Visual Exploration of Multivariate Data in Breast Cancer by Dimensional Reduction

Claudio Varini

Der Technischen Fakultät der Universität Bielefeld

vorgelegt zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

March 2006

Gedruckt auf alterungsbeständigem Papier ∞ ISO 9706

Acknowledgment

This work is the result of a collaboration between the Applied Neuroinformatics Group (Technical Faculty, Bielefeld University), the Condensed Matter Theory Group (Department of Physics, Bielefeld University) and the Clinical Hospital of the Munich University. The financial support was provided by the DFG graduate school *Structure Formation Processes* and the Condensed Matter Theory Group.

The people I would like to thank are:

- Juniorprof. Dr. Ing. Tim W. Nattkemper and PD Dr. Andreas Degenhard for their devotion to this work, their enthusiasm and the many fruitful discussions.
- Prof. Dr. Friederike Schmid for giving me the opportunity of working with the Condensed Matter Theory Group. Thanks also to all the member of this group for their support and help in everyday life.
- Prof. Dr. Wolf-Jürgen Beyn for giving me the opportunity of being a member of the *Structure Formation Processes* group.
- Axel Wismüller and Oliver Lange for providing the DCE-MRI data.
- Axel Saalbach and Thorsten Twellmann for their help with the software libraries.
- Christian Müller, Ning Wei and Dirk Stößel for proofreading parts of the manuscript.
- my parents, my siblings, my nephews and nieces for their continuous support and deep affection.

Publications

Parts of the results of this thesis have been published in:

- Biomedical Signal Processing and Control (2006), *Visual Exploratory Analysis of DCE-MRI Data in Breast Cancer by Dimensional Data Reduction: a Comparative Study*, C. Varini, A. Degenhard and T. W. Nattkemper, note: first version accepted.
- Neurocomputing (2006), *ISOLLE: LLE with Geodesic Distance*, C. Varini, A. Degenhard and T. W. Nattkemper, note: in press.
- Proceedings of the SPIE Medical Imaging Conference (2006), *Histological Characterization of DCE-MRI Breast Tumors with Dimensional Data Reduction*, C. Varini, A. Degenhard and T. W. Nattkemper.
- Proceedings of the Practice and Knowledge Discovery in Databases (PKDD) conference (2005), *ISOLLE: Locally Linear Embedding with Geodesic Distance*, C. Varini, A. Degenhard and T. W. Nattkemper.
- Proceedings of the Medical Imaging Understanding and Analysis (MIUA) conference (2004), *Visualisation of Breast Tumours DCE-MRI Data using LLE*, C. Varini, T. W. Nattkemper, A. Degenhard and A. Wismüller.
- Proceedings of the International Joint Conference on Neural Networks (2004), *Breast MRI Analysis by LLE*, C. Varini, T. W. Nattkemper, A. Degenhard and A. Wismüller.

List of Symbols

Some of the symbols most frequently used in this work are listed in the following. Note that the list is not exhaustive as several terms, such as some symbols used only for explaining a certain theoretical part, have been deliberately omitted.

- v: number of samples of the training set
- D: dimension of the original data space
- d: dimension of the embedded space.
- $\mathbf{x}$: input vector
- $\mathbf{y}$: output vector
- m: number of prototypes in clustering algorithms
- W: reconstruction weights in LLE
- n: number of neighbors in LLE
- $\mathbf{u}$: reference vector in SOM
- $N \times N$: dimension of the two-dimensional grid
- r_i: node in the grid linked with $\mathbf{u}_i$
- $\mathbf{m}$: orthonormal vector in PCA
- t: number of neighbors in neighborhood preservation and trustworthiness.
- Δ: regularization term in LLE.

Contents

1 Introduction

Breast cancer is the most common form of cancer in females worldwide, with a high incidence especially in Europe and North America. Albeit scientific research has contributed to reducing the mortality rate from this disease considerably in recent decades, many of its aspects such as causes and contributing factors remain unknown. Nowadays, technological progress offers new opportunities for studying the mechanisms involved with breast cancer.

However, the complexity of such mechanisms involves the monitoring of many variables and parameters, resulting in high-dimensional multivariate data, i. e. many measurements of one or more quantities are acquired from one or more sources and recorded as an ordered string of variables. Multivariate data in breast cancer comprise three categories, namely image, molecular and clinical data.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) and DNA microarrays are two novel instruments for the study of breast cancer that provide image and molecular multivariate data, respectively.

One indicator for a malignant tumor is the change in the tissue vascularity. To visualize these changes, DCE-MRI represents a powerful imaging technique for tumor detection and diagnosis that records both temporal and spatial information. The DCE-MRI technique involves the imaging of a region of interest immediately before and repeatedly after the administration of a paramagnetic contrast agent, resulting in a temporal sequence of high-resolution images or volumes. Each pixel or voxel is associated with a time-series of signal intensity values, whose kinetics provides information about the vascular structure of the tissue. In turn, the analysis of the time-series can allow the physicians not only to detect suspect lesions, but also the possibility to discriminate between benign and malignant tissue.

The genesis of cancer begins when the DNA of a normal cell changes, or "mutates", making normal cells grow uncontrollably. These changes need to be analyzed and DNA microarrays can be particularly useful for this purpose. DNA microarrays are a major tool in genome analysis that have opened a new era of unprecedented opportunity for uncovering the molecular basis of cancer, thereby possibly shedding light on its development mechanism and evolution. DNA microarrays can be used to develop a new taxonomy of cancer, including major insights into the genesis, progression, prognosis, and response to therapy on the basis of gene expression profiles. DNA microarrays measures the expression levels of numerous genes, often thousands at a time.

To extract useful information from DCE-MRI and microarray measurements, the human user has to explore high-dimensional data structures, in which meaningful interrelationships among variables might be hidden by the large amount of data. Therefore, automatic tools that facilitate the data exploration are strongly desired.

1.1 Scope and goals

The scope of this PhD thesis is the application and improvement of computational techniques based on dimensional data reduction for the visual exploration of DCE-MRI and DNA microarray data in breast cancer. Algorithms for dimensional data reduction aim to compute low-dimensional projections of high-dimensional data while best preserving the data topology. In this work several algorithms for dimensional data reduction are used to project the experimental multi-dimensional data sets (DCE-MRI and microarray) into a two-dimensional space for the visual exploration of the similarities between single items. Indeed, similar items in the high-dimensional space are expected to be mapped to neighboring points in the projected space. Therefore, from the visualization of the embedding one can infer information concerning the similarity between items in the high-dimensional data.

1.2 Outline of the manuscript

The manuscript comprises nine chapters. The DCE-MRI and microarray technologies along with the corresponding experimental data sets analyzed in this work are summarized in chapters 2 and 3 respectively; chapter 4 introduces the principles of information visualization and some established methods for the visual exploration of DCE-MRI and microarray data; chapter 5 gives an introduction to dimensional reduction and the algorithms used in this work; chapter 6 illustrates the results from the DCE-MRI breast cancer data analysis; chapter 7 introduces two proposed modifications of the LLE algorithm, an algorithm for dimensional reduction that is largely used in this work; chapter 8 discusses the results obtained from the microarray data analysis; chapter 9 regards the conclusions.

2 Theory of Magnetic Resonance Imaging

This chapter introduces the basic theory of *magnetic resonance imaging* (MRI), an imaging technique used extensively in medical settings to produce high-quality images of interior parts of the human body. The principles of MRI are introduced from section 2.1 to section 2.3. The content of these sections is, to great extent, based on (Webb, 1996), (Wehrli et al., 1988), (Twellmann, 2005), and (Brown and Semelka, 1999).

In the second part (section 2.4) the principles of *dynamic contrast enhanced magnetic resonance imaging* (DCE-MRI) applied to breast cancer are introduced. The understanding of the basic principles of DCE-MRI is fundamental to comprehending the DCE-MRI data analysis explained in chapter 6. The DCE-MRI data investigated in this work are introduced in section 2.4.1.

2.1 Production of Net Magnetization

Magnetic resonance imaging (MRI) has demonstrated to be a valuable diagnostic instrument for the treatment of various disorders, including brain diseases, spinal illness, angiography and tumors.

The physical phenomenon of *magnetic resonance* (MR) is based upon the interaction between an external magnetic field and a nucleus that possesses spin. As an example consider the ^{1}H nucleus, consisting of a single proton. The ^{1}H nucleus is the most abundant isotope for hydrogen. Its choice is motivated by its very high sensitivity to the applied magnetic field and its abundance in the human body. The quantum theory predicts for ^{1}H a spin-quantum number of $m = \frac{1}{2}$.

Concurrent with the spinning there is a local magnetic field. A nucleus with spin can be described, according to the semi-classical approach, by a vector having an axis of rotation with a definite orientation and magnitude to this axis. The associated magnetic field is parallel to the axis of rotation for the nucleus. This orientation of the nuclear spin and the changes induced due to the experimental manipulations that the nucleus undergoes when excited by an external magnetic field provide the basis for the MR signal.

Consider an arbitrary volume of tissue containing protons possessing randomly orientated spins in all directions with respect to the semi-classical approach. Performing a vector addition of these spin vectors produces zero sum, i. e. no net magnetization is observed in the tissue. If the tissue is placed inside a magnetic field $\mathbf{B}_0$, the individual protons begin to rotate about the magnetic field. The frequency of precession

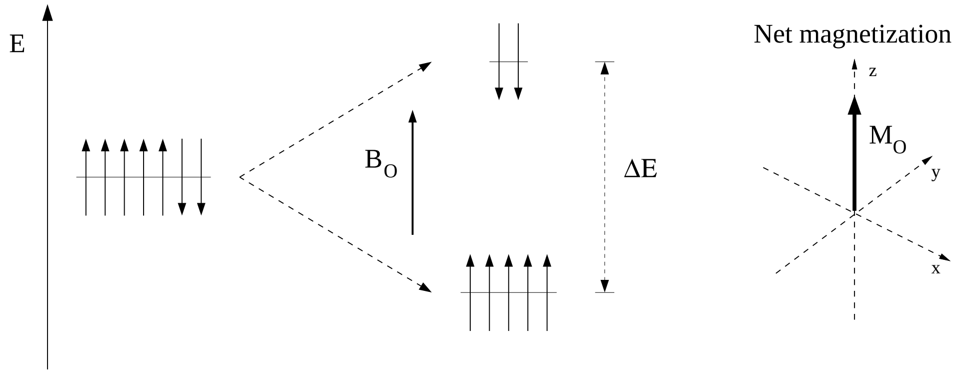

Figure 2.1: (Left) Zeeman diagram. In the absence of a magnetic field, a collection of protons will have the configuration of z components equal in energy so that there is no preferential alignment between the spin up and spin down orientations. In the presence of a magnetic field $\mathbf{B}_0$, the spin up orientation (parallel to $\mathbf{B}_0$) is of lower energy and its configuration will contain more protons than the higher energy, spin down configuration. The higher number of spins in the lower energy level means that the vector sum of spins will be nonzero and the tissue will become magnetized with a value $\mathbf{M}_0$ known as net magnetization as displayed by a vector in the z-direction (right) according to the semi-classical approach.

is proportional to the strength of the magnetic field and is expressed by the *Larmor equation*:

$$\omega_0 = \gamma B_0/2\pi, \tag{2.1}$$

where ω_0 is the *Larmor frequency* in MHz, B_0 is the magnetic field strength in tesla (T) that the proton experiences, and γ (s^{-1} T^{-1}) is a particle-specific constant known as the *gyromagnetic ratio*.

By convention, $\mathbf{B}_0$ is defined to be oriented in the z direction of the coordinate system. In the semi-classical approach, the motion of each proton is then described within an unique set of x, y, z coordinates. The x and y coordinates vary with time as the proton precesses, but the z component, the one parallel to $\mathbf{B}_0$, is constant with time.

In the directions perpendicular to $\mathbf{B}_0$ there is still no net magnetization. However, in the direction parallel to $\mathbf{B}_0$, there is a constant, nonzero interaction between the proton and $\mathbf{B}_0$, known as the *Zeeman effect*. In a quantum theoretical approach, the result of the Zeeman interaction is that spins in the two orientations, parallel (also known as *spin up*) and antiparallel (*spin down*), have different energies (Fig. 2.1). This energy difference ΔE is proportional to $\mathbf{B}_0$:

$$\Delta E = \frac{h\gamma\mathbf{B}_0}{2\pi}, \tag{2.2}$$

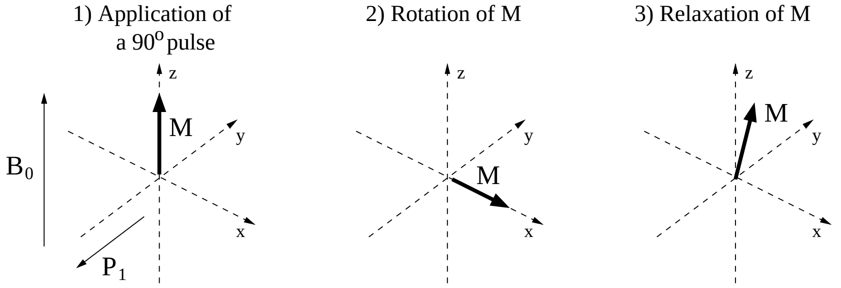

Figure 2.2: Effect of a $90°$ rf pulse. 1) The rf pulse broadcast at the frequency ω_0 can be treated as an additional magnetic field $\mathbf{P}_1$ oriented perpendicular to $\mathbf{B}_0$. 2) The protons absorb the applied energy and $\mathbf{M}$ rotates perpendicular to both $\mathbf{B}_0$ and $\mathbf{P}_1$. 3) After switching $\mathbf{P}_1$ off, a return of $\mathbf{M}$ to its equilibrium state will be observed as the protons release their energy.

where h is Plank's constant. The ratio between the concentration of spins ($N_{\text{spin-up}}$ and $N_{\text{spin-down}}$) in the two energy levels is governed by the *Boltzmann distribution*

$$\frac{N_{\text{spin-up}}}{N_{\text{spin-down}}} = e^{-\Delta E/kT}, \tag{2.3}$$

where k is Boltzmann's constant. Because of the different number of protons in the two energy levels at room temperature, the tissue will become magnetized in the presence of $\mathbf{B}_0$ with a value $\mathbf{M}_0$, known as the *net magnetization* (Fig. 2.1 right). The orientation of this net magnetization will be in the same direction as $\mathbf{B}_0$ and will be constant with time. This configuration has the lowest energy and is the arrangement to which the protons will naturally try to return following any perturbation such as energy absorption. The manipulation of this induced magnetization $\mathbf{M}_0$ is the source of measurable signal for all the MR experiments.

2.2 T1/T2 relaxation times

The simplest manipulation of $\mathbf{M}_0$ involves the application of a short burst, or radio frequency (rf) pulse. If a $90°$ pulse $\mathbf{P}_1$ having the Larmor frequency ω_0 is applied to a sample perpendicular to the z-axis, there will be a coupling between the rf pulse and $\mathbf{M}_0$ so that energy can be transferred to the protons. Because of absorption of the rf energy of frequency ω_0, $\mathbf{M}_0$ will rotate in the xy-plane ($\mathbf{m}_z = 0$). After some time, the protons release their energy and the magnetization returns to its equilibrium value (Fig. 2.2). This return of magnetization follows an exponential growth process with time t

Tissue Type	T1$(\mathbf{B}_0 = 1.5T)$	T2$(\mathbf{B}_0 = 1.5T)$
Fat	$\approx$ 260 ms	$\approx$ 85 ms
Muscle	$\approx$ 860 ms	$\approx$ 45 ms
White matter of the brain	$\approx$ 780 ms	$\approx$ 90 ms
Gray matter of the brain	$\approx$ 920 ms	$\approx$ 100 ms

Table 2.1: T1 and T2 of several tissue types.

$$\mathbf{M}_z(t) = \mathbf{M}_0 \left[1 - \text{exp}\left(-\frac{t}{\text{T1}}\right)\right], \tag{2.4}$$

where T1 is the time required for $\mathbf{M}_z$ to return to 63% of its original value following an excitation pulse. T1 is also known as the *spin-lattice relaxation time*. This term refers to the fact that the excited proton ("spin") transfers its energy to its surroundings ("lattice") rather than to another spin. This energy no longer contributes to spin excitation.

The second observable effect of the application of $\mathbf{P}_1$ is the dephasing of the net magnetization. Absorption of energy from $\mathbf{P}_1$ causes $\mathbf{M}$ to rotate into the xy-plane. This effect gradually disappears once $\mathbf{P}_1$ is turned off and the return of the transverse net magnetization $\mathbf{M}_{xy}$ to zero is described by

$$\mathbf{M}_{xy} = \mathbf{M}_{x_0y_0}\text{exp}\left(-\frac{t}{\text{T2}^*}\right), \tag{2.5}$$

where $\mathbf{M}_{x_0y_0}$ is the magnetization in the xy-plane immediately following the excitation pulse, and the time constant T2* depends on three factors:

$$1/\text{T2}^* = 1/\text{T2} + 1/\text{T2}_M + 1/\text{T2}_{MS}. \tag{2.6}$$

The time T2, known as the *spin-spin relaxation time* or *transverse relaxation time*, is the time required for $\mathbf{M}_{xy}$ to decay to 37% of its initial value. T2$_M$ denotes the dephasing time due to the finite degree of inhomogeneity of $\mathbf{B}_0$ and T2$_{MS}$ is the dephasing time due to the magnetic susceptibility differences between different adjacent tissue types. By the application of a $180°$ rf pulse, the effects of the $\mathbf{B}_0$ inhomogeneity and magnetic susceptibility differences can be eliminated, so that T2* is well approximated by T2.

The T1 and T2 times are biological parameters that are tissue dependent (see table 2.1) and can therefore be used to discriminate among different tissue types in the human body.

The application of an rf pulse to a tissue segment produces a time-varying magnetic flux that induces a current in an rf coil. This current acts as a sine function which decays over time as a consequence of the dephasing of the spins. The temporal signal can be converted from a function of time to a function of frequency by the *inverse Fourier transform*. The frequency spectrum typically exhibits one peak at the resonance frequency ω_0 (see Fig. 2.3).

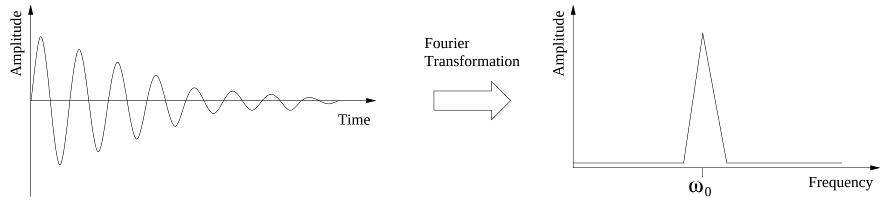

Figure 2.3: The transverse magnetization rotating about the z-axis can be measured by placing a rf coil around the x-axis. The current induced in the coil by the time-varying magnetic flux represents a decaying sine wave that in the frequency domain exhibits a single peak at the Larmor frequency ω_0.

2.3 Principles of Magnetic Resonance Imaging

The principles on how to obtain an MR image from the signal induced by the manipulation of $\mathbf{M}_0$ are outlined in this section.

2.3.1 Spatial Decomposition of MR Signals

According to section 2.2, the frequency of the energy that a proton absorbs depends on the magnetic field that it experiences. Therefore, the application of a spatially dependent magnetic field gradient $\mathbf{G}$ causes a resonance frequency that depends on the spatial coordinates of the spin. MRI uses this dependence to localize the protons to different regions of space.

In presence of a magnetic field gradient $\mathbf{G}$, the frequency ω_i of the energy absorbed by a proton located at position $\mathbf{r}_i$ is given by

$$\omega_i = \gamma(\mathbf{B}_0 + \mathbf{G} \cdot \mathbf{r}_i). \tag{2.7}$$

This formula represents the extension of eq. (2.1). In case, with the above gradient is switched on, a further rf pulse having a certain central frequency is applied, only a narrow region of tissue perpendicular to the gradient will absorb the rf energy. The central frequency of the pulse determines the position of the excited region when a selection gradient is present. Different slice positions are achieved by changing the central frequency. To obtain image information, nuclear magnetic resonance (NMR) spectra are recorded for varying directions of $\mathbf{G}$.

Three magnetic field gradients $\mathbf{G}_x$, $\mathbf{G}_y$, $\mathbf{G}_z$, placed orthogonally to one another, are required to encode information in three dimensions. The gradients need to be applied in an appropriate temporal order as illustrated in Fig. 2.4. A magnetic field gradient (for example $\mathbf{G}_z$), simultaneously applied with a 90°-pulse, excites the spins situated in a slide perpendicular to the magnetic field gradient. After turning off these two signals, the other two magnetic field gradients (in this case $\mathbf{G}_x$ and $\mathbf{G}_y$) are broadcast and the

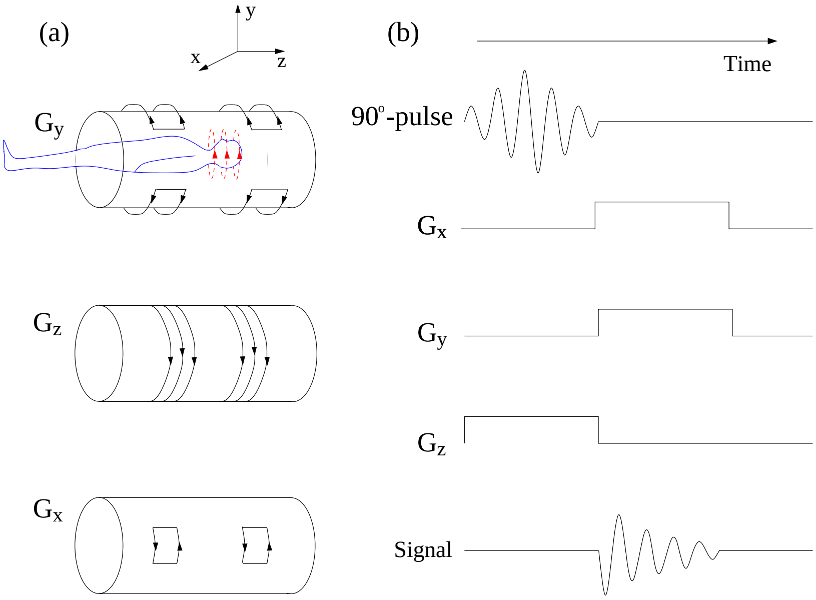

Figure 2.4: (a) The magnetic fields $\mathbf{G}_x$, $\mathbf{G}_y$, $\mathbf{G}_z$ are generated by a current passing through appropriately designed coils. The $90°$-pulse is generated by an additional coil (in this example the bird cage coil in red). (b) An appropriate $90°$-pulse together with a particular magnetic field gradient (in this case $\mathbf{G}_z$) excite only the spins that are located in a certain slice perpendicular to the gradient. The spatial resolution in the plane is determined by the magnetic field gradient obtained by the linear combination of the remaining gradients (in this case $\mathbf{G}_x$ and $\mathbf{G}_y$).

current induced in the RF coil is recorded. This is used to reconstruct the image of the particular slide.

In practice, an MR image is obtained by more complex pulse sequences such as the *spin echo sequence*, which is the one most commonly used. This sequence comprises two rf pulses, the $90°$ pulse is generated each TR seconds (repetition time), creating the detectable magnetization, and the $180°$ pulse that refocuses it at TE (echo time). The selection of TE and TR determines the resulting image contrast.

2.3.2 T1/T2-Weighted Imaging Sequences

The pixel intensity $I(x,y)$ of an MR image obtained by a spin echo sequence is governed by

$$I(x,y) \propto \rho(x,y) \underbrace{\left[1 - \exp\left(-\frac{\mathrm{TR}}{\mathrm{T1}}\right)\right]}_{\text{T1-weighting}} \underbrace{\exp\left(-\frac{\mathrm{TE}}{\mathrm{T2}}\right)}_{\text{T2-weighting}} \tag{2.8}$$

where $\rho(x,y)$ denotes the proton density. The TE and TR times module the contributions of T1 and T2 to the pixel intensity, thereby affecting the final MR image. If TR is much larger than all values of T1 relative to the tissue of a certain region of interest (ROI), the sensitivity of the pixel intensity to the T1 relaxation is reduced as the T1-weighting term converges to one. The obtained image is termed *T2-weighted* image. The same effect for T2 can be obtained by setting TE to a value that is considerably lower than all values of T2 relative to tissue of a given ROI. The resulting image is named *T1-weighted* image.

In general, in T1-weighted images, tissues that have short T1 relaxation times (such as fat) are characterized by bright signal. Tissues with long T1 relaxation times (such as cysts, cerebrospinal fluid and edema) appear as dark signal. In T2-weighted images, tissues that have long T2 relaxation times (such as fluids) appear bright.

It is also possible to eliminate the contributions of both T1 and T2. In such a case the pixel intensity depends only on the proton density $\rho(x,y)$ and the obtained image is named *proton density image*.

2.3.3 Multispectral Magnetic Resonance Imaging

Multispectral magnetic resonance imaging consists in acquiring a sequence of two- or three-dimensional MR images from the same ROI and each image of the sequence is generated with a different T1/T2-weighting. In order to guarantee the correct registration of the images, the patient is supposed not to move during the acquisition time. As a result, each spatial position in the ROI is associated with a vector of intensity values, each obtained with different settings of T1 and T2. Multispectral images allow the discrimination of different tissue types and are used to distinguish gray matter, white matter or cerebrospinal fluid from pathological structures such as multiple sclerosis or carcinoma in the brain.

2.3.4 Contrast agents

The contrast between certain tissue types can be enhanced by contrast agents (CA), i. e. intravenous injected fluids that alter the relaxation times. After the administration, the CA molecules rapidly mix with the blood and their concentration increases within tissue with high vascularity or affinity to the CA molecules, thereby leading to a signal enhancement. The analysis of the CA concentration over time is useful for extracting physiological parameters that characterize the tissue locally.

Paramagnetic contrast agents such as *gadolinium based chelate* (Gd-DTPA) shorten the time T1. Tissue that accumulates CA rapidly is displayed with increased brightness in T1-weighted images, which in turn can be recorded faster by virtue of the shortened T1 value.

In contrast to the paramagnetic contrast agent, superparamagnetic CA shorten the T2 and $T2^*$ values that are typically accumulated in healthy tissue. Therefore, healthy tissue appears with suppressed brightness in T2-weighted spin echo sequences, while the tumor signal intensity does not change significantly.

2.3.5 Multislice and 3D Imaging

A crucial aspect of MR imaging is the time needed for the acquisition of a single MR image. The acquisition time largely depends on the desired resolution and the repetition time TR. Thus, there exists a trade-off between the acquisition time and the image resolution. In general, the resolution should be as high as possible in order to allow for the exploration of small anatomical structures and minimize the *partial volume effect*. This describes the fact that the signal measured for each voxel arises from the entire tissue within a small three-dimensional cuboid, and in turn different tissue types might be combined together. The higher the resolution, the smaller the voxel and the lower the partial volume effect.

On the other hand, depending on the region of the body to be imaged, the image resolution is limited by the acquisition time. For instance the imaging of the abdomen has to be conducted within one breath-hold (5 to 25 seconds) in order to avoid motion artifacts.

To accelerate the image acquisition, several techniques have been proposed. In a single slice sequence, only the relaxation of spins located in a single plane determined by the slice selective gradient are recorded. However, MRI allows to simultaneously record the relaxation of spins in several parallel slices (*multislice imaging*). Instead of affecting only spins with a resonance frequency in the range of $\nu_{pulse} \pm \Delta\nu_{pulse}$, in multislice imaging, spins with a resonance frequency in one of the n non-overlapping frequency bands $[\nu_{pulse_1} \pm \Delta\nu_{pulse}]$, ..., $[\nu_{pulse_n} \pm \Delta\nu_{pulse}]$ are affected in quick succession and the T1 relaxation of the spins in the corresponding n slices is measured simultaneously. The maximum number of slices that can be examined simultaneously is limited by the ratio TR/TE. In addition, the slices have to be placed with a sufficient spatial distance to each other in order to avoid partial excitation of the spins in adjacent slices.

Whereas the multislice imaging technique acquires the image of a three-dimensional volume as a sequence of one or more tomographic slices, thereby possibly leading to a temporal shift between slices, 'real' three-dimensional images can be recorded by more sophisticated techniques such as the *three-dimensional fast low-angle-shot* (3D flash) sequence.

2.4 Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI)

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) consists in repeatedly imaging a region of interest (ROI) (in this work the female breast) over time in order to monitor the dynamic of an injected paramagnetic contrast agent in the tissue.

As a result, a temporal sequence of m MR 2D-images or 3D-volumes is recorded. Typically, this temporal sequence is composed of at least one precontrast image or volume, recorded before the administration of the contrast agent, and several postcontrast images or volumes, recorded after the administration of the contrast agent. Each pixel $\mathbf{p}(x, y)$ or voxel $\mathbf{p}(x, y, z)$ can be connected with a time-series $\mathbf{s_p} = (s_{\mathbf{p}1}, s_{\mathbf{p}2}, ...s_{\mathbf{p}m})$ of m intensity values, where $s_{\mathbf{p}i}$ denotes the signal intensity of the pixel (or voxel) $\mathbf{p}$ in the i-th image (or volume). This time-series can be interpreted as a time-point belonging to a signal space $\mathbb{R}^m$ (see Fig. 2.5).

The temporal course of the signal intensity values is affected by the concentration of contrast agent within the respective tissue. Specifically, high-vascularized tissue will absorb more contrast agent, resulting in a signal enhancement. By contrast, low-vascularized tissue will be characterized by negligible or absent signal enhancement as a consequence of the low quantity of absorbed contrast agent.

The analysis of the time-series can thus allow medical experts to discriminate between different tissue types, including potential breast tumor lesions. Indeed, the latter are commonly characterized by highly-vascularized levels of tissue, and, consequently, they

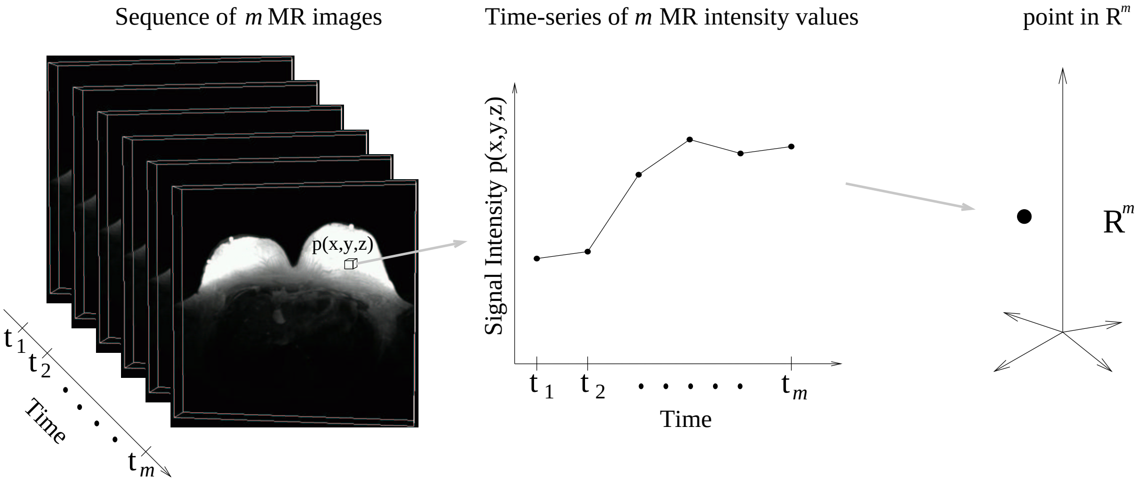

Figure 2.5: Given a sequence of m volumes in DCE-MRI, each voxel $\mathbf{p}(x, y, z)$ can be associated with a time-series of m signal intensity values. This time-series can be regarded as a data point in $\mathbb{R}^m$.

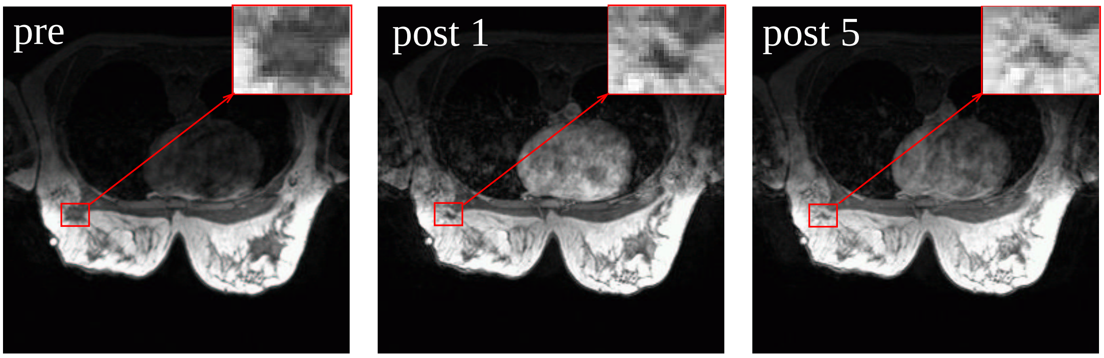

Figure 2.6: Three 2D-sections taken from a sequence of DCE-MRI 3D-volumes of a female breast. The volumes involved are the precontrast, the first and fifth postcontrast ones. The breast has a tumor lesion that is magnified. Prior to the injection of the contrast agent, the lesion appears dark (pre image) One can see (post 1 image) that the signal intensity relative to the lesion tissue enhances as a consequence of the absorption of the contrast agent. The enhancement is even stronger after some time (post 5 image).

are supposed to enhance after the administration of the contrast agent (see Fig 2.6).

In addition, the analysis of the time-series can allow medical experts to discriminate between benign and malignant tumor lesions. In fact, according to (Heywang-Köbrunner and Beck, 1996) and (Brown et al., 2000), benign and malignant lesions are expected to differ with respect to the observed uptake characteristics of the contrast agent as schematically shown in Fig. 2.7. Specifically, malignant tumors are supposed to be highly vascularized. For this reason the processes of absorption (wash-in) and wash-out of the contrast agent into the lesion are expected to be significantly fast. By contrast, benign tumors are usually less vascularized and this leads in general to a slow wash-in followed by a modest wash-out of the contrast agent.

DCE-MRI in breast cancer has demonstrated high sensitivity, with values of 93-100% (Brown et al., 2000). Moreover, 17-34% of breast cancers that are visible on DCE-MRI are not detected on mammography (Weatherall et al., 2001). In addition, the anatomic extent of lesions can be more accurately defined with MRI compared with mammography (98% vs 55%) (Esserman et al., 1999). However, the differentiation of benign from malignant lesions in DCE-MRI can be difficult, with the result that many studies reported specificities ranging from 37% to 97% (Kelcz et al., 2002).

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) has proved to be a particularly valuable multivariate intramodular imaging technique in the diagnosis of breast cancer for young women (Brown et al., 2000), whose breast tissue is particularly dense. Indeed, in the case of the latter, a standard technique such as x-ray

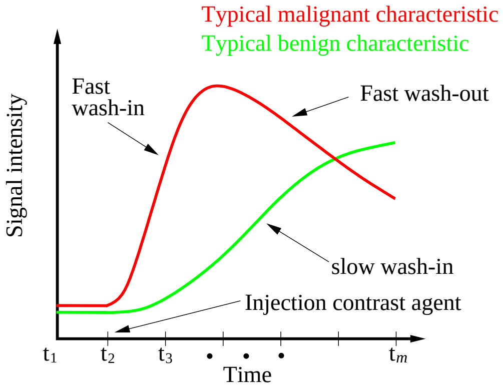

Figure 2.7: Model-based contrast agent characteristics in DCE-MRI of benign and malignant breast tumor tissue. The contrast agent is injected at time t_2. The malignant curve is characterized by a fast wash-in and wash-out phases by virtue of the highly vascularized tissue. The tissue of benign lesions is typically less vascularized and the corresponding contrast characteristics exhibit a slower wash-in. The wash-out phase is often not visible (as in figure) because it takes place after the time point t_m.

mammography can fail to detect abnormal breast lesions (Rankin, 2000). By contrast, DCE-MRI has demonstrated considerable potential for detection and monitoring of abnormalities in young women (Brown et al., 2000). In this regard, several studies (Brown et al., 2000; Kriege et al., 2001) have been conducted in recent years in order to test the efficacy of DCE-MRI as a screening technique for young women at high risk of developing breast cancer (e. g. women with a strong family history of breast cancer or with mutation in one of the breast cancer (BRCA) genes).

2.4.1 DCE-MRI Data Set of this Work

The DCE-MRI data set investigated in this work was provided by the City Center Hospital of the Munich University. This study considers images data from seven women with malignant breast tumors and seven women with benign lesions, 14 tumor cases in total.

The data of each patient comprises six three-dimensional ($m = 6$) volumes of the full breast acquired with a temporal resolution of 110 s. The first volume was acquired before the bolus injection of a paramagnetic contrast agent, thereafter followed by the remaining five measurements. The imaging process was performed with a 1,5 T system (Magneton Vision, Siemens, Erlangen, Germany) equipped with a dedicated surface coil to enable simultaneous imaging of both breasts. First, transversal images were acquired with a STIR (short tau inversion recovery) sequence (TR$= 5600$ ms, TE$= 60$ ms, FA$= 90°$, TI$= 150$ ms, matrix size of 256×256 pixels, slice thickness 4mm), then a dynamic T1 weighted gradient echo sequence (3D FLASH) was performed (TR$= 12$ ms, TE$= 5$ ms,

case ID	Histologic Type	Size
malignant cases		
M1	ductal carcinoma	207
M2	scirrhous carcinoma	68
M3	DCIS (ductal carcinoma in situ)	49
M4	status post mastectomy, multilocullar recurrent duc. carc.	743
M5	ductal papillomatosis, transition into papillary carc.	169
M6	ductal carcinoma	284
benign cases		
B1	fibroadenoma	169
B2	fibrous mastopathy	497
B3	scar	26
B4	lymph node	113
B5	granuloma with signs of inflammation	99
B6	chronic mastitis	25

Table 2.2: The tumors of the 12-cases data set.

FA$= 25^\circ$) in transversal slice orientation with a matrix size of 256×256 pixels and an effective slice thickness of 4 mm.

The tumor lesions were manually marked out by an experienced radiologist using a cursor on a screening device. The classification of each lesion was histologically proven by a surgical biopsy. In order to validate the diagnosis, all patients were also included in a clinical follow-up for more than one year.

The size (number of voxels) of the 14 marked out lesions along with their histological characterization are shown in table 2.2 and table 2.3. The tumors are listed in two different tables because the cases M7 and B7 have been made available much later than the 12 cases listed in table 2.2. For this reason two different data sets are considered in this work: a 12-cases data set given by the tumors in table 2.2, and a 14-cases data

case ID	Lesion Classification	Size
malignant case		
M7	DCIS (ductal carcinoma in situ)	163
benign case		
B7	scar	10

Table 2.3: The 2 new tumor cases with their histological type.

Data set	# Benign Points	# Malignant Points	Total size
12-cases	1520	929	2449
14-cases	1683	939	2622

Table 2.4: The two data sets analyzed in this work.

set given by all the lesions. Both data sets are summarized in table 2.4. Most of all the work presented in this thesis regards the analysis of the 12-cases data set. In particular, it is analyzed in chapter 6 and chapter 7. By contrast, the 14-cases data set is analyzed only in section 6.4.

2.5 Summary

In this chapter, the magnetic resonance imaging (MRI) and dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) techniques have been introduced. MRI is a non ionizing imaging technique that allows for the acquisition of two- and three-dimensional images with high spatial resolution and excellent soft-tissue contrast. The MRI signal arises from the spin-lattice and spin-spin relaxation time of hydrogen nuclei in a static magnetic field after being excited by a sequence of electromagnetic pulses at the resonance frequency of hydrogen. The spatial information is encoded as a signal by superimposing three orthogonal magnetic field gradients that lead to a spatially dependent resonance frequency and phase of hydrogen nuclei. The signal is measured as a current that is induced in a tuner receiver coil due to the time varying flux caused by the relaxing nuclei. Finally, this signal is converted into image information by the inverse Fourier transform.

The DCE-MRI technique involves the MR imaging of a region of interest prior to the injection of a contrast agent and repeatedly thereafter, resulting in a temporal sequence of 2D-images or 3D-volumes. DCE-MRI has demonstrated promising potential for the screening of young women at high risk of breast cancer.

The chapter is concluded by presenting the DCE-MRI data investigated in this work.

3 Introduction to Microarray Technology

This chapter introduces the principles of the microarray technology. For more information see (Kambhampati, 2004; Schena, 2000) The microarray data analyzed in this study are described in section 3.3. At the end of the chapter, a glossary with some terms typical of the microarray technology is included.

3.1 DNA Microarrays

DNA microarrays offer the latest technological advancement for multi-gene detection and diagnostics. This novel technology, which started to appear during the second half of the 1990s, has historically evolved from the initial experimental reports published in the mid 1970s, which indicated that labeled nucleic acids could be used to monitor the expression of nucleic acid molecules attached to a solid support. However, it was not until 1995 that the first article describing the application of DNA microarray technology to expression analysis was published in the scientific literature by Patrick Brown and his colleagues at Stanford University (Schena et al., 1995).

Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. In addition, DNA microarrays are also used for DNA sequence analysis, immunology, genotyping and diagnostics. Moreover, the flexibility and high throughput capabilities of microarrays hold tremendous potential for pathogen detection, identification, and genotyping in molecular diagnostic laboratories.

Microarray technology has also become a major tool for the investigation of cancer, as indicated by the rapidly increasing number of publications describing DNA microarray analysis of cancers since 2000 (Ochs and Godwin, 2003). One of the first studies correlating gene expression profiles with tumor classification is (Golub et al., 1999). Gene expression pattern analysis have been used for classifying breast tumors in (Perou et al., 2000; Sorlie et al., 2001).

3.2 Microarray Principles

DNA microarrays are typically composed of DNA "probes" that are bound to a solid substrate such as glass, plastic or silicon chip forming an array. Each identified sequence gene on the substrate corresponds to a fragment of genomic DNA, complementary DNA (cDNA), polymerase chain reaction (PCR) products or chemically synthesized oligonucleotides and represents a single gene.

Usually a single DNA microarray slide/chip may contain thousands of spots, each representing a single gene and collectively the entire genome of an organism.

3.2.1 Types of microarray

A schematic diagram of the two most commonly used DNA microarray formats to date, the glass DNA microarrays and high-density oligonucleotides microarrays (often referred to as a "chip"), are shown in Fig 3.1 and Fig. 3.2, respectively.

Glass DNA microarray was the first type of DNA microarray technology developed. It is produced by using a robotic device, which deposits (spots) a nanoliter of DNA (50-150 μm in diameter) onto a coated microscope glass slide surface in serial order with a distance of approximately 200-250 μm from each other, one spot-one gene. These moderate sized glass cDNA microarrays also bears about 10.000 spots or more on an area of 3.6 cm^2.

In situ (on chip) oligonucleotides array format (Fig. 3.2) is a sophisticated platform of microarray technology which is manufactured by using the technology of in situ chemical synthesis. This type of technology has also been employed to manufacture so-called GeneChips which refer to its high density oligonucleotide based DNA arrays. Presently, the commercial versions of Affymetrix GeneChips hold up to 500,000 spots in a 1.28 cm^2 chip area, and due to such very high information content, they are finding widespread use in the hybridization-based detection and analysis of mutations and polymorphisms, such as single nucleotide polymorphisms or disease-relevant mutations

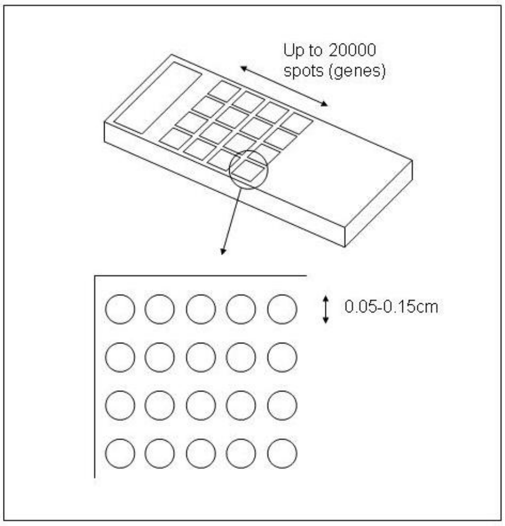

Figure 3.1: Glass complementary DNA (cDNA) microarray. This picture is taken from (Gundogdu and Elmi, 2005).

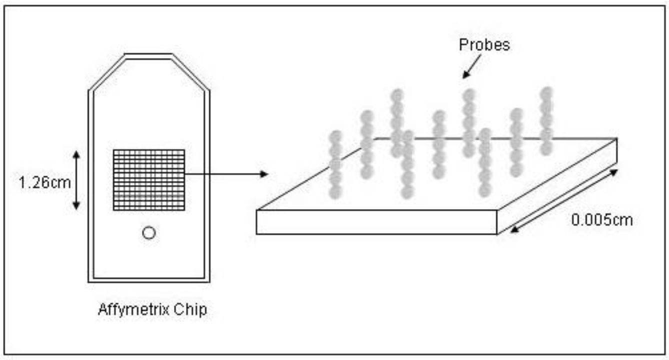

Figure 3.2: Illustration of a DNA GeneChip (Affymetrix). The picture is taken from (Gundogdu and Elmi, 2005).

analysis ("genotyping"), as well as a wide range of other applications such as gene expression studies, to mention a few.

3.2.2 Principles of DNA Microarray Experiments

A typical microarray experiment comprises four major steps that are illustrated in Fig. 3.3

1. Target preparation and labeling
2. Hybridization
3. Washing
4. Image acquisition and data analysis

Target preparation Initially, the sample preparation starts by extracting RNA containing messenger RNA (mRNA) from two cell cultures (treated cells and control cells, also termed test sample and control sample, respectively) whose expression levels need to be compared. This step is crucial, simply because the overall success of any microarray experiment depends on the quality of the RNA.

The sample mRNA extracted from the test and control samples are then separately converted into complementary DNA (cDNA) using a reverse-transcriptase enzyme. This step also requires a short primer to initiate cDNA synthesis. On this stage, DNA from the control sample are labeled with a green dye (Cy3), whereas DNA from the test sample are labeled with a red dye (Cy5).

Hybridization Hybridization is the process of joining two complementary strands of DNA to form a double-stranded molecule. Here, the labeled cDNA (Sample and Control) are mixed together, and then purified to remove contaminants such as primers, unincorporated nucleotides, cellular proteins, lipids, and carbohydrates.

After purification, the mixed labeled cDNA is competitively hybridized against denatured PCR product or cDNA molecules spotted on a glass slide. Ideally, each molecule in the labeled cDNA will only bind to its appropriate complementary target sequence on the immobilized array.

Washing The slides are washed after hybridization, first to remove any labeled cDNA that did not hybridize on the array, and secondly to increase stringency of the experiment to reduce cross hybridization. The latter is achieved by either increasing the temperature or lowering the ionic strength of the buffers.

Image acquisition and data analysis This is the final step of microarray experiments. The aim is to produce an image of the surface of the hybridized array. Here the slide

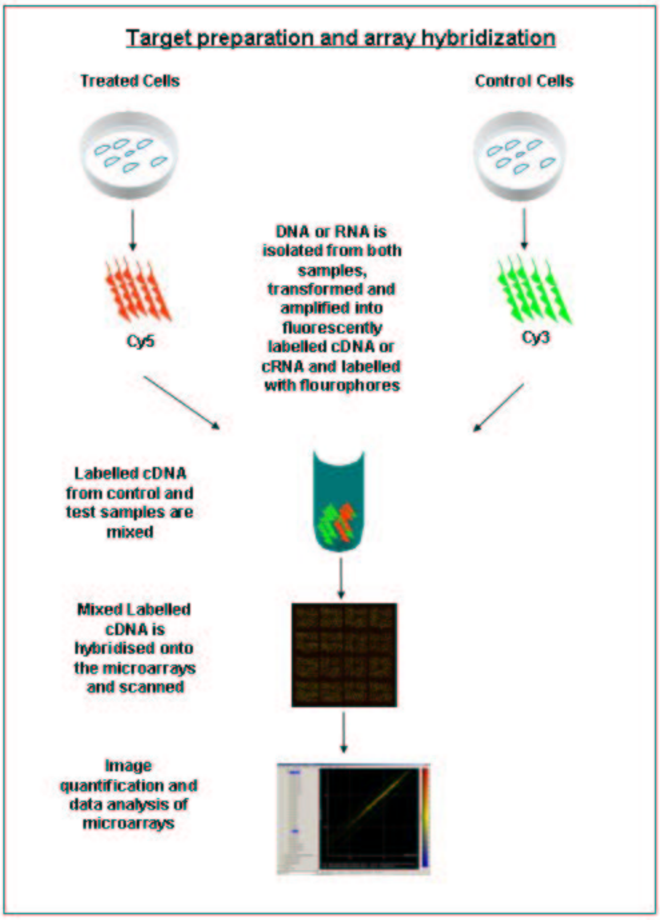
Target preparation and array hybridization
Treated Cells
Control Cells
DNA or RNA is isolated from both samples, transformed and amplified into fluorescently labelled cDNA or cRNA and labelled with flourophores
Cy5
Cy3
Labelled cDNA from control and test samples are mixed
Mixed Labelled cDNA is hybridised onto the microarrays and scanned
Image quantification and data analysis of microarrays

Figure 3.3: Microarray experimental principles. The picture is from (Gundogdu and Elmi, 2005).

Data set	Number of Patients (v)	Number of Genes for each Patient (D)
70-genes	78	70
231-genes	78	231
5223-genes	78	5223

Table 3.1: The three microarray data sets analyzed in this work.

is dried and placed into a laser scanner to determine how much labeled cDNA (probe) is bound to each target spot. Laser excitation of the incorporated targets yields an emission with characteristic spectra, which is measured using a confocal laser microscope. Classically, microarray software often uses red spots on the microarray to represent genes upregulated compared to the control sample, green to represent those genes that are downregulated in the test sample, and black to represent those genes of equal abundance in both test and control samples. An example of an image obtained in this way is shown in Fig. 4.4.

3.3 Microarray Data sets of this Work

The microarray data analyzed in this work were acquired from 78 patients suffering from breast cancer. 44 patients have continued to be disease-free after a period of five years, while 35 patients developed metastases within five years. The former patients are treated as benign cases, while the latter represent malignant cases. Starting from a set of 24482 genes, relevant genes suspected to be connected with metastases were progressively filtered in groups of 5223, 231 and 70 genes by van't Veer et al. (van't Veer et al., 2002), resulting in three data sets that are listed in table 3.1. These data sets represent a two-class problem and are analyzed in this work in terms of differentiation between the benign and malignant cases.

Glossary

RNA A nucleic acid found in all living cells. Plays a role in transferring information from DNA to the protein-forming system of the cell.

Messenger RNA (mRNA) Single stranded RNA molecule that specifies the amino acid sequence of one or more polypeptide chains.

Complementary DNA (cDNA) DNA that is synthesized from a messenger RNA template, the single-stranded form is often used as a probe in physical mapping to locate the gene or can be cloned in the double stranded form. Viral reverse transcriptase can be used to synthesize DNA that is complementary to RNA (for example an isolated mRNA).

Polymer Chain Reaction (PCR) The first practical system for in vitro amplification of DNA and as such one of the most important recent developments in molecular biology.

Oligonucleotide Linear sequence of up to 20 nucleotides joined by phosphodiester bonds. Above this length the term polynucleotide begins to be used.

4 Information Visualization

This chapter provides an introduction to the concept of information visualization for the visual exploration of multidimensional data. After giving a general overview on this topic in section 4.1 and section 4.2, some examples of information visualization systems for the exploration of DCE-MRI and microarray data are presented in section 4.3.

4.1 Introduction to Information Visualization

This section is based on the content of (Keim, 1992) and (Spence, 2001).

Nowadays, the modern instruments in information technology make it possible to store massive amounts of data. The data is typically automatically recorded via sensors and monitoring systems. Usually, as many parameters are recorded, the data is high-dimensional. However, the high dimensionality of the data hampers the extraction of valuable information from that data. In particular, it is almost impossible for a human user to explore exhaustively large data sets containing even millions of data items without the support of computer-based techniques for data exploration. The computer-based data exploration may be accomplished by methods from statistics and machine learning, thereby providing numerical results.

Techniques for information visualization are also a powerful instrument for the analysis of large high-dimensional data sets. The main purpose of information visualization is to present the data in some visual form, in order to allow the human user to gain insight into that data (Spence, 2001). Visual data exploration has several advantages over statistics and machine learning techniques (Keim, 1992):

- it can easily deal with highly inhomogeneous and noisy data
- it is intuitive and does not require understanding of complex mathematics or statistical algorithms
- it presents data directly to human eye, whose extremely high capability of pattern recognition is still superior to other analytical techniques (Li, 2004)

4.2 Data Dimensionality

A real-world data set typically consists of a large number of records, each consisting of a large number of variables or attributes. The number of these variables, that is often referred to as the dimensionality of the data, influences strongly the difficulty of representing the data set in an interactive display.

If each record is characterized by one, two or three variables, the data is referred to as univariate, bivariate or trivariate, respectively. For all these types of data the task of visualization is relatively straightforward as the records can be represented as points in a one-, two- or three-dimensional scatter plot.

When the number of variables exceeds three, the data is said to be multivariate. In this case the data visualization is more challenging since the attributes can not be mapped directly to a scatter plot. Therefore, more sophisticated visualization approaches are required. Many approaches have been proposed for this purpose but an exhaustive list is out of the scope of this work. Some of these techniques are described in chapter 3 in (Spence, 2001).

The DCE-MRI and microarray data sets analyzed in this work, that are described in section 2.4.1 and section 3.3 respectively, are multivariate data.

The records of the DCE-MRI data sets are given by the voxels of the marked tumor lesions and each voxel is characterized by six attributes, i. e. a time-series of six signal intensity values, and, in turn, the DCE-MRI data set is six-dimensional.

A record in the microarray data set corresponds to a patient. Each patient is characterized by a series of genes. In particular, there are three kinds of gene series for each patient, a gene series with 70 genes, a gene series with 231 genes and a gene series with 5223 genes. It follows that there are three data sets that are 70-, 231- and 5223-dimensional, respectively.

The approach to their visual exploration adopted in this work is based on dimensional reduction. Specifically, algorithms for dimensional reduction are used to compute low-dimensional projections of the multivariate data sets. The visualization of the low-dimensional projections can allow the human user to gain insight with regard to the similarity between the data records in the data sets. Indeed, records characterized by similar attributes are expected to be mapped to neighboring data points in the projected space. An introduction to algorithms for dimensional data reduction is given in chapter 5.

4.3 Visual exploration of multivariate breast cancer data

In this section, some established techniques for the visual representation of information extracted from breast DCE-MRI and microarray data are presented. In this work, only the technique shown in section 4.3.2 is employed and compared with one of the techniques proposed in the experimental part (see section 6.3).

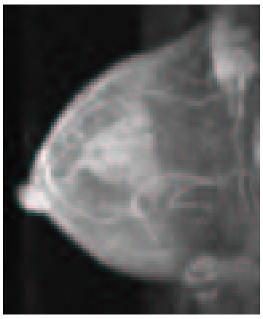

Figure 4.1: Example of the image of a female breast obtained by MIP. At each voxel position the signal intensity corresponds to the maximal component of the correlate time-series.

4.3.1 Maximum Intensity Projection (MIP) in DCE-MRI

Maximum Intensity Projection, or MIP, is used to detect highly vascularized structures in DCE-MRI images or volumes from a single patient. The MIP produces a single image or volume in which the highest intensity value among the components of the respective time-series is depicted at each pixel or voxel.

For example, given a sequence of m 3D volumes, each voxel $\mathbf{p}(x, y, z)$ can be connected with a time-series $\mathbf{s_p} = (s_{\mathbf{p}1}, s_{\mathbf{p}2}, ...s_{\mathbf{p}m})$ of m intensity values (see Fig. 2.5). MIP applied to the entire sequence produces one volume in which the signal intensity $s_\mathbf{p}(\text{MIP})$ of each voxel $\mathbf{p}(x, y, z)$ is given by:

$$s_\mathbf{p}(\text{MIP}) = \max[s_{\mathbf{p}1}, s_{\mathbf{p}2}, ...s_{\mathbf{p}m}].$$

The MIP image highlights the spatial distribution of enhancing lesions for surgery treatment planning (Iacconi et al., 2005). An image obtained by MIP is shown in Fig. 4.1.

The MIP technique represents a conceptually simple approach to the visualization of multivariate DCE-MRI data. However, this technique does not provide information about the dynamic of the time-series. It may occur that the maximal components of two time-series are equal but the dynamic of the time-series is completely different, for example suppose they are characterized by the two behaviors shown in Fig. 2.7. The voxels related to these time-series would have the same signal intensity in the MIP image. Therefore, the different dynamic of the time-series would not be visualized and, in turn, in this case the MIP technique does not prove useful for the discrimination between benign and malignant lesions.

4.3.2 Three Time Point Method (3TP) in DCE-MRI

The three-time-points (3TP) method (Degani et al., 1997) is a model-based technique used in DCE-MRI data analysis that characterizes the dynamic of the time-series by analyzing the changes in the signal intensity at three selected time points t_0, t_1 and t_2 along the contrast characteristic. The volumes related to these particular time points are one pre-contrast, one early post-contrast and one late post-contrast volumes.

The wash-in and wash-out phases related to a certain voxel with spatial coordinate $\mathbf{p}(x, y, z)$ are evaluated from the three signal intensity values $I_\mathbf{p}(t_0)$, $I_\mathbf{p}(t_1)$, $I_\mathbf{p}(t_2)$ and visually represented in terms of color intensity $i_\mathbf{p}$ and color hue $h_\mathbf{p}$, respectively (Weinstein et al., 1999).

The wash-in is estimated from the first two time points and is coded as color intensity as follows:

$$i_\mathbf{p} = \frac{I_\mathbf{p}(t_1) - I_\mathbf{p}(t_0)}{t_1 - t_0}. \tag{4.1}$$

The wash-out is estimated from the values of the two post-contrast intensities $I_\mathbf{p}(t_1)$, $I_\mathbf{p}(t_2)$ and coded by color hues according to one of the three patterns:

$$h_\mathbf{p} = \begin{cases} \text{blue}: & \text{if } I_\mathbf{p}(t_2) > I_\mathbf{p}(t_1) \wedge |I_\mathbf{p}(t_2) - I_\mathbf{p}(t_1)| > \sigma I_\mathbf{p}(t_1) \\ \text{green}: & \text{if } |I_\mathbf{p}(t_2) - I_\mathbf{p}(t_1)| < \sigma I_\mathbf{p}(t_1) \\ \text{red}: & \text{if } I_\mathbf{p}(t_2) < I_\mathbf{p}(t_1) \wedge |I_\mathbf{p}(t_2) - I_\mathbf{p}(t_1)| > \sigma I_\mathbf{p}(t_1) \end{cases}$$

The parameter σ regulates the tolerance for the comparison of $I_\mathbf{p}(t_1)$ and $I_\mathbf{p}(t_2)$. Typically it is set to 0.1, i. e. only a signal change of a at least 10% of the value of $I_\mathbf{p}(t_1)$ is considered to be indicative of the presence or absence of a significant wash-out phase. The color hue blue denotes a slow wash-out, indicating benign behavior. Moderate wash-out is encoded as green, indicating a suspect case. Malignant behavior is encoded as red and is characterized by a fast wash-out. The three patterns of the wash-out cha-

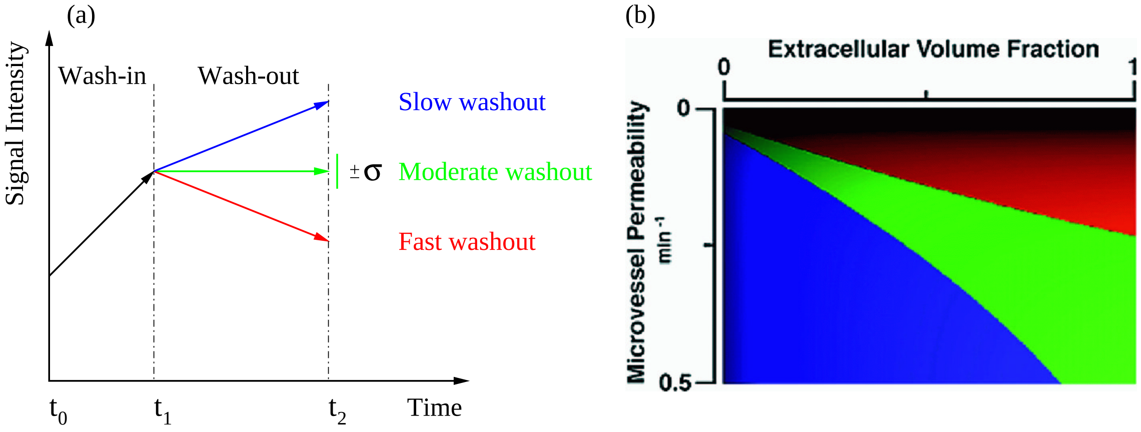

Figure 4.2: (a) Pattern estimation of the contrast agent characteristic. (b) Color map of 3TP. This map allows the user to interpret each pair of values ($i_\mathbf{p}, h_\mathbf{p}$) in terms of microvessel permeability and extracellular volume fraction. This picture is taken from (Weinstein et al., 1999).

racteristics are shown in Fig. 4.2 left. Two of these curves (the benign and malignant ones) are explained in section 2.4 and visualized in Fig. 2.7.

By a calibration map (Fig. 4.2 right) based on modeling tracer kinetics adapted for MRI (Tofts and Kermode, 1991), the color intensity and color hue of each voxel are related to two pathophysiological features: *microvascular permeability* (K^{trans}), which is equal to the microcapillary surface area times permeability, and the *extracellular volume fraction* (v_e), which determines the volume fraction accessible to the contrast agent (Degani et al., 1997). The calibration map is divided into three regions. Tissue with high K^{trans} and low v_e is indicative of benign tissue and is visualized in blue. A high value of v_e and a low value of K^{trans} are indicative of malignant tissue and encoded as

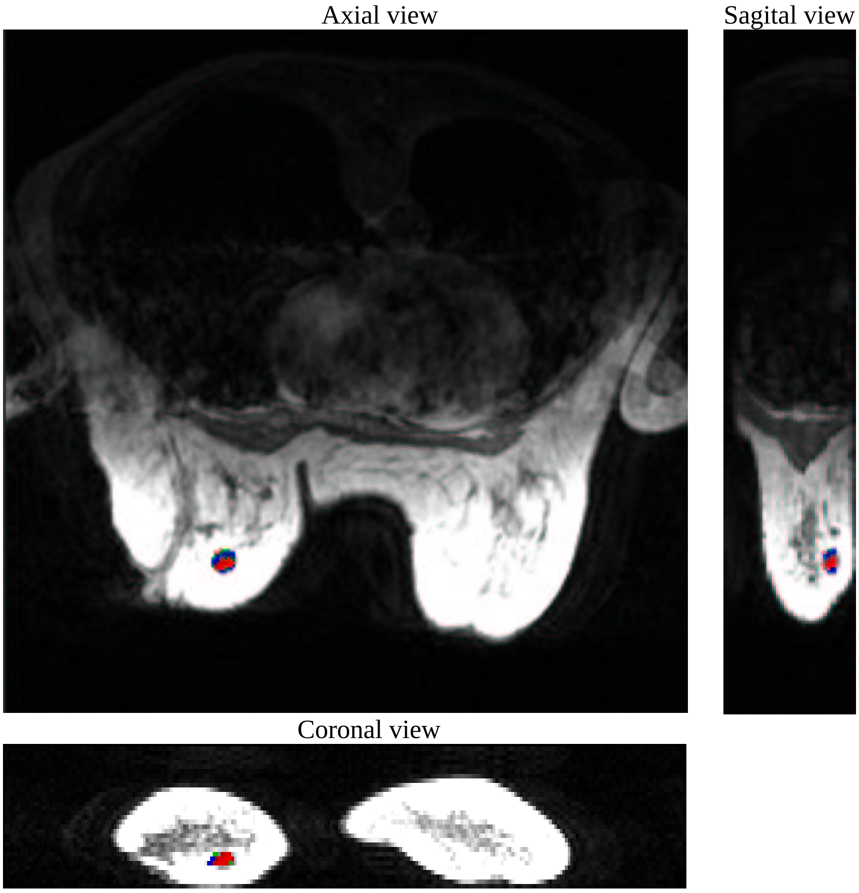

Figure 4.3: Example of a tumor lesion visualized according to the 3TP color scheme. The lesion was manually segmented by a physician from a sequence of DCE-MRI volumes. The presence of the colors red, green and blue reveals the heterogeneity of the tumor tissue. A magnification of the lesion can be seen in Fig. 6.20 (lesion M5).

red. These two regions are separated by a green region that characterizes tissue with intermediate values of v_e and K^{trans}.

The three time-points (t_0, t_1, t_2) are chosen by an iterative process aiming at reaching a calibration map in which the $K^{trans} - v_e$ plane is divided into three approximately equal areas, one for each color (Degani et al., 1997).

An example of lesion colored by 3TP is shown in three different sections (axial, sagital and coronal views) in Fig. 4.3. Although this lesion was pathologically classified as malignant, one can see that it also comprises blue and green parts, thereby revealing the presence of local regions with benign and suspect dynamic characteristics.

4.3.3 Visualization of Microarray Data

One of the most widely used techniques for the analysis of microarray gene-expression data is *hierarchical clustering* (Quackenbush, 2001).

Hierarchical clustering is an agglomerative approach in which single gene expression profiles are joined to form clusters such that those within each cluster are more closely related to one another than gene expression profiles assigned to different clusters. The clusters are further joined until the process has been carried to completion, forming a single hierarchical tree (Quackenbush, 2001). The term hierarchical clustering is mentioned also in section 5.3. Hierarchical clustering has the advantage that it is simple and the result can be easily visualized (Eisen and Brown, 1999).

The process of hierarchical clustering proceeds in a simple manner. First, the pairwise distance matrix among all the genes to be clustered is calculated. Second, the distance matrix is searched for the two most similar genes or clusters; initially each cluster consists of a single gene. This is the first stage in the clustering process. If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives. Third, the two selected clusters are merged to produce a new cluster that now contains at least two objects. Fourth, the distances are calculated between this new cluster and all other clusters. There is no need to calculate all distances as only those involving the new cluster have changed. Last, steps 2–4 are repeated until all objects are in one cluster (Quackenbush, 2001).

Two examples of hierarchical clustering applied to microarray data are shown in Fig. 4.4. All the samples are assembled into a single tree in which each sample corresponds to one leaf. Conventionally, all the leaves are shown at the same level of the drawing. The ordering of the leaves is arbitrary, as is their horizontal position. The heights of the internal nodes may be arbitrary, or may be related to the metric information used to form the clustering.

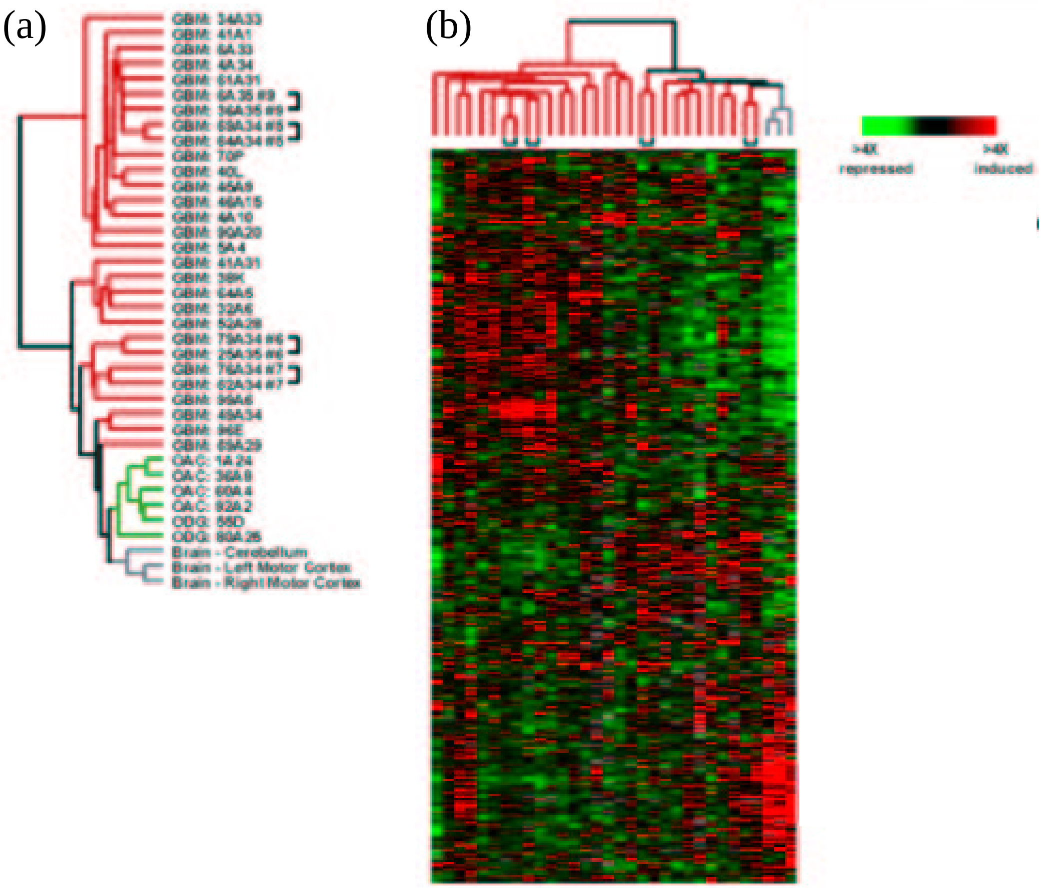

Figure 4.4: (A) Unsupervised hierarchical clustering of 38 samples. (B) Unsupervised hierarchical clustering of 32 samples. The data are displayed as a hierarchical cluster where rows represent genes and columns represent experimental samples. Colored pixels capture the magnitude of the response for any gene, where shades of red and green represent induction and repression, respectively, relative to the median for each gene. Black pixels reflect no change from the median and gray pixels represent missing data. The pictures were taken from (Liang et al., 2005).

5 Dimensional Reduction

This chapter introduces the theory of dimensional reduction and is composed of five parts. The first part (section 5.1) introduces the general idea of dimensional reduction and mentions the two principal categories of algorithms for dimensional reduction, namely (a) vector quantization and clustering techniques and (b) dimensional data reduction techniques. The latter are described in section 5.2 together with two algorithm used in this work, principal component analysis (PCA) and locally linear embedding (LLE). Section 5.3 introduces the concepts of vector quantization and clustering and describes a further algorithm used in this work, self-organizing maps (SOM). Section 5.4 shows the quantitative measurements that are used in this work for the evaluation of the low-dimensional projections. Finally, some properties of high-dimensional data are introduced in section 5.5.

5.1 Introduction to Data and Dimensional Reduction

With the increased abilities for automated data collection made possible by modern technology, the size of data collections in real-world applications has been growing in recent years. As a result, many real-world data comprise sets of records characterized by many variables, the number of which determines the dimension of the data (see also section 4.2). Typically, the data is said to be multivariate when the data dimension is larger than three. Let D denote the data dimension and v denote the number of records (also termed data points, items or entities) of a multivariate data set.

In many practical applications, multivariate data in $\mathbb{R}^D$ usually have a true (or intrinsic) dimensionality much lower than D. This fact is intuitively illustrated in Fig. 5.1, where a data set with $D = 3$ and $v = 4$ is visualized. All the points have the same value of the z-coordinate. In other words there is no variance in the z direction. The points differ from each other only with respect to the x and y coordinates. It follows that this data set has an intrinsic dimensionality equal to two.

Another example is given by the synthetic face data set taken from (Tenenbaum et al., 2001). It comprises 698 images of 64×64 pixels that were obtained by varying the angles of rotation of a synthetic face from up to down and from left to right. This data set can be treated as a set of 698 data points in a 64×64-dimensional data space. Because t his data space is obtained by varying two angles of rotation, i. e. two degrees of freedom, one can expect that the data set possesses an intrinsic dimensionality much smaller than that of the data space. The data set is projected onto two dimensions using Isomap, an

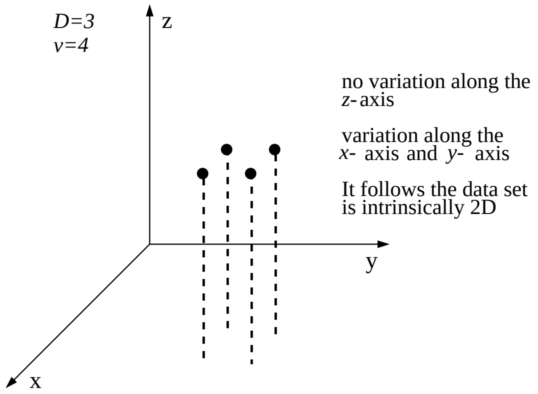
D=3
v=4
z
y
x
no variation along the z-axis
variation along the x- axis and y- axis
It follows the data set is intrinsically 2D

Figure 5.1: Simple data set with dimension $D = 3$ and number of points $v = 4$. The coordinates of the points do not vary along the z-axis but only along the x-axis and y-axis. Therefore, the intrinsic dimension of the data set is two.

algorithm for nonlinear dimensional reduction (Tenenbaum et al., 2000), and the result is visualized in Fig. 5.2. Here the images are plotted as points in a two-dimensional space. One can see that the projected points form a smooth plane which is highly correlated with the face rotation. Specifically, the rotation from left to right of the face is mapped along the horizontal axis, while the rotation from up to down is mapped along the vertical axis. It is evident that the data set is intrinsically two-dimensional although it is embedded in the 64×64 data space. This is equivalent to saying that the changes induced in the images by the rotations are highly correlated and described by two variables.

In general, it may hence be sensible to map high-dimensional data into a lower-dimensional space in order to eliminate redundant relationships between variables. Consider a data set of v data points $\{\mathbf{x}\}_v = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_v\}$ lying in a D-dimensional space $\mathbb{R}^D$ and forming some distribution. The goal is to approximate the unknown distribution spanned by the data points in $\mathbb{R}^D$ such that data points produced in a lower-dimensional space by some dimensional reduction process are "close" (in some well-defined sense) to data points from the generating distribution (Cherkassky and Mulier, 1998). For this purpose there exist two types of methods:

- **Dimensional reduction techniques** These algorithms are used to find low-dimensional representations of high-dimensional data. Each data point is mapped to a point in a lower-dimensional space while the global statistical information is preserved as much as possible. In general, given a set of v data points $\{\mathbf{x}\}_v = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_v\}$ in D-dimensional space $\mathbb{R}^D$, algorithms for dimensional reduction aim at mapping the data to a d-dimensional space $\mathbb{R}^d$, where $d < D$,

$$G(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^d \tag{5.1}$$

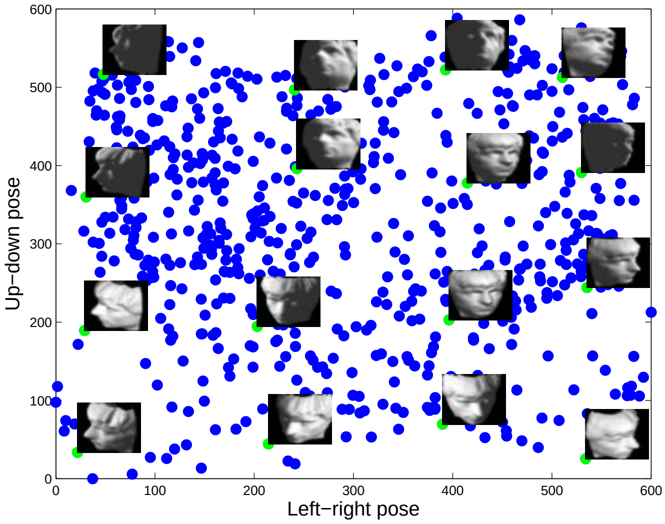

Figure 5.2: Set of 698 images (each with 64×64 pixels) of a rotating face projected into a two-dimensional space. Some points are visualized together with the corresponding images. One can see that the rotation of the face spans a two-dimensional structure in the 64×64 image space by virtue of the two degrees of freedom, the left-right and up-down rotation movements. The face dataset was taken from (Tenenbaum et al., 2001).

producing a low-dimensional encoding $\mathbf{y}_i = G(\mathbf{x}_i)$ for each input vector $\mathbf{x}_i$.

- **Vector quantization and clustering** Here, the objective is to approximate the distribution of a given D-dimensional data set $\{\mathbf{x}\}_v$ using m reference vectors (also called prototypes) $\{\mathbf{u}\}_m = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m\}$, where typically $m < v$.

  One can distinguish between *vector quantization* (VQ) and *clustering*. VQ methods aim to minimize a well-defined approximation (quantization) error when the number of prototypes m is fixed a priori. Clustering methods have a more vague objective of finding interesting non-overlapping grouping of data points. Each group or cluster should contain similar data points and data points from different groups should not be similar. Often clustering algorithms also represent each group by a reference vector, and such methods have a strong similarity to VQ.

5.2 Dimensional Data Reduction

The goal of dimensional reduction is to obtain a compact, accurate representation of the data that reduces or eliminates drastically redundant components (Hinton and Sejnowski, 1999). At the same time, the reduction of the number of dimensions should not result in loss of information relevant to the task at hand. Thus, there is a trade-off between the advantages of the reduced dimensionality and the loss of information due to dimensional reduction.

The most important applications of dimensional data reduction can be divided into three groups (Verbeek, 2004):

- **Visualization of high-dimensional data for explorative data analysis.** Reducing the dimension of the data to two or three dimensions enables the user to plot the data on a computer screen. This can allow to extract the relevant information in the data and detect dominant relationships between variables.

- **Data compression.** Using fewer dimensions to express the data allows for a more compact data storage and data transmission using less bandwidth.

- **Increasing efficiency and performance of subsequent steps of data analysis.** Dimensional data reduction can increase the efficiency of operation and training of automatic classifiers and regression functions, which typically have an execution time and training time at least linear with the number of dimensions of the data.

  In addition, a lower number of dimensions may increase the performances of the system. This effect, which may appear counter-intuitive, can be explained by the *curse of dimensionality* (Bellmann, 1961). This refers to the fact that the number of data points required to estimate a data distribution grows exponentially with the dimensionality of the data. Consequently, the data distribution of a certain number of data points in a low-dimensional space may be estimated more reliably that the distribution of the same number of points in a high-dimensional space.

In this work the methods for dimensional reduction are exclusively used for the first purpose, the visual exploration of high-dimensional data.

The algorithms for dimensional data reduction can be divided into *linear* and *nonlinear* ones. Linear methods attempt to reveal linear combination of variables best approximating the original data structure. The classical linear techniques are Principal Component Analysis (PCA) (Jollife, 1986) and Multi-Dimensional Scaling (MDS) (Kruskal and Wish, 1978). PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space. Classical MDS finds an embedding that preserves the inter-point distances and is equivalent to PCA when those distances are Euclidean (Tenenbaum et al., 2000).

Methods for nonlinear data reduction have received growing attention in the literature in recent years. They are able to reveal nonlinear relationships between variables in a high dimensional data set. Besides traditional techniques such as nonlinear MDS (Kruskal, 1964) and Sammon's nonlinear mapping (Sammon, 1969), several new algorithms for nonlinear dimensional reduction have recently been published, including Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Isomap (Tenenbaum et al., 2000), Curvilinear Component Analysis (CCA) (Demartines and Herault, 1997), Curvilinear Distance Analysis (CDA) (Lee et al., 2000), Distance-preserving Projection (Yang, 2004), Hessian Eigenmaps (Donoho and Grimes, 2003), Laplacian Eigenmaps (Belkin and Niyogi, 2003) and charting (Brand, 2002).

Some of these methods can be divided into *local* and *global* approaches. Local approaches (e. g. LLE, Laplacian Eigenmaps) attempt to preserve the local geometry of the data. Specifically, they seek to map nearby points on the manifold to nearby points in the low-dimensional representation. Global approaches (e. g. Isomap) attempt to preserve geometry at all scales, mapping nearby points on the manifold to nearby points in low-dimensional space, and faraway points to faraway points.

The principal advantages of the global approach are that it tends to give a more faithful representation of the data's global structure, and that its metric-preserving properties are better understood theoretically. The local approaches have two principal advantages: (1) computational efficiency as they involve only sparse matrix computations which may yield a polynomial speed-up; (2) representational capacity: they may give useful results on a broader range of manifolds, whose local geometry is close to Euclidean, but whose global geometry may be not (de Silva and Tenenbaum, 2002).

In this work two algorithm for dimensional reduction are employed for visual data exploration of DCE-MRI and microarray data sets, namely LLE and PCA.

The choice of employing PCA is motivated by the fact that this is a well established linear technique that has been employed in a broad range of different applications. In addition, its application is straightforward as the PCA algorithm does not require any parameter to be set.

The LLE is a recently proposed algorithm able to handle nonlinear data structures and is characterized by some attractive properties such that uniqueness of the solution, low number of input parameters and of preservation of the data topology in the projected space. These properties, in particular the latter one, make LLE interesting for the purpose of visual data exploration. In addition, the applications of the LLE algorithm in real-world domains are still relatively few. LLE has already successfully been applied to microarray data (Chao and Lihui, 2004), membrane protein data (Wang et al., 2005), and human face data analysis (Zhang et al., 2004b; Graf and Wichmann, 2002; Zhang et al., 2004a).

The PCA and LLE algorithms are described in the following two sections.

