Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

vorgelegt von
Dipl.-Ing. M.Sc. Bernd Michael Kelm
aus Heilbronn

Tag der mündlichen Prüfung: ____________

Auswertung vektorwertiger klinischer Bilddaten mittels probabilistischer graphischer Modelle: Quantifizierung und Mustererkennung

Gutachter: Prof. Dr. Fred A. Hamprecht

Evaluation of Vector-Valued Clinical Image Data Using Probabilistic Graphical Models: Quantification and Pattern Recognition

by
Bernd Michael Kelm

Submitted to the Combined Faculties for the Natural Sciences and for Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Science

at the Rupertus Carola University.

Heidelberg, ____________.

Referees: Prof. Dr. Fred A. Hamprecht

____

Zusammenfassung

Mit der fortschreitenden technologischen Entwicklung bildgebender Verfahren mittels magnetischer Kernspinresonanz (MR) gewinnen Aufnahmetechniken, welche Aufschluss über physiologische Prozesse geben, immer mehr an Bedeutung für die klinische Anwendung. Beispielsweise können mittels der *dynamischen kontrastverstärkten MR Bildgebung* räumlich aufgelöst Rückschlüsse über Perfusion und Permeabilität des Gewebes gezogen werden oder mittels der *MR spektroskopischen Bildgebung* die räumliche Verteilung gewisser Zellmetabolite studiert werden.

Diesen Techniken gemein ist die Gewinnung vektorwertiger Bilddaten, welche für eine effiziente diagnostische Auswertung ungeeignet sind und zunächst algorithmisch aufbereitet werden müssen. Die möglichst völlig automatische Aufbereitung hat dabei zum Ziel, den Informationsgehalt der hochdimensionalen vektorwertigen Ausgangsdaten auf eine aussagekräftige und überschaubare Anzahl von Größen zu reduzieren welche sodann gut visualisiert werden können. Hierbei können generell zwei Herangehensweisen unterschieden werden, nämlich der datenorientierte *Mustererkennungsansatz* und der modellorientierte *Quantifizierungsansatz*. Beide werden in der vorliegenden Arbeit behandelt und führen im Ergebnis zu *Wahrscheinlichkeitskarten* im einen Fall und zu *Parameterkarten* im anderen Fall.

Üblicherweise beschränken sich bestehende Methoden der Einfachheit halber auf eine voxelweise Auswertung. Der grundsätzlich neuartige Ansatz in der vorliegenden Arbeit behandelt dagegen die Einbeziehung räumlicher Bildinformation. Diesbezüglich werden moderne statistische Methoden vorgeschlagen welche es ermöglichen, den räumlichen Kontext in konsistenter Weise zu berücksichtigen. Zum Einsatz kommen hier insbesondere probabilistische graphische Modelle, welche sowohl eine intuitive Modellierung erlauben als auch eine attraktive Auswahl effizienter Inferenz-Algorithmen anbieten. Auf dem Gebiet des maschinellen Lernens und maschinellen Sehens gewinnen diese Verfahren augenblicklich zunehmend an Bedeutung und sind Gegenstand aktiver Forschung.

Die im Rahmen der vorliegenden Arbeit durchgeführten Experimente zeigen, dass die Ausnutzung von räumlichem Kontext mittels graphischer Modelle zu deutlich verbesserten Resultaten führt, sowohl für den Quantifizierungsansatz als auch den Mustererkennungsansatz.

Abstract

With the advancing technological development of imaging techniques based on nuclear magnetic resonance (MR), modalities that carry information about physiological processes gain ever more importance for clinical purposes. *Dynamic contrast-enhanced MR imaging*, for example, provides spatially resolved insight into tissue perfusion and permeability. The spatial distribution of certain cell metabolites can be studied using *MR spectroscopic imaging*.

These techniques have in common that the acquired vector-valued image data is not amenable to an efficient diagnostic evaluation and requires algorithmic preprocessing. The objective of an ideally fully automatic preprocessing is to reduce the information contained in the original high-dimensional vector-valued data to a meaningful and concise set of values that can well be visualized. Two general approaches can be distinguished, namely data oriented *pattern recognition* and model oriented *quantification* approaches. Both are examined in the present work and result in so-called *probability maps* and *parameter maps*, respectively.

For simplicity, existing approaches usually evaluate such data in a voxel-wise fashion. Beyond that, the present work examines the usage of spatial image information. To this end, modern statistical methods are proposed that consistently include spatial context. In particular, the application of probabilistic graphical models is proposed, which allow for intuitive modelling and also provide a attractive pool of inference algorithms. Currently, these models increasingly gain importance in machine learning as well as in computer vision.

The experiments conducted in the present thesis demonstrate that the exploitation of spatial context by means of graphical models leads to significantly better results for quantification as well as for pattern recognition approaches.

Acknowledgments

First of all I would like to thank Prof. Dr. Fred A. Hamprecht, who has provided a great place for research at the Interdisciplinary Center for Scientific Computing (IWR). I am grateful for his unlimited support throughout my time at the IWR, that he gave me the freedom to pursue my research interests and that he provided me with financial support, computing resources and ample opportunities for collaboration. I very much appreciated his interdisciplinary research approach that allowed me to learn a lot about a diversity of exciting applications for statistical signal processing and machine learning methods in the fields of industrial process control and the life sciences. Besides Prof. Hamprecht I would like to thank Prof. Dr. Gerhard Reinelt, Prof. Dr. Reinhard Männer and Prof. Dr. Barbara Paech for serving on my committee. I am grateful for the doctoral stipend and support by the German National Academic Foundation.

My time at the IWR could not have been so successful without the many scientific and fundamental discussions with my colleagues. In particular, I want to thank Björn Menze for sharing his thoughts and experience in the application of pattern recognition methods to spectral data and Linus Görlitz for his help on clarifying some mathematical concepts. I would also like to thank all the people in the image processing groups who gave me an enjoyable time in Heidelberg, in particular Marc Kirchner, Daniel Kondermann, Claudia Nieuwenhuis, PD Dr. Christoph Garbe, Pavel Pavlov, Martin Schmidt, Xin Li, Dr. Pau Montes, Mark Jäger, Dr. Jochen Schmähling, Bjoern Andres, Bernhard Renard, Mario Frank, Stefan Trittler, Matthias Wieler, Dr. Michael Hissmann and Prof. Dr. Bernd Jähne. I also enjoyed advising and working with Natalie Müller, Felix Müller, Michael Wormit, Rahul Nair, Bastien Saquet, Lars Feistner and Nico Splitthoff.

The present work would not have been possible without numerous collaborators who not only provided data and domain knowledge but also devoted a considerable amount of their time to discussions and special purpose experiments.

