Density and Copula Estimation using Penalized Spline Smoothing

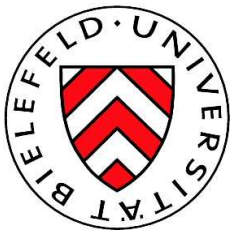
Dissertation

zur Erlangung des Grades eines Doktors

der Wirtschaftswissenschaften (Dr. rer. pol.)

der Fakultät für Wirtschaftswissenschaften

der Universität Bielefeld



vorgelegt von

Dipl.-Wirt. Math. Christian Schellhase

Bielefeld, im Mai 2012

Dekan: Prof. Dr. Herbert Dawid

Gutachter: Prof. Dr. Göran Kauermann (LMU München)

Gutachterin: Prof. Dr. Tatyana Krivobokova (Georg-August Universität Göttingen)

Tag der mündlichen Prüfung: 25.09.2012

Acknowledgments

I would like to thank Prof. Dr. Göran Kauermann (LMU Munich) for his excellent instructions and awesome support. I am thankful for the opportunity to work with him in a most friendly relationship and a very pleasant and productive collaboration. His comprehensive and great knowledge have effectively inspired my statistic understanding and curiosity. I appreciate his extensive statistics ideas and constructive feedbacks. I thank also Prof. Dr. Tatyana Krivobokova (University of Göttingen) as my second supervisor for the expertise about my thesis and Prof. David Ruppert (Cornell University) for the effective and constructive collaboration in the context of copula estimation.

Special thanks go to my wife Maximiliane and my little daughter Marlene, who have appreciatively encouraged and supported me with all their power and love throughout the work on my thesis.

I am grateful for the financal support provided by the Deutsche Forschungsgemeinschaft (DFG Project-Nr. KA 1188/5-1 and KA 1188/5-2).

Many thanks go to my colleagues for talks, ideas and friedly support. Especially to apl. Prof. Dr. Peter Wolf for his indescribable great support in special cases using `R`.

Contents

List of Figures

List of Tables

1 Introduction

Estimating the unknown probability distribution and density functions of univariate or multivariate data is a demanding task in sciences, e.g. statistics or biometrics, for many years. Of course, observed data appear without providing their theoretical distributions. Starting with univariate observed data, it is the aim of density estimation to find any continuous density function $f(\cdot)$, such that

$$\int f(x)\ \mathrm{d}(x) = 1 \tag{1.1}$$

with $f(x) \geq 0$. Hence, a non-negative probability mass is assigned to each observed x. There are parametric and non-parametric approaches to model the density function. Fitting the parameters of any known distribution function to the observed data, using e.g. maximum likelihood theory, is possible, but may be misleading as data usually appear different to any theoretical parametric distribution function, e.g. normal distribution. That is, these approaches estimate the optimal distribution parameters, e.g. mean and variance in this case of the normal distribution. It is the idea of nonparametric estimation approaches to describe the empirical distribution of data without any a priori knowledge of the theoretical distribution. A famous nonparametric estimation method is the kernel density estimation approach, which will also be considered in this thesis.

Usually a univariate analysis of real world phenomena is not satisfying as one is also interested in dependence structures and causal relationships. A first step towards this direction is the extension of univariate density estimation to multivariate density estimation. For a p-variate random variable, the multivariate density is given by

$$\int \ldots \int f(x_1, \ldots, x_p)\ \mathrm{d}(x_1) \ldots\ \mathrm{d}(x_p) = 1.$$

Even though this formula is a straightforward extension of (1.1), the statistical implications are much more complex. Especially, due to the increasing amount of huge datasets becoming available during the last decades, e.g. from financial markets, population development or biological experiments, many applications in the multivariate case focus on discovering interactions and dependencies between marginal observations. In this

thesis penalized smoothing splines, also denoted as P-splines or penalized splines, are the main tool for non-parametric density estimation, as they allow for flexible and smooth estimation of univariate and multivariate density and distribution functions.

1.1 Motivation

Penalized smoothing splines have developed rapidly in scientific literature during the past decades. A major benefit of penalized smoothing splines is, that the estimation approaches can be constructed without any a priori assumptions on distribution functions and thus without any restriction with respect to the latter, although some regularity conditions (e.g. smoothness) have to be fulfilled. This is also valid for other non-parametric approaches, e.g. kernel density estimation. Hence, the investigated approaches in this thesis do not estimate any parameters of given distribution functions, but estimate a univariate density by maximizing a constructed likelihood function in combination with a penalization approach.
This thesis also covers an extension of the univariate case by investigation of copula distribution and copula density, which are used to analyse dependencies of observed data. The established estimation approaches of multivariate copula distributions estimate parameters using maximum likelihood theory, that are correlation parameters and e.g. degrees of freedom in the case of a multivariate t-distribution. Additionally, the margins of copula distributions are often estimated parametrically in a foregoing separated estimation step. In the approach of this thesis, the marginal distributions and the joint copula distribution can be estimated in one step using penalized smoothing splines in combination with quadratic programming with respect to some side constraints. Multivariate densities can be decomposed into a product of marginal and conditional densities as

$$f(x_1, \ldots, x_p) = f(x_p|x_1, \ldots, x_{p-1}) \cdot f(x_1, \ldots, x_{p-1}) = \cdots = \prod_{i=2}^{p} f(x_i|x_1, \ldots, x_{i-1}) \cdot f(x_1).$$

Sklar (1959) provided a theorem, that allows for a decomposition of this joint p-variate density into bivariate copula density functions, which are often denoted as pair-copula densities. This idea is the foundation for dependence vines, that is each bivariate density function has to be specified to describe the joint (copula) density function, following a given decomposition. In common literature, parametric procedures, based on maximum likelihood theory are used to estimate the optimal parameter(s) for each possible copula family, where also the determination of the optimal copula family is not negligible. Each pair-copula density can be determined using penalized smoothing

splines without restrictions on any theoretical copula distribution function.
Especially, the combination of nonparametric univariate density estimators with nonparametric copula density estimators is investigated, whereas the approaches are based on penalized smoothing splines. That is, the marginal distributions are estimated separately in a foregoing step and the copula density is estimated using the latter results. To the best of my knowledge, this combined application of penalized spline smoothing techniques is new to literature.

1.2 Outline

Beside the introduction, this thesis consists of six chapters. The second chapter covers the statistical methods and concepts used in the following chapters. Penalized spline smoothing is explained. This part focuses on using B-splines as basis functions for penalized smoothing splines as well as on presenting penalized splines as a linear mixed model. Additionally, an overview of kernel density estimation and the underlying ideas in the univariate and multivariate case, is given. The degree of smoothness of kernel density functions is determined by a smoothing parameter. The smoothing parameter selecting by cross validation is also exemplified in this thesis. All these techniques are used in the simulation studies in Chapter 3 and Chapter 4 to compare the performance of the penalized smoothing splines density estimation approach. Moreover, Chapter 2 describes the concept and idea of copula theory, presenting the best known copula families and their parametric estimation approach. The last part of the chapter introduces dependence vines and the corresponding parametric estimation, required for Chapter 5.
Chapter 3 introduces an application of penalized splines to estimate univariate density functions, representing the unknown density by a convex mixture of basis densities. The weights of the basis functions are estimated in a penalized form. The considered approach is compared with classical kernel density estimation and further estimation approaches. Penalized smoothing splines provide by an integration of the basic functions also the estimated distribution of the corresponding estimated density. Moreover, the approach is extended to grouped data depending on categorical covariates. This allows for a test of equality of the grouped densities as an alternative to the classical Kolmogorov-Smirnov test. Simulations compare the investigated approach with existing univariate approaches and show promising results.
Chapter 4 discusses an approach to estimate multivariate copula density functions using penalized smoothing splines. The estimate of high-dimensional density functions using full tensor products of B-spline basis functions is introduced. The concept of sparse

grids (see Zenger 1991) is applied, which equals to a reduced tensor product. The spline coefficients are accordingly penalized to achieve a smooth fit. It is the innovative aspect of the presented approach to estimate the marginal and joint density in one step, using quadratic programming with linear constraints for the spline coefficients. Simulation studies for samples from Archimedean and elliptical copula families compare the introduced approach with the classical multivariate kernel density estimator. The results of the penalized splines outperform the competitor.

In Chapter 5, dependence vines are investigated, especially D-vines which follow a special decomposition of the joint density. In this chapter a modification of the penalized high-dimensional copula estimator, presented in Chapter 4, is used in the bivariate case. That is the joint density is estimated by estimating the pair-copula densities, due to the recursive dependence structure given by a D-vine. Additionally, simulations compare the parametric estimation of D-vines with the presented approach and show an equivalent behaviour.

Chapter 6 presents an extension, combining the univariate density estimation approach from Chapter 3 and the copula density estimator investigated in Chapter 4 for exchange rate data, which are also used in Chapter 4. This application outperforms the approaches considered in Chapter 4.

This thesis uses the software `R` (see R Development Core Team 2011) for the simulation studies. Furthermore, the investigated approaches using penalized spline smoothing in Chapter 3, 4 and 5 are implemented in `R` packages.

2 Theoretical Background

The main focus within this thesis is the estimation of densities or distributions of univariate and multivariate data using the technique of penalized splines. First, the idea and principle of penalized splines are presented in Section 2.1. Then, density and copula estimation in general are described in Section 2.2 and Section 2.3. Finally the idea of dependence vines, especially D-vines are, discussed in Section 2.4.

2.1 Penalized Splines

This chapter presents the principle of penalized splines, following Green and Silverman (1994), Ruppert, Wand, and Carroll (2003), Fahrmeir, Kneib, and Lang (2007) and Krivobokova (2006). The underlying idea is explained by starting with a response $y = (y_1, \ldots, y_n)$ and a single covariate $x = (x_1, \ldots, x_n)$. This concept is easily extended to a multivariate setup, which is often called (generalized) additive model (see Wood 2006). The extension to a generalized model is not mentioned in this introduction, because the applications in the following chapters of this thesis do not use any specific distributional assumptions.
In the context of classical linear models, the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ describes a linear relationship between x and y. Penalized splines offer a technique to model a more flexible smooth function $f(x)$, such that

$$y_i = f(x_i) + \epsilon_i \tag{2.1}$$

with $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$. A function f is usually called smooth, when it is at least twice continuously differentiable. The main idea is to separate the observed range of data $x \in [a, b]$, into sections, fitting a twice continuously differentiable spline function in each section. The intersecting points of these sections are called knots, noted as a=$\mu_1 < \cdots < \mu_m = b$. Their number m determines the amount of flexibility, allowed in the functional relationship. In addition, a spline of degree l consists of polynomials of degree l or less, that means l determines the degree of differentiability of f. That is, polynomial splines $\phi_k, k = 1, \ldots, m$, fulfilling these constraints are used for the estimation. Usually, quadratic or cubic polynomial splines are used in many

applications.
Within this framework, f in (2.1) can be written as weighted sum of basis functions $\phi_k, k = 1, \ldots, m$, that is

$$f(x) = \sum_{k=1}^{m} c_k \phi_k(x), \tag{2.2}$$

where $c_k,\ k = 1, \ldots, m$ are called basis coefficients. The model equation (2.1) can be rewritten as

$$y = f(x) + \epsilon = \Phi(x)c + \epsilon \tag{2.3}$$

with $c = (c_1, \ldots, c_m)^T$ as vector of the coefficients, the design matrix $\Phi(x) = (\phi_1(x), \ldots, \phi_m(x))$ and vector of the residuals $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. The model (2.3) is a parametric model, that is optimal weights c_k can be estimated using the ordinary least-squares estimator. Hence, the optimal weights results as

$$\hat{c} = (\Phi(x)^T \Phi(x))^{-1} \Phi(x)^T y.$$

Assuming a normal distribution of the response y, we use the following model

$$y \sim N(\Phi(x)c, \sigma_\epsilon^2 I_n)$$

with the $n \times n$ identity matrix I_n and a constant σ_ϵ^2.

2.1.1 Spline Bases

There are several possibilities to choose a type of basis functions for ϕ_k in (2.2). Penalized splines as referred to Eilers and Marx (1996) are based on B-splines basis functions, introduced by de Boor (1978) and described later on. B-spline bases are constructed easily and have numerical and practical advantages compared with other basis functions as e.g. truncated polynomials. Wood (2006) gives an introduction to so called thin plate splines, which have some advantages when estimating high dimensional functions, but will not be discussed in detail in this thesis. Moreover, there exist radial basis functions or natural cubic splines (see Ruppert, Wand, and Carroll 2003), which are also not considered in detail in this thesis.
The easiest extension of a parametric linear model is done using the basis of truncated polynomials. That is, the model using truncated polynomials of degree l for m knots, separating the support [a,b] of x, such that $a = \mu_1 < \cdots < \mu_m = b$ is given by

$$y_i = c_0 + c_1 x_i + \cdots + c_{l+1} x_i^l + c_{l+2}(x_i - \mu_2)_+^l + \cdots + c_{l+m-1}(x_i - \mu_{m-1})_+^l + \epsilon_i \tag{2.4}$$

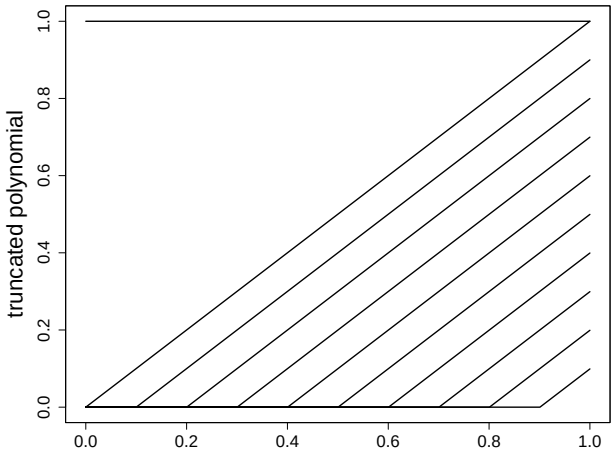


Figure 2.1: Truncated polynomials basis of degree $l = 1$ with equidistant knots.

with the truncated polynomials

$$(x - \mu_j)_+^l = \begin{cases} (x - \mu_j)^l & x \geq \mu_j \\ 0 & \text{else} \end{cases}.$$

So, the model consists of $l + 1$ polynomials and $m - 2$ truncated polynomials, such that $d = l + m - 1$ basis functions exist. Analogously to the linear model, the basis functions are noted as design matrix

$$\Phi(x) = \begin{pmatrix} 1 & x_1 & \dots & x_1^l & (x_1 - \mu_2)_+^l & \dots & (x_1 - \mu_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & x_n & \dots & x_n^l & (x_n - \mu_2)_+^l & \dots & (x_n - \mu_{m-1})_+^l \end{pmatrix}$$

of dimension $n \times d$ with corresponding coefficient vector $c = (c_1, \dots, c_d)$. Within this framework, the truncated polynomials are easily implemented, but they are not always numerically stable, when penalization concepts are introduced later. Figure 2.1 shows an example of linear truncated polynomials with equidistant knots.
An alternative to truncated polynomials are B-splines. Following de Boor (1978), the j-th B-spline basis of degree $l + 1$ is defined as

$$B_j^l(x) = \frac{x - \mu_j}{\mu_{j+1} - \mu_j} B_j^{l-1}(x) + \frac{\mu_{j+l+1} - u}{\mu_{j+l+1} - \mu_{j+1}} B_{j+1}^{l-1}(x),$$

with $B_j^0(x) = 1_{[\mu_j, \mu_{j+1})}(x)$ and knots $\mu_j, j = 1, \dots, m$. Eilers and Marx (2010) show, that B-splines can be computed by differencing of corresponding truncated polynomials.

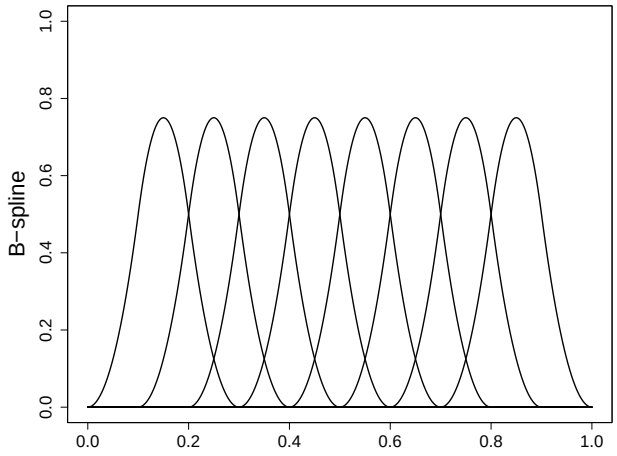


Figure 2.2: B-spline basis of degree $l = 2$ with equidistant knots.

B-splines are considered, because they have many desirable attributes (see de Boor 1978 or Eilers and Marx 1996). First for a B-spline of degree l, only $l + 2$ knots build the support of a single B-spline. That is, the support is bounded, in contrast to e.g. truncated polynomials. The polynomial pieces join at q knots and at the joining points, derivatives up to order $l - 1$ are continuous. Moreover, B-splines create a partition of 1 and each B-spline overlaps only with $2l + 2$ neighbouring B-splines. So, for the construction of a B-spline basis of degree l, there are $m + 2l + 1$ knots needed. Furthermore, the co-domain of B-splines is limited and derivatives of the j-th B-spline are easily calculated as

$$\frac{\partial}{\partial x} B_j^l(x) = l \cdot \left(\frac{1}{\mu_{j+l} - \mu_j} B_j^{l-1}(x) - \frac{1}{\mu_{j+l+1} - \mu_{j+1}} B_{j+1}^{l-1}(x) \right).$$

B-splines are constructed, such that the piecewise polynomials are fitted smoothly in the knots. That is, a B-spline basis consists of $l + 1$ polynomials of degree l, which are $l - 1$ times continuously differentiable, see Eilers and Marx (1996).
These facts have numerical and therefore computational advantages compared with other types of basis functions. The location and the amount of knots mu for a B-spline basis have to be chosen adequately. In the context of penalized splines, Ruppert, Wand, and Carroll (2003) suggest to set 20 up to 40 knots. This amount of knots assures enough flexibility to describe the data. For the number of knots Ruppert, Wand, and Carroll (2003) suggest to use the rule

$$m = \min \left(\frac{1}{4} \times \text{number of unique } x_i, 35 \right)$$

and recommend to place the knots μ by

$$\mu_k = \left(\frac{k+1}{m+2}\right) \text{th sample quantile of the unique } x_i,$$

for $k = 1, \ldots, m$. These rules suggest choosing the knots depending on the data x. The amount of the knots steers the estimation, so that the fit is flexible enough to describe the structure of data x, whereas a sparse amount of knots may not be flexible enough. Of course, the placement of knots can be done in different ways. In many applications, the locations are chosen equidistantly, what allows numerical inferences in further applications. The presented approaches in the further chapter of this thesis use equidistant knots, too. Figure 2.2 shows an example of B-splines with degree 2 with equidistant knots.
The corresponding design matrix for B-spline basis functions B_j^l is given by

$$\Phi(x) = \begin{pmatrix} B_1^l(x_1) & \ldots & B_d^l(x_1) \\ \vdots & & \vdots \\ B_1^l(x_n) & \ldots & B_d^l(x_n) \end{pmatrix},$$

which consists of $d = l + m - 1$ basis functions. To show the construction principle of B-splines by differencing corresponding truncated polynomials, we have to add $2l$ truncated polynomials. We need $2l$ additional knots outside the support of y, due to the recursive definition for the construction of a complete B-spline basis. Further details are available in Ruppert, Wand, and Carroll (2003) and Eilers and Marx (2010).
Bernstein polynomials are another possible class of basis functions for spline smoothing. The Bernstein polynomial of degree K is defined as

$$\tilde{\phi}_{Kk}(u) = \binom{K}{k} u^k (1-u)^{K-k} \tag{2.5}$$

for $k = 0, \ldots, K$ and $u \in [0, 1]$. Considering the $K + 1$ Bernstein polynomials (2.5) of degree K for $k = 0, \ldots, K$, they form a partition of unity, that is they sum to one for all values of u. Any Bernstein polynomial of degree K can be written in the terms of the power basis $\{1, u, u^2, u^3, \ldots, u^K\}$, that is (see Doha, Bhrawy, and Saker 2011)

$$\tilde{\phi}_{Kk}(u) = \sum_{i=k}^{K} (-1)^{i-k} \binom{K}{i} \binom{i}{k} u^i.$$

Especially, the B-spline basis function $B_j^K(u)$ coincides with Bernstein polynomial $\tilde{\phi}_{Kk}(u)$ for $j = 0, \ldots, K$ and $u \in [0, 1]$, if the B-spline basis is constructed with $2n$

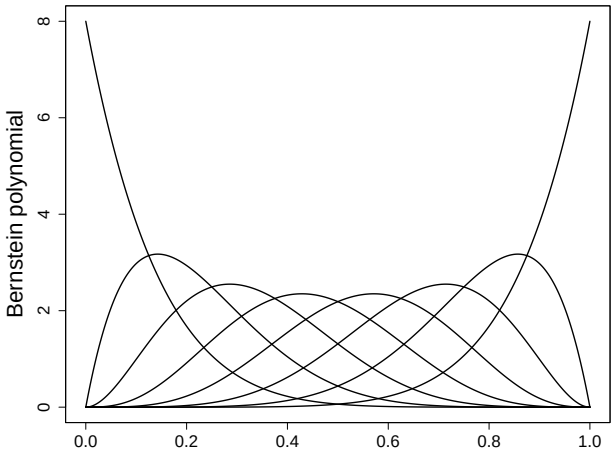


Figure 2.3: Standardized Bernstein polynomials with $K = 7$.

knots $\mu_1 = \cdots = \mu_n = 0$ and $\mu_{n+1} = \cdots = \mu_{2n} = 1$ (see Prautzsch, Boehm, and Paluszny 2002). The integration in the range of $[0, 1]$ of Bernstein polynomial (2.5) of order K results in the definite integral, that is (see Doha, Bhrawy, and Saker 2011)

$$\int_0^1 \tilde{\phi}_{Kk}(u) = \frac{1}{K+1} \quad \text{for } k = 0, \ldots, K.$$

Normalization of (2.5) with factor $(K + 1)$ leads to the basis $\boldsymbol{\phi}_K(u) = (\phi_{K0}(u), \ldots, \phi_{KK}(u))$ of standardized Bernstein polynomials, defined as

$$\phi_{Kk}(u) = (K+1)\binom{K}{k} u^k (1-u)^{K-k}. \tag{2.6}$$

That is $\phi_{Kk}(u)$ is non-negative and normalized to be a density. Moreover, it follows that (2.6) is a Beta distribution and $\int_0^1 \phi_{Kk}(u)\, \mathrm{d}u = 1$. Figure 2.3 shows normalized Bernstein polynomials of degree $K = 7$.

2.1.2 Penalization

The fit of (2.2) may be wiggly, due to a large number of basis functions. To ensure a smooth and nice fit of the data in (2.2), a roughness penalty is introduced. The penalty for the truncated polynomials (2.4) is defined as $\sum_{j=l+2}^{d} c_j^2$, that is penalizing too much variability of the truncated polynomials. Adding this penalty term into (2.2),

the penalized least squares function minimizes

$$\sum_{i=1}^{n}\{y_i - \sum_{k=1}^{d}\phi_k(x_i)c_k\}^2 + \lambda \sum_{j=l+2}^{d} c_j^2,$$

for penalty parameter λ, controlling the amount of smoothing. The penalty term is usually noted as

$$\lambda \sum_{j=l+2}^{d} c_j^2 = \lambda c^T D c$$

with penalty matrix $D = \text{blockdiag}(0_{(l+1)\times(l+1)}, I_{(m-2)})$. Commonly, the integrated squared second order derivative of f is used as penalty for B-splines basis functions, because the second order derivative is a suitable measure for the curvature of a f, that is the penalty term results as

$$\lambda \int (f''(z))^2 \, \mathrm{d}z.$$

This idea of penalized spline smoothing traces back to O'Sullivan (1986). A penalization concept for B-splines, based on penalizing differences of the basis coefficients, is presented in Eilers and Marx (1996). They proposed to base the penalty on second order differences of the coefficients. The difference operator of order a, is defined as

$$\begin{aligned}
\Delta^1 c_k &= c_k - c_{k-1} \\
\Delta^2 c_k &= \Delta^1\Delta^1 c_k = \Delta^1(c_k - c_{k-1}) = c_k - 2c_{k-1} + c_{k-2} \\
&\vdots = \vdots \\
\Delta^a c_k &= \Delta^{a-1} c_k - \Delta^{a-1} c_{k-1}.
\end{aligned}$$

For $a = 2$, the second order difference matrix L^2 for a B-spline basis with d basis functions equals

$$L^2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}, \tag{2.7}$$

with L^2 is $(d-2) \times d$ dimensional. The penalty term for second order differences is given by

$$\lambda \sum_{k=l+1}^{d} (\Delta^2 c_k)^2 = \lambda c^T D c \tag{2.8}$$

with $d \times d$ dimensional penalty matrix

$$D = (L^2)^T L^2. \tag{2.9}$$

Adding the penalty term (2.8) to (2.2), the penalized least squares function results as

$$\sum_{i=1}^{n} \{y_i - \sum_{k=1}^{d} \phi_k(x_i) c_k\}^2 + \lambda c^T D c.$$

In summary, the corresponding penalized least squares function for truncated polynomials and B-splines arise identically. That is, the estimator for the optimal coefficients $\hat{c}$ using truncated polynomials or B-splines results in

$$\hat{c} = (\Phi(x)^T \Phi(x) + \lambda D)^{-1} \Phi(x)^T y. \tag{2.10}$$

Penalized splines are often titled as non-parametric models to highlight the flexibility of the approach in contrast to the classical linear model. Comparing the B-splines with the truncated polynomials, the locally bounded support of the B-spline functions may be advantageous, e.g. for numerical implementations. Additionally, for a large number of knots and a smoothing parameter close to zero $(\Phi(x)^T \Phi(x) + \lambda D)^{-1}$ can be incomputable, see Ruppert, Wand, and Carroll (2003) for an algorithm, tackling this problem.
For further considerations, the concept of the hat matrix from the linear model is extended to penalized smoothing splines. The smoother matrix S_λ due to (2.10) results as

$$S_\lambda = \Phi(x)(\Phi(x)^{-1} \Phi(x) + \lambda D)^{-1} \Phi(x)^T. \tag{2.11}$$

The fitted values $\hat{f}$ result by using (2.11) as

$$\hat{f} = S_\lambda y = \Phi(x)(\Phi(x)^{-1} \Phi(x) + \lambda D)^{-1} \Phi(x)^T y \tag{2.12}$$

with penalized log-likelihood function

$$l(c) = \log \left\{ \sum_{i=1}^{n} \{y_i - \sum_{k=1}^{d} \phi_k(x_i) c_k\}^2 + \lambda c^T D c \right\}. \tag{2.13}$$

Maximizing of (2.13) results in the optimal coefficients c of the penalized spline for a given penalty parameter λ. The selection of an optimal λ is discussed in the next subsection. The definition of the degrees of freedom is adopted to describe the effective number of fitted parameters. For penalized splines, the following relation can be shown

(see Fahrmeir, Kneib, and Lang 2007)

$$\mathrm{df}_{fit}(S_\lambda) = \mathrm{tr}(S_\lambda) = \mathrm{tr}\left(\Phi(x)^T\Phi(x)(\Phi(x)^T\Phi(x) + \lambda D)^{-1}\right). \tag{2.14}$$

For a penalized spline with $\lambda = 0$, m knots and splines of degree l, it follows $\mathrm{tr}(S_0) = l + 1 + m$, whereas $\mathrm{tr}(S_\lambda) \to l + 1$ as $\lambda \to \infty$. So, $l + 1 \leq \mathrm{df}_{fit}(S_\lambda) \leq l + 1 + m$ (see Ruppert, Wand, and Carroll (2003)). Alternatively, the residual degrees of freedom are defined as

$$\mathrm{df}_{res} = n - 2\mathrm{tr}(S_\lambda) + \mathrm{tr}(S_\lambda {S_\lambda}^T)$$

which is equivalently transformed to

$$n - \mathrm{df}_{res} = 2\mathrm{tr}(S_\lambda) - \mathrm{tr}(S_\lambda {S_\lambda}^T). \tag{2.15}$$

Both measures (2.14) and (2.15) coincide for parametric regression models fitted by ordinary least squares, because $S_\lambda {S_\lambda}^T = S_\lambda$. But these measures differ for nonparametric models for 'mid-size' smoothing, whereas for low or high penalization both definitions tend to coincide. In the case of none penalization and infinite penalization, the fits incline to parametric regression fits.

2.1.3 Smoothing Parameter Selection

The quality and preciseness of the estimation (2.12) depends considerably on the penalty term λ. Therefore, the selection of the optimal smoothing parameter λ is discussed in this subsection. Intuitively, the mean squared error (MSE) is a well-known measure for the goodness of an estimated function $\hat{f}(x)$, that is the MSE is defined as

$$\mathrm{MSE}(\hat{f}(x)) = \left(\mathrm{E}(\hat{f}(x) - f(x))\right)^2 + \mathrm{var}\left(\hat{f}(x)\right). \tag{2.16}$$

In (2.16), the first term reflects the squared bias and the second the variance of $\hat{f}(x)$. But squared bias and variance in (2.16) can not be simultaneously minimized, reflecting the bias-variance trade-off for penalized spline smoothing. Choosing larger values of λ leads to a smaller variance, but increased bias. Reducing the value of λ results in the converse, so a greater variance and smaller bias. Therefore, approaches for the optimal selection of the smoothing parameter λ are discussed. First, minimizing the residual sum of squares (RSS) of $\hat{f}(x)$, that is $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$ results in the trivial interpolate estimator for c_k. Therefore, minimizing the cross-validation criterion

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{(-i)}(x_i))^2$$

is used for selection of λ, where $\hat{f}^{(-1)}(x_i)$ notes the fit omitting the ith observation. Using the smoother matrix S_λ (2.11), the cross validation criterion can be approximated (see Ruppert, Wand, and Carroll 2003) as

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - s_{ii}}\right)^2 \tag{2.17}$$

with s_{ii} is the ith element of the diagonal of S_λ. Craven and Wahba 1979 replace s_{ii} by their average $\frac{1}{n}\sum_{i=1}^{n} s_{ii} = \frac{1}{n}\mathrm{tr}(S_\lambda)$. This replacement in (2.17) is known as generalized cross-validation criterion (GCV) given by

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - \mathrm{tr}(S_\lambda)/n}\right)^2. \tag{2.18}$$

Both measures (2.17) and (2.18) imply a grid search, selecting that λ with minimal fit criterion, that is with minimal CV or rather GCV. Another approach to select optimal parameters is minimizing the Kullback-Leibler information (see Kullback and Leibler 1951)

$$I(f, g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\,\mathrm{d}x \tag{2.19}$$

between the true density $f(x)$ and estimated density $g(x)$, which are both continuous functions. The interpretation of $I(f, g)$ is the distance from g to f. In the case of discrete distributions p_i and q_i for $i = 1, \ldots, n$, (2.19) is defined as

$$I(f, g) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right).$$

The Kullback-Leibler information is only computable with full knowledge about f and g, but that is unrealistic. Akaike (1974) described the information loss, based on the empirical log-likelihood function at its maximum point. Akaike (1974) defined the Akaike information criterion (AIC) as

$$\mathrm{AIC} = \log(\mathrm{RSS}(\lambda)) + 2 \cdot K/n \tag{2.20}$$

with RSS is the residual sum of squares $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ of the estimated model and K is the number of used parameters in the model, see (2.14) for a possible choice of K. Hurvich and Tsai (1989) presented an improved AIC with respect to the sample size n, called corrected AIC, which is given by

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2K(K+1)}{n - K - 1}. \tag{2.21}$$

The number of parameters K in (2.20) and (2.21) can be approximated by the trace of the smoothing matrix S_λ, depending on the selected penalty parameter λ, that is $K = df(\lambda) = tr(S_\lambda)$. At this point, a grid search is useful to find the optimal smoothing parameter λ, minimizing AIC or rather AIC_c. In the case of different candidate models, the difference

$$\Delta(\text{AIC})_i = \text{AIC}_i - \text{AIC}_{min} \tag{2.22}$$

estimate the relative expected Kullback-Leibler difference between the candidate model i and the model with minimal AIC or rather AIC_c (see Burnham and Anderson 2010). These relative values allow an easy ranking and comparison of candidate models, the absolute value is not the main important detail. Selecting the optimal model using the AIC measures (2.20), (2.21) and (2.22), implies a grid search fitting different models with different penalty parameters λ.
A direct calculation of an optimal penalty parameter λ is possible, representing the penalized smoothing spline as linear mixed model (see e.g. Wand 2003).

2.1.4 Link to Linear Mixed Models

This subsection discusses linear mixed models, following Ruppert, Wand, and Carroll (2003) and Fahrmeir, Kneib, and Lang (2007). The classical linear mixed model is given by

$$y = X\beta + U\gamma + \epsilon \tag{2.23}$$

with X and U are the model matrices, β is called vector of fixed effects and γ is the vector of individual- or cluster-specific random effects in the model and ϵ the usual vector of residuals. The assumptions for β and γ are

$$\begin{pmatrix} \gamma \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \right) \tag{2.24}$$

with G and R are block diagonal covariance matrices. The underlying distribution of y given γ following from (2.23) and (2.24) is

$$y|\gamma \sim N(X\beta + U\gamma, R), \quad \gamma \sim N(0, G). \tag{2.25}$$

Estimating of the fixed effects is easily done, solving

$$y = X\beta + \epsilon^*, \quad \epsilon^* = U\gamma + \epsilon.$$

This yields the classical linear model $y \sim N(X\beta, R+UGU^T)$. Defining $V = R+UGU^T$, using the least squares estimator for the fixed effects β for known matrix V results in the best linear unbiased predictor (BLUP) given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \tag{2.26}$$

The BLUP for γ, based on β results as

$$\hat{\gamma} = GU^T V^{-1}(y - X\hat{\beta}). \tag{2.27}$$

The proof for $\hat{\gamma}$ is given in McCulloch and Searle (2001). If R and G are known, the estimator (2.27) results as the best linear unbiased predictor (BLUP) (see Robinson 1991). Henderson (1950) uses the assumptions $y|\gamma \sim N(X\beta + U\gamma, R), u \sim N(0, G)$ to maximize the likelihood of (y, γ) over the unknowns β and γ, using the joint density of y and γ. This results in the penalized least squares criterion

$$(y - X\beta - U\gamma)^T R^{-1}(y - X\beta - U\gamma) + \gamma^T G^{-1}\gamma. \tag{2.28}$$

It is easy to prove from (2.28) that the BLUP of (β, γ) can be formulated such that

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = (C^T R^{-1} C + B)^{-1} C^T R^{-1} y$$

with $C = [X\ U]$ and $B = \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} \end{pmatrix}$. The fitted values are given by

$$\hat{y} = X\hat{\beta} + U\hat{\gamma} = C(C^T R^{-1} C + B)^{-1} C^T R^{-1} y. \tag{2.29}$$

Usually, R and G in (2.24) are unknown, such that a maximum likelihood (ML) estimator and in an extension a restricted maximum likelihood estimator are used for the prediction of R and G. First, the unknown parameters are named with ϑ, such that $V(\vartheta) = UG(\vartheta)U^T + R(\vartheta)$. (2.25) changes to

$$y \sim N(X\beta, V(\vartheta))$$

and the corresponding log-likelihood equals except some additive constants

$$l(\beta, \vartheta) = -\frac{1}{2}\{\log(|V(\vartheta)| + (y - X\beta)^T V(\vartheta)(y - X\beta)\}. \tag{2.30}$$

Maximizing (2.30) with respect to β yields the estimator for fixed effects, that is

$$\hat{\beta} = (X^T V(\vartheta)^{-1} X)^{-1} X^T V(\vartheta)^{-1} y. \tag{2.31}$$

Inserting (2.31) into (2.29) yields the profile-log-likelihood given by

$$l_P(\vartheta) = -\frac{1}{2}\{\log(|V(\vartheta)| + (y - X\beta(\vartheta))^T V(\vartheta)(y - X\beta(\vartheta))\}. \tag{2.32}$$

Analogously, the restricted log-likelihood l_R is achieved, integrating out β in the marginal log-likelihood $l_R(\vartheta) = \log\left(\int L(\beta, \vartheta)\, \mathrm{d}\beta\right)$ (see Searle, Casella, and McCulloch 1992), that is

$$l_R(\vartheta) = l_P(\vartheta) - \frac{1}{2}\log|X^T V(\vartheta)^{-1} X|. \tag{2.33}$$

Maximizing of (2.33) yields the estimator $\hat{\vartheta}_{REML}$, which minimizes the bias compared to $\hat{\vartheta}_{ML}$, achieved from maximizing of (2.32) with respect to ϑ. Computation of $\hat{\vartheta}_{REML}$ is done iteratively, using e.g. Newton-Raphson-algorithm or Fisher-Scoring-algorithm. Replacing the estimated covariance matrices $\hat{G}$ and $\hat{V}$ in the BLUPs (2.26) and (2.27) results in the estimated best linear unbiased predictors (EBLUP)

$$\begin{aligned} \tilde{\beta} &= (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y \quad \text{and} \\ \tilde{\gamma} &= \hat{G} U^T \hat{V}^{-1} (y - X\hat{\beta}). \end{aligned}$$

2.1.5 Linear Mixed Model Representation of Penalized Splines

The fitted penalized spline (2.12) can be formulated as linear mixed model (2.29) (see Wand 2003, Kauermann 2005 or recent work by Reiss and Ogden 2009 and Wood 2011). Assuming the coefficient γ in (2.12) to be random and define X as matrix containing the polynomials and U as matrix containing the truncated polynomial basis functions, the following model results

$$y|\gamma \sim N(X\beta + U\gamma, \sigma_\epsilon^2 I_n), \quad \gamma \sim N(0, \sigma_\gamma^2 I_d).$$

With respect to (2.29), with $R = \sigma_\epsilon^2 I_n$ and $G = \sigma_\gamma^2 I_d$ the fitted values $\hat{y}$ results as

$$\hat{y} = C(C^T C + \frac{\sigma_\epsilon^2}{\sigma_\gamma^2} D)^{-1} C^T y, \tag{2.34}$$

with $D = \text{blockdiag}(0_{(l+1)\times(l+1)}, I_d^{-1})$. The ratio $\sigma_\epsilon^2/\sigma_\gamma^2$ in (2.34) represents the smoothing parameter λ in the context of penalized splines. The inverse of penalty matrix D in (2.34) has to be symmetric and positive definite, which is the case for truncated poly-

nomials. Other basis functions have to be adapted to reach a symmetric and positive definite penalty matrix D. Green (1987) and Fahrmeir, Kneib, and Lang (2004) discuss this topic in detail. At this point, the changes in the case of B-splines are summarized, following Krivobokova (2006).
Considering the difference matrix (2.7) for B-splines of degree l, based on difference penalty of order a and m knots, D has the dimension $(m+1+l)\times(m+1+l-a)$. That is, the corresponding penalty matrix, defined by $(L^a)^T L^a$ (see (2.9)), is singular with rank $m+1+l-a$. Using a singular value decomposition results in $(L^a)^T L^a = Z\text{diag}(z)Z^T$ with Z are the eigenvectors and z are the eigenvalues in decreasing order, such that the first $m+1+l-a$ eigenvalues are non negative and the remaining a equals zero. The matrix Z and the eigenvalues z can be decomposed into $Z = [Z_+ \; Z_0]$ and $z = (z_+, z_0)$, such that

$$\begin{aligned} \Phi(x)c &= \Phi(x)ZZ^Tc = \Phi(x)[Z_0{Z_0}^Tc + Z_+\text{diag}(z_+^{-1/2})\text{diag}(z_+^{1/2}){Z_+}^Tc] \\ &= \Phi(x)[Z_0\beta + Z_+\text{diag}(z_+^{-1/2})c] \\ &= X\beta + U_\Phi\gamma. \end{aligned} \tag{2.35}$$

However, it yields

$$c^T(L^a)^TL^ac = c^TZ\text{diag}(z)Z^Tc = c^TZ_0\text{diag}(0_a)Z_0^Tc + c^TZ_+\text{diag}(z_+)Z_+^Tc = \gamma^T\gamma.$$

That is, only the coefficients γ are penalized, using the penalty matrix $I_{m+1+l-a}$ and a mixed model presentation is possible. The mixed model results as

$$y|\gamma \sim N(X\beta + U_\Phi\gamma, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \sigma_\gamma^2 I_{m+1+l-a}).$$

The singularity of $(L^a)^TL^a$ causes, that the representation (2.35) is not unique. Matrices B_β and B_γ of dimensions $(m+1+l)\times a$ and $(m+1+l)\times(m+1+l-a)$ do any one-to-one transformations, such that $\Phi(x)c = \Phi(x)[B_\beta\beta + B_\gamma\gamma]$.
Therefore, B_β and B_γ are selected, such that

- $[B_\beta \; B_\gamma]$ has full rank (uniqueness of transformation);
- $B_\beta^TB_\gamma = B_\gamma^TB_\beta = 0$;
- $B_\beta^T(L^a)^TL^aB_\beta = 0$ and
- $B_\gamma^T(L^a)^TL^aB_\gamma = I_{m+1+l-a}$.