5.2.1 Principal Component Analysis (PCA)

This description of the PCA algorithm is based on the content of (Bishop, 1995). Imagine having v vectors $\{\mathbf{x}\}_v$ in a D-dimensional space. The goal is to map such vectors to a d-dimensional space, where $d < D$. Without loss of generality, a generic vector $\mathbf{x}$ can be represented by a linear combination of D orthonormal vectors $\{\mathbf{m}\}_D$

$$\mathbf{x} = \sum_{j=1}^{D} z_j \mathbf{m}_j \tag{5.2}$$

where explicit expressions for the coefficients z_j are given by

$$z_j = \mathbf{m}_j^T \mathbf{x}. \tag{5.3}$$

Now only a subset $d < D$ of the basis vectors $\{\mathbf{m}\}_D$ are retained, so that only d coefficients $\{z\}_d$ are used. The remaining coefficients are replaced by constants $\{b\}_{(D-d)}$ so that the vector $\mathbf{x}$ is approximated by the expression

$$\tilde{\mathbf{x}} = \sum_{j=1}^{d} z_j \mathbf{m}_j + \sum_{j=d+1}^{D} b_j \mathbf{m}_j = \mathbf{z} + \sum_{j=d+1}^{D} b_j \mathbf{m}_j. \tag{5.4}$$

This represents a form of dimensionality reduction since the vector $\mathbf{x}$ is approximated by the d-dimensional vector $\mathbf{z}$. If one considers the whole data set, one wishes to find the basis vectors $\{\mathbf{m}\}_D$ and the coefficients $\{b\}_{(D-d)}$ such that the vectors $\{\mathbf{z}\}_v$ give the best approximation to the original vectors $\{\mathbf{x}\}_v$. The error in the vector $\mathbf{x}$ introduced by

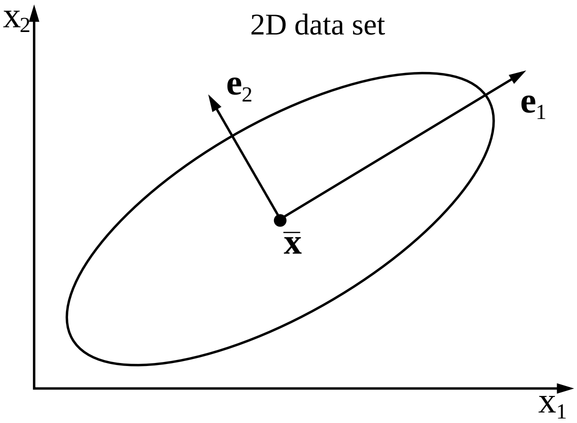

Figure 5.3: Schematic illustration of a two-dimensional data set along with the two principal components $\mathbf{e}_1$ and $\mathbf{e}_2$. The directions of these eigenvectors are those along which the variance of the cluster is maximal. The one-dimensional projection of the data set is obtained by first subtracting off the mean $\bar{\mathbf{x}}$, and then projecting onto $\mathbf{e}_1$.

the dimensional reduction is given by

$$\mathbf{x} - \tilde{\mathbf{x}} = \sum_{j=d+1}^{D} (z_j - b_j)\mathbf{m}_j \tag{5.5}$$

One considers as best approximation the one which minimizes the sum of the squares of the errors over the whole data set. Hence, one minimizes

$$E_d = \frac{1}{2}\sum_{n=1}^{v} ||\mathbf{x}_n - \tilde{\mathbf{x}}_n||^2 = \frac{1}{2}\sum_{n=1}^{v}\sum_{j=d+1}^{D} (z_j^n - b_j)^2. \tag{5.6}$$

If one sets the derivative of E_d with respect to b_j to zero one finds

$$b_j = \frac{1}{v}\sum_{n=1}^{v} z_j^n = \mathbf{m}_j^T \bar{\mathbf{x}} \tag{5.7}$$

where one has defined the mean vector $\bar{\mathbf{x}}$ to be

$$\bar{\mathbf{x}} = \frac{1}{v}\sum_{i=1}^{v} \mathbf{x}_i. \tag{5.8}$$

Using eq. (5.3) and (5.7) one can write E_d as

$$\frac{1}{2}\sum_{j=d+1}^{D}\sum_{i=1}^{v}\{\mathbf{m}_j^T(\mathbf{x}_i - \bar{\mathbf{x}})\}^2 = \frac{1}{2}\sum_{j=d+1}^{D} \mathbf{m}_j^T \mathbf{\Sigma} \mathbf{m}_j \tag{5.9}$$

where $\mathbf{\Sigma}$ is the covariance matrix of the set of vectors $\{\mathbf{x}\}_v$ and is given by

$$\mathbf{\Sigma} = \sum_{i=1}^{v}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{5.10}$$

One can demonstrate that the minimum of E_d occurs when vectors $\{\mathbf{m}\}_D$ are equal to the eigenvectors of the covariance matrix $\mathbf{\Sigma}$.

Finally, the d dimensional projection $\mathbf{y}_i$ of the input data point $\mathbf{x}_i$ is obtained by projecting ($\mathbf{x}_i - \bar{\mathbf{x}}$) along those eigenvectors $\mathbf{e}_i$ ($i = 1, .., d$) (also called principal components) corresponding to the d largest eigenvalues of $\mathbf{\Sigma}$, i. e.

$$\mathbf{y}_i = [(\mathbf{x}_i - \bar{\mathbf{x}}) \cdot \mathbf{e}_1, (\mathbf{x}_i - \bar{\mathbf{x}}) \cdot \mathbf{e}_2, ..., (\mathbf{x}_i - \bar{\mathbf{x}}) \cdot \mathbf{e}_d]. \tag{5.11}$$

The PCA technique is illustrated schematically in Fig. 5.3 for the case of a two-dimensional data set.

5.2.2 Locally Linear Embedding (LLE)

This section describes the Locally Linear Embedding (LLE) algorithm and the content is based on the information contained in (Lawrence and Roweis, 2003). The LLE algorithm recovers global nonlinear structure from locally linear fits. The LLE algorithm, summarized in Fig. 5.4, is based on simple geometric intuitions. LLE assumes that each data point and its neighbors lie on a locally linear patch and then applies the patch in a low space to generate a neighbor-preserving embedding. Suppose the input data consists of v real-valued vectors $\{\mathbf{x}\}_v$, each of dimensionality D, sampled from some underlying manifold. The LLE algorithm is composed of three steps:

Step 1

Provided that there is sufficient data (such that the manifold is well-sampled), one expects each data point and its neighbors to lie on or close to a locally linear patch of

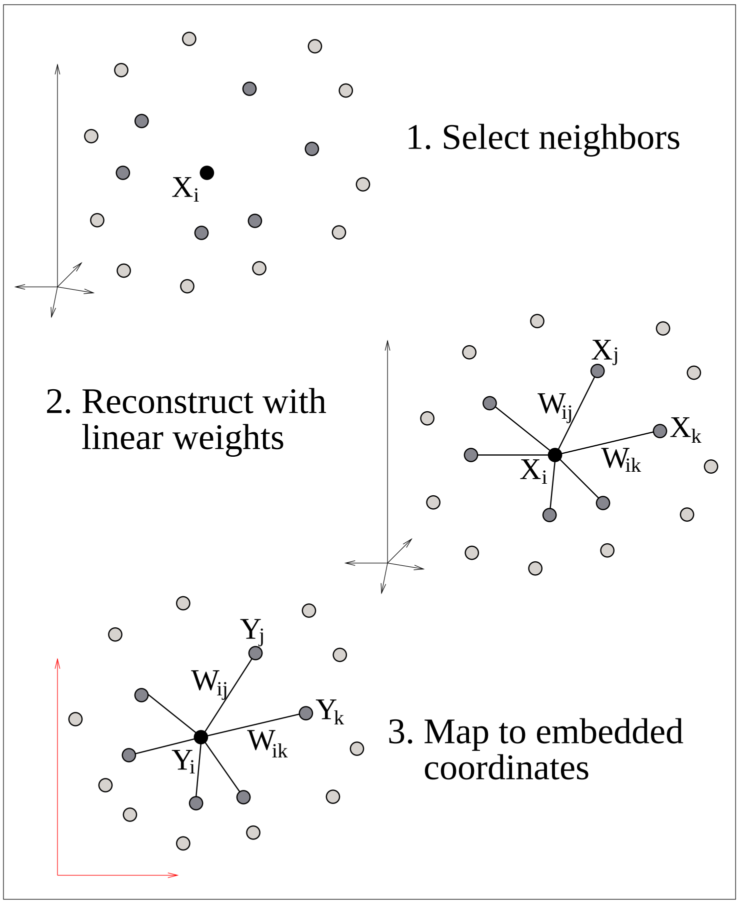
1. Select neighbors
X_i
2. Reconstruct with linear weights
X_j
W_ij
X_k
W_ik
X_i
Y_j
W_ij
Y_k
W_ik
Y_i
3. Map to embedded coordinates

Figure 5.4: Schema of the three steps of LLE.

the manifold, i. e. the relationships between neighbors are locally linear. The first step consists in calculating the n nearest neighbors of each data point $\mathbf{x}_i$.

Step 2

The local geometry of these patches is characterized by linear coefficients that reconstruct each data point from its neighbors. Reconstruction errors are measured by the cost function

$$\Psi(W) = \sum_{i=1}^{v} |\mathbf{x}_i - \sum_{j=1}^{n} W_{ij}\mathbf{x}_j|^2 \tag{5.12}$$

which adds up the squared distances between all the data points and their reconstructions. The weight W_{ij} summarizes the contribution of the jth neighbor to the reconstruction of $\mathbf{x}_i$. To compute the weights, one minimizes the cost function $\Psi(W)$ subject to two constraints: first, that each data point $\mathbf{x}_i$ is reconstructed only from its n neighbors, enforcing $W_{ij} = 0$ if $\mathbf{x}_j$ does not belong to the set of neighbors of $\mathbf{x}_i$; second, that the rows of the weights matrix sum to one: $\sum_{j=1} W_{ij} = 1$. The optimal weights are found by solving a least-squares problem. With the given constraints, eq. (5.12) can be simplified to a linear system and the weights can be computed in closed form as follows:

$$\Psi(W) = \sum_{i=1}^{v} |\mathbf{x}_i - \sum_{j=1}^{n} W_j\mathbf{x}_j|^2 = \sum_{jk} W_j W_k C_{jk}. \tag{5.13}$$

In the second identity, the term

$$C_{jk} = (\mathbf{x}_i - \mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_k) \tag{5.14}$$

is the local covariance matrix. The weights which minimize the error function of eq. (5.12) are given by:

$$W_j = \frac{\sum_{k=1}^{n} C_{jk}^{-1}}{\sum_{lm} C_{lm}^{-1}},\ l, m \in \{1, .., n\}. \tag{5.15}$$

In some cases, for example if the number of neighbors is greater than the input dimension $(n > D)$, it arises that the matrix $\mathbf{C}$ is singular or nearly singular and the solution of eq. (5.13) is not unique. In this case the matrix $\mathbf{C}$ must be conditioned by adding a small multiple of the identity matrix (Lawrence and Roweis, 2003):

$$C_{ij} \leftarrow C_{ij} + \delta_{ij}\Gamma \tag{5.16}$$

where Γ is defined as

$$\Gamma = \frac{\mathrm{Tr}(C)}{n}\Delta^2. \tag{5.17}$$

The term Δ is a regularization term set by the user and its value must be much smaller than 1.

The constrained weights that minimize the reconstruction errors obey an important symmetry: for any particular data point, they are invariant to rotations, rescalings, and translations of that data point and its neighbors. By symmetry, it follows that the reconstruction weights characterize geometric properties of each neighborhood. Note that the invariance to translations is specifically enforced by the sum-to-one constraint on the rows of the weight matrix.

Suppose the data lie on or near a smooth nonlinear manifold of lower dimensionality $d < D$. To a good approximation then, there exists a linear mapping, consisting of a translation, rotation and rescaling, that maps the high-dimensional coordinates of each neighborhood to global internal coordinates on the manifold.

By design, the reconstruction weights $\{W_i\}_n$ reflect intrinsic geometric properties of the data that are invariant to exactly such transformations. Therefore, one expects their characterization of local geometry in the original data space to be equally valid for local patches on the manifold. In particular, the same weights $\{W_{ij}\}(j = 1, ..n)$ that reconstruct $\mathbf{x}_i$ in D dimensions should also reconstruct its embedded manifold coordinates in d dimensions. LLE constructs a neighborhood-preserving mapping based on the above idea.

Step 3

In the final step of the algorithm, each high-dimensional data point $\mathbf{x}_i$ is mapped to a low-dimensional vector $\mathbf{y}_i$ representing global internal coordinates on the manifold. This is done by choosing d-dimensional coordinates $\{\mathbf{y}\}_v$ to minimize the embedding cost function

$$\Phi(\mathbf{y}) = \sum_{i=1}^{v} |\mathbf{y}_i - \sum_{j=1}^{n} W_{ij}\mathbf{y}_j|^2. \tag{5.18}$$

This cost function, like the previous one, is based on locally linear reconstruction errors, but here the weights W_{ij} are fixed while optimizing the coordinates $\{\mathbf{y}\}_v$. The embedding cost in eq. (5.18) defines a quadratic form in the vectors $\{\mathbf{y}\}$. Subject to the constraints $\frac{1}{v}\sum_{i=1}^{v}\mathbf{y}_i\mathbf{y}_i^T = I$ and $\sum_{i=1}^{v}\mathbf{y}_i = 0$, the solution of this quadratic function is given by the bottom $d+1$ eigenvectors of the sparse matrix

$$\mathbf{S} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W}). \tag{5.19}$$

These eigenvectors are associated with the $d+1$ smallest eigenvalues of $\mathbf{S}$. The bottom eigenvector is related to the smallest eigenvalue whose value is closest to zero. This eigenvector is the unit vector with all equal components and is discarded. Let $\mathbf{E}$ be the $v \times d$ matrix of the eigenvectors of matrix $\mathbf{S}$ associated with the d smallest eigenvalues (excluding the smallest one). The projection $\mathbf{y}_i$ of the data point $\mathbf{x}_i$ is given by the ith line of matrix $\mathbf{E}$.

5.3 Vector Quantization and Clustering

The problem of clustering is that of dividing a data set into a number of groups, called clusters, based on some measure of similarity (Cherkassky and Mulier, 1998). The goal is to find a set of clusters such that samples within a cluster are more similar than samples from different clusters.

Cluster analysis differs from vector quantization design in that the similarity measure for clustering is chosen subjectively based on its ability to create meaningful clusters. The clusters can be organized *hierarchically* and described in terms of a hierarchical tree structure, or they can be purely *partitional* (Cherkassky and Mulier, 1998). Hierarchical clustering, that has already been mentioned in section 4.3.3, can be either *agglomerative* (bottom up) or *divisive* (top down). An agglomerative hierarchical method initially regards each sample as one cluster and gradually merges these clusters into larger clusters until all samples are in a single cluster (the root node in the hierarchical cluster tree). A divisive hierarchical method starts with a single cluster containing all the data and recursively splits clusters into subclusters. Partitional methods can be further classified into two groups, namely *hard* and *soft clustering*. In hard clustering, each sample is assigned to one and only one cluster. In soft clustering, each sample is associated with several clusters with certain probabilities.

There exist many clustering algorithms and an exhaustive list is out of the scope of this work. A comprehensive survey on clustering algorithms can be found in (Xu and Wunsch, 2005).

Vector quantization algorithms assign a set of data points into m clusters by minimizing a certain error function. The sum of squared errors in one of the most widely used criteria. Suppose one has a set of v data points $\{\mathbf{x}\}_v$ and one wants to organize them into m clusters $\{C_1, ..., C_k\}$. The squared error criterion is defined as (Xu and Wunsch, 2005)

$$J(\mathbf{\Upsilon}, \mathbf{U}) = \sum_{i=1}^{m} \sum_{j=1}^{v} \upsilon_{ij} ||\mathbf{x}_j - \mathbf{u}_i||^2 \tag{5.20}$$

where $\mathbf{\Upsilon}$ denotes a partition matrix in which each element is

$$\upsilon_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in C_j \\ 0 & \text{otherwise.} \end{cases} \quad \text{with } \sum_{i=1}^{m} \upsilon_{ij} = 1 \; \forall j;$$

$\mathbf{u}_i$ (ith reference vector) denotes the mean vector of the ith cluster and $\mathbf{U}$ is the reference vector matrix.

The K-means algorithm is the best-known squared error-based algorithms for vector quantization (Forgy, 1965). In this work, the Self-Organizing maps (SOM) algorithm (Kohonen, 2000), a very popular neural network-based vector quantization technique, is employed. The rationale for this choice is that SOM is particularly suitable for the exploratory visualization of multivariate data (Similä, 2005). For an overview about methods for the visualization of SOM data see (Vesanto, 1999). The principal steps of the SOM algorithm are introduced in the following section.

5.3.1 Self-Organizing Map (SOM)

The objective of self-organizing maps (SOM) is to represent high-dimensional data $\{\mathbf{x}\}_v$ with a set of reference vectors $\{\mathbf{u}\}_m$ where each reference vector $\mathbf{u}_i$ is linked with a node r_i in a typically two-dimensional grid. The grid can be chosen to be rectangular, hexagonal or even irregular. Each data point $\mathbf{x}_i$ is mapped to the node related to its nearest reference vector. A particularly important property of SOM is to preserve the topology of the data space, i. e. close points in the original space are mapped onto nearby nodes in the grid (Yin, 2002).

In each learning step, an input vector $\mathbf{x}_i$ is selected and its nearest reference vector $\mathbf{u}_i$ is identified by evaluating

$$i = \arg\min_j \{||\mathbf{u}_j - \mathbf{x}_i||^2\}. \tag{5.21}$$

The node in the grid linked to $\mathbf{u}_i$ is called *winner node* and is denoted by r_i. Let N_i denote the reference vector $\mathbf{u}_i$ and its nearest reference vectors, determined with respect to the grid topology. Specifically, the nodes of the nearest reference vectors are the closest to r_i in the low-dimensional grid (see Fig. 5.5) The reference vector $\mathbf{u}_i$ and its neighbors are then updated using the following formula:

$$\mathbf{u}_k(t+1) = \mathbf{u}_k(t) + h_{ik}(t)[\mathbf{x}_i - \mathbf{u}_k], \; k \in N_i. \tag{5.22}$$

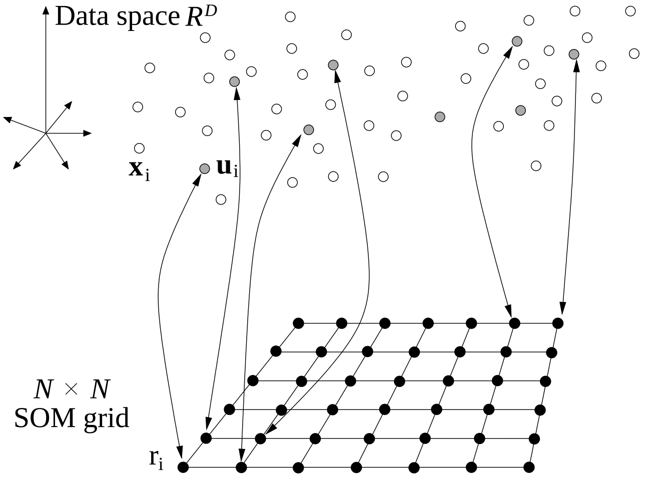

Figure 5.5: The data set (group of white points) $\{\mathbf{x}\}_v$ is approximated by a set of reference vectors (gray points) $\{\mathbf{u}\}_m$, where $m = N \times N$. Each reference vector $\mathbf{u}_i$ is connected with a node r_i in the SOM grid. During the training phase, for each data point $\mathbf{x}_i$, the nearest reference vector $\mathbf{u}_i$ and its topological neighbors according to the SOM grid are moved towards $\mathbf{x}_i$, as described by eq. (5.22).

where $t = 0, 1, 2, ...$ is an integer coordinate which represents the number of the current iteration in the training process. The total number of iterations is set by the user. The function $h_{ik}(t)$ acts as a neighborhood function on the grid. For the solution to converge, it is necessary that $h_{ik} \to 0$ when $t \to \infty$. In the literature, h_{ik} is frequently defined in terms of the Gaussian function,

$$h_{ik}(t) = \alpha(t) exp\left(-\frac{||r_i - r_k||^2}{2\sigma^2(t)}\right), \; k \in N_i \tag{5.23}$$

where r_i and r_k denote the nodes relative to the reference vectors $\mathbf{u}_i$ and $\mathbf{u}_k$. The function $\alpha(t)$ is the *learning rate factor*, and the parameter $\sigma(t)$ defines the width of the neighborhood function. Both $\alpha(t)$ and $\sigma(t)$ are monotonically decreasing functions of time.

5.4 Evaluation of the Dimensional Reductions

This section illustrates the numerical quantities that are used in this work for the quantitative evaluation of the dimensional reductions obtained using the PCA, SOM and LLE algorithms. In particular, the dimensional reductions obtained by PCA and LLE are quantitatively evaluated by the measurements from section 5.4.1 to section 5.4.4, while the SOM projection is evaluated by an ROC curve analysis (section 5.4.5).

5.4.1 Fisher Criterion

The following part is based on the content of (Bishop, 1995). The Fisher criterion quantifies the compactness and the distance between two clusters in a two-dimensional space. In this work, the DCE-MRI and microarrays data sets (both are two-classes data sets) are projected onto a two-dimensional space for visual data exploration and the Fisher criterion is employed to quantify the separation between the benign and malignant clusters.

Consider a problem in which there are v_1 points of class $\mathcal{C}_1$ and v_2 points of class $\mathcal{C}_2$. Let $\mathbf{m}_1$ and $\mathbf{m}_2$ be the mean vectors of the two classes, i. e.

$$\mathbf{m}_i = \frac{1}{v_i} \sum_{k \in \mathcal{C}_i} \mathbf{x}_k, \; i = 1, 2. \tag{5.24}$$

Let $\mathbf{j}$ be the vector that can best separate the two clusters. It is defined as

$$\mathbf{j} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \tag{5.25}$$

where $\mathbf{S}_W^{-1}$ is the total within-class covariance matrix which is given by

$$\mathbf{S}_W = \sum_{\mathbf{x} \in \mathcal{C}_1} (\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{\mathbf{x} \in \mathcal{C}_2} (\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T. \tag{5.26}$$

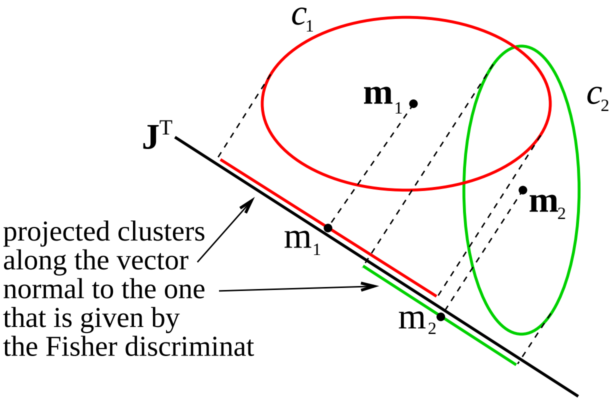
C_1
C_2
$\mathbf{m}_1$
$\mathbf{m}_2$
$\mathbf{J}^{\mathrm{T}}$
m_1
m_2
projected clusters along the vector normal to the one that is given by the Fisher discriminat

Figure 5.6: The clusters C_1 and C_2 are projected along $\mathbf{j}^T$. The vector $\mathbf{j}$ is defined in eq. (5.25). By computing the means and variances of the projected clusters as shown in eq. (5.27) and (5.28), respectively, one can compute the value of the Fisher criterion by eq. (5.29).

Eq. (5.25) is known as Fisher's linear discrimant.

Considering the projections of the two clusters along $\mathbf{j}^T$ (see Fig. 5.6), their mean and variance are respectively given by

$$m_i = \mathbf{j}^T \mathbf{m}_i,\ i = 1,2 \text{ and} \tag{5.27}$$

$$\sigma_i^2 = \sum_{\mathbf{x} \in \mathcal{C}_i} (\mathbf{j}^T \mathbf{x}_k - m_i)^2,\ i = 1,2. \tag{5.28}$$

Finally, the Fisher criterion $F_{1,2}$ between the two clusters is defined as (Bishop, 1995):

$$F_{1,2} = \frac{(m_2 - m_1)^2}{\sigma_1^2 + \sigma_2^2} \tag{5.29}$$

The larger the value of $F_{1,2}$, the less the two clusters overlap. This is desirable as C_1 and C_2 typically represent data of different entities that a computer should distinguish as much as possible.

5.4.2 Stress

Stress quantifies the overall deviation between the distances (i.e. the extent to which they differ) in the original and dimensionally reduced space (Hjaltason and Samet, 2003). Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two data points and $\mathbf{y}_1$ and $\mathbf{y}_2$ their projection in the embedding space; their distance in the original and in the embedding space are indicated by $d(\mathbf{x}_1, \mathbf{x}_2)$ and $\delta(\mathbf{y}_1, \mathbf{y}_2)$, respectively. Stress is typically defined in terms of variance as

$$\mathsf{ST} = \frac{\sum_{\mathbf{x}_1,\mathbf{x}_2} (\delta(\mathbf{y}_1, \mathbf{y}_2) - d(\mathbf{x}_1, \mathbf{x}_2))^2}{\sum_{\mathbf{x}_1,\mathbf{x}_2} d(\mathbf{x}_1, \mathbf{x}_2)^2}. \tag{5.30}$$

Note that the scales of the original and embedded spaces need to be equal. The value of ST lies between zero and one. The lower the value of stress, the better the distances among data points are preserved in the embedding, i. e. the process of dimensional reduction is more accurate.

5.4.3 Neighborhood Preservation

Neighborhood Preservation (NP) quantifies the average percentage of neighbors in the original space which are preserved after the dimensional reduction. It is defined as

$$\mathsf{NP}(t) = \frac{1}{v} \sum_{i=1}^{v} \mathsf{p}_t(\mathbf{x}_i) \tag{5.31}$$

where $\mathsf{p}_t(\mathbf{x}_i)$ is the percentage of the t-nearest neighbors of point $\mathbf{x}_i$ in the original space which are preserved in the low-dimensional space. For example, if only 25% of its t-nearest neighbors are preserved in the embedding, then $\mathsf{p}_t(\mathbf{x}_i)$ will equal 0.25. The value of NP lies between zero and one. A high value of NP (close to one) denotes a good preservation of the local relations between data points in the low-dimensional space and this is therefore desirable.

5.4.4 Trustworthiness

The trustworthiness measure has been proposed in (Kaski et al., 2003) and also employed in (Venna and Kaski, 2005). It quantifies how trustworthy is a projection of a high-dimensional data set onto a low-dimensional space, specifically a projection is trustworthy if the set of the t nearest neighbors of each data point in the low-dimensional space are also close-by in the original space.

More formally, the trustworthiness measure M is defined as (Venna and Kaski, 2005)

$$M(t) = 1 - \frac{2}{vt(2v - 3t - 1)} \sum_{i=1}^{v} \sum_{j \in U_t(i)} (r(i, j) - t), \tag{5.32}$$

where $r(i, j)$ is the rank of the data point j in the ordering according to the distance from i in the original data space, and $U_t(i)$ denotes the set of those data points that are among the t-nearest neighbors of the data point i in the low-dimensional space but not in the original space. The maximal value that trustworthiness can take is equal to one. The closer $M(t)$ is to one, the better the low-dimensional space describes the original data.

5.4.5 Receiver Operating Characteristics (ROC)

Receiver operating characteristic (ROC) graphs are a useful technique for visualizing and comparing the performance of different classifiers (Fawcett, 2004). Their usage is very common in medical decision making (Hanley and McNeil, 1982). ROC analysis has appeared to offer more robust evaluation of the relative performance of different classifier than traditional comparison of relative errors. Rather than considering raw errors, ROC analysis decomposes performance into true and false positive rates (Webb and Ting, 2005).

Note that in this work the ROC analysis is not used to compare the performances of different classifiers, but to evaluate the overlap between the benign and malignant clusters (of DCE-MRI data) in the projected space obtained by PCA, SOM and LLE, as described in section 6.2.2, second approach.

The principles of the ROC curve are explained considering a two class problem (classes P (positive) and N (negative)). To evaluate classifiers performance by ROC, first a set of samples in a test data set has to be labeled by a human user according to a certain property, i. .e. each sample is labeled as belonging to either class P or class N. For example, in this work the samples are the time-series associated with each voxel (DCE-MRI data) and the patients characterized by a sequence of gene expressions (microarray data) These time-series are labeled as benign or malignant if they belong to a benign or malignant lesion, respectively (see section 6.2.2.2), In the microarray data, the patients are labeled as benign or malignant according to the absence or presence of metastasis.

This set of samples together with their labels are called the *gold standard*. Next, each of these samples is labeled by an automatic classifier and the ROC analysis of the classifier performance is based on observation of four types of outcome:

True Positives (TP) : number of positive samples (according to the gold standard) that are correctly classified as positive

False Positives (FP) : number of negative samples (according to the gold standard) that are wrongly classified as positive

True Negatives (TN) : number of negative samples (according to the gold standard) that are correctly classified as negative

False Negatives (FN) : number of positive samples (according to the gold standard) are wrongly classified as negative.

These four quantities can be used to compute four indices useful for comparing the classifier performance.

Sensitivity (SE) or True-Positive-Fraction (TPF) : percentage of positive samples among all positive samples that were correctly labeled as positive by the classifier

$$\mathrm{SE} = \mathrm{TPF} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Specificity (SP) : percentage of negative samples among all negative samples that were correctly labeled as negative by the classifier

$$\mathsf{SP} = \frac{\mathsf{TN}}{\mathsf{TN} + \mathsf{FP}}.$$

False-Positive-Fraction (FPF) : percentage of negative samples among all negative samples that were wrongly labeled as positive by the classifier

$$\mathsf{FPR} = \frac{\mathsf{FP}}{\mathsf{FP} + \mathsf{TN}} = 1 - \frac{\mathsf{TN}}{\mathsf{TN} + \mathsf{FP}} = 1 - \mathsf{SP}.$$

Positive-Predictive-Value (PPV) : ratio of the number of positive samples labeled correctly as positive by the classifier and the total number of samples classified as positive

$$\mathsf{PPV} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}.$$

The classifier typically classifies the samples according to a certain decision rule that depends on a parameter t_c. Each (SE,SP) pair is function of t_c and, in turn, different values of t_c produce different (SE,SP) pairs.

5.4.5.1 ROC Graph

This section is based on the content of (Fawcett, 2004). The ROC graph is obtained by plotting SE on the y-axis (vertical) and 1−SP (horizontal) on the x-axis. An ROC curve depicts relative trade-offs between benefits (SE) and costs (1−SP) (Fawcett, 2004). A ROC graph with five classifiers labeled A through E is shown in Fig. 5.7.

There are several important points in ROC space. The point (0,0) represents the strategy of never issuing a positive classification; such a classifier labels correctly all the negative samples as negative (1−SP$= 0$) but labels erroneously all the positive samples as negative (SE$= 0$).

The point (0,1) represents the perfect classification. All the positive samples are correctly classified as positive (SE$= 1$) and no negative samples are classified as positive (1−SP$= 0$).

Informally, one point in ROC space is better than an other if it is located towards the northwest (SE is higher or (1−SP) is lower or both) of the first one. In Fig. 5.7, classifier A has better performance than classifier B. Classifiers appearing on the left-hand side of an ROC graph, near the x-axis, may be thought of as "conservative": they make positive classifications only with strong evidence so they make few false positive errors, but often they have low SE values as well. Classifiers on the upper right-hand side of an ROC graph may be thought of as "liberal": they make positive classifications with weak evidence so they classify nearly all positives correctly, but classifies many

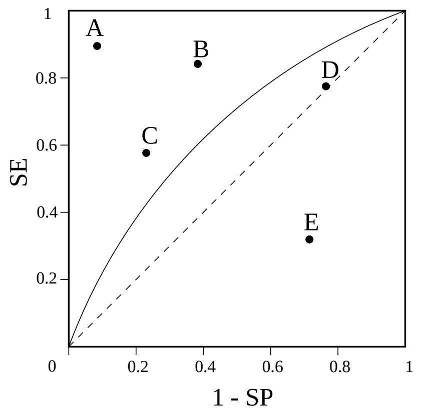

Figure 5.7: An ROC graph with five points.

negatives wrongly (Fawcett, 2004). In Fig. 5.7, classifier A is more conservative than classifier B and classifier D is more liberal than classifier C.

If the tuning parameter t_c of the decision rule of a classifier changes continuously, plotting all the corresponding pairs (SE, 1−SP) typically results in an arc-shaped curve like the one shown in Fig. 5.7.

5.4.5.2 Random Performance

The classifiers along the diagonal $y = x$ are those that randomly guess a class (like the point D in Fig. 5.7). For example, if a classifier randomly guesses the positive class 50% or 70% of the time, it can be expected to get 50% or 70% of the positives correct but its false positive rate will increase to 50% or 70%, yielding the point (0.5, 0.5) or (0.7, 0.7) in ROC space. Any classifier appearing under the diagonal $y = x$ performs worse than random guessing. Therefore, the classifiers usually appear in the upper left triangle (Fawcett, 2004). However, notice that any classifier that produces a point in the lower right triangle can be negated (i. e. its true positive classifications become false negative mistakes and its false positives become true negatives) to produce a point in the upper left triangle.

5.5 Properties of High-Dimensional Spaces

In this section some properties of high-dimensional data spaces are introduced. This section provides a theoretical background to the microarray data analysis carried out in chapter 8.

The introduced properties of high-dimensional spaces are somewhat unexpected as they are counter-intuitive when compared to what happens in two or three dimensions.

The following section is based on the content of (Verleysen, 2003) (for more information see also (Verleysen and Francois, 2005)) where two key questions in the analysis of high-dimensional data are addressed:

- what is the limit between low and high-dimensional spaces?
- what is the number of points required for learning in high-dimensional spaces?

5.5.1 Limit between low and high-dimensional spaces

Textbooks and scientific articles illustrate methods and other data analysis tools on one-, two- or three-dimensional examples, and measure their performances on higher-dimensional problems for which no representation is possible. It is also widely accepted that most real-world problems are high-dimensional. But where is the limit between low- and high-dimensional data?

To answer this question, let us take a few examples of concepts that are intuitive in dimensions up to three and expandable to higher ones. The dimension where our intuitive view is not valid anymore is the answer to our question. Consider the volume of a sphere in dimension d. This volume is given by

$$V(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d \tag{5.33}$$

where r is the radius. Looking at the graph of $V(d)$ when $r = 1$ (Fig. 5.8) leads to the surprising observation that the volume rapidly decreases towards zero when d increases. Our intuitive view of the volume of a sphere is thus misleading in high dimensions.

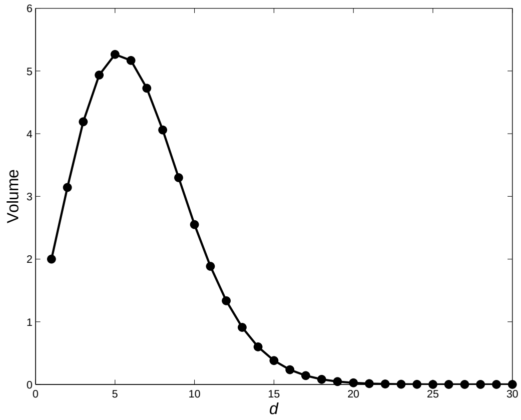

Figure 5.8: Volume of the sphere with $r = 1$ versus the dimension of the space.

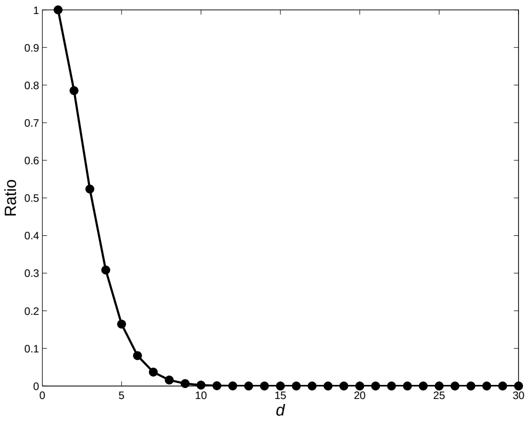

Figure 5.9: Ratio between the volume of a sphere and the volume of a cube (the length of an edge of the cube equals the diameter of the sphere) versus the dimension of the space.

One should compare this volume to a value that seems "natural" to humans. One way to do this is to plot the ratio between the volume of a sphere and the volume of a cube having edge equal to the diameter of the sphere. This ratio is shown in Fig. 5.9. Having in mind a segment, a circle and a sphere in one, two and three dimensions, respectively, in Fig. 5.9 one can see that the ratio will decrease with the dimension of the sphere. What is more surprising is that this ratio is below 10% when the dimension is lower than 5.

Another way to consider this problem is to plot the ratio between the volume of a sphere with radius 0.9 and a sphere with radius 1, versus the dimension (Fig. 5.10). Obviously, this ratio is equal to 0.9 raised to the power of d. The values plotted in Fig. 5.10 mean that 90% of the volume of a sphere in dimension greater than 20 is contained in the spherical shell whose thickness is 10% of the initial radius.

Another comment concerns Gaussian functions in high dimensions. Intuitively, Gaussian functions are used for their local properties: most of the integral of the function is contained in a limited volume around its center. It is well known that 90% of the samples of a normalized scalar Gaussian distribution are statistically distributed in the interval [-1.65, 1.65]. What is less obvious is that this percentage rapidly decreases to 0 with respect to the dimension of the space. In Fig. 5.11 the percentage of samples of a Gaussian distribution falling in the sphere of radius 1.65, versus the dimension of the space is shown. In dimension 10 this percentage is below 1%. In other words, when the dimension increases, most of the volume of a Gaussian function is contained in the tails instead of being located near the center, in total disagreement with the commonly accepted view of locality.

More than geometrical properties, these four examples show that data even uni-

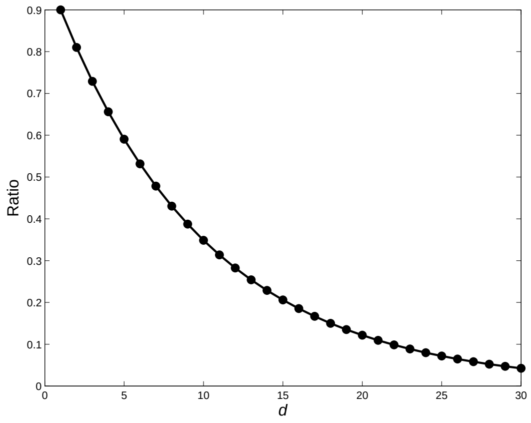

Figure 5.10: Ratio between the volume of a sphere with radius 0.9 and the volume of a sphere with radius 1, versus the dimension of the space.

formly distributed concentrate in unexpected parts of high-dimensional spaces and that function considered as local are not local anymore in high dimensions.

5.5.2 Number of learning points in high-dimensional settings

The discussion about the problems related to high dimensions is also intended to point out the necessity for a growing number of samples when the dimension increases. It is difficult to tackle the problem of finding the number of samples required to reach a predefined level of precision in approximation tasks. The reason is that results tackling level of approximation tasks are usually rather theoretical ones, derived from asymptotic developments where the number of samples tends to infinity.

Intuitively, the number of samples should increase exponentially with the dimension of the space. If 100 learning points are necessary to obtain one defined precision in a two-dimensional scalar function approximation problem, the same level of precision would require 10^d learning points in a d-dimensional space. As an example, Silverman (Silverman, 1986) addressed the problem of finding the required number of samples for a specific problem in the context of the approximation of a Gaussian distribution with fixed Gaussian kernels. Silverman's results are summarized in Fig. 5.12. Silverman's results can be approximated by (Comon et al., 1994)

$$\log_{10} N(d) \approx 0.6(d - 0.25). \tag{5.34}$$

At this point the reader might think that one never has enough samples in high dimensions. Consider indeed a problem where 10 samples would be required in dimension 1 and 100 in dimension 2 (imagine a segment and a square approximated by 10 and

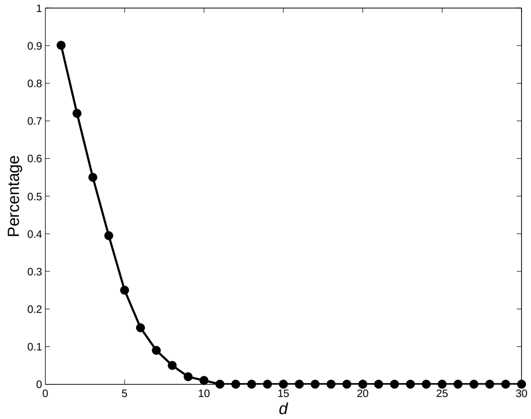

Figure 5.11: Percentage of samples from a Gaussian distribution falling in the sphere of radius 1.65, versus the dimension of the space.

100 samples, respectively). This would mean that 10^{20} would be required in dimension 20. This is obviously impossible in real-world problems while dimensions much greater than 20 are common.

Fortunately, it appears that real-world problems do not suffer so severely from the "curse of dimensionality" problem, as in most situations data are located near a manifold of dimension lower than d. This fact is also mentioned at the beginning of chapter 6.