First of all I want to thank Dr. Christian Zechmann from the German Cancer Research Center (DKFZ, Heidelberg) and his colleagues Dr. Oliver Nix, Dr. Klaus Baudendistel, Dr. Marc-André Weber, Dr. Bram Stiltjes and Dr. Thomas Neff. Also thanks to Prof. Dr. Peter Bachert and Prof. Dr. Dr. Peter Huber who have always been supportive. The clear research focus on clinical relevance helped a lot in understanding and identifying the most important weaknesses and needs in current MR processing approaches. In this context I also would like to thank PD Dr. Heinz-Peter Schlemmer and Dr. Matthias P. Lichy from the University Hospital Tübingen (UKT, Tübingen) who provided MRSI data that I could start working on.

I am deeply grateful to Anke Henning and Prof. Dr. Ulrike Dydak from the Swiss Federal Institute of Technology (ETH, Zürich) for acquiring high resolution MRSI data. In particular, I enjoyed learning about the technical pitfalls of MRSI and to fathom its current technological limits.

Many thanks go to Prof. Dr. Andrew McCallum and the Information Extraction and Synthesis Laboratory (IESL) at the University of Massachusetts (UMASS, Amherst) for the cordial hospitality and for providing an inspiring working environment during the three months of my visit. I also had the pleasure and opportunity for intense cooperation with Dr. Chris Pal which gave me many new insights into basic issues of machine learning and I want to thank Charles Sutton for explaining algorithmic details of graphical models. It was a pleasure to attend the reading group of the Computer Vision Laboratory and I want to thank its members for many interesting discussions, in particular Prof. Dr. Erik Learned-Miller and Jerod Weinman.

Thanks to Dr. Peter Verweer from the European Molecular Biology Laboratory (EMBL, Heidelberg) who provided fluorescence lifetime imaging data (FLIM) which allowed me to examine the transfer of the proposed algorithms to another interesting domain of life science imaging.

Finally, I would like to thank Prof. Dr. Thomas G. Dietterich who first pointed me to the exciting field of graphical models and infected me with his enthusiasm for machine learning methods.

Thanks to Seong-Min and my parents for their love and support.

Contents

Chapter 1.

Introduction

1.1. Motivation

In 2003, Paul C. Lauterbur and Sir Peter Mansfield jointly received the "Nobel Prize in Physiology or Medicine for their discoveries concerning magnetic resonance imaging". Their seminal work enabled modern magnetic resonance imaging (MRI) as it is available in medical diagnostics today and initiated manifold research into an exciting, useful and versatile imaging modality.

Magnetic resonance imaging has always been an interdisciplinary challenge [116] and the success of clinical MR tomography has only been possible because of the mutual exchange of ideas between physicists, engineers and physicians. Interdisciplinarity is still of crucial importance for the field and is reflected in the diversity of researchers that meet at conferences like the annual meetings of the International Society for Magnetic Resonance in Medicine (ISMRM, `http://www.ismrm.org/`) or the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB, `http://www.esmrmb.org/`).

Measuring magnetic properties of atomic nuclei *in vivo* has several advantages over other clinical imaging modalities, such as for example Computer Tomography (CT) or Positron Emission Tomography (PET). First of all, MR does not pose any radiation risk to the patient and images can be acquired completely *non-invasively*. It has proved very useful for imaging soft tissue and using fast imaging techniques such as Echo-Planar Imaging (EPI), anatomical cross sections of the human body can be acquired in less than 100ms on modern clinical MR scanners. Last but not least, magnetic resonance is extremely versatile and allows for a whole family of modalities that record different anatomical and physiological properties.

Most often, MR is used to record resonance signals from the protons (^{1}H) of water molecules using a suitable sequence [115, 125]. Various magnetic properties of the water protons can be recorded such as the spin-grid-relaxation time T_1, the spin-spin-relaxation time T_2 or simply the echo strength revealing proton density (PD) [135, 66]. Nowadays, these techniques are so well-understood that even whole-body scans can be performed in a few minutes on commercially available MR scanners.

Beyond these standard MR sequences other, more advanced MR modalities are used for research in an *in vivo* situation. In functional MRI (fMRI), for example, the BOLD (blood-oxygenation-level dependent) effect is used to visualize brain activity. The diffusivity of tissue can be visualized with diffusion tensor imaging (DTI) which provides the basis for tractography. Although very interesting for research, these modalities are meaningful mostly in the brain and of limited utility for clinical purposes.

Two MR modalities that are highly relevant for the diagnosis of pathophysiologies, in particular tumors, are dynamic contrast-enhanced MR imaging (DCE-MRI) and magnetic resonance spectroscopic imaging (MRSI).

Dynamic Contrast-Enhanced MR Imaging (DCE-MRI) is used to track the diffusion of a paramagnetic contrast medium (CM) such as Gd-DTPA and study tissue perfusion and vascular permeability *in vivo* [135, 41]. During the intravenous injection of the CM, a sequence of several T1-weighted MR image volumes is recorded at intervals of a few seconds (Fig. 1.1). Hence, a T1 signal-time curve is obtained for every voxel (Fig. 1.2). A diagnostic evaluation of this signal-time curve is usually based on a pharmacokinetic model whose parameters characterize the uptake and the washout of the CM from the underlying tissue (*e.g.* [41]). Since these parameters change characteristically in pathologic tissue, DCE-MRI can be used to detect and localize tumors. Its application comprises but is not limited to the diagnosis of breast [40], bone-marrow [86] and prostate cancer [223, 178, 104].

Magnetic Resonance Spectroscopic Imaging (MRSI) is used to obtain a spatially resolved information on the concentration of certain biomolecules [22, 170] (cf. Fig. 1.3). It can be acquired completely noninvasively using standard clinical MR scanners. In ^{1}H NMR spectroscopic imaging, resonance signals from the water are suppressed and one is interested in the resonance signals from protons that are bound to cell metabolites instead. These can be detected by a characteristical shift of their Larmor frequency. Since the chemical environment of molecules can damp the magnetic field, protons bound to different metabolites resonate at slightly different frequencies which is called *chemical shift* (cf. 1.4). However, the concentration of metabolites is much smaller than that of water which results in metabolite resonances that are orders of magnitudes ($\sim 10^5$) smaller than for water. Thus, *in vivo* ^{1}H NMR spectroscopic imaging yields spectral images with information about the local cell metabolism but usually with very poor signal-to-noise ratio (SNR). This poses problems to any signal processing procedure. Additional problems can arise

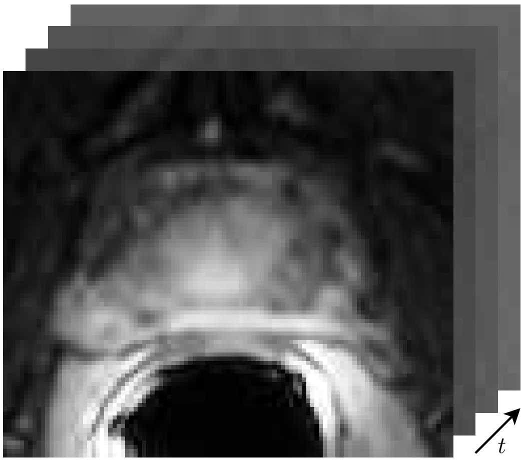


Figure 1.1.: In Dynamic Contrast-Enhanced MR imaging (DCE-MRI) a sequence of T_1-weighted MR images is recorded after the injection of a paramagnetic contrast medium. Its abundance can be monitored as intensity time-curves in every voxel which allows to draw conclusions about tissue perfusion and permeability.