The last three conditions ensure, that only γ is penalized with identity matrix (for more information see Green 1987 and Fahrmeir, Kneib, and Lang 2004). Using $B_\beta =$

$[1, b, \ldots, b^{a-1}]$ with $b = (1, 2, \ldots, m + l + 1)$ and $W_\gamma = (L^a)^T (L^a (L^a)^T)^{-1}$ have become a common choice (see Krivobokova 2006). The final transformation is given by

$$\Phi(x)c = \Phi(x)[B_\beta \beta + (L^a)^T (L^a (L^a)^T)^{-1} \gamma] =: X\beta + U_\Phi \gamma,$$

whereas X results in a polynomial of degree a.

2.1.6 Bivariate Penalized Splines

This section discusses the extension of the univariate penalized spline approach into the bivariate case. This is done as contribution to the investigations presented in Chapter 4 and 5. The estimation of bivariate smooth functions f, with respect to two marginal variables x_1 and x_2 is motivated by using penalized B-splines. That is, we define a tensor products of univariate B-spline bases $\Phi^{(1)}(x_1)$ and $\Phi^{(2)}(x_2)$ as

$$\Phi_{jk}(x_1, x_2) = \Phi_j^{(1)}(x_1) \cdot \Phi_k^{(2)}(x_2), \quad j = 1, \ldots, d_1, \quad k = 1, \ldots, d_2,$$

with d_1 and d_2 are the dimensions of the univariate B-spline bases $\Phi^{(1)}(x_1)$ and $\Phi^{(2)}(x_2)$. The smooth function f results as weighted sum, that is

$$f(x_1, x_2) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} c_{jk} \Phi_{jk}(x_1, x_2), \tag{2.36}$$

with $c_{jk}, j = 1, \ldots, d_1$ and $k = 1, \ldots, d_2$ are the corresponding basis coefficients. Defining the design matrix M with rows as

$$m_i^T = (\Phi_{11}(x_{i1}, x_{i2}), \ldots, \Phi_{d_1 1}(x_{i1}, x_{i2}), \ldots, \Phi_{1 d_2}(x_{i1}, x_{i2}), \ldots, \Phi_{d_1 d_2}(x_{i1}, x_{i2}))$$

and the vector of the corresponding coefficients as

$$c = (c_{11}, \ldots, c_{d_1 1}, \ldots, c_{1 d_2}, \ldots, c_{d_1 d_2})^T,$$

resulting the equation

$$y = Mc + \epsilon.$$

Analogously to the univariate case, a penalty is introduced in (2.36) to achieve a smooth fit for a suitable amount of basis functions. First, we define marginal first difference matrices L_1 and L_2 as in the univariate case (see (2.7)) in the direction of x_1 and x_2. These matrices are extended line by line and column by column, using Kronecker

products, that is the line by line penalty term is constructed as

$$c^T(I_{d_2} \otimes L_1)^T(I_{d_2} \otimes L_1)c = \sum_{k=1}^{d_2}\sum_{j=2}^{d_1}(c_{jk} - c_{j-1,k})^2,$$

whereas the column by column penalty term is given by

$$c^T(L_2 \otimes I_{d_1})^T(L_2 \otimes I_{d_1})c = \sum_{j=1}^{d_1}\sum_{k=2}^{d_2}(c_{jk} - c_{j,k-1})^2.$$

The whole penalty term results as

$$\lambda c^T Dc = \lambda c^T[(I_{d_2} \otimes L_1)^T(I_{d_2} \otimes L_1) + (L_2 \otimes I_{d_1})^T(L_2 \otimes I_{d_1})]c,$$

which can reformulated using rules for Kronecker products as quadratic penalty term

$$\lambda c^T Dc = \lambda c^T[I_{d_2} \otimes D_1 + D_2 \otimes I_{d_1}]c$$

with $D_1 = L_1^T L_1$ and $D_2 = L_2^T L_2$. Due to this fact, the selection procedures for the optimal penalty parameter λ discussed for the univariate case in Section 2.1.3 can be applied.
In Chapter 4 and 5 of this thesis, the concept of univariate penalized splines is extended to higher dimensions, using tensor products of univariate B-spline bases and the difference penalty as described in foregoing parts of this chapter. But the full tensor product is neglected, due to the curse of dimensionality for an extensive amount of basis functions and the so called sparse grids are introduced in Chapter 4. Bivariate estimations based on the full tensor product are done in Chapter 5.

2.2 Kernel Density Estimation

Observed data never disclose their probability distribution, neither their probability density. Scientists have been looking for methods to explain behaviour and attributes of observations of any noticed statistics. Since the last century, density estimation has been one of the most challenging and ambitious tasks in theoretical and applied statistics. This section presents techniques of kernel density estimation, which will be used in further chapters of this thesis.

2.2.1 Univariate Kernel Density Estimation

The topic of univariate kernel density estimation is introduced, following Silverman (1986). From the beginning we assume, that the n observations $x_1, \ldots, x_n$ are independent, identically distributed observations from a continuous univariate distribution with probability density function f, see (1.1).
Estimates of the unknown density are denoted as $\hat{f}$. The main ideas of kernel density estimation go back to Nadaraya (1964) and Watson (1964), see also Nadaraya (1974), which is probably one of the best known approaches estimating unknown density functions. Silverman (1986), Scott (1992) and Li and Racine (2007) give overviews about the development and motivation of kernel density estimation. Pearson (1938) mentioned how to describe data by graphical tools, e.g. by using histograms. Until today, the histogram is of one the easiest and best known statistical tools estimating distribution of data. Usually, it is done by separating the observed range of data x into classes $[\mu_0, \mu_1), [\mu_1, \mu_2), [\mu_2, \mu_3), \ldots, [\mu_{k-1}, \mu_k)$. The area under the histogram on each class shall reflect the number of elements, defined as f_j, in each class. Since the total area of the histogram equals 1, the histogram corresponds to the total number of elements n in the dataset. Defining the width of each class as $w_j = c_j - c_{j-1}$, the area of each class of the histogram is equal to the proportion of elements in class c_j, that is the height of each class is defined as f_j/w_j. Obviously, the classes c_j determine the accuracy and the form of the histogram, but there is no general optimal rule how to choose them. Of course, histograms are not continuous, because jump discontinuities appear at each point c_j. The histogram does not fulfill the conditions of (1.1), obviously. Furthermore, the existence of many or less points in neighbouring bins does not effect the current bin.
Histograms with sliding widths of the classes c_j are the first step to improve the histogram, defining a range h, that provides points on both sides of points x_i affecting the current bin of the histogram. The idea is to move the interval $[x_i - h, x_i + h)$ over the range of x. Then the estimate of the density $\hat{f}(x_i)$ is given with $\hat{f}(x_i) = \frac{\text{number of events in } [x_i - h, x_i + h)}{n \cdot 2h}$. Based on this idea, the kernel density estimator for any kernel function $K(\cdot)$ is defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.37}$$

with h is called bandwidth or smoothing parameter. Histograms with sliding widths are still discontinuous, so different continuous kernel functions $K(\cdot)$ have been explored in the literature. Some famous kernel functions are presented in Table 2.1.
Fundamentally, h in (2.37) has to been chosen adequately. If h becomes very large,

Kernel	*K(u)*
Epanechnikov	$\frac{3}{4\sqrt{5}}(1-\frac{u^2}{5})$ for $-1 \leq u < 1$ 0 else
Biweight	$\frac{15}{16}(1-u^2)^2$ for $\lvert u\rvert < 1$ 0 else
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right), u \in \mathbb{R}$
Rectangular	$\frac{1}{2}$ for $-1 \leq u < 1$ 0 else

Table 2.1: Kernel functions

all details of the density disappear, while for a very small h, the density estimation function $\hat{f}(\cdot)$ jumps turbulently at each observation x_i. Now, the optimal h should be chosen, depending on some criteria. The difference between the unknown true density $f(\cdot)$ and the estimated density $\hat{f}(\cdot)$ should be minimal. A possible measure, considering this question, is the (MSE) (2.16). But the MSE is not applicative, due to the trade-off between reducing the bias with increasing variance or vice versa when choosing the optimal h. Moreover the MSE is depending on the investigated bandwidth h. The expectation, variance and the following results are given by (see Silverman 1986)

$$\begin{aligned} \mathrm{E}(\hat{f}(x)) &= \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)\ \mathrm{d}x \\ \mathrm{var}(\hat{f}(x)) &= \frac{1}{n}\int \frac{1}{h^2} K\left(\frac{x-u}{h}\right)^2 f(u)\ \mathrm{d}(u) - \left\{\frac{1}{h}\int K\left(\frac{x-u}{h}\right) f(u)\ \mathrm{d}(u)\right\}^2 \end{aligned}$$

Using a Taylor-series expansion of $\mathrm{E}(\hat{f}(x))$, the bias at any point x results as (see Silverman 1986)

$$\mathrm{bias}\{\hat{f}(x)\} = \frac{1}{2}\sigma_K^2 h^2 f''(x) + O(h^4).$$

Moreover, the expectation of $\hat{f}(x)$ equals $f(x)$ to order $O(h^2)$, if the kernel function K in (2.37) satisfies the following three conditions

$$\begin{aligned} \int K(u)\ \mathrm{d}u &= 1 \\ \int uK(u)\ \mathrm{d}u &= 0 \\ \int u^2 K(u)\ \mathrm{d}u &\equiv \sigma_K^2 > 0 \text{ for any constant } \sigma_K^2. \end{aligned}$$

The Epanechnikov kernel minimizes the MSE (2.16) optimally, compared with other common kernel functions (see Epanechnikov 1969). Rosenblatt (1956) has introduced the mean integrated squared error (MISE), an improved uniform measure of the accuracy of the whole estimation $\hat{f}(\cdot)$, whereas the MSE (2.16) is a point measure of the estimation $\hat{f}(\cdot)$, evaluated in a point x. The MISE is given by

$$\text{MISE}(\hat{f}) = \text{E} \int \{\hat{f}(x) - f(x)\}^2 \, \text{d}x. \tag{2.38}$$

Silverman (1986, p. 35) mentions, that 'the MISE is by far the most tractable global measure'. In the literature exists also the integrated mean squared error (IMSE), which coincides with the MISE (see Scott 1992).
Estimating the optimal bandwidth h can be done with minimizing an approximate integrated squared error (AMISE), because the exact integral in (2.38) can be solved only numerically. Based on (2.38), the AMISE of (2.37) is calculated as sum of the integrated squared bias $\int \text{bias}\{\hat{f}(x)\}^2 \text{d}x$ and the approximated integral of the estimated variance $\int \text{var}\hat{f}(x) \text{d}x$. The approximated AMISE is given by

$$\text{AMISE}(h) = \frac{1}{4} h^4 \sigma_K^4 R(f'') + \frac{R(K)}{nh} \tag{2.39}$$

with $R(g) = \int g(x)^2 \text{d}x$ and $\sigma_g^2 = \int x^2 g(x) \text{d}x$ for any function $g(\cdot)$. The optimal bandwidth h with respect to (2.39) results as $h = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{(1/5)} n^{-1/5}$. The sole unknown component in (2.39) is $R(f'')$, so rewriting (2.39) depending on an kernel-based estimate $S(\alpha)$ of $R(f'')$ results in

$$\widehat{\text{AMISE}}(h) = \frac{1}{4} h^4 \sigma_K^4 S(\alpha) + \frac{R(K)}{nh}.$$

Minimizing (2.39) gives an equation for an optimal bandwidth h. For the Gaussian kernel, it follows (see Scott 1992)

$$h = \frac{4}{3}^{(1/5)} \sigma n^{-1/5} \approx 1.06 \hat{\sigma} n^{-1/5}$$

with $\hat{\sigma}^2$ as estimated variance σ^2 of the normal distribution. Choosing the optimal bandwidth h for any kernel function $K(\cdot)$ is often done automatically e.g. using a cross-validation approach. Therefore Scott and Terrell (1987) present the general formula for an unbiased cross-validation scheme, that is

$$\text{UCV}(h) = \frac{R(K)}{nh} + \frac{2}{n^2 h} \sum_{i<j} \gamma(\Delta_{ij}) \tag{2.40}$$

with $\gamma(\Delta) = \int K(u)K(u+\delta)\mathrm{d}u$ and $\Delta_{ij} = (x_i - x_j)/h$. Park and Marron (1990) present an estimator $\hat{S}(\alpha)$, that results in a consistently good simulation performance for the selection of h. Sheather and Jones (1991) improve this selection criteria using an improved estimator of $R(f'')$, called $\hat{S}_D(\alpha)$, contributing a positive amount to the bias in estimating $R(f'')$. Further details are presented in Sheather and Jones (1991). Another method to estimate the optimal bandwidth h is likelihood cross-validation, introduced by Duin (1976), that is maximizing

$$\log L = \sum_{i=1}^{n} \log \hat{f}_{-i}(x_i) \tag{2.41}$$

with respect to h, where $\hat{f}_{-i}(x_i)$ is the leave-one-out kernel estimator of $f(x_i)$ defined as

$$\hat{f}_{-i}(x_i) = \frac{1}{(n-1)/h} \sum_{j=1, j\neq i}^{n} K\left(\frac{x_i - x_j}{h}\right).$$

In Chapter 3 of this thesis, univariate kernel density estimations with bandwidth selection based on (2.40) and based on the improved version of Sheather and Jones (1991) are done.

2.2.2 Multivariate Kernel Density Estimation

In this section, the univariate kernel density estimation is extended to the multivariate case, following Li and Racine (2007). The kernel density estimator for multivariate data of dimension p is a natural extension of (2.37) and given by

$$\hat{f}(x) = \frac{1}{nh_1 \dots h_p} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), \tag{2.42}$$

with $K\left(\frac{x_i - x}{h}\right) = k\left(\frac{x_{i1} - x_1}{h_1}\right) \times \cdots \times k\left(\frac{x_{ip} - x_p}{h_p}\right)$ and $k(\cdot)$ is an univariate kernel function (see examples in Table 2.1). As in the univariate case, it can be shown, that $\lim_{n\to\infty} \mathrm{MSE}(\hat{f}(x)) = 0$. The bias of (2.42) results as (see Li and Racine 2007)

$$\mathrm{bias}(\hat{f}(x)) = \frac{\sigma_K}{2} \sum_{i=1}^{p} h_i^2 \frac{\partial^2 f(x)}{\partial x \partial x} + O\left(\sum_{i=1}^{p} h_i^3\right)$$

with $\sigma_K = \int u^2 k(u)\ \mathrm{d}u$. Li and Racine (2007) present the variance of $\hat{f}(x)$ as follows

$$\mathrm{var}(\hat{f}(x)) = \frac{1}{nh_1 \dots h_p}\left[\kappa^p f(x) + O\left(\sum_{i=1}^{p} h_i^2\right)\right] = O\left(\frac{1}{h_1 \dots h_p}\right)$$

with $\kappa = \int u^2(u)\,\mathrm{d}(u)$. Combining the results above, the order of $\mathrm{MSE}(\hat{f}(x))$ results as

$$\mathrm{MSE}(\hat{f}(x)) = O\left(\left(\sum_{i=1}^{p} h_i^2\right)^2 + (nh_1 \dots h_p)^{-1}\right).$$

For $n \to \infty, \max_{1\leq i\leq p} h_i \to 0$ and $nh_1 \dots h_p \to \infty$, it follows $\hat{f}(x) \to f(x)$ in MSE, that is $\hat{f}(x) \to f(x)$ in probability. Analogously to the univariate case, the optimal parameters h_i should balance bias and variance terms, i.e. $h_i^4 = O((nh_1 \dots h_p)^{-1})$ and the optimal parameters result as $h_i = c_i n^{-1/(p+4)}$ for positive constant $c_i, i = 1, \dots, p$. Least squares cross-validation in the multivariate case can optimally determine the h_i, Li and Racine (2007) determine the leading term of the cross-validation criterion as follows

$$\mathrm{CV}(h_1, \dots, h_p) = \int \left[\sum_{i=1}^{p} B_i(u) h_i^2\right] \mathrm{d}(u) + \frac{\kappa^p}{nh_1 \dots h_p}, \tag{2.43}$$

where $B_i(u) = (\sigma_K/2) f_{ii}(u)$. One can show, that the values, minimizing (2.43) are optimal smoothing parameters also minimize the leading term of the IMSE.
In Chapter 4, an application of multivariate kernel density estimation is done and the bandwidths $h = (h_1, \dots, h_p)$ are selected following the multivariate analogon of (2.41).

2.3 Copulae

Copula modelling and estimation have become extremely popular over the last decade for modelling the dependence of random variables and their interrelation. This section follows Rank (2007), Nelsen (2006) and Durante and Sempi (2010) introducing the concept and parametric estimation approaches of copulae. At the very beginning, Hoeffding (1940) studied multivariate distributions under 'arbitrary changes of scale', but he did not introduce copulas. Originally introduced by Sklar (1959), the idea of a copula is attractive since it allows to decompose a multivariate distribution into its univariate margins and its interaction structure, expressed through the copula. Assuming the p-dimensional random vector $(x_1, \dots, x_p)$ with univariate marginal distribution $F_j(x_j)$ for $j = 1, \dots, p$, Sklar's theorem states that the joint distribution equals

$$F(x_1, \dots, x_p) = C\big(F_1(x_1), \dots, F_p(x_p)\big), \tag{2.44}$$

where $C(\cdot, \cdot)$ is the copula which is a p-dimensional distribution function $C : [0,1]^p \to [0,1]$ with uniform univariate margins. While $C(\cdot, \cdot)$ is a distribution function, furthermore $C(\cdot, \cdot)$ is monotone increasing in each component u_j . The marginal component i is obtained with $u_j = 1$ for all $j \neq i$, that is $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$. Due to

(2.44), we obtain for continuous F_i and $u = (u_1, \ldots, u_p)$ the copula function, that is

$$C(u) = F(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)) \tag{2.45}$$

with $F_i^{-1}(u_i)$ is the pseudo inverse of $F_i(u_i)$. According to the fact, that (2.45) is a cumulative distribution function, the copula density $c(u)$ can be computed for sufficient differentiable copulas, that is

$$c(u) = \frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1 \cdots \partial u_p}. \tag{2.46}$$

Using the chain rule yields

$$c(u) = \frac{f(F^{-1}(u_1), \ldots, F_p^{-1}(u_p))}{f_1(F_1^{-1}(u1)) \cdots f_p(F_p^{-1}(u_p))}$$

with f is the joint density and f_i are the marginal densities, for $i = 1, \ldots, p$. Describing dependencies, thus analyzing conditional distributions between random variables with known copula $C(\cdot, \cdot)$, is easily done, because the conditional cumulative distribution function may be derived directly from the copula itself. For two random variables U_1 and U_2 and known copula $C(\cdot, \cdot)$, assuming sufficient regularity, the cumulative distribution function results as

$$\begin{aligned} P(U_2 \leq u_2 | U_1 = u_1) &= \lim_{\delta \to 0} \frac{P(U_2 \leq u_2, U_1 \in (u_1 - \delta, u_1 + \delta])}{P(U_1 \in (u_1 - \delta, u_1 + \delta])} \\ &= \lim_{\delta \to 0} \frac{C(u_1 + \delta, u_2) - C(u_1 - \delta, u_2)}{2\delta} \\ &= \frac{\partial}{\partial u_1} C(u_1, u_2). \end{aligned}$$

Each copula $C(\cdot, \cdot)$ lies between certain bounds, named Fréchet-Hoeffding bounds. Hoeffding (1940) and Fréchet (1951) showed, that

$$\max\left\{\sum_{i=1}^{p} u_i + 1 - p, 0\right\} \leq C(u) \leq \min\{u_1, \ldots, u_p\}.$$

The Fréchet-Hoeffding bounds are related to copulas. The comonotonicity copula is given by

$$M(u) = \min\{u_1, \ldots, u_p\} \tag{2.47}$$

and refers the case of perfect positive dependence. Increasing transformations $T_2, \ldots, T_p$ are defined as $U_i = T_i(U_i)$ for $i = 2, \ldots, p$. Using (2.44), these random variables follows the comonotonicity copula. The countermonotonicity copula describes the opposite

extreme. It is defined for two random variables U_1 and U_2 as

$$W(u) = \max\{u_1 + u_2 - 1, 0\}. \tag{2.48}$$

This copula describes negative dependence, as $U_2 = T(U_1)$ with strictly increasing function T. Both copulas (2.47) and (2.48) are not differentiable, thus they do not have densities.

2.3.1 Copula Families

Copulas are used to describe various dependencies for building stochastic models. Therefore, different copula families are investigated in the literature, beyond the comonotonicity and countermonotonicity copula families. Joe (1997) gave some inspirations about properties of a 'good' copula family. He mentioned interpretability, a flexible and wide range of dependence and an easy handling. First of all, the independence copula

$$\Pi(u) = \prod_{i=1}^{p} u_i \tag{2.49}$$

describes the case of no dependence beyond the considered data. Using (2.44), random variables are independent, if and only if, their copula is the independence copula, thus the associated copula density is constant. Further copula families are investigated in the literature. The so called elliptical copulas are derived from multivariate distributions. $U = (U_1, \ldots, U_p)$ is said to have an elliptical distribution with mean $\mu \in \mathbb{R}^p$, covariance matrix Σ and generator $g : [0, +\infty[\rightarrow [0, +\infty[$, if U can be expressed in the form $U = \mu + RAW$, with AA^T is the Cholesky decomposition of $\Sigma = (\sigma_{ij})$, W is a p-dimensional random vector uniformly distributed on the sphere $S^{p-1} = \{w \in \mathbb{R}^p : w_1^2 + \cdots + w_p^2 = 1\}$ and R is a positive random variable independent of W with density, given for every $r > 0$, by

$$f_g(r) = \frac{2\pi^{p/2}}{\Gamma(p/2)} r^{p-1} g(r^2).$$

The first class of copula distribution, considered later in this thesis follows an elliptical distribution. The multivariate Gaussian and multivariate t-distribution contain to this class. If the density of an elliptical copulas distribution exists, it is given for $x \in \mathbb{R}^p$ by

$$h_g(x) = |\Sigma|^{-1/2} g((x - \mu)^T (x - \mu)). \tag{2.50}$$

Using the generator function $g(t) = (2\pi)^{-p/2} \exp(-t/2)$ in (2.50), U has a multivariate Gaussian distribution. U follows a multivariate t-distribution with ν degrees of freedom, if $g(t) = c(1 + t/\nu)^{-1(p+\nu)/2}$ is used in (2.50) for a suitable constant c. Considering p

normally distributed random variables $U_1, \ldots, U_p$, the multivariate Gaussian copula is defined as

$$C_{\Sigma}^{Ga}(u) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)) \tag{2.51}$$

with Φ as the cumulative distribution function of a standard normal distribution, while Φ_{Σ} is the cumulative distribution function for a p-variate normal distribution with zero mean and covariance matrix Σ. Analogously, the t-copula describes the multivariate case for p random variables, following a t-distribution. The t-copula is given by

$$C_{\nu,\Sigma}^{t}(u) = t_{\nu,\Sigma}(t_{\nu}^{-1}(u_1), \ldots, t_{\nu}^{-1}(u_p)), \tag{2.52}$$

with Σ is a correlation matrix, t_ν is the cumulative distribution function of the one-dimensional t_ν distribution with ν degrees of freedom and $t_{\nu,\Sigma}$ is the cumulative distribution function of the multivariate $t_{\nu,\Sigma}$ distribution with ν degrees of freedom.
The second class of copula distribution, considered later in this thesis, are the Archimedean copulas, introduced following McNeil and Neslehová (2009). The Archimedean generator is any decreasing and continuous function $\psi : [0, \infty[\rightarrow [0, 1]$ and satisfying $\psi(0) = 1, \lim_{t\rightarrow\infty} \psi(t) = 0$, which is strictly decreasing on $[0, \inf\{t|\psi(t) = 0\}[$. Moreover, by definition $\psi(+\infty) = 0$ and $\psi^{-1}(0) = \inf\{t \geq 0|\psi(t) = 0\}$, denoting with $\psi(t)^{-1}$ the pseudo-inverse of $\psi(t)$. So, a p-dimensional copula C is called Archimedean copula, if

$$C(u) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_p)) \tag{2.53}$$

for some Archimedean generator ψ. McNeil and Neslehová (2009) stated for an Archimedean generator ψ and for C_ψ given in (2.53), that C_ψ is a p-dimensional copula, if and only if, the restriction of ψ to $]0, \infty[$ is p-monotone, i.e. it satisfy

a) ψ is differentiable up to the order $p-2$ in $]0, \infty[$ and the derivatives satisfy $(-1)^k \psi^{(k)}(t) \geq 0$ for $k \in 0, \ldots, d-2$ for every $t > 0$

b) $(-1)^{p-2}\psi^{(p-2)}$ is decreasing and convex in $]0, +\infty[$.

Well known Archimedean copulas are the Gumbel, Frank and Clayton copula. Originally, the Gumbel-Hougaard copula is considered in Gumbel (1960) and extended in Hougaard (1986). Very often this copula family is named Gumbel copula and is given by

$$C_{\theta}^{GH}(u) = \exp\left(-\left(\sum_{i=1}^{p}(-\log(u_i))^{\theta}\right)^{1/\theta}\right) \tag{2.54}$$

with $\theta \geq 1$. The corresponding generator function is $\psi(t) = \exp(-t^{1/\theta})$. For $\theta = 1$ in (2.54), the independence copula (2.49) is obtained as special case. For $\theta \rightarrow +\infty$, the

limit of (2.54) is the comonotonicity copula (2.47) (see Durante and Sempi 2010). The Mardia-Takahasi-Clayton copula is defined as

$$C_\theta^{MTC}(u) = \max\left\{\left(\sum_{i=1}^{p} u_i^{-\theta} - (p-1)\right)^{-1/\theta}, 0\right\} \tag{2.55}$$

with $\theta \geq \frac{-1}{p-1}, \theta \neq 0$. For $\theta \to 0$, (2.55) coincide with (2.49), that is the independence copula. The generator for (2.55) is given by $\psi_\theta(t) = (\max\{1+\theta t, 0\})^{-1/\theta}$. McNeil and Neslehová (2009) proved, that for every p-dimensional Archimedean copula C and for every $u \in \mathbb{R}^p$ $C_{\theta_L}^{MTC}(u) \leq C(u)$ for $\theta_L = \frac{-1}{p-1}$. (2.55) can be derived from the pareto distribution by Mardia (1962). Also the Burr distribution by Takahasi (1965) is associated with the Clayton's model (see Clayton 1978). So, the copula family is often named Clayton copula.

Another Archimedean copula family is the Frank copula (see Frank 1979), given by

$$C_\theta^{Fr}(u) = -\frac{1}{\theta}\log\left(1 + \frac{\prod_{i=1}^{p}(\exp(-\theta u_i) - 1)}{(\exp(-\theta) - 1)^{p-1}}\right), \tag{2.56}$$

where $\theta > 0$. For $\theta \to 0$ (2.56) equals (2.49), that is the independence copula and for $p = 2$, θ can also be selected as $\theta < 0$. The Archimedean generator for (2.56) is $\psi_\theta(t) = -\frac{1}{\theta}\log(1 - (1 - \exp(-\theta))\exp(-t))$.

Tail dependence measures the correlation between the variables in the upper-right quadrant and in the lower-left quadrant of $[0,1]^2$. These correlations are of special interest in many applications, analyzing dependencies in the extreme cases. For two random variables U_1 and U_2 with cumulative distribution functions $F_i, i = 1, 2$, the coefficient of upper tail dependence is defined as

$$\lambda_u = \lim_{w \to 1} P(U_2 > F_2^{-1}(w) | U_1 > F_1^{-1}(w)) = \lim_{w \to 1} \frac{1 - 2w + C(w,w)}{1 - w},$$

if the limit exists and $\lambda_u \in [0,1]$. Intuitively, for large values of U_1, also large values of U_2 are expected. The coefficient of lower tail dependence is defined as

$$\lambda_l = \lim_{w \to 0} P(U_2 \leq F_2^{-1}(w) | U_1 \leq F_1^{-1}(w)) = \lim_{w \to 0} \frac{C(w,w)}{w}.$$

Similarly, for small values of U_1, small values of U_2 are also expected. Nelsen (2006) calculates λ_u and λ_l for the families of Archimedean copulas. Rank (2007) calculates the tail dependence coefficients for the bivariate t-copula (2.52) with $\Sigma = \rho$ in the bivariate case. Some of these results are listed in Table 2.2. The Gumbel copula (2.54) has no lower tail dependence, but upper tail dependence. In contrast, the Clayton

copula family	λ_l	λ_u
Gumbel	0	$2 - 2^{1/\theta}$
Clayton	$2^{-1/\theta}$	0
Frank	0	0
t-copula $t_{\nu,\rho}$	$2t_{\nu+1}\left(-\sqrt{(}\frac{\nu+1)(1-\rho)}{1+\rho}\right)$	$2t_{\nu+1}\left(-\sqrt{(}\frac{\nu+1)(1-\rho)}{1+\rho}\right)$

Table 2.2: Tail dependence for various copula families.

copula (2.55) has lower tail dependence, but no upper tail dependence. The Frank copula has no tail dependences. Joe (1997) and Nelsen (2006) give overviews about further classes of copula families which are not mentioned in this thesis.
Exemplary plots of some copula families are presented in Figure 2.4, observing different characteristics for each copula family. Beginning with a) Gumbel copula in Figure 2.4, we observe upper tail dependence, thus a peak around $(1, 1)$. The Clayton copula b) shows lower tail dependence, thus a peak around $(0, 0)$. McNeil, Frey, and Embrechts (2005) computed the tail dependence for the Gaussian copula with the result of asymptotical independence in upper and lower tails. Therefore, the Gaussian copula do not have any tail dependence, independent of its correlation parameter. Correlation of copulas is often described using Kendell's tau and Spearman's rho. For random variables $U = \{U_1, \ldots, U_p\}$ with marginals $F_1, \ldots, F_p$, respectively, Spearman's rho matrix is defined by

$$\rho_S(U) = \mathrm{Corr}(F_1(U_1), \ldots, F_p(U_p)),$$

with $\rho_S(U)_{i,j} = \mathrm{Corr}(F_i(U_i), F_j(U_j))$. Alternatively, Kendell's tau for two random variables U_1 and U_2 and two random variables $\tilde{U}_1$ and $\tilde{U}_2$ with the same joint distribution, but independent of U_1 and U_2, is defined as

$$\rho_\tau(U_1, U_2) = E[\mathrm{sign}((U_1 - \tilde{U_1}) \cdot (U_2 - \tilde{U_2}))].$$

That is, if we plot two points from these random variables on a graph, connecting them by a line, the line is increasing for positive dependence and decreasing otherwise. For $(U_1 - \tilde{U}_1) \cdot (U_2 - \tilde{U}_2)$ a positive sign indicates an increase, while a negative sign would denote a decrease. If both probabilities are equal, that is upward and downward slopes are expected with the same probability, Kendell's tau is $\rho_\tau = 0$. If $\rho_\tau > 0$, a higher probability of upward slope is expected, for a negative value of ρ_τ a downward slope. In the p-dimensional case, for a random variable U and an independent copy $\tilde{U}$, Kendell's tau is defined as

$$\rho_\tau(X) = \mathrm{Cov}[\mathrm{sign}(U - \tilde{(}U)].$$

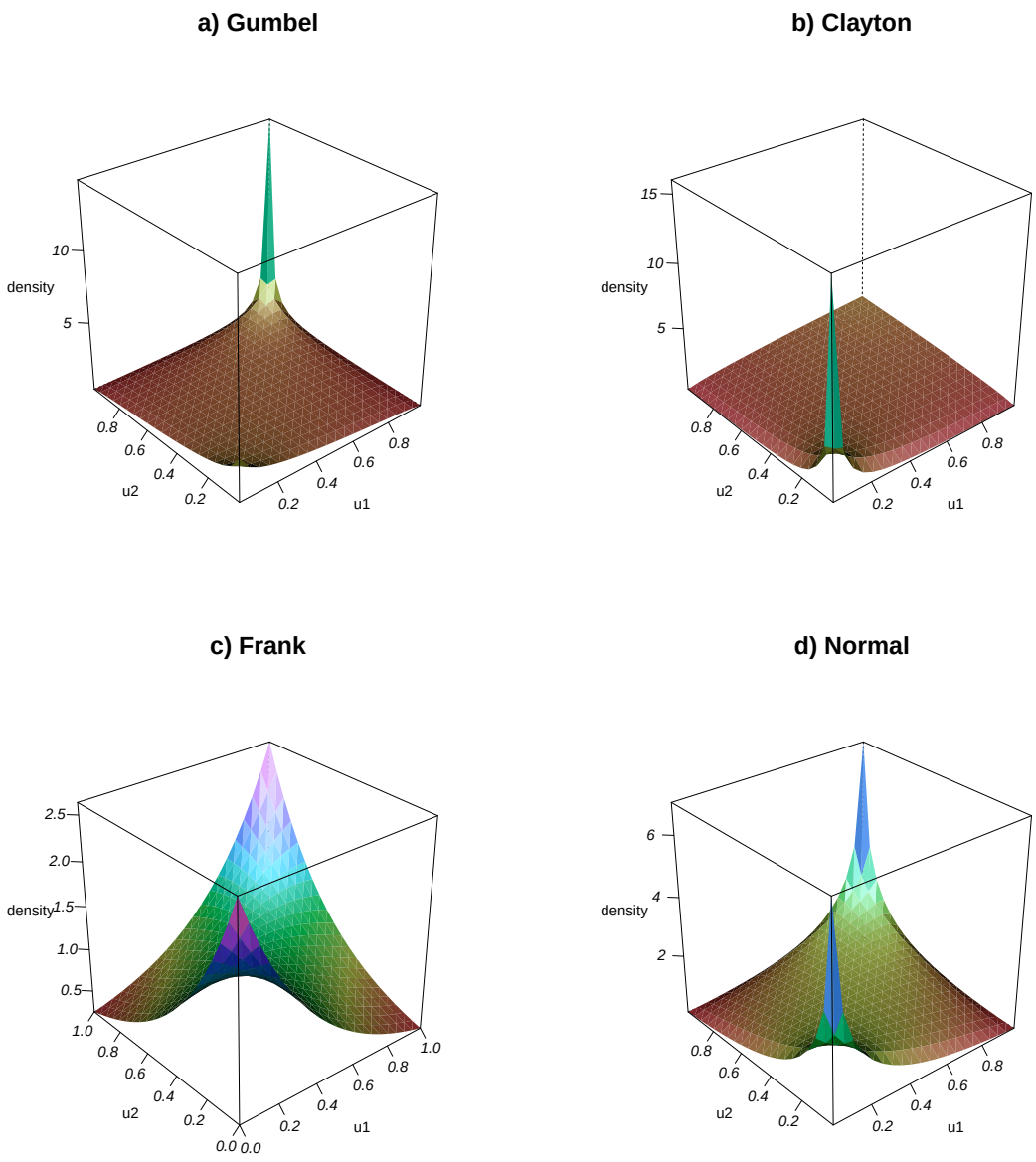


Figure 2.4: Exemplary copula plots: a) Gumbel copula with $\theta = 1.33$, b) Clayton copula with $\theta = 2/3$, c) Frank copula with $\theta = 2.39$ and d) Gaussian copula with $\theta = 0.5$.

2.3.2 Copula Estimation

Estimation methods for copula models, using maximum likelihood estimation (MLE), are considered in this paragraph, following Choros, Ibragimov, and Permiakova (2010) and Joe (1997). This parametric estimation approach is used in the simulation studies in Chapter 4 and 5. Due to Sklar's theorem (2.44), the likelihood function of

a p-dimensional copula density (2.46) is given by

$$l = \sum_{j=1}^{n} \log f(x_1^{(j)}, \ldots, x_p^{(j)}) \tag{2.57}$$

for an (i.i.d.) random sample $x^{(j)} = (x_1^{(j)}, \ldots, x_p^{(j)}), j = 1, \ldots, n$ with density f. For random samples with dependent margins, decomposing the log-likelihood, with respect to the dependence structure represented by copula C, that is

$$l_C = \sum_{j=1}^{n} \log c(F_1(x_1^{(j)}), \ldots, F_p(x_p^{(j)}))$$

and the marginal log-likelihoods

$$l_i = \sum_{j=1}^{n} \log f(x_i^{(j)})$$

results in $l = l_C + \sum_{i=1}^{p} l_i$. The copula C depends on a (vector) parameter θ and each margin f_i on (vector) parameters α_i, that is maximum likelihood estimators $(\hat{\alpha}_1^{MLE}, \hat{\alpha}_2^{MLE}, \ldots, \hat{\alpha}_p^{MLE}, \hat{\theta}_d^{MLE})$ result simultaneously by maximization of (2.57):

$$\begin{aligned}
(\hat{\alpha}_1^{MLE}, \hat{\alpha}_2^{MLE}, \ldots, \hat{\alpha}_p^{MLE}, \hat{\theta}_d^{MLE}) &= \\
\arg \max_{\alpha_1, \ldots, \alpha_p, \theta} l_C(\alpha_1, \ldots, \alpha_p, \theta) + \sum_{i=1}^{p} l_i(\alpha_i) &= \\
\arg \max_{\alpha_1, \ldots, \alpha_p, \theta} \sum_{j=1}^{n} \log c(F_1(x_1^{(j)}; \alpha_1), F_2(x_2^{(j)}; \alpha_2), \ldots, F_p(x_p^{(j)}, \alpha_p); \theta) + & \\
\sum_{i=1}^{p} \sum_{j=1}^{n} \log f_i(x_i^{(j)}; \alpha_i) \qquad .&
\end{aligned}$$

Alternatively, Joe (1997) discusses the method of inference functions for margins (IFM). In a first step, the marginal coefficients α_i are estimated from the log-likelihood l_i of each margin, that is $\hat{\alpha}_i^{IFM} = \arg\max_{\alpha_i} l_i(\alpha_i)$. Replacing α by their estimations $\hat{\alpha}_i^{IFM}$ in the copula likelihood l_C, the estimator $\hat{\theta}^{IFM}$ is computed by maximizing $l_C(\hat{\alpha}_1^{IFM}, \ldots, \hat{\alpha}_p^{IFM}, \theta)$. The MLE estimator solves, under regularity conditions,

$$(\partial l/\partial \alpha_1, \partial l/\partial \alpha_2, \ldots, \partial l/\partial a_p, \partial l/\partial \theta) = 0,$$

while the IFM estimator solves

$$(\partial l_1/\partial \alpha_1, \partial l_2/\partial \alpha_2, \ldots, \partial l_p/\partial a_p, \partial l/\partial \theta) = 0.$$

Joe (1997) shows, that MLE and IFM estimations are equivalent in the special cases of multivariate normal distribution functions. Moreover, Choros, Ibragimov, and Permiakova (2010) mention, that the IFM estimator is consistent and asymptotically normal under the usual regularity conditions and that the IFM estimator provides a highly efficient alternative to the MLE estimator.
Genest, Ghoudi, and Rivest (1995) discuss the semi-parametric estimation as an alternative to the inference discussed above, estimating the univariate margins F_i nonparametrically, e.g. by the empirical distribution functions $\hat{F}_i$ in the first step. Given $\hat{F}_i$, the copula parameter θ are estimated as

$$\hat{\theta} = \arg\max_{\theta} L_C(\theta) = \arg\max_{\theta} \sum_{j=1}^{n} \log c(\hat{F}_1(x_1^{(j)}), \ldots, \hat{F}_p(x_p^{(j)}); \theta).$$

Genest, Ghoudi, and Rivest (1995) show, that the estimated parameters $\hat{\theta}$ of θ are consistent and asymptotically normal under suitable regularity conditions. Furthermore, the authors assume same regularity assumptions for bivariate copulas, which are fulfilled by many copula families, and show, that the estimator $\hat{\theta}$ is fully efficient at independence.
Alternatively, nonparametric inference for copula estimation is applied (see Choros, Ibragimov, and Permiakova 2010), while an estimator $\hat{C}(u_1, \ldots, u_p)$ of a p-dimensional copula $C(u_1, \ldots, u_p)$ is usually an empirical inversion, that is

$$\hat{C}(u_1, \ldots, u_p) = \hat{F}(\hat{F}_1^{-1}(u_1), \ldots, \hat{F}_p^{-1}(u_p))$$

with $\hat{F}$ is a nonparametric estimator of the distribution function F and $\hat{F}_1^{-1}, \ldots, \hat{F}_p^{-1}$ are nonparametric estimators of the pseudo-invers $F_i^{-1}(s) = \{t | F_i(t) \geq s\}$ of the univariate margins. $\hat{F}$ is usually taken to be the empirical univariate distribution function and $\hat{F}_i^{-1}(s)$ is estimated by the pseudo-invers of the empirical distribution function. This empirical process is consistent and asymptotic normal for general copulas C with continuous partial derivatives (see Fermanian, Radulovic, and Wegkamp 2004 and Fermanian and Scaillet 2003). Fermanian, Radulovic, and Wegkamp (2004) show also, that smoothed copula processes like $\hat{C}(u_1, u_2) = \hat{F}(\hat{F}_1^{-1}(u_1), \hat{F}_2^{-1}(u_2))$ are also asymptotic normal under regularity conditions. Fermanian, Radulovic, and Wegkamp (2004) use nonparametric kernel estimators $\hat{F}(x_1, x_2) = \sum_{t=1}^{T} K\left(\frac{x - X_t}{h_T}, \frac{y - Y_t}{h_T}\right)$ of the joint distributions functions for some bivariate kernel function K for bandwidths h_T, satisfying

$h_T \to 0$ as $T \to \infty$.