5.5.3 Concentration of measure phenomenon

While this is not directly related to the two questions introduced at the beginning of this section, the *concentration of measure phenomenon* is another surprising result in high-dimensional spaces which is more related to data analysis. This phenomenon in the case of Gaussian distributions (with standard deviation equal to 1) is illustrated in Fig. 5.13. For several dimensions of the space (1, 2, 3, 5, 10 and 20), the figure shows the probability density function (pdf) of finding a point drawn according to a Gaussian distribution, at distance r from the center of that distribution. In dimension 1, this pdf is a monotonically decreasing function. In dimension 2, it has a bell shape with a peak around 1, that illustrates the fact that there are more points at distance 1 from the center than at distance 0.2 or 2. When the dimension increases, the bell shape remains, but is shifted to the right. In dimension 20 the percentage of data lying at a distance of less than 2 from the center is so low that it cannot be seen at the scale of the figure, despite the fact that the standard deviation of the Gaussian distribution is 1. This means that the distances between all points and the center of the distribution are concentrated in a small interval. Relative differences between these distances vanish. Therefore, these distances become less and less discriminative when the dimension increases and this is

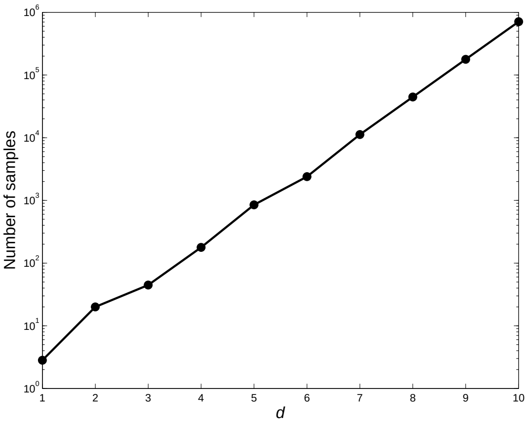

Figure 5.12: Number of samples required to approximate a Gaussian distribution with fixed Gaussian kernels, with an approximation error of about 10% (according to (Comon et al., 1994)) versus the dimension of the space.

relevant for data analysis.

The above discussion issues that, under soft conditions on the distribution of samples (uniform distribution is thus not required), the variance of any measure -distance, norm- remains fixed while its average increases with the dimension of the space.

The concentration of norm phenomenon is more precisely detailed in several mathematical works. In (Demartines, 1994) it was shown that for random vectors with independent and identically distributed components, the mean of their Euclidean norm increases as the square root of the dimension of the space, while the variance of their norm does not increase. He concludes that if the dimension is high, all vectors are normalized, as the error resulting from taking the mean of their norm instead of their actual norm becomes negligible.

Independently of Demartines' results, in (Beyer et al., 1999) it was proved that when the dimension increases, the relative difference between the largest and smallest norm in a dataset converges to zero in probability. The relative difference is the difference between the largest and smallest norms, divided by the smallest one. The result is valid for arbitrary distance measures, under mild conditions on the distribution. Beyer concludes that, in a nearest neighbor search context, all the points converge to approximately the same distance from the query point. Thus, the notion of nearest neighbor becomes less intuitive in high-dimensional spaces.

Other metrics may be used in high-dimensional data analysis in place of the Euclidean distance. One of these metrics is based on the Minkowski norm. Given two D-dimensional vectors $\mathbf{x}, \mathbf{q}$, their distance according to the Minkowski metric is de-

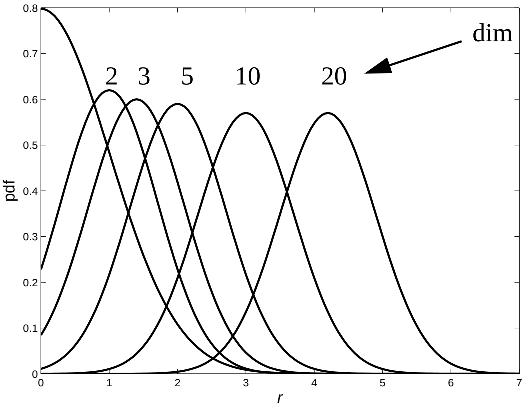

Figure 5.13: Probability of a point from a Normal distribution to be at distance r from the center for several space dimensions.

fined as

$$||(\mathbf{x},\mathbf{q})||_p = \left(\sum_{i=1}^{D} |x_i - q_i|^p \right)^{\frac{1}{p}}, \tag{5.35}$$

where p is the order of the metric. When $p = 2$ this distance equals the Euclidean metric. In (Aggarwal et al., 2001) it is suggests to use in general $p < 2$. On the other hand, in (Francois et al., 2005) it was shown the optimal distance to be used for neighbor search also depends on the type of noise on the data. For example, fractional norms are preferable in the case of colored noise, but if Gaussian noise is assumed, then the Euclidean metric is more robust to than the fractional one.

In this work the Minkowski metric is utilized in section 8.4.

6 DCE-MRI Data Analysis

In this section the experimental results concerning the analysis of DCE-MRI data using techniques of unsupervised dimensional reduction are detailed. The data set analyzed in this chapter is the 12-cases data set (see table 2.4) with the exception of section 6.4, where the 14-cases data set is analyzed.

These data sets contain time-series belonging to benign and malignant lesions. From the medical point of view, a vital question is how the benign and malignant signals differ. The answer to this question involves the comparison of tumors from different patients. This is however a demanding task for the human observer because of the multi-dimensional nature of the data. Computer-assisted techniques supporting the human observer in the exploration of DCE-MRI data are therefore desirable.

Several computed-based approaches to the analysis of DCE-MRI data have been proposed in recent years. They regard tumor classification (Gilhuijs et al., 1998; Lucht et al., 2001b; Abdolmaleki et al., 2001; Knowles et al., 2000; Degenhard et al., 2002; Arbach et al., 2004), tumor detection (Subramanian et al., 2004), tissue characterization (Twellmann et al., 2005) and tissue segmentation (Lucht et al., 2001a). However, all these approaches allow the human observer to analyze one sole tumor lesion at a time, while radiologists, in order to investigate the structural differences between benign and malignant tumors, also need techniques allowing for the exploration of DCE-MRI data from a set of different patients simultaneously. But a simultaneous exploration of several tumor lesions in the anatomical space is infeasible as subject to inter- and intra-observer variability (Orel and Schnall, 2001), and the visualization of the entire signal space with conventional techniques is not possible since the signal dimension exceeds three.

In this work, to explore the DCE-MRI data, the set of the time-series is projected into a lower-dimensional space, whose visualization may allow the human observer to discern patterns and regularities in the multi-dimensional data. In particular the signal distributions of different tumor types can be visualized, thereby allowing for a multi-case analysis.

This chapter is divided into four parts. The first part describes the application of LLE on the DCE-MRI data set for visual data exploration. The data is reduced down to two dimension by LLE with different values of its input parameters (n and Δ), whose effects on the output projection are analyzed. The two-dimensional projected spaces obtained by LLE are visualized by a scatter plot with customized colors for the exploration of the degree of similarity between benign and malignant tumors.

The performances of LLE are compared with those of PCA and SOM in the second part of the chapter. Here the structural properties of the algorithms are mentioned, as

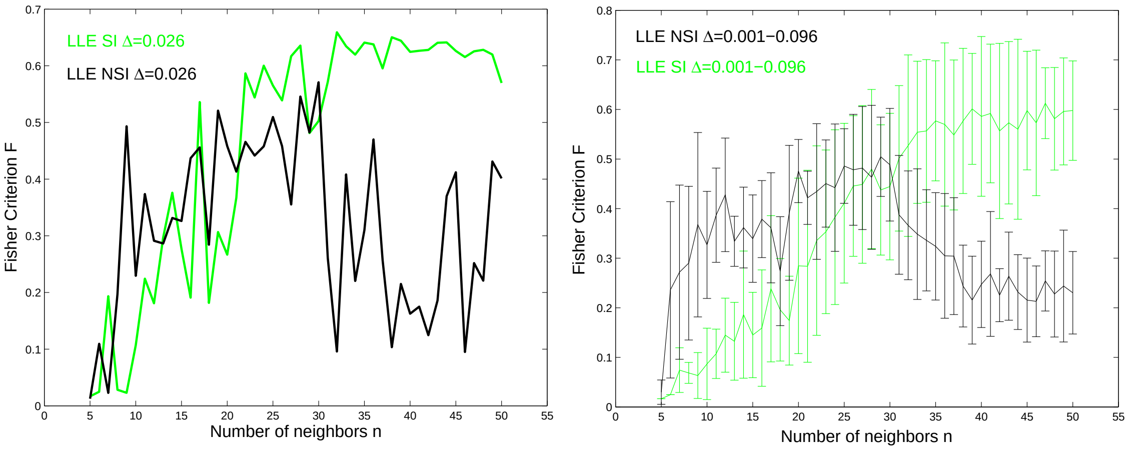

Figure 6.1: (Left) Plot of F of the two-dimensional projections of the DCE-MRI dataset with (NSI) and without normalized time-series (SI) computed by LLE with n between 5 and 50 and $\Delta = 0.026$. (Right) Plots with error bars that represent the variance induced on F by Δ varying.

well as their advantages and shortcomings.

The third part of the chapter illustrates the customized color visualization of tumor lesions by LLE. This is done by projecting the DCE-MRI data set into a 3D space by LLE and encoding the new 3D coordinates within the RGB color space. As a result, a tumor lesion can be visualized by customized colors that describe the dynamic of the contrast agent. The images obtained by LLE are compared with those obtained by the Three-Time Point method (3TP), which is the state-of-the-art technique for the customized visualization of breast tumor lesions in DCE-MRI (see section 4.3.2).

The fourth and last part of the chapter describes the usage of the 2D projection obtained by LLE for the histological characterization of the breast lesions. Specifically, the scatter plot is visualized with colors encoding the histologic families of the tumors. This can be of use for investigating the degree of similarity of different families and for characterizing histologically new tumor cases.

6.1 Visual exploration of DCE-MRI data by LLE

In this section the DCE-MRI dataset is explored by projecting it onto a two-dimensional space by LLE and by visualizing the new coordinates in a scatter plot.

This section is organized into three parts. In the first part the LLE algorithm is applied to both the original and normalized DCE-MRI data set with different values of its input parameters n and Δ. The obtained projections are evaluated in terms of separation between the benign and malignant cluster by the Fisher criterion F (section 5.4.1). Four projections of the original DCE-MRI data and four projections of the normalized data are also visualized in a scatter plot, in which mapped time-series of benign (malignant) tumors are visualized in green (red).

In the second part the projections obtained by LLE applied to the original DCE-MRI data are evaluated using three further quantities, namely stress (ST) (section 5.4.2), neighborhood preservation (NP) (section 5.4.3) and trustworthiness (section 5.4.4).

In the third and last part of this section the visualization of the LLE embedding is expanded to include information regarding the wash-in phase of the time-series.

6.1.1 Effects on the Projections of Varying the LLE Inputs

In this section, the effects on the LLE projections of varying both input parameters of LLE, the number of neighbors n and the correction term Δ, are investigated. The dimension of the input space equals 6 and therefore Δ is required when $n > 6$ (see eq. (5.17) in section 5.2.2). The different projections are evaluated with the Fisher criterion F. The higher the value of F, the less the clusters overlap.

This analysis is conducted on two different data, the 12-cases DCE-MRI dataset and the same dataset with time-series scaled to [0,1], i. e. each time-series is scaled to the range [0,1]. Thereafter they will also be called as original and normalized DCE-MRI data set, respectively.

At first, only the variations of one parameter are taken into account. Specifically, n is varied between 5 and 50 and Δ is fixed to 0.026, a value set empirically. The plot of the respective Fisher values are visualized in Fig. 6.1 left. The SI (signal intensity) and NSI (normalized signal intensity) curves do not differ significantly for $n \leq 30$, while for $n > 30$ the SI curve is the predominant one with quite stable values of F, while the NSI curve exhibits lower and quite changeable values of F.

Now the variations of both input parameters are taken into account. Specifically, for each n, the value of Δ is varied from 0.001 to 0.096 (21 values with interval 0.005) and the standard deviation induced to F is visualized as an error bar. The curves obtained for the DCE-MRI data with and without time-series normalization are shown in Fig. 6.1 right. Also in this case the LLE projections of the original DCE-MRI data achieve the highest values of F for $n > 30$. It follows that the normalization of the time-series leads to a loss of information that, in turn, causes the benign and malignant clusters to be less separated in the projections.

The values of standard deviation due to Δ changing are overall comparable in both curves and are quite stable with respect to n.

Four two-dimensional projections related to the graphs of Fig 6.1 left are visualized by scatter plots in Fig. 6.2 (original DCE-MRI data) and Fig. 6.3 (DCE-MRI with normalized time-series). Points that are the projections of time-series belonging to benign (malignant) tumors are visualized in green (red) and they will thereafter be named benign (malignant) points. In Fig. 6.2 the best separation between benign and malignant points can be observed for $n \geq 30$, in agreement with the SI curve in Fig. 6.1 left, where

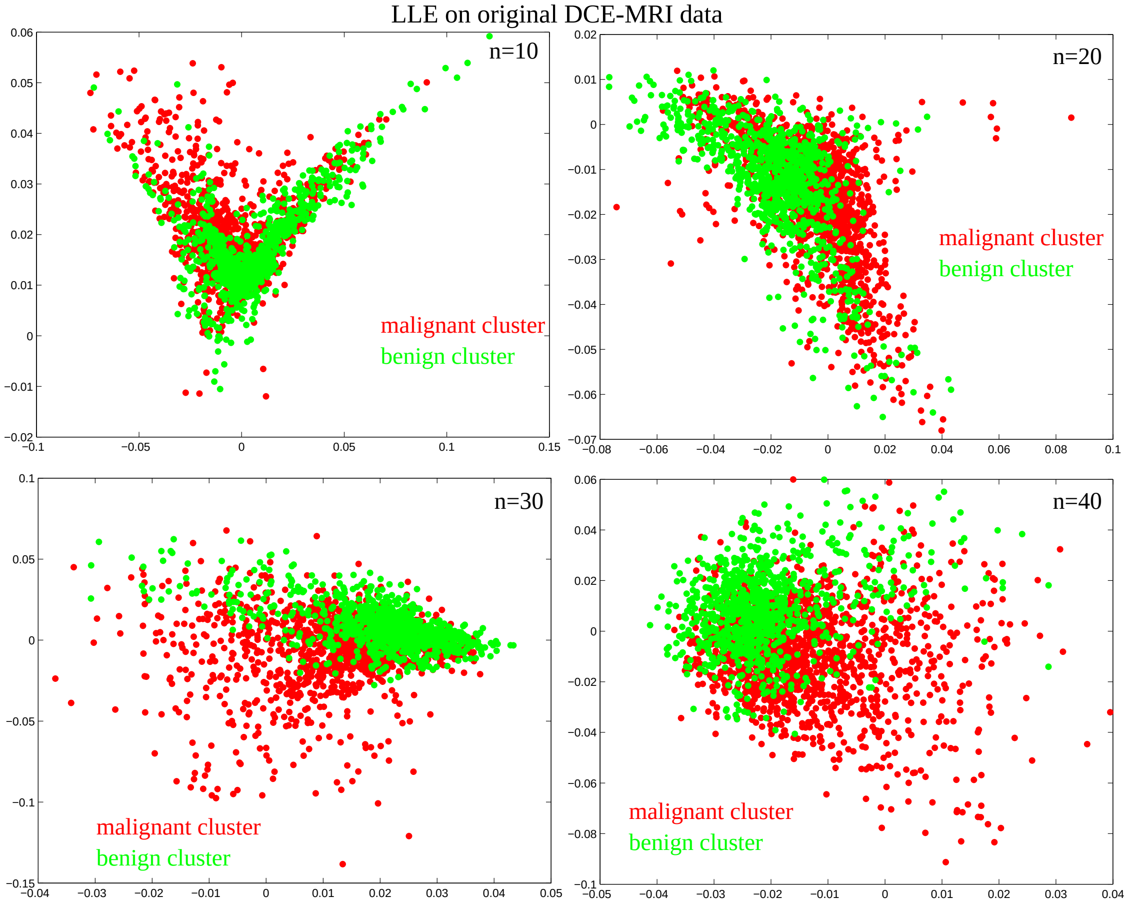

Figure 6.2: Four LLE mappings of the original DCE-MRI data. The projection obtained with $n = 10$ exhibits a poor cluster separation. The cluster separation is slightly more observable with $n = 20$. The mappings obtained with $n = 30$ and $n = 40$ exhibit the best cluster separation. In particular, in these cases the benign points are localized in the upper part of the projections and the lower part contains only malignant points.

the largest values of F are achieved for $n > 20$. In particular in the projections obtained with $n = 30$ and $n = 40$, two distinct regions are observable. In one of these only malignant points are present, in the other there are both benign and malignant points. Overall, the benign and malignant clusters are far less separated in the projections of the normalized DCE-MRI data that are visualized in Fig. 6.3. This is also in agreement with the facts that the respective values of F are lower than the ones computed for the original DCE-MRI dataset (Fig. 6.1 left).

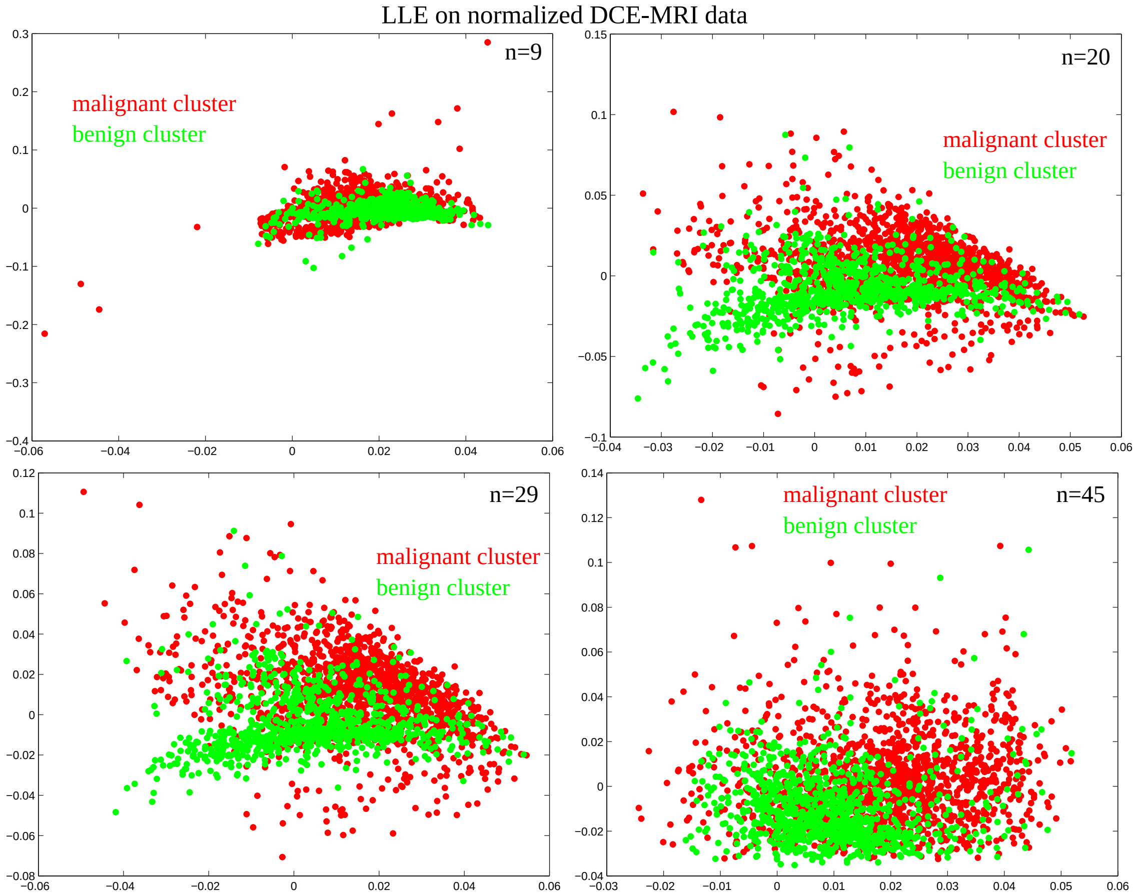

Figure 6.3: Four LLE mappings of the DCE-MRI data with each time-series scaled to [0,1]. In the projection obtained with $n = 9$, the cluster separation is very poor, while it is more observable in the remaining three projections. However, in all cases the cluster separations are worser than those in Fig. 6.2, in agreement with the curves in Fig. 6.1 left, i. e. lower values of F correspond to lower separation between the clusters.

6.1.2 Further Evaluation of the projections of the original DCE-MRI data

In this section, the projections providing overall the largest values of F, i. e. the projections obtained by LLE applied to the original DCE-MRI data, are further analyzed. In particular the respective LLE embeddings are further evaluated in terms of stress (section 5.4.2), neighborhood preservation (section 5.4.3) (NP) and trustworthiness (section 5.4.4). In particular, each value of trustworthiness and NP is the average of 16 values computed for t between 5 and 20. In this way the preservation of both the local (neighborhood preservation and trustworthiness) and global (stress) data structure of

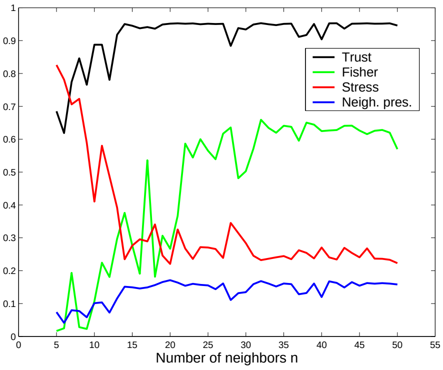

Figure 6.4: Curves of the trustworthiness, Fisher criterion, stress and neighborhood preservation related to the LLE embeddings of the DCE-MRI data.

the original data into the low-dimensional projection is quantified.

The plots of the quantities computed for n between 5 and 50 are visualized in Fig. 6.4. All the curves exhibit two different behaviors, i. e. the values are quite changeable with $n < 20$, and are far more stable with $n > 20$. In addition, with $n > 20$ the curves of the trustworthiness, Fisher criterion and neighborhood preservation achieve their maximal values (i. e. the best performances), while the stress curve exhibits its lowest values (i. e. the best performances). It is evident that the best embeddings, i. e. the embeddings in which structural properties of the original DCE-MRI data are best preserved, are those obtained with $n > 20$. According to the graphs, these embeddings (with $n > 20$) exhibit comparable characteristics, neglecting the small fluctuations of the curves. Two examples of these embeddings (computed with $n = 30$ and $n = 40$) are visualized in Fig. 6.2.

6.1.3 Visualization of the LLE Projection with expanded Information

In this section the goal is to investigate how correlated the low-dimensional coordinates and the dynamic characteristics of the time-series are. For this purpose the LLE embedding is visualized including further information, namely the wash-in characteristic of the time-series (see section 2.4). Specifically, the size of each point is proportional to the difference between the signal intensity values at the first post-contrast and pre-contrast of the respective time-series. It was found empirically that a particularly evident correlation between them is observable in the 2D projection obtained with $n = 25$. This projection is shown in Fig. 6.5. One can observe that the dimension of the points grows steadily along the direction of the blue arrow. At the same time, the biggest points are malignant ones, while the smallest ones are benign, in accordance with the model-based characteristics of benign and malignant tumors as described in Fig. 2.7. It appears thus evident a correlation between the wash-in characteristic and the degree of

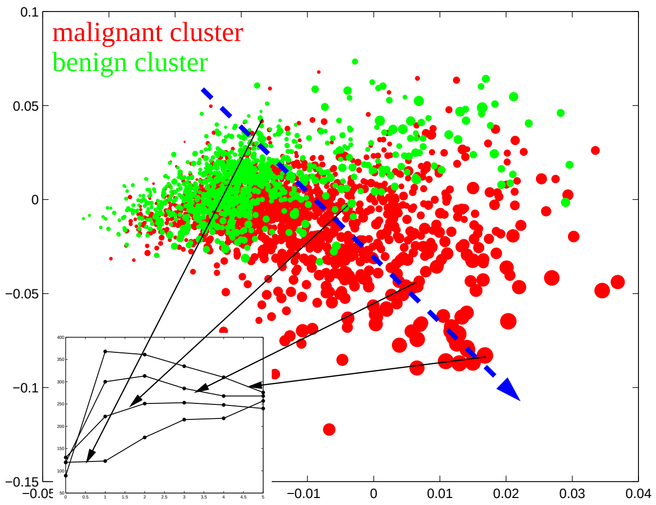

Figure 6.5: Two-dimensional reduction of the twelve-tumors data set obtained by LLE with $n = 25$ and $\Delta = 0.026$. The point size is proportional to the pre-contrast intensity of the mapped time-series. The blue arrow gives the gradient of the point size. Four points chosen along this direction are correlated with their respective time-series. These time-series are characterized by progressively larger intensities of wash-in.

malignancy of the tissue. The blue arrow seems to determine a direction for a gradient of malignancy, i .e. the time-series of the points mapped along that direction exhibit progressively stronger malignant behaviors (growing wash-in). This fact is evidenced by the visualization of the time-series of four representative points in the scatter plot (Fig. 6.5).

By virtue of the evident correlation between the coordinates in this LLE embedding and the wash-in, the value $n = 25$ is adopted as an optimal one for later experiments in this chapter.

6.2 Visual exploration of the DCE-MRI Data by SOM, PCA and LLE

So far the LLE algorithm has been employed to explore the signal dynamic of the DCE-MRI dataset. In this section the PCA and SOM algorithms are also used for the same purpose and their performances are compared with those of LLE. PCA and SOM are two of the most frequently used algorithms for the analysis of high-dimensional data and have been resumed in section 5.2.1 and section 5.3.1, respectively.

This section is organized into two parts. The first part regards the optimization of dimension of the 2D-grid of SOM. In the second part the performances of SOM, PCA and LLE applied to the DCE-MRI dataset are compared, while discussing the advantages and shortcomings of each algorithm. The second part also includes four examples of mappings of single tumor lesions.

6.2.1 Selection of the optimal dimension of the SOM grid

The SOM requires the setting of several parameters. Most of the values adopted in this work are those employed in (Nattkemper and A., 2005), where the same DCE-MRI dataset is processed by SOM in order to produce a customized color visualization of the lesions. Specifically, the learning rate factor $\alpha(t)$ is decreased linearly from an initial value equal to N to a final value equal to $0.2 \cdot N$; the number of iterations of the SOM algorithm is set equal to $10^3 \cdot N^2$.

In this work only the dimension of the two-dimensional grid (that equals the number of reference vectors $\mathbf{u}$) is tuned. In particular, as a squared grid of dimension $N \times N$ is utilized, the parameter to optimize is N. For this purpose the SOM algorithm is executed with different values of N ($5 \leq N \leq 25$) and the corresponding SOM projections are evaluated in terms of *average quantization error* (AQE) and *topology preservation* (TP).

AQE quantifies how accurately the trained reference vectors approximate the data set. This measure is defined as the sum of the distances between each data point $\mathbf{x}_i$ and its nearest reference vector $\mathbf{u}_i$, normalized by the total number of data points:

$$\text{AQE} = \frac{1}{v} \sum_{i=1}^{v} ||\mathbf{x}_i - \mathbf{u}_i||. \tag{6.1}$$

The less AQE, the more accurately the distribution of the reference vectors approximates the distribution of the DCE-MRI data set. Note that the symbolism used for eq. (6.1) may be misleading, as the reader may think the number of reference vector is v because the index i is also used for the reference vector $\mathbf{u}_i$. In reality, their number is $N \times N$ and it can happen that more than one point have the same nearest reference vector.

The topology preservation measure describes how well SOM describes the topology of the data set, i. e. the relations between neighboring data points. TP can be calculated by (Kiviluoto, 1996):

$$\text{TP} = \frac{1}{v} \sum_{i=1}^{v} p(\mathbf{x}_i), \tag{6.2}$$

$$\text{where } p(\mathbf{x}_i) = \begin{cases} 1 & \text{if the first and second nearest reference} \\ & \text{vectors of } \mathbf{x}_i \text{ are not next to each other} \\ 0 & \text{otherwise.} \end{cases}$$

TP is also an error function and the lower its value, the better the topology of the data is preserved in the SOM grid. In conclusion, the dimension N of the grid should be chosen in such a way to minimize the values of AVQ and TP.

6.2.2 Comparison of the performances of PCA, SOM and LLE

The comparison of the performances of SOM, PCA and LLE is performed by quantifying to which extend benign and malignant data are separated when projected in the respective projected spaces. Of course, the best result arises when the benign and malignant clusters are best separated, which would mean the particular algorithm is best capable of distinguishing between benign and malignant time-series.

Two different approaches are used to carry on the comparison. The approaches share the idea that each data point $\mathbf{y}_i$ in the projected space is given a label ρ_i that depends on its K nearest neighbors. Each label provides information regarding the overlap between benign and malignant cluster in the neighborhood given by the K nearest neighbors. The analysis of all the labels quantifies the global overlap between the benign and malignant cluster in the projected space.

Note that the approaches differ when applied to the PCA and LLE on the one hand, and to SOM on the other hand. This is due to the fact that the projections obtained by PCA and LLE are intrinsically different from the projections obtained by SOM. Specifically, in the PCA and LLE projections each input point $\mathbf{x}_i$ is mapped to an output point $\mathbf{y}_i$ and each of them is visualized in a scatter-plot. By contrast, in the SOM projection each $\mathbf{x}_i$ is mapped to a node r_i on a two-dimensional grid and the number of nodes ($N \times N$) is lower than the number of input vectors v, which means that more than one point are mapped to the same node. As a result, the SOM projection depicts the data distribution by a schematic grid in which the single points $\{\mathbf{y}\}_v$ are not visible.

6.2.2.1 Approach 1 based on data density estimation

With regard to the LLE and PCA projections, each point $\mathbf{y}_i$ is assigned a label $\rho_i(K)$, a real value that is computed as follows: let $\mathbf{y}_i$ be the mapping of a point $\mathbf{x}_i$ belonging to a malignant (benign) lesion. This means $\mathbf{y}_i$ is regarded as malignant (benign) point and belongs to the malignant cluster. Consider now the K nearest neighbors of $\mathbf{y}_i$ (K=7 for instance) and how many are malignant (benign) (suppose there are four malignant (benign) neighbors). The label $\rho_i(K)$ is given by the percentage of malignant (benign) points among the K-nearest neighbors (in this case it is $\rho_i(7) = 4/7 = 0.57$, i. e. 57$\%$ of the 7-nearest neighbors of $\mathbf{y}_i$ belong to its same class). This approach is also illustrated in Fig. 6.6 left.

The label $\rho_i(K)$ is computed in a different way in the case of the SOM mapping. In this case, each point $\mathbf{y}_i$ is given by the coordinates of the node r_i in the grid. This node is linked to the reference vector $\mathbf{u}_i$, that is the closest reference vector to $\mathbf{x}_i$ (see section 5.3.1). The methodology for computing $\rho_i(K)$ is the following: at first the reference vector $\mathbf{u}_i$ is given a label that equals the percentage of the predominant class among its K-nearest neighbors. For instance, let K be seven as in Fig. 6.6 right, and suppose four neighbors belong to the malignant cluster and three to the benign one. This means the points of the malignant cluster form the majority of the seven nearest neighbors of $\mathbf{u}_i$

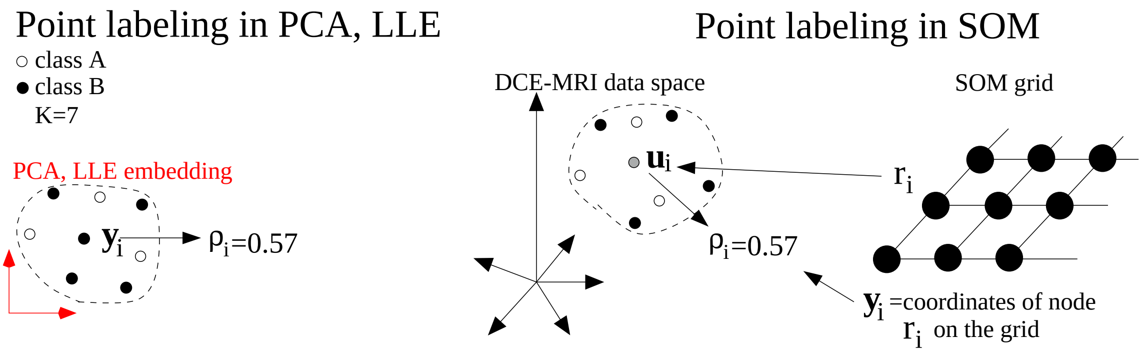

Figure 6.6: This figure illustrates how the label $\rho_i(K)$ is computed in the PCA, LLE (left) and SOM (right) projections according to the first approach. Let $K = 7$. On the left it is shown a portion of LLE (or PCA) embedding. Point $\mathbf{y}_i$ (the projection of $\mathbf{x}_i$) of class B has four neighbors belonging to class B and three to class A. This means that 57% of the seven-nearest neighbors of $\mathbf{y}_i$ belong to its same class. It follows that $\rho_i(7) = 0.57$. The labeling of the points in SOM is shown on the right. Point $\mathbf{y}_i$ has $\mathbf{u}_i$ as respective reference vector. Of the seven nearest neighbors of $\mathbf{u}_i$, three belong (43%) to class A and four (57%) to class B. The latter percentage is the largest one, thus the label of $\mathbf{u}_i$ equals 0.57. Point $\mathbf{y}_i$ is given the same label of $\mathbf{u}_i$, that is $\rho_i(7)$=0.57.

(57% malignant and 43%). The reference vector $\mathbf{u}_i$ is thus assigned a label that equals 0.57 and, in turn, the point $\mathbf{y}_i$ is then assigned the same label of $\mathbf{u}_i$, i. e. $\rho_i(K) = 0.57$. To summarize, $\mathbf{y}_i$ is assigned the label of the respective reference vector $\mathbf{u}_i$ that is given by the percentage of points the points belonging to the predominant class, as computed among the K-nearest neighbors of $\mathbf{u}_i$. This methodology is illustrated in Fig. 6.6 right.

Finally, for all three algorithms, the *clusters separation* $\rho(K)$ is given by

$$\rho(K) = \frac{1}{v}\sum_{i=1}^{v} \rho_i(K) \tag{6.3}$$

where $0 \leq \rho(K) \leq 1$. The closer $\rho(K)$ is to one, the more the benign and malignant clusters are separated from each other in the projected space. Conversely, values of $\rho(K)$ close to zero signify that the benign and malignant clusters largely overlap.

6.2.2.2 Approach 2 based on an ROC curve

The other approach adopted for evaluating the cluster overlap is also based on the K-nearest neighbor search and is quite similar to the first approach. However, the second approach differs in the fact that each point $\mathbf{y}_i$ is labeled as either benign or malignant.

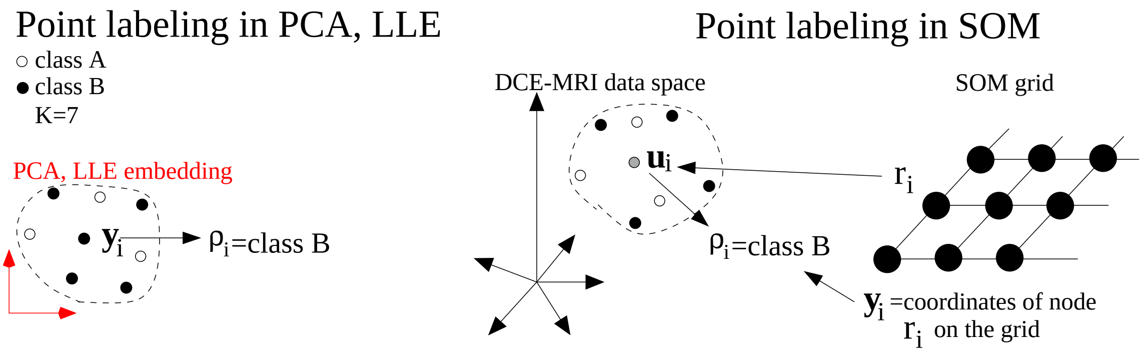

Figure 6.7: This figure illustrates how the label $\rho_i(K)$ is computed in the PCA, LLE (left) and SOM (right) projections in the second approach. Let K be equal to 7. On the left it is shown a portion of LLE (or PCA) embedding. Point $\mathbf{y}_i$ (the projection of $\mathbf{x}_i$) of class B has four neighbors belonging to class B and three to class A. This means that the majority of its seven-nearest neighbors belong to class B. It follows that $\rho_i(7)$ =class B. The labeling of the points in SOM is shown on the right. Point $\mathbf{y}_i$ has $\mathbf{u}_i$ as respective reference vector. The majority of the seven nearest neighbors of $\mathbf{u}_i$ belong to class B and, in turn, it is labeled as belonging to class B. As a result $\mathbf{y}_i$ is labeled in the same way, i. e. $\rho_i(7)$=class B.

In other words, ρ_i is a real value in the first approach and a boolean value in the second approach.

Also in the second approach, the labeling process is different in the PCA, LLE and SOM projections.

In the LLE and PCA projections, if the majority of the K nearest neighbors of $\mathbf{y}_i$ are malignant (benign), then $\mathbf{y}_i$ will be labeled as malignant (benign) (see Fig. 6.7 left).

In the SOM projection each point $\mathbf{y}_i$ is given the same label of the correlated reference vector $\mathbf{u}_i$. The reference vector $\mathbf{u}_i$ is labeled as malignant (benign) if the majority of its K nearest neighbors are malignant (benign) (see Fig. 6.7 right).

After labeling all the data points $\{\mathbf{y}\}_v$ it is possible to compute the percentage of malignant points correctly classified (true positive fraction or TPF) and the percentage of benign points wrongly classified (false positive fraction or FPF) (see section 5.4.5). If this calculation is performed with respect to different values of K, one obtains a set of (TPF,FPF) values, one for each value of K, that can be plotted in an ROC curve. (section 5.4.5). The values of TPF and FPF equal the sensitivity (SE) and $1-$ specificity (SP), respectively.

6.2.3 Results

In this section the obtained results of the comparison are shown and evaluated. At first, the optimization of the dimension of the SOM grid is discussed. Next, the two-dimensional projections obtained by PCA, LLE and SOM are visualized and the performances of the algorithms are evaluated and compared in terms of separation between benign and malignant clusters in the projected spaces. Finally, for a single case study, the points of single lesions are highlighted in the low-dimensional mappings, whereby each lesion can be compared with all the remaining ones simultaneously.

6.2.3.1 Dimension of the SOM grid

Concerning the dimension of the SOM grid, the values of N between 3 and 25 are investigated. The respective values of AQE and TP are plotted in Fig. 6.8. Here one can observe that both quantities progressively decrease as N increases, but, at the same time, their values, in first approximation, can be regarded as constant for $N > 15$. Thus, any value of N within the range comprises between 15 and 25 appears a reasonable choice with regard to the values of AQE and TP. On the other hand, a larger value of N leads to a higher number of empty nodes, i. e. nodes whose reference vectors are without mapped data points. In general, their presence is not desirable, as it is symptom of overfitting of the data by SOM. Finally, the value $N = 19$ is chosen as it provides a good trade-off between low values of AQE and TP and low number of empty nodes.

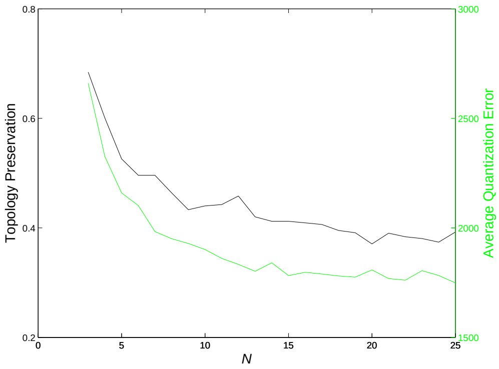

Figure 6.8: Topology preservation (TP) and average quantization error (AQE) calculated for different dimensions of the grid. The grid has dimension $N \times N$. Note that with $N > 15$ values of both TP and AQE are minimal and undergo negligible variations.

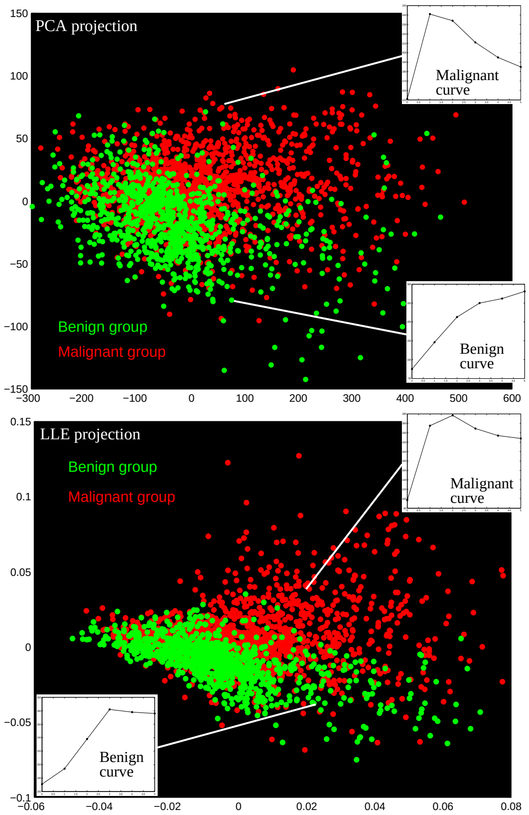

Figure 6.9: Visualization of the PCA and LLE projections of the DCE-MRI dataset. Two examples of benign and malignant time-series along with their respective projections are also shown. Two distinct regions are clearly observable. The upper region contains only malignant points. The lower region contains both benign and malignant points. The latter are scarcely visible because the clusters overlap.