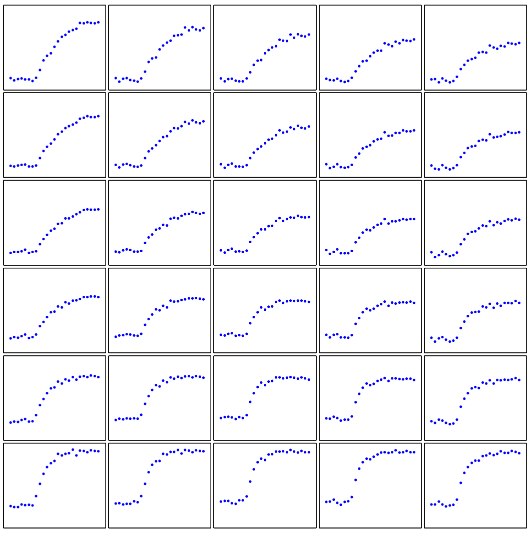
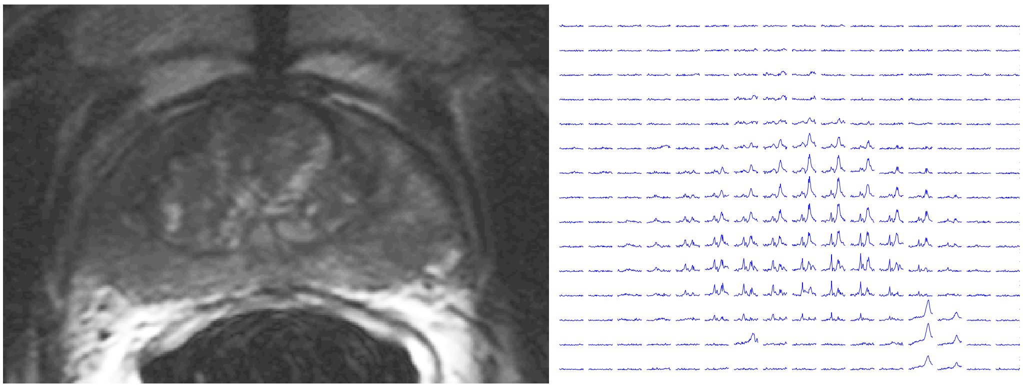


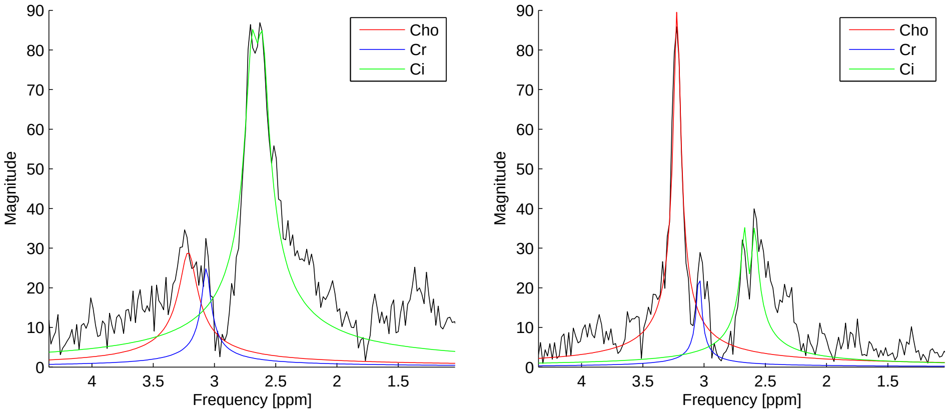
Figure 1.2.: Signal-time curves from adjacent voxels of a DCE-MR image of the prostate. Without postprocessing of the vector-valued DCE-MR image its diagnostic content could only be assessed by an experienced radiologist and would require inspecting each voxel individually. Certainly this way of extracting diagnostic information lacks objectivity.



(a) T_2-weighted MR image.

(b) ^{1}H-MRS image.

Figure 1.3.: T_2-weighted MR and ^{1}H-MRSI image of the same slice of a prostate tumor. The ^{1}H-MRS image carries metabolic information which, however, can only be accessed after appropriate postprocessing of the spectral image. One such approach is shown in Figs. 1.5 and 1.6.



(a) Spectrum from healthy voxel

(b) Spectrum from tumor voxel

Figure 1.4.: In ^{1}H-MRSI each voxel contains a spectral signal that reveals metabolic information. The shown examples are typical prostate spectra and show the three resonances of choline (Cho), creatine (Cr) and citrate (Ci). Tumors can be identified by high choline and reduced citrate concentrations.

from artifacts such as susceptibility artifacts, foldover artifacts, overlapping peaks or broad baselines stemming from macromolecules [106].

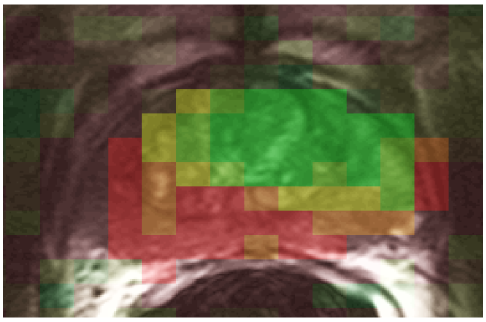


Figure 1.5.: Patient from Fig. 1.3 evaluated with a pattern recognition method. Red voxels indicate high probability for tumor whereas green voxels indicate low probability. Instead of evaluating every voxel individually, probability maps can be generated fully automatically with pattern recognition methods proposed in the present thesis.

On an abstract level, both modalities thus produce *vector-valued MR image data* that captures information about *physiological processes*. Currently, both are actively researched for their applicability and clinical utility.

In contrast to common MRI modalities which can be depicted as gray valued images, both DCE-MRI and MRSI produce data that require automatic postprocessing in order to reduce their dimensionality for display. If no such dimensionality reduction was performed, the signal in every voxel would have to be analyzed individually which, at high spatial resolutions, quickly becomes impractical. Clearly there is a need for automatic and reliable evaluation strategies for vector-valued MR image data such as, for example, DCE-MRI and MRSI.