2.4 Dependence Vines

Dependence vines, especially D-vines are investigated in Chapter 5. In this subsection, the concept and estimation of D-vines is introduced. The principle of dependence vines is modelling flexible multivariate distributions as discussed in this section, following Kurowicka and Cooke (2006) and Czado (2010). As recent overview about this topic is also given by Kurowicka and Joe (2010). Both references focus on the analysis of dependence structures in multivariate data, introducing vines. Let $x = (x_1, \ldots, x_p)$ be a p-dimensional continuous random vector with continuously differentiable marginal distribution functions $F_j(x_j), j = 1, \ldots, p$. Let $f(x_1, \ldots, x_p)$ be the corresponding multivariate density, which with Sklar's (1959) theorem can be written as

$$f(x_1, \ldots, x_p) = c\{F_1(x_1), \ldots, F_p(x_p)\} \prod_{j=1}^{p} f_j(x_j), \tag{2.58}$$

where $c(.)$ is the copula density. To simplify notation, we denote with $u_j = F_j(x_j)$ so that the copula density is written as $c(u_1, \ldots, u_p)$. For dimension $p = 2$, the conditional density of X_1 given X_2, using (2.58) yields

$$f(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2)) f(x_1), \tag{2.59}$$

where c_{12} is a bivariate copula, which is often called pair-copula. Extending (2.59) to the multivariate case with distinct indices $i, j, i_1, \ldots, i_p$ with $i < j$ and $i_1 < \cdots < i_p$, the conditional density $c_{i,j|i_1,\ldots,i_p}$ is defined as

$$c_{i,j|i_1,\ldots,i_p} = c_{i,j|i_1,\ldots,i_p}(F(x_i|x_{i_1}, \ldots, x_{i_k}, F(x_j|x_{i_1}, \ldots, x_{i_p})).$$

The density $f(x_t|x_1, \ldots, x_{t-1})$ results recursively, using (2.59) for the conditional distribution of (X_1, X_t) given $X_2, \ldots, X_{t-1}$, is given as

$$f(x_t|x_1, \ldots, x_{t-1}) = \left[\prod_{s=1}^{t-2} c_{s,t|s+1,\ldots,t-1} \right] c_{(t-1),t} f_t(x_t). \tag{2.60}$$

That is, the conditional density $f(x_t|x_1, \ldots, x_{t-1})$ is constructed by different pair-copulas $c_{i,j|i_1,\ldots,i_p}$. Bedford and Cooke (2002) introduced the class of regular vines. To describe dependences structures in high-dimensional distributions, a dependence tree as an acyclic undirected graph is used. Each tree consists of nodes $N = 1, \ldots, n$

and edges E, where E is an unordered subset of N with no cycle. Each regular vine on n variables consists of nested trees, where the edges of tree j are the nodes of the tree $j+1$ and each tree exhibits the maximum number of nodes. In a regular vine V on n variables, each pair of two edges in tree j are connected by an edge in tree $j+1$, if these edges assign a common node. V is called vine on n elements, if $V = (T_1, \ldots, T_{n-1})$ and T_1 is a connected tree with nodes $N_1 = 1, \ldots, n$ and edges E_1 and for $i = 2, \ldots, n-1$, T_i is tree with nodes $N_i = E_{i-1}$. V is called regular vine, if additionally the proximity condition is fulfilled, that is if c and d are nodes of T_i connected by an edge in T_i, where $c = \{c_1, c_2\}$ and $d = \{d_1, d_2\}$, then exactly one of the a_i equals one of the b_i.
A regular vine is called a D-vine, if the number of edges attached to a node equals at most 2. Figure 5.1 shows a D-vine for $p = 5$. Fitting (2.60) in (2.58) with $s = i, t = i+j$, the multivariate density f results as

$$f(x_1, \ldots, x_p) = \left[\prod_{t=2}^{p}\prod_{s=1}^{t-2} c_{s,t|s+1,\ldots,t-1}\right]\left[\prod_{t=2}^{p} c_{(t-1),t}\right]\left[\prod_{k=1}^{p} f_k(x_k)\right]. \tag{2.61}$$

(2.61) consists of pair-copula densities $c_{i,j|i_1,\ldots,i_p}$ and marginal densities f_k and (2.61) is the distribution of a D-vine. This principle is called the pair-copula construction principle. A regular vine is called a canonical or C-vine, if each tree T_i has a unique node with $n-i$ number of edges attached to the node. The node with maximal number of edges attached to the node in T_1 is the root, that is the node with $p-1$ edges in tree T_1. If one applies (2.59) to the conditional distribution of (X_{t-1}, X_t) given $X_1, \ldots, X_{t-2}$ to express $f(x_t|x_1, \ldots, x_{t-1})$ recursively, we get

$$f(x_t|x_1, \ldots, x_{t-1}) = c_{t-1,t|1,\ldots,t-2} f(x_t|x_1, \ldots, x_{t-2}). \tag{2.62}$$

Fitting (2.62) into (2.58) for $j = t-k, j+1 = t$ yields

$$f(x_1, \ldots, x_p) = \left[\prod_{j=1}^{p-1}\prod_{i=1}^{d-j} c_{j,j+1|1,\ldots,j-1}\right]\prod_{k=1}^{p} f_k(x_k), \tag{2.63}$$

which is the distribution of a canonical vine. Denoting the edges in tree T_i by $jk|D$ where $j < k$ and D is the conditioning set, the notation of the edges e in tree T_i depends on the two edges in tree T_{i-1}, which have a common node in tree T_{i-1}. The edges are noted by $a = j(a), k(a)|D(a)$ and $b = j(b), k(b)|D(b)$ with $V(a) = \{j(a), k(a), D(a)\}$ and $V(b) = \{j(b), k(b), D(b)\}$. Therefore, nodes a and b are joined

by edge $e = j(e), k(e)|D(e)$, where

$$\begin{aligned} j(e) &= \min\{i : i \in (V(a) \cup V(b)) \setminus D(e)\} \\ k(e) &= \max\{i : i \in (V(a) \cup V(b)) \setminus D(e)\} \\ D(e) &= V(a) \cap V(b). \end{aligned}$$

Fitting a regular vine with node set $\mathcal{N} = \{N_1, \ldots, N_{p-1}\}$ and edge set $\mathcal{E} = \{E_1, \ldots, E_{d-1}\}$, each edge $e = j(e), k(e)|D(e)$ in E_i is associated with a bivariate copula density $c_{j(e),k(e)|D(e)}$. $X_{D(e)}$ denotes the sub random vector of $X = (X_1, \ldots, X_p)$ indicated by indices $D(e)$. A vine distribution of the random vector X with marginal densities $f_k, k = 1, \ldots, p$ and the conditional density of $(X_{j(e)}, X_{k(e)})$ given $x_{D(e)}$ is defined as $c_{j(e),k(e)|D(e)}$ for the regular vine tree with node set $\mathcal{N}$ and edge set $\mathcal{E}$. Kurowicka and Cooke (2006) proved, that the joint density of X is uniquely determined and given by

$$f(x_1, \ldots, x_p) = \prod_{j=1}^{p} f(x_j) \prod_{i=1}^{p-1} \prod_{e \in E_i} c_{j(e),k(e)|D(e)}(F(x_{j(e)}|x_{D(e)}), F(x_{k(e)}|x_{D(e)}))$$

with $x_{D(e)}$ denotes the sub-vector of x indicated by $D(e)$. Exemplarily, the vine distribution of the D-vine in Figure 5.1 has the joint density given by

$$f(x_1, \ldots, x_5) = \prod_{k=1}^{5} f_k(x_k) \cdot c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{45} \cdot c_{13|2} \cdot c_{24|3} \cdot c_{35|4} \cdot \cdot c_{14|23} \cdot c_{25|34} \cdot c_{15|234}.$$

Using the pair-copula construction principle, any bivariate copula family (see Section 2.3.1) may be optimal any node of the dependence vines. Due to the bivariate case, the parameter of each possible copula family are easily estimated using e.g. maximum likelihood theory (see Section 2.3.2).

2.4.1 Estimation of Regular Vine Copulas

Aas, Czado, Frigessi, and Bakken (2009) talk firstly about stepwise estimation and maximum likelihood estimation for the vine copula parameters. The joint density for a C-vine (2.63) or D-vine (2.61) is explicitly given, so the likelihood is easily derived. The main task is to consider the involved conditional distribution functions. Joe (1996) shows for $v \in D$ and $D_{-v} = D \setminus v$

$$F(x_j|x_D) = \frac{\partial C_{x_j,x_v|D-v}(F(x_j|x_{D_{-v}}), F(x_v|x_{D_{-v}}))}{\partial F(x_v|x_{D_{-v}})}. \tag{2.64}$$

If D consists only of one element, that is $D = \{v\}$, it follows that

$$F(x_j|x_v) = \frac{\partial C_{x_j,x_v}(F(x_j), F(x_v))}{\partial F(x_v)}.$$

For uniform margins, using a parameterized copula conditional distribution function $C_{jv}(x_j, x_v) = C_{jv}(x_j, x_v|\theta_{jv})$ one can write

$$h(x_j|x_v, \theta_{jv}) = \frac{\partial C_{j,v}(x_j, x_v|\theta_{jv})}{\partial x_v}. \tag{2.65}$$

Conditional distribution functions where D contains more than one element can be expressed using (2.64). Czado (2010) presents the recursive relation

$$F(x_j|x_D) = h(F(x_j|x_{D_{-v}})|F(x_v|x_{D_{-v}}), \theta_{jv|D_{-v}}). \tag{2.66}$$

So, the conditional distribution functions with conditioning set D can be calculated recursively using the h-function, following from lower dimensional conditional set as given by (2.66). Thereby, the number of parameters of a pair-copula construction to be estimated grow quadratically in the dimension p, $p \cdot (p-1)/2$ different pair-copulas have to be parameterized. Therefore, the parameters corresponding to the pair-copulas should be estimated sub-sequentially from the first tree to the last tree.
It exists $p!/2$ distinct C-vines or D-vines for a decomposition on p nodes (see Aas, Czado, Frigessi, and Bakken 2009). Therefore, additional information are needed to select suitable vine trees. In the case of a D-vine Aas, Czado, Frigessi, and Bakken (2009) order the first level of the D-vine due to the strongest bivariate dependencies, which might be measured by Kendell's τ or tail dependencies (see Section 2.3). If the order of the first level has been chosen, the parameters are selected, applying a goodness-of-fit test for each pair, varying the copula families and selecting the copula family with the best fit. If the first tree is fitted, using the recursive formula (2.65) allows to calculate the next tree of the vine. If there are M possible copula families, there are $M \cdot p \cdot (p-1)/2$ different pair-copulas to be selected and compared during the estimation of the vine. Applying goodness-of-fit tests on the full p dimensional sample, would involve fitting $M^{p\cdot(p-1)/2}$ models, but the computational effort would increases excessively even for small M and small p. Alternatively, Bayesian approaches with applications of Markov chain Monte Carlo method (MCMC) exist (see Smith, Min, Almeida, and Czado 2010 or Min and Czado 2011), which are not considered in this thesis in detail.

2.4.2 Sampling from D-vines

Once the optimal D-vine has been completely estimated, sampling is interesting for further uses of fitted models. Sampling from a fitted D-vine is done with the standard sampling algorithm for D-vines (see Kurowicka and Cooke 2006 or Aas, Czado, Frigessi, and Bakken 2009). We illustrate the algorithm of Kurowicka and Cooke (2006) for four variables in Figure 2.5. At the beginning, we sample four independent uniform (0,1) variables $u_1, \ldots, u_4$. Within the algorithm, the values of the conditional distribution functions $\hat{F}(\cdot)$ are determined, using equation (2.65) with the estimated coefficients $\hat{v}$ of the corresponding D-vine. In the following, the inverse of each conditional distribution function $\hat{F}^{-1}(\cdot)$ is numerically approximated. At the start x_1 is given and x_2 is easily calculated by inverting the conditional distribution function of $F(u_2|x_1)$. If x_2 has been calculated, $F(x_1|x_2)$, $\hat{F}^{-1}(u_3|\hat{F}(x_1|x_2))$ and then $\hat{F}^{-1}(\hat{F}^{-1}(u_3|\hat{F}(x_1|x_2))|x_2)$ must be evaluated to obtain an estimate of x_3. Of course, this is easily done in higher dimensions. So, computational demand for sampling of a D-vine increases with extended dimension of the D-vine.

Sample $w_1, \ldots, w_4$ independent uniform on $[0, 1]$.
$x_1 = w_1$
$x_2 = u_{2|1}^{-1} = \hat{F}^{-1}(w_2|x_1)$
$x_3 = u_{3|2}^{-1}(u_{3|12}^{-1}) = \hat{F}^{-1}(\hat{F}^{-1}(w_3|\hat{F}(x_1|x_2))|x_2)$
$x_4 = u_{4|3}^{-1}(u_{4|23}^{-1}(u_{4|123}^{-1})) = \hat{F}^{-1}(\hat{F}^{-1}(\hat{F}^{-1}(w_4|\hat{F}(x_1|x_2, x_3)|\hat{F}(x_2|x_3)|x_3)))$

Figure 2.5: Sampling algorithm for D-vine

3 Density Estimation and Comparison with a Penalized Mixture Approach

This chapter is joint work with Göran Kauermann (LMU Munich). It is forthcoming in *Computational Statistics*, compare Schellhase and Kauermann (2012).

The focus of Chapter 3 is an application of penalized smoothing splines to estimate univariate density functions. The idea is to represent the unknown density by a convex mixture of basis densities, where the weights are estimated in a penalized form. The proposed method extends the work of Komárek and Lesaffre (2008) and allows for general density estimation. Simulations show that the proposed approach outperforms existing density estimation approaches. The idea is extended to allow the density to depend on some (factorial) covariate. Additionally, we can test on equality of the densities in the groups, assuming a binary group indicator. This provides a smooth alternative to the classical Kolmogorov-Smirnov test or an Analysis of Variance and it shows stable behaviour.

3.1 Introduction

Density estimation has a long standing tradition in statistics and the different routines can be roughly categorized in four partly overlapping approaches. (a) First and most prominent there is kernel density estimation which traces back to ideas of Nadaraya (1964) and Watson (1964), see also Nadaraya (1974). The method is well established and extensively discussed in e.g. Wand and Jones (1995) or Simonoff (1996). (b) A second approach results by writing the unknown density as

$$\hat{f}(y) = \exp\{\eta(y)\} \,/ \int \exp\{\eta(z)\}\, dz \tag{3.1}$$

with $\eta(\cdot)$ unknown but smooth function which is estimated using spline technology. This approach traces back to Good and Gaskins (1971), see also Silverman (1982) and the idea has been further developed by Gu (1993) or Dias (1998), see also Gu and Wang (2003). (c) A third approach results by extending and smoothing the classical

histogram as originally suggested by Boneva, Kendall, and Stefanov (1971). Following this idea Lindsey (1974a, 1974b) suggests density estimation by transferring the density estimation problem to a regression estimation scenario, with the number of observations per bin in the histogram as Poisson count, see also Efron and Tibshirani (1996). Eilers and Marx (1996) make use of the idea using penalized spline smoothing, see also Ruppert, Wand, and Carroll (2003). The spline approach and the Poisson approach (c) are thereby closely related which results by approximating the integral in (3.1) with a rectangular method. (d) A fourth line of density estimation has been suggested by using a mixture approach. In this case, the unknown density results by finite mixture of densities components. These mixture components are usually built from known distributions (e.g. normal) with unknown parameters. This yields the classical mixture models discussed extensively in McLachlan and Peel (2000), see also Young, Hunter, Chauveau, and Benaglia (2009), Li and Barron (1999) or Fraley and Raftery (2002). (e) Another approach to estimate the unknown density is the log-spline approach (see Koo, Kooperberg, and Park 1999), modelling the log-density function by (almost cubic) splines using maximum likelihood estimation and Newton-Raphson method to compute optimal coefficients. (f) A sixth idea to estimate densities is tackled using wavelets, expanding the unknown density in terms of a wavelet expansion (see e.g. Hall and Patil 1995, Nason and Silverman 1999 or Nason 2008). Our approach (g) presented in this paper distinguishes from the classical mixture model in two ways. First, we take completely specified mixture components, that is not only the distribution type, but also the parameters are fixed. Secondly, the number of mixture components is chosen in a lavish way and we impose a penalty to achieve smooth density fits. Ghidey, Lesaffre, and Eilers (2004) have proposed to use a finite but penalized mixture of Gaussian densities for the estimation of a random effect distribution in a linear mixed model. The idea has been extended and further developed in a number of papers which include Komárek, Lesaffre, and Hilton (2005), Komárek (2006) and Komárek and Lesaffre (2008). The idea of Komárek (2006) shows also similarities to the approach of Babu, Canty, and Chaubey (2002), who approximate the density with a mixture of Bernstein polynomials. In this paper we generalize the original idea of Komárek and Lesaffre (2008) to univariate density estimation. Extending the mixture to a continuous mixture has recently been proposed by Liu, Levine, and Zhu (2009).

In this paper we follow (g) using finite mixture densities for the smooth estimation of an unknown density. The collection of the densities used in the mixture in fact plays the role of a basis and the weights correspond to basis coefficients. The weights itself can be fitted with penalized techniques to obtain a smooth density fit. In principle, any type of mixture density can be used and there is no requirement for Gaussian

mixtures. In this paper we make use of a mixture of B-spline basis functions normed to be densities. This allows to theoretically investigate the properties of the fit and also guarantees stable numerical performance. To achieve smoothness we make use of penalized spline smoothing in the style of Ruppert, Wand, and Carroll (2003), see also O'Sullivan (1986) and Eilers and Marx (1996). With the link between penalized spline smoothing and mixed models (see Wand 2003) the method shows its full flexibility and versatility as demonstrated in the commendable survey recently composed by Ruppert, Wand, and Carroll (2009).

A general question in penalized spline smoothing concerns the number of splines used for fitting. A rule of thumb has been suggested in Ruppert (2002) who shows that the number of splines does not affect the fit once sufficient splines have been chosen, which is usually a small number compared to the sample size regardless of the form of the function to be fitted. The same conclusion is drawn in Kauermann and Opsomer (2011) who make use of the link between mixed models and penalized spline smoothing. Allowing the spline dimension to depend on the sample size provides an asymptotic framework which has been investigated in Li and Ruppert (2008), Kauermann, Krivobokova, and Fahrmeir (2009) and Claeskens, Krivobokova, and Opsomer (2009). Though these results shed some light on the theoretical properties of penalized spline estimation, there is hardly any practical impact and the rule of thumb for choosing the spline dimension (see Ruppert 2002) is still recommendable.

We also extend the classical density estimation problem by allowing the density to depend on some covariates x, say. That is to say we let the mixture weight depend on exogenous quantities. We restrict this modelling exercise to factorial quantities x, which allows us to compare densities in two (or more) groups. As example we look at the return of stocks of different companies and different years. The idea may be seen as nonparametric Analysis of Variance (ANOVA) and follows closely the testing framework for the Kolmogorov-Smirnov test.

The scientific contributions of the paper are twofold. First, we show how a density can be estimated with a penalized mixture of basis densities. The novel routine is contrasted in simulations to the various competitors described above, that is (a) kernel density estimation, (b) spline based density estimation, (c) Poisson approximated density estimation and (d) classical mixture density estimation, (e) log-spline density estimation and (f) wavelet density estimation. As will be seen, the performance of the available routines is quite diverse and the penalized mixture approach performs promising. The second contribution of the paper is to explore penalized mixture density estimation in testing scenarios when comparing distributions in two (or more) groups.

This paper is organized as follows. In Section 2 we introduce the idea of density estimation with penalized splines. Section 3 demonstrates the fitting in simulations and an example. In Section 4 we extend the idea by allowing the density to depend on covariate x, which is demonstrated in a simulation and an example in Section 5. Section 6 concludes the paper.

3.2 Penalized Density

3.2.1 Mixture Modelling and Penalized Estimation

We are interested in nonparametric estimation of the density of the univariate random variable y. We therefore approximate the density of y as a mixture of densities

$$f_K(y) = \sum_{k=-K}^{K} c_k \boldsymbol{\phi}_k(y), \tag{3.2}$$

where $\boldsymbol{\phi}_k(y)$ are subsequently called basis densities. The weights c_k in (3.2) are parameterized as

$$c_k(\boldsymbol{\beta}) = \frac{\exp(\beta_k)}{\sum_{k=-K}^{K} \exp(\beta_k)} \tag{3.3}$$

with $\beta_0 \equiv 0$ for identifiability and $\boldsymbol{\beta} = (\beta_{-K}, \ldots, \beta_{-1}, \beta_1, \ldots, \beta_K)$ so that $\int f_K(y)\mathrm{d}y = 1$. The basis densities are thereby known and fixed density functions with specified parameters. We assume that $\boldsymbol{\phi}_k(y)$ is continuous on its support and converges to zero at the boundary of the support. A possible choice for the basis densities is to take $\boldsymbol{\phi}_k(y)$ as Gaussian density with fixed mean μ_k and variance σ_k^2, where the mean values μ_k may be called the knots of the basis. Numerically more stable and theoretically more appealing are B-spline densities which are standard B-splines (see de Boor 1978) normed to be densities. We will subsequently notate the knots at which the basis densities are located as μ_k with k running from $-K$ to K for convenience. We assume, that the knots μ_k cover the range of observed values of y and their location is fixed. A typical and simple setting is to have equidistant knots which will be assumed subsequently. Apparently, the number of knots plays an important role in terms of bias and variance and a small number K will lead to biased estimates while for large values of K the estimates will be wiggled. We will therefore utilize the idea of penalized spline smoothing by choosing the number of knots K in a lavish and generous way and impose a penalty to achieve smoothness. The penalty is put on the basis coefficients β_k by penalizing the variation of c_k over k. Assuming independent

observations $y_i, i = 1, ..., n$, the log likelihood takes the form

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\log \sum_{k=-K}^{K} c_k(\boldsymbol{\beta}) \boldsymbol{\phi}_k(y_i) \right]. \tag{3.4}$$

The log likelihood is now supplemented by adding a quadratic penalty term to the likelihood which yields the penalized log likelihood

$$l_p(\boldsymbol{\beta}, \lambda) = l(\boldsymbol{\beta}) - \frac{1}{2} \lambda \boldsymbol{\beta}^T D_m \boldsymbol{\beta} \tag{3.5}$$

where the penalty matrix D_m induces smoothness and λ is the penalty parameter. With respect to the choice of D_m we follow the idea of penalized splines (see Eilers and Marx 1996) and we want the variation of weights c_k to be penalized. This holds if β_k does not differ abruptly from β_{k-1} or β_{k+1}, respectively. We therefore penalize m-th order differences. Let $\tilde{L}_m$ denote the m-th order difference matrix, where e.g. $\tilde{L}_1$ is

$$\tilde{L_1} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

Note that $\tilde{L}_m$ is $(\tilde{K} - m) \times \tilde{K}$ dimensional with $\tilde{K} = 2K + 1$. Since $\beta_0 \equiv 0$ by definition, we can omit the linear combination with β_0. Let therefore $L_m = \tilde{L}_m[, \{-K, \ldots, -1, 1, \ldots, K\}]$ denote the matrix by omitting the redundant middle column in L_m corresponding to β_0, where the notation $[, A]$ refers to extracting the columns given by the index set A. The penalty D_m now results as $L_m^T L_m$.
Finally we sketch how to maximize (3.5) with respect to β using a Newton-Raphson approach. Denote with $\mathcal{C}(\boldsymbol{\beta})$ the $(2K + 1) \times (2K)$ matrix with elements

$$\frac{\partial c_k(\boldsymbol{\beta})}{\partial \beta_j}, \quad k = -K, ..., K, \quad j = -K, \ldots, -1, 1, \ldots, K$$

which results as

$$\mathcal{C}(\boldsymbol{\beta}) = \left(\text{diag}(\tilde{\mathbf{c}}) - \tilde{\mathbf{c}} \tilde{\mathbf{c}}^T \right)[, \{-K, ..., -1, 1, ..., K\}] \ ,$$

where $\tilde{\mathbf{c}} = (c_{-K}(\boldsymbol{\beta}), \ldots, c_0(\boldsymbol{\beta}), \ldots, c_K(\boldsymbol{\beta}))^T$. The derivative of (3.5) with respect to $\boldsymbol{\beta}$ now equals

$$s_p(\boldsymbol{\beta}; \lambda) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \lambda D_m \boldsymbol{\beta} = \sum_{i=1}^{n} \frac{\mathcal{C}^T(\boldsymbol{\beta}) \tilde{\boldsymbol{\phi}}_i}{f(y_i)} - \lambda D_m \boldsymbol{\beta} \tag{3.6}$$

with $\tilde{\boldsymbol{\phi}}_i = (\phi_{-K}(y_i), \ldots, \phi_0(y_i), \ldots, \phi_K(y_i))^T$ and $f(y)$ as defined in (3.2). The negative second order derivative of (3.5) with respect to $\boldsymbol{\beta}$ may be approximated by

$$J_p(\boldsymbol{\beta}; \lambda) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}} + \lambda D_m \approx \sum_{i=1}^{n} \frac{\mathcal{C}^T(\boldsymbol{\beta}) \tilde{\boldsymbol{\phi}}_i \tilde{\boldsymbol{\phi}}_i^T \mathcal{C}(\boldsymbol{\beta})}{f(y_i)^2} + \lambda D_m. \tag{3.7}$$

Newton-scoring is done for estimating $\boldsymbol{\beta}$, using a fixed λ.

3.2.2 Selecting the Penalty Parameter

The penalty parameter λ steers the amount of smoothness of the fitted density and it needs to be selected data driven. A straight forward approach is the Akaike Information Criterion (AIC) (see Akaike 1974) selecting λ by minimizing

$$\text{AIC}(\lambda) = -l(\hat{\boldsymbol{\beta}}) + df(\lambda) \tag{3.8}$$

where

$$df(\lambda) = \text{tr}\left(J_p^{-1}(\hat{\boldsymbol{\beta}}; \lambda)\ J_p(\hat{\boldsymbol{\beta}}; \lambda = 0) \right) \tag{3.9}$$

approximate the degree of the fit. Note that $df(\lambda = 0) = 2K$ is giving the number of parameters. Alternatively one may apply Generalized Cross Validation (GCV). Apparently, selecting λ by minimizing (3.8) requires a grid search and fitting the density for a set of λ values, which is usually quite time consuming. Alternatively, in penalized spline smoothing it has been shown useful to make use of the link to mixed models (see Wand 2003, Kauermann 2005 or recent work by Reiss and Ogden 2009 and Wood 2011). To do so, we adopt a Bayesian viewpoint and comprehend the penalty as *a priori* distribution in the sense that the coefficient vector is assumed to be random with

$$\boldsymbol{\beta} \sim N(0, \lambda^{-1} D_m^-) \tag{3.10}$$

where D_m^- denotes the generalized inverse of D_m. The prior (3.10) is degenerated, which needs to be corrected as follows. We decompose $\boldsymbol{\beta}$ into the two components $\boldsymbol{\beta}^\sim$ and $\boldsymbol{\beta}^\perp$, respectively, such that $\boldsymbol{\beta}^\sim$ is a normally distributed random vector with non degenerated variance and $\boldsymbol{\beta}^\perp$ are the remaining components treated as parameters, see also Wand and Ormerod (2008). In fact based on a singular value decomposition we have

$$D_m = U^\sim \Lambda^\sim U^{\sim T}$$

with $\Lambda^\sim$ as diagonal matrix with positive eigenvalues and $U^\sim \in \mathbb{R}^{p \times h}$ with corresponding eigenvectors where $p = 2K$ is the number of elements in $\boldsymbol{\beta}$ and $h = p - m$ is the rank

of D_m with m as degree of the difference matrix $\tilde{L}_m$. Extending $U^{\sim}$ to an orthogonal basis by $U^{\perp}$ gives $\boldsymbol{\beta}^{\sim} = U^{\sim T}\boldsymbol{\beta}$ with the a priori assumption $\boldsymbol{\beta}^{\sim} \sim N(0, \lambda^{-1}\Lambda^{\sim -1})$ and with $U = (U^{\sim}, U^{\perp})$ as orthogonal basis, we get $\boldsymbol{\beta}^{\perp} = U^{\perp T}\boldsymbol{\beta}$. Conditioning on $\boldsymbol{\beta}^{\sim}$, we have y being distributed according to (3.2) and with (3.10) we get the mixed model log likelihood

$$l_m(\lambda, \boldsymbol{\beta}^{\perp}) = \log \int |\lambda\Lambda^{\sim}|^{\frac{1}{2}} \exp\{l_p(\boldsymbol{\beta}, \lambda)\} \, \mathrm{d}\boldsymbol{\beta}^{\sim}. \tag{3.11}$$

The integral can be approximated by a Laplace approximation (see also Rue, Martino, and Chopin 2009)

$$l_m(\lambda, \hat{\boldsymbol{\beta}}^{\perp}) \approx \frac{1}{2}\log|\lambda\Lambda^{\sim}| + l_p(\hat{\boldsymbol{\beta}}, \lambda) - \frac{1}{2}\log|U^{\sim T} J_p(\hat{\boldsymbol{\beta}}; \lambda) U^{\sim}|. \tag{3.12}$$

where $\hat{\boldsymbol{\beta}}$ denotes the penalized maximum likelihood estimate. We can now differentiate (3.12) with respect to λ which gives

$$\begin{aligned} \frac{\partial l_m(\lambda, \hat{\beta}^{\perp})}{\partial \lambda} &= -\frac{1}{2}\hat{\boldsymbol{\beta}}^T D_m \hat{\boldsymbol{\beta}} \\ &\quad + \frac{1}{2\lambda}\mathrm{tr}\left\{(U^{\sim T} J_p(\hat{\boldsymbol{\beta}}; \lambda = 0) U^{\sim} + \lambda\Lambda^{\sim})^{-1} U^{\sim T} J_p(\hat{\boldsymbol{\beta}}; \lambda = 0) U^{\sim}\right\} \end{aligned} \tag{3.13}$$

For practical implementation we approximate the trace component in (3.13) by $df(\lambda) - (m-1)$ with $df(\lambda)$ as in (3.9). In fact with this simplification, we can construct an estimating equation from (3.13) via

$$\hat{\lambda}^{-1} = \frac{\hat{\boldsymbol{\beta}}^T D_m \hat{\boldsymbol{\beta}}}{df(\hat{\lambda}) - (m-1)}. \tag{3.14}$$

Apparently, both sides of equation (3.14) depend on λ. An iterative solution is possible by fixing λ on the right hand side in (3.14), update λ on the left hand side and iterate this step by updating the right hand side of (3.14). This estimation scheme has been suggested in generalized linear mixed models by Schall (1991), see also Searle, Casella, and McCulloch (1992). For penalized spline smoothing Wood (2011) shows that the selection of smoothing parameter λ based in the mixed model approach behaves superior compared to AIC selected values, see also Reiss and Ogden (2009).

We can also use the marginal likelihood (3.12) to check or select the number of knots used in the basis. In fact the maximized $l_m(\lambda, \hat{\boldsymbol{\beta}}^{\perp})$ depends on K which may be denoted as $l_m(\lambda, \hat{\boldsymbol{\beta}}^{\perp}; K)$. Considering K itself as a parameter we can maximize the marginal likelihood. In simulations we well see later that the actual choice of K has little influence on the performance which exactly mirror Ruppert's (2002) findings in standard smooth regression models.

3.2.3 Properties of the Estimate

We show further theoretical properties, (i) that the estimated density has minimal Kullback-Leibler distance to the unknown true density and (ii) the asymptotic normality of the estimated coefficients $\boldsymbol{\beta}$. Moreover, we present results about bias and variance of the estimation.

Looking at theoretical properties of the estimation we focus on two questions. First, how well can the mixture density (3.2) approximate an unknown true density and secondly, what are the estimation properties of the penalized estimate. Let $f_K(y, \hat{\boldsymbol{\beta}})$ denote the mixture density (3.2) with weights $c_k(\hat{\boldsymbol{\beta}})$ defined through (3.3). Moreover, let $f_0(y)$ denote the true continuous unknown density. We define $\boldsymbol{\beta}^{(0)} = (\beta_{-K}^{(0)}, \ldots, \beta_K^{(0)})$ as the true parameter in the sense that $f_K(y, \boldsymbol{\beta}^{(0)})$ and $f_0(y)$ have minimal Kullback-Leibler distance based on the true density. So, we intent to minimize $E_{f_0(y)}\left\{\log\left(\frac{f_K(y,\hat{\boldsymbol{\beta}})}{f_0(y)}\right)\right\}$ with respect to $\hat{\boldsymbol{\beta}}$, which is equivalent to $0 = E_{f_0(y)}(\frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \log f_K(y, \hat{\boldsymbol{\beta}}))$. This means that $\boldsymbol{\beta}^{(0)}$ is implicitly defined through

$$0 = E_{f_0(y)}\left\{\frac{\mathcal{C}(\boldsymbol{\beta}^{(0)})^T \tilde{\boldsymbol{\phi}}(y)}{f_K(y, \boldsymbol{\beta}^{(0)})}\right\} \tag{3.15}$$

where $\tilde{\boldsymbol{\phi}}(y) = \big(\phi_{-K}(y), \ldots, \phi_0(y), \ldots, \phi_K(y)\big)^T$. Note that $\boldsymbol{\beta}^{(0)}$ depends on K, the number of knots, which is suppressed in our notation for simplification. Let $r(y, \boldsymbol{\beta}) = f_0(y)/f_K(y, \boldsymbol{\beta})$ be the ratio of the true and approximate density and define $H_k = H_k(\boldsymbol{\beta}) = \int \phi_k(y)\ r(y, \boldsymbol{\beta})\ \mathrm{d}y$. Note that $\sum_{k=-K}^{K} c_k(\boldsymbol{\beta}^{(0)}) H_k = 1$. Based on (3.15) and reflecting the definition of matrix $\mathcal{C}(\boldsymbol{\beta})$ we derive $H_k = 1$ for $k = -K, \ldots, K$. This allows with the well-known mean value theorem for integration to show the existence of $\xi_k \in [\mu_k, \mu_{k+1}]$ with $f_0(\xi_k) = f_K(\xi_k, \boldsymbol{\beta}^{(0)})$ for $k = -K, \ldots, K-1$. It follows with the mean value theorem for integration $\int \phi_k(y) r(y) dy\ = 1 = \int \phi_k(y) dy\ r(\xi_k)$. So, there exists ξ_k, such that $r(\xi_k) = 1$. Assuming now that the knots are placed densely in the sense $\mu_k - \mu_{k+1} = O(K^{-1}),\ k = -K, \ldots, K-1$ we obtain for $\delta_k(y) = f_0(y) - f_K(y, \boldsymbol{\beta}^{(0)})$ with simple Taylor series expansion the order $\delta_k(y) = O(K^{-1})$ for $\mu_{-K} \leq y \leq \mu_K$. We will call $\delta_k(y)$ subsequently the approximation bias. Using B-splines as basis densities allows us to obtain an even smaller asymptotic order for the approximation bias. In fact, if $f_0(y)$ is q-times differentiable and $\phi_k(y)$ is a B-spline density of degree q, we obtain for $q \geq 1$ the order $\delta(y) = O\left(K^{-q}\right)$. A proof is given later, Section 3.2.4. It is therefore practically as well as theoretically advisable to set ϕ_k as B-splines. To this end we have derived the approximation bias, so that we have answered the question how well the mixture density (3.2) can approximate the true unknown density $f_0(y)$. The next step is to investigate the properties of the penalized estimate of parameter

$\boldsymbol{\beta}^{(0)}$. In principle this boils down to standard penalized likelihood estimation so that simple and standard expansions yield (see Kauermann, Krivobokova, and Fahrmeir 2009) the necessary results. In fact we obtain

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} \approx J_p^{-1}(\boldsymbol{\beta}^{(0)}; \lambda)\ s_p(\boldsymbol{\beta}^{(0)}; \lambda)$$

which allows to formulate the asymptotic normality

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\big(\boldsymbol{\beta}^{(0)} + \text{bias}(\boldsymbol{\beta}^{(0)}, \lambda), V(\boldsymbol{\beta}^{(0)}, \lambda)\big) \tag{3.16}$$

with

$$\text{bias}(\boldsymbol{\beta}^{(0)}, \lambda) = -\lambda I_p^{-1}(\boldsymbol{\beta}^{(0)}, \lambda) D_m \boldsymbol{\beta}^{(0)} \tag{3.17}$$

$$V(\boldsymbol{\beta}^{(0)}, \lambda_0) = I_p^{-1}(\boldsymbol{\beta}^{(0)}, \lambda) I_p(\boldsymbol{\beta}^{(0)}, \lambda = 0) I_p^{-1}(\boldsymbol{\beta}^{(0)}, \lambda) \tag{3.18}$$

where $I_p(\boldsymbol{\beta}^{(0)}, \lambda) = E_{f_0(y)}\big\{J_p(\boldsymbol{\beta}^{(0)}; \lambda)\big\}$. In Section 2.3, we will use the above-mentioned well known link between penalized spline smoothing and mixed models. In the context of mixed models (3.18) is justified by Kass and Steffey (1989) and extended by Searle, Casella, and McCulloch (1992). The final step is now to transfer (3.16) to properties of the density estimate $f_K(y, \hat{\beta}) = \sum c_k(\hat{\boldsymbol{\beta}})\ \phi_k(y) = \tilde{\boldsymbol{\phi}}^T(y)\ \tilde{c}(\hat{\boldsymbol{\beta}})$. We get

$$f_0(y) - f_K(y, \hat{\boldsymbol{\beta}}) \overset{a}{\sim} N\big(\text{bias}\big(f_K(y, \hat{\boldsymbol{\beta}})\big), \text{Var}\big(f_K(y, \hat{\boldsymbol{\beta}})\big)\big)$$

with

$$\begin{aligned}
\text{bias}\big(f_K(y, \hat{\boldsymbol{\beta}})\big) &= \tilde{\boldsymbol{\phi}}^T(y)\ \mathcal{C}(\boldsymbol{\beta}^{(0)})\ \text{bias}(\boldsymbol{\beta}^{(0)}, \lambda_0) \\
\text{Var}\big(f_K(y, \hat{\boldsymbol{\beta}})\big) &= \tilde{\boldsymbol{\phi}}^T(y)\ \mathcal{C}(\boldsymbol{\beta}^{(0)})\ V(\boldsymbol{\beta}^{(0)}, \lambda_0)\ \ \mathcal{C}^T(\boldsymbol{\beta}^{(0)}) \tilde{\boldsymbol{\phi}}^T(y)
\end{aligned}$$

Since the penalized Fisher information $I_p(\boldsymbol{\beta}^{(0)}, \lambda)$ is difficult to calculate we replace it by its observed version $J_p(\boldsymbol{\beta}^{(0)}; \lambda)$ to calculate confidence intervals. Komárek, Lesaffre, and Hilton (2005) argue, that there is no guarantee that the middle matrix of (3.18), $J_p(\boldsymbol{\beta}^{(0)}; \lambda = 0)$ is positive semidefinite. In this case one may use $J_p^{-1}(\boldsymbol{\beta}^{(0)}; \lambda)$ instead of (3.18) for calculating confidence intervals. The latter can also be justified following the mixed model framework discussed subsequently, as derived in Ruppert, Wand, and Carroll (2003, page 140).