6.2.3.2 Visualization of the projections obtained by PCA, SOM and LLE

The two-dimensional projections of the DCE-MRI data set obtained by PCA and LLE are visualized in Fig. 6.9. The color scheme encodes the labels (benign/malignant) of the tumors to which the respective time-series belong. Both mappings show qualitative agreement, in particular concerning the benign cluster, which in both cases is localized in the lower part of the scatter plot. In addition, the benign cluster overlaps with many malignant data points, suggesting that many of them exhibit benign characteristics. This is in agreement with the clinical experience of physicians, i. e. malignant tumor lesions can be highly heterogeneous, in particular can include large benign regions, while benign lesions are typically more homogeneous (Kelcz et al., 2002). In particular, the latter statement is confirmed by the higher compactness of the benign cluster as compared to the malignant one.

The SOM mapping of the DCE-MRI data is shown in Fig. 6.10. A node is visualized in red (green) if the majority of the mapped points to it is malignant (benign). The brighter the color, the larger the majority. One can observe that red nodes are mainly located in the upper right part of the grid, while the majority of the green nodes are located in the bottom left part. This suggests that benign and malignant data points are mainly located in different regions of the DCE-MRI data space, although various

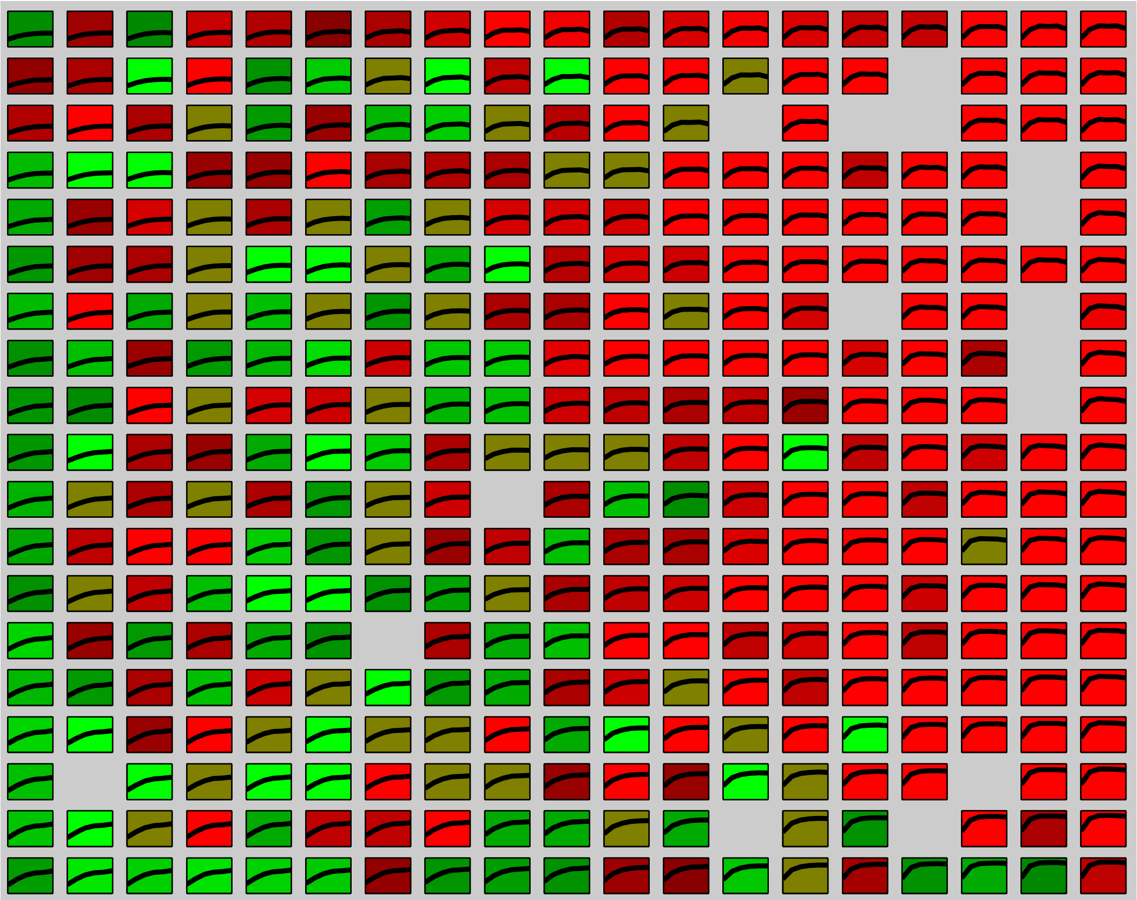

Figure 6.10: Visualization of the DCE-MRI data by a 19×19 SOM grid. Each reference vector is plotted inside the respective node. The color of each node describes the labels of the points mapped into it. Specifically, if the majority of the points are malignant (benign), then the node is visualized in red (green). The brighter the color, the larger is the number of points belonging to this majority.

malignant points are mapped close to benign ones, as evidenced by some red/dark-red nodes located in the bottom left part of the grid, close to green nodes. This analysis is in substantial agreement with the PCA and LLE projections, i. e. two different regions can be seen in all the projections, a region with only malignant points, and a region with both benign and malignant points. Each node r_i in the SOM grid contains the plot of the corresponding reference vector $\mathbf{u}_i$. One can observe that the values of the signal intensity increases from left to right and from up to down. In addition, the curves on the right half of the grid (mostly red nodes) exhibit a wash-out phase (a decrease in the signal intensity that is typical of malignant tissue) that decreases until vanishing from up to down. By contrast, the curves on the green nodes on the left part of the grid

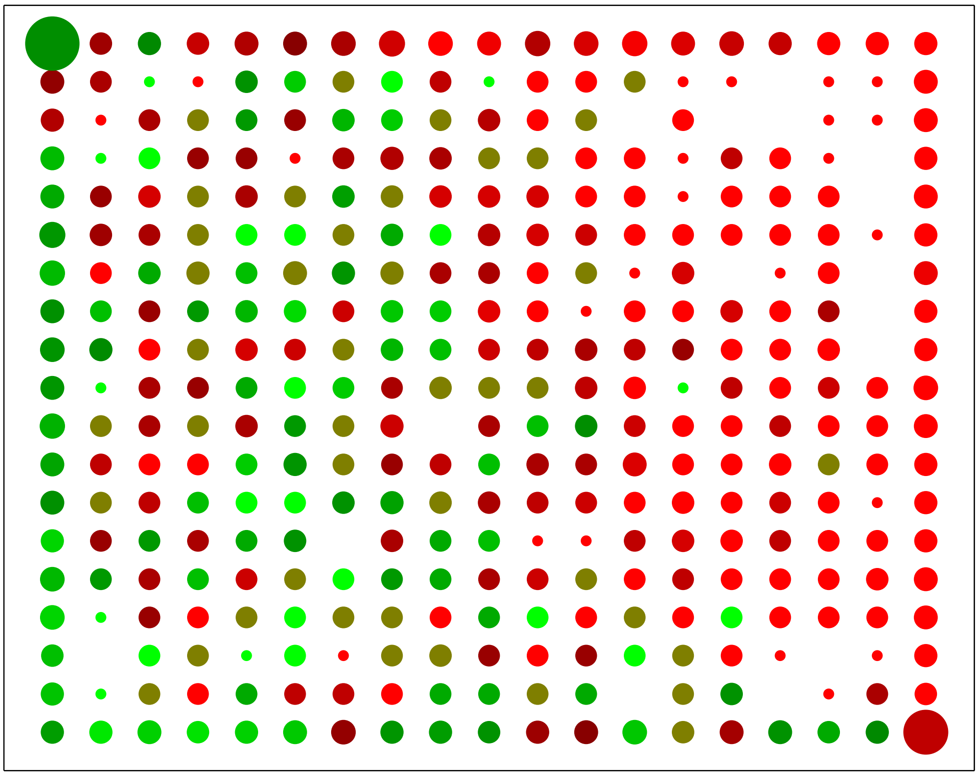

Figure 6.11: Alternative visualization of the DCE-MRI data set by a 19×19 SOM grid. The area of the node is proportional to the number of both benign and malignant points mapped to it, while the colors are the same as those in Fig. 6.10.

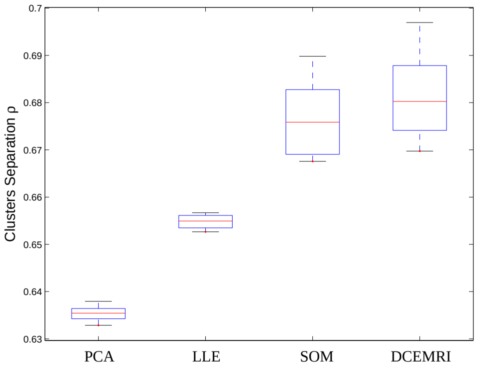

Figure 6.12: Cluster separation ρ computed for different values of K. SOM clearly shows the best performance ($\rho_{\text{SOM}} = 0.676$)) in terms of separation between the benign and malignant clusters. ρ_{SOM} is the closest value to the one computed over the original DCE-MRI data set ($\rho_{\text{DCE-MRI}} = 0.680$)). SOM can also be interpreted as the algorithm which best preserves the salient information contained in the original space, that is the differentiation between benign and malignant data. The other values of average clusters separation are $\rho_{\text{PCA}} = 0.635$ and $\rho_{\text{LLE}} = 0.655$.

typically do not exhibit any wash-out phase, resembling the typical benign curve of Fig. 2.7.

An alternative visualization of the SOM grid is presented in Fig. 6.11. The colors are the same as those in Fig. 6.10. What is different is that each node is visualized as a circle whose area is proportional to the number of points mapped onto the node itself.

6.2.3.3 Comparison of the performance of the algorithms

The next phase of this study consists in comparing the performances of PCA, SOM and LLE in terms of best separation between the benign and malignant clusters in the respective projected spaces. As described in section 6.2.2, two different approaches are adopted for this purpose.

In the first approach, the values of ρ (eq. (6.3)) are computed for values of K between 20 and 39. The results are shown in the box plot in Fig. 6.12. A box has lines at the lower quartile, median, and upper quartile values computed over all the values of K, while the whiskers are lines extending from each end of the box to show the extent of the rest of the values. The SOM mapping provides the best separation between the benign and malignant clusters ($\rho_{\text{SOM}} = 0.676$). At the same time, LLE presents higher average value of ρ than PCA ($\rho_{\text{LLE}} = 0.655$, $\rho_{\text{PCA}} = 0.635$). This is in agreement with the scatter plots in Fig. 6.8, where one can observe that the benign cluster in the LLE projection is more compact and localized as compared to the one in the PCA projection

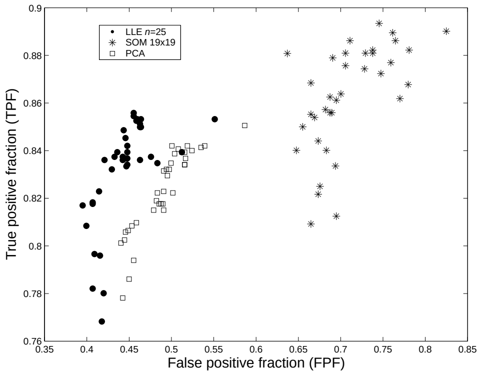

Figure 6.13: Pseudo-ROC curve of the the (TPF, FPF) values obtained for K varying between 3 and 39. LLE shows superior performance as compared to PCA in terms of both TPF and FPF as the LLE values of TPF and FPF are respectively higher and lower than the PCA ones. SOM significantly outperforms LLE with regard to TPF (i. e. the SOM values of TPF are higher). On the other hand, the lowest FPF (or equivalently the highest SP) is obtained by LLE. This is particularly valuable as DCE-MRI can suffer from high values of FPF (or equivalently low values of SP).

and, as a results, the clusters are less separated. In Fig. 6.12 it is also shown the box of ρ calculated in the original DCE-MRI data space ($\rho_{\text{DCE-MRI}} = 0.680$). This value quantifies the separation between benign and malignant data in the original space, that is the salient information one wants to preserve as much as possible in the dimensional reduced space. The closer a value of ρ related to a certain algorithm is to $\rho_{\text{DCE-MRI}}$, the better the particular algorithm can preserve this salient information. In this case SOM is thus the algorithm which best preserves this information.

In the second approach the comparison of the performances of PCA, LLE and SOM is conducted by computing pair of values (TPF, FPF) for values of K between 3 and 39. With respect to the terminology of section 5.4.5, the malignant class is the positive class, while the benign class is the negative class. It follows that TPF is the fraction of malignant points correctly labeled and FPF is the fraction of benign points wrongly labeled according to the K-nearest neighbor rule. The values of TPF and FPF are plotted in a pseudo-ROC curve shown in Fig. 6.13. The LLE algorithm clearly outperforms PCA in terms of both TPF and FPF. Indeed, LLE provides higher values of TPF (i. e. the malignant points $\mathbf{y}_i$ with the majority of their K-nearest neighbors that are malignant are less numerous in the LLE embedding than in the PCA embedding) and lower values of FPF (i. e. the benign points $\mathbf{y}_i$ whose majority of their K-nearest neighbors are malignant are more numerous in the LLE embedding than in the PCA embedding). One can also observe that the largest values of TPF are achieved by SOM. Yet, SOM is also characterized by the largest values of false positive fraction. To summarize, the lowest

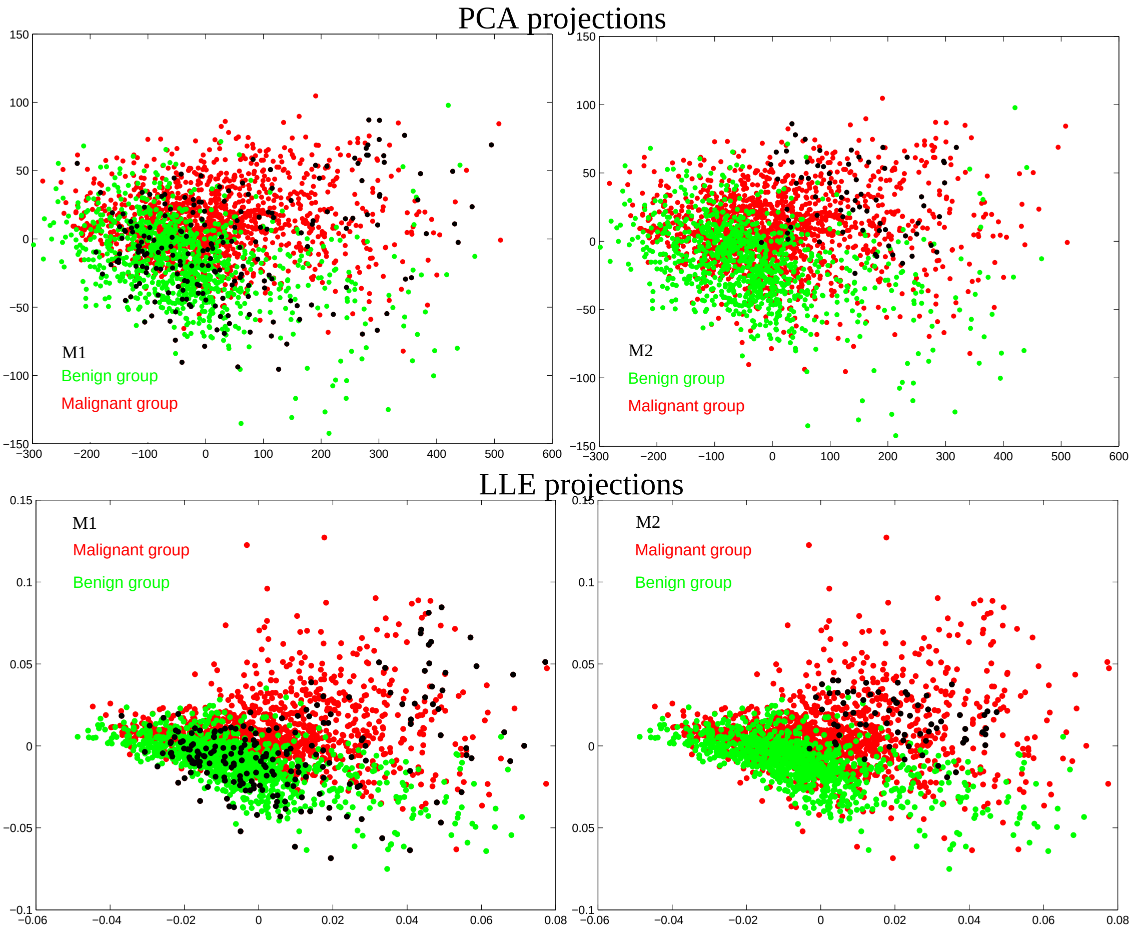

Figure 6.14: Mapping of tumors M1 and M2. M1 appears to be largely heterogeneous, as its data points are scattered over all the embeddings. M2, by contrast, appears more homogeneous, as its data points are relatively clustered and localized near other malignant points.

fraction of benign data points classified as malignant is obtained using LLE, and SOM allows to obtain the highest true positive fraction. DCE-MRI is known to provide high sensitivity (SE=TPF) (Brown et al., 2000) and possibly low specificity (Kelcz et al., 2002) (FPF $=1-$SP) (which is not desirable). Therefore, an algorithm (in this case LLE) which provides higher specificity (or equivalently lower FPF) is particularly appealing. For this reason, from a medical point of view, the performance obtained by LLE is regarded as the best one.

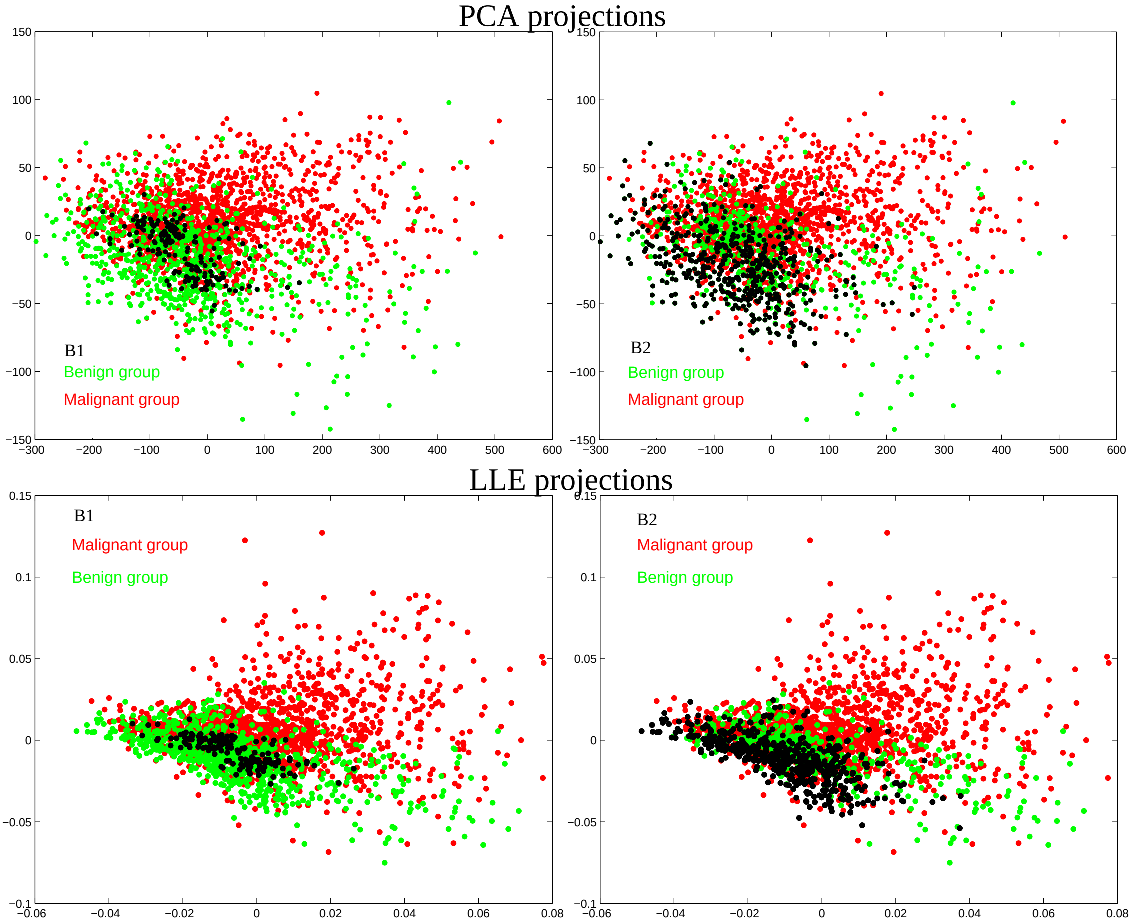

Figure 6.15: Mappings of benign lesion 1 and benign lesion 2 by PCA, SOM and LLE, respectively. In the PCA and LLE projections, the points of both lesions are localized in the lower left part, close to the other benign data points.

6.2.3.4 Analysis of Single Lesions

The low-dimensional mappings can also be used to analyze single tumor lesions. Specifically, the projections of Fig. 6.9 and Fig. 6.11 are visualized highlighting the mapping of the time-series of a single lesion at a time. The visualization of the points of a single tumor lesion allows the user to qualitatively analyze its signal dynamics in relation to all the other lesions contained in the data set simultaneously.

The mappings of tumors M1 and M2 in the PCA and LLE projections are visualized in black in Fig. 6.14. The mappings of each tumor in the PCA and LLE projections are qualitatively similar. The points of M1 are scattered over large parts of the embeddings, thereby revealing a high degree of heterogeneity in the respective DCE-MRI time-series.

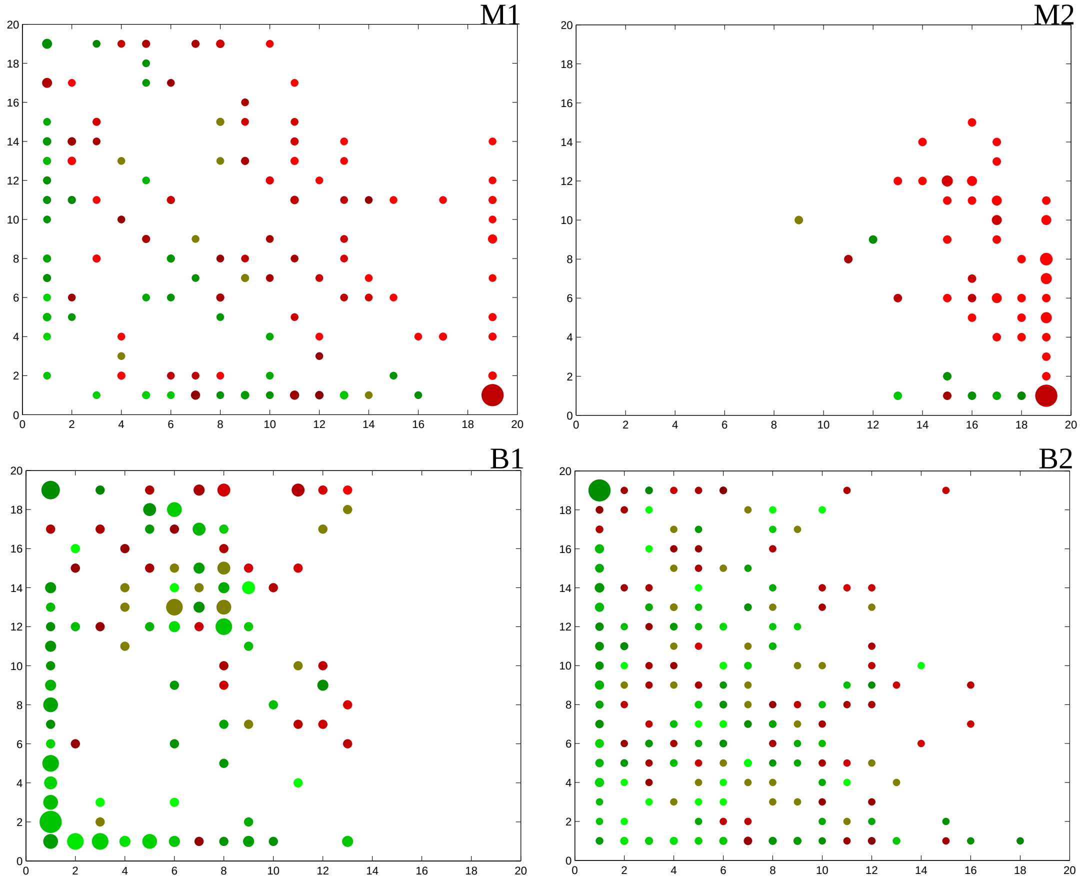

Figure 6.16: Mappings of malignant lesion 1, malignant lesion 2, benign lesion 1 and benign lesion 2 by SOM.

By contrast, the points of M2 are more clustered and almost not overlapping with benign points.

The mappings of two benign tumors, namely B1 and B2, in the PCA and LLE projections are visualized in Fig. 6.15. Also in this case the mapping of a single tumor in the PCA projection is qualitatively similar to the respective mapping in the LLE projection. The mappings of both B1 and B2 are localized in the lower part of the display.

The mappings of tumors M1, M2, B1 and B2 in the SOM grid are visualized in Fig. 6.16. The nodes visualized are only those to which points from a particular lesion are mapped. The area of the node reflects the number of mapped data points restricted to that particular tumor lesion. The colors of the nodes are the same as those in Fig. 6.11, i. e. the colors reflect the SOM training over the entire DCE-MRI data set. In this way it is possible to evaluate the signal dynamic of a single tumor lesion with respect to the signals of the other lesions from the data set.

The data points of M2 are mainly mapped in the right part of the grid (red nodes), while, by contrast, points from M1 are mapped to nodes on both sides of the grid. These mappings are in agreement with those visualized in Fig. 6.14, i. e. tumor M1 appears quite heterogeneous, while tumor M2 is more homogeneous. With regard to the benign lesions B1 and B2, their respective time-series are principally mapped to green nodes localized in the left-bottom part of the grid. This is consistent with the PCA and LLE projections, where the data points of B1 and B2 are mapped close to the data points from the other benign lesions.

6.2.4 Discussion

The LLE algorithm projects the DCE-MRI data onto a two-dimensional space in which the benign and malignant data show a higher degree of separability than in the PCA embedding. On the other hand, SOM quantitatively outperforms LLE in this regard, although the two algorithm yield two structurally different visualization modalities which make a quantitative comparison somewhat questionable. The choice between SOM and LLE depends on the task at hand since the algorithms have different characteristics.

The SOM output is a two-dimensional grid that depicts symbolically the data distribution. This data representation gives an idea about the cluster separation but does not allow for the comparison of single data points. The usage of a grid has the advantage that the node coloring solves the problem of points overlapping (i. e. points in the high-dimensional space may be highly overlapping in the low-dimensional space, thereby hampering the visual analysis), that is common in the LLE projections (Similä, 2005). In addition, the SOM algorithm can also summarize massive data sets. However, it has several input parameters to be set which make its tune phase more difficult. Furthermore, the SOM output is somewhat abstract (the data points are not visualized directly) and its interpretation requires some knowledge on how the SOM algorithm works.

The LLE algorithm produces a more intuitive output in which all the points are visualized, allowing for a comparison between pair of points. The algorithm has one or two input (depending on the data set dimension) parameters. On the other hand it can not be applied to large data sets. Indeed, let v be the number of points in the data set, the LLE algorithm constructs a $v \times v$ matrix, and this can be computationally prohibitive for many systems if v is in the order of several thousands (for instance 5000).

6.3 Color Visualization of Breast Lesions and Comparison with 3TP

In this section the LLE coordinates are used for the visualization of single tumor lesions. Specifically, the DCE-MRI dataset is projected onto three dimensions by LLE, and

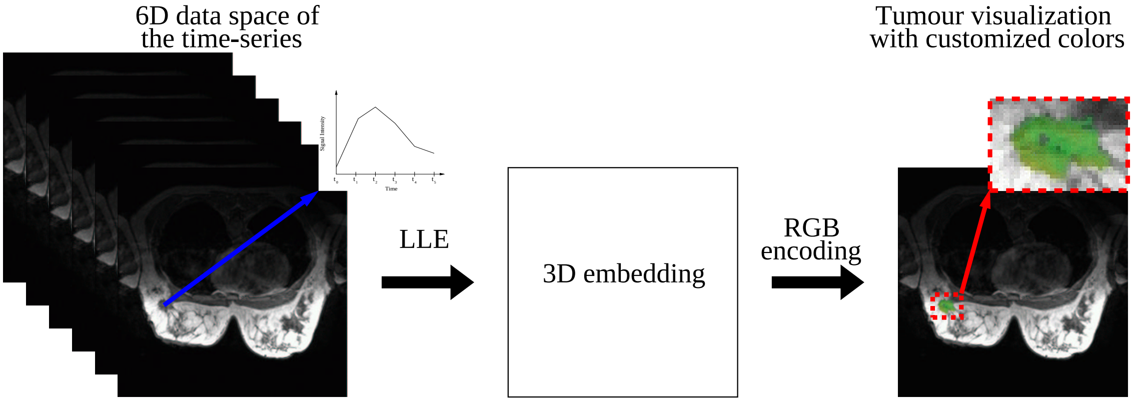

Figure 6.17: Scheme of the tumor visualization with customized colors utilizing LLE

the three coordinates obtained for each time-series are encoded within the RGB color scheme in order to visualize the tumor lesions with customized colors (see Fig 6.17).

These colors describe the dynamic of the contrast agent within the tissue. The information of the time-series is thus compacted in one single image and this allows for a fast analysis of the dynamic of the injected contrast agent, thereby providing information on the vascularization of the tumor tissue.

The proposed LLE-based approach is compared with the three-time point method (3TP), which actually is considered the state-of-the-art for the color visualization of breast tumor lesions in DCE-MRI (see section 4.3.2). The three time-points taken into account are those of the pre-contrast, the first post-contrast and the fifth post-contrast volumes.

6.3.1 Results

The correlation between the 3TP and LLE color mapping is shown in Fig. 6.18 and Fig. 6.19. In Fig. 6.18, the 2D LLE embedding (in fact the same as shown in Fig. 6.5) is plotted with colors given by the 3TP method, i. e. the color of each point is determined by the 3TP method applied to the corresponding time-series. One can observe that most of the time-series are mapped by LLE in agreement with the 3TP analysis. In fact, three different clusters (red, green and blue points) can be easily detected, suggesting that there exists a relationship between the LLE coordinates and the 3TP colors and, in turn, the LLE coordinates reflect pathophysiological features of the tissue. Points with slow wash-out (blue points) are localized in the upper part of Fig. 6.18, while points related to faster wash-out (red points) are progressively mapped to the lower region. The marker of each point specifies whether the corresponding time-series belongs to a benign or a malignant lesion. One can see that all points belonging to benign tumors are clustered by LLE only in the upper part of the figure. In the lower region there are some

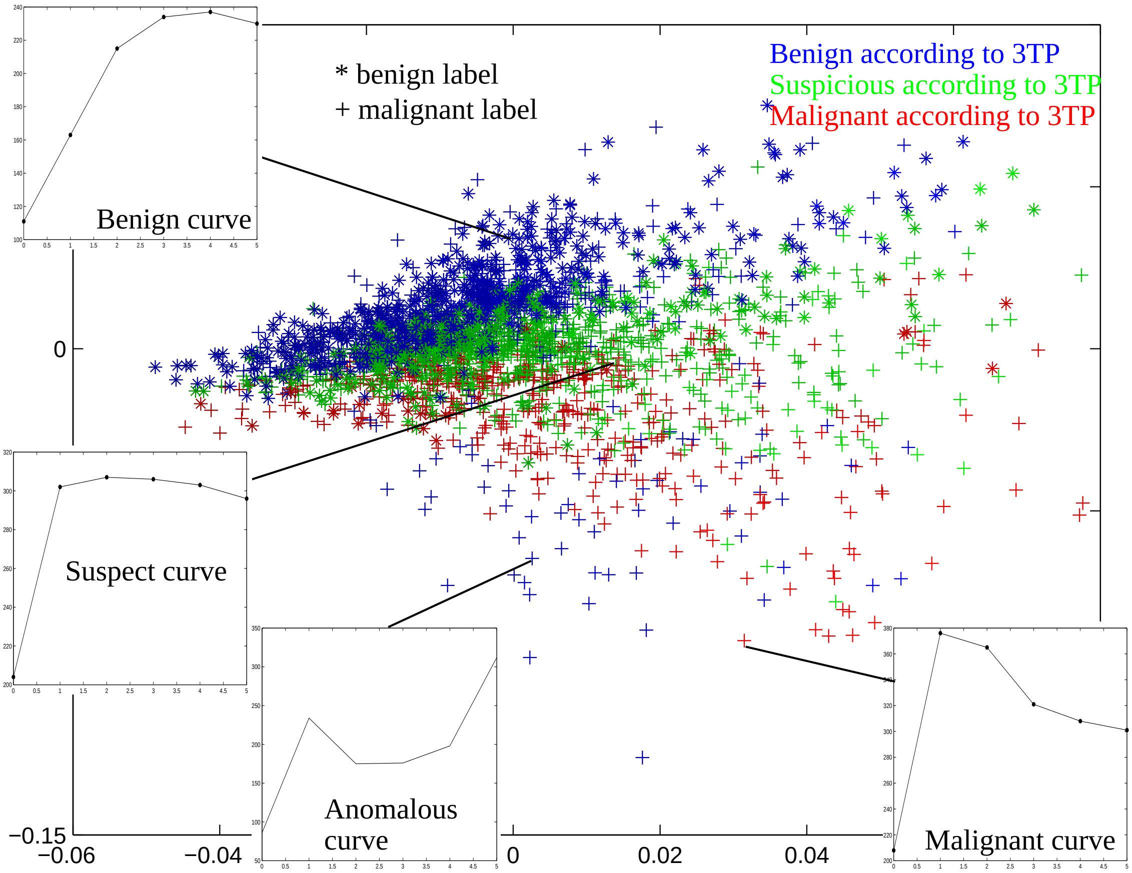

Figure 6.18: 2D Embedding coordinates from the LLE algorithm. Colors correspond to those obtained by the 3TP method considering the corresponding time-series in the DCE-MRI data space. Most of the benign points are localized in the upper part, in agreement with the 3TP analysis (encoded by the blue color according to 3TP). The majority of the points with moderate wash-out (suspicious behavior) are approximately in the central region of the mapping (encoded by the green color according to 3TP). Most of the points showing fast wash-out (malignant behavior) are localized in the lower part of the picture (encoded by the red color according to 3TP). Some points classified as benign by 3TP are instead mapped close to data points belonging to malignant lesions using LLE (blue points in the lower part of the mapping). One contrast characteristics is shown and one can see it is characterized by an untypical wash-out phase due to patient motion.

malignant points with blue color, i. e. according to the 3TP method they are labeled as benign. Conversely, the LLE algorithm maps them to the malignant part. By showing

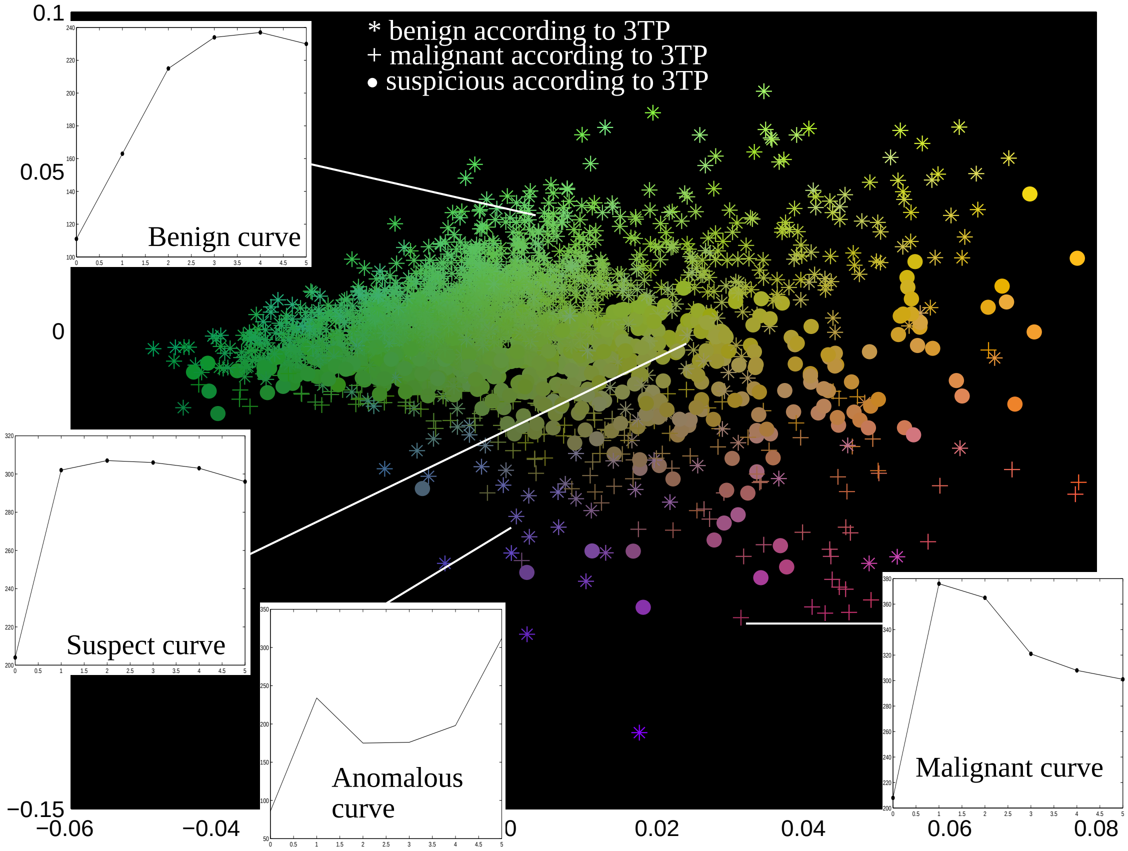

Figure 6.19: 2D LLE embedding with colors determined by the coordinates of the 3D LLE embedding. The marker of each point gives information about the color that such point has with respect the 3TP method.

one related contrast characteristics, one can observe that such points are characterized by an anomalous wash-out phase. The plot shows a clear wash-out behavior, however followed by a signal increase. This was caused by a slight movement of the patient during the acqusition of the sequence of MR volumes and was not noticed before. In this case, 3TP fails to interpret these time-series characteristics, confirming a limitation of the method that has already been mentioned in (Kelcz et al., 2002); in contrast, LLE is able to detect such abnormalities by virtue of taking into account the entire time-series and not only three components.

In Fig. 6.19 the 2D LLE coordinates with the correlated colors given by the 3D LLE embedding are shown. The marker of each point reflects the labeling scheme of the 3TP method: star-shaped points are blue in 3TP; round-shaped points are green in 3TP and plus-shaped ones correspond to red points in 3TP. According to LLE, points characterized by slow wash-out (blue points in 3TP) are encoded as green, green-yellow.

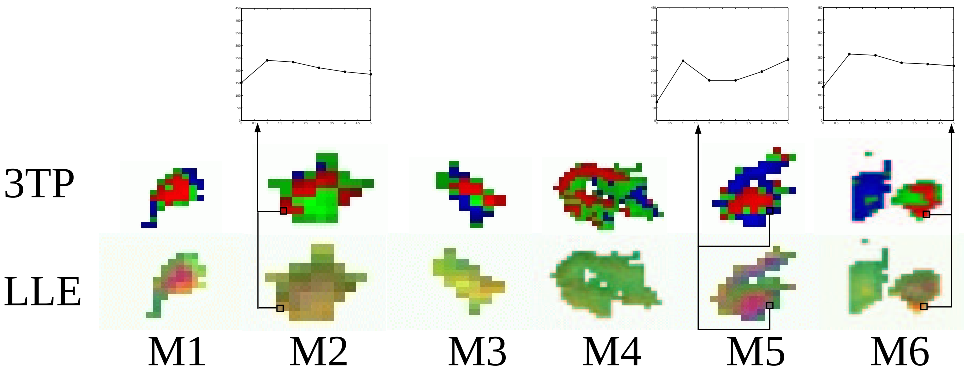

Figure 6.20: Color visualization of the tumor lesions using 3TP and LLE. Overall, the images obtained by LLE are consistent with those obtained by 3TP. In particular, the colors of the images of M1, M5 and M6 obtained by LLE are strongly correlated with the colors obtained by 3TP. Three time-series are also visualized, one for each tumor M2, M5 and M6. The visualized time-series of M5 is encoded as blue (benign) by 3TP but as blue by LLE (which means malignant according to the mapping in Fig. 6.19). This time-series is characterized by an anomalous wash-out probably due to patient motion. This anomalous behavior is interpreted correctly by LLE and wrongly by 3TP. Specifically, 3TP does not detect a wash-out phase because it takes into account only the first, second and sixth time-points along the time-series. LLE, by contrast, processes the entire time-series and is therefore able to classify it as malignant.

