand, computational inference for such models may be more challenging.

While Figure 3.5b shows that MAP estimation for the TV prior is able to recover the edges of $u^{\dagger,\infty}$ within the computational grid, it also reveals a crucial drawback of standard TV-based reconstructions: The estimates increasingly suffer from *contrast loss*. This well-known phenomena is more than a simple scaling problem as suggested by our example. In Osher et al. (2006), *Bregman iterations* were developed to compensate for it: A series of MAP estimates is computed for a prior that is iteratively updated. While this idea is conceptually close to inference procedures used in hierarchical Bayesian modeling (which we will introduce in Section 3.3), the rigorous formalization of the Bregman iteration within the Bayesian framework is not straightforward. See Burger and Osher (2013) for further details and a possible Bayesian interpretation.

3.2.5. Heavy-tailed Prior Models

The decay of a 1D distribution $p(u)$ for $|u| \longrightarrow \infty$ (the *tails* of the distribution) determines the likelihood of u taking exceptionally large absolute values. For many phenomena, empirical distributions have been observed which cannot be described by distributions that are *exponentially bounded*: They cannot be dominated by $C \exp(-\alpha|u|)$ for any choice of C and α. Such distributions are called *heavy-tailed*. In the multivariate case, the decay of the distribution into any particular direction (the radial profile) determines whether a distribution is heavy-tailed or not. For instance, all ℓ_p^q priors with $q \geqslant 1$ are not heavy-tailed as they are exponentially bounded for $\|u\|_2 \longrightarrow \infty$. For general Gibbs priors, the growth of $\mathcal{J}(u)$ for this limit is important.

While ℓ_p^q priors are heavy-tailed for $q < 1$, we will also examine a class of distributions with an even slower tail decay, namely, a power law decay: *Fat-tailed* distributions are characterized by an asymptotic decay of

$$p(u) \sim \|u\|^{-(1+\alpha)} \quad as \quad \|u\| \longrightarrow \infty, \quad \alpha > 0. \tag{3.26}$$

They occur in the description of extreme events such as earthquakes. Fat-tailed distributions often do not have finite moments, in particular no finite covariance for $0 < \alpha < 2$, and are not log-concave. A popular example of a fat-tailed distribution in

1D is the (centered) *Cauchy distribution*:

$$p(u) \propto \left(1 + \frac{u^2}{\theta}\right)^{-1} \tag{3.27}$$

which decays like $|u|^{-2}$. Here, θ determines the scale of the distribution, but should not be confused with the (non-existent) variance. The Cauchy distribution is a special case of the generalized (but centralized) *Student's t-distribution*:

$$p(u) \propto \left(1 + \frac{1}{\nu}\left(\frac{u^2}{\theta}\right)\right)^{-\frac{\nu+1}{2}} \tag{3.28}$$

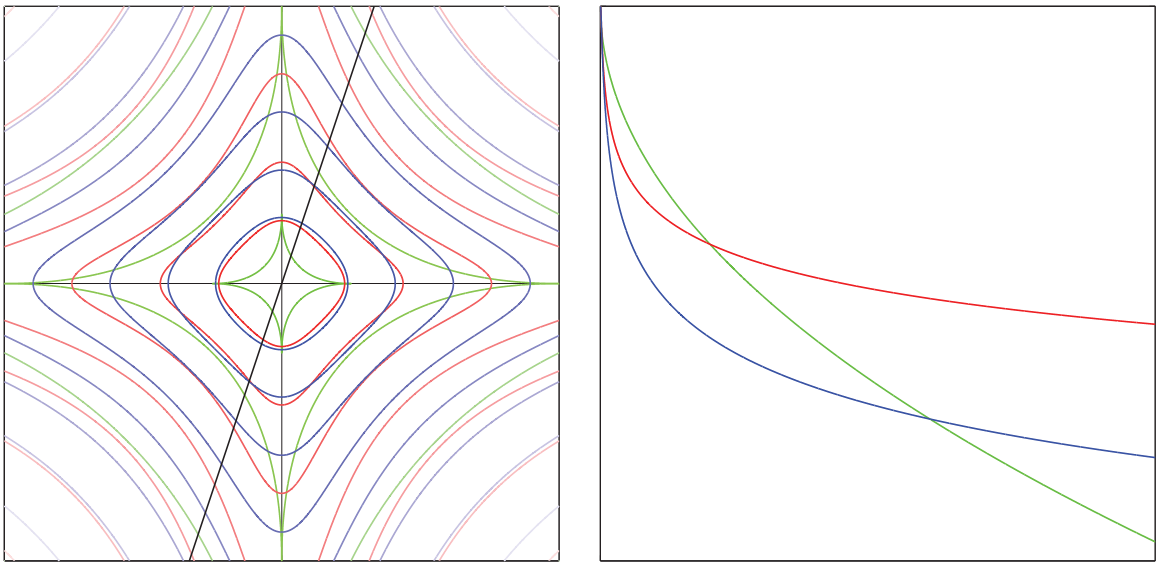
with $\nu > 0$ *degrees of freedom*. The Student's t-distribution decays like $|u|^{-(\nu+1)}$. From (3.28) we can construct a multivariate prior of the form:

$$\begin{aligned} p_{prior}(u) &\propto \prod_i^h \left(1 + \frac{1}{\nu}\left(\frac{(D_i^T u)^2}{\theta}\right)\right)^{-\frac{\nu+1}{2}} \\ &= \exp\left(-\frac{\nu+1}{2}\sum_i^h \log\left(1 + \frac{(D_i^T u)^2}{\nu\,\theta}\right)\right), \end{aligned} \tag{3.29}$$

where $D_i^T \in \mathbb{R}^n, i = 1, \ldots, h$ corresponds to the i-th row of a matrix D (similar to the construction of ℓ_p^q priors). We will refer to (3.29) as *product t-prior* and to the special case of $\nu = 1$ as *product Cauchy prior*. As demonstrated in (3.29) this prior can, of course, be written as a Gibbs prior with a logarithmic energy term (cf. (3.10)). However, compared to other prior models, this is not a "natural" description. One should also note that the parameterization does not allow for a linear scaling of the energy as in $\exp(-\lambda\mathcal{J}(u))$. Figure 3.8 shows the level-sets and radial profiles of three heavy-tailed distributions. An interesting feature of the product t-priors compared to ℓ_p^q priors is that their level-sets change their shape from a convex, ℓ_2-like shape near the origin to a non-convex shape. Figure 3.9 illustrates the use of the product Cauchy prior in Bayesian inversion.

As in the case of ℓ_p^q priors we might want to preserve certain sub-structures in the components of u. We can define *product t-block priors* as

$$p_{prior}(u) \propto \prod_i^h \left(1 + \frac{1}{\nu}\left(\frac{\left\|D_{[i]}^T u\right\|_2^2}{\theta}\right)\right)^{-\frac{\nu+1}{2}} \tag{3.30}$$



(a) (b)

Figure 3.8.: (a) Level sets of different (unnormalized) non-log-concave priors for the values $\{e^0, e^{-1}, e^{-2}, \dots\}$. Green lines: ℓ_p prior with $p = 1/2$, $\lambda = 2$. Red lines: Product Cauchy prior with $\theta = 0.03$. Blue lines: Product t-prior with $\theta = 0.03$, $\nu = 2$. (b) Radial profiles thereof along the black line in (a), with a logarithmic scaling of the vertical axis.

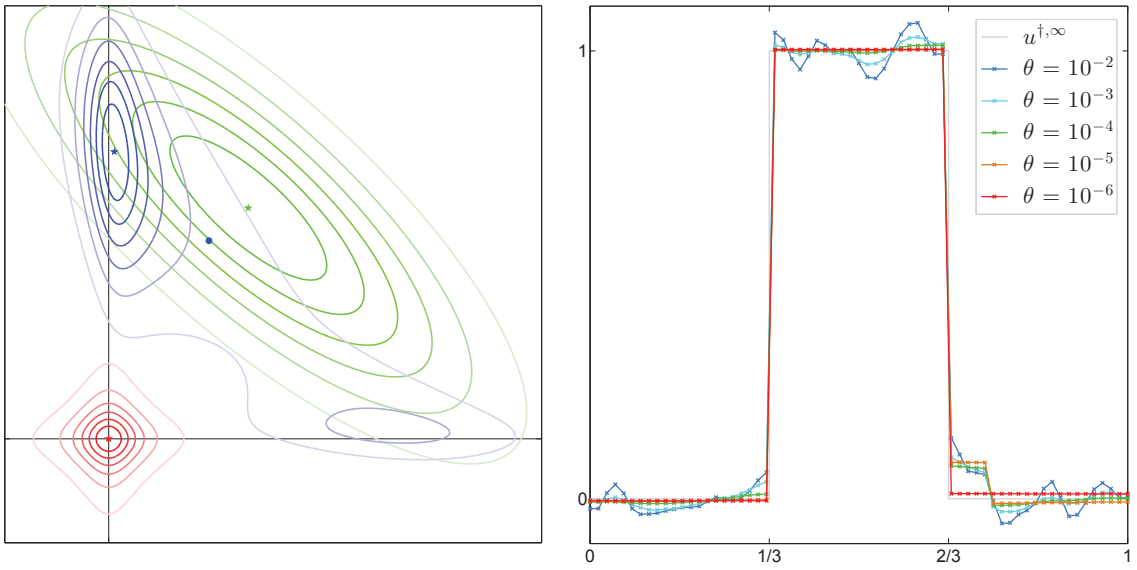
where $D_{[i]}^T \in \mathbb{R}^{h_i \times n}$, $i = 1, \dots, h$. In the case of EEG/MEG, $D_{[i]}^T$ extracts the $h_i = 3$ vector components at the i-th location and $\left\| D_{[i]}^T u \right\|_2$ would compute the corresponding current amplitude.

Notes and Comments

The parameterization of the product t-priors used in this thesis is rather non-standard but allows for an easy comparison with the other prior models. Note that multivariate t-distributions are commonly not defined by (3.29) but rather as

$$p(u) \propto \left(1 + \frac{1}{\nu}(u-\mu)^T \Sigma^{-1} (u-\mu)\right)^{-\frac{\nu+n}{2}} \tag{3.31}$$

with a scale matrix Σ. In the context of variational regularization, the properties of the product Cauchy energy as a regularization term were analyzed in Offtermatt and Kaltenbacher (2011), where also promising numerical results were presented.
In correspondence to ℓ_p priors, it is tempting to generalize the product t-prior by replacing $(D_i^T u)^2$ with $|D_i^T u|^p$. While such distributions have not been considered in statistics (to the best of our knowledge), we will see that we can derive them from a hierarchical Bayesian model in Section 3.3.3.



(a) (b)

Figure 3.9.: (a) Illustration of Bayesian inference with a Cauchy prior: Level sets of likelihood (green), prior (red) and resulting posterior (blue). The star markers indicate the corresponding maxima; the dot marker the CM estimate of the posterior. (b) MAP estimates for the "Boxcar" scenario ($n = 63$) using a Cauchy prior with different θ and D being the forward difference operator.

3.2.6. Normalization and Improper Priors

Prior distributions do not necessarily have to be proper probability distributions in the sense that they are normalizable, i.e., that they are in $L^1(\mathbb{R}^n)$. They only have to complement the likelihood distribution as a function of u in such a way that the posterior is normalizable. For linear inverse problems they have to be proper distributions only on the sub-spaces spanned by the singular vectors of A corresponding to singular values that are very small or even zero. Applied to the prior models introduced above, this is achieved if the condition of D^T restricted to these sub-spaces is sufficiently large. In particular, the null-space of D^T should not overlap with them. Another example is to use only hard constraints as a prior. While the Lebesgue measure of $\mathcal{C}$ may be infinite or zero, their usage can lead to a proper posterior. Prior distributions that are not normalizable are called *improper priors*. As one normally tries to incorporate as little a-priori knowledge as possible to rather let the data determine the solution, their usage is very common.

3.3. Hierarchical Bayesian Modeling

Hierarchical Bayesian modeling (*HBM*) is an extension of classical Bayesian modeling. The aim is to construct complex prior models that comprise different levels for the embedding of a-priori information of different origin and kind, organized in a top-down structure. In this thesis, we will mainly use HBM to develop alternative prior models for sparse Bayesian inversion that have interesting properties compared to ℓ_p priors. We will first develop HBM as an intuitive extension of Gaussian priors, and then generalize this construction to ℓ_p priors.

In Figure 3.4b, we saw that modeling all increments of a function as Gaussian random variables,

$$p_{prior}(u) \propto \prod_i \exp\left(-\lambda(u_{i+1} - u_i)^2\right), \tag{3.32}$$

is not suited for recovering functions with sparse increments. From a stochastic perspective, this is not surprising: Gaussian variables take their values on a characteristic length scale defined by their standard deviation; they are not *scale invariant*. If all increments have the same standard deviation $1/\sqrt{2\lambda}$, it is rather likely that their amplitudes are also similar. A sparse (or rather compressible, cf. Section 1.3) increment vector would consist of a few large-scale and a lot of small-scale increments. Within the Gaussian model, such vectors are extremely unlikely. The key idea is now to replace the fixed, uniform standard deviation of the Gaussian model by an individual, flexible γ_i for every component,

$$p_{prior}(u) \propto \prod_i^{n-1} \exp\left(-\frac{(u_{i+1} - u_i)^2}{\gamma_i}\right), \tag{3.33}$$

and to estimate γ_i from the data as well. Such parameters that determine the prior and have to be estimated from the data as well are called *hyperparameters* or *latent variables*. The latter term emphasizes that, in contrast to u, they were artificially introduced, often do not have a concrete physical meaning, and cannot be observed directly. In the Bayesian approach, they are modeled as random variables as well and their uncertainty is, again, modeled by a prior distribution $p_{hyper}(\gamma)$ on the vector of γ_i's (*hyperprior*). The joint prior over u and γ is usually expressed in the conditional form as

$$p_{prior}(u, \gamma) \propto p_{prior}(u|\gamma)\, p_{hyper}(\gamma) \tag{3.34}$$

and the full posterior for both u and γ becomes

$$p_{post}(u, \gamma|f) \propto p_{like}(f|u)\, p_{prior}(u|\gamma)\, p_{hyper}(\gamma). \tag{3.35}$$

As the posterior now depends on two different kinds of unknowns, more possibilities for Bayesian estimation are available (see "Notes and Comments"). We will concentrate on *fully-Bayesian* estimates, which treat u and γ in an equal way:

$$(\hat{u}_{\mathrm{MAP}}, \hat{\gamma}_{\mathrm{MAP}}) := \underset{(u,\gamma)}{\operatorname{argmax}} \{\, p_{post}(u, \gamma | f)\} \tag{3.36}$$

$$(\,\hat{u}_{\mathrm{CM}}\,,\,\hat{\gamma}_{\mathrm{CM}}\,) := \mathbb{E}\left[(u, \gamma)|f\right] = \int (u, \gamma)\; p_{post}(u, \gamma|f)\,\mathrm{d}u\,\mathrm{d}\gamma \tag{3.37}$$

In Figure 3.10, the extension of the Gaussian increment model to an HBM using an *inverse gamma distribution*,

$$p_{hyper}(\gamma) \propto \prod_i \gamma_i^{-(\alpha+1)} \exp\left(-\frac{\beta}{\gamma_i}\right), \tag{3.38}$$

as a hyperprior is illustrated. With increasing spread of the hyperprior, the individual γ_i's are allowed to vary more widely. As a result, two components of the increment vector can take significantly larger values than the rest and accurately reproduce the block function.

In principle, any parameters of the prior can be declared a hyperparameter and be estimated from the data. The next section describes the specific construction we examine in this thesis. Further models can be found in the "Notes and Comments" section. Hyperprior modeling is presented in Section 3.3.3.

3.3.1. Conditionally ℓ_p Hypermodels

The construction scheme used to extend the Gaussian increment model to the HBM (3.33) can easily be generalized to ℓ_p priors,

$$\exp\left(-\lambda \left\| D^T u \right\|_p^p\right) = \prod_i^h \exp\left(-\lambda \left| D_i^T u \right|^p\right), \tag{3.39}$$

by replacing the uniform scale parameter $1/\lambda$ with an individual γ_i:

$$p_{prior}(u|\gamma) = \prod_i^h \frac{1}{\mathcal{N}(\gamma_i)} \exp\left(-\frac{\left| D_i^T u \right|^p}{\gamma_i}\right) \tag{3.40}$$

The normalization factor $\mathcal{N}(\gamma_i)$ depends on γ. In principle, this dependence has to be computed in order to build (3.35). However, any particular dependence can be compensated for by choosing a specific hyperprior. Even if this leads to an improper prior, the posterior can still be proper (cf. Section 3.2.6). Therefore, we will only

approximate the dependence of normalization, but in a way that is exact for invertible D. We can easily compute that

$$\int \exp\left(-\frac{1}{\gamma_i}|v|^p\right) \mathrm{d}v = \int \exp\left(-\left|\frac{v}{\gamma_i^{1/p}}\right|^p\right) \mathrm{d}v \overset{sub.}{=} \gamma_i^{1/p} \int \exp\left(-|w|^p\right) \mathrm{d}w.$$

Hence, we can define the *(conditionally) ℓ_p hypermodels* as

$$\begin{aligned} p_{prior}(u|\gamma) :\propto \prod_i^h \gamma_i^{-1/p} \exp\left(-\frac{\left|D_i^T u\right|^p}{\gamma_i}\right) &= \exp\left(-\sum_i^h \frac{\left|D_i^T u\right|^p}{\gamma_i} - \frac{1}{p}\log(\gamma_i)\right) \\ = \exp\left(-\left\|\Gamma^{-1/p} D^T u\right\|_p^p - \frac{1}{p}\log(\det(\Gamma))\right), &\quad \text{with} \quad \Gamma := \mathrm{diag}\left(\gamma_1, \ldots, \gamma_h\right) \end{aligned} \tag{3.41}$$

The generalization to *(conditionally) ℓ_p-block hypermodels* can be done similar to (3.30), only the normalization has to be adapted:

$$p_{prior}(u|\gamma) \propto \prod_i^h \gamma_i^{-h_i/p} \exp\left(-\frac{\left\|D_{[i]}^T u\right\|_2^p}{\gamma_i}\right) \tag{3.42}$$

3.3.2. Hyperprior Modeling

Prior models for the positive *scale parameters* γ in ℓ_p hypermodels are usually different from those introduced for u in Section 3.2. For instance, a prior for u_i would be *non-informative* if it is translation invariant:

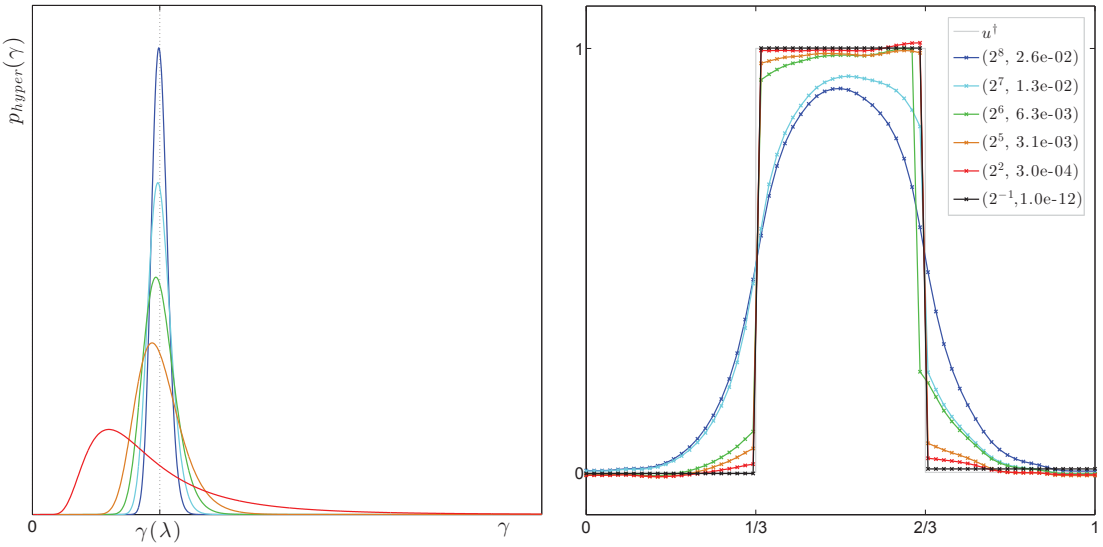
$$\mathbb{P}(u_i \in [a, b]) \overset{!}{=} \mathbb{P}(u_i \in [a + c, b + c]) \quad \forall\, c \in \mathbb{R} \tag{3.43}$$

Obviously, only the flat prior $p_{prior}(u_i) \propto 1$ fulfills this requirement. Using this prior for all components u_i basically means that u is solely determined by the likelihood, i.e., by the data. For a scale variable γ_i, a hyperprior would be non-informative if the probability that the variable lives on a certain scale is invariant:

$$\mathbb{P}(u_i \in [e^a, e^b]) \overset{!}{=} \mathbb{P}(u_i \in [e^{a+c}, e^{b+c}]) \quad \forall\, c \in \mathbb{R} \tag{3.44}$$

A flat hyperprior would not fulfill this requirement. Instead, the non-informative hyperprior is given by the fat-tailed $p_{hyper}(\gamma_i) \propto \gamma_i^{-1}$:

$$\mathbb{P}(u_i \in [e^{a+c}, e^{b+c}]) = \int_{e^{a+c}}^{e^{b+c}} \frac{1}{\gamma_i} \mathrm{d}\gamma_i = \log\left(e^{b+c}\right) - \log\left(e^{a+c}\right) = b - a. \tag{3.45}$$



(a) **(b)**

Figure 3.10.: Extension of the Gaussian prior (3.32) to an HBM (3.33). (a) Inverse gamma hyperpriors with mean $\gamma(\lambda) = 1/100^2$ and increasing variance. (b) Corresponding MAP estimates using an HBM. The legend displays the hyperprior parameter (α, β), cf. Section 3.3.2. The first hyperprior (blue line) is very narrow. As a result, the corresponding MAP estimate does not differ from the corresponding Gaussian model with $\lambda = 100^2$, cf. Figure 3.4b. For growing variance, the estimates change. Black line: MAP estimate using a parametrization of the hyperprior chosen to achieve the best reconstruction results, but with a different mean. The corresponding hyperprior would not fit well into plot (a).

In HBM, we often do not have too much a-priori knowledge about γ except that its components should be mutually independent. Therefore, we would like to use the non-informative hyperprior on all γ_i to let the data determine them. However, the improper distribution γ_i^{-1} leads to an improper posterior (Gelman 2006) and cannot be used. Instead, approximations are used that preserve certain properties of γ_i^{-1} while suppressing others. One possibility is given by the *inverse gamma distribution*:

$$p_{hyper}(\gamma_i) \propto \gamma_i^{-(\alpha+1)} \exp\left(-\frac{\beta}{\gamma_i}\right) = \exp\left(-\frac{\beta}{\gamma_i} - (\alpha+1)\log(\gamma_i)\right), \tag{3.46}$$

with the *shape parameter* $\alpha > 0$ and the *scale parameter* $\beta > 0$. The ratio behind this choice is that $\alpha > 0$ is introduced to increase the decay towards infinity. Now, the tail is integrable, but at the price of blowing up the integral towards zero. This is fixed by introducing the mollifier $\exp(-\beta\gamma_i^{-1})$, a positive function that decreases fast to zero for small values of γ_i and approaches 1 for large values of γ_i. The mollifier leads to a cut-off at a certain scale, determined by β, while not interfering with the power law decay. Figure 3.11 illustrates this construction. Inverse gamma distributions have the further

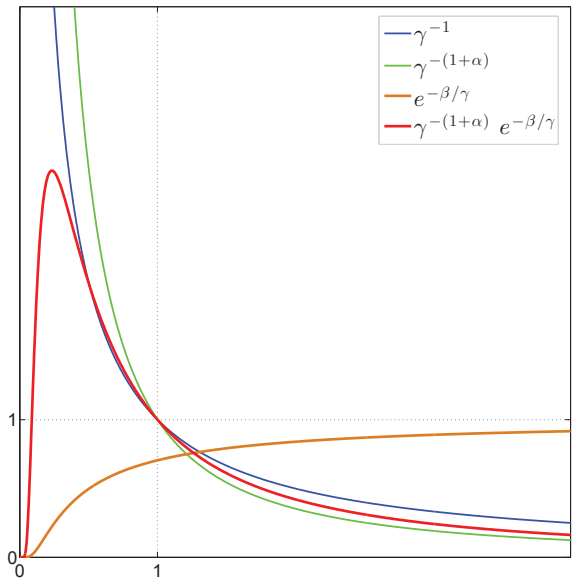


Figure 3.11.: Illustration of the approximation of γ_i^{-1} by the inverse gamma distribution.

advantage that they are *conjugate* to the ℓ_p hypermodel $p_{prior}(u|\gamma)$: Conditioned on u, the prior is also a product of inverse gamma distributions

$$p_{prior}(\gamma|u) \propto \prod_i^h \exp\left(-\frac{\left|D_i^T u\right|^p + \beta}{\gamma_i} - (\alpha + 1 + 1/p)\log(\gamma_i)\right) \tag{3.47}$$

with the parameters $\bar{\alpha}_i = \alpha + 1/p$ and $\bar{\beta}_i = \left|D_i^T u\right|^p + \beta$. A potential drawback of using the fat-tailed inverse gamma hyperprior is the non-convexity of its energy, i.e., the hyperprior is not log-concave. A related approximation of γ_i^{-1} is given by the *gamma distribution*,

$$p_{hyper}(\gamma_i) \propto \gamma_i^{\alpha-1} \exp\left(-\frac{\gamma_i}{\beta}\right) = \exp\left(-\frac{\gamma_i}{\beta} + (\alpha - 1)\log(\gamma_i)\right), \tag{3.48}$$

which is log-concave and exponentially bounded for $\alpha \geqslant 1$ (for $\alpha = 1$ it reduces to the exponential distribution). Section A.1.7 in Lucka (2011) contains a detailed comparison between both distributions.

Another popular approximation is given by the *log-normal distribution* describing a random variable constructed as $\gamma_i = \exp(z)$ where $z \sim \mathcal{N}(\mu, \sigma)$:

$$p_{hyper}(\gamma_i) \propto \gamma_i^{-1} \exp\left(-\frac{(\log(\gamma_i) - \mu)^2}{2\sigma^2}\right) \tag{3.49}$$

Here, γ_i^{-1} is modulated around a mean scale μ. Choosing a large value of σ leads to a better approximation.

3.3.3. Implicit Priors

The hyperparameter were only introduced to construct priors by a certain scheme. As already noted, they do not have a physical meaning by themselves. One might therefore ask, which *implicit prior* on u,

$$p_{prior}(u) = \int p_{prior}(u|\gamma) p_{hyper}(\gamma)\,\mathrm{d}\gamma, \tag{3.50}$$

results from their use. For ℓ_p hypermodels with an inverse gamma hyperprior, we can explicitly compute it:

$$\int p_{prior}(u|\gamma) p_{hyper}(\gamma)\,\mathrm{d}\gamma \overset{u}{\propto} \prod_i^h \int \exp\left(-\frac{\left|D_i^T u\right|^p + \beta}{\gamma_i} - (\alpha + 1 + 1/p)\log(\gamma_i)\right)\mathrm{d}\gamma_i \tag{3.51}$$

The integrands are inverse gamma distributions with parameters $\bar{\alpha}_i = \alpha + 1/p$ and $\bar{\beta}_i = \left|D_i^T u\right|^p + \beta$. The integrals are therefore given by the normalization of these distributions:

$$\begin{aligned}\prod_i^h \int \exp\left(-\frac{\bar{\beta}_i}{\gamma_i} - (\bar{\alpha}_i + 1)\log(\gamma_i)\right)\mathrm{d}\gamma_i &= \prod_i^h \frac{\Gamma(\bar{\alpha}_i)}{\bar{\beta}_i^{\bar{\alpha}_i}} = \prod_i^h \frac{\Gamma(\alpha + 1/p)}{\left(\left|D_i^T u\right|^p + \beta\right)^{(\alpha+1/p)}} \\ &\overset{u}{\propto} \prod_i^h \left(\left|D_i^T u\right|^p + \beta\right)^{-(\alpha+1/p)} \overset{u}{\propto} \prod_i^h \left(1 + \frac{\left|D_i^T u\right|^p}{\beta}\right)^{-(\alpha+1/p)}\end{aligned} \tag{3.52}$$

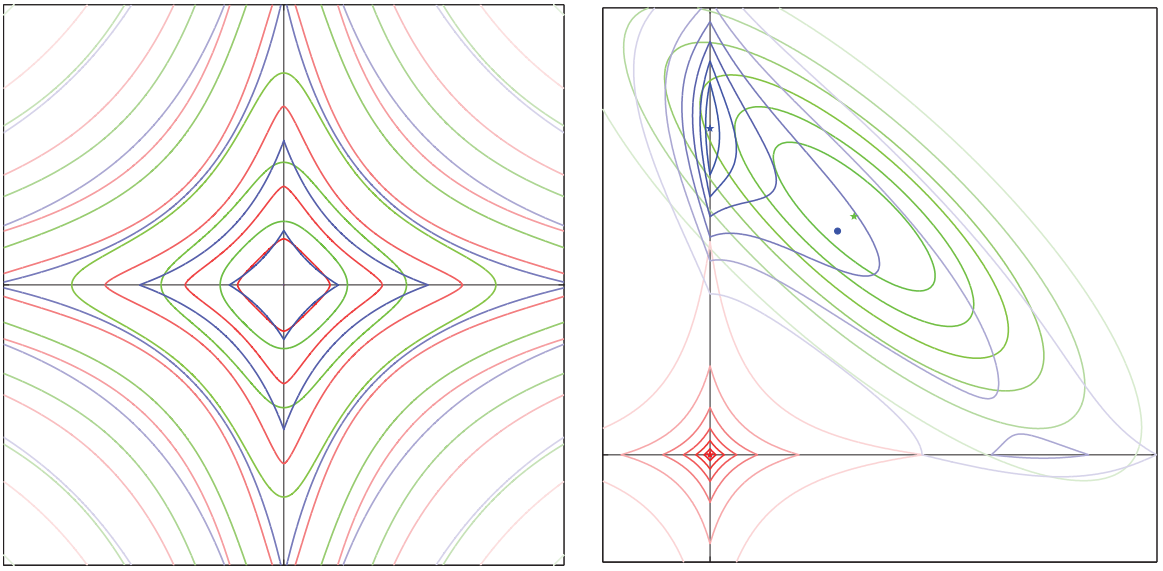
Here, $\Gamma(x)$ denotes the Gamma function. It turns out that the implicit prior is a heavy-tailed distribution that looks like a re-parameterized generalization of the product t-prior defined in (3.29) which we obtain for $p = 2$, $\alpha = \nu/2$ and $\beta = \nu\theta$. Using a similar parameterization as (3.29), we will therefore define the *product t_p-prior* as:

$$p_{prior}(u) \propto \prod_i^h \left(1 + \frac{1}{\nu}\left(\frac{\left|D_i^T u\right|^p}{\theta}\right)\right)^{-\frac{\nu+1}{p}} \tag{3.53}$$

Figure 3.12a compares the level-sets of different product t_p priors and Figure 3.12b illustrates the use of the product t_1 prior.

3.3.4. Notes and Comments

To the best of our knowledge, neither ℓ_p hypermodels nor product t_p priors have been considered for Bayesian inversion apart from $p = 1, 2$ (see, e.g., Garrigues and Olshausen 2010, for an ℓ_1 hypermodel).



(a) (b)

Figure 3.12.: (a) Level sets of different (unnormalized) product t_p-priors for the values $\{e^0, e^{-1}, e^{-2}, \ldots\}$. Green lines: $p = 2$, $\nu = 2$, $\theta = 0.03$. Red lines: $p = 1.5$, $\nu = 1$, $\theta = 0.06$. Blue lines: $p = 1$, $\nu = 1$, $\theta = 0.3$. (b) Illustration of Bayesian inference with a product t_1 prior: Level sets of likelihood (green), prior (red) and resulting posterior (blue). The star markers indicate the corresponding maxima, the dot marker the CM estimate of the posterior.

Note that the HBM construction here is different from the one used in Lucka (2011), where *Gaussian scale mixture models* were defined as

$$p_{prior}(u) \sim \mathcal{N}(0, \Sigma_u), \quad \Sigma_u = \sum_i^h \gamma_i C_i$$
$$\implies p_{prior}(u) \propto \exp\left(-\frac{1}{2} u^T \left(\sum_i^h \gamma_i C_i\right)^{-1} u\right)$$

with semi-positive definite *covariance components* C_i. This can easily be extended to other multivariate distributions that are constructed from 1D distributions by replacing the uni-variate quantity u^2/σ^2 by $u^T \Sigma^{-1} u$ with a *scale matrix* Σ (*elliptical distributions*, cf. the notes to Section 3.2.5). However, it cannot be extended to ℓ_p^q prior models with a general D for $p \neq 2$: In this case, we do not have an explicit analogue to a scale matrix. Therefore, we use a hierarchical extension of D as described above and use the term "ℓ_p hypermodel" to differentiate it from scale mixture models. In the Gaussian case, we can transform

$$p_{prior}(u) \propto= \exp\left(-\frac{1}{2}\left\|\sqrt{2}\Gamma^{-1/2} D^T u\right\|_2^2\right) = \exp\left(-\frac{1}{2} u^T \left(2D\Gamma^{-1} D^T\right) u\right), \qquad (3.54)$$

so we indirectly model the inverse covariance matrix $\left(2D\Gamma^{-1}D^T\right)$ by a direct modeling of $\Gamma^{-1/2}D^T$. However, in general, both HBM formulations for the Gaussian case are different. An exception here is the case $D = I_n$.

Fully-Bayesian inference as applied in this thesis relies on treating u and γ equally. As u corresponds to the physical quantity of primary interest while γ is just a mathematical parameter we introduced to formulate the prior model, this is not the only reasonable choice. *Type I approaches* first integrate the joint posterior with respect to γ and then maximize it with respect to u, usually in an iterative *expectation maximization* (*EM*) scheme. *Type II* or *empirical Bayesian approaches* first integrate over u and maximize it with respect to γ to then use the prior model corresponding to this optimal γ to infer u in a classical Bayesian way. Again, EM schemes are typically used for the computational tasks. Both methods are normally referred to as *semi-Bayesian* methods. *Variational Bayesian* approaches assume that the joint posterior factorizes with respect to u and γ and approximate the single-variable posteriors by distributions that facilitate the use of alternating updating schemes. They are normally referred to as *approximate-Bayesian* methods.

While we use the term "hierarchical Bayesian model" in this thesis for a very specific type of prior construction, there is no sharp general definition of it. Hierarchical models can be considered as a specific type of *Bayesian (belief) networks*, also known as *latent variable* or *graphical* models, which are among the most popular inference models used in contemporary statistics, in particular in the wider field of machine learning. A general reference is given by MacKay (2003). A recent overview of hierarchical models used in EEG/MEG source reconstruction can be found in Lucka (2011), further examples in other inverse problems include Bardsley et al. (2010), Calvetti and Somersalo (2007, 2008), Helin (2010b), Helin and Lassas (2009), Wang et al. (2013).

3.4. Bayesian Estimation and Decision Theory

3.4.1. Bayesian Estimators

Apart from the two popular point estimates $\hat{u}_{\mathrm{MAP}}$ and $\hat{u}_{\mathrm{CM}}$, the information contained in the posterior can be exploited in other ways:

- Point estimates can be based on other criteria. In the next section, we will develop a theoretical framework to derive alternative point estimates that are optimal for certain criteria. Another example are heuristically defined estimates. For instance,

we could define $\hat{u}$ such that $\hat{u}_i$ is the median of the *marginal posterior* on u_i:

$$p^i_{post}(u_i|f) := \int p_{post}(u|f)\,\mathrm{d}u_1 \ldots \mathrm{d}u_{i-1}\,\mathrm{d}u_{i+1} \ldots \mathrm{d}u_n \tag{3.55}$$

- *Credible interval* estimates are a Bayesian analogue to confidence intervals in frequentist statistics: For a component u_i, one searches for intervals $[a, b]$ that contain a certain fraction p of the probability mass of the marginal posterior on u_i. There are different possible choices of such intervals. One could search for the narrowest interval, an interval such that $p^i_{post}(a|f) = p^i_{post}(b|f)$, such that it is centered around the mean, or that it fulfills other reasonable constraints.
- *Credible region* estimates are a multivariate version of credible intervals: One searches for superlevel sets of the full posterior containing a certain fraction p of the total posterior probability mass.
- *Extreme value probabilities* try to estimate the a-posteriori probability that a feature $g(u)$, $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ exceeds some critical value.
- *Conditional covariance* (*CCov*) estimates try to assess the spatial distribution of the variance and explore the dependencies between the components of u:

$$\mathbb{C}\mathrm{ov}\,[u|f] = \int (u - \hat{u}_{\mathrm{CM}})(u - \hat{u}_{\mathrm{CM}})^T\, p_{post}(u|f)\,\mathrm{d}u \tag{3.56}$$

  Conditional variance (*CVar*) and *conditional standard deviation* (*CStd*) are defined accordingly.

The computational challenges of the different estimates vary considerably. We will revisit this issue in the next Chapter.

3.4.2. Bayes Cost Method

In the Bayesian framework, an estimator $\hat{u}$ for $u^\dagger$ is a random variable as well: It relies on the random variables ε and u. As such, it might perform well in certain cases while giving catastrophic results in others. *Bayesian decision theory* (or more general, *statistical estimation theory*) examines the general behavior of estimators to find optimal estimators for a given task. A common approach to measure the desired and undesired properties of an estimator $\hat{u}$ is to define a *cost or loss function* $\Psi(u, \hat{u})$. This is the basis of the *Bayes cost method*: The *Bayes cost* is defined by the expected cost or average performance:

$$BC_\Psi(\hat{u}) := \mathbb{E}\,[\Psi(u, \hat{u}(f))|f] \;=\; \int\int \Psi(u, \hat{u}(f))\, p(u, f)\,\mathrm{d}u\,\mathrm{d}f$$

$$= \int\int \Psi(u, \hat{u}(f))\, p_{like}(f|u)\, \mathrm{d}f\, p_{prior}(u)\, \mathrm{d}u$$
$$\overset{(3.1)}{=} \int\int \Psi(u, \hat{u}(f))\, p_{post}(u|f)\, \mathrm{d}u\, p(f)\, \mathrm{d}f \tag{3.57}$$

The *Bayes estimator* $\hat{u}_\Psi$ is the estimator, which minimizes $BC_\Psi(\hat{u})$:

$$\hat{u}_\Psi := \underset{\hat{u}}{\operatorname{argmin}} \{BC_\Psi(\hat{u})\} \tag{3.58}$$

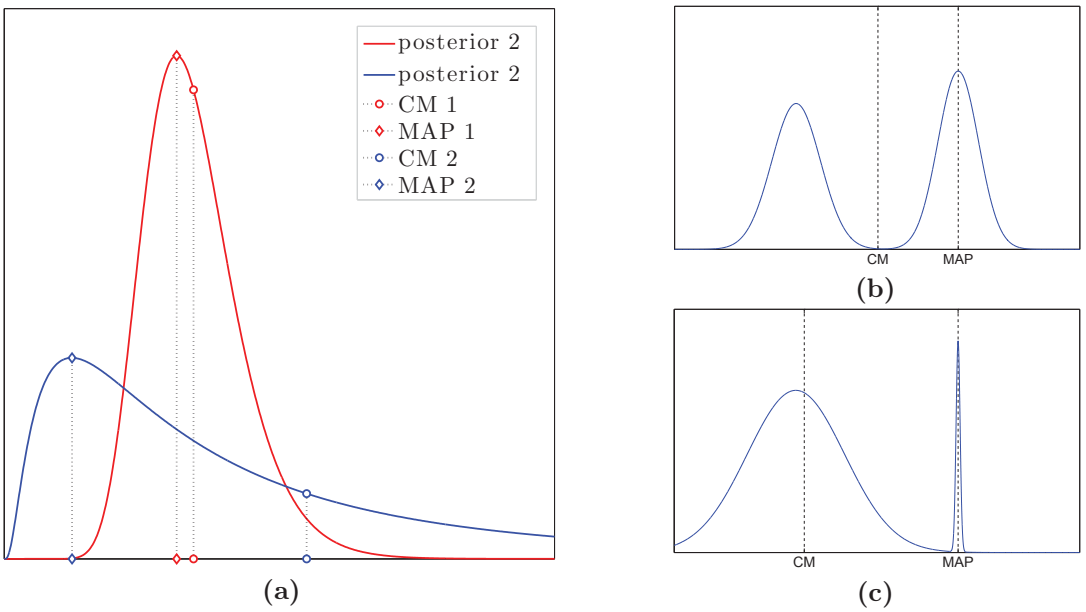
In (3.57), $\hat{u}(f)$ only depends on f and the marginal density $p(f)$ is non-negative. Thus, $\hat{u}_\Psi$ also minimizes

$$\hat{u}_\Psi(f) = \underset{\hat{u}}{\operatorname{argmin}} \left\{ \int \Psi(u, \hat{u}(f))\, p_{post}(u|f)\, \mathrm{d}u \right\}. \tag{3.59}$$

The Bayes cost method is one way to design estimators for a given task. Conversely, it can be used to examine estimators that were defined in other ways by deriving and comparing their Bayes cost functions. A focus of this thesis is to compare MAP and CM estimates for sparse prior models in high dimensional scenarios. Once defined (cf. (3.2) and (3.3)), an immediate and obvious question arises: What are their differences and which of them is "better"? This is not only a matter of constant debate within the field of Bayesian inversion, but due to the direct correspondence of $\hat{u}_{\mathrm{MAP}}$ to the solution of (3.12), it is also a dispute with the field of variational regularization theory. In the next section, we will review the "classical" view on this issue as found in all standard presentations on Bayesian inference (see, e.g., Gelman et al. 2003, Kaipio and Somersalo 2005, Kay 1993).

3.4.3. MAP or CM Estimation: The Classical View

The CM estimate is the mean of the posterior, while the MAP estimate is the (highest) mode of the posterior (see Figure 3.13a). However, this does not provide any intuition why one of them should be better suited to represent a distribution. Hence, a lot of presentations of the topic provide plots of hypothetical distributions like Figures 3.13b and 3.13c to show that none of them is better in general. However, one might argue that the CM estimator as the mean value is an intuitive choice: It is the "center of (probability) mass" and corresponds to the average of a sample, familiar from every-day descriptive statistics.
As the illustrative comparison does not give any useful intuition, the Bayes cost formalism is usually used to provide a decisive theoretical argument:



(a) (b) (c)

Figure 3.13.: (a) Comparison of MAP and CM estimates for two posterior densities. (b)-(c) Hypothetical, bimodal posterior distributions to show that none of the point estimates is better in general.

- The CM estimate is the Bayes estimator for the mean squared error,

$$\Psi_{\text{MSE}}(u, \hat{u}) = \|u - \hat{u}\|_2^2, \tag{3.60}$$

  which seems to be a very natural and reasonable choice for Ψ. Interpreted geometrically, one also speaks of a "well-centeredness" of $p_{post}(u|f)$ around $\hat{u}_{\text{CM}}$. As it is by default unbiased with respect to $p_{post}(u|f)$, one can further show that $\hat{u}_{\text{CM}}$ is also the *minimum error variance estimator*.

- On the other hand, the MAP estimate can only be seen as an *asymptotic* Bayes estimator of

$$\Psi_\delta(u, \hat{u}) = \begin{cases} 0 & \text{if } \|u - \hat{u}\|_\infty \leqslant \delta \\ 1 & \text{otherwise} \end{cases} \tag{3.61}$$

  for $\delta \to 0$ (*uniform cost* or *0-1 loss*). Therefore, it is usually not considered a proper Bayes estimator. This characterization also does not seem to allow for an intuitive geometrical interpretation of $\hat{u}_{\text{MAP}}$ akin to the one for $\hat{u}_{\text{CM}}$.

In summary, the classical view favors the CM estimate and discriminates the MAP estimate on the basis of the Bayes cost method.
The theoretical difference between MAP and CM estimates seems to fit to the different kind of operation by which they are defined: One as an optimization, the other as an integration task (cf. (3.2), (3.3)). Our computational studies will involve the

computation of many MAP and CM estimates. In light of these results, we will revisit the "MAP or CM?" question in Chapter 6, where we will also provide new theoretical results.

3.4.4. Notes and Comments

Due to computational challenges, the full range of Bayesian estimation in high dimensional inverse problems is rarely explored up to now. Exemplary applications to remote sensing, algae population dynamics, image deblurring and geothermal reservoir modeling can be found in Calvetti and Somersalo (2007), Cui et al. (2011), Haario et al. (2004, 2006).

3.5. Recovery Conditions for MAP Estimates

In this section, we will introduce concepts from compressed sensing (cf. Section 1.3) to examine the performance of MAP estimates for ℓ_1 priors in the *noise-free limit*, i.e., we assume $\operatorname{tr}(\Sigma_\varepsilon) \longrightarrow 0$:

$$p_{like}(f|u) \propto \mathbb{1}_{\{Au=f\}}(u) \tag{3.62}$$

We are particularly interested whether the *exact recovery* of sparse $u^\dagger$ is possible, i.e., $\hat{u}_{\mathrm{MAP}} = u^\dagger$. These examinations can be seen as complementary to Bayesian decision theory. Examining $\hat{u}_{\mathrm{MAP}}$ in this limit is related to the concept of *consistency* of estimators in frequentist statistics, and can be seen as assessing the "best-case" performance: In practical applications, the noise-free limit often corresponds to optimal measurement conditions like the limit of long recording times or high radiation doses in CT or the limit of many averaged trials in the EP/EF analysis with EMEG. The assumption of $u^\dagger$ being really sparse rather than compressible should also be understood in a "best-case" sense: For the continuous prior models we use, the probability of $|u|_0 < n$ is zero.
In this section and the following chapters, basic concepts from convex analysis will be used which are summarized in Section A.1. An extensive reference containing the proofs to all presented theorems is given by Foucart and Rauhut (2013).

Basic Properties We are interested in recovering $u^\dagger$ being the *unique k-sparse solution* to $Au = f$: $\{u \,|\, Au = Au^\dagger, |u|_0 \leqslant k\} = \{u^\dagger\}$. It turns out that $u^\dagger$ is the unique solution to

$$\min |u|_0, \quad s.\,t. \quad Au = f. \tag{L0}$$

While solving (L0) is difficult we recall that computing $\hat{u}_{\text{MAP}}$ for an ℓ_1 prior and (3.62) would correspond to solving

$$\min\|u\|_1, \quad s.\,t. \quad Au = f. \tag{L1}$$

This problem (also called *basis pursuit*) is considerably less difficult to solve due to the convexity of the ℓ_1-norm. Hence, conditions that guarantee that $u^\dagger$ and $\hat{u}_{\text{MAP}}$ are equal would be very advantageous. For the most fundamental one we have to define:

Definition 3.1. A matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy the *null space property* (*NSP*) relative to a set $I \subset \{1,\dots,n\}$ if

$$\|v_I\|_1 < \|v_{I^c}\|_1 \qquad \forall\, v \in \ker(A) \setminus \{0\}, \tag{NSP}$$

where v_I/V_{I^c} are the vectors of all components of v belonging to I/I^c. A satisfies the null space property of order k if it satisfies the NSP relative to all sets with $\text{card}(I) \leqslant k$.

The NSP basically requires the kernel of A to be spread out in $\mathbb{R}^n$; in particular, it should not contain any sparse elements.

Theorem 3.1. Given $A \in \mathbb{R}^{m\times n}$, every $u^\dagger \in \mathbb{R}^n$ with support I is the unique solution to (L1) with $f = Au^\dagger$ if and only if A satisfies the NSP relative to I.

Theorem 3.2. Given $A \in \mathbb{R}^{m\times n}$, every k-sparse $u^\dagger \in \mathbb{R}^n$ is the unique solution to (L1) with $f = Au^\dagger$ if and only if A satisfies the NSP of order k.

Theorem 3.1 is an example of a *non-uniform* or *local recovery condition* as it depends on the concrete $u^\dagger$, while Theorem 3.2 is a *uniform* or *global recovery condition* that guarantees the recovery for *all* k-sparse $u^\dagger$. While both NSP conditions are optimal in the sense that they are necessary and sufficient, they are hard to verify in practice. In the next two sections, we will examine stronger conditions that will imply that the NSP condition holds but are easier to compute.

Normalization Some of these conditions will require A to be column-normalized: $\|A_i\|_2 = 1$. Let $W := \text{diag}\,(\|A_1\|_2 \dots, \|A_n\|_2)$, $A^\sharp := AW^{-1}$, $v := Wu$. Then, $Au = A^\sharp v$ and we can also examine the normalized problems:

$$v^\dagger \quad := \text{argmin}\,|v|_0, \quad s.\,t. \quad A^\sharp v = f \tag{nL0}$$

$$\hat{v}_{\text{MAP}} := \text{argmin}\|v\|_1, \quad s.\,t. \quad A^\sharp v = f \tag{nL1}$$

Note that $\hat{v}_{\text{MAP}}$ corresponds to the MAP estimate for an ℓ_1 prior with diagonal weighing matrix $D = W$. We can also re-transform $\hat{u}_v := W^{-1}\hat{v}_{\text{MAP}}$, but this estimate does not

correspond to a MAP estimate in an obvious way anymore. We will examine the differences between normalized and un-normalized recovery in more detail in Section 5.4.6. If we formulate a recovery condition using $A^\sharp$ it means that this condition is only valid for the normalized setting. If we use A instead, it can be used in both cases.

3.5.1. Uniform Recovery Conditions

The strongest uniform condition is related to the *coherence* of a matrix A,

$$\mu(A) := \max_{i \neq j} \frac{|A_i^T A_j|}{\|A_i\| \|A_j\|}, \tag{3.63}$$

which measures the maximal similarity of the matrix columns by their scalar product. One can show that the bounds

$$\sqrt{\frac{1}{m}} \leqslant \sqrt{\frac{n-m}{m(n-1)}} \leqslant \mu(A) \leqslant 1 \tag{3.64}$$

hold. Two matrix columns A_i and A_j correspond to the signals Av_1, Av_2 produced by two 1-sparse solutions $v_1 = e_i$, $v_2 = e_j$. Intuitively, if the coherence is small, they cannot be too similar and it should be possible to separate v_1 and v_2 correctly. The next theorem formalizes this:

Theorem 3.3. Let $A^\sharp \in \mathbb{R}^{m \times n}$ have ℓ_2 normalized columns. If

$$k \leqslant \tfrac{1}{2}(\mu(A^\sharp)^{-1} + 1), \tag{Coh}$$

then every k-sparse $u^\dagger$ can be recovered as the unique solution to (L1) with $f = A^\sharp u^\dagger$.

While the coherence is easy to compute, its recovery guarantees are rather restrictive and can be improved a lot by conditions that rely on the *restricted isometry constant* of $A^\sharp$, which is the smallest number δ_k such that

$$(1 - \delta_k)\|u\|_2^2 \leqslant \left\|A^\sharp u\right\|_2^2 \leqslant (1 + \delta_k)\|u\|_2^2 \qquad \forall u : |u|_0 \leqslant k. \tag{3.65}$$

There are different theorems relying on δ_k that give sufficient conditions under which further assumptions (such as noise or modeling errors) which algorithm will recover any k-sparse $u^\dagger$. They typically take the form

$$\delta_{pk} < \delta_*, \tag{3.66}$$

where p is an integer. For our purpose, we have:

Theorem 3.4. Let $A^\sharp \in \mathbb{R}^{m\times n}$ have ℓ_2 normalized columns. If

$$\delta_{2k} \leqslant 4/\sqrt{41} \approx 0.6246, \tag{RIP}$$

then every k-sparse $u^\dagger$ can be recovered as the unique solution to (L1) with $f = A^\sharp u^\dagger$.

3.5.2. Non-uniform Conditions

If uniform recovery conditions for a given k are not fulfilled, we can still examine local recovery conditions for a large number of k-sparse vectors to make statements like "x% of all k-sparse vectors $u^\dagger$ can be recovered as the unique solution to (L1) with $f = Au^\dagger$". In Tropp (2004), the following sufficient recovery condition is established:

Theorem 3.5. Given $A \in \mathbb{R}^{m\times n}$, a k-sparse $u^\dagger$ with support I can be recovered as the unique solution to (L1) with $f = Au^\dagger$ if A_I is injective and

$$\left\|A_I^+ a_j\right\|_1 < 1 \quad \forall j \notin I, \tag{Tr}$$

where A_I^+ denotes the Moore-Penrose pseudo-inverse of A_I.

One can easily show that (Coh) implies (Tr). In Fuchs (2004), the optimality conditions of (L1) and its dual problem were analyzed to derive the following weaker condition, which turns out to be sufficient and necessary (cf. Theorem 4.30 in Foucart and Rauhut 2013):

Theorem 3.6. Given $A \in \mathbb{R}^{m\times n}$, a k-sparse $u^\dagger$ with support I can be recovered as the unique solution to (L1) with $f = Au^\dagger$ if and only if A_I is injective and a *dual vector* $w \in \mathbb{R}^m$ exists such that

$$|w^T A_j| < 1 \quad \forall j \notin I \qquad \text{and} \qquad A_I^T w = \operatorname{sign}(u_I^\dagger). \tag{FuB}$$

In addition to (FuB), a stronger but easier to verify sufficient condition is also given in Fuchs (2004):

Theorem 3.7. Given $A \in \mathbb{R}^{m\times n}$, a k-sparse $u^\dagger$ with support I can be recovered as the unique solution to (L1) with $f = Au^\dagger$ if A_I is injective and

$$|w_+^T A_j| < 1 \quad \forall j \notin I \qquad \text{for} \qquad w_+ := (A_I^T)^+ \operatorname{sign}(u_I^\dagger) \tag{FuA}$$

Comparing (FuA) and (FuB), we note that the idea behind (FuA) is to restrict the dual vector from being any solution to $A_I^T w = \operatorname{sign}(u_I^\dagger)$ to the minimum-norm solution $w_+ := (A_I^T)^+ \operatorname{sign}(u_I^\dagger)$. Hence, (FuA) implies (FuB) (which explains the notation using

A and B). As one can easily check that (Tr) implies (FuA), we have a chain of stronger to weaker recovery conditions:

$$(\text{Coh}) \Rightarrow (\text{Tr}) \Rightarrow (\text{FuA}) \Rightarrow (\text{FuB})$$

Unfortunately, the verification of the conditions also becomes increasingly computationally demanding, as we will see in Section 4.4.

3.5.3. Stable and Robust Conditions

The above conditions hold for the noise-free limit under the assumption of exact sparsity. They can be turned into stronger conditions that also cover compressible $u^\dagger$ and measurement noise. Naturally, these conditions are harder to fulfill for a matrix A. *Stable* conditions cover the case of compressible $u^\dagger$. Instead of exact recovery guarantees they bound the ℓ_1-error $\left\|u^\dagger - u_{\ell_1}\right\|_1$ of the solution u_{ℓ_1} of (L1) by the *best k-term approximation* of $u^\dagger$ which is defined as

$$\sigma_k(u)_p := \inf_{|v|_0 \leqslant k} \|u - v\|_p \tag{3.67}$$

We have

$$\sigma_k(u)_q \leqslant \frac{c_{p,q}}{k^{1/p-1/q}} \|u\|_p, \qquad p \leqslant 2, \quad c_{p,q} \leqslant 1, \tag{3.68}$$

in particular

$$\sigma_k(u)_2 \leqslant \frac{1}{2\sqrt{k}} \|u\|_1. \tag{3.69}$$

As one example, one can obtain the *stable NSP with constant* ρ by replacing $\|v_I\|_1 < \|v_{I^c}\|_1$ in (NSP) with $\|v_I\|_1 < \rho\|v_{I^c}\|_1$, $0 < \rho < 1$, including the following error estimate:

$$\left\|u^\dagger - u_{\ell_1}\right\|_1 \leqslant \frac{2(1+\rho)}{1-\rho} \sigma_k(u^\dagger)_1, \tag{3.70}$$

Robust recovery conditions also cover the case of measurement noise. The equality constraint $Au = f$ in (L1) is replaced by $\|Au - f\|_2 \leqslant \delta$, where δ is an a-priori bound on $\|\varepsilon\|_2$:

$$\min \|u\|_1, \quad s.\,t. \quad \|Au - f\|_2 \leqslant \delta, \tag{3.71}$$

The *robust NSP with constants* ρ *and* τ is obtained by replacing $\|v_I\|_1 < \|v_{I^c}\|_1$ in (NSP) with $\|v_I\|_1 < \rho\|v_{I^c}\|_1 + \tau\|Av\|_2$, $0 < \rho < 1$, $\tau > 0$. The solution $u^\delta_{\ell_1}$ of (3.71) then fulfills:

$$\left\|u^\dagger - u^\delta_{\ell_1}\right\|_1 \leqslant \frac{2(1+\rho)}{1-\rho} \sigma_k(u)_1 + \frac{4\tau}{1-\rho}\delta \tag{3.72}$$

For matrices A fulfilling (RIP), one can obtain

$$\left\|u^\dagger - u^\delta_{\ell_1}\right\|_1 \leqslant C\sigma_k(u)_1 + D\sqrt{k}\delta \tag{3.73}$$

$$\left\|u^\dagger - u^\delta_{\ell_1}\right\|_2 \leqslant \frac{C}{\sqrt{k}}\sigma_k(u)_1 + D\delta \tag{3.74}$$

with constants $C, D > 0$ depending only on δ_{2k}.

3.5.4. Source Conditions

Regularization theory (cf. Section 1.2) is also concerned with deriving error estimates but the setting is usually different from the one considered in typical compressed sensing applications. One would start in the infinite dimensional setting (1.2) and analyze the convergence of solutions $u^\delta_{\lambda(\delta,f^\delta)}$ of (1.5) in the noise-less limit, which is in some way parameterized as $\delta \searrow 0$. The regularization parameter λ should be chosen depending on δ, requiring that $\lambda(\delta, f^\delta) \searrow 0$ as $\delta \searrow 0$. Compared to typical compressed sensing scenarios, $\mathcal{A}$ is (severely) ill-conditioned. The error between $u^\delta_{\lambda(\delta,f^\delta)}$ and $u^{\dagger,\infty}$ is measured by a non-negative error measure $\mathcal{E}(u^\delta_{\lambda(\delta,f^\delta)}, u^{\dagger,\infty})$ and one tries to derive *convergence rates* of the form

$$\mathcal{E}(u^\delta_{\lambda(\delta,f^\delta)}, u^{\dagger,\infty}) = \mathcal{O}(\phi(\delta)) \qquad \text{for} \qquad \delta \searrow 0 \tag{3.75}$$

for an *index function* ϕ, i.e., ϕ is a positive, strictly increasing function on $\mathbb{R}_+$ with $\lim_{t\to 0}\phi(t) = 0$. A central observation is that such rates cannot be established without assuming some kind of smoothness of $u^{\dagger,\infty}$ with respect to $\mathcal{A}$ and its adjoint $\mathcal{A}^*$. In the discrete setting and for ℓ_p-norm data fidelities $\mathcal{H}_f = \frac{1}{p}\|Au - f\|_p^p$ (which covers the Gaussian noise model), such a smoothness assumption is given by:

Definition 3.2. An element u meets a *source condition* (SC) with respect to a regularization functional $\mathcal{J}$ if there exists a *source element* $w \in \mathbb{R}^m$ such that $A^T w \in \partial\mathcal{J}(u)$ (cf. Section A.1).

Put differently, u satisfies a SC if a *subgradient* $\xi \in \partial\mathcal{J}(u)$ exists that is in the range of A^T. As A^T is typically a smoothing operator in inverse problems, this implies that ξ is smooth. In addition to a SC, one has to measure the error in the *Bregman distance* induced by $\mathcal{J}(u)$ (cf. Section A.1) to derive convergence rates, in particular in the case of non-smooth $\mathcal{J}(u)$:

$$\mathcal{E}(u^\delta_{\lambda(\delta,f^\delta)}, u^{\dagger,\infty}) := D_J^\xi(u^\delta_{\lambda(\delta,f^\delta)}, u^{\dagger,\infty}) = \mathcal{J}(u^\delta_{\lambda(\delta,f^\delta)}) - \mathcal{J}(u^{\dagger,\infty}) - \left\langle \xi, u^\delta_{\lambda(\delta,f^\delta)} - u^{\dagger,\infty}\right\rangle \tag{3.76}$$

For the ℓ_1-norm, the i-th component of the subdifferential is given as the set-valued function:

$$(\partial\mathcal{J}(u))_i = \begin{cases} 1 & \text{for } u_i > 0 \\ [-1,1] & \text{for } u_i = 0 \\ -1 & \text{for } u_i < 0 \end{cases} \qquad \text{(cf. Figure A.1b)} \tag{3.77}$$

The SC then takes the form

$$\exists\, w: \quad \left|(A^T w)_j\right| \leqslant 1 \quad \forall j \notin I \quad \text{and} \quad A_I^T w = \operatorname{sign}(u_I). \tag{SC}$$

with $|(A^T w)_j| = |w^T A_j|$ we note the close resemblance to (FuB). However, (SC) cannot guarantee the uniqueness of the solution to (L1) unless we modify the inequality to $|(A^T w)_j| < 1$. This variant is then equivalent to (FuB) and was called *strong source condition* (*SSC*) in Moeller (2012), where also further analysis and interpretation of source conditions in the compressed sensing framework can be found.
For the noisy case, if a source condition with source element w such that

$$(\Sigma_\varepsilon^{-1/2} A)^T w = \xi \in \partial\mathcal{J}(u) \tag{3.78}$$

is met, we can get the following error estimates:

$$D_{\mathcal{J}}^{\xi}(\hat{u}_{\text{MAP}}, u) \leqslant \frac{1}{2\lambda}\|\varepsilon\|^2_{\Sigma_\varepsilon^{-1}} + \frac{\lambda}{2}\|w\|_2^2 \tag{3.79}$$

$$\frac{1}{2}\|A\hat{u}_{\text{MAP}} - Au\|^2_{\Sigma_\varepsilon^{-1}} + \lambda D_{\mathcal{J}}^{sym}(\hat{u}_{\text{MAP}}, u) \leqslant \frac{1}{2}\|\varepsilon\|^2_{\Sigma_\varepsilon^{-1}} + \frac{\lambda^2}{2}\|w\|_2^2 \tag{3.80}$$

They are adopted by applying those given in Benning (2011), Burger et al. (2007) to the *whitened forward equation:*

$$\bar{f} = \bar{A}u + \bar{\varepsilon}, \qquad \text{where} \qquad \bar{x} := \Sigma_\varepsilon^{-1/2} x \tag{3.81}$$

Treating sparsity defects requires the introduction of *variational inequalities* as smoothness conditions (Burger et al. 2013).
For EEG/MEG source reconstruction, we will also examine the addition of a positivity constraint to the ℓ_1 prior (cf. Section 3.2.2). The strong source condition for this case is given by:

$$\exists\, w: \quad (A^T w)_j < 1 \quad \forall j \notin I \quad \text{and} \quad A_I^T w = \operatorname{sign}(u_I). \tag{SSC$_+$}$$

3.5.5. Block-Sparsity

As for the prior modeling, we might only be interested in the sparsity of $u^\dagger$ with respect to certain sub-structures. To examine recovery conditions for such situations, we need to introduce some notation: We will only examine partitions of the n components of $u^\dagger$ into N blocks of equal size l, i.e, $n = Nl$. Let $u_{[i]} := (u_{(i-1)l+1}, \dots, u_{il})^T \in \mathbb{R}^l$ be the i-th block vector, $A_{[i]} := \left[A_{(i-1)l+1}, \dots, A_{il}\right] \in \mathbb{R}^{m\times l}$ the corresponding columns of A and $\mathcal{I} := \{i | u_{[i]} \neq 0\}$ as the *block support*. We define $|u|_{[0]} := \operatorname{card}(\mathcal{I})$ as number of non-zero blocks, and $\|u\|_{[1]} := \sum_i^N \left\|u_{[i]}\right\|_2$ as the ℓ_1 norm of the block amplitudes $\left\|u_{[i]}\right\|_2$ (which can be expressed as an $\ell_{2,1}$-matrix norm, cf. (3.15)). We are now interested in recovering $u^\dagger$ being the *unique k-block-sparse solution* to $Au = f$, as

$$\min |u|_{[0]}, \quad s.\,t. \quad Au = f. \tag{BlkL0}$$

Again, we rather compute $\hat{u}_{\text{MAP}}$ for an ℓ_1-block prior and (3.62) which corresponds to solving the convex problem

$$\min \|u\|_{[1]}, \quad s.\,t. \quad Au = f, \tag{BlkL1}$$

All presented recovery conditions can be transformed to the block sparse case. We start with (Coh). The key idea is to replace the absolute value of the scalar product $A_i^T A_j$ by the spectral norm of $A_{[i]}^{\sharp\,T} A_{[j]}^\sharp$ to define the *block coherence*

$$\mu_{blk}(A^\sharp) := \max_{i\neq j} \frac{\rho\left(A_{[i]}^{\sharp\,T} A_{[j]}^\sharp\right)}{l}. \tag{3.82}$$

It measures how well signals generated by linear combinations of the i-th block can be approximated by those of the j-th block. The maximal coherence within the single blocks is measured by the *sub coherence*

$$\mu_{sub}(A^\sharp) := \max_i \mu\left(A_{[i]}^\sharp\right) \tag{3.83}$$

The block recovery condition corresponding to (Coh) is then given by

Theorem 3.8. Let $A^\sharp \in \mathbb{R}^{m\times Nl}$ have ℓ_2 normalized columns. If

$$kl \leqslant \tfrac{1}{2}\left(\mu_{blk}(A^\sharp)^{-1} + l - (l-1)\mu_{sub}(A^\sharp)\mu_{blk}(A^\sharp)^{-1}\right) \tag{BlkCoh}$$

then every k-block sparse $u^\dagger$ can be recovered as the unique solution to (BlkL1) with $f = A^\sharp u^\dagger$.

If we define the *restricted block isometry constant* of $A^\sharp$, as the smallest number $\delta_{[k]}$ such that

$$(1-\delta_{[k]})\|u\|_2^2 \leqslant \left\|A^\sharp u\right\|_2^2 \leqslant (1+\delta_{[k]})\|u\|_2^2 \qquad \forall\, u : |u|_{[0]} \leqslant k \tag{3.84}$$

one can show (see Eldar and Mishali 2009):

Theorem 3.9. Let $A^\sharp \in \mathbb{R}^{m\times Nl}$ have ℓ_2 normalized columns. If

$$\delta_{[2k]} \leqslant \sqrt{2}-1, \tag{BlkRIP}$$

then every k-block sparse $u^\dagger$ can be recovered as the unique solution to (BlkL1) with $f = A^\sharp u^\dagger$.

For a $kl \times l$ matrix B build by stacking k blocks $B_1, \ldots, B_k$ of size $l \times l$, we can define

$$\rho_l(B) := \sum_i \rho\left(B_i\right) \tag{3.85}$$

With this, we can extend (Tr) to the block sparse case (adopted from Eldar et al. 2010):

Theorem 3.10. Given $A^\sharp \in \mathbb{R}^{m\times Nl}$, a k-block sparse $u^\dagger$ with block support $\mathcal{I}$ can be recovered as the unique solution to (BlkL1) with $f = A^\sharp u^\dagger$ if $A^\sharp_{[\mathcal{I}]}$ is injective and

$$\rho_l\left(A^{\sharp\,+}_{[\mathcal{I}]} A^\sharp_{[j]}\right) < 1 \quad \forall j \notin \mathcal{I} \tag{BlkTr}$$

For extending (FuA) and (FuB) we use the equivalence to the strong source condition. The subgradient is characterized by

$$\xi \in \partial|u|_{[1]} \iff \begin{cases} \left\|\xi_{[i]}\right\|_2 < 1 & \text{if } i \notin \mathcal{I} \\ \xi_{[i]} \quad = \nabla\left\|u_{[i]}\right\|_2 = \dfrac{u_{[i]}}{\left\|u_{[i]}\right\|_2} & \text{if } i \in \mathcal{I} \end{cases}, \tag{3.86}$$

which means that for a single block $[i]$ with $u_{[i]} \neq 0$ the subgradient $\xi_{[i]}$ is the unit vector describing the slope of the tangent plane to the l-dim. cone given by the epigraph of $\|\cdot\|_2$ in $u_{[i]}$. For $u_{[i]} = 0$, it is the set of the slopes of all planes passing through the tip of cone that stay beneath it (see also Tellen 2013). The source condition $A^T w \in \partial|u|_{[1]}$ can then be formulated as

Theorem 3.11. Given $A \in \mathbb{R}^{m\times Nl}$, a k-block sparse $u^\dagger$ with block support $\mathcal{I}$ can be recovered as the unique solution to (BlkL1) with $f = Au^\dagger$ if a dual vector $w \in \mathbb{R}^m$

exists such that

$$\left\|(A^T w)_{[j]}\right\|_2 < 1 \quad \forall j \notin \mathcal{I} \qquad \text{and} \qquad A_{[\mathcal{I}]}^T w = \xi_{[\mathcal{I}]}, \tag{BlkFuB}$$

where $\xi_{[\mathcal{I}]}$ is defined as in (3.86).

The restriction of w to a particular dual vector gives

Theorem 3.12. Given $A \in \mathbb{R}^{m \times Nl}$, a k-block sparse $u^\dagger$ with block support $\mathcal{I}$ can be recovered as the unique solution to (BlkL1) with $f = Au^\dagger$ if $A_{[\mathcal{I}]}$ is injective and

$$\left\|(A^T w^+)_{[j]}\right\|_2 < 1 \quad \forall j \notin \mathcal{I} \qquad \text{with} \qquad w^+ = (A_{[\mathcal{I}]}^T)^+ \xi_{[\mathcal{I}]}. \tag{BlkFuA}$$

3.5.6. Notes and Comments

The field of compressed sensing was established around 2006 by the works of Candes et al. (2006) and Donoho (2006) who linked various earlier works, ideas and developments and coined the term *compressed sensing*. Although being a relatively young field, there is an extensive amount of publications already. Fortunately, a number of topical reviews appeared lately. Foucart and Rauhut (2013) is an exhaustive reference that was found most helpful for writing this section. Two further reviews used are Fornasier (2010) and Eldar and Kutyniok (2012).

From the Bayesian point of view, the analysis in this section can be seen as an examination of the MAP estimate *conditioned* on a concrete realization $u^\dagger$ of the prior and in the noise-free limit. As noted, one problem of this is that the priors used cannot really represent exactly sparse elements. A further problem is that the recovery conditions are all formulated for $u^\dagger$, not for $u^{\dagger,\infty}$. Unless we commit the inverse crime assumption $u^\dagger = u^{\dagger,\infty}$, i.e., the prior model is defined on the correct space (cf. Section 2.1), the recovery conditions for $u^\dagger$ might be of limited use for $u^{\dagger,\infty}$. This also affects the validity of the noise-free limit: ε does not only consist of a noise term, which might vanish in a best-case scenario, but also of the model error term $Au^\dagger - P\mathcal{A}u^{\dagger,\infty}$, which will not vanish. Again, this underlines that the examination of the recovery conditions presented here should be understood as an assessment of the best performance of the MAP estimate that is theoretically possible, ignoring noise, sparsity defects, model errors and neglecting an inverse crime.

The error estimates for a non-vanishing noise term all relied on the concrete realization of ε or an a-priori bound on it. In the statistical sense, we examined error estimates conditioned on the subset $\{\varepsilon \mid \|\varepsilon\|_2 < \delta\}$. Deriving unconditional error estimates and convergence rates is much more involved. In particular, an appropriate notion of convergence that allows for deriving such rates is required. References are Bissantz et al.

(2004, 2007), Engl et al. (2005), Hofinger (2006), Hofinger and Pikkarainen (2007, 2009), Kekkonen et al. (2014), Mathé and Tautenhahn (2011). For further reading on deterministic convergence rates in Bregman distances we refer to Benning (2011), Burger and Osher (2004), Grasmair (2010), Schuster et al. (2012).
Apart from Moeller (2012), the connection between source conditions and exact recovery conditions was also examined in Benning (2011), Lorenz et al. (2011), Trede (2009). We developed (FuA) from (FuB) by restricting the dual vector w to only one element of the $(m-k)$-dim affine subspace defined by $A_I^T d_I = \operatorname{sign}(u_I^\dagger)$. Of course, a continuum of recovery conditions between (FuA) and (FuB) can be formulated by restricting w to other subsets of the subspace. This could be used to obtain conditions that guarantee further properties of the MAP estimates, for instance demanding $\|w\|_2 < t$ can be used to fix the rates (3.79) and (3.80).

3.6. Selected Advanced Topics

In this section, we discuss some advanced topics in Bayesian inversion that are closely related to the topics examined in this thesis.

3.6.1. Infinite Dimensional Bayesian Inversion

While regularization theory is traditionally formulated and analyzed in an infinite dimensional function space setting (cf. Section 1.2), Bayesian inversion is traditionally formulated as an application of normal, finite dimensional Bayesian inference to discretized inverse problems. In recent years, a lot of research on extending the Bayesian approach to the function space setting has been carried out. The underlying motivations are manifold:

- The properties of high dimensional objects are often different than intuition based on low dimensional illustrations, such as Figures 3.3, 3.5a or 3.13, might predict. For large n, their properties might in fact be closer to those of their infinite dimensional limits. Therefore, one should also study those. Related to this issue is the problem of formulating prior information in a *discretization invariant* way: We usually have a-priori information about the infinite dimensional $u^{\dagger,\infty}$ only, but need to encode them into a prior for the n-dim. $u^\dagger$. As n can typically be chosen freely, we want to construct the prior in such a way that the same a-priori information is expressed for all n and estimates and the posterior converge to well defined limits for $n \to \infty$. The search for discretization invariant non-Gaussian priors is an active field of research initiated by Lassas and Siltanen (2004), who found out that the conventional TV prior cannot be formulated in a discretization

invariant way. We will further examine this phenomena in the computational studies in Chapter 5.

- Apart from the prior, the discretization has also influence on the likelihood, where it manifests in the difference between A and $P\mathcal{A}$. Examining or compensating for (see below) the influence of using coarse discretizations or model reduction techniques requires an infinite dimensional Bayesian model as a reference. Another topic is how estimates and posterior behave in the infinite dimensional limits of measurement projection P and noise model Noi (cf. Section3.1).

- A discretization invariant prior and a consistent approximation of the forward operator are important for designing *multi-level* algorithms. These can accelerate the high-dimensional non-linear computations that some estimates require. Such algorithms rely on a correspondence between the computed estimates for different n. One way to achieve this is to derive all priors and likelihoods from one infinite dimensional model in a consistent way.

- Related to the last point is the observation that many algorithms that were formulated and developed in a discrete setting work well for small n, but loose efficiency for $n \to \infty$. We will encounter examples of this behavior in Section 5.1.2. One idea to overcome this problem is to design and formulate algorithms in the infinite dimensional setting first and to discretize them in a second step. While this is an established approach for developing optimization techniques for variational regularization problems like (1.5), and, thereby, for MAP estimation, it is less common for developing *sampling schemes* for computing other Bayesian estimators. An example of such an approach can be found in Cotter et al. (2013).

- It would offer the possibility to further investigate the relations between variational regularization and Bayesian techniques.

Bayesian inversion with infinite dimensional unknowns involves several difficulties. A particular challenge is that there is no analogue of the Lebesgue measure on an infinite dimensional Banach space. As noted at the beginning of this chapter, the way we presented Bayesian inversion here crucially relied on the existence of probability densities/*Radon-Nikodym derivatives* with respect to the Lebesgue measure. In infinite dimensions, priors have to be constructed in a different way.

Notes and Comments

In the introduction of Lassas et al. (2009), a detailed review on previous work on the topic is given. Comelli (2011), Hämäläinen et al. (2013), Kolehmainen et al.

(2012), Lassas et al. (2009), Lassas and Siltanen (2004) are of particular importance to this thesis. In Helin (2010a, b), Helin and Lassas (2009), infinite dimensional hierarchical Bayesian models were also investigated as a consequence of the missing discretization invariance of the TV prior. A new line of work initiated by Stuart et al. mainly targets non-linear inverse problems with prior measures that have a density with respect to an infinite dimensional Gaussian measure,. See Stuart (2010) for an introduction and Agapiou et al. (2013), Dashti et al. (2012, 2013), Hairer et al. (2011) for further works.

3.6.2. Bayesian Treatment of Nuisance Parameters

In addition to u, the likelihood may contain further parameters that are uncertain in practical applications but, unlike the unknowns, are not of interest. We call such parameters *nuisance parameters*:

- Most often, the noise model itself is only an approximation (cf. Section 3.1). In addition, its parameters (e.g. Σ_ε^{-1}) have to be estimated in some way and therefore usually carry uncertainty.
- Solving the forward problem usually involves some kind of discretization errors. They can be regarded as nuisance parameters as well.
- In addition, the forward problem typically depends on parameters, for instance PDE coefficients in linear problems (such as the conductivity σ in (2.12)) or sensor parameters (such as the exact location of the electrodes in EEG).

Their choice might be crucial but difficult. The Bayesian approach offers several techniques to mitigate the effects of nuisance parameters. Again, the central step is, to model all uncertain parameters as random variables and to encode any prior knowledge about their values into a probability distribution. Then, they become subject to Bayesian inference as well.

Noise Parameters

As an example, consider a Gaussian noise model with $\Sigma_\varepsilon = \sigma^2 \mathcal{I}_m$, where the variance σ^2 is unknown, but some prior knowledge about its mean and variance is available. Similar as in Section 3.3, a convenient prior model for σ^2 able to encode this information is given by the inverse gamma distribution (3.46) (cf. Figures 3.10a, 3.11):

$$p_{prior}(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \tag{3.87}$$

The joint posterior over u and σ^2 given f is then given by

$$p_{post}(u, \sigma^2|f) \propto p_{like}(f|u, \sigma^2)p_{prior}(\sigma^2)p_{prior}(u) \tag{3.88}$$

We now have different possible ways to account for the uncertainty in σ^2:

- *Marginalization*: If we integrate over σ^2, its uncertainty propagates to the other variables, in this case to u. This is a form of *generalized error propagation*.

- *Model selection*: We can also integrate over u to obtain the posterior distribution $p_{post}(\sigma^2|f)$ and compute a MAP estimate $\hat{\sigma}^2_{\text{MAP}}$ from it. In a second step, we fix $\sigma^2 = \hat{\sigma}^2_{\text{MAP}}$ in the likelihood and carry out inference for u. Practically, this procedure involves the computation of the conditional *model evidence* $p(f|\sigma^2)$:

$$p(\sigma^2|f) \propto p_{prior}(\sigma^2) \int p_{like}(f|u, \sigma^2)p_{prior}(u)\,\mathrm{d}u = p_{prior}(\sigma^2)p(f|\sigma^2) \tag{3.89}$$

  For a fixed σ^2, $p(f|\sigma^2) = p(f)$ is the normalization factor appearing in (3.1) and is of no further importance. For model selection, it becomes the central object. In general, given two models $\mathcal{M}_1$ and $\mathcal{M}_2$, model selection compares the *posterior odds*

$$\frac{p(\mathcal{M}_1|f)}{p(\mathcal{M}_2|f)} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \frac{p(f|\mathcal{M}_1)}{p(f|\mathcal{M}_2)}. \tag{3.90}$$

  The posterior odds is the product of the *prior odds* and the *Bayes factor*, which compares the conditional model evidences $p(f|\mathcal{M}_{1,2})$ (also called *marginalized likelihoods*).

- *Joint inference*: We use the full posterior to jointly estimate u and σ^2. The estimate of σ^2 can be used to calibrate subsequent measurements or as a fixed parameter in subsequent inversions.

Note that while the introduction of a new variable seems similar to hierarchical Bayesian modeling, the difference is that the prior does not depend on σ^2.

Approximation Error Modeling

Let us assume that the forward operator $P\mathcal{A}$ depends on parameters c and that $A[c, n_{dof}](u)$ is the discrete forward operator built using these parameters in a numerical forward computation with n_{dof} degrees of freedom. Furthermore, we denote the real but unknown parameters by $c^\dagger$, and assume that N_{dof} corresponds to a discretization fine enough such that the discretization error is negligible. Then, we can describe the

approximation error η of using $A(c, n_{dof})$ instead of $A(c^\dagger, N_{dof})$ as

$$\begin{aligned} f &= A[c^\dagger, N_{dof}](u) + \varepsilon \\ &= A[c^\dagger, N_{dof}](u) + A[c, n_{dof}](u) - A[c, n_{dof}](u) + \varepsilon \\ &= A[c, n_{dof}](u) + \underbrace{A[c^\dagger, N_{dof}](u) - A[c, n_{dof}](u)}_{:=\eta} + \varepsilon \end{aligned} \tag{3.91}$$

The concrete realization of the approximation error is not known as it depends on the unknown parameters $c^\dagger$, the unknown discretization error, and the solution u itself. In many applications, ignoring the approximation error leads to systematic inversion errors often called *artifacts*. In EMEG for instance, specifying a wrong conductivity σ (cf. (2.12)) can lead to a systematic mislocalization of focal source configurations (see Lanfer et al. 2012). *Approximation error modeling* (*AEM*) is a Bayesian technique that computes η's a-priori statistics by using the prior on u and assuming a prior on c (similar as in the previous section), and incorporates them into the inversion. This way, the approximation error is compensated for, and systematic inversion errors are mitigated. In principle, one could infer a posterior distribution for u from (3.91) by marginalizing c. This is called *complete error model* but is often computationally not feasible. Instead, the *enhanced error model* assumes that η and u are mutually independent and approximates the distribution of η by a Gaussian: $\eta \sim \mathcal{N}(\mu_\eta, \Sigma_\eta)$. The mean μ_η can be computed as

$$\mu_\eta = \int \left(A[c', N_{dof}](u) - A[c, n_{dof}](u)\right) p_{prior}(c') p_{prior}(u) \, \mathrm{d}c' \, \mathrm{d}u, \tag{3.92}$$

the covariance Σ_η accordingly. This way, accounting for the approximation error can be achieved by modifying the measurement noise statistics (which explains the name of the approach):

$$f = A[c, n_{dof}](u) + \bar{\varepsilon}, \qquad \bar{\varepsilon} := \eta + \varepsilon \sim \mathcal{N}(\mu_\eta + \mu_\varepsilon, \Sigma_\varepsilon + \Sigma_\eta) \tag{3.93}$$

Notes and Comments

Model selection can also be used to determine parameters of the prior model, i.e., hyperparameters. This approach is very popular in the machine learning community (see, e.g., Hastie et al. 2009). For the application to EEG/MEG, see Sato et al. (2004) for the choice of the source space model, and Henson et al. (2009a), Strobbe et al. (2014) for the choice of the head model. Model selection is one way to exploit the information given by *model comparison* (3.90). Another possibility is to perform *model averaging*. See Hastie et al. (2009), Toussaint (2011) for general references

and Trujillo-Barreto et al. (2004) for an application to EEG/MEG. Finally, model comparison can be used to give data-based evidence for the advantage of using one forward model instead of others. For instance, see Henson et al. (2010, 2009b) for the validation of the benefits of multimodal integration over single modality-based imaging. General references for approximation error modeling are given by Kaipio and Somersalo (2005, 2007). AEM has been applied to a couple of applications with very promising results. See Nissinen (2011), Nissinen et al. (2008, 2009, 2011) and references therein for *electrical impedance tomography* (*EIT*), Arridge et al. (2006), Heiskala et al. (2012), Kolehmainen et al. (2011), Tarvainen et al. (2010, 2013) for optical and photo-acoustic applications, Lipponen et al. (2013) for cloud modeling and Cui et al. (2011) for geothermal reservoir modeling.
In our presentation of AEM, we omitted the technical difficulty that the discretization of the unknowns to $\mathbb{R}^n$ might produce an approximation error and that in some inverse problems (including EIT), the discretization of the unknowns is coupled to n_{dof}. To account and compensate for this requires a discretization invariant formulation of the prior (cf. Section 3.6.1). For the inverse problems scenarios we consider, our implementations of the forward mapping do not depend on n. Therefore, we do not face the coupling problem.