3.2.4 Asymptotic Behaviour of B-spline Densities

Let $\phi_k(y) = b_{q,k}(y)$ be a normed B-spline basis of order q defined on the support $[\mu_k, \mu_{k+q+1}]$ such that $\int b_{q,k}(y)\mathrm{d}y = 1$. Let $f_{K,q}(y, \boldsymbol{\beta}) = \sum_k c_k(\boldsymbol{\beta}) b_{q,k}(y)$ be the mixture B-spline density and let $r_q(y) = r_q(y, \boldsymbol{\beta}) = f_0(y)/f_{K,q}(y, \boldsymbol{\beta})$ be the ratio of the true and mixture density. Let $\mu_{-K}, \ldots, \mu_0, \ldots, \mu_K, \ldots, \mu_{K+q+1}$ be the knots located equidistantly with order $\mu_k - \mu_{k-1} = O(K^{-1})$. Note that our B-spline basis is q times differentiable within each interval $[\mu_k, \mu_{k+1}]$ and in particular, boundary splines are continuous. With (3.15) we get

$$\int_{\mu_k}^{\mu_{k+q+1}} b_{q,k}(y) r_q(y) \mathrm{d}y = 1 \tag{3.19}$$

so that there exists a $\xi_k \in (\mu_k, \mu_{k+q+1})$ with $r_q(\xi_k) = 1$ for $k = -K, \ldots, K$. With the recursive formula for derivatives of B-splines (see Butterfield 1976) we get for $q \geq 2$ with partial integration and making use of (3.19) for $k = -K, \ldots, K-1$

$$\begin{aligned}
\int_{\mu_k}^{\mu_{k+q+2}} b_{q+1,k}(y) r_q'(y)\, dy &= b_{q+1,k}(y) r_q(y) \Big|_{\mu_{k+q+2}}^{\mu_k} \\
&+ K \left\{ \int_{\mu_k}^{\mu_{k+q+1}} b_{q,k}(y) r_q(y)\, dy \right. \\
&- \left. \int_{\mu_{k+1}}^{\mu_{k+q+1}} b_{q,k+1}(y) r_q(y)\, dy \right\} \\
&= 0
\end{aligned}$$

This in turn shows with the mean value theorem that there exists a $\xi_k^{(1)} \in [\mu_k, \mu_{k+q+2}]$ with $r_q'(\xi_k^{(1)}) = 0$. Considering the derivative of $r_q(y)$ it is easily derived that $f_{K,q}'(\xi_k^{(1)}) = f_0'(\xi_k^{(1)}) + O(K^{-1})$. With the same arguments as above we can show that there exists $\xi_k^{(l)}$ with $1 \leq l \leq q-1$ and $k = -K, \ldots, K-l$ such that $f^{(l)}(\xi_k^{(l)}) = f_{K,q}^{(l)}(\xi_k^{(l)}) + O(K^{-1})$. This allows to conclude with iterative arguments that for $q \geq 1$ and for $l \leq q-1$

$$f_{K,q}^{(l)}(y) = f^{(l)}(y) + O(K^{-q+l})$$

so that for $l = 0$ we get the approximation error

$$f_{K,q}(y) = f(y) + O(K^{-q}).$$

3.2.5 Practical Settings, Numerical Implementation and Extensions

The fitting requires a number of practical settings which are implemented in the `R` package `pendensity` (see Schellhase 2010). First, we need to allocate the basis density given a set of observations $y_1, \ldots, y_n$. We suggest to use B-splines allocated at equidistant knots μ_k with the most left knot μ_L, fulfilling $\mu_L \leq \min(y_i)$ and the most right knot $\mu_R \geq \max(y_i)$. The performance of the estimations can be improved using additional equidistant knots beyond $[\mu_L, \mu_R]$. Therefore, the used penalization of neighbouring weights c_k in interaction with additional knots can achieve a better fit of the densities at the boundaries. In our simulations (see Section 3.3) we run estimations with one additional knot placed with the same distance used for the knots in the support at each end of $[\mu_L, \mu_R]$ and observe an improved result for several distributions.
As starting value we found that assuming a uniform distribution is useful, i.e. we set $\beta_k = 0$ to start the Newton procedure. We also experimented with different starting values but observed that the uniform start is preferable in terms of iteration steps to reach the maximum of the penalized likelihood. To avoid terminating the algorithm in a local instead of global maximum, it is advisable to fit the density for a number of different starting values and take the fit with the maximum value of the likelihood. It should be noted, however, that the problem of local maxima occurs if the penalty is not strong enough, since the penalty in (3.5) works towards the concavity of the penalized likelihood. It is therefore recommendable to start the Newton procedure with a large λ. Finally, the number of knots, i.e. the dimension of the density basis needs to be selected. Generally, we suggest to use a large K, where we have decided upon the default setting $K = 20$, which corresponds to a 41 dimensional basis. This mirrors the rule of thumb suggested in Ruppert (2002). Increasing $K \gg 20$ does not lead to an improved performance of the fit. But K should not be selected too small, due to the appearance of an approximation bias of not ignorable size (see Kauermann, Krivobokova, and Fahrmeir 2009). We show the influence of K on the fit in the next section and we confirm the impression of Ruppert (2002) in that the actual choice of K has little influence on the fit.
Conceptually, the approach is easily extended to multivariate density estimation. In this case we replace basis densities $\phi_k(\cdot)$ in (3.2) by Tensor products of univariate fixed basis densities. The index k is then running over a grid and the penalty should be formulated in each direction of the grid, that is row- and columnwise for two dimensions.

3.3 Simulations and Example

3.3.1 Simulations

Univariate Density Estimation

To demonstrate the performance of the penalized density estimate we run a number of simulations. We use (i) a normal distribution $F_0(y) \sim N(0,1)$, a mixture of normals (ii) $F_0(y) \sim \frac{1}{2}N(-\frac{1}{2},\frac{1}{4})+\frac{1}{2}N(\frac{1}{2},\frac{1}{4})$, two bimodal mixtures (iii) as $F_0(y) \sim \frac{1}{2}N(-\frac{3}{2},1)+\frac{1}{2}N(\frac{3}{2},1)$ and (iv) with $F_0(y) = \frac{3}{4}N(-\frac{3}{2},1)+\frac{1}{4}N(\frac{3}{2},1)$, mixture of five normal densities (v) as $F_0(y) \sim \frac{13}{20}N(-1,\frac{1}{2})+\frac{2}{20}N(-\frac{1}{2},\frac{1}{2})+\frac{1}{20}N(0,1)+\frac{3}{20}N(\frac{1}{2},\frac{1}{2})+\frac{1}{20}N(1,\frac{1}{2})$, a normal variance mixture as (vi) with $F_0(y) \sim \frac{1}{2}N(0,1)+\frac{1}{2}N(0,10)$, (vii) a gamma distribution $\Gamma(3,1)$ and (viii) a beta distribution Beta$(10,10)$. To compare our results labelled with $\hat{f}_K(\cdot)$ with alternative routines we use, (a) classical kernel density estimates (see Wand and Jones 1995), (b) the density estimation proposal of Gu and Wang (2003), (c) the approach of density estimation of Eilers and Marx (1996), (d) a mixture density approach, (e) the log-spline routine and (f) a wavelet approach, respectively. For the traditional kernel density estimate (a) labelled as $\hat{f}_{kernel}(\cdot)$, we utilize two approaches for selecting the bandwidth. First we use cross validation (bw=ucv) and secondly we choose the bandwidth by the approach of Sheather and Jones (1991) (bw=SJ). Both kernel routines are implemented in the `density()` routine in `R`. For (b) one estimates the unknown density $f(\cdot)$ by the logistic density transform (3.1) with a roughness penalty imposed on $\eta(y)$ which penalizes integrated squared order derivatives. This routine is implemented in `R` in the `gss` package (see Gu 2009) and we label the resulting estimated density with $\hat{f}_{spline}(\cdot)$. For the third approach (c) we divide the support of the data points in a large number of bins. Following Ruppert, Wand, and Carroll (2003) we use $B = 200$ equidistant subintervals (bins) and notate with b_j the number of observations in the j-th bin, $j = 1,\ldots,200$. With m_j as bin center and d_j as bin width we fit the Poisson model $b_j \sim \text{Poisson}(f(m_j)nd_j)$. One can now fit the density function $f(\cdot)$ using for instance the `gam()` procedure in `R`, see Wood (2006). For the fourth approach (d) we make use of the `R` package `mixtools` (see Young, Hunter, Chauveau, and Benaglia 2009) and select the number of mixture components using a Bayesian Information Criterion (BIC) and the entropy criterion suggested in Celeux and Soromenho (1996). We thereby increased K successively starting from $K = 1$ until the criterion reaches its optimum. The fifth approach, the log-spline density estimation (e) is implemented in `R` package `logspline` (see Kooperberg 2009). Finally, the wavelet density estimation (f) is done with `R` package `wavethresh` (see Nason 2010), with finest resolution level equal to one and Daubechies least asymmetric wavelets. For comparison with our penalized density estimate (g) we use $2K+1$ bins with $K = 20$

and $K = 30$, respectively and label the resulting density estimate with $\hat{f}_{bin,K}(\cdot)$. We also select K data driven to maximize the likelihood derived in Section 3.2.2.
To evaluate the performance of the fit we run $N = 500$ replicates of the simulation for different sample sizes n and different K and calculate the integrated Mean Squared error. Therefore we first calculate the Mean Squared Error

$$\mathrm{MSE}(\hat{f}(\tilde{y}_k)) = \frac{1}{N}\sum_{j=1}^{N}\left\{\hat{f}_{(j)}(\tilde{y}_k) - f_0(\tilde{y}_k)\right\}^2,$$

where the calculated estimated densities $\hat{f}_{(j)},\ j = 1,\ldots,N$ and the true densities f_0 are evaluated at fixed and equidistant values $\tilde{y}_k,\ k = 1,\ldots,1000$, say. The IMSE results as follows

$$\widehat{\mathrm{IMSE}}(\hat{f}(\tilde{y})) = \frac{1}{1000}\sum_{k=1}^{1000}\left\{\mathrm{MSE}(\hat{f}(\tilde{y}_k))\right\}.$$

Accordingly the results of the competing density estimations $\hat{f}_K(\cdot)$, $\hat{f}_{kernel}(\cdot)$, $\hat{f}_{spline}(\cdot)$, $\hat{f}_{bin,K}(\cdot)$, $\hat{f}_{mixture}(\cdot)$, $\hat{f}_{log}(\cdot)$ and $\hat{f}_{wave}(\cdot)$are shown in Table 3.2. Note that for simulation scenario i) we used for the mixture (d) the true one component normal distribution with fitted parameters which maybe considered as artificial benchmark in this case. In general it appears that the approach with a penalized mixture performs promisingly well in comparison with the six competitors, even though no method is uniformly superior. In general, however, in scenarios where the penalized mixture approach is not optimal its optimal IMSE is not more than 62% larger than the IMSE of the best density estimate, while this number is larger for all other competitors. For small n but even more for large n we observe the well established fact that the quality of the fit remains the same and K does not influence the performance of the fit. We notice an improved performance in some examples, if one adds one additional knot at each end outside of the support. In Table 3.2, the results of the penalized mixture approach are done with one additional knot at each end. Overall, the density estimation with a penalized mixture appears as reasonable competitor for density estimation.

3.3.2 Example: Daily Returns

We give a short example which will be picked up again in the next section. We look at the return of the two Germans stocks *Deutsche Bank AG* and *Allianz AG* in 2006. The corresponding density estimates of the penalized mixture approach are given in Figure 3.1 and Figure 3.2. We show the penalized mixture estimate and the difference in the density estimates to competitors (a) kernel density estimate, (b) spline based approach, (c) the binning based approach, (d) the finite mixture estimation, (e) the log-spline

approach and (f) the wavelet estimate. Apparently, the kernel density estimate, the Eilers & Marx estimate and as well as the mixture estimation show for the Deutsche Bank data some peak structure in the center and additional structure for values around -1, while the result of the spline approach is nearly similar to the penalized mixture estimation. Again for the Allianz data, the kernel density estimate and the mixture estimate show some peak structure in the center and additional structure for values around 2 and -2, while the result of the spline approach is nearly similar to the penalized mixture estimation. Clearly, in both scenarios, the true function is unknown, but in the simulations the penalized density estimate performs comparable to the spline approach so that the structure shown by the other five estimates might be spurious.

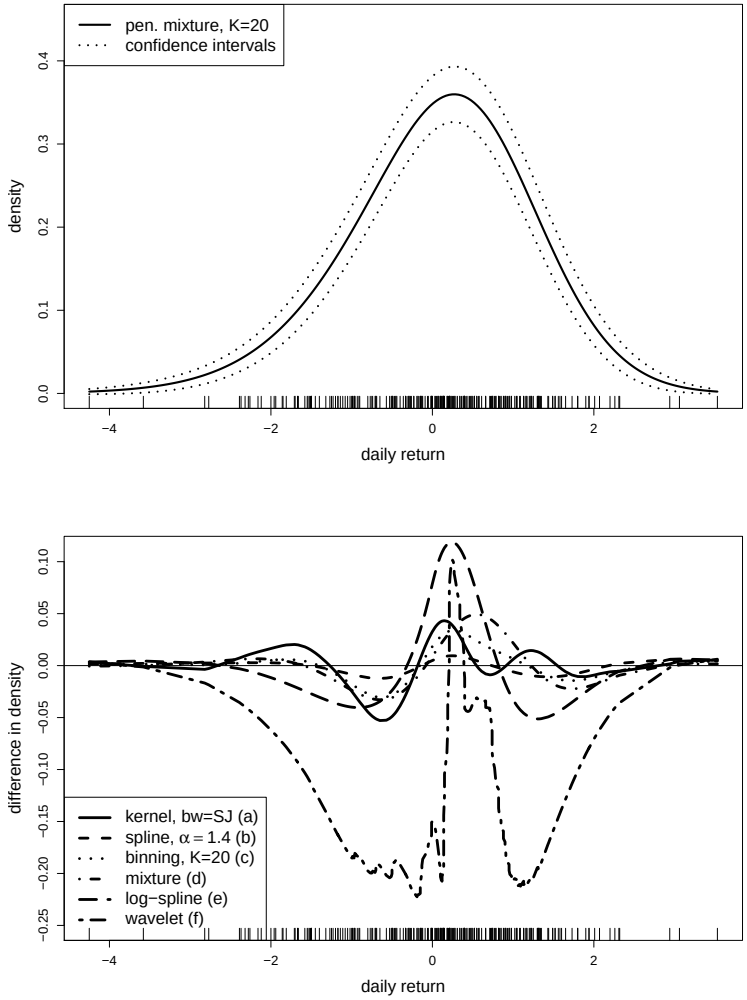


Figure 3.1: Top: Penalized mixture density $\hat{f}$ of the return of Deutsche Bank AG in 2006. Bottom: Difference in density estimates of penalized mixture to alternative density estimation routines, (a) kernel density estimation, (b) spline estimation, (c) binning estimation, (d) mixtures, (e) log-spline estimation and (f) wavelet estimation.

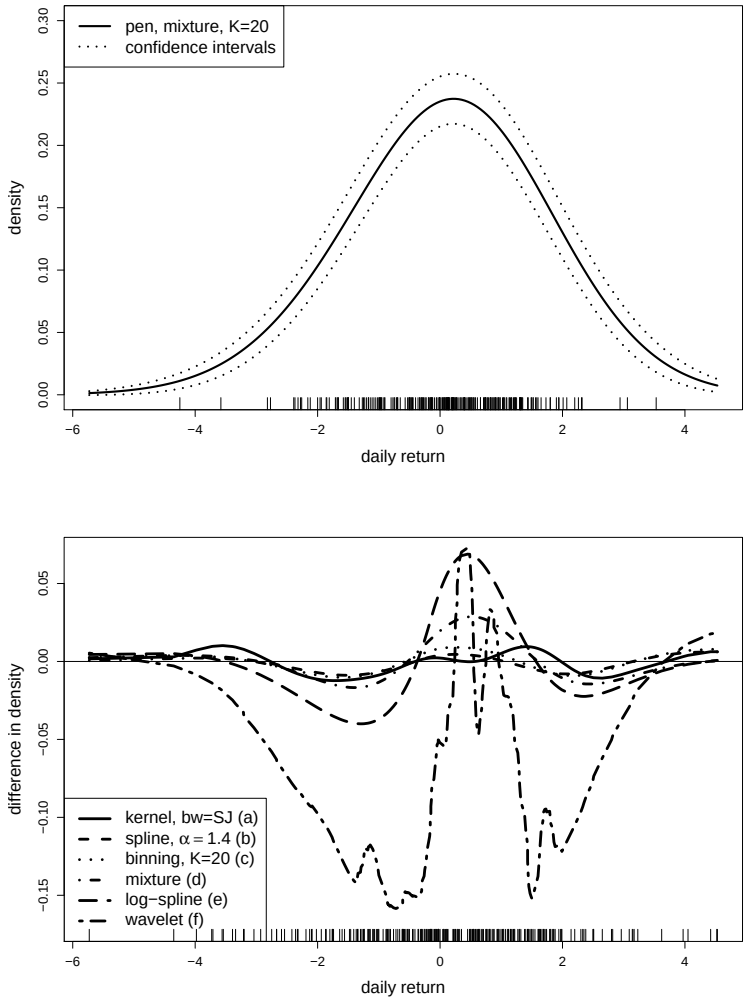


Figure 3.2: Top: Penalized mixture density $\hat{f}$ of the return of Allianz AG in 2006. Bottom: Difference in density estimates of penalized mixture to alternative density estimation routines, (a) kernel density estimation, (b) spline estimation, (c) binning estimation, (d) mixtures, (e) log-spline estimation and (f) wavelet estimation.

3.4 Nonparametric Comparison of Densities

3.4.1 Covariate Dependent Density

We can extend the above density estimation by allowing the density to depend on some covariates x, say. We intend to estimate the conditional density $f(y|x)$. Let $y_i|x_i$ denote a random sample (with x_i either random or fixed) and $x_i = (x_{i1}, \ldots, x_{is})$ is a vector of covariates. We now assume that the weights c_k depend on x which is modelled as

$$c_k(x, \boldsymbol{\beta}) = \frac{\exp(Z(x)\boldsymbol{\beta}_k)}{\sum_{j=-K}^{K} \exp(Z(x)\boldsymbol{\beta}_j)} \tag{3.20}$$

where $Z(x)$ is a design matrix, e.g. $Z(x_i) = (1, x_{i1}, \ldots, x_{is})$. Let $\boldsymbol{\beta} = (\beta_{-K}^T, \ldots, \beta_{-1}^T, \beta_1^T, \ldots, \beta_K^T)^T$ be the parameter vector and $\beta_0 \equiv 0$ for identifiability reasons. The approach can be compared to finite mixture models with mixture weights depending on covariates, see e.g. Bishop 2006, Chapter 14.5 or Müller, Quintana, and Rosner (2009). In contrast to the finite mixture, however, we again assume that K is large and will impose penalties on the weights. Let p be the dimension of $Z(x)$, i.e. the number of columns. In principle, we could have a different design for the different knots, but it is convenient and practical to assume that $Z(x)$ does not depend on k and let $\mathcal{Z}(x) = I_{2K} \otimes Z(x)$, where I_{2K} is the $2K$-dimensional unit matrix and $\otimes$ denotes the tensor product. The log likelihood then becomes

$$\boldsymbol{l}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\log\{ \sum_{k=-K}^{K} c_k(x_i, \boldsymbol{\beta}) \boldsymbol{\phi}_k(y_i) \} \right] \tag{3.21}$$

with $c_k(x, \beta)$ as in (3.20). Similar to (3.5) we add a quadratic penalty term to (3.21) so that the penalized likelihood results as follows. Looking for instance at first order differences, i.e. $m = 1$, we have $\alpha_k(x) - \alpha_{k-1}(x) = Z(x)(\beta_k - \beta_{k-1}),\ k = -K+1, \ldots, K$. Utilizing matrix notation we can write the m-th order difference as $\Delta_m \beta := (1_{\tilde{K}-m} \otimes Z(x))(\tilde{L}_m \otimes I_p)\beta$ with I_p as p dimensional identity matrix. This yields the penalty as squared m-th order difference through $\boldsymbol{\beta}^T \Delta_m^T \Delta_m \boldsymbol{\beta}$. Note that the penalty depends on the particular values of the covariates x. Taking the average over the observed values we obtain the final penalty $\boldsymbol{\beta}^T \boldsymbol{D}_m \boldsymbol{\beta}$ where

$$\boldsymbol{D}_m = (L_m^T \otimes I_p^T)(I_{\tilde{K}-m} \otimes \frac{Z^T Z}{n})(L_m \otimes I_p)$$

with $Z = (Z^T(x_1), ..., Z^T(x_n))^T \in \mathbb{R}^{n \times p}$. The penalized likelihood results now as $\boldsymbol{l}_p(\boldsymbol{\beta}, \lambda) = \boldsymbol{l}(\boldsymbol{\beta}) - \frac{1}{2}\lambda \boldsymbol{\beta}^T \boldsymbol{D}_m \boldsymbol{\beta}$. Based on (3.6) the penalized first derivative equals

$\boldsymbol{s}_p(\boldsymbol{\beta};\lambda) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta};\lambda)$ where

$$\boldsymbol{s}_i(\boldsymbol{\beta};\lambda) = \mathcal{Z}^T(x_i)\mathcal{C}^T(x_i,\boldsymbol{\beta})\frac{\tilde{\boldsymbol{\phi}}_i}{\hat{f}(y_i|x_i)} - \lambda \boldsymbol{D}_m\boldsymbol{\beta}$$

with obvious definition for $\mathcal{C}(x_i,\beta)$. Analogously to (3.7) we approximate the negative penalized second order derivative through

$$\boldsymbol{J}_p(\boldsymbol{\beta};\lambda) = -\frac{\partial^2 l_p(\boldsymbol{\beta},\lambda)}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T} \approx \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta};\lambda)\boldsymbol{s}_i{}^T(\boldsymbol{\beta};\lambda) + \lambda\boldsymbol{D}_m.$$

Estimation can now be carried out in the same way as done in the previous sections. This also applies to the estimation of the penalty parameter λ. Assuming the prior distribution (3.10) allows with the same arguments used in Section (3.2.2) to calculate the penalty parameter from the mixed model resulting as

$$\hat{\lambda}^{-1} = \frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{D}_m\hat{\boldsymbol{\beta}}}{df(\hat{\lambda}) - p(m-1)}.$$

Moreover, all other results concerning the asymptotic distribution of the estimate extend from the previous section so that we do not explicitly list them here for the sake of space.

3.4.2 Testing Densities on Equality

We can employ the idea above now to test the hypotheses on equality of densities. We formulate this by testing

$$H_0 : f(y|x_{(1)}) = f(y|x_{(0)}),\ y \in \mathbb{R} \tag{3.22}$$

for two specific values of $x_{(1)} = (x_{(1)1},\ldots,x_{(1)s})$ and $x_{(0)} = (x_{(0)1},\ldots,x_{(0)s})$. For instance, if $s = 1$ and $x_{i1} \in \{0,1\}$ indicates two groups, we may test with (3.22) whether the distribution of y_i is the same in the two groups instead of comparing densities. We look at differences in the distribution functions and define the test statistics

$$T_{max} = \max\{|T(\tau_k)|\ , k = -K,\ldots,K\}$$

with

$$T(y) = \hat{F}(y|x_1) - \hat{F}(y|x_0) = \sum_{k=-K}^{K}(c_k(x_1,\hat{\boldsymbol{\beta}}) - c_k(x_0,\hat{\boldsymbol{\beta}}))\Phi_k(y),$$

and $\tau_{-K}, \ldots, \tau_0, \ldots, \tau_K$ are denoting the knots of the basis functions and $\Phi_k(y)$ are distribution functions to basis densities $\phi_k(y)$. Under H_0 we have $E\{T(y)\} = 0$ for all y and based on the asymptotic arguments used before we can show that $\tilde{\boldsymbol{T}} = (T(\tau_{-K}), \ldots, T(\tau_0), \ldots, T(\tau_K))^T$ follows the asymptotic distribution

$$\tilde{\boldsymbol{T}} \overset{a}{\sim} N(0, \boldsymbol{W}) \tag{3.23}$$

with

$$\boldsymbol{W} = \tilde{\Phi}[\mathcal{C}_1 - \mathcal{C}_0]V(\beta^{(0)}, \lambda)[\mathcal{C}_1 - \mathcal{C}_0]^T\tilde{\Phi}^T$$

where $\mathcal{C}_j = \mathcal{C}(x_j, \hat{\boldsymbol{\beta}})\mathcal{Z}(x_j)$ for $j = 0, 1$ and $\tilde{\Phi} \in \mathbb{R}^{(2K+1)\times(2K+1)}$ as matrix with entries $\Phi_k(\tau_l)$ where (row) index k and (column) index l with $l, k = -K, \ldots, K$. Finally matrix $V(\boldsymbol{\beta}^{(0)}, \lambda)$ is the variance matrix (3.18) extended to the case of covariate dependent densities. Note that matrix $\boldsymbol{W}$ is easily calculated which allows to simulate the distribution of T_{max} in a straight forward way by sampling $\tilde{T}$ from (3.23). This can be done relatively fast after some spectral decomposition of $\boldsymbol{W}$ so that any approximate calculation of the distribution of T_{max} is numerically easy.

3.5 Simulation and Example

3.5.1 Simulation

We run a small simulation to check the performance of the fit, particularly of the testing idea based on T_{max}. To do so we simulate $n = 100$ and $n = 400$ data points from the following distributions. We assume a univariate covariate (group indicator) with $x_i = 0$ for $n/2$ and $x_i = 1$ for the remaining $n/2$ observations. We simulate y given x from the following scenarios. First, (i) we draw y from a standard normal for both $x = 0$ and $x = 1$, i.e. $y|x \sim N(0, 1)$, (ii) we draw $y|x = 0 \sim N(0, 1)$ and $y|x = 1 \sim N(\frac{1}{5}, 1)$ that is we shift the mean by $\frac{1}{5}$ for $x = 1$, and finally (iii) $y|x = 0 \sim N(0, 1)$ and $y|x = 1 \sim \frac{1}{2}N(-\frac{1}{2}, \frac{1}{4}) + \frac{1}{2}N(\frac{1}{2}, \frac{1}{4})$. For all three scenarios we calculate for each simulation the p-value resulting for T_{max}. We repeat the simulation 1000 times and give in Table 3.1 the number of p-values smaller than a nominal level α. Bear in mind that for scenario (i) the null hypothesis is true so that the p-value should be uniformly distributed on $[0, 1]$. As reference we also calculate both, the p-value resulting for a Kolmogorov-Smirnov test based on comparing the sample for $x = 0$ against $x = 1$ as well as the p-value resulting from the linear model $y = \beta_0 + x\beta_x$ and a t-test on $H_0 : \beta_x = 0$. As can be seen from the simulated numbers the test on the equalities of densities works convincingly well which supports the idea of density estimation with a penalized mixture.

level		simulation	Test on T_{max}	Kolmogorov-Smirnov Test	Test on $\beta_x = 0$ in Linear Model
$\alpha = 0.01$	(i)	$n = 100$	0.010	0.011	0.009
		$n = 400$	0.009	0.011	0.007
	(ii)	$n = 100$	**0.063**	0.042	0.057
		$n = 400$	**0.288**	0.182	**0.288**
	(iii)	$n = 100$	**0.031**	0.005	0.003
		$n = 400$	**0.377**	0.080	0.003
$\alpha = 0.05$	(i)	$n = 100$	0.058	0.041	0.058
		$n = 400$	0.052	0.049	0.053
	(ii)	$n = 100$	**0.163**	0.116	0.155
		$n = 400$	0.504	0.397	**0.526**
	(iii)	$n = 100$	**0.134**	0.051	0.030
		$n = 400$	**0.735**	0.313	0.036

Table 3.1: Proportion of p-values smaller than α, based on 1000 simulations. Optimal performance is set in bold.

3.5.2 Example

As example we look again at the daily returns for the two stocks considered in Section 3.3.2. We look at data from 2006 and 2007, and our focus of interest is to test the hypothesis that the distribution of the returns is the same in the two years. The corresponding plot is shown in Figure 3.3. Applying the test based on T_{max} to this example yields the p-values of 0.048 for *Deutsche Bank AG* and 0.019 for *Allianz AG*. Hence, there is indication that the returns in the two years differ in distribution.

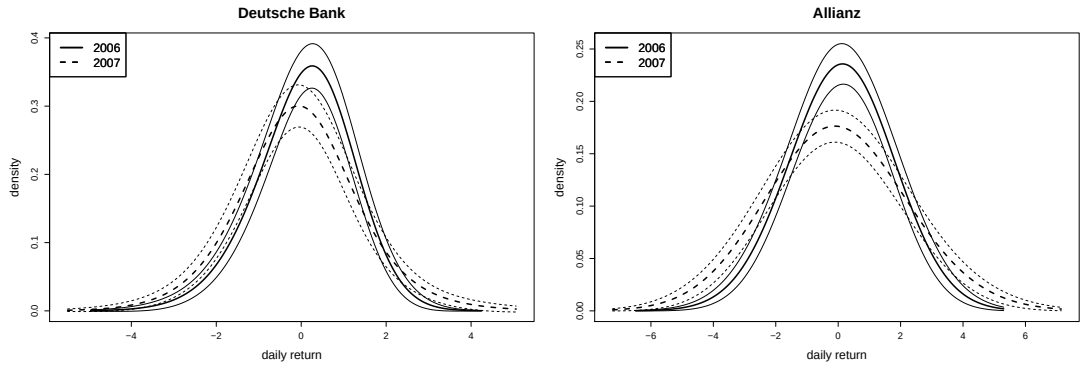


Figure 3.3: Density of the return of Deutsche Bank AG and Allianz AG in 2006 and 2007.

3.6 Conclusion

In this paper we tackled the classical problem of density estimation. Our approach picked up the idea of Komárek and Lesaffre (2008) and extended this to regular as well as covariate dependent density estimation. We examined density estimation scheme based on penalized B-spline bases using the direct link from penalized smoothing splines to mixed models. Simulations showed promising results when comparing our density estimation to competitors. First, in simple density estimation it appears that the penalized mixture approach proposed here behaves better or at least similarly compared to the common alternatives (a) kernel density estimation, (b) spline based density estimation (c) binning based estimation, (d) mixture densities, (e) log-spline density estimation and (f) wavelet density estimation. Moreover, our density estimation approach performed almost as the best, regarding the IMSE, while the classical approach (c) binning did not operate optimally in any considered density case. Secondly, extending the procedure towards covariate dependent density estimation allows for testing on the equality of densities in different groups. The approach showed superior behaviour in our simulations when compared to the classical Kolmogorov-Smirnov test. This test on equality of densities in different groups carries some omnibus power, which is seen especially in cases, where the standard tests do not announce inequality of the groups (see Table 3.1).

The approach is in principle easy to extend to multivariate density estimation. In the multivariate case, though, the numerical requirements of the penalized mixture approach do however exponentially increase due to the increasing number of B-spline basis functions. Because of this curse of dimensionality multivariate density estimation remains a difficult task.

density approach		(g) penalized mixture			(a) kernel		(b) spline	(c) binning		(d) mixture		(e) log-spline	(f) wavelet	
	rel. IMSE	$\hat{f}_K(y)$ $K=20$	$\hat{f}_K(y)$ $K=30$	$\hat{f}_K(y)$ K_{opt}	$\hat{f}_{kernel}(y)$ bw=ucv	$\hat{f}_{kernel}(y)$ bw=SJ	$\hat{f}_{spline}(y)$ Gu	$\hat{f}_{bin,K}(y)$ EM $K=20$	$\hat{f}_{bin,K}(y)$ EM $K=30$	$\hat{f}_{mixture}(y)$ BIC	$\hat{f}_{mixture}(y)$ entropy	$\hat{f}_{log}(y)$ Koo	$\hat{f}_{wave}(y)$ Nason	best absolute IMSE
(i)	$n=100$	**1.000**	1.040	1.149	2.485	2.056	1.472	2.124	2.247	3.374	4.515	4.677	6.449	0.396
	$n=400$	**1.000**	1.042	1.069	3.986	3.208	1.694	2.444	2.514	3.583	3.264	6.181	6.722	0.072
(ii)	$n=100$	1.186	1.002	**1.000**	1.637	1.292	1.128	1.599	1.615	2.527	2.445	4.027	4.561	1.074
	$n=400$	1.429	1.397	1.492	1.532	1.169	1.032	1.238	1.241	1.399	1.354	2.799	**1.000**	0.378
(iii)	$n=100$	**1.000**	1.176	1.136	1.434	1.156	1.358	1.601	1.624	1.526	3.220	2.358	4.965	0.346
	$n=400$	1.097	1.009	1.035	1.478	1.212	**1.000**	1.124	1.124	**1.000**	1.327	2.159	3.381	0.113
(iv)	$n=100$	1.052	1.085	1.076	1.291	**1.000**	1.061	1.231	1.247	1.238	3.007	2.110	3.939	0.446
	$n=400$	1.061	1.111	1.091	1.889	1.495	1.131	1.323	1.333	**1.000**	1.808	2.475	3.818	0.099
(v)	$n=100$	**1.000**	1.088	1.138	1.433	1.148	1.090	1.227	1.253	1.387	2.440	2.353	4.339	0.980
	$n=400$	1.025	1.062	1.062	1.864	1.574	1.124	1.326	1.322	**1.000**	1.483	2.657	2.541	0.242
(vi)	$n=100$	1.145	1.079	1.150	1.053	1.020	**1.000**	1.007	1.015	1.170	1.192	1.167	1.987	0.454
	$n=400$	**1.000**	1.017	**1.000**	1.030	1.006	**1.000**	1.003	1.003	1.058	1.030	1.141	1.288	0.361
(vii)	$n=100$	**1.000**	1.371	1.142	1.364	1.056	1.023	1.381	1.427	2.238	2.907	2.358	3.871	0.302
	$n=400$	1.664	1.626	1.785	1.224	**1.000**	1.159	1.196	1.196	2.841	2.841	1.748	2.542	0.107
(viii)	$n=100$	1.097	1.039	1.142	1.685	1.360	**1.000**	1.354	1.446	2.632	2.525	3.444	10.612	44.197
	$n=400$	1.199	1.123	**1.000**	2.676	2.073	1.344	2.197	2.243	2.575	2.255	4.631	51.972	9.012

Table 3.2: Relative Integrated Mean Squared Error. Optimal performance is set equal to one and in bold. The best absolute IMSE is times 10^3.

4 Flexible Copula Density Estimation with Penalized Hierarchical B-Splines

This essay is joint work with Göran Kauermann (LMU Munich) and David Ruppert (Cornell University). It is submitted to *Scandinavian Journal of Statistics*, compare Kauermann, Schellhase, and Ruppert (2012).

Chapter 4 investigates an approach to estimate multivariate copula density functions using penalized smoothing splines. In the chapter a new method for flexible spline fitting for copula density estimation is introduced, that is spline coefficients are penalized to achieve a smooth fit. To weaken the curse of dimensionality, instead of a full tensor spline basis, a reduced tensor product based on sparse grids (see Zenger 1991) is used. To achieve uniform margins of the copula density, linear constraints are placed on the spline coefficients and quadratic programming is used to fit the model. Simulations and practical examples accompany the presentation.

4.1 Introduction

Copulas allow for stochastic modelling of multivariate distributions beyond the classical normal distribution. The idea traces back to Sklar (1959), though Hoeffding (1940) might be consulted as earlier reference, see Nelsen (2006). Copulas have experienced general interest in the last years, primarily in the area of finance, see for instance McNeil, Frey, and Embrechts (2005), though the idea has been applied in other contexts as well, see for example Bogaerts and Lesaffre (2008) or Song, Mingyao, and Yuan (2009) for bio-statistical applications or Danaher and Smith (2011) for the use of copulas in marketing. A general overview and survey of recent contributions in copula modelling is found e.g. in Härdle and Okhrin (2009) or, from a more personal viewpoint, in Embrechts (2009); see also Kolev, Anjos, and Mendes (2006). A comprehensive collection of new approaches in copula estimation is provided in Jaworski, Durante, Härdle, and

Rychlik (2010). This includes, inter alia, hierarchical modelling of Archimedean copulas as suggested in Okhrin, Okhrin, and Schmid (2009) and Savu and Trede (2010). Lambert (2007) uses Bayesian spline smoothing for estimating the generator function of a Archimedean copula. Joe (1996) pursues the use of so called pair-copulas, where multiple interaction is reduced to bivariate copula modelling, see also Bedford and Cooke (2002) or Czado (2010).

While the above literature on copula estimation is vast and extensive, this does not apply to non- and semi-parametric routines for copula estimation which is tackled in this chapter. This is surprising at a first glance but can in our opinion be explained with the following two reasons. First, a copula has the property that its univariate margins are uniform. Such side constraints are however difficult to accommodate in available non-parametric estimation routines. Secondly, copulas have the potential to work in high dimensional problems, while classical non-parametric techniques suffer from the so called curse of dimensionality if the dimension exceeds two (or three). Our approach presented in this part solves the first problem by directly including constraints on the margins in the optimization routine. It turns out that the requirement of uniform margins can be easily formulated as linear constraints on spline coefficients. Moreover, we tackle the second problem, the curse of dimensionality, by making use of so-called sparse grids. This means instead of a full tensor product of splines as basis, a reduced form is used to achieve numerical feasibility in dimensions beyond two (or three).

Considering the literature on non-parametric copula estimation we refer to kernel density methods proposed in Gijbels and Mielniczuk (1990) which are further discussed in Fermanian and Scaillet (2003), Fermanian, Radulovic, and Wegkamp (2004) and Chen and Huang (2007). In these papers, the copula itself is fitted using a smoothed version of the empirical copula. Omelka, Gijbels, and Veraverbeke (2009) modify the estimate by correcting the "corner" bias of the kernel density estimates. More recently the use of wavelet based estimation has been suggested by Morettin, Toloi, Chiann, and Miranda (2010) for copula estimation or Genest, Masiello, and Tribouley (2009) for copula density estimation, see also for a more theoretical investigation Autin, Pennec, and Tribouley (2010). As an alternative to wavelets, the use of Bernstein polynomials has been proposed in Sancetta and Satchell (2004); see also Qu, Qian, and Xie (2009) and Pfeifer, Straßburger, and Philipps (2009). Instead of Bernstein polynomials one may also use linear B-splines, as pursued in this chapter, see also Shen, Zhu, and Song (2008). Replacing the copula density itself by a piecewise constant function has been pursued by Qu, Qian, and Xie (2009) or in Qu and Yin (2012). The use of Wavelets, piecewise constraints, Bernstein polynomials and B-splines allows to accommodate the constraint that univariate marginal distributions are uniform. In practice, however,

none of these methods do directly extend to higher dimensions due to the above-mentioned curse of dimensionality. That is to say, numerically it is hardly feasible to apply the routines to more than two (or three) dimensions, so that the major focus in all cited papers lies on the bivariate case. In our approach, we make use of B-splines to model the copula density itself. To do so, we replace the copula density by a (linear) combination of tensor products of univariate B-splines on $[0, 1]$. This idea builds upon Marx and Eilers (2005); see also Koo (1996). With simple linear constraints on the spline coefficients we can guarantee that the univariate margins of the copula are uniform, that is the spline estimate itself is a copula density. To achieve smoothness of the fitted copula, we impose a penalty on the spline coefficients as suggested by Eilers and Marx (1996), see also Ruppert, Wand and Carroll (2003, 2009).