The green points in the 3TP method are encoded by LLE as dark-green, orange. Most of the points with a fast wash-out (red color in 3TP) are then mapped with red and blue hues. In particular, in LLE the blue points are those presenting an anomalous wash-out phase as described above.

The images of the 12 lesions obtained by 3TP and LLE are shown in Fig. 6.20 (malignant cases) and Fig. 6.21 (benign cases). Both methods generally exhibit qualitative agreement, and results show that benign tumors are largely homogeneous (low colow variation), while the malignant ones show a higher degree of heterogeneity (high color variation), in agreement with the experience of radiologists (Furman-Haran et al., 2001). In addition, a strong correlation between the variance of the colors for both methods is observed. In particular, the images of M1, M5 and M6, and of B1, B2 and B3 are particularly consistent. By observing the images of B4 and B6 in Fig. 6.21, one can see that some voxels are mapped as malignant by 3TP (red hue). The respective images obtained by LLE are in contrast more homogeneous, thereby showing higher specificity. This is particularly valuable, since the DCE-MRI technique is known to present high sensitivity, but possibly low values of specifity (as low as 30%) (Su et al., 2003).

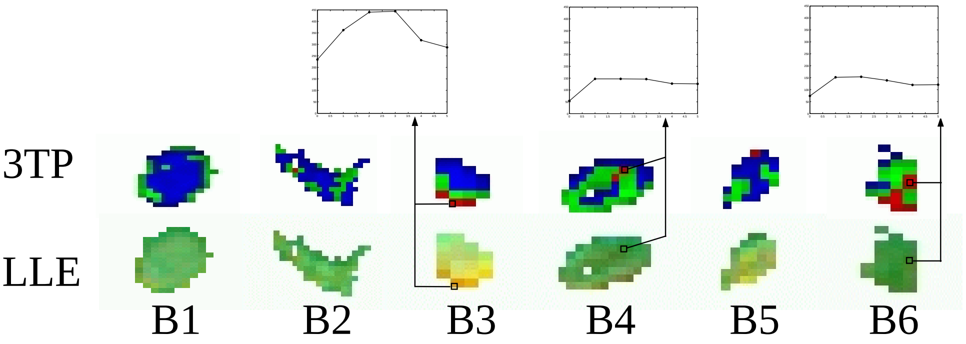

Figure 6.21: Color visualization obtained by LLE and 3TP of the benign lesions. According to 3TP, the benign lesions have some red voxels (i. e. tissue with malignant properties) which by contrast are visualized in green by LLE (which means benign according to the mapping in Fig. 6.19). The visualization of two time-series of tumors B4 and B6 shows that the absorption of contrast agent is very limited and thus the vascularity, which is an indicator of the degree of malignancy, of the respective tissue is low. This validated the LLE color scheme.

6.4 Histologic characterization of the breast lesions

In this section the information contained in the LLE embedding is expanded in order to characterize the tumor lesions histologically. Specifically, both the label (benign/malignant) and the histologic characterization of the lesions are encoded in the LLE embedding visualization. The color of a data point denotes the histological family of the respective lesion while its shape encodes the label. The experiment is carried out on the 14-cases data set composed of the tumors listed in table 2.2. and table 2.3.

6.4.1 Experiments

The dataset contains three pairs of tumors of the same histologic type, namely M1-M6 (ductal carcinoma), M3-M7 (ductal carcinoma in situ) and B3-B7 (scar). One wants to investigate whether the projections of the time-series of the tumors of each pair considered alone are located close-by in the two-dimensional space. If this turned out to be the case, then the two tumors of that particular pair would possess similar time-series.

For this purpose each of the tumors is left out at a time to generate a training dataset of the time-series of 13 tumors. This training set is reduced down to two dimensions by LLE with $n = 25$. Next, the time-series of the left-out tumor are mapped to the projected space by using the two-dimensional coordinates of the training dataset computed previously. This method is described in Fig. 6.22. In other words, the two-dimensional projections of the time-series of the left-out tumor are computed without recomputing

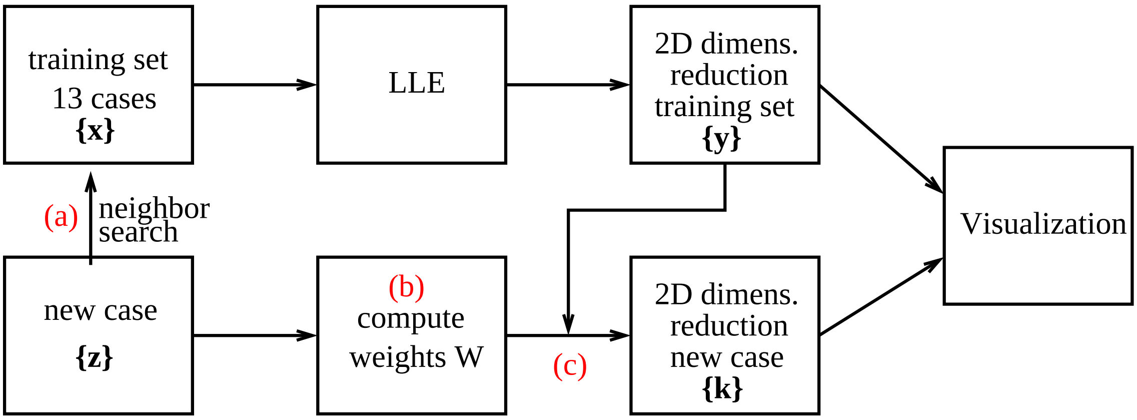

Figure 6.22: Schema of the procedure of work. The training set (time-series of 13 tumors) denoted by $\{\mathbf{x}\}$ is projected onto two dimensional vectors $\{\mathbf{y}\}$ by LLE. The time-series $\{\mathbf{z}\}$ of the new case are projected onto two dimensions in the following way: (a) identify the n nearest neighbors of $\mathbf{z}_i$ among the training vectors $\{\mathbf{x}\}$; (b) compute the weights $\{W\}$; (c) the projection of $\mathbf{z}_i$ is given by $\mathbf{k}_i = \sum_{j=1}^{n} W_j \mathbf{y}_j$, where the sum is performed over the n projected points correlated to the n neighbors of $\mathbf{z}_i$.

the LLE embedding, that is, the LLE projection of the 13 tumors is generalized as described in section 6.1 of (Lawrence and Roweis, 2003) in order to analyze a new tumor case.

6.4.2 Results

Here the mappings of the tumors of the same histologic type are analyzed and discussed. The mapped tumor (termed as new case in Fig. 6.22) is visualized in white, while the black points denote the points of the tumor from the training set and of the same histologic type of the new case. The clusters related to these two tumors are highlighted by traced ellipses. Furthermore, three representative time-series along with their projections are visualized for both tumors.

The mappings of the tumors M1-M6 (ductal carcinoma) are visualized in Fig. 6.23. Their clusters largely overlap, thereby suggesting their respective time-series exhibit similar signal dynamic. The visualization of three time-series for each tumor confirms this fact. In this case tumors of the same histologic type are thus mapped to the same part of the embedding.

The mappings of the tumors M3-M7 (ductal carcinoma in situ (DCIS)) are visualized in Fig. 6.24. The mapping of M3 lies on the opposite side of where the points of M7 are. Therefore, these lesions appear to differ with respect to the dynamic of the contrast agent. The visualization of three representative time-series show that those of M3 are

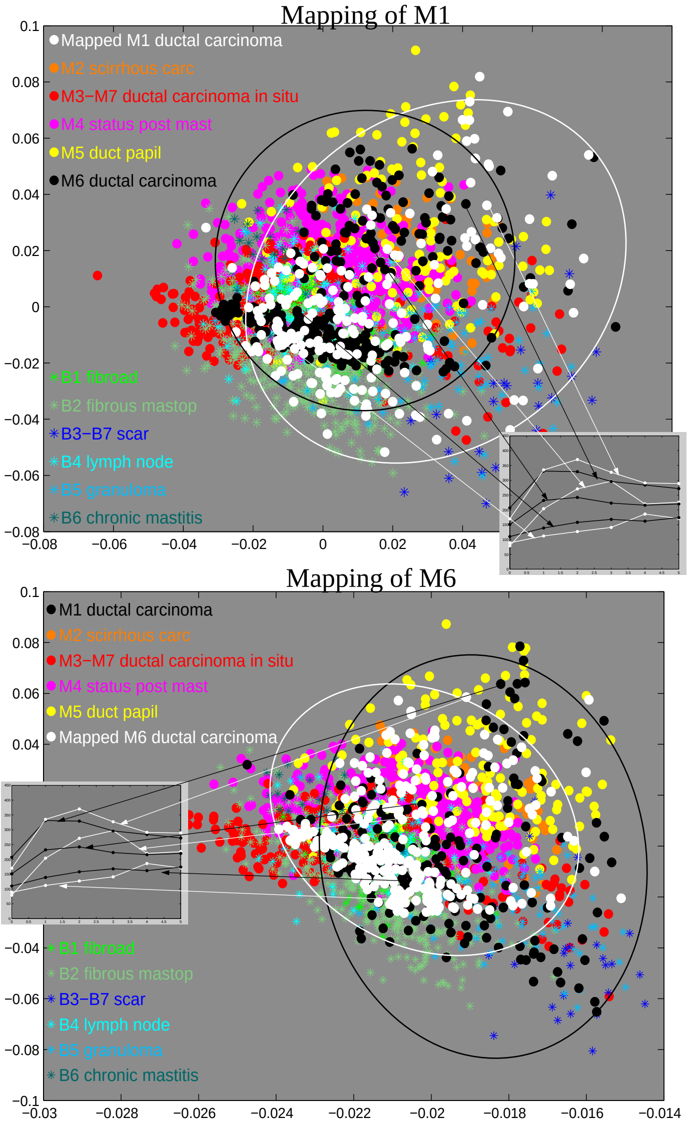

Figure 6.23: Visualization of the mappings of tumors M1 (top) and M6 (bottom). In both cases the clusters of M1 and M6 largely overlap, thereby suggesting a high degree of similarity with respect to the DCE-MRI time-series. In addition, three couples of points are correlated with the respective time-series, whose visualization confirms their similarity.

characterized by a higher signal intensity. This agrees with results documented in the literature (Tan, 2001), i. e. breast DCIS lesions can be heterogeneous.

The mappings of the tumors B3-B7 (scar) are shown in Fig. 6.25. The clusters related

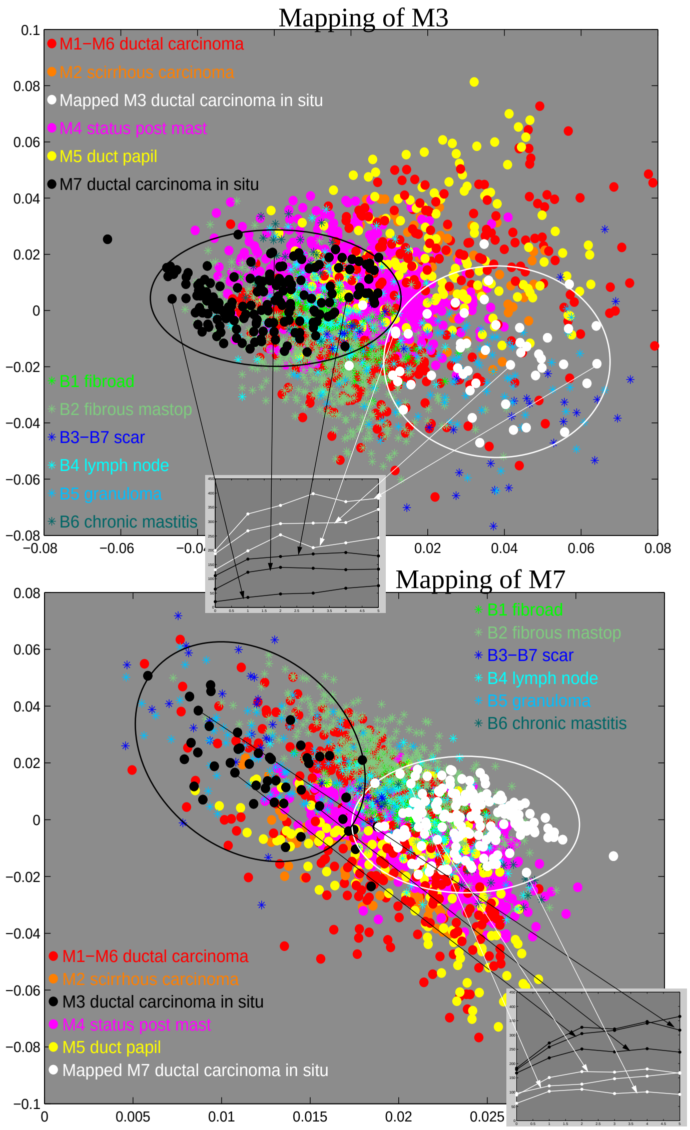

Figure 6.24: Visualization of the mappings of tumors M3 (top) and M7 (bottom). In both cases the clusters are located in different region of the projected space. Thus, tumors M3 and M7 exhibit different time-series behaviors, as confirmed by the visualization of three of them.

to the two tumors do not overlap, thereby revealing a certain degree of divergence among the respective time-series. The visualization of three time-series for each tumor shows that the time-series of B3 are characterized by larger values of signal intensity.

If one imagines that the histologic family of the initially left out tumor is not known,

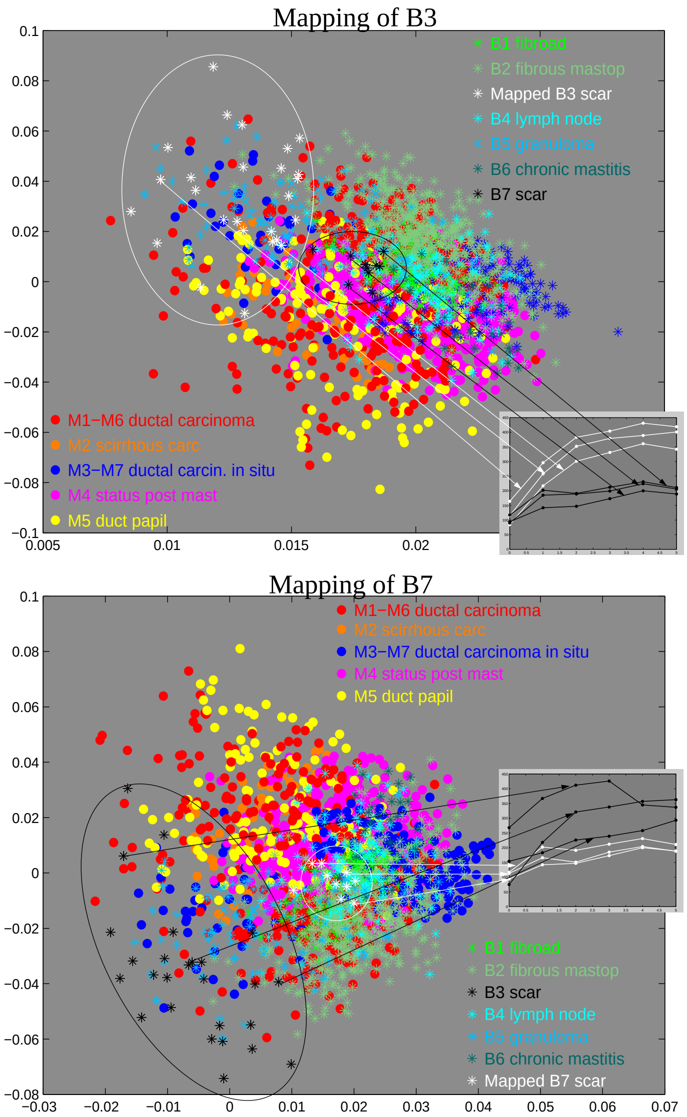

Figure 6.25: Visualization of the mappings of tumors B3 (top) and B7 (bottom). Like in Fig. 6.24, the clusters do not overlap in both mappings. This means the two tumors are different with respect to the time-series dynamic. The visualization of three time-series for each tumor evinces that those of B3 exhibit the largest values of signal intensity.

only in the cases of tumors M1-M6 one can infer their histologic family from the observation of the visualized LLE projection. Let M1 be a tumor one wants to characterize histologically, i. e. one supposes its histologic family is not known. If the points of M1 are projected into the LLE embedding of the 13 remaining tumors, one can note its points are localized close to those of M6, which is a ductal carcinoma. One could thus speculate that also M1 is a ductal carcinoma, as it actually is. This is true also if one considers M6 as the tumor one wants to characterize histologically instead of M1.

However, this procedure is not valid for tumors M3-M7-B3-B7. Indeed, their mappings may lead to draw misleading conclusions, as the points of tumors of the same histologic type are mapped to different regions in the LLE projection. In other words, there is not a unique way to histologically characterize the tumors M3-M7-B3-B7 from the observation of the respective LLE projections because the M3 (B3) cluster does not overlap with the M7 (B7) one and vice versa.

Note that the visual comparison of some histologic families in the display may prove difficult because of other overlapping clusters. In this case clusters that are not easily visible can be put in the foreground by clicking on the correlated text on the legend.

6.5 Conclusions of this Chapter

This chapter describes the analysis of DCE-MRI data based on LLE. This analysis can be divided into three parts, namely comparison between benign and malignant time-series, tumor visualization with customized colors and histologic characterization of tumors. In the first part the performance of LLE algorithm has also been compared with those of SOM and PCA.

The LLE algorithm has proved to be a powerful instrument for the visual exploration of DCE-MRI data having time-series of different tumors, as it allows to visualize the signal dynamic of all time-series in a single scatter plot. This makes it possible to detect similarities and relationships between the time-series.

The LLE algorithm has also proved useful for the visualization of tumor lesions with customized colors as it yields images that are comparable with those obtained by 3TP. The advantage is that the LLE approach is, in contrast to the 3TP, model-free. Consequently, the usage of LLE is not constrained by the existence of a model and the algorithm can in principle be used for the color visualization of other organ diseases, such as liver cancer, for which it is very difficult to model the dynamic of an injected contrast agent.

7 Modifications of LLE

This chapter illustrates two modifications of the LLE algorithm. These two modifications regard the first step (neighbor search) and second step (weight computation) of LLE, respectively (see section 5.2.2 for the explanation of these steps). These modifications attempt to overcome two possible weaknesses of LLE.

The first weakness regards the fact that the n nearest neighbors are typically computed according to the Euclidean distance. In section 7.1 it is explained that this fact can lead to some problems if n is high. The proposed variation of LLE for solving these problems consists in using the geodesic distance instead of the Euclidean metric.

The second weakness is that LLE has, in addition to n, a further input parameter (Δ), that is required to compute the weights W, in case $n > D$ (see eq. (5.17)). The value of Δ can influence the LLE output and therefore needs to be tuned accurately. The second variation of LLE represents an innovative manner of computing the weights W with the advantage that the parameter Δ is not required any longer. This variation is explained in section 7.2.

7.1 ISOLLE: LLE with Geodesic Distance

The LLE algorithm assumes linearity in the local area centered on each data point. Each area is mathematically characterized by a set of coefficients (weights) that correlate the particular data point with its n nearest neighbors. The aggregation of all areas can be intuitively thought as an assemblage of linear patches which approximates the nonlinear data structure. The high-dimensional data is then projected into a lower-dimensional space while preserving the weights between neighboring data points (see also section 5.2.2).

The number of neighbors n strongly influences the accuracy of the linear approximation of the nonlinear data structure. Intuitively, supposing the data is sufficiently dense, the smaller n, the smaller the area, the more faithful is the linear approximation. This is similar to considering a 2D curve approximated by a set of joined segments: the shorter the segments, the more faithful is the linear approximation of the curve.

However, the value of n must be sufficiently high, otherwise the local areas may become too small and, in turn, disjoint. This is equivalent to saying that the graph constructed by connecting each point $\mathbf{x}_i$ with its n nearest neighbors is disjoint (see Fig. 7.1 left). Disjoint areas can be obtained especially when the data is sparse or spread among multiple clusters. In case the areas are disjoint, LLE can fail to detect the global

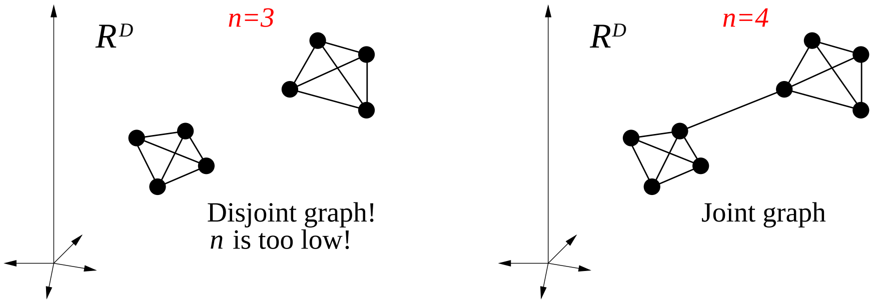

Figure 7.1: Given a data set with eight points, if $n = 3$ (left), the graph constructed by connecting each point with its n nearest neighbors is disjoint. In such a case LLE would fail to detect the global structure of the data with detrimental effects on the dimensional reduction (Polito and Perona, 2001). With $n = 4$ (right) the graph is joint.

structure of the data (Polito and Perona, 2001), thereby resulting in an unsatisfying dimensional reduction. To address this problem, in (Vlachos et al., 2002) it is proposed to search for the $n/2$ nearest and $n/2$ farthest neighbors of each data point. Another approach is given in (Hadid and Pietikäinen, 2003), where the authors suggest to connect the disjoint manifold or interpolating the embeddings of some samples.

In general, for larger values of n the linear areas are more likely to overlap. This is equivalent to saying that the graph of the neighboring points is more likely to be joint (see Fig. 7.1 right). The number of neighbors n therefore needs to be sufficiently high to satisfy this condition.

On the other hand, as the neighbors search is typically conducted using the Euclidean distance, this may lead to the condition that a data point will have neighbors which are instead very distant with regard to the intrinsic geometry of the data. More intuitively, one can imagine this fact as a short circuit (see Fig. 7.2). The presence of short circuits is undesirable, as they can cause LLE to misinterpret the real data structure.

To address the above outlined problems occurring to LLE when employed with a high number of neighbors n, the usage of LLE with geodesic distance is proposed. More specifically, the n nearest neighbors are searched with respect to the geodesic distance. The combination of LLE and geodesic distance is termed as ISOLLE. This geodesic distance has already been employed in other methods for nonlinear dimensional data reduction such as Isomap (Tenenbaum et al., 2000), Curvilinear Distance Analysis (Lee et al., 2000) and Self-organizing Maps (Wu and Takatsuka, 2005). The geodesic distance between two data points can be intuitively thought as their distance along the contour of an object (see Fig. 7.2 right). For instance, consider the distance between Paris and New York. Their geodesic distance is the distance along the curvature of the Earth. Their Euclidean distance instead is the length of the straight line connecting the two cities which is below the level of the ground. Points faraway from

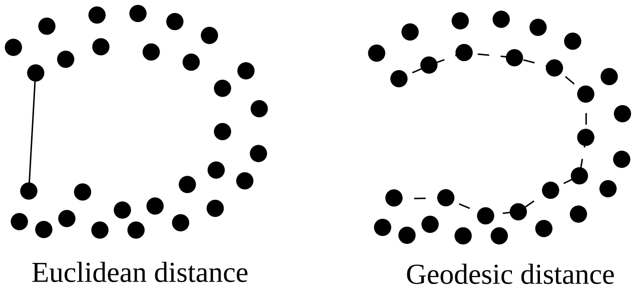

Figure 7.2: The short circuit induced by Euclidean distance is shown on the left. In case the number of neighbors n is set to a relatively high value, the two points in the figure may be treated as neighbors, although they belong to opposite parts of the horseshoe. This may cause LLE to fail to detect the real global structure of the data. The benefits of the geodesic distance are shown on the right. In this case the two points are not neighbors, as they are faraway according to the geodesic distance.

each other, as measured by the geodesic distance, may appear deceptively close in the high-dimensional input space as measured by the Euclidean distance.

In this section it is shown that the employment of the geodesic distance can lower the probability to create short circuits during the neighbors search, thereby allowing for a more accurate dimensional reduction. The comparison of the performances of ISOLLE and LLE is conducted by analyzing two data sets. The first is a synthetic data set, namely a three-dimensional swissroll (see Fig. 7.3), which was also used in (Tenenbaum et al., 2000) and (Roweis and Saul, 2000). The second data set is the DCE-MRI data set with 12 tumor cases that has already been analyzed in chapter 6. The number of points of the datasets along with their data dimension are displayed in table 7.1.

Both data sets are reduced to two dimensions for different values of the number of neighbors n. The dimensional reduction of the swissroll is evaluated qualitatively, while the analysis of the DCE-MRI data set requires a statistical approach because of the complexity of the data. In addition, in the final part the running times of LLE and ISOLLE are also compared.

Table 7.1: Data sets analyzed by LLE and ISOLLE.

Data set	Number of points v	Dimension D
Swissroll	1000	3
12-cases DCE-MRI data	2449	6

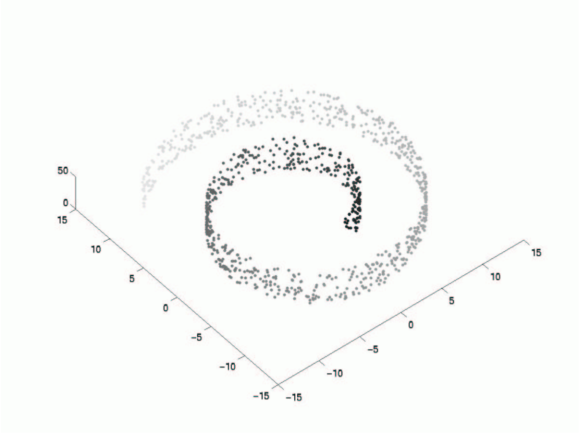

Figure 7.3: Three-dimensional swissroll composed of 1000 points. Its dimensional reduction onto two dimensions should ideally result in a rectangle. The reason for this fact is intuitive if one thinks that a swissroll can be obtained by folding a sheet of paper. If you unfold the swissroll, in the same way you roll out a sleeping bag, then you get the sheet of paper, i e. a two-dimensional structure.

7.1.1 The ISOLLE algorithm

The ISOLLE algorithm differs from LLE only in the first step, i.e. the neighbors search. More specifically, ISOLLE computes the n nearest neighbors of each data point according to the geodesic distance. For this purpose a small variation of Dijkstra's algorithm (Dijkstra, 1959) is employed. Given a graph where each node is a data point, this algorithm computes the shortest paths from a particular node to all remaining nodes. In this work the computation is restricted to the n shortest paths, i. e. for each node only the n nearest points according to the geodesic distance are computed. An animated example of Dijkstra's algorithm can be seen at (Pemmaraju and Skiena, 2003).

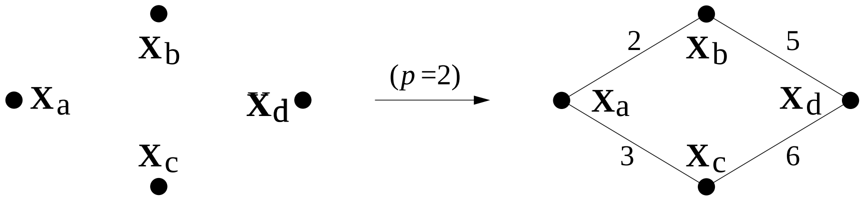

Figure 7.4: Construction of the neighborhood graph for a data set of four points. The graph is obtained by connecting each point with its p-nearest neighbors ($p = 2$) according to the Euclidean distance. The weight of each edge is given by the Euclidean distance between the two respective points.

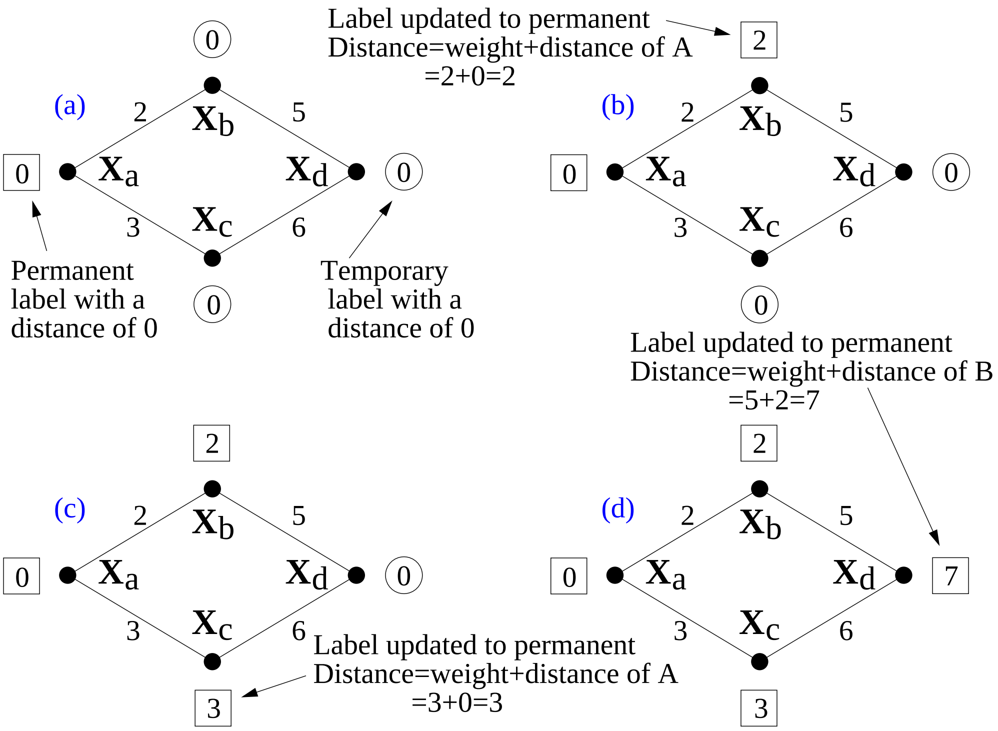
Label updated to permanent
Distance=weight+distance of A
=2+0=2
(a)
(b)
(c)
(d)
0
2
3
7
2
5
3
6
Xa
Xb
Xc
Xd
Permanent label with a distance of 0
Temporary label with a distance of 0
Label updated to permanent
Distance=weight+distance of B
=5+2=7
Label updated to permanent
Distance=weight+distance of A
=3+0=3

Figure 7.5: Scheme of Dijkstra's algorithm. (a) The algorithm begins by initializing any vertex in the graph (vertex $\mathbf{x}_a$, for example) with a permanent label with the value of 0, and all other vertices with a temporary label with the value of 0. (b) The algorithm then proceeds to select the least cost edge connecting a vertex with a permanent label (currently vertex $\mathbf{x}_a$) to a vertex with a temporary label (vertex $\mathbf{x}_b$, for example). Vertex $\mathbf{x}_b$'s label is then updated from a temporary to a permanent label. Vertex $\mathbf{x}_b$'s value is then determined by the addition of the cost of the edge with vertex $\mathbf{x}_a$'s value. (c) The next step is to find the next least cost edge extending to a vertex with a temporary label from either vertex $\mathbf{x}_a$ or vertex $\mathbf{x}_b$ (vertex $\mathbf{x}_c$, for example), change vertex $\mathbf{x}_c$'s label to permanent, and determine its distance to vertex $\mathbf{x}_a$. This process is repeated until the labels of n vertices in the graph are permanent.

The process of finding the geodesic neighbors is composed of two phases. The first phase consists in constructing a weighted graph G over the data set where neighboring data points are connected. In principle, any similarity measure d_E can be adopted to determine neighboring relations, and probably the Euclidean distance is the most common choice. Two points are neighbors if they are closer than a fixed distance ε (ε-graph), or one is the q nearest point of the other (q-graph). These relations between neighbors are represented by edges of weights $d_E(\mathbf{X}_i, \mathbf{X}_j)$ (Tenenbaum et al., 2000).

An example of graph construction is shown in Fig. 7.4.

In the second phase the n nearest neighbors of each data point are found according to the geodesic distance computed by Dijkstra's algorithm. This algorithm begins at a specific node (source vertex) and extends outward within the graph until all the vertices have been reached (in this work only the n nearest nodes). Dijkstra's algorithm creates labels associated with vertices. These labels represent the distance (cost) from the source vertex to that particular vertex. Within the graph, there exist two kinds of labels: temporary and permanent. The temporary labels are given to vertices that have not been reached. The value given to these temporary labels can vary. Permanent labels are given to vertices that have been reached and their distance (cost) to the source vertex is known. The value given to these labels is the distance (cost) of that vertex to the source vertex, that is equal to the geodesic distance. For any given vertex, there must be a permanent label or a temporary label, but not both. An example of computation of geodesic distance is shown in Fig. 7.5.

Both steps of the neighbor search according to the geodesic distance are detailed as follows:

Construct the neighborhood graph: define the graph G over all data points by connecting points $\mathbf{x}_i$ and $\mathbf{x}_j$ if (as measured by $d_E(\mathbf{x}_i, \mathbf{x}_j)$) they are closer than ε (ε-graph), or if $\mathbf{x}_i$ is one of the q nearest neighbors of $\mathbf{x}_j$ (q-graph). Set edge lengths equal to $d_E(\mathbf{x}_i, \mathbf{x}_j)$.

Compute n nearest points with Dijkstra's algorithm: given a graph G=(V,E) where V is a set of vertices and E a set of edges, Dijkstra algorithm keeps two sets of vertices:

S $-$the set of vertices whose shortest paths from the source vertex have already been determined. These vertices have a permanent label

V$-$S $-$the remaining vertices. These have a temporary label

The other data structures needed are:

$\mathbf{x}_0$ $-$initial beginning vertex (source vertex)

N $-$number of vertices in G

D $-$array of estimates of shortest path to $\mathbf{x}_0$.

The basic mode of operation of Dijkstra's algorithm is:

```
1 S={$\mathbf{x}_0$}
2 For i=1 to $N$
      D[i]=E[$\mathbf{x}_0$,i]
```

```
3 For i=1 to $N-1$
    Choose a vertex w in V-S such that D[w] is minimum
    Add w to S
    For each vertex v in V-S
      D[v]=min(D[v],D[w]+E[w,v])
```

The construction of the graph G requires a further parameter (ε or q) to be set by the user. In (Tenenbaum et al., 2000) it is pointed out that the scale-invariant parameter q is typically easier to set than ε, but may yield misleading results when the local dimensionality varies across the data set. A sensible way to set this parameter can be to choose the minimal value such that all the pairwise geodesic distances are finite.

7.1.2 Procedure of Work

The difference between LLE and ISOLLE are illustrated by considering the three-dimensional swissroll. At first, the graphs obtained by connecting each data point with its n neighbors by an edge are visualized. This is done for different values of n. The n nearest neighbors are computed with respect to the Euclidean distance in LLE and the geodesic distance in ISOLLE. The visualization of the graphs makes it possible to check for the possible presence of short circuits induced by the Euclidean metric. The effects of these short circuits are then evaluated qualitatively by visualizing the respective two-dimensional projections, i. e. the projections of the swissroll obtained by LLE and ISOLLE for those particular values of n.

The DCE-MRI data set is reduced down to two dimensions by both LLE and ISOLLE for n varying between 5 and 40. The evaluation of the dimensional reduction of the DCE-MRI data set requires a statistical analysis, as the output can not be predicted a priori because of the complexity and multi-dimensional nature of the data, and consequently it is not possible to visually evaluate the accuracy of the low-dimensional projection. For this reason the quality of the DCE-MRI projections is evaluated by means of two numerical quantities, namely neighborhood preservation (NP) (section 5.4.3) with $t=5$ and stress (ST) (section 5.4.2).

Prior to the computation of the value of stress, both the original and embedded coordinates are scaled to [0,1] in order to allow for a correct comparison between different embeddings.

The neighborhood graphs of the two data sets for computing the geodesic distances are ε-graphs with the values of ε fixed to the minimal possible value such that all the pairwise distances are finite, which is equivalent to saying that the neighborhood graph is joint. The values empirically found for each data set are: ε(swissroll)=5; ε(DCE-MRI data)=90.

7.1.3 Results

This section begins with the analysis of the swissroll data set. The DCE-MRI data set is analyzed in the second part. In the final part the LLE and ISOLLE algorithms are compared in terms of running time.

In Fig. 7.6 one can see six graphs obtained by connecting each point of the swissroll with its n nearest neighbors for $n = 6$, $n = 15$, and $n = 40$. These neighbors were computed with respect to the Euclidean and geodesic distance. Note that the graphs are three-dimensional but are visualized in section in order to make the short circuits more visible.

Already with $n = 15$, the graph obtained with the Euclidean distance has many short circuits. With $n = 40$ the number of short circuits increases considerably. By contrast, the graphs obtained with the geodesic distance do not present short circuit effects, even when the number of neighbors n equals 40. This shows that the usage of the geodesic distance can drastically reduce the number of short circuits. This means that the neighbor search conducted in LLE with $n = 15$ or $n = 40$ is afflicted by many short circuits. Conversely, the neighbor search conducted in ISOLLE does not suffer from this problem, irrespective of n.

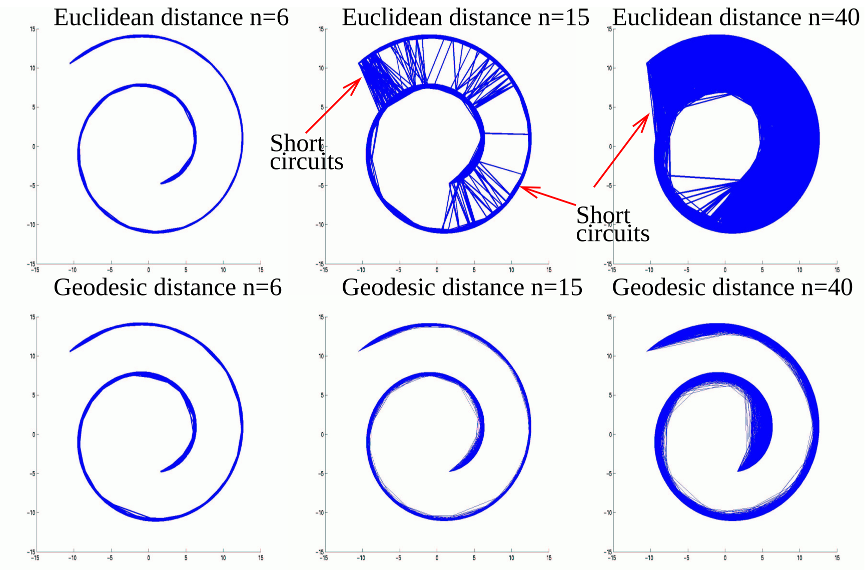

Figure 7.6: Neighbors graphs of the swissroll data set obtained according to the Euclidean and geodesic distances. In the Euclidean graph with $n = 15$ there are already short circuits. Their number considerably increases with $n = 40$. Conversely, in all the geodesic graphs there are no short circuits. Note that the three-dimensional swissroll is shown in section in order to ease the detection of the short circuits.

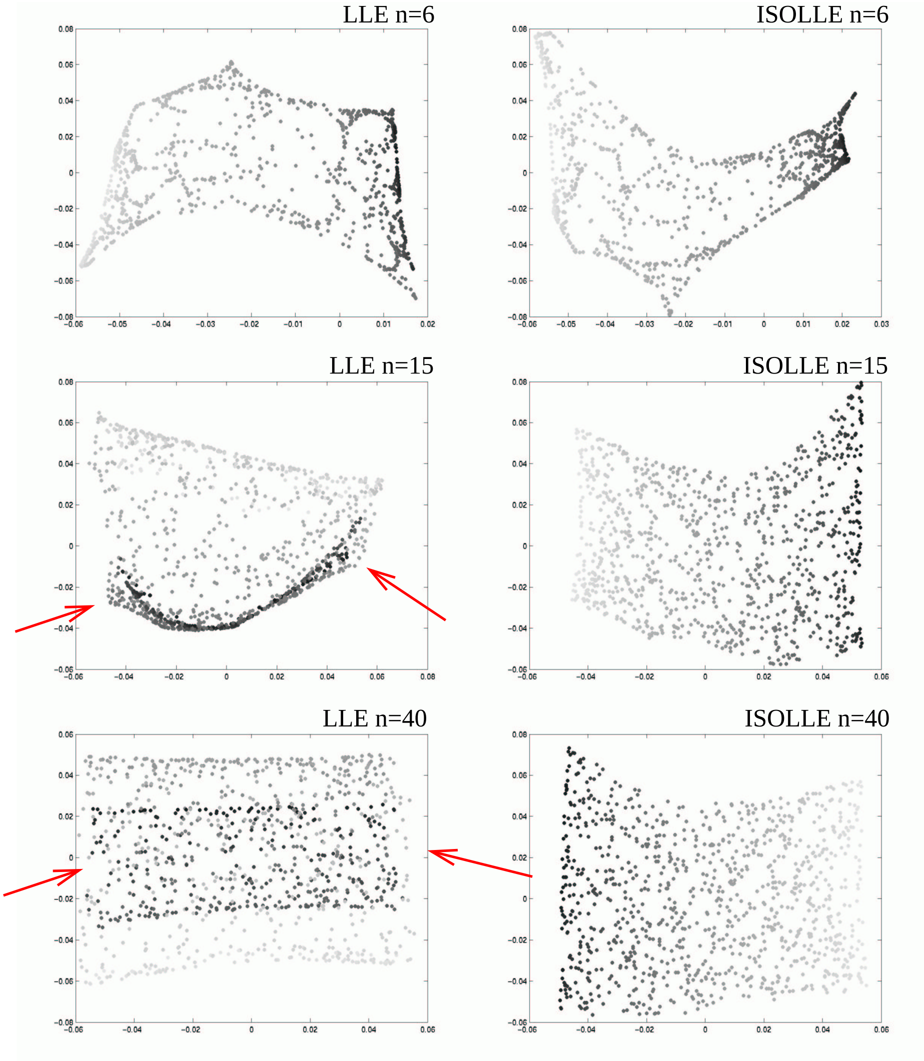

Figure 7.7: Two-dimensional reductions of the swissroll data set obtained by LLE and ISOLLE for n given in Fig. 7.6. While LLE fails to preserve the structure of the swissroll with $n \geq 15$, ISOLLE yields a good projection of the data in all cases. The arrows show where the original structure of the swissroll is preserved wrongly.