3.7. Notes and Comments

Bayesian inference is often treated as a straightforward extension of statistical inference or only used to motivate the choice of a particular variational regularization scheme (1.5). However, both approaches obscure its main potential which comes from the radically different concept of probability employed in its reasoning. Jaynes and Bretthorst (2003) is an exceptional reference for a deeper introduction into this topic. For the work and conception of this thesis, Kaipio and Somersalo (2005) was the main inspiration. Further notable references for Bayesian inversion include Stuart (2010), Tarantola (2005) while Gelman et al. (2003), Kay (1993) provide a broader overview on Bayesian inference beyond inverse problems.

4 Bayesian Computation

In this chapter, we will present and develop computational methods required for applying Bayesian inversion. In particular, sparse, high-dimensional imaging scenarios pose specific challenges not encountered in other areas of Bayesian inference. Although some of the algorithms presented here are applicable to arbitrary noise, forward and prior models, the presentation is tailored towards linear inverse problems with Gaussian noise and the prior models described in the last chapter.
In general, the estimators introduced in Section 3.4.1 either rely on optimization or integration tasks, or even a mix of both. For instance, computing the narrowest credible interval containing the probability mass q requires solving

$$\hat{I}_{cr} = \underset{[a,b],b>a}{\operatorname{argmin}}(b-a) \qquad s.\,t. \quad \int_a^b p^i_{post}(u_i|f)\,\mathrm{d}u_i = q \tag{4.1}$$

In the first section of this chapter, we will examine *posterior sampling* methods that allow to integrate the posterior by *Monte Carlo* integration. A lot of the work for this thesis was devoted to developing fast sampling schemes that can be applied in the typical scenarios we examine, especially in the experimental data scenarios. Therefore, this section is more detailed than the following Section 4.2 about optimization methods. In Section 4.3, we will examine the similarities between sampling and optimization methods, and illustrate how techniques developed to accelerate optimization methods can also be used to speed up sampling schemes. Finally, Section 4.4 will discuss computational schemes to verify the recovery conditions presented in Section 3.5.
A particular challenge for the implementation is that both forward operator A and prior operator D may not be available in an explicit form, i.e., as a matrix. Instead, we need to use *matrix-free* algorithms that only involve multiplications with A or D (or their transposes) with a vector. To keep the presentation concise, details of the

implementation have been moved to Section A.2 in the appendix. However, we stress here that they often constitute the most challenging and tedious works for this thesis.

Whitening All algorithms presented in this chapter are formulated (and implemented) for $\varepsilon \sim \mathcal{N}(0, I_m)$. This can be achieved by a *pre-whitening/decorrelation* of the forward equation (1.4):

$$\Sigma_\varepsilon^{-1/2} f = \Sigma_\varepsilon^{-1/2} A + \Sigma_\varepsilon^{-1/2} \varepsilon. \tag{4.2}$$

If $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$, then $\Sigma_\varepsilon^{-1/2}\varepsilon \sim \mathcal{N}(0, I_m)$. Therefore, one can use $\Sigma_\varepsilon^{-1/2} f$ and $\Sigma_\varepsilon^{-1/2} A$ instead of f and A in the algorithms.

4.1. Posterior Sampling Methods

This section will develop methods to draw random samples u^i from the posterior $p_{post}(u|f)$. Such samples can be used to compute integrals over the posterior by *Monte Carlo* (*MC*) integration.

4.1.1. Monte Carlo Integration

Suppose we want to compute integrals like

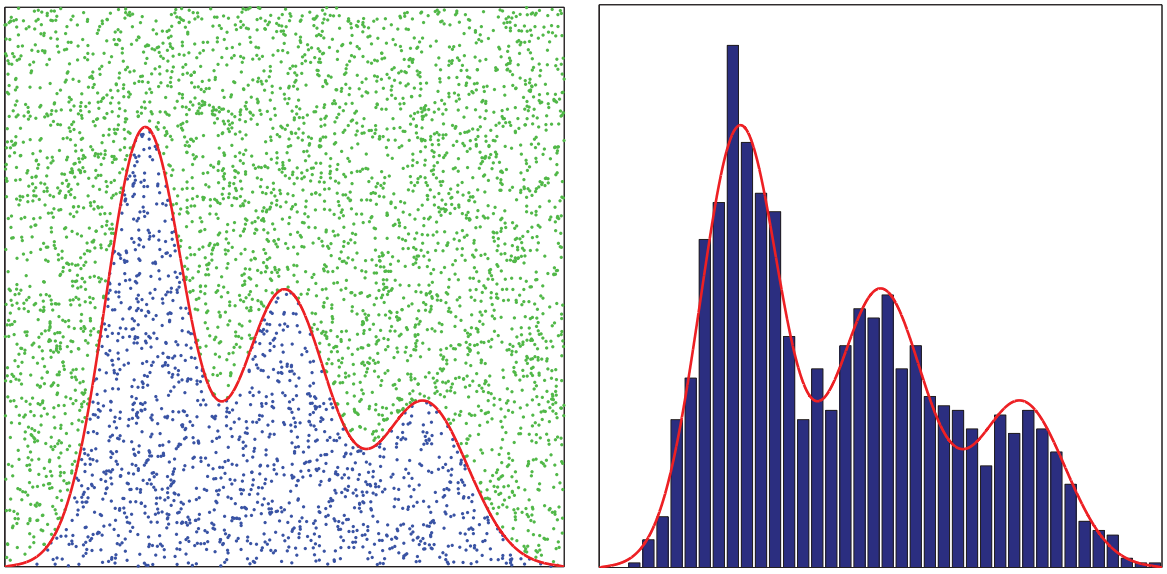
$$\mathbb{E}\left[g(x)\right] = \int g(x)p(x)\,\mathrm{d}x, \tag{4.3}$$

where $p(x)$ is a probability density, and $g \in L^1(\mathbb{R}^n)$ is a feature of interest. For $n = 1$ and a given number of function evaluations K, traditional *Gauss-type quadratures* would first compute a suitable grid $\{x^i\}$, $i = 1, \ldots, K$, and corresponding weights ω^i, and then approximate

$$\int g(x)p(x)\,\mathrm{d}x \approx \sum_i^K \omega^i g(x^i). \tag{4.4}$$

The grid points and the weights are only determined by $p(x)$ and are typically chosen in such a way that (4.4) is exact for polynomials g up to a high degree. Unfortunately, such a procedure is infeasible in high dimensions: Extending a K-point rule to $\mathbb{R}^n$ requires K^n integration points and computing these points and the weights requires a good knowledge of $p(x)$. In our case, $p(x)$ is the Bayesian solution to the inverse problem: It is exactly what we do not know well.

Intuitively, if we cannot compute weights, our computational grid $\{x^i\}$ should be dense where $p(x)$ is relatively large and thus, a large contribution to (4.3) is to be expected. The idea of Monte Carlo integration is that such a grid is automatically generated if



(a) (b)

Figure 4.1.: An illustration of accept-reject methods. (a) To sample from the density $p(x)$ (red line), uniform samples (x_i, y_i) (blue and green dots) are generated in a region enclosing its graph. All samples fulfilling $y_i \leqslant p(x_i)$ (blue dots) are accepted. (b) Histogram computed from the x values of all accepted samples.

we chose x^i to be i.i.d. samples of $p(x)$. The *law of large numbers* guarantees that this stochastic strategy works:

$$\lim_{K\to\infty} \frac{1}{K} \sum_{i}^{K} g(x^i) = \mathbb{E}\left[g(x)\right] = \int g(x)p(x)\,\mathrm{d}x, \qquad a.s., \tag{4.5}$$

with a rate $\mathcal{O}(1/\sqrt{K})$, which is independent of n (Klenke 2008). While the rate of convergence is poor compared to deterministic approaches, the independence of n is a striking advantage of MC integration. To implement MC integration, we need to be able to produce random realizations of $p(x)$. Algorithms for this purpose are called *samplers* or *sampling/simulation schemes*. In the next section, we will introduce some basic samplers.

4.1.2. Direct Sampling Methods

All computational sampling methods rely on a *pseudorandom number generator* (*PRNG*), which produces a sequence of numbers that cannot be distinguished from real random numbers. Formally, one usually demands that the numbers pass a certain hypothesis test for being i.i.d. samples of the underlying distribution. The most widely used PRNG is the *Mersenne twister* (Matsumoto and Nishimura 1998), which is also

used in MATLAB (version R2014a). It generates integers in a certain range, which can be re-scaled to provide uniformly distributed random numbers in $[0, 1]$: $r^i \sim \text{unif}(0, 1)$. Uniform samples are the basis of all other samplers. For instance, the most direct way to utilize them is given by the *inverse cumulative distribution* (*icd*) method: Let

$$F(y) = \int_{-\infty}^{y} p(x)\,\mathrm{d}x \tag{4.6}$$

be the *cumulative distribution function* (*cdf*) of $p(x)$ and $F^{-1}(r) : [0, 1] \to \mathbb{R}$ its (generalized) inverse. If $r \sim \text{unif}(0, 1)$ then $F^{-1}(r) \sim p(x)$. For example, consider the exponential distribution $p(x) = \lambda \exp(-\lambda x)$ on $\mathbb{R}_+$. $F(y)$ is given by $1 - \exp(-\lambda y)$ and $F^{-1}(r) = -\log(1 - r)/\lambda$. However, $F^{-1}(r)$ is often not available in a closed form or its numerical evaluation or approximation is too expensive or unstable. The idea of *transformation methods* is to transform uniform samples such that the distribution of result is $p(x)$. One example is the *Box-Muller* transform (BOX AND MULLER 1958), which transforms two unif$(0, 1)$ numbers r and r' into two independent standard normal distributed numbers z and z' by

$$z = \sqrt{-2\log(r)}\cos(2\pi r'), \qquad z' = \sqrt{-2\log(r)}\sin(2\pi r'). \tag{4.7}$$

We can then generate $x \sim \mathcal{N}(\mu, \sigma^2)$ from the above by $x := \sigma z + \mu$. Another example is to generate a random sign by $\text{sign}(r - 0.5)$. With these ingredients, we can already construct a sampler for an ℓ_1 prior with $D = I_n$:

$$u_i = \text{sign}(r_i - 0.5)\,\frac{\log(1 - r_i')}{\lambda}, \qquad i = 1 \ldots, n, \quad r_i, r_i' \sim \text{unif}(0, 1) \tag{4.8}$$

The random draws from (3.25) in Figures 3.6a and 3.6b were also generated by such a scheme.

A more general class of sampling methods relies on a simple observation: Sampling from a distribution $p(x)$ is equivalent to sampling uniformly from the area under the graph of $p(x)$: $\mathcal{G}_p := \{(x, z) | 0 \leqslant z \leqslant p(x)\}$. This finding, formalized as the *Fundamental Theorem of Simulation*, is the basis for *accept-reject methods*, which draw uniform samples (x, z) from a region enclosing $\mathcal{G}_p$ and only accept the sample if it fulfills $z \leqslant p(x)$. Figure 4.1 shows an illustration of this principle. For instance, the method we use to sample (inverse) gamma distributions is a combination of a transformation of a uniform and a standard normal random number with an accept-reject step (MARSAGLIA AND TSANG 2000).

Unless the single components are mutually independent, sampling multivariate random variables is considerably more difficult. One exception is the multivariate normal

distribution: If $x \sim \mathcal{N}(\mu, \Sigma)$, then $Ax + b \sim \mathcal{N}(b + A\mu, A\Sigma A^T)$. Converse, if $AA^T = \Sigma$ (e.g., by *Cholesky decomposition*) and $x \sim \mathcal{N}(0, I_n)$, we can generate $y \sim \mathcal{N}(\mu, \Sigma)$ by setting $y = Ax + \mu$. Note that the computation can also be performed by solving a linear system. For example, consider sampling from a general ℓ_2 prior $p_{prior}(u) \propto \exp(-\lambda \left\| D^T u \right\|_2^2)$: In principle, the covariance of the prior is given as $\Sigma_u = (2\lambda DD^T)^{-1}$, and one could compute and decompose it to draw a sample v. However, often it is preferable to generate $x \sim \mathcal{N}(0, I_h)$ and let v be the least squares solution to the linear equation $\sqrt{2\lambda} D^T v = x$, given by $v = (2\lambda DD^T)^{-1}\sqrt{2\lambda} Dx$. This way, v also has the covariance matrix

$$\left((2\lambda DD^T)^{-1}\sqrt{2\lambda} D\right)\left((2\lambda DD^T)^{-1}\sqrt{2\lambda} D\right)^T = (2\lambda DD^T)^{-1}(2\lambda DD^T)(2\lambda DD^T)^{-1} = (2\lambda DD^T)^{-1} \quad (4.9)$$

Likewise, one can draw samples from the posterior that results from using such an ℓ_2 prior. It is a Gaussian with mean and covariance given by

$$\mathbb{E}[u|f] = \left(2\lambda DD^T + A^T A\right)^{-1} A^T f \quad (4.10)$$

$$\mathbb{C}\mathrm{ov}[u|f] = \left(2\lambda DD^T + A^T A\right)^{-1} \quad (4.11)$$

(see Section A.1.4 in Lucka 2011 or Section 3.4 in Kaipio and Somersalo 2005). One can sample from it by generating $x \sim \mathcal{N}(0, I_{m+h})$ and solving

$$\begin{bmatrix} A \\ \sqrt{2\lambda} D^T \end{bmatrix} v \overset{ls}{=} \begin{bmatrix} f \\ 0 \end{bmatrix} + x \quad (4.12)$$

in a least-squares sense.

4.1.3. Markov Chain Monte Carlo Methods

Often, direct samplers generating *independent samples* from $p(x)$ are not known. However, the *strong ergodic theorem* ensures that Monte Carlo integration (4.5) still converges if the sequence $\{x^i\}$ is dependent, but originates from an *ergodic Markov chain* that has $p(x)$ as its *equilibrium distribution*. A proper introduction of ergodicity theory and Markov chains that have an infinite dimensional state space is rather technical and is omitted here. We will give an informal introduction and refer to Klenke (2008), Liu (2008), Robert and Casella (2005) for further reading.

An illustrative example of a Markov chain is a Gaussian random walk:

$$x^{i+1} = x^i + \varepsilon^{i+1}, \quad i \in \mathbb{N}_0 \quad x_0 = 0; \quad \varepsilon^{i+1} \sim \mathcal{N}(0, \sigma_i^2) \quad (4.13)$$

Every x^i is the realization of a random variable X^i. Such an ordered series of random variables $\{X^i\}_{i=0}^{\infty}$ is called *stochastic process*; the discrete index i stands for a time step. Conditioned on all *past states* $X^j = x^j$, $j = 0, \ldots, i$, the distribution of X^{i+1} is given as $\mathcal{N}(x^i, \sigma_i^2)$. Therefore, it only depends on the *current state* x^i:

$$\mathbb{P}\left(X^{i+1} = x | X^1 = x^1, \ldots, X^i = x^i\right) = \mathbb{P}\left(X^{i+1} = x | X^i = x^i\right) \tag{4.14}$$

This property is called *Markov property*. In common language, one might put it as "the future depends on the past only through the present", or "the present already contains all past information about the future". One can also say that the process has "no memory". A Markov chain is a stochastic process that possess the Markov property. A *time-homogeneous* Markov chain is further characterized by

$$\mathbb{P}\left(X^{i+1} = y | X^i = x\right) = \mathbb{P}\left(X^{i+k+1} = y | X^{i+k} = x\right), \qquad \forall\, k \in \mathbb{N}, \tag{4.15}$$

which means that the *transition probability distribution* is stationary in time. In the random walk (4.13), this would correspond to $\sigma_i^2 = \sigma^2\ \forall\, i$. Markov chains can be characterized by a *transition kernel* $T^i(x, B)$, which generates the i-th step:

$$T^i(x, B) = \mathbb{P}\left(X^{i+1} \in B | X^i = x\right) \tag{4.16}$$

In the random walk (4.13), $T^i(x, B)$ can be described by the *transition density* $t^i(x, y) = \mathcal{N}(y; x, \sigma_i^2)$. A chain is time-homogeneous if $T^i(x, B) = T(x, B)\ \forall\, i$. In this case, the following properties of the transition kernel are important to determine the behavior of the chain:

- Assume that the realization of the current state X^i can be described by the probability measure μ^i. Using the transition kernel, we can define a linear propagation operator $\mathcal{T}$ mapping μ^i to the probability measure μ^{i+1} of X^{i+1}:

$$\mu^{i+1}(B) = \mathcal{T}[\mu^i](B) := \int T(x, B)\, \mu^i(\mathrm{d}x) \tag{4.17}$$

  In our random walk example (4.13), assume that the realization $X^i = x$ can be described by the density $p^i(x)$. Then, $\mathcal{T}$ generates the density of X^{i+1} by a convolution with the kernel $\mathcal{N}(0, \sigma^2)$:

$$p^{i+1}(y) = \int \mathcal{N}(y; x, \sigma^2)\, p^i(x)\, \mathrm{d}x = \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) p^i(x)\, \mathrm{d}x = \mathcal{N}(0, \sigma^2) * p^i \tag{4.18}$$

Eigenfunctions of $\mathcal{T}$ are called *invariant/stationary/equilibrium* distributions of the kernel T. The random walk obviously does not have such a distribution.

- A kernel is *irreducible* if there is a positive probability that, regardless of the starting point x_0, every set with positive measure is reached after a finite number of steps.

- A kernel is *aperiodic* if the probability that the chain gets trapped in a periodic loop is zero.

Although the formal definition of these properties for Markov chains with infinite dimensional state space is rather involved, and further technical conditions have to be met, they provide the conceptual basis on which the *ergodicity* of the Markov chain can be established: Independent of the starting point x_0, the distribution of X^i converges to the equilibrium distribution μ of the Markov chain and the time average $\frac{1}{K}\sum_i^K g(x^i)$ (*ergodic average*) of a function $g(x)$ converges to the space-average $\int g(x)\mu(\mathrm{d}x)$.
While Markov chain theory is devoted to analyzing a given Markov chain to find its equilibrium distribution, *Markov chain Monte Carlo* (*MCMC*) theory tries to construct a Markov chain such that it has a desired equilibrium distribution. If we are given a density $p(x)$ to sample from, this is the MCMC strategy:

Algorithm 4.1. Markov chain Monte Carlo

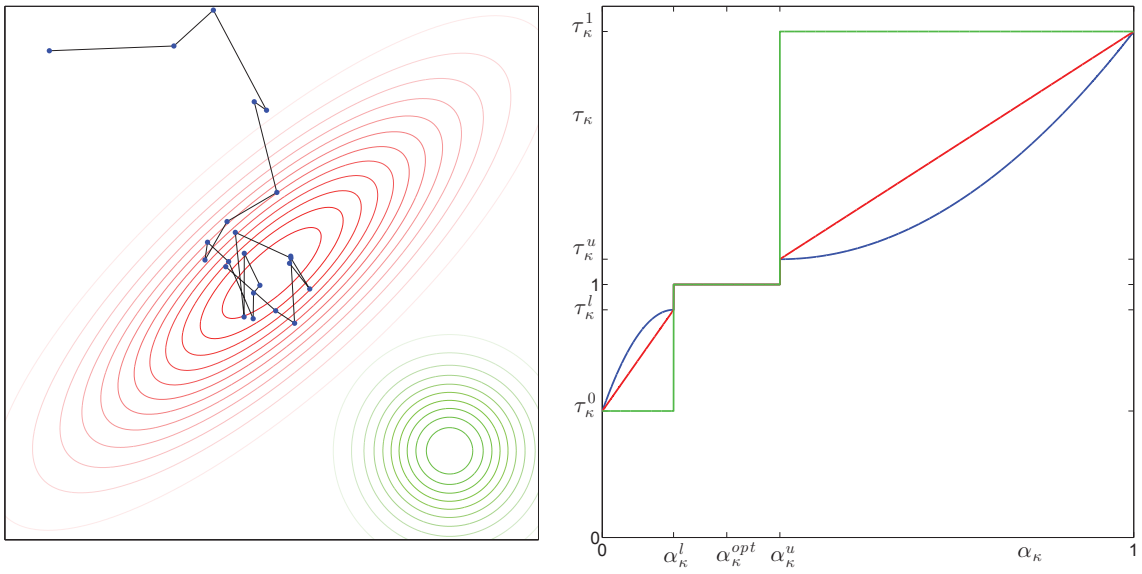
1. Construct an ergodic transition kernel T such that $p(x)$ is the density of its equilibrium distribution.

2. Choose an initial state x_0 and define a *burn-in time* K_0.

3. Simulate $x^{i+1} \sim T(x^i, \cdot)$ for $i = 0, \ldots, K_0 + K - 1$.

4. Discard $\{x^i\}_{i=0}^{K_0}$ and use $\{x^i\}_{i=K_0+1}^{K_0+K}$ as a sample of $p(x)$.

One realization of this procedure is called a *run*. The time steps $i = 1, \ldots, K_0$ are called *burn-in phase*. There are generic schemes to construct an ergodic transition kernel for every given $p(x)$. The problem is rather to construct a kernel that is *efficient*. We will return to this point after introducing the two basic schemes on which most MCMC methods rely. Often, the chain $\{x^i\}$ is thinned by a certain *sub-sampling rate* (SSR) to decrease the statistical dependence between subsequent samples and to safe memory. Theoretically, the sub-sampled chain corresponds to a Markov chain whose transition kernel is given as the SSR-fold convolution of the single-step transition kernel of the original chain.



(a) (b)

Figure 4.2.: (a) Example of an MCMC run by a Metropolis-Hastings sampler. Red lines: Level-sets of $p(x)$. Green lines: Level-sets of the Gaussian proposal kernel. (b) Examples of scaling functions used in the automatic κ-adaptation scheme.

4.1.4. Metropolis Hastings Sampling

The *Metropolis-Hastings* (*MH*) method (Hastings 1970, Metropolis et al. 1953) is a very simple construction of the transition kernel:

Algorithm 4.2. (Metropolis-Hastings Sampling)
Let $q(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be a function satisfying $\int q(x, y)\, \mathrm{d}y = 1$ for all $x \in \mathbb{R}^n$ (proposal distribution). Step 3. in Algorithm 4.1 is implemented as:

3.1. Draw y from the proposal distribution $q(x^i, y)$.

3.2. Compute the *acceptance ratio*

$$\mathrm{acc}(x^i, y) = \frac{p(y)\, q(y, x^i)}{p(x^i)\, q(x^i, y)}. \tag{4.19}$$

3.3. Draw $r \sim \mathrm{unif}(0, 1)$.

3.4. If $\mathrm{acc}(x^i, y) \geqslant r$, set $x^{i+1} = y$, else set $x^{i+1} = x^i$.

Note that the requirements on $p(x)$ for this scheme are minimal: As only ratios of probabilities are used, we only have to know $p(x)$ up to a scaling factor. Furthermore, we only need to be able to evaluate $p(x)$ for any given x. Each sampling step requires one such evaluation (in inverse problems, the computational demanding part of this evaluation is usually applying the forward mapping A). In this respect, MH can be

considered a "black-box sampler", which explains its success in many different application areas (Liu 2008). However, the whole difficulty in MH is shifted to the construction of the proposal function $q(x, y)$: While the scheme works for a lot of proposal distributions in theory, its application is only feasible if $q(x, y)$ leads to a chain that moves "fast". This way, all the important regions of the sampling space are explored in a reasonable amount of computational time and consecutive samples are as uncorrelated as possible. It is easy to see that these demands lead to a dilemma: The optimal proposal function is given by $q(x, y) = p(y)$, which turns the MH scheme into a direct sampler for $p(x)$. However, if we would know a direct sampler for $p(x)$, we would not consider performing MCMC in the first place. As a rule of thump, the more information about $p(x)$ is incorporated into the design of $q(x, y)$, the better. This is an obvious contradiction to the "black-box" character of MH. As a consequence of this dilemma, a huge number of different MH-based schemes exist for various types of $p(x)$ that try to improve upon the basic approaches that construct $q(x, y)$ without any reference to $p(x)$. However, we will only consider these basic, but most commonly applied MH schemes in this thesis: The *symmetric random-walk* MH schemes (*SRWMH*) which generate y by

$$y = x + \vartheta, \qquad \mathbb{E}[\vartheta] = 0, \qquad \vartheta \sim \tilde{q}(\|\vartheta\|_2), \tag{4.20}$$

for a suitable probability distribution $\tilde{q}$ on $\mathbb{R}_+$. This means that a new proposal is generated by perturbing the current state x in a random, unbiased, symmetric way. Therefore, $q(x, y) \propto \tilde{q}(\|x - y\|_2)$, and q vanishes from the acceptance ratio (4.19). The two models for ϑ we will mainly use are:

1. *MH-Iso*: All components of x are updated: $\vartheta_i \sim \mathcal{N}(0, \kappa^2)$, $\forall\, i$.
2. *MH-Si*: One component i_* of x is randomly chosen and updated while all other components remain unchanged: $\vartheta_{i_*} \sim \mathcal{N}(0, \kappa^2)$, $\vartheta_{-i_*} = 0$.

Here, ϑ_{-i} denotes all components of ϑ except the i-th one. See Figure 4.2a for an illustration of an MH-Iso chain for a 2D Gaussian distribution $p(x)$.
For the SRWMH schemes, the proper choice of κ is essential. If it is very small, the proposals will always be accepted since $p(x)$ is usually continuous. In return, the exploration of the sampling space takes a long time. On the contrary, if κ is too large, the differences in probability may be huge because the tails of the Gaussian likelihood decay very fast. Consequently, new proposals will hardly be accepted. A good overview on this topic is given in Neal and Roberts (2006), Roberts and Rosenthal (2001). The remarkable result is that in high dimensions, having a total *acceptance rate* α_κ of new proposals of about 0.234 leads to an optimal efficiency, independent of the distribution to sample from (we will discuss what efficiency for MCMC sampling means

in more detail in Section 4.1.6). Furthermore, this optimal efficiency hardly drops in the range $\alpha_\kappa \in [0.1, 0.4]$. This yields an easy-to-implement rule to tune κ: One could find a κ that leads to such an α_κ rate in a preliminary run and initialize the real run with it. However, it turns out that this κ is typically only optimal once the chain has left the burn-in phase and reached the main support of the distribution: It can hinder the chain from ever getting there if used right from the start. For these reasons, online adaptation of κ is usually used. The empirical acceptance rate α_κ is monitored, and κ is increased if it is higher than some α_κ^u and decreased if it is lower than some α_κ^l. In theory, the resulting chain is not a Markov chain anymore (but it is still ergodic). Nevertheless, in practice, a large enough interval $[\alpha_\kappa^l, \alpha_\kappa^u] \subset [0.1, 0.4]$ can be chosen such that κ hardly ever changes once the burn-in phase is over. Thereby, the real chain is not affected by the adaptation. The concrete adaptation scheme we use is that for every 1000 samples, κ is multiplied with a scaling function $\tau_\kappa(\alpha_\kappa)$. This function is monotonically increasing, 1 in $[\alpha_\kappa^l, \alpha_\kappa^u]$ and a shifted monomial with a little offset in the other intervals. Furthermore, we demand that $\tau_\kappa(0) = \tau_\kappa^0 > 0$, $\tau_\kappa(1) = \tau_\kappa^1 > 1$. Figure 4.2b shows examples of such functions. Every time κ changes (i.e., $\tau_\kappa \neq 1$), the acceptance count is reset.

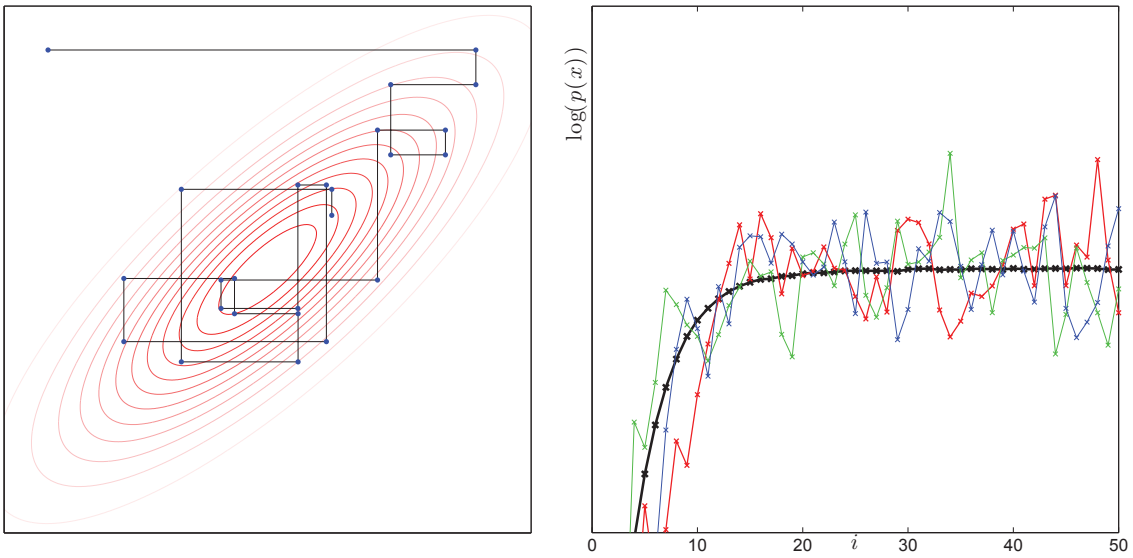
Notes and Comments

While the numerical implementation of basic SRWMH scheme is trivial, the Metropolis-Hastings proposal acceptance-rejection scheme is the basis for very sophisticated algorithms:

- *Adaptive MH* methods adapt the proposal distribution based on the sampling history, for instance by using scaled versions of the sample covariance matrix as a distribution on ϑ in SRWMH schemes. Obviously, this destroys the Markov property of the resulting stochastic process, but under some conditions, the process remains ergodic. The introduction of Latuszynski et al. (2013) provides a recent literature overview while more specific aspects can be found in Andrieu and Thoms (2008), Haario et al. (2004, 2001, 2005), Roberts and Rosenthal (2009). Note that this is a more radical and explicit online adaptation than the tuning scheme for κ that we constructed with the explicit intention that it is no longer active in the main phase of the run.

- *Delayed Rejection MH* methods design proposal distributions that can locally adapt to the target distribution (unlike Adaptive MH, which globally adapts the proposal). See Haario et al. (2006), Mira (2001).

- *Delayed Accepetance MH* was developed for large-scale non-linear inverse problems



(a) (b)

Figure 4.3.: (a) Example of an MCMC run by a SC Gibbs sampler. Red lines: Level-sets of $p(x)$. (b) Plots of $\log(p_{post}(u^i|f))$ using the RSG sampler in the Boxcar scenario with $n = 1023$ for a TV prior with $\lambda = 800$. Red, green and blue plots: Three independent realizations. Black plot: The average of 5000 independent realizations.

that come with a high computational cost for evaluating the forward model. New proposals are first "tested" with a reduced forward model. Only accepted proposals are then evaluated with the full model. See Christen and Fox (2005), Cui et al. (2011).

4.1.5. Gibbs Sampling

The basic idea of *Gibbs sampling* (Geman and Geman 1984) is to construct the transition kernel directly from conditioned, lower dimensional versions of $p(x)$. One often encounters a partition of the n components of x into blocks $I_1, \dots, I_N \subset \{1, \dots, n\}$ such that fast samplers for the conditional density of x_{I_j} given $x_{[-I_j]}$ are available. For instance, consider the posterior $p_{post}(u, \gamma|f)$ of the conditionally ℓ_2 hypermodel (3.41) with inverse gamma hyperprior (3.46):

$$p_{post}(u, \gamma|f) \propto \exp\left(-\frac{1}{2}\|f - Au\|^2_{\Sigma_\varepsilon^{-1}} - \sum_i^h \frac{(d_i^T u)^2 + \beta}{\gamma_i} - (\alpha + 3/2)\log(\gamma_i)\right) \quad (4.21)$$

By construction, $x = (u, \gamma)$ is a partition such that $p_{post}(u|\gamma, f)$ is a multivariate Gaussian distribution and $p_{post}(\gamma|u, f) = p_{post}(\gamma|u)$ is a product of inverse gamma distributions (see (3.47)). For both distributions, fast direct samplers are known (see Section 4.1.2). A Gibbs sampler alternates between updating one of them through an

explicit sampler while keeping the other fixed. This can be generalized:

Algorithm 4.3. (Gibbs Sampling)
Let $[1], [2], \ldots, [N]$ denote a partition of $\{1, \ldots, n\}$, and $\mathrm{Ind}_N : \mathbb{N} \longrightarrow \{1, \ldots, N\}$ a *block index choice function*. Step 3. in Algorithm 4.1 is implemented as:

3.1. Choose a block index $j = \mathrm{Ind}_N(i)$.

3.2. Draw $y \sim p\left(x_{[j]} \,\middle|\, x^i_{[-j]}\right)$.

3.3. Set $x^{i+1}_{[j]} = y$, and $x^{i+1}_{[-j]} = x^i_{[-j]}$.

The most basic variant of this scheme (often simply called *the* Gibbs sampler) is the *single component (SC) Gibbs sampler*: $[j] = \{j\}$. Starting at $i = 1$, every N subsequent steps of the sampler are called a *sweep*. Ind_N determines, in which order the blocs are updated. We will use two variants in this thesis:

1. *Systematic scan Gibbs* (*SSG*): A fixed order is repeated over and over again, for instance, $\mathrm{Ind}_N(i) = \mod(i, N) + 1$.

2. *Random scan Gibbs* (*RSG*): In each step, a block is chosen uniformly at random.

A Gibbs sampler is determined by the blocking and the updating scheme. The blocking scheme is usually more or less predetermined by $p(x)$ and the choice of the updating scheme is not too important for performance of the sampler (RSG always works, but occasionally, other schemes are a little faster). In particular, there are no parameters like κ in MH whose tuning is essential for obtaining an acceptable performance. Therefore, it is often said that "the Gibbs sampler automatically adapts to $p(x)$". The drawback of Gibbs sampling is that it cannot be considered a "black-box" sampler like MH: An efficient implementation of a Gibbs sampler needs to

(a) compute the conditional, low dimensional densities in an explicit, parameterized form in a fast way.

(b) employ a fast, robust and exact sampling scheme for the parameterized form of the low dimensional densities.

Especially point (a) rules out Gibbs sampling for Bayesian inversion if the posterior has a complicated structure. Non-Gaussian noise models, non-linear forward operators and non-Gaussian priors are such complications. For these reasons, SRWMH schemes are commonly used in such situations. A main contribution of this thesis was to develop SC Gibbs samplers for various non-Gaussian priors. Section 4.1.7 covers point (a) for the likelihood and prior models we use. The sampling in point (b) then needs a 1D sampler. Direct methods will be developed for some priors in Section 4.1.8, but cannot be derived for all priors. In Section 4.1.9, we will introduce a fast and flexible 1D sampling scheme

called *slice sampling*, which we will implement for the remaining non-hierarchical prior models in Section 4.1.10. In Section 4.1.11, we will construct Gibbs samplers for ℓ_p hypermodels based on the samplers for ℓ_p priors developed before.

Notes and Comments

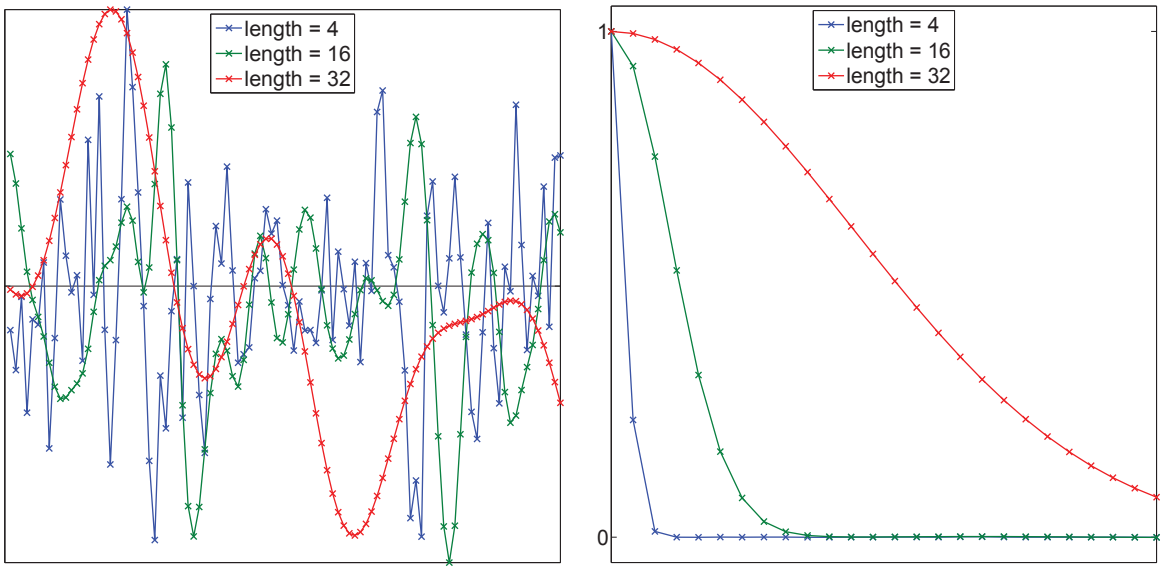
Note that the main idea of Gibbs sampling is to move the current state along certain low dimensional affine subspaces: For instance in SC Gibbs, the state moves along the line $x^i + te_j$ with $j = \mathrm{Ind}_N(i)$. However, these affine subspaces do not need to be aligned to the coordinate axes. One could also move along $x^i + tv$, by drawing t from the 1D density $p_v(t) \propto p(x^i + tv)$. This corresponds to conditioning $p(x)$ *orthogonal* to v. A move along $v = e_1 + e_2$ amounts to updating x_1 and x_2 simultaneously by the same amount: We collapsed two dimensions into one by this move. This observation is the basis of *multigrid Monte Carlo* techniques that sample along directions indicated by a *restriction operator* similar to those used in multigrid methods in numerical analysis (see Goodman and Sokal 1989, Liu 2008, Liu and Sabatti 2000). A further generalization is to allow more general moves in $\mathbb{R}^n$ than only linear shifts, see Liu and Sabatti (2000).

Parallel to the development of adaptive MH samplers, *adaptive Gibbs samplers* were proposed. There are two principled options for adapting the Gibbs sampler: The first is to adjust the directions v in which the updates are performed, while the other is to adjust the selection function Ind_N. A recent overview of such techniques can be found in Latuszynski et al. (2013), we will further discuss this issue in Chapter 7.

4.1.6. MCMC Convergence Diagnostics

Defining and assessing the efficiency of a sampling algorithm for a general purpose rather than a specific aim is a difficult task (Liu 2008). Two types of *convergence diagnostics* are usually applied: *Qualitative diagnostics* rely on the visual inspection of some property of the chain $\{x^i\}$. In contrast, *quantitative diagnostics* try to compute characteristics that can be used to tune the sampler in an automated fashion. This should allow unexperienced users to perform "black box" Bayesian inference. Despite a lot of research on theses topics (Brooks and Roberts 1998, Cowles and Carlin 1996, Roberts and Sahu 1997, Thompson 2010), no universal method is known. The approach we take here is more or less the current gold standard in the field.

Heuristically, it is easy to identify two key ingredients of a good MCMC algorithm: $\{x^i\}$ should be as close as possible to independent samples of $p(x)$. Therefore, it should have a short burn in time (otherwise, the samples do not come from $p(x)$) and subsequent samples should become uncorrelated as fast as possible.



(a) (b)

Figure 4.4.: (a) Three stochastic processes and (b) their autocorrelation functions.

Burn-in analysis The sufficient length of the burn-in phase, K_0, can be deduced from observing $\log(p(x^i))$. Once it starts oscillating around a constant value, the distribution of x^i is close enough to the equilibrium distribution and the chain reached the *stationary phase*. Averaging $\log(p(x))$ over a large number of independent chains that all started at the same initialization removes the oscillations and allows to determine K_0 in an easy fashion. See Figure 4.3b for an example of such a plot.

Autocorrelation analysis To measure the average correlation between subsequent samples, we choose a test function $g : \mathbb{R}^n \to \mathbb{R}$ and estimate the *autocorrelation function* (*acf*) $R : \mathbb{N} \longrightarrow [-1, 1]$ of the series $g^i := g(x^i)$:

$$R(\tau) := \frac{\langle (g^i - \mu)\,(g^{i+\tau} - \mu) \rangle_i}{\varrho^2}, \quad \mu := \langle g^i \rangle_i, \quad \varrho := \langle (gi - \mu)^2 \rangle_i \tag{4.22}$$

Here, $\langle \cdot \rangle_i$ denotes the time-average over the infinite series. $R(\tau)$ is called *lag-τ autocorrelation* w.r.t. g. For ergodic Markov chains, $R(\tau)$ is positive and strictly decreasing. A fast decay of $R(\tau)$ indicates that consecutive samples get mutually uncorrelated fast (for uncorrelated g^i we would have $R(\tau) = \delta_{(\tau,0)}$). Figure 4.4 shows $R(\tau)$ for three stochastic processes. Such a visual comparison is often most instructive to compare different samplers and, as we will see in the numerical studies, can furthermore reveal additional properties. The *integrated autocorrelation time* τ_{int} is the integral of $R(\tau)$

extended to a piecewise linear function on $\mathbb{R}_+$:

$$\tau_{int} = \frac{1}{2} + \sum_{\tau=1}^{\infty} R(\tau). \tag{4.23}$$

τ_{int} is a quantitative measure of the amount of autocorrelation contained in a chain. An ergodic Markov chain of length K in its stationary phase has effectively the same statistical power for MC integration as $K_{eff} = K/(2\tau_{int})$ independent samples of its stationary distribution. K_{eff} is often called *effective sample size* and can guide the choice of the sub-sampling rate SSR. In practice, we have to estimate $R(\tau)$ by

$$\hat{R}(\tau) := \frac{1}{(K-\tau)\hat{\varrho}^2} \sum_{i=1}^{K-\tau} \left(g^i - \hat{\mu}\right)\left(g^{i+\tau} - \hat{\mu}\right) \tag{4.24}$$

$$\hat{\varrho}^2 := \frac{1}{K}\sum_{i=1}^{K}\left(g^i - \hat{\mu}\right)^2, \quad \hat{\mu} := \frac{1}{K}\sum_{i=1}^{K} g^i \tag{4.25}$$

(there are other possibilities to define $\hat{R}$, but we need $\hat{R}(0) = 1$). We estimate $R(\tau)$ on the basis of $(K - \tau)$ samples. As a result, the error of the estimation grows with τ and the tails of $\hat{R}(\tau)$ typically oscillate and also become negative. For these reasons, we should also estimate the error of $\hat{R}(\tau)$. In addition, estimating τ_{int} by

$$\hat{\tau}_{int} = \frac{1}{2} + \sum_{\tau}^{K} \hat{R}(\tau). \tag{4.26}$$

turns out to be unstable. A robust estimation of $\hat{R}(\tau)$, τ_{int} and their error is an involved topic (Thompson 2010). Here, we will use the approach presented in Wolff (2004), which also allows to reduce the estimation error by incorporating multiple independent chains.

For practical comparisons, the decrease of autocorrelation with respect to the raw number of samples (*statistical efficiency*) is not decisive. A sampler with a slow decrease might still outperform other samplers if it computes new samples considerably faster (*computational efficiency*). To address this, the acf and τ_{int} can be scaled by the computation time per sample t_s: $R^*(t = \tau/t_s) := R(\tau)$, $t_{int} := t_s\tau_{int}$. This facilitates the comparison of conceptually different sampling methods. $R^*(t)$ and t_{int} measure how fast a sampler can produce a certain loss in autocorrelation, which is of main interest for practical applications. However, while these measures are more decisive to compare different samplers, they rely on their concrete implementation.

Typically, the test function g is chosen with respect to the specific aim of inference. For instance, one could use the distance to the empirical mean of the whole chain if CM

estimation is performed, or the projection onto a specific coordinate if that coordinate should be marginalized. Then, the rate of autocorrelation decrease is a measure of the efficiency of the chain for the specific inference aim. Often, $g(x) = \log(p(x))$ is regarded as a generic choice, but we experienced that it might be rather uninformative in Bayesian inversion.

Instead of relying on 1D projections of the chains, it would arguably be better to develop and monitor a multivariate extension of the acf, e.g., the *autocorrelation matrix function* $\mathrm{R} : \{0, \dots, K-1\} \to \mathbb{R}^{n \times n}$

$$\mathrm{R}_{k,l}(\tau) := \frac{1}{(K-\tau)\hat{\varrho}_k \hat{\varrho}_l} \sum_{i=1}^{K-\tau} \left(x_k^i - \hat{\mu}_k\right) \left(x_l^{i+\tau} - \hat{\mu}_l\right) \tag{4.27}$$

$$\hat{\varrho}_k^2 := \frac{1}{K} \sum_{i=1}^{K} \left(x_k^i - \hat{\mu}_k\right)^2, \quad \hat{\mu}_k := \frac{1}{K} \sum_{i=1}^{K} x_k^i. \tag{4.28}$$

One could then define an integrated autocorrelation time as a measure of the integrated autocorrelation matrix,

$$\frac{1}{2} I_n + \sum_i \mathrm{R}(\tau), \tag{4.29}$$

for instance, its determinant or trace. However, its computation requires the storage of the whole chain $\{x^i\}$, not only of its projection $\{g^i\}$, which is infeasible in typical imaging applications.

Autocorrelation analysis for multimodal posteriors, especially for those originating from fat-tailed priors (cf. Section 3.2.5), is way more difficult to carry out and should be interpreted with care: As it is based on second order statistics, $R(\tau)$ might not be too meaningful to characterize the chain, or it might not even exist (the classical estimators for sample mean and variance will diverge). But also practically, estimating $R(\tau)$ by (4.24) and (4.25) becomes considerably more difficult. One reason is that $\hat{R}(\tau)$ is very sensitive to the accuracy of $\hat{\mu}$ (4.25). Consider a bi-modal distribution in 1D consisting of two Gaussians with equal variance but different means $\tilde{\mu}$ and $-\tilde{\mu}$: An MH sampler initialized in $x = \mu$ will stay in the first mode for some time before crossing zero and entering the second mode. Let's assume that a short chain is used for estimating $R(\tau)$ for $g(u) = u$. If the chain did not yet cross zero, all estimated quantities only reflect the properties of the local movement of the chain within the mode. As such, $\hat{\mu} \approx \tilde{\mu}$ and $\hat{\tau}_{int}$ will be small, as the samples are more or less evenly distributed around $\hat{\mu}$. However, once the chain crosses zero to enter the other mode, the estimate for $\hat{\mu}$ will shift and $\hat{\tau}_{int}$ will increase fast. All estimates are now dominated by the switch between the modes, and no longer by the local movement of the chain within a particular mode. If we have a second estimate for μ, $\hat{\mu}^{ref}$ (possible from an independent, longer MCMC run), we

can replace $\hat{\mu}$ by $\hat{\mu}^{ref}$ in (4.24) and (4.25) to compute $\hat{\tau}_{int}^{ref}$ as a second estimate for τ_{int}. If the difference between $\hat{\tau}_{int}^{ref}$ and $\hat{\tau}_{int}$ is large, we probably face the problems sketched above.
In general, autocorrelation analysis might not be the most suitable way to assess the performance of MCMC samplers for multimodal distributions. One should rather try to measure the ability of the sampler to switch between different modes of the posterior in a more direct way.

4.1.7. SC Gibbs Posterior Sampling

For sampling $p_{post}(u|f)$ with SC Gibbs sampling, we first need simple, parameterized representations of the conditional 1D densities. In this section, we will derive such representations for the prior models we introduced. The aim is to find a basis $\{v_1, \dots, v_n\}$ of $\mathbb{R}^n$ to represent $u = \sum \xi_i v_i =: V\xi$ such that $p_{post}(\xi_i|\xi_{-i}, f)$ can be described using as few parameters as possible. Once such a basis is found, the part of $p_{post}(\xi_i|\xi_{-i}, f)$ coming from the likelihood is easy to derive: We define $\Psi := AV$ and $\varphi(i) := f - \Psi_{-i}\xi_{-i}$. Then, we find that

$$\begin{aligned}\frac{1}{2}\|f - Au\|_2^2 = \frac{1}{2}\|f - AV\xi\|_2^2 = \frac{1}{2}\|f - \Psi\xi\|_2^2 = \frac{1}{2}\|f - (\Psi_{-i}\xi_{-i} + \Psi_i\xi_i)\|_2^2 \\ = \frac{1}{2}\|(\varphi(i) - \Psi_i\xi_i\|_2^2 \overset{\xi_i}{\propto} \frac{1}{2}\|\Psi_i\|_2^2\xi_i^2 + \Psi_i^T\varphi(i)\xi_i =: ax^2 - bx, \quad (4.30)\end{aligned}$$

where we introduced $x := \xi_i$, $a := \frac{1}{2}\|\Psi_i\|_2^2$, and $b := \Psi_i^T\varphi(i) = \Psi_i^T f - (\Psi_i^T\Psi_{-i})\xi_{-i}$ to ease the notation for the following sections. Note that while a and $\Psi_i^T f$ can be precomputed, $(\Psi_i^T\Psi_{-i})\xi_{-i}$ relies on the current state of the ξ-chain and has to be computed in every step of the sampler. Especially for complicated forward operators in high dimensional scenarios, this operation is the computational bottle-neck of SC Gibbs samplers. Therefore, a careful, scenario-dependent implementation is important to obtain a fast sampler. The details of this step can be found in Appendix A.2.
Now we proceed to determine V and the part of $p_{post}(\xi_i|\xi_{-i}, f)$ coming from the prior. The energies of the ℓ_p^q and the t_p priors we introduced can be written as

$$\mathcal{J}(u) = \left(\sum_j^h \phi\left(\left|D_j^T u\right|\right)\right)^\alpha = \left(\sum_j^h \phi\left(\left|\sum_l (D_j^T v_l)\xi_l\right|\right)\right)^\alpha, \quad (4.31)$$

with a function $\phi(z)$ fulfilling $\phi(0) = 0$ and an exponent α. To obtain simple conditional densities for all ξ_i, we thus have to choose V such that

$$\max_i |D^T v_i|_0 \quad (4.32)$$

is as small as possible. In this thesis, we will mainly consider the particular case of $D^T \in \mathbb{R}^{h \times n}$ having full rank and $h \leqslant n$. This includes the case where the columns D are elements of a basis, (3.25), and the increment prior in 1D with Neumann boundary conditions, (3.18). Due to the full rank, we can choose $v_1, \ldots, v_h$ such that $D^T v_l = e_l$ for $l = 1, \ldots, h$, and $v_{h+1}, \ldots, v_n$ such that $D^T v_l = 0$ for $l = h+1, \ldots, n$ (for D being a basis, we have $V = D$). With this transformation, $\mathcal{J}(\xi)$ simplifies to

$$\mathcal{J}(\xi) \propto \left(\sum_l^h \phi\left(|\xi_l|\right) \right)^\alpha = \left(\phi\left(|\xi_i|\right) + \sum_{l \neq i}^h \phi\left(|\xi_l|\right) \right)^\alpha . \tag{4.33}$$

As above, we will define $x := \xi_i$. For ℓ_p^q priors, we have $\phi(z) = |z|^p$ and $\alpha = q/p$. The conditional posterior can then be written as

$$p(x) \propto \exp\left(-ax^2 + bx - c\left(|x|^p + d\right)^{q/p} \right), \quad c := \lambda \mathbb{1}_{\{i \leqslant h\}}, \quad d := \sum_{l \neq i}^h |\xi_l|^p, \tag{4.34}$$

which simplifies to

$$p(x) \propto \exp\left(-ax^2 + bx - c|x|^p \right), \quad c := \lambda \mathbb{1}_{\{i \leqslant h\}}, \tag{4.35}$$

for ℓ_p priors. For the product t_p-priors (3.53), we have $\phi(z) = \log(1 + |z|^p/(\nu\theta))$ and $\alpha = 1$. The conditional posterior can be written as

$$p(x) \propto \exp\left(-ax^2 + bx - c \log\left(1 + \frac{|x|^p}{d} \right) \right), \quad c := \frac{\nu + 1}{p} \mathbb{1}_{\{i \leqslant h\}}, \quad d := \nu\theta. \tag{4.36}$$

The block prior energies we introduced can be written as

$$\mathcal{J}(u) \propto \left(\sum_j^h \phi\left(\left\| D_{[j]}^T u \right\|_2 \right) \right)^\alpha = \left(\sum_j^h \phi\left(\left\| \sum_l (D_{[j]}^T v_l) \xi_l \right\|_2 \right) \right)^\alpha . \tag{4.37}$$

If we denote the block to which the i-th component belongs by $[l_i]$ and assume that a similar transformation as above is available, we can simplify (4.37) to

$$\left(\sum_l^h \phi\left(\left\| \xi_{[l]} \right\|_2 \right) \right)^\alpha = \left(\phi\left(\sqrt{\xi_i^2 + \sum_{j \in [l_i], j \neq i} \xi_j^2} \right) + \sum_{l \neq l_i}^h \phi\left(\sqrt{\sum_{j \in [l]} \xi_j^2} \right) \right)^\alpha . \tag{4.38}$$

The conditional posterior for the ℓ_p^q-block prior can then be written as

$$p(x) \propto \exp\left(-ax^2 + bx - c\left(\left(x^2 + g\right)^{p/2} + d \right)^{q/p} \right),$$

$$\text{where} \qquad c := \lambda \mathbb{1}_{\{i \leqslant h\}}, \qquad g := \sum_{j \in [l_i],\, j \neq i} \xi_j^2, \qquad d = \sum_{l \neq l_i}^{h} \left(\sum_{j \in [l]} \xi_j^2 \right)^{p/2}. \quad (4.39)$$

For the t_p-block prior, we obtain

$$p(x) \propto \exp\left(-ax^2 + bx - c \log\left(1 + \frac{(x^2+g)^{p/2}}{d} \right) \right),$$
$$\text{where} \qquad c := -\frac{\nu+1}{p} \mathbb{1}_{\{i \leqslant h\}}, \qquad d := \nu\theta, \qquad g := \sum_{j \in [l_i],\, j \neq i} \xi_j^2. \quad (4.40)$$

In other cases that cannot be treated by the above scheme, a simple parameterization has to be derived explicitly. For instance, consider the isotropic TV prior in 2D, (3.22). Every pixel $u_{(i,j)}$ only appears in tree terms of the energy:

$$\begin{aligned} \mathcal{J}_{iTV}\left(u_{(i,j)} \,\middle|\, u_{-(i,j)} \right) \overset{(i,j)}{\propto} & \sqrt{(u_{(i+1,j)} - u_{(i,j)})^2 + (u_{(i,j+1)} - u_{(i,j)})^2} \\ & + \sqrt{(u_{(i,j)} - u_{(i-1,j)})^2 + (u_{(i-1,j+1)} - u_{(i-1,j)})^2} \\ & + \sqrt{(u_{(i+1,j-1)} - u_{(i,j-1)})^2 + (u_{(i,j)} - u_{(i,j-1)})^2}. \end{aligned} \quad (4.41)$$

Therefore, we can write the conditional posterior as

$$p(x) \propto \exp\left(-ax^2 + bx - c \sum_{j=1}^{3} \sqrt{d_j (x - e_j)^2 + g_j} \right), \quad d_j \in \{0,1,2\}, \quad g_j \geqslant 0, \quad (4.42)$$

with appropriately computed parameters d_j, e_j, g_j. For a general ℓ_1, prior where D does not fulfill the above requirements, an explicit form is given by

$$p(x) \propto \exp\left(-ax^2 + bx - c \sum_{j \in \operatorname{supp}(D^T v_i)} |d_j x - e_j| \right),$$
$$\text{where} \qquad c := \lambda \mathbb{1}_{\{i \leqslant h\}}, \qquad d_j := \left(D^T v_i \right)_j, \qquad e_j := \left(D^T V_{-i} \xi_{-i} \right)_j \quad (4.43)$$

The difficulty of incorporating additional hard constraints (cf. Section 3.2.2) depends on the shape of the feasible set $\mathcal{C}$ and the transformation V applied. In the following, we assume that they lead to a feasible (semi-)finite interval $[x_{min}, x_{max}]$ to which the continuous densities computed above can be restricted. In the case of $\mathcal{C}$ being convex, such an interval always exists and there are computationally efficient ways to compute it.

4.1.8. Direct SC Gibbs Posterior Sampling

After having derived the parameterized SC densities $p(x)$, we can now turn to the second challenge in designing a fast SC Gibbs sampler: Developing fast, robust and exact 1D samplers for $p(x)$. For some $p(x)$, we can rely on direct samplers (cf. Section 4.1.2).

ℓ_2-prior For (4.35) and $p = 2$, we find that $p(x)$ is, as expected, a Gaussian density:

$$\exp\left(-ax^2 + bx - cx^2\right) \propto \exp\left(-\frac{(x-\mu_{sc})^2}{2\sigma_{sc}^2}\right), \tag{4.44}$$

where $\mu_{sc} := b/(2a+2c)$, and $\sigma_{sc}^2 = 1/(2a+2c)$. Hence, direct samplers can be employed. For the constrained (*truncated*) case $x \in [x_{min}, x_{max}]$, various direct samplers were developed. We will use a modified, more robust, version of Chopin (2011) throughout this thesis.

ℓ_1-prior For (4.35) and $p = 1$, a direct, icd-based method was developed in Lucka (2012): First, we need to compute the normalization factor for $p(x) \propto \exp(-a\,x^2 + b\,x - c\,|x|)$. Splitting the integral from $-\infty$ to ∞ into two parts (from $-\infty$ to 0 and the rest) yields subproblems that can be treated like the normalization of the normal distribution (completing the square and a linear integral transformation). This leads to:

$$\begin{aligned}\mathcal{N} &:= \int_{-\infty}^{\infty} \exp(-ax^2 + b\,x - c|x|)\,\mathrm{d}x \\ &= \frac{1}{2}\sqrt{\frac{\pi}{a}}\left(e^{\frac{(b+c)^2}{4a}}\operatorname{erfc}\left(\frac{b+c}{2\sqrt{a}}\right) + e^{\frac{(c-b)^2}{4a}}\operatorname{erfc}\left(\frac{c-b}{2\sqrt{a}}\right)\right) \\ &=: \chi\left(\tilde{e}_+\operatorname{erfc}(\alpha_+) + \tilde{e}_-\operatorname{erfc}(\alpha_-)\right),\end{aligned} \tag{4.45}$$

where $\operatorname{erfc}(y) := \frac{2}{\sqrt{\pi}}\int_y^\infty e^{-t^2}\,\mathrm{d}t$ denotes the *complementary error function*. The cdf $F(y)$ is given by:

$$\begin{aligned}F(y) &= \frac{1}{\mathcal{N}}\int_{-\infty}^{y} \exp(-a\,x^2 + b\,x - c\,|x|)\,\mathrm{d}x \\ &= \frac{\chi}{\mathcal{N}}\cdot\begin{cases}\tilde{e}_+\operatorname{erfc}\left(-\sqrt{a}\,y + \alpha_+\right), & \text{if } y < 0, \\ \tilde{e}_+\operatorname{erfc}(\alpha_+) + \tilde{e}_-\left(\operatorname{erfc}(\alpha_-) - \operatorname{erfc}\left(\sqrt{a}\,y + \alpha_-\right)\right), & \text{if } y > 0.\end{cases}\end{aligned} \tag{4.46}$$

Inverting this cdf for a given $r \sim \mathrm{unif}(0,1)$ is simple. To find $y = F^{-1}(r)$, we first check if $y < 0$ by using the cdf for this domain: Let

$$z := \operatorname{erfcinv}\left(\frac{r\,\mathcal{N}}{\chi\,\tilde{e}_+}\right) = \operatorname{erfcinv}\left(\frac{r\,\chi\,\left(\tilde{e}_+\operatorname{erfc}(\alpha_+) + \tilde{e}_-\operatorname{erfc}(\alpha_-)\right)}{\chi\tilde{e}_+}\right)$$

$$= \text{erfcinv}\left(r\left(\text{erfc}\left(\alpha_+\right) + \frac{\tilde{e}_-}{\tilde{e}_+}\,\text{erfc}\left(\alpha_-\right)\right)\right)$$
$$= \text{erfcinv}\left(r\left(\text{erfc}\left(\alpha_+\right) + \exp\left(-\frac{b\,c}{a}\right)\,\text{erfc}\left(\alpha_-\right)\right)\right). \qquad (4.47)$$

Then, y is given by $y = -(z - \alpha_+)/\sqrt{a}$. If $y > 0$, the other half of the cdf has to be inverted: Let

$$z := \text{erfcinv}\left(\left(-\frac{r\,\mathcal{N}}{\chi} + \tilde{e}_+\text{erfc}\left(\alpha_+\right) + \tilde{e}_-\text{erfc}\left(\alpha_-\right)\right)\tilde{e}_-^{-1}\right)$$
$$= \text{erfcinv}\left((1-r)\left(\exp\left(\frac{b\,c}{a}\right)\text{erfc}\left(\alpha_+\right) + \text{erfc}\left(\alpha_-\right)\right)\right). \qquad (4.48)$$

Then, y is given by $y = (z - \alpha_-)/\sqrt{a}$. The constrained case $x \in [x_{min}, x_{max}]$ is, in principle, easy to handle using the icd method: Instead of drawing $r \sim \text{unif}(0,1)$ we compute $F(x_{min})$ and $F(x_{max})$ by (4.46) and draw $r \sim \text{unif}(F(x_{min}), F(x_{max}))$.
The complementary error function and its inverse are difficult to handle numerically, because there are no identities that allow to rescale or shift their evaluation to other intervals. Therefore, a robust numerical implementation of formulas (4.46), (4.47) and (4.48) is rather involved. For the sake of a concise presentation, we present all details in Section A.3.

ℓ_1^2-prior As

$$\exp\left(-ax^2 + bx - c\left(|x| + d\right)^2\right) \propto \exp\left(-(a+c)x^2 + bx - 2cd|x|\right), \qquad (4.49)$$

we can also use the above formulas for a direct sampler for (4.34) with $p = 1$ and $q = 2$.

4.1.9. Slice Sampling

In this section, we introduce an MCMC technique to sample from the SC density $p(x)$ for those cases where no fast and robust direct samplers are known. One could, of course, employ SRWMH schemes with a univariate Gaussian proposal $\mathcal{N}(0, \kappa)$ for this purpose. The whole sampler would then be called *Metropolis-within-Gibbs* sampler. Compared to MH-Si, a Metropolis-within-Gibbs sampler explicitly computes the SC densities and uses MH to sample from those, not for the whole multivariate density. However, the proper tuning of κ is way more difficult than in the multivariate case: The SC densities can be very different between components, and even for a fixed component, they may vary dramatically over the run. This is particularly true for multimodal posteriors from sparse, non-log-concave priors. An example is given in Figure 4.5a, where the shape and spatial spread of a SC density vary considerably for a small variation in one of its

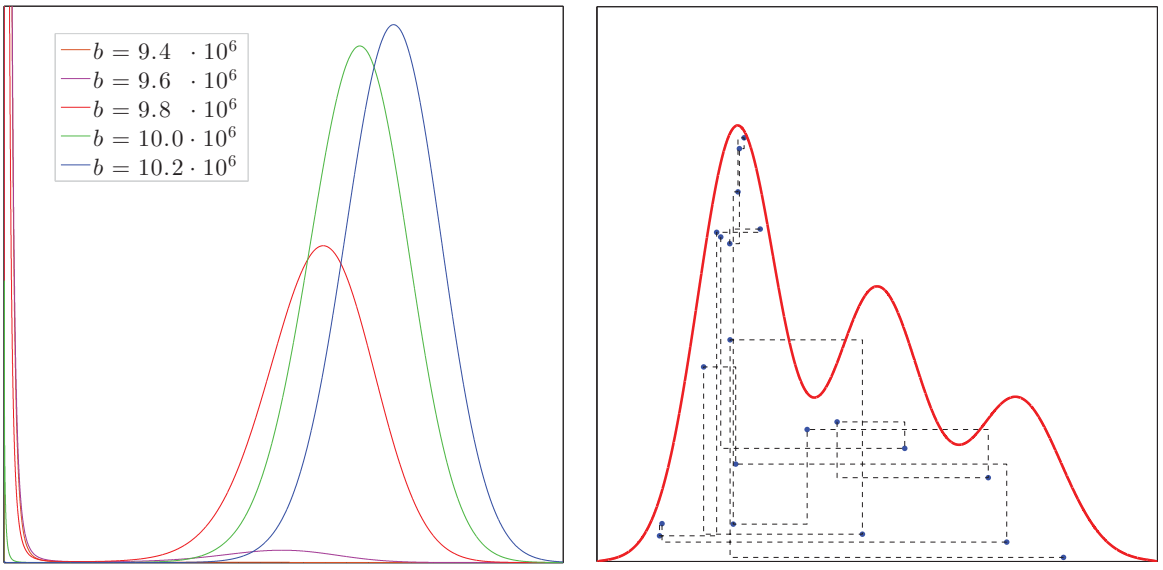


Figure 4.5.: (a) The SC density (4.35) for $p = 0.8$ with $a = 6.6 \cdot 10^6$, $c = 10^6$ and several, slightly varying values for b (such values occur, e.g., in the point source reconstruction scenario, cf. Section 2.2.2). In addition, non-negativity constraints are used. (b) Slice sampling. The x coordinates of the blue dots are samples of $p(x)$ (red line), the dashed black line illustrates the path of the sampler.

parameters. A fixed κ cannot be tuned to yield similar and stable acceptance rates for all components. Using an individual κ_i for each component might fail due to potential multimodality, and tuning it automatically would require n times more samples than tuning one κ for all components.

The Gibbs sampler automatically adapts to a distribution, but cannot be applied to a 1D density $p(x)$. However, it can be used to sample uniformly from its 2D subgraph $\mathcal{G}_p$ which will also generate a sample of $p(x)$ (see Section 4.1.2 and Figure 4.1). This is the basic idea of *slice sampling* (Neal 2003):

Algorithm 4.4. (Basic Slice Sampling)
For a univariate density $p(x)$, Step 3. in Algorithm 4.1 can be implemented as:

3.1. Draw y uniform from $[0, p(x^i)]$ (vertical move).

3.2. Draw x uniform from $S^y := \{z \mid p(z) \geqslant y\}$ (horizontal move).

3.3. Set $x^{i+1} = x$.

An illustration is given in Figure 4.5b. In general, the difficulty of slice sampling is determining S^y in Step 3.2. If it does not allow for an explicit formulation, numerical root-finding algorithms have to be used to determine all $\{z \mid p(z) - y = 0\}$ and S_y has to be constructed from them. For non-log-concave $p(x)$, S^y may consist of multiple

intervals. For the SC posterior densities we derived in Section 4.1.7, determining S^y numerically is not a feasible option. We will rather need to generalize the principle behind slice sampling: Slice sampling is a variant of *auxiliary variables algorithms* that introduce an additional variable y with a suitable density $p(y|x)$. Then, samples (x^i, y^i) from $p(x, y) = p(x)p(y|x)$ are obtained by a Gibbs sampler (which relies on $p(y|x)$ and $p(x|y)$), and only the x^i are kept. For the basic slice sampler, $p(y|x)$ is chosen as

$$p(y|x) = \frac{1}{p(x)} \mathbb{1}_{\{[0,p(x)]\}}(y), \tag{4.50}$$

i.e., as unif$(0, p(x^i))$. We then have

$$p(x, y) = p(x) \frac{1}{p(x)} \mathbb{1}_{[0,p(x)]}(y) \tag{4.51}$$

$$p(x|y) \propto \mathbb{1}_{\{[0,p(x)]\}}(y) = \mathbb{1}_{\{x \mid p(x) \geqslant y\}}(x) \tag{4.52}$$

If $p(x)$ can decomposed as $p(x) \propto p_1(x)p_2(x)$ - for instance, for $p(x) \propto \exp(-\mathcal{J}_1(x) - \mathcal{J}_2(x))$ - we can define

$$p(y|x) = \frac{1}{p_2(x)} \mathbb{1}_{\{[0,p_2(x)]\}}(y), \tag{4.53}$$

which leads to

$$p(x, y) = p(x)p(y|x) = p(x) \frac{1}{p_2(x)} \mathbb{1}_{\{[0,p_2(x)]\}}(y) = p_1(x) \mathbb{1}_{\{[0,p_2(x)]\}}(y), \tag{4.54}$$

$$p(x|y) = p_1(x) \mathbb{1}_{\{x \mid p_2(x) \geqslant y\}}(x). \tag{4.55}$$

This split is appealing if

$$S_2^y := \{z \mid p_2(z) \geqslant y\} \tag{4.56}$$

is a single interval and easy to determine and $p_1(x)$ constrained to an interval is easy to sample from.

Algorithm 4.5. (Slice Sampling)
For a univariate density $p(x) \propto p_1(x)p_2(x)$, Step 3. in Algorithm 4.1 can be implemented as:

3.1. Draw y uniform from $[0, p_2(x^i)]$ (vertical move).

3.2. Draw x from $p_1(x) \mathbb{1}_{S_2^y}(x)$ (weighted horizontal move).

3.3. Set $x^{i+1} = x$.

4.1.10. Slice Sampling Withing SC Gibbs Posterior Sampling

We want to use the slice sampler to sample from the SC densities, which we might call *slice-within-Gibbs* sampling. For the SC densities, a split of $p(x)$ into likelihood $p_1(x) = \exp(-ax^2 + bx)$ and prior parts $p_2(x)$ is advantageous: As most prior terms are unimodal and even symmetric to zero, S_2^y is a single interval and can be determined explicitly:
For (4.34), we have $p_2(x) \propto \exp\left(-c\left(|x|^p + d\right)^{q/p}\right)$ and

$$\exp\left(-c\left(|x|^p + d\right)^{q/p}\right) \geqslant y \iff |x| \leqslant \left(\left(-\frac{\log(y)}{c}\right)^{p/q} - d\right)^{1/p}. \tag{4.57}$$

For (4.39), we have $p_2(x) \propto \exp\left(-c\left(\left(x^2 + g\right)^{p/2} + d\right)^{q/p}\right)$ and

$$\exp\left(-c\left(\left(x^2 + g\right)^{p/2} + d\right)^{q/p}\right) \geqslant y \iff |x| \leqslant \sqrt{\left(\left(-\frac{\log(y)}{c}\right)^{p/q} - d\right)^{2/p} - g}. \tag{4.58}$$

For (4.36), we have $p_2(x) \propto \left(1 + \frac{|x|^p}{d}\right)^{-c}$ and

$$\left(1 + \frac{|x|^p}{d}\right)^{-c} \geqslant y \iff |x| \leqslant d^{1/p}\left(y^{-1/c} - 1\right)^{1/p}. \tag{4.59}$$

For (4.40) we have $p_2(x) \propto \left(1 + \frac{(x^2+g)^{p/2}}{d}\right)^{-c}$ and

$$\left(1 + \frac{|x|^p}{d}\right)^{-c} \geqslant y \iff |x| \leqslant \sqrt{d^{1/p}\left(y^{-1/c} - 1\right)^{2/p} - g}. \tag{4.60}$$

For the TV prior, (4.42), we need to compute S_2^y numerically. However, the energy of $p_2(x)$ is convex. Therefore, S_2^y is a single interval given by the solutions to $p_2(x) = y$. As the energy of $p_2(x)$ is also piecewise smooth and can be bounded from below, we can easily find starting points for fast, derivative-based root-finding-algorithms. The details are given in Appendix A.4. A generalization to other convex, piecewise-smooth energies is straight-forward. The piecewise linear energy from (4.43) is a special case: $p_2(x) = y$ can be solved explicitly in a simple way.
The likelihood part is a Gaussian with $\mu_{SS} = b/(2a)$ and $\sigma_{SS}^2 = 1/(2a)$. As noted in Section 4.1.8, fast and robust samplers for truncated Gaussians exist. Incorporating hard constraints in slice sampling is very easy: Instead of sampling $p_1(x)$ truncated to S_2^y, we sample it truncated to $S_2^y \cap [x_{min}, x_{max}]$.

In principle, the slice sampler will generate a full Markov chain, but practically, we only need one sample from it. We will initialize it with the current value ξ_i of the component we want to update. Then, we only have to determine the length of the burn-in phase k_0 and choose the first sample of the real run as a sample of $p(x)$.

4.1.11. Posterior Sampling for Hierarchical Bayesian Models

As mentioned in the introduction of Section 4.1.5, the construction of hierarchical Bayesian models by conditional distributions is appealing for carrying out Gibbs sampling over the partition into u and γ: $p_{post}(u|f,\gamma)$ usually belongs to a class of distributions for which a sampler is known, and $p_{post}(\gamma|f,u) = p_{post}(\gamma|u)$ often factorizes into univariate distributions over γ_i. In fact, hierarchical modeling is often only used *because* it allows for Gibbs posterior sampling.
The conditionally ℓ_p hypermodels (3.41) we use in this thesis were constructed such that $p_{post}(u|f,\gamma)$ is a posterior resulting from using an ℓ_p prior. For these posteriors, samplers were developed in the previous sections. As we usually use a factorizing hyperprior of the type

$$p_{hyper}(\gamma) \propto \prod_i^h \gamma_i^{-\delta} \exp\left(-\varphi_i(\gamma_i)\right), \tag{4.61}$$

the conditional posterior $p_{post}(\gamma|u)$ also factorizes:

$$p_{post}(\gamma|u) \propto \prod_i^h \exp\left(-\frac{\left|D_i^T u\right|^p}{\gamma_i} - (\delta + 1/p)\log(\gamma_i) - \varphi_i(\gamma_i)\right) \tag{4.62}$$

In certain cases, for instance for the inverse gamma distribution we mainly use in this thesis, direct univariate samplers can be used (cf. (3.47)). As shown in Section 3.3.3, ℓ_p hypermodels with inverse gamma hyperpriors can be used as *surrogate* prior models for product t_p priors. Thereby, we have a second sampling scheme for product t_p priors. To distinguish this sampler from the SC Gibbs sampler, we will refer to it as *blocked Gibbs sampler*.

4.1.12. Notes and Comments

Random number generation and Monte Carlo strategies are a vast topic. In particular, since samplers for more complex tasks usually have to be constructed by combining various simpler samplers, the number of possible sampling techniques is huge. However, no universal method is known which exhibits a good performance for all types of distributions. For a comprehensive overview, we refer to Liu (2008), Robert and Casella (2005).

4.2. Posterior Optimization Methods

In this section, we will present algorithms that can be used to compute MAP estimates for different prior distributions.

4.2.1. Least Squares Methods for Gaussian Priors

As mentioned in Section 4.1.2, the posterior using a Gaussian prior is also a Gaussian. Therefore, the MAP and the CM estimate are given by

$$\hat{u}_{\mathrm{MAP}} = \hat{u}_{\mathrm{CM}} = \mathbb{E}[u|f] = \left(2\lambda DD^T + A^TA\right)^{-1} A^Tf \tag{4.63}$$

The most efficient way to solve this problem depends on the properties of A and D. In Lucka (2011), various techniques to reformulate and solve (4.63) are discussed. If A or D cannot be used as an explicit matrix, only iterative methods that compute $\hat{u}_{\mathrm{MAP}}$ by solving

$$\begin{bmatrix} A \\ \sqrt{2\lambda}D^T \end{bmatrix} v \overset{ls}{=} \begin{bmatrix} f \\ 0 \end{bmatrix} \tag{4.64}$$

in a least-squares sense can be used. We will use the *conjugate gradient least squares* (*CGLS*) method for this purpose, see Section 3.5 in Lucka (2011). A general reference is given by Saad (2003).

4.2.2. ADMM Methods for Log-Concave Priors

The optimization problem arising from computing the MAP estimate for log-concave priors (cf. Section 3.2.3),

$$\min_u \quad \mathcal{E}(u) := \frac{1}{2}\|f - Au\|_2^2 + \lambda\mathcal{J}(u), \tag{4.65}$$

is convex. Convex optimization problems, which comprise least-squares and linear optimization problems, are a fundamental class of optimization problems. Using concepts from convex analysis such as *duality* and *subgradient calculus*, they can be solved in a very efficient and generic way (Boyd and Vandenberghe 2004). A further advantage of (4.65) is that the posterior energy $\mathcal{E}(u)$ naturally decomposes into likelihood and prior energy terms, which are both convex. Due to this separation, it is possible to introduce an auxiliary variable $v \in \mathbb{R}^l$ to reformulate (4.65) as

$$\min_{u,v} \quad h(u) + g(v) \quad s.\,t. \quad Eu + Fv = b, \tag{4.66}$$

where $E \in \mathbb{R}^{l\times n}, F \in \mathbb{R}^{l\times k}, b \in \mathbb{R}^l$. For instance, $h(u) = \frac{1}{2}\|f - Au\|^2_{\Sigma_\varepsilon^{-1}}$, $g(u) = \lambda\mathcal{J}(u)$, $u - v = 0$ is always possible. For ℓ_p^q priors with a matrix D^T, $v = \lambda^{p/q}D^T u$ would be a reasonable split. The split problem (4.66) can be solved efficiently with the *alternating direction method of multipliers* (*ADMM*). An extensive reference is given by Boyd et al. (2011). ADMM builds on *dual ascent*, *augmented Lagrangian techniques*, and the *method of multipliers*: Solving the dual problem to (4.66) by gradient ascend would consist of the iteration

$$(u^{i+1}, v^{i+1}) := \operatorname*{argmin}_{(u,v)} \mathcal{L}(u, v, y^i) \tag{4.67}$$

$$y^{i+1} := y^i + \alpha\left(Eu^{i+1} + Fv^{i+1} - b\right), \tag{4.68}$$

where

$$\mathcal{L}(u, v, y) := h(u) + g(v) + y^T(Eu + Fv - b) \tag{4.69}$$

is the *Lagrangian* for problem (4.66), y the dual variable and α a step size. Augmented Lagrangian techniques replace $\mathcal{L}(u, v, y)$ by

$$\mathcal{L}_\rho(u, v, y) = h(u) + g(v) + y^T(Eu + Fv - b) + \frac{\rho}{2}\|Eu + Fv - b\|_2^2, \tag{4.70}$$

with a *penalty parameter* ρ. While the primal problem for (4.70) is equivalent to (4.66), the dual problem is easier to solve. Applying dual ascend to the modified problem would consist of the iteration

$$(u^{i+1}, v^{i+1}) := \operatorname*{argmin}_{u,v} \mathcal{L}_\rho(u, v, y^i) \tag{4.71}$$

$$y^{i+1} := y^i + \rho\left(Eu^{i+1} + Fv^{i+1} - b\right), \tag{4.72}$$

whereby, ρ is now the step size α. This is known as the method of multipliers for solving (4.66). It has the nice property that all iterates (u^i, v^i, y^i) are dual feasible. ADMM consists of replacing the joint minimization in the first step by a single alternation between u and v:

Algorithm 4.6. (Alternating Direction Method of Multipliers)
Given ρ, $\mathcal{L}_\rho(u, v, y)$, v^0,y^0, repeat

$$u^{i+1} := \operatorname*{argmin}_{u} \mathcal{L}_\rho(u, v^i, y^i) \tag{4.73}$$

$$v^{i+1} := \operatorname*{argmin}_{v} \mathcal{L}_\rho(u^{i+1}, v, y^i) \tag{4.74}$$

$$y^{i+1} := y^i + \rho\left(Eu^{i+1} + Fv^{i+1} - b\right) \tag{4.75}$$

In principle, the alternation between u and v can be carried out multiple times before the dual variable y is updated. We will mainly use the *scaled form* of ADMM, which we obtain by replacing $w := y/\rho$. Then, the explicit formulation of ADMM is given as

Algorithm 4.7. (ADMM, explicit scaled form)
Given ρ, $h(u)$, $g(u)$, v^0,w^0, repeat

$$u^{i+1} := \operatorname*{argmin}_u \left(h(u) + \frac{\rho}{2} \left\| Eu + Fv^i - b + w^i \right\|_2^2 \right) \tag{4.76}$$

$$v^{i+1} := \operatorname*{argmin}_v \left(g(v) + \frac{\rho}{2} \left\| Eu^{i+1} + Fv - b + w^i \right\|_2^2 \right) \tag{4.77}$$

$$w^{i+1} := w^i + \left(Eu^{i+1} + Fv^{i+1} - b \right) \tag{4.78}$$

Posterior Optimization

We will now apply ADMM to the posterior energy (4.65) with a prior energy of the form $\lambda \mathcal{J}(D^T u)$. We will split by $D^T u = v$, i.e., $E = D^T$, $F = -I_h$, $b = 0$ and obtain

$$u^{i+1} := \operatorname*{argmin}_u \left(\frac{1}{2} \|f - Au\|_{\Sigma_\varepsilon^{-1}}^2 + \frac{\rho}{2} \left\| D^T u - (v^i - w^i) \right\|_2^2 \right) \tag{4.79}$$

$$v^{i+1} := \operatorname*{argmin}_v \left(\mathcal{J}(v) + \frac{\rho}{2\lambda} \left\| v - (D^T u^{i+1} + w^i) \right\|_2^2 \right) \tag{4.80}$$

$$w^{i+1} := w^i + D^T u^{i+1} - v^{i+1}. \tag{4.81}$$

The split decouples the operators A and D from the potentially non-quadratic $\mathcal{J}(v)$. Problem (4.79) is a least-squares problem which we can solve as in Section 4.2.1. If we define the *proximity operator* by

$$\mathbf{prox}_{\mathcal{J},\alpha}(x) := \operatorname*{argmin}_z \left(\mathcal{J}(z) + \frac{\alpha}{2} \|z - x\|_2^2 \right), \tag{4.82}$$

the solution of (4.80) is given by:

$$v^{i+1} = \mathbf{prox}_{\mathcal{J},\rho/\lambda} \left(D^T u^{i+1} + w^i \right). \tag{4.83}$$

The proximity operators of many functionals can be evaluated using closed-form expressions (see Combettes and Pesquet 2011, for an extensive overview). For $\mathcal{J}(z) = \|z\|_1$, $z^* := \mathbf{prox}_{\|\cdot\|_1,\alpha}(x)$ is given by component-wise *soft thresholding/shrinkage*:

$$z_i^* = \operatorname{sign}(x_i) \cdot \max\{0, |x_i| - 1/\alpha\} = \begin{cases} x_i - 1/\alpha & \text{if} \quad x_i > 1/\alpha \\ 0 & \text{if} \quad |x_i| \leqslant 1/\alpha \\ x_i + 1/\alpha & \text{if} \quad x_i < -1/\alpha \end{cases} \tag{4.84}$$

In Appendix A.5, we generalize the above to other ℓ_1-norm based prior energies we use in this thesis.