With the spline approach suggested, we are, however, still faced with the problem of the curse of dimensionality. This implies that the dimensionality of the spline basis increases exponentially with the dimension of the variables and, in fact, can reach the order of a million even for 4 or 5 dimensional random vectors. To adapt the spline approach to higher dimensions, we make use of so called "sparse grids". The idea traces back to Zenger (1991) and is extensively discussed and motivated in Bungartz and Griebel (2004); see also Garcke (2006). Sparse grids make use of hierarchical B-splines as discussed, for instance, in Forsey and Bartels (1988). The idea is to represent a B-splines basis by B-splines of lower dimension, that is, built upon fewer knots. Figure 4.1 shows how a linear B-spline [plot (a)] can be represented by B-splines constructed at fewer knots [plots (b) to (d)]. More details are provided in the following parts. The idea of sparse grids is now to replace the full tensor product by a reduced form including only products of hierarchical splines up to a limited hierarchy order. This reduces the numerical effort tremendously and allows us to weaken the curse of dimensionality. Practically it means we are able to fit 4 (or even 5) dimensional copulas with a fully semi-parametric approach.

The novel contributions of the chapter are (a) copula density estimation which guarantees uniform margins and allows for fast numerical fitting by imposing simple linear constraints on the parameters and (b) proposing the use of sparse grids in the field of nonparametric copula estimation which allows to weaken the curse of dimensionality to fit models in 3, 4 or 5 dimensions. The following sections are organized as follows. In Section 2 we introduce the estimation routine with hierarchical B-splines and sparse grids. At the end of Section 2, we discuss the numerical implementation including the incorporation of constraints on the marginal densities. In Section 3, we investigate the performance of our copula estimator using simulations and two examples.

4.2 Penalized B-Spline Estimation of a Copula Density

4.2.1 B-Spline Density Basis

Following Sklar's (1959) theorem, we can write the distribution of the p dimensional random vector $X = (X_1, \ldots, X_p)$ as

$$F(x_1, \ldots, x_p) = C\{F_1(x_1), \ldots, F_p(x_p)\}, \tag{4.1}$$

where $C(., \ldots, .)$ is the copula corresponding to $F(\cdot)$. We assume that copula $C(., \ldots, .)$ is a distribution function on the p-dimensional cube $[0,1]^p$, with uniform marginal distributions and copula density $c(., \ldots, .)$ which is related to the density $f(x_1, \ldots, x_p)$ through

$$f(x_1, \ldots, x_p) = c\{F_1(x_1), \ldots, F_p(x_p)\} \prod_{j=1}^{p} f_j(x_j). \tag{4.2}$$

Our intention is to estimate the copula density $c(.)$ itself, either assuming the marginal distribution $F_j(x_j)$ to be known or being estimated separately. Let therefore $u_j = F(x_j)$ so that $c(u_1, \ldots, u_p)$ is a density on $[0,1]^p$ with the p margin-constraints

$$\int_{\substack{p \\ \times \\ i \neq j}} c(u_1, \ldots, u_p) \prod_{i \neq j}^{p} \mathrm{d}u_i = 1, \text{ for } j = 1, \ldots, p. \tag{4.3}$$

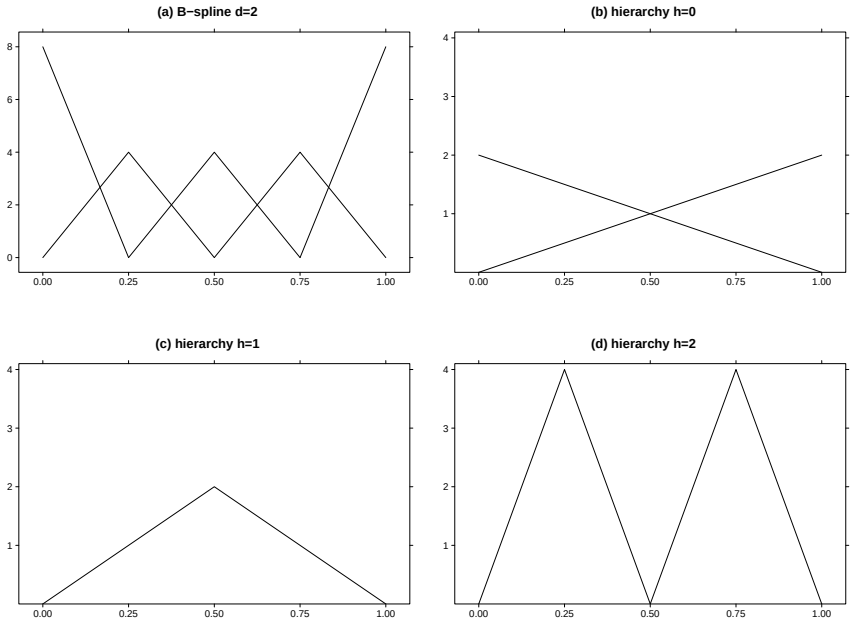


Figure 4.1: (a) B-spline density basis and corresponding hierarchical B-spline density basis ((b),(c),(d)) with different hierarchy levels.

We estimate $c(\cdot)$ in a flexible semi-parametric way by taking the p constraints (4.3) into account. To do this, we will approximate $c(\cdot)$ by a mixture of basis densities. Let therefore $\phi_k(u)$ be a regular linear univariate B-spline normalized to be a density, i.e., $\int \phi_k(u)\, \mathrm{d}u = 1$ with $u \in [0,1]$ and denote with $\Phi(\cdot) = \{\phi_l(\cdot),\ l = 1, \ldots, K\}$ the univariate B-spline density basis of dimension K, see Figure 4.1 (a). We construct the full tensor product as $\mathbf{\Phi}(u_1, \ldots, u_p) = \bigotimes_{j=1}^{p} \Phi(u_j)$ and reexpress $\mathbf{\Phi}(\cdot)$ by letting $\mathbf{k} = (k_1, \ldots, k_p)$ be a p-tuple with $\mathbf{k} \in \mathcal{K} = \{1, \ldots, K\}^p$. The components of $\mathbf{\Phi}(\cdot)$ are then

$$\phi_{\mathbf{k}}(u_1, \ldots, u_p) = \phi_{k_1,\ldots,k_p}(u_1, \ldots, u_p) = \prod_{j=1}^{p} \phi_{k_j}(u_j),$$

where $k_j \in \{1, \ldots, K\}$ for $j = 1, \ldots, p$. The idea is now to approximate the copula density through the B-splines such that

$$c(u_1, \ldots, u_p) \approx \sum_{\mathbf{k} \in \mathcal{K}} b_{\mathbf{k}} \phi_{\mathbf{k}}(u_1, \ldots, u_p) =: c(u_1, \ldots, u_p; \mathbf{b}). \tag{4.4}$$

The goodness of the approximation depends thereby on the richness of the basis, that is, on the number of elements in $\mathcal{K}$. We discuss this point later. The elements of $\mathbf{b} = (b_{\mathbf{k}}, \mathbf{k} \in \mathcal{K})$ are subsequently called the spline basis coefficients and with each single basis spline being a density itself we obtain with conditions

$$\sum_{\mathbf{k} \in \mathcal{K}} b_{\mathbf{k}} = 1, \quad c(\mathbf{u}; \mathbf{b}) \geq 0 \tag{4.5}$$

that $c(\mathbf{u}; \mathbf{b})$ in (4.4) is a density. For simplicity we ignore at this point that $c(\cdot; \mathbf{b})$ is not guaranteed to be copula density in that univariate margins are not guaranteed to be uniform. We will come back to this condition later.

To construct the likelihood for spline coefficients $\mathbf{b}$, assume we have a random sample $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ with $i = 1, \ldots, n$ from which we construct $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip})$ through $u_{ij} = \hat{F}_j(x_{ij})$. Here, $\hat{F}_j(.)$ is a $\sqrt{n}$ consistent estimate of the marginal distribution function, which in the simplest case is just the empirical distribution function and hence nu_{ij} are the ranks. Based on $\mathbf{u}_i, i = 1, \ldots, n$, the log likelihood for $\mathbf{b}$ is then

$$l(\mathbf{b}) = \sum_{i=1}^{n} \log \left\{ \sum_{\mathbf{k} \in \mathcal{K}} b_{\mathbf{k}} \phi_{\mathbf{k}}(u_{i1}, \ldots, u_{ip}) \right\}, \tag{4.6}$$

which needs to be maximized subject to the constraints (4.5). The accuracy of the spline approximation in (4.4) improves for large K, but the corresponding fit will suffer from estimation variability due to over-parameterization of the data. Entertaining the ideas of penalized splines (see also Ruppert, Wand, and Carroll 2003), we impose a

penalty on spline coefficients $b_{\mathbf{k}}$ to achieve a smooth fit for $c(\cdot)$. Eilers and Marx (1996) suggest to penalize r-th order differences for the B-spline coefficients. This easily extends to the multivariate setting as shown in Marx and Eilers (2005). Let $L \in \mathbb{R}^{(K-r)\times K}$ be a difference matrix of order r, e.g. for $r = 1$ we get

$$L = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

and let $W = \text{diag}(w_1, \ldots, w_K)$ be the weight matrix linking a regular B-spline basis to a B-spline density basis, i.e. w_l is the integral from 0 to 1 of the l-th regular B-spline. With matrix L we can now penalize differences in neighbouring spline coefficients and define the penalty matrix $P = WL^TLW$; see also Wand and Ormerod (2008) and Ruppert, Wand, and Carroll (2003). This penalty applies only to a single dimension. To achieve smoothness of the fitted copula density for all variables, we use the Kronecker product yielding the entire penalty matrix

$$\mathbf{P}(\boldsymbol{\lambda}) = \sum_{j=1}^{p} \lambda_j \mathbf{P}_j.$$

with $\mathbf{P}_j = \left(\bigotimes_{l=1}^{j-1} I_K\right) \otimes P \otimes \left(\bigotimes_{l=j+1}^{p} I_K\right)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ where I_K is the K dimensional identity matrix and $\bigotimes_{l=1}^{j-1}$ denotes component-by-component tensor products (where $\bigotimes_{l=1}^{0} I_K = 1 = \bigotimes_{l=p+1}^{p} I_K$). The coefficient λ_j is the penalty parameter for the j-th variable which needs to be selected in a data driven manner, as discussed later. Incorporating the penalty into the log likelihood gives the penalized log likelihood

$$l_p(\mathbf{b}, \boldsymbol{\lambda}) = l(\mathbf{b}) - \frac{1}{2}\mathbf{b}^T\mathbf{P}(\boldsymbol{\lambda})\mathbf{b}, \tag{4.7}$$

which is maximized for given $\boldsymbol{\lambda}$ with respect to $\mathbf{b}$. Note that $\boldsymbol{\lambda}$ determines the amount of smoothness for the fitted coefficients and setting $\boldsymbol{\lambda} = 0$ gives the unpenalized ML estimate.

4.2.2 Hierarchical B-splines and Sparse Grids

The modelling approach proposed above becomes numerically infeasible if the dimension p exceeds 2 or 3, since the dimension of the tensor product basis grows exponentially in p. To illustrate this curse of dimensionality, Table 4.1 gives the dimension of a full tensor product based on a linear B-spline basis of dimension $K = 2^d + 1$ for

$d = D$	basis	$p = 2$	$p = 3$	$p = 4$	$p = 5$
3 ($K = 9$)	tensor prod. ($D = dp$)	81	729	6,561	59,049
	sparse ($D = d$)	37	123	368	1,032
4 ($K = 17$)	tensor prod. ($D = dp$)	289	4,913	83,521	1,419,857
	sparse ($D = d$)	81	297	961	2,882
5 ($K = 33$)	tensor prod. ($D = dp$)	1,089	35,937	1,185,921	39,135,393
	sparse ($D = d$)	177	705	2,441	7,763

Table 4.1: Dimension of tensor product basis $\tilde{\mathbf{\Phi}}_{(d)}(u_1, \ldots, u_p)$ (full tensor product) and reduced sparse hierarchical basis $\tilde{\mathbf{\Phi}}^{(D)}_{(d)}(u_1, \ldots, u_p)$ with D set equal to d for $q = 1$, i.e., linear B-splines.

different dimensions of $\mathbf{u}$ ranging from $p = 2$ to $p = 5$. Even for a $p = 3$ dimensional vector u and $K = 17$, one ends up with nearly 5000 parameters, which is at the limit of numerical feasibility. We therefore suggest reducing the spline dimension for numerical purposes by not taking a full tensor product but, instead, using a reduced form to guarantee numerical feasibility. Our approach makes use of Zenger's (1991) so called 'sparse grids'. To apply the idea we first transform the univariate B-spline density into its hierarchical form. Let the linear univariate B-spline density basis be built upon $2^d + 1$ equidistant knots $\tau_k = k2^{-d}, k = 0, \ldots, 2^d$. The basis has dimension $K = 2^d + \mathbf{1}$ and is denoted subsequently as $\Phi_{(d)}(u) = \left\{\phi_{(d)l}(u), l = 1, \ldots, K\right\}$. We can reexpress this basis in hierarchical terms as derived in Forsey and Bartels (1988, 1995); see also Garcke (2006). Let $\mathcal{I}_0 = \{1, \mathbf{2}\}$ and $\mathcal{I}_h = \{2j, \text{for } 1 \leq j \leq 2^{h-1}\}$ for $h = 1, \ldots, d$ denote hierarchical index sets. The hierarchical B-spline basis linearly equivalent to $\Phi_{(d)}(u)$ is then defined through

$$\tilde{\Phi}_{(d)}(u) = \left\{\phi_{(h)l}(u), l \in \mathcal{I}_h, h = 0, \ldots, d\right\} = \left\{\Phi_{(h)\mathcal{I}_h}, h = 0, \ldots, d\right\}. \tag{4.8}$$

Figure 4.1 illustrates the hierarchical spline in plots (b) to (d) with B-spline basis $\phi_{(0)1}(.), \phi_{(0)2}(.)$ for (b), $\phi_{(1)2}(.)$ for (c) and $\phi_{(2)2}(.), \phi_{(2)4}(.)$ for (d). It is not difficult to show that both bases, (a) and (b) to (d), span the same space so that $\Phi_{(d)}(u) = \tilde{\Phi}_{(d)}(u)\tilde{A}$ for some invertible $K \times K$ matrix $\tilde{A}$. We now reformulate the penalized likelihood (4.7) by replacing the B-spline bases in (4.4) with their hierarchical form. To do this, let the complete tensor product based on the hierarchical B-spline basis $\tilde{\Phi}_{(d)}(\cdot)$ be denoted with

$$\tilde{\mathbf{\Phi}}_{(d)}(u_1, \ldots, u_p) = \bigotimes_{j=1}^{p} \tilde{\Phi}_{(d)}(u_j) = \mathbf{\Phi}(u)\tilde{\mathbf{A}}^{-1}$$

and $\tilde{\mathbf{A}}^{-1} = \bigotimes_{j=1}^{p} \tilde{A}^{-1}$. Let $\tilde{\mathbf{b}} = \tilde{\mathbf{A}}^{-1}\mathbf{b}$ denote the corresponding spline coefficient vector for basis $\tilde{\mathbf{\Phi}}_{(d)}(\cdot)$. The penalized likelihood (4.7) can then be rewritten in terms of $\tilde{\mathbf{b}}$

taking the form

$$\tilde{l}_p(\tilde{\mathbf{b}}, \boldsymbol{\lambda}) = \tilde{l}(\tilde{\mathbf{b}}) - \frac{1}{2}\tilde{\mathbf{b}}^T\tilde{\mathbf{P}}(\boldsymbol{\lambda})\tilde{\mathbf{b}}$$

with $\tilde{l}(\tilde{\mathbf{b}}) = \sum_{i=1}^{n} \log\left\{\tilde{\boldsymbol{\Phi}}_{(d)}(\mathbf{u}_i)\tilde{\mathbf{b}}\right\}$ and $\tilde{\mathbf{P}}(\boldsymbol{\lambda}) = \sum_{j=1}^{p} \lambda_j \tilde{\mathbf{P}}_j$ where

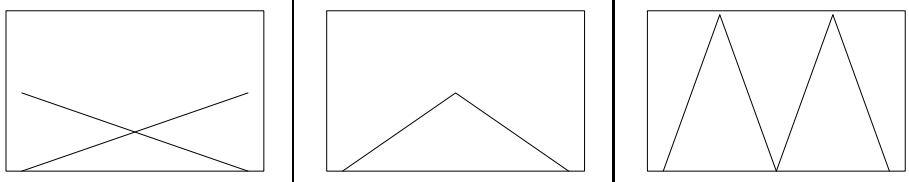
$$\tilde{\mathbf{P}}_j = \left(\bigotimes_{l=1}^{j-1} \tilde{I}_{(d)}\right) \otimes \{(\tilde{A}^{-1})^T P \tilde{A}^{-1}\} \otimes \left(\bigotimes_{l=j+1}^{p} \tilde{I}_{(d)}\right)$$

and $\tilde{I}_{(d)} = (W\tilde{A}^{-1})^T(W\tilde{A}^{-1})$.

The parameterization with hierarchical B-splines allows us to tackle the curse of dimensionality by making use of a so-called sparse grid approach. The underlying idea is to consider spline tensor products up to a cumulated hierarchy order D only. Figure 4.2 illustrates the idea for dimension $p = 2$ and $D = 2$ using a linear B-spline basis. To be specific, we define the sparse grid tensor product as

$$\tilde{\boldsymbol{\Phi}}_{(d)}^{(D)}(u_1, \ldots, u_p) = \left(\bigotimes_{j=1}^{p} \Phi_{(h_j)\mathcal{I}_{h_j}}(u_j), \sum_{j=1}^{p} h_j \leq D\right). \tag{4.9}$$

The upper index D refers to the maximum hierarchy level and the lower index d is the hierarchy level of the marginal hierarchical B-spline basis. Note that $d \leq D \leq pd$ is a useful range for D and $\tilde{\boldsymbol{\Phi}}_{(d)}^{(pd)}(\cdot) = \tilde{\boldsymbol{\Phi}}_{(d)}(\cdot)$. The reduction of the basis reduces the numerical effort tremendously as can be seen from Table 4.1 where we show the dimension of $\tilde{\boldsymbol{\Phi}}_{(d)}$ and $\tilde{\boldsymbol{\Phi}}_{(d)}^{(D)}$ for various values of d and D. For $p = 3$ and $d = D = 4$ (i.e. $K = 2^d + 1$) we get a 297 dimensional basis instead of 4913 dimensional. Note that the reduced basis is created by extracting columns of the complete tensor product basis. This means we can write



	$\Phi_{(0)\mathcal{I}_0}(u_1)$	$\Phi_{(1)\mathcal{I}_1}(u_1)$	$\Phi_{(2)\mathcal{I}_2}(u_1)$
$\Phi_{(0)\mathcal{I}_0}(u_2)$	$\Phi_{(0)\mathcal{I}_0}(u_1) \otimes \Phi_{(0)\mathcal{I}_0}(u_2)$	$\Phi_{(1)\mathcal{I}_1}(u_1) \otimes \Phi_{(0)\mathcal{I}_0}(u_2)$	$\Phi_{(2)\mathcal{I}_2}(u_1) \otimes \Phi_{(0)\mathcal{I}_0}(u_2)$
$\Phi_{(1)\mathcal{I}_1}(u_2)$	$\Phi_{(0)\mathcal{I}_0}(u_1) \otimes \Phi_{(1)\mathcal{I}_1}(u_2)$	$\Phi_{(1)\mathcal{I}_1}(u_1) \otimes \Phi_{(1)\mathcal{I}_1}(u_2)$	**omitted**
$\Phi_{(2)\mathcal{I}_2}(u_2)$	$\Phi_{(0)\mathcal{I}_0}(u_1) \otimes \Phi_{(2)\mathcal{I}_2}(u_2)$	**omitted**	**omitted**

Figure 4.2: Representation of $\tilde{\boldsymbol{\Phi}}_{(2)}^{(2)}(u_1, u_2)$ for two dimensions ($p = 2$).

$$\tilde{\boldsymbol{\Phi}}_{(d)}^{(D)}(u_1, \ldots, u_p) = \tilde{\boldsymbol{\Phi}}_{(d)}(u_1, \ldots, u_p)\mathbf{J}_{(d)}^{(D)}$$

where $\mathbf{J}_{(d)}^{(D)}$ is an indicator matrix with entries 0 and a single entry 1 per column for extracting the matching columns of $\tilde{\mathbf{\Phi}}_{(d)}$. Note that with this definition $\mathbf{J}_{(d)}^{(pd)}$ is the identity matrix. Let $\tilde{\mathbf{b}}^{(D)}$ denote the basis coefficients corresponding to the sparse splines basis. We define the sparse penalized log likelihood by extracting the corresponding elements from the complete penalty matrix, that is

$$\tilde{l}_p^{(D)}(\tilde{\mathbf{b}}^{(D)}, \boldsymbol{\lambda}) = \tilde{l}^{(D)}(\tilde{\mathbf{b}}^{(D)}) - \frac{1}{2}\tilde{\mathbf{b}}^{(D)^T}\tilde{\mathbf{P}}^{(D)}(\boldsymbol{\lambda})\tilde{\mathbf{b}}^{(D)} \tag{4.10}$$

with obvious definition for $\tilde{l}^{(D)}(\tilde{\mathbf{b}}^{(D)})$ and $\tilde{\mathbf{P}}^{(D)}(\boldsymbol{\lambda}) = \mathbf{J}_{(d)}^{(D)^T}\tilde{\mathbf{P}}(\boldsymbol{\lambda})\mathbf{J}_{(d)}^{(D)}$. Note that since $\tilde{\mathbf{b}}^{(pd)} = \tilde{\mathbf{b}}$ we have $\tilde{l}^{(pd)}(\cdot) = \tilde{l}(\cdot)$.
Now that we have reduced the basis dimension to make copula density estimation feasible even beyond the bivariate case, it remains to tackle the question of how well we can approximate an arbitrary copula density $c(u)$ by a sparse grid representation $c(\mathbf{u}, \mathbf{b}^{(D)}) = \mathbf{\Phi}_{(d)}^{(D)}(\mathbf{u})\mathbf{b}^{(D)}$.

4.2.3 Approximation Error

Let $c^{(D)}(\mathbf{u}; \mathbf{b}) = \tilde{\mathbf{\Phi}}_{(d)}^{(D)}(\mathbf{u})\tilde{\mathbf{b}}^{(D)}$ denote the sparse grid B-spline representation of the true copula density $c(\mathbf{u})$. We assume that $c(\mathbf{u})$ is continuously differentiable, and we denote with $\tilde{\mathbf{b}}_0^{(D)}$ the true parameter in the sense that $c^{(D)}(\mathbf{u}; \tilde{\mathbf{b}}_0^{(D)})$ and $c(\mathbf{u})$ have smallest Kullback-Leibler distance with $\tilde{\mathbf{b}}_0^{(D)}$ fulfilling constraint (4.22). This implies that vector $\tilde{\mathbf{b}}_0^{(D)}$ minimizes the Lagrange function

$$E\left\{\log\big(c^{(D)}(\mathbf{u}; \tilde{\mathbf{b}}^{(D)})\big)\right\} + \rho(\mathbf{1}^T\tilde{\mathbf{b}}^{(D)} - 1) \tag{4.11}$$

with ρ as the Lagrange multiplier. Differentiation of (4.11) with respect to $\tilde{\mathbf{b}}^{(D)}$ yields

$$\int \tilde{\mathbf{\Phi}}^T(\mathbf{u})\frac{c(\mathbf{u})}{c^{(D)}(\mathbf{u}; \tilde{\tilde{\mathbf{b}}}^{(D)})}\mathrm{d}\mathbf{u} = \rho\mathbf{1} \tag{4.12}$$

where $\rho = 1$ results from multiplying (4.12) from the left hand side with $\tilde{\mathbf{b}}_0^{(D)^T}$. Using definition (4.9), we find the components of $\tilde{\mathbf{\Phi}}(\mathbf{u})$ to have the form $\prod_{j=1}^p \phi_{(h_j)l_j}(u_j)$ with $l_j \in \mathcal{I}_{(h_j)}$ and $\sum_{j=1}^p h_j \leq D$ where $h_j \geq 0$. We naturally assume that $D \geq d$ and define with $r(\mathbf{u}) = c(\mathbf{u})/c^{(D)}(\mathbf{u}; \tilde{\mathbf{b}}_0^{(D)})$ the ratio of the true and the approximate copula

density. With (4.12) we get for a single component in $\tilde{\mathbf{\Phi}}(\mathbf{u})$

$$\begin{aligned} 1 &= \int_0^1 \phi_{(h_1)l_1}(u_1) \left\{ \int_{\times_{j=2}^p [0,1]} \prod_{j=2}^p \phi_{(h_j)l_j}(u_j) r(u_1, \ldots, u_p) \mathrm{d}u_2 \ldots \mathrm{d}u_p \right\} \mathrm{d}u_1 \\ &= \int_{U_{(h_1)l_1}} \phi_{(h_1)l_1}(u_1) r_1(u_1) \mathrm{d}u_1 \end{aligned} \tag{4.13}$$

where $r_1(u_1)$ denotes the bracketed term in (4.13) and $U_{(h_j)l_1}$ is the support of basis $\phi_{(h_1)l_1}$. Following the mean value theorem for integration we find a value $\tilde{u}_{(h_1)l_1} \in U_{(h_1)l_1}$ so that $r_1(\tilde{u}_{(h_1)l_1}) = 1$. This allows to recursively apply the same argument to the bracketed term in (4.13). Let $h_1^{(D)} = D - \sum_{j=2}^p h_j$, then condition (4.13) holds for all $h_1 \leq h_1^{(D)}$. Since $\left\{ \phi_{(h_1)\mathcal{I}_{h_1}}(u_1), h_1 \leq h_1^{(D)} \right\}$ spans the linear space of $\Phi_{\left(h_1^{(D)}\right)}(u_1)$ of the non hierarchical B-spline basis of order h_1 we obtain that for all $u_1 \in [0,1]$ there exists a $\tilde{u}_1$ with $|u_1 - \tilde{u}_1| \leq 2^{-h_1^{(D)}}$ and $r_1(\tilde{u}_1) = 1$. Applying the same argument recursively we get the final result that for all $\mathbf{u} \in [0,1]^p$ there exists a $\tilde{\mathbf{u}}$ with $||\mathbf{u} - \tilde{\mathbf{u}}|| \leq 2^{-D}$ and $c(\mathbf{u}) = c^{(D)}(\tilde{\mathbf{u}}; \mathbf{b}^{(D)})$. With simple Taylor approximation we therefore obtain

$$c(\mathbf{u}) = c^{(D)}(\tilde{\mathbf{u}}; \mathbf{b}^{(D)}) + O(2^{-D}). \tag{4.14}$$

Thus, the cumulated hierarchy D determines the order of the approximation error, so, not surprisingly, accuracy and numerical feasibility are in competition.

4.2.4 Statistical Properties of the Estimate

We discuss the statistical properties of the estimate. Let $\hat{\tilde{\mathbf{b}}}$ denote the penalized Maximum Likelihood estimate based on (4.10) and let $\tilde{\mathbf{s}}_p^{(D)}(\tilde{\mathbf{b}}^{(D)}, \boldsymbol{\lambda})$ and $\tilde{\mathbf{H}}_p^{(D)}(\tilde{\mathbf{b}}^{(D)}, \boldsymbol{\lambda})$ be the first and second order derivatives of $\tilde{l}_p^{(D)}(\tilde{\mathbf{b}}^{(D)}, \boldsymbol{\lambda})$, respectively, i.e.,

$$\tilde{\mathbf{s}}_p^{(D)}(\tilde{\mathbf{b}}^{(D)}, \boldsymbol{\lambda}) = \sum_{i=1}^n \frac{\mathbf{\Phi}_{(d)}^{(D)}(u_i)}{c^{(D)}(u_i, \tilde{\mathbf{b}}^{(D)})} - \tilde{\mathbf{P}}^{(D)}(\boldsymbol{\lambda}) \tilde{\mathbf{b}}^{(D)} \tag{4.15}$$

$$\tilde{\mathbf{H}}_p^{(D)}(\tilde{\mathbf{b}}^{(D)}, \boldsymbol{\lambda}) = -\sum_{i=1}^n \frac{\mathbf{\Phi}_{(d)}^{(D)}(u_i) \mathbf{\Phi}_{(d)}^{(D)^T}(u_i)}{c^{(D)}(u_i, \tilde{\mathbf{b}}^{(D)})} - \tilde{\mathbf{P}}^{(D)}(\boldsymbol{\lambda}) \tag{4.16}$$

where $c^{(D)}(u_i, \tilde{\mathbf{b}}^{(D)}) = \mathbf{\Phi}_{(d)}^{(D)}(u_i) \tilde{\mathbf{b}}^{(D)}$. Denote with $\tilde{\mathbf{b}}_0^{(D)}$ the 'true' spline coefficient vector, in the sense that the true copula density $c(\mathbf{u})$ and $c^{(D)}(\mathbf{u}, \tilde{\mathbf{b}}_0^{(D)})$ have smallest Kullback-Leibler distance. This defines $\tilde{\mathbf{b}}_0^{(D)}$ implicitly through $E\left\{ \tilde{\mathbf{s}}_p(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda} = 0) \right\} = 0$. For the solution of $\tilde{s}_p^{(D)}(\hat{\tilde{\mathbf{b}}}^{(D)}, \boldsymbol{\lambda}) = 0$, we get with simple regular expansion techniques

(see e.g., Kauermann, Krivobokova, and Fahrmeir 2009)

$$\hat{\tilde{\mathbf{b}}}^{(D)} - \tilde{\mathbf{b}}_0^{(D)} = -\boldsymbol{H}_p^{(D)^{-1}}(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda})\mathbf{s}_p^{(D)}(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda}) + \dots$$

which allows us to derive asymptotic statements about the estimates for $n \to \infty$ and D fixed. In fact, applying the central limit theorem we can derive asymptotic normality of $\hat{\tilde{\mathbf{b}}}^{(D)}$ with mean and variance asymptotically equal to

$$\mathrm{E}(\hat{\tilde{\mathbf{b}}}^{(D)}) = \tilde{\mathbf{b}}_0^{(D)} + \left\{\mathbf{H}_p^{(D)}(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda})\right\}^{-1} \tilde{\mathbf{P}}^{(D)}(\boldsymbol{\lambda})\tilde{\mathbf{b}}_0^{(D)} \tag{4.17}$$

$$\mathrm{Var}(\hat{\tilde{\mathbf{b}}}^{(D)}) = \left\{\mathbf{H}_p^{(D)}(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda})\right\}^{-1} \mathbf{H}_p^{(D)}\left(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda} = 0\right)\left\{\mathbf{H}_p^{(D)}(\tilde{\mathbf{b}}_0^{(D)}, \boldsymbol{\lambda})\right\}^{-1}. \tag{4.18}$$

4.2.5 Constraints on the Parameters and Penalization

Until now we have not incorporated the constraints that univariate margins of the copula density $c(\mathbf{u})$ are uniform. To have the estimate $c(\mathbf{u}, \hat{\tilde{\mathbf{b}}}^{(D)})$ be a proper copula density we need to impose uniform, univariate margins. First, we need to calculate the marginal density from $\tilde{\boldsymbol{\Phi}}_{(d)}^{(D)}(u_1, \dots, u_p)\tilde{\mathbf{b}}^{(D)}$. Looking for example at Figure 4.2 we can appreciate that the univariate margins are represented with the univariate spline basis $\tilde{\Phi}_{(d)}^{(D)}(u_j)$ and the corresponding marginal basis coefficient vector $\tilde{\mathbf{b}}_{(j)}^{(D)}$, say, with elements being calculated as the sum over a set of elements of $\tilde{\mathbf{b}}^{(D)}$. In the bivariate case this results from summing up row-wise (for u_2) or column-wise (for u_1) the corresponding spline coefficients in the basis representation shown in Figure 4.2. Let the marginal hierarchical basis $\tilde{\Phi}_{(d)}(u)$ in (4.8) be indexed by $\{\tilde{\phi}_{(d)l}(\cdot), l = 1, \dots, K\}$, and let $\tilde{h} = (\tilde{h}_l, l = 1, \dots, K)$ denote the hierarchy level of $\tilde{\phi}_{(d)l}(u)$, that is, $\tilde{\phi}_{(d)l}(u)$ is element of $\tilde{\Phi}_{(\tilde{h}_l)\mathcal{I}_{\tilde{h}_l}}$. For instance, looking at Figure 4.1 (or 4.2), the hierarchy levels for the hierarchical bases built from (b), (c), and (d) are 0, 1, and 2, respectively. The sparse grid basis $\tilde{\boldsymbol{\Phi}}_{(d)}^{(D)}(\mathbf{u})$ in (4.9) can now be indexed as

$$\left\{\prod_{j=1}^{p} \tilde{\phi}_{(d)l_j}(u_j), \sum_{j=1}^{p} \tilde{h}_{l_j} \le D,\ l_j = 1, \dots, K\right\}$$

and accordingly we index the spline coefficient vector with $\tilde{\mathbf{b}}^{(D)} = (\tilde{b}_{l_1,\dots,l_p}^{(D)}; \sum_{j=1}^{p} \tilde{h}_{l_j} \le D)$. As a result, the marginal density for u_j is as follows. Let $\mathrm{d}u_{-j}$ denote the integral measure $\prod_{m \ne j} \mathrm{d}u_m$, then

$$\int_{\substack{p \\ \times \\ i \ne j}\ [0,1]} \tilde{\boldsymbol{\Phi}}_{(d)}^{(D)}(u_1, \dots, u_p)\tilde{\mathbf{b}}^{(D)}\mathrm{d}u_{-j} = \sum_{l_j=1}^{K} \tilde{\phi}_{(d)l_j}(u_j)\tilde{b}_{(j)l_j}^{(D)} =: \tilde{\Phi}_{(d)}(u_j)\tilde{b}_{(j)}^{(D)} \tag{4.19}$$

with $\tilde{\Phi}_{(d)}(\cdot)$ as hierarchical marginal basis defined in (4.8). The elements of coefficient vector $\tilde{b}_{(j)}^{(d)}$ result from the $p-1$ dimensional sum

$$b_{(j)l_j}^{(D)} = \sum_{l_{-j}:\sum_{m\neq j} l_m \leq D} \tilde{b}_{l_1,\ldots,l_p}^{(D)} \tag{4.20}$$

where l_{-j} denote the sum over all l_m with $m \neq j$. Note that (4.20) is a simple linear calculation. Note that this is a simple linear calculation and hence fast and straight forward, so that the marginal density is numerically easy to obtain. To guarantee that the marginal density is uniform, we now simply impose the constraints on the coefficients evaluated at the knots τ_k

$$\tilde{\Phi}_{(d)}(\tau_k)\hat{\tilde{b}}_{(j)}^{(D)} = 1, k = 1, \ldots, K,\ j = 1, \ldots, p. \tag{4.21}$$

We need two further constraints to have $c(\mathbf{u}, \tilde{\mathbf{b}}^{(\mathbf{D})})$ being a density. First, the fitted curve $c(\mathbf{u}; \hat{\tilde{\mathbf{b}}}^{(D)}) := \tilde{\mathbf{\Phi}}_{(d)}^{(D)}(u_1, \ldots, u_p)\hat{\tilde{\mathbf{b}}}^{(D)}$ is required to be a density. Since all columns in the hierarchical basis $\tilde{\mathbf{\Phi}}_{(d)}^{(D)}$ are B-spline densities over $u_1 \ldots, u_p$ we therefore need to guarantee that the sum of the components of $\hat{\tilde{\mathbf{b}}}^{(D)}$ equals 1, i.e.,

$$\mathbf{1}^T\hat{\tilde{\mathbf{b}}}^{(D)} = 1. \tag{4.22}$$

We also need that the fitted density is nonnegative which yields the additional constraint

$$c(u_1, \ldots, u_p; \hat{\tilde{\mathbf{b}}}^{(D)}) \geq 0,\ u_j \in [0, 1],\ j = 1, \ldots, p. \tag{4.23}$$

The constraints (4.21), (4.22) and (4.23) can be accommodated as side conditions in a quadratic programming tool to maximize the likelihood (4.10). We made use of the implemented version in `R` in the `quadprog` package. As a starting value for $\tilde{\mathbf{b}}$, we use a uniform distribution on the the cube $[0, 1]^p$. This is easily obtained with the hierarchical B-spline basis. The knots are placed equidistantly. The entire procedure is implemented in the `R` package `pencopula` (see Schellhase 2012) available on the `CRAN` server (see *http://cran.r-project.org/*).

Note that (4.21) and (4.22) are simple equations. To satisfy constraint (4.23), we require the condition to hold at the $(2^d + 1)^p$ equidistant knots locations of the tensor product B-spline density basis. If p and d increase, the number of conditions and hence the computational effect of the quadratic program increase enormously, e.g. a full tensor product for $p = 4$ and $d = 4$ contains 83521 entries. With the following trick, we can reduce the calculation time without any loss of accuracy. The idea is, when calculating the constraints, to omit knot locations of the full tensor product where

the density itself is high. This is incorporated in the algorithm in two ways. First, in the initial step we omit knot locations for the calculation of the constraint (4.23) which are close to the observations. In the subsequent steps, when density estimates in the iteration are available, we omit knot locations with a high value of the fitted density. Such reduction of the constraints accelerates the computation of the quadratic programming step.
The final thing to adjust is the amount of penalization. In practice, we need to choose λ in a data-driven manner and in principle we need to select a separate λ_j for each dimension. To limit the numerical effort, however, we let $\lambda_1 = \lambda_2 = \cdots = \lambda_p$ and minimize the corrected Akaike information criterion (Hurvich and Tsai 1989, see also Burnham and Anderson 2010) defined as

$$\mathrm{AIC}_c(\boldsymbol{\lambda}) = -2\tilde{l}(\hat{\tilde{\mathbf{b}}}^{(D)}, \boldsymbol{\lambda}) + 2\mathrm{df}(\boldsymbol{\lambda}) + \frac{2\mathrm{df}(\boldsymbol{\lambda})(\mathrm{df}(\boldsymbol{\lambda})+1)}{n - \mathrm{df}(\boldsymbol{\lambda}) - 1} \tag{4.24}$$

where $\mathrm{df}(\boldsymbol{\lambda})$ is the degree of the model defined through

$$\mathrm{df}(\boldsymbol{\lambda}) = \mathrm{tr}\left[\left\{\tilde{\mathbf{H}}_p^{(D)}(\hat{\tilde{\mathbf{b}}}^{(D)}, \boldsymbol{\lambda})\right\}^{-1} \tilde{\mathbf{H}}_p^{(D)}\left(\hat{\tilde{\mathbf{b}}}^{(D)}, \boldsymbol{\lambda} = 0\right)\right].$$

where $\tilde{H}_p^{(D)}(.)$ is the second order derivative of the likelihood, see formula (4.16) for details.