Possible effects of these short circuits on the two-dimensional projection of the swissroll data set can be seen in Fig. 7.7. Here it is clear that LLE fails to preserve the global structure of the data with $n = 15$ and in particular $n = 40$, as in both cases the darkest points are mapped close to brighter points. On the contrary, ISOLLE can correctly unfold the swissroll in all three cases, and the structure of the data is clearly preserved. In particular, the ISOLLE projection is also accurate with $n = 40$, while the respective LLE projection results completely incorrect.

The evaluation of the dimensional reduction of the DCE-MRI data set is conducted by taking into account the neighborhood preservation and stress measures. Their average values along with the respective variances computed with respect to n between 5 and 40 are displayed in table 7.2.

The projections by ISOLLE result better with respect to both quantities. Indeed, the average ST value is lower than the one by LLE, suggesting that ISOLLE better preserves the metric of the tumor data. The higher value of the average NP by ISOLLE gives evidence that this algorithm also leads to a better preservation of the topology of the data. Two scatter plots of the DCE-MRI breast data embeddings obtained by LLE and ISOLLE with $n = 20$ are shown in Fig. 7.8. Interestingly, the benign cluster in the projection obtained by ISOLLE appears more localized and compact than in the projection obtained by LLE. Moreover, benign and malignant data exhibit a larger overlap in the projection obtained by LLE. This indicates that ISOLLE can better separate benign from malignant data and this is of considerable value from the medical point of view.

Finally, the performances of LLE and ISOLLE are compared in terms of running time. Both algorithms were performed with different n on a Pentium IV 2.8 GHz. The respective values of running times are shown in table 7.3. The ISOLLE algorithm involves a larger computation time and the divergence of speed becomes more marked as n increases. The higher computational time of ISOLLE is somewhat expected as the algorithm requires a further step as compared to LLE, i. e. the construction of the neighborhood graph over all data points (see Fig. 7.4).

7.1.4 Discussion

In general, the usage of ISOLLE should be preferred to LLE in particular when n needs to be relatively high (for example in case of sparse or clustered data) and in turn short circuits are more likely to occur. One way to determine if a certain data set requires a relatively high value of n is to perform an analysis of the smallest eigenvalues of matrix $\mathbf{S}$ from eq. (5.19). Specifically, in standard conditions matrix $\mathbf{S}$ has only one eigenvalue close to zero. However, if n is so small that the linear areas are disjoint, then matrix $\mathbf{S}$ will have more than one *close-to-zero* eigenvalue (Polito and Perona, 2001). Therefore, the minimum n for which $\mathbf{S}$ has only one eigenvalue close to 0 can be taken into account

Table 7.2: Average and variance values of stress (ST) and neighborhood preservation(NP) computed for the DCE-MRI data set.

ST(LLE)	ST(ISOLLE)	NP(LLE)	NP(ISOLLE)
0.454±0.034	$\mathbf{0.337 \pm 0.025}$	0.081±0.001	$\mathbf{0.115 \pm 0.002}$

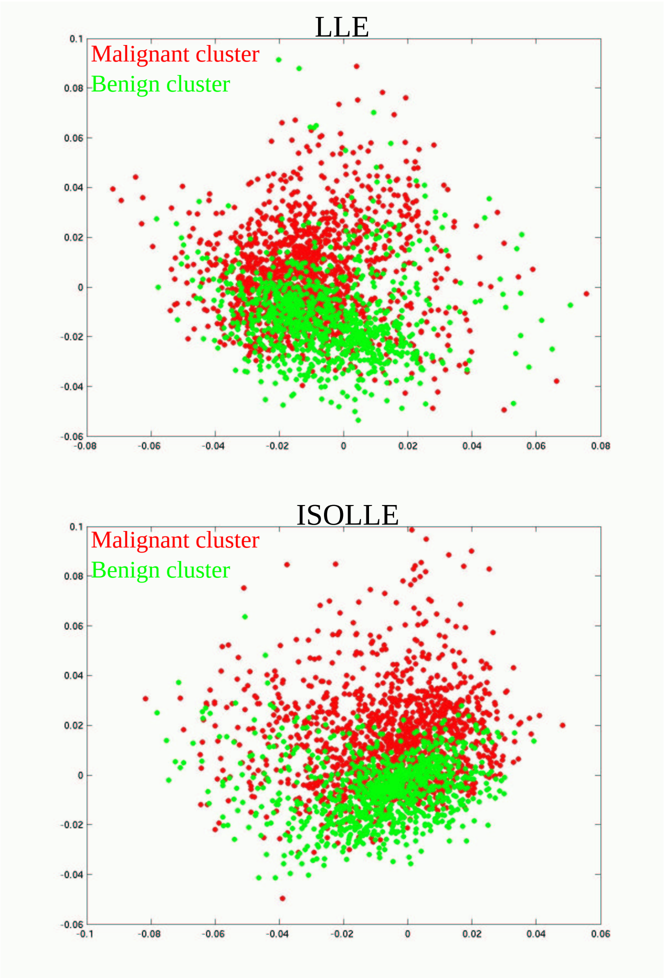

Figure 7.8: Two scatter plots of the two-dimensional embeddings of the DCE-MRI breast data set obtained by LLE and ISOLLE. In both cases n equals 20. Note that the benign and malignant clusters overlap less in the ISOLLE embedding. In particular, here the benign cluster is more compact and localized.

in order to evaluate which algorithm is more suited for the analysis of the data.

7.1.5 Conclusions concerning ISOLLE

In section 7.1 a new approach to the neighbor search in LLE based on the geodesic distance is proposed. Its usage can reduce the number of short circuits considerably, thereby improving the preservation of the data structure. This is shown by investigating the neighbors graphs obtained by LLE and ISOLLE on a synthetic three-dimensional

Table 7.3: Table of the running times in seconds.

n	Swissroll		DCE-MRI	
	LLE	ISOLLE	LLE	ISOLLE
10	0.20	2.66	1.26	16.55
20	0.23	6.32	1.39	39.24
30	0.27	11.25	1.62	69.31
40	0.30	17.29	1.75	106.37

swissroll. The ISOLLE graphs do not exhibit short circuits, even when the number of neighbors is high. By contrast, the standard neighbors search with Euclidean distance in LLE causes many short circuits. As a consequence, ISOLLE can detect the intrinsic two-dimensional structure of the swissroll with both small and large values of the number of neighbors n. Conversely, LLE fails to unfold the swissroll with $n \geq 15$.

Regarding the dimensional reduction of the DCE-MRI data set, the results clearly show that ISOLLE significantly outperforms LLE in terms of both stress and neighborhood preservation.

Experiments concerning the running times revealed that ISOLLE is slower than LLE and this becomes more noticeable as n increases.

In conclusion, ISOLLE exhibits a superior ability to project the investigated data sets into a two-dimensional space while preserving the original data structure but at the cost of a larger running time.

7.2 Distance-Weighted LLE (DWLLE)

The LLE algorithm is typically described as having one input parameter, the number of neighbors n. Actually, this is not completely correct. Indeed, a second parameter, the regularization term Δ, is required in case $n > D$ (see eq. (5.17)). In the literature this term is rarely mentioned and in (Lawrence and Roweis, 2003) is generically suggested to set its value much lower than 1. In practice, this parameter can significantly affect the final result. Three two-dimensional projections of the three-dimensional S-manifold data set visualized in Fig. 7.12 are shown in Fig. 7.9. These projections are obtained by LLE with $n = 26$ and Δ varying between 0.031 and 0.056. Despite the fact that the Δ varies within a very small range, the respective two-dimensional projections appear very different. In particular, only the projection obtained with $\Delta = 0.041$ reflects faithfully the original data structure of the S-manifold while the other two are wrong. This shows clearly that Δ can considerably affect the LLE projection.

It is obvious that getting rid of Δ would have immediate utility to the LLE user. Specifically, an alternative version of the LLE algorithm that does not require Δ even with $n > D$ and, at the same time, yields results consistent with those obtained by

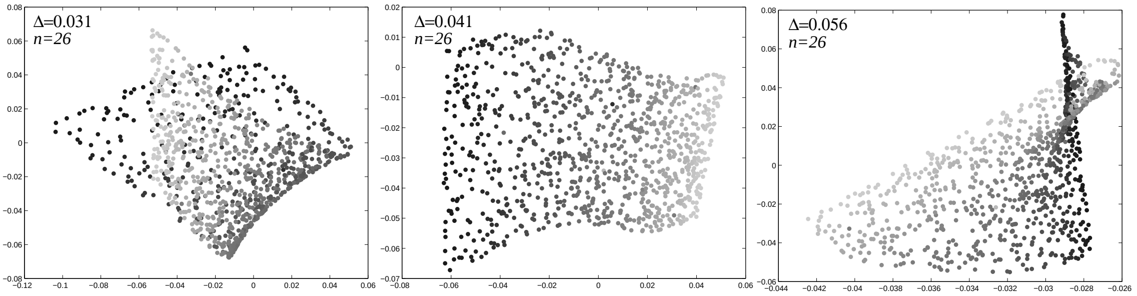

Figure 7.9: Three projections of the S-manifold of Fig. 7.12 obtained by LLE with $n = 26$ and different values of Δ. One can observe that apparently small changes in Δ lead to extremely different results. In particular only the projection relative to $\Delta = 0.041$ is accurate while the other ones do not reflect the original structure of the S-manifold.

standard LLE would be desirable.

In this section a modification of the LLE algorithm, termed Distance-Weighted LLE (DWLLE) is introduced. The main characteristic of DWLLE is that it does not require the parameter Δ. Indeed, DWLLE computes the weights W between neighboring data points without a linear system and, in turn, Δ is not needed any longer. The basic idea of DWLLE is to assign the largest weights to the closest neighbors of a certain data point. The closer the neighbor, the larger its weight. This approach is based on the fact that in LLE each data point $\mathbf{x}_i$ is approximated by the sum of its n neighbors multiplied by the respective linear weights (see eq. (5.12)). These linear weights are computed by eq. (5.15). In DWLLE, the closest neighbors are forced to have the largest weights. In this way, the closest neighbors neighbors play the major role in the data point approximation. The closer a neighbor $\mathbf{x}_j$ is to the data point $\mathbf{x}_i$, the larger will be the linear weight W_{ij}.

An experimental justification for DWLLE is shown in Fig. 7.10. The plots visualize the average weights computed for the swissroll data set by standard LLE versus the ranking of the neighbors. Specifically, the weights correlating each point $\mathbf{x}_i$ with its n-closest neighbor have been averaged over the whole data set. The result is the average value of the weight between a point and its n-nearest neighbor. Four different plots of the average weights versus the neighbor ranking are visualized in Fig. 7.10. They refer to the 5-, 10-, 20- and 30-nearest neighbors, respectively. In all four cases the closest neighbors have, on average, the largest weights as computed by eq. (5.15), i. e. the closest neighbors have the most predominant contribution to the point approximation, that is the basic principle behind DWLLE. This principle is illustrated in Fig. 7.11.

The DWLLE algorithm is introduced in the next section. Then, the data and experiments conducted to compare LLE and DWLLE are explained in section 7.2.2. This

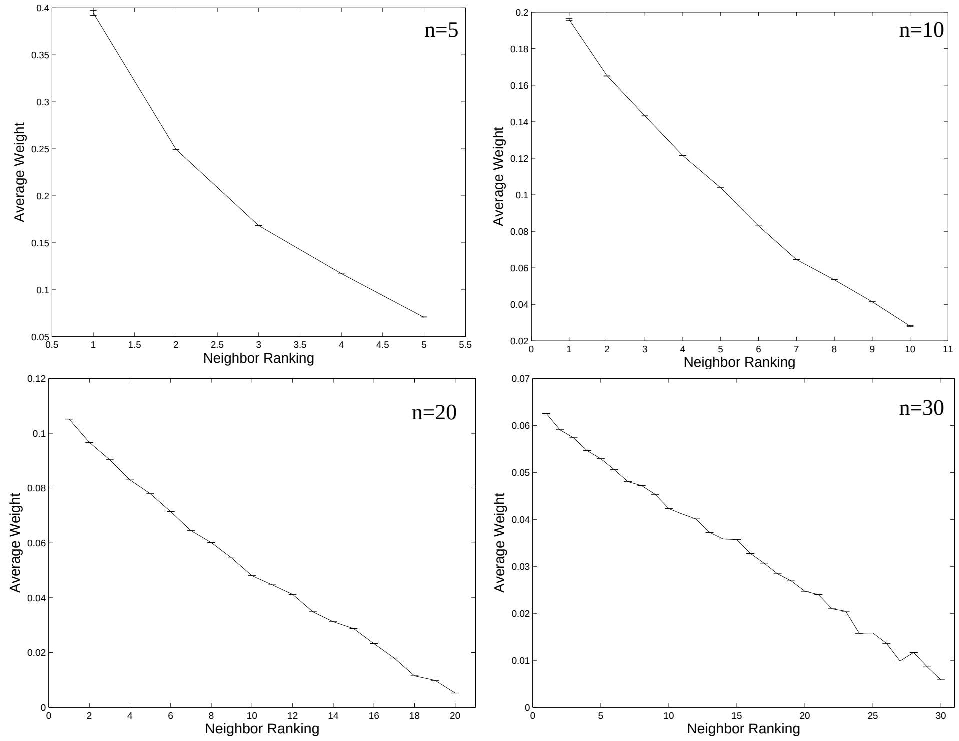

Figure 7.10: The average values of the weights computed by standard LLE are plotted versus the neighbor ranking. The graphs refer to the swissroll data set and were obtained with different values of n. The weights were computed for values of Δ between 0.001 and 0.096. The variance values are visualized as error bars but are negligible in all cases. In all plots the closest neighbors have the largest weights, providing justification for the DWLLE approach.

section is followed by the results and the conclusions.

7.2.1 The DWLLE Algorithm

The DWLLE algorithm is based on a novel approach for computing of the weights W, whose advantage over standard LLE is not to require the regularization term Δ when $n > D$. As already stated before, in LLE each data point $\mathbf{x}_i$ is approximated by the weighted sum of its n nearest neighbors. The weights are computed in eq. (5.12) such that the error function $\Psi(W)$ is minimized. The DWLLE algorithm assigns the largest weights to neighbors that are closest to $\mathbf{x}_i$ (Fig. 7.11). In this way each $\mathbf{x}_i$ is mainly approximated by the closest among the n neighbors. The weight W_{ij} between two

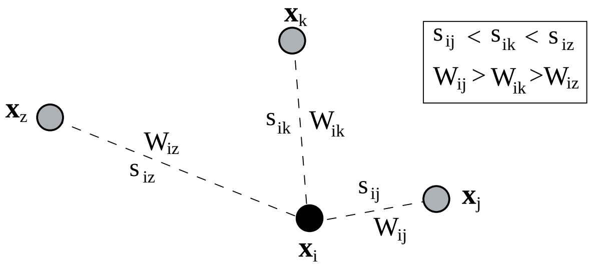

Figure 7.11: The DWLLE algorithm consists in assigning the largest weights to the closest neighbors. In this figure, a data point $\mathbf{x}_i$ having three neighbors $\mathbf{x}_j$, $\mathbf{x}_k$ and $\mathbf{x}_z$ is shown. The distance between $\mathbf{x}_i$ and its neighbors are denoted by s_{ij}, s_{ik} and s_{iz}, respectively. $\mathbf{x}_j$ ($\mathbf{x}_z$) is the closest (farthest) neighbor and, in turn, the respective weight W_{ij} (W_{iz}) will be the largest (smallest) one.

neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ can then be computed as follows:

$$W_{ij} = 1/\sqrt{|\mathbf{x}_i - \mathbf{x}_j|}. \tag{7.1}$$

Here, the value of W_{ij} is taken inversely proportional to the root of the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. In this way the closest points will have the largest weights.

The function of eq. 7.1 does not represent the unique possibility. In principle, any other function showing a monotonic decrease for the weights with increasing distance can be adopted.

7.2.2 Data and Experiments

The DWLLE is compared with the standard LLE by considering two three-dimensional synthetic data sets ($D = 3$), a swissroll and a S-manifold (both with $v = 1000$) that are visualized in Fig. 7.12, and a real-world data set, the 12-cases DCE-MRI data set ($D = 6$, $v = 2449$). From section 6.1.1 it is known that LLE computes accurate embeddings of the 12-cases DCE-MRI data sets only for $n > 20$, that means the regularization term Δ is required as $n > D = 6$ (recall eq.(5.17)).

The three data sets are projected onto a two-dimensional space by DWLLE and standard LLE. The quality of the obtained projections is evaluated by the neighborhood preservation NP (see section 5.4.3). In particular, each value of NP is the average of 16 values computed for t between 5 and 20.

7.2.3 Results

The S-manifold and swissroll data sets and their two-dimensional projections obtained by standard LLE and DWLLE are shown in Fig. 7.12. The input parameters (Δ and n

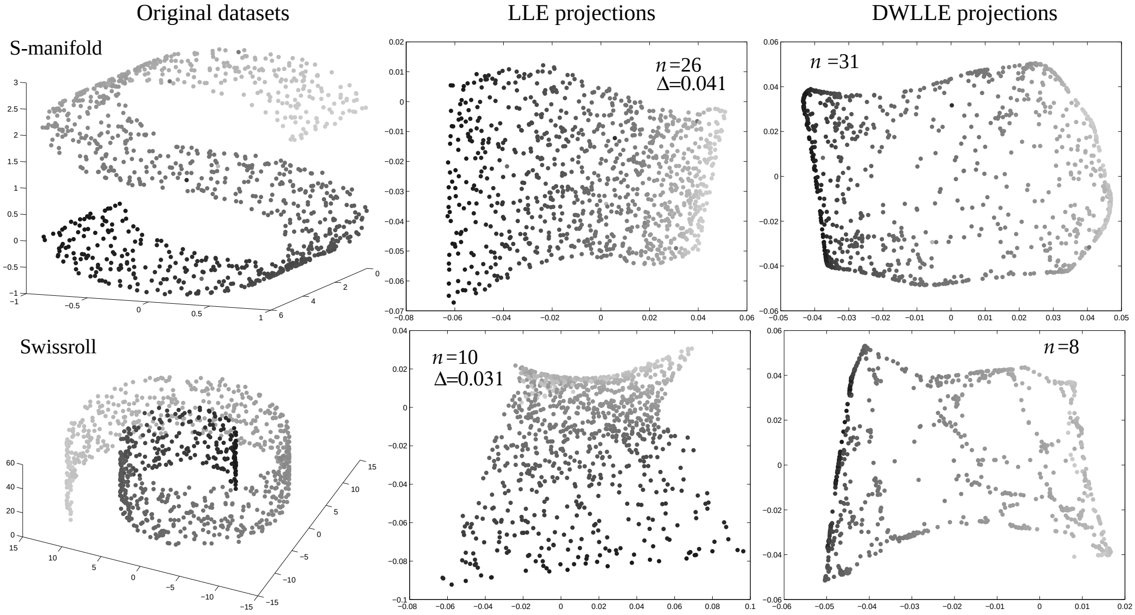

Figure 7.12: The data sets used in this study are shown on the left, while their two-dimensional projections obtained by LLE and DWLLE are shown in the middle and on the right, respectively. Considerably, for both data sets the DWLLE algorithm yields results that are qualitatively comparable with the LLE ones. In particular, the order of the colors in the original data sets is maintained in the projections, and this gives evidence that the topology of the data is preserved.

for LLE, n for DWLLE) are tuned by maximizing NP. All the projections exhibit a good preservation of the data topology, as the order of the colors is maintained. The DWLLE applied on the synthetic data sets thus provides results that are qualitatively comparable with those obtained by standard LLE.

Several two-dimensional projections of the DCE-MRI data set obtained by DWLLE with different values of n are shown in Fig. 7.13. In all the projections, the benign and malignant clusters largely overlap and a cluster separation can hardly be observed. Compared to the dimensional reductions obtained by LLE (see Fig. 6.2), the one obtained by DWLLE on the DCE-MRI data are inferior in terms of cluster separation. This is probably due to the higher dimensionality of the DCE-MRI data set. In particular, the DWLLE works sufficiently well on the two synthetic data sets, that are three-dimensional, and the weights computed by DWLLE approximate the ones obtained by LLE quite accurately. This approximation becomes worse when the input data dimension increases as in the case of the DCE-MRI data that is six-dimensional.

The next step of the analysis consists in investigating how the number of neighbors

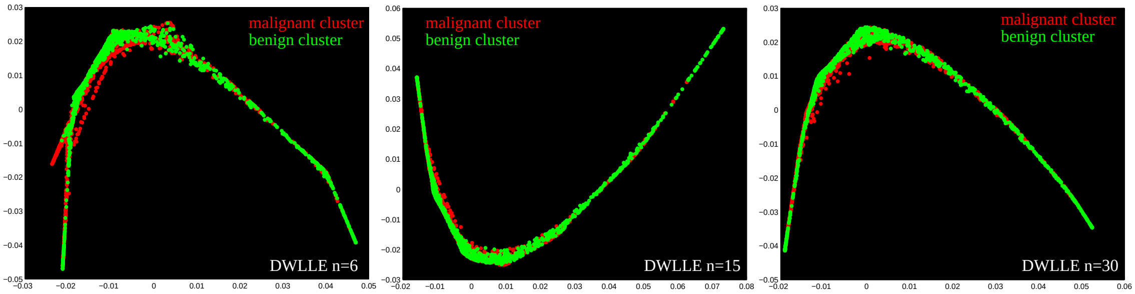

Figure 7.13: Three two-dimensional projections of the DCE-MRI data set obtained by DWLLE. In all three cases the benign and malignant clusters largely overlap. The projections computed by standard LLE are clearly superior (see for instance Fig. 6.5).

n does affect the performances of DWLLE as compared to the LLE ones. The values of NP are computed for the projections computed by LLE (with n between 5 and 40 and Δ between 0.001 and 0.096 (21 values of Δ)) and DWLLE (n between 5 and 40).

The plots relative to the swissroll data set are shown in Fig. 7.14, where the error bars represent the variance of NP due to Δ. The performances of the DWLLE are comparable with the LLE ones for $n \leq 24$. With $n > 24$ the error induced by the DWLLE, which is due to the computation of suboptimal linear weights, is likely to be higher than the LLE one and, in turn, NP decreases. The value of the variance is very low, thus Δ does not influence considerably the two-dimensional projection obtained by LLE.

The plots relative to the S-manifold are shown in Fig. 7.15. The LLE plot shows that the LLE output changes strongly with n and Δ varying. In particular, the variance relative to Δ is not negligible. By contrast, the curve relative to DWLLE is very stable with respect to n and is comparable with the best values of the LLE plot.

The curves concerning the DCE-MRI data are shown in Fig. 7.16. The performances of DWLLE are comparable with the LLE ones for $n < 13$. The LLE curve is characterized by larger values for $n \geq 13$ and, at the same time, the variance induced by Δ is very small. The plotted curves do not differ significantly and this may suggest that the visualization of the two-dimensional projections obtained by LLE and DWLLE should be not too different. In reality, the projections appear very different, as shown in Fig. 6.2 (LLE projections) and Fig. 7.13 (DWLLE projections). In particular, the DWLLE projections are far less informative with regard to the comparison between benign and malignant time-series.

7.2.4 Discussion and Conclusions

The DWLLE algorithm is based on a modification of the LLE algorithm. Both algorithms comprise three steps and the first and third steps are the same. By contrast, the second

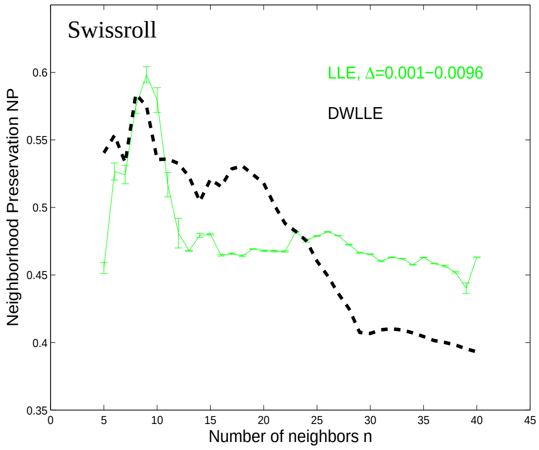

Figure 7.14: Plots of the neighborhood preservation versus n between 5 and 40 computed for the swissroll data set. The error bars in the LLE curve represent the variance of NP induced by Δ varying between 0.001 and 0.096. The variance in the LLE curve induced by Δ is negligible.

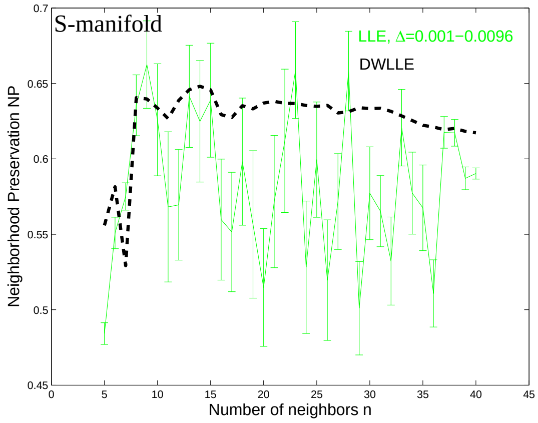

Figure 7.15: Plots of the neighborhood preservation versus n between 5 and 40 relative to the S-manifold. The LLE curve is very instable and this means the 2D projection of the S-manifold depends considerably on the values of n and Δ. By contrast, the DWLLE curve is quite stable with respect to n.

step is different. Specifically, in DWLLE the weight between two neighboring points depends only on their distance. The value of the weight is inversely proportional to the distance between the points. This approach has the advantage over the LLE one that the parameter Δ is not required.

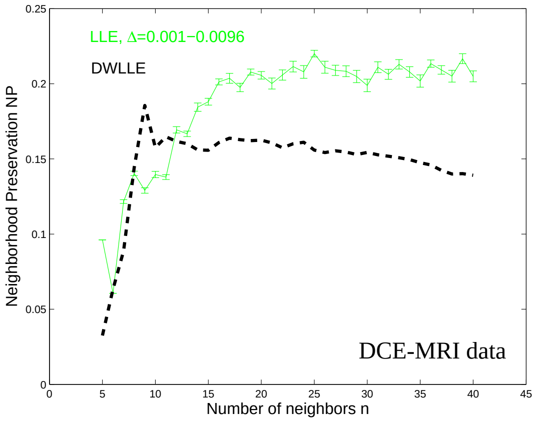

Figure 7.16: Plots of the neighborhood preservation over n between 5 and 40 for the DCE-MRI data set. The variance of the LLE curve induced by Δ varying is very small. Overall, the values of the DWLLE curve are lower than the LLE ones.

The DWLLE algorithm provides similar results to the LLE ones when applied to two synthetic data sets, a swillroll and a S-manifold. While the dimensional reduction of the swissroll obtained by LLE is not much influenced by Δ, Δ influences considerably the projection of the S-manifold. Thus, the employment of DWLLE on the S-manifold is particularly sensible.

DWLLE exhibits poor performances when applied to the DCE-MRI data set. Indeed, the visualization of the projected spaces results in highly overlapping benign and malignant clusters and, at the same time, the values of neighborhood preservation relative to DWLLE are overall lower than the ones relative to the LLE projections.

7.3 Conclusions of this chapter

The LLE algorithm presents desirable properties such as nonlinear data reduction and uniqueness of the computed solution, but there are cases in which its application may lead to misleading results. One possible source of problems is the usage of the Euclidean distance for the neighbor search. This metric is the standard choice for this purpose. In section 7.1 it is shown that the Euclidean distance can produce short circuits, i. e. faraway points may appear deceptively close. Short-circuits can cause LLE to compute wrong dimensional reductions. To address this problem, in section 7.1 the usage of the geodesic distance (ISOLLE) as alternative metric is proposed, and its efficacy is demonstrated on two data sets, a swissroll and the 12-cases DCE-MRI data set.

Another possible source of problems in LLE may be to set not properly the value of

Δ. This parameter is required only when $n > D$ in order to guarantee the unicity of the LLE solution. In section 7.2 it is introduced one example of data set, namely the S-manifold, whose dimensional reduction is strongly dependent on the value of Δ. The DWLLE algorithm is introduced as an alternative to LLE and has the advantage of not requiring Δ for the computation of the dimensional reduction. The performances of the LLE and DWLLE algorithms are compared taking into account a swissroll, an S-manifold and the 12-cases DCE-MRI data set and the DWLLE algorithm works particularly well on the S-manifold data set.

8 Analysis of Microarray Data in Breast Cancer

This chapter illustrates the analysis conducted by PCA (section 5.2.1) and LLE (section 5.2.2) of the breast cancer microarray data described in section 3.3. The analysis is divided into two parts. The first part is described in section 8.3 and concerns the application of PCA and LLE. In particular, LLE is employed with three different metrics, the dot-product, the Euclidean metric and the geodesic distance. The second part of the analysis consists in investigating the usage of the LLE algorithm with the Minkowski metric (see eq. (5.35)). This part is described in section 8.4.

8.1 Target of the Analysis

The visualization of microarray data is typically based on hierarchical clustering algorithms (see section 4.3.3). The purpose of the current analysis is to visually explore the microarray data using algorithms for dimensional reduction.

The microarray data sets analyzed in this work represent a two-class problem (patients with and without metastases that are also seen as malignant and benign cases, respectively) and the essential information is how different these classes are expressed by the data. In particular, the knowledge of this information might shed light on how determinant the considered genes are to the development of metastases.

In order to explore the degree of similarity between the classes, the microarray data sets are projected onto a two-dimensional space and visualized with customized colors encoding the class labels (benign and malignant). The gene series of each patient is represented as a point in the display, and the closeness of any pair of points reflects the similarity of the corresponding gene-series.

8.2 Experiments

The set of D genes of each patient is treated as a D-dimensional data point. It follows that the three data sets are composed of 78 data points whose dimension is 70, 231 and 5223, respectively.

The data sets are projected onto a two-dimensional space by PCA and LLE. The LLE has already been applied for microarray data analysis in (Chao and Lihui, 2004), although here the main purpose was automatic classification and not visual data exploration. In LLE it is important to set the number of neighbors n properly. The values of n that are investigated in this study vary between 5 and 60. The metric with respect to which the neighbors are found also plays an important role. The Euclidean distance is probably the most common used metric. However, it has been concluded in (Aggarwal et al., 2001) that the Euclidean metric may not be an optimal choice in high-dimensional spaces as it may prove to be poorly discriminative. For this reason two other metrics, the dot-product and the geodesic distance, are investigated. The dot-product is often used as a similarity measure between different gene expressions, and the geodesic distance has been proposed as a metric for LLE in section 7.1. All the metrics are applied to both the original and the normalized data scaled to [0,1].

The low-dimensional projections are quantitatively evaluated firstly in terms of separation between the clusters of the patients with and without metastases by the Fisher criterion F (section 5.4.1), secondly by the trustworthiness (section 5.4.4). In particular, each value of trustworthiness is the average of 16 values computed for t between 5 and 20.

In section 8.4 the LLE algorithm is employed with a further metric, the Minkowski metric. This is an alternative metric for the analysis of high-dimensional spaces. This metric requires a parameter, the order p, to be set (see eq. (5.35)). With $p = 2$ the Minkowski metric reduces to the Euclidean metrics. The Minkowski metric is applied with both $p \leq 2$ and $p \geq 2$ in order to investigate how sensitive the neighbor search (and consequently the LLE output) is to the value of p. The quality of the corresponding dimensional reductions are evaluated by the Fisher criterion F.

8.3 Use of LLE with Different Metrics

The values of F and trustworthiness related to the projections obtained by LLE of the three data sets are plotted versus n for different metrics in Fig. 8.1. One can observe that, in general, the values of F decrease progressively with the number of genes increasing. The largest separation between the benign and malignant classes is achieved in the two-dimensional projection of the 70-genes data set while the worst separation is observable in the projection of the 5223-genes data set. The decrease of F indicates that the points of the benign and malignant classes are harder to be separated in the projected space when the number of genes increases. This is probably correlated with the growing dimension of the data space. Indeed, the dimension of the data space increases from 70 to 5223 while the number of patients (i. e. the number of data points) does not change. As a consequence, the data space becomes more and more sparse and, in turn, the structures of the classes may appear more confused to LLE. This fact may appear more clear by recalling the so-called curse of dimensionality effect (see section 5.5.2). This refers to the fact that the number of data points required to estimate a data distribution with a certain error grows exponentially with the dimension of the data space. In the current case, the dimension of the data space increases but not the

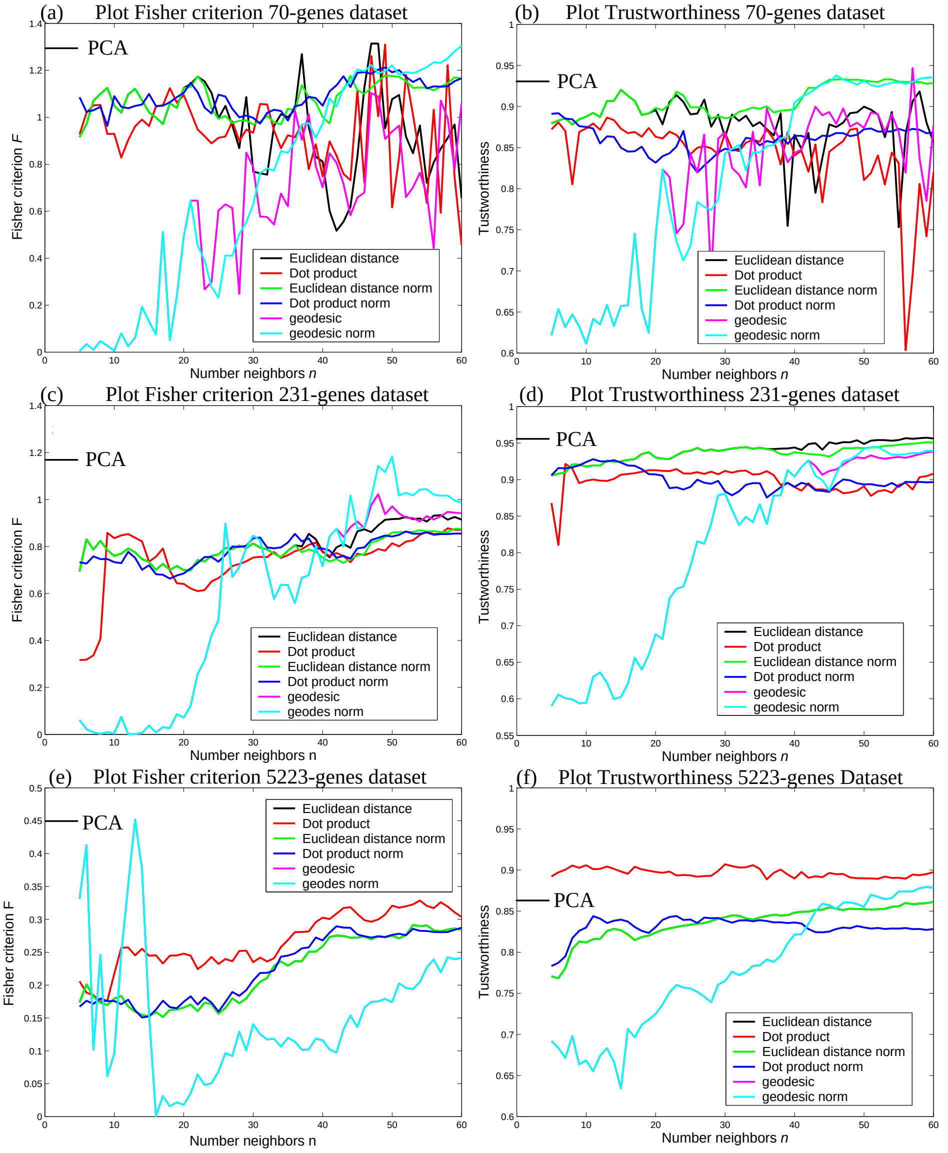

Figure 8.1: Plots of the values of the Fisher criterion F (left) and trustworthiness (right) related to the projections of the three data sets obtained by LLE with varying n and different metrics. In each plot the corresponding value computed for the projection obtained by PCA is also visualized.

number of data points and, as a consequence, one can expect that the error estimation of the data distribution increases.

Overall, the best separation between the classes (i. e. the highest values of F) is achieved for $n > 30$. The Euclidean and dot-product metrics provide similar results with

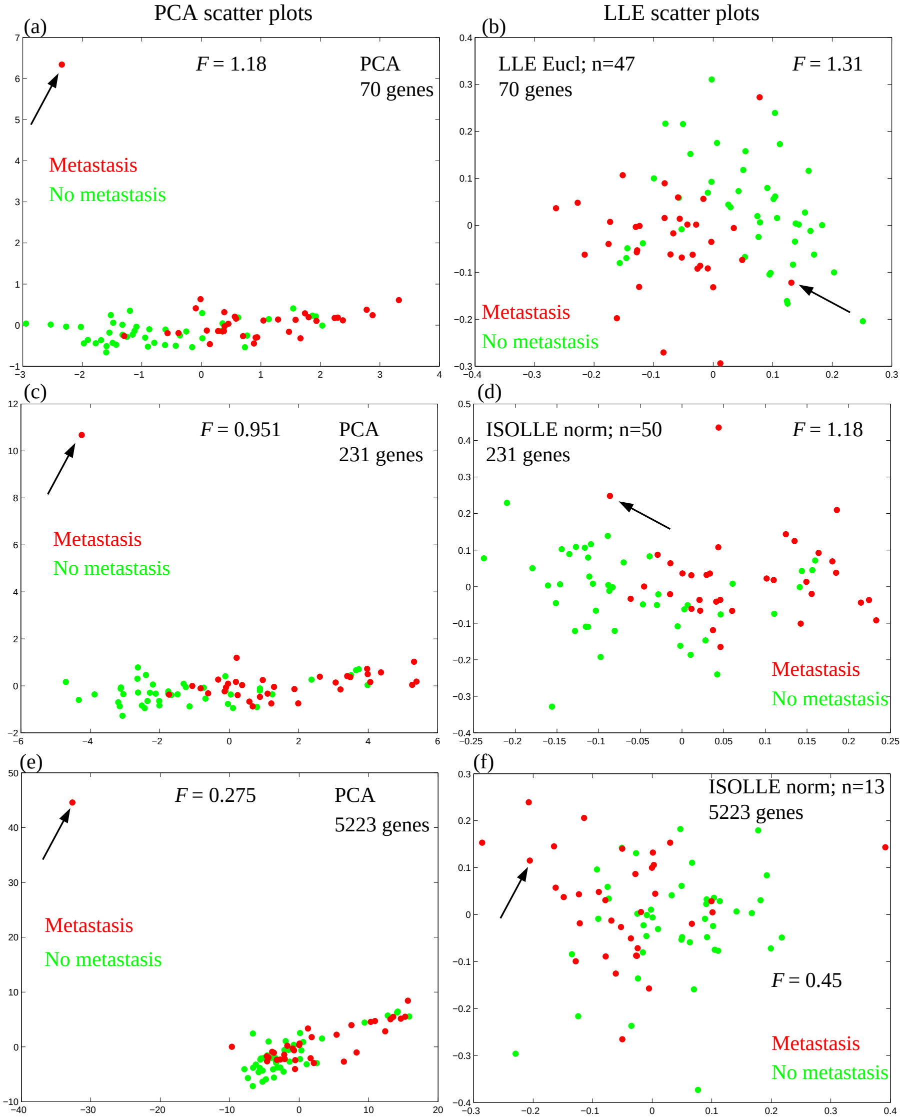

Figure 8.2: Scatter plots of six projections obtained by PCA (left) and LLE (right) of the microarray datasets. The latter projections are obtained by LLE employed with those values of n that provide the maximal values of F in Fig. 8.1.

and without data normalization and the corresponding curves are quite stable with respect to n. By contrast, the curves related to the geodesic distances are characterized by a different behavior that can be observed in all three data sets. More specifically, the values of both F and trustworthiness are lower than those related to the two other metrics for $n < 30$. This may be explained by recalling the fact that the neighbors computed according to the geodesic distance are not searched in all directions in the data space as in the case of the Euclidean distance and dot-product, but are searched along a contour that is the contour of the data set in the data space (see Fig. 7.2 for the sake of clarity). Therefore, the configuration of the neighboring graphs (i. e. the graphs obtained by linking each point to its n nearest neighbor) obtained with the geodesic distance are probably completely different from those obtained with the Euclidean distance and dot-product, and it may happen that the graphs obtained with the geodesic distance best approximate the data structure only for $n > 30$.