Stopping criterion

The *primal feasibility* of (u^i, v^i) can be measured by the norm of the *primal residuum* $r^{i+1} := Eu^{i+1} + Fv^{i+1} - b$ of the equality constraint. The *dual feasibility* can measured by the norm of the quantity $s^{i+1} := \rho E^T F(v^{i+1} - v^i)$, which we will call *dual residuum*. The necessary and sufficient optimality conditions for the ADMM problem are fulfilled if both residuals are zero. Therefore, one can derive a suitable stopping criterion for ADMM based on their norms (see Boyd et al. 2011, for details):

$$\left\|r^i\right\|_2 \leqslant \epsilon^{pri} \quad \text{and} \quad \left\|s^i\right\|_2 \leqslant \epsilon^{dual}, \tag{4.85}$$

where

$$\epsilon^{pri} = \sqrt{l}\epsilon^{abs} + \epsilon^{rel} \max\left\{\left\|Eu^i\right\|_2, \left\|Fv^i\right\|_2, \|b\|_2\right\} \tag{4.86}$$

$$\epsilon^{dual} = \sqrt{n}\epsilon^{abs} + \epsilon^{rel}\rho\left\|E^T w^i\right\|_2, \tag{4.87}$$

and $\epsilon^{abs} > 0$ and $\epsilon^{rel} > 0$ are predefined absolute and relative tolerances.

Penalty parameter adaptation

In practice, the speed of convergence of the ADMM method strongly depends on the proper choice of the penalty parameter ρ. For these reasons, an automatic online adaptation of ρ would be advantageous. Given that ρ becomes stationary after a finite number of iterations, ADMM is still guaranteed to convergence. We use a simple scheme that is based on the convergence criterion we use: One should keep primal and dual residual norms $\|r^i\|_2$ and $\|s^i\|_2$ close to each other while they both converge to zero. If the primal residuum is too large, we should increase ρ, which penalizes deviations from the equality constraint:

$$\rho^{i+1} := \begin{cases} \tau^{inc}\rho^i & \text{if} \quad \|r^i\|_2 > \mu\|s^i\|_2 \\ \rho^i/\tau^{dec} & \text{if} \quad \|s^i\|_2 > \mu\|r^i\|_2 \\ \rho^i & \text{otherwise} \end{cases} \tag{4.88}$$

Here, $\tau^{inc} > 1$, $\tau^{dec} > 1$ and $\mu > 1$ are parameters that control the adaptation. Whenever ρ is changed, the scaled dual variable w needs to be adjusted: $w^{i+1} := (\rho^i/\rho^{i+1})w^{i+1}$. While adopting ρ, one has to keep in mind that it also controls the condition of the least-squares problem (4.79), which becomes ill-conditioned in the limit of $\rho \to 0$.

Thereby, iterative methods such as CGLS will need considerably more iterations to solve it to a given tolerance. As this step is the computational bottle-neck of the ADMM method, it limits the ability of ρ-adaptation to speed up the convergence of ADMM in terms of computational time. In these situations, a lower bound for ρ should be chosen.

Constraints

To account for additional constraints of the form $u \in \mathcal{C}$ (cf. Section 3.2.2), ADMM has to be modified. Depending on the concrete form of $\mathcal{C}$, several options are available:

- The constraints can be incorporated in $h(u)$. Then, (4.79) becomes a constrained least squares problem. For certain constraints, like non-negativity, tailored algorithms have been developed (Chen and Plemmons 2009).

- The constraints can be incorporated into $g(v)$. Then, the proximity operator for solving (4.80) has to be modified. This only works if the constraints can be transferred from u to v.

- If $\mathcal{C}$ has a complicated or non-explicit structure, an additional splitting

$$\min_{u,v,t} \quad h(u) + g(v) + c(t) \quad s.\,t. \quad Eu + Fv + Gt = b, \tag{4.89}$$

  with

$$c(t) = \begin{cases} 0 & \text{if} \quad t \in \mathcal{C} \\ \infty & \text{else} \end{cases} \tag{4.90}$$

  has to be applied.

Notes and Comments

The appealing property of ADMM is its generality. Provided that they are convex, it offers a principled but simple way to treat various prior energies. Furthermore, it works for all combinations of operators A and D. In contrast, faster and even simpler methods are often available that only work for a specific combination of A and D.
Interestingly, ADMM can also be interpreted as an implementation of the *Bregman iteration* (Bregman 1967) for (4.66): For a convex optimization problem

$$\min_{u} \quad \mathcal{E}(u) \quad s.\,t. \quad Fu = b \tag{4.91}$$

the Bregman iteration is given by

$$\begin{aligned} u^{i+1} &= \operatorname*{argmin}_u \left\{ D_{\mathcal{E}}^p(u, u^i) + \frac{\rho}{2}\|Fu - b\|_2^2 \right\} \\ &= \operatorname*{argmin}_u \left\{ \mathcal{E}(u) - \left\langle \xi^i, u - u^i \right\rangle + \frac{\rho}{2}\|Fu - b\|_2^2 \right\}, \end{aligned} \tag{4.92}$$

where $D_{\mathcal{E}}^p(u, v)$ is, again, the Bregman distance encountered in (3.76) (cf. Section A.1). Using this iteration for image reconstruction is analyzed in Osher et al. (2006). Applied to the split problem (4.66), the Bregman iteration is called *split Bregman* method (Goldstein and Osher 2009). In most scenarios, the split Bregman method is equivalent to ADMM (Esser 2009). More general relations to other methods can be found in Boyd et al. (2011), Zhang et al. (2011).

4.2.3. Parameter Fitting for ℓ_p^q Priors

Consider a non-negative, convex functional $\mathcal{J}(u)$ and let

$$\hat{u}_{\text{MAP}}^r(\mu) := \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2}\|f - A(u)\|_{\Sigma_\varepsilon^{-1}}^2 + \mu \mathcal{J}(u)^r \right\}. \tag{4.93}$$

The optimality condition (cf. Section A.1) for $\hat{u}_{\text{MAP}}^r(\mu)$ is given by

$$\begin{aligned} & 0 \in \partial \left(\mu \mathcal{J}(\hat{u}_{\text{MAP}}^r(\mu))^r \right) + A^T \left(A\hat{u}_{\text{MAP}}^r(\mu) - f \right) \\ \implies \quad & 0 \in \mu r \mathcal{J}(\hat{u}_{\text{MAP}}^r(\mu))^{r-1} \partial \mathcal{J}(\hat{u}_{\text{MAP}}^r(\mu)) + A^T \left(A\hat{u}_{\text{MAP}}^r(\mu) - f \right) \end{aligned} \tag{4.94}$$

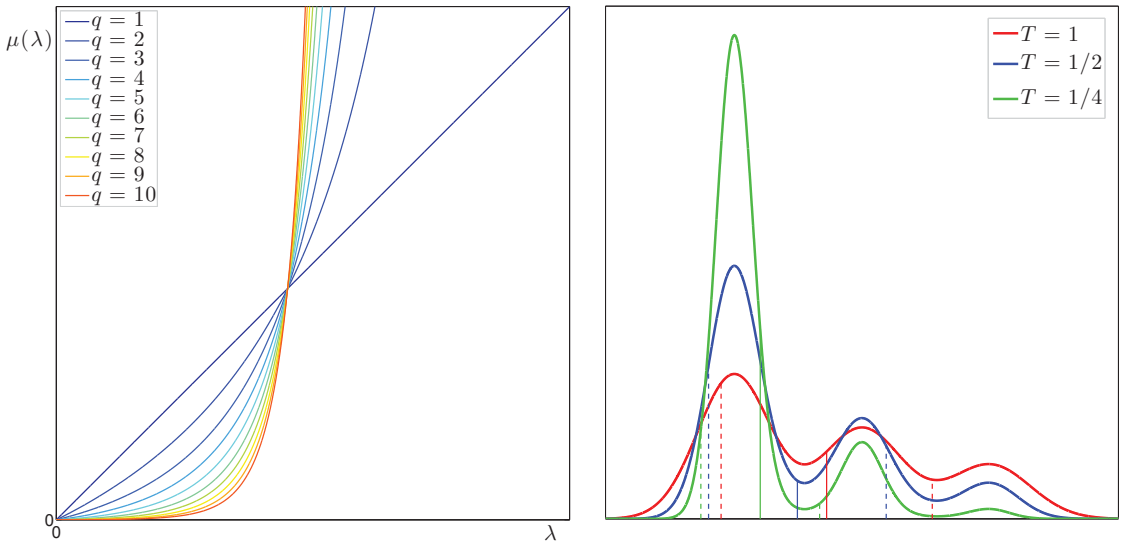
If we define $\lambda_* := \mu r \mathcal{J}(\hat{u}_{\text{MAP}}^r(\mu))^{r-1}$, we see that $\hat{u}_{\text{MAP}}^r(\mu)$ also fulfills the optimality condition for

$$\hat{u}_{\text{MAP}}(\lambda_*) := \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2}\|f - A(u)\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda_* \mathcal{J}(u) \right\}, \tag{4.95}$$

and hence $\hat{u}_{\text{MAP}}(\lambda_*) = \hat{u}_{\text{MAP}}^r(\mu)$. Therefore, by computing a MAP estimate $\hat{u}_{\text{MAP}}(\lambda)$ for a given λ, we obtain $\hat{u}_{\text{MAP}}^r(\mu(\lambda))$ for

$$\mu(\lambda) = \frac{\lambda}{r \mathcal{J}(\hat{u}_{\text{MAP}}(\lambda))^{(r-1)}}, \tag{4.96}$$

given that $\mathcal{J}(\hat{u}_{\text{MAP}}(\lambda)) > 0$. Let $\lambda^0 \in (\mathbb{R}_+ \cup \{\infty\})$ be the smallest λ such that $\mathcal{J}(\hat{u}_{\text{MAP}}(\lambda)) = 0$. As $\mathcal{J}(\hat{u}_{\text{MAP}}(\lambda))$ is decreasing on $(0, \lambda^0)$, (4.96) is a strictly increasing, super-linear function on $(0, \lambda^0)$. Therefore, one can easily compute $\hat{u}_{\text{MAP}}^r(\mu_*)$ for a given μ_* by inverting $\mu(\lambda)$ numerically, i.e., by iteratively fitting λ such that $\mu(\lambda_i)$ is sufficiently close to μ_*. Simple bisection or secant methods can be used for this purpose. In general, the *regularization path* $\{(\lambda_r, \hat{u}_{\text{MAP}}^r(\lambda_r) \,|\, \lambda_r > 0\}$ is equal for all $r \geqslant 1$: For



(a) (b)

Figure 4.6.: (a) $\mu(\lambda)$ for ℓ_1^q priors in the "Boxcar" scenario ($n = 63$). (b) Simulated annealing of a density to optimize (red line). The vertical soild lines indicates the location of the mean, the dashed lines the locations of mean $\pm$ standart deviation. With decreasing temperature more and more probability mass is concentrated in the highest mode, the mean converges to the mode and the standart deviation decreases.

every $\hat{u}^r_{\text{MAP}}(\lambda_r)$, there is a $\lambda_{r'}$ such that $\hat{u}^{r'}_{\text{MAP}}(\lambda_{r'}) = \hat{u}^r_{\text{MAP}}(\lambda_r)$. One could easily generalize the above to $g(\mathcal{J}(u))$, for more general functions $g(\cdot)$ with certain properties.
We will use the above procedure to compute MAP estimates for ℓ_p^q priors $p_{prior}(u) \propto \exp\left(-\mu \left\|D^T u\right\|_p^q\right)$ with $q > p$. In this case, we have $\mathcal{J}(u) = \left\|D^T u\right\|_p^p$ and $r = q/p$. For $p \geqslant 1$, MAP estimates for ℓ_p priors can be computed using ADMM. Figure 4.6a shows $\mu(\lambda)$ for $p = 1$ and various values of q.

4.2.4. Simulated Annealing

The idea of *simulated annealing* (*SA*, Kirkpatrick et al. 1983) is simple: The location of the MAP estimate is invariant to a rescaling of the posterior energy $\mathcal{E}(u)$ by a scalar *temperature* T,

$$\operatorname*{argmax}_u \{p_{post}(u|f)\} = \operatorname*{argmax}_u \{\exp(-\mathcal{E}(u))\} = \operatorname*{argmax}_u \{\exp(-\mathcal{E}(u)/T)\}. \quad (4.97)$$

However, if we define the *tempered posterior* $p^T_{post}(u|f)$ as $\exp(-\mathcal{E}(u)/T)$, the normalization of the distributions changes as well and induces a second, uniform, rescaling of the whole distribution. This interplay between energy rescaling and probability normalization asymptotically concentrates the whole probability in the MAP estimate

of $p_{post}(u|f)$ (see Figure 4.6b). In particular, we have

$$\lim_{T\to 0} \mathbb{E}_{p_{post}^T(u|f)}[u] = \lim_{T\to 0} \frac{\int u \exp(-\mathcal{E}(u)/T)\, \mathrm{d}u}{\int \exp(-\mathcal{E}(u)/T)\, \mathrm{d}u} = \operatorname*{argmax}_u \{\exp(-\mathcal{E}(u))\}. \tag{4.98}$$

This yields a simple stochastic optimization scheme:

Algorithm 4.8. (Simulated Annealing)
Given $p_{post}(u|f)$, an initial point u^0 and an *annealing schedule* $\{T_i\}_{i=1}^{N_T}$, $T_i \in \mathbb{R}_+$, $T_{i+1} < T_i$, repeat for $i = 1, \ldots, N_T$

1. Run a Markov chain for $p_{post}^{T_i}(u|f)$ initialized at u^{i-1} for K_i steps.
   Output: $\{u^{i,j}\}_{j=1}^{K_i}$
2. Set u^i to the last sample u^{i,K_i} of the chain.

Compute

$$\hat{u} := \operatorname*{argmax}_{i,j} \left\{p_{post}(u^{i,j}|f)\right\} \tag{4.99}$$

as an approximation to $\hat{u}_{\mathrm{MAP}}$.

In this formulation, K_i has to be chosen such that the Markov chain reaches its equilibrium phase after cooling. Its choice depends on the annealing schedule $\{T_i\}_{i=1}^{N_T}$. A slow annealing schedule will lead to better results, in particular for multimodal posteriors. Discrete cooling schedules use large values for K_i and a large decrease in T. Neal (1993) advocates the use of a slow but continuous cooling: $K_i = 1$ for all i and a slow decrease in T. We will use such a scheme:

$$T_i = q^i \cdot T_0, \quad \text{where} \quad q := \left(\frac{T_{end}}{T_0}\right)^{(1/N_T)} < 1, \qquad K_i = 1 \quad \forall\, i \tag{4.100}$$

Here, T_0 and T_{end} are predefined start and end temperatures. For unimodal posteriors, $T_0 = 1$ will be chosen, but for multimodal posteriors, $T_0 > 1$ can be used to avoid getting trapped in sup-optimal local modes at the beginning.
All samplers used in this thesis can easily be reformulated to incorporate the energy scaling by T. Details can be found in Appendix A.6.
Simulated annealing is different from deterministic optimization techniques. It is usually applied in combination with MH sampling as a probabilistic, "black-box" metaheuristic for global optimization of complicated, discrete energy functions (Liu 2008, Robert and Casella 2005). Applying it to continuous optimization using Gibbs sampling is rather uncommon, in particular in high dimensional settings like image reconstruction. We will especially use it in situations where no other optimization techniques can be applied, for instance when using non-convex, non-smooth prior energies such as the ℓ_p

energies for $p < 1$. In Section 5.1.5, we will compare its performance to deterministic optimization.

Notes and Comments

Simulated annealing is inspired by the annealing technique in metallurgy: A solid material is exposed to a heat bath with a temperature high enough to melt it. During a subsequent cooling, the particles will spatially arrange in a state of minimal energy. The energy typically features many local minima, reflecting various possible defects of the optimal lattice structure. The temperature corresponds to mobility of the material to switch to new energy states. A fast cooling (*quenching*) will very likely result in the system being trapped in a sub-optimal minimum (which might, however, correspond to a state with desirable material properties). Slow annealing aims to optimize the *crystallinity* of the material.
Simulated tempering (Liu 2008, Robert and Casella 2005) is a related MCMC sampling technique: Aside the main chain, multiple auxiliary chains with $T > 1$ are run in parallel. The chains occasionally switch states which allows the main chain to escape local modes.

4.2.5. Alternating Optimization for Hierarchical Bayesian Models

For ℓ_p hypermodels, optimizing the joint posterior $p_{post}(u, \gamma|f)$ with respect to both u and γ will be based on the same considerations as sampling from it (cf. Section 4.1.11): The conditional construction is exploited in a *block coordinate descent* scheme.

Algorithm 4.9. (Alternating HBM Optimization)
Given $p_{post}(u, \gamma|f)$ and an initial γ^0, repeat for $i = 1, 2, 3, \ldots$

1. $u^i := \operatorname{argmax}_u \{p_{post}(u|f, \gamma^{i-1})\}$
2. $\gamma^i := \operatorname{argmax}_\gamma \{p_{post}(\gamma|f, u^i)\} = \operatorname{argmax}_\gamma \{p_{post}(\gamma|u^i)\}$

until $p_{post}(u^{i+1}, \gamma^{i+1}|f) < p_{post}(u^i, \gamma^i|f)$ or a maximal number of iterations is reached.

Step 1. corresponds to computing a MAP estimate for an ℓ_p prior and the methods introduced in the previous sections can be used. For step 2., we, again, consider the case of factorizing hyperpriors of the form (4.61). This leads to the factorization of the conditional posterior $p_{post}(\gamma|u^i)$; cf. (4.62). Therefore, the optimization over γ consists of a component-wise optimization of all 1D conditional densities:

$$\gamma_j^i = \operatorname*{argmax}_{\gamma_j} \left\{ \exp\left(-\frac{\left|D_j^T u^i\right|^p}{\gamma_j} - (\delta + 1/p)\log(\gamma_j) - \varphi_i(\gamma_j) \right) \right\} \tag{4.101}$$

For the inverse gamma distribution, $p_{prior}(\gamma_j|u^i)$ is, again, an inverse gamma distribution with the parameters $\bar{\alpha}_j := \alpha + 1/p$, $\bar{\beta} := \left|D_j^T u^i\right|^p + \beta$; cf. (3.47). Its mode (4.101) is given by

$$\gamma_j^i = \frac{\bar{\beta}}{\bar{\alpha}+1} = \frac{\left|D_j^T u^i\right|^p + \beta}{\alpha + 1 + 1/p}; \tag{4.102}$$

cf. Section A.1.7 in Lucka (2011). As the posterior with an inverse gamma hyperprior is not log-concave, the alternating optimization Algorithm 4.9 will only converge to a local mode. For the case of $p = 2$, this was examined in Lucka (2011), Lucka et al. (2012). Therefore, heuristic initialization approaches were developed for computing full-MAP estimates with this scheme: It was observed that using full-CM estimates for γ^0 resulted in local modes corresponding to reconstructions of good quality. The result of this heuristic will be called *near-mean* (*NM*) estimate. Another approach is to compute several full-CM estimates from small chains, use all of them in the alternating optimization scheme, and pick the result with the highest posterior probability. In Lucka (2011), it was shown that this strategy most often outperforms NM estimates with respect to the posterior probability. Therefore, the result of this strategy will be referred to as "the" MAP estimate.
For the gamma distribution (3.48), (4.101) is given by

$$\underset{\gamma_j}{\operatorname{argmin}} \left\{ \frac{\left|D_j^T u^i\right|^p}{\gamma_j} + \frac{\gamma_j}{\beta} - (\alpha - 1 - 1/p)\log(\gamma_j) \right\}. \tag{4.103}$$

We can easily compute the optimality conditions:

$$1^{st} \text{ order:} \qquad 0 = \gamma_j^2 - (\alpha - 1 - 1/p)\beta\gamma_j - \left|D_j^T u^i\right|^p \beta \tag{4.104}$$

$$2^{nd} \text{ order:} \qquad 0 \leqslant \frac{2\left|D_j^T u^i\right|^p}{\gamma_j^3} + \frac{\alpha - 1 - 1/p}{\gamma_j^2} \tag{4.105}$$

The second order condition is fulfilled for $\alpha \geqslant 1 + 1/p$. The positive solution of the first order condition is given by

$$\gamma_j = \frac{(\alpha - 1 - 1/p)\beta}{2} + \sqrt{\frac{(\alpha - 1 - 1/p)^2\beta^2}{4} + \left|D_j^T u^i\right|^p \beta} \tag{4.106}$$

Note that if $p \geqslant 1$ and $\alpha > 1$, the whole posterior is log-concave. Therefore, the alternating optimization scheme 4.9 will converge to a global minimum.

4.3. Iterative Optimization and MCMC Sampling

Traditionally, deterministic optimization and stochastic sampling are two very distinct fields of applied mathematics with their own terminology and very little exchange. Therefore, iterative optimization methods such as ADMM seem conceptually very different from the basic MCMC sampling methods such as MH at first glance. While both produce a series $\{x^i\}$ of points in $\mathbb{R}^n$, iterative optimization schemes seem to construct this series in a clear and determined way, while MCMC chains seem to be characterized by a fuzzy, random-walk-like behavior. However, in principle, both series $\{x^i\}$ are constructed to lead to the convergence of a certain quantity in a computationally efficient way:

- In iterative optimization, $\{x^i\}$ is constructed such that $x^i \to \hat{x} \in \operatorname{argmin}\{-\log p(x)\}$ as fast as possible.
- MCMC schemes construct $\{x^i\}$ such that $\frac{1}{K}\sum_i^K g(x_i) \to \int g(x)p(x)\,\mathrm{d}x$ as fast as possible.

The fuzzy, random-walk-like behavior of MCMC chains is clearly not the main aim of MCMC samplers, it is rather an undesired by-product of its construction scheme. Remind that the two "MC's" in "MCMC" refer to two different kinds of randomness:

- Markov chain: The *unwanted*, random-walk-like randomness that one has to tolerate because direct methods to draw independent samples for $p(x)$ are not known.
- Monte Carlo: The *wanted*, independent-samples-like randomness that leads to the convergence of the integral.

Designing fast MCMC schemes essentially means getting rid of the first "MC".
Practically, one can observe a lot of similarities between iterative optimization and sampling:

- Both suffer from similar problems, for instance from strong dependencies between single components. This is a natural feature of inverse problems as the compact forward operator $\mathcal{A}$ "wraps up and compresses" many dimensions, cf. Section 1.2.
- Many algorithms for sampling and optimization are surprisingly similar: We saw in Sections 4.1.2 and 4.2.1 that sampling or optimizing a Gaussian distribution only requires a slight modification of the right hand side of a linear system. In Sections 4.1.11 and 4.2.5, we saw that both sampling and optimization for HBM follow the same alternation scheme (cf. Section 3.2 in Lucka 2011). Both ADMM and the general slice sampler are splitting schemes that decouple the posterior

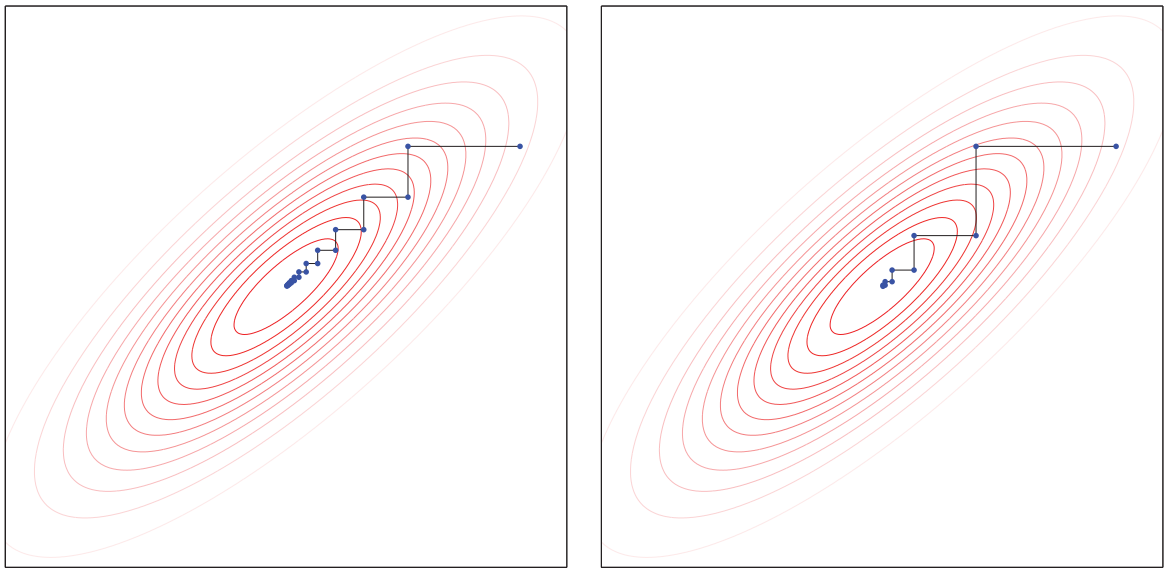
into likelihood and prior parts by introducing auxiliary variables, and alternate between updating them (although we use the slice sampler only for the conditional SC densities, one could also split the complete posterior). Simulated annealing shows how to employ MCMC sampling to construct an optimization scheme.

- In the computational studies in Chapter 5, we will see that both are only computationally efficient if they exploit the analytical structure of $p(x)$, for instance, both ADMM and slice sampling are only efficient if the split is chosen in a way that explicit solutions to the updates of all variables are available.

As a result, it could be fertile to transfer techniques from one field to the other. The field of optimization is older and way better explored compared to computational sampling, which was established in METROPOLIS et al. (1953) and only recently gained practical importance. Hence, transferring optimization techniques to sampling, especially to suppress superfluous randomness, is an important future field of research.

4.3.1. Over-relaxation Techniques

As an example for a transfer from optimization to sampling, we examine how *over-relaxation* can be applied to speed up Gibbs sampling: For a symmetric, positive-definite $G \in \mathbb{R}^{n\times n}$ and a given $c \in \mathbb{R}^n$, consider a Gaussian density $p(x) \sim \mathcal{N}(\nu, \Sigma)$ where $\nu = G^{-1}c$ and Σ are not known or cannot be computed directly. Instead, we can only compute mean $\mu(x, j)$ and standard deviation $\sigma(x, j)$ of the conditional single component densities $p(x_j \big| x_{-j})$ (which are also Gaussian). Therefore, we can only optimize or sample over the single component densities. The iterative optimization is the well-known *Gauss-Seidel method* to solve the linear system $Gx = c$: In each step, x_j is replaced by $\mu(x, j)$ and the iteration repeatedly runs over all components (see Figure 4.7a). The sampling is just the SC Gibbs sampler: In each step, x_j is replaced by $\mu(x, j) + \sigma(x, j)z$, where $z \sim \mathcal{N}(0, 1)$ (cf. Figure 4.3a and Section 4.1.2). Obviously, the convergence of both methods is strongly affected by strong correlations between single components. As noted above, this occurs naturally in typical under-determined inverse problems. For the Gauss-Seidel solver, *successive over-relaxation* (*SOR*, see SAAD 2003) is a well known technique to counteract this coupling between single components in order to increase the convergence rate: In each step, x_j is replaced by $\mu(x, j) + \alpha(x_j - \mu(x, j))$, where $-1 < \alpha < 1$ (often, the equivalent parameterization by $\omega = 1 - \alpha$ is used). If $\alpha < 0$, x_j is *over-relaxed* to the other side of the mean, while $\alpha > 0$ leads to *under-relaxation* or *damping*. Figures 4.7b and 4.8a illustrate how SOR speeds up the convergence of the Gauss-Seidel solver. ADLER (1981) showed that the Gibbs sampler can be accelerated by the same idea: Replacing x_j by $\mu(x, j) + \alpha(x_j - \mu(x, j)) + \sqrt{1 - \alpha^2}\sigma(x, j)z$ leads to an over-relaxed Gibbs sampler that converges faster (see Figure 4.8b). Note that with



(a) (b)

Figure 4.7.: Illustration of successive over-relaxation in the Gauss-Seidel solver: (a) No over-relexation, $\alpha = 0$, (b) over-relaxation, $\alpha = -0.25$

increasing $|\alpha|$, the random part of the update is more and more suppressed while the chain is still ergodic. Hence, this technique is getting rid of the first "MC" in MCMC as discussed above. In Neal (1995), this idea is generalized to arbitrary distributions using *order statistics*:

Algorithm 4.10. (Ordered Over-relaxation, OOR)
The SC sampling step 3.2. in Algorithm 4.3 is replaced by

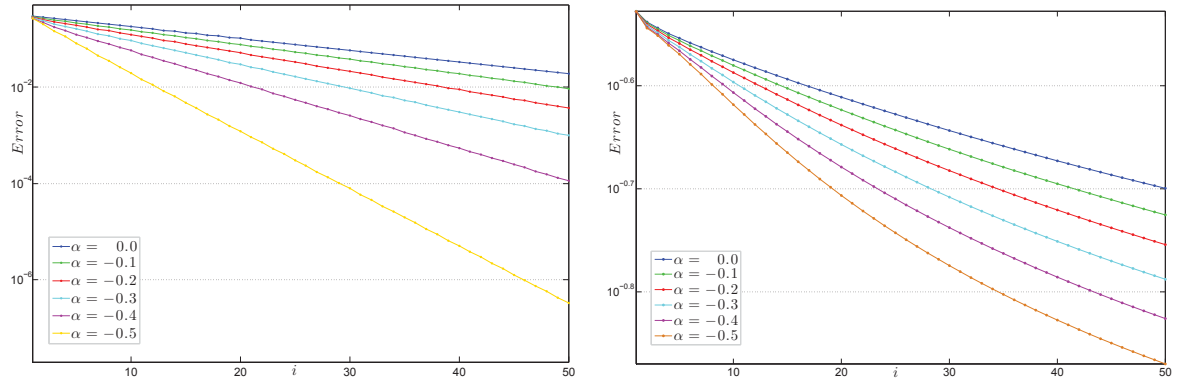
3.2.1. Draw N_O random values $s^1, \ldots, s^{N_O}$ from the conditional, 1D density $p(x_j|x^i_{-j})$, where $N_O \in \mathbb{N}$ is odd.

3.2.2. Sort $s^1, \ldots, s^{N_O}$ plus the current value x^i_j in non-decreasing order, labeling them as follows:

$$s^{(0)} \leqslant s^{(1)} \leqslant \cdots \leqslant s^{(t)} = x^i_j \leqslant \cdots \leqslant s^{(N_O)} \tag{4.107}$$

3.2.3. Set $y = s^{(N_O - t)}$.

Here, N_O functions like an over-relaxation parameter: A large N_O leads to stronger over-relaxation and suppresses more randomness of the sampling process. In the limit of $N_O \to \infty$, we have that $F(y) = 1 - F(x^i_j)$: The update is completely deterministic and mirrors the cdf value of x^i_j around 0.5, i.e., around the median of the distribution. For SC densities that are symmetric around the mean (and mean and median coincide), x^i_j is mirrored at the mean and $p(y|x^i_{-j}) = p(x^i_j|x^i_{-j})$. This gives some intuition why



(a) SOR **(b)** Over-relaxed Gibbs sampler by Adler

Figure 4.8.: Error of (a) SOR iterate, (b) chain mean for different values of α

OOR accelerates the convergence: The mean of $p(\cdot|x^i_{-j})$ can be computed from only two samples: It is given by $(y + x^i_j)/2$.

We can use OOR in all Gibbs samplers that work with univariate distributions:

- For sampling the SC densities in SC Gibbs sampling (see Section 4.1.5).
- For vertical and horizontals moves in the slice sampler (see Sections 4.1.9 and 4.1.10).
- For sampling the univariate conditional hyperparameter posteriors $p_{post}(\gamma_i|u)$ in the HBM-Gibbs sampler (see Section 4.1.11).

The basic Algorithm 4.10 requires N_O times more computation time for these sampling steps. Therefore, its benefits might be overcompensated by the increase of computational time. However, one can often either avoid the linear increase of computational time by a different implementation or neglect it:

Whenever drawing y consists of a monotone, invertible transformation $g(z)$ of another random variable z, OOR can also be performed on z using $z^i_j := g^{-1}(x^i_{-j})$. For instance, in Neal (1995) an implementation of OOR for the inverse cumulative distribution method is given which renders the computation time nearly independent of N_O:

Algorithm 4.11. (ICD Implementation of OOR)
The SC sampling step 3.2. in Algorithm 4.3 is replaced by

3.2.1. Compute $r = F\left(x^i_j\right)$, $r \in [0, 1]$.

3.2.2. Let r' be the ordered over-relaxation of r with respect to the uniform distribution on $[0, 1]$ and N_O (computed with Algorithm 4.10).

3.2.3. Replace y by $F^{-1}(r')$.

Usually, the first and the third step are computationally most demanding. As they

do not depend on N_O, the computation time is nearly independent of N_O. In Lucka (2012), this scheme was used to implement OOR for the direct ℓ_1 sampler (cf. Section 4.1.8).
Even if using Algorithm 4.11 is not possible, one can often neglect the additional computational load: In SC Gibbs sampling, computing the parameters of the SC densities requires way more computation time than sampling from them. As a result, one can spend more time on this sampling step without increasing the total computational time in a significant way. In the HBM-sampler, the same is true for updating the hyperparameters: Sampling $p_{post}(u|f,\gamma)$ is the computational bottleneck as it involves the forward operator. In slice sampling, the situation is different, but for multimodal densities, OOR might increase the probability of the sampler to escape from a mode.
We will denote the over-relaxed versions of SSG, RSG and RPSG by by appending "-ON_O": For instance, "SSG-O7" denotes the systematic scan Gibbs Sampler with ordered over-relaxation using $N_O = 7$. A detailed, computational examination of over-relaxation will be carried out in Section 5.1.4.

4.3.2. Notes and Comments

The example of using over-relaxation to speed up the optimization or sampling of Gaussian densities was used for illustrative reasons only. In most situations, SOR (or other stationary solvers) are no longer used to solve linear systems, in particular large, sparse ones. Instead, conjugate gradient methods are often employed (cf. Section 4.2.1). Interestingly, the concepts behind conjugate gradient optimization can also be transferred to sample high dimensional Gaussians with a sparse correlation matrix, see Schneider and Willsky (2003) and Parker and Fox (2012).

4.4. Computation of Recovery Conditions

In this section, we will discuss how the recovery conditions introduced in Section 3.5 can be tested computationally. The null space property (NSP) can usually not be verified by direct computations, only a falsification by Monte Carlo simulation is possible. The other recovery conditions were developed because of this shortcoming.
Computing the coherence numbers $\mu(A)$, $\mu_{blk}(A)$ and $\mu_{sub}(A)$ is trivial if A is given as a matrix of moderate size. In other cases (which are not relevant to this thesis) the computations need to be arranged in a suitable way.
It is hard to compute the RIP constant δ_k in a direct way, but we can bound it by brute-force Monte Carlo computations:

Algorithm 4.12. (RIP Bounds)
For A with coherence $\mu(A)$, the RIP constant δ_k can be bounded by $[\delta_k^{lb}, \delta_k^{ub}]$, where $\delta_k^{ub} = \mu(A)(k-1)$ and the lower bound δ_k^{lb} can be refined iteratively:
Set $\delta_k^{lb} = 0$, $n_{last} = 0$, $\delta_{diff} = 0$. For $i = 1, 2, 3, \ldots$ do

1. Generate $u \in \mathbb{R}^n$ with $|u|_0 = k$ and $\|u\|_2 = 1$ uniform at random.
2. Compute $\delta_u = |1 - \|Au\|_2|$
3. If $\delta_u > \delta_k^{lb}$, set $n_{last} := 0$; $\delta_{diff} := \delta_u - \delta_k^{lb}$; $\delta_k^{lb} := \delta_u$. Else, set $n_{last} := n_{last} + 1$.

One can stop the iteration once $[\delta_k^{lb}, \delta_k^{ub}]$ is sufficiently narrow, the recovery condition to be checked is falsified or if n_{last}, the number of samples since the last change of δ_k^{lb}, is sufficiently large.

Although this brute-force algorithm is not very elegant, it will suffice for our computational studies. The block variant, where one can further use that $\delta_k \leqslant \delta_{[k]} \leqslant \delta_{dk}$, is straightforward to derive.
Conditions (Tr), (FuA), (BlkTr), (BlkFuA) are easy to compute. Condition (FuB) is a *feasibility problem* with linear equality and inequality constraints:

$$(\text{FuB}) \overset{comp.}{\Longleftrightarrow} \begin{bmatrix} A_{I^c}^T \\ -A_{I^c}^T \end{bmatrix} w \leqslant 1 \quad \text{and} \quad A_I^T w = \operatorname{sign}(u_I^\dagger), \tag{4.108}$$

where "$\leqslant$" is understood in a component-wise sense and the "computational equivalence" should indicate that the strict distinction between "$\leqslant$" and "$<$" is difficult. For computing (SSC_+), the inequality constraint simplifies to $A_{I^c}^T w \leqslant 1$. We will use *primal-dual interior point methods* to tackle these problems (see Boyd and Vandenberghe 2004, for a general reference). Depending on the study design, a direct Matlab implementation (through `linprog.m`) or a MOSEK or SeDuMi implementation (interfaced through CVX) are used. See Section A.8 for an overview of the software used. In addition to testing (FuB) by (4.108), we may also want to minimize $\left\|A_{I^c}^T w\right\|_\infty$ or $\|w\|_2$ on the set of feasible points (the latter for the error estimates (3.79) and (3.80)). In Matlab, we can use `quadprog.m` for minimizing $\|w\|_2$ and reformulate the minimization of $\left\|A_{I^c}^T w\right\|_\infty$ to a linear program in standard form (solved by `linprog.m`). Using CVX, both problems can be formulated in an easy way. In (BlkFuB), we have *conic* instead of linear inequality constraints:

$$(\text{BlkFuB}): \qquad \left\|(A^T w)_{[j]}\right\|_2 < 1 \quad \forall j \notin \mathcal{I} \qquad \text{and} \qquad A_{[\mathcal{I}]}^T w = \xi_{[\mathcal{I}]} \tag{4.109}$$

As conic constraints are also convex, interior point methods can be used to tackle such a *second-order cone* feasibility problem (Boyd and Vandenberghe 2004) as well.

We use MOSEK or SeDuMi implementations (interfaced through CVX). Additional convex functions such as $\max_{j \notin \mathcal{I}} \left(\left\| (A^T w)_{[j]} \right\|_2 \right)$ or $\|w\|_2$ can be incorporated as well. Both optimization problems are very ill-conditioned. To stabilize them, imposing additional upper and lower bound constraints $lb \leqslant w \leqslant ub$ is often necessary. Still, one may also need to try a sequence of different solver implementations as no single implementation works for all $u^\dagger$. For instance, the *active-set* and even the *simplex* implementations of the linear solver `linprog.m` in MATLAB have to be used if the interior point implementation fails.

5 Computational Studies

In this chapter, we will examine the aspects discussed in the last chapters by numerical studies. The computational methods introduced in the previous chapter will be investigated in Section 5.1. Section 5.2 contains a collection of simulated data studies on various conceptual aspects of Bayesian inversion while Sections 5.3 and 5.4 contain application specific studies for CT and EMEG. In particular, they will discuss the challenges of experimental data scenarios.

5.1. Evaluation of the Computational Methods

Whenever computation times are discussed, special attention was paid that implementations and computational platforms used were as comparable as possible. In addition, MATLAB was limited to a single computational thread (parallelization is discussed in Section 7.4). However, all computation time comparisons should give a general idea about the behavior and applicability of the algorithms rather than absolute figures.

5.1.1. Prior Sampling

We start by using the Gibbs samplers to sample different prior distributions. This illustrates which kind of results they promote. Figure 5.1 shows the results.

5.1.2. Comparison of MCMC Samplers for ℓ_1 priors

In this section, a comparison between MH and SC Gibbs samplers for ℓ_1 priors is carried out. The computations are based on more extensive studies published in LUCKA (2012), but were partly rearranged and recomputed for this thesis.

Table 5.1.: Burn-in length K_0 for the Boxcar scenario using a TV prior, $u^0 = 0$ and different combinations of (n, λ). All samplers use an SSR of n.

	(63,100)	(63,200)	(63,400)	(127,280)	(255,400)	(511,560)	(1023,800)
MH-Iso	4.0e2	8.0e2	4.0e3	4.0e3	1.3e4	6.0e4	2.0e5
MH-Si	4.0e2	1.0e3	5.0e3	5.0e3	1.5e4	6.0e4	2.0e5
RSG	2.0e2	2.0e2	2.0e2	8.0e1	5.0e1	3.0e1	2.0e1
SSG	4.0e2	5.0e2	5.0e2	1.5e2	1.0e2	1.5e2	1.5e2

Visual Comparison

We start by a visual impression of the convergence of the MCMC samplers in the "Spots" scenario. We use $n = 513 \times 513 = 263\,169$, and a simple ℓ_1 prior ($D = I_n$) with $\lambda = 2 \cdot 10^7$. In Figure 5.2, the CM estimates obtained after 1, 4 and 16 hours of computational time are shown for MH-Iso, MH-Si, RSG and SSG. The burn-in length K_0 (cf. Section 4.1.6) used to compute the CM estimates was determined in a pre-study: For RSG, $K_0 = 15$ and for SSG, using $K_0 = 25$ is sufficient. The burn-in analysis for the MH samplers revealed that even after 20 hours of computation, the chain was still far away from the stationary phase. Therefore, each CM estimate shown was computed discarding the first half of the samples generated so far. Choosing a good color scale to compare the results for a single sampler is not easy because of outliers in the 1h image: These outliers would lower the contrast if a simple linear min-max scaling based on all images is chosen. Using an individual scaling for each image would, instead, lead to the impression that the value of $\hat{u}_{\mathrm{CM}}$ in certain regions changes although it is actually constant. Therefore, we use the following scaling: For each method, we merged and sorted the absolute pixel values of all three CM estimates. From this sorted set, the largest 0.1% elements are discarded. All estimates are divided by the maximum of the remaining values and clipped to $[-1, 1]$.
In these first, qualitative results, Gibbs and MH samplers show a very different performance. In the following sections, we try to quantify this impression.

Burn-In Analysis

We perform the quantitative studies using the Boxcar scenario and the TV prior. Two different combinations of n and λ_n are examined here:

1. $n = 63$ in combination with $\lambda = 100$, 200 and 400, respectively. This focuses on increasing the impact of the prior. The posterior becomes less and less Gaussian because the weight of the ℓ_1 prior is increased.

2. $n = 2^N - 1$ for $N = 7, 8, \ldots$ with $\lambda_n = 25 \cdot \sqrt{n+1}$. This scaling is related to the

Table 5.2.: Estimated integrated autocorrelation time $\hat{\tau}_{int}$ for the "Boxcar" scenario using a TV prior and different combinations of (n, λ). All samplers use an SSR of n.

	(63,100)	(63,200)	(63,400)
MH-Iso	(1.77±0.06)e2	(5.27±0.28)e2	(1.14±0.09)e3
MH-Si	(1.27±0.01)e1	(1.89±0.02)e1	(3.22±0.05)e1
RSG	(4.59±0.13)e2	(3.39±0.08)e2	(2.39±0.05)e2
SSG	(2.43±0.05)e2	(1.75±0.03)e2	(1.19±0.02)e2

	(127,280)	(255,400)	(511,560)	(1023,800)
MH-Iso	(2.63±0.21)e3	(4.73±0.48)e3	(1.33±0.23)e4	(9.01±3.74)e4
MH-Si	(6.78±0.15)e1	(2.07±0.05)e2	(1.97±0.39)e4	(6.18±2.75)e4
RSG	(7.18±0.33)e2	(9.78±0.25)e1	(3.65±0.04)e0	(0.53±0.00)e0
SSG	(3.61±0.12)e2	(5.67±0.14)e1	(1.99±0.02)e0	(0.26±0.00)e0

study of the discretization invariance of the TV prior (cf. Section 3.6.1) which will follow in Section 5.2.2.

Table 5.1 lists the burn-in length K_0, determined as described in Section 4.1.6. The burn-in length is an important factor for the practicability of the algorithms, but it is a difficult measure for a fair and definite comparison of the sampling methods: First, it crucially relies on the initialization of the chain. Hence, one would have to compare all methods for various common initialization strategies, which would be infeasible and too application specific. Second, K_0 also relies on the κ-adaptation strategy for the MH schemes (cf. Section 4.1.4). We always use a piecewise linear scaling function $\tau_\kappa(\alpha_\kappa)$ such as the red plot in Figure 4.2b. The parameters are $\alpha_\kappa^{opt} = 0.2341$, $\alpha_\kappa^l = 0.1841$, $\alpha_\kappa^u = 0.2841$, $\tau_\kappa^1 = 2.5$, $\tau_\kappa^0 = 0.375$ $\tau_\kappa^l = 0.95$, $\tau_\kappa^u = 1.05$. While these additional tuning parameters render the problem of a meaningful comparison worse, their influence on K_0 decreases with increasing n: For $n = 1023$, the κ adaptation is usually finished after about 10^2 samples, i.e., after a small fraction of K_0. Despite these problems, Table 5.1 indicates a clear trend, which we also observed in Figure 5.2: The burn-in length of MH samplers increases with n and λ while it stays constant or even decreases for the SC Gibbs samplers. For $n = 1023$, the computation time for completing the burn-in phase is already about 18 hours for the MH samplers while it is about half a second for the RSG sampler.

Autocorrelation Analysis

As discussed in Section 4.1.6, we need to define a suitable test function $g(u)$ to perform an autocorrelation analysis. As we want to examine different n, it should be a function

Table 5.3.: Estimated integrated autocorrelation time $\hat{\tau}_{int}$ of the RSG sampler in the "Spots" scenario using a simple ℓ_1 prior and different combinations of $(N,\ \lambda)$: $n = (2^N - 1)^2$; SSR $= (2^{N-2} - 1)^2$.

(5,8e5)	(6,1e6)	(7,5e6)	(8,1e7)
12.68±0.26	7.68±0.13	5.75±0.12	6.56±0.29

that can be consistently defined for all n and captures the main variability of the posterior in some way. For the "Boxcar" scenario, we will project onto the direction of the largest variance, i.e., the first eigenvector ν_1 of the covariance matrix of the posterior:

$$g(u) := \langle \nu_1, u \rangle \tag{5.1}$$

In general, it should be challenging for all MCMC samplers to reduce the correlation of subsequent samples in this direction ν_1. For each scenario we examine, the covariance matrix of the posterior was estimated from a long chain of the RSG sampler, as this sampler will turn out to be the most reliable at a high performance. Note that this choice does not give an advantage to the RSG sampler in the autocorrelation analysis: It is rather a disadvantage if other samplers would have different directions of highest variance. We checked that this is not the case in a test scenario examined in preliminary studies.

Table 5.2 lists the estimates $\hat{\tau}_{int}$ of the integrated autocorrelation times. Note that the accurate estimation of a large τ_{int} naturally requires computing a very long chain (some of the acf computations in this thesis took up to two weeks). For $n = 1023$ and SSR $= n$, the integrated autocorrelation time τ_{int} of the RSG sampler corresponds to about 15ms computation time whereas τ_{int} of MH-Iso corresponds to about 3h. Thereby, the results confirm the visual impression and the trends observed in the burn-in analysis: The efficiency of MH samplers dramatically decreases when n or λ is increased. In contrast, the Gibbs samplers show the opposite behavior. While the results for the MH samplers could have been anticipated from previous studies (e.g., Comelli 2011, Kolehmainen et al. 2012, Lassas and Siltanen 2004), the results for the Gibbs samplers come as a surprise. To gain confidence in them, we also performed autocorrelation studies in other imaging scenarios. For the "Spots" scenario using a simple ℓ_1 prior ($D = I_n$), we also found that the burn-in length decreases to around 20 with n increasing. For defining the test function $g(u)$ to estimate τ_{int}, we cannot compute the covariance matrix anymore. Instead, we simply took the projection u onto the green box in Figure 5.2, i.e., $g(u)$ sums up all components of u inside this image area. The estimates $\hat{\tau}_{int}$ can be found in Table 5.3. The λ's for the different n were chosen by visual inspection. We will later examine the use of Besov space priors with Haar wavelets and $p = 1$ (cf. Section

3.2.4) in the "Phantom-CT" scenario. Therefore, we also performed an extensive burn-in and autocorrelation analysis of the RSG sampler for scenarios from $n = 4\,096$ up to $n = 1\,048\,576$ using different values for λ. A detailed record of the results is omitted here: The burn-in length again decreases to around 10-30 with increasing n. As a test function $g(u)$, a projection of the wavelet coefficients of u onto the wavelet coefficients of $u^{\dagger,\infty}$ was used. For the values of λ considered later on, and an SSR of n, the estimate for τ_{int} was always close to 1. Using a lower SSR, the trend that τ_{int} is decreasing with increasing λ could also be replicated.

Discussion

More detailed results and discussions can be found in Lucka (2012). Here, we only stress the points important for the following studies:

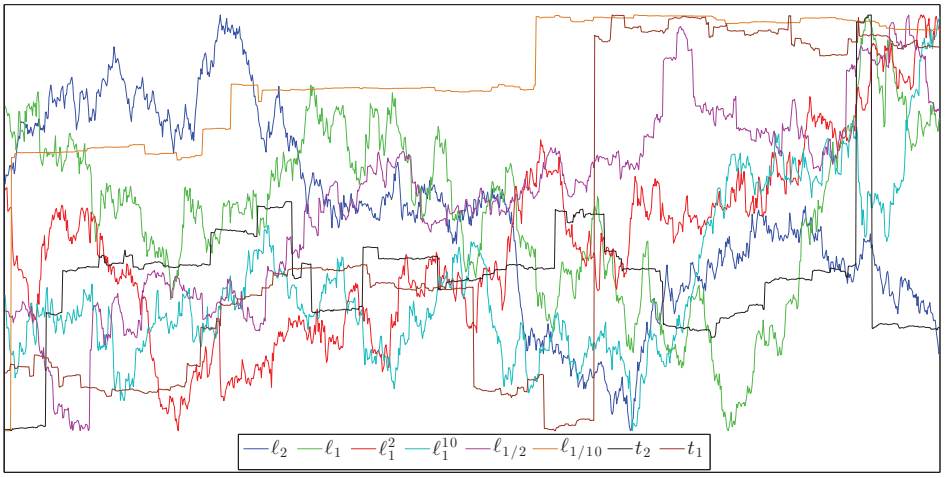
MH The quantitative studies showed that the efficiency of the basic SRWMH samplers dramatically decreases when either the influence of the ℓ_1 prior increases (i.e., λ is increased) or the dimension of the unknowns, n, is increased. For the largest examined number of unknowns, $n = 1023$ (which is still moderate for typical inverse problem scenarios), both the burn-in and integrated autocorrelation time are in the order of a few hours. The visual results in Figure 5.2 clearly confirm this. Furthermore, our findings are in line with those of former applications of SRWMH samplers to similar inverse problems scenarios with ℓ_1 priors (e.g., in Comelli 2011, Kolehmainen et al. 2012, Lassas and Siltanen 2004).

Gibbs The visual results in Figure 5.2 already suggested that the SC Gibbs samplers are applicable to high dimensional scenarios. The quantitative studies further reveal the surprising trend that their efficiency actually increases with increasing n and λ. While the burn-in length for high dimensions saturates around 10-30 complete sweeps (SSR $= n$), the integrated autocorrelation time gets so small that using SSR $< n$ becomes reasonable. Comparing RSG and SSG, Tables 5.1 and 5.2 show that the SSG sampler usually requires a longer burn-in time while having a shorter $\hat{\tau}_{int}$. A more detailed examination in Section 5.1.4 will reveal the cause for this phenomenon: $\hat{R}(\tau)$ becomes negative for SSG, i.e., it produces *anti-correlated* samples.

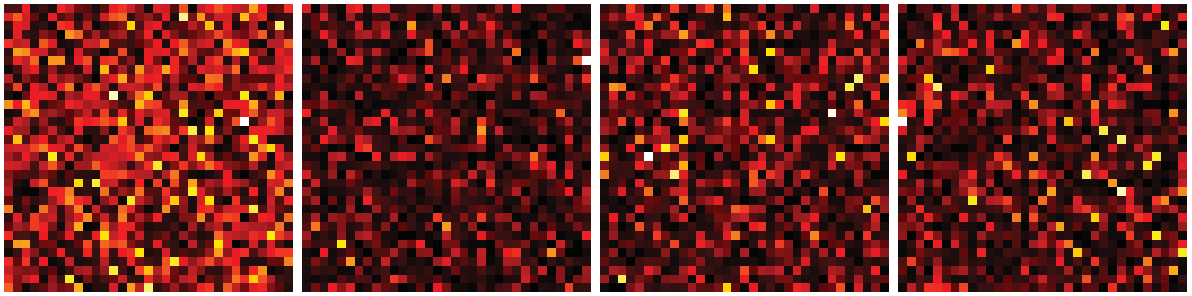
General There are multiple reasons for the loss of performance of the basic SRWMH samplers compared to the SC Gibbs samplers in the specific scenarios examined here. The crucial part for an MH sampler is the design of a good proposal distribution (cf. Section 4.1.4). The basic MH samplers we applied are "black-box sampler" algorithms. In the design of their proposal distributions, no specific information about the posterior is

taken into account. In return, they exhibit short computation times. They are designed to sample from low dimensional, Gaussian-like distributions. However, high dimensional posteriors from sparsity promoting priors seem to have very different properties. As a result, the SRWMH-samplers have to take very small steps to obtain a good acceptance rate. This leads to long burn-in times and a slow decrease in autocorrelation. SC Gibbs samplers incorporate more posterior-specific information into the sampling procedure at the costs of a larger computation time (cf. Section 4.1.5). The conditional SC densities are the optimal transition densities for the conditional move but have to be computed and sampled explicitly. Incorporating this small extra amount of problem specific information seems to be sufficient to generate an efficient sampling procedure even for high dimensional scenarios: In the upcoming studies, we will use Gibbs sampling up to $n > 10^6$. This is far beyond any previously reported use of MCMC for similar scenarios (typically, $n = 10^3 - 10^4$ is regarded as the feasible limit).

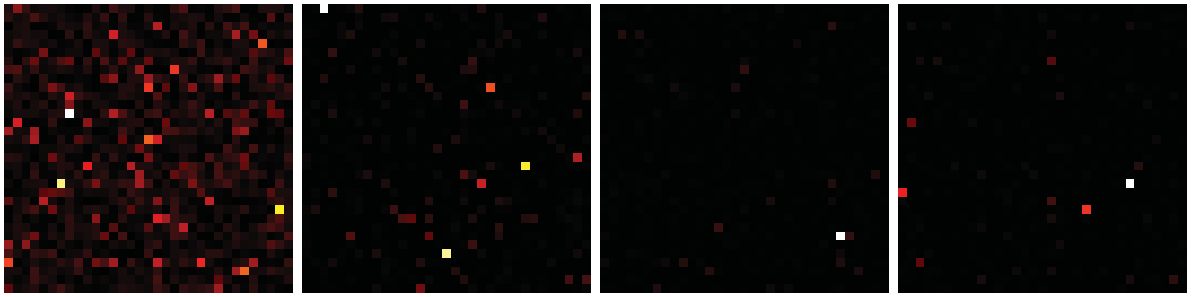
One should stress that our findings only apply for the specific scenarios we examined: In general, both MH and Gibbs sampling have advantages and disadvantages and will outperform the other given a specific scenario. However, the surprising properties of the SC Gibbs sampler enabled and motivated many of the investigations in this thesis. In particular, it motivated the extension of the direct sampler for specific ℓ_1 priors as introduced in Lucka (2012) to other important prior models by the slice sampler (cf. Section 4.1.10).



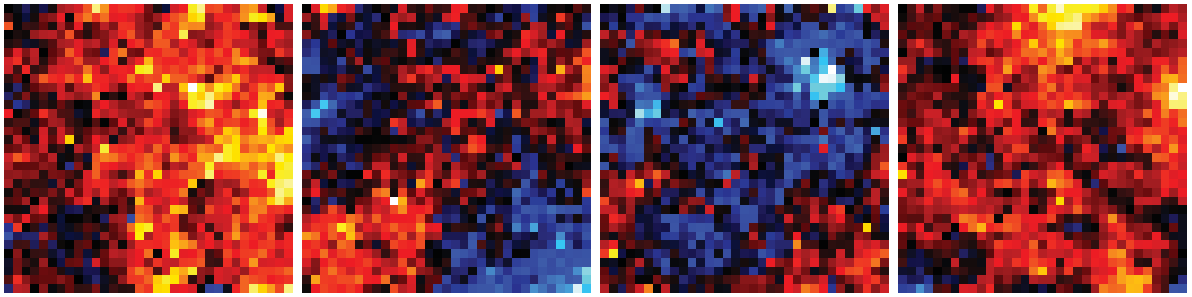
(a) Increment priors



(b) ℓ_2 prior **(c)** ℓ_1 prior **(d)** ℓ_1^2 prior **(e)** ℓ_1^{10} prior

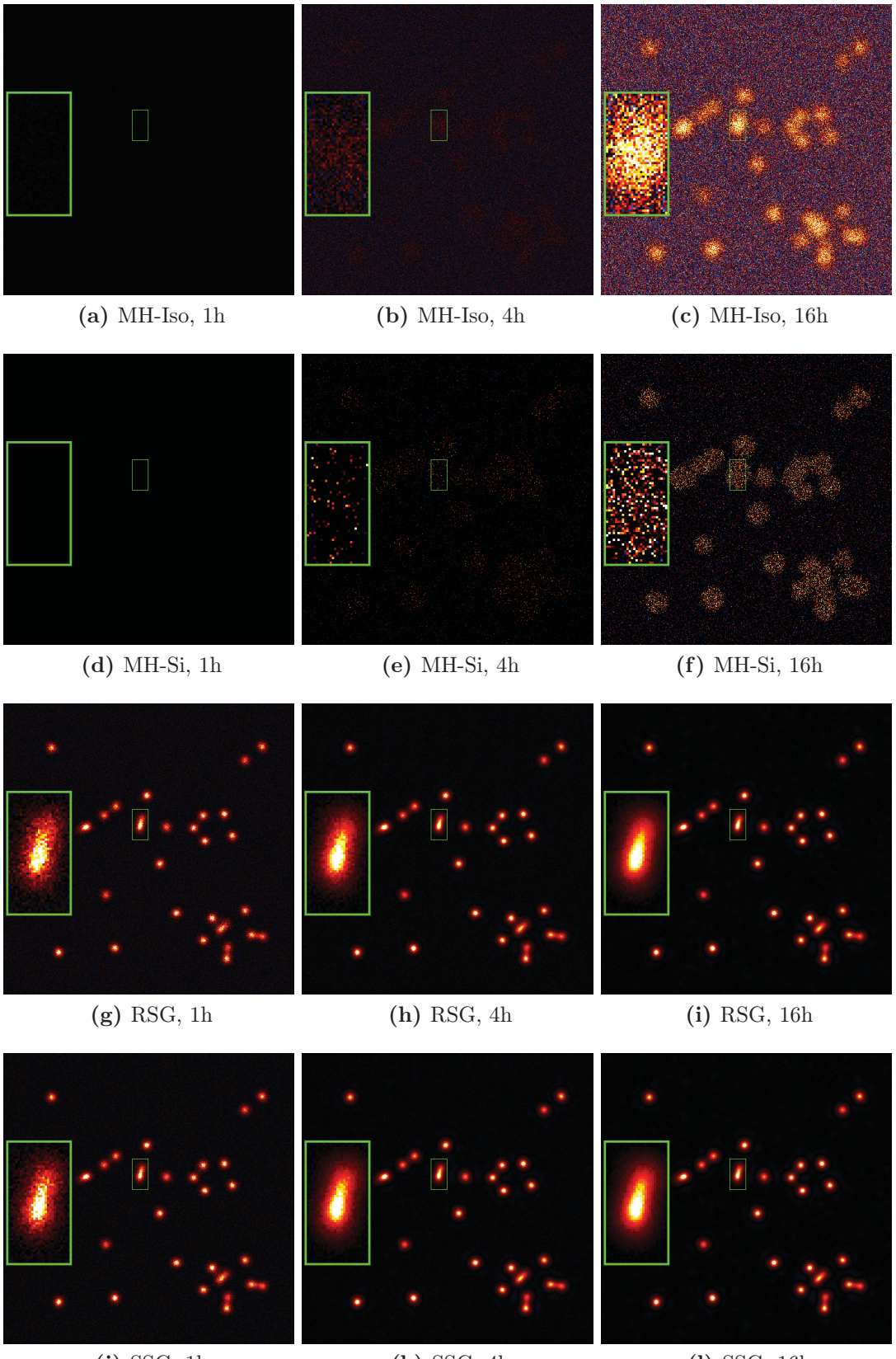


(f) $\ell_{1/2}$ prior **(g)** $\ell_{1/10}$ prior **(h)** product t_2-prior **(i)** product t_1-prior



(j) TV prior, sample 1 **(k)** TV prior, sample 2 **(l)** TV prior, sample 3 **(m)** TV prior, sample 4

Figure 5.1.: (a) Prior samples in the "Boxcar" scenario using $n = 1023$ and D as in (3.18). The functions were rescaled and recentered such that $\max(u) = 1$, $\min(u) = -1$. (b)-(i) Prior samples in the "Spots" scenario using $n = 33 \times 33$, $D = I_n$ and non-negativity constraints. The functions were rescaled such that $\max(u) = 1$. (j)-(m) Isotropic TV prior samples in the "Spots" scenario using $n = 33 \times 33$ and the constraint $u_1 = 0$. The functions were rescaled such that $\|u\|_\infty = 1$.



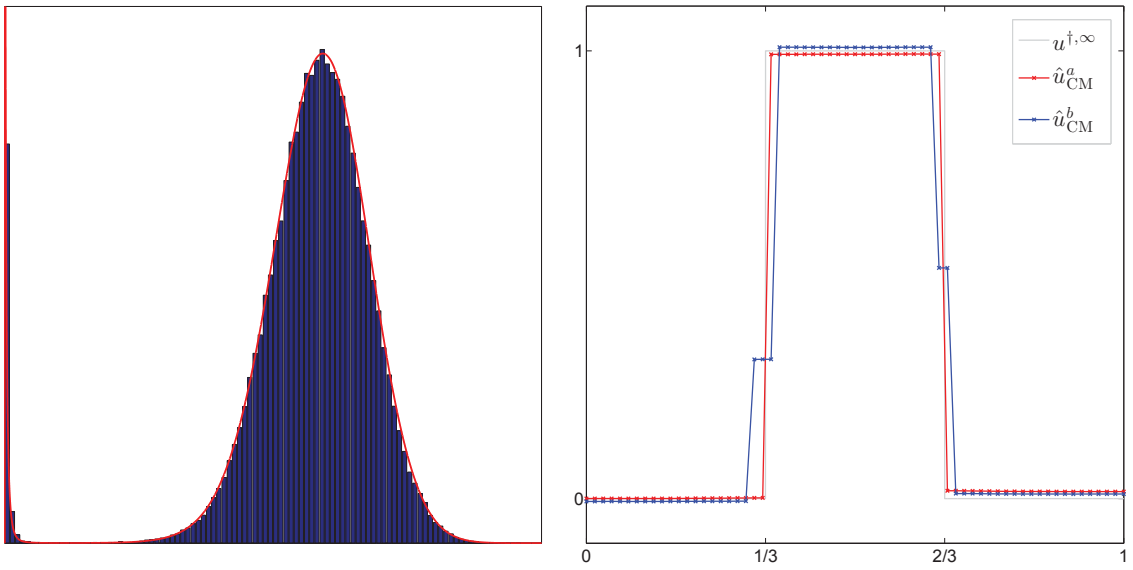
(a) MH-Iso, 1h **(b)** MH-Iso, 4h **(c)** MH-Iso, 16h

(d) MH-Si, 1h **(e)** MH-Si, 4h **(f)** MH-Si, 16h

(g) RSG, 1h **(h)** RSG, 4h **(i)** RSG, 16h

(j) SSG, 1h **(k)** SSG, 4h **(l)** SSG, 16h

Figure 5.2.: Visual results for the "Spots" scenario using $n = 513 \times 513$ and a simple ℓ_1 prior with $\lambda = 2 \cdot 10^7$. The inset shows a zoom into the marked area in the original figure. The scaling used in these images is explained in the text.



(a) (b)

Figure 5.3.: (a) Histogram (blue bars) of the slice sampler compared to targeted SC density (red line here and in Figure 4.5a). (b) For the "Boxcar" scenario, two approximations to the CM estimate using an $\ell_{1/2}$ increment prior were computed from two independent chains (each about 16h of computational time).

Table 5.4.: Comparison of τ_{int} for direct and slice-within-RSG samplers using different burn-in lengths for the slice sampler. The "Boxcar" scenario with $n = 255$, a TV prior with $\lambda = 400$, an SSR of n and a projection onto the largest eigenvector of the covariance matrix as $g(u)$ was used.

direct	$K_0^{ss} = 200$	$K_0^{ss} = 100$	$K_0^{ss} = 40$	$K_0^{ss} = 20$	$K_0^{ss} = 10$
97.8±2.5	101.3±2.6	102.0±2.6	109.4±2.9	149.2±4.6	231.4±8.6

5.1.3. Examination of the Slice Sampler

First, extensive studies were carried out to verify that the slice samplers developed in Section 4.1.10 accurately reproduce all the SC densities they aim to sample from. For this, the convergence of the sample histograms to the underlying SC densities was checked visually. Figure 5.3a shows such a comparison.

When using slice sampling as a 1D sampler within a SC Gibbs sampler, the length of the burn-in phase of the slice sampler, K_0^{ss}, determines how well a direct 1D sampler is resembled. We studied its influence on the efficiency of the SC Gibbs sampler in terms of burn-in length K_0 and integrated autocorrelation time τ_{int} as in the previous section. Table 5.4 lists the results for τ_{int} for a scenario, where the direct ℓ_1 sampler examined in the last section can be used as a reference. One can observe that already for small K_0^{ss}, the differences between direct and slice sampler are negligible in practice. Concerning

K_0, a slightly longer burn-in length was only detected for $K_0^{ss} = 10$. From $K_0^{ss} \geqslant 20$ on, no visual difference was observed (cf. Section 4.1.6). Similar examinations using ℓ_2 priors (where, again, a direct sampler can be used as a reference) showed that in this case, significant differences vanish for even smaller values of K_0^{ss}. Tables 5.6a, 5.6b and 5.6c show the results of similar examinations for an ℓ_p prior with $p = 1.2$, an ℓ_p^q prior with $p = 1$, $q = 10$ and the isotropic TV prior in 2D, respectively. While we do not have a direct sampler as a reference here, one can clearly see that τ_{int} is converging to a limit for increasing K_0^{ss} (the results in Table 5.6c seem to suggest that in some cases, even using $K_0^{SS} = 1$ might be sufficient).

While the application to such log-concave priors was the main motivation behind the development of slice-within-Gibbs samplers, the application to non-log-concave priors like ℓ_p for $p < 1$ or product t_p priors seems tempting. However, the examination of the slice-within-Gibbs samplers in this situation is considerably more difficult: While tests such as shown in Figure 5.3a suggest that the slice sampler is able to reproduce the SC densities, they also show that the number of steps required for switching between two modes of the SC density might be quite large. This difficulty of switching between *conditional* SC modes adds to the general difficulty of the SC Gibbs sampler to switch between the *unconditional* modes of the full posterior in $\mathbb{R}^n$. The latter would also occur if direct samplers would be used for the SC densities. Both difficulties interact in a non-trivial way. Often, visual results can be deceiving about the true extent of the problem: In Figure 5.3b, we compare two CM estimates using an ℓ_p increment prior with $p = 0.5$, each computed using 16 hours of computation time. The first, red plot is an almost perfect reconstruction, which gives confidence in the sampler as well. However, the second, blue plot clearly indicates that the sampler got stuck in a sub-optimal mode and could not escape from it. As discussed in Section 4.1.6, autocorrelation analysis may fail in such a case as well. For instance, we computed CM and CCov estimates for the "Boxcar" scenario with $n = 63$ and a product Cauchy increment prior ($\theta = 10^{-6}$) using one day of computational time. The first eigenvector was used for an autocorrelation analysis as in the previous section. While $\hat{\tau}_{int}$ estimated τ_{int} to 12.06 ± 0.25, $\hat{\tau}_{int}^{ref}$, which used the projection of the pre-computed CM estimate as $\hat{\mu}$, estimated τ_{int} to $78\,788 \pm 35\,676$.

Discussion

The slice-within-Gibbs samplers were mainly developed for ℓ_p priors with $1 < p < 2$, ℓ_p^q priors with $p = 1$, $1 < q$, the TV prior in 2D, (3.22), and other ℓ_1-block priors. The results show that for these log-concave priors, a good performance can already be obtained when using only a few steps of the slice sampler. Thereby, the computational

Table 5.5.: Comparison of τ_{int} for slice-within-RSG samplers using different burn-in lengths for the slice sampler.

(a) The "Boxcar" scenario with $n = 255$, an ℓ_p increment prior with $p = 1.2$, $\lambda = 400$, an SSR of n and a projection onto the component with the largest variance as $g(u)$ is used.

$K_0^{ss} = 1$	$K_0^{ss} = 2$	$K_0^{ss} = 4$	$K_0^{ss} = 8$	$K_0^{ss} = 16$	$K_0^{ss} = 32$	$K_0^{ss} = 64$
41.9±1.1	33.3±0.8	23.4±0.5	18.3±0.3	15.8±0.4	14.6±0.3	14.8±0.3

(b) The "Boxcar" scenario with $n = 255$, an ℓ_p^q increment prior with $p = 1$, $q = 10$, $\lambda = 0.02$, an SSR of n and a projection onto the largest eigenvector of the covariance matrix as $g(u)$ is used.

$K_0^{ss} = 1$	$K_0^{ss} = 2$	$K_0^{ss} = 4$	$K_0^{ss} = 8$	$K_0^{ss} = 16$	$K_0^{ss} = 32$	$K_0^{ss} = 64$
638±46	425±26	307±16	198±9	161±6	155±7	135±6

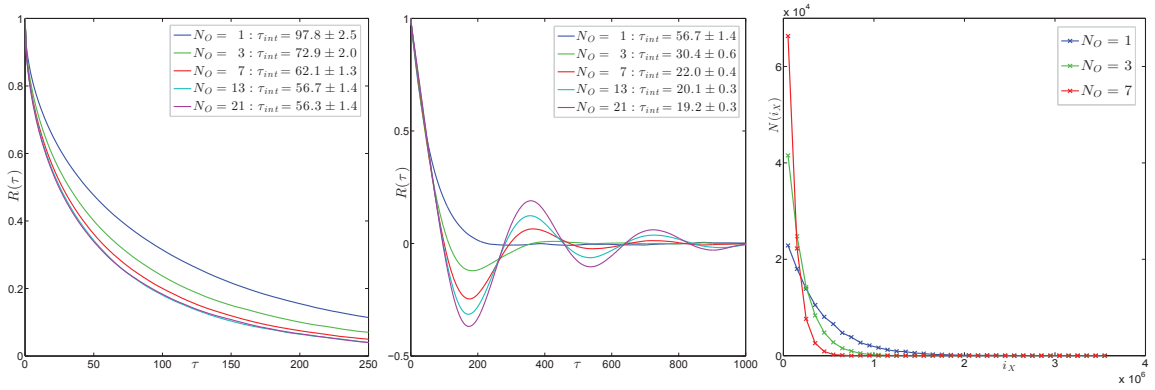
(c) The "Phantom-CT" scenario with $n = 129 \times 129$, an isotropic TV prior with $\lambda = 500$, an SSR of 4096 and a projection onto the green box in Figure 5.10e as $g(u)$ is used.

$K_0^{ss} = 0$	$K_0^{ss} = 1$	$K_0^{ss} = 2$	$K_0^{ss} = 4$	$K_0^{ss} = 8$	$K_0^{ss} = 16$	$K_0^{ss} = 32$
23.8±1.4	21.2±1.1	21.3±1.2	22.3±1.2	20.9±1.1	19.5±1.0	20.8±1.1

time it takes to sample the SC density is still considerably shorter than the time it takes to compute its parameters in high dimensional scenarios (cf. Section 4.1.7). The application of slice-within-Gibbs sampling to non-log-concave prior models is in an experimental state up to now. For such scenarios, the question of whether slice sampling is an adequate technique to sample the SC densities is in fact of secondary importance. First, the general potential of SC Gibbs sampling for non-log-concave prior models has to be examined. For product t_p priors with $1 \leqslant p$, using a blocked Gibbs sampler for an ℓ_p hypermodel (cf. Sections 4.1.11 and 3.3.3) is an attractive alternative which can be used to tackle this question. In any case, new performance measures need to be developed to replace MCMC autocorrelation analysis for multi-modal target distributions (cf. Section 4.1.6).

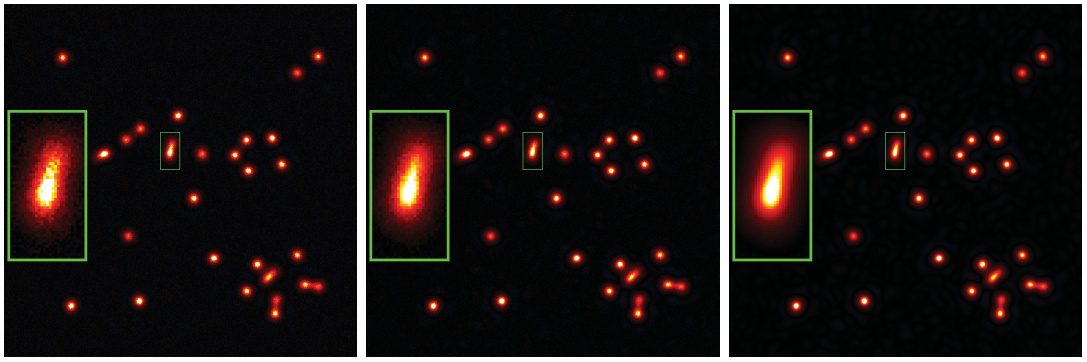
5.1.4. Oriented Overrelaxation Studies

Direct Sampler For the direct SC Gibbs sampler and the ℓ_1 prior, OOR can be implemented using Algorithm 4.11 (details can be found in Lucka 2012). As discussed in Section 4.3.1, this renders the additional computational time nearly independent of N_O. However, as noted above, the computational effort of the SC sampling step in high dimensional scenarios is negligible, anyhow. In Lucka (2012), visual convergence, burn-in and autocorrelation analysis were performed in the same way as in Section 5.1.2 for both RSG and SSG using $N_O = 3$ and $N_O = 7$ in addition to the normal samplers



(a) RSG **(b)** SSG **(c)**

Figure 5.4.: (a)-(b) Influence of OOR on $R(\tau)$ and τ_{int} (SSR $= n$). The "Boxcar" scenario with $n = 255$ and a TV prior with $\lambda = 400$ were used. (c) Influence of OOR on the cross statistic (5.2).



(a) SSG-O1, 2h **(b)** SSG-O7, 2h **(c)** Reference solution, 7d

Figure 5.5.: Comparison between normal and overrelaxed SSG sampler after two hours of computation time and a reference solution. All figure settings are the same as in Figure 5.2.

(i.e., $N_O = 1$). Here, we only summarize the most notable results: The burn-in length of RSG samplers is unaffected by OOR, while it slightly increases for SSG samplers (e.g., from 150 to 200 for $N_O = 7$ in the $(1023, 800)$ case in Table 5.1). Figures 5.4a and 5.4b show the results of an autocorrelation analysis. For the RSG sampler, OOR leads to a noticeable decrease in autocorrelation. For large N_O, τ_{int} saturates at almost half of its value for $N_O = 1$. For SSG, using OOR seems to produce a $R(\tau)$ which oscillates around zero: Samples after a certain number of steps get anti-correlated to each other. The estimation of τ_{int} we use (Wolff 2004) is not designed to be used in situations where $R(\tau) < 0$ (in contrast, it assumes $\hat{R}(\tau) < 0$ is the result of a too short chain length). However, OOR actually only amplifies this behavior: In the $(1023, 800)$ case in Table 5.2, a closer look reveals that $R(1) = -0.24$ already for the un-over-relaxed SSG. This also explains that $\tau_{int} = 0.26$ is lower than $\tau_{int} = 0.5$ of an uncorrelated series of random

Table 5.6.: $\hat{\tau}_{int}$ of the blocked Gibbs sampler using the ℓ_2 hypermodel with an inverse gamma hyperprior ($\alpha = 0.5$, $\beta = 10^{-4}$) for the recovery of a single source in the "simEEG" scenario. As a projection $g(u, \gamma)$ we use the first eigenvector of the full covariance matrix for (u, γ) after applying a specific block weighting to it. OOR was applied to the sampling of the univariate conditional hyperparameter posteriors.

$N_O = 1$	$N_O = 3$	$N_O = 5$	$N_O = 7$	$N_O = 9$
8550±283	6504±191	5977±169	5829±163	5396±145

variables. A visual comparison is given in Figure 5.5 (the study design was the same as in Section 5.1.2). However, note that Figures 5.4a and 5.4b show the combination of sampler and computation time that produced the most distinctive difference between normal and overrelaxed samplers. In this scenario, using OOR also seems to pay off in terms of computational efficiency ($N_O = 7$ only takes about 1.03 times longer than $N_O = 1$).

Slice Sampler For the slice sampler examined in Table 5.4 we introduced OOR for $K_0^{ss} = 20$ and computed the corresponding $\hat{\tau}_{int}$. Without OOR, we have $\hat{\tau}_{int} = (149.2 \pm 4.6)$. Using $N_O = 3$, we obtain $\hat{\tau}_{int} = (108.0 \pm 2.9)$ which is better than using $K_0 = 40$ without OOR. Using $N_O = 7$, we obtain $\hat{\tau}_{int} = (101.1 \pm 2.6)$ which is better than using $K_0 = 200$ without OOR. Currently, OOR for the slice sampler is implemented in the naive way using Algorithm 4.10 and not in an efficient way. However, using $K_0 = 20$ and $N_O = 7$ is already faster than using $K_0 = 200$ without OOR.
We also examined the influence of OOR for slice sampling with a non-log-concave ℓ_p prior with $p = 0.8$. We chose the specific SC density corresponding to the green plot in Figure 4.5a. Instead of relying on autocorrelation analysis, we computed a statistic that relates to the ability of the sampler to switch between the two modes: One can easily derive that the modes must be located in $[0, \mu]$, where $\mu = b/(2a) > 0$ is the mode of the Gaussian likelihood part on the SC posterior. Therefore, if one starts one chain, $\{x_1^i\}$, in 0, and another one, $\{x_2^i\}$, in μ, it will take a while until both chains cross for the first time:

$$i_X := \min \left\{ i \,|\, x_1^i > x_2^i \right\} \tag{5.2}$$

The probability distribution of i_x carries information about the ability of the chains to switch between the two modes. Figure 5.4c shows empirical histograms of this *cross-statistic* using different values of N_O. One can clearly see that using OOR leads to earlier crossings. However, when interpreting the absolute numbers, one should bear in mind that the initialization with $x^0 = \mu$ is far outside (μ = 7.6e-5, while the second mode is actually at 2.6e-5).