4.3 Simulations and Examples

4.3.1 Simulation

To get an impression of the performance of the routine, we simulated data from a given copula $c_0(\cdot)$, say, using the `copula` package in `R`; see Yan (2007). We thereby simulate data from different copulas in two correlation scenarios with Kendall's tau $\tau = 0.25$ and with $\tau = 0.5$, respectively. With respect to the copulas, we simulate data from (i) a Clayton copula, (ii) a Frank copula, (iii) a Gumbel copula and two different t-copulas, (iv) a t-copula with 3 degrees of freedom, and (v) a t-copula with 4 degrees of freedom, each with sample size n=500. We simulate data in $p = 2, 3$ and 4 dimensions.
We fit the simulated data following our procedure and the performance is validated by analyzing the simulation mean of the corrected Akaike information criterion AIC_c of non-parametric estimators, denoted by $\widehat{\mathrm{AIC}}_{np}$. The results are based on 200 simulations for $p = 2, 3$ and 100 simulations for $p = 4$ and shown in Table 4.4 for different values of d, the spline dimension, and D, the hierarchy order. The optimal smoothing parameter

λ is selected with a simple grid search. Note that $d = 3$, $D = 6$ as well as $d = 4$, $D = 8$ refer to a full tensor product for $p = 2$. For comparison, we fit the data with a kernel density estimator using the quadratic Epanechnikov-kernel and optimal bandwidth selected with likelihood cross-validation. For fitting we use the R package `np` (see Hayfield and Racine 2008). The corresponding AIC_c is denoted as $\widehat{\text{AIC}}_{kernel}$, where we use the multivariate analogon of the univariate Akaike information criterion by Loader (1999). Furthermore, we fit the data with Bernstein polynomials as basis functions but without any penalization (see Sancetta and Satchell 2004). We use quadratic programming with the same side constraints as in our routine, that is imposing uniform margins. As basis dimensions of the Bernstein polynomial we use $3, 4, 5, \ldots, 10$. To avoid over-fitting we select the dimension of the basis again by the use of the corrected Akaike information criterion AIC_c. The corresponding AIC_c is denoted as $\widehat{\text{AIC}}_{bern}$. As an ultimate benchmark, we calculated the AIC_c value for the true copula from which we simulated the data but with their parameter replaced by its Maximum Likelihood fitted value, as implemented in `R` using the `copula` package. This value is denoted as $\widehat{\text{AIC}}_{true}$.

Let us now look at the results in Table 4.4. First we investigate the two dimensional setting, i.e. $p = 2$, which is visualized in Figure 4.3 by plotting the distance to the optimal AIC_c for the different competitors. We start with the low correlation case, i.e. $\tau = 0.25$. The results of the full tensor product kernel $d = 4, D = 8$ yield optimal results for each copula scenario. Furthermore the sparse grid ($d = 3, D = 3$ and $d = 4, D = 4$) is slightly less efficient for this scenario, but shows comparably distances to the optimal AIC_c as the optimal full tensor product does. The kernel density approach shows the largest difference to the optimal AIC_c in this case. Also, the Bernstein polynomials are outperformed with respect to the difference to the optimal AIC_c in this case. The picture changes slightly when looking at the stronger correlation $\tau = 0.5$. Again, the full tensor product for $d = 4, D = 8$ yields the best results with respect to the distance to optimal AIC_c followed by the the full tensor product for $d = 3, D = 6$ with slightly increased differences. Moreover the sparse grid ($d = 3, D = 3$ and $d = 4, D = 4$) performs weaker but still better than the kernel approach and the Bernstein polynomials, which have the highest distance to optimal AIC_c.

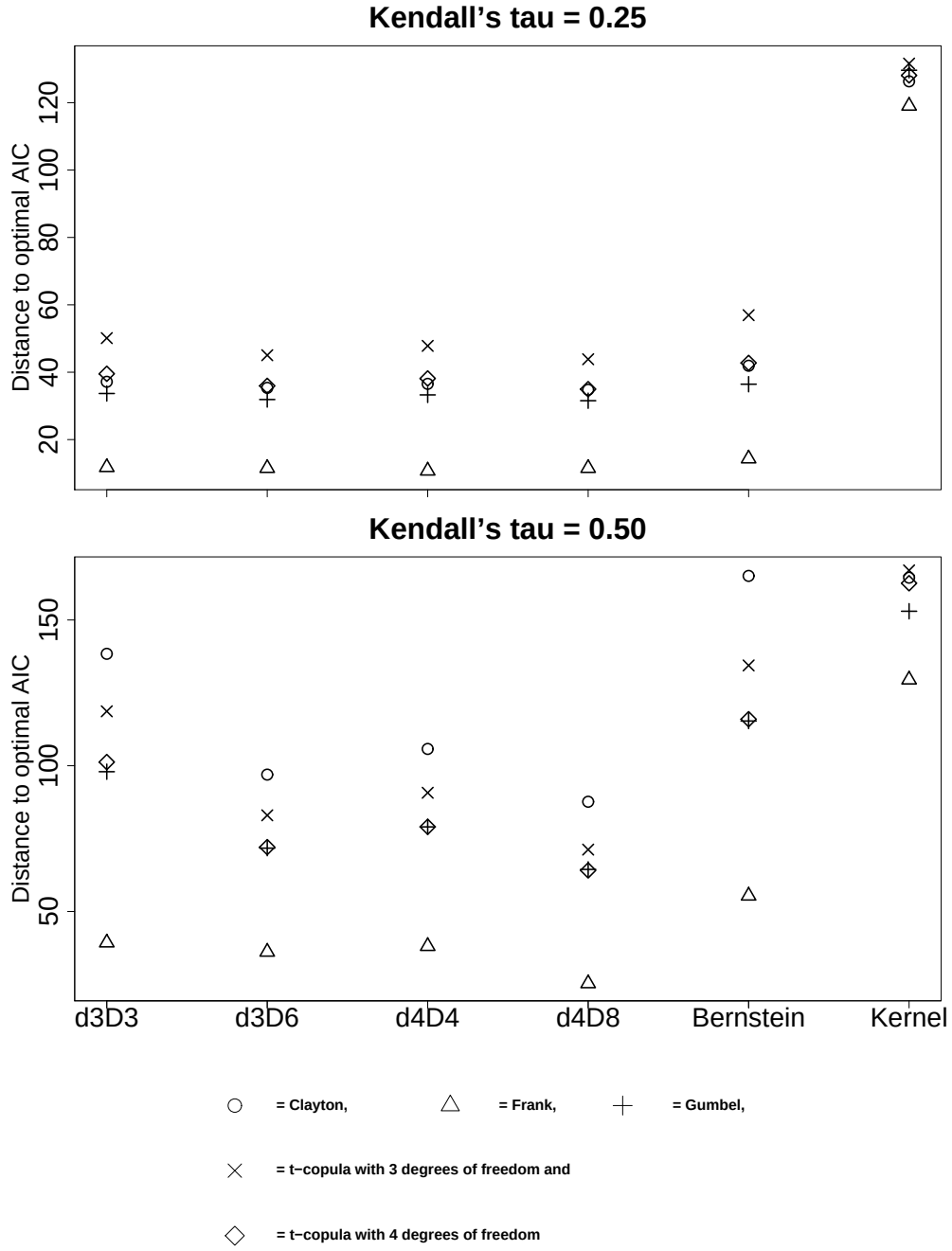


Figure 4.3: Simulated AIC difference $\widehat{\text{AIC}} - \text{AIC}_{true}$ for $p = 2$. From left to right: $\widehat{\text{AIC}}_{np} - \text{AIC}_{true}$ for $d = 3, D = 3$ and $d = 3, D = 6$ and $d = 4, D = 4$ and $d = 4, D = 8$, respectively, $\widehat{\text{AIC}}_{bernstein} - \text{AIC}_{true}$ and finally $\widehat{\text{AIC}}_{kernel} - \text{AIC}_{true}$

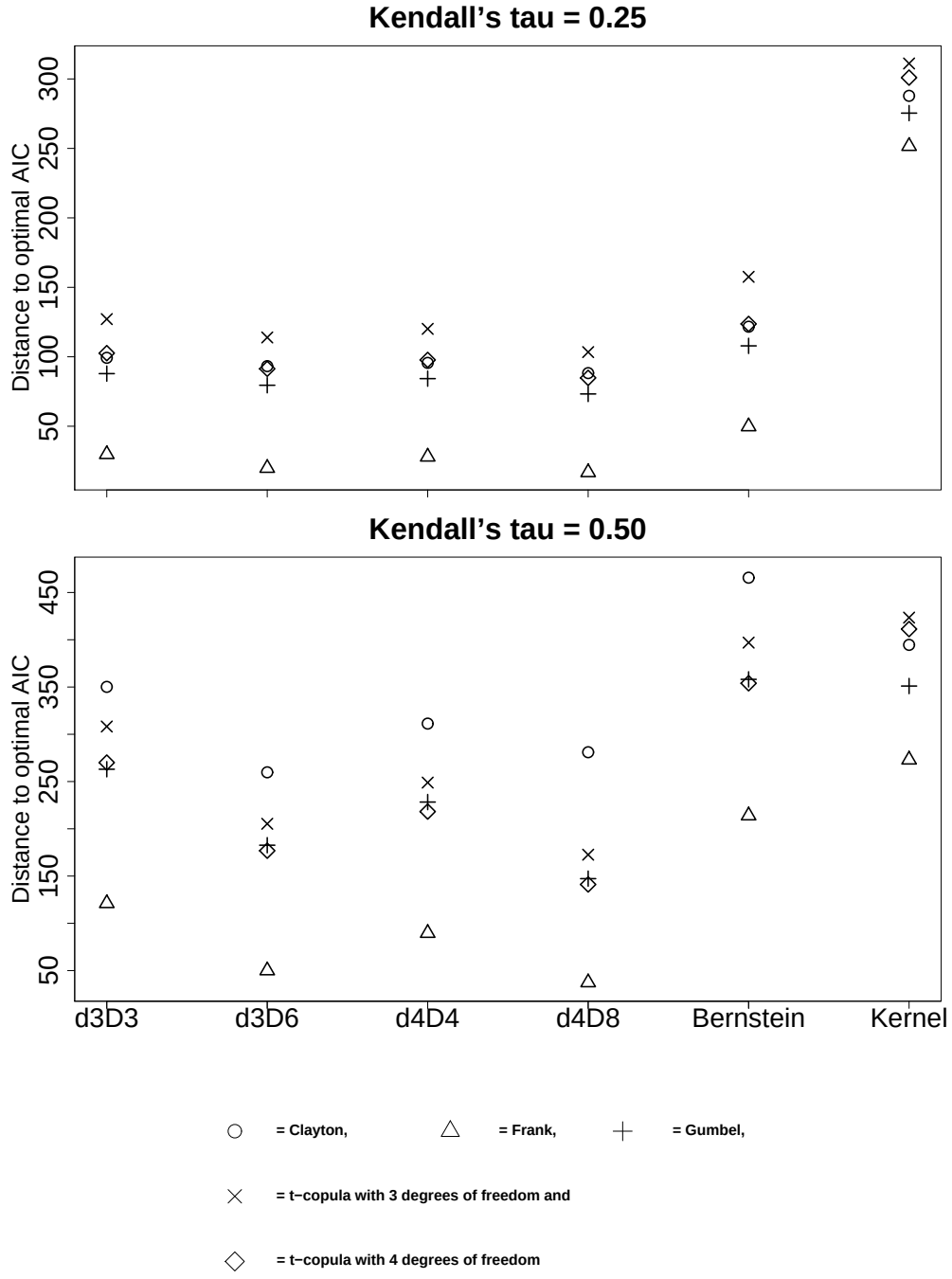


Figure 4.4: Simulated AIC difference $\widehat{\text{AIC}} - \text{AIC}_{true}$ for $p = 3$. From left to right: $\widehat{\text{AIC}}_{np} - \text{AIC}_{true}$ for $d = 3, D = 3$ and $d = 3, D = 6$ and $d = 4, D = 4$ and $d = 4, D = 8$, respectively, $\widehat{\text{AIC}}_{bernstein} - \text{AIC}_{true}$ and finally $\widehat{\text{AIC}}_{kernel} - \text{AIC}_{true}$

Next we look at dimension $p = 3$. The results are visualized in Figure 4.4. Note that for $p = 3$ all cases of our approach are sparse grids and the full tensor pruduct with e.g. $d = 4, D = 12$ would be numerically demanding, see also Table 4.1. Generally, for the small correlation case $\tau = 0.25$ (top) we see a tendency that the sparse grid fit outperforms both, the Bernstein polynomial fit and the kernel based fit. Looking at sparse grids using $d = 3, D = 6$ and $d = 4, D = 8$, we obtain the smallest distance to the optimal AIC_c. A similar picture is seen for the strong correlation case, i.e. $\tau = 0.5$. The spars grids using $d = 3, D = 3$ and $d = 4, D = 4$ show comparable differences to optimal AIC_c.
Finally, considering the four dimensional case $p = 4$, we simulate from the Clayton, Frank and t-copula with 4 degrees of freedom. For the low correlation case $\tau = 0.25$ we observe the lowest distances to the optimal AIC_c for the sparse grids and the Bernstein polynomials and the kernel approach are outperformed. Looking at the stronger correlation $\tau = 0.5$ we observe a similar behaviour. Overall we can conclude that the sparse grid behaves competitive, in particular for dimensions beyond 2.
Finally, looking at the computing time we list in Table 4.2 the CPU time for the sparse grid approach for different values of $d(= D)$ and dimensions $p = 2, 3, 4$. Again, though the computing time increases with p, calculation is still feasible for dimension $p = 4$.

$d = D$	$p = 2$	$p = 3$	$p = 4$
3 ($K = 9$)	1.063	2.020	13.652
4 ($K = 17$)	4.017	11.081	175.251

Table 4.2: Elapsed `system.time` for a Frank copula with $N = 500$ observations.

4.3.2 Example

Finally, we illustrate the applicability of the procedure with two examples. In both examples, we use t-distribution as univariate margins with maximum-likelihood theory estimated parameters. We present the results with smoothing parameter λ, chosen by AIC_c in Table 4.3.2.
First, we look at monthly interest rate data from the `R` package `Ecdat` using the data set `Capm`. The raw data are monthly risk-free interest rates which could be used to fit a Capital Asset Pricing Model (CAPM). We have jittered the data somewhat and created a bivariate sample by computing lagged rates and changes in rates. The data and the contour plot of the sparse grid-based fitted copula (left) and the corresponding copula density (right) are plotted in Figure 4.5, for $d = 5$ and $D = 5$. Note that the copula distribution function on $\bigtimes_{j=1}^{p}[0, 1]$ is easily calculated by taking the integrated B-spline

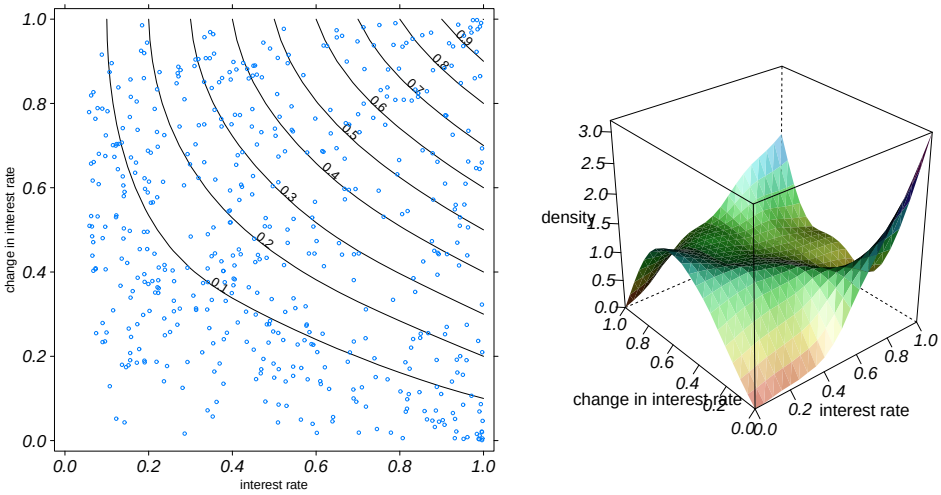


Figure 4.5: Copula (left) and copula density (right) for the interest rate data from the data set `Capm` in the `R` package `Ecdat` with $d = 5$ and $D = 5$.

densities weighted with the spline coefficients. The density shows a strong positive association between the lagged rate and the volatility of the rate change. Specifically, the density is high where the lagged rate and the magnitude of the rate change are either both small or both large. For comparison, we fitted the copula for different spline dimensions and also with a full tensor product and list the results in Table 4.3.2 (left). We show the maximum likelihood $\hat{l}$ and the Akaike Information criterion. Moreover we fit classical copula families to the data with maximum-likelihood theory estimated parameters. Also, we use Bernstein polynomials to construct the copula and choose the dimension of them by the Akaike Information Criterion. The results are shown in Table 4.3.2. Apparently, none of the parametric models are close to the results of the non-parametric approach and among the latter, the penalization spline estimators outperform the Bernstein polynomial estimators, using the AIC_c as the criterion.
As a second example, we investigate three daily world currency indices from January 3rd, 2000 until May 6th, 2011. The dataset includes values of $n = 2854$ business days compared to the US-dollar. The data set includes the Australian dollar (AUS), the Euro (EUR) and the Japanese yen (YEN). We analyze the log-return from day t to day $t + 1$. We present the results for this data set in Table 4.3.2 (right). For comparison we also fit parametric copula models to the data, also listed in Table 4.3.2. Note, a full tensor product for $p = 3$ is constructed with $d = 3, D = 9$ or $d = 4, D = 12$, but at least for $d = 4, D = 12$ the approach is not feasible due to the curse of dimensionality. Therefore, we fit the data with a compromise between the smallest sparse grid and the full tensor product, using $d = 3, D = 6$ and $d = 4, D = 8$. The greater sparse grids with $d = 3, D = 6$ and $d = 4, D = 8$ result with higher log-likelihood compared

		`Capm` data		exchange rate data	
d	D	log-likelihood $\hat{l}$	AIC_c	log-likelihood $\hat{l}$	AIC_c
3	3	40.343	-51.162	873.980	-1610.068
3	6	50.932	-55.714	1007.578	-1725.735
4	4	43.983	-52.412	978.359	-1707.725
4	8	57.361	-57.077	1117.326	-1774.491
5	5	46.202	-53.209	-	-
5	10	60.598	-58.556	-	-
Clayton		19.008	-36.007	83.410	-164.819
Frank		2.811	-3.654	2.707	-3.412
Gumbel		1.391	-0.775	31.649	-61.296
Normal		3.990	-5.972	27.654	-53.307
Bernstein		34.417	-36.833	886.640	-1523.279

Table 4.3: Results for various combinations of d and D for data examples in Section 4.3.2, compared with results fitting maximum likelihood based optimal parameters for classical copula families and Bernstein polynomials choosing the dimension by the Akaike Information Criterion.

with the cases $d = 3, D = 3$ and $d = 4, D = 4$. Overall, the fits are better than the competing models including the Bernstein polynomials. Our approach allows to analyze the bivariate margins of this estimated three dimensional copula. The contour plot of the fitted bivariate margins (left) with minimal AIC_c and the corresponding copula density (right) are plotted in Figure 4.6 with $d = 4$ and $D = 8$. We observe different dependencies among the bivariate margins. Obviously, the high peaks in $(0, 0)$ and $(1, 1)$ in the bivariate marginal copula of the Euro and the Australian dollar (Figure 4.6, right in the top row) indicate dependence between these currencies in the observation period, both currencies have risen or have fallen if one of them have risen or have fallen. The bivariate marginal copula of the Euro and the Japanese yen (Figure 4.6, right in the middle row) shows a different dependency. The bivariate marginal copula of the Australian dollar and the Japanese yen (Figure 4.6, right in the middle row) shows more complex behaviour, which is mirrored in the non-parametric fit.

4.4 Discussion

We propose in this chapter how to fit copula densities with penalized B-splines. Our approach thereby accommodates side constraints like uniform univariate margins so that the fitted density is a copula density itself. The use of a reduced tensor product basis allows to extend the approach to higher dimensions by maintaining numerical feasibility. Apparently, the approach does not circumvent the curse of dimensionality,

but it shifts it a little bit so that calculation on 3, 4 (or 5) dimensions is possible. Moreover, we show (see Table 4.3.2), that the choices of d and D are not crucial, if they are chosen large enough to avoid substantial bias. The approach can be extended to higher dimensions by making use of further techniques as for instance pair copula estimation. Generally, the semi-parametric approach suggested in the chapter contributes to the weakly development field of non- and semi-parametric copula estimation.

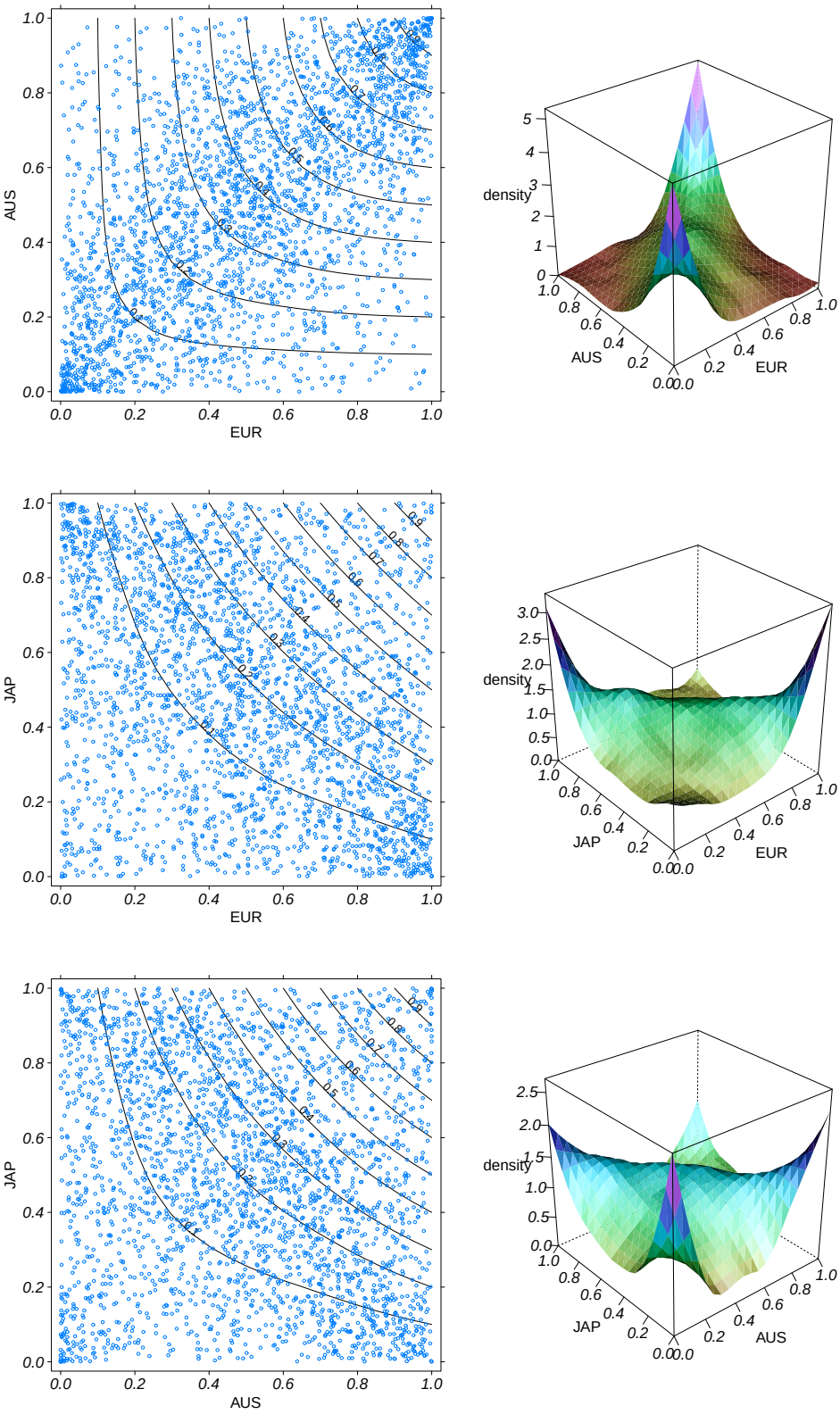


Figure 4.6: Bivariate marginal copula distribution (left) and copula density (right) between Euro (EUR), Australian Dollar (AUS) and Japanese Yen (JAP) compared to the US-dollar from January 3rd, 2000 until May 6th, 2011 with $d = 4$ and $D = 8$.

	copula	true AIC_{true}	$d=3$ $D=3$	$d=3$ $D=6$	$d=4$ $D=4$	$d=4$ $D=8$	Bernstein AIC_{bern}	kernel Epanechnikov AIC_{kernel}
$p=2$	(i) Clayton $\tau=0.25$	-107.01 (22.38)	-69.85 (16.67)	-71.66 (17.40)	-70.47 (17.34)	**-72.22 (17.89)**	-65.09 (15.51)	19.33 (23.05)
	(i) Clayton $\tau=0.50$	-427.26 (38.25)	-288.87 (23.01)	-330.31 (32.40)	-321.50 (29.39)	**-339.61 (46.75)**	-262.14 (22.89)	-262.74 (38.64)
	(ii) Frank $\tau=0.25$	-72.70 (15.94)	-60.93 (14.33)	-61.19 (14.37)	**-61.97 (15.06)**	-61.19 (14.55)	-58.37 (16.03)	46.34 (25.67)
	(ii) Frank $\tau=0.50$	-315.72 (28.86)	-276.43 (26.21)	-279.60 (32.83)	-277.65 (25.89)	**-290.42 (28.25)**	-260.35 (23.65)	-186.22 (35.77)
	(iii) Gumbel $\tau=0.25$	-94.19 (19.38)	-60.51 (15.25)	-62.30 (15.62)	-60.92 (15.51)	**-62.66 (15.87)**	-57.76 (14.86)	35.44 (21.04)
	(iii) Gumbel $\tau=0.50$	-374.33 (34.29)	-276.38 (25.09)	-302.60 (30.24)	-295.31 (29.20)	**-309.89 (31.91)**	-258.99 (24.66)	-221.33 (34.27)
	(iv) tcop $df=3, \tau=0.25$	-119.29 (25.59)	-69.17 (20.39)	-74.25 (21.27)	-71.48 (20.95)	**-75.44 (21.52)**	-62.36 (17.93)	12.32 (25.16)
	(iv) tcop $df=3, \tau=0.50$	-390.74 (43.55)	-272.07 (29.82)	-307.75 (37.40)	-300.02 (36.21)	**-319.51 (39.84)**	-256.34 (29.66)	-223.74 (42.62)
	(v) tcop $df=4, \tau=0.25$	-102.52 (21.86)	-62.97 (17.10)	-66.56 (17.63)	-64.36 (17.46)	**-67.53 (17.80)**	-59.75 (15.71)	25.60 (25.18)
	(v) tcop $df=4, \tau=0.50$	-376.86 (38.80)	-275.60 (29.48)	-304.79 (34.86)	-297.81 (33.47)	**-312.76 (42.46)**	-260.91 (28.51)	-214.26 (41.80)
$p=3$	(i) Clayton $\tau=0.25$	-273.40 (37.94)	-174.14 (25.85)	-180.14 (28.15)	-177.79 (26.69)	**-185.16 (29.68)**	-151.76 (23.89)	14.45 (37.46)
	(i) Clayton $\tau=0.50$	-974.26 (67.70)	-624.11 (39.54)	**-714.50 (47.27)**	-662.98 (113.29)	-693.25 (123.01)	-508.62 (33.75)	-579.69 (60.68)
	(ii) Frank $\tau=0.25$	-192.99 (27.84)	-163.18 (25.81)	-173.13 (26.95)	-165.01 (26.25)	**-176.21 (27.80)**	-143.22 (25.94)	58.60 (38.50)
	(ii) Frank $\tau=0.50$	-747.08 (48.07)	-625.93 (40.03)	-697.03 (43.33)	-657.44 (49.28)	**-709.70 (74.74)**	-533.13 (32.25)	-474.06 (52.78)
	(iii) Gumbel $\tau=0.25$	-247.64 (35.96)	-159.77 (25.31)	-168.13 (24.44)	-163.37 (26.03)	**-174.40 (25.86)**	-139.78 (23.91)	27.76 (35.72)
	(iii) Gumbel $\tau=0.50$	-876.30 (59.10)	-613.29 (39.78)	-693.67 (45.92)	-648.01 (53.34)	**-729.04 (55.94)**	-518.21 (34.66)	-525.27 (55.88)
	(iv) tcop $df=3, \tau=0.25$	-299.08 (37.95)	-171.96 (26.35)	-185.16 (26.53)	-178.98 (27.53)	**-195.71 (29.29)**	-141.52 (23.23)	12.05 (37.86)
	(iv) tcop $df=3, \tau=0.50$	-896.82 (62.32)	-588.61 (41.44)	-691.52 (54.46)	-647.94 (48.92)	**-724.21 (99.85)**	-499.84 (38.24)	-473.40 (67.63)
	(v) tcop $df=4, \tau=0.25$	-261.47 (36.83)	-158.86 (26.85)	-170.10 (32.20)	-163.74 (27.83)	**-176.65 (34.63)**	-137.83 (25.17)	39.48 (36.90)
	(v) tcop $df=4, \tau=0.50$	-859.93 (64.03)	-589.89 (44.97)	-683.12 (58.14)	-641.70 (49.90)	**-718.80 (71.10)**	-505.96 (41.39)	-448.67 (67.88)
$p=4$	(i) Clayton $\tau=0.25$	-462.05 (56.02)	-278.72 (36.68)	-	**-288.16 (37.57)**	-	-164.85 (37.33)	24.82 (54.77)
	(i) Clayton $\tau=0.50$	-1576.78 (95.39)	-886.38 (47.02)	-	**-916.73 (89.81)**	-	-716.17 (44.24)	-885.47 (81.09)
	(ii) Frank $\tau=0.25$	-346.10 (34.68)	-276.96 (30.91)	-	**-285.13 (32.66)**	-	-162.40 (31.57)	80.21 (46.67)
	(ii) Frank $\tau=0.50$	-1232.39 (61.67)	**-959.32 (50.03)**	-	-940.42 (125.61)	-	-765.26 (42.94)	-773.73 (72.53)
	(v) tcop $df=4, \tau=0.25$	-456.45 (54.97)	-263.87 (34.45)	-	**-280.39 (37.33)**	-	-155.55 (32.49)	78.12 (58.53)
	(v) tcop $df=4, \tau=0.50$	-1409.14 (87.22)	-895.12 (52.63)	-	**-914.93 (121.91)**	-	-717.53 (41.36)	-657.75 (110.54)

Table 4.4: Reported is the mean (sd) of the AIC_c. The optimal results are set in bold.

5 Flexible Pair-Copula Estimation in D-vines with Penalized Splines

This essay is joint work with Göran Kauermann (LMU Munich). It is a working paper, compare Kauermann and Schellhase (2012).

In this chapter a new method for flexible fitting of dependences vines, especially for D-vines is investigated. Therefore, pair-copulas are estimated semi-parametrically using penalized Bernstein polynomials or linear B-splines, respectively, as spline bases in each knot of the D-vine throughout each level. A penalty induce smoothness of the fit while the high dimensional spline basis guarantees flexibility. To ensure uniform univariate margins of each pair-copula, linear constraints are placed on the spline coefficients and quadratic programming is used to fit the model. The amount of penalizations for each pair-copula is driven by a penalty parameter which is selected in a numerically efficient way. Simulations and practical examples accompany the presentation.

5.1 Introduction

Copula modelling and estimation has become extremely popular over the last decade. Originally introduced by Sklar (1959) the idea of a copula is attractive since it allows to decompose a multivariate distribution into its univariate margins and its interaction structure, expressed through the copula. Assuming the p-dimensional random vector $(x_1, \ldots, x_p)$ with univariate marginal distributions $F_j(x_j)$ for $j = 1, \ldots, p$ Sklar's theorem states that the joint distribution can be written as

$$F(x_1, \ldots, x_p) = C\big(F_1(x_1), \ldots, F_p(x_p)\big). \tag{5.1}$$

Here $C(.)$ is called the copula which can be comprehended as distribution function on $[0, 1]^p$ with the additional property of having uniform univariate margins. We refer to McNeil, Frey, and Embrechts (2005), Nelsen (2006) or Kolev, Anjos, and Mendes (2006) for a general discussion on copulas. For a recent overview and introduction see Härdle and Okhrin (2009) or Jaworski, Durante, Härdle, and Rychlik (2010).

Numerous strategies to model copulas have been suggested in the last years, this includes Archimedean copulas (see e.g. Okhrin, Okhrin, and Schmid 2009 or Savu and Trede 2010), elliptical copulas (see Frahm, Junker, and Szimayer 2003) or so called pair-copulas as originally proposed by Joe (1996). The idea of the latter is to model a multivariate copula by a collection of pairwise, that is two dimensional copulas. The pair-copula uses conditional distributions as arguments but the copula itself is independent of any conditioning variables. This is a restriction but it makes the approach numerically very powerful and handy as demonstrated in Czado (2010) or Aas, Czado, Frigessi, and Bakken (2009). The collection of paired copulas can be structured in a set of trees, defined as vines in Bedford & Cooke (2001, 2002). Assuming a hierarchical or sequential factorization of the distribution leads to a so called D-vine focused also in this chapter, see e.g. Kurowicka and Cooke (2006) or Smith, Min, Almeida, and Czado (2010). Though pair-copulas yield flexibility, the approach leaves the user with the task of model selection, see e.g. Min and Czado (2011). In fact not only the vine structure needs to be determined but also for each node in the D-vine a specific copula model has to be selected, such as Archimedean or elliptical copula, etc. We aim to further develop this point by employing flexible, semi-parametric copula estimation for each pair.

Assuming a continuous distribution function $F(x_1, \ldots, x_p)$ we can differentiate (5.1) to get the density, where for $p = 2$ we get

$$f(x_1, x_2) = c\big(F_1(x_1), F_2(x_2)\big) f_1(x_1) f_2(x_2)$$

with $f_j(.)$ as marginal densities and $c(.)$ as the copula density. Our aim is to estimate the copula density $c(\cdot)$ in a flexible, that is semi-parametric way by refraining from any strong parametric assumptions on the structure of the pairs. To do so we use penalized splines with Bernstein polynomials and linear B-splines as spline basis. Bernstein polynomials for copula estimation have been used before for instance in Sancetta and Satchell (2004) or Bouezmarni, Rombouts, and Taamouti (2010). B-splines are discussed thoroughly e.g. in Ruppert, Wand, and Carroll (2003). Both, Bernstein polynomials and linear B-splines can reproduce the uniform distribution in $[0, 1]$, which is the reason why using them here.

Generally, the number of splines determines the flexibility of the model, thus taking high degree Bernstein polynomials or a high dimensional B-spline basis, yields sufficient modelling flexibility. On the other hand, like in regular spline smoothing, a high dimensional basis exhibits a large amount of estimation variability yielding non smooth, wiggled estimation. We therefore borrow the idea of penalization from the spline smoothing literature, see e.g. Wahba (1990). That is we impose a penalty on the spline

basis which guarantees numerical stability and provides a smooth well behaved fit. The following sections are organized as follows. The estimation scheme using Bernstein Polynomials and linear B-splines for the pair-copula construction is presented in Section 2. The penalization concept and the practical settings are described in the second part of Section 2. Section 3 gives a practical example and simulation studies. We finalize the chapter with an discussion in Section 4.

5.2 Pair-Copula Construction

5.2.1 D-Vines

Let $x = (x_1, \ldots, x_p)$ be a p-dimensional continuous random vector with continuously differentiable marginal distribution functions $F_j(x_j), j = 1, \ldots, p$. Let $f(x_1, \ldots, x_p)$ be the corresponding multivariate density, which with Sklar's (1959) theorem can be written as

$$f(x_1, \ldots, x_p) = c\{F_1(x_1), \ldots, F_p(x_p)\} \prod_{j=1}^{p} f_j(x_j) \tag{5.2}$$

where $c(.)$ is the copula density. To simplify notation we denote with $u_j = F_j(x_j)$ so that the copula density is written as $c(u_1, \ldots, u_p)$. We decompose $c(.)$ to pair-copulas, where we restrict ourselves to so called D-vines (see Bedford and Cooke 2002). The presentation of pair-copulas thereby follows closely the motivating introduction in Czado (2010) so that we will be concise here. The underlying idea is that we can factorize any densities to

$$f(x_1, \ldots, x_p) = \prod_{j=2}^{p} f(x_j|x_1, \ldots, x_{j-1}) f(x_1) \tag{5.3}$$

for a given index order of the variables. For $1 < t \leq p$ we can use (5.2) and write

$$\begin{aligned} f(x_t|x_1, \ldots, x_{t-1}) &= c\{F(x_t|x_1, \ldots, x_{t-2}), F(x_{t-1}|x_1, \ldots, x_{t-2})|x_1, \ldots, x_{t-2}\} \\ &\quad \times f(x_t|x_1, \ldots, x_{t-2}) \end{aligned} \tag{5.4}$$

with $c(., .|x_1, \ldots, x_{t-2})$ as conditional copula. The driving idea of pair-copulas is now that the conditional copula in (5.4) does not depend on the variables we condition on, that is in (5.4) we assume

$$\begin{aligned} &c\{F(x_t|x_1, \ldots, x_{t-2}), F(x_{t-1}|x_1, \ldots, x_{t-2})|x_1, \ldots, x_{t-2}\} \\ &\equiv c\{F(x_t|x_1, \ldots, x_{t-2}), F(x_{t-1}|x_1, \ldots, x_{t-2})\} \end{aligned} \tag{5.5}$$

To simplify notation let $c_{i,j|D} = c\{F(x_i|x_D), F(x_j|x_D)\}$ for some index set D with $i, j \notin D, i \neq j$ and $x_D = (x_k : k \in D)$. Then, assuming the pair-copula assumption (5.5) we can rewrite (5.3) to

$$f(x_1, \ldots, x_p) = \left(\prod_{j=1}^{p-1} \prod_{i=1}^{p-j} c_{i,i+j|D_{ij}} \right) \left(\prod_{j=1}^{p} f_j(x_j) \right) \tag{5.6}$$

where $D_{ij} = \{i+1, \ldots, i+j-1\}$ (see Czado 2010). The construction principle can be visualized by a set of nested trees coined as vines by Bedford and Cooke (2002). Exemplary for $p = 5$ a D-vine based on factorization (5.3) takes the form as shown in Figure 5.1.

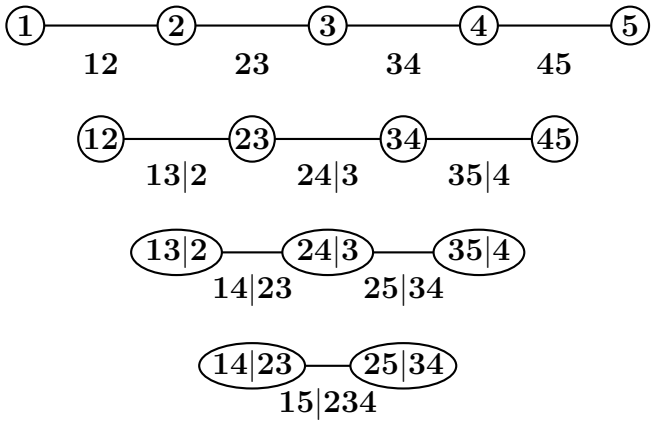


Figure 5.1: A D-vine with five covariates.

5.2.2 Approximation of Pair-Copulas

Looking at formula (5.6) we see that the entire distribution is built from bivariate copulas of the form $c_{ij|D} = c\{F(x_i|x_D), F(x_j|x_D)\}$. Our intention is now to estimate $c_{ij|D}$ in a flexible, that is semi-parametric manner. To do so we first replace the copula by a weighted sum of $K+1$ normed basis splines ϕ_{Kk_i} with $\int \phi_{Kk_i}(u)\, \mathrm{d}u = 1$ for $k_i = 0, \ldots, K$. A bivariate basis is easily constructed building a Tensor product of the basis functions ϕ_{Kk_i}. Let therefore $u_{i|D} = F(x_i|x_D)$. We now approximate $c_{ij|D}$ with the representation $\tilde{c}_{ij|D}$, say, defined through

$$\begin{aligned} \tilde{c}_{ij|D}(u_{i|D}, u_{j|D}, \boldsymbol{v}^{(i,j|D)}) &:= \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \phi_{Kk_1}(u_{i|D}) \phi_{Kk_2}(u_{j|D}) \boldsymbol{v}_{k_1,k_2}^{(i,j|D)} \\ &= \{\boldsymbol{\phi}_K(u_{i|D}) \otimes \boldsymbol{\phi}_K(u_{j|D})\} \boldsymbol{v}^{(i,j|D)} \end{aligned} \tag{5.7}$$

where $\boldsymbol{v}^{(i,j|D)} = (v_{00}^{(i,j|D)}, \ldots, v_{0K}^{(i,j|D)}, \ldots, v_{KK}^{(i,j|D)})$ is subsequently called the coefficient

vector and $\boldsymbol{\phi}_K(u) = (\phi_{K0}(u), \ldots, \phi_{KK}(u))$. We postulate positive coefficients

$$v_{k_1,k_2}^{(i,j|D)} \geq 0 \tag{5.8}$$

which in turn guarantees that $\tilde{c}_{ij|D}$ is positive. Moreover we require

$$\sum_{k_1,k_2} \boldsymbol{v}^{(i,j|D)} = 1 \tag{5.9}$$

which in turn guarantees that $\tilde{c}_{ij|D}$ in (5.7) is a density since each single component of the Tensor product is a density. Note that in order to guarantee that $\tilde{c}_{ij|D}$ is in fact a bivariate copula density we additionally need that its two univariate marginal densities are uniform. That is we need $\tilde{c}_{i|D} = \int c_{ij|D}\ \mathrm{d}u_{j|D} \equiv 1$ and accordingly $\tilde{c}_{j|D} \equiv 1$. This condition can be formulated as simple linear constraint on the coefficient vector as will be shown subsequently for the different bases used.