Overall, the curves of F and trustworthiness are in substantial agreement, as they exhibit similar behaviors and the relations between the curves of the three metrics are similar. For instance, the red curves (related to LLE with dot-product) in Fig. 8.1(e) and Fig. 8.1(f) are in both cases the predominant ones.

In Fig. 8.1 the values of the Fisher criterion F and trustworthiness related to the projections obtained by PCA are also visualized. Note that in all plots, except for the plot in Fig. 8.1(f), the values of the PCA projections are comparable with the best values obtained by LLE.

Note that in Fig. 8.1(e) and Fig. 8.1(f) only four curves are visible instead of six. Specifically, the curves related to the Euclidean distance (black curve) and geodesic distance (pink curve) can not be seen as they overlap with the Euclidean distance norm (green curve) and geodesic norm (sky-blue curve) respectively. Thus, in the case of the 5223-genes data set the normalization of the data has no influence on LLE performed with the Euclidean and geodesic distances.

The projections obtained by LLE that provide the highest values of F are shown in Fig. 8.2 (right), along with the projections obtained by PCA (left). The points of patients with (without) metastasis are visualized in red (green). The PCA projections reveal that the microarray data are intrinsically one-dimensional as the data variance along the second principal component (y-axis) is small as compared with the one along the largest eigenvalue (x-axis). The metastases / non metastases classes are easy to be recognized and quite well separated in the PCA projections of the 70- (Fig. 8.2(a)) and 231-genes (Fig. 8.2(c)) data sets. Yet, the classes lie more close-by in the PCA projection of the 5223-genes dataset (8.2(e)). By contrast, the benign and malignant classes are easier to be recognized in the LLE projections in all cases (Fig. 8.2(b)-(d)-(f)). It is particularly interesting the fact that the two classes are separated also in the projection of the 5223 genes dataset.

Despite the fact that these values of F are similar, the visualization of the corresponding projections results in visually different patterns. In particular, the points in the LLE projections appear more scattered than in the PCA ones. This is due to the presence of an outlier, which is indicated by an arrow, in all the three PCA projections (see Fig. 8.2). The same points are also visualized in the corresponding LLE plots and surprisingly they do not appear to be outliers, i. e. they are mapped close to the remaining points. This may be explained by the fact that LLE aims to preserve the neighboring interrelationships more than the interpoint distances and one can presume the outlier point is mapped by LLE close to its neighbors without preserving their distances.

In conclusion, the performances of PCA are statistically comparable with the best ones of LLE On the other hand, in the LLE projections the points are more scattered and this might ease the visual exploration of the data.

8.4 Use of LLE with Minkowski metric

The microarray data set is composed of a relatively low number v of patients and a large series of D genes for each patient. In this work the data sets are interpreted as a D-dimensional data space with v data points, i. e. a high-dimensional data space with relatively few points. In the analysis of microarray data it is important to know that high-dimensional spaces can have counter-intuitive geometrical properties (see section 5.5). One of these properties concerns the fact that the concept of nearest neighbor is less intuitive in high-dimensional spaces. In particular, the Euclidean distance may not be the most appropriate choice for nearest neighbor search in high-dimensional data as this metric may become less discriminative.

An alternative metric for the analysis of high-dimensional spaces is the Minkowski norm, which is defined in section 5.5.3. In the current section the Minkowski metric is used for the neighbor search in LLE. The value of the order p is set to values both lower and higher than two in order to investigate to which extent this value affects the neighbor search and whether there are some changes that can be correlated with the increase in the dimension of the data space. In these experiments the number of neighbors n is varied from 5 to 60.

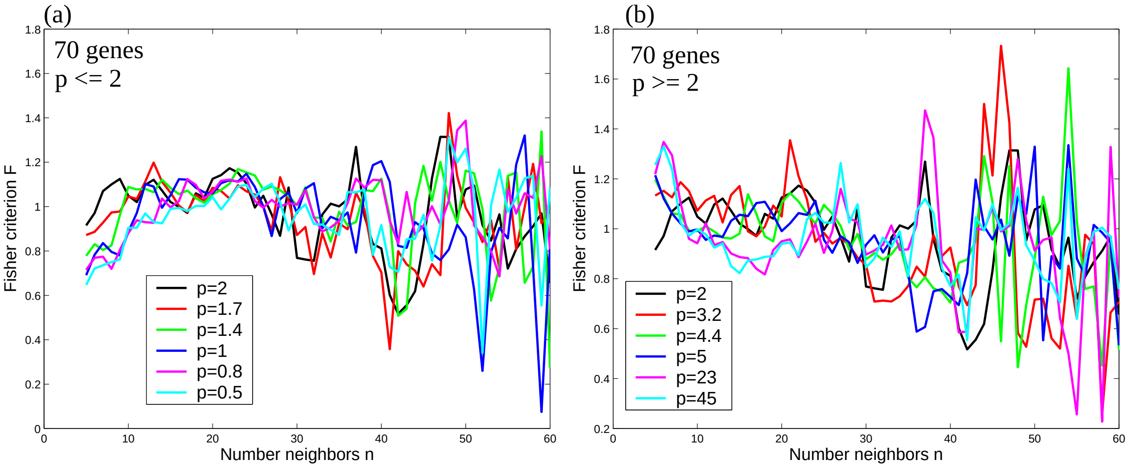

Figure 8.3: Plots of Fisher values obtained with different orders p of Minkowski metric on the 70-genes data set. The curves obtained with $p \leq 2$ are shown on the left, while those obtained with $p \geq 2$ are shown on the right. In both figures the values of p seem not to affect the curves.

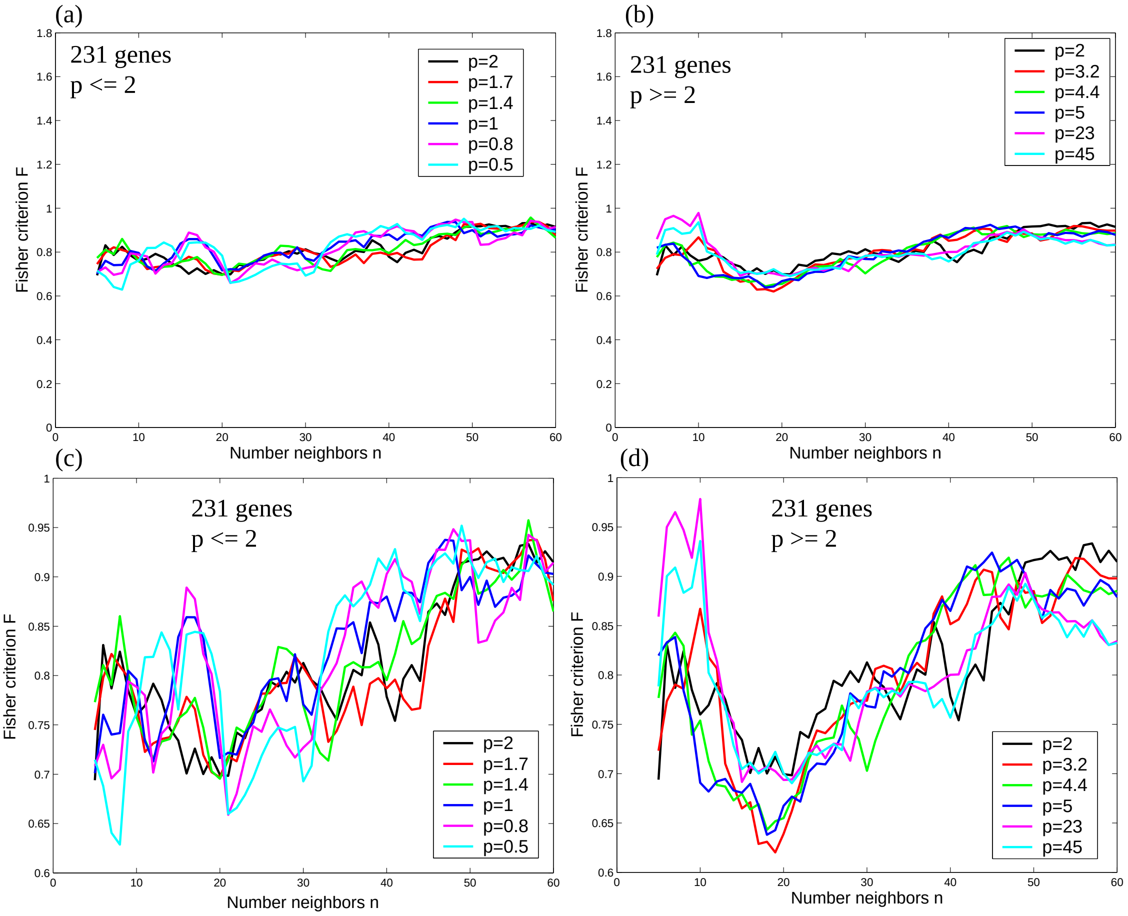

Figure 8.4: Plots of Fisher values obtained with different orders p of Minkowski metric. The microarray dataset is 231-dimensional. The plots in (c) and (d) are the enlargements of the curves plotted in (a) and (b) respectively.

Several curves of Fisher values related to the 70-dimensional dataset are visualized in Fig. 8.3 that refer to the curves obtained with $p \leq 2$ (left) and $p \geq 2$ (right) respectively.

In both graphs the curves are quite changeable with respect to n. A dependence on the value of p is not observable and a predominant curve is not present. It follows that the value of p does not seem to affect the final dimensional reduction of the 70-dimensional data set. It is interesting to note that the largest values of F are obtained with $n > 40$ and $p \geq 2$.

The curves of the Fisher criterion F related to the 231-dimensional dataset are shown in Fig. 8.4. Fig. 8.4 (a) and (b) contain the curves related to $p \leq 2$ and $p \geq 2$ respectively plotted with the same scale of Fig. 8.3. Fig. 8.4 (c) and (d) represent the enlargements of the two previous figures and allow to observe the curves more in detail. From the

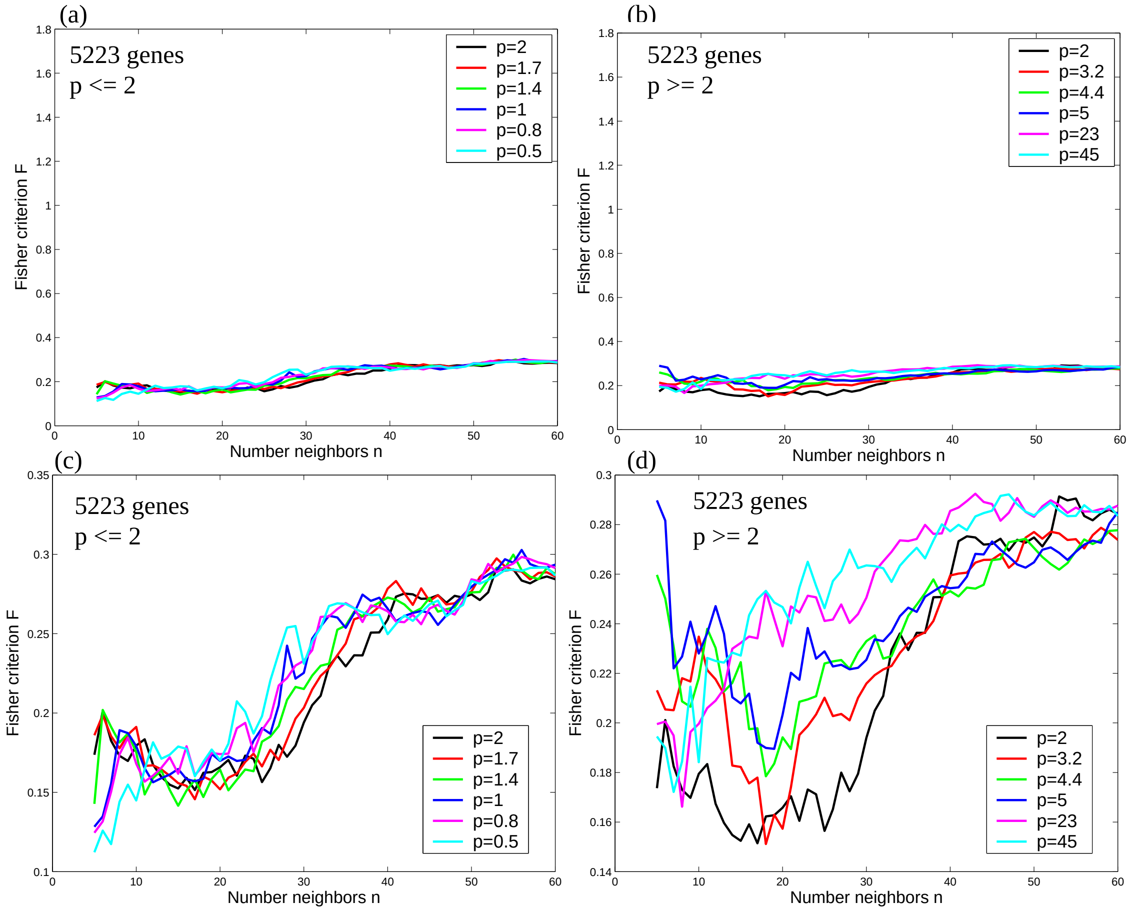

Figure 8.5: Plots of Fisher values obtained for different orders ($p \leq 2$ left and $p \geq 2$ right) of Minkowski metric. The microarray dataset is 5223-dimensional. The curves in (a) and (b) are magnified in (c) and (d) respectively.

observation of Fig. 8.4 (a)-(b) one can deduce that the value of p has a low influence on the neighbor search, as all the the curves appear quite similar. It is however interesting to observe that the curves in Fig. 8.4 (d) (enlargement of the curves $p \geq 2$) tend to cross less as compared with the curves in Fig. 8.3 (b) (curves $p \geq 2$ related to the 70-genes data set). This suggests that the value of p seems to have more influence, although little, on the neighbor search of LLE applied to the 231-genes data set.

The curves related to the 5223-dimensional dataset are shown in Fig. 8.5. As above, in Fig.8.5 (a)-(b) the curves are plotted with the same scale of Fig. 8.3, and in Fig. 8.5 (c)-(d) the corresponding enlargements are visualized. Also in this case the curves plotted in Fig. 8.5 (a)-(b) appear quite similar and the changes induced by p are very small. Particularly interesting is to observe Fig. 8.5 (d), i. e. the enlargement of the curves obtained with $p \geq 2$. Indeed, all the curves exhibit larger values of F than the ones of the curve obtained with $p = 2$. Moreover, the larger p, the larger the values of F, i. e. a dependency between the curves and p can be observed. The values of p therefore seem to affect more the final projection of the 5223-dimensional dataset as compared to the 70- and 231-dimensional datasets. The values of p might gain influence as the dimension of the dataset increases, in particular the values $p \geq 2$. This is somewhat surprising as in (Hinneburg et al., 2000) it is stated the nearest neighbor search with the Minkowski metric when $p > 2$ are meaningless in high-dimensional spaces. In these experiments the values $p > 2$ influence more the final dimensional reduction and this becomes more evident as the data dimension increases.

8.5 Discussion

The PCA and LLE algorithms are employed for the visual exploration of microarray data. The PCA algorithm provides results that are statistically comparable this the best obtained by LLE. On the other hand, the visualization of the PCA and LLE projections results in different patterns. In the LLE projections the points are quite uniformly scattered and the benign and malignant clusters are clearly visible for all data sets. In the PCA projections the clusters appears elongated along the x-axis also because of the presence of an outlier which hampers the visual analysis of the data sets. In addition, the benign and malignant clusters are not well separated in the PCA projection of the 5223-genes data set.

The visualization of microarray data using algorithms for dimensional data reduction can prove useful for the comparison of gene series from different patients. This information can also be extracted by hierarchical clustering algorithms, but they do not provide indication of the degree of similarity between single gene sequences. On the contrary, this information is contained in the PCA and LLE projections as distance between the points.

Furthermore, the visualization of the clusters with customized colors encoding the class labels can reveal how different the clusters are with respect to a particular set of genes, This may contribute to determine the role of a set of gene in the development of metastasis.

9 Summary, Conclusions and Outlook

The aim of this PhD thesis was to investigate the potential of algorithms for dimensional reduction for the visual exploration of multivariate data in breast cancer. In multivariate data, each entity is characterized by a feature vector comprising many attributes, where "many" is typically any number greater than three. In this work, two types of multivariate data were explored, namely **dynamic contrast enhanced magnetic resonance imaging** (DCE-MRI) data and **microarray** data.

In the DCE-MRI data, an entity is represented by a time-series connected with a certain voxel and the number of attributes is given by the number of MR volumes recorded. In the microarray data, an entity is given by a patient that is characterized by the expression values of a series of genes. Both the experimental data sets were acquired from female patients with benign and malignant breast lesions. Detecting the interrelationships between the many variables characterizing the benign and malignant patients is of vital interest to physicians but it is also a challenging task because of the multi-dimensional nature of the data.

In this work, methods for **dimensional reduction** were used for **visual exploration** of the data sets with regard to the investigation of the degree of similarity of data related to benign and malignant tumors. In this context, each entity contained in the data sets was treated as a data point in a multidimensional data space, whose dimension equals the number of attributes characterizing the entity itself. Entities with similar attributes are neighboring data points in this data space. Algorithms for dimensional data reduction can help to reveal these relationships by projecting the data space onto a two- or three-dimensional space while best preserving the information contained. In turn, the visualization of the projected space can provide insight into the similarity among entities.

9.1 Summary

An algorithm particularly suited for visualizing similarities between entities in high-dimensional data is **locally linear embedding (LLE)**, as it aims to compute low-dimensional projections of high-dimensional data while best preserving the local data geometry. Specifically, neighboring data points in the high-dimensional data space are projected by LLE onto neighboring data points in a lower-dimensional space, i. e. the interrelationships between neighboring data points in the high-dimensional data space are maintained in the lower-dimensional space. Because of this appealing property, the

LLE algorithm was the method for dimensional data reduction most used in this work.

The LLE algorithm was employed to explore the similarity among DCE-MRI time-series of benign and malignant tumors. The results obtained by LLE were compared with those obtained by *self-organizing maps* (SOM) and *principal component analysis* (PCA), that are two of the most widely utilized techniques for dimensional reduction. The LLE algorithm demonstrated superior capability, as compared to PCA, to extract essential information from the DCE-MRI data set. On the other hand, PCA does not require any input parameter, while LLE has one input parameter (in some cases two input parameters) that must be tuned. According to the statistical analysis, the SOM algorithm outperformed LLE in terms of separation between benign and malignant time-series. However, the SOM has several parameters that must be set and this can be extremely time-consuming for the human user and can cause the results to vary strongly.

Since LLE provided good results on the DCE-MRI data, the LLE algorithm was investigated more in detail. In particular, two modifications of the LLE algorithm were developed to overcome two shortcomings that can limit its performance. The first modification, termed as ISOLLE, consisted in using LLE with the geodesic distance instead of the Euclidean metric. It was shown that this allows for a better preservation of the data structure in the dimensionally reduced space. This was validated on both synthetic data and the DCE-MRI data set. The second modification, termed as DWLLE, introduced an innovative approach to the weight computation that allows the elimination of one input parameter of LLE, the regularization term Δ. The results obtained by DWLLE on synthetic data were very promising and consistent with those obtained by standard LLE.

The analysis of the microarray data set in terms of discrimination between gene expression relative to benign and malignant breast tumors was conducted using PCA and LLE. LLE was performed with several metrics, including the Minkowski metric with different orders as the order to the Minkowski metric can play a major role when the dimension of the data space is considerably high, as in the case of the microarray data.

In all the experiments, the LLE algorithm was applied for different values of the input parameters and the quality of the respective projections was evaluated accurately using more than one quantity.

9.2 Conclusions

The LLE algorithm has shown considerable potential for the visual exploration of multivariate data in the biomedical domain. The dimensional reduction of the data and its visualization allows the user to explore the similarity among different entities characterized by possibly numerous attributes. Revealing pattern in the data provides information on the interrelationships between these attributes, thereby allowing the human user to gain understanding from the data.

Depending on the task at hand, the visualization of the dimensional-reduced data can be integrated with supplemental information, typically labels, characterizing the entities. For instance this information can be encoded in terms of color or shape patterns.

In chapter 6, the dimensional reduction of the DCE-MRI data set was visualized using three different color schemes. The encoding of the class labels (benign / malignant) on the display as color pattern provides information on how good the adopted DCE-MRI protocol can help distinguish between benign and malignant lesions, and this can represent a feedback for technicians and physicians.

The second color scheme consisted in visualizing the entire experimental DCE-MRI data set while highlighting the points from a single particular tumor lesion. In this way the signal dynamic related to this tumor lesion can be compared with the time-series from several different lesions simultaneously, thereby allowing for a multi-case comparison. The comparison of different lesions can provide precious information to physicians. This task, however, might prove difficult if the comparison is performed directly by taking into account the MR volumes because of the multi-dimensional nature of the data. It is thus beneficial to physicians to represent the dynamic of many different time-series into a single display.

Concerning the third color scheme, the histological labels of the tumors were encoded as color information in order to permit physicians to compare the signal dynamic of different histologic families of breast tumors.

Besides proving useful for the visual exploration of the degree of similarity of time-series from different tumor lesions, LLE has shown considerable potential for visualizing the breast lesions with customized colors reflecting the temporal course of the time-series. The colors of the tumor visualizations relative to LLE were consistent with those obtained by the model-based *three-time-point method* (3TP), with the advantage that in LLE, a model of the contrast agent dynamic is not required and, in turn, the LLE approach may in principle also be applied to data concerning tumors of other organs, such as liver, for which it is very difficult to model the dynamic of the processes involved. Algorithms for dimensional reduction may be an alternative to model-based techniques, such as the 3TP, when the number of different clinical cases in the experimental data set is sufficiently high, such that the algorithms can learn meaningful behaviors from the data (e. g. typical benign and malignant time-series), thereby compensating for the lack of a model.

9.3 Outlook

There are two problems that afflict LLE, which restricts its field of application. The first problem is that the algorithm can not handle data sets with more than a certain number (in the order of 10^4) of data points because of the excessive computer memory required. This problem is mentioned also in section 6.5, where a possible solution is discussed.

The second problem that afflicts LLE, but more in general all the algorithms for dimensional data reduction, may occur when the data dimension is very high and the number of samples relatively low. Provided these conditions, the data space is very sparse and it may be difficult for LLE to detect a global structure in the data. To face this problem, in the framework of this PhD it was proposed to pre-process the data by the *discrete wavelet transform* (DWT) (Daubechies, 1991) in order to reduce the dimension of the data space. The DWT enables the assessment of localized and scale-dependent information in signals by a set of features (Mallat and Zhong, 1992). DWT computes for each data point a set of descriptive features whose number is much lower than the data dimension. In turn, the DWT feature space is more dense and applying LLE to this space may improve the detection of a global structure. This approach was successfully applied to explore a data set of brain tumor images (Varini et al., 2006). This might serve as an example for the use of hybrid methods including LLE.

In this work only the first two coordinates discovered by LLE were visualized, i. e. these coordinates were derived from the two smallest eigenvector of matrix S (eq. (5.19)). In (Lawrence and Roweis, 2003) it is stated that clusters that overlap in the first two dimensions are typically separated in others. It would be interesting to investigate if this is also valid for the case of the DCE-MRI and microarray data sets analyzed in this work.

Bibliography

Abdolmaleki, P., Buadu, L. D., and Naderimansh, H. (2001). Feature Extraction and Classification of Breast Cancer on Dynamic Resonance Imaging using Artificial Neural Network. *Cancer Letters*, 171:183–191.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *ICDT Conference*, London, England.

Arbach, L., Stolpen, A., and Reinhardt, J. M. (2004). Classification of breast MRI lesions using a backpropagation neural network (BNN). In *2004 International Symposium on Biomedical Imaging*, pages 253–256.

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15:1373–1396.

Bellmann, R. (1961). *Adaptive Control Processes: A guided Tour*. Princeton University Press.

Beyer, K. S., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "Nearest Neighbor" meaningful? In *In C. Beeri, P. Buneman eds.: Database Theory ICDT '99, 7th International Conference. Volume 1540 of Lecture Notes in Computer Science, Springer*, pages 217–235, Jerusalem, Israel.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York.

Brand, M. (2002). Charting a Manifold. In *Neural Information Proceeding Systems: Natural and Synthetic*, Vancouver, Canada.

Brown, J., Buckley, D., Coulthard, A., Dixon, A. K., Dixon, J. M., Easton, D. F., Eeles, R. A., Evans, D. G., Gilbert, F. G., Graves, H., Hayes, C., Jenkins, J., Jones, A. P., Keevil, S. F., Leach, M. O., Liney, G. P., Moss, S. M., Padhani, A. R., Parker, G. J., Poiton, L. J., Ponder, B. A., Redpath, T. W., Sloane, J. P., Turnbull, L. W., Walker, L. G., and Warren, R. M. (2000). Magnetic Resonance Imaging Screening in Women at genetic Risk of Breast Cancer: Imaging and Analysis Protocol for the UK Multicenter Study. *Magnetic Resonance Imaging*, 18:765–776.

Brown, M. A. and Semelka, R. C. (1999). *MRI: Basic Principles and Applications*. Wiley Liss Publication.

Chao, S. and Lihui, C. (2004). Feature Dimension Reduction for Microarray Data Analysis using Locally Linear Embedding. In *3rd Asia-Pacific Bioinformatic Conference*.

Cherkassky, V. and Mulier, F. (1998). *Learning from Data*. Wiley and Sons.

Comon, P., Voz, J. L., and Verleysen, M. (1994). Estimation of Performance Bounds in Supervised Classification. In *European Symposium on Artificial Neural Networks*, pages 37–42, Brussel, Belgium.

Daubechies, I. (1991). *Ten Lectures on Wavelets*. CBMS-NFS Series Appl. Math., SIAM.

de Silva, V. and Tenenbaum, J. B. (2002). Global versus local Methods in nonlinear Dimensionality Reduction. *Advances in Neural Information Processing Systems*, 15.

Degani, H., Gusis, V., Weinstein, D., Field, S., and Strano, S. (1997). Mapping pathophysiological features of breast tumors by mri at high spatial resolution. *Nature Medicine*, 3(7):780–782.

Degenhard, A., Tanner, C., Hayes, C., Hawkes, D. J., and Leach, M. O. (2002). Comparison between Radiological and Artificial Neural Network Diagnosis in Clinical Screening. *Phys. Meas: Special Issue: Simulation and Modelling applied to Medicine*, 23:727–739.

Demartines, P. (1994). *Analyse de donnée par réseaux de neurones auto-organisées.* PhD thesis, Institut National Polytechnique de Grenoble (France).

Demartines, P. and Herault, J. (1997). A self-organizing Neural Network for nonlinear Mapping of Data Sets. *IEEE Trans. Neural Network*, 8(1):148–154.

Dijkstra, E. W. (1959). A Note on two Problems in Connection with Graphs. *Numer. Math*, 1:269–271.

Donoho, D. L. and Grimes, C. (2003). Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-Dimensional Data. *Proceedings of the National Academy of Sciences* , 100(10):5591–5596.

Eisen, M. B. and Brown, P. O. (1999). DNA Arrays for Analysis of Gene Expression. *Meth. Enzymol.*, 303:179–205.

Esserman, L., Hylton, N., Yassa, L., Barclay, J., Frankel, S., and Sickles, E. (1999). Utility of Magnetic Resonance Imaging in the Management of Breast Cancer: Evidence for Improved Preoperative Staging. *Journal of Clinical Oncology*, 17(1):110–119.

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, HP Laboratories, Palo Alto, CA.

Forgy, E. (1965). Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications. *Biometrics*, 21:768–780.

Francois, D., Wertz, V., and Verleysen, M. (2005). Noneuclidean Metrics for Similarity Search in Noisy Datasets. In *ESANN 2005, European Symposium on Artificial Neural Networks*, Bruges, Belgium.

Furman-Haran, E., Grobgeld, D., Kelcz, F., and Degani, H. (2001). Critical Role of Spatial Resolution in Dynamic Contrast-Enhanced Breast MRI. *Journal of Magnetic Resonance Imaging*, 13:862–867.

Gilhuijs, K. G. A., Giger, M. L., and Bick, U. (1998). Computerized Analysis of Breast Lesions in Three Dimensions using Dynamic Magnetic-Resonance Imaging. *Medical Physics*, 25(9):1647–1654.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., and Loh, M. L. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537.

Graf, A. B. A. and Wichmann, F. A. (2002). Gender Classification of Human Faces. In Verlag, S., editor, *proc. Biologically Motivated Computer Vision*, pages 491–501.

Gundogdu, O. and Elmi, A. (2005). www.lshtm.ac.uk/itd/grf/microarrayoverview.htm.

Hadid, A. and Pietikäinen, M. (2003). Efficient Locally Linear Embeddings of Imperfect Manifolds. In *MLDM*.

Hanley, J. A. and McNeil, B. J. (1982). The Meaning and Usee of the Area under a Receiving Operating Characteristic (ROC) curve. *Radiology*, 143:29–36.

Heywang-Köbrunner, S. H. and Beck, R. (1996). *Contrast-Enhanced MRI of the Breast*. Springer, Berlin.

Hinneburg, A., Aggarwal, C., and Keim, D. (2000). What is the nearest Neighbor in High-Dimensional Spaces. In *Proc. of the VLDB Conference*, pages 506–515.

Hinton, G. E. and Sejnowski, T. J. (1999). *Unsupervised Learning and Map Formation: Foundations of Neural Computation*. MIT Press, MA.

Hjaltason, G. R. and Samet, H. (2003). Properties of Embeddings Methods for Similarity Searching in Metric Spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):530–549.

Iacconi, C., Cilotti, A., Marini, C., Moretti, M., Mazzola, D., Odoguardi, F., Cardillo, F. A., and Starita, A. (2005). Maximum Intensity Projection in Contrast-Enhanced Magnetic Resonance of the Breast: Current Applications and Prospectives. In *proc. of the European Conference on Emergent Aspects in Clinical Data Analysis (EACDA)*, Pisa, Italy.

Jollife, I. T. (1986). *Principal Component Analysis*. Springer Verlag, New York.

Kambhampati, D., editor (2004). *Protein Microarray Technology*. Wiley-VCH Verlag.

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., and Castrén, E. (2003). Trustworthiness and Metrics in Visualizing Similarity of Gene Expression. *BMC Bioinformatics*, 4(48).

Keim, D. A. (1992). Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 7(1):100–107.

Kelcz, F., Furman-Haran, E., Grobgeld, D., and Degani, H. (2002). Clinical Testing of High-Spatial-Resolution Parametric Contrast Enhanced MR Imaging of the Breast. *AJR*, 179:1485–1492.

Kiviluoto, K. (1996). Topology Preservation in Self-Organizing Maps. In *Proc. of Int. Joint Conf. on Neural Networks*, Piscataway, New Jersey, USA.

Knowles, A., Gibbs, P., and Turnbull, L. (2000). Improved Classification of Breast DCE-MRI using Neural Network Ensemble. In *Proc. Int. Soc. Mag. Reson. Med. (ISMRM)*, page 2163.

Kohonen, T. (2000). *Self-Organizing Maps*. Springer Verlag, Berlin.

Kriege, M., Brekelmans, C. T. M., Boetes, C., Rutgers, E. J. T., Oosterwijk, J. C., Tollenaar, R. A. E. M., Manoliu, R. A., Holland, R., de Koning H. J., and Klijn, J. G. M. (2001). MRI screening for breast cancer in women with familiar or genetic predisposition: design of the Dutch National Study (MRISC). *Familial Cancer*, 1:163–168.

Kruskal, J. (1964). Multidimensional Scaling by optimizing Goodness-of-Fit to a non-metric Hypothesis. *Psichometrika*, 29:1–27.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage Publications, Beverly Hills.

Lawrence, J. B. and Roweis, S. T. (2003). Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. *Journal of Machine Learning Research*, 4:119–155.

Lee, J., Lendasse, A., Donckers, N., and Verleysen, M. (2000). A robust nonlinear Projection Method. In *Proceedings of ESANN 2000*, pages 13–20, Bruges, Belgium.

Li, J. X. (2004). Visualization of High-dimensional Data with Relational Perspective Map. *Information Visualization*, 3:49–59.

Liang, Y., Diehn, M., Aldape, K. D., Nicholas, M. K., Bollen, A. W., Lamborn, K. R., Berger, M. S., Botstein, D., Brown, P. O., and Israel, M. A. (2005). http://microarray-pubs.stanford.edu/gbm/images/Figure1.html.

Lucht, R., Delorme, S., and Brix, G. (2001a). An artificial Neural Network for the Segmentation of dynamic MR mammography image series. In *Proc. Int. Soc. Mag. Reson. Med. (ISMRM)*, volume 9, page 839.

Lucht, R., Knopp, M., and Brix, G. (2001b). Classification of signal-times Curves from Dynamic MR Mammography by Neural Networks. *Magnetic Resonance Imaging*, 19(1):51–57.

Mallat, S. and Zhong, S. (1992). Characterization of Signals from Multiscale Edges. *Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732.

Nattkemper, T. W. and A., W. (2005). Tumour feature visualization with unsupervised learning. *Medical Image Analysis*. accepted.

Ochs, M. F. and Godwin, A. K. (2003). Microarrays in Cancer: Research and Applications. *BioTechniques*, 34:4–15.

Orel, S. G. and Schnall, M. D. (2001). MR Imaging of the Breast for the Detection, Diagnosis and Staging of Breast Cancer. *Radiology*, 220:13–30.

Pemmaraju, S. and Skiena, S. (2003). http://www.cs.sunysb.edu/∼skiena/combinatorica/-animations/dijkstra.html.

Perou, C. M., Sorlie, T., Eisen, M. B., van den Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., H., J., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular Portraits of Human Breast Tumours. *Science*, 406:747–752.

Polito, M. and Perona, P. (2001). Grouping and Dimensionality Reduction by Locally Linear Embedding. In *Neural Information Processing Systems NIPS*.

Quackenbush, J. (2001). Computational Analysis of Microarray Data. *Nature Reviews Genetics*, 18:418–427.

Rankin, S. C. (2000). MRI of the Breast. *The British Journal of Radiology*, 73:806–818.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326.

Sammon, J. W. (1969). A nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.*, C-18(5):401–409.

Schena, M., editor (2000). *Microarray Biochip Technology*. Eaton Publishing.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Similä, T. (2005). Self-organizing Map Learning Nonlinearly Embedded Manifolds. *Information Visualization*, 4:22–31.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., B., E. M., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A. L. (2001). Gene Expression Patterns of Breast Carcinomas distinguish Tumor Subclasses with Clinical Implication. In *Proc. Natl. Acad. Sci. USA98*, pages 10869–10874.

Spence, R. (2001). *Information Visualization*. Addison-Wesley, ACM Press.

Su, M. Y., Cheung, Y. C., Fruehauf, J. P., Yu, H., Nalcioglu, O., Mechetner, E., Kyshtoobayeva, A., Chen, S. C., Hsueh, S., McLaren, C. E., and Wan, Y. L. (2003). Correlation of Dynamic Contrast Enhancement MRI Parameters with Microvessel Density and VEGF for Assessment of Angiogenesis in Breast Cancer. *Journal of Magnetic Resonance Imaging*, 18:467–477.

Subramanian, K. R., Brockway, J. P., and Carruthers, W. B. (2004). Interactive detection and visualization of breast lesions from dynamic contrast enhanced mri volumes. *Computerized Medical Imaging and Graphics*, 28(8):435–444.

Tan, P. H. (2001). Pathology of Ductal Carcinoma in Situ of the Breast: a Heterogeneous Entity in Need of Greater Understanding. *Ann. Acad. Med. Singapore*, 30(6):671–676.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction . *Science*, 290:2319–2322.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2001). http://isomap.stanford.edu/datasets.html.

Tofts, P. S. and Kermode, A. G. (1991). Measurement of the Blood-Brain Barrier Permeability and Leakage Space Using DynamicMR Imaging. 1. Fundamental Concepts. *Magnetic Resonance in Medicine*, 17:357–367.

Twellmann, T. (2005). *Data-Driven Analysis of Dynamic Contrast-Enhanced Magnetic Resonance Imaging Data in Breast Cancer Diagnosis*. PhD thesis, University of Bielefeld, Germany.

Twellmann, T., Lichte, O., and Nattkemper, T. W. (2005). An Adaptive Tissue Characterization Network for Model-Free Visualization of Dynamic Contrast-Enhanced Magnetic Resonance Imaging Data. *IEEE Trans. on Medical Imaging*, 24(10):1256–1266.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Perterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene Expression Profiling predicts Clinical Outcome of Breast Cancer. *Nature*, 415(6871):530–536.

Varini, C., Lessmann, B., Degenhard, A., Hans, V., and Nattkemper, T. W. (2006). Visual Exploration of Pathology Images by a Discrete Wavelet Transform Preprocessed Locally Linear Embedding. In *proc. of the BVM Conference*, Hamburg, Germany.

Venna, J. and Kaski, S. (2005). Local Multidimensional Scaling with Controlled Tradeoff between Trustworthiness and Continuity. In *in proc. of WSOM*, pages 695–702, Paris, France.

Verbeek, J., J. (2004). *Mixture Models for Clustering and Dimension Reduction*. PhD thesis, University of Amsterdam, The Netherlands.

Verleysen, M. (2003). *Learning high-dimensional Data*. Limitations and future Trends in Neural Computatio, IOS Press, Amsterdam, The Netherlands.

Verleysen, M. and Francois, D. (2005). The Curse of Dimensionality in Data Mining and Time Series Prediction. In *IWANN'05, International Work-Conference on Artificial Neural Networks*, Barcelona, Spain.

Vesanto, J. (1999). SOM-based Data Visualization Methods. *Intelligente Data Analysis*, 3(2):111–126.

Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G., and Koudas, N. (2002). Non-Linear Dimensionality Reduction Techniques for Classification and Visualization. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wang, M., Yang, J., Xu, Z., and Chou, K. (2005). SLLE for predicting Membrane Proteins Types. *Journal of Theoretical Biology*, 232:7–15.

Weatherall, P. T., Evans, G. F., Metzger, G. J., Saborrian, M. H., and Leitch, A. M. (2001). MRI vs. histologic measurement of breast cancer following chemotherapy: Comparison with x-ray mammography and palpation. *Journal of Magnetic Resonance Imaging*, 13(6):868–875.

Webb, G. I. and Ting, K. M. (2005). On the Application of ROC Analysis to Predict Classification Performance under Varying Class Distribution. *Machine Learning*, 59:25–32.

Webb, S. (1996). *The Physics of Medical Imaging*. Medical Science Series, IOP Publishing.

Wehrli, F. H., Shaw, D., and Kneeland, B. J. (1988). *Biomedical Magnetic Resonance Imaging*. VCH Publishers.

Weinstein, D., Strano, S., Cohen, P., Field, S., Gomori, J. M., and Degani, H. (1999). Breast Fibroadenoma: Mapping of pathophysiologic Features with three-time-point, contrast enhanced MR imaging-Pilot study. *Radiology*, 210:233–240.

Wu, Y. and Takatsuka, M. (2005). The geodesic self-organizing Map and its Error Analysis. In *Australian Computer Science Conference*, volume 38, pages 343–352, Newcastle, Australia.

Xu, R. and Wunsch, D. I. (2005). Survey on Clustering Algorithms. *IEEE Tran. on Neural Networks*, 16(3):645–678.

Yang, L. (2004). Distance-Preserving Projection of High-Dimensional Data for Nonlinear Dimensionality Reduction. *Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1243–1246.

Yin, H. (2002). Data Visualisation and Manifold Mapping using ViSOM. *Neural Networks*, 15:1005–1016.

Zhang, C., Wang, J., Zhao, N., and Zhang, D. (2004a). Reconstruction and Analysis of Multi-Pose Face Images based on Nonlinear Dimensionality Reduction. *Pattern Recognition*, 37:325–336.

Zhang, J., Shen, H., and Zhou, Z. H. (2004b). Unified Locally Linear Embedding and Linear Discriminant Analysis Algorithm (ULLELDA) for Face Recognition. In Springer Verlag, editor, *SINOBIOMETRICS*, pages 296–304, Guangzhou, China.