1.2. Scope

Depending on what kind of information is to be gained from the data, two principally different evaluation strategies can be pursued which are termed *quantification* and *pattern recognition* in the present thesis:

Quantification denotes the process of deriving physically meaningful parameter estimates from the given data. It usually consists of constructing a *parametric model* for the vector-valued signal and subsequently determining parameters in every voxel that best explain the observed data. The spatially resolved data is then represented by *parameter maps* (or images) that can be depicted and interpreted by the physician.

Pattern Recognition (PR) attempts to tackle the (diagnostic) decision problem directly. Based on a *training data set*, created by an expert, PR reduces the amount of physical modelling and concentrates on the available data instead. It does this

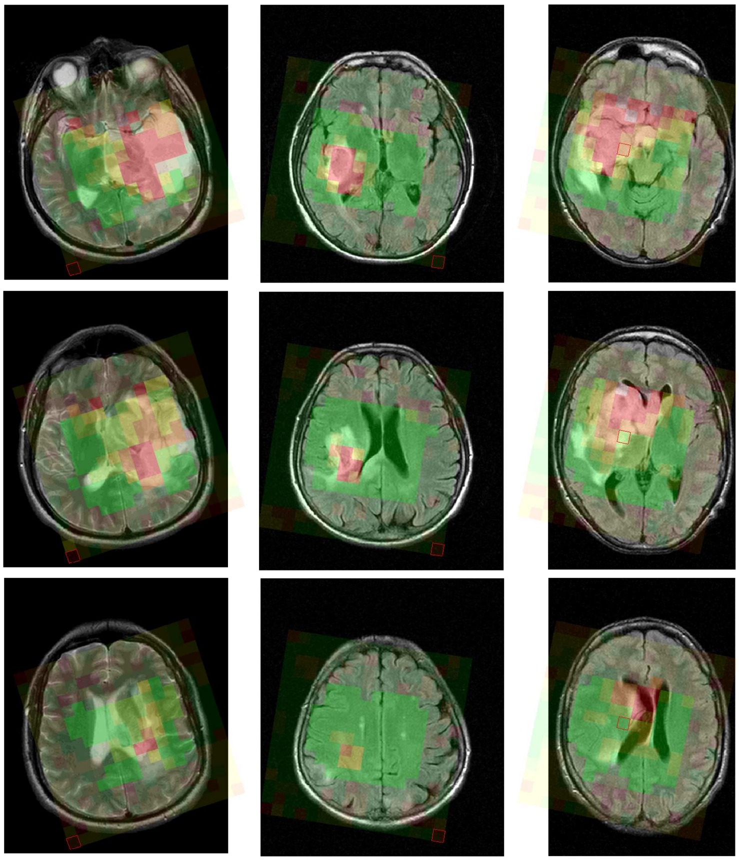


Figure 1.6.: ^{1}H-MRSI data from a patient with brain tumor evaluated with pattern recognition methods. After the automatic processing of the whole MRSI data volume the physician can quickly browse through all slices and concentrate on suspicious regions only (image courtesy of Björn Menze [16]).

by *mimicking* or *imitating the human expert* (a physician) in deriving diagnostic information.

Both approaches, quantification and pattern recognition are usually applied in a voxel-wise fashion, thus neglecting the spatial nature of the MR data. In the present thesis, methods are proposed and investigated that incorporate information from a local neighborhood in processing the data in each voxel. Thus, the focus is shifted from evaluating a *collection of voxels* to evaluating *image data*. Based on the theory of *graphical models* the proposed methods pay special attention to the application of global and sound probabilistic models. It is shown that both quantification and pattern recognition can significantly gain from utilizing spatial context.

1.3. Outline

All chapters of the present thesis are largely self-contained in that they may be read and understood individually though the thesis is organized around the described distinction between *quantification* and *pattern recognition*. Theoretical background on methods and mathematical tools is collected in the appendices.

Chapter 2 sets off with a comparison of quantification based and pattern recognition approaches applied to the estimation of tumor probability from prostate MRSI based on the independent evaluation of single voxels. Parts of chapter 2 have been published in [4] and [9].

The following two chapters describe ways to employ spatial context for improving quantification and pattern recognition, respectively. In chapter 3, the application of a generalized Gaussian Markov random field (GGMRF) is proposed to introduce spatial prior knowledge in the fitting of nonlinear model functions. An efficient blocked version of the iterated conditional modes algorithm is proposed for tackling the resulting high-dimensional but sparse optimization problem. Parts of chapter 3 are found in [10, 2] and [1].

In chapter 4, approaches for pattern recognition using spatial context are proposed. For this purpose the sought label map is represented by a discrete-valued random field and generative as well as discriminative probabilistic graphical models for simultaneous segmentation and classification are derived. A conditional random field (CRF) with Tikhonov-like parameter prior is shown to provide a sound and interpretable approach for spectral data. Results on simulated MRSI data show that both accuracy and the area under the receiver operator characteristic can be significantly improved over the single voxel approach, in particular with increasing signal noise. Thus, using spatial prior knowledge in form of the proposed CRF will allow to record MRSI at even higher spatial resolutions. Parts of chapter 4 are found in [3].

The thesis concludes in chapter 5 with a short summary of the most important results.

Chapter 2.

Estimating Tumor Probability From Magnetic Resonance Spectra

2.1. Introduction

Clinical studies have shown significant diagnostic value of ^{1}H magnetic resonance spectroscopic imaging (MRSI) for the detection of tumorous tissue in the prostate [173, 51, 221, 175, 190, 149]. Despite the promising results of these and other studies, the integration of MRSI in the clinical routine remains difficult. Among the reasons are

- Technical Problems. The acquisition of high quality MRSI data *in vivo* is not an easy task and still requires a lot of experience. In general, the acquired signals have a very low signal-to-noise ratio and frequently the signal is distorted in a way that completely obscures the metabolic information of interest (e.g. peak broadening, baselines from residual water and lipid resonances).
- Signal Processing. Inaccurately quantified metabolite ratios resulting from noisy or distorted signals lead to questionable results in the statistical evaluation. A fully automatic evaluation, based on either quantified signals or on spectral patterns, cannot be trusted if information is provided without a clue on its reliability.
- Medical Interpretation. It is not exactly clear according to which rule the metabolic information should be evaluated. A lot of knowledge and training is required in order to correctly interpret the spectra. Even then, the result is not objective and might differ from person to person.

- Efficiency. The manual evaluation and visual classification of multiple single spectra in MRSI data sets is very time-consuming, especially with increasing spatial resolutions. In the clinical routine, MRSI also competes with other imaging modalities. If the evaluation is too time-consuming and the risk of gaining only little diagnostic information is high, other modalities might become more attractive. Certainly there is a non-trivial cost-benefit tradeoff to consider.