First, we consider Bernstein polynomials (Lorentz 1953 or Rivlin 1969) as basis functions. Let therefore $\boldsymbol{\phi}_K(u)$ be the basis of normed Bernstein polynomials of degree K, where

$$\phi_{Kk}(u) = (K+1)\binom{K}{k} u^k (1-u)^{K-k}. \tag{5.10}$$

Note that $\phi_{Kk}(u)$ is normed to be a density, i.e.(5.10) is a Beta distribution and $\int_0^1 \phi_{Kk}(u)\ \mathrm{d}u = 1$. Based on properties of Bernstein polynomials $\tilde{c}_{i|D} = \int c_{ij|D}\ \mathrm{d}u_{j|D} \equiv 1$ holds if the marginal coefficients fulfill

$$v_{k_1.}^{(i,j|D)} = \sum_{k_2} v_{k_1,k_2}^{(i,j|D)} = 1/(K+1) \tag{5.11}$$

for all $k_1 = 0, \ldots, K$. These constraints can be easily formulated in matrix notation yielding the linear constraints

$$A_K \boldsymbol{v}^{(i,j|D)} = \mathbf{1} \tag{5.12}$$

where A_K sums up the elements of $v_{k_1,k_2}^{(i,j|D)}$ column-wise (i.e. over k_2) and row-wise (i.e. over k_1), i.e. $A_K^T = ((I_K \otimes \mathbf{1}_K^T)), (\mathbf{1}_K^T \otimes I_K))$, where $\mathbf{1}_K$ is the column vector of dimension K with elements 1 and I_K is the K dimensional identity matrix. Alternatively, we use linear B-splines ϕ_{Kk_i} (see de Boor 1978), normalized to be a density, i.e. $\int \phi_{Kk_i}(u)\ \mathrm{d}u = 1$ and denote with $\boldsymbol{\phi}_K(u) = (\phi_{Kl}(u), l = 0, \ldots, K)$ the univariate B-spline density of dimension $K+1$. To guarantee that the marginal density is uniform, we now simply impose the constraints on the coefficients evaluated at the knots τ_k, so $A_K = \Phi_K(\tau)$ with $\tau = \tau_1, \ldots, \tau_K$..

From the copula density (5.7) we can easily calculate the copula $\tilde{C}(.)$ itself by noting

that

$$\tilde{C}_{ij|D}(u_{i|D}, u_{j|D}) = \int_0^{u_{i|D}} \int_0^{u_{j|D}} \tilde{c}_{ij|D}(z_i, z_j)\, \mathrm{d}z_i\, \mathrm{d}z_j.$$

Letting $\Phi_{Kk}(u) = \int_0^u \phi_{Kk}(z)\, \mathrm{d}z$ be the integrated Bernstein polynomial, i.e. the Beta distribution, or the integrated B-spline basis. Then from (5.7) we get the explicit form

$$\tilde{C}_{ij|D}(u_{i|D}, u_{j|D}|D) = \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \Phi_{Kk_1}(u_{i|D}) \Phi_{Kk_2}(u_{j|D}) v_{k_1,k_2}^{(i,j|D)}.$$

Considering copula density (5.7) we recognize that the arguments of the pair-copula, i.e. $u_{i|D}$ and $u_{j|D}$, are itself calculated from lower dimensional conditional distributions, the latter being represented by lower dimensional knots in the vine. Our approach thereby easily allows to calculate the arguments $u_{i|D}$ and $u_{j|D}$. To exemplify this note for $r \in D$ we have (see Joe 1996)

$$\begin{aligned} u_{i|D} &= F(x_i|x_D) = \frac{\partial C_{ir|D_{-r}}\{F(x_i|x_{D_{-r}}), F(x_r|x_{D_{-r}})\}}{\partial F(x_r|x_{D_{-r}})} \\ &= \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \Phi_{Kk_1}(u_{i|D_{-r}}) \phi_{Kk_2}(u_{r|D_{-r}}) v_{k_1,k_2}^{(i,r|D_{-r})}. \end{aligned} \tag{5.13}$$

where $D_{-r} = D \setminus \{r\}$. Hence, with the knowledge of coefficient vector $\boldsymbol{v}^{(i,r|D_{-r})}$ it is easy to calculate $u_{i|D}$. Iterative application of (5.13) finally allows to completely specify the pair-copula density for all variables.

5.2.3 Estimation

In the above presentation we left the specification of the univariate marginal distribution $F_i(x_j), i = \ldots, p$ so far undiscussed. This is a conventional and appealing approach by separating univariate marginal density estimation from copula density estimation, see Rank (2007, Section 2) or Jaworski, Durante, Härdle, and Rychlik (2007, Section 3). We therefore subsequently assume that the univariate margins $F_i(.)$ are either known, or they are estimated separately for instance by their empirical distribution function. Let $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{p,t})$ be an i.i.d. sample with $t = 1, \ldots, n$ and define with $\hat{u}_{i,t} = \hat{F}_i^{-1}(x_{i,t})$, where $\hat{F}_i(.)$ is either the fitted univariate margin or $\hat{F}_i(\cdot)$ is the empirical distribution function. In the latter case $\hat{u}_{i,t}$ is the (empirical) rank of $x_{i,t}$. Assume now that distributions $F(x_i|x_D)$ and $F(x_j|x_D)$ are already fitted and let $\hat{u}_{i,t|D} := \hat{F}(x_{i,t}|x_D)$, where $\hat{F}(x_i|x_D)$ denotes the fitted version of $F(x_i|x_D)$ and corresponding definition for $\hat{u}_{j,t|D}$.

With the specification of the margins it remains to estimate the set of coefficient vectors

$\boldsymbol{v}^{(i,j|D)}$ to obtain the entire distribution. With $\hat{u}_{i,t|D}$ as defined before we get the log-likelihood contribution for the pair-copula of i and j with (5.7) through

$$l_{ij|D}(\boldsymbol{v}^{(i,j|D)}) = \sum_{t=1}^{n} \log\left[\{\boldsymbol{\phi}_K(\hat{u}_{i,t|D}) \otimes \boldsymbol{\phi}_K(\hat{u}_{j,t|D})\}\boldsymbol{v}^{(i,j|D)}\right]. \tag{5.14}$$

This likelihood contribution is easily maximized with respect to $\boldsymbol{v}^{(i,j|D)}$ subject to the linear side constraints (5.8), (5.9) and (5.12). In fact simple quadratic programming can be used to solve this problem. To estimate the pair-copula we make use of the `quadprog` package in `R` which allows to solve the quadratic program. Let therefore $\mathbf{s}^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda^{(i,j|D)})$ and $\mathbf{H}^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda^{(i,j|D)})$ denote the first and second order derivatives of (5.19) yielding

$$\mathbf{s}^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda^{(i,j|D)}) = \sum_{t=1}^{T} \frac{\boldsymbol{\phi}_K(\hat{u}_{it|D}) \otimes \boldsymbol{\phi}_K(\hat{u}_{jt|D})}{\tilde{c}_{ij|D}(\hat{u}_{it|D}, \hat{u}_{jt|D}, \boldsymbol{v}^{(i,j|D)})} - \lambda^{(i,j|D)}\mathbf{P}\boldsymbol{v}^{(i,j|D)}. \tag{5.15}$$

$$\mathbf{H}^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda^{(i,j|D)}) =$$
$$-\sum_{t=1}^{T} \frac{(\boldsymbol{\phi}_K(\hat{u}_{it|D}) \otimes \boldsymbol{\phi}_K(\hat{u}_{jt|D}))(\boldsymbol{\phi}_K(\hat{u}_{it|D}) \otimes \boldsymbol{\phi}_K(\hat{u}_{jt|D}))^T}{\tilde{c}_{ij|D}(\hat{u}_{it|D}, \hat{u}_{jt|D}, \boldsymbol{v}^{(i,j|D)})} - \lambda^{(i,j|D)}\mathbf{P}. \tag{5.16}$$

We approximate the penalized likelihood $l^p_{ij|D}$ in (5.19) through a second order Taylor expansion yielding

$$l^p_{ij|D}\big(\mathbf{v}^{(ij|D)} + \boldsymbol{\delta}^{(ij|D)}, \lambda^{ij|D}\big) \approx l^p_{ij|D}\big(\mathbf{v}^{(ij|D)}, \lambda^{(ij|D)}\big)\boldsymbol{\delta}^{(ij|D)^T}\mathbf{s}^p_{ij|D}\big(\mathbf{v}^{ij|D}, \lambda^{(ij|D)}\big)$$
$$+ \frac{1}{2}\boldsymbol{\delta}^{(ij|D)^T}\mathbf{H}^p_{ij|D}\big(\mathbf{v}^{(ij|D)}, \lambda^{(ij|D)}\big)\boldsymbol{\delta}^{(ij|D)}, \tag{5.17}$$

where $\boldsymbol{\delta}^{(ij|D)}$ is the iteration step selected by maximizing (5.17) subject to the linear constraints (5.8), (5.9) and (5.12). This optimization is carried out iteratively, by approximating the likelihood as in (5.17) in each iteration step. To start the algorithm an admissible starting value for $\mathbf{v}^{(i,j|D)}$ is required. We use a uniform distribution on the the cube $[0,1]^2$ which defines the starting value in unique way.

Considering now a D-vine structure shown exemplary in Figure 5.1 we see that we can fit the entire copula by successively fitting pair-copulas by maximizing log-likelihoods of type (5.14). In fact we fit on each level the knots of the tree and calculate the fitted coefficients $\hat{u}_{i|D}$ with (5.13) from previously fitted copulas. In particular, if parallel computing is possible, the entire procedure can be calculated parallel on each tree level.

5.2.4 Penalization

Though the approach above is flexible, it may not be parsimonious at the same time since we parameterize each bivariate copula by a set of $(K+1)^2$ parameters. As a consequence the fitted copula may be wiggled and not desirably smooth. This problem is well known from the smoothing literature (see Wahba 1990) and can be easily solved by imposing an appropriate penalty on the log-likelihood. In fact, assuming smooth copula densities it seems natural to postulate that the integrated squared second order derivatives are small, see e.g. Wood (2006). We therefore formulate a penalty matrix of the form

$$\int \left(\frac{\partial^2 \tilde{c}_{ij|D}(u_i, u_j)}{(\partial u_i)^2} \right)^2 + \left(\frac{\partial^2 \tilde{c}_{ij|D}(u_i, u_j)}{(\partial^2 u_j)^2} \right)^2 \, \mathrm{d}u_i \, \mathrm{d}u_j. \tag{5.18}$$

We can rewrite (5.18) for the Bernstein polynomials. For the marginal penalties in u_i and u_j in (5.18) follows with (5.7) and transformations

$$\begin{aligned}
&\int \left(\frac{\partial^2 \tilde{c}_{ij|D}(u_i, u_j)}{(\partial u_i)^2} \right)^2 \, \mathrm{d}u_i \, \mathrm{d}u_j \\
&= (\boldsymbol{v}^{(i,j|D)})^T \int \left[\frac{\partial^2 \phi_K(u_{i|D})}{(\partial u_i)^2} \otimes \phi_K(u_{j|D}) \right]^T \left[\frac{\partial^2 \phi_K(u_{i|D})}{(\partial u_i)^2} \otimes \phi_K(u_{j|D}) \right] \, \mathrm{d}u_i \, \mathrm{d}u_j \boldsymbol{v}^{(i,j|D)} \\
&= (\boldsymbol{v}^{(i,j|D)})^T \underbrace{\int \left[\left(\frac{\partial^2 \phi_K(u_{i|D})}{(\partial u_i)^2} \right)^T \frac{\partial^2 \phi_K(u_{i|D})}{(\partial u_i)^2} \right] \, \mathrm{d}u_i \otimes \left[(\phi_K(u_{j|D}))^T \phi_K(u_{j|D}) \right]}_{:= P_{u_i}} \boldsymbol{v}^{(i,j|D)}
\end{aligned}$$

The integral of the second order derivatives of Bernstein polynomials are calculated easily. The second order derivative of (5.10) equals (see Doha, Bhrawy, and Saker 2011)

$$\frac{\partial^2 \phi_{Kk}(u)}{(\partial u)^2} = \frac{(K+1)!}{(K-2)!} \sum_{m=\max(0,k+2-K)}^{\min(k,2)} (-1)^{m+2} \binom{2}{m} \phi_{K-2,k-m}(u).$$

This is rewritten as

$$\frac{\partial^2 \phi_{Kk}(u)}{(\partial u)^2} = (\phi_{K-2,k}(u) B) w$$

with

$$B = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}, B \in \mathbb{R}^{(K-2)\times(K+1)}$$

and $w = \frac{(K+1)!}{(K-2)!}$. Therefore, the matrix P_{z_i} and P_{z_j} are equivalent to

$$P_{u_i} = (wB^T \int \phi_{K-2,k}(u_{i|D})\phi_{K-2,k}(u_{i|D})\ \mathrm{d}u_i Bw) \otimes \left[(\phi_K(u_{j|D}))^T \phi_K(u_{j|D})\right]$$
$$P_{u_j} = \left[(\phi_K(u_{i|D}))^T \phi_K(u_{i|D})\right] \otimes (wB^T \int \phi_{K-2,k}(u_{j|D})\phi_{K-2,k}(u_{j|D})\ \mathrm{d}u_j Bw).$$

So, the penalty can be written as quadratic form $\lambda^{(i,j|D)} \boldsymbol{v}^{(i,j|D)^T} P_{int} \boldsymbol{v}^{(i,j|D)}$ where $\lambda^{(i,j|D)}$ is the penalty parameter steering the amount of smoothness and $P_{int} := P_{u_i} + P_{u_j}$. It follows, we can rewrite (5.18) for the Bernstein polynomials as quadratic form $\boldsymbol{v}^{(i,j|D)^T} \mathbf{P} \boldsymbol{v}^{(i,j|D)}$ with $\mathbf{P}$ as penalty matrix. Note that $\mathbf{P}$ needs to be calculated only once for all bivariate copulas. We therefore suggest to replace the log-likelihood (5.14) by its penalized version

$$l^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda^{(i,j|D)}) = l_{ij|D}(\boldsymbol{v}^{(i,j|D)}) - \frac{1}{2}\lambda^{(i,j|D)} \boldsymbol{v}^{(i,j|D)^T} \mathbf{P} \boldsymbol{v}^{(i,j|D)}, \tag{5.19}$$

where $\lambda^{(i,j|D)}$ is the penalty parameter steering the amount of penalization.
Though penalizing the integrated squared second order derivatives is standard in the spline smoothing literature it might not be the best penalty choice for copula estimation. In fact, using the integrated squared second order derivatives as penalty and due to the side constraints (5.8), (5.9) and (5.12) we obtain a quadratic copula if we set the penalty parameter $\lambda^{(i,j|D)}$ to infinity. Intuitively, it might therefore better to work with a difference penalty of first or second order differences of the coefficients as suggested for spline smoothing in Eilers and Marx (1996). We define the difference based penalty matrix $\mathbf{P_{diff}}$ for the m-order differences through

$$\mathbf{P^m_{diff}} := (\mathbf{1}_{K+1} \otimes L_m)^T (L_m \otimes \mathbf{1}_{K+1}) \tag{5.20}$$

with e.g.

$$L_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

Now, with $\mathbf{P}$ in (5.19) replaced by $\mathbf{P^m_{diff}}$ we obtain the independence copula, if we set the penalty parameter $\lambda^{(i,j|D)}$ to infinity. As before, we maximize (5.19) using quadratic programming, which makes use of the first (5.15) and second order derivatives (5.16) of (5.19).

5.2.5 Selecting the Penalty Parameter

The penalty parameter $\lambda^{(i,j|D)}$ in (5.19) needs to be selected adequately, that is data driven based on the data at hand. To simplify notation, let us write λ instead of $\lambda^{(i,j|D)}$ in this section. Given the quadratic form of the penalty in (5.19) we again borrow results from the spline smoothing literature. The idea is to comprehend the penalty as normal prior imposed on the spline coefficient vector as proposed for smoothing spline coefficient by Wahba (1985), Stein (1990) or Efron (2001). The idea has been extended to penalized spline estimation presented in Ruppert, Wand & Carroll (2003, 2009) and is being used here as well. To do so we adopt a Bayesian viewpoint and comprehend the penalty as *'a priori'* normal distribution on the spline coefficient in that

$$\boldsymbol{v}^{(i,j|D)} \sim N(0, \lambda^{-1}\mathbf{P}^{-}) \tag{5.21}$$

where $\mathbf{P}^{-}$ denotes the (generalized) inverse of the used penalty matrix $\mathbf{P}$. The penalty parameter now plays the role of a (hyper) parameter in the prior distribution which can be estimated by maximizing the resulting likelihood. The latter is equivalent to following empirical Bayes arguments. The prior (5.21) is degenerated, which needs to be corrected as follows. We decompose $\boldsymbol{v}^{(i,j|D)}$ into the two components $\boldsymbol{v}^{(i,j|D)^{\sim}}$ and $\boldsymbol{v}^{(i,j|D)^{\perp}}$, respectively, such that $\boldsymbol{v}^{(i,j|D)^{\sim}}$ is a normally distributed random vector with non degenerated variance and $\boldsymbol{v}^{(i,j|D)^{\perp}}$ are the remaining components treated as parameters, see also Wand and Ormerod (2008). In fact based on a singular value decomposition we have

$$\mathbf{P} = \mathbf{U}^{\sim}\mathbf{\Lambda}^{\sim}\mathbf{U}^{\sim T}$$

with $\mathbf{\Lambda}^{\sim}$ as diagonal matrix with positive eigenvalues and $U^{\sim} \in \mathbb{R}^{(K+1)\times h}$ with corresponding eigenvectors where $K+1$ is the number of elements in $\boldsymbol{v}^{(i,j|D)}$ and $h = K+1-4$ is the rank of P. Extending $U^{\sim}$ to an orthogonal basis by $U^{\perp}$ gives $\boldsymbol{v}^{(i,j|D)^{\sim}} = U^{\sim T}\boldsymbol{v}^{(i,j|D)}$ with the a priori assumption $\boldsymbol{v}^{(i,j|D)^{\sim}} \sim N(0, \lambda^{-1}\Lambda^{\sim -1})$ and with $U = (U^{\sim}, U^{\perp})$ as orthogonal basis, we get $\boldsymbol{v}^{(i,j|D)^{\perp}} = U^{\perp T}\boldsymbol{v}^{(i,j|D)}$. Conditioning on $\boldsymbol{v}^{(i,j|D)^{\sim}}$, we have x being distributed according to (5.6) and with (5.21) we get the mixed model log likelihood

$$l^{m}_{ij|D}(\lambda, \boldsymbol{v}^{(i,j|D)^{\perp}}) = \log \int |\lambda\Lambda^{\sim}|^{\frac{1}{2}} \exp\left\{ l^{p}_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda) \right\} \mathrm{d}\boldsymbol{v}^{(i,j|D)^{\sim}}. \tag{5.22}$$

The integral can be approximated by a Laplace approximation (see also Rue, Martino,

and Chopin 2009)

$$l^m_{ij|D}(\lambda, \hat{\boldsymbol{v}}^{(i,j|D)^\perp}) \approx \frac{1}{2}\log|\lambda\Lambda^\sim| + l^p_{ij|D}(\hat{\boldsymbol{v}}^{(i,j|D)}, \lambda) - \frac{1}{2}\log|U^{\sim T}H^p_{ij|D}(\hat{\boldsymbol{v}}^{(i,j|D)}, \lambda)U^\sim| \tag{5.23}$$

where $\hat{\boldsymbol{v}}^{(i,j|D)}$ denotes the penalized maximum likelihood estimate. We can now differentiate (5.23) with respect to λ which gives

$$\begin{aligned}\frac{\partial l^m_{ij|D}(\lambda, \hat{\boldsymbol{v}}^{(i,j|D)^\perp})}{\partial\lambda} &= -\frac{1}{2}\hat{\boldsymbol{v}}^{(i,j|D)^T}P\hat{\boldsymbol{v}}^{(i,j|D)} \\ &+ \frac{1}{2\lambda}\mathrm{tr}\underbrace{\left\{(U^{\sim T}H^p_{ij|D}(\hat{\boldsymbol{v}}^{(i,j|D)}, \lambda)U^\sim + \lambda\Lambda^\sim)^{-1}U^{\sim T}H^p_{ij|D}(\hat{\boldsymbol{v}}^{(i,j|D)}, \lambda=0)U^\sim\right\}}_{:=S(\lambda)}.\end{aligned} \tag{5.24}$$

We can construct the estimating equation for the difference penalty through

$$\hat{\lambda}^{-1} = \frac{\hat{\boldsymbol{v}}^{(i,j|D)^T}\mathbf{P}\hat{\boldsymbol{v}}^{(i,j|D)}}{\mathrm{tr}(S(\lambda))} \tag{5.25}$$

with $S(\lambda)$ as equivalent to a smoothing matrix. Apparently, both sides of equation (5.25) depend on λ but an iterative solution is possible by fixing λ on the right hand side in (5.25), update λ on the left hand side and iterate this step by updating the right hand side of (5.25). This estimation scheme has been suggested in generalized linear mixed models by Schall (1991), see also Searle, Casella, and McCulloch (1992). For penalized spline smoothing Wood (2011) shows that the selection of smoothing parameter λ based in the mixed model approach behaves superior compared to AIC selected values, see also Reiss and Ogden (2009).

5.2.6 Practical Settings and Specifying the Vine

To maximize the likelihood we need to specify starting values of the coefficients. We suggest to take $\boldsymbol{v}_0^{(i,j|D)}$ mirroring an independence density and set the penalty parameter $\lambda_0^{(i,j|D)}$ to a moderate size. In each step we estimate new weights $\hat{\boldsymbol{v}}^{(i,j|D)}$, keeping $\lambda^{(i,j|D)}$ fixed and then refit $\lambda^{(i,j|D)}$ using (5.25). This estimation scheme is repeated until convergence.

Most importantly now is that we need to specify the vine structure to estimate the entire copula for all variables. For D-vines this implies that the order of variables in the first tree level completely specifies the vine. The intention is therefore that the first level tree with the pairwise knots (see Figure 5.1) captures the majority of (pairwise) dependencies. We use statistical model selection, based on the pair-wise estimated corrected Akaike information criterion (cAIC) (Hurvich and Tsai 1989, see

also Burnham and Anderson 2010)

$$\mathrm{AIC}_c(\lambda) = -l^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda) + \mathrm{df}(\lambda) + \frac{2\mathrm{df}(\lambda)(\mathrm{df}(\lambda)+1)}{n - \mathrm{df}(\lambda) - 1} \tag{5.26}$$

with $\mathrm{df}(\lambda)$ is the degree of the model defined through

$$\mathrm{df}(\lambda) = \mathrm{tr}\left[\left\{\mathbf{H}^p_{ij|D}(\boldsymbol{v}^{(i,j|D)}, \lambda)\right\}^{-1} \mathbf{H}^p_{ij|D}\left(\boldsymbol{v}^{(i,j|D)}, \lambda = 0\right)\right]. \tag{5.27}$$

to select the order of the D-vine. Beginning in the top tree level of a D-vine, we calculate all $\binom{p}{2}$ marginal pairwise copulas fitted by penalized splines. For each pair (i, j) this gives the fitted maximized likelihood value $l_{ij}(\hat{\mathbf{v}}^{(i,j)})$ with $\hat{\mathbf{v}}^{(i,j)}$ as penalized estimate resulting from (5.19) and penalty parameter selected data driven as discussed above. Note that $l_{ij}(\hat{\mathbf{v}}^{(i,j)}) \geq 0$, where $l_{ij}(\hat{\mathbf{v}}^{(i,j)}) = 0$ indicates independence amongst the variable pair (i, j). We order the variable pairs, subject to their increasing estimated pairwise AIC_c and start with the pair of covariates with lowest estimated AIC_c. We now select the pairs of variables such that the resulting selection gives a tree, as sketched in Figure 5.1 on the first level. The problem of finding this selection is equivalent to solve a traveler salesman problem (see Applegate 2006) by interpreting the AIC_c as distance measure between two variables (see Brechmann 2010). Once this problem is solved, the specification of the first tree level completely defines the D-vine.

The complexity of D-vines increases exponentially with an increasing number of variables and it seems advisable to simplify, that is truncate a D-vine. We therefore suggest to truncate the vine by using the independence copula for higher order tree levels of the vine. Brechmann, Czado, and Aas (2012) suggest an equivalent principle of truncation, based on changes of Information Criteria like AIC or BIC between levels. In our approach an independent copula is indicated if the estimated penalty parameter λ tends to infinity for this copula, so the penalty dominates the estimation. In fact, penalizing first order differences of $\mathbf{v}^{(i,j)}$ results for $\lambda \to \infty$ exactly in an independence copula density. This indicates the level of truncation.

In (5.18), we penalizes second order derivatives of Bernstein polynomials of each margin and accordingly we achieve a quadratic fit at each margin. If $l_{ij}(\hat{\mathbf{v}}^{(i,j)}) \to 0$ and $\lambda \to \infty$, an independent copula is reached and the $\mathrm{AIC}_c \to 4$. Due to numerical difficulties to calculate an accurate equal distribution of the coefficients $\hat{\mathbf{v}}^{(i,j)}$ in this case, we calculate the present AIC_c and replace $\hat{\mathbf{v}}^{(i,j)}$ with equal weighted coefficients, if λ increases monotonously and the present AIC_c is greater than 4. If all pair-copulas in a level of the tree are estimated with nearly equal weighted coefficients, all missing pair-copulas in higher levels are independent copulas. This indicates the level of truncation for the

penalty of integrated squared second order derivatives.
The entire routine is presented in an R-package `penDvine`, which will be available on the `CRAN` server soon.

5.3 Simulations and Examples

5.3.1 Simulations

In order to demonstrate the performance of our approach, we run some simulations of our approach. We simulate data from a a) Frank copula, b) Clayton copula and c) t-copula with $df = 3$, each with Kendall's τ set to $\tau = 0.25$ and $\tau = 0.5$. As sample size we take of size $N = 100$ and $N = 500$, respectively and the simulations size is $n = 100$. This gives 12 simulation scenarios (3 different copulas, 2 values for τ, 2 sample sizes). As basis dimension we work with $K = 14$. The simulated data are fit with three different spline settings. First, we use Bernstein polynomials, penalizing second order differences of the coefficients. Second, we use Bernstein polynomials, but penalize second order derivatives as in (5.18). The third estimation is done with B-splines, penalizing second order differences of the spline coefficients. As benchmark, we also calculate the AIC_c value for the true copula from which we simulated the data but with their parameter replaced by its Maximum Likelihood fitted value, as implemented in `R` using the `copula` package.
Table 5.3 reports the results for a bivariate simulation. Up to exceptions, the B-spline approach using the second order penalty results with minimal AIC_c, closely followed by the Bernstein polynomials with penalized second order difference. In the scenarios of Kendell's tau $\tau = 0.25$ and $N = 500$, the Bernstein polynomials with penalized second order difference behave better than the B-spline approach. Often, the Bernstein polynomials with integral penalty yield the poorest fit, especially for $N = 500$.
We extend the previous setup and sample four-dimensional data using the same simulation scenarios from above. For comparison and somewhat as competition to our routine we use the function `CDVineCopSelect` from the R-package `CDVine` (see Schepsmeier and Brechmann 2011) to estimate a D-vine. `CDVineCopSelect` thereby fits a D-vine copula model, selecting appropriate copula families estimating bivariate copula in each node using maximum likelihood estimation. The program calculates the corresponding AIC for all available copula families in the R-package, e.g. Gaussian, Student t-copula, Clayton, Frank, Gumbel or Joe. A complete list of supported copula families by `CDVineCopSelect` is given in Table 5.2. Finally the family with the minimum value is chosen in each node sequentially. We report the AIC_c value of the `CDVine` package but stress, that the degree of freedom is not calculated appropriately, since it omits the

selection of the copula family. We do not emphasize this point too much. The results are presented in Table 5.4. Like in Table 5.3, the smallest AIC_c value is selected by the `CDVine` package, which is not surprising since we are simulating from implemented copulas, that is the true copula is within the list of fitted copulas.
Throughout the whole simulation study (see Table 5.4), the Bernstein polynomials penalized with second order differences behave not optimal. Like above, the linear B-splines results with the best performance amongst the spline fitted copulas.

5.3.2 Examples

As first practical example we investigate the maximum daily wind-speed in Germany, measured at 12 locations distributed over Germany: a) BRE: Bremen, b) MS-OS: Münster-Osnabrück, c) LEI: Leipzig-Halle, d) BER: Berlin, e) ARK: Arkona, f) CUX: Cuxhaven, g) KAS: Kassel, h) FRA: Frankfurt, i) MUC: München, j) KEM: Kempten, k) FEL: Feldberg and l) KOE: Köln-Bonn from 1st January 2000 to 31st December 2011 and the dataset consists of $n = 4139$ observations. We estimate a D-vine, using our approach with $K = 12$ for the cases i) Bernstein polynomials penalizing second order differences, ii) Bernstein polynomials penalizing squared integral of second order derivatives iii) B-splines penalizing second order differences and as competitor iv) the routine `CDVineCopSelect` from the R-package `CDVine`. The results are reported in Table 5.1 (left). Our approach with B-splines penalizing second order differences results with lowest AIC_c and with the highest log-likelihood. We observe the optimal D-vine with minimal AIC_c for the B-spline approach, presented in Figure 5.2. Three estimated pair-copulas, marked in Figure 5.2 with a red triangle, are exemplary visualized in Figure 5.3. Interestingly, the conditional copula density in Figure 5.2 (bottom) indicates less dependence between the maximal windspeed in Leipzig-Halle and Arkona, given the maximal windspeed measured in Berlin. These results indicates a better performance using our semi-parametric approach compared with `CDVineCopSelect` from the R-package `CDVine`, which selects only one copula family as the optimal one.
In the second example, we consider the daily sunshine duration in Germany, measured at the same 12 locations as in the first example. Again, the data are measured from 1st January 2000 to 31st December 2011 and the dataset consists of $n = 4139$ observations. We estimate a D-vine, using the same approaches as in the first example and report the results in Table 5.1 (right). The approach with B-splines penalizing second order differences results with lowest AIC_c and with the highest log-likelihood. The fitted D-vine is presented in Figure 5.4 and behaves optimally compared to the model selected by `CDVineCopSelect`.

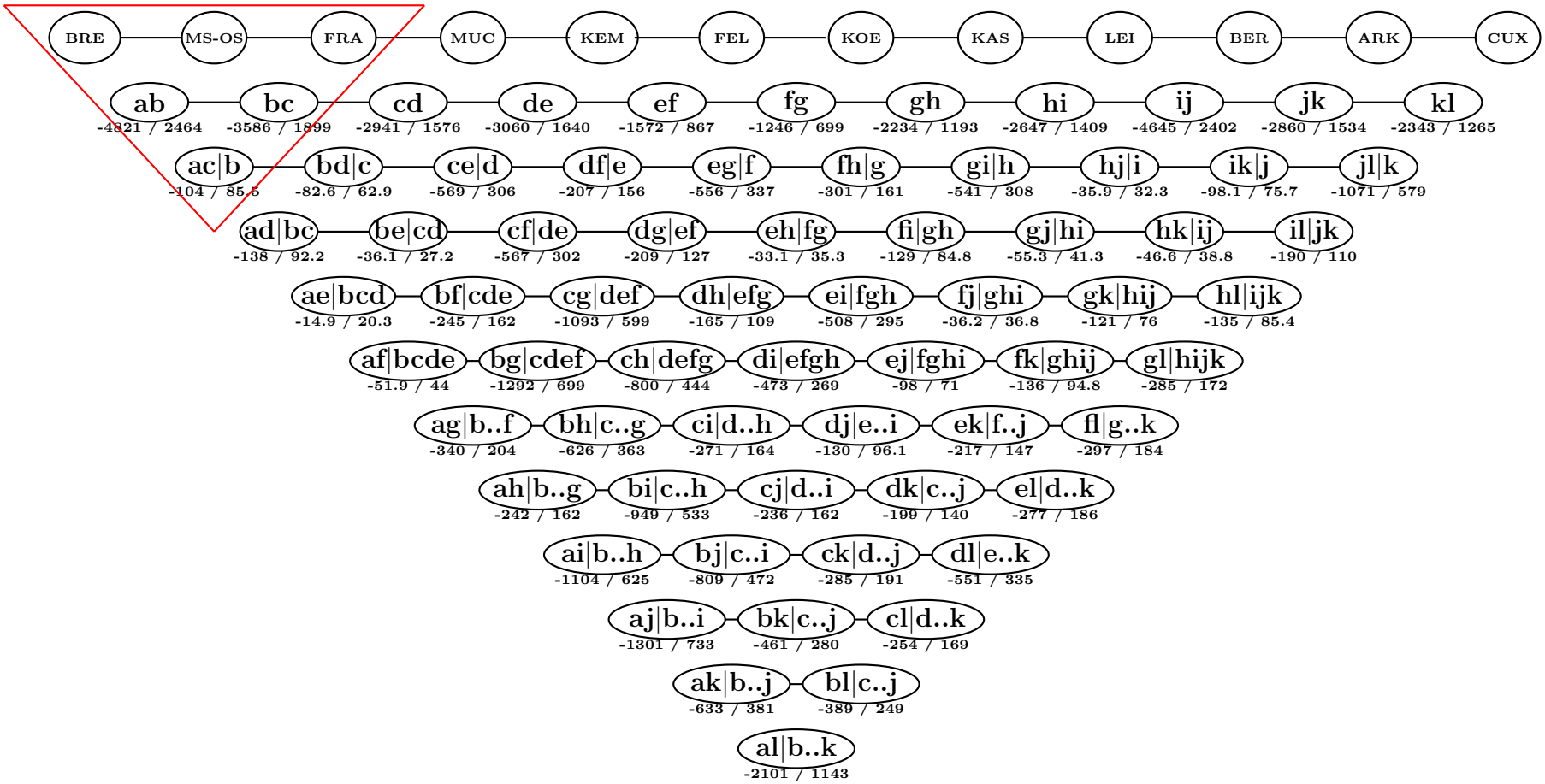


Figure 5.2: Fitted D-Vine for the wind data with $K = 12$ and B-splines, penalizing second order differences with a) BRE=Bremen, b) MS-OS Münster-Osnabrück, c) FRA: Frankfurt, d) MUC: München, e) KEM: Kempten, f) FEL: Feldberg, g) KOE: Köln-Bonn, h) KAS: Kassel, i) LEI: Leipzig-Halle, j) BER: Berlin, k) ARK: Arkona and l) CUX: Cuxhaven. Reported are AIC_c / log-likelihood.

	wind data		sun data	
approach	AIC_c	log-likelih.	AIC_c	log-likelih.
i) Bernstein polyn., Difference pen.	-45032.65	23753.08	-67789.69	35736.81
ii) Bernstein polyn., Derivative pen	-44582.42	23950.99	-68098.80	36462.65
iii) B-splines, Difference pen.	-54050.01	30006.22	-93007.73	51597.96
iv) CDVineCopSelect	-48958.39	24590.20	-74902.65	37573.33

Table 5.1: Example of wind and sun data: reported is corrected Akaike Information Criterion (AIC_c) and the log-likelihood for i) our approach with Bernstein polynomials, penalizing second order differences, ii) our approach with Bernstein polynomials, penalizing squared integral of second order derivatives, iii) our approach with B-splines, penalizing second order differences and iv) CDVineCopSelect.

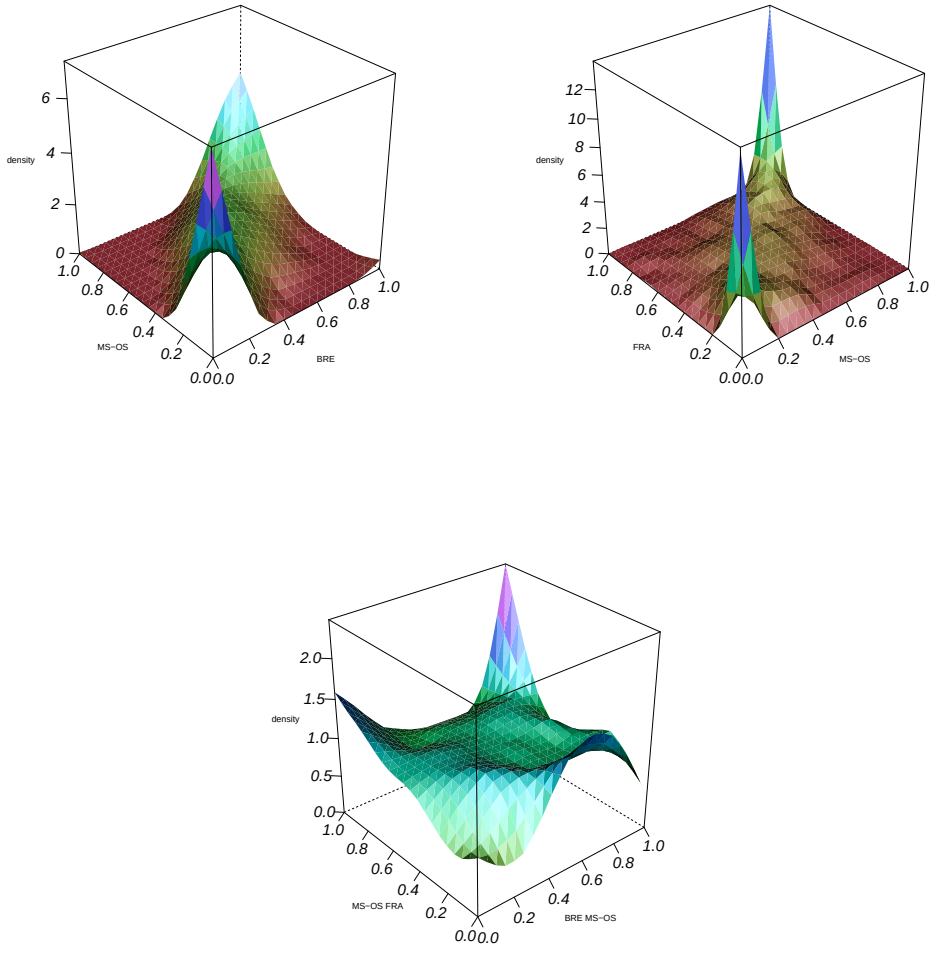


Figure 5.3: Copula density of Bremen and Münster (top left), copula density of Münster and Frankfurt (top right) and the conditional copula density of Bremen and Frankfurt, given Münster (bottom).