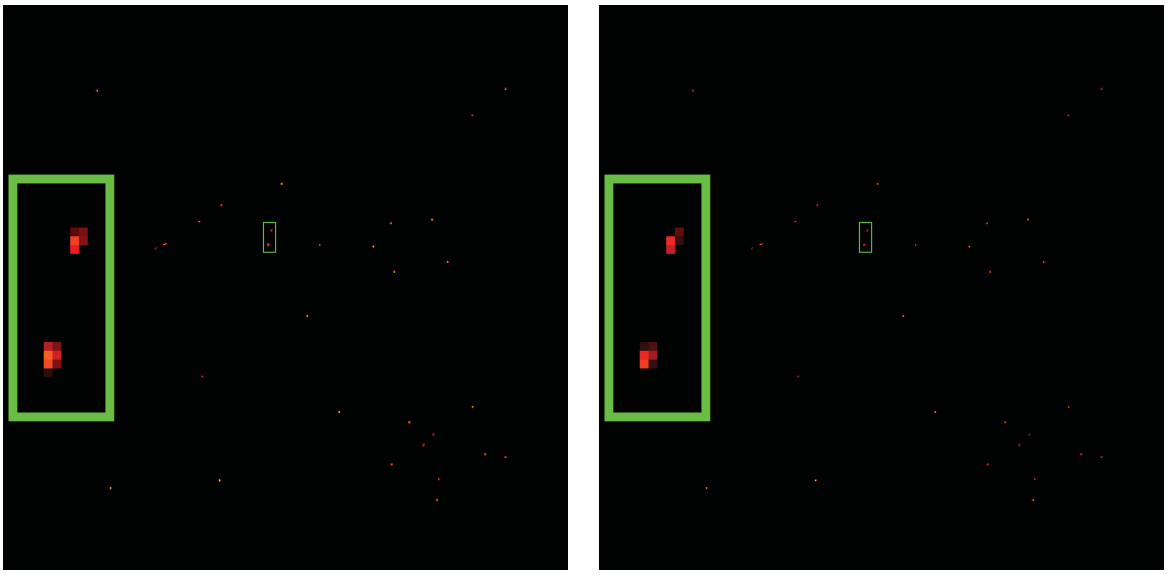
HBM Sampler For ℓ_p hypermodels, OOR can be applied to the sampling of the univariate conditional hyperparameter posteriors $p_{post}(\gamma_i|u)$ in the HBM-Gibbs sampler (see Sections 4.1.11 and 4.3.1). Again, OOR is currently implemented in the naive way using Algorithm 4.10. Table 5.6 lists $\hat{\tau}_{int}$ using the ℓ_2 hypermodel with an inverse gamma hyperprior in the "simEEG" scenario. As discussed in Section 4.1.6, performing autocorrelation analysis is difficult for multimodal posteriors. Our analysis is based on reference statistics of the posterior computed in about 300 days of single CPU computation time. The estimates $\hat{\tau}_{int}$ are each based on six independent chains ($4 \cdot 10^7$ samples for each chain, which corresponds to about one week of computation time). Using this much computational effort, the difference between $\hat{\tau}_{int}$ and $\hat{\tau}_{int}^{ref}$ was small enough so that autocorrelation analysis can provide reasonable results. Again, we see that using OOR can significantly reduce $\hat{\tau}_{int}$. In the concrete scenario, the dimensions of sensor and source space are low. Therefore, the decrease in τ_{int} is compensated for by the increase of computational time. However, for the full EMEG sensor configuration used in the EP/EF studies and a source grid with 3mm instead of 6mm spacing, the additional computational effort is negligible and the increase of the statistical efficiency comes without the cost of a significant decrease in computational efficiency.

Discussion

The results consistently show that using OOR reduces the autocorrelation of MCMC samplers and thereby increases their statistical efficiency. Using cross statistics, we showed it can also enhance the ability of slice samplers to switch between single modes of a distribution. If the sampling step in which OOR is used is the computational bottle neck of the whole sampler, efficient implementations of OOR such as Algorithm 4.11 are required to translate the superior statistical efficiency of OOR into superior computational efficiency. However, if other parts of the whole sampler are the computational bottle neck, even naive implementations of OOR can increase the overall computational efficiency of the sampler. This is often the case in high dimensional scenarios which, unfortunately, do not allow for similar extensive, quantitative evaluation studies as performed in this section.

5.1.5. Simulated Annealing Studies

Once a fast and robust sampler is available, adding simulated annealing to its implementation is easy (cf. Sections 4.2.4 and A.6). Therefore, using SA for MAP estimation was originally intended as a fill-in for low-dimensional situations where ADMM is not applicable (due the non-convexity of the prior energy) or not implemented yet (such as for ℓ_p priors with $1 < p < 2$). Typically, SA is used in combination with MH and



(a) ADMM, 16d 13h **(b)** SA with RSG, 1d 12h

Figure 5.6.: Visual comparison of the MAP estimates computed by ADMM and SA. The "Spots" scenario with $n = 513 \times 513$ and a simple ℓ_1 prior with $\lambda = 2 \cdot 10^7$ was used. The inset shows a zoom into the marked area in the original figure.

is not even considered for high-dimensional scenarios if no deterministic alternative is available. The reason is the extremely long burn-in time of MH for large n (cf. Section 5.1.2): After the initial burn-in time required to sample the un-tempered posterior, each cooling step necessitates a new burn-in phase of the chain. In addition, the proposal distribution in MH needs to be adapted after every cooling step as well. Before this adaptation has taken place (e.g., by an auto-tuning procedure such as we are using), the chain hardly moves. In total, these issues render SA with MH practically infeasible for high dimensional scenarios.

However, the surprising properties of SC Gibbs samplers suggest that using SA with Gibbs sampling might be more promising: Its short burn-in times might enable the Gibbs samplers to follow even fast cooling schedules and its "auto-adaptation" to the target density might be another crucial advantage over MH in this respect. Therefore, we compared SA with RSG to ADMM for ℓ_1 priors. First, we checked in the "Boxcar" scenario using a TV prior that SA and ADMM really converge to the same result. Then, we compared both methods in the "Spots" scenario with $N = 9$, i.e., $n = 513 \times 513$. For ADMM, $\epsilon^{abs} = \epsilon^{rel} = 10^{-4}$ were defined as stopping tolerances. The adaptation of ρ was not stable for these tolerances: $\|r^i\|_2$ and $\|s^i\|_2$ started to oscillate. As discussed in Section 4.2.2, this might be a problem of the iterative solution of the least-squares problem. Instead, a number of values for ρ were tested, of which $\rho = 1$ was finally used for all ADMM computations in the "Spots" scenario (this is also advocated in

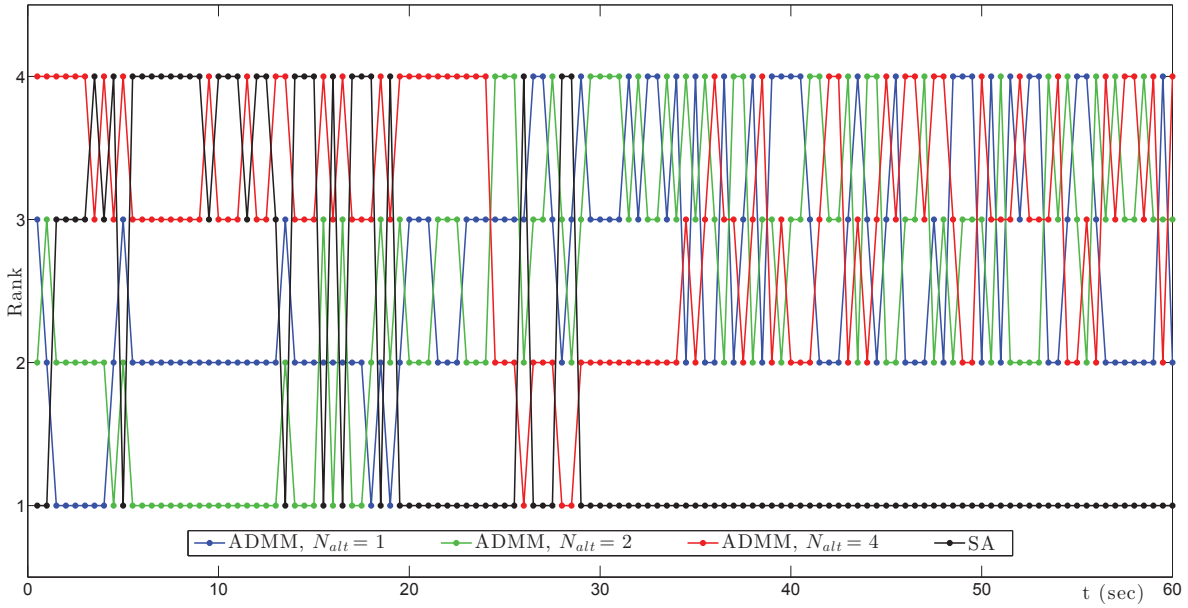


Figure 5.7.: Comparison of posterior optimization by ADMM and SA: Different Methods are ranked by the posterior energy obtained after a certain amount of computational time. A lower rank corresponds to a superior performance.

Goldstein and Osher 2009). As in Boyd et al. (2011), Goldstein and Osher (2009), we used only one alternation between u and v updates (we will denote this number by N_{alt}). The computation stopped after 111 464 iterations which took 16 days and 13 hours of computation time on a single CPU. SA with RSG was used with $K = 5000$, $K_0 = 200$, $T_0 = 1$ and $T_{end} = 10^{-50}$. The computation took 1 day and 12 hours of computation time on a single CPU. Figure 5.6 shows the final results of both methods. While they seem very similar at first glance, the zoom reveals that the SA result is sparser. An examination of the chain of posterior energies produced by SA shows that the energies of u^i are already smaller than the energy of the final ADMM result after half of the total chain[1].

In a more detailed study, we compare the posterior energies reached by the different methods after a certain amount of computational time. We use the "Boxcar" scenario with $n = 255$ and a TV prior with $\lambda = 400$. First, we fixed a number of computation times t_i for which we want to compare the different methods. For ADMM, we then only have to fix N_{alt} (ρ is automatically adapted) and let the algorithm run for a certain number of iterations. The result is a vector of posterior energies obtained after a certain amount of computational time. We can define the posterior energy reached at t_i by a linear interpolation from this vector. For SA, an inherent problem with such a comparison is that the concept behind its parameterization is usually different: One

[1] A potential complication in the comparison is that the forward operator used by ADMM is implemented using ffts, whereas the SC Gibbs sampler relies on a direct application of the convolution kernel. However, we checked and this is not the reason.

fixes K_0, T_0 and T_{end} and chooses K (and thereby, q) depending on the amount of computation time available. With this parameterization, results with shorter or longer chains are not really representative for the method. Therefore, we ran SA for each computation time t_i independently (although with the same random number generation seed). We initialized SA with a sample from the un-tempered posterior to get rid of the burn-in phase which further complicates a meaningful comparison. To facilitate the interpretation of the results, Figure 5.7 shows a ranking of posterior energies obtained by the different methods at t_i rather than the absolute energy values. The ADMM results show that using only one inner iteration is preferable for short computation times less than 5 seconds. After that, the additional computational effort of using more inner iterations pays off: First using two iterations ranks best, then using four. Finally, after about 35 seconds, there is no clear order among the ADMM methods anymore. SA clearly ranks before the ADMM methods for long computation times. For short times, the situation is less clear, but the results show that SA is at least competitive to ADMM.

Discussion

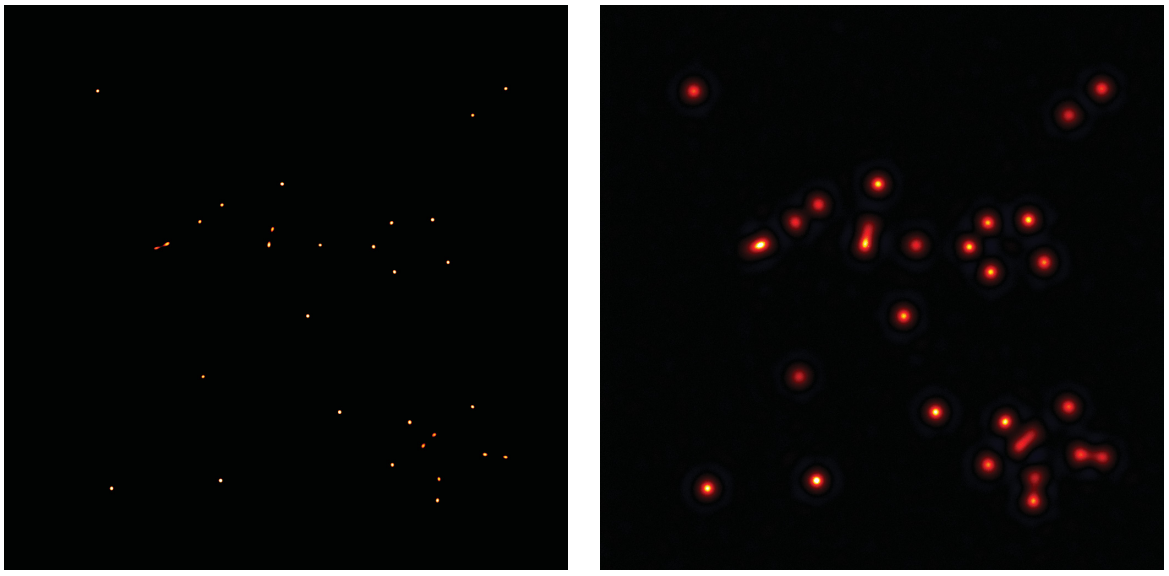
In total, the first results obtained for using SA in combination with Gibbs sampling are quite promising. In particular, they challenge common beliefs about the applicability of SA for high-dimensional optimization. However, more detailed studies need to be done. It should also be pointed out that although SC Gibbs samplers for various scenarios and priors have been developed, deterministic optimization strategies such as ADMM are still way more generally applicable.

5.2. General Bayesian Inversion Studies

In this section, we will present a collection of simulated data studies that addressed different topics in Bayesian inversion. Most often, they aimed to verify or illustrate certain theoretical results. The key topics are discretization invariance (cf. Section 3.6.1), the comparison of MAP and CM estimates and sparsity.

5.2.1. "Spots" Reconstruction with an ℓ_1 Prior

Figure 5.8 compares MAP and CM estimates for the "Spots" scenario. While the MAP estimate yields a sparse reconstruction, the CM estimate does not.



(a) MAP by ADMM **(b)** CM by RSG-O7

Figure 5.8.: MAP and CM estimate for the "Spots" scenario ($n = 1023 \times 1023$) using the standart ℓ_1 prior with $\lambda = 4 \cdot 10^7$.

5.2.2. The Discretization Dilemma of the TV Prior

The edge-preserving properties of the MAP estimate using a TV prior (cf. Figure 3.5b) stimulated interest in the general properties of Bayesian inversion with such non-Gaussian prior models, with a focus on the comparison between CM and MAP estimates (cf. Section 3.4.3). As mentioned in Section 3.6.1, Lassas and Siltanen (2004) showed that the TV prior in 1D is not discretization invariant. In particular, the posterior cannot converge to a well-defined, edge-preserving limit for $n \to \infty$. To summarize their results:

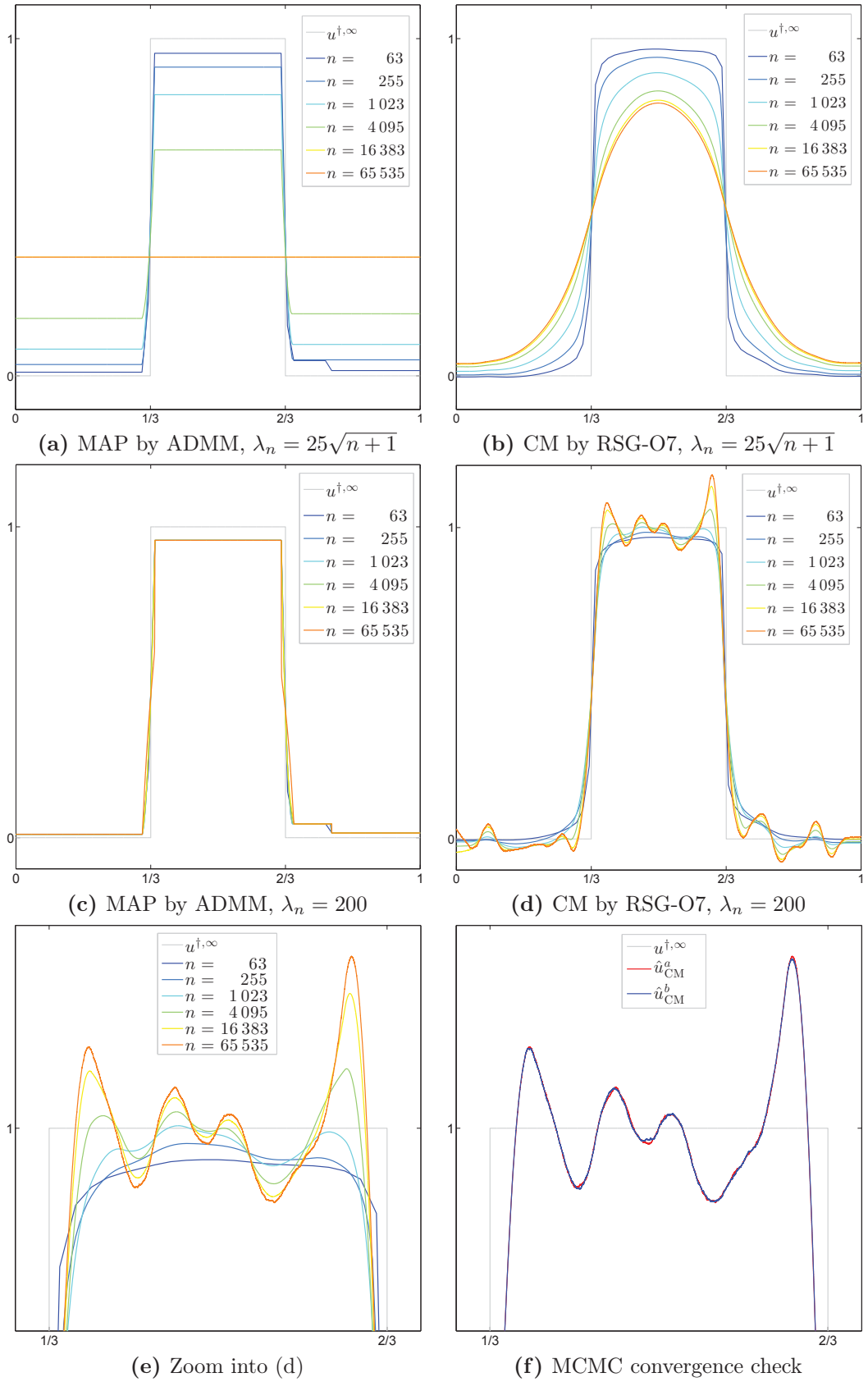
- For $n \to \infty$, the posterior only converges to a non-trivial limit if $\lambda_n \propto \sqrt{n+1}$. However, this limit is a Gaussian smoothness prior and the CM estimate converges to a smooth limit while the MAP estimate converges to constant function.
- For $n \to \infty$ and $\lambda_n = \text{const.}$, both posterior and CM estimate diverge while the MAP estimate converges to an edge-preserving limit.

The limit $n \to \infty$ is a severe challenge for the computational verification of these results. In Lassas and Siltanen (2004), an MH sampler similar to MH-Iso or MH-Si (it updates a fraction of components in every step) was used to compute CM estimates in the "Boxcar" scenario for $n = 63, 255, 1023, 4095$. Although the whole computation took about a month of time on a desktop PC equipped with a 2.8 GHz single core CPU, the authors admitted that the results for $n = 4095$ were only partly satisfying. The examinations in Section 5.1.2 were tailored to this scenario and explain this failure of the

MH sampler (the more detailed examinations in Lucka 2012 also included the sampler used in Lassas and Siltanen 2004). However, due to its surprising properties, the SC Gibbs samplers is the right tool to carry out a satisfactory computational examination of the limit $n \to \infty$: Figures 5.9a and 5.9b compare MAP and CM estimates for $\lambda_n = 25\sqrt{n+1}$ up to $n = 65\,535$. The results clearly match the theoretical predictions. The CM estimate for $n = 65\,535$ can be obtained in about 3 hours of computational time on a CPU comparable to the one used in Lassas and Siltanen (2004). Figures 5.9c and 5.9d show the same comparison for $\lambda_n = \text{const}$. Figure 5.9e shows a zoom into the plot to clarify the divergence of the CM estimate. By comparing two CM estimates computed from independent MCMC chains, Figure 5.9f demonstrates that this is not an error of the RSG sampler to compute it.

For image dimensions higher than one, no theoretical results are available yet (to the best of our knowledge). Using slice sampling within the RSG-O7 sampler as developed in Section 4.1.10, we can explore such a case computationally. The details of the slice sampler implementation used to sample from the SC densities (cf. Section 4.1.10) are given in Section A.4. Figure 5.10 shows MAP and CM estimates for the "Phantom-CT" scenario using an isotropic TV prior with $\lambda = 500$. Contrary to the 1D case, the CM estimates get smoother for a constant value of λ as the resolution increases. In Section 5.3, we will use the "Walnut-CT" scenario to compare CM and MAP estimates for the TV prior in 2D computed with experimental data .

In total, the CM estimates either get smooth when $n \to \infty$ or diverge. For the 1D case, these results confirm the theoretical predictions. Thereby, the gradient of the CM estimates is never sparse and the edge-preserving properties of the MAP estimate are lost. For denoising using a TV prior, i.e., $A = I_n$, this lack of sparsity has also been examined theoretically in Louchet (2008). These problems of the TV prior motivated research on whether and how it is possible to formulate edge-preserving priors in a consistent, discretization invariant way. In the following sections, we will illustrate some of these developments.



(a) MAP by ADMM, $\lambda_n = 25\sqrt{n+1}$

(b) CM by RSG-O7, $\lambda_n = 25\sqrt{n+1}$

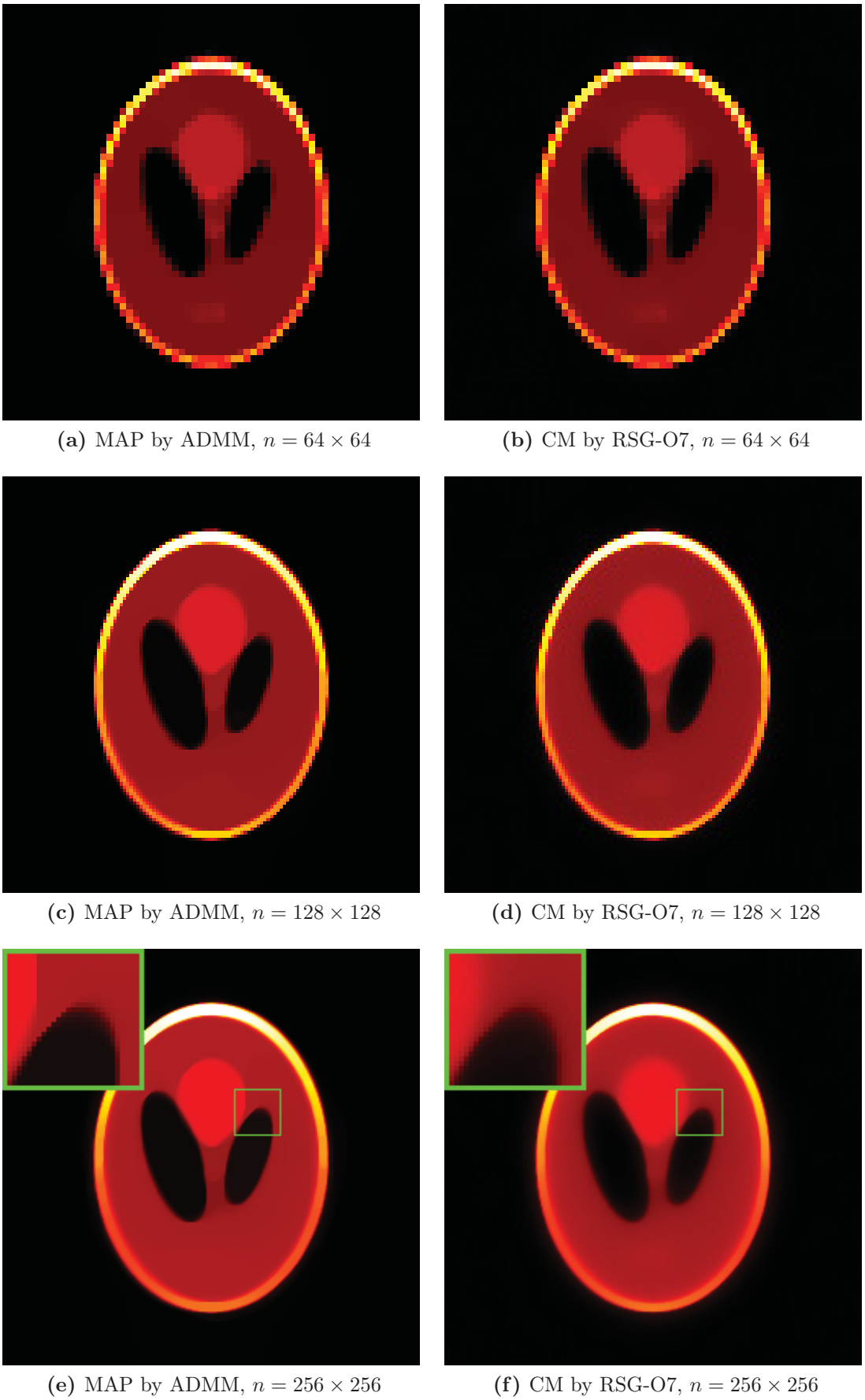
(c) MAP by ADMM, $\lambda_n = 200$

(d) CM by RSG-O7, $\lambda_n = 200$

(e) Zoom into (d)

(f) MCMC convergence check

Figure 5.9.: (a)-(d) MAP and CM estimates for the "Boxcar" scenario using the TV prior with different scalings of λ_n. (e) A zoom into (d). (f) Two approximations to the CM estimate for $n = 65\,535$ computed from independent MCMC chains to demonstrate the oscillations of the CM estimate are not caused by the MCMC error.



(a) MAP by ADMM, $n = 64 \times 64$

(b) CM by RSG-O7, $n = 64 \times 64$

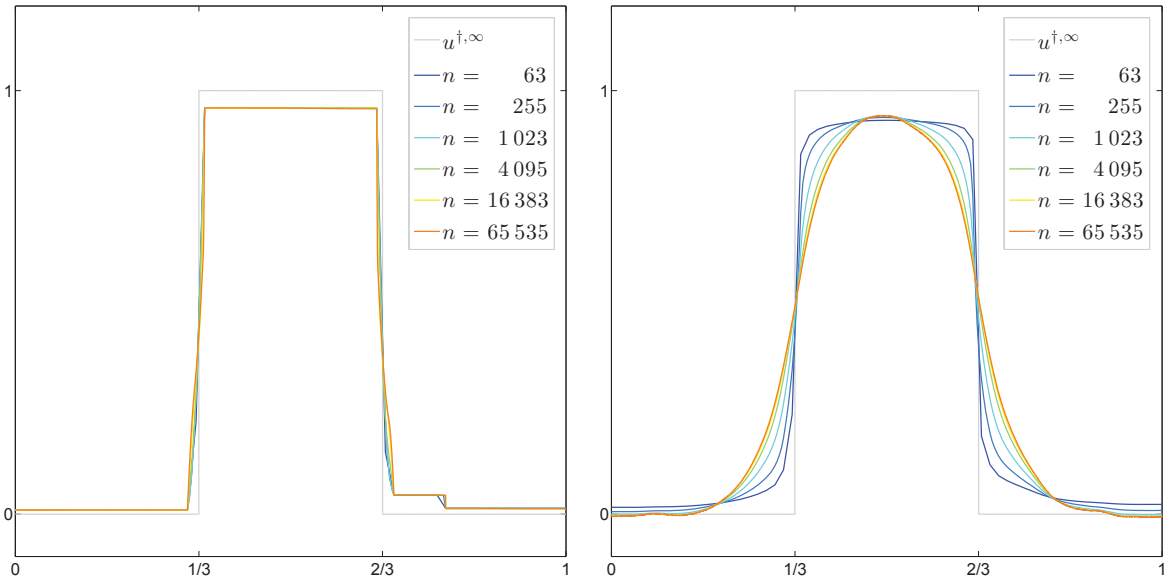
(c) MAP by ADMM, $n = 128 \times 128$

(d) CM by RSG-O7, $n = 128 \times 128$

(e) MAP by ADMM, $n = 256 \times 256$

(f) CM by RSG-O7, $n = 256 \times 256$

Figure 5.10.: MAP and CM estimates for the "Phantom-CT" scenario using an isotropic TV prior with $\lambda = 500$. In the highest resolution, a zoom inset is added.



(a) MAP by ADMM and parameter fitting. **(b)** CM by RSG-O7.

Figure 5.11.: MAP and CM estimates for the "Boxcar" scenario using the q-TV prior with $q = 2$, $\lambda_n = 100$.

5.2.3. q-TV Priors

In Comelli (2011), *q-TV* priors were defined as ℓ_p^q increment priors with $p = 1$, which means that we modify the TV prior by raising the whole TV functional to the power of q. Theoretically, it was shown that using $q = 2$ and $\lambda_n = \text{const.}$, both MAP and CM estimate converge to a limit function, while choosing $\lambda_n \propto \sqrt{n+1}$, both converge to zero. No theoretical characterization of the non-zero limit of the CM estimates for $\lambda_n = \text{const.}$ was given, but numerical studies using MH samplers were conducted. The results suffered from the same problems as those in Lassas and Siltanen (2004) and, in fact, initiated the development of the SC Gibbs samplers in Lucka (2012). Figure 5.11 compares MAP and CM estimates for $\lambda_n = 100$ up to $n = 65\,535$. The MAP estimates were computed as described Section 4.2.3 whereas the CM estimates were computed by the RSG-O7 sampler. We clearly see that both estimates converge to a non-trivial limit function. The MAP estimates do not differ from those for the normal TV prior (cf. Figure 5.9c). The CM estimates converge to a smooth limit which differs from the $\lambda_n \propto \sqrt{n+1}$ limit for the normal TV prior (cf. Figure 5.9b). From the visual impression, the limit for $q = 2$ is more convincing than the limit for the normal TV prior, i.e., $q = 1$. Finally, Figure 5.12a shows CM estimates for $q = 10$ and $\lambda_n =$const. Comelli (2011) did not find a theoretical prediction for this situation, but the numerical results suggest that the CM estimates converge to a constant function in this case.

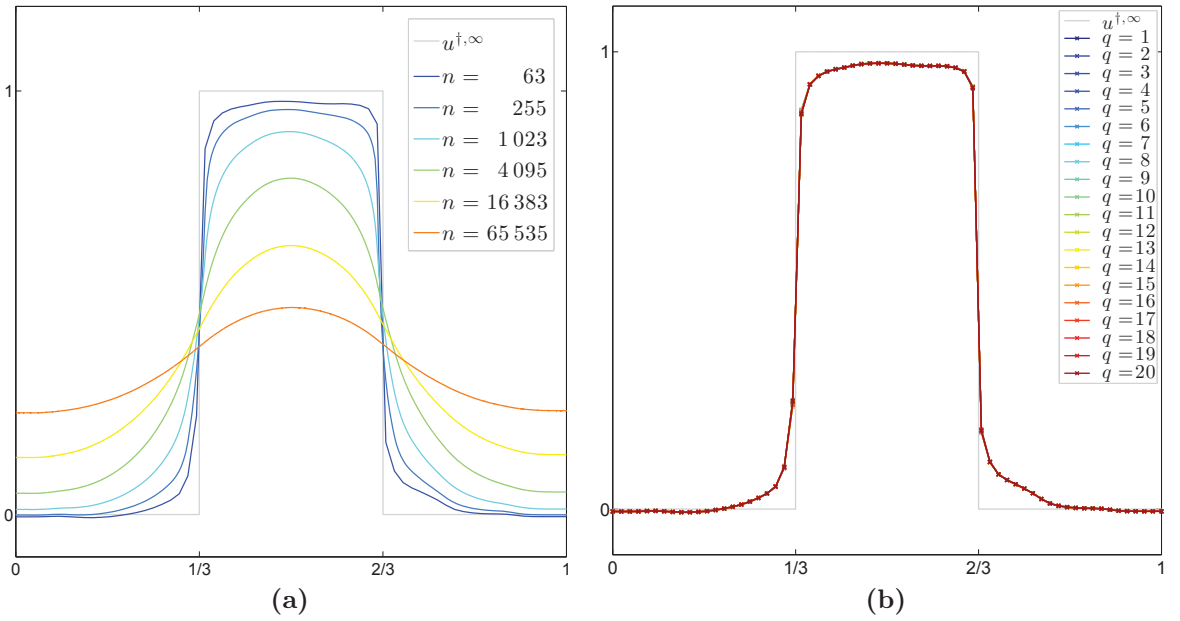
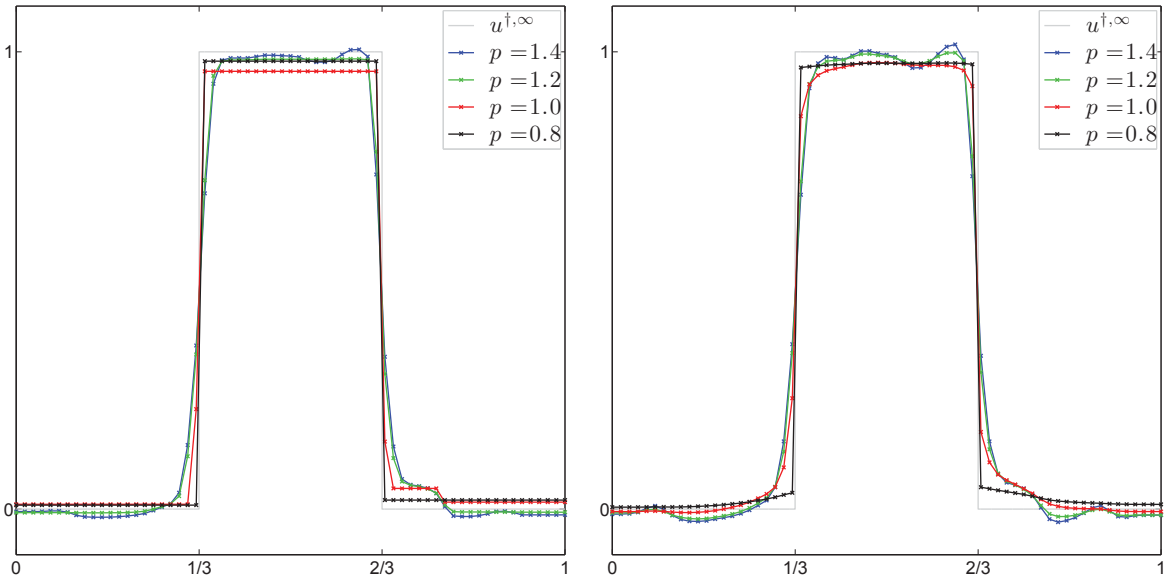


Figure 5.12.: (a) CM estimates for the "Boxcar" scenario using the q-TV prior with $q = 10$, $\lambda_n = 1.625\text{e-}2$ (cf. Table A.2). (b) CM estimates using the q-TV prior for the "Boxcar" scenario ($n = 63$), different q and λ_q from Table A.2.

In a second study, we wanted to compare MAP and CM estimates for $n = 63$ and $q = 1, \ldots, 20$. For a meaningful comparison, we needed to adjust λ dependent on q: A *parameter choice rule* related to the *discrepancy principle* (Kaipio and Somersalo 2005) is to demand that the likelihood energies $\|f - A\hat{u}\|^2_{\Sigma_\varepsilon^{-1}}$ of all estimates should be equal to the likelihood energy of the MAP estimate for $q = 1$ and $\lambda = 200$. The implementation of this criterion is straight-forward for MAP estimates computed with deterministic optimization, but needs some care for CM estimates computed by MCMC. However, we achieved that the differences in relative likelihood energy, i.e., $\|f - A\hat{u}\|^2_{\Sigma_\varepsilon^{-1}} / \|f\|^2_{\Sigma_\varepsilon^{-1}}$, are all below 1.2%. Table A.2 lists the resulting λ_q. One can see that λ_{CM} and λ_{MAP} diverge from each other for increasing q. By intention, the parameter choice rule forces all MAP estimates to be equal (cf. Section 4.2.3). More surprisingly, Figure 5.12b suggests that the same seems to hold true for the CM estimates. The implications of this finding have to be examined in more detail. If the results hold for arbitrary n and λ, one could reproduce any CM estimate for q_1 for a second $q_2 \neq q_1$. Thereby, all limits for q_1 could be reproduced for q_2. For instance, it could be possible to reproduce the limit for $q = 2$, $\lambda_n =$const. with $q = 1$. As this limit clearly differs from the $q = 1$, $\lambda_n \propto \sqrt{n+1}$ limit, it would mean that there is a limit for $q = 1$ not considered in Lassas and Siltanen (2004).



(a) MAP by SA **(b)** CM estimates by RSG

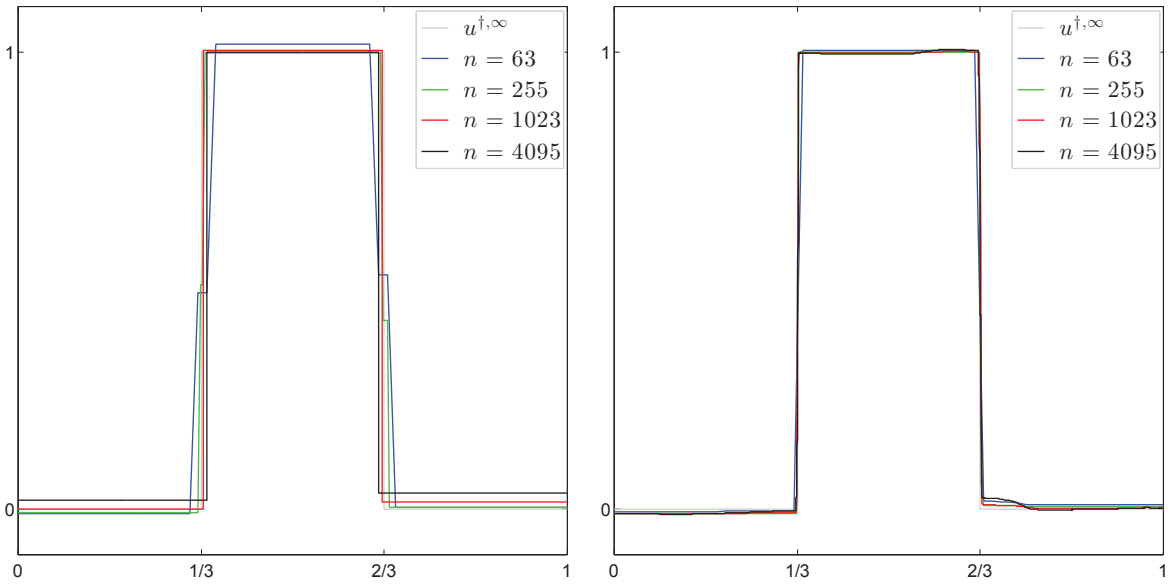
Figure 5.13.: MAP and CM estimates for the "Boxcar" scenario using the p-TV prior and $n = 63$.

5.2.4. p-TV Priors

Another apparent modification of the original, ℓ_1-based TV prior is to consider ℓ_p increment priors. We will refer to these priors as p-TV priors. Figure 5.13 compares MAP and CM estimates for different values of p. As in the previous section, λ was chosen in such a way that all likelihood energies are equal and that $\lambda = 200$ for $p = 1$. The results suggest that using $p < 1$ leads to superior results for both MAP and CM estimates compared to $p = 1$. The MAP estimate is closer to the real solution as it is both sparser in the increment basis and the contrast loss is reduced. The CM estimate for $p = 0.8$ looks way more convincing compared to those for $p \geqslant 1$: It has clear pronounced edges that separate smooth, denoised parts. However, as discussed in Section 5.1.3, using the slice-within SC Gibbs samplers for $p < 1$ needs to be examined carefully (and possibly improved). Therefore, we did not yet compute results for smaller values of p or for larger values of n to examine the limit $n \to \infty$ as in the $p = 1$ case (cf. Figure 5.3b for the inherent dangers of such computations).

5.2.5. ℓ_p Hypermodels

Various hierarchical Bayesian models have been considered for edge-preserving image reconstruction (see, e.g., Bardsley et al. 2010, Calvetti and Somersalo 2007, 2008, Helin 2010b, Helin and Lassas 2009), in particular with the intention that such models might posses the discretization invariance the TV prior lacks. A simple, heuristic



(a) NM estimates (cf. Section 4.2.5) **(b)** CM estimates

Figure 5.14.: NM and CM estimates for the "Boxcar" scenario using the ℓ_2 hypermodel for the increments with $\alpha = 0.5$, $\beta = 10^{-10}$.

explanation why the TV prior is not discretization invariant is given by Calvetti and Somersalo (2007): In the 1D "Boxcar" scenario, we can construct a continuous estimation u_n^∞ for $u^{\dagger,\infty}$ by extending $u \in \mathbb{R}^n$ to a piecewise constant function with $u^\infty(x) = u_i$ for $x \in [x_{i-1}^n, x_i^n]$. Using the increment basis $\xi_j = u_{j+1} - u_j$, we have that

$$u_i = u_0 + \sum_{j=0}^{i-1} \xi_j \tag{5.3}$$

and thereby:

$$u_n^\infty(x) = u_0 + \sum_{j=0}^{\lfloor xn \rfloor} \xi_j, \tag{5.4}$$

where $\lfloor xn \rfloor$ is the integer part of xn. For increasing n, the number of summands increases for a fixed location x. The discrete TV prior encodes that the increments ξ_j are all mutually independent, identically Laplace distributed random variables, $\xi_j \propto \exp(-\lambda|\xi_j|)$. As a result, $u_n^\infty(x)$ is the sum of mutually independent, identically distributed random variables with finite second moments. Due to the central limit theorem, its distribution will becomes more and more Gaussian with increasing n: The edge-preserving TV prior converges to a Gaussian smoothness prior.
This explanation motivates examining ℓ_p hypermodels as increment priors: Their use implicitly leads the to fat-tailed t_p priors on the increments which decay as $|xi_j|^{-\tilde{\alpha}-1}$

Figure 5.15.: Filtered back projection of the "PhantomCT" data for $n = 256 \times 256$.

with $\tilde{\alpha} = \alpha + \frac{1}{p} - 1$ (cf. Section 3.3.3). As they do not have a finite second moment for $0 < \tilde{\alpha} < 2$, the standard central limit theorem does not hold for them anymore. Instead, a generalization by Gnedenko and Kolmogorov (1954) asserts that they converge to an *alpha-stable* distribution with the *stability parameter* $\tilde{\alpha}$ (see also Klenke 2008). For instance, using $p = 2$ and $\alpha = 0.5$, we obtain a Cauchy prior on the increments, and the sum of Cauchy-distributed random variables is, again, a Cauchy distribution which does not converge to a Gaussian (in contrast, the sum of Poisson distributed random variables is Poisson distributed but, with a certain scaling, converges to a Gaussian). However, to the best of our knowledge, there is no detailed theoretical examination of this topic yet. In particular, it is not clear how α and β have to be scaled with n. In Figure 5.14, we compared NM and CM estimates for keeping α and β constant. In general, the results look quite promising compared to those obtained with ℓ_p priors: They look sparse and do not seem to suffer from contrast loss. However, while the MAP estimates stay sparse, the CM estimates tend to develop variations on the small scales using constant values for α and β.

5.2.6. Besov Priors

In Kolehmainen et al. (2012), Lassas et al. (2009), Besov space priors with $p = 1$ (cf. Section 3.2.4) were proposed and examined as edge-preserving, ℓ_1-based alternatives to TV priors. By theoretical arguments, it was shown that they are discretization invariant and that the CM and MAP estimate both converge to non-trivial limit functions for $\lambda_n = \text{const}$. In Kolehmainen et al. (2012), these findings were confirmed in a low-dimensional 1D image deblurring scenario. Haar wavelets (see Figures 3.7a and

A.3) were found to be most appropriate to represent piecewise constant functions. In Hämäläinen et al. (2013), these numerical studies were extended to high-dimensional 2D problems corresponding to our two CT scenarios. However, only MAP estimates were computed. For this thesis, an efficient implementation of the SC Gibbs sampler for the specific combination of Haar wavelets v_i and the CT forward operator $\mathcal{A}$ (both for para- and fan-beam geometry) was developed that allows to carry out sample-based posterior inference in this scenario. Details of the implementation can be found in Section A.2. In this section, we will use these developments to tie in with the discretization invariance studies of the previous sections and extend the results of Kolehmainen et al. (2012) to the 2D "Phantom-CT" scenario. Section 5.3 will contain the application to the real-data "Walnut-CT" scenario.

In Figure 5.16, single samples of the posterior are shown. The visual impression suggests that most variability of the posterior is found at small image scales. In Figure 5.17, CM and MAP estimates for increasing n are compared. In line with Kolehmainen et al. (2012), Lassas et al. (2009), we can confirm that both estimates converge to edge-preserving limits for $n \to \infty$ and that they are sparse/compressible in the wavelet basis. We can also confirm the surprising result of Kolehmainen et al. (2012) that the differences between MAP and CM estimates decreases for increasing λ: As the growing impact of the non-Gaussian prior renders the posterior less and less Gaussian, one would rather expect the opposite. Figure 5.18 compares the regularization paths $\{(\hat{u}_{\mathrm{CM}}, \lambda) \,|\, \lambda > 0\}$ and $\{(\hat{u}_{\mathrm{MAP}}, \lambda) \,|\, \lambda > 0\}$. We see that small values of λ lead to different kinds of errors for MAP and CM estimates: The MAP estimates look as if noise consisting of small scale wavelets was added to the regularized solution for a larger λ whereas the CM estimates look more like the filtered back projection (see Figure 5.15 and cf. Section 2.3). In particular, they start to feature the typical stripe-like artifacts. Despite the interesting theoretical properties for Bayesian inversion, one has to admit that the MAP estimates using a TV prior (cf. Figure 5.10) are visually more convincing for this scenario.

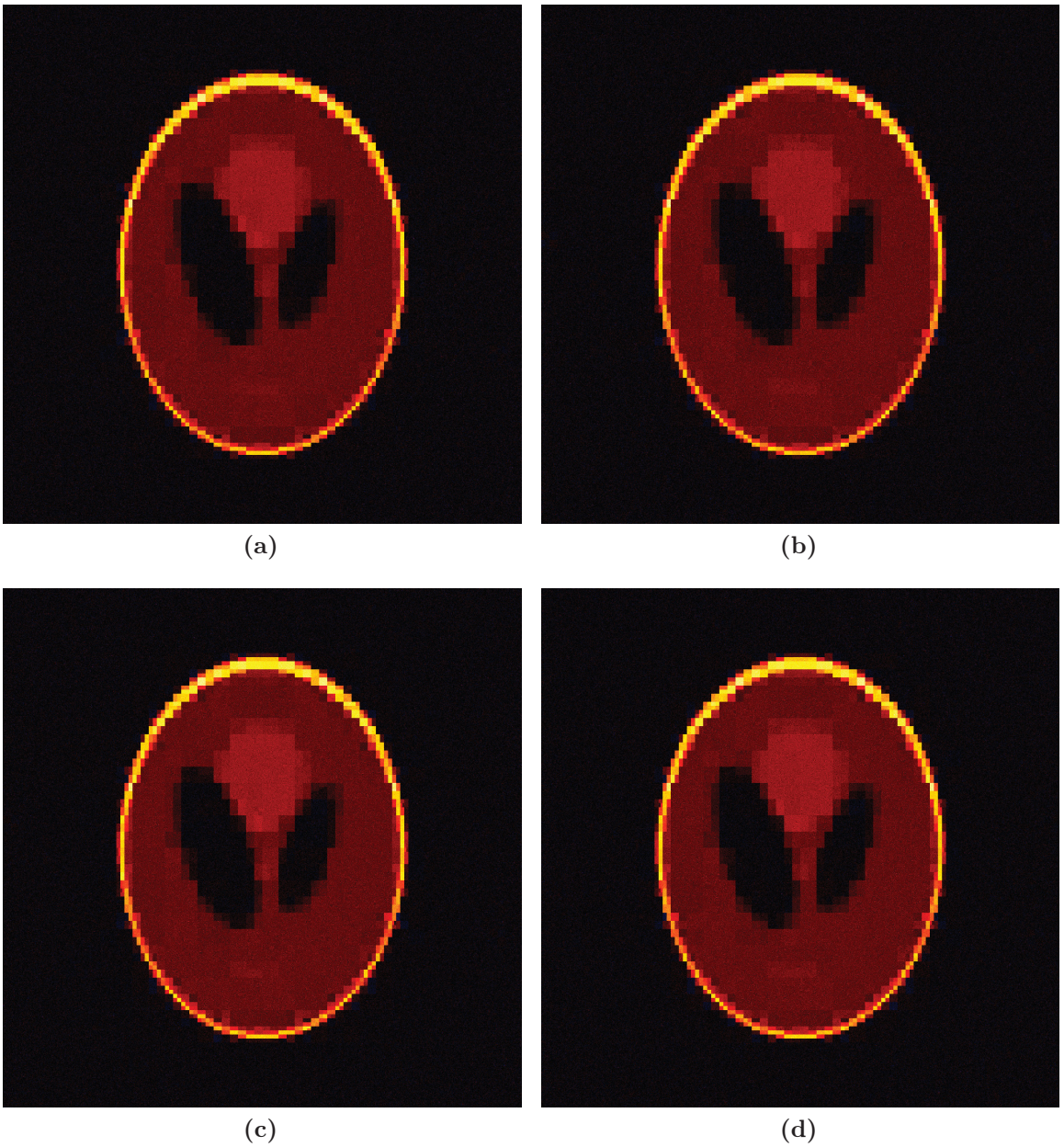


Figure 5.16.: Four posterior samples in the "PhantomCT" scenario using $n = 1024 \times 1024$ and a Haarwavelet Besov prior with $\lambda = 2 \cdot 10^4$

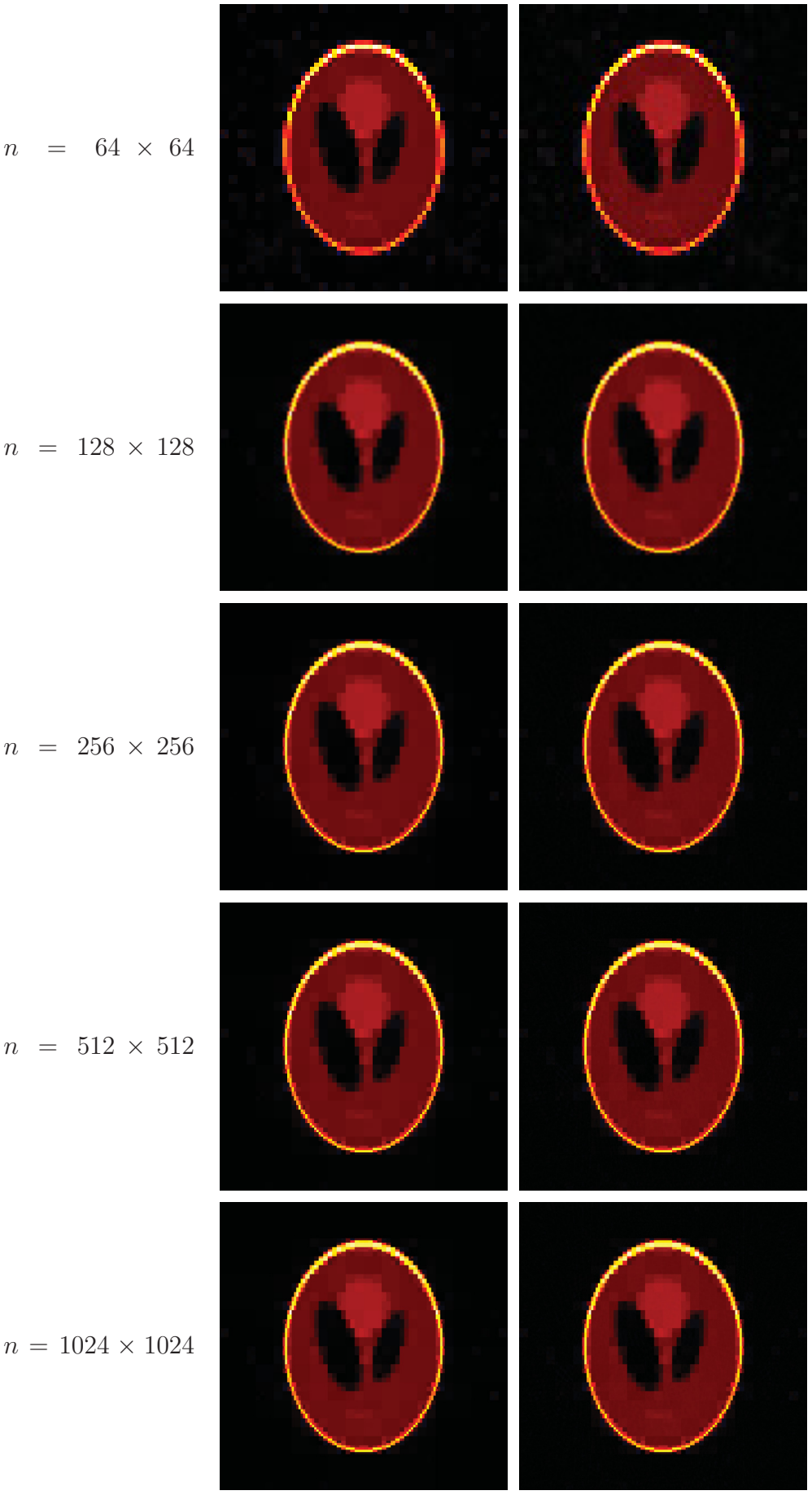
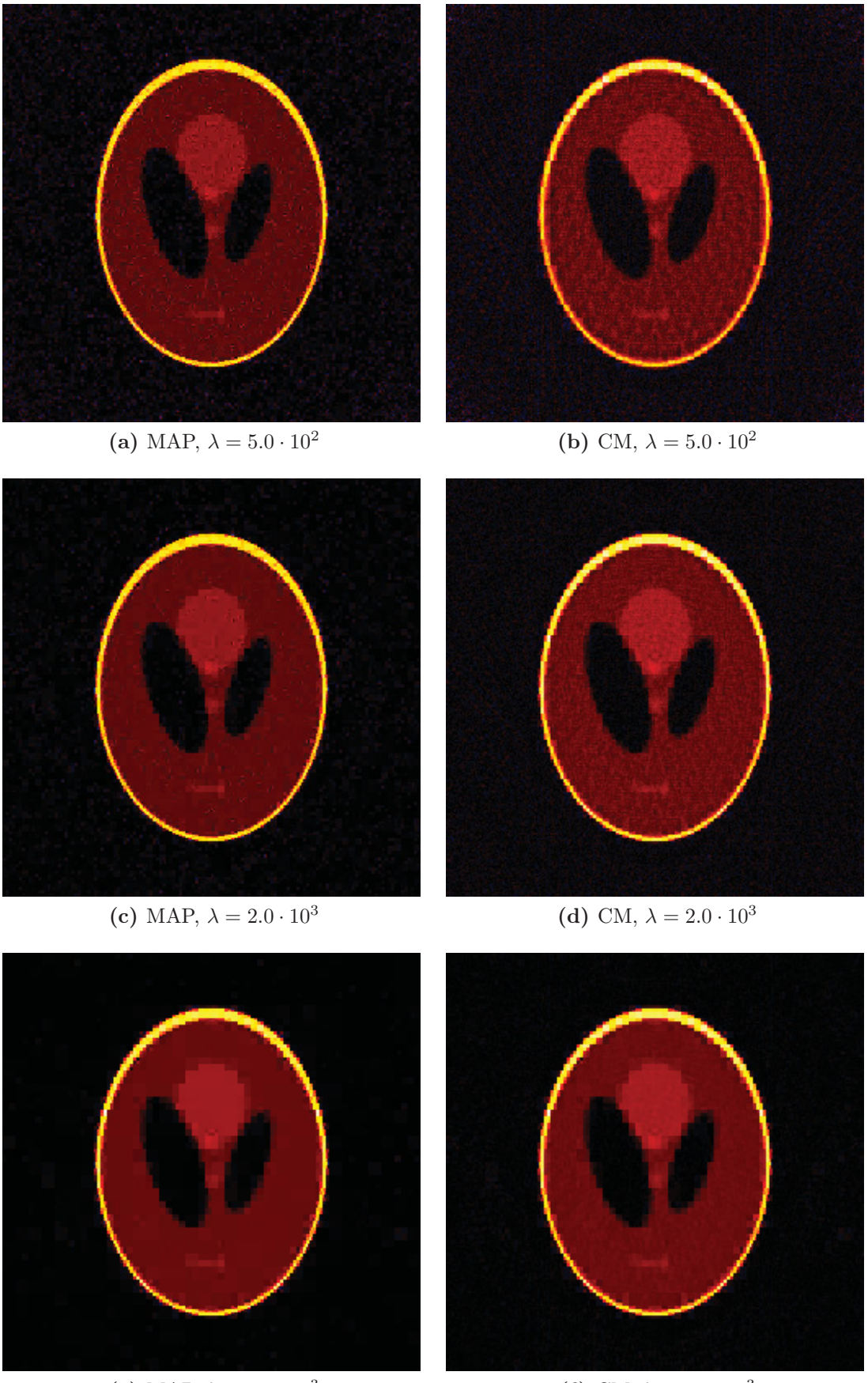


Figure 5.17.: "Phantom-CT" reconstructions using a Besov prior and $\lambda = 2 \cdot 10^4$. MAP (left column) and CM (right column) estimates are shown for increasing n.



(a) MAP, $\lambda = 5.0 \cdot 10^2$ **(b)** CM, $\lambda = 5.0 \cdot 10^2$

(c) MAP, $\lambda = 2.0 \cdot 10^3$ **(d)** CM, $\lambda = 2.0 \cdot 10^3$

(e) MAP, $\lambda = 8.0 \cdot 10^3$ **(f)** CM, $\lambda = 8.0 \cdot 10^3$

Figure 5.18.: "Phantom-CT" reconstructions using a Besov prior with $n = 256 \times 256$.

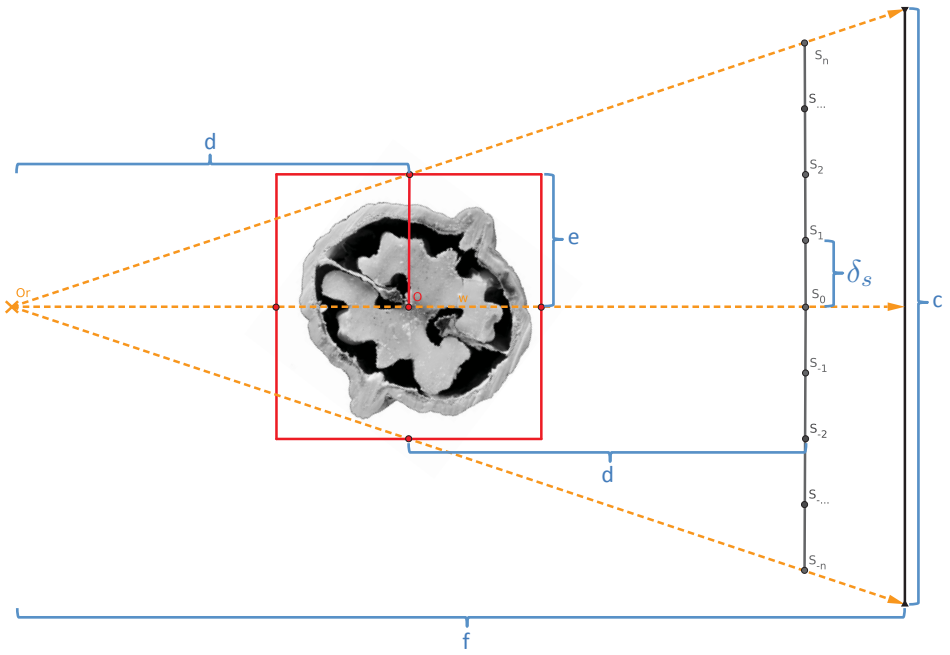


Figure 5.19.: Sketch of the measurement setup: d = 110mm, f = 300mm, c = 144.8mm, e = 21.05mm, δ_s = 0.0367mm

5.3. Computed Tomography Studies

In this section, we will show how to apply the methods developed in this thesis to the real-data "Walnut-CT" scenario. For this, we will first have to discuss details of the measurement setup and the noise modeling. A general review of Bayesian inversion applied to CT can be found in Kolehmainen et al. (2003), Siltanen et al. (2003).

5.3.1. Measurement Setup

As described in Section 2.3.3, the X-ray source faces a fixed, planar detector and the target is placed on a rotatable bar in-between (see Figure 2.8a). In the following, we will only describe those aspects of the measurement setup that are important for the 2D reconstruction of the central slice of the walnut:
The detector is composed of $M = 2296$ pixel with 0.05mm width, yielding a total detector width of 144.8mm. The distance between source and detector is 300mm and the distance between source and target is 110mm. We assume that a virtual detector is placed 110mm away from the target instead of the real detector. Thereby, we eliminate one parameter from the description of the fan-beam forward operator (although incorporating two different distances is actually trivial in our implementation approach). The transformed pixel-width of the virtual detector is $\delta_s = 0.0367$. The use of the virtual detector leads to a linear scaling of the reconstructed activity which could

(a) (b) (c)

Figure 5.20.: (a) TV reconstruction (128×128 pixel) used in the noise modeling procedure. (b) Binary source mask derived from it. (c) TV reconstruction restricted to the mask.

easily be removed after reconstruction. We choose the target to define the center of the physical coordinate system. The area used for the reconstruction is the square $[-e, e]^2$, where $e = 21.05$mm. See Figure 5.19 for a geometrical sketch of the setup. The length e will define the unit-length in the mathematical coordinate system used to formulate and implement the inversion algorithms: $[-e, e]^2$ will correspond to $[-1, 1]^2$. Full angle recordings with an angular spacing of 3° were recorded (see Figure 2.8c).

5.3.2. Noise Modeling

As discussed in Section 3.1, additive Gaussian noise models, $\varepsilon \sim \mathcal{N}(\mu_\varepsilon, \Sigma_\varepsilon)$, provide good approximations in many real data scenarios. For typical CT applications, a detailed discussion, review and justification of the Gaussian noise model can be found in Siltanen et al. (2003). Unfortunately, we do not have reference measurements that would allow to estimate μ_ε and Σ_ε independently from the data f we use for the reconstruction. Therefore, we have to use strongly simplifying assumptions on μ_ε and Σ_ε and estimate them relying on a kind of boot-strap procedure using f only (which contains only a single realization of ε!): We will make use of the a-priori knowledge that the measurement setup was chosen such the walnut covers only a fraction of the field of view of every angular projection. Therefore, there will be sensor pixels where $f^\dagger$ is 0 and only ε determines f. To identify those pixels, we need a rough estimate of the support of $u^{\dagger,\infty}$. We obtain this by exploiting that the walnut has a clear defined boundary. Therefore, edge-preserving reconstruction techniques should be able to identify this boundary reasonably well, even if $\varepsilon \sim \mathcal{N}(0, I_m)$ is assumed as a noise model. We computed the MAP estimate for a TV prior on a coarse $n = 128 \times 128$ computational grid using a high value of λ to ensure that background and support are well separated.

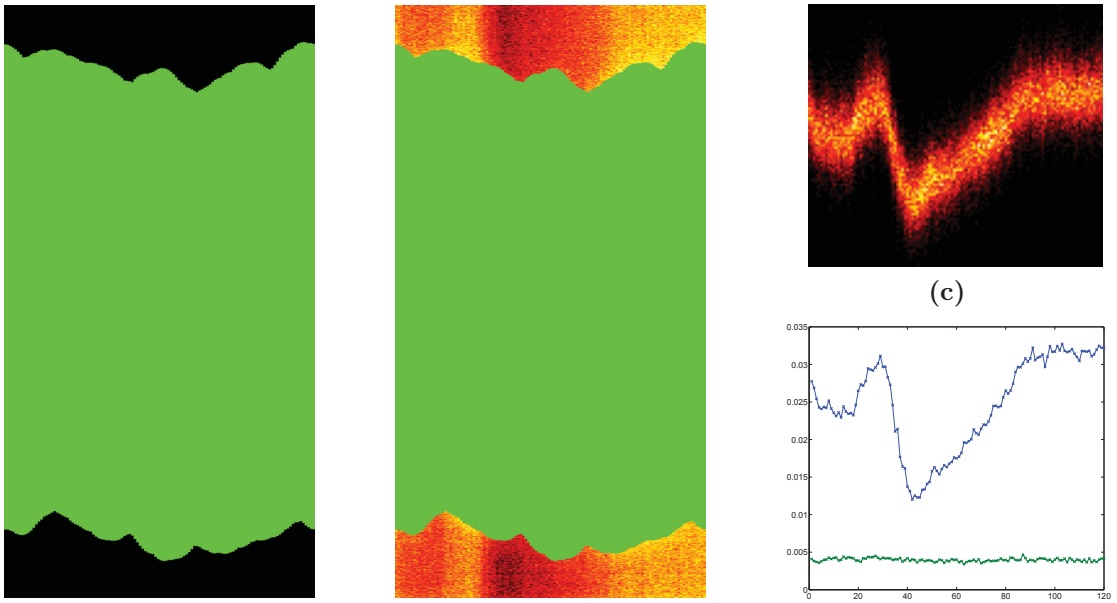


Figure 5.21.: (a) Data mask derived from a forward mapping of the source map. (b) Rescaled full sinogram restricted to the complement of the data mask. In both images, the resolution in angular direction was increased by a factor of 10. (c) Anglewise histograms of the noise pixels. The horizontal axis corresponds to the angle. (d) Anglewise empirical mean (blue line and marks) and standard deviation (green line and marks) of the noise pixels.

Figure 5.20a shows the solution used. This solution was subsequently thresholded and the resulting binary mask was filled and dilated by one pixel in all directions. Figure 5.20b shows the source mask and Figure 5.20c shows the MAP estimate restricted to it. The support of the sinogram of the source mask was computed to construct a data mask (see Figure 5.21a). In any pixels not belonging to this mask we can assume that $f^\dagger$ is 0. These 55 463 pixels (see Figure 5.21b) are now used to estimate μ_ε and Σ_ε. The simplest noise modeling would consist of computing the global empirical mean and variance of all noise pixels, μ_{gl} and σ_{gl}^2, and assuming an i.i.d. Gaussian noise model on every single pixel with these global statistics: $(\mu_\varepsilon)_i = \mu_{gl}$ for all i and $\Sigma_\varepsilon = \sigma_{gl}^2 I_m$. However, Figure 5.22a shows a (normalized) histogram of the noise pixels with respect to the 107 different discrete values they take compared to a normal distribution $\mathcal{N}(\mu_{gl}, \sigma_{gl}^2)$. One can see that this global approximation leads to a considerable misfit between model and reality. From Figure 5.21b, one can already guess that the noise statistic seems to depend on the projection angle θ. Figure 5.21c shows a 2D histogram in which the angular dependence is resolved, while 5.21d shows the angle-wise empirical means and standard deviations: While the standard deviation stays almost constant over angles, the mean shifts. A possible explanation for this might be that the noise in the image is mainly caused by

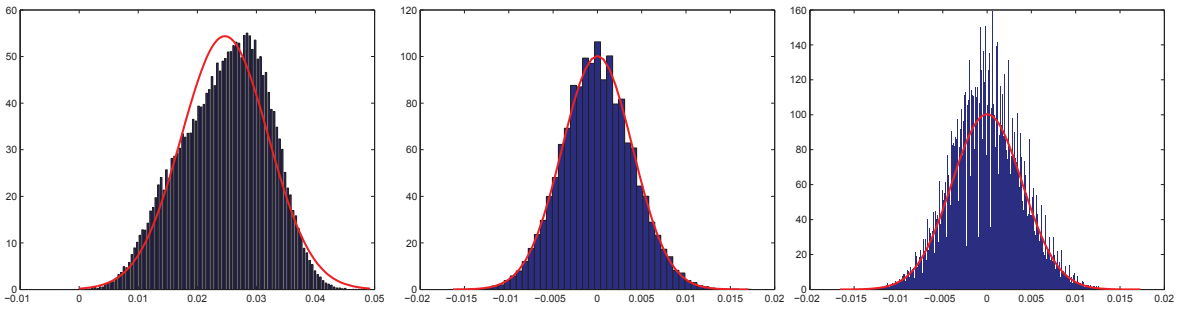


Figure 5.22.: Normalized histograms of all noise pixels (blue bars) and normal pdfs (red lines) fitted to them: (a) without correction, $\mathcal{N}(\mu_{gl}, \sigma_{gl})$, (b) after subtraction of the angle-wise mean, $\mathcal{N}(0, \sigma_{gl2})$, using 50 bins and (c) using 500 bins.

photons scattered inside the walnut or on the interface between air and walnut (notice that there is a halo around the walnut in Figure 2.8b). Then, the angular dependence would be reasonable as the scattering depends on the concrete position of the walnut with respect to the X-ray source. Scattering is a modeling error or nuisance term rather than a classical measurement noise term (cf. Sections 2.3 and 3.6.2). As it strongly relies on the object to to recover, calibration measurements to pre-estimate μ_ε and Σ_ε would not be of too much help in our situation. Based on these findings, we refine the noise model by assuming that $\varepsilon(i, \theta) \sim \mathcal{N}(\mu_\theta, \sigma_{gl2})$, i.e., we use an angle depended mean and a modified global standard deviation. For estimating the empirical angle-wise mean, between 387 and 543 measurement values per angle can be used. Figure 5.22b shows a histogram of all pixel values after their angle-wise mean has been subtracted. The red line shows a normal distribution with zero mean and a standard deviation estimated from all 55 463 mean-corrected pixels. Assuming that the standard deviation has no angular dependence seems like a reasonable approximation from Figure 5.21d. While extending the noise model to an angle-wise standard deviation would easily be possible, it would unnecessarily increase the number of parameters to estimate from the single noise realization in this boot-strap fashion. In such a situation, a conservative modeling approach involving less parameters is to be preferred. Figure 5.22b seems to suggest that the chosen normal approximation is nearly optimal. However, if one increases the number of bins used to compute the histogram from 50 to 500, one can see in Figure 5.22c that it fits less well on smaller scales as the discrete nature of the measurements is still present. As the angular spacing of 3° corresponds to a rather sparse angle setup, i.e., a low resolution in θ-direction, we will also decrease the sensor resolution, i.e., the resolution in s-direction: Every four subsequent pixels of the 2296 original pixels were added up (see Figure 2.8c). The noise model for these 574 virtual pixels has to be adjusted accordingly.

5.3.3. Reconstruction Results

Full Angle

First, we examine reconstructions using a full (but sparse) set of 120 angles from 0° to 357°. Using a TV prior with λ chosen by manual inspection and a resolution of $n = 256 \times 256$ pixel, we computed MAP estimates by ADMM and used the RSG sampler to draw posterior samples. Compared to the "Phantom-CT" scenario, the sampler needs a lot more burn-in steps when initialized by $u^0 = 0$: Early samples show ring-like artifacts on the boundaries that seem to stem from the fan-beam geometry. The problem can easily be avoided by choosing another initialization u^0, for instance the result of a few ADMM iterations for computing the corresponding MAP estimate. In Figures 5.24a and 5.24b, MAP and CM estimates are compared. From the visual impression, they seem to coincide. In Figures 5.23c and 5.23d, pixel values above 20 % of the maximal intensity were set to white and the color scale was adjusted to the remaining values. Thereby, we can see that on smaller scales, the MAP estimate features the typical "staircase" artifact of producing artificial jumps while the CM estimate does not seem to suffer from it. This confirms the findings of Louchet and Moisan (2013), who examined a denoising scenario ($A = I_n$). However, we know from our results in the "Phantom-CT" scenario (cf. Figure 5.10), that in return, the CM estimate also blurs real edges. Figure 5.24c shows the CStd estimate. We see that most of the variability of the posterior (which reflects uncertainty) seems to concentrate on feature boundaries. Within the boundaries, there are hot-spots of high uncertainty.

Images of the ℓ_2 norm of the gradient of a structural image u are of interest for a variety of applications, in particular to enhance the inversion of functional imaging data in multimodal imaging (cf. Section 1.1 and Figure 1.3). In Figure 5.23f, we computed the CM estimate of the gradient image:

$$\int \|\nabla u\|_2 \, p_{post}(u|f) \, \mathrm{d}u \tag{5.5}$$

Although $\|\cdot\|_2$ is not a linear function, the relative error to simply taking the gradient image of $\hat{u}_{\mathrm{CM}}$ (cf. Figure 5.24b) is only 0.3%. With 2.3 %, the relative error to the gradient image of $\hat{u}_{\mathrm{MAP}}$ (cf. Figure 5.24a) is only slightly larger. Therefore, we did not include these images here.

As a second prior model, we examine Besov space priors with $p = 1$, formulated in the Haar wavelet basis (cf. Sections 3.2.4 and 5.2.6). Our results complement Hämäläinen et al. (2013), who used the same data to compare MAP estimates to filtered back projections for a decreasing number of angles and to examine a sparsity-based criterion to choose λ. Figure 5.24 shows MAP, CM, and CStd computed for $\lambda = 750$ (found by

manual inspection) and $n = 512 \times 512$. Similar to our results in the "Phantom-CT" scenario, and to the TV results presented in the last paragraph, MAP and CM visually coincide. The CStd result shows that also using the Besov prior, most of the uncertainty concentrates in feature boundaries.

Limited Angle

Now, we examine reconstructions using a limited and sparse set of only 41 angles from 30° to 150°. In Figure 5.25, results corresponding to those computed in the last section are show. Certain structures, such as parts of the shell, cannot be reconstructed with this setup. While one would assume that such regions carry a larger uncertainty than the regions that can be recovered, the CStd estimates show that the opposite is true. Compared to the full angle setup, another artifact of the reconstructions is more apparent: With both priors, the MAP and CM estimates of the non-negative mass absorption coefficient u are negative in several pixels. To avoid this, we can incorporate the a-priori information of non-negativity as a hard constraint (cf. Section 3.2.2). In Figure 5.25, CM and CStd estimates for a TV prior with $\lambda = 0.1$ are compared with or without non-negativity constraints. Besides the negative regions, other image artifacts are removed as well. The CStd estimates (which share the same color scale) show that a significant reduction of uncertainty is achieved, in particular in the background regions. Comparing Figures 5.25b and 5.26b, we see that without constraints, λ has to be chosen four times larger to achieve a similar reduction of noise and artifacts. But this comes at the cost of smoothing out more image details as well.

5.3.4. Discussion

A major aim of this section was to demonstrate the application of various Bayesian inversion techniques presented and developed in this thesis to a challenging, high-dimensional real-world problem:

- In Section 5.3.1 the physical measurement setup was converted into a mathematical forward model.
- Section 5.3.2 examined a boot-strap approach to define the noise model, i.e., the likelihood distribution.
- Finally, we computed and examined different estimates for different prior models in Section 5.3.3. Besides the classical MAP and CM estimates, CStd estimates were computed to assess the spatial distribution of the posterior variance, which reflects the uncertainty of the reconstructions. In addition, the CM estimate of

the ℓ_2 norm of the gradient was computed as an example of an estimator of a feature $g(u)$ instead of u itself.

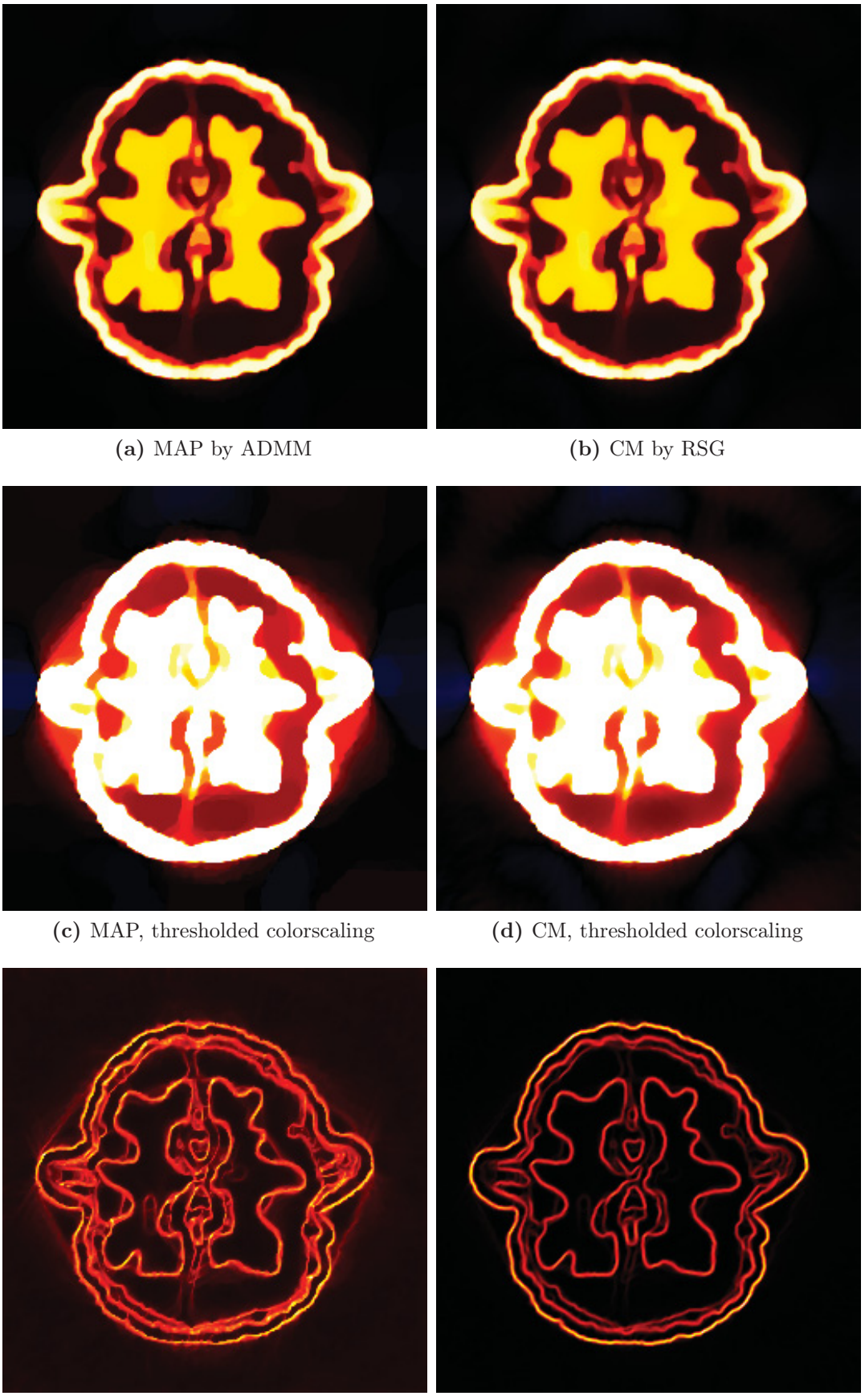
Several observations about the concrete results can be made:

- If we compare the results for TV and Besov priors, we again have to admit that the visual impression of the TV results is more convincing (cf. Section 5.2.6). In the Besov results, the blocky shape of the Haar wavelets (cf. Figure A.3) is clearly visible. Figure 5.26e compares the ability of different wavelet families to compress the photograph of the walnut, Figure 1.6a. We see that testing other wavelets as a basis for the Besov prior might be an interesting topic for future studies. Another improvement might be the development of *isotropic* or *structured* Besov priors, which exploit the scale-space relationships between the different wavelet coefficients more efficiently.

- Using the TV prior, MAP estimation produces results with sharper edges compared to CM estimation. However, not all of these edges need to be feature edges which leads to the well-known staircase artifact. Not surprisingly, the size of the differences depends on the amount of information given by the data: While the differences in the full angle setup are hardly visible using a normal color scaling, they are more pronounced in the limited angle case.

- Our results confirm that the inclusion of hard constraints such as non-negativity can enhance the reconstruction quality significantly. Although this finding is already quite well known, it is often neglected in designing inversion techniques due to the additional implementation effort.

- One motivation behind computing CStd estimates was that we expected them to provide an uncertainty image that allows to identify the regions missed by limited in contrast to full angle tomography easily. Unfortunately, this expectation was not met.



(a) MAP by ADMM

(b) CM by RSG

(c) MAP, thresholded colorscaling

(d) CM, thresholded colorscaling

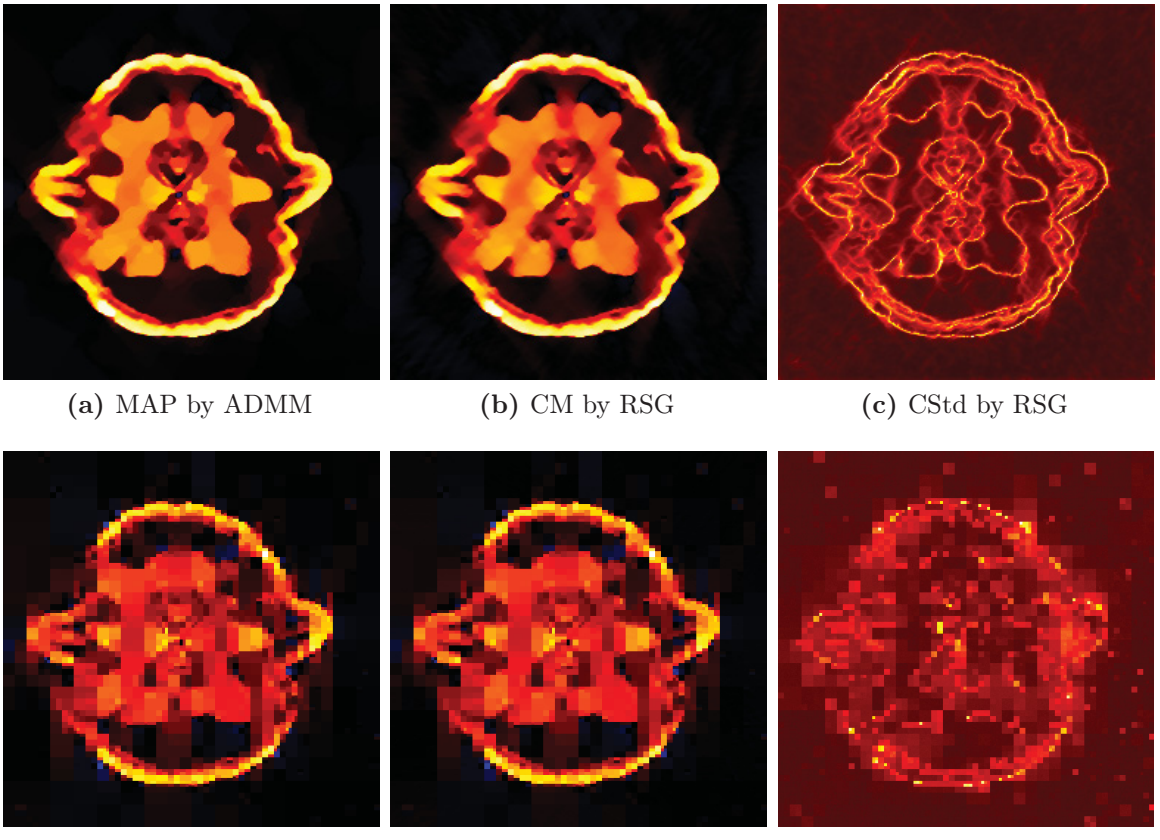
(e) CStd by RSG

(f) CM of $\|\nabla u\|_2$ by RSG

Figure 5.23.: Different estimates in the full angle "Walnut-CT" scenario using a TV prior with $\lambda = 3$ and a resolution of $n = 256 \times 256$ pixel.

(a) MAP by ADMM **(b)** CM by RSG **(c)** CStd by RSG

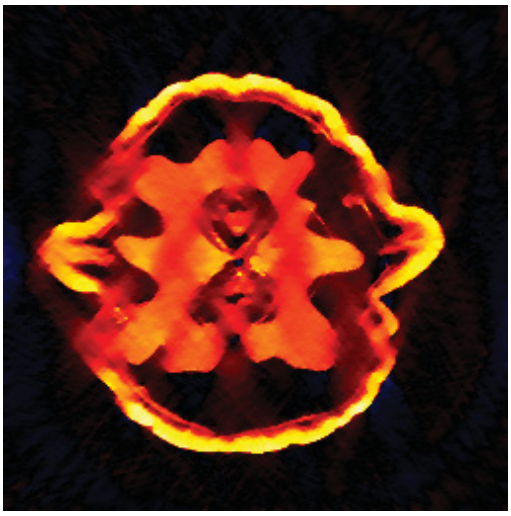
Figure 5.24.: Different estimates in the full angle "Walnut-CT" scenario using a Besov prior with $\lambda = 750$ and a resolution of $n = 512 \times 512$ pixel.



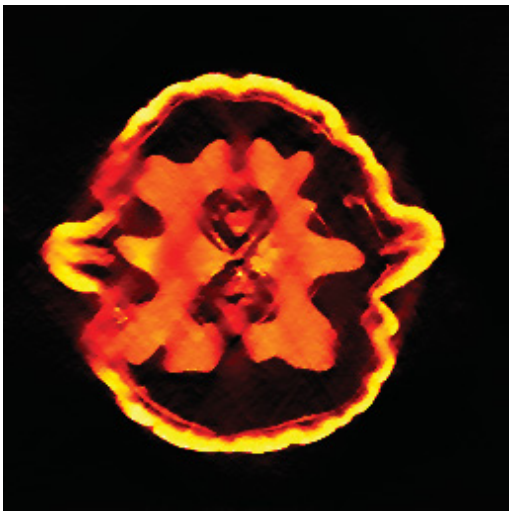
(a) MAP by ADMM **(b)** CM by RSG **(c)** CStd by RSG

(d) MAP by ADMM **(e)** CM by RSG **(f)** CStd by RSG

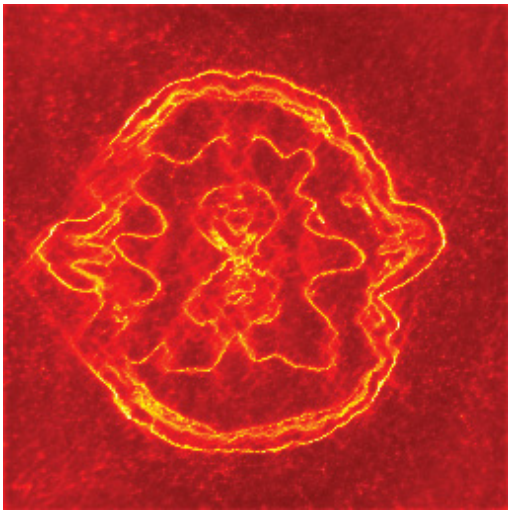
Figure 5.25.: Different estimates in the limited angle "Walnut-CT" scenario using (a)-(c) a TV prior with $\lambda = 0.4$, $n = 256 \times 256$ and (d)-(f) a Besov prior with $\lambda = 250$, $n = 512 \times 512$.



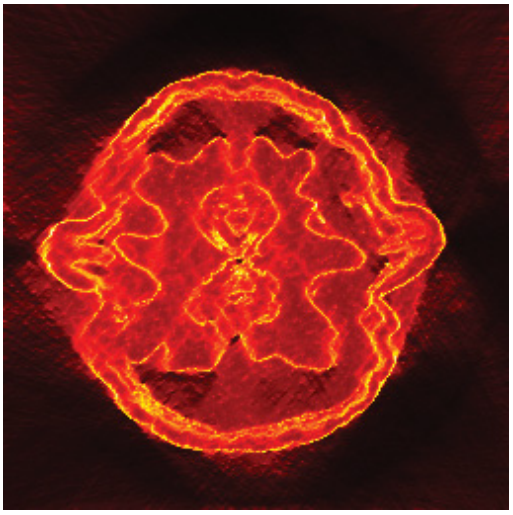
(a) CM, no constraints



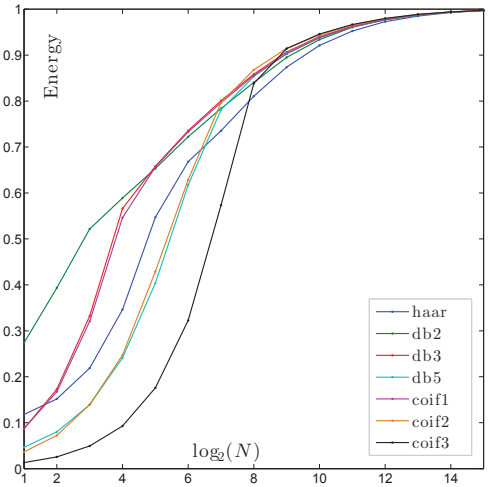
(b) CM, non-negativity constraints



(c) CStd, no constraints



(d) CStd, non-negativity constraints



(e)

Figure 5.26.: (a)-(d) CM and CStd in the limited angle "Walnut-CT" scenario using a TV prior with $\lambda = 0.1$ with or without non-negativity constraints. The color scales of both CM estimates and both CStd estimates are equal. (e) Compression rates of different wavelet representations of Figure 1.6a: The fractional energy $\|X_N\|_2^2/\|X\|_2^2$ of the N largest wavelet coefficients is plotted vs. the truncation index N in powers of 2.

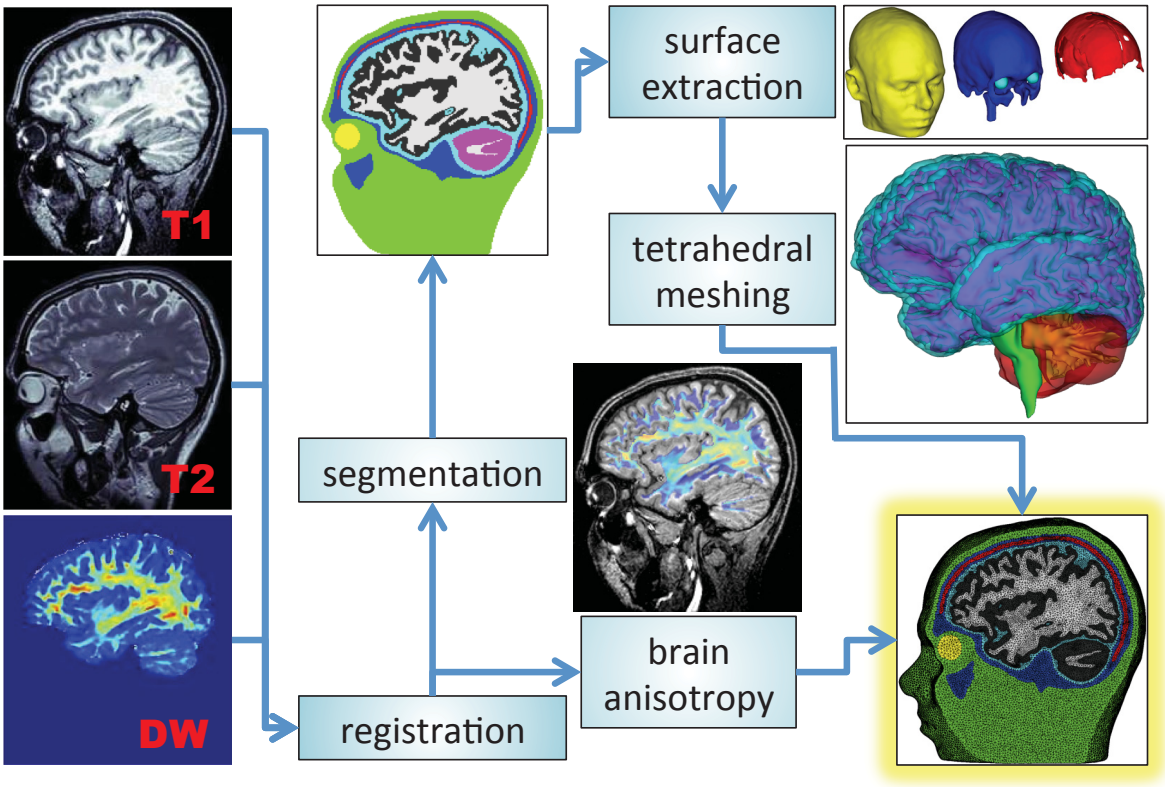


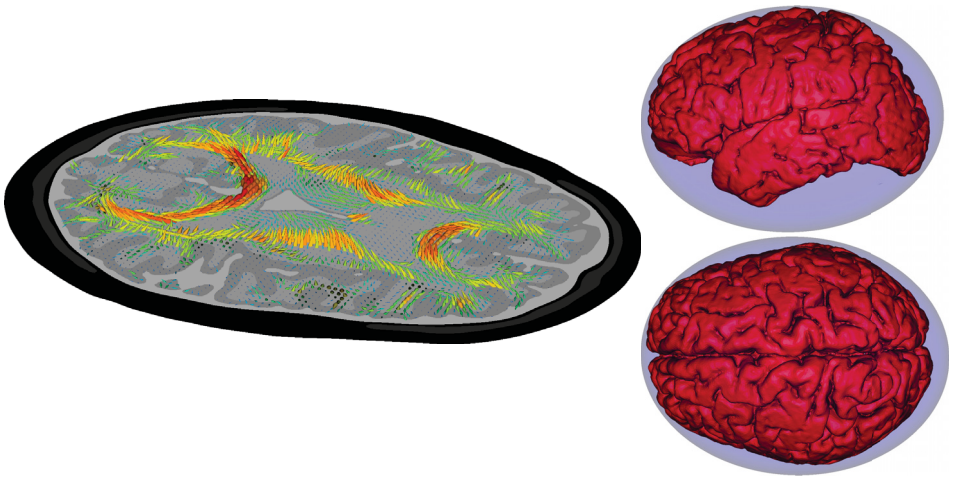
Figure 5.27.: Sketch of the head model generation pipeline.

5.4. EEG/MEG Source Reconstruction Studies

In the last section of this chapter, we will present the results for the EEG/MEG source reconstruction studies introduced in Section 2.4.2. Compared to the other applications considered in this thesis, source reconstruction comes with the most practical effort: Sophisticated head modeling, source space construction, real data processing, noise modeling and inverse reconstruction are often tedious to carry out and implement and require the integration of different software tools and visualization techniques. The scope and extent of this thesis does not permit an exhaustive description of all procedures and topics relevant to this work. Therefore, many details will be cited from other publications.

5.4.1. Head Model Generation

For the specific aims of our studies, we need a highly realistic, individual, anisotropic head model and cannot rely on standard head models (semi-)automatically provided by currently available software solutions. Figure 5.27 sketches the steps required to obtain our model. First, different MRI scans are acquired: *T1*- and *T2-weighted* scans provide anatomical information about the head's soft tissues while *diffusion-weighted* (*DW*)



(a) (b)

Figure 5.28.: (a) The first eigenvectors of the anisotropic conductivity tensors of GM and WM scaled by the corresponding FA values and visualized by bi-directional color-coded cones in one transversal slice of the data set. The spatial resolution of the tensor data was decreased, a cross-section of the FEM model with gray scale color-coded elements was added. The corpus callosum connecting left and right cerebral hemispheres and the pyramidal tracts connecting upper motor neurons and the brainstem are most noticeable. (b) Comparison between the GM surface (red) and the ellipsoid surface representing the intercranial tissues (light blue).

scans provide information on the local diffusion properties of water molecules. For a joint usage of the information obtained, all MR images need to be transformed into the same coordinate system. This is an *image registration* problem; see Modersitzki (2004) for a general overview. Especially registering the DW-MR images is a non-trivial task (Ruthotto et al. 2012).

Once this is achieved, the anatomical images can be used to classify every image voxel into a predefined set of tissues that one would like to differentiate. This problem is called *image segmentation*; see Section 4 in Aubert and Kornprobst (2006) for a general reference and Lanfer (2014) for a recent overview on head tissue segmentation for EEG source reconstruction. The different tissue contrast of T1 and T2 images can facilitate this problem. The result is a labeled voxel-image (cf. Figure 5.27). For our studies, we segmented the image into ten tissue compartments: Skin, eyes, cortical/compact and cancellous/spongious bones of the skull (called "skull compacta" and "skull spongiosa" from now on), cerebrospinal fluid (*CSF*), brainstem and gray and white matter of the cerebrum and the cerebellum. In the following, we will refer to the cerebrum gray- and white matter by *GM/WM* (we will not perform source reconstruction of cerebellum activity).

The geometry of a tissue compartment is defined by the cluster of voxels that carry

its label. In principle, using slight adaptations, this voxel-based geometry description can be converted directly into a hexahedral finite element mesh. See Section 1.4.2. in Aydin et al. (2014) and references therein for such an approach. Here, we aim to describe the geometries of the tissue compartments by surface representations. For this, we extracted high resolution triangular surface meshes from the voxelized compartments. After some further processing, these meshes can be used to generate tetrahedral finite element meshes fulfilling certain geometrical constraints (*constraint Delaunay tetrahedralizations*). A fine spatial meshing is required to achieve a high level of detail and to fulfill constraints on the volume of the elements (*volume constraints*), which is required to obtain a satisfactory numerical accuracy. Details will be discussed in the next section. Note that the resulting FEM mesh is labeled, i.e., every tetrahedron belongs to a tissue compartment.

The DW-MRI data can be used to estimate the water diffusion tensor for every voxel (*diffusion tensor imaging, DTI*). Its eigenvectors and eigenvalues reflect the local direction of the tissue. If the eigenvalues are very different from each other, the tissue in the corresponding voxel is highly directed. The *fractional anisotropy* (*FA*) measures this difference, and thereby, the degree of anisotropy, by a scalar value between zero and one in every voxel and provides a quantitative image-based measure for clinical diagnostics (see Basser and Pierpaoli 1996, for a formal definition). The DW-MRI data is symbolized by such an image in Figure 5.27: The image above the "brain anisotropy" box is an overlay of the thresholded FA image with the T1 images. There are several physical models to convert the water diffusion tensor into a conductivity tensor that we can use for modeling GM and WM as anisotropic. For this thesis, we relied on the approach proposed in Rullmann et al. (2009). Figure 5.28a visualizes the first eigenvectors of the resulting conductivity tensors.

Notes and Comments

This head model was also used in Janssen et al. (2013), Lucka (2011), Lucka et al. (2012), Pursiainen et al. (2012), Rampersad et al. (2014), Vorwerk et al. (2014). The most thorough description of the technical details omitted in this section is given in the supplementary material[2] of Janssen et al. (2013). However, note that the procedure used to convert the water diffusion tensors into conductivity tensor as described in Section "Anisotropic conductivity tensors" therein, namely the one of Opitz et al. (2011), differs from the approach used in this thesis. In Janssen et al. (2013), the head model was used to simulate *transcranial magnetic stimulation* (*TMS*). Therefore, the importance of many modeling steps was motivated by references from the field of

[2] stacks.iop.org/PMB/58/4881/mmedia

Table 5.7.: Compartments and conductivities used for the different head models. An "i" indicates that this compartment is included in the model and modeled isotropic whereas an "a" means that it is modeled anisotropic.

Compartment	σ (S/m)	HM1	HM2	HM3	HM4	HM5	HM6
Skin	0.43	i	i	i	i	i	-
Eyes	0.505	i	i	-	-	-	-
Skull	0.01	-	-	i	i	i	-
Skull comp.	0.0064	i	i	-	-	-	-
Skull spong.	0.02865	i	i	-	-	-	-
CSF	1.79	i	i	i	-	-	-
GM (cerebrum)	0.33	a	i	i	-	-	-
WM (cerebrum)	0.14	a	i	i	-	-	-
Cerebellum GM	0.33	i	i	i	-	-	-
Cerebellum WM	0.14	i	i	i	-	-	-
Brainstem	0.33	i	i	i	-	-	-
Intracranial	0.33	-	-	-	i	i	i

electro-magnetic brain stimulation. Lucka et al. (2012), Pursiainen et al. (2012), Vorwerk et al. (2014) contain references that motivate them for EEG/MEG source reconstruction. Section A.8 contains a listing of the software packages used.

5.4.2. Head Model Cascade

We used the data set described in the last section to construct a series of different head models reflecting various degrees of realism encountered in EEG/MEG source reconstruction (cf. Section 2.4.2 and Figure 2.13). For the construction of the most realistic head models, called HM1 and HM2, all surfaces were meshed. The element size in the skull compartments and in the GM was restricted to 1mm^3, in all other compartments it was restricted to 3mm^3. This results in a triangulation consisting of 984 569 vertices and 6 107 561 elements. A compartment representing the CSF was constructed using a closed inner skull surface. In HM1, GM and WM are anisotropic as described in the last section whereas they are isotropic and homogeneous in HM2. For head model HM3, only the surfaces skin, skull compacta, brainstem, gray and white matter of the cerebrum and the cerebellum were used. For meshing, the same volume constraints were used which resulted in a triangulation consisting of 931 564 nodes and 5 782 609 elements. A CSF compartment was included as well. For head model HM4, only the surfaces skin and skull compacta were used. For meshing, a volume constraint of 1.2mm^3 was used which resulted in a triangulation consisting of 885 655 nodes and 5 591 986 elements. A compartment representing all intracranial tissues was constructed using a closed inner-skull surface. This three-compartment head model corresponds to

the head models most commonly used in BEM approaches (cf. Section 2.4.1).
Head models HM1-HM4 are realistically shaped. As described in the previous section, the construction of such head models requires the acquisition of anatomical MRI images. If such data is not available, simplified geometries like spheres are fitted to the sensor positions. An advantage of such models is that semi-analytical formulas for solving the forward problem are known, which can be evaluated fast (e.g., Munck and Peters 1993). We want to mimic the use of such models and compare them to the realistically shaped ones. For this, we will later need to construct source space positions $r_1, \ldots, r_N$ that are valid source locations in all head models. It turned out that using spherically shaped head models fitted to the sensor positions is too restrictive for this purpose. Instead, an ellipsoid was used as the basic geometric shape. In addition, the fit needed to incorporate the actual GM surface in addition to the sensor positions to result in a compatible head model. In the following, we outline the construction of the ellipsoid model: In a first step, a triangular surface of the outer CSF compartment boundary (which corresponds to the inner skull boundary) was extracted from HM1. All vertices inside the convex hull of the EEG sensor positions were determined and their convex hull was constructed as a triangulated surface. The number of faces of this surface was reduced to 100 by Matlab's `reducepatch.m` routine, resulting in a smoothed surface consisting of almost equal size faces. The 52 vertices of this surface were then projected to the actual GM surface vertices, resulting in 52 equally sampled locations in the superior GM surface areas (in the convex hull of the EEG sensors). An ellipsoid was fitted to them (details are given in Section A.8). This initial fit was then refined to enclose all GM surface vertices: Iteratively, the ellipsoid was re-centered in the transversal plane and then all semi-axes were simultaneously slightly enlarged or shrunk. Figure 5.28b shows the final ellipsoid representing the intracranial tissue surface. The ellipsoids representing skull and skin surfaces were computed by assuming a ratio of 1:0.93:0.85 between skin, skull and brain semi-axes lengths. After that, triangulated surfaces of the different compartments were generated using Matlab's `isosurface.m` routine and were meshed like the realistically shaped head models. The number of surface triangles and the volume constraints used for meshing were chosen to match the number of FEM vertices and elements of the other head models. Note that it would be possible to solve the forward problem for this geometry using explicit formulas (Dassios 2009). However, for the sake of compatibility with the other models, the same FEM-based forward approach was used for the ellipsoidal models as well. While head model HM5 consists of the three compartments skin, skull and intracranial tissue as described above, head model HM6 consists of only one compartment, generated by the ellipsoidal skin surface.
Table 5.7 lists the compartments and their conductivities used for creating the different

head models. References for their choice can be found in Baumann et al. (1997), Dannhauer et al. (2011), Ramon et al. (2006), Rullmann et al. (2009).

5.4.3. Source Space Construction and Forward Computation

The location of the source space nodes is a crucial choice for EEG/MEG source reconstruction by CDR methods (cf. Section 2.4.2). In principle, the neural generators of the EEG/MEG signal are the pyramidal cells located in parts of the GM (Hämäläinen et al. 1993, Murakami and Okada 2006, Nunez and Srinivasan 2005, Okada et al. 1997). *Volume-based* source space constructions aim at a discretization of this compartment into voxels and place three orthogonal dipoles at the center of each voxel. The reconstructed source activity can be represented as a 3D image which can easily be aligned with other 3D images. For instance, it can be compared with fMRI activation or visualized overlayed on an anatomical T1-weighted MRI. Due to the deep but thin sulci and the strong folding of the cortex, this approach requires a very accurate segmentation. In addition, the 3D character of the spatial discretization leads to a cubic growth of n with decreasing spatial resolution of the source space. These difficulties often lead to the usage of *surface-based* source space constructions. These approaches rely on the fact that the pyramidal cells are organized in thin layers oriented normal to the cortical surface: They aim to discretize a 2D surface representing such a layer rather than the whole volume. This leads to a quadratic growth of n with decreasing spatial resolution. Furthermore, the directedness of the pyramidal cells with respect to the surface can be used to constrain the orientation of the reconstructed current vectors: In the most extreme case, only one normally oriented dipole is placed at each location, which reduces n by a factor of three and simplifies the vector reconstruction to a scalar reconstruction problem. This is called *normal constraint* (NC) or *cortical orientation constraint*. As the forward operator is very sensitive to changes in dipole orientation, this constraint requires a very accurate surface segmentation and a very fine spatial resolution. *Loose orientation constraints* (LOC) (Lin et al. 2006) circumvent these problems by computing local orientations statistics to allow for a certain variability of the current direction around the local averaged normal direction. Thereby, the effective dimension of the current vector at a given location is in-between one and three. LOC can be realized using block-based prior models with weightings. The physiological a-priori knowledge about the neural generators actually allows for incorporating a further constraint on u, which is rarely discussed in source reconstruction literature: The pyramidal cells not only determine the orientation of the currents but also its direction. Using the NC, this can be modeled by using an additional non-negativity constraint on u (cf. Section 3.2.2). We will examine the additional benefit of this

constraint in Section 5.4.6.
Both volume and surface based approaches have advantages and disadvantages. Surface based approaches often rely on a flattened and smoothed representation of the cortical surface, which does not include the deep-lying GM areas or areas encased by WM (e.g., the insular, the cingulate cortex, the hippocampus or the thalamus). Nevertheless, working with such surface representations is reasonable and even advantageous for a wide range of experimental designs. However, other brain networks often involve deep-lying sources as well. For instance, the evoked potentials and fields examined in this thesis also contain components reflecting such activity (Parkkonen et al. 2009, Sander et al. 2010, Scherg and Buchner 1993) and the analysis of both networks is a common clinical application of EEG/MEG.

Volume-based Construction

In principle, this construction approach is straight-forward to implement. A regular spatial grid is constructed such that it encases the head model. The grid points will represent the center of the source space voxels and the grid spacing their size. The Venant approach we use for the forward computation (cf. Section 2.4.2) imposes a technical constraint on possible source locations: All FEM elements attached to the nearest FEM node to a source location have to be GM elements (*Venant constraint*). Therefore, all grid points not meeting this constraint are discarded. All remaining points that are located inside a GM FEM element build the source grid. A fast implementation of this construction requires to pre-compute which elements are attached to a node, which elements are connected via common faces and the usage of convex hulls and *k-d trees* for nearest neighbor searches.
If a source space with a fixed number of voxels instead of a fixed voxel-size should be constructed , the above procedure has to be iterated until a suitable grid is found. As the number of voxels obtained with the above procedure is a complicated, discrete and non-monotonic function of grid spacing and location, a robust fitting heuristic had to be implemented.
It is often advantageous to have a parceling of the gray matter volume with respect to the source locations found (*source volumes*). For instance, for defining neighborhood relations or visualization purposes. Using FEM head models, we can construct the source volumes by first assigning all FEM nodes to a source location. Thereby, each FEM element is assigned to one or more source volumes by its nodes. If it is assigned to more than one source volume, it lies on the boundary between source volumes. The initial assignment of all FEM nodes to a source location should not be done by nearest neighbor searches based on the normal, euclidean distance: As the gray matter volume

is very non-convex due to the strong folding of the cortex, this can result in scattered, disconnected source volumes. Instead, we first map each source location to the nearest FEM node based on the euclidean distance and assign all remaining FEM nodes to these source volume centers based on the euclidean graph distance on the triangulation: The distance between two nodes of the triangulation is given by the shortest sequence of edges connecting both nodes. Practically, the assignment can be computed using Dijkstra's algorithm (Cormen et al. 2001, Dijkstra 1959)

Surface-based Construction

This construction approach is particularly popular for BEM head models, for which the geometries are described by triangulated surfaces anyway (cf. Section 2.4.2). The GM surface, the interface between GM and WM or a surface constructed in-between the former two is used for the source space construction. The nodes and the normal vectors of the facets of that surface can directly be used to define the source space. For the volume-based FEM head models, surface-based source space construction is more difficult: For all forward approaches, the locations of the source space nodes should lie in the interior of the GM compartment, not on its boundary. Therefore, one has to find a way to choose source locations inside the GM with a meaningful correspondence to the GM's surface. Meanwhile, the method should also lead to an equal parceling of the surface. In the following, we outline the final approach we developed:

1. A pool of candidate locations $C = \{c_1, \ldots, c_r\}$ that are inside a GM FEM element and fulfill the Venant constraint is generated. At minimum, one should take all interior FEM nodes of the GM compartment for this purpose.

2. A triangulated surface Υ of the interface between GM and WM is extracted. The normal vectors $\Gamma = \{\nu_1, \ldots, \nu_h\}$ in the surface vertices $S = \{s_1, \ldots, s_h\}$ are computed.

3. A grid size is chosen and a regular 3D voxel grid is constructed to cover the GM compartment.

4. All voxels $\{V_1, \ldots, V_l\}$ containing surface vertices are determined. Of all surface vertices $S_j = \{s_{j_1}, \ldots, s_{j_k}\}$ inside a given voxel V_j, the one closest to the center of the voxel is computed and added to a set $\tilde{S} \subset S$. This procedure aims at constructing a regular sampling of Υ. However, as $\tilde{S}$ is build from nodes of the GM-WM interface, these locations are not valid for the forward computation.

5. Optimally, one would shift each $\tilde{s}_i \in \tilde{S}$ from the GM-WM interface in normal direction to a position right in the middle of the GM. However, the curvature of Υ is very strong in the sulci of the cortex. This can result in shifting a lot of $\tilde{s}_i$

towards a common center, resulting in a clustering of source locations. In addition, there might be no locations fulfilling the Venant constraint on the direct beam from $\tilde{s}_i$ in the direction of ν_i. The pragmatic solution we chose is to project every $\tilde{s}_i$ to the set of valid locations C by the use of a local metric: We split $d_{ij} = \tilde{s}_i - c_j$ into parts that are locally normal and tangential: $d_{ij} = d_{ij}^{\nu_i,\perp} + d_{i,j}^{\nu_i,\parallel}$ and defined

$$dist_\theta(i,j) = \sqrt{(1-\theta)^2 \left\| d_{i,j}^{\nu_i,\parallel} \right\|_2^2 + \theta^2 \left\| d_{i,j}^{\nu_i,\perp} \right\|_2^2}. \tag{5.6}$$

Then, we chose $\tilde{c}_i$ as the element c_j in C, which minimizes $\mathrm{dist}_\theta(i,j)$ for $\theta = 2/3$ and defined $\tilde{C} = \{\tilde{c}_1, \ldots, \tilde{c}_N\}$ as the source locations.

6. Steps 3.-5. can be iterated to find the grid size g, which leads to a source space with the desired number of sources N. As in the volume-based construction, a robust fitting heuristic is required for this task.

7. The normal orientation for each $\tilde{c}_i$ is defined as a specific local average of Γ: Let σ_i be half the distance of $\tilde{c}_i$ to the nearest $\tilde{c}_{j\neq i}$. Then, $\tilde{\nu}_i$ is computed as

$$\tilde{\nu}_i = \sum_j^t \nu_j \exp\left\{ -\frac{1}{2\sigma_i^2} \|\tilde{c}_i - s_j\|_2^2 \right\}, \tag{5.7}$$

and re-normalized.

Using the normal constraint, $\{\tilde{c}_i, \tilde{\nu}_i\}$, $i = 1, \ldots, N$ describes the source space, thus, $n = N$. If a full or weighted vector reconstruction should be performed, two dipoles spanning the tangential plane at each location are added and $n = 3N$.

Forward Computation

For the forward computations in all of our studies, we used the Venant direct method (cf. Section 2.4.2) with piecewise linear basis functions. For sufficiently regular meshes, recent studies (Lew et al. 2009, Vorwerk 2011, Vorwerk et al. 2012) show that this approach yields suitable accuracy over all realistic source locations. Furthermore, this approach has a high computational efficiency when used in combination with the FE transfer matrix approach (Wolters et al. 2004). The computations were performed with SimBio (cf. Section A.8).

5.4.4. Hierarchical Bayesian Inversion Studies for EEG, MEG and EMEG

Background

In Lucka (2011), Lucka et al. (2012), we compared EEG source reconstruction using an ℓ_2 hypermodel ($D = I_n$) with an inverse gamma hyperprior to established, ℓ_2 prior based inverse methods. For multiple focal (i.e., sparse) source scenarios, fully-Bayesian inference with the HBM prior improved upon the ℓ_2 based methods in many aspects. In particular, the results showed good localization properties for single dipoles and did not suffer from systematic depth mislocalization: Compared to source spaces based on smoothed and flattened surfaces, source spaces based on an accurate cortical surface segmentation or volume-based source spaces contain many more deep-lying locations. In this case, a phenomenon called *depth bias* is of fundamental importance for the correct localization of source activity. Many inverse methods fail to reconstruct deep-lying sources at the correct depth; rather, the sources are reconstructed too close to the skull. This is a well-known systematic error (see the references in Lucka et al. 2012), which can be crucial in clinical applications. One example is given by the pre-surgical functional mapping of the eloquent cortex (Schiffbauer et al. 2002). Our studies showed that the HBM-based reconstruction results did not suffer from this error.
Another effect related to the depth bias is the *masking* of deep-lying sources by superficial sources: If the true source configuration consists of multiple, spatially separated sources with different depths, many inverse methods only recover the sources close to the skull (see, e.g., Wagner et al. 2004). This effect complicates the analysis of networks of interacting brain areas which is a recent topic of interest in brain imaging (Kiebel et al. 2009). Furthermore, several clinical applications require a correct detection and separation of multiple sources, for instance, the reconstruction of the auditory pathway (Parkkonen et al. 2009) or specific cases of epileptiform discharges (Hufnagel et al. 1994, Janszky et al. 2000). In contrast to the established inverse methods, HBM-based methods were less likely to miss single sources in multiple source scenarios and were more often able to reconstruct the correct number of sources.

Motivation

In this study, we addressed two points that were left for future work in the outlook of Lucka et al. (2012):

- In the article, only EEG was investigated. Here, we investigate whether the findings also apply for MEG and how they compare to EEG/MEG combination, i.e., EMEG. The question of whether EEG or MEG is better suited for source reconstruction in general or for a specific experiment is an old but still ongoing

discussion in the neuroimaging field of research. As many different aspects need to be considered, no general answer was given yet (and may not exist). A common objection against using EEG and MEG recordings for a combined, multi-modal reconstruction by EMEG is that the combined modality might rather reflect the deficits of the single modalities and not their strengths. Aydin et al. (2014) give a recent overview on several of these aspects. Concerning the inverse problem, the difference between EEG and MEG source reconstruction has mainly been examined by ℓ_2-based inverse methods up to date (see, e.g., Molins et al. 2008).

- To facilitate the interpretation of our results, Lucka et al. (2012) used a head model with a homogeneous intracranial compartment. Especially for EEG/MEG combination, the use of a realistic, individual, anisotropic and calibrated (see Section 5.4.5) head model is mandatory (see Aydin et al. 2014, for a recent overview).

Study Design

For this study, we used the "simEMEG" scenario described in Section 2.4.2, i.e., the most realistic head model and artificial EEG/MEG sensor configurations that enable a fair comparison between both modalities. A volume-based source space with a grid spacing of 6 mm was used which leads to 1336 source locations and $n = 4008$ unknowns to recover. For $u^{\dagger,\infty}$, source configurations consisting of one, two or three dipoles with 100nAm amplitude were used. This current amplitude corresponds to the expected source strength in the evoked potentials/fields scenario. The locations of the dipoles were randomly chosen in the GM compartment, with the restriction that they fulfill the Venant constraint. The orientations were chosen randomly. Simulated measurements were computed and i.i.d. Gaussian noise with a *signal-to-noise ratio* (*SNR*) of 20 was added to all measurements: $\sigma = \|f\|_2/(m \cdot 20)$.
For combined EMEG, $f^{EEG/MEG}$ and $\sigma^{EEG/MEG}$ were computed for the single modalities, first. Then, they were stacked to construct the combined inversion model. The combined lead-field A^{EMEG} was constructed in the same fashion from A^{EEG} and A^{MEG}. Note that the different units and scales for EEG and MEG vanish from the inverse model once the pre-whitening (4.2) is performed: For the i.i.d. noise model, both $f^{EEG/MEG}$ and $A^{EEG/MEG}$ are divided by $\sigma^{EEG/MEG}$. The resulting variables do not carry a physical unit anymore but correspond to statistical significance measured in multiples of the standard deviation of a normal distribution. We considered the following inverse methods:

- Fully-Bayesian point estimates for an inverse gamma hyperprior with $\alpha = 0.5$, $\beta = 10^{-4}$:

  - *HBM-CM*: Full-CM estimate computed with the blocked Gibbs sampler (cf. Section 4.1.11). The burn-in length, K_0, was 1000 for single dipole recovery and 5000 for multiple dipole recovery. The size of the real run, K, was 50 000 for single dipole recovery and 200 000 for multiple dipole recovery.
  - *HBM-NM*: Full-NM estimate (cf. Section 4.2.5) initialized with the full-CM estimate.
  - *HBM-MAP*: Full-MAP estimates computed with the multiple-seed heuristic explained in Section 4.2.5. For computing the full-CM estimate seeds, we used $K_0 = 25$, $K = 200$. For the single dipole recovery, 128 of such seeds were tested; for the multiple dipoles scenarios, 256 were used.

- MAP estimates for ℓ_2 priors with different diagonal weightings $D \in \mathbb{R}^{n\times n}$ (*weighted minimum-norm estimates*, see Fuchs et al. 1998a, Lucka et al. 2012, for an overview and further references):
  - *MNE*: $D = I_n$.
  - *WMNE-ℓ_2*: $D = \mathrm{diag}_i\left(\|A_i\|_2\right)$.
  - *WMNE-ℓ_2^{reg}*: $D = \mathrm{diag}\left(\frac{\chi_i^2+\beta^2}{\chi_i}\right)$, with $\chi_i = \|A_i\|_2$, $\beta = \max_i\{\chi_i\}\,\frac{m\sigma^2}{\|f\|_2^2}$.
  - *WMNE-ℓ_∞^{reg}*: $D = \mathrm{diag}\left(\frac{\chi_i^2+\beta^2}{\chi_i}\right)$, with $\chi_i = \|A_i\|_\infty$, $\beta = \max_i\{\chi_i\}\,\frac{m\sigma^2}{\|f\|_2^2}$.

  Each D was normalized such that $\det(D) = 1$. The prior parameter λ was chosen by the *discrepancy principle* (Kaipio and Somersalo 2005). Several other weighting strategies were examined as well but the results are omitted here.

- *sLORETA* (Pascual-Marqui 2002): This inverse method consits of using an ℓ_2 prior with $D = I_n$ and standardizing the MAP estimate for the current vector at a given location, $u_{[i]}$, by the posterior covariance:

$$F = u_{[i]}^T\left(\mathbb{C}\mathrm{ov}\left[u[i]|f\right]\right)^{-1}u_{[i]}, \quad \mathbb{C}\mathrm{ov}\left[u[i]|f\right] = \left(A^T\left(AA^T + \lambda I_m\right)^{-1}A\right)_{[i,i]} \tag{5.8}$$

  This yields a pseudo statistic of F-type for every source space node.

We will refer to sLORETA and the WMNEs as "the established methods". As in Lucka et al. (2012), we validated inverse methods by computing the statistics of *dipole localization error* (*DLE*) and *earth mover's distance* (*EMD*) between reference and reconstructed activity. Both measures are explained in the appendix in Section A.7.

Results and Discussion

EEG vs. MEG Table 5.8a lists the results for the single dipole recovery and Table 5.8b for the recovery of two or three dipoles. The DLE results for WMNE-ℓ_2 and WMNE-ℓ_2^{reg}

Table 5.8.: Results of the simulated EEG, MEG and EMEG inversion studies.

(a) Statistics of DLE and EMD (displayed as "mean ± std") of different inverse methods for the recovery of 1000 source configurations consisting of one dipole.

Method	DLE			EMD		
	EEG	MEG	EMEG	EEG	MEG	EMEG
MNE	21.3 ± 11.7	20.0 ± 13.2	17.8 ± 10.1	59.5 ± 5.1	60.1 ± 5.5	55.2 ± 5.3
WMNE-ℓ_2	26.5 ± 16.3	45.3 ± 21.6	21.2 ± 14.5	59.6 ± 3.9	58.3 ± 5.8	55.0 ± 5.1
WMNE-ℓ_2^{reg}	23.2 ± 12.5	18.8 ± 11.3	15.7 ± 8.6	59.6 ± 3.9	59.0 ± 4.7	55.0 ± 4.9
WMNE-ℓ_∞^{reg}	17.6 ± 9.1	13.6 ± 8.2	11.3 ± 6.1	59.1 ± 4.4	59.0 ± 4.7	54.3 ± 5.2
sLORETA	7.2 ± 3.9	6.3 ± 3.6	5.0 ± 2.4	44.6 ± 4.7	48.8 ± 5.3	34.8 ± 4.9
HBM-CM	10.8 ± 5.8	10.6 ± 6.7	8.6 ± 4.4	12.9 ± 5.8	12.9 ± 6.2	11.0 ± 4.6
HBM-NM	9.9 ± 5.2	9.9 ± 6.7	7.8 ± 4.2	10.2 ± 5.6	10.5 ± 6.5	9.0 ± 4.6
HBM-MAP	7.4 ± 5.0	8.1 ± 7.3	6.4 ± 4.4	7.8 ± 5.8	8.5 ± 7.1	7.8 ± 5.6

(b) Statistics of EMD (displayed as "mean ± std") of different inverse methods for the recovery of k source configurations consisting of l dipoles.

Method	$l = 2$, $k = 500$			$l = 3$, $k = 100$		
	EEG	MEG	EMEG	EEG	MEG	EMEG
MNE	49.2 ± 3.6	49.9 ± 4.3	45.4 ± 3.7	44.1 ± 3.6	45.0 ± 4.0	40.9 ± 3.1
WMNE-ℓ_2	49.5 ± 3.2	49.0 ± 4.4	45.8 ± 3.4	44.1 ± 3.2	43.9 ± 4.2	40.9 ± 3.0
WMNE-ℓ_2^{reg}	49.4 ± 3.2	49.2 ± 3.9	45.7 ± 3.3	44.1 ± 3.2	44.0 ± 3.8	40.9 ± 3.0
WMNE-ℓ_∞^{reg}	49.0 ± 3.4	49.2 ± 3.9	45.1 ± 3.4	43.8 ± 3.3	44.1 ± 3.7	40.5 ± 3.0
sLORETA	41.1 ± 4.0	44.7 ± 4.8	35.3 ± 4.5	39.0 ± 4.6	41.5 ± 5.4	34.8 ± 5.3
HBM-CM	19.5 ± 7.3	22.8 ± 10.3	13.3 ± 5.5	27.1 ± 8.3	28.4 ± 9.1	18.2 ± 6.6
HBM-NM	19.3 ± 9.0	22.2 ± 11.6	12.2 ± 6.2	28.7 ± 9.3	29.1 ± 10.2	17.9 ± 7.3
HBM-MAP	21.4 ± 8.8	27.4 ± 13.3	16.3 ± 6.9	30.7 ± 7.3	35.3 ± 11.1	22.9 ± 7.2

(c) Fraction of 1000 single dipoles that are reconstructed too deep.

Method	EEG	MEG	EMEG	Method	EEG	MEG	EMEG
MNE	0.300	0.182	0.240	sLORETA	0.522	0.524	0.539
WMNE-ℓ_2	0.467	0.616	0.526	HBM-CM	0.518	0.529	0.533
WMNE-ℓ_2^{reg}	0.406	0.515	0.455	HBM-NM	0.527	0.516	0.529
WMNE-ℓ_∞^{reg}	0.336	0.410	0.459	HBM-MAP	0.532	0.526	0.525

demonstrate that introducing the regularization of the weighting by Fuchs et al. (1998a) is particularly important for MEG. Without the WMNE-ℓ_2 results, the results of the established methods suggest that MEG has better localization properties compared to EEG. However, the HBM results do not confirm this but rather suggest that EEG and MEG have very similar localization properties for single dipoles. The EMD results confirm that the established methods are not able to reconstruct focal sources. In the multiple dipole scenarios, the HBM-based methods show better reconstruction performance for EEG than for MEG.
The comparison to similar studies on this topic (see, e.g., Babiloni et al. 2004, Baillet et al. 1999, Fuchs et al. 1998b, Liu et al. 2002, Molins et al. 2008) is difficult: Other studies often used realistic sensor setups, which means that the number of MEG sensors is considerably higher than the number of EEG sensors (yet, conclusions about the *general* properties of the single modalities were made). In addition, other studies often committed the inverse crime of placing the reference sources on the computational grid, used other error metrics and inverse methods or analyzed single source scenarios, only. For these reasons, there is no consensus to which we can compare our results.

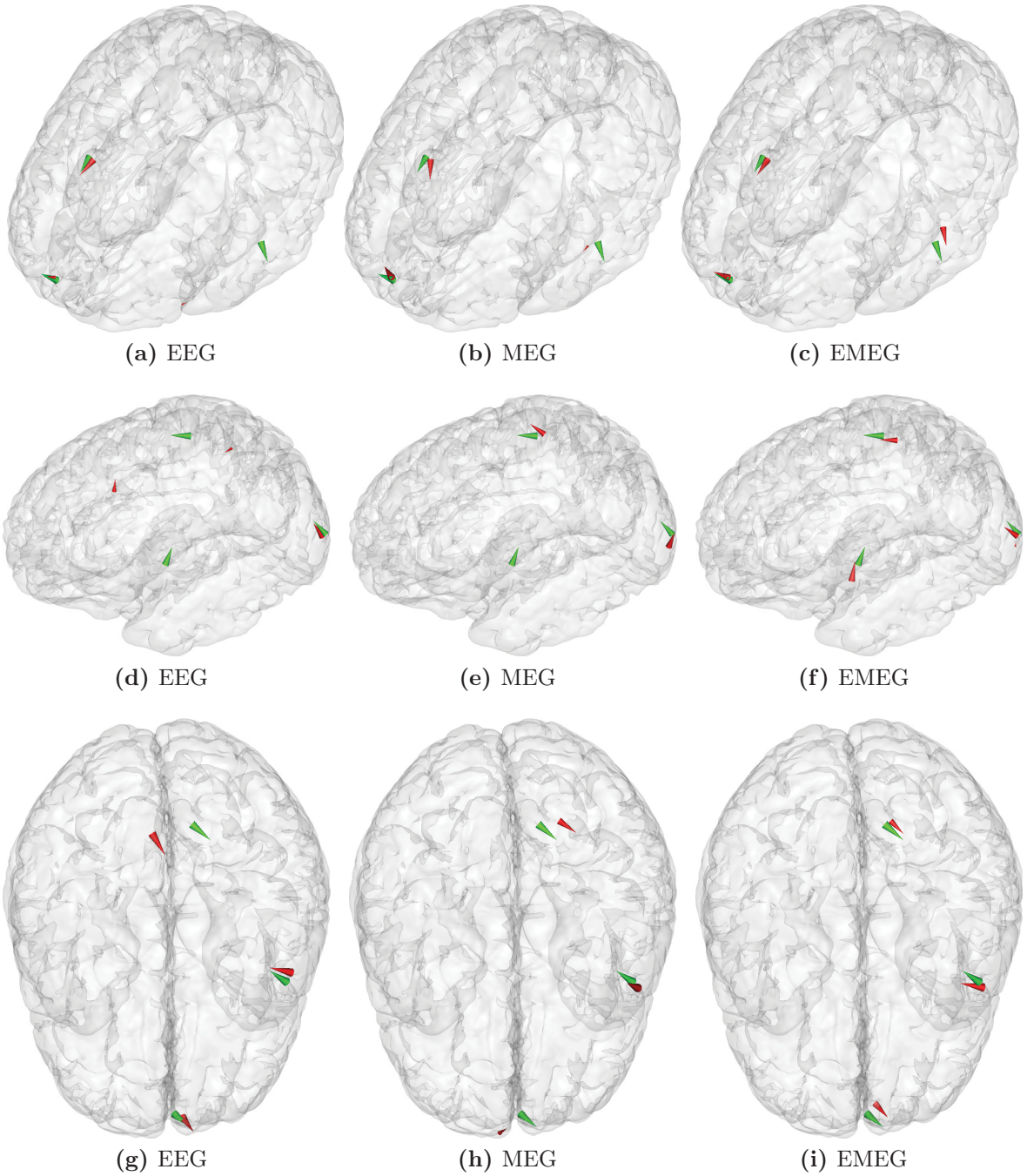
EEG/MEG Combination Tables 5.8a and 5.8b show that the combination of EEG and MEG improves the average performance of all methods. In particular, the improvement of the HBM-based methods in the multiple dipole scenarios is noticeable. A closer look at the statistics of the EMD for these scenarios shows that especially extremely erroneous reconstructions can be avoided by EEG/MEG combination. Figure 5.29 discusses the HBM-NM results for three interesting source scenarios. The results for single dipoles are in line with former studies investigating the combination for other inverse methods (see, e.g., the references given in the previous paragraph). However, as those studies most often did not investigate multiple-dipole scenarios due to the lack of a suitable error measure like the EMD, our results stress the particular value of EEG/MEG combination for multiple-source configurations.

Depth-Bias For examining the depth bias, we defined the depth of a point in the realistic head model as its minimal distance to the superior skin surface nodes (all nodes whose z-coordinate is larger than the minimal z-coordinate of the sensor positions). Then, we compared the depth of the single dipole sources with the depth of the source space node with the largest reconstructed amplitude. Table 5.8c shows the fraction of dipole sources that were reconstructed too deep. A value close to 0.5 indicates that the inverse method does not suffer from systematic depth mislocalization. The results show that especially the MNE fails to reconstruct sources in the correct depth. The different weightings introduced in the WMNE schemes aim to compensate for this. However, our

results show that a single weighting cannot prevent from depth-bias in all modalities. Furthermore, a comparison with Table 5.8a shows that while the depth localization may profit from the weighting compared to the MNE, the overall localization does not necessarily improve accordingly. The HBM-based methods and sLORETA did not show a depth bias in any modality.

Conclusions

The differences between established and HBM methods for the comparison between EEG and MEG show that it is difficult to make general statements about the localization properties of a single modality. The ill-posed nature of the EEG/MEG inverse problem prohibits a separation of the properties of the modality from the properties of the inverse method used. Meaningful statements are only possible for their combination. Our results show that EMEG seems to rather combine the strengths and not the weaknesses of both modalities: The source reconstruction results are stabilized and improved to a considerable amount for all inverse methods. Again, a more detailed comparison shows that especially the HBM methods profit from the combination, in particular for source separation in multiple source scenarios. This further underlines the potential of these methods for complex source scenarios in real applications.



(a) EEG (b) MEG (c) EMEG

(d) EEG (e) MEG (f) EMEG

(g) EEG (h) MEG (i) EMEG

Figure 5.29.: HBM-NM reconstructions (red cones) of three different source scenarios with three dipolar sources (green cones). Note that the real sources lie in-between the source space nodes. Therefore, a perfect recovery is not possible. (a)-(c) EEG failed to reconstruct all three dipoles satisfactorily. MEG was only able to reconstruct the tangential part of the rightmost dipole. EMEG was able to reconstruct location, amplitude and orientation of all sources satisfactorily. (d)-(f) Among other deficits, both EEG and MEG failed to reconstruct the bottom left source. Again, the EMEG result is much more convincing. (g)-(i) The EEG result was already quite good while the MEG had problems to reconstruct the radial, frontal source. EMEG was not disturbed by the weakness of MEG but was able to produce an even slightly better result than EEG alone.

5.4.5. Auditory and Somatosensory Evoked Potentials and Fields

Motivation: From Simulated to Real Data

The results of the simulation studies clearly demonstrated the potential of fully-Bayesian inference for HBM for the reconstruction of focal source configurations. Therefore, one main goal of this thesis was to process real, experimental data. As introduced in Section 2.4.2, we will use auditory and somatosensory evoked potentials and fields. The early components of these responses are well-studied and are commonly assumed to correspond to rather focal brain activations. In contrast to our expectations, the first results were quite unsatisfactory: Especially the full-MAP estimates were very unstable with regard to parameter choices and too dependent on the randomness in the initialization by the CM estimate. The CM estimate seemed more robust, but the results were also less good than expected. These first results (which we will not present in detail here) motivated the development and implementation of a new processing pipeline and a more careful examination of its different steps; in particular, whether the non-linear, non-convex optimization employed for computing the full-MAP estimate is sensitive to the uncertainties these steps involve:

- While some obvious inverse crimes (cf. Section 2.1) were avoided in the simulation studies, the uncertainty concerning the forward model is certainly larger for real data: Even sophisticated head models rely on simplifying assumptions, are prone to MRI errors and artifacts as well as registration and segmentation errors and use standard values for the conductivities of many compartments. In addition, the sensor positions also carry uncertainty (see below). HBM inference may be too sensitive to the resulting error of A.

- The raw SEP/SEF data is a complex spatio-temporal mixture of various signals of which most are actually unrelated to the stimulation paradigm (cf. Section 3.1). As explained in Section 2.4.2, trial-averaging is used to reduce the impact of unrelated signals on the evoked response. However, averaging is a Monte Carlo technique (cf. Section 4.1.1). Therefore, even with very optimistic assumptions on the nuisance signals, averaging can reduce the error only with the rate $\mathcal{O}(1/\sqrt{K})$. As a consequence, various additional signal processing techniques are used to improve the SNR. In the end, the noise model has to account for the nuisance signals remaining after all the different preprocessing steps. As a result, the Gaussian model we use might not be a valid approximation or its estimation may be too inaccurate. Again, HBM inference may be too sensitive to this error of the likelihood distribution.

- As EEG/MEG is severely ill-conditioned and under-determined, all inverse solutions are necessarily *prior-dominated*, i.e., they reflect more information from the prior than from the likelihood. Therefore, inverse results are very sensitive to a miss-specification of the prior. While the hierarchical prior model we use is strongly based on the assumption of sparsity/focality, the stimulation-unrelated brain activity is still present in the trial-averaged signal as a distributed, low-amplitude source. Thereby, our prior model is not fully correct.

To account for these issues, we will conduct sensitivity studies at certain points of the development of the processing pipeline. Finally, we will present the source reconstructions obtained with the new processing pipeline and revisit topics that we previously examined in simulation studies.
General guidelines for data acquisition and preprocessing are given by Picton et al. (2000) for EEG and by Gross et al. (2013) for MEG. All technical details and terms not explained in the following sections can be found in these articles and the references therein.

Data Acquisition

All recordings were performed in a magnetically shielded room and in supine position. In total, the whole head MEG system (CTF, VSM MedTech Ltd.) comprises 595 SQUID coils of which 548 are arranged to form physical gradiometers while the rest are commonly used to define reference channels. The combination of the coils into different measurement channel layouts can be described by a $m \times 595$ matrix applied to the full data. The position of the subject's head inside the dewar (cf. Figure 2.9) is continuously tracked by recording the locations of three small head localization coils placed at the anatomical *fiducials* nasion, left and right ear canals. The same fiducials are also used in the MRI acquisition. This can be used to register the MEG sensors to the head model in the later analysis. The technical details of this procedure are omitted here. Simultaneous EEG recordings were performed using 74 electrodes arranged in the *10-20 system* using the 10% division (*10-10 system*). The electrode positions are measured and digitized using a Polhemus device. In addition, EOG and ECG channels were recorded but are not used in this work. All data was acquired at a sampling rate of 1200Hz and filtered online with a 300Hz low pass filter. Section A.8 lists the software used for the different processing tasks.

Empty Room Recordings Besides the evoked responses, we recorded 5 minutes of *empty room recordings* of MEG only. These recordings can be regarded as samples of the measurement noise and will be used in the noise modeling studies. Two gradiometer

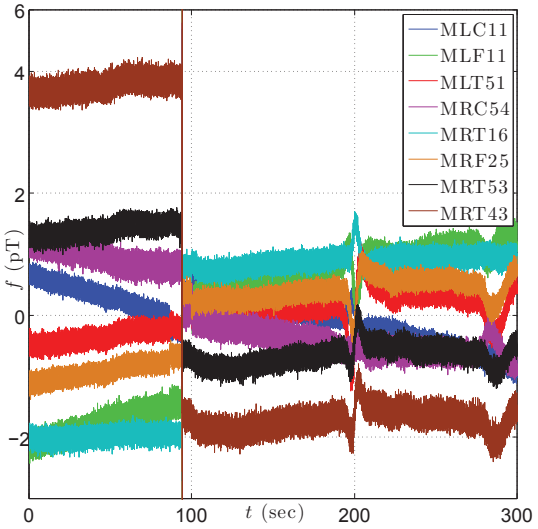


Figure 5.30.: Time courses of a selection of MEG channels before pre-processing (the mean was removed to fit all courses into one plot).

channels were broken and were not recorded.

Somatosensory Stimuli Somatosensory responses can be evoked by tactile or electrical stimulation of the median nerve of the subject. The data processed here stems from electrical stimulation: Electrical square pulses with 0.5ms duration were applied to either the left or the right wrist. The stimulation side was varied randomly between left and right and the *inter-stimulus interval* (*ISI*) was varied randomly in-between 350ms to 450ms. One reason for applying this procedure is to avoid habituation. Another reason is that stimulation-locked responses from earlier stimulations that might fall into a following pre-stimulus interval should average out. The electrical stimulation current creates an artifact, which declines fast. To further reduce its influence, the polarity of the stimulation is switched in the second half of the recording session. In total, 978 left-hand trials and 972 right-hand trials were recorded. In this thesis, we will only examine the left-hand trials.

Auditory Stimuli Prior to the measurement, a hearing test is conducted to determine the hearing threshold for a pure sinusoidal tone with 350 Hz frequency. Then, this tone is presented bilaterally, 55dB above this threshold, in 125 trials. The ISI was varied randomly between 3.5s and 4.5s.

Signal Processing

For the first reconstructions of real data, we used rather standard, conservative pre-processing steps. There are two main strategies to clear the data from artifacts and

other unwanted signal components: The first aims at identifying contaminated trials to fully exclude them from the further analysis. The second strategy tries to clear the trial from the nuisance signal by applying signal processing techniques. We will use a combination of both approaches.

Empty Room Recordings The MEG data was processed in the synthetic first order gradiometer channel layout. This means that the reference channels are used in a specific way to correct for the finite distance between the two coils that constitute the main physical gradiometer channels. Figure 5.30 shows the time courses of several single channels. The mean of each channel (*baseline*) was subtracted to fit them into one plot. One can clearly see that all channels jump shortly before $t = 100$s. This is an example of the well-known *SQUID jump artifacts*. Around $t = 200$s and $t = 290$s, two further artifacts with a smooth temporal evolution are visible that most probably correspond to a field caused by an event outside the recording chamber. We can further see that the baseline of the time courses seems to change over time. This *drift artifact* can be modeled by a linear regression.
To clear the data from these artifacts, we first chopped the continuous data into 60 trials of 5s length. Based on a visual inspection the trials corresponding to $85\text{s} < t \leqslant 105\text{s}$ (SQUID jump), $195\text{s} \leqslant t < 205\text{s}$ and $280\text{s} \leqslant t < 295\text{s}$ (external artifacts) were removed. Then, a specific DTF-based line noise removal procedure was applied and the baseline of each trial was corrected for constant and linear trends.

Somatosensory Stimuli The EEG channels FC1 and F1 were very noisy and showed shifts on long temporal scales, suggesting that the electrodes were not attached properly and the electric contact between electrode and skin was varying. Therefore, they are excluded from the further analysis. The electric potential at one electrode can only be measured with respect to a reference electrode. Therefore, EEG recordings have one degree of freedom that can be defined by the user. We transform the EEG channels to *common average reference*: The sum of all channels is always zero. The MEG data was processed in the synthetic third order gradiometer channel layout. In this scheme, the reference channels are also used to reduce the contribution of far away but strong magnetic fields to the measurements. The data was chopped into trials of 200ms length before and after each stimulus onset. Within each trial, the stimulus onset corresponds to $t = 0$ms. First, the baseline of the MEG channels of each trial was reset to zero based on the mean values in the pre-stimulus interval, i.e., $-200\text{ms} \leqslant t \leqslant 0\text{ms}$. As a second step, we applied band-pass filtering from 1Hz to 30Hz using a zero-phase forward and reverse Butterworth filter of order four. Thereby, we may also remove stimulus-related, high-frequency signals reflecting short-lived brain activity, but for the localization of the

N20 response, signals from outside this frequency band are mainly noise contributions. To avoid filtering artifacts resulting from the short trial length, each trial was padded by the 2s of data before and after the trial before the filtering was applied. As a final step, we exclude those trials from averaging that are severely contaminated by artifacts: In general, trial rejection is a difficult task as one has to decide whether a trial can improve the quality of the unknown signal or not. This choice also relies on the total number of trials available which is rather large in our case. Therefore, we used a simple, conservative, automatic trial rejection procedure: The variance of each channel for each trial is computed and the 20% trials with the largest maximal channel variance are rejected, leaving 782 of 978 trials for averaging. Figure 5.31 shows butterfly and topography plots. Note that unfortunately, the strong band-pass filtering we use in this first examination leads to a shift in the timings of the physiological components, which therefore have to be interpreted with care. We will reconstruct the peak of the N20(m) signal component found around $t = 15$ms. In un-filtered data, this component is found around $t = 23$ms after stimulus onset (see BUCHNER et al. 1995, 1994, FUCHS et al. 1998b, and references therein).

Auditory Stimuli We only report the differences to the SSEP/SSEF data here: In addition to the EEG channels FC1 and F1, also the MEG channels MLO32 and MLC32 were excluded from the further analysis. A trial length of 1s before and after each stimulus onset is used. The variance of each MEG channel for each trial is computed and visualized. Based on the maximal channel variance per trial, 36 of the 125 trials are rejected by visual inspection. For EEG, the trials rejected for MEG were rejected as well. Using a visual inspection of the remaining trials, only one further trial was rejected. We will reconstruct the signal at 92ms at the rising flank of the N100(m) component. Figures 2.12 and 5.31 show butterfly and topography plots.

Noise Modeling

Similar to CT, we want to use an additive Gaussian noise model, $\varepsilon \sim \mathcal{N}(\mu_\varepsilon, \Sigma_\varepsilon)$. The empty room recordings were carried out to assess the statistics of the noise contributions that are not due to the measurement of a subject. Figure 5.32a shows a scatter plot of the two MEG channels with the largest covariance after pre-processing and Figure 5.32b a histogram of one of them compared to a Gaussian fit. By visual inspection, using a Gaussian model seems to be very adequate for the noise contributions present in empty room recordings. Due to the baseline correction, $\mu_\varepsilon = 0$. Figures 5.32c and 5.32d visualize Σ_ε estimated from the whole empty room recording data.
The pre-stimulus intervals of the real recordings only contain non-stimulus related signals. In contrast to the empty room recordings, this includes also nuisance signals produced

by the subject such as heart beat signals and un-evoked brain activity. Figures 5.32e - 5.32l visualize the covariance (and correlation) matrices of the pre-stimulus data (before averaging). Compared to the empty room recordings, inter-channel correlations are more pronounced and the covariance structure is more degenerate: The condition number of the empty room covariance matrix is about 50 while it is about 6839 for the pre-stimulus matrix. The correlation pattern seems to reflect the spatial arrangement and grouping of the sensors. This suggests that spatially smooth nuisance fields dominate the signal in the pre-stimulus interval, most likely caused by other forms of brain activity.

Sensitivity Study In our first attempt to produce HBM results for experimental MEG data, we used the diagonal part of the pre-stimulus covariance matrix *after* averaging. We will now examine whether HBM is sensitive to such noise simplifications by a simulation study. For this, we use the same single dipole recovery scenario as in Section 5.4.4 (using MEG only) but with four different noise models. We will again use an SNR of 20, which is a typical value for evoked potentials, and therefore define $\sigma = \|f\|_2/(m \cdot 20)$. Then, the empirical covariance matrix Σ_{emp} (empty room or pre-stimulus data) is re-scaled to define the *full noise model*:

$$\Sigma_f = \frac{\sigma^2 m}{\mathrm{tr}\,(\Sigma_{emp})}\Sigma_{emp} \tag{5.9}$$

For generating the *permuted noise model*, we take the singular value decomposition of $\Sigma_f = VSV^T$, randomly permute the columns of V to obtain V_p, and define $\Sigma_p = V_p S V_p^T$. Obviously, Σ_p has the same algebraic properties as Σ_f, but its correlation pattern does not reflect the spatial correlation between the sensors anymore. The *diagonal noise model* is defined by $\Sigma_d = \mathrm{diag}\,(\Sigma_f)$ and the *i.i.d. noise model* by $\Sigma_i = \sigma^2 I_m$. By construction,

$$\mathrm{tr}\,(\Sigma_f) = \mathrm{tr}\,(\Sigma_p) = \mathrm{tr}\,(\Sigma_d) = \mathrm{tr}\,(\Sigma_i) = \sigma^2 m, \tag{5.10}$$

and thereby, the noise levels of the models are comparable. The sensitivity study examines what happens, if the noise model chosen for the reconstruction does not match the one used to generate the data. Table 5.9a lists the results if the empty room covariance is used as Σ_{emp}. We see that in general, there are only small differences in localization performance with the exception of using the permuted noise model for data generation and the full model for inversion with the established methods. For the HBM methods, all variations are small and consistent in the sense that using the same noise model for data generation and inversion should yield the best results. Table 5.9b lists the results if the pre-stimulus covariance is used as Σ_{emp}. Now all methods are severely disturbed if the permuted noise model is used for data generation and the full model

Table 5.9.: Mean DLE of different inverse methods for the recovery of 1000 source configurations consisting of one dipole when different noise models for data generation and inversion are used.

(a) The empirical covariance matrix Σ_{emp} is derived from the empty room recordings.

WMNE-ℓ_2^{reg}		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	13.38	13.41	13.43	13.61
	Σ_d		13.69	13.60	14.25
	Σ_f			14.45	32.72
	Σ_p				13.72

HBM-CM		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	5.49	5.68	5.88	5.58
	Σ_d		5.57	5.77	5.74
	Σ_f			5.66	5.89
	Σ_p				5.48

sLORETA		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	5.20	5.18	5.22	5.40
	Σ_d		5.18	5.18	5.93
	Σ_f			5.22	27.21
	Σ_p				5.12

HBM-NM		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	5.47	5.64	5.87	5.56
	Σ_d		5.59	5.76	5.71
	Σ_f			5.57	5.73
	Σ_p				5.52

(b) The empirical covariance matrix Σ_{emp} is derived from the pre-stimulus data.

WMNE-ℓ_2^{reg}		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	13.38	13.33	14.17	17.36
	Σ_d		13.56	14.12	22.15
	Σ_f			17.87	88.46
	Σ_p				36.20

HBM-CM		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	5.68	5.76	6.17	5.60
	Σ_d		5.84	6.29	5.53
	Σ_f			5.47	76.20
	Σ_p				5.01

sLORETA		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	5.20	5.18	5.45	8.80
	Σ_d		5.18	5.43	12.71
	Σ_f			5.00	82.23
	Σ_p				4.82

HBM-NM		data noise Σ_i	Σ_d	Σ_f	Σ_p
inv noise	Σ_i	5.73	5.70	6.30	5.53
	Σ_d		5.78	6.18	5.55
	Σ_f			5.64	72.98
	Σ_p				5.24

for inversion. Otherwise, the variations for HBM and sLORETA are, again, very small, while WMNE-ℓ_2^{reg} is affected more strongly.

Sensitivity to Background Activity

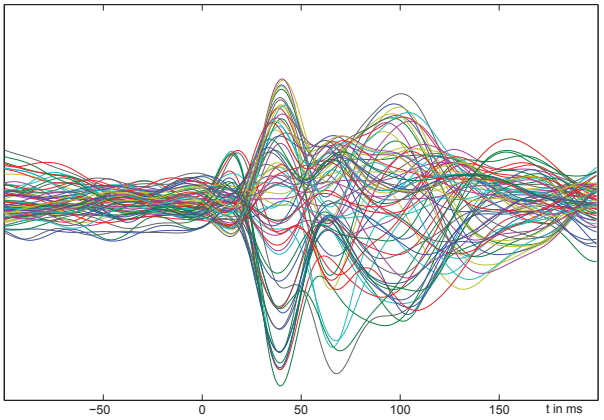
As described above, another potential disturbance for sparse inverse methods might be the presence of the signals originating from the averaged un-evoked brain activity u_{bkg}^{∞} in f:

$$f = \mathcal{A}(u^{\dagger,\infty} + u_{bkg}^{\infty}) + \varepsilon \tag{5.11}$$

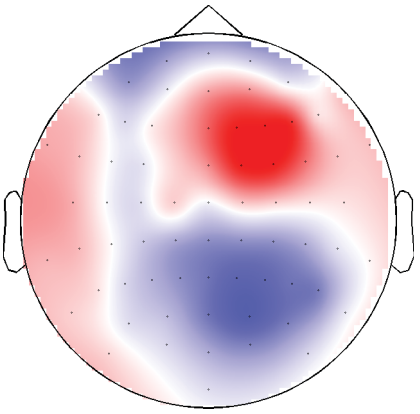
Table 5.10.: Mean DLE/EMD of different inverse methods for the recovery of 1000 source configurations consisting of one dipole when background activity is added to f.

Method	none	focal, 10%	focal, 20%	Gauss, 10%	Gauss, 20%
WMNE-ℓ_2^{reg}	13.38/55.66	13.39/56.29	13.67/57.82	13.39/56.43	13.96/58.27
sLORETA	5.20/41.67	5.29/42.20	5.57/43.76	5.26/42.26	5.59/43.93
HBM-CM	5.55/7.12	5.90/8.85	6.12/14.30	5.72/8.05	6.62/14.35
HBM-NM	5.58/5.81	5.81/6.19	6.23/9.68	5.69/5.95	6.80/7.35

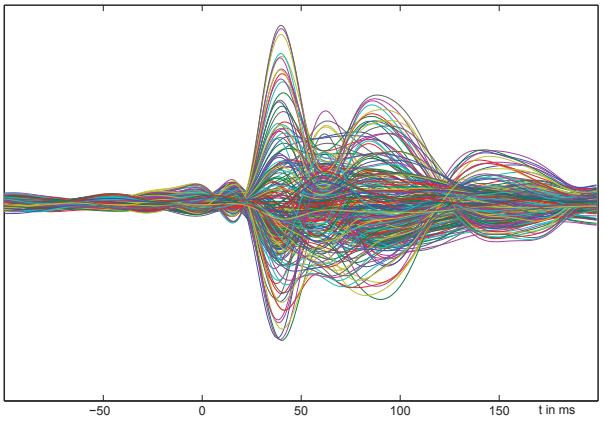
If necessary, one could discretize u_{bkg}^{∞}, introduce a prior model for it, and apply one of the techniques introduced in Section 3.6.2. However, we will first conduct another sensitivity study using the same single dipole recovery scenario as in Section 5.4.4 (with MEG) to test if it is necessary: We want to examine how the reconstruction results change from $u_{bkg}^{\infty} = 0$ if we assume that u_{bkg}^{∞} either consists of two focal sources different from the main one to recover, or it consists of a Gaussian random field at the source space nodes ($\mathcal{A}u_{bkg}^{\infty} = Az$, $z \sim \mathcal{N}(0, I_n)$). In both cases, we scale $\mathcal{A}u_{bkg}^{\infty}$ to either 10% or 20% of $\mathcal{A}u^{\dagger,\infty} + \varepsilon$ (measured in ℓ_∞-norm). Table 5.10 lists the results. We see that adding background activity only leads to a moderate impairment of the reconstruction properties of all methods.



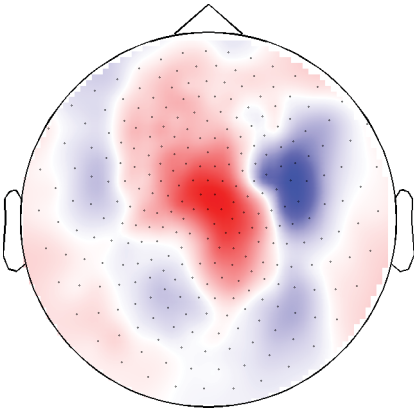
(a) SSEP (EEG) butterfly



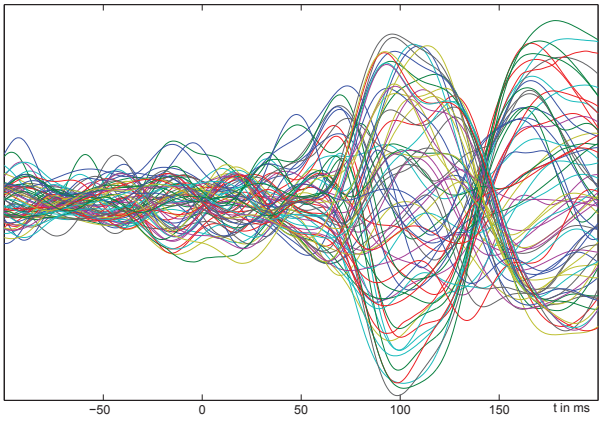
(b) SSEP (EEG) topography, 15ms



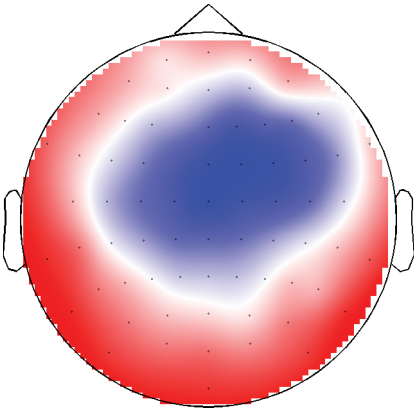
(c) SSEF (MEG) butterfly



(d) SSEF (MEG) topography, 15ms

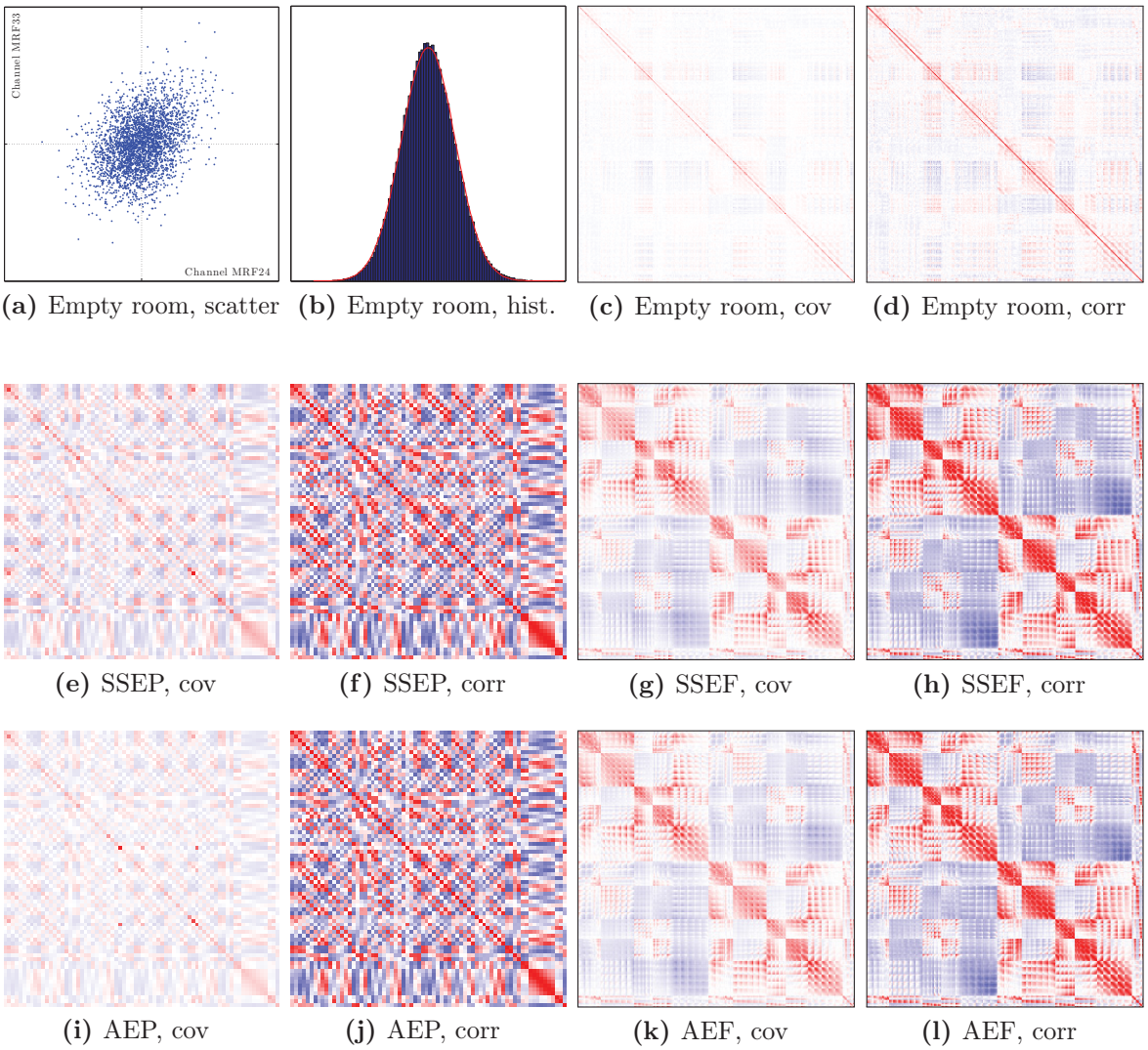


(e) AEP (EEG) butterfly



(f) AEP (EEG) topography, 92ms

Figure 5.31.: Butterfly and topography plots for SSEP, SSEF and AEP data (cf. Figure 2.12).

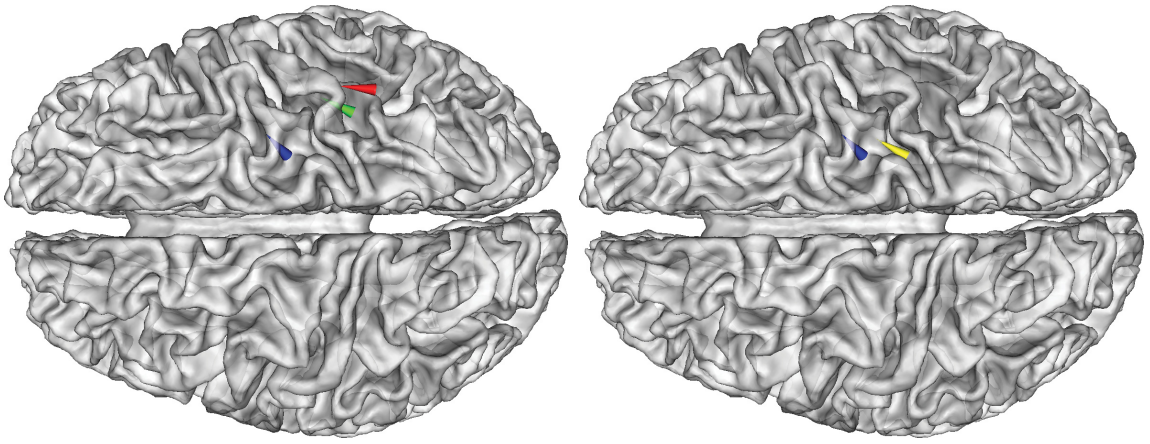


(a) Empty room, scatter **(b)** Empty room, hist. **(c)** Empty room, cov **(d)** Empty room, corr

(e) SSEP, cov **(f)** SSEP, corr **(g)** SSEF, cov **(h)** SSEF, corr

(i) AEP, cov **(j)** AEP, corr **(k)** AEF, cov **(l)** AEF, corr

Figure 5.32.: (a) Empty room data after pre-processing: Scatter plot of two (subsampled) MEG channels. (b) Normalized histogram of empty room channel MRF24 after pre-processing (blue bars) and normal approximation (red line). (c)-(l) Visualizations of covariance and correlation matrices.



(a) SDS **(b)** HBM-NM

Figure 5.33.: HBM and SDS results for the SSEP data, computed for a volume-based source space with a grid spacing of 6mm. (a) SDS results using the i.i.d. (red cone), diagonal (green cone) or full (blue cone) noise model. (b) HBM-NM results using the i.i.d. or diagonal noise model (yellow cone) or the full noise model (blue cone).

Source Reconstruction Results

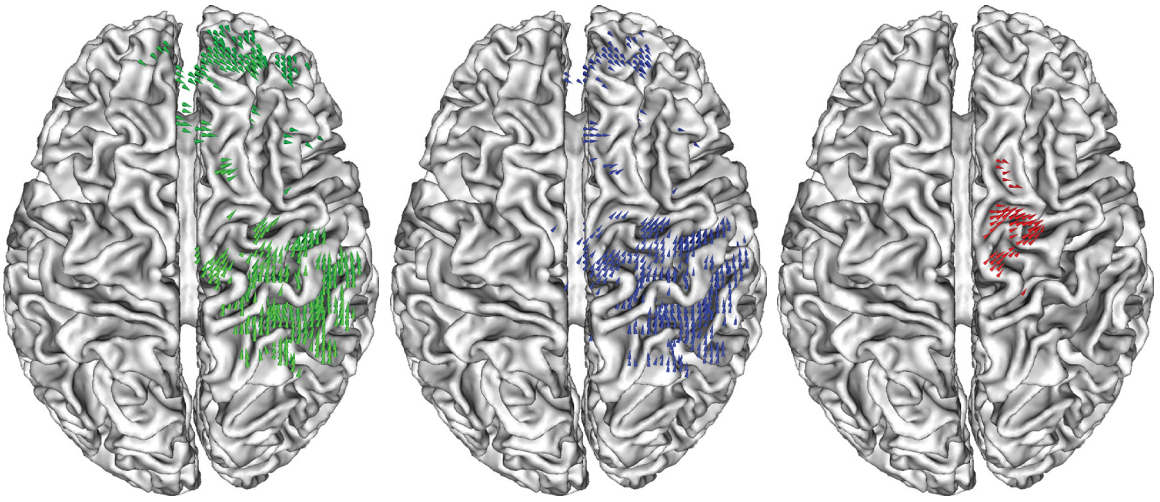
Somatosensory Data We first process the somatosensory data because the generator of the N20(m) component of left-hand stimulation is supposed to be a source configuration with a simple structure: It is commonly assumed to be well-approximated by a single equivalent current dipole in the right-hemispheric somatosensory 3b area, i.e., in a superficial location and with a quasi-tangential orientation (see Buchner et al. 1995, 1994, Fuchs et al. 1998b, Hämäläinen et al. 1993, and references therein). The EEG and MEG topographies (Figures 5.31b and 5.31d) support this assumption: They are both dominated by two opposing poles, which suggests a single source. The midpoints between both poles are roughly above the location in both modalities. Finally, the EEG poles are orthogonal to the MEG ones. Compare also Figure 2.11b, which shows artificial topographies for such a source scenario.

As we know that a single source is a reasonable source model, we will compare HBM results for SSEP data to a *single dipole scan* (*SDS*):

$$\hat{u}_{SDS} := \underset{u}{\operatorname{argmin}} \|f - Au\|^2_{\Sigma_\varepsilon^{-1}} \quad s.\,t. \quad |u|_{[0]} = 1, \tag{5.12}$$

Practically, we minimize $\left\|f - A_{[i]}v_i\right\|^2_{\Sigma_\varepsilon^{-1}}$, $v_i \in \mathbb{R}^3$, for each source space node i and choose the dipole v_i which achieves the best fit as the solution. Using this inverse method can be seen as the current "gold-standard" for this source reconstruction problem.

We use this scenario to revisit the noise modeling and sensitivity studies. Three noise models will be used to process the SSEP data: The *full noise model* will use the



(a) i.i.d. noise model **(b)** diagonal noise model **(c)** full noise model

Figure 5.34.: MNE for the SSEP data, thresholded at 75 % and computed for different noise models and a volume-based source space with a grid spacing of 2 mm.

pre-stimulus covariance matrix Σ_{pre} (see Figure 5.32e) but with a slight regularization:

$$\Sigma_{reg} = (1 - \delta)\Sigma_{pre} + \delta I_m, \tag{5.13}$$

where we choose the smallest δ such that the condition of Σ_{reg} is below 10^4. The *diagonal noise model* uses diag (Σ_{reg}), and the *i.i.d. noise model* uses $\bar{\sigma}^2 I_m$, where $\bar{\sigma}^2$ is the geometrical mean of the diagonal of Σ_{reg}.

Figure 5.33 shows the results, computed with the same volume-based source space (grid spacing: 6mm) used in Section 5.4.4. Both methods show the same result for the full covariance model which is assumed to be the most accurate one. However, note that HBM does not limit the number of active sources explicitly! We also see that HBM does not seem to be more sensitive to noise simplification than SDS, which is commonly regarded a very robust source reconstruction technique.

In Figure 5.34, we compare MNEs, computed for a volume-based source space with a grid spacing of 2mm. While the Gaussian prior model used by the MNE is not appropriate to describe the single, dipolar activity we expect, we see from the results that accurate noise modeling is also important for MNEs: The result for the full noise model is much more focused than the others and its support is in good correspondence with the location found by HBM and SDS.

While all three inverse methods show coherent results with respect to location and influence of noise modeling, the location found does not exactly correspond to the expected somatosensory 3b area but is slightly shifted towards frontal regions.

Auditory Data Compared to the somatosensory N20(m), the auditory N100(m) component is presumably generated by a more complicated source configuration: Commonly, two dipolar, bi-lateral sources near the planum temporale are assumed to be an accurate description of it (Pantev and Lutkenhoner 2000). An examination of the topographies (see Figures 5.31f and 2.12b) reveals why MEG is commonly considered a more reliable modality for auditory source reconstruction: Figure 2.12b shows two separated single dipole patterns, each consisting of two opposing poles, and one in each hemisphere. For EEG, the specific arrangement of the two sources leads to a topography, which does not give such a clear indication: Both sources approximately point into the same direction. Due to the head geometry, their positive poles overlay (big blue area in Figure 5.31f), while their negative poles are not well-captured by the electrode cap. Therefore, we use this scenario to revisit the comparison of the single modalities and their combination examined in Section 5.4.4.
For the single modalities, we use the full noise model with the regularized pre-stimulus covariance matrices (see Figures 5.32i, 5.32k) as above. For EMEG, a block diagonal noise model $\Sigma_{reg}^{EMEG} = \operatorname{diag}\left(\Sigma_{reg}^{EEG}, \Sigma_{reg}^{MEG}\right)$ is build from the single modalities. Thereby, we do not account for the covariance between EEG and MEG channels, which we leave for future work. Figure 5.35 shows the results computed with the same source space as before. For MEG, we used the same hyperprior parameters α and β as in the simulation studies, as those were chosen to reconstruct multiple dipole scenarios with similar source amplitudes. The result fits very well to our expectations. For EEG, we needed to increase α to obtain a reasonable result. However, the corresponding result also fits surprisingly well to our source hypothesis. The slight differences in location and orientation compared to the MEG-based reconstructions may result from the insufficient coverage of the EEG cap in the lower parts of the head: For an accurate estimation of location and orientation, a coverage of the maxima of the poles by the EEG cap would be necessary. Finally, the EMEG result resembles the MEG result. If we assume that the MEG result is more reliable than the EEG one, this behavior is consistent with the simulation studies we performed in Section 5.4.4: If one modality yields a better result than the other, the combined reconstruction is not disturbed by the weaker modality but can profit from both information (cf. Figure 5.29).

Discussion and Outlook

In this section, we examined the inversion of experimental data for a quite challenging inverse problem:

- Compared to CT, realistic and accurate forward modeling and computation is much more involved (cf. Section 5.4.1). Geometrically, our head model features

a degree of realism that is seldom found in EEG/MEG source reconstruction. However, concerning EEG/MEG combination, a considerable improvement could still be achieved if *calibrated* conductivities would be used: The conductivities we assigned (cf. Table 5.7) were standard values found in the literature. While certain compartments like the CSF show little inter- and intra-subject variability, it is known that, for instance, the skull conductivity can vary considerably. This uncertainty was found to be a major problem for EEG/MEG combination. Therefore, calibration procedures were developed that try to identify the skull conductivity or the ratio of the two skull compartments by using reference SSEP/SSEF measurements. A recent overview of this topic can be found in Aydin et al. (2014) and references therein. While the principled computations required by the calibration procedure were carried out for the SSEP/SSEF data examined in this section, the results were not included in this thesis anymore.

- We investigated the aspect of data preprocessing, which we ignored up to now. While we only carried out very elementary, conservative preprocessing procedures, further improvements can be expected if signal unmixing strategies are applied: *Principal component analysis* (*PCA*) and *independent component analysis* (*ICA*, Makeig et al. 1996) are most commonly used to clear signal components such as the heart's signal. Model-based *factor analysis* approaches such as *stimulus-evoked independent factor analysis* (SEIFA, Nagarajan et al. 2006) have been developed to unmix stimulus-related and background brain activity (factor analysis is closely related to *low-rank approximation*).

- Even after improving upon the previous two aspects, uncertainties will remain. While Section 3.6.2 introduced principled ways to address them, we conducted sensitivity studies to examine how potential modeling errors will affect the reconstruction results. We found that HBM was surprisingly robust to noise mis-specification and residual background activity.

- We computed the first fully-Bayesian HBM source reconstructions for experimental data (to the best of our knowledge) and could demonstrate that HBM does not only give promising results in simulated data studies but can also provide reasonable source estimates in real data scenarios. To consolidate our results, further evaluations have to be carried out:

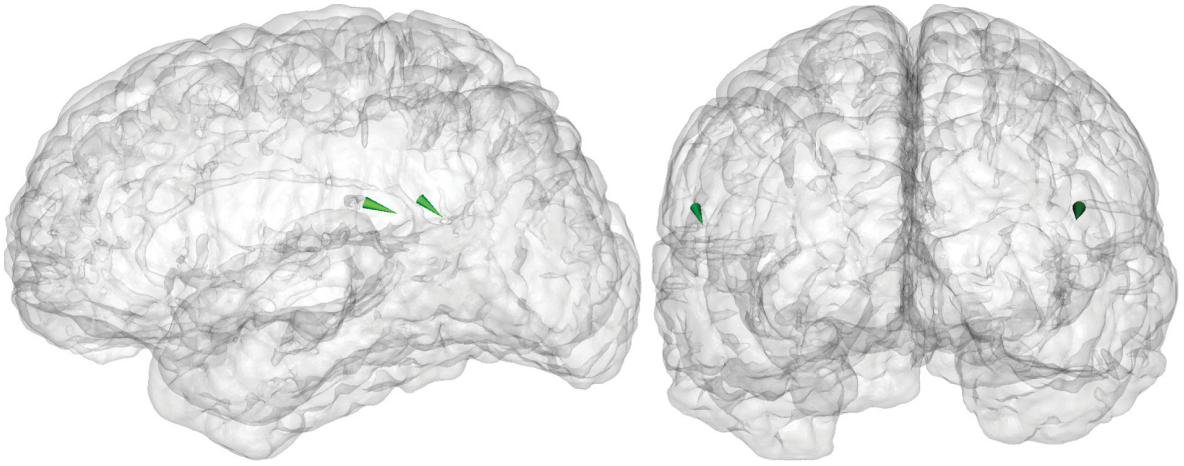
  - For the SSEP data, the HBM results matched those of the well-established, robust SDS. However, as both results did not fully match our a-priori expectations on the location of the underlying brain activity, further examinations have to be carried out: A problem of the current studies may be the strong

band-pass filtering we used. Aside altering the timings of the physiological components, it may also affect the spatial topographies in a non-trivial way. In addition, the reconstructed locations have to be compared to an atlas-based cortical segmentation matched to the head model.

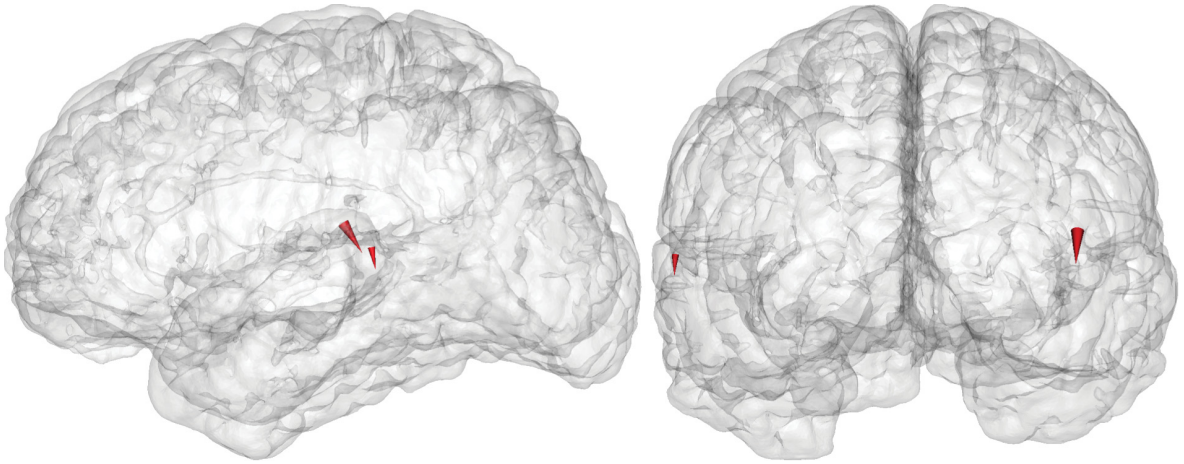
– For the AEP/AEF data, the location of the HBM results fitted well to our a-priori expectations on the underlying brain activity and the results of EMEG compared to EEG or MEG alone fitted well to our simulation studies in Section 5.4.4. However, the impact on the orientation of the sources has to be investigated in more detail.

In addition to the above points, the stability of the results to the SNR of the data has to be examined; for instance by reducing the number of trials selected for trial averaging. Most importantly, different subjects and source scenarios have to be investigated.

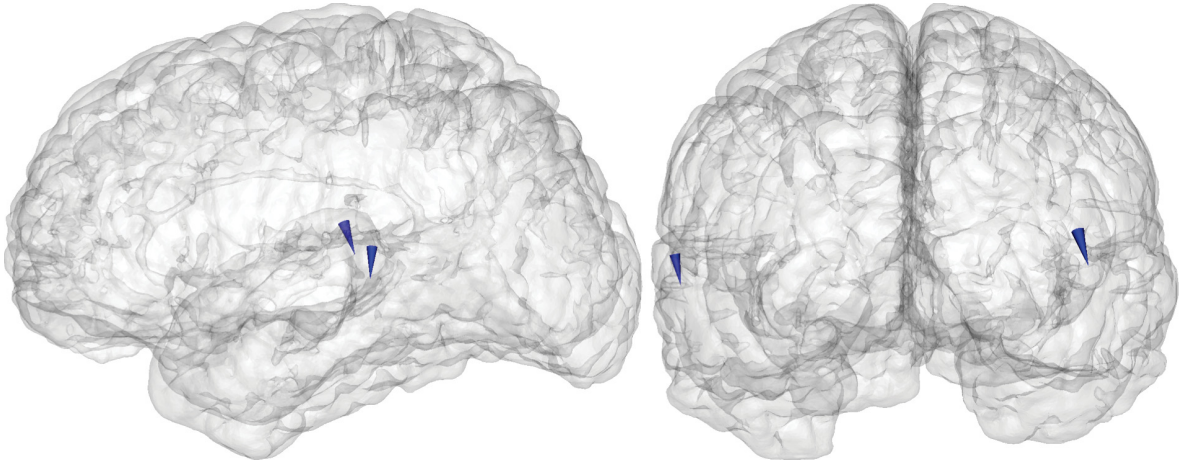
- We ignored the temporal aspect of EEG/MEG inversion up to now: Its high temporal resolution is one of the main arguments for using EEG/MEG in many situations. For source reconstruction, the temporal information can further be used to supplement the poor spatial information of a single time-point measurement. The section "Static and dynamic inverse problems" in Lucka et al. (2012) contains a discussion of possible *spatio-temporal* inversion techniques, in particular in combination with HBM.



(a) EEG, HBM parameters: $\alpha = 1.2, \beta = 10^{-4}$



(b) MEG, HBM parameters: $\alpha = 0.5, \beta = 10^{-4}$



(c) EMEG, HBM parameters: $\alpha = 0.5, \beta = 10^{-4}$

Figure 5.35.: HBM-NM results for the AEP/AEF data visualized by colored cones inside the gray matter surface.

5.4.6. Sparse Recovery Conditions

Motivation

Using sparsity constraints has become popular in source reconstruction as well (we give references at the end of this section). Besides the HBM-based approaches we examined in the previous sections, various kinds of ℓ_1 (block) priors were proposed. In Section 3.5, we discussed that under certain "best-case" assumptions, the performance of MAP estimates for such prior models can be examined using different recovery conditions. Here, we test whether and which of these conditions may be suitable to investigate topics in source reconstruction such as

- the interplay between realistic forward and sparse inverse modeling: How do the intrinsic recovery properties of A evolve with modeling complexity? This is different from examining which kind of reconstruction errors the use of a simplified head model induces (see, e.g., Lanfer et al. 2012). It rather tries to assess how predictive the reconstruction performance of an inverse method in a simplified head model is for its performance in a more realistic one.

- the number source space locations N needed for sparse inversion: Clearly, the spatial resolution is ultimately limited by the imaging modality and not by the user-defined N. Choosing a small N for computational reasons might potentially lower the spatial resolution achievable. On the other hand, choosing a large N does not mean that the source grid size actually reflects the real spatial resolution and leads to unnecessary computational effort. The dependence of the recovery conditions on N may shed new light on this question.

- the comparison between EEG, MEG and EMEG was often examined by simulation studies like the one in Section 5.4.4, only. Sparse recovery conditions may provide a new perspective on this topic.

Study Design and Results

We examined all the head models HM1-HM6 described in Section 5.4.2. Figure A.5 shows the sensor configuration used. It consists of 74 EEG electrodes and 273 MEG gradiometer channels and corresponds to a realistic setup similar to the one used in the real data studies. We constructed surface-based source spaces consisting of $N = 62$, 125, 250, 500, 1000, 2000, 4000, 8000 and 16000 source locations. For source reconstruction in real data scenarios, N is typically at least 1000. Therefore, we will refer to $N = 62 - 500$ as "small" and to $N \geqslant 1000$ as "realistic". We examined two sparse reconstruction approaches (cf. Section 5.4.3):

1. Scalar reconstruction using the normal constraint: In this setting, a simple ℓ_1 prior ($D = I_n$) is used and the normal recovery conditions apply for the MAP-estimates.

2. Full vector reconstruction without the normal constraint: In this setting, an ℓ_1-block prior, (3.16), is used and the block recovery conditions discussed in Section 3.5.5 apply.

The following computations were carried out for EEG, MEG and EMEG:

- All coherence measures, (3.63), (3.82) and (3.83) were computed for all N.
- The lower bound δ_2^{lb} on δ_2 was computed by Algorithm 4.12 using 10^8 2-sparse samples for all N.
- All non-uniform recovery conditions were tested for 1000 source configurations $u^\dagger$ with $k = 2$ and $k = 3$ active locations. In the full vector reconstruction, only the normally oriented of the three source dipoles at each active location was active. We excluded source configurations where two of the active locations were neighboring each other as one would not speak of two separated sources in this case. For the scalar conditions (Tr), (FuA), (FuB)/(SSC) and (SSC_+), computations were limited to $N = 8000$; for the block conditions (BlkTr), (BlkFuA), (BlkFuB)/(SSC), computations were limited to $N = 1000$. Whenever a condition can be applied to both A and $A^\sharp$, both variants were tested.

The lead-field matrices A_{EEG}, A_{MEG} were re-scaled by their spectral norm prior to these computations. A_{EMEG} was build by stacking the re-scaled A_{EEG}, A_{MEG}. This corresponds to assuming that both modalities have the same signal-to-noise ratio (cf. Sections 5.4.4, 5.4.5). Note that the column-normalization required by some conditions is performed after this re-scaling and is not related to it.
All results are listed in tables A.3-A.19 in the appendix. For the non-uniform conditions, the empirical probability and confidence intervals for a significance level of $\alpha = 5\%$ are displayed.

Discussion

Coherence Note that the source space locations for smaller N are not a subset of those for a larger N. Therefore, a monotonic increase of the coherence cannot be expected. Tables A.3 and A.4 show that even for the small N, the coherence values $\mu(A)$ are so close to 1 that (Coh) only guarantees exact recovery in the trivial case of 1-sparse solutions $u^\dagger$. Tables A.5 and A.6 show that the situation is even worse for full vector reconstruction: Although the block coherence is notably smaller than the normal one, (BlkCoh) does not even give exact recovery guarantees for 1-block-sparse solutions. A

particular reason for this is that the sub coherence is very close to 1, which means that there are certain locations where the topographies of the three dipoles at these locations are very similar. Note, however, that (BlkCoh) would guarantee for the exact recovery of both location and orientation. In source reconstruction, we are often only interested in the exact location.

The results do not yield a clear trend for the comparison between EEG and MEG. However, for their combination, the coherence results suggest that EMEG is stronger than the stronger one of the single modalities for small N, while it is weaker for large N: $\min(\mu(A^{EEG}, A^{MEG}) \leqslant \mu(A^{EMEG})$ is never true for $N \leqslant 500$ but true in 16 of 18 cases for $N \geqslant 4000$. The latter seems to confirm the common objection that combining EEG and MEG might rather compensate the particular strengths of one modality with weaknesses of the other (cf. Section 5.4.4). However, as the results, above all, show that the coherence condition is too strong for examining exact recovery for source reconstruction, it is not yet clear whether drawing such conclusions is really valid.

RIP Table A.7 lists δ_2^{lb}, while δ_2^{ub} is given by $\mu(A)$ (cf. Table A.3). Comparing both values, we see that the upper bound is quite tight. Unfortunately, the results for δ_2^{lb} also already prohibit (RIP) from being true for $k = 1$. As all other RIP (block) constants δ_k, $\delta_{[k]}$ are even larger than δ_2, condition (RIP) will never be fulfilled in our scenario.

Concerning the comparison between EEG, MEG and EMEG, the RIP results suggest the same conclusions as the coherence results. However, one should, again, be cautious as the results primarily show that uniform recovery conditions seem to be too strong to analyze sparse source reconstruction.

Non-uniform, scalar We first examine the results for the normalized lead-field $A^\sharp$: Comparing the results for (Tr) (Table A.8), (FuA) (Table A.10) and (FuB)/(SSC) (Table A.12), we see that the empirical probabilities differ dramatically. This suggests that the sufficient conditions (Tr) and (FuA) are also too strong and only the sufficient and necessary condition (FuB)/(SSC) is adequate for examining exact recovery in source reconstruction. Therefore, we omitted the results for $k = 3$ for the other conditions. The results also clearly show that using the normalized lead-field has a dramatic influence on the recovery performance; see for instance Table A.13. As discussed in Section 3.5, the implications of this difference are not fully examined yet. Comparing the results with or without an additional non-negativity constraint (e.g., Tables A.11 and A.14) shows that the inclusion of such constraints can significantly enhance the recovery performance and motivates further research in this direction.

In contrast to the uniform recovery conditions, the non-uniform conditions (FuB)/(SSC) and (SSC_+) clearly show that the MEG sensor configuration is superior to the EEG

one for sparse recovery. As the number of MEG channels is considerably larger than the number of EEG electrodes, this is also the expected result. A fair comparison like the one carried out in Section 5.4.4 is difficult for realistic sensor configurations: One should at least compare this MEG sensor configuration to a *high-resolution EEG cap* consisting of 256 channels (Oostenveld and Praamstra 2001). However, for this study we chose to use the same sensor configurations as in the real data recordings. Concerning the combination of EEG and MEG, the results for (FuB)/(SSC) and (SSC_+) clearly support the findings of Section 5.4.4: The combination significantly increases the average reconstruction performance compared to the single modalities. These findings underline that the results of sufficient but not necessary conditions like (Coh) should be interpreted with care and might be misleading.

Non-uniform, vectorial The full vector reconstruction results are similar to the scalar ones: There is a dramatic difference between conditions (BlkTr) and (BlkFuA) and the condition BlkFuB/(SSC). (BlkTr) was actually never full-filled in this study. With 1000 samples tested, this amounts to a 5%-confidence interval of $[0.000, 0.004]$. Concerning the comparison between EEG, MEG and EMEG, the results also show that the MEG sensor configuration used is superior to the EEG one and that combining EEG and MEG can significantly improve over the single modalities: Using $N = 1000$ and HM1 is for instance a situation comparable to the scenarios used in the studies in Sections 5.4.4 and 5.4.5. Table A.18 shows that in this case, the empirical recovery probability increases from 1.5% for EEG alone or 13% for MEG alone to 46.1% in the combined case.

Head Model Comparison Based on the results for (FuB)/(SSC) in Tables A.11-A.13, we will draw some preliminary conclusions on the comparison between the different head models HM1-HM6:

- HM1 and HM2 differ by modeling the GM and WM anisotropic or isotropic. Without the normalization of A, both EEG and MEG show decreased recovery probabilities for HM1 compared to HM2. The decrease for EEG is more pronounced. Using normalization, this effect disappears. Therefore, a better understanding of the effect of normalization on the recovery conditions is also needed to fully understand this phenomenon.
- The difference between HM2 and HM3, which consists of neglecting the eye compartment and replacing the two-layered skull by a single-layered one does not seem to have a major impact on the recovery probabilities.
- From HM4 to HM5, the realistic geometry of the three compartment model is

replaced by the ellipsoidal shape. While EEG and EMEG do not show significant differences in recovery probability, MEG is strongly effected by it.

- HM6 is somewhat special in the comparison in the sense that it would not be considered an appropriate head model for EEG in practice. The large differences for EEG between HM5 and HM6 also partly reflect this. For MEG, we see only subtle changes. This suggests that from the sparse recovery perspective, both models are similar.

In general, the interpretation of the results using the normalized $A^\sharp$ is complicated by the high recovery probabilities found in this case (*ceiling effect*). For distinguishing probabilities very close to one, a larger number of samples would have to be drawn.
Compared to the scalar reconstruction, the results for the vectorial case, BlkFuB, are more difficult to interpret: They guarantee the exact recovery of location and orientation, but we are typically rather interested in the correct location and would allow for an imprecision in orientation. Therefore, we did not include the vectorial results into the discussion above.

Conclusions

Uniform recovery conditions are easy to compute, give very general recovery guarantees and are therefore most often considered in compressed sensing theory. However, our results show that they are too strong for the inverse problem of EEG/MEG and cannot be used to tackle the type of questions that motivate our work. From the non-uniform recovery conditions, only the weakest one, (FuB)/(SSC), gives promising recovery guarantees in a range, where, for instance, a meaningful comparison between different head models is possible (although the preliminary conclusions we drew in the last paragraph need to be refined). However, the computational verification of this condition is more difficult and time-consuming.

Extensions

As discussed in Section 5.4.3, given the inherent uncertainties of the MRI-derived surface representations, using the flexible *loose* instead of the fixed *normal* orientation constraint is probably a better way to include direction information. As these constraints correspond to weighted block priors, the studies carried out in this section can easily be repeated for them.
Since

$$\min \|w\|_2^2 \qquad s.\,t. \qquad w \quad \text{fulfills (FuB)/(SSC)} \tag{5.14}$$

is related to obtaining optimal error estimates by (3.79) and (3.80) in the case of non-vanishing noise, computing its statistics can help to extend the current results beyond the noise-free assumption. Preliminary studies of such kind were already carried out but are not included in this thesis anymore. First results suggest that normalizing the lead-fields does not only lead to improved exact recovery rates in the noise-free case but also lead to significantly lower mean values of (5.14).

Outlook

The aim of these first, elementary studies was to implement and identify the right tools to tackle relevant questions in source reconstruction. For this purpose, a rather general study design was used and only empirical recovery statistics were computed. Future studies can tackle more application-specific questions:

- *Experimental design*: If a hypothesis about the source configuration to recover is available, we can use this methodology to examine which modality is best suited to recover it or how to alter the configuration to improve its reconstruction, for instance by using additional electrodes.

- The spatial resolution of sparse source reconstruction: For a given source location i, BlkFuB/(SSC) is tested for all 2-sparse $u^\dagger$ consisting of i and another site j (using the NC). The result is a binary map which can be visualized on the cortical surface. It displays in which spatial configuration the second source has to be placed to be separable from the source at location i. It should be possible to define a measure of local spatial resolution from this binary map. This leads to a map of local spatial resolution which can be used to classify the different brain regions depending on whether the modality can offer a good or bad resolution. The comparison between EEG, MEG and EMEG could then be focused to certain brain regions, for instance, "should one use EEG or MEG for examining auditory activity and which benefit would EMEG provide for this scenario?".

- While the previous point contained ideas to define a local spatial resolution for a fixed source space, the next step would be to use this methodology to compare source spaces with different N to obtain a reasonable balance between the number of locations used and the spatial resolution they actually achieve.

- Designing source spaces would be a more sophisticated use of these ideas: The task would be to arrange a fixed number of source locations N in order to exploit the limits of the local spatial resolution. This might result in a more dense covering of the superficial cortical layers compared to the deep-lying regions. While such a paradigm may seem odd for applications where source reconstruction is performed

after the measurement and is not time-critical (such as the analysis of evoked potentials), a recent field of research with exiting potential applications is in EEG/MEG online source reconstruction (Pieloth et al. 2013).

- Using this methodology to quantify the reliability of a reconstructed solution would be interesting for clinical applications: For instance, for the pre-surgical diagnosis of epilepsy patients, it is important to assure that the source reconstruction did not miss any active source.

Notes and Comments

Important contributions to the use of ℓ_1-based approaches to EEG/MEG source reconstruction include Chang et al. (2010), Gramfort et al. (2013), Haufe et al. (2008), Huang et al. (2006), Matsuura and Okabe (1995), Ou et al. (2009), Uutela et al. (1999). A recent overview also covering the use of orientation constraints can be found in Chang et al. (2013).

6 Classical Bayesian Theory Revisited

In this chapter, we will revisit some topics discussed in Chapter 3 in the light of the computational results we presented in the previous chapter.

6.1. MAP or CM Estimation Revisited

In Section 3.4.3, we discussed the classical view on the "MAP or CM" topic, which favors the CM estimate over the MAP estimate on the basis of the Bayes cost formalism. This point of view is not only challenged by our computational results, but also by recent work of others. We will summarize these observations and results in the next section. Then, we will introduce new theoretical ideas, which are not longer contradictory, but fit to all of these results, disprove certain common myths, and will lead to new insights and perspectives for the comparison of variational regularization and Bayesian inference.

6.1.1. Converse Observations and Results

We start off with a very basic observation: As discussed in Sections 3.1 and 3.2.4, Gaussian priors are the most popular and arguably the most fundamental class of priors, due to various reasons such as their maximum entropy property, alpha-stability and the central limit theorem. However, for this most fundamental class of priors, the seemingly fundamentally different MAP and CM estimate happen to be equal. From the classical view, this can only be interpreted as a meaningless coincidence, which is arguably not

fully satisfactory.

Another observation is that the theoretical discrimination of the MAP estimate contrasts its popularity and success in practical applications, in particular in high-dimensional scenarios. Its popularity may be due to pragmatical reasons: Without MCMC techniques such as those developed in this thesis, computing CM estimates is infeasible in high-dimensional scenarios. Therefore, one often encounters a strange contrariness in publications about high-dimensional Bayesian inversion: Usually, a careful prior modeling is presented and the CM estimate is regarded as the optimal inference technique. However, for computational reasons, often only a MAP estimate can be computed. This circumstance is usually regretted and excused for. If the computational results are not fully satisfactory, shortcomings of the MAP estimate are discussed as a potential reason for it. However, even if the results are really good, concern is expressed that computing MAP estimates is not a proper Bayesian technique, and that CM estimates may even be superior.

The MCMC techniques developed in this thesis allowed us to bridge this gap between theoretical considerations and practical experience for various high-dimensional scenarios and prior models, in particular for sparsity-promoting priors. Our computational studies in Chapter 5 contained several comparisons between CM and MAP estimates:

- Figures 5.8, 5.9c, 5.9a, 5.10, 5.11a, 5.13, 5.17, 5.18, 5.23, 5.24 and 5.25 visually compare MAP and CM estimates for ℓ_p^q priors. In most cases, the MAP estimates are clearly more convincing. The only exceptions are those cases, where MAP and CM estimate visually coincide, for instance for the Besov priors with a large λ, and the TV prior results for the "Walnut-CT" scenario, Figures 5.23 and 5.25. In the latter case, it is not yet clear which estimate one should prefer for a subsequent use of the images produced (cf. Section 1.1).

- For the HBM-based source reconstructions in EEG/MEG, we already showed in Lucka (2011), Lucka et al. (2012) that full-MAP estimates appeared superior to full-CM estimates by visual impression and comparable by error measure statistics if they are suitably computed. The studies in Section 5.4 confirm this. In particular, we only showed images of HBM-NM results (which can be regarded as a suitable approximation to full-MAP estimates) because they were either visually identical, or superior to HBM-CM estimates.

Recently, Gribonval (2011), Gribonval and Machart (2013), Louchet and Moisan (2013) revealed that every CM estimate for a prior $p_{prior}(u)$ is also a MAP estimate for a different prior $\tilde{p}_{prior}(u)$. Their intention was to warn against the common "reverse reading" of designing a particular $\mathcal{J}(u)$ for recovering certain classes of real solutions with (3.12) and then claiming to perform Bayesian MAP estimation with the

prior $p_{prior}(u) = \exp(-\lambda\mathcal{J}(u))$. In Gribonval et al. (2012), it was shown that this MAP estimate is usually not very well suited to recover solutions $u^\dagger$ that are really distributed like $\exp(-\lambda\mathcal{J}(u))$. In our studies, the true solutions did also not correspond well to the priors used: This is apparent when comparing the random draws in Section 5.1.1 with the true solutions in Chapter 2. For the topic of this section, these results mean that a general discrimination of MAP estimates based on the Bayes cost formalism would only make sense if one strongly believes that the chosen prior most accurately models the distribution of the real solution. Otherwise, one ends up in the contradiction that the appraised CM estimate is simultaneously a discredited MAP estimate (just for another prior). The next sections resolve these contradictions between the observations and the classical view.

6.1.2. A Novel Characterization of the MAP Estimate

In this section, we will present a novel Bayes cost approach to MAP estimates for log-concave Gibbs priors (cf. Section 3.2.3). Throughout this section we will assume that $\mathcal{J} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a Lipschitz-continuous convex functional, such that for $\lambda > 0$ the function $u \mapsto \|Au\|^2 + \lambda\mathcal{J}(u)$ has at least linear growth at infinity. Due to Rademacher's theorem (cf. Evans and Gariepy 1991), this implies that $p_{post}(u|f)$ is log-concave and differentiable almost everywhere in $\mathbb{R}^n$. The main ingredient will be the (generalized) Bregman distance (cf. Section A.1).

New Bayes Cost Functions

The classical discrimination of the MAP estimate as only being asymptotically a Bayes estimator for the uniform cost (3.61) (cf. Section 3.4.3) has a crucial flaw: It does not mean that the MAP estimate cannot be a proper Bayes estimator for a different cost function. This suggests that one should search for alternative costs better suited to the asymptotic Banach space structure such as Bregman distance costs:

Definition 6.1. Let $L \in \mathbb{R}^{n\times n}$ be regular and $\beta > 0$. Define

$$\Psi_{\mathrm{LS}}(u, \hat{u}) := \|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} + \beta\|L(\hat{u} - u)\|_2^2 \tag{6.1}$$

$$\Psi_{\mathrm{Brg}}(u, \hat{u}) := \|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} + 2\lambda D_{\mathcal{J}}(\hat{u}, u). \tag{6.2}$$

Both $\Psi_{\mathrm{LS}}(u, \hat{u})$ and $\Psi_{\mathrm{Brg}}(u, \hat{u})$ are proper, convex (with respect to $\hat{u}$) cost functions. In the following, we will need the decay property

$$\lim_{R\to\infty} \int_{\partial\mathcal{B}_R(0)} p_{post}(u|f)\,\mathrm{d}u = 0, \tag{6.3}$$

which is fulfilled under the linear growth assumption above, which yields constants a, b independent of R such that

$$p_{post}(u|f) \leq a \exp\left(-\frac{b}{R}\right) \qquad \text{on} \quad \mathcal{B}_R(0). \tag{6.4}$$

Theorem 6.1. Let $\mathcal{J}$ be as above and let $\lambda > 0$ and $\beta \geqslant 0$. Then, the CM estimate is a Bayes estimator for $\Psi_{\text{LS}}(u, \hat{u})$ and the MAP estimate is a Bayes estimator for $\Psi_{\text{Brg}}(u, \hat{u})$.

Proof. We start from (3.59) and insert the definition of $\Psi_{\text{LS}}(u, \hat{u})$:

$$\hat{u}_{\Psi_{\text{LS}}}(f) = \operatorname*{argmin}_{\hat{u}} \left\{ \int \left(\|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} + \beta \|L(\hat{u} - u)\|_2^2 \right) p_{post}(u|f) \, \mathrm{d}u \right\} \tag{6.5}$$

We can rewrite the above by inserting $\hat{u}_{\text{CM}}$ and expanding squares

$$\begin{aligned} \hat{u}_{\Psi_{\text{LS}}}(f) = \operatorname*{argmin}_{\hat{u}} \Bigg\{ & \int \left(\|A(\hat{u} - \hat{u}_{\text{CM}})\|^2_{\Sigma_\varepsilon^{-1}} + \beta \|L(\hat{u} - \hat{u}_{\text{CM}})\|_2^2 \right) p_{post}(u|f) \, \mathrm{d}u \\ & + \int \left(\|A(u - \hat{u}_{\text{CM}})\|^2_{\Sigma_\varepsilon^{-1}} + \beta \|L(u - \hat{u}_{\text{CM}})\|_2^2 \right) p_{post}(u|f) \, \mathrm{d}u \\ - 2 \int & \left(\langle A(\hat{u} - \hat{u}_{\text{CM}}), A(u - \hat{u}_{\text{CM}}) \rangle_{\Sigma_\varepsilon^{-1}} + \beta \langle L(\hat{u} - \hat{u}_{\text{CM}}), L(u - \hat{u}_{\text{CM}}) \rangle_2 \right) p_{post}(u|f) \, \mathrm{d}u \Bigg\} \end{aligned} \tag{6.6}$$

Due to the linearity and the definition of the CM estimate (3.3) the last integral vanishes and hence, $\hat{u} = \hat{u}_{\text{CM}}$ is obviously a minimizer. For the MAP estimate, we again start from (3.59) and insert the definition of $\Psi_{\text{Brg}}(u, \hat{u})$:

$$\hat{u}_{\Psi_{\text{Brg}}}(f) = \operatorname*{argmin}_{\hat{u}} \left\{ \int_{\mathbb{R}^n} \left(\|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} + 2\lambda D_{\mathcal{J}}(\hat{u}, u) \right) p_{post}(u|f) \, \mathrm{d}u \right\} \tag{6.7}$$

Now, we can exclude the null-set where $\mathcal{J}(u)$ is not Fréchet-differentiable,

$$\mathcal{S} := \{ u \in \mathbb{R}^n | \ \text{card}\,(\partial \mathcal{J}(u)) \neq 1 \}, \tag{6.8}$$

from the integration and insert the definition of $D_{\mathcal{J}}(\hat{u}, u)$ on $\mathcal{S}^c$:

$$\begin{aligned} \hat{u}_{\Psi_{\text{Brg}}}(f) = \operatorname*{argmin}_{\hat{u}} \Bigg\{ & \int_{\mathcal{S}^c} \big(\|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} \\ & + 2\lambda \left(\mathcal{J}(\hat{u}) - \mathcal{J}(u) - \langle \mathcal{J}'(u), \hat{u} - u \rangle \right) \big) p_{post}(u|f) \mathrm{d}u \Bigg\} \end{aligned} \tag{6.9}$$

The squared norm can be developed as in the case of the CM estimate, while for the

Bregman distance we use the following elementary identity:

$$D_{\mathcal{J}}(\hat{u},u) = D_{\mathcal{J}}(\hat{u},\hat{u}_{\mathrm{MAP}}) + D_{\mathcal{J}}(\hat{u}_{\mathrm{MAP}},u) + \langle \hat{p}_{\mathrm{MAP}} - \mathcal{J}'(u), \hat{u} - \hat{u}_{\mathrm{MAP}}\rangle, \tag{6.10}$$

where $\hat{p}_{\mathrm{MAP}} \in \partial\mathcal{J}(\hat{u}_{\mathrm{MAP}})$. Thus, on $\mathcal{S}^c$ we have

$$\begin{aligned}\|A(\hat{u}-u)\|^2_{\Sigma_\varepsilon^{-1}} + 2\lambda D_{\mathcal{J}}(\hat{u},u) = \\ \|A(\hat{u}-\hat{u}_{\mathrm{MAP}})\|^2_{\Sigma_\varepsilon^{-1}} + 2\lambda D_{\mathcal{J}}(\hat{u},\hat{u}_{\mathrm{MAP}}) + \|A(\hat{u}_{\mathrm{MAP}}-u)\|^2_{\Sigma_\varepsilon^{-1}} + 2\lambda D_{\mathcal{J}}(\hat{u}_{\mathrm{MAP}},u) \\ + 2\langle A(\hat{u}-\hat{u}_{\mathrm{MAP}}), A(\hat{u}_{\mathrm{MAP}}-u)\rangle_{\Sigma_\varepsilon^{-1}} + 2\lambda\langle \hat{p}_{\mathrm{MAP}} - \mathcal{J}'(u), \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle. \end{aligned} \tag{6.11}$$

The first two terms in the second line are obviously minimal for $\hat{u} = \hat{u}_{\mathrm{MAP}}$, while the other terms in this line are independent of $\hat{u}$. In the last line we can insert the subgradient from the optimality condition for $\hat{u}_{\mathrm{MAP}}$,

$$\hat{p}_{\mathrm{MAP}} = -\frac{1}{\lambda}A^*\Sigma_\varepsilon^{-1}(A\hat{u}_{\mathrm{MAP}} - f) \in \partial\mathcal{J}(\hat{u}_{\mathrm{MAP}}), \tag{6.12}$$

and rewrite

$$\begin{aligned} 2\langle A(\hat{u}-\hat{u}_{\mathrm{MAP}}), A(\hat{u}_{\mathrm{MAP}}-u)\rangle_{\Sigma_\varepsilon^{-1}} + 2\lambda\langle \hat{p}_{\mathrm{MAP}} - \mathcal{J}'(u), \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle \\ = -2\langle A(\hat{u}-\hat{u}_{\mathrm{MAP}}), Au - f\rangle_{\Sigma_\varepsilon^{-1}} - 2\lambda\langle -\mathcal{J}'(u), \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle \\ = -2\langle A^*\Sigma_\varepsilon^{-1}(Au-f) + \lambda\mathcal{J}'(u), \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle \\ = 2\langle \nabla_u \log p_{post}(u|f), \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle. \end{aligned} \tag{6.13}$$

Using the logarithmic derivative $\nabla_u p_{post}(u|f) = (\nabla_u \log p_{post}(u|f))p_{post}(u|f)$, the posterior expectation of the latter equals

$$2\int_{\mathcal{S}^c} \langle \nabla_u \log p_{post}(u|f), \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle\, p_{post}(u|f)\,\mathrm{d}u = 2\langle \int_{\mathcal{S}^c} \nabla_u p_{post}(u|f)\,\mathrm{d}u, \hat{u}-\hat{u}_{\mathrm{MAP}}\rangle. \tag{6.14}$$

With Gauss' theorem and (6.3) we finally obtain:

$$\begin{aligned} \left\| \int \nabla_u p_{post}(u|f)\,\mathrm{d}u \right\| &= \lim_{R\to\infty} \left\| \int_{\mathcal{B}_R(0)} \nabla_u p_{post}(u|f)\,\mathrm{d}u \right\| \\ &= \lim_{R\to\infty} \left\| \int_{\partial\mathcal{B}_R(0)} p_{post}(u|f)\frac{u}{R}\,\mathrm{d}u \right\| \leqslant \lim_{R\to\infty} \int_{\partial\mathcal{B}_R(0)} p_{post}(u|f)\,\mathrm{d}u = 0 \end{aligned} \tag{6.15}$$

□

First, we apply Theorem 6.1 to the fundamental case of Gaussian priors. We can parameterize any (centered) Gaussian energy as $\mathcal{J}(u) = \beta/(2\lambda)\|Lu\|_2^2$. For this choice,

$2\lambda D_{\mathcal{J}}(\hat{u}, u) = \beta\|L(\hat{u} - u)\|_2^2$, and $\Psi_{\text{LS}}(u, \hat{u}) = \Psi_{\text{Brg}}(u, \hat{u})$: The equality of MAP and CM estimate in the Gaussian case is no longer a strange coincidence but follows naturally from the properties of the Bregman distance.

In the non-Gaussian case, the domain of $\mathcal{J}$ usually defines a Banach space or a subset thereof in the limit $n \to \infty$. For instance, the discrete total variation prior will define the space of functions of bounded variation in the limit (Burger and Osher 2013). In such a space, there is no natural Hilbert space norm that one should obtain as the limit of $\|Lu\|^2$. Even worse, it is questionable whether any Hilbert space norm is a meaningful measure for functions of bounded variation. The only reasonable choice might be $L = 0$, which means that $\Psi_{\text{LS}}(u, \hat{u})$ measures purely in the output space, which will be a Hilbert space. However, for ill-posed inverse problems with noisy data it is well-established that one should not just minimize a criterion related to the output Au.

A MAP-Centered Form of the Posterior

As pointed out in Section 3.4.3, one classical geometrical argument was that the CM estimate is in the center of mass of $p_{post}(u|f)$, while the MAP estimate does not allow for such an interpretation. Using Bregman distances, we can rewrite $p_{post}(u|f)$ in a MAP-centered form, which also disqualifies this argument. We use the optimality condition of the MAP-estimate (3.12),

$$A^*\Sigma_\varepsilon^{-1}(A\hat{u}_{\text{MAP}} - f) + \lambda\hat{p}_{\text{MAP}} = 0, \qquad \hat{p}_{\text{MAP}} \in \partial\mathcal{J}(\hat{u}_{\text{MAP}}), \tag{6.16}$$

to rewrite $A^*\Sigma_\varepsilon^{-1}f$ in the posterior energy:

$$\begin{aligned}
\frac{1}{2}\|A\,u - f\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda\mathcal{J}(u) =& \frac{1}{2}\|A\,u\|_{\Sigma_\varepsilon^{-1}}^2 - \langle A^*\Sigma_\varepsilon^{-1}f, u\rangle + \lambda\mathcal{J}(u) + \frac{1}{2}\|f\|_{\Sigma_\varepsilon^{-1}}^2 \\
=& \frac{1}{2}\|A\,u\|_{\Sigma_\varepsilon^{-1}}^2 - \langle A^*\Sigma_\varepsilon^{-1}A\hat{u}_{\text{MAP}} + \lambda\hat{p}_{\text{MAP}}, u\rangle + \lambda\mathcal{J}(u) + \frac{1}{2}\|f\|_{\Sigma_\varepsilon^{-1}}^2 \\
=& \frac{1}{2}\|A\,u\|_{\Sigma_\varepsilon^{-1}}^2 - \langle\Sigma_\varepsilon^{-1}A\hat{u}_{\text{MAP}}, Au\rangle + \frac{1}{2}\|A\hat{u}_{\text{MAP}}\|_{\Sigma_\varepsilon^{-1}}^2 \\
&+ \lambda\left(\mathcal{J}(u) - \mathcal{J}(\hat{u}_{\text{MAP}}) - \langle\hat{p}_{\text{MAP}}, u - \hat{u}_{\text{MAP}}\rangle\right) \\
&- \frac{1}{2}\|A\hat{u}_{\text{MAP}}\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda\left(\mathcal{J}(\hat{u}_{\text{MAP}}) - \langle\hat{p}_{\text{MAP}}, \hat{u}_{\text{MAP}}\rangle\right) + \frac{1}{2}\|f\|_{\Sigma_\varepsilon^{-1}}^2 \\
=& \frac{1}{2}\|A\,(u - \hat{u}_{\text{MAP}})\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda D_{\mathcal{J}}^{\hat{p}_{\text{MAP}}}(u, \hat{u}_{\text{MAP}}) + \text{const.},
\end{aligned} \tag{6.17}$$

where const. sums all terms not depending on u. Hence, we can write the posterior as

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|A(u - \hat{u}_{\text{MAP}})\|_{\Sigma_\varepsilon^{-1}}^2 - \lambda D_{\mathcal{J}}^{\hat{p}_{\text{MAP}}}(u, \hat{u}_{\text{MAP}})\right). \tag{6.18}$$

Now, the posterior energy is the sum of two convex functionals both minimized by $\hat{u}_{\mathrm{MAP}}$. Thereby, $\hat{u}_{\mathrm{MAP}}$ is the center of $p_{post}(u|f)$ with respect to the distance induced by (6.17).

Average Optimality of the CM Estimate

To further compare MAP and CM estimates, we derive an "average optimality condition" for the CM estimate. Let

$$\hat{p}_{\mathrm{CM}} := \mathbb{E}\left[\mathcal{J}'(u)\right] = \int \mathcal{J}'(u)\, p_{post}(u|f)\,\mathrm{d}u \tag{6.19}$$

be the CM estimate for the (sub)gradient of $\mathcal{J}(u)$. We have:

$$\begin{aligned} A^*\Sigma_\varepsilon^{-1}(A\hat{u}_{\mathrm{CM}} - f) + \lambda\hat{p}_{\mathrm{CM}} &= A^*(A\Sigma_\varepsilon^{-1}\,\mathbb{E}[u] - f) + \lambda\,\mathbb{E}[\mathcal{J}'(u)] \\ &= \mathbb{E}\left[A^*\Sigma_\varepsilon^{-1}(Au - f) + \lambda\mathcal{J}'(u)\right] \\ &= \int_{\mathcal{S}^c} A^*\Sigma_\varepsilon^{-1}(Au - f) + \lambda\mathcal{J}'(u)\, p_{post}(u|f)\,\mathrm{d}u \\ &= \int_{\mathcal{S}^c} \nabla_u p_{post}(u|f)\,\mathrm{d}u = 0, \end{aligned} \tag{6.20}$$

where the integral term, again, vanishes. Comparing (6.20) to (6.16), we see that the CM estimate fulfills an optimality condition "on average", i.e., with respect to the average gradient, $\hat{p}_{\mathrm{CM}} = \mathbb{E}[\mathcal{J}'(u)]$, but not with respect to the gradient $\mathcal{J}'(\hat{u}_{\mathrm{CM}}) = \mathcal{J}'(\mathbb{E}[u])$. The difference between MAP and CM estimate here manifests in $\mathcal{J}'(\mathbb{E}[u]) \neq \mathbb{E}[\mathcal{J}'(u)]$, which, again, vanishes for the Gaussian case where $\mathcal{J}'(u)$ is linear.

New Inequalities

Finally, we show that when measured in the Bregman distance $D_J(\hat{u}, u)$, which is a more reasonable error measure than norms in the case of a non-quadratic $\mathcal{J}(u)$ (cf. Section 3.5.4), the MAP estimate performs better than the CM estimate. In return, the CM estimate out-performs the MAP estimate when the error is measured in a quadratic distance:

Theorem 6.2. Let $L \in \mathbb{R}^{n\times n}$ be regular, then we have

$$\mathbb{E}\left[\|L(\hat{u}_{\mathrm{CM}} - u)\|_2^2\right] \leqslant \mathbb{E}\left[\|L(\hat{u}_{\mathrm{MAP}} - u)\|_2^2\right] \tag{6.21}$$

$$\mathbb{E}\left[D_{\mathcal{J}}(\hat{u}_{\mathrm{MAP}}, u)\right] \leqslant \mathbb{E}\left[D_{\mathcal{J}}(\hat{u}_{\mathrm{CM}}, u)\right]. \tag{6.22}$$

Proof. The first inequality directly follows from the fact that $\hat{u}_{\mathrm{CM}}$ is also the Bayes estimator for $\Psi(u, \hat{u}) = \|L(\hat{u}_{\mathrm{CM}} - u)\|_2^2$, which follows from the proof to Theorem 6.1.

For the second inequality, we use the minimizing properties of MAP and CM estimates:

$$\begin{aligned}\int \left(\|A(\hat{u}_{\mathrm{MAP}}-u)\|^2_{\Sigma_\varepsilon^{-1}}+2\lambda D_{\mathcal{J}}(\hat{u}_{\mathrm{MAP}},u)\right)\, p_{post}(u|f)\,\mathrm{d}u \\ \leqslant \int \left(\|A(\hat{u}_{\mathrm{CM}}\ -u)\|^2_{\Sigma_\varepsilon^{-1}}+2\lambda D_{\mathcal{J}}(\hat{u}_{\mathrm{CM}},u)\right)\, p_{post}(u|f)\,\mathrm{d}u \\ \leqslant \int \left(\|A(\hat{u}_{\mathrm{MAP}}-u)\|^2_{\Sigma_\varepsilon^{-1}}+2\lambda D_{\mathcal{J}}(\hat{u}_{\mathrm{CM}},u)\right)\, p_{post}(u|f)\,\mathrm{d}u \\ +\beta\int \left(\|L(\hat{u}_{\mathrm{MAP}}-u)\|_2^2-\|L(\hat{u}_{\mathrm{CM}}-u)\|_2^2\right) p_{post}(u|f)\,\mathrm{d}u \qquad (6.23)\end{aligned}$$

Since $\beta > 0$ is arbitrary, we can consider $\beta \to 0$ and obtain

$$\int D_{\mathcal{J}}(\hat{u}_{\mathrm{MAP}},u)\, p_{post}(u|f)\,\mathrm{d}u \leqslant \int D_{\mathcal{J}}(\hat{u}_{\mathrm{CM}},u)\, p_{post}(u|f)\,\mathrm{d}u \qquad (6.24)$$

□

Notes and Comments

A potential irritation might be that the cost function for the MAP estimate depends on the chosen prior while the one for the CM estimate does not. However, this is usually not a drawback but rather an advantage: The prior energy $\mathcal{J}(u)$ is chosen such that it grasps the most distinctive features of u (cf. Section 3.2). Often, one is consequently also most interested in estimating these features correctly, which is measured by $D_{\mathcal{J}}(u,v)$ better than in some squared error metric. For instance, in the "Spots" scenario, one is mainly interested in the correct separation and location of the intensity spots while their absolute amplitudes might be of minor interest. In such situations, the standard squared error is a poor indicator of reconstruction quality (see also the discussions in Benning 2011, Burger and Osher 2004, Burger et al. 2007, Schuster et al. 2012). On the other hand, the induced Bregman distance $D_{\mathcal{J}}(u,v)$ is 0 if the sign pattern of u and v coincide (cf. Table A.1) and grows only linearly, otherwise.

6.2. Sparsity in Bayesian Inversion

In the computational results, we saw that both ℓ_1 priors and conditionally ℓ_2 hypermodels can lead to sparse or at least compressible estimates. In this section, we will discuss the differences between the two approaches to encode sparsity as a-priori information in the Bayesian framework. For simplicity, we assume that all prior operators are the identity, $D = I_n$.

6.2.1. The ℓ_p Approach to Sparsity

The starting point for this approach is the observation that the solution to

$$\hat{u}_{\mathrm{MAP},2} = \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ \frac{1}{2}\|f - Au\|^2_{\Sigma_\varepsilon^{-1}} + \lambda\|u\|_2^2 \right\} \tag{6.25}$$

is easy to compute but not sparse, while the solution to

$$\hat{u}_{\mathrm{MAP},0} = \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ \frac{1}{2}\|f - Au\|^2_{\Sigma_\varepsilon^{-1}} + \lambda|u|_0 \right\} \tag{6.26}$$

is sparse but hard to compute. Obviously, (6.25) can be interpreted as the MAP estimate for using an ℓ_p prior with $p = 2$, while (6.26) can be interpreted as the limit thereof for $p \to 0$. This suggests to use the MAP estimate with the smallest p that allows for an efficient computation. Due to the fact that convex optimization problems can be solved very efficiently (cf. Section 4.2.2), this happens to be $p = 1$. While this procedure produces excellent results in a variety of applications, labeling it as a Bayesian inversion technique has certain flaws: In Bayesian inversion, the choice of the prior should reflect our a-priori knowledge about the solution, not the limitations of our computational abilities. If sparsity or compressibility is what we expect, the prior samples in Figure 5.1 illustrate that an ℓ_1 prior is not the correct model. In addition, from the Bayesian point of view, the sparsity of the MAP estimate can almost be considered a defect. The only reason for it is the non-differentiability of the convex ℓ_1 norm at 0, i.e, on a null-set: As the optimality condition is given by

$$-\frac{1}{\lambda}A^T(A\hat{u}_{\mathrm{MAP}} - f) \in \partial|\hat{u}_{\mathrm{MAP}}| = \begin{cases} \{1\} & \text{for } (\hat{u}_{\mathrm{MAP}})_i > 0 \\ [-1,1] & \text{for } (\hat{u}_{\mathrm{MAP}})_i = 0 \\ \{-1\} & \text{for } (\hat{u}_{\mathrm{MAP}})_i < 0 \end{cases}, \tag{6.27}$$

(cf. Section A.1), sparse u are most likely to fulfill it (cf. Figure 3.5a). For these reasons, one should rather characterize this approach as "reverse reading" as already discussed in Section 6.1.1: First, an optimization problem is designed such that it produces a solution with certain features. Then, the solution is interpreted as a MAP estimate in a Bayesian framework. The latter is problematic for the discussed reasons.

In general, the prior models used in this thesis all rely on probability densities and are therefore not well-suited to express sparsity measured in this binary sense of components being either exactly zero or having an arbitrary non-zero value: The set of u where at least one component is exactly zero is a null-set in all density-based prior models. For modeling this kind of sparsity, one should actually use semi-discrete prior models such

as a Bernoulli-Gauss model:

$$u_i = \xi_i \cdot v_i, \qquad v_i \sim \mathcal{N}(0, \sigma_u^2), \quad \mathbb{P}(\xi_i = 0) = q;\ \mathbb{P}(\xi_i = 1) = (1 - q), \tag{6.28}$$

and define q as the expected fraction of non-zero components of u. However, such models are extremely difficult to handle computationally.

6.2.2. The HBM Approach to Sparsity

In Section 3.3, we introduced hierarchical prior models as an alternative to the ℓ_p-based ones. Section 3.3.3 further showed that using the non-log-concave inverse gamma distribution (3.46) as a hyperprior leads to a class of heavy-tailed implicit priors on u, which we called product t_p priors:

$$p_{prior}(u) \propto \prod_i^h \left(1 + \frac{\left|D_i^T u\right|^p}{\nu\theta}\right)^{-\frac{\nu+1}{p}} \tag{6.29}$$

The parameter p is typically not varied but fixed by computational considerations: Choosing $p = 2$ is most common as the conditional updating of u in alternating algorithms (cf. Sections 4.1.11 and 4.2.5) can then be carried out by solving a linear system. Since θ is a scaling parameter with a similar meaning as λ for ℓ_p priors, the only parameter that can still influence the shape of the prior is ν. Its function is similar to the function of p in the ℓ_p based approach to model sparsity:

- There is a Gaussian limit: For $\nu \to \infty$, we can use that $\log(1+x) \approx x$ for $|x| \ll 1$:

$$\frac{\nu+1}{2} \sum_i^n \log\left(1 + \frac{u_i^2}{\nu\theta}\right) \approx \frac{\nu}{2} \sum_i^n \frac{u_i^2}{\nu\theta} = \frac{1}{2\theta}\|u\|_2^2 \tag{6.30}$$

  In the general case, the limit is given by $(p\theta)^{-1}\|u\|_p^p$. For finite values of ν, the above approximation holds only in a region around 0 (cf. Figure 3.12a) and the complete prior is never log-concave. One can easily compute that the region of convexity is characterized by $\|u\|_\infty < \sqrt{\nu\theta}$.

- There is a sparse limit: We can re-write the energy as

$$\begin{aligned}\frac{\nu+1}{2} \sum_i^n \log\left(1 + \frac{u_i^2}{\nu\theta}\right) &\overset{u}{\propto} \frac{\nu+1}{2} \sum_i^n \log\left(\nu\theta + u_i^2\right) \\ &= \frac{1}{2} \sum_i^n \nu \log\left(\nu\theta + u_i^2\right) + \frac{1}{2} \sum_i^n \log\left(\nu\theta + u_i^2\right).\end{aligned} \tag{6.31}$$

For $\nu \searrow 0$, the limit of the first summand is always 0. The limit of the second summand is given as

$$\lim_{\nu \searrow 0} \left(\frac{1}{2} \sum_i^n \log\left(\nu\theta + u_i^2\right) \right) = \begin{cases} -\infty & \text{if } u_i = 0 \text{ for any } i \\ \sum_i^n \log\left(|u_i|\right) & \text{else} \end{cases} \tag{6.32}$$

This limit is universal for all p. As (6.32) is not bounded from below, it is more intuitive to examine $\mathcal{J}_\varepsilon(u) = \sum \log(\varepsilon + |u_i|)$ for a small but finite $\varepsilon \approx \sqrt{\nu\theta}$. $\mathcal{J}_\varepsilon(u)$ sums up the *scales* of the components of u with respect to the reference scale ε. Thereby, it can be considered a measure of compressibility: If the components of u take their values on very different scales, u can be well-approximated by a sparse u_0. The corresponding prior samples in Figure 5.1 where generated using $\nu = 1$. Another intuition for the sparsifying properties of (6.32) is given by recognizing that

$$\sum_i^n \log\left(|u_i|\right) = \log\left(\prod_i^n |u_i| \right), \tag{6.33}$$

which means that the prior energy is given by the logarithm of the n-dim volume spanned by the components of u. By its multiplicative nature, it is more effectively minimized by collapsing single dimensions to zero than by shrinking all components isometrically.

Note that in the full HBM prior model parameterized by α and β, one has to choose $\alpha = \nu/2$, $\beta = \nu\theta$ to obtain the above limits for $p = 2$.

6.2.3. Comparison and Fusion

Table 6.1 compares the prior energies of the two approaches. Most notably, the type of sparsity induced differs: ℓ_p-based approaches lead to a binary type of sparsity by the non-differentiability in the tip of the prior density at zero. HBM-based approaches lead to a scale-based compressibility by the slow decay of the tails of the prior density combined with the multiplicative way in which the single components interact. This combination leads to the non-convex shape of its level sets. Note that the non-log-concavity of the prior alone is not generating compressible solutions: If we would define a multivariate Cauchy prior by

$$p_{prior}(u) \propto \left(1 + \frac{\left\| D^T u \right\|_2^2}{\theta} \right)^{-1} \tag{6.34}$$

instead of the product-type way we used in (3.29), the posterior would only be bi-modal, with a second mode in zero: Figure 6.1 shows the level sets of the resulting posterior.

Given these two different mechanisms that lead to compressibility/sparsity, it would be interesting to examine ℓ_1 hypermodels, which combine both properties. Computing full-MAP estimates via Algorithm 4.9 would require to compute a series of MAP estimates for changing ℓ_1 priors. While this is computationally feasible (e.g., by ADMM), our experience with ℓ_2 hypermodels suggests that a suitable initialization for the iteration is essential to avoid getting stuck in sub-optimal modes as discussed in Section 4.2.5. Initialization strategies based on full-CM estimates require the development of efficient sampling techniques. The ℓ_1 sampler developed in this thesis could be particularly well-suited to be used within the blocked Gibbs scheme also used for sampling ℓ_2 hypermodels (cf. Section 4.1.11): As the conditional posterior to be sampled from changes after every γ update, a burn-in phase for the u-sampler is necessary. In addition, the new sample for u has to get uncorrelated to the current sample u which is used as an initialization. Our studies in Section 5.1.2 suggest that the direct ℓ_1 sampler might be able to achieve those demands efficiently.

Table 6.1.: Comparison between the ℓ_p-based and the HBM-based approach to sparsity (for an ℓ_2 hypermodel).

feature	ℓ_p model	ℓ_2 hypermodel
$\mathcal{J}(u)$	$\|u\|_p^p$	$\frac{\nu+1}{2}\sum \log\left(1+\frac{u^2}{\nu\theta}\right)$
sparsifying parameter	$p>0$	$\nu>0$
quadratic limit	$p=2$	$\nu\to\infty$
sparse limit	$p\to 0$	$\nu\to 0$
limit functional	$\|u\|_0$	$\sum_i^n \log(\|u_i\|)$ if all $u_i\neq 0$, $-\infty$ else
solutions	sparse	compressible
differentiable	$p>1$	always
convex	everywhere for $p\geqslant 1$	$\|u\|_\infty < \sqrt{\nu\theta}$
homogeneous	yes	no

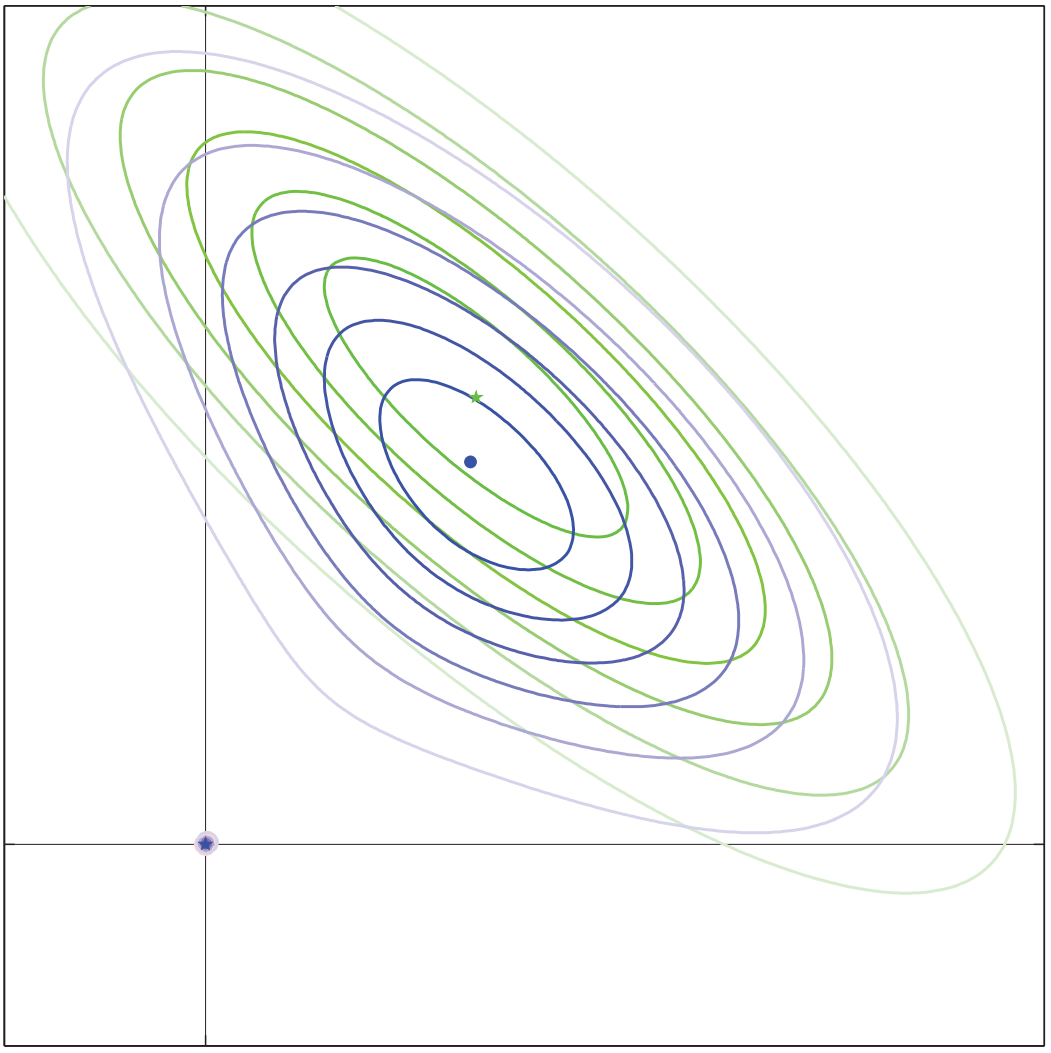


Figure 6.1.: Illustration of Bayesian inference with prior (6.34): Level sets of likelihood (green), prior (red), and resulting posterior (blue). The star markers indicate the corresponding maxima; the dot marker the CM estimate of the posterior.

7 Conclusion, Outlook and Perspectives

While each of the numerical studies included a separate discussion pointing to possible extensions, this chapter aims to summarize and reflect upon this thesis as a whole. In addition, it will point to possible future directions of research.

7.1. Bayesian Inversion as a General Framework for Biomedical Imaging

We demonstrated that Bayesian inversion can be applied to realistic, challenging imaging scenarios in two experimental data studies:

- Using appropriate MCMC algorithms, we carried out sample-based Bayesian inference in very high dimensional ($n > 10^6$) computed tomography scenarios.
- Hierarchical Bayesian modeling is a popular prior modeling paradigm that is usually applied in considerable lower-dimensional and less ill-posed statistical inference problems. We demonstrated that it can also be used to obtain physiologically plausible source reconstructions in the notoriously ill-posed EEG/MEG inverse problem.

Both studies had a "proof-of-concept" character: The aim was to produce reasonable, first results with a tolerable amount of computational effort. To explore the full potential of Bayesian inversion, it would be interesting to apply these techniques to a concrete, more specific imaging task, for which the uncertainty representation by the posterior

density can really add to the information given by a simple point estimate. Similarly interesting and relevant for many applications is the explicit incorporation of the various kinds of model uncertainties into the inversion (cf. Section 3.6.2).

7.2. MAP and CM Estimation

One main aim of the thesis was to shed new light on the "MAP or CM?" question, both from computational and theoretical perspectives. Our results were quite surprising in many ways: In certain situations, for instance in the 2D "Walnut-CT" scenario, MAP and CM estimates were almost identical even when using the TV prior. Despite the fact that one would typically not expect this to happen when using a non-Gaussian prior, also the theoretical results and computed examples in 1D suggested the opposite. The similarity of MAP and CM estimate when using a Besov prior is also not fully understood yet, especially since it increases when the impact of the non-Gaussian prior is increased (by λ). While examining these phenomena from a theoretical perspective is an important future direction of research, the closeness of MAP and CM estimates supports the practical relevance of the Bayesian approach in applications: If MAP and CM would always be as different as in the 1D "Boxcar" scenario using the TV prior and the CM estimate would always correspond to a very sub-optimal solution, the general use and relevance of posterior-based inference could not be justified. For instance, CStd estimates to characterize the spatial distribution of the posterior variance like those shown in Figures 5.26c and 5.26c would, as they characterize the spread of the posterior around the CM estimate, be of limited value if the MAP estimate would be chosen as the primary estimate of interest.

The theoretical results presented in Section 6.1 are of more fundamental nature: The rehabilitation of the MAP estimate as a proper Bayes estimator by the use of Bregman distances justifies its popularity in practical applications, resolves a number of otherwise converse results and observations, and disproves common misconceptions about the nature of MAP estimation. This opens a new perspective to relate variational regularization and Bayesian inference: The "MAP or CM?" question was always seen as the key question of such a comparison. However, while it might be an obvious question, it puts the focus on a direct comparison between point estimates and suggests that one should choose between one of the two approaches. The real strength of Bayesian approaches is to model and quantify uncertainty and information at all stages of the problem, *beyond* point estimates. In this direction, Bayesian techniques can very well *complement* variational approaches.

7.3. Prior Models

In the general studies on Bayesian inversion in Section 5.2, we examined various prior models to replace the conventional TV prior due to its lack of discretization invariance in 1D. In particular, we found that non-log-concave prior models are an interesting topic for future investigations, although this will require various methodical developments.
For the experimental data studies, we focused on the ℓ_1-based TV and Besov priors for CT and on ℓ_2 hypermodels with inverse gamma hyperpriors for EEG/MEG.
The TV estimates appeared visually more convincing than the Besov estimates with Haar wavelets, but as discussed in Section 5.3.4, several possible modifications of the Besov prior should be investigated in future studies. In Figure 5.23, we saw that CM estimates suffer less from the staircasing artifact than MAP estimates. Recently, also alternative regularization approaches that aim to reduce this artifact while preserving the desired properties if the TV functional have been proposed. Benning et al. (2013) give an overview of these *higher-order TV* methods. As they correspond to MAP estimates for alternative prior models, it would be interesting to examine them from a Bayesian perspective as well.
Concerning HBM priors, various future investigations and developments are possible:

- Potentially most interesting is the examination of the proposed ℓ_p hypermodels for $p \neq 2$, in particular for $p = 1$. While the methodical and algorithmical requirements for such an examination were developed in this thesis, a computational examination was not carried out yet.

- As discussed in Section 3.3.4, we chose a specific construction scheme different from the more commonly used Gaussian scale mixture models. It would be interesting to compare both prior models.

- We only considered the inverse gamma distribution as a hyperprior model in this thesis, although there are interesting alternatives (cf. Section 3.3.2).

- The fully-Bayesian inference techniques we employed in this thesis should be compared to semi- and variational Bayesian inference approaches (cf. Section 3.3.4).

Both ℓ_1-based and HBM-based prior models aim to encode sparsity as a-priori information. While we compared their theoretical properties in Section 6.2 and computed examples for the "Boxcar" scenario, a direct comparison for an application with experimental data was not carried out yet.

Prior Parameter Choice

A topic neglected in this thesis was the choice of the prior parameters λ, θ, α or β. We chose them by visual inspection, which is arguably neither convincing with respect to the Bayesian modeling paradigm (they should reflect our a-priori knowledge on the scales and shapes of the distributions), nor practical in most real applications.
Choosing them by our a-priori knowledge as suggested by the Bayesian philosophy has two flaws:

- Often, our prior model is only a surrogate for the real stochastic model we would prefer to describe the unknowns. This complicates the choice of its parameters. One example is the "Spots" scenario: $u^{\dagger,\infty}$ was generated by a simple stochastic model. While one could, in principle, derive a pdf for u from this and use it as a prior, this approach would lead to an intractable posterior. Using an ℓ_1 prior as a surrogate is clearly not justified by its statistical properties (cf. the random sample in Figure 5.1c) and therefore, choosing λ such that it reflects any concrete a-priori knowledge on the original image is difficult.

- Even if our prior model would reflect an accurate stochastic description of the unknowns, using it might result in undesired solutions: The ill-posedness of the inverse problem usually requires to choose the prior parameters more conservative than our a-priori knowledge would suggest.

In such a situation, the Bayesian philosophy suggests to consider λ (or the corresponding parameters of the other prior models) as a hyperparameter to be estimated from the data as well. This leads to a hierarchical prior model and potentially to a non-log-concave posterior.
In the non-Bayesian literature, several *parameter choice rules* $\lambda(\varepsilon, f)$ were proposed, either inspired by analytical, statistical or heuristical considerations: See Engl et al. (1996), Kaipio and Somersalo (2005) for a general overview and Almeida and Figueiredo (2013), Deledalle et al. (2012), Eldar (2009), Giryes et al. (2011), Kolehmainen et al. (2012) for concrete procedures.
A heuristic parameter choice rule for Besov and TV priors in 2D could actually be formulated and examined by using the surprising phenomenon that MAP and CM estimate are very similar if λ is large enough, and start to develop different artifacts if it is too small (cf. Figure 5.18): One can choose the smallest λ such that the relative difference between MAP and CM estimate is below some threshold.
A important contribution would be to establish efficient parameter choice rules for the HBM parameter (α,β); a problem which is more or less completely open up to now.

7.4. Bayesian Computation

Another main objective of this thesis was to demonstrate that sample-based Bayesian inversion is feasible in high dimensions if suitable computational tools are available: We developed and examined fast MCMC sampling techniques for various prior distributions and applied them to various imaging scenarios, including studies with experimental. The dimensions of u in the computed examples were often chosen to be very high (up to $n = 1\,048\,576$) just to demonstrate the capabilities of the new algorithms. To the best of our knowledge, $n > 10^6$ is far beyond the dimensions realized for sample-based Bayesian inference in similar imaging scenarios.

While new and fast sampling techniques were established with this thesis, not all of them were evaluated carefully enough and their full potential for practical applications is yet to be explored: Sampling the posteriors arising from the use of ℓ_p hypermodels for $p \neq 2$ as suggested in the previous section will, for instance, require to combine the slice-within-Gibbs sampling developed in Section 4.1.10 with the blocked Gibbs sampling for HBM presented in Section 4.1.11, and is a very interesting future direction of research.

While the algorithms developed here already enabled us to perform investigations that were not possible before, one should not forget that they are conceptually still extremely simple: We only perform random scan SC Gibbs sampling (the challenging part is to derive and implement its sub-steps). Therefore, major improvements could be realized by implementing more sophisticated variants of the SC Gibbs scheme. A promising future direction would be to develop *adaptive Gibbs sampling* for our scenarios: In the RSG sampler, the new component to update is randomly chosen by a *selection probability* $p_i > 0$, $\sum p_i = 1$. In our studies, we fixed it as $p_i = 1/n$; thereby, the component was chosen uniformly at random. This may be extremely sub-optimal: In sparse imaging scenarios such as the "Spots" scenario, using a sparse prior will result in most of the components u_i having little variability and low correlations while only a few show large fluctuations and significant correlations (for the "spots" scenario, an image of CStd essentially looks like the CM estimate shown in Figure 5.8b). For the Gibbs sampler to converge fast, it would be advantageous to update the few components with a large variance more often than the others. As we typically do not know them in advance, we would adapt the selection probabilities used at step $i + 1$ on the fly, i.e., based on the chain history $\{u^j\}_{j=1}^i$. Essentially, we would take advantage of the inherent sparsity of the problem to let the algorithm's performance depend on the sparsity level k rather than on n. The ideas developed in Latuszynski et al. (2013) could be a starting point for such developments.

Another significant speed-up can be achieved by *parallel computing*: The most simple

and straight forward way is to run several MCMC chains independently from each other, each on a single CPU core, and to pool the samples afterwards. This does not require any sophisticated implementation and can result in the optimal linear speed-up if the chains mix is fast (i.e., short burn-in and integrated autocorrelation times). Fortunately, our results suggest that this is the case for the samplers developed in this thesis. In fact, many of the computations were already performed based on multiple, parallel chains, but we did not discuss this detail in the corresponding sections. More sophisticated parallelization techniques rely on interactions between the parallel chains in order to increase the mixing time (see Liu 2008, for further details). In particular, they try to avoid that the main chain gets stuck in a local mode of the posterior. Their use for multimodal posteriors resulting from non-log-concave priors should be examined in future studies.

Finally, multigrid strategies as discussed in Section 4.1.5 could be particularly advantageous to avoid getting trapped in local modes of multimodal posteriors.

The surprising results obtained by using simulated annealing with Gibbs instead of MH samplers have to be investigated more thoroughly: They need to be confirmed for more imaging applications and other prior models. As SA was originally designed for non-log-concave distributions, the application to HBM would be extremely interesting.

We investigated over-relaxation as one example of optimization techniques transferred to sampling schemes, and found that it can significantly enhance their statistical, and often also their computational efficiency. This motivates the search for other transferable optimization concepts.

The inverse problems examined in this thesis are well-modeled by a linear forward operator and additive Gaussian noise. For several other inverse problems, this is not the case. Therefore, an important future development would be to extend our results and studies to non-linear inverse problems and other noise models, for instance, to Poisson noise (cf. Section 3.1).

A
Appendix

A.1. Subdifferentials and Bregman Distances

Subdifferential and Optimality

In this section, we summarize some basic concepts of convex analysis on $\mathbb{R}^n$ that will be needed throughout the thesis. A concise introduction tailored to convex, variational regularization of inverse problems in the general (infinite dimensional) case can be found in Chapters 2 and 3 of Benning (2011). A more general presentation of convex analysis and optimization is given in Boyd and Vandenberghe (2004).

Definition A.1. For a proper, convex functional $\mathcal{J} : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$, the *subdifferential* $\partial\mathcal{J}(u)$ at u is defined as

$$\partial\mathcal{J}(u) := \{p \in R^n \,|\, \mathcal{J}(v) \geqslant \mathcal{J}(u) + \langle p, v - u\rangle \,,\, \forall v \in \mathbb{R}^n\} \,. \tag{A.1}$$

The subdifferential is always a non-empty, convex and compact set and an element $p \in \partial\mathcal{J}(u)$ is called a *subgradient* of $\mathcal{J}$ in u. If $\mathcal{J}$ is differentiable in u, $\partial\mathcal{J}(u) = \{\mathcal{J}'(u)\}$. Thereby, subdifferentiability extends (Fréchet-)differentiability for the important class of convex functionals. In 1D, the subgradient (or *subderivative*) has a simple illustrative meaning: $\mathcal{J}(u) + p(v - u)$ describes a line through $(u, \mathcal{J}(u))$ with slope p. The set of all slopes p such that this line is either touching or below the graph of $\mathcal{J}(u)$ is the subderivative $\partial\mathcal{J}(u)$. It is a non-empty, closed interval $[p_-, p_+]$, where

$$p_- = \lim_{h \searrow 0} \frac{\mathcal{J}(u) - \mathcal{J}(u-h)}{h}, \qquad p_+ = \lim_{h \searrow 0} \frac{\mathcal{J}(u+h) - \mathcal{J}(u)}{h}. \tag{A.2}$$

Both limits exist and fulfill $p_- \leqslant p_+$. Apparently, if the subderivative contains only one element, i.e., $p_- = p_+$, then $\mathcal{J}$ is differentiable at u and $\mathcal{J}'(u) = p_- = p_+$. A classical example where subdifferentiability extends the normal differentiability is given by the absolute value function $\mathcal{J}(u) = |u|$:

$$\partial|u| = \begin{cases} 1 & \text{for } u > 0 \\ [-1, 1] & \text{for } u = 0 \\ -1 & \text{for } u < 0 \end{cases} \tag{A.3}$$

The minima of such convex functionals $\mathcal{J}(u)$ cannot be characterized using normal derivatives but using subgradients:

Theorem A.1. A point $u \in R^n$ is a minimum of a proper, convex functional $\mathcal{J} : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$ if and only if $0 \in \partial\mathcal{J}(u)$.

The proof of this *optimality condition* is simple and instructive: If $0 \in \partial\mathcal{J}(u)$, we have that

$$0 = \langle 0, v - u\rangle \leqslant \mathcal{J}(v) - \mathcal{J}(u) \qquad \forall v \in \mathbb{R}^n, \tag{A.4}$$

and thereby, u is a global minimizer of $\mathcal{J}$. If $0 \notin \partial\mathcal{J}(u)$, there must be at least one $v \in R^n$ such that

$$\mathcal{J}(v) > \mathcal{J}(u) + \langle 0, v - u\rangle = \mathcal{J}(u), \tag{A.5}$$

and thereby, u cannot be a global minimizer of $\mathcal{J}$. The uniqueness of the minimizer can only be guaranteed if $\mathcal{J}(u)$ is *strictly convex*. To apply these concepts to MAP estimation (e.g., for ℓ_1 priors), we further note that

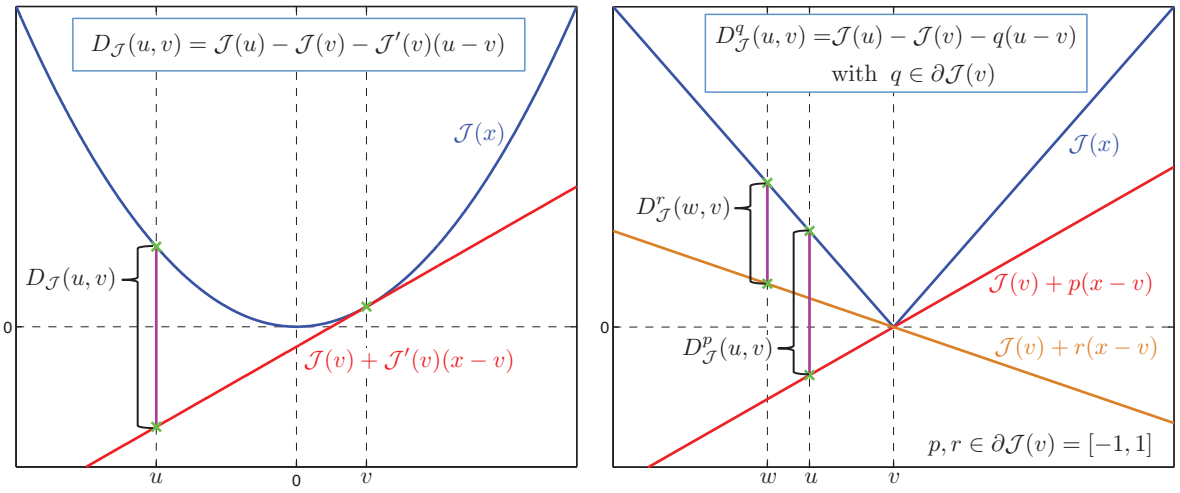
$$\partial\left(\frac{1}{2}\|f - Au\|_2^2 + \lambda\mathcal{J}(u)\right) = A^T(Au - f) + \lambda\partial\mathcal{J}(u). \tag{A.6}$$

Bregman Distances

Definition A.2. For a proper, convex functional $\mathcal{J} : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$, the *Bregman distance* $D^p_{\mathcal{J}}(u, v)$ between $u, v \in \mathbb{R}^n$ for a subgradient $p \in \partial\mathcal{J}(v)$ is defined as

$$D^p_{\mathcal{J}}(u, v) = \mathcal{J}(u) - \mathcal{J}(v) - \langle p, u - v\rangle, \qquad p \in \partial\mathcal{J}(v) \tag{A.7}$$

We will often simplify the notation to $D_{\mathcal{J}}(u, v)$, and use $D^p_{\mathcal{J}}(u, v)$ only if we want to stress the potential ambiguity arsing from the set-valued character of the subdifferential. Table A.1 lists the Bregman distances induced by some Gibbs energies $\mathcal{J}(u)$. Figure A.1 gives an illustration: Basically, $D_{\mathcal{J}}(u, v)$ measures the difference between $\mathcal{J}$ and its linearization in u at another point v. Further, $D_{\mathcal{J}}(u, v) \geqslant 0$ and for strictly convex



(a) $\mathcal{J}(x) = x^2$ **(b)** $\mathcal{J}(x) = |x|$

Figure A.1.: Illustrative explanation of the Bregman distance.

Table A.1.: Bregman distances induced by some Gibbs energies $\mathcal{J}(u)$ commonly used for prior modeling. Note that if $\mathcal{J}$ is separable, so is $D_{\mathcal{J}}(u,v)$. In these cases, the scalar expression is listed, only.

$\mathcal{J}(u)$	$\mathrm{dom}(\mathcal{J})$	$D_{\mathcal{J}}(u,v)$
$\frac{1}{2}\|Lu\|_2^2$	$\mathbb{R}^n$	$\frac{1}{2}\|L(u-v)\|_2^2$
$\|u\|^p$, $(1 < p < \infty)$	$\mathbb{R}$	$\|u\|^p - p\,u\,\mathrm{sign}(v)\|v\|^{p-1} + (p-1)\|v\|^p$
$\|u\|$	$\mathbb{R}$	$(\mathrm{sign}(u) - \mathrm{sign}(v))\,u$
$u \log u - u$	$\mathbb{R}_{\geqslant 0}$	$u \log \frac{u}{v} + v - u$ (Kullback-Leibler divergence)

$\mathcal{J}(u)$, $D_{\mathcal{J}}(u,v) = 0$ implies $u = v$. However, the Bregman distance is not a distance in the usual mathematical sense (i.e., a metric) as it is, in general, neither symmetric nor satisfies the triangle inequality. We will further use that $D_{\mathcal{J}}(u,v)$ is convex in u. Bregman distances have become an important tool in variational regularization, e.g., to derive error estimates and convergence rates (Benning 2011, Burger et al. 2013, Burger and Osher 2004, Burger et al. 2007), to enhance inverse methods by Bregman iterations (Burger et al. 2007, Moeller 2012) or to develop optimization schemes like the Split-Bregman algorithm (Goldstein and Osher 2009) which is closely related to the ADMM algorithm we use for computing MAP estimates (cf. "Notes and Comments" in Section 4.2.2).

A.2. Application Specific Implementation Details

In this section, we discuss all details that are required to implement the algorithms described in this thesis in MATLAB for the each specific imaging scenario separately. Details about the functions used can be found in MATLAB's documentation[1]. While tailored to the algorithms used in this thesis, the techniques presented can easily be adopted to implement many other algorithms for the specific scenarios. In particular, those used for SC Gibbs sampling can be used for implementing fast *greedy* algorithms as used in compressed sensing (FOUCART AND RAUHUT 2013). Although this section was moved to the appendix to keep the main presentation concise, parts of it contain the most challenging and tedious works for this thesis.

Boxcar

For generating the measurement data, (2.1) is directly implemented using `quad.m` on a continuous representation of $u^{\dagger,\infty}$.

ADMM A and D^T can be explicitly constructed as (sparse) matrices by (2.2) and (3.18). Then, the least squares system resulting from (4.79) can also be formed explicitly and solved by the `backslash` operator.

MH We only need to implement the matrix-vector products Au and $D^T u$. While we simply use the explicitly constructed matrix A to compute Au, we implement the application of the difference operator $D^T u$ by `diff.m`.

SC Gibbs D^T has full rank $h = n - 1$. Therefore, we can find $v_1, \dots, v_{n-1}$ such that $D^T v_i = e_i$ and v_n such that $D^T v_n = 0$. It is easy to see that complementing the *step functions* $(v_i)_j = \mathbb{1}_{\{j>i\}}$, $i = 1, \dots, n-1$, by the constant function $v_n = 1$ fulfills these requirements. If we reorder them and define $V := [v_n, v_1, \dots, v_{n-1}]$, we can write V as

$$V_{(i,j)} = \begin{cases} 1 & \text{if} \quad i \geqslant j \\ 0 & \text{else} \end{cases} \tag{A.8}$$

Vu can be implemented as `cumsum(u,1)` and V^{-1} is given by `[u(1);diff(u,1,1)]`. In a similar way, $\Psi = AV$ can be computed explicitly by applying the `cumsum` function to A and some reordering of the result. Using this, we can pre-compute $a := \frac{1}{2}\|\Psi_i\|_2^2$. For computing $b := \Psi_i^T \varphi(i) = \Psi_i^T f - (\Psi_i^T \Psi_{-i})\xi_{-i}$, we pre-compute $\Psi_i^T f$ and $\|\psi_i\|_2^2$ for all i

[1] online available at http://www.mathworks.co.uk/help/matlab/

and build the $n \times n$ matrix $\Phi := \Psi^t \Psi$. Then, computing $(\Psi_i^T \Psi_{-i})\xi_{-i}$ can be performed by using

$$(\Psi_i^t \, \Psi_{[-i]}) \, \xi_{[-i]} = \xi^t \Phi_{(\cdot,i)} - \xi_i \|\psi_i\|_2^2, \tag{A.9}$$

which involves a scalar product of dimension n as the most extensive operation.

Point source reconstruction

For generating the measurement data, (2.3) is implemented by applying `imfilter.m` to a discretization of $u^{\dagger,\infty}$ on a spatial grid with a 4 times higher resolution than the one used in the inversion.

ADMM and MH As the Gaussian kernel is symmetric, $A = A^T$. We will implement the matrix-vector multiplication Au (and $A^T u = Au$) using the *convolution theorem*

$$\mathcal{F}[g * u] = \mathcal{F}[g] \cdot \mathcal{F}[g], \tag{A.10}$$

and *fast Fourier transforms* (*ffts*). For this, we pre-compute $\mathcal{F}[g]$ by applying `fft2.m` to a discretization of g, `g_dis`, on the computational grid. Then, Au is basically given as `ifft2(fft2(u) .* g_ft))`, i.e., a 2D-fft of u followed by a point-wise multiplication and an inverse 2D-fft. For achieving a higher accuracy, images and kernels are padded with zeros before applying the ffts.
As we only use $D^T = I_n$ in this scenario, using the procedures to compute Au (and $A^T u$) are all we need to implement MH in straight forward way. The details of implementing ADMM are explained in Section A.5.

SC Gibbs Since $V = I_n$, $\Psi = A$. A_i is simply the kernel function g, centered at the i-th pixel. We again pre-compute $\Psi_i^T f$ and $\|\psi_i\|_2^2$. For computing b, we can then derive that $(\Psi_i^T \Psi_{-i})\xi_{-i}$ is given by

$$(\Psi_i^t \, \Psi_{[-i]}) \, \xi_{[-i]} = \left[\left(A^T A\right) \cdot \xi\right]_i - \xi_i \|A_i\|_2^2. \tag{A.11}$$

Here, $A^T A$ is a twofold Gaussian convolution which could be replaced by a single Gaussian convolution with a larger variance, but we only need its discrete kernel which is given by `imfilter(g_dis,g_dis,'conv')`. The value of the i-th pixel of a discrete convolution applied to an image u can be computed by the scalar product of the discrete convolution kernel centered at this pixel with the image. As such, A.11 can be implemented in a fast, direct way, exploiting that the spatial width of the double Gaussian kernel is still considerable smaller than the image size.

Computed Tomography

MATLAB comes with a couple of functions to simulate CT, such as `radon.m`, `iradon.m`, `fanbeam.m` and `ifanbeam.m`. However, we eventually decided to use the algorithms sketched in Section 2.3.2 only: They rely on very basic geometrical operations like computing the crossings of two lines which can be implemented in a fast and robust way using `.mex` files in MATLAB. In addition, they are extremely easy to parallelize as the computations for the different angles are completely independent from each other. The geometric nature of the operations involved and their massive parallelizability suggest that an implementation on a GPU might be of orders faster than the serial CPU implementation developed and used for the studies in this thesis.
For generating the measurement data, $u^{\dagger,\infty}$ is discretized on a spatial grid that is about 3 finer than the grid used in the inversion. For this $u^\dagger$, our algorithms are used to compute $g^{(i,j)} = P\mathcal{A}v^{(i,j)}$ for all rectangles $v_{(i,j)}, i, j = 1, \dots, 2^N$ representing non-zero pixels, i.e., $u^\dagger_{(i,j)} \neq 0$. Here, $v_{(i,j)}$ is the indicator function of the square representing the (i, j)-th pixel. Then, we compute f by

$$f = \sum_{(i,j)} u^\dagger_{(i,j)} g^{(i,j)} \tag{A.12}$$

ADMM and MH We explicitly construct the matrix A by computing $g^{(i,j)} = P\mathcal{A}v_{(i,j)}$ for all pixel as above and stacking the sinograms $g^{(i,j)}$ into column vectors. For the isotropic 2D TV prior, (3.22), we need to implement the difference operators in x and y direction and their transposes. The details of this can be found in Section A.5. For the Besov prior, (3.25), using Haar wavelets, we need to construct the 2D multiresolution analysis and compute the weights ω_i. The details of this construction can be found in Section 2.1 of HÄMÄLÄINEN et al. (2013). Here, we only note that for $p = 1$ and a 2D image, $\omega_i = 1 \,\forall\, i$ and that the wavelets can be indexed by (j, l, k_1, k_2), where $j = 0, 1, 2, \dots, N$ determines the scale $(n = 2^N \cdot 2^N)$, $l \in \{1, 2, 3\}$ the shape, and (k_1, k_2) the location of the wavelet. In Figure A.3, the first sixteen Haarwavelets are shown to illustrate the construction principle. Using multiresolution wavelet constructions is computationally attractive because decomposition and analysis (i.e., Vc, $V^T u$) can be performed by *fast wavelet transforms*.

SC Gibbs Using a TV prior, we also construct the matrix A as above and compute a and b as in "Boxcar" scenario. The parameters of the prior part of the SC density, (4.42), are in principle simple to compute. However, the concrete implementation is tedious due to the correct handling of the boundary pixels.
As the Besov prior is an ℓ_1 prior on the coefficients of a basis, we can simply use $V = D$,

i.e., we transform the posterior into the wavelet basis. $\Psi_i = (AV)_i$ is the integrated Radon transform of the i-th wavelet: $\Psi_{(j,l,k_1,k_2)} = P\mathcal{A}v_{(j,l,k_1,k_2)}$. As Haar wavelets can be described as the sum of the indicator functions of one to four rectangles, we can, again, use our algorithms to compute $\Psi_{(j,l,k_1,k_2)}$ as well as the scalar product $\Psi^T_{(j,l,k_1,k_2)}g$ for a given tupel (j,l,k_1,k_2) and $g \in \mathbb{R}^m$. Figure A.4 shows $\Psi_{(j,l,k_1,k_2)}$ for the first sixteen Haar wavelets. We can now compute b by using

$$b = \Psi^T_{(j,l,k_1,k_2)}f - \Psi^T_{(j,l,k_1,k_2)}(\Psi\xi) + \xi_{(j,l,k_1,k_2)}\left\|\Psi_{(j,l,k_1,k_2)}\right\|_2^2 \tag{A.13}$$

in the following way:

- We again pre-compute $\Psi^T_{(j,l,k_1,k_2)}f$ and $\left\|\Psi_{(j,l,k_1,k_2)}\right\|_2^2$ for all (j,l,k_1,k_2). Then, we store the measurement that the current state ξ would cause as f_ξ and initialize it by Au^0. In principle, f_ξ is given as $\Psi\xi$, and can be directly computed at any time but this computation is too expensive to be performed at every SC update.
- For a given wavelet coefficient $\xi_{(j,l,k_1,k_2)}$ that is to be updated, we construct $\Psi_{(j,l,k_1,k_2)}$ and compute the scalar product $\Psi^T_{(j,l,k_1,k_2)}f_\xi$ to update b by the above formula (note that $\Psi^T_{(j,l,k_1,k_2)}(f - \Psi\xi)$ is just a projection of $\Psi_{(j,l,k_1,k_2)}$ onto the current residual of $f_\xi = \Psi\xi$). With the constructed $\Psi_{(j,l,k_1,k_2)}$ and the change, $\delta_{(j,l,k_1,k_2)}$, in $\xi_{(j,l,k_1,k_2)}$ caused by the sampling step, we can then update $f_\xi = f_\xi + \delta_{(j,l,k_1,k_2)}\Psi_{(j,l,k_1,k_2)}$.
- While this iterative updating of f_ξ is fast, inaccuracies can accumulate over time, leading to a misfit between f_ξ and $\Psi\xi$. Therefore, we reset f_ξ to the exact $\Psi\xi$ every n steps.

Discussion Using MATLAB's `radon.m` function turned out to be problematic for the following reasons:

- `iradon.m` is not the exact adjoint to `radon.m`, at least a scaling needs to be added to fix this, but this scaling is not explicitly given.
- It further involves an offset of its center with respect to the coordinate system which complicates comparisons between different n.
- It only implements the Radon transform, the integration over the sensor pixels needs to be done subsequently (and the adjoint of the integration needs to be computed as well).
- By default, the number of points on the sensor grid for which the Radon transform is computed is fixed and depends on n. This is a tedious complication when results for different n should be compared and m_s should remain fixed.

- Implementing the SC Gibbs sampler for the Besov prior with `radon.m` would be too slow anyhow. Given this, using the same implementation of the forward operator in all algorithms is advantageous, for instance for comparing MAP and CM estimates.

The procedure described above, i.e., using the algorithms described in Section 2.3.2 to construct A explicitly turned out to be both feasible and advantageous:

- A is of size $(m_s m_\theta) \times n$ but is very sparse as the columns are the sinograms of single pixels. For the "Phantom-CT" scenario ($m_s = 500$, $m_\theta = 45$) with $n = 1024 \times 1024$, it has only 0.3% non-zeros and its size is 1GB.
- The setup time for this configuration takes about 3 minutes (using `radon.m`, the setup would take several hours). However, as discussed above, parallelization could significantly reduce this time.
- Both the center of the coordinate system and the sensor size can be chosen freely.
- It is fully compatible with the implementation used by the Gibbs samplers.
- Applying Au by the matrix is usually way faster than calling `radon.m`. For the example above, it is about 7 times faster (although one has to bear in mind that the output of `radon.m` is of size $m_s = 1453$ in sensor space and not of size $m_s = 500$.

While we derived the algorithms and implementation for 2D only, the basic operations are standard problems of computer graphics. Therefore, an extension to cone beam scanning geometry for 3D reconstruction should not be a principled problem: A voxel would be projected onto a 2D surface by a diverging bundle of rays.

EMEG

The challenging part in EMEG is the setup of the lead-field matrix A as described in Section 5.4. Once assembled, its small size ($m \ll n$) allows for an easy implementation of all algorithms. In particular, the least-squares problems that arise in ADMM or HBM sampling or optimization can be solved fast and explicitly. Details are given in Lucka (2011).

A.3. Implementation Details of the ℓ_1 Sampler

In this section, we discuss how to implement formulas (4.46), (4.47) and (4.48). The complementary error function and its inverse are difficult to handle numerically because there are no identities that allow to rescale or shift their evaluation to other intervals. For the applications we address, problems due to limited precision occur if formulas

(4.46), (4.47) and (4.48) are implemented directly (formula (4.46) is only required for applying ordered overrelaxation). Dependent on the signs of α_+ and α_-, we use different alternative formulas that allow for a stable numerical evaluation. Additionally, we express $\mathrm{erfc}(x)$ in terms of the scaled complementary error function $\mathrm{erfcx}(x) = \exp(x^2)\mathrm{erfc}(x)$, which decays slower for $x \to +\infty$. As the corresponding transformations are elementary but lengthy to write down, we only list the results here. Because $c \geqslant 0$, not both α_+ and α_- can be negative, which leaves three different cases to examine:

$\underline{\alpha_+ > 0,\ \alpha_- > 0:}$ Let $\gamma_{++} := \mathrm{erfcx}\,(\alpha_+) + \mathrm{erfcx}\,(\alpha_-)$. Then, the parts of (4.46) are given by:

$$y < 0: \qquad \exp\left(-ay^2 + 2\sqrt{a}y\alpha_+\right)\mathrm{erfcx}\left(-\sqrt{a}y + \alpha_+\right)/\gamma_{++} \tag{A.14}$$

$$y > 0: 1 - \exp\left(-ay^2 - 2\sqrt{a}y\alpha_-\right)\mathrm{erfcx}\left(\ \sqrt{a}y + \alpha_-\right)/\gamma_{++} \tag{A.15}$$

The arguments of erfcinv in (4.47) and (4.48) are given by:

$$\text{In (4.47)}: \qquad r\,\exp\left(-\alpha_+^2\right)\,\gamma_{++} \tag{A.16}$$

$$\text{In (4.48)}: \quad (1-r)\exp\left(-\alpha_-^2\right)\,\gamma_{++} \tag{A.17}$$

$\underline{\alpha_+ < 0,\ \alpha_- > 0:}$ Since erfcx increases very fast for $x \to -\infty$, one has to use the identity $\mathrm{erfcx}(-x) = 2\,\exp(x^2) - \mathrm{erfcx}(x)$. Let $\gamma_{-+} := \mathrm{erfcx}\,(-\alpha_+) - \mathrm{erfcx}\,(\alpha_-)$. Then, the formulas to implement the parts of (4.46) are given by:

$$\begin{matrix} y < 0 \\ -\sqrt{a}y + \alpha_+ > 0 \end{matrix} : \quad \frac{\exp\left(-\left(\sqrt{a}y - \alpha_+\right)^2\right)\mathrm{erfcx}(-\sqrt{a}y + \alpha_+)}{2 - \exp\left(-\alpha_+^2\right)\gamma_{-+}} \tag{A.18}$$

$$\begin{matrix} y < 0 \\ -\sqrt{a}y + \alpha_+ < 0 \end{matrix} : \quad \frac{2 - \exp\left(-\left(\sqrt{a}y - \alpha_+\right)^2\right)\mathrm{erfcx}(\sqrt{a}y - \alpha_+)}{2 - \exp\left(-\alpha_+^2\right)\gamma_{-+}} \tag{A.19}$$

$$\begin{matrix} y > 0 \\ \sqrt{a}y + \alpha_- > 0 \end{matrix} : \quad 1 - \frac{\exp\left(-\left(\sqrt{a}y - \alpha_-\right)^2\right)\mathrm{erfcx}(\sqrt{a}y + \alpha_-)}{2\exp\left(\frac{bc}{a}\right) - \exp\left(-\alpha_-^2\right)\gamma_{-+}} \tag{A.20}$$

$$\begin{matrix} y > 0 \\ \sqrt{a}y + \alpha_- < 0 \end{matrix} : \quad 1 - \frac{2 - \exp\left(-\left(\sqrt{a}y - \alpha_-\right)^2\right)\mathrm{erfcx}(-\sqrt{a}y - \alpha_-)}{2\exp\left(\frac{bc}{a}\right) - \exp\left(-\alpha_-^2\right)\gamma_{-+}} \tag{A.21}$$

The arguments of erfcinv in (4.47) and (4.48) are given by:

$$\text{In (4.47)}: \qquad r\quad \left(2 - \exp(-\alpha_+^2)\gamma_{-+}\right) \tag{A.22}$$

$$\text{In (4.48)}: \quad (1-r)\left(2\exp\left(\frac{bc}{a}\right) - \exp(-\alpha_-^2)\gamma_{-+}\right) \tag{A.23}$$

$\underline{\alpha_+ > 0,\, \alpha_- < 0:}$ Let $\gamma_{+-} := \operatorname{erfcx}(\alpha_+) - \operatorname{erfcx}(-\alpha_-)$. Then, the parts of (4.46) are given by:

$$y < 0: \quad \frac{\exp\left(-\left(\sqrt{a}y - \alpha_+\right)^2\right)\operatorname{erfcx}(-\sqrt{a}y + \alpha_+)}{2\exp\left(-\frac{bc}{a}\right) + \exp\left(-\alpha_+^2\right)\gamma_{+-}} \tag{A.24}$$

$$\begin{matrix} y > 0 \\ \sqrt{a}y + \alpha_- > 0 \end{matrix}: \quad 1 - \frac{\exp\left(-\left(\sqrt{a}y + \alpha_-\right)^2\right)\operatorname{erfcx}(\sqrt{a}y + \alpha_-)}{2 + \exp\left(-\alpha_-^2\right)\gamma_{+-}} \tag{A.25}$$

$$\begin{matrix} y > 0 \\ \sqrt{a}y + \alpha_- < 0 \end{matrix}: \quad 1 - \frac{2 - \exp\left(-\left(\sqrt{a}y + \alpha_-\right)^2\right)\operatorname{erfcx}(-\sqrt{a}y - \alpha_-)}{2 + \exp\left(-\alpha_-^2\right)\gamma_{+-}} \tag{A.26}$$

The arguments of erfcinv in (4.47) and (4.48) are given by:

$$\text{In (4.47)}: \qquad r\left(2\exp\left(\frac{-bc}{a}\right) + \exp(-\alpha_+^2)\gamma_{+-}\right) \tag{A.27}$$

$$\text{In (4.48)}: \quad (1-r)\left(2 + \exp(-\alpha_-^2)\gamma_{+-}\right) \tag{A.28}$$

Using the above expressions directly can still lead to stability issues, because very large numbers are often multiplied with very small numbers. It is preferable to compute the logarithms of the expressions, first. For this, let $x > 0$, $(x + y) > 0$, then:

$$\log(x+y) = \log(x) + \log(1 + \operatorname{sign}(y)\exp(\log(|y|) - \log(x)) \tag{A.29}$$

Using this identity we can compute the logarithms of expressions (A.14)-(A.28). We note that $\operatorname{sign}(\pm\operatorname{erfcx}(\cdot)) = \pm 1$.

$$\log((\text{A.14})) = -ay^2 + 2\sqrt{a}y\alpha_+ + \log\left(\operatorname{erfcx}(-\sqrt{a}y + \alpha_+)\right) - \log(\gamma_{++}) \tag{A.30}$$

$$\log(1 - (\text{A.15})) = -ay^2 - 2\sqrt{a}y\alpha_- + \log\left(\operatorname{erfcx}(\sqrt{a}y + \alpha_-)\right) - \log(\gamma_{++}) \tag{A.31}$$

$$\log((\text{A.16})) = \log(r) - \alpha_+^2 + \log(\gamma_{++}) \tag{A.32}$$

$$\log((\text{A.17})) = \log(1-r) - \alpha_-^2 + \log(\gamma_{++}) \tag{A.33}$$

$$\begin{aligned}\log((\text{A.18})) = &-\left(-\sqrt{a}y + \alpha_+\right)^2 + \log\left(\operatorname{erfcx}(-\sqrt{a}y + \alpha_+)\right) - \log(2) \\ &- \log\left(1 - \operatorname{sign}(\gamma_{-+})\exp\left(-\alpha_+^2 + \log(|\gamma_{-+}|) - \log(2)\right)\right)\end{aligned} \tag{A.34}$$

$$\begin{aligned}\log((\text{A.18})) = &\log\left(1 - \exp\left(\log\left(\operatorname{erfcx}(\sqrt{a}y - \alpha_+)\right) - \log(2) - \left(\sqrt{a}y - \alpha_+\right)^2\right)\right) \\ &- \log\left(1 - \operatorname{sign}(\gamma_{-+})\exp\left(-\alpha_+^2 + \log(|\gamma_{-+}|) - \log(2)\right)\right)\end{aligned} \tag{A.35}$$

$$\begin{aligned}\log(1 - (\text{A.20})) = &-\left(\sqrt{a}y + \alpha_-\right)^2 + \log\left(\operatorname{erfcx}(\sqrt{a}y + \alpha_-)\right) - \log(2) - \frac{bc}{a} \\ &- \log\left(1 - \operatorname{sign}(\gamma_{-+})\exp\left(-\alpha_-^2 + \log(|\gamma_{-+}|) - \log(2) - \frac{bc}{a}\right)\right)\end{aligned} \tag{A.36}$$

$$\log\left(1-(\text{A.21})\right) = \log\left(1-\exp\left(\log\left(\operatorname{erfcx}(-\sqrt{a}y-\alpha_-)\right)-\log(2)-\left(\sqrt{a}y+\alpha_-\right)^2\right)\right) - \frac{bc}{a} - \log\left(1-\operatorname{sign}(\gamma_{-+})\exp\left(-\alpha_-^2+\log(|\gamma_{-+}|)-\log(2)-\frac{bc}{a}\right)\right) \tag{A.37}$$

$$\log\left((\text{A.22})\right) = \log(r)+\log(2) + \log\left(1-\operatorname{sign}(\gamma_{-+})\exp\left(-\alpha_+^2+\log(|\gamma_{-+}|)-\log(2)\right)\right) \tag{A.38}$$

$$\log\left((\text{A.23})\right) = \log(1-r)+\log(2)+\frac{bc}{a} + \log\left(1-\operatorname{sign}(\gamma_{-+})\exp\left(-\alpha_-^2+\log(|\gamma_{-+}|)-\log(2)-\frac{bc}{a}\right)\right) \tag{A.39}$$

$$\log\left((\text{A.24})\right) = -\left(-\sqrt{a}y+\alpha_+\right)^2+\log\left(\operatorname{erfcx}(-\sqrt{a}y+\alpha_+)\right)-\log(2)+\frac{bc}{a} - \log\left(1+\operatorname{sign}(\gamma_{+-})\exp\left(-\alpha_+^2+\log(|\gamma_{+-}|)-\log(2)+\frac{bc}{a}\right)\right) \tag{A.40}$$

$$\log\left((\text{A.25})\right) = -\left(\sqrt{a}y+\alpha_-\right)^2+\log\left(\operatorname{erfcx}(\sqrt{a}y+\alpha_-)\right)-\log(2) - \log\left(1+\operatorname{sign}(\gamma_{+-})\exp\left(-\alpha_-^2+\log(|\gamma_{+-}|)-\log(2)\right)\right) \tag{A.41}$$

$$\log\left((\text{A.26})\right) = \log\left(1-\exp\left(\log\left(\operatorname{erfcx}(-\sqrt{a}y-\alpha_-)\right)-\log(2)-\left(\sqrt{a}y+\alpha_-\right)^2\right)\right) - \log\left(1+\operatorname{sign}(\gamma_{+-})\exp\left(-\alpha_-^2+\log(|\gamma_{+-}|)-\log(2)\right)\right) \tag{A.42}$$

$$\log\left((\text{A.27})\right) = \log(r)+\log(2)-\frac{bc}{a} + \log\left(1+\operatorname{sign}(\gamma_{+-})\exp\left(-\alpha_+^2+\log(|\gamma_{+-}|)-\log(2)+\frac{bc}{a}\right)\right) \tag{A.43}$$

$$\log\left((\text{A.28})\right) = \log(1-r)+\log(2) + \log\left(1+\operatorname{sign}(\gamma_{+-})\exp\left(-\alpha_-^2+\log(|\gamma_{+-}|)-\log(2)\right)\right) \tag{A.44}$$

To obtain the numerical stability required for performing simulated annealing, one needs to use three further identities in the above expressions:

$$\left(-\sqrt{a}y+\alpha_+\right)^2 = \alpha_+^2+y\left(ay-c-b\right) \tag{A.45}$$

$$\left(\sqrt{a}y+\alpha_-\right)^2 = \alpha_-^2+y\left(ay+c-b\right) \tag{A.46}$$

$$\alpha_+^2 = \alpha_-^2+\frac{bc}{a} \tag{A.47}$$

For (4.47) and (4.48), if w denotes the logarithm of the argument of erfcinv, one can compute $\operatorname{erfcinvlog}(w) := \operatorname{erfcinv}\left(\exp(w)\right)$ using a standard implementation of erfcinv if w is not too small (the loss of precision using $\exp(w)$ instead of computing the full argument of erfcinv is negligible since the variation of erfcinv is very small even on logarithmic scale). However, even using 64 bit precision is not sufficient for the applications we address. Therefore, we use an asymptotic approximation of $\operatorname{erfcinvlog}(w)$

for $w < -680$ from the *Digital Library of Mathematical Functions*[2]:
An approximation of $z = \operatorname{erfcinv}(\exp(w))$ for $w \longrightarrow -\infty$ is given by:

$$\begin{aligned}
\theta &:= -\log(\pi) - \log(-w) \\
v &:= (-\theta - 2) \\
s &:= 2/(\theta - 2w) \\
a_2 &:= \frac{1}{8} v \\
a_3 &:= -\frac{1}{32}(v^2 + 6v - 6) \\
a_4 &:= \frac{1}{384}(4v^3 + 27v^2 + 108v - 300) \\
z &\approx s^{-1/2} + a_2 s^{3/2} + a_3 s^{5/2} + a_4 s^{7/2}
\end{aligned} \tag{A.48}$$

The discrepancy of this approximation to the implementation of erfcinv in MATLAB is $2.34 \cdot 10^{-12}$ for $w = -690$ and as it is an asymptotic formula, the error further decreases for $w \to -\infty$.

A.4. Implementation Details of the TV Slice Sampler

We have

$$p_2(x) = \exp\left(-c \sum_{j=1}^{3} \sqrt{d_j (x - e_j)^2 + g_j}\right), \quad d_j \in \{0, 1, 2\}, \quad g_j \geqslant 0, \tag{A.49}$$

and have to solve

$$y = p_2(x) \quad \Longleftrightarrow \quad -\frac{\log(y)}{c} = \sum_{j=1}^{3} \sqrt{d_j (x - e_j)^2 + g_j}, \tag{A.50}$$

where $y \in (0, p_2(x^i))$ with probability 1 and $p_2(x^i) \leqslant 1$. Assume that $\{e_1, e_2, e_3\}$ are sorted and define $J_j(x) := \sqrt{d_j (x - e_j)^2 + g_j}$ and $h := -\log(y)/c$. Then, $J(x) := \sum_j J_j(x)$ is convex and smooth in $I_1 := (-\infty, e_1)$, $I_2 := (e_1, e_2)$, $I_3 := (e_2, e_3)$ and $I_4 := (e_3, \infty)$. It is monotonic in I_1 and I_4 and is bounded from below by $b(x) := \sum_j \sqrt{d_j}|x - e_j|$. Define $\left[x_-^*, x_+^*\right] = \operatorname{argmin} J(x)$ as the interval of minimizers and x_-, x_+ as the solutions to $y = p_2(x)$. We have $x_- < x_-^*$, $x_+ > x_+^*$, $x_- \in I_1 \cup I_2 \cup I_3$ and $x_+ \in I_2 \cup I_3 \cup I_4$ with probability 1 and $[x_-^*, x_+^*] \subset [e_1, e_3]$. See Figure A.2 for two illustrations.

[2] National Institute of Standards and Technology, http://dlmf.nist.gov/, 2011

We will compute x_- by a Newton's method:

$$x_-^i = x_-^{i-1} - \frac{J(x_-^{i-1}) - h}{J'(x_-^{i-1})}, \tag{A.51}$$

initialized in a point x_-^0 such that $x_-^0 \leqslant x_-$ and $J(x)$ is smooth on $[x_-^0, x_-]$. In each step, the Newton's method approximates $J(x)$ by a tangent in x_-^{i-1}. Due to the convexity of $J(x)$ and $x_-^0 \leqslant x_- < x_-^*$ the iterates never overshoot: $x_-^0 \leqslant x_-^i \leqslant x_-$ for all i. Thereby, they stay in $[x_-^0, x_-]$ and the derivative exists. Finding such an initialization x_-^0 requires some simple considerations:
The subdifferential $\partial J(x)$ is given as the sum of the subdifferentials of $J_i(x)$ (in the set-valued sense of addition):

$$\partial J_j(x) = \begin{cases} \left\{ \dfrac{d_j(x - e_j)}{\sqrt{d_j(x-e_j)^2 + g_j}} \right\}, & \text{if} \quad x \neq e_j \quad \text{or} \quad g_j > 0 \\ \left[-\sqrt{d}, \sqrt{d} \right], & \text{if} \quad x = e_j \quad \text{and} \quad g_j = 0. \end{cases} \tag{A.52}$$

Now, let $J_e^* := \min_j J(e_j)$. We can distinguish two cases:
$h > J_e^*$: In this case, we check the following conditions in sequence:

- If $h > J(e_1)$, x_- is in I_1. We use the lower bound $b(x)$ to determine x_-^0 such that $b(x_-^0) = h$:

$$x_-^0 = e_1 + \frac{J(e_1) - h}{\sum_j \sqrt{d_j}} \tag{A.53}$$

  As $b(x) \leqslant J(x)$, and both are monotonic in I_1, we have that $x_-^0 < x_-$.

- Else if $h > J(e_2)$, x_- is in I_2. We perform one Newton step from e_1 using the maximal subgradient in e_1:

$$x_-^0 = e_1 - \frac{J(e_1) - h}{\max(\partial J(e_1))} \tag{A.54}$$

  This way, $x_-^0 \leqslant x_-$ and $[x_-^0, x_-] \subset I_2$, i.e., $J(x)$ is differentiable for all iterates.

- Else, $h > J(e_3)$ and x_- is in I_3. With a similar reasoning, we set

$$x_-^0 = e_2 - \frac{J(e_2) - h}{\max(\partial J(e_2))} \tag{A.55}$$

For finding x_+, a similar reasoning can be applied. In the locations of non-differentiability, the minimal subgradient has to be used.

$h < J_e^*$: In this case, $J(x)$ is not piecewise linear (cf. the yellow line in Figure A.2a)

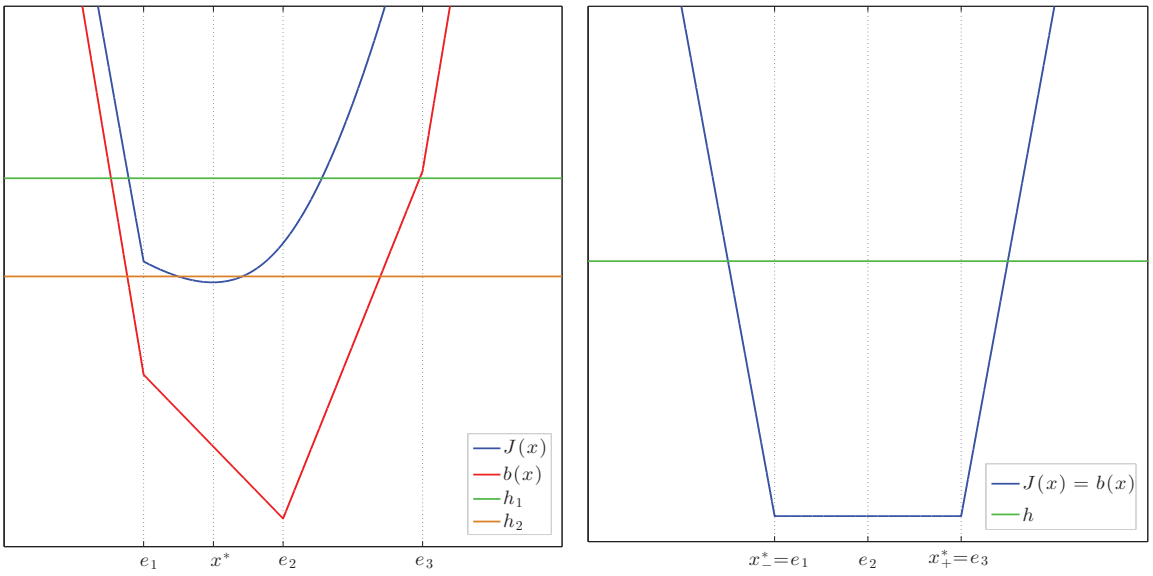


Figure A.2.: $J(x)$ (blue line), $b(x)$ (red line) and h (green and yellow lines) for $(e_1, e_2, e_3) = (-1, 0, 1)$ and (a) $(d_1, d_2, d_3) = (2, 1, 1)$, $(g_1, g_2, g_3) = (0, 0.5, 1)$, (b) $(d_1, d_2, d_3) = (1, 0, 1)$, $(g_1, g_2, g_3) = (0, 0, 0)$

and the unique minimizer x^* is not in $\{e_1, e_2, e_3\}$. The convexity ensures that $x_- < x^* < x_+$ are all either in I_2 or I_3. If $\max(\partial J(e_1)) < 0$ and $\min(\partial J(e_2)) > 0$ we have that x^* (and thereby x_- and x_+) are in I_2. Otherwise, they are in I_3. As above, initial points x_-^0 and x_+^0 fulfilling the conditions can be found by performing one Newton step from the corners of the interval using the maximal/minimal subgradient.

The case $h = J_e^*$ has probability zero.

A.5. Implementation Details of ADMM

In this section, we provide details on the MATLAB implementation used for all ADMM computations in this thesis. We want to solve

$$\min_u \left\{ \frac{1}{2}\|Au - f\|_2^2 + \mathcal{J}(u) \right\}, \tag{A.56}$$

with

$$\mathcal{J}(u) := \sum_i^N \sum_k^{l_i} \sqrt{\sum_j^{M_i} (\Theta_{i,j} D_{ij}^T u)_k^2}, \tag{A.57}$$

where $D_{ij}^T \in \mathbb{R}^{h_i \times n}$ is a fixed linear mapping and $\Theta_{i,j} : \mathbb{R}^{h_i} \to \mathbb{R}^{h_i}$ is a linear weighting. For both of them, matrix-vector and transpose-matrix-vector multiplications must be

available as function handles. For the normal ℓ_1 prior, $\Theta_{i,j}(v) = \lambda v$. However, by introducing $\Theta_{i,j}$ as a function handle, we can use this formulation to treat varying weightings, for instance as appearing in ℓ_p hypermodels. Examples for (A.57) are given by

- Standard ℓ_1 prior (e.g., in the "Spots" scenario):

$$D_{11}^T u = u, \qquad D_{11} v = v$$

- ℓ_1-block priors in EMEG:

$$D_{1i}^T u = \texttt{u(i:3:end)}, \qquad D_{1i} v = \texttt{stretch(v,i,3,n)},$$

  where $i = 1, 2, 3$ and `strech(v,i,j,n)` results in a vector of length `n`, which has the fields of `u` on every `j`-th field, starting with the `i`-th field.

- TV prior in 1D with NB boundary conditions (e.g., in the "Boxcar" scenario):

$$D_{11}^T u = \texttt{diff(u)}, \qquad D_{11} v = \texttt{[-v(1);-diff(v);v(end)]}$$

- Isotropic TV prior (NB) in a 2D $N \times N$ grid (e.g., in the "Phantom-CT" scenario):

$$D_{11}^T u = \texttt{[diff(u,1,2),zeros(N,1)]}, \quad D_{12}^T u = \texttt{[diff(u,1,1);zeros(1,N)]}$$

$$D_{11} v = \texttt{[-v(:,1),-diff(v,1,2)]}, \quad D_{12} v = \texttt{[-v(1,:);-diff(v,1,1)]}$$

- Product of the previous prior and an ℓ_1 Besov prior in 2D: $D_{11}^T, D_{12}^T, D_{11}, D_{12}$ as above and

$$D_{21}^T u = \texttt{s * wavedec2(u)}, \qquad D_{21} v = \texttt{s * waverec2(v)},$$

  where `s` is a scaling factor.

To treat the problem (A.56) with the ADMM formalism, we split by $\Theta_{i,j} D_{ij}^T u - v_{ij} = 0$, which can be written as $D^T u - v = 0$ by vertically stacking all $\Theta_{i,j} D_{ij}^T$. Then, we obtain

$$u^{k+1} = \underset{u}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Au - f\|_2^2 + \sum_{i,j}^{N,M_i} \frac{\rho}{2} \left\| v_{ij} - \Theta_{i,j} D_{ij}^T u - w_{ij} \right\|_2^2 \right\} \tag{A.58}$$

for the update of u^k, cf. (4.76). This least-squares problem is solved by an own CGLS implementation, which solves the problem

$$\underset{x}{\operatorname{argmin}} \left\{ \sum_{i}^{N} \beta_i \|G_i x - c_i\|_2^2 \right\} \tag{A.59}$$

if function handles for $G_i x$ and $G_i^T y$ are provided.
The update of v, (4.77), is given by:

$$v_{ij}^{k+1} = \max\left(a_i^k - \frac{1}{\rho}, 0\right) \frac{\Theta_{i,j} D_{ij}^T u^{k+1} + w_{ij}^k}{a_i^k}, \tag{A.60}$$

where

$$a_i^k = \sqrt{\sum_{j}^{M_i} \left(\Theta_{i,j} D_{ij}^T u^{k+1} + w_{ij}^k\right)^2}. \tag{A.61}$$

Finally, the update of w, (4.78), is given by :

$$w_{ij}^{k+1} = w_{ij}^k + \left(\Theta_{i,j} D_{ij}^T u^{k+1} - v_{ij}^{k+1}\right) \tag{A.62}$$

The norms of primal residuum r and dual residuum s are given by

$$\left\|r^{k+1}\right\|_2^2 = \left\|D^T u^{k+1} - d\right\|_2^2 = \sum_{i,j}^{N,M_i} \left\|\Theta_{i,j} D_{ij}^T u^{k+1} - v_{ij}^{k+1}\right\|_2^2 \tag{A.63}$$

$$\left\|s^{k+1}\right\|_2^2 = \rho^2 \left\| \sum_{i,j}^{N,M_i} D_{ij} \Theta_{i,j}^T (v_{ij}^k - v_{ij}^{k+1}) \right\|_2^2 \tag{A.64}$$

A.6. Implementation Details of Simulated Annealing

In the SC Gibbs sampling, only the parameters of the SC densities have to be changed. However, as the annealing proceeds, the distributions get more and more concentrated and singular. As a result, extremely robust implementations of the SC samplers are required. For ℓ_2-hypermodels, an elementary computation shows that sampling from the tempered conditional Gaussian part of the posterior can easily be implemented by replacing (4.12) by

$$\begin{bmatrix} A \\ \sqrt{2\lambda} D^T \end{bmatrix} v \overset{ls}{=} \begin{bmatrix} f \\ 0 \end{bmatrix} + \sqrt{T}\, x. \tag{A.65}$$

Using an inverse gamma hyperprior, the tempered part of the posterior depending on γ is given by

$$p_{prior}(\gamma|u) \propto \prod_i^h \exp\left(-\frac{\left\|D_i^T u\right\|_2^2 + \beta}{T\gamma_i} - \frac{\alpha + 1 + 1/p}{T}\log(\gamma_i)\right), \tag{A.66}$$

which is, again, an inverse gamma distribution with the parameters

$$\bar{\beta} = \frac{\alpha + 3/2}{T} - 1 \qquad \bar{\beta} = \frac{\left\|D_i^T u\right\|_2^2 + \beta}{T}. \tag{A.67}$$

A.7. Validation Measures for EMEG Studies

In this section, we present the performance measures we use to validate the EMEG source reconstruction studies with simulated data. A detailed discussion can be found in Section 1.3.3 in Lucka (2011) and in Lucka et al. (2012). The *dipole localization error* (*DLE*) can be used to validate the reconstruction u of a single source $u^{\dagger,\infty} = q\delta(r - r^*)$, cf. (2.16). It measures the spatial distance between r^* and the source space node j with the largest estimated current vector $u[j]$:

$$\mathrm{DLE}(u, u^{\dagger,\infty}) := \|r_j - r^*\|, \qquad \text{with} \qquad j = \operatorname*{argmax}_i \left\{\|u[i]\|_2\right\}.$$

While the DLE is the most commonly used validation measure, it suffers from several drawbacks:

- Its generalization to multiple source scenarios is difficult: The source grid $\{r_i\}$ is often arranged in an irregular fashion. Thereby, the definition of local maxima of u is difficult.
- It is only sensitive to the mode of the estimated current; its spatial extend is not accounted for.

To overcome these limitations, the *earth mover's distance* (*EMD*) was independently proposed in Haufe et al. (2008) and Lucka (2011) as a new validation measure for EMEG source reconstruction. It can be computed for arbitrary $u^{\dagger,\infty}$ and u, is sensitive to location, relative amplitude and spatial extend and does not rely on committing the obvious inverse crime of identifying $u^{\dagger,\infty} = u^{\dagger}$. The EMD is a *Wasserstein metric*, which are distance measures between probability distributions originating from the theory of *optimal transport* (Ambrosio et al. 2008):

Definition A.3 (Wasserstein metric)**.** Let μ and ν be two probability measures on a Radon space (Ω, d) that have a finite p^{th} moment for some $p \geq 1$. Then the p^{th}

Wasserstein distance $W_p(\mu, \nu)$ is defined as:

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} dist(r,r')^p \, \mathrm{d}\gamma(r,r') \right)^{1/p}, \tag{A.68}$$

where $\Gamma(\mu, \nu)$ denotes the class of all transport maps, i.e., measures on $\Omega \times \Omega$ with marginals μ and ν.

The EMD is the $p = 1$ Wasserstein distance for the 3D-Euclidean distance $dist(r, r') = \|r - r'\|_2$ between the probability measures induced by the amplitudes of $u^{\dagger,\infty}$ and u:

$$\mathrm{EMD}(u, u^{\dagger,\infty}) = W_1(\mu_u, \mu_{u^{\dagger,\infty}}) \tag{A.69}$$

where

$$\mu_u(B) := \frac{\sum_i \|u[i]\|_2 \mathbb{1}_B(r_i)}{\sum \|u[i]\|_2}, \qquad \mu_{u^{\dagger,\infty}}(B) := \frac{\int_B \left\|u^{\dagger,\infty}(r)\right\|_2 \mathrm{d}\lambda(r)}{\int \|u^{\dagger,\infty}(r)\|_2 \, \mathrm{d}\lambda(r)}. \tag{A.70}$$

Section 3.7 in Lucka (2011) discusses how to compute this quantity.
The name "earth mover's distance" comes from the intuitive explanation of this quantity given by Monge in 1781: The first probability measure is considered as an amount of sand piled on a space Ω and the second measure as a hole with the same size. For a given distance function *dist*, the minimum-cost transport of the sand into the holes has to be determined (where the cost of a single assignment is understood as classical physical work in terms of distance times amount of sand). This minimal cost is the Wasserstein distance between the two measures.
While the EMD can be computed for arbitrary complex u and $u^{\dagger,\infty}$, it reduces to intuitive measures in simple scenarios. For instance, if both u and $u^{\dagger,\infty}$ are given by a single dipole, it yields the spatial distance between them, and thereby, the DLE. If both consist of the same number of single dipoles and all dipoles have the same amplitude, the EMD is the minimal-distance assignment of the dipoles of $u^{\dagger,\infty}$ to the dipoles of u.

A.8. Software

In this section, we give a brief overview of the software used for this thesis.

External

MATLAB (www.mathworks.com/products/matlab)
is a high-level language and interactive environment for numerical computation, visualization and programming. Most of the computations performed in this thesis were

carried out with MATLAB, mainly through two toolboxes which are described in the next section. In addition, most of the figures were produced with MATLAB, for instance, such as Figures 1.6, 2.1, 2.2, 2.10b, 2.12a, 3.3a, 3.4, 3.10, 3.13, 4.1, 4.2, 5.1, 5.2, 5.7 and 5.23.

CVX (cvxr.com/cvx)
is a MATLAB-based modeling system for convex optimization. It allows for an easy and intuitive formulation of convex optimization problems, converts them internally and calls external solvers to solve the reformulated problems. We used it to compute the recovery conditions (FuB)/(SSC) and (BlkFuB). As external solvers, we used MOSEK (www.mosek.com), a multipurpose large-scale optimization software, and SeDuMi (sedumi.ie.lehigh.edu), a software for optimization over symmetric cones.

GeoGebra (www.geogebra.org)
is a dynamic mathematics software for all levels of education that joins geometry, algebra, tables, graphing, statistics and calculus in one package. We used it to produce the geometrical drawings in Figures 2.4, 2.5, 2.6, 5.19.

SimBio (https://www.mrt.uni-jena.de/simbio)
is a software for forward and inverse computations in EEG/MEG. We used the NeuroFEM sub-package for FEM-based EEG/MEG forward computations (cf. Section 5.4.3).

TetGen (www.tetgen.org)
is a program to generate high quality tetrahedral meshes of 3D polyhedral domains fulfilling certain constraints (so called *constrained Delaunay tetrahedralizations*). We used it in our EEG/MEG head modeling pipeline described in Section 5.4.1: From the triangulated compartment surfaces, we generated the tetrahedral finite element volume meshes for the FEM-based forward computations.

FieldTrip (fieldtrip.fcdonders.nl, Oostenveld et al. 2011)
is a MATLAB toolbox for EEG/MEG data analysis that is mainly developed at the Donders Institute (Nijmegen, Netherlands). We used it for the pre-processing of the experimental EEG/MEG data described in Section 5.4.5 and for producing the EEG/MEG topography plots such as Figure 2.12b.

SCIRun (www.scirun.org)
is a problem solving environment, for modeling, simulation and visualization of scientific

problems. It mainly targets bioelectrical applications and is developed by the SCI Institute (Utah, USA). For this thesis, we only used its 3D volume rendering capacities to visualize different objects in EMEG source reconstruction (Figures 2.10c, 2.11, 2.13, 5.28, 5.29, 5.33, 5.34, 5.35 and A.5)

Misc For EEG/MEG head modeling, CURRY (www.compumedicsneuroscan.com), FSL (www.fmrib.ox.ac.uk/fsl) and BESA (www.besa.de) were used. A detailed description of their usage is given in the supplementary material[3] of Janssen et al. (2013). The ellipsoid fit used in the head model cascade scenario described in Section 5.4.2 was performed using MATLAB code written by Yury Petrov and distributed through the Matlab central file exchange portal:
www.mathworks.de/matlabcentral/fileexchange/24693-ellipsoid-fit

Own Developments

Two MATLAB toolboxes were developed that will be made available on the author's homepage[4] alongside the publication of this thesis.

BayesInversion is a toolbox that implements Bayesian inversion for various inverse problems scenarios in a generic, conceptual, pseudo-object-orientated way: In a first step, an instance of an inverse problems scenario is created, which includes all the information about sensor, noise and forward modeling and the choice of the discretization approach. For simulated data scenarios, a continuous representation of $u^{\dagger,\infty}$ is created as well. Then, non-inverse-crime data for $u^{\dagger,\infty}$ is generated or pre-processed experimental data is imported. Next, a prior model is chosen. Finally, an inference procedure is determined and the computational inversion is carried out.
More specifically, the toolbox contains:

- All the inverse problems scenarios examined in Chapter 2. For the EMEG scenarios, the second toolbox described below is interfaced.
- All the prior models examined in Chapter 3.
- All the computational routines developed and examined in Chapter 4. MCMC-based computations can be parallelized.
- Functions to analyze the inversion results for carrying out studies such as those in Chapter 5. For instance, functions that compute EMD and DLE for EMEG source reconstructions or perform autocorrelation analysis of MCMC chains.

[3] stacks.iop.org/PMB/58/4881/mmedia
[4] currently: http://wwwmath.uni-muenster.de/num/Arbeitsgruppen/ag_burger/organization/lucka/

- Various functions to create and export visualizations of the results in a simple way. Examples include Figures 2.1, 2.2, 2.7, 3.4b, 5.2, 5.23 5.8, 5.10, 5.20 and 5.32c.
- Various template scripts that combine the above procedures to carry out extensive computational studies such as presented in this thesis in an efficient, script-based and user-friendly way.
- Optimized implementations of several computationally demanding routines by compiled fortran code (interfaced through `.mex` files). Some examples of the speed-ups that can be obtained include:
  - Using the MATLAB implementation of slice-within-RSG sampler for the isotropic TV prior (cf. Section A.4) to sample the posterior in the "Phantom-CT" scenario with $n = 64 \times 64$ takes about 90 times longer than the `.mex` version.
  - Using the MATLAB implementation of the RSG sampler to sample the posterior in the "Walnut-CT" scenario with $n = 64 \times 64$ using the Besov prior takes about 40 times longer than the `.mex` version.
  - Data generation in the "Phantom-CT" scenario using $n = 1024 \times 1024$ takes about 3-4 times longer using a MATLAB implementation of the CT-simulation function described in Section A.2 compared to using the `.mex` version.

SimBioInterface is a toolbox to work with EMEG head models. It contains functions

- to read and write the file formats used by SimBio and TetGen (see above). This includes FEM volume meshes consisting of tetrahedra or hexahedra, triangular surface meshes, conductivity tables, sensor configurations and source configurations to be used with SimBio.
- to perform various operations with the FEM meshes. Examples include identifying the element in which a given point lies, computing adjacency matrices of the mesh's faces, computing the volume of the elements or extracting and refining compartment surfaces.
- to construct different kinds of source spaces (cf. Section 5.4.3), to assign the FEM elements or surface facets to the source space nodes or to compute mesh-based or surface-based distances between the nodes.
- to generate different kinds of random source configurations to be used as reference sources in simulation studies.

- to carry out forward computations using SimBio as described in Section 5.4.3. For this purpose, the toolbox generates all the input files for SimBio, including the parameter file, calls SimBio by the `system.m` command, and imports the results into MATLAB after SimBio's termination. It allows to distribute forward computations to multiple CPU cores by splitting the sensor configurations.
- to visualize head models, sensor configurations, measurement data, source spaces and inverse reconstructions: Figures 2.10c, 2.11, 2.13, 5.29, 5.33, 5.34, 5.35 and A.5 were generated by these functions. All functions can either be used for visualization within MATLAB, which is most convenient for most practical tasks, or to call SCIRun, which is preferable for high quality images.

A.9. Publications and Presentations Related to the Thesis

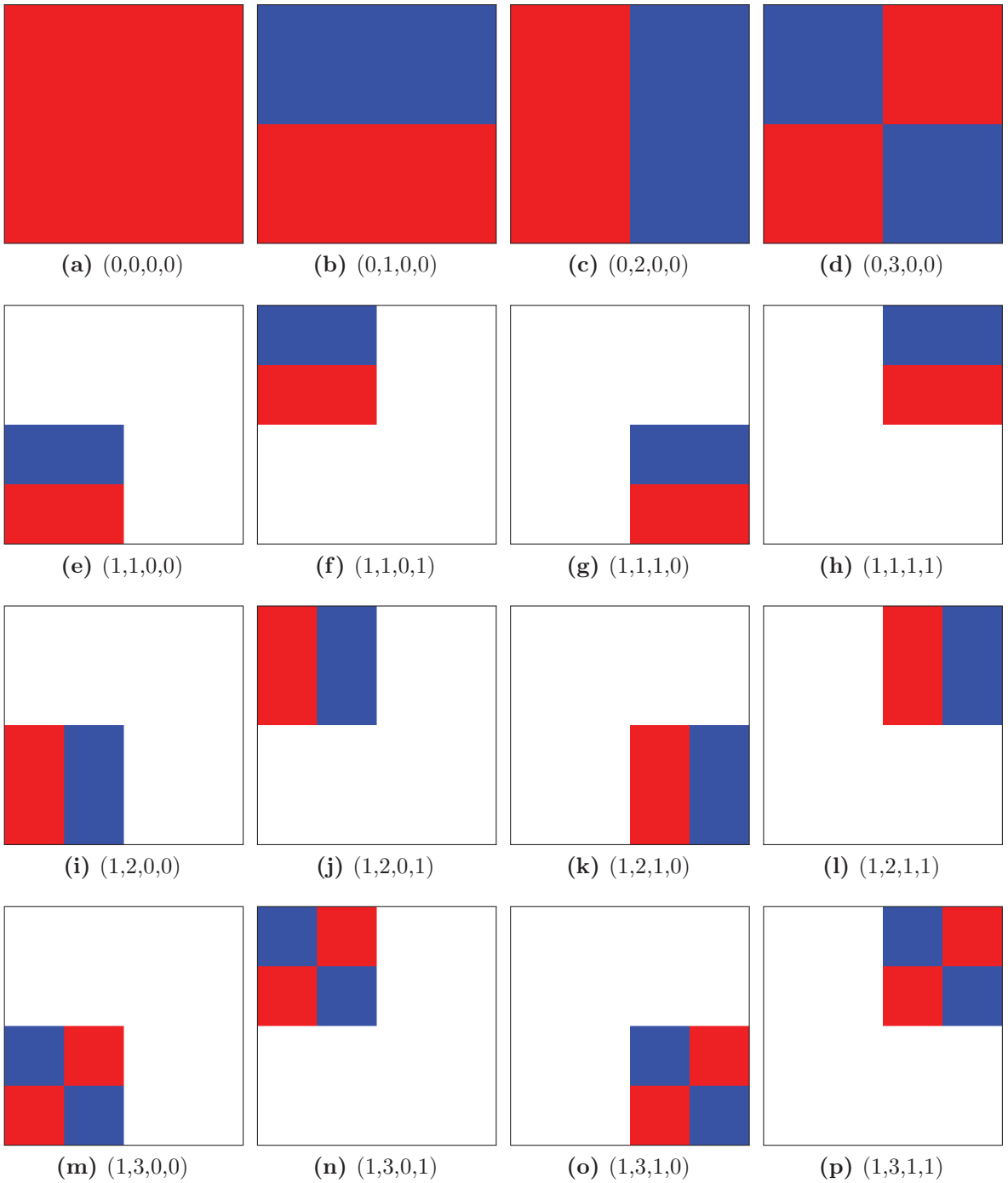
In this section, we summarize which publications and presentations are related to this thesis, in particular, which of the passages or results contained in this thesis were published/presented before and/or also contain major contributions from other authors.

- The derivation and examination of the direct SC Gibbs sampler for ℓ_1 priors was published in Lucka (2012). The computational studies therein were partly rearranged and recomputed for this thesis. In particular, the visual presentation of the results was improved and the techniques of Wolff (2004) were used in the autocorrelation analysis (thanks to the anonymous reviewer for this suggestion).
- The EEG/MEG/EMEG comparison studies in Section 5.4.4 were first presented on the "18-th International Conference on Biomagnetism" in Paris, 2012 and on several other occasions thereafter.
- Parts of the reconstructions of the SEP/SEF data and the sensitivity studies (Section 5.4.5) were first presented at the "Applied Inverse Problems Conference" in Daejeon, 2013. The full results were first presented on the "International Conference on Basic and Clinical Multimodal Imaging" in Geneva, 2013, and on several other occasions thereafter.
- The examination of the sparse recovery conditions for EEG/MEG (Section 5.4.6) started as the Master's thesis Tellen (2013), which was co-supervised by the author. The results of the Master's thesis were then extended and refined and first presented at the "Matheon Workshop on Compressed Sensing and its Applications" in Berlin, 2013, and on several other occasions thereafter.
- The novel Bayes cost functions presented in Section 6.1.2 were developed by Martin Burger. Together with some of the computational comparisons between

MAP and CM estimates carried out in this thesis (cf. the overview in Section 6.1.1), they were first presented at the "Applied Inverse Problems Conference" in Daejeon, 2013 and on several occasions thereafter. They were subsequently published in Burger and Lucka (2014).

- In Pursiainen, Lucka and Wolters (2012), the head model described in this thesis was used to examine the effects of including a model of the measurement electrodes in the EEG forward computation (cf. Section 2.4.3)
- The head model was also used in Janssen, Rampersad, Lucka, Lanfer, Lew, Aydin, Wolters, Stegeman and Oostendorp (2013) and Rampersad, Janssen, Lucka, Aydin, Lanfer, Lew, Wolters, Stegeman and Oostendorp (2014) to study the effects of realistic head modeling on simulating electro-magnetic brain stimulation techniques.

A.10. Additional Figures



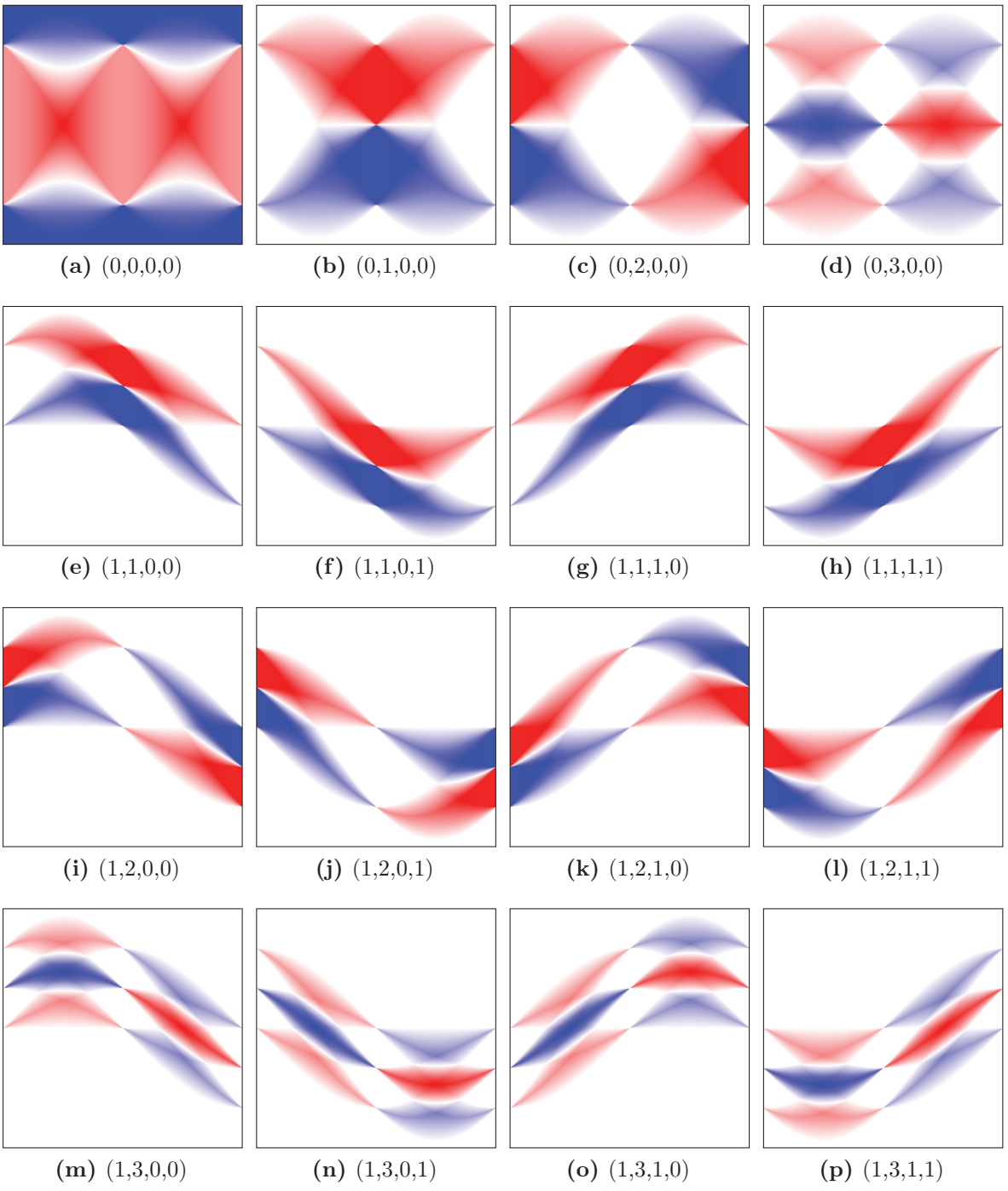
(a) (0,0,0,0) **(b)** (0,1,0,0) **(c)** (0,2,0,0) **(d)** (0,3,0,0)

(e) (1,1,0,0) **(f)** (1,1,0,1) **(g)** (1,1,1,0) **(h)** (1,1,1,1)

(i) (1,2,0,0) **(j)** (1,2,0,1) **(k)** (1,2,1,0) **(l)** (1,2,1,1)

(m) (1,3,0,0) **(n)** (1,3,0,1) **(o)** (1,3,1,0) **(p)** (1,3,1,1)

Figure A.3.: The first sixteen Haarwavelets in 2D, labeled as (j, l, k_1, k_2) and in blue-to-red color coding.



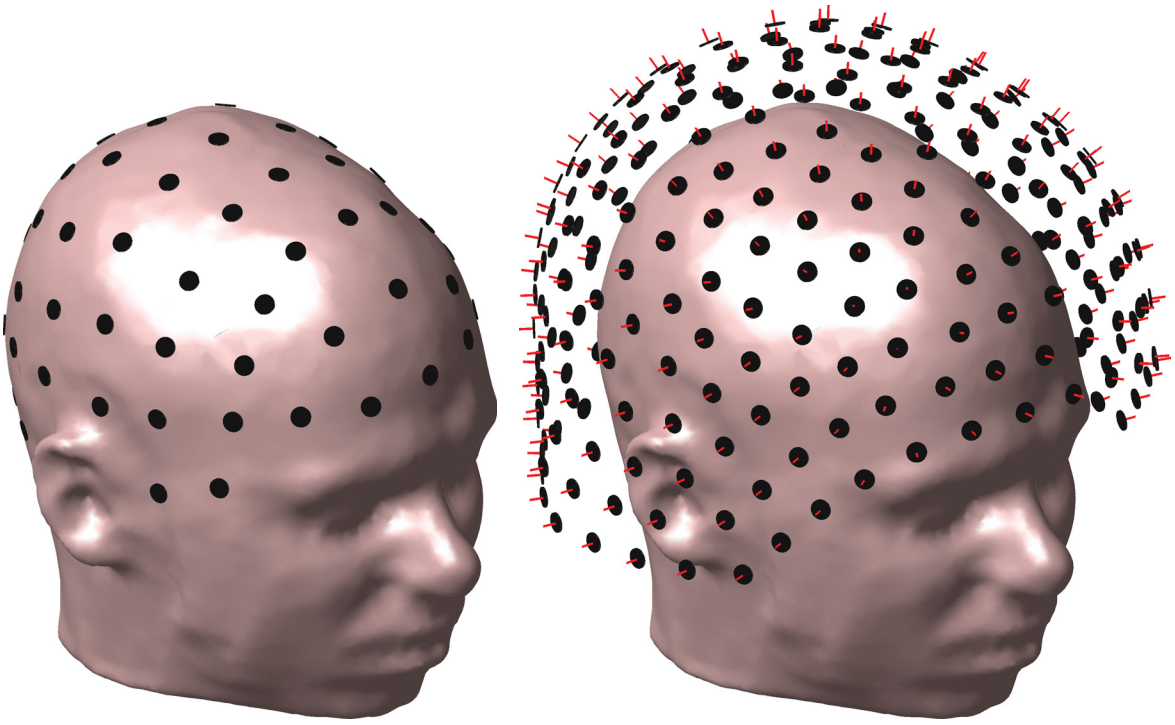
(a) (0,0,0,0) **(b)** (0,1,0,0) **(c)** (0,2,0,0) **(d)** (0,3,0,0)

(e) (1,1,0,0) **(f)** (1,1,0,1) **(g)** (1,1,1,0) **(h)** (1,1,1,1)

(i) (1,2,0,0) **(j)** (1,2,0,1) **(k)** (1,2,1,0) **(l)** (1,2,1,1)

(m) (1,3,0,0) **(n)** (1,3,0,1) **(o)** (1,3,1,0) **(p)** (1,3,1,1)

Figure A.4.: The Radon transformation of the first sixteen Haarwavelets in 2D, labeled as (j, l, k_1, k_2) and in blue-to-red color coding.



(a) EEG (b) MEG

Figure A.5.: Realistic EEG/MEG sensor configuration.

A.11. Additional Tables

Table A.2.: Values for λ defined by the parameter choice rule described in Section 5.2.3

q	λ_{CM}	λ_{MAP}	q	λ_{CM}	λ_{MAP}
1	1.714e2	1.997e2	11	7.170e-3	3.497e-1
2	3.976e1	1.059e2	12	2.826e-3	1.854e-1
3	1.222e1	5.611e1	13	1.278e-3	9.826e-2
4	4.268e0	2.974e1	14	5.379e-4	5.208e-3
5	1.553e0	1.577e1	15	2.386e-4	2.761e-2
6	5.913e-1	8.357e0	16	1.048e-4	1.463e-2
7	2.356e-1	4.430e0	17	4.503e-5	7.757e-3
8	9.825e-2	2.348e0	18	1.937e-5	4.112e-3
9	3.931e-2	1.245e0	19	8.174e-6	2.180e-3
10	1.625e-2	6.597e-1	20	3.856e-6	1.155e-3

Table A.3.: The value $1 - \mu(A)$ for all combinations of head model and N using the normal constraint.

EEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	4.30e-2	9.30e-3	1.01e-2	3.63e-2	5.31e-2	7.91e-2
125	1.10e-2	5.87e-3	7.22e-3	3.01e-2	2.26e-2	1.93e-2
250	3.92e-3	9.38e-3	9.72e-3	8.76e-3	5.84e-3	1.58e-2
500	5.32e-3	4.64e-3	6.95e-3	4.55e-3	3.57e-3	7.12e-3
1000	4.78e-4	1.21e-3	8.72e-4	1.04e-3	1.75e-3	2.18e-3
2000	2.35e-4	9.05e-4	1.25e-3	1.01e-3	9.44e-4	1.32e-3
4000	4.09e-6	2.32e-5	3.22e-5	1.85e-7	7.88e-7	1.44e-6
8000	2.58e-6	1.47e-6	1.52e-6	9.24e-9	5.58e-9	1.56e-8
16000	8.84e-7	1.28e-6	1.52e-6	9.24e-9	4.44e-9	1.56e-8

MEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	8.46e-2	6.06e-2	6.65e-2	6.50e-2	4.54e-2	8.03e-2
125	2.34e-2	4.50e-2	4.24e-2	1.78e-2	2.21e-2	2.11e-2
250	2.33e-2	2.09e-2	2.27e-2	1.32e-2	1.52e-2	1.44e-2
500	4.02e-3	3.23e-3	3.35e-3	5.82e-3	5.26e-3	5.63e-3
1000	1.09e-3	1.73e-3	1.35e-3	1.39e-3	1.31e-3	1.24e-3
2000	1.84e-3	2.49e-3	1.86e-3	8.57e-4	7.85e-4	7.69e-4
4000	9.63e-5	8.58e-5	6.83e-5	4.59e-5	5.47e-5	4.59e-5
8000	2.08e-7	9.77e-7	1.01e-6	1.01e-7	9.25e-8	7.63e-8
16000	2.08e-7	9.77e-7	1.01e-6	4.40e-8	9.25e-8	7.63e-8

EMEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.23e-1	9.66e-2	1.06e-1	1.21e-1	1.10e-1	1.44e-1
125	4.22e-2	5.19e-2	3.65e-2	7.40e-2	3.70e-2	3.51e-2
250	4.06e-2	4.47e-2	4.13e-2	4.32e-2	1.11e-2	1.93e-2
500	6.86e-3	6.94e-3	6.99e-3	1.36e-2	1.76e-2	1.79e-2
1000	1.20e-3	2.28e-3	1.90e-3	1.19e-3	2.00e-3	1.89e-3
2000	2.44e-3	2.08e-3	1.66e-3	1.05e-3	1.14e-3	1.26e-3
4000	5.76e-5	6.35e-5	5.91e-5	2.05e-5	2.22e-5	2.14e-5
8000	1.63e-6	1.45e-6	1.51e-6	7.86e-8	7.23e-8	7.18e-8
16000	7.90e-7	1.45e-6	1.51e-6	6.03e-8	6.63e-8	7.14e-8

Table A.4.: The value $1 - \mu(A)$ for all combinations of head model and N using three orthogonal dipoles per location.

EEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	2.45e-2	9.30e-3	1.01e-2	2.99e-2	1.53e-2	4.23e-2
125	1.10e-2	5.87e-3	7.22e-3	2.11e-2	1.35e-2	1.93e-2
250	3.92e-3	5.67e-3	6.07e-3	5.66e-3	5.84e-3	1.54e-2
500	3.65e-3	4.63e-4	3.52e-4	4.43e-3	2.80e-3	3.33e-3
1000	4.78e-4	1.21e-3	8.72e-4	1.04e-3	1.75e-3	2.18e-3
2000	2.35e-4	2.93e-4	3.84e-4	1.01e-3	6.79e-4	1.32e-3
4000	4.09e-6	2.32e-5	3.22e-5	1.85e-7	7.88e-7	1.44e-6
8000	2.58e-6	1.47e-6	1.52e-6	9.24e-9	5.58e-9	1.56e-8
16000	8.84e-7	1.28e-6	1.52e-6	9.24e-9	4.44e-9	1.56e-8

MEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.61e-3	2.09e-3	1.51e-3	9.96e-3	6.43e-3	1.16e-2
125	4.77e-3	4.16e-4	5.76e-4	1.23e-3	2.25e-3	8.41e-3
250	6.34e-4	2.88e-3	3.02e-3	1.29e-3	1.65e-3	3.13e-3
500	3.46e-4	8.45e-4	1.92e-3	1.06e-3	1.27e-3	3.66e-3
1000	1.09e-3	5.92e-4	5.00e-4	3.82e-4	4.93e-4	1.07e-3
2000	4.17e-4	8.82e-4	8.04e-4	3.50e-4	4.87e-4	6.33e-4
4000	9.63e-5	8.58e-5	6.83e-5	4.59e-5	5.47e-5	4.59e-5
8000	2.08e-7	9.77e-7	1.01e-6	1.01e-7	9.25e-8	7.63e-8
16000	2.08e-7	9.77e-7	1.01e-6	4.40e-8	9.25e-8	7.63e-8

EMEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	6.23e-2	7.10e-2	7.57e-2	8.62e-2	6.33e-2	7.00e-2
125	4.34e-2	2.80e-2	2.77e-2	2.91e-2	3.17e-2	4.25e-2
250	1.67e-2	1.14e-2	1.04e-2	1.13e-2	7.67e-3	1.82e-2
500	4.28e-3	2.31e-3	2.50e-3	5.85e-3	4.04e-3	4.46e-3
1000	1.04e-3	1.93e-3	1.49e-3	1.14e-3	2.19e-3	2.15e-3
2000	2.12e-3	1.99e-3	1.63e-3	1.04e-3	1.13e-3	1.27e-3
4000	5.55e-5	6.09e-5	5.68e-5	1.93e-5	2.17e-5	1.98e-5
8000	1.65e-6	1.45e-6	1.52e-6	7.88e-8	7.33e-8	7.20e-8
16000	7.91e-7	1.45e-6	1.52e-6	6.11e-8	6.86e-8	7.22e-8

Table A.5.: The value $\mu_{blk}(A)$ for all combinations of head model and N using three orthogonal dipoles per location.

EEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.64	0.63	0.61	0.54	0.52	0.53
125	0.69	0.68	0.65	0.56	0.55	0.59
250	0.76	0.72	0.70	0.62	0.58	0.63
500	0.78	0.82	0.80	0.62	0.58	0.66
1000	0.85	0.85	0.85	0.66	0.59	0.68
2000	0.82	0.81	0.80	0.69	0.60	0.67
4000	0.88	0.88	0.87	0.73	0.59	0.66
8000	0.89	0.90	0.90	0.73	0.61	0.69
16000	0.90	0.91	0.91	0.72	0.61	0.70

MEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.65	0.69	0.69	0.63	0.61	0.61
125	0.68	0.70	0.69	0.65	0.66	0.66
250	0.73	0.73	0.73	0.77	0.67	0.67
500	0.83	0.84	0.84	0.74	0.72	0.72
1000	0.84	0.78	0.78	0.82	0.73	0.72
2000	0.89	0.86	0.85	0.86	0.74	0.73
4000	0.88	0.90	0.89	0.82	0.73	0.73
8000	0.92	0.89	0.89	0.90	0.74	0.74
16000	0.92	0.89	0.89	0.90	0.74	0.73

EMEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.54	0.56	0.56	0.53	0.52	0.47
125	0.62	0.65	0.62	0.53	0.53	0.52
250	0.73	0.69	0.67	0.58	0.58	0.59
500	0.75	0.80	0.78	0.65	0.61	0.62
1000	0.79	0.78	0.79	0.66	0.61	0.60
2000	0.80	0.79	0.77	0.73	0.64	0.64
4000	0.83	0.83	0.81	0.71	0.64	0.65
8000	0.87	0.85	0.84	0.80	0.64	0.65
16000	0.88	0.86	0.87	0.75	0.64	0.66

Table A.6.: The value $1 - \mu_{sub}(A)$ for all combinations of head model and N using three orthogonal dipoles per location.

EEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	2.05e-1	1.25e-1	1.45e-1	2.51e-1	2.87e-1	1.42e-1
125	6.05e-2	1.26e-1	1.22e-1	1.95e-1	1.94e-1	9.86e-2
250	4.35e-2	4.78e-2	6.90e-2	1.38e-1	2.35e-1	8.15e-2
500	5.92e-2	7.02e-2	8.19e-2	2.09e-1	2.06e-1	1.01e-1
1000	4.99e-2	6.45e-2	7.52e-2	1.59e-1	1.89e-1	1.17e-1
2000	2.72e-2	4.22e-2	4.87e-2	1.70e-1	1.89e-1	7.67e-2
4000	3.53e-2	4.35e-2	4.85e-2	8.80e-2	2.03e-1	7.51e-2
8000	1.45e-2	2.16e-2	2.87e-2	1.12e-1	1.92e-1	5.97e-2
16000	1.43e-2	1.79e-2	2.22e-2	9.37e-2	1.73e-1	4.73e-2

MEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.61e-3	2.09e-3	1.51e-3	9.96e-3	6.43e-3	1.16e-2
125	4.77e-3	4.16e-4	5.76e-4	1.23e-3	2.25e-3	8.41e-3
250	6.34e-4	2.88e-3	3.02e-3	1.29e-3	1.65e-3	3.13e-3
500	3.46e-4	8.45e-4	1.92e-3	1.06e-3	1.27e-3	3.66e-3
1000	1.25e-3	5.92e-4	5.00e-4	3.82e-4	4.93e-4	1.07e-3
2000	5.08e-4	8.82e-4	8.04e-4	7.42e-4	6.32e-4	8.42e-4
4000	1.27e-4	9.50e-4	1.04e-3	4.57e-4	5.21e-4	1.38e-3
8000	1.33e-4	4.21e-4	5.35e-4	5.45e-4	2.65e-4	6.03e-4
16000	3.86e-4	3.07e-4	4.75e-4	2.41e-4	1.32e-4	4.79e-4

EMEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.27e-1	1.44e-1	1.79e-1	2.62e-1	4.42e-1	3.56e-1
125	6.92e-2	1.12e-1	1.48e-1	1.72e-1	2.70e-1	1.52e-1
250	2.97e-2	2.31e-2	3.56e-2	1.56e-1	2.48e-1	1.80e-1
500	4.30e-2	8.17e-2	9.14e-2	1.40e-1	2.59e-1	1.12e-1
1000	3.94e-2	5.85e-2	7.52e-2	1.29e-1	2.29e-1	1.43e-1
2000	2.20e-2	3.13e-2	4.07e-2	1.42e-1	2.29e-1	1.52e-1
4000	2.65e-2	4.16e-2	4.24e-2	6.45e-2	1.39e-1	6.71e-2
8000	6.64e-3	1.05e-2	1.46e-2	9.71e-2	1.59e-1	7.99e-2
16000	7.42e-3	9.24e-3	1.31e-2	9.04e-2	1.22e-1	5.35e-2

Table A.7.: The value $1 - \delta_2^{lb}(A)$ for all combinations of head model and N computed by Algorithm 4.12 using 10^8 2-sparse samples.

EEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	4.30e-2	9.30e-3	1.01e-2	3.63e-2	5.31e-2	7.91e-2
125	1.10e-2	5.87e-3	7.22e-3	3.01e-2	2.26e-2	1.93e-2
250	3.92e-3	9.38e-3	9.72e-3	8.76e-3	5.84e-3	1.58e-2
500	5.35e-3	4.64e-3	6.95e-3	4.55e-3	3.57e-3	7.12e-3
1000	5.25e-4	1.26e-3	9.19e-4	1.09e-3	1.78e-3	2.18e-3
2000	2.58e-4	9.28e-4	1.40e-3	1.41e-3	1.50e-3	1.89e-3
4000	8.71e-4	6.61e-4	6.34e-4	1.69e-4	1.95e-4	1.63e-4
8000	3.96e-4	1.47e-3	1.13e-3	7.91e-4	1.03e-3	1.22e-3
16000	5.16e-4	5.17e-4	5.40e-4	1.15e-3	7.81e-4	1.21e-3

MEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	8.46e-2	6.06e-2	6.65e-2	6.50e-2	4.54e-2	8.03e-2
125	2.34e-2	4.50e-2	4.24e-2	1.78e-2	2.21e-2	2.11e-2
250	2.33e-2	2.09e-2	2.27e-2	1.32e-2	1.52e-2	1.44e-2
500	4.02e-3	3.23e-3	3.35e-3	5.82e-3	5.26e-3	5.64e-3
1000	1.15e-3	1.80e-3	1.43e-3	1.46e-3	1.37e-3	1.31e-3
2000	2.28e-3	2.50e-3	2.25e-3	1.24e-3	1.17e-3	1.16e-3
4000	5.11e-4	5.00e-4	4.83e-4	5.48e-4	5.31e-4	5.08e-4
8000	5.60e-6	6.37e-6	6.40e-6	5.49e-6	5.48e-6	5.47e-6
16000	4.70e-4	9.39e-4	9.51e-4	7.87e-4	4.15e-4	4.77e-4

EMEG						
N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.23e-1	9.66e-2	1.06e-1	1.21e-1	1.10e-1	1.44e-1
125	4.22e-2	5.19e-2	3.65e-2	7.40e-2	3.70e-2	3.51e-2
250	4.06e-2	4.47e-2	4.13e-2	4.32e-2	1.11e-2	1.93e-2
500	6.86e-3	6.97e-3	7.02e-3	1.36e-2	1.76e-2	1.79e-2
1000	1.24e-3	2.32e-3	1.94e-3	1.23e-3	2.04e-3	1.93e-3
2000	2.56e-3	2.08e-3	1.66e-3	1.18e-3	1.67e-3	1.97e-3
4000	1.44e-4	1.50e-4	1.45e-4	1.22e-4	1.45e-4	1.50e-4
8000	5.94e-4	6.19e-4	5.35e-4	3.13e-4	4.02e-4	8.07e-4
16000	1.14e-3	1.42e-3	1.48e-3	8.29e-4	7.71e-4	7.93e-4

Table A.8.: Empirical probabilities and confidence intervals (significance level $\alpha = 5\%$) of (Tr) being true for $A^\sharp$ and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.065	0.078	0.105	0.212	0.101	0.717
125	0.016	0.027	0.037	0.079	0.035	0.533
250	0.005	0.002	0.007	0.019	0.006	0.259
500	0.000	0.000	0.000	0.004	0.000	0.151
1000	0.000	0.000	0.000	0.001	0.000	0.073
2000	0.000	0.000	0.000	0.000	0.000	0.020
4000	0.000	0.000	0.000	0.000	0.000	0.017
8000	0.000	0.000	0.000	0.000	0.000	0.012

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.051,0.082]	[0.062,0.096]	[0.087,0.126]	[0.187,0.239]	[0.083,0.121]	[0.688,0.745]
125	[0.009,0.026]	[0.018,0.039]	[0.026,0.051]	[0.063,0.097]	[0.024,0.048]	[0.502,0.564]
250	[0.002,0.012]	[0.000,0.007]	[0.003,0.014]	[0.011,0.030]	[0.002,0.013]	[0.232,0.287]
500	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.001,0.010]	[0.000,0.004]	[0.129,0.175]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.006]	[0.000,0.004]	[0.058,0.091]
2000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.012,0.031]
4000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.010,0.027]
8000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.006,0.021]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.551	0.539	0.536	0.449	0.482	0.536
125	0.312	0.353	0.331	0.212	0.230	0.245
250	0.141	0.123	0.126	0.101	0.101	0.112
500	0.051	0.066	0.064	0.056	0.036	0.046
1000	0.019	0.019	0.018	0.019	0.013	0.020
2000	0.005	0.006	0.006	0.006	0.003	0.007
4000	0.000	0.000	0.000	0.000	0.002	0.002
8000	0.000	0.002	0.002	0.001	0.000	0.000

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.520,0.582]	[0.508,0.570]	[0.505,0.567]	[0.418,0.480]	[0.451,0.513]	[0.505,0.567]
125	[0.283,0.342]	[0.323,0.384]	[0.302,0.361]	[0.187,0.239]	[0.204,0.257]	[0.219,0.273]
250	[0.120,0.164]	[0.103,0.145]	[0.106,0.148]	[0.083,0.121]	[0.083,0.121]	[0.093,0.133]
500	[0.038,0.067]	[0.051,0.083]	[0.050,0.081]	[0.043,0.072]	[0.025,0.049]	[0.034,0.061]
1000	[0.011,0.030]	[0.011,0.030]	[0.011,0.028]	[0.011,0.030]	[0.007,0.022]	[0.012,0.031]
2000	[0.002,0.012]	[0.002,0.013]	[0.002,0.013]	[0.002,0.013]	[0.001,0.009]	[0.003,0.014]
4000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.007]	[0.000,0.007]
8000	[0.000,0.004]	[0.000,0.007]	[0.000,0.007]	[0.000,0.006]	[0.000,0.004]	[0.000,0.004]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.802	0.801	0.839	0.817	0.832	0.969
125	0.527	0.550	0.582	0.561	0.556	0.837
250	0.319	0.339	0.355	0.383	0.327	0.626
500	0.182	0.212	0.233	0.254	0.195	0.462
1000	0.129	0.132	0.146	0.165	0.111	0.345
2000	0.071	0.080	0.087	0.093	0.062	0.245
4000	0.034	0.046	0.050	0.051	0.023	0.156
8000	0.040	0.036	0.040	0.046	0.026	0.135

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.776,0.826]	[0.775,0.825]	[0.815,0.861]	[0.792,0.841]	[0.807,0.855]	[0.956,0.979]
125	[0.496,0.558]	[0.519,0.581]	[0.551,0.613]	[0.530,0.592]	[0.525,0.587]	[0.813,0.859]
250	[0.290,0.349]	[0.310,0.369]	[0.325,0.386]	[0.353,0.414]	[0.298,0.357]	[0.595,0.656]
500	[0.159,0.207]	[0.187,0.239]	[0.207,0.260]	[0.227,0.282]	[0.171,0.221]	[0.431,0.493]
1000	[0.109,0.151]	[0.112,0.155]	[0.125,0.169]	[0.143,0.189]	[0.092,0.132]	[0.316,0.375]
2000	[0.056,0.089]	[0.064,0.099]	[0.070,0.106]	[0.076,0.113]	[0.048,0.079]	[0.219,0.273]
4000	[0.024,0.047]	[0.034,0.061]	[0.037,0.065]	[0.038,0.067]	[0.015,0.034]	[0.134,0.180]
8000	[0.029,0.054]	[0.025,0.049]	[0.029,0.054]	[0.034,0.061]	[0.017,0.038]	[0.114,0.158]

Table A.9.: Empirical probabilities and confidence intervals (significance level $\alpha = 5\%$) of (FuA) being true for A and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.051	0.055	0.075	0.197	0.181	0.380
125	0.015	0.011	0.020	0.077	0.087	0.233
250	0.001	0.009	0.012	0.060	0.046	0.096
500	0.001	0.002	0.004	0.021	0.022	0.043
1000	0.000	0.000	0.000	0.005	0.006	0.011
2000	0.000	0.000	0.000	0.005	0.002	0.006
4000	0.000	0.000	0.000	0.002	0.001	0.003
8000	0.000	0.000	0.000	0.000	0.000	0.001

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.038,0.067]	[0.042,0.071]	[0.059,0.093]	[0.173,0.223]	[0.158,0.206]	[0.350,0.411]
125	[0.008,0.025]	[0.006,0.020]	[0.012,0.031]	[0.061,0.095]	[0.070,0.106]	[0.207,0.260]
250	[0.000,0.006]	[0.004,0.017]	[0.006,0.021]	[0.046,0.077]	[0.034,0.061]	[0.078,0.116]
500	[0.000,0.006]	[0.000,0.007]	[0.001,0.010]	[0.013,0.032]	[0.014,0.033]	[0.031,0.057]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.002,0.012]	[0.002,0.013]	[0.006,0.020]
2000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.002,0.012]	[0.000,0.007]	[0.002,0.013]
4000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.007]	[0.000,0.006]	[0.001,0.009]
8000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.006]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.107	0.115	0.101	0.056	0.088	0.100
125	0.020	0.022	0.019	0.010	0.029	0.033
250	0.014	0.016	0.013	0.017	0.017	0.017
500	0.004	0.005	0.005	0.008	0.008	0.008
1000	0.001	0.002	0.002	0.003	0.004	0.003
2000	0.000	0.001	0.000	0.001	0.002	0.002
4000	0.000	0.001	0.001	0.000	0.000	0.000
8000	0.000	0.000	0.000	0.000	0.000	0.000

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.089,0.128]	[0.096,0.136]	[0.083,0.121]	[0.043,0.072]	[0.071,0.107]	[0.082,0.120]
125	[0.012,0.031]	[0.014,0.033]	[0.011,0.030]	[0.005,0.018]	[0.020,0.041]	[0.023,0.046]
250	[0.008,0.023]	[0.009,0.026]	[0.007,0.022]	[0.010,0.027]	[0.010,0.027]	[0.010,0.027]
500	[0.001,0.010]	[0.002,0.012]	[0.002,0.012]	[0.003,0.016]	[0.003,0.016]	[0.003,0.016]
1000	[0.000,0.006]	[0.000,0.007]	[0.000,0.007]	[0.001,0.009]	[0.001,0.010]	[0.001,0.009]
2000	[0.000,0.004]	[0.000,0.006]	[0.000,0.004]	[0.000,0.006]	[0.000,0.007]	[0.000,0.007]
4000	[0.000,0.004]	[0.000,0.006]	[0.000,0.006]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
8000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.323	0.436	0.447	0.492	0.609	0.698
125	0.084	0.128	0.119	0.255	0.334	0.443
250	0.030	0.056	0.056	0.177	0.160	0.298
500	0.011	0.012	0.011	0.070	0.062	0.148
1000	0.003	0.006	0.006	0.026	0.021	0.067
2000	0.000	0.002	0.003	0.011	0.009	0.046
4000	0.001	0.002	0.002	0.005	0.003	0.022
8000	0.000	0.000	0.000	0.003	0.002	0.008

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.294,0.353]	[0.405,0.467]	[0.416,0.478]	[0.461,0.523]	[0.578,0.639]	[0.668,0.726]
125	[0.068,0.103]	[0.108,0.150]	[0.100,0.141]	[0.228,0.283]	[0.305,0.364]	[0.412,0.474]
250	[0.020,0.043]	[0.043,0.072]	[0.043,0.072]	[0.154,0.202]	[0.138,0.184]	[0.270,0.327]
500	[0.006,0.020]	[0.006,0.021]	[0.006,0.020]	[0.055,0.088]	[0.048,0.079]	[0.127,0.172]
1000	[0.001,0.009]	[0.002,0.013]	[0.002,0.013]	[0.017,0.038]	[0.013,0.032]	[0.052,0.084]
2000	[0.000,0.004]	[0.000,0.007]	[0.001,0.009]	[0.006,0.020]	[0.004,0.017]	[0.034,0.061]
4000	[0.000,0.006]	[0.000,0.007]	[0.000,0.007]	[0.002,0.012]	[0.001,0.009]	[0.014,0.033]
8000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.001,0.009]	[0.000,0.007]	[0.003,0.016]

Table A.10.: Empirical probabilities and confidence intervals (significance level $\alpha = 5\%$) of (FuA) being true for $A^\sharp$ and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.184	0.218	0.274	0.380	0.236	0.826
125	0.091	0.115	0.138	0.206	0.122	0.660
250	0.066	0.057	0.072	0.115	0.071	0.412
500	0.032	0.035	0.042	0.050	0.032	0.272
1000	0.015	0.013	0.018	0.028	0.017	0.160
2000	0.006	0.009	0.011	0.011	0.006	0.070
4000	0.001	0.001	0.001	0.001	0.000	0.050
8000	0.001	0.001	0.002	0.001	0.002	0.041

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.160,0.209]	[0.193,0.245]	[0.247,0.303]	[0.350,0.411]	[0.210,0.264]	[0.801,0.849]
125	[0.074,0.111]	[0.096,0.136]	[0.117,0.161]	[0.181,0.232]	[0.102,0.144]	[0.630,0.689]
250	[0.051,0.083]	[0.043,0.073]	[0.057,0.090]	[0.096,0.136]	[0.056,0.089]	[0.381,0.443]
500	[0.022,0.045]	[0.024,0.048]	[0.030,0.056]	[0.037,0.065]	[0.022,0.045]	[0.245,0.301]
1000	[0.008,0.025]	[0.007,0.022]	[0.011,0.028]	[0.019,0.040]	[0.010,0.027]	[0.138,0.184]
2000	[0.002,0.013]	[0.004,0.017]	[0.006,0.020]	[0.006,0.020]	[0.002,0.013]	[0.055,0.088]
4000	[0.000,0.006]	[0.000,0.006]	[0.000,0.006]	[0.000,0.006]	[0.000,0.004]	[0.037,0.065]
8000	[0.000,0.006]	[0.000,0.006]	[0.000,0.007]	[0.000,0.006]	[0.000,0.007]	[0.030,0.055]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.700	0.721	0.716	0.650	0.673	0.703
125	0.526	0.568	0.546	0.450	0.472	0.472
250	0.312	0.308	0.304	0.242	0.266	0.294
500	0.168	0.185	0.186	0.157	0.136	0.161
1000	0.088	0.096	0.089	0.076	0.082	0.092
2000	0.035	0.036	0.035	0.030	0.025	0.028
4000	0.008	0.012	0.011	0.016	0.011	0.012
8000	0.009	0.008	0.009	0.008	0.010	0.009

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.671,0.728]	[0.692,0.749]	[0.687,0.744]	[0.620,0.680]	[0.643,0.702]	[0.674,0.731]
125	[0.495,0.557]	[0.537,0.599]	[0.515,0.577]	[0.419,0.481]	[0.441,0.503]	[0.441,0.503]
250	[0.283,0.342]	[0.279,0.338]	[0.276,0.334]	[0.216,0.270]	[0.239,0.295]	[0.266,0.323]
500	[0.145,0.193]	[0.161,0.210]	[0.162,0.212]	[0.135,0.181]	[0.115,0.159]	[0.139,0.185]
1000	[0.071,0.107]	[0.078,0.116]	[0.072,0.108]	[0.060,0.094]	[0.066,0.101]	[0.075,0.112]
2000	[0.024,0.048]	[0.025,0.049]	[0.024,0.048]	[0.020,0.043]	[0.016,0.037]	[0.019,0.040]
4000	[0.003,0.016]	[0.006,0.021]	[0.006,0.020]	[0.009,0.026]	[0.006,0.020]	[0.006,0.021]
8000	[0.004,0.017]	[0.003,0.016]	[0.004,0.017]	[0.003,0.016]	[0.005,0.018]	[0.004,0.017]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.880	0.868	0.890	0.867	0.880	0.977
125	0.657	0.694	0.725	0.675	0.690	0.888
250	0.501	0.530	0.547	0.560	0.510	0.746
500	0.330	0.360	0.374	0.403	0.338	0.592
1000	0.250	0.243	0.257	0.285	0.224	0.477
2000	0.141	0.151	0.160	0.167	0.126	0.345
4000	0.082	0.095	0.108	0.119	0.076	0.252
8000	0.078	0.071	0.077	0.086	0.048	0.196

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.858,0.899]	[0.845,0.888]	[0.869,0.909]	[0.844,0.887]	[0.858,0.899]	[0.966,0.985]
125	[0.627,0.686]	[0.664,0.722]	[0.696,0.752]	[0.645,0.704]	[0.660,0.719]	[0.867,0.907]
250	[0.470,0.532]	[0.499,0.561]	[0.516,0.578]	[0.529,0.591]	[0.479,0.541]	[0.718,0.773]
500	[0.301,0.360]	[0.330,0.391]	[0.344,0.405]	[0.372,0.434]	[0.309,0.368]	[0.561,0.623]
1000	[0.223,0.278]	[0.217,0.271]	[0.230,0.285]	[0.257,0.314]	[0.199,0.251]	[0.446,0.508]
2000	[0.120,0.164]	[0.129,0.175]	[0.138,0.184]	[0.144,0.192]	[0.106,0.148]	[0.316,0.375]
4000	[0.066,0.101]	[0.078,0.115]	[0.089,0.129]	[0.100,0.141]	[0.060,0.094]	[0.225,0.280]
8000	[0.062,0.096]	[0.056,0.089]	[0.061,0.095]	[0.069,0.105]	[0.036,0.063]	[0.172,0.222]

Table A.11.: Empirical probabilities and confidence intervals (significance level $\alpha = 5\%$) of (FuB)/(SSC) being true for A and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	0.973	0.961	0.960	0.932	0.990	0.881
250	0.753	0.774	0.766	0.810	0.863	0.657
500	0.634	0.648	0.643	0.672	0.702	0.544
1000	0.452	0.559	0.551	0.564	0.602	0.404
2000	0.312	0.428	0.429	0.414	0.430	0.268
4000	0.182	0.259	0.257	0.282	0.312	0.199
8000	0.110	0.135	0.139	0.204	0.233	0.136

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.961,0.982]	[0.947,0.972]	[0.946,0.971]	[0.915,0.947]	[0.982,0.995]	[0.859,0.900]
250	[0.725,0.779]	[0.747,0.800]	[0.738,0.792]	[0.784,0.834]	[0.840,0.884]	[0.627,0.686]
500	[0.603,0.664]	[0.617,0.678]	[0.612,0.673]	[0.642,0.701]	[0.673,0.730]	[0.513,0.575]
1000	[0.421,0.483]	[0.528,0.590]	[0.520,0.582]	[0.533,0.595]	[0.571,0.632]	[0.373,0.435]
2000	[0.283,0.342]	[0.397,0.459]	[0.398,0.460]	[0.383,0.445]	[0.399,0.461]	[0.241,0.297]
4000	[0.159,0.207]	[0.232,0.287]	[0.230,0.285]	[0.254,0.311]	[0.283,0.342]	[0.175,0.225]
8000	[0.091,0.131]	[0.114,0.158]	[0.118,0.162]	[0.179,0.230]	[0.207,0.260]	[0.115,0.159]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	0.996	0.994
1000	0.991	0.995	0.995	0.998	0.965	0.961
2000	0.896	0.927	0.922	0.942	0.796	0.819
4000	0.728	0.796	0.790	0.828	0.629	0.644
8000	0.582	0.633	0.625	0.690	0.491	0.504

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
500	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.990,0.999]	[0.987,0.998]
1000	[0.983,0.996]	[0.988,0.998]	[0.988,0.998]	[0.993,1.000]	[0.952,0.976]	[0.947,0.972]
2000	[0.875,0.914]	[0.909,0.942]	[0.904,0.938]	[0.926,0.956]	[0.770,0.821]	[0.794,0.842]
4000	[0.699,0.755]	[0.770,0.821]	[0.763,0.815]	[0.803,0.851]	[0.598,0.659]	[0.613,0.674]
8000	[0.551,0.613]	[0.602,0.663]	[0.594,0.655]	[0.660,0.719]	[0.460,0.522]	[0.473,0.535]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	0.999	1.000	0.999	0.999	0.999	0.998
2000	0.946	0.967	0.974	0.994	0.989	0.981
4000	0.844	0.890	0.896	0.970	0.964	0.942
8000	0.747	0.822	0.818	0.943	0.942	0.902

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
500	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
1000	[0.994,1.000]	[0.996,1.000]	[0.994,1.000]	[0.994,1.000]	[0.994,1.000]	[0.993,1.000]
2000	[0.930,0.959]	[0.954,0.977]	[0.962,0.983]	[0.987,0.998]	[0.980,0.994]	[0.970,0.989]
4000	[0.820,0.866]	[0.869,0.909]	[0.875,0.914]	[0.957,0.980]	[0.951,0.975]	[0.926,0.956]
8000	[0.719,0.774]	[0.797,0.845]	[0.793,0.841]	[0.927,0.957]	[0.926,0.956]	[0.882,0.920]

Table A.12.: Empirical probabilities and confidence intervals (significance level $\alpha = 5\%$) of (FuB)/(SSC) being true for $A^\sharp$ and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	0.998	0.997	0.997	0.998	1.000	1.000
250	0.969	0.968	0.969	0.974	0.963	0.986
500	0.948	0.936	0.941	0.949	0.938	0.979
1000	0.888	0.878	0.887	0.908	0.884	0.943
2000	0.776	0.769	0.780	0.818	0.787	0.886
4000	0.676	0.665	0.678	0.740	0.670	0.849
8000	0.652	0.642	0.664	0.709	0.653	0.810

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.993,1.000]	[0.991,0.999]	[0.991,0.999]	[0.993,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.956,0.979]	[0.955,0.978]	[0.956,0.979]	[0.962,0.983]	[0.949,0.974]	[0.977,0.992]
500	[0.932,0.961]	[0.919,0.950]	[0.925,0.955]	[0.933,0.962]	[0.921,0.952]	[0.968,0.987]
1000	[0.867,0.907]	[0.856,0.898]	[0.866,0.906]	[0.888,0.925]	[0.863,0.903]	[0.927,0.957]
2000	[0.749,0.801]	[0.742,0.795]	[0.753,0.805]	[0.793,0.841]	[0.760,0.812]	[0.865,0.905]
4000	[0.646,0.705]	[0.635,0.694]	[0.648,0.707]	[0.712,0.767]	[0.640,0.699]	[0.825,0.871]
8000	[0.622,0.682]	[0.611,0.672]	[0.634,0.693]	[0.680,0.737]	[0.623,0.683]	[0.784,0.834]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	0.998	1.000	0.999
2000	0.995	0.994	0.995	0.992	0.989	0.992
4000	0.984	0.984	0.985	0.982	0.980	0.979
8000	0.979	0.975	0.972	0.969	0.962	0.962

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
500	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
1000	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.993,1.000]	[0.996,1.000]	[0.994,1.000]
2000	[0.988,0.998]	[0.987,0.998]	[0.988,0.998]	[0.984,0.997]	[0.980,0.994]	[0.984,0.997]
4000	[0.974,0.991]	[0.974,0.991]	[0.975,0.992]	[0.972,0.989]	[0.969,0.988]	[0.968,0.987]
8000	[0.968,0.987]	[0.963,0.984]	[0.960,0.981]	[0.956,0.979]	[0.948,0.973]	[0.948,0.973]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	0.999	1.000	0.999	0.998	0.999	1.000
4000	0.997	0.994	0.994	0.997	0.995	0.996
8000	0.989	0.990	0.989	0.988	0.991	0.993

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
500	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
1000	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
2000	[0.994,1.000]	[0.996,1.000]	[0.994,1.000]	[0.993,1.000]	[0.994,1.000]	[0.996,1.000]
4000	[0.991,0.999]	[0.987,0.998]	[0.987,0.998]	[0.991,0.999]	[0.988,0.998]	[0.990,0.999]
8000	[0.980,0.994]	[0.982,0.995]	[0.980,0.994]	[0.979,0.994]	[0.983,0.996]	[0.986,0.997]

Table A.13.: Empirical probabilities of (FuB)/(SSC) being true for $A/A^\sharp$ and $k = 3$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	0.955	0.943	0.942	0.897	0.966	0.791
250	0.580	0.606	0.595	0.609	0.679	0.421
500	0.400	0.457	0.450	0.418	0.451	0.298
1000	0.295	0.392	0.380	0.282	0.332	0.180
2000	0.144	0.232	0.234	0.166	0.186	0.097
4000	0.064	0.104	0.101	0.069	0.090	0.036
8000	0.023	0.044	0.041	0.041	0.052	0.022

(b) EEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	0.995	0.994	0.993	0.999	0.998	1.000
250	0.912	0.908	0.907	0.933	0.917	0.956
500	0.777	0.768	0.777	0.832	0.785	0.897
1000	0.674	0.660	0.672	0.725	0.678	0.811
2000	0.497	0.497	0.517	0.583	0.507	0.736
4000	0.386	0.362	0.388	0.441	0.372	0.624
8000	0.283	0.260	0.272	0.328	0.274	0.528

(c) MEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	0.993	0.984
1000	0.979	0.990	0.992	0.996	0.931	0.928
2000	0.807	0.870	0.860	0.878	0.704	0.701
4000	0.580	0.675	0.660	0.703	0.479	0.478
8000	0.387	0.455	0.428	0.493	0.306	0.330

(d) MEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	0.999	1.000	0.999	0.996	0.997
2000	0.987	0.985	0.991	0.986	0.973	0.975
4000	0.971	0.972	0.965	0.957	0.937	0.951
8000	0.920	0.913	0.917	0.902	0.879	0.891

(e) EMEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	0.997	0.998	0.997	0.999	0.998	0.995
2000	0.904	0.930	0.939	0.982	0.976	0.966
4000	0.760	0.842	0.845	0.925	0.922	0.885
8000	0.564	0.709	0.706	0.845	0.831	0.783

(f) EMEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	0.998	0.996	0.995	0.996	0.996	0.997
4000	0.993	0.993	0.993	0.990	0.985	0.993
8000	0.971	0.971	0.968	0.970	0.970	0.976

Table A.14.: Empirical probabilities of (SSC_+) being true for $A/A^\sharp$ and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	0.987	0.998	1.000	1.000	1.000
250	0.947	0.923	0.923	0.928	0.977	0.919
500	0.791	0.817	0.814	0.879	0.935	0.821
1000	0.652	0.730	0.715	0.779	0.849	0.684
2000	0.469	0.597	0.599	0.665	0.702	0.540
4000	0.333	0.442	0.433	0.521	0.562	0.416
8000	0.205	0.241	0.239	0.395	0.443	0.298

(b) EEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	0.996	0.996	0.996	0.997	0.996	1.000
500	0.984	0.987	0.989	0.989	0.987	0.993
1000	0.948	0.955	0.955	0.962	0.955	0.978
2000	0.891	0.890	0.898	0.915	0.900	0.946
4000	0.827	0.817	0.824	0.852	0.820	0.904
8000	0.773	0.765	0.772	0.799	0.761	0.862

(c) MEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	0.992	0.993
2000	0.978	0.981	0.984	0.988	0.914	0.920
4000	0.871	0.894	0.898	0.938	0.769	0.802
8000	0.763	0.799	0.793	0.861	0.616	0.657

(d) MEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	1.000	1.000	1.000	0.999	1.000	0.998
4000	0.994	0.995	0.995	0.996	0.994	0.994
8000	0.990	0.992	0.992	0.990	0.987	0.982

(e) EMEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	0.990	0.992	0.994	0.998	0.997	0.997
4000	0.934	0.965	0.969	0.994	0.991	0.987
8000	0.882	0.917	0.930	0.978	0.978	0.966

(f) EMEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	1.000	1.000	1.000	1.000	1.000	1.000
4000	0.999	0.999	1.000	1.000	0.999	0.999
8000	0.998	0.998	0.997	0.994	0.996	0.995

Table A.15.: Empirical probabilities of (SSC_+) being true for $A/A^\sharp$ and $k = 3$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded.

(a) EEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	0.984	0.994	1.000	1.000	1.000
250	0.901	0.841	0.842	0.850	0.955	0.852
500	0.644	0.671	0.662	0.729	0.806	0.622
1000	0.502	0.576	0.570	0.577	0.653	0.464
2000	0.281	0.417	0.415	0.440	0.462	0.328
4000	0.150	0.239	0.233	0.249	0.283	0.167
8000	0.074	0.122	0.114	0.160	0.171	0.094

(b) EEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	0.985	0.982	0.982	0.990	0.982	0.997
500	0.936	0.932	0.932	0.952	0.936	0.971
1000	0.844	0.833	0.841	0.871	0.848	0.912
2000	0.715	0.712	0.734	0.763	0.734	0.855
4000	0.617	0.602	0.610	0.640	0.590	0.754
8000	0.467	0.452	0.461	0.488	0.446	0.644

(c) MEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	0.984	0.988
2000	0.944	0.956	0.957	0.975	0.873	0.875
4000	0.798	0.839	0.842	0.870	0.654	0.675
8000	0.574	0.650	0.653	0.711	0.460	0.509

(d) MEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	0.998	0.998	0.998	0.999	0.996	0.996
4000	0.997	0.997	0.996	0.992	0.978	0.988
8000	0.972	0.969	0.968	0.967	0.942	0.948

(e) EMEG empirical probabilities for A.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	0.999	1.000
2000	0.984	0.981	0.983	0.997	0.995	0.994
4000	0.906	0.941	0.951	0.976	0.977	0.968
8000	0.794	0.850	0.850	0.945	0.944	0.916

(f) EMEG empirical probabilities for $A^\sharp$.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	1.000	0.999	1.000	1.000	0.998	1.000
4000	0.998	0.998	0.999	0.996	0.997	0.998
8000	0.987	0.989	0.990	0.985	0.987	0.992

Table A.16.: Empirical probabilities and confidence intervals of (BlkFuA) being true for A and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded and the orientation of the real sources was always chosen normal to the cortical surface.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.000	0.000	0.000	0.056	0.056	0.152
125	0.000	0.000	0.000	0.024	0.026	0.054
250	0.000	0.000	0.000	0.008	0.012	0.032
500	0.000	0.000	0.000	0.003	0.002	0.009
1000	0.000	0.000	0.000	0.000	0.000	0.001

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.043,0.072]	[0.043,0.072]	[0.130,0.176]
125	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.015,0.036]	[0.017,0.038]	[0.041,0.070]
250	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.003,0.016]	[0.006,0.021]	[0.022,0.045]
500	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.001,0.009]	[0.000,0.007]	[0.004,0.017]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.006]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.000	0.000	0.000	0.000	0.000	0.000
125	0.000	0.000	0.000	0.000	0.000	0.000
250	0.000	0.000	0.000	0.000	0.000	0.000
500	0.000	0.000	0.000	0.000	0.000	0.000
1000	0.000	0.000	0.000	0.000	0.000	0.000

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
125	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
250	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
500	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.000	0.000	0.000	0.060	0.014	0.142
125	0.000	0.000	0.000	0.013	0.014	0.049
250	0.000	0.000	0.000	0.008	0.005	0.024
500	0.000	0.000	0.000	0.002	0.001	0.006
1000	0.000	0.000	0.000	0.001	0.000	0.001

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.046,0.077]	[0.008,0.023]	[0.121,0.165]
125	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.007,0.022]	[0.008,0.023]	[0.036,0.064]
250	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.003,0.016]	[0.002,0.012]	[0.015,0.036]
500	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.007]	[0.000,0.006]	[0.002,0.013]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.006]	[0.000,0.004]	[0.000,0.006]

Table A.17.: Empirical probabilities and confidence intervals of (BlkFuA) being true for $A^\sharp$ and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded and the orientation of the real sources was always chosen normal to the cortical surface.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.012	0.022	0.025	0.032	0.028	0.184
125	0.001	0.005	0.004	0.006	0.004	0.045
250	0.001	0.000	0.001	0.003	0.001	0.022
500	0.000	0.002	0.001	0.000	0.000	0.001
1000	0.000	0.000	0.000	0.000	0.000	0.001

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.006,0.021]	[0.014,0.033]	[0.016,0.037]	[0.022,0.045]	[0.019,0.040]	[0.160,0.209]
125	[0.000,0.006]	[0.002,0.012]	[0.001,0.010]	[0.002,0.013]	[0.001,0.010]	[0.033,0.060]
250	[0.000,0.006]	[0.000,0.004]	[0.000,0.006]	[0.001,0.009]	[0.000,0.006]	[0.014,0.033]
500	[0.000,0.004]	[0.000,0.007]	[0.000,0.006]	[0.000,0.004]	[0.000,0.004]	[0.000,0.006]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.006]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.000	0.000	0.000	0.000	0.000	0.000
125	0.000	0.000	0.000	0.000	0.000	0.000
250	0.000	0.000	0.000	0.000	0.000	0.000
500	0.000	0.000	0.000	0.000	0.000	0.000
1000	0.000	0.000	0.000	0.000	0.000	0.000

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
125	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
250	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
500	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]
1000	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]	[0.000,0.004]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.055	0.063	0.085	0.272	0.216	0.669
125	0.026	0.044	0.045	0.178	0.111	0.375
250	0.009	0.013	0.017	0.078	0.045	0.176
500	0.000	0.003	0.002	0.041	0.020	0.075
1000	0.001	0.004	0.005	0.013	0.005	0.029

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.042,0.071]	[0.049,0.080]	[0.068,0.104]	[0.245,0.301]	[0.191,0.243]	[0.639,0.698]
125	[0.017,0.038]	[0.032,0.059]	[0.033,0.060]	[0.155,0.203]	[0.092,0.132]	[0.345,0.406]
250	[0.004,0.017]	[0.007,0.022]	[0.010,0.027]	[0.062,0.096]	[0.033,0.060]	[0.153,0.201]
500	[0.000,0.004]	[0.001,0.009]	[0.000,0.007]	[0.030,0.055]	[0.012,0.031]	[0.059,0.093]
1000	[0.000,0.006]	[0.001,0.010]	[0.002,0.012]	[0.007,0.022]	[0.002,0.012]	[0.020,0.041]

Table A.18.: Empirical probabilities and confidence intervals of (BlkFuB)/(SSC) being true for A and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded and the orientation of the real sources was always chosen normal to the cortical surface.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.648	0.602	0.602	0.606	0.616	0.489
125	0.353	0.294	0.282	0.356	0.368	0.301
250	0.135	0.122	0.105	0.214	0.220	0.183
500	0.051	0.037	0.032	0.113	0.133	0.086
1000	0.015	0.004	0.003	0.055	0.074	0.036

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.617,0.678]	[0.571,0.632]	[0.571,0.632]	[0.575,0.636]	[0.585,0.646]	[0.458,0.520]
125	[0.323,0.384]	[0.266,0.323]	[0.254,0.311]	[0.326,0.387]	[0.338,0.399]	[0.273,0.330]
250	[0.114,0.158]	[0.102,0.144]	[0.087,0.126]	[0.189,0.241]	[0.195,0.247]	[0.159,0.208]
500	[0.038,0.067]	[0.026,0.051]	[0.022,0.045]	[0.094,0.134]	[0.113,0.156]	[0.069,0.105]
1000	[0.008,0.025]	[0.001,0.010]	[0.001,0.009]	[0.042,0.071]	[0.059,0.092]	[0.025,0.049]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	0.996	0.982
250	0.799	0.774	0.779	0.721	0.494	0.543
500	0.304	0.242	0.237	0.268	0.120	0.162
1000	0.130	0.075	0.093	0.147	0.044	0.067

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.990,0.999]	[0.972,0.989]
250	[0.773,0.823]	[0.747,0.800]	[0.752,0.804]	[0.692,0.749]	[0.463,0.525]	[0.512,0.574]
500	[0.276,0.334]	[0.216,0.270]	[0.211,0.265]	[0.241,0.297]	[0.101,0.142]	[0.140,0.186]
1000	[0.110,0.152]	[0.059,0.093]	[0.076,0.113]	[0.126,0.170]	[0.032,0.059]	[0.052,0.084]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	0.941	0.909	0.929	0.941	0.952	0.921
500	0.686	0.653	0.624	0.778	0.746	0.692
1000	0.461	0.416	0.430	0.575	0.595	0.518

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.925,0.955]	[0.889,0.926]	[0.911,0.944]	[0.925,0.955]	[0.937,0.964]	[0.903,0.937]
500	[0.656,0.715]	[0.623,0.683]	[0.593,0.654]	[0.751,0.803]	[0.718,0.773]	[0.662,0.721]
1000	[0.430,0.492]	[0.385,0.447]	[0.399,0.461]	[0.544,0.606]	[0.564,0.626]	[0.487,0.549]

Table A.19.: Empirical probabilities and confidence intervals of (BlkFuB)/(SSC) being true for $A^\sharp$ and $k = 2$. Computed from 1000 samples $u^\dagger$ for all combinations of head model and N. Solutions $u^\dagger$ with neighboring sources were excluded and the orientation of the real sources was always chosen normal to the cortical surface.

(a) EEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	0.695	0.785	0.780	0.910	0.917	0.972
125	0.474	0.533	0.512	0.627	0.613	0.668
250	0.250	0.296	0.297	0.366	0.363	0.446
500	0.139	0.156	0.154	0.202	0.220	0.248
1000	0.060	0.082	0.073	0.088	0.095	0.105

(b) EEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.665,0.723]	[0.758,0.810]	[0.753,0.805]	[0.891,0.927]	[0.898,0.933]	[0.960,0.981]
125	[0.443,0.505]	[0.502,0.564]	[0.481,0.543]	[0.596,0.657]	[0.582,0.643]	[0.638,0.697]
250	[0.223,0.278]	[0.268,0.325]	[0.269,0.326]	[0.336,0.397]	[0.333,0.394]	[0.415,0.477]
500	[0.118,0.162]	[0.134,0.180]	[0.132,0.178]	[0.178,0.228]	[0.195,0.247]	[0.222,0.276]
1000	[0.046,0.077]	[0.066,0.101]	[0.058,0.091]	[0.071,0.107]	[0.078,0.115]	[0.087,0.126]

(c) MEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	0.999	1.000
250	0.960	0.952	0.961	0.937	0.718	0.865
500	0.642	0.620	0.615	0.537	0.367	0.475
1000	0.310	0.344	0.359	0.256	0.146	0.197

(d) MEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.994,1.000]	[0.996,1.000]
250	[0.946,0.971]	[0.937,0.964]	[0.947,0.972]	[0.920,0.951]	[0.689,0.746]	[0.842,0.886]
500	[0.611,0.672]	[0.589,0.650]	[0.584,0.645]	[0.506,0.568]	[0.337,0.398]	[0.444,0.506]
1000	[0.281,0.340]	[0.315,0.374]	[0.329,0.390]	[0.229,0.284]	[0.125,0.169]	[0.173,0.223]

(e) EMEG empirical probabilities.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	1.000	1.000	1.000	1.000	1.000	1.000
125	1.000	1.000	1.000	1.000	1.000	1.000
250	0.984	0.983	0.988	0.994	0.995	0.999
500	0.904	0.937	0.920	0.965	0.967	0.989
1000	0.748	0.826	0.854	0.885	0.896	0.938

(f) EMEG confidence intervals.

N	HM1	HM2	HM3	HM4	HM 5	HM6
62	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
125	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]	[0.996,1.000]
250	[0.974,0.991]	[0.973,0.990]	[0.979,0.994]	[0.987,0.998]	[0.988,0.998]	[0.994,1.000]
500	[0.884,0.922]	[0.920,0.951]	[0.901,0.936]	[0.952,0.976]	[0.954,0.977]	[0.980,0.994]
1000	[0.720,0.775]	[0.801,0.849]	[0.831,0.875]	[0.864,0.904]	[0.875,0.914]	[0.921,0.952]

Bibliography

Adler, S.L., *Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions*, Physical Review D: Particles and fields (1981), 23: pp. 2901–2904. 117

Agapiou, S., Larsson, S. and Stuart, A.M., *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stochastic Processes and their Applications (2013), 123(10): pp. 3828 – 3860. 76

Almeida, M.S.C. and Figueiredo, M.a.T., *Parameter estimation for blind and non-blind deblurring using residual whiteness measures.*, IEEE Transactions on Image Processing (2013), 22(7): pp. 2751–63. 220

Ambrosio, L., Gigli, N. and Savaré, G., *Gradient Flows in Metric Spaces and in the Spaces of Probability Measures*, Birkhäuser, Basel, 2nd edition (2008). 35, XVII

Andrieu, C. and Thoms, J., *A tutorial on adaptive MCMC*, Computational Statistics (2008), 18(4): pp. 343–373. 90

Arridge, S.R., *Optical tomography in medical imaging*, Inverse Problems (1999), 15(2): p. R41. 23

Arridge, S.R., Kaipio, J.P., Kolehmainen, V., Schweiger, M., Somersalo, E., Tarvainen, T. and Vauhkonen, M., *Approximation errors and model reduction with an application in optical diffusion tomography*, Inverse Problems (2006), 22(1): p. 175. 79

Arridge, S.R. and Schotland, J.C., *Optical tomography: forward and inverse problems*, Inverse Problems (2009), 25(12): p. 123010. 23

Aubert, G. and Kornprobst, P., *Mathematical Problems in Image Processing*, volume 147 of *Applied Mathematical Sciences*, Springer New York, 2nd edition (2006). 164

Aydin, U., Vorwerk, J., Küpper, P., Heers, M., Kugel, H., Galka, A., Hamid, L., Wellmer, J., Kellinghaus, C., Rampp, S. and Wolters, C.H., *Combining EEG and MEG for the Reconstruction of Epileptic Activity Using a Calibrated Realistic Volume Conductor Model*, PLoS ONE (2014), 9(3): p. e93154. 31, 165, 173, 192

Babiloni, F., Babiloni, C., Carducci, F., Romani, G.L., Rossini, P.M., Angelone, L.M. and Cincotti, F., *Multimodal integration of EEG and MEG data: a simulation study with variable signal-to-noise ratio and number of sensors.*, Human Brain Mapping (2004), 22(1): pp. 52–62. 176

Baillet, S., Garnero, L., Marin, G. and Hugonin, J.P., *Combined MEG and EEG Source Imaging by Minimization of Mutual Information*, IEEE Transactions on Biomedical Engineering (1999), 46: pp. 522–534. 176

Bardsley, J., Calvetti, D. and Somersalo, E., *Hierarchical regularization for edge-preserving reconstruction of PET images*, Inverse Problems (2010), 26: p. 035010 (16pp). 60, 146

Basser, P. and Pierpaoli, C., *Microstructural and physiological features of tissues elucidated by quantitive-diffusion-tensor MRI*, Journal of Magnetic Resonance. Series B (1996), 111: pp. 209–219. 165

Baumann, S.B., Wozny, D.R., Kelly, S.K. and Meno, F.M., *The electrical conductivity of human cerebrospinal fluid at body temperature.*, IEEE Transactions on Biomedical Engineering (1997), 44(3): pp. 220–3. 168

Benning, M., *Singular Regularization of Inverse Problems*, Ph.D. thesis, University of Muenster (2011). 70, 74, 210, I, III

Benning, M., Brune, C., Burger, M. and Müller, J., *Higher-Order TV Methods—Enhancement via Bregman Iteration*, Journal of Scientific Computing (2013), 54(2-3): pp. 269–310. 219

Berkels, B., Burger, M., Droske, M., Nemitz, O. and Rumpf, M., *Cartoon extraction based on anisotropic image classification*, in *Vision, Modeling, and Visualization Proceedings* (2006) pp. 293–300. 46

Bissantz, N., Hohage, T. and Munk, A., *Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise*, Inverse Problems (2004), 20(6): p. 1773. 73

Bissantz, N., Hohage, T., Munk, A. and Ruymgaart, F., *Convergence rates of general regularization methods for statistical inverse problems and applications*, SIAM Journal on Numerical Analysis (2007), 45(6): pp. 2610–2636. 74

Box, G.E.P. and Muller, M.E., *A Note on the Generation of Random Normal Deviates*, The Annals of Mathematical Statistics (1958), 29(2): pp. 610–611. 84

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J., *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Foundations and Trends in Machine Learning (2011), 3(1): pp. 1–122. 107, 109, 111, 138

Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge University Press, New York (2004). 106, 121, I

Brazier, M.A.B., *A study of the electric field at the surface of the head*, Electroencephalography and Clinical Neurophysiology (1949), pp. 38–52. 26

Bregman, L., *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics (1967), 7(3): pp. 200 – 217. 110

Brooks, S. and Roberts, G., *Assessing convergence of Markov chain Monte Carlo algorithms*, Computational Statistics (1998), 8(4): pp. 319–335. 93

Buchner, H., Adams, L., Müller, A., Ludwig, I., Knepper, A., Thron, A., Niemann, K. and Scherg, M., *Somatotopy of human hand somatosensory cortex revealed by dipole source analysis of early somatosensory evoked potentials and 3D-NMR tomography*, Electroencephalography and Clinical Neurophysiology (1995), 96(2): pp. 121–134. 183, 189

Buchner, H., Fuchs, M., Wischmann, H.A., Dössel, O., Ludwig, I., Knepper, A. and Berg, P., *Source analysis of median nerve and finger stimulated somatosensory evoked potentials: Multichannel simultaneous recording of electric and magnetic fields combined with 3d-MR tomography*, Brain Topography (1994), 6(4): pp. 299–310. 183, 189

Buchner, H., Knoll, G., Fuchs, M., Rienäcker, A., Beckmann, R., Wagner, M., Silny, J. and Pesch, J., *Inverse localization of electric dipole current sources in finite element models of the human head.*, Electroencephalography and Clinical Neurophysiology (1997), 102: pp. 267–278. 27

Burger, M., Flemming, J. and Hofmann, B., *Convergence rates in ℓ^1-regularization if the sparsity assumption fails*, Inverse Problems (2013), 29(2): p. 025013. 70, III

Burger, M. and Lucka, F., *Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators*, Inverse Problems (2014), 30(11): p. 114004. XXIII

Burger, M. and Osher, S., *Convergence rates of convex variational regularization*, Inverse Problems (2004), 20: pp. 1411–1421. 74, 210, III

Burger, M. and Osher, S., *A Guide to the TV Zoo*, in *Level Set and PDE Based Reconstruction Methods in Imaging*, Lecture Notes in Mathematics, pp. 1–70, Springer International Publishing (2013). 45, 49, 208

Burger, M., Resmerita, E. and He, L., *Error estimation for Bregman iterations and inverse scale space methods in image restoration*, Computing (2007), 81(2-3): pp. 109–135. 70, 210, III

Buzug, T., *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*, Springer (2008). 22

Calvetti, D., Hakula, H., Pursiainen, S. and Somersalo, E., *Conditionally Gaussian hypermodels for cerebral source localization.*, SIAM Journal on Imaging Sciences (2009), 2(3): pp. 879–909. 31

Calvetti, D. and Somersalo, E., *A Gaussian hypermodel to recover blocky objects*, Inverse Problems (2007), 23(2): pp. 733–754. 60, 64, 146, 147

Calvetti, D. and Somersalo, E., *Hypermodels in the Bayesian imaging framework*, Inverse Problems (2008), 24(3): p. 034013 (20pp). 60, 146

Candes, E., Romberg, J. and Tao, T., *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory (2006), 52(2): pp. 489–509. 10, 73

Chang, W.T., Ahlfors, S.P. and Lin, F.H., *Sparse current source estimation for MEG using loose orientation constraints*, Human Brain Mapping (2013), 34(9): pp. 2190–2201. 201

Chang, W.T., Nummenmaa, A., Hsieh, J.C. and Lin, F.H., *Spatially sparse source cluster modeling by compressive neuromagnetic tomography*, NeuroImage (2010), 53(1): pp. 146–160. 201

Chen, D. and Plemmons, R.J., *Nonnegativity constraints in numerical analysis*, in *in A. Bultheel and R. Cools (Eds.), Symposium on the Birth of Numerical Analysis*, World Scientific Press, Singapore (2009) . 110

Chopin, N., *Fast simulation of truncated Gaussian distributions*, Statistics and Computing (2011), 21(2): pp. 275–288. 100

Christen, J.A. and Fox, C., *MCMC using an Approximation,*, Journal of Computational and Graphical Statistics (2005), 14(4): pp. 795–810. 91

Colton, D. and Kress, R., *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer Berlin Heidelberg (1992). 13

Combettes, P. and Pesquet, J.C., *Proximal splitting methods in signal processing*, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, editors H.H. Bauschke, R.S. Burachik, P.L. Combettes, V. Elser, D.R. Luke and H. Wolkowicz, Springer Optimization and Its Applications, pp. 185–212, Springer New York (2011). 108

Comelli, S., *A Novel Class of Priors for Edge-Preserving Methods in Bayesian Inversion*, Master's thesis, University of Milan, Italy (2011). 75, 126, 127, 144

Cormen, T.H., Stein, C., Rivest, R.L. and Leiserson, C.E., *Introduction to Algorithms*, MIT Press and McGraw-Hill, 2nd edition (2001). 170

Cotter, S., Roberts, G., Stuart, A. and White, D., *MCMC methods for functions: modifying old algorithms to make them faster*, Statistical Science (2013), 28: pp. 424–446. 75

Cowles, M. and Carlin, B., *Markov chain Monte Carlo convergence diagnostics: a comparative review*, Journal of the American Statistical Association (1996), 91(434): pp. 883–904. 93

Cui, T., Fox, C. and O'Sullivan, M.J., *Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm*, Water Resources Research (2011), 47. 64, 79, 91

Dannhauer, M., Lanfer, B., Wolters, C. and Knösche, T., *Modeling of the human skull in EEG source analysis*, Human Brain Mapping (2011), 32(9): pp. 1383–1399. 168

Dashti, M., Harris, S. and Stuart, A., *Besov priors for Bayesian inverse problems*, Inverse Problems and Imaging (2012), 6: pp. 183–200. 76

Dashti, M., Law, K.J.H., Stuart, A.M. and Voss, J., *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems (2013), 29(9): p. 095017. 76

Dassios, G., *Electric and Magnetic Activity of the Brain in Spherical and Ellipsoidal Geometry*, in *Mathematical Modeling in Biomedical Imaging I*, editor H. Ammari, Lecture Notes in Mathematics, pp. 133–202, Springer Berlin Heidelberg (2009). 167

Dassios, G. and Fokas, A.S., *The definite non-uniqueness results for deterministic EEG and MEG data*, Inverse Problems (2013), 29(6): p. 065012. 30

de Munck, J.C., van Dijk, B.W. and Spekreijse, H., *Mathematical dipoles are adequate to describe realistic generators of human brain activity.*, IEEE Transactions on Biomedical Engineering (1988), 35(11): pp. 960–966. 26

Deledalle, C.A., Vaiter, S., Peyre, G., Fadili, J. and Dossal, C., *Unbiased risk estimation for sparse analysis regularization*, in *2012 19th IEEE International Conference on Image Processing*, IEEE (2012) pp. 3053–3056. 220

Dijkstra, E., *A note on two problems in connexion with graphs*, Numerische Mathematik (1959), 1(1): pp. 269–271. 170

Donoho, D.L., *Compressed sensing*, IEEE Transactions on Information Theory (2006), 52(4): pp. 1289–1306. 10, 73

Eldar, Y.C., *Generalized SURE for Exponential Families: Applications to Regularization*, IEEE Transactions on Signal Processing (2009), 57(2): pp. 471–481. 220

Eldar, Y.C., Kuppinger, P. and Bolcskei, H., *Block-Sparse Signals: Uncertainty Relations and Efficient Recovery*, IEEE Transactions on Signal Processing (2010), 58(6): pp. 3042–3054. 72

Eldar, Y.C. and Kutyniok, G. (editors) *Compressed Sensing - Theory and Applications*, Cambridge University Press, New York, 1st edition (2012). 73

Eldar, Y.C. and Mishali, M., *Robust Recovery of Signals From a Structured Union of Subspaces*, IEEE Transactions on Information Theory (2009), 55(11): pp. 5302–5316. 72

Eltoft, T., Kim, T. and Lee, T.W., *On the multivariate Laplace distribution*, IEEE Signal Processing Letters (2006), 13(5): pp. 300–303. 48

Engl, H., Hanke-Bourgeois, M. and Neubauer, A., *Regularization of Inverse Problems*, Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht (1996). 11, 220

Engl, H.W., Hofinger, A. and Kindermann, S., *Convergence rates in the Prokhorov metric for assessing uncertainty in ill-posed problems*, Inverse Problems (2005), 21(1): p. 399. 74

Esser, E., *Applications of Lagrangian-based Alternating Direction Methods and connections to Split Bregman*, Technical report, University of California, Irvine (2009). 111

Evans, L.C. and Gariepy, R.F., *Measure Theory and Fine Properties of Functions*, volume 5, CRC press (1991). 205

Evans, S.N. and Stark, P.B., *Inverse problems as statistics*, Inverse Problems (2002), 18(4): p. R55. 11

Fornasier, M., *Theoretical Foundations and Numerical Methods for Sparse Recovery*, De Gruyter, Berlin, Boston (2010). 73

Foucart, S. and Rauhut, H., *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, Basel (2013). 10, 64, 67, 73, IV

Frikel, J., *Reconstructions in limited angle X-ray tomography: Characterization of classical reconstructions and adapted curvelet sparse regularization*, Ph.D. thesis, Technische Universität München (2013). 19

Fuchs, J., *On sparse representations in arbitrary redundant bases*, IEEE Transactions on Information Theory (2004), 50(6): pp. 1341–1344. 67

Fuchs, M., Drenckhahn, R., Wischmann, H. and Wagner, M., *An improved boundary element method for realistical volume conductor modeling.*, IEEE Transactions on Biomedical Engineering (1998a), 45(8): pp. 980–997. 174, 176

Fuchs, M., Wagner, M., Wischmann, H.A., Köhler, T., Theissen, A., Drenckhahn, R. and Buchner, H., *Improving source reconstructions by combining bioelectric and biomagnetic data*, Electroencephalography and Clinical Neurophysiology (1998b), 107(2): pp. 93–111. 176, 183, 189

Garrigues, P. and Olshausen, B.A., *Group Sparse Coding with a Laplacian Scale Mixture Prior*, in *NIPS'10* (2010) pp. 676–684. 58

Gelman, A., *Prior distributions for variance parameters in hierarchical models*, Bayesian Analysis (2006), 1(3): pp. 515–533. 56

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B., *Bayesian Data Analysis*, Chapman and Hall/CRC Texts in Statistical Science, CRC Press, 2nd edition (2003). 62, 79

Geman, S. and Geman, D., *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.*, IEEE Transactions on Pattern Analysis and Machine Intelligence (1984), PAMI-6(6): pp. 721–741. 91

Giryes, R., Elad, M. and Eldar, Y.C., *The projected GSURE for automatic parameter tuning in iterative shrinkage methods*, Applied and Computational Harmonic Analysis (2011), 30(3): pp. 407–422. 220

Gnedenko, B. and Kolmogorov, A., *Limit distributions for sums of independent random variables*, Addison-Wesley, Cambridge (1954). 148

Goldstein, T. and Osher, S., *The Split Bregman method for L1-regularized problems*, SIAM Journal on Imaging Sciences (2009), 2: pp. 323–343. 111, 138, III

Goodman, J. and Sokal, A.D., *Multigrid Monte Carlo method. Conceptual foundations*, Physical Review D: Particles and Fields (1989), 40: pp. 2035–2071. 93

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M. and Kowalski, M., *Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations*, NeuroImage (2013), 70(0): pp. 410 – 422. 201

Grasmair, M., *Generalized Bregman distances and convergence rates for non-convex regularization methods*, Inverse Problems (2010), 26(11): p. 115014. 74

Gribonval, R., *Should Penalized Least Squares Regression be Interpreted as Maximum A Posteriori Estimation?*, IEEE Transactions on Signal Processing (2011), 59(5): pp. 2405–2410. 204

Gribonval, R., Cevher, V. and Davies, M., *Compressible Distributions for High-Dimensional Statistics*, IEEE Transactions on Information Theory (2012), 58(8): pp. 5016–5034. 205

Gribonval, R. and Machart, P., *Reconciling "priors" & "priors" without prejudice?*, Research Report RR-8366, INRIA (2013). 204

Gross, J., Baillet, S., Barnes, G.R., Henson, R.N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., Parkkonen, L., Taylor, J.R., van Wassenhove, V., Wibral, M. and Schoffelen, J.M., *Good practice for conducting and reporting MEG research*, NeuroImage (2013), 65(0): pp. 349 – 363. 180

Haario, H., Laine, M., Lehtinen, M., Saksman, E. and Tamminen, J., *Markov Chain Monte Carlo Methods for High Dimensional Inversion in Remote Sensing*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2004), 66(3): pp. 591–607. 64, 90

Haario, H., Laine, M., Mira, A. and Saksman, E., *DRAM: Efficient adaptive MCMC*, Computational Statistics (2006), 16(4): pp. 339–354. 64, 90

Haario, H., Saksman, E. and Tamminen, J., *An Adaptive Metropolis Algorithm*, Bernoulli (2001), 7(2): pp. 223–242. 90

Haario, H., Saksman, E. and Tamminen, J., *Componentwise adaptation for high dimensional MCMC*, Computational Statistics (2005), 20(2): pp. 265–273. 90

Hadamard, J., *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New York (1923). 6

Hämäläinen, K., Harhanen, L., Hauptmann, A., Kallonen, A., Niemi, E. and Siltanen, S., *Total variation regularization for large-scale X-ray tomography*, International Journal of Tomography and Simulation (2014), 25(1): pp. 1–25. 22

Hämäläinen, K., Kallonen, A., Kolehmainen, V., Lassas, M., Niinimaki, K. and Siltanen, S., *Sparse Tomography*, SIAM Journal on Scientific Computing (2013), 35(3): pp. B644–B665. 22, 75, 149, 157, VI

Hämäläinen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J. and Lounasmaa, O.V., *Magnetoencephalography - Theory, instrumentation, and applications to noninvasive studies of the working human brain*, Reviews of Modern Physics (1993), 65(2): pp. 413–497. 168, 189

Hairer, M., Stuart, A. and Voss, J., *Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods*, in *The Oxford Handbook of Nonlinear Filtering*, editors D. Crisan and B. Rozovsky, pp. 833–873, Oxford University Press (2011). 76

Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Mathematics and Statistics, Springer New York, 2nd edition (2009). 78

Hastings, W.K., *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika (1970), 57(1): pp. 97–109. 88

Haufe, S., Nikulin, V.V., Ziehe, A., Müller, K.R. and Nolte, G., *Combining sparsity and rotational invariance in EEG/MEG source reconstruction*, NeuroImage (2008), 42(2): pp. 726 – 738. 201, XVII

Heiskala, J., Kolehmainen, V., Tarvainen, T., Kaipio, J.P. and Arridge, S.R., *Approximation error method can reduce artifacts due to scalp blood flow in optical brain activation imaging*, Journal of Biomedical Optics (2012), 17(9): pp. 096012–1–096012–7. 79

Helin, T., *Discretization and Bayesian Modeling in Inverse Problems and Imaging*, Ph.D. thesis, Aalto University School of Science and Technology (2010a). 76

Helin, T., *On infinite-dimensional hierarchical probability models in statistical inverse problems*, Inverse Problems and Imaging (2010b), 3: pp. 567–597. 60, 76, 146

Helin, T. and Lassas, M., *Hierarchical Models in Statistical Inverse Problems and the Mumford–Shah Functional*, Technical Report arXiv:0908.3396 (2009). 60, 76, 146

Helin, T., Lassas, M. and Siltanen, S., *Infinite Photography: New Mathematical Model for High-Resolution Images*, Journal of Mathematical Imaging and Vision (2010), 36(2): pp. 140–158. 39

Henson, R.N., Flandin, G., Friston, K.J. and Mattout, J., *A Parametric Empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction.*, Human Brain Mapping (2010), 31(10): pp. 1512–1531. 79

Henson, R.N., Mattout, J., Phillips, C. and Friston, K.J., *Selecting forward models for MEG source-reconstruction using model-evidence.*, NeuroImage (2009a), 46(1): pp. 168–176. 78

Henson, R.N., Mouchlianitis, E. and Friston, K.J., *MEG and EEG data fusion: simultaneous localisation of face-evoked responses.*, NeuroImage (2009b), 47(2): pp. 581–589. 79

Hofinger, A., *Ill-posed problems: Extending the deterministic theory to a stochastic setup*, Ph.D. thesis, Universität Linz (2006). 74

Hofinger, A. and Pikkarainen, H.K., *Convergence rate for the Bayesian approach to linear inverse problems*, Inverse Problems (2007), 23(6): p. 2469. 74

Hofinger, A. and Pikkarainen, H.K., *Convergence rates for linear inverse problems in the presence of an additive normal noise*, Stochastic Analysis and Applications (2009), 27(2): pp. 240–257. 74

Huang, M.X., Dale, A.M., Song, T., Halgren, E., Harrington, D.L., Podgorny, I., Canive, J.M., Lewis, S. and Lee, R.R., *Vector-based spatial–temporal minimum L1-norm solution for MEG*, NeuroImage (2006), 31(3): pp. 1025 – 1037. 201

Hufnagel, A., Elger, C., Pels, H., Zentner, J., Wolf, H., Schramm, J. and Wiestler, O., *Prognostic significance of ictal and interictal epileptiform activity in temporal lobe epilepsy.*, Epilepsia (1994), 35(6): pp. 1146–1153. 172

Janssen, A.M., Rampersad, S.M., Lucka, F., Lanfer, B., Lew, S., Aydin, U., Wolters, C.H., Stegeman, D.F. and Oostendorp, T.F., *The influence of sulcus width on simulated electric fields induced by transcranial magnetic stimulation*, Physics in Medicine and Biology (2013), 58(14): p. 4881. 165, XX, XXIII

Janszky, J., Jokeit, H., Schulz, R., Hoppe, M. and Ebner, A., *EEG predicts surgical outcome in lesional frontal lobe epilepsy.*, Neurology (2000), 54(7): pp. 1470–1476. 172

Jaynes, E. and Bretthorst, G., *Probability Theory: The Logic of Science*, Cambridge University Press (2003). 79

Kaipio, J.P. and Somersalo, E., *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*, Springer New York (2005). 43, 62, 79, 85, 145, 174, 220

Kaipio, J.P. and Somersalo, E., *Statistical inverse problems: Discretization, model reduction and inverse crimes*, Journal of Computational and Applied Mathematics (2007), 198(2): pp. 493–504. 13, 79

Kay, S., *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice-Hall PTR (1993). 62, 79

Kekkonen, H., Lassas, M. and Siltanen, S., *Analysis of regularized inversion of data corrupted by white Gaussian noise*, Inverse Problems (2014), 30(4): p. 045009. 39, 74

Kiebel, S.J., Garrido, M.I., Moran, R., Chen, C.C. and Friston, K.J., *Dynamic causal modeling for EEG and MEG*, Human Brain Mapping (2009), 30(6): pp. 1866–1876. 172

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., *Optimization by Simulated Annealing*, Science (1983), 220(4598): pp. 671–680. 112

Kirsch, A., *An Introduction to the Mathematical Theory of Inverse Problems*, Springer New York (1996). 11

Klenke, A., *Probability Theory: A Comprehensive Course*, Springer London, 1st edition (2008). 35, 37, 83, 85, 148

Kolehmainen, V., Lassas, M., Niinimäki, K. and Siltanen, S., *Sparsity-promoting Bayesian inversion*, Inverse Problems (2012), 28(2): p. 025005 (28pp). 75, 126, 127, 148, 149, 220

Kolehmainen, V., Siltanen, S., Järvenpää, S., Kaipio, J.P., Koistinen, P., Lassas, M., Pirttilä, J. and Somersalo, E., *Statistical inversion for medical x-ray tomography with few radiographs: II. Application to dental radiology.*, Physics in Medicine and Biology (2003), 48(10): pp. 1465–90. 153

Kolehmainen, V., Tarvainen, T., Arridge, S.R. and Kaipio, J.P., *Marginalization of uninteresting distributed parameters in inverse problems; Application to diffuse optical tomography*, International Journal for Uncertainty Quantification (2011), 1(1): pp. 1–17. 79

Lanfer, B., *Automatic Generation of Volume Conductor Models of the Human Head for EEG Source Analysis*, Ph.D. thesis, University of Muenster (2014). 164

Lanfer, B., Scherg, M., Dannhauer, M., Knösche, T., Burger, M. and Wolters, C., *Influences of skull segmentation inaccuracies on EEG source analysis*, NeuroImage (2012), 62(1): pp. 418 – 431. 78, 195

Lassas, M., Saksman, E. and Siltanen, S., *Discretization invariant Bayesian inversion and Besov space priors.*, Inverse Problems and Imaging (2009), 3(1): pp. 87–122. 47, 75, 76, 148, 149

Lassas, M. and Siltanen, S., *Can one use total variation prior for edge-preserving Bayesian inversion?*, Inverse Problems (2004), 20: pp. 1537–1563. 14, 74, 76, 126, 127, 140, 141, 144, 145

Latuszynski, K., Roberts, G.O. and Rosenthal, J.S., *Adaptive Gibbs samplers and related MCMC methods*, The Annals of Applied Probability (2013), 23(1): pp. 66–98. 90, 93, 221

Lew, S., Wolters, C.H., Dierkes, T., Röer, C. and MacLeod, R.S., *Accuracy and run-time comparison for different potential approaches and iterative solvers in finite element method based EEG source analysis.*, Applied Numerical Mathematics (2009), 59(8): pp. 1970–1988. 171

Lin, F.H., Belliveau, J.W., Dale, A.M. and Hämäläinen, M.S., *Distributed current estimates using cortical orientation constraints*, Human Brain Mapping (2006), 27(1): pp. 1–13. 168

Lipponen, A., Kolehmainen, V., Romakkaniemi, S. and Kokkola, H., *Correction of approximation errors with Random Forests applied to modelling of cloud droplet formation*, Geoscientific Model Development (2013), 6(6): pp. 2087–2098. 79

Liu, A.K., Dale, A.M. and Belliveau, J.W., *Monte Carlo simulation studies of EEG and MEG localization accuracy.*, Human Brain Mapping (2002), 16(1): pp. 47–62. 176

Liu, J., *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics, Springer New York (2008). 85, 89, 93, 105, 113, 114, 222

Liu, J. and Sabatti, C., *Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation*, Biometrika (2000), 87(2): pp. 353–369. 93

Lorenz, D.A., Schiffler, S. and Trede, D., *Beyond convergence rates: exact recovery with the tikhonov regularization with sparsity constraints*, Inverse Problems (2011), 27(8): p. 085009. 74

Louchet, C., *Variational and Bayesian models for image denoising: from total variation towards non-local means*, Ph.D. thesis, Université Paris Descartes (2008). 141

Louchet, C. and Moisan, L., *Posterior Expectation of the Total Variation model: Properties and Experiments*, SIAM Journal on Imaging Sciences (2013), 6(4): pp. 2640–2684. 157, 204

Lucka, F., *Hierarchical Bayesian Approaches to the Inverse Problem of EEG/MEG Current Density Reconstruction*, Master's thesis, University of Muenster (2011). 24, 30, 44, 57, 59, 60, 85, 106, 115, 116, 165, 172, 204, VIII, XVII, XVIII

Lucka, F., *Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors*, Inverse Problems (2012), 28(12): p. 125012. 100, 120, 123, 127, 128, 133, 141, 144, XXII

Lucka, F., Pursiainen, S., Burger, M. and Wolters, C.H., *Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents*, NeuroImage (2012), 61(4): pp. 1364–1382. 28, 115, 165, 166, 172, 173, 174, 193, 204, XVII

MacKay, D., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 1st edition (2003). 60

Makeig, S., Bell, A., Jung, T. and Sejnowski, T., *Independent component analysis of electroencephalographic data*, Advances in Neural Information Processing Systems (1996), pp. 145–151. 192

Marsaglia, G. and Tsang, W.W., *A Simple Method for Generating Gamma Variables*, ACM Transactions on Mathematical Software (2000), 26(3): pp. 363–372. 84

Mathé, P. and Tautenhahn, U., *Regularization under general noise assumptions*, Inverse Problems (2011), 27(3): p. 035016. 74

Matsumoto, M. and Nishimura, T., *Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator*, ACM Transactions on Modeling and Computer Simulation (1998), 8(1): pp. 3–30. 83

Matsuura, K. and Okabe, Y., *Selective minimum-norm solution of the biomagnetic inverse problem.*, IEEE Transactions on Biomedical Engineering (1995), 42(6): pp. 608–15. 201

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E., *Equation of state calculations by fast computing machines*, Journal of Physical Chemistry (1953), 21: pp. 1087–1092. 88, 117

Michel, C.M. and Murray, M.M., *Towards the utilization of EEG as a brain imaging tool*, NeuroImage (2012), 61(2): pp. 371 – 385. 3, 31

Mira, A., *On Metropolis-Hastings algorithms with delayed rejection*, Metron (2001), LIX(3-4): pp. 231–241. 90

Modersitzki, J., *Numerical Methods for Image Registration*, Oxford University Press (2004). 164

Moeller, M., *Multiscale Methods for Generalized Sparse Recovery and Applications in High Dimensional Imaging*, Ph.D. thesis, University of Muenster (2012). 70, 74, III

Molins, A., Stufflebeam, S.M., Brown, E.N. and Hämäläinen, M., *Quantification of the benefit from integrating MEG and EEG data in minimum l2-norm estimation.*, NeuroImage (2008), 42(3): pp. 1069–1077. 173, 176

Mueller, J.L. and Siltanen, S., *Linear and Nonlinear Inverse Problems with Practical Applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2012). 12

Munck, J. and Peters, M., *A fast method to compute the potential in the multisphere model*, IEEE Transactions on Biomedical Engineering (1993), 40(11): pp. 1166–1174. 167

Murakami, S. and Okada, Y., *Contributions of principal neocortical neurons to magnetoencephalography and electroencephalography signals*, The Journal of Physiology (2006), 575(3): pp. 925–936. 168

Nagarajan, S.S., Portniaguine, O., Hwang, D., Johnson, C. and Sekihara, K., *Controlled support MEG imaging.*, NeuroImage (2006), 33(3): pp. 878–885. 192

Natterer, F., *The Mathematics of Computerized Tomography*, Teubner-Wiley (1986). 18, 22

Neal, P. and Roberts, G., *Optimal scaling for partially updating MCMC algorithms*, Annals of Applied Probability (2006), 16(2): pp. 475–515. 89

Neal, R.M., *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical report (1993). 113

Neal, R.M., *Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation*, Technical Report 9508, Learning in Graphical Models (1995). 118, 119

Neal, R.M., *Slice Sampling*, Annals of Statistics (2003), 31(3): pp. 705–767. 102

Nicolas-Alonso, L.F. and Gomez-Gil, J., *Brain Computer Interfaces, a Review*, Sensors (2012), 12(2): pp. 1211–1279. 31

Niemi, E., Lassas, M. and Siltanen, S., *Dynamic X-ray tomography with multiple sources*, in *Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on* (2013) pp. 618–621. 24

Nissinen, A., *Modelling Errors in Electrical Impedance Tomography*, Ph.D. thesis, University of Eastern Finland (2011). 79

Nissinen, A., Heikkinen, L. and Kaipio, J., *The Bayesian approximation error approach for electrical impedance tomography - experimental results*, Measurement Science and Technology (2008), 19: p. 015501. 79

Nissinen, A., Heikkinen, L., Kolehmainen, V. and Kaipio, J., *Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography*, Measurement Science and Technology (2009), 20: p. 105504. 79

Nissinen, A., Kolehmainen, V. and Kaipio, J., *Compensation of Modelling Errors Due to Unknown Domain Boundary in Electrical Impedance Tomography*, IEEE Transactions on Medical Imaging (2011), 30(2): pp. 231–242. 79

Nunez, P.L. and Srinivasan, R., *Electric Fields of the Brain: The Neurophysics of EEG*, Oxford University Press, 2nd edition (2005). 168

Offtermatt, J. and Kaltenbacher, B., *A convergence analysis of Tikhonov regularization with the Cauchy regularization term*, Technical report, Stuttgart Research Centre for Simulation Technology (SRC SimTech) (2011). 51

Okada, Y.C., Wu, J. and Kyuhou, S., *Genesis of MEG signals in a mammalian CNS structure.*, Electroencephalography and Clinical Neurophysiology (1997), 103(4): pp. 474–85. 168

Oostenveld, R., Fries, P., Maris, E. and Schoffelen, J.M., *FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data*, Computational Intelligence and Neuroscience (2011), 2011: pp. 1:1–1:9. XIX

Oostenveld, R. and Praamstra, P., *The five percent electrode system for high-resolution EEG and ERP measurements*, Clinical Neurophysiology (2001), 112(4): pp. 713 – 719. 198

Opitz, A., Windhoff, M., Heidemann, R.M., Turner, R. and Thielscher, A., *How the brain tissue shapes the electric field induced by transcranial magnetic stimulation*, NeuroImage (2011), 58(3): pp. 849 – 859. 165

Osher, S., Burger, M., Goldfarb, D., Xu, J. and Yin, W., *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling and Simulation (2006), 4(2): pp. 460–489. 49, 111

O'Sullivan, F., *A Statistical Perspective on Ill-Posed Inverse Problems*, Statistical Science (1986), 1(4): pp. pp. 502–518. 11

Ou, W., Hämäläinen, M.S. and Golland, P., *A distributed spatio-temporal EEG/MEG inverse solver*, NeuroImage (2009), 44(3): pp. 932–946. 201

Pantev, C. and Lutkenhoner, B., *Magnetoencephalographic studies of functional organization and plasticity of the human auditory cortex.*, Journal of Clinical Neurophysiology (2000), 17(2): pp. 130–142. 191

Parker, A. and Fox, C., *Sampling Gaussian distributions in Krylov spaces with conjugate gradients*, SIAM Journal on Scientific Computing (2012), 34(3). 120

Parkkonen, L., Fujiki, N. and Mäkelä, J., *Sources of auditory brainstem responses revisited: contribution by magnetoencephalography*, Human Brain Mapping (2009), 30(6): pp. 1772–1782. 169, 172

Pascual-Marqui, R.D., *Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details.*, Methods and Findings in Experimental and Clinical Pharmacology (2002), 24 Suppl D(Suppl D): pp. 5–12. 174

Picton, T., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson, R., Miller, G., Ritter, W., Ruchkin, D., Rugg, M. and Taylor, M., *Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria*, Psychophysiology (2000), 37(2): pp. 127–152. 180

Pieloth, C., Pizarro, J., Knosche, T., Maess, B. and Fuchs, M., *An online system for neuroelectromagnetic source imaging*, in *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013 IEEE 7th International Conference on*, volume 01 (2013) pp. 270–274. 201

Pursiainen, S., Lucka, F. and Wolters, C.H., *Complete electrode model in EEG: relationship and differences to the point electrode model*, Physics in Medicine and Biology (2012), 57(4): pp. 999–1017. 31, 165, 166, XXIII

Radon, J., *Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten*, Berichte über die Verhandlungen der Königlich-Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse (Leipzig: Teubner) (1917), 69: pp. 262–277. 16

Radon, J., *On the Determination of Functions from Their Integral Values along Certain Manifolds*, IEEE Transactions on Medical Imaging (1986), 5(4): pp. 170–176, translation by P.C. Parks. 16

Ramon, C., Schimpf, P.H. and Haueisen, J., *Influence of head models on EEG simulations and inverse source localizations.*, BioMedical Engineering OnLine (2006), 5: p. 10 (13pp). 168

Rampersad, S., Janssen, A., Lucka, F., Aydin, U., Lanfer, B., Lew, S., Wolters, C., Stegeman, D. and Oostendorp, T., *Simulating Transcranial Direct Current Stimulation With a Detailed Anisotropic Human Head Model*, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2014), 22(3): pp. 441–452. 165, XXIII

Robert, C.P. and Casella, G., *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer New York (2005). 85, 105, 113, 114

Roberts, G.O. and Rosenthal, J.S., *Optimal scaling for various Metropolis-Hastings algorithms*, Statistical Science (2001), 16(4): pp. 351–367. 89

Roberts, G.O. and Rosenthal, J.S., *Examples of Adaptive MCMC*, Journal of Computational and Graphical Statistics (2009), 18(2): pp. 349–367. 90

Roberts, G.O. and Sahu, S.K., *Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler*, Journal of the Royal Statistical Society. (1997), 59(2): pp. pp. 291–317. 93

Rudin, L.I., Osher, S. and Fatemi, E., *Nonlinear total variation based noise removal algorithms*, Physica D Nonlinear Phenomena (1992), 60: pp. 259–268. 10, 45

Rullmann, M., Anwander, A., Dannhauer, M., Warfield, S.K., Duffy, F.H. and Wolters, C.H., *EEG source analysis of epileptiform activity using a 1 mm anisotropic hexahedra finite element head model.*, NeuroImage (2009), 44(2): pp. 399–410. 165, 168

Ruthotto, L., Kugel, H., Olesch, J., Fischer, B., Modersitzki, J., Burger, M. and Wolters, C.H., *Diffeomorphic susceptibility artifact correction of diffusion-weighted magnetic resonance images*, Physics in Medicine and Biology (2012), 57(18): p. 5715. 164

Saad, Y., *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2003). 106, 117

Sander, T.H., Knösche, T.R., Schlögl, A., Kohl, F., Wolters, C.H., Haueisen, J. and Trahms, L., *Recent advances in modeling and analysis of bioelectric and biomagnetic sources.*, Biomedizinische Technik / Biomedical engineering (2010), 55(2): pp. 65–76. 169

Sato, M., Yoshioka, T., Kajihara, S., Toyama, K., Goda, N., Doya, K. and Kawato, M., *Hierarchical Bayesian estimation for MEG inverse problem.*, NeuroImage (2004), 23(3): pp. 806–826. 78

Scherg, M. and Buchner, H., *Somatosensory evoked potentials and magnetic fields: separation of multiple source activities*, Physiological Measurement (1993), 14(4A): p. A35. 169

Schiffbauer, H., Berger, M., Ferrari, P., Freudenstein, D., Rowley, H. and Roberts., T., *Preoperative magnetic source imaging for brain tumor surgery: a quantitative comparison with intraoperative sensory and motor mapping.*, Journal of Neurosurgery (2002), 97(6): pp. 1333–1342. 172

Schimpf, P., Ramon, C. and Haueisen, J., *Dipole Models for the EEG and MEG*, IEEE Transactions on Biomedical Engineering (2002), 49(5): pp. 409–418. 27

Schneider, M.K. and Willsky, A.S., *A Krylov Subspace Method for Covariance Approximation and Simulation of Random Processes and Fields*, Multidimensional Systems and Signal Processing (2003), 14(4): pp. 295–318. 120

Schuster, T., Kaltenbacher, B., Hofmann, B. and Kazimierski, K., *Regularization Methods in Banach Spaces*, Radon Series on Computational and Applied Mathematics, De Gruyter, Berlin, Boston (2012). 11, 74, 210

Shepp, L. and Logan, B., *The Fourier reconstruction of a head section*, IEEE Transactions on Nuclear Science (1974), 21(3): pp. 21–43. 21

Siltanen, S., *MAT-52506 Inverse Problems* (2009), lecture notes; retrieved from: matriisi.ee.tut.fi/courses/MAT-52500/IPnotes12.pdf; last accessed: 19.11.2014. 13

Siltanen, S., Kolehmainen, V., Järvenpää, S., Kaipio, J.P., Koistinen, P., Lassas, M., Pirttilä, J. and Somersalo, E., *Statistical inversion for medical x-ray tomography with few radiographs: I. General theory.*, Physics in Medicine and Biology (2003), 48(10): pp. 1437–63. 153, 154

Strobbe, G., van Mierlo, P., Vos, M.D., Mijović, B., Hallez, H., Huffel, S.V., López, J.D. and Vandenberghe, S., *Bayesian model selection of template forward models for EEG source reconstruction*, NeuroImage (2014), 93, Part 1(0): pp. 11 – 22. 78

Stuart, A.M., *Inverse problems: A Bayesian perspective*, Acta Numerica (2010), 19: pp. 451–559. 76, 79

Tanzer, O., Järvenpää, S., Nenonen, J. and Somersalo, E., *Representation of bioelectric current sources using Whitney elements in the finite element method.*, Physics in Medicine and Biology (2005), 50(13): pp. 3023–39. 31

Tarantola, A., *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial Mathematics, Philadelphia, PA, USA (2005). 79

Tarvainen, T., Kolehmainen, V., Pulkkinen, A., Vauhkonen, M., Schweiger, M., Arridge, S.R. and Kaipio, J.P., *An approximation error approach for compensating for modelling errors between the radiative transfer equation and the diffusion approximation in diffuse optical tomography*, Inverse Problems (2010), 26(1): p. 015005. 79

Tarvainen, T., Pulkkinen, A., Cox, B., Kaipio, J. and Arridge, S., *Bayesian Image Reconstruction in Quantitative Photoacoustic Tomography*, IEEE Transactions on Medical Imaging (2013), 32(12): pp. 2287–2298. 79

Tellen, S., *Sparse Reconstruction and Realistic Head Modeling in EEG/MEG*, Master's thesis, University of Muenster (2013). 72, XXII

Thompson, M., *A comparison of methods for computing autocorrelation time*, Arxiv preprint arXiv:1011.0175 (2010). 93, 95

Toussaint, U., *Bayesian inference in physics*, Reviews of Modern Physics (2011), 83(3): pp. 943–999. 78

Trede, D., *Inverse Problems with Sparsity Constraints: Convergence Rates and Exact Recovery*, Ph.D. thesis, University of Bremen (2009). 74

Tropp, J., *Greed is good: algorithmic results for sparse approximation*, IEEE Transactions on Information Theory (2004), 50(10): pp. 2231–2242. 67

Trujillo-Barreto, N.J., Aubert-Vázquez, E. and Valdés-Sosa, P.A., *Bayesian model averaging in EEG/MEG imaging.*, NeuroImage (2004), 21(4): pp. 1300–1319. 79

Uutela, K., Hämäläinen, M. and Somersalo, E., *Visualization of magnetoencephalographic data using minimum current estimates.*, NeuroImage (1999), 10(2): pp. 173–180. 201

Vogel, C.R., *Computational Methods for Inverse Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2002). 11

Vorwerk, J., *Comparison of Numerical Approaches to the EEG Forward Problem*, Master's thesis, University of Muenster (2011). 31, 171

Vorwerk, J., Cho, J.H., Rampp, S., Hamer, H., Knösche, T.R. and Wolters, C.H., *A guideline for head volume conductor modeling in EEG and MEG*, NeuroImage (2014), 100(0): pp. 590 – 607. 31, 165, 166

Vorwerk, J., Clerc, M., Burger, M. and Wolters, C.H., *Comparison of boundary element and finite element approaches to the EEG forward problem*, Biomedizinische Technik / Biomedical engineering (2012), 57. 27, 31, 171

Wagner, M., Fuchs, M. and Kastner, J., *Evaluation of sLORETA in the presence of noise and multiple sources.*, Brain Topography (2004), 16(4): pp. 277–280. 172

Wang, Y., Jiang, X., Yu, B. and Jiang, M., *A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data*, Journal of the American Statistical Association (2013), 108(502): pp. 483–493. 60

Wolff, U., *Monte Carlo errors with less errors*, Computer Physics Communications (2004), 156(2): pp. 143 – 153. 95, 134, XXII

Wolters, C.H., Grasedyck, L. and Hackbusch, W., *Efficient computation of lead field bases and influence matrix for the FEM-based EEG and MEG inverse problem*, Inverse Problems (2004), 20(4): pp. 1099–1116. 171

Wolters, C.H., Köstler, H., Möller, C., Härtlein, J., Grasedyck, L. and Hackbusch, W., *Numerical mathematics of the subtraction method for the modeling of a current dipole in EEG source reconstruction using finite element head models.*, SIAM Journal on Scientific Computing (2007), pp. 24–45. 27

Yu, L., Liu, X., Leng, S., Kofler, J.M., Ramirez-Giraldo, J.C., Qu, M., Christner, J., Fletcher, J.G. and McCollough, C.H., *Radiation dose reduction in computed tomography: techniques and future perspective.*, Imaging in Medicine (2009), 1(1): pp. 65–84. 19

Zhang, X., Burger, M. and Osher, S., *A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration*, Journal of Scientific Computing (2011), 46(1): pp. 20–46. 111