Two basic approaches to the evaluation of MRSI can be distinguished: the *quantification based* approach and the *pattern recognition* (PR) approach. Quantification aims at estimating relative metabolite concentrations as accurately as possible. For that purpose, the most likely parameter estimate for a given signal model is usually determined with a nonlinear least squares (NLS) approach. However, quantification may fail for various reasons. In particular, in the presence of artifacts and severe noise the NLS objective can have many local optima and the result becomes very sensitive to the choice of initial values. Prior knowledge about the expected signal shape can help to alleviate these problems [204], but it also leads to an estimation bias and can be harmful in unanticipated cases where the employed prior knowledge is inadequate. A subsequent statistical analysis which gains diagnostic information from the spectral data relies on these parameter estimates and therefore inherits all problems associated with the quantification.

Pattern recognition approaches do not require an explicit quantification step. Although the same methods and classifiers (*e.g.* logistic regression [85], artificial neural networks [85, 67], support vector machines (SVM) [179], etc.) can be used for both, quantified signals and spectral patterns, only methods applied to the latter will be referred to as "pattern recognition" (PR) approaches in accordance with, for example, [84]. The PR approach is characterized by minimal preprocessing, thus avoiding errors introduced by feature calculation steps. It is left to the classifier to construct features and extract the relevant information to distinguish random effects from significant changes in the spectral pattern. Since it is not exact quantification that is the main goal in clinical applications but accurate diagnostic information, it is suggested to address the diagnostic problem directly without prior quantification (cf. Fig. 2.1).

In the following, related work is briefly reviewed in order to emphasize common ideas and highlight differences with the proposed approach. Recently, encouraging results have been reported [193, 61, 123, 62, 184, 124] from studies on the automated classification of brain tumor spectra in the context of the INTERPRET project (`http://carbon.uab.es/INTERPRET/`). Tate *et al.* [193] show that the influence of acquisition parameters (manufacturer, sequence, TE, TR) on the spectral pattern is small enough to allow for stable classification results across multiple centers.

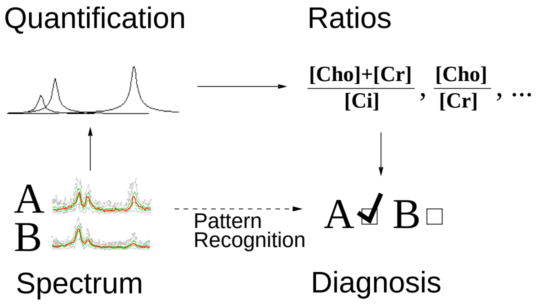


Figure 2.1.: Two approaches to the diagnostic evaluation of NMR spectroscopic data. After quantification of the signal, concentration ratios are used to answer a diagnostic question. In contrast, pattern recognition approaches address the diagnostic question directly based on the spectral pattern.

In [61] and [123], Devos and Lukas *et al.* examine and compare different preprocessing strategies and classifiers for long and short echo time brain spectra respectively. They show that the best results are obtained with L_2-normalized magnitude spectra, omitting for example baseline and phase corrections. Although a nonlinear classifier has been employed, no improvement over linear classifiers could be observed in both studies, which the authors attribute to the limited amount of available data.

Laudadio *et al.* [113] propose a PR approach using magnitude spectra that incorporates spatial context. It is applied to simulated as well as *in vivo* prostate MRSI data and focuses on evaluating the benefit of incorporating spatial information.

In [7], Menze *et al.* examine classifiers for the discrimination of recurrent tumor and brain lesions after radiotherapy based on single voxel MRS. An exhaustive combination of feature extraction methods and classifiers is benchmarked according to several error measures. Regularized linear classifiers with preceding dimensionality reduction (binning) are found to perform best on the given data set.

A similar comparative study has not been performed on prostate MRSI data yet. In the present chapter an extensive collection of linear subspace methods and a representative set of state-of-the-art nonlinear classifiers are evaluated on prostate data. For the first time also the influence of different quantification algorithms on the classification results is examined. Furthermore, experiments are conducted comparing the use of magnitude and real spectra. Since it is common practice in prostate MRSI to analyze the acquired data based on quantification [173, 51, 221, 175, 190, 149], the comparison of quantification based approaches with PR approaches is emphasized.

2.2. Methods

Only approaches that can be used for a fully automated analysis of MRSI data are considered in this study because extensive user interaction is not acceptable in clinical routine use. An emphasis is put on methods that can provide tumor *probability* estimates, a much richer description of classification results than hard class labels. Finally, all selected methods have either been proposed for NMR spectroscopic data before or are closely related to such methods.

The section starts with a short description of the employed data set. Subsequently, the used feature extraction and classification methods are concisely summarized with ample references to the literature. The last subsection is devoted to the error measure used to compare the different approaches.

2.2.1. Data

^{1}H-NMR spectroscopic image volumes from an ongoing prostate MRSI study have been collected at the German Cancer Research Center (dkfz, Heidelberg). The data was acquired on a clinical 1.5T scanner (Magnetom Symphony; Siemens Medical Solutions, Erlangen, Germany) with a disposable endorectal coil (MRInnervu; Medrad Inc., Indianola, PA, USA) and the protocol described in [174, 173]. 512 datapoints with a bandwidth of 1000-1250 Hz were acquired (TE/TR=120/650 ms). The field of view (FOV) and the volume of interest (VOI; selected with PRESS pulses) were adapted to the size of the individual prostates. Typical FOVs were around 60-66 × 78-84 × 66-78 mm. An elliptical k-space acquisition scheme and apodization with a Hanning filter was employed [174]. The total acquisition time was limited to 10 minutes and the spectral data was interpolated to yield a volume of 16^3 voxels. Along with the MRSI data, T_2-weighted axial MR images (turbo-spin echo, TE=129, TR=4000-4800 ms, FOV= 140 × 140 mm, matrix size 512 × 512, 20-25 slices, slice thickness = 4 mm) were acquired. Two exemplary spectra are shown in Fig. 2.2.

For 12 of the 36 recorded patients, poor shimming, ineffective fat suppression or problems with the endorectal coil resulted in corrupted MRSI data. These patients have been excluded from the data set. For several patients, results from a histologic step-section examination were available. These could be used as "gold standard" for a qualitative evaluation. The training set was created using a semimanual analysis of the spectra according to standard decision rules based on the metabolite resonances of Cho, Cr and Ci [149, 190, 220] . Altogether, 76 slices with 256 voxels each have been labeled with respect to their spectral pattern class (healthy, undecided, tumor) and the signal quality (not evaluable, poor, good). In judging the signal quality both, low signal-to-noise ratios and artifacts (nuisance resonances, heavy baselines) have

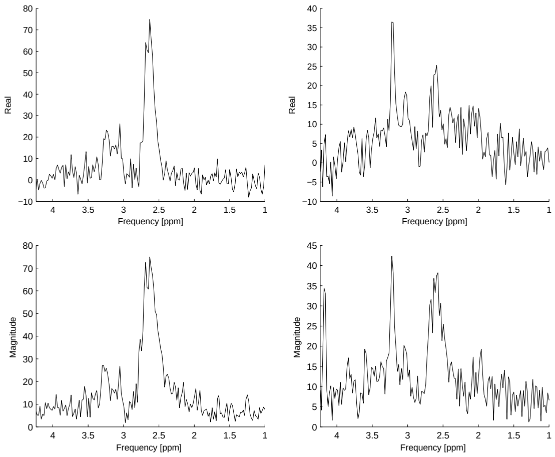


Figure 2.2.: Two example spectra (left: healthy, right: tumor) after HSVD water/lipid removal and zerofilling to 1024 datapoints. The top row shows manually phased real absorption spectra whereas the bottom row shows the corresponding magnitude spectra. A slight increase in line width can be observed when switching from real to magnitude spectra.

Table 2.1.: Distribution of labels in the prostate data set (76 slices from 24 patients).

quality \ class	healthy	undecided	tumor	all
not evaluable	–	–	–	15268
poor	721	437	284	1442
good	1665	629	452	2746
all	2386	1066	736	19456

been considered. An overview of the collected data is given in Tab. 2.1. Only spectra that are evaluable (signal quality “poor” and “good”) have been used in this study. The large number of “not evaluable” voxels is due to outer volume suppression and the limiting coil sensitivity profile in prostate MRSI. Only about one fourth of the voxels in the FOV actually lie within the prostate.

2.2.2. Preprocessing and Feature Extraction

Both quantification and PR profit from the prior removal of nuisance peaks and baselines in the spectra. Therefore, prior to further processing, the residual water and lipid resonances were removed by time-domain selective HSVD filtering (cf. appendix A), i.e. by removing all signal components with poles outside the interesting frequency range of 2.4 to 3.6 ppm.

Table 2.2.: FID components used for quantifying prostate MRSI. The parameters have been initialized with the given value and constrained to the range given in brackets.

Metabolite	Model	Frequency [ppm]	Line Width [Hz]
Choline	Lorentzian	3.22 [±.03]	6.25 $[0, 31.25]$
Creatine	Lorentzian	3.04 [±.03]	6.25 $[0, 31.25]$
Citrate-1	Lorentzian	2.65 [±.03]	6.25 $[0, 31.25]$
Citrate-2	Lorentzian	2.60 [±.03]	6.25 $[0, 31.25]$

Quantification. Three different methods have been used for the quantification: QUEST [160] and AMARES [204] from the jMRUI tool [139] and a custom implementation of a constrained VARPRO approach which used an interior trust region algorithm for optimization [52] (cf. appendix A). Quantification was performed with four Lorentzian components (cf. Tab. 2.2). Besides small frequency shifts of ±.03 ppm for the individual components, a common shift of up to ±.625 ppm was allowed for in the VARPRO approach. Furthermore, the zero-order phases of all components have been tied. Similar constraints have been used for AMARES. Since AMARES does not support constraints on the overall frequency shift, the individual components have been constrained to ±.625 ppm. In addition, the amplitudes of the two citrate peaks have been tied. For QUEST, three metabolite templates have been constructed by simulating noise-free Lorentzian lines according to Tab. 2.2.

Spectral Patterns. For the PR approach, zerofilling yielded an interpolated spectrum at 1024 frequencies. Automatic zero-order phase correction was performed based on the first recorded data point. From both magnitude and real spectra, 40 values at equidistant frequencies between 3.34 ppm and 2.36 ppm have been calculated by linear interpolation to account for differences in the imaging frequency and the bandwidth. Finally the spectral patterns have been L_1-normalized, *i.e.* each channel was divided by the sum of the absolute values over all channels. Fig. 2.3 shows robust statistics of the extracted spectral magnitude patterns as obtained from the evaluable spectra in the prostate data set. The general tumor pattern of elevated Cho + Cr peak (channels 8/14) versus a reduced Ci peak (channel 31) is clearly recognizable.

Different subspace methods have been used for dimensionality reduction of the spectral patterns. They are particularly appropriate for prostate MRSI since, ideally, only three metabolites contribute to the spectral shape. In particular, four subspace methods have been considered: principal components analysis (PCA), partial least squares (PLS), independent component analysis (ICA) and nonnegative matrix fac-

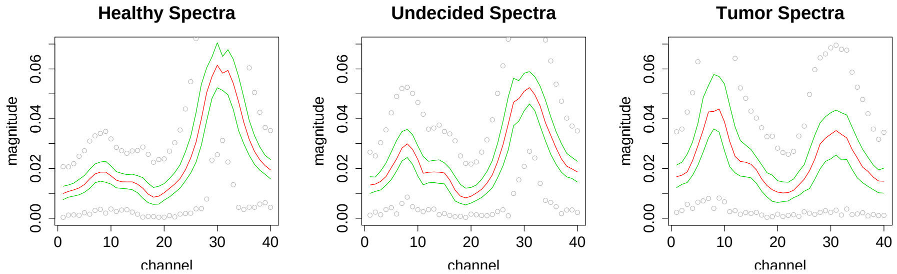


Figure 2.3.: Spectral patterns in the prostate data (3.34-2.36 ppm). From left to right typical patterns of healthy, undecided and tumor tissue can be recognized with their characteristic choline (channel 8), creatine (channel 14) and citrate (channel 31) ratios. In the spirit of a box-and-whiskers plot [129], the median (red), the hinges (green) and extreme points (circles) are shown for each channel.

torization (NMF) which are briefly described in the following (more details are found in appendix B).

- PCA seeks K uncorrelated latent variables $z_k(x) = \alpha_k^T x$ (factors, score variables) that capture all relevant information of the original predictors x. The loadings α_k are obtained as the directions of maximum variance. PCA is described, for example, in [85] and has successfully been used for MRS in [61, 123, 184, 7].
- PLS also seeks uncorrelated factors but additionally considers the given classification task. The latent variables are determined by maximizing both the variance and the correlation with the class label [85]. The determined subspace can thus be expected to better capture the information relevant for classification. PLS has originally been proposed in chemometrics [214] and is therefore designed for spectral data. Its good performance in clinical MRSI has been demonstrated in [7].
- ICA is a subspace method that has been used for MR spectra for example in [184]. As opposed to PLS and PCA, ICA not only requires uncorrelated but statistically independent components. After centering, prewhitening and dimensionality reduction, ICA reduces to a search over rotations that minimize the mutual information between the components or equivalently maximize the negentropy [85, p.498]. In this study, the FastICA algorithm has been used with the logcosh approximation to negentropy [92].
- NMF has also recently been proposed for the extraction of spectral components [168]. It enforces nonnegative loadings and scores which is a reasonable constraint for magnitude spectra. Here a robust version of the alternating nonnegative least squares algorithm has been used.

An important advantage of linear subspace methods is their amenability to interpretation. The weighting of the spectral channels expressed in the constructed components or loadings can be visualized and helps to understand the decision process of the trained classifier.

2.2.3. Classification

Linear classifiers model the decision boundary as a hyperplane in the space of the explanatory variables. Several studies on MRS classification have applied linear discriminant analysis (LDA) [61, 123, 193] which models the feature distributions as Gaussians with common covariance matrix. Instead, *logistic regression* (LR) which can be derived from the same probabilistic model by using conditional likelihood [85, pp.103ff] has been used. LR is designed for discriminating classes instead of modeling feature distributions which is appropriate for classification tasks [85, p.105]. It easily generalizes to the multi-class problem without requiring additional tools such as, for example, error-correcting output codes [65].

Furthermore, two linear classifiers which have explicitly been designed for spectral data were considered. *Generalized PLS* (GPLS) can be used to perform LR and PLS in one step [127]. *P-spline signal regression* (PSR) exploits the prior knowledge that neighboring spectral channels are correlated by modeling the coefficient profile as a cubic spline function [128].

Nonlinear classifiers are more powerful than linear classifiers in that nonlinear decision boundaries can be constructed. However, this also leads to “black-box” methods which, in general, are hardly interpretable. Here, three nonlinear classifiers are considered: *random forests* (RF) [39], an ensemble method, and *support vector machines* (SVM) and *Gaussian processes* (GP), two kernel methods [179].

In short, the RF classifier learns a collection of a few hundred slightly different decision trees [39]. The diversity of the trees is encouraged by using bootstraps of the given sample and by randomly selecting a subset of feature variables considered in each node when growing the decision trees. A new example is classified according to the majority vote of the trees in the forest. Thus, the RF classifier employs ideas common with bagging and boosting [85].

Kernel methods perform an implicit mapping to a high-dimensional feature space. The constructed linear decision boundary (a hyperplane) in this high-dimensional feature space corresponds to a nonlinear decision boundary in the original feature space. By using a positive definite kernel instead of the usual dot-product, most linear classifiers can be “kernelized” to yield nonlinear classifiers. In this study two kernel methods have been used, support vector machines (SVM) and Gaussian processes (GP) [179]. The least-squares SVM used for MRS classification for example in [61, 123] can be viewed as a kernelized ridge regression and, except for an additional bias term, is identical to the GP method used in this study [74].

2.2.4. Error Measure

The area under curve (AUC) of the receiver operator characteristic was used to measure classification performance. It is determined as the area under the graph obtained by plotting $sensitivity$ against $1 - specificity$. Since it does not depend on the chosen threshold that determines the tradeoff between the true positive and true negative rates, it is independent of class priors and misclassification costs. It is therefore an appropriate performance measure for comparing binary classifiers. The AUC attains its maximum value of 1 for perfect separation, whereas it is .5 for random predictions.

Cross-validation (CV) has been used to obtain reliable estimates for the AUCs. In using CV, it should be considered that spectra obtained from the same patient are certainly correlated, violating the i.i.d. assumption in CV. Therefore a “leave-one-patient-out” scheme which determines the performance measure (AUC) for every patient with the classifier trained on all other patients was employed.

2.3. Results

All reasonable combinations of feature extraction methods and classifiers have been evaluated. The tested combinations are listed in Fig. 2.4 where the methods have been abbreviated as described in the previous section. The employed box-and-whiskers plots [129] are robust summaries of the 24 AUC values obtained from leave-one-patient-out cross-validation. The thick line within the box marks the median value and the box itself is bounded by the two hinges which are versions of the first and third quartiles. The whiskers extend to the most extreme data points which are no more than 1.5 times the interquartile range from the box.

2.3.1. Linear PR methods vs. quantification.

Fig. 2.4a compares quantification approaches based on VARPRO (v), AMARES (a), QUEST (q) and two PR approaches. In addition to various classifiers, results from the conventional metabolite ratio rule $(\mathrm{Cho}+\mathrm{Cr})/\mathrm{Ci}$ (and $\mathrm{Ci}/(\mathrm{Cho}+\mathrm{Cr}+\mathrm{Ci})$ in the case of VARPRO) are provided. Since, given a particular quantification algorithm, all classifiers performed similarly, not all results are depicted for AMARES and QUEST. It should be noted that only spectra for which at least one of the peaks was found ($a_k > 0$) have been used in the evaluation of AMARES and QUEST. For AMARES this was about 74% and for QUEST 97% of the data set. The performance of the two PR approaches PCA/LR and PLS/LR (PLS and PCA with logistic regression) based on magnitude (m) spectra is comparable with that of QUEST-based quantification approaches.

2.3.2. Linear vs. nonlinear PR methods.

Fig. 2.4b compares linear and nonlinear PR approaches based on magnitude spectra. For comparison, the first compartment repeats the results for QUEST. The second compartment summarizes linear and the third compartment nonlinear PR approaches.

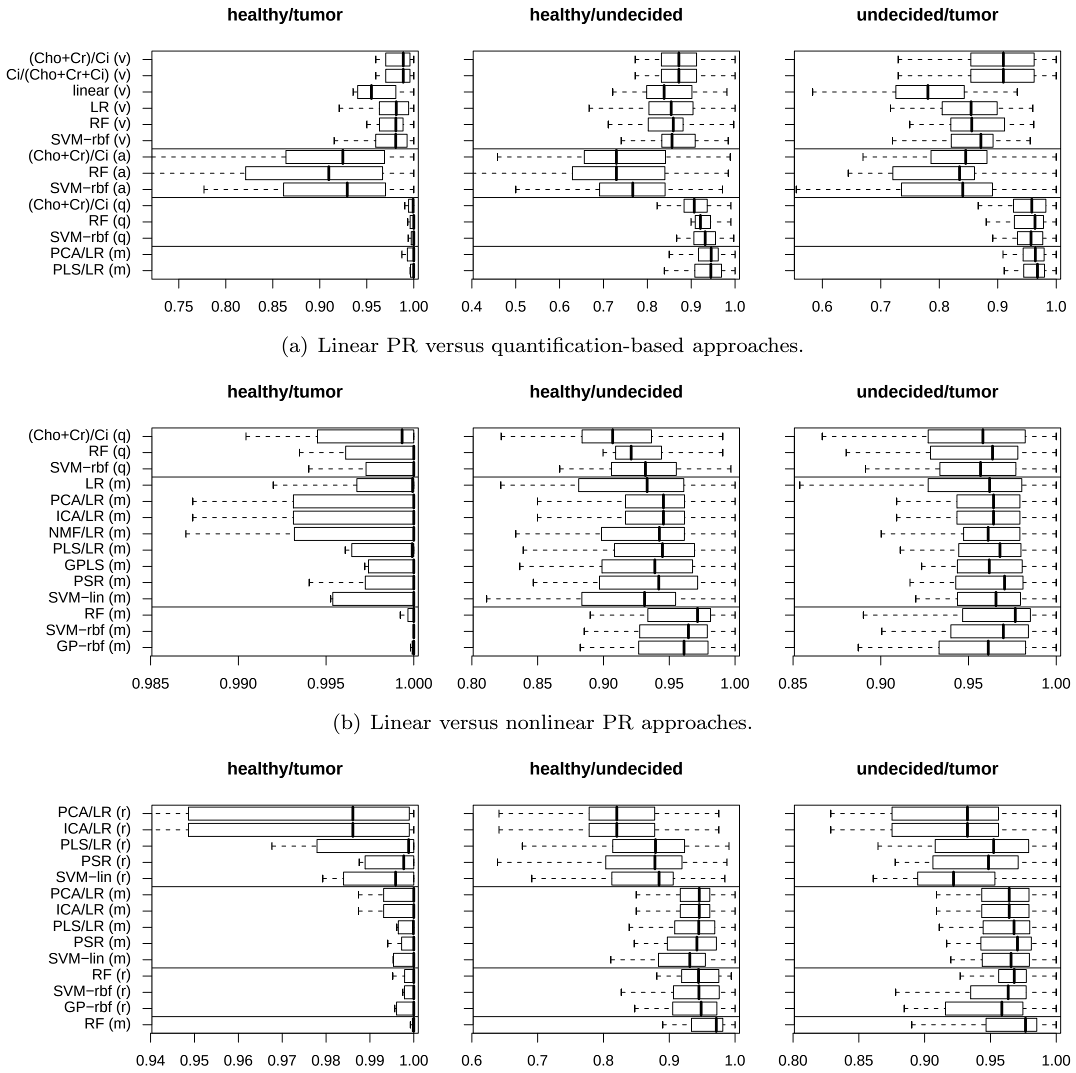
LR (m) shows results with (unregularized) logistic regression based on all 40 spectral channels. Then, results for the five subspace methods PCA, ICA, NMF, PLS and GPLS are given. In the conducted experiments the four most important loadings have been used which, in the case of PCA and ICA, covered about 80% of the variance and seemed sufficient according to a scree plot (not shown). For PSR, a generalized linear model (GLM) with logistic link function and binomial posterior has been used, the same GLM which yields LR. The SVM with linear kernel (SVM-lin) is listed as a linear method since the decision boundary remains a hyperplane in the original feature space.

Finally, results for the nonlinear PR methods are provided. The random forest (RF) classifier has been trained with 500 trees, nodesize 1 and a subset of 13 considered variables in each split. For the SVM as well as for the GP method, the width of the employed radial basis function (RBF) kernel has been estimated from a fraction of the respective training data set (procedure sigest, cf. [102]).

2.3.3. Real vs. magnitude spectra.

In Fig. 2.4c, PR methods using magnitude and real spectra are compared. First, results for the subspace methods PCA, ICA and PLS are provided (NMF does not make sense for real spectra), followed by PSR and the linear SVM. The corresponding results for magnitude spectra are repeated for comparison.

The last two compartments show results obtained with nonlinear classifiers. Based on real spectra, results for the SVM and GP classifiers with RBF kernel and the



(a) Linear PR versus quantification-based approaches.

(b) Linear versus nonlinear PR approaches.

(c) Real versus magnitude spectra.

Figure 2.4.: Comparison of various approaches: (v)-VARPRO, (a)-AMARES, (q)-QUEST quantification based approaches versus PR approaches based on (m)agnitude and (r)eal spectra. The box-and-whiskers plots show the median, the hinges and the extreme points of the area under curve (AUC) values of the receiver operator characteristic obtained from leave-one-patient-out cross-validation. Linear PR approaches combining a subspace method X with logistic regression (X/LR) easily achieve the same performance as the best quantification approaches (*i.e.* QUEST). Even slightly better results are obtained with nonlinear PR approaches (RF, SVM, GP) applied to raw magnitude spectra (m). Details of the different methods are described in the text. Note that the individual plot scales differ.

Table 2.3.: Cross-validation results for a selection of the tested methods. The given test error values are one minus the AUC of the receiver operator characteristic when training is performed on all but the tested patient, *i.e.* better performance is indicated by smaller values.

	Pattern Recognition				**Quantification**	
patient	**SVM-rbf (m)**	**GP-rbf (m)**	**RF (m)**	**PLS/LR (m)**	**SVM-rbf (q)**	**(Cho+Cr)/Ci (q)**
1	5.00e-04	2.00e-04	2.00e-04	6.00e-04	2.70e-03	9.60e-03
2	0	0	0	0	0	0
3	0	0	0	2.49e-02	0	0
4	0	0	0	0	0	1.50e-03
5	0	0	0	0	1.00e-04	1.30e-03
6	0	0	0	0	0	0
7	1.90e-03	1.00e-03	2.00e-03	1.00e-02	6.00e-03	4.90e-03
8	1.70e-03	1.70e-03	0	1.70e-03	7.10e-03	1.78e-02
9	0	0	0	0	0	0
10	0	0	0	2.00e-04	0	2.60e-03
11	0	0	8.00e-04	2.30e-03	3.00e-04	1.40e-03
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	3.20e-03	0	0
15	0	1.90e-03	5.00e-04	2.00e-04	3.31e-02	3.56e-02
16	0	0	1.44e-02	8.85e-02	2.87e-02	4.07e-02
17	3.46e-02	5.63e-02	6.49e-02	6.49e-02	0	0
18	0	0	0	0	0	0
19	0	0	0	3.90e-03	3.90e-03	6.00e-03
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	3.00e-04	4.30e-03	4.27e-02	0	2.00e-03
23	0	0	0	0	2.70e-03	0
24	0	0	0	0	0	8.30e-03
mean	**1.60e-03**	**2.60e-03**	**3.60e-03**	**1.01e-02**	**3.50e-03**	**5.50e-03**

RF classifier are provided. Representative for the nonlinear classifiers applied to magnitude spectra, RF (m) is repeated.

2.3.4. Detailed comparison.

Detailed cross-validation results for six of the tested methods are provided in Tab. 2.3. For each of the 24 patients one minus the AUC of the respective classifier is given when trained on all other patients. In the first four columns, results for three nonlinear classifiers (SVM, GP and RF) and LR with PLS-subspace based on magnitude (m) spectra are listed. Then, two quantification approaches based on QUEST (q) follow. Since no training is required for the ratio rule, (Cho + Cr)/Ci (q) just reflects the AUC results when broken down to individual patients. The last row provides mean values for the respective methods.

Although the performance differences between the classifiers in Tab. 2.3 seem to be small, statistical significance of some differences can be established using a Wilcoxon signed rank test. Concerning the question whether nonlinear classifiers can improve results over linear methods, it is observed that based on magnitude spectra, SVM-rbf (m), GP-rbf (m) and RF (m) significantly outperform PLS/LR (m) (p = .0002/.0012/.0006). Also, based on QUEST the SVM-rbf (q) performs sig-