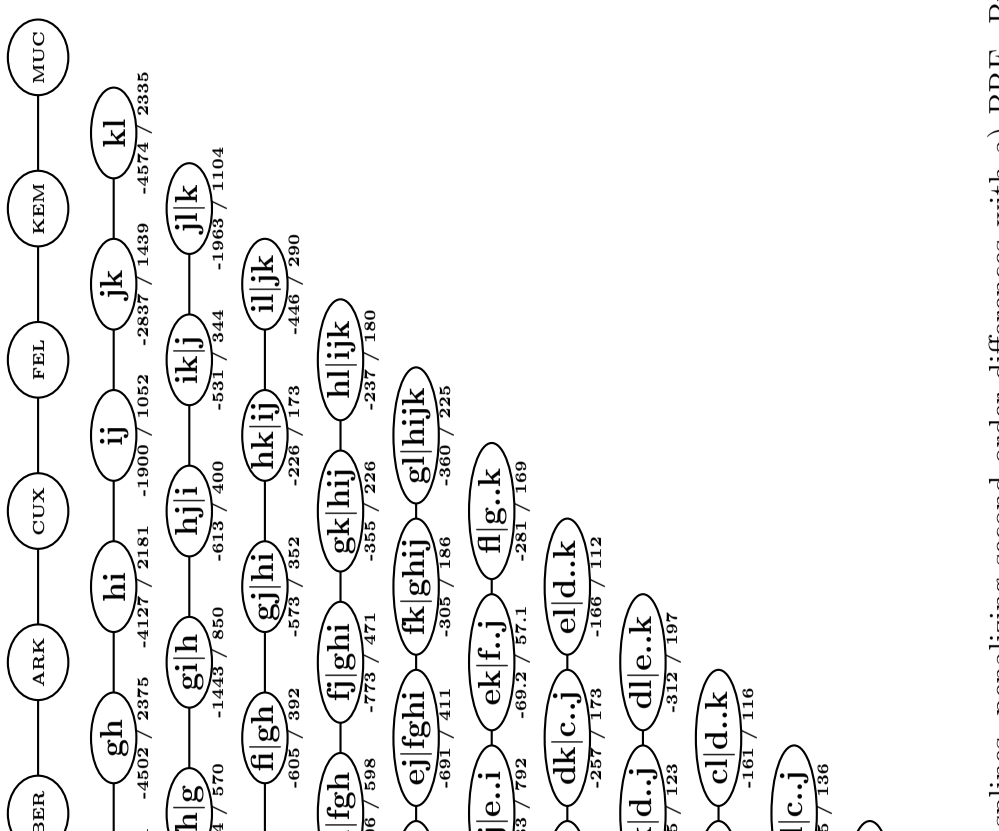


Figure 5.4: Fitted D-Vine for the sum data with $K = 12$ and B-splines, penalizing second order differences with a) BRE=Bremen, b) MS-OS Münster-Osnabrück, c) KOE: Köln-Bonn, d) FRA: Frankfurt, e) KAS: Kassel, f) LEI: Leipzig-Halle, g) BER: Berlin, h) ARK: Arkona, i) CUX: Cuxhaven, j) FEL: Feldberg, k) KEM: Kempten and l) MUC: München. Reported are AIC_c / log-likelihood.

5.4 Discussion

In this chapter we propose how to fit D-vines with penalized Bernstein polynomials or penalized B-splines respectively, estimating pair-copulas in each knot of the D-vine. Our approach thereby accommodates side constraints like uniform univariate margins so that the fitted density in each knot of the D-vine is a copula density itself. We consider two different established penalty approaches, which work both well. Probably there exist more efficient methods, but this is not the focus of this chapter. Generally, we can estimate a D-vine without any defaults to the entire distribution functions of the pair-copulas. Each estimation procedure for a pair-copula requires only a low computational demand and the computational time for the whole D-vine can be reduced using parallel computing approaches. Furthermore we do not need to test at each knot whether the pair-copula is from any known copula family. Our routine behaves acceptably in the sense of the corrected Akaike information criterion. The results in Section 3 exhibit the applicability of our approach.

code number	type
0	independence copula
1	Gaussian copula
2	Student t copula (t-copula)
3	Clayton copula
4	Gumbel copula
5	Frank copula
6	Joe copula
7	BB1 copula
8	BB6 copula
9	BB7 copula
10	BB8 copula
13	rotated Clayton copula (180 degrees; "survival Clayton")
14	rotated Gumbel copula (180 degrees; "survival Gumbel")
16	rotated Joe copula (180 degrees; "survival Joe")
17	rotated BB1 copula (180 degrees; "survival BB1")
18	rotated BB6 copula (180 degrees; "survival BB6")
19	rotated BB7 copula (180 degrees; "survival BB7")
20	rotated BB8 copula (180 degrees; "survival BB8")
23	rotated Clayton copula (90 degrees)
24	rotated Gumbel copula (90 degrees)
26	rotated Joe copula (90 degrees)
27	rotated BB1 copula (90 degrees)
28	rotated BB6 copula (90 degrees)
29	rotated BB7 copula (90 degrees)
30	rotated BB8 copula (90 degrees)
33	rotated Clayton copula (270 degrees)
34	rotated Gumbel copula (270 degrees)
36	rotated Joe copula (270 degrees)
37	rotated BB1 copula (270 degrees)
38	rotated BB6 copula (270 degrees)
39	rotated BB7 copula (270 degrees)
40	rotated BB8 copula (270 degrees)

Table 5.2: Codes for copula families in CDVineCopSelect.

	Example	Bernstein Difference Penalty	Bernstein Derivative penalty	B-spline Difference penalty	true
a)	Clayton, $N = 100, \tau = 0.25$	-6.63 (7.86) / 9.21 (4.81)	-6.03 (7.89) / 7.73 (4.60)	**-6.53 (7.79) / 9.21 (5.15)**	-20.73 (10.30) / 11.39 (5.15)
	Clayton, $N = 500, \tau = 0.25$	**-74.28 (18.44) / 47.88 (10.27)**	-73.18 (18.89) / 47.72 (11.74)	-72.63 (18.68) / 48.84 (11.56)	-107.62 (23.38) / 54.82 (11.69)
	Clayton, $N = 100, \tau = 0.5$	-48.32 (12.72) / 33.43 (6.88)	-49.37 (14.27) / 34.42 (8.61)	**-51.99 (14.02) / 38.43 (8.10)**	-84.33 (18.15) / 43.18 (9.07)
	Clayton, $N = 500, \tau = 0.5$	-325.03 (29.48) / 180.40 (15.28)	-305.54 (57.73) / 172.39 (35.02)	**-340.69 (32.13) / 200.27 (17.78)**	-430.25 (39.45) / 216.13 (19.73)
b)	Frank, $N = 100, \tau = 0.25$	-6.07 (6.68) / 7.89 (3.78)	-6.11 (6.65) / 7.45 (3.58)	**-6.11 (6.59) / 7.75 (3.90)**	-13.79 (7.68) / 7.92 (3.84)
	Frank, $N = 500, \tau = 0.25$	-62.67 (16.21) / 37.63 (8.76)	-62.50 (16.10) / 36.35 (8.68)	**-62.71 (16.13) / 37.12 (9.31)**	-73.15 (16.61) / 37.58 (8.31)
	Frank, $N = 100, \tau = 0.5$	-45.76 (11.69) / 31.50 (6.35)	-45.88 (11.99) / 31.42 (7.13)	**-48.37 (12.95) / 35.01 (7.52)**	-62.02 (13.91) / 32.03 (6.95)
	Frank, $N = 500, \tau = 0.5$	-292.54 (30.53) / 161.20 (15.71)	-287.34 (38.76) / 160.48 (22.09)	**-298.38 (31.69) / 169.87 (16.74)**	-318.07 (30.96) / 160.04 (15.48)
c)	t-copula, $df = 3, N = 100, \tau = 0.25$	**-5.78 (7.33) / 10.02 (4.96)**	-4.41 (7.91) / 7.42 (5.42)	-5.32 (7.30) / 9.89 (5.85)	-21.84 (11.11) / 12.98 (5.56)
	t-copula, $df = 3, N = 500, \tau = 0.25$	**-75.21 (19.85) / 50.87 (10.80)**	-73.91 (20.13) / 52.22 (11.98)	-72.81 (21.33) / 53.10 (13.48)	-118.93 (24.90) / 61.48 (12.45)
	t-copula, $df = 3, N = 100, \tau = 0.5$	-45.56 (13.16) / 32.73 (7.09)	-45.77 (14.12) / 33.26 (8.70)	**-48.94 (14.30) / 37.80 (8.52)**	-76.25 (18.39) / 40.19 (9.19)
	t-copula, $df = 3, N = 500, \tau = 0.5$	-308.29 (34.72) / 173.79 (18.11)	-295.85 (45.55) / 171.02 (27.46)	**-316.48 (37.93) / 189.67 (21.34)**	-391.03 (41.61) / 197.53 (20.81)

Table 5.3: Bivariate examples: reported is the mean of the corrected Akaike Information Criterion (AIC_c) / log-likelihood for $K = 14$. The bracketed terms give the standard deviations.

	Example	Bernstein Difference Penalty	Bernstein Derivative penalty	B-spline Difference penalty	CDVine
a)	Clayton, $N = 100, \tau = 0.25$	-19.40 (17.70) / 43.73 (10.84)	-18.73 (18.21) / 37.77 (10.77)	**-20.71 (18.17) / 44.49 (12.27)**	-93.75 (24.04) / 53.00 (12.10)
	Clayton, $N = 500, \tau = 0.25$	**-307.27 (52.08) / 209.21 (29.69)**	-304.04 (52.30) / 206.76 (32.26)	-307.10 (53.59) / 216.68 (33.42)	-463.07 (62.88) / 237.99 (31.49)
	Clayton, $N = 100, \tau = 0.5$	-168.89 (33.82) / 131.59 (18.90)	-169.02 (34.08) / 129.31 (20.35)	**-183.98 (38.95) / 149.55 (24.34)**	-307.41 (44.05) / 160.21 (22.13)
	Clayton, $N = 500, \tau = 0.5$	-1214.99 (80.07) / 695.25 (41.51)	-1159.33 (112.29) / 675.98 (66.99)	**-1278.77 (91.37) / 772.30 (51.57)**	-1582.18 (103.46) / 797.63 (51.78)
b)	Frank, $N = 100, \tau = 0.25$	-14.70 (14.80) / 36.94 (8.72)	-15.01 (14.69) / 33.95 (7.72)	**-16.54 (15.26) / 36.36 (9.07)**	-65.00 (17.22) / 39.17 (8.74)
	Frank, $N = 500, \tau = 0.25$	-254.76 (36.45) / 163.85 (19.89)	-255.94 (36.23) / 158.31 (19.68)	**-261.25 (36.91) / 162.71 (20.78)**	-317.39 (40.45) / 166.64 (20.27)
	Frank, $N = 100, \tau = 0.5$	-145.56 (23.87) / 115.25 (12.78)	-146.81 (24.11) / 113.25 (13.50)	**-155.82 (25.94) / 125.69 (14.38)**	-225.89 (26.94) / 119.99 (13.42)
	Frank, $N = 500, \tau = 0.5$	-1053.63 (66.26) / 597.84 (34.02)	-1032.8 (76.76) / 591.00 (43.95)	**-1087.13 (72.52) / 631.86 (38.39)**	-1190.02 (68.54) / 602.81 (34.28)
c)	t-copula, $df = 3, N = 100, \tau = 0.25$	-10.73 (16.50) / 42.99 (11.46)	-6.83 (16.20) / 32.49 (10.43)	**-11.48 (16.91) / 42.64 (12.93)**	-95.16 (23.99) / 57.13 (12.11)
	t-copula, $df = 3, N = 500, \tau = 0.25$	-331.25 (48.19) / 236.69 (27.33)	-322.94 (48.22) / 237.67 (30.16)	**-332.10 (50.27) / 254.06 (31.77)**	-525.57 (59.94) / 274.68 (29.98)
	t-copula, $df = 3, N = 100, \tau = 0.5$	-157.62 (32.49) / 129.94 (18.44)	-155.80 (33.53) / 125.13 (20.44)	**-168.77 (37.25) / 144.22 (23.91)**	-282.79 (41.10) / 151.48 (20.68)
	t-copula, $df = 3, N = 500, \tau = 0.5$	-1166.30 (84.03) / 679.96 (44.15)	-1140.43 (88.81) / 683.87 (51.11)	**-1206.35 (91.51) / 744.82 (51.27)**	-1474.85 (96.26) / 749.41 (48.13)

Table 5.4: Fourdimensional examples: reported is the mean of the corrected Akaike Information Criterion (AIC_c) / log-likelihood for $K = 14$. The bracketed terms give the standard deviations.

6 Extension

This chapter presents an extension of the considered approaches combining the concepts of univariate penalized density estimation (see Chapter 3) and penalized copula density estimation (see Chapter 4). We re-investigate the currency example presented in Chapter 4, but the univariate distributions are estimated using the approach presented in Chapter 3.

The data set includes $n = 2854$ observations of the Australian dollar (AUS), the Euro (EUR) and the Japanese yen (JAP) from January 3rd, 2000 until May 6th, 2011. Again, we analyze the log-return from day t to day $t+1$ and estimate the density of each dataset using the approach presented in Chapter 3 with $K = 20$. Then we estimate the copula density for the same values of d and D as in the example in Chapter 4. The results are presented in Table 6.1 (left) and compared to the estimated results in Chapter 4. In Chapter 4, the marginal data were separately fitted to t-distributions and the corresponding results of the copula density estimations are repeated in Table 6.1 (right).

Analyzing Table 6.1, we observe increased log-likelihood and decreased AIC_c values for each scenario, whenever the marginal data are estimated with the approach of Chapter 3. Of course, the absolute difference between corresponding values of AIC_c is not interpretable. Moreover, the AIC_c does not consider the foregoing estimations of marginal distributions. Estimating the univariate distributions with the penalized splines approach outperforms the competitor.

The contour plot of the fitted bivariate margins (left) with the minimal AIC_c and the corresponding copula density (right) are plotted in Figure 6.1 with $d = 4$ and $D = 8$. Comparing the plots in Figure 6.1 with the corresponding plots in Figure 4.6 shows remarkable differences between the estimations. First, the bivariate copula densities in Figure 6.1 (right) look smoother then in Figure 4.6 (right), probably due to the univariate penalized estimation. Second, the contour plots show different marginal distributions. The contour plots of the copula distribution of EUR and JAP in Figure 6.1 (left, mid) show an agglomeration at the margins, where at least one of both values is close to 1. That behaviour was not observed in Figure 4.6 (left, mid). Of course, these facts indicate a different copula density, see Figure 6.1 (right, mid) and Figure

		exchange rate data Chapter 4			
		`pendensity`		t-distribution	
d	D	log-likelihood $\hat{l}$	AIC_c	log-likelihood $\hat{l}$	AIC_c
3	3	996.088	-1856.782	873.980	-1610.068
3	6	1046.201	-1959.769	1007.578	-1725.735
4	4	1088.495	-1968.252	978.359	-1707.725
4	8	1121.137	-2029.449	1117.326	-1774.491
Clayton		167.242	-332.483	83.410	-164.819
Frank		85.862	-169.722	2.707	-3.412
Gumbel		70.530	-139.059	31.649	-61.296
Normal		105.978	-209.955	27.654	-53.307
Bernstein		977.908	-1705.816	886.640	-1523.279

Table 6.1: Results for various combinations of d and D for exchange rate data example in Chapter 4 using (left) `pendensity` from Chapter 3 for the marginal distribution and (right) repeated results using marginal t-distribution (see Chapter 4).

4.6 (right, mid). Also the comparison of the contour plots of AUS and JAP in Figure 6.1 (right, bottom) and Figure 4.6 (right, bottom) indicate differences, which are also visible in different copula densities for both time series, see Figure 6.1 (left, bottom) and Figure 4.6 (left, bottom).

Using this combination of penalized splines approaches is an appealing new extension of the ideas presented in the preceding chapters of this thesis. Further research may tackle this combination in detail.

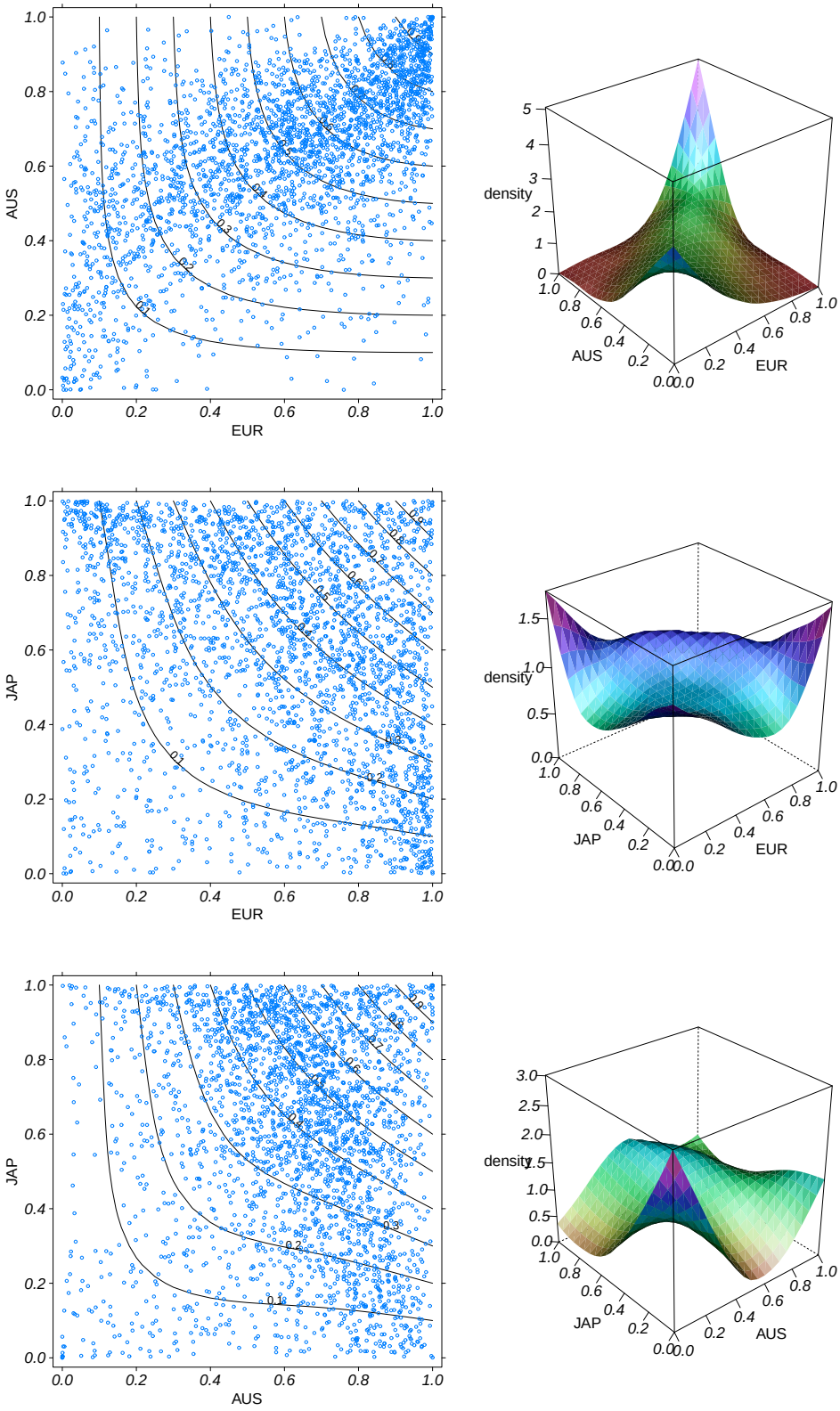


Figure 6.1: Bivariate marginal copula distribution (left) and copula density (right) between Euro (EUR), Australian Dollar (AUS) and Japanese Yen (JAP) compared to the US-dollar from January 3rd, 2000 until May 6th, 2011 with $d = 4$ and $D = 8$ using `pendensity` from Chapter 3 for estimating the marginal distribution.

7 Summary

This thesis discussed applications of penalized smoothing splines for univariate density and copula density estimation. We presented different types of basis functions, preferring the B-spline bases. To get smooth density fits, we penalized huge differences of neighboring basis coefficients, both in the univariate and multivariate cases. The link between P-splines and linear mixed models was used for iterative estimation of the optimal smoothing parameter λ. The application of quadratic programming, also in combination with sparse grids worked satisfactorily for the estimation of (high-dimensional) copula densities. In the context of dependence vines, Bernstein polynomials were investigated as spline basis, but the usage of different penalties did not yield optimal results. The fits using penalized B-spline outperformed the other approaches. As theoretical starting point, Chapter 2 discussed the substantial theory for applications of the following chapters. Chapter 3 presented the univariate density estimation approach with penalized smoothing splines and theoretical results of the estimator were presented. First, the estimator had minimal Kullback-Leibler distance to the unknown density and secondly, we showed asymptotic normality of estimated coefficients. We calculated the integrated mean squared error (IMSE) for several density scenarios in simulations studies. The corresponding results were satisfactory for our density estimation approach, which performed usually best. The extension to a covariate dependent density estimation approach allowed for tests of equality of grouped densities. This test is powerful, especially when the standard tests did not announce inequality of the groups. We implemented this approach in the `R` package `pendensity`, available on CRAN.

The presented copula density estimator in Chapter 4 was constructed using sparse grids based on linear B-spline functions to circumvent the curse of dimensionality. Furthermore, quadratic programming was used for simultaneous estimation of marginal and joint copula densities. Accordingly, we penalized differences of the basis coefficients in this context, but the penalty parameter λ was determined by a grid search, such that λ minimized AIC_c. This penalized copula density estimation approach allowed for estimation in up to five or even six dimensions. Moreover, calculated AIC_c values in the simulation studies for samples of various copula families presented better results of the copula density approach using penalized B-splines compared to kernel density

estimation and Bernstein polynomials. Additionally, the approach allowed for an analysis of bivariate dependence in the context of high dimensional copula densities. These marginal copula densities were presented in the examples of Chapters 4 and 6. The entire estimation concept was implemented in the `R` package `pencopula`, available on CRAN.

For the estimation of dependence vines, discussed in Chapter 5, we used a modified idea of the copula density estimation approach from Chapter 4 in the bivariate case. Throughout this chapter, the pair-copula construction principle was considered, especially in the case of D-vines. The estimation of D-vines was done by estimation of pair-copulas using penalized splines in each node of the dependence tree. We additionally considered penalized Bernstein polynomials as possible basis functions, but they did not outperform penalized B-splines. We presented ideas for ordering the first level of the D-vine based on AIC_c values, which determined the structure of the complete D-vine. Furthermore, we presented concepts to truncate the D-vine at a given level in the case that only independent pair-copulas were estimated. The simulation studies showed comparable results with respect to AIC_c for the penalized spline approach to the true copula density. But the examples of wind and sun data showed powerful results in contrast to the established parametric estimation approaches. This approach of flexible pair-copula estimation will be available on CRAN in the package `penDvine` soon. Further perspectives consider further dependence vines, e.g. C-Vines, which follow a different decomposition of the joint density. Probably, results of estimated C-Vines can be comparably good as in the case of D-vines.

Finally, the usage of penalized smoothing splines resulted in comparable or rather better models for univariate and copula densities compared to established parametric or non-parametric estimators. Moreover, the combination of the penalized univariate density estimator and the penalized copula density estimator in Chapter 6 provided an increased performance.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance Mathematics and Economics 44*(2), 182–198.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions of Automatic 19*(6), 716–723.

Applegate, D. L. (2006). *The traveling salesman problem*. Princeton series in applied mathematics. Princeton Univ. Press.

Autin, F., E. L. Pennec, and K. Tribouley (2010). Thresholding methods to estimate copula density. *Journal of Multivariate Analysis 101*(1), 200 – 222.

Babu, G. J., A. J. Canty, and Y. P. Chaubey (2002). Application of bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference 105*(2), 377 – 392.

Bedford, T. and R. Cooke (2001). Probability density decomposition for conditionally dependent random variables modeled by Vines. *Annals of Mathematics and Artificial Intelligence 1*(32), 245–268.

Bedford, T. and R. M. Cooke (2002). Vines: A new graphical model for dependent random variables. *The Annals of Statistics 30*(4), 1031–1068.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Bogaerts, K. and E. Lesaffre (2008). Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in medicine 27*(30), 6379–6392.

Boneva, L. I., D. Kendall, and I. Stefanov (1971). Spline transformations: Three new diagnostic aids for the statistical data- analyst. *Journal of the Royal Statistical Society. Series B 33*(1), 1–71.

Bouezmarni, T., J. Rombouts, and A. Taamouti (2010). Asymptotic properties of the bernstein density copula estimator for [alpha]-mixing data. *Journal of Multivariate Analysis, Elsevier 101*(1), 1–10.

Brechmann, E., C. Czado, and K. Aas (2012). Truncated regular vines in high dimensions with applications to financial data. *Canadian Journal of Statistics 40*(1), 68–85.

Brechmann, E. C. (2010). Truncated and simplified regular vines and their applications. diploma thesis. Master's thesis, Technische Universität München.

Bungartz, H.-J. and M. Griebel (2004). Sparse grids. *Acta Numerica 13*, 147–269.

Burnham, K. and D. R. Anderson (2010). *Model selection and multimodel inference - A practical Information-Theoretic Approach*. Springer, Berlin Heidelberg.

Butterfield, K. (1976). The computation of all the derivatives of a b-spline basis. *IMA Journal of Applied Mathematics 17*(1), 15–25.

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification 13*, 195–212. 10.1007/BF01246098.

Chen, S. X. and T.-M. Huang (2007). Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics 35*(2), 265–282.

Choros, B., R. Ibragimov, and E. Permiakova (2010). Copula estimation. In P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik (Eds.), *Copula Theory and Its Applications*, Volume 198 of *Lecture Notes in Statistics*, pp. 77–91. Springer Berlin Heidelberg.

Claeskens, G., T. Krivobokova, and J. Opsomer (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*(3), 529–544.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*(1), pp. 141–151.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377–403.

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger, P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik (Eds.), *Copula Theory and Its Applications*, Volume 198 of *Lecture Notes in Statistics*, pp. 93–109. Springer, Berlin Heidelberg.

Danaher, P. J. and M. S. Smith (2011). Modeling multivaraite distributions using copulas: applications in marketing. *Marketing Science 30*(1), 4–21.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

Dias, R. (1998). Density estimation via hybrid splines. *Journal of Statistical Computation and Simulation 60*(4), 277–293.

Doha, E. H., A. H. Bhrawy, and M. A. Saker (2011). On the Derivatives of Bernstein Polynomials: An Application for the Solution of High Even-Order Differential Equations. *Boundary Value Problems Volume 2011*.

Duin, R. (1976). On the choice of smoothing parameters for parzan estimators of probability density functions. *IEEE Transactions on Computing C-25*, 1175–1179.

Durante, F. and C. Sempi (2010). Copula theory: An introduction. In P. Jaworski, F. Durante, W. K. Haerdle, and T. Rychlik (Eds.), *Copula Theory and Its Applications*, Volume 198 of *Lecture Notes in Statistics*, pp. 3–31. Springer Berlin Heidelberg.

Efron, B. (2001). Selection criteria for scatterplot smoothers. *Ann. Statist. 29*, 470–504.

Efron, B. and R. Tibshirani (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics 24*(6), 2431–2461.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

Eilers, P. H. C. and B. D. Marx (2010). Splines, knots and penalties. *WIREs Computational Statistics 2*(6), 637–653.

Embrechts, P. (2009). Copulas: A personal view. *Journal of Risk and Insurance 76*(3), 639–650.

Epanechnikov, V. (1969). Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl. 14*, 156–161.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: A bayesian perspective. *Statistica Sinica 14*, 731–761.

Fahrmeir, L., T. Kneib, and S. Lang (2007). *Regression*. Berlin [u.a.]: Springer.

Fermanian, J.-D., D. Radulovic, and M. Wegkamp (2004). Weak convergence of empirical copula processes. *Bernoulli 10*(5), 847–860.

Fermanian, J.-D. and O. Scaillet (2003). Nonparametric estimation of copulas for time series. *Journal of Risk 5*(4).

Forsey, D. R. and R. H. Bartels (1988). Hierarchical B-spline refinement. In *SIGGRAPH '88: Proceedings of the 15th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, pp. 205–212. ACM.

Forsey, D. R. and R. H. Bartels (1995). Surface fitting with hierarchical splines. *ACM Trans. Graph. 14*(2), 134–161.

Frahm, G., M. Junker, and A. Szimayer (2003). Elliptical copulas: Applicability and limitations. *Statistics and Probability Letters* (63), 275–286.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Frank, M. (1979). On the simultaneous associativity of f(x,y) and x+y-f(x,y). *Aequationes Mathematicae 19*, 194–226. 10.1007/BF02189866.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont donnés. *Annales de l'Université de Lyon 14*(3), 53–77.

Garcke, J. (2006). Sprase grid tutorial. Technical report, Centre for mathematics and its applications, Mathematical Sciences Institute, Australien National University, Canberra.

Genest, C., K. Ghoudi, and L.-P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika 82*(3), 543–552.

Genest, C., E. Masiello, and K. Tribouley (2009). Estimating copula densities through wavelets. *Insurance: Mathematics and Economics 44*(2), 170–181.

Ghidey, W., E. Lesaffre, and P. H. C. Eilers (2004). Smooth random effects distribution in a linear mixed model. *Biometrics 60*(4), 945–953.

Gijbels, I. and J. Mielniczuk (1990). Estimation the density of a copula function. *Communication in Statistics: Theory an Methods 19*(2), 445–464.

Good, I. J. and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika 58*(2), 255–277.

Green, D. J. and B. W. Silverman (1994). *Nonparametric Regression and generalized linear models*. Chapman & Hall.

Green, J. (1987). Penalized likelihood for general semiparametric regression models. *International Statistical Review 55*, 245–259.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association 88*(422), 495–504.

Gu, C. (2009). *gss: General Smoothing Splines*. R package version 1.0-5.

Gu, C. and J. Wang (2003). Penalized likelihood density estimation: direct cross-validation and scalable approximation. *Statistica Sinica 13*(3), 811–826.

Gumbel, E. J. (1960). Distributions del valeurs extremes en plusieurs dimensions. *Publ. l'Inst. de Statistique, Paris 9*, 171–173.

Hall, P. and P. Patil (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *The Annals of Statistics 23*(3), pp. 905–928.

Härdle, W. and O. Okhrin (2009). De copulis non est disputandum - copulae: an overview. *AStA - Advances in Statistical Analysis 94*(1), 1–31.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5).

Henderson, C. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics 21*, 309–10.

Hoeffding, W. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin 5*, 179–233.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika 73*(3), 671–678.

Hurvich, C. M. and c.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

Jaworski, P., F. Durante, W. Härdle, and T. Rychlik (2010). *Copula Theory and Its Applications*. Lecture notes in statistics ; 198 : Proceedings. Springer.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series 28*, 120–141.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Kass, R. E. and D. Steffey (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association 84*(407), 717–726.

Kauermann, G. (2005). A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planning and Inference 127*(1–2), 53–69.

Kauermann, G., T. Krivobokova, and L. Fahrmeir (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B 71*(2), 487–503.

Kauermann, G. and J. Opsomer (2011). Data-driven selection of the spline dimension in penalized spline regression. *Biometrika 98*(1), 225–230.

Kauermann, G. and C. Schellhase (2012). Flexible Pair-Copula Estimation in D-vines with Penalized Splines. Working paper.

Kauermann, G., C. Schellhase, and D. Ruppert (2012). Flexible Copula Density Estimation with Penalized Hierarchical B-Splines. *Scandinavian Journal of Statistics*. (submitted).

Kolev, N., U. Anjos, and B. Mendes (2006). Copulas: a review and recent developments. *Stochastic Models 22*(4), 617–660.

Komárek, A. (2006). Accelerated failure time models for multivariate doubly-interval-censored data with flexible distributional assumptions. *Ph.D. thesis, Leuven: Katholieke Universiteit Leuven, Faculteit Wetenschappen*.

Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized gaussian mixture as a random-effects distribution. *Computational Statistics and Data Analysis 52*(7), 3441–3458.

Komárek, A., E. Lesaffre, and J. Hilton (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational & Graphical Statistics 14*(3), 726–745.

Koo, J.-Y. (1996). Bivariate B-splines for Tensor logspline density estimation. *Computational Statistics & Data Analysis 21*(1), 31–42.

Koo, J.-Y., C. Kooperberg, and J. Park (1999). Logspline density estimation under censoring and truncation. *Scandinavian Journal of Statistics 26*(1), pp. 87–105.

Kooperberg, C. (2009). *logspline: Logspline density estimation routines*. R package version 2.1.3.

Krivobokova, T. (2006). *Theoretical and Practical Aspects of Penalized Spline Smoothing*. Ph. D. thesis, Universität Bielefeld.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), pp. 79–86.

Kurowicka, D. and R. Cooke (2006). *Uncertainty analysis with high dimensional dependence modelling*. Chichester: Wiley.

Kurowicka, D. and H. Joe (Eds.) (2010). *Dependence modeling: Vine Copula Handbook*. World Scientific Publishing.

Lambert, P. (2007). Archimedean copula estimation using Bayesian splines smoothing techniques. *Computational Statistics & Data Analysis 51*(12), 6307–6320.

Li, J. Q. and A. R. Barron (1999). Mixture density estimation. In *In Advances in Neural Information Processing Systems 12*, pp. 279–285. MIT Press.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics*. Princeton, NJ [u.a.]: Princeton Univ. Press.

Li, Y. and D. Ruppert (2008). On the asymptotics of penalized splines. *Biometrika 95*(2), 415–436.

Lindsey, J. K. (1974a). Comparison of probability distributions. *Journal of the Royal Statistical Society. Series B 36*(1), 38–47.

Lindsey, J. K. (1974b). Construction and comparison of statistical models. *Journal of the Royal Statistical Society. Series B 36*(3), 418–425.

Liu, L., M. Levine, and Y. Zhu (2009). A functional EM algorithm for mixing density estimation via nonparametric penalized likelihood maximization. *Journal of Computational & Graphical Statistics 18*(2), 481–504.

Loader, C. (1999). *Local Regression and Likelihood*. Springer.

Lorentz, G. (1953). Bernstein polynomials. *Mathematical Expositions, no. 8, Univ. of Toronto Press, Toronto*.

Mardia, K. V. (1962). Multivariate pareto distributions. *The Annals of Mathematical Statistics 33*(3), pp. 1008–1015.

Marx, B. and P. H. C. Eilers (2005). Multidimensional penalized signal regression. *Technometrics 47*, 13–22.

McCulloch, C. and S. Searle (2001). *Generalized, Linear and Mixed Models*. New York: Wiley.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management*. Princeton University Press, Princeton Series in Finance.

McNeil, A. and J. Neslehová (2009). Multivariate Archimedean copulas, d-monotone functions and l1-norm symmetric distributions. *Ann. Stat. 37*(5B), 3059–3097.

Min, A. and C. Czado (2011). Bayesian model selection for d-vine pair-copula constructions. *Canadian Journal of Statistics 39*(2), 239–258.

Morettin, P., C. Toloi, C. Chiann, and J. C. Miranda (2010). Wavelet smoothed empirical copula estimations. *Brazilian Review of Finance 8*(3), 263–281.

Müller, P., F. Quintana, and G. Rosner (2009). Bayesian Clustering with Regression. University of Texas M.D. Anderson Cancer Center, Houston, TX 77030 U.S.A.

Nadaraya, E. (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theory of Probability and Its Applications 19*(1), 133–141.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9*(1), 141–142.

Nason, G. (2010). *wavethresh: Wavelets statistics and transforms.* R package version 4.5.

Nason, G. P. (2008). *Wavelet Methods in Statistics with R.* Springer. ISBN 978-0-387-75960-9.

Nason, G. P. and B. W. Silverman (1999). Wavelets for regression and other statistical problems. In M. G. Schimek (Ed.), *Smoothing and Regression: Approaches, Computation, and Application*, Series in Probability and Statistics. Wiley.

Nelsen, R. (2006). *An introduction to copulas* (second ed.). Berlin: Springer.

Okhrin, O., Y. Okhrin, and W. Schmid (2009). Properties of the hierarchical archimedean copulas. Technical report, Humboldt Universität zu Berlin, SFB 649 Discussion Paper 2009-014.

Omelka, M., I. Gijbels, and N. Veraverbeke (2009). Improved kernel estimation of copulas: Weak convergence and goodness-of-fit testing. *Annals of Statistics 37*(5B), 3023–3058.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science 1*(4), 502–518.

Park, B. U. and J. S. Marron (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association 85*(409), pp. 66–72.

Pearson, E. S. (1938). *Karl Pearson; an appreciation of some aspects of his life and work, by E.S. Pearson.* The University press, Cambridge [Eng.].

Pfeifer, D., D. Straßburger, and J. Philipps (2009). Modelling and simulation of dependence structures in nonlife insurance with Bernstein copulas. *Preprint*.

Prautzsch, H., W. Boehm, and M. Paluszny (2002). *Bezier and B-Spline Techniques.* Springer.

Qu, L., Y. Qian, and H. Xie (2009). Copula density estimation by total variation penalized likelihood. *Communications in Statistics - Simulation and Computation 38*(9), 1891–1908.

Qu, L. and W. Yin (2012). Copula density estimation by total variation penalized likelihood with linear equality constraints. *Computational Statistics & Data Analysis 56*(2), 384 – 398.

R Development Core Team (2011). *R: A Language and Environment for Statistical*

Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rank, J. (Ed.) (2007). *Copulas*. London: Risk Books.

Reiss, T. and R. Ogden (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society. Series B 71*(2), 505–523.

Rivlin, T. (1969). *An Introduction to the Approximation of Functions*. Blaisdell Publishing Co., Ginn and Co., Waltham.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science 6*(1), 15–32.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics 27*(3), pp. 832–837.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B 71*(2), 319–392.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Ruppert, D., M. P. Wand, and R. J. Carroll (2009). Semiparametric regression during 2003 – 2007. *Electronic Journal of Statistics 3*, 1193–1256.

Sancetta, A. and S. Satchell (2004). Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory 20*(3), 535–562.

Savu, C. and M. Trede (2010). Hierarchies of Archimedean copulas. *Quantitative Finance 10*(3), 295–304.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Schellhase, C. (2010). *pendensity: Density Estimation with a Penalized Mixture Approach*. R package version 0.2.3.

Schellhase, C. (2012). *pencopula: Flexible Copula Density Estimation with Penalized Hierarchical B-Splines*. R package version 0.3.2.

Schellhase, C. and G. Kauermann (2012). Density Estimation and Comparison with a Penalized Mixture Approach. *Computational Statistics*. (to appear).

Schepsmeier, U. and E. C. Brechmann (2011). *CDVine: Statistical inference of C- and D-vine copulas*. R package version 1.1-4.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Scott, D. W. and G. R. Terrell (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association 82*(400), pp. 1131–1146.

Searle, S., G. Casella, and C. McCulloch (1992). *Variance Components*. Wiley.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B 53*(3), 683–690.

Shen, X., Y. Zhu, and L. Song (2008). Linear B-spline copulas with applications to nonparametric estimation of copulas. *Computational Statistics & Data Analysis 52*(7), 3806–3819.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics 10*(3), 795–810.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis / B.W. Silverman*. Chapman and Hall, London ; New York :.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer Verlag.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Smith, M., A. Min, C. Almeida, and C. Czado (2010). Modeling longitudinal data using a pair-copula construction decomposition of serial dependence. *Journal of the American Statistical Association 105*, 1467–1479.

Song, P., L. Mingyao, and Y. Yuan (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics 65*, 60–68.

Stein, M. L. (1990). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *The Annals of Statistics 18*, 1139–1157.

Takahasi, K. (1965). Note on the multivariate Burr's distributions. *Annals of the Institute of Statistical Mathematics 17*, 257–260.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics 13*, 1378–1402.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics 18*(2), 223–249.

Wand, M. and M. C. Jones (1995). *Kernel smoothing*. Chapman and Hall.

Wand, M. P. and J. T. Ormerod (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian & New Zealand Journal of Statistics 50*(2), 179–198.

Watson, G. (1964). Smooth regression analysis. *Sankhya Series A 26*, 359–372.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B 73*(1), 3–36.

Wood, S. N. (2006). *Generalized additive models*. Chapman and Hall/CRC.

Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software 21*(4), 1–21.

Young, D., D. Hunter, D. Chauveau, and T. Benaglia (2009). mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software 32*(6).

Zenger, C. (1991). Sparse grids. *Wolfgang Hackbusch (Ed.): Parallel algorithms for partial differential equations, volume 31 of Notes on Numerical Fluid Mechanics, Vieweg*, 241–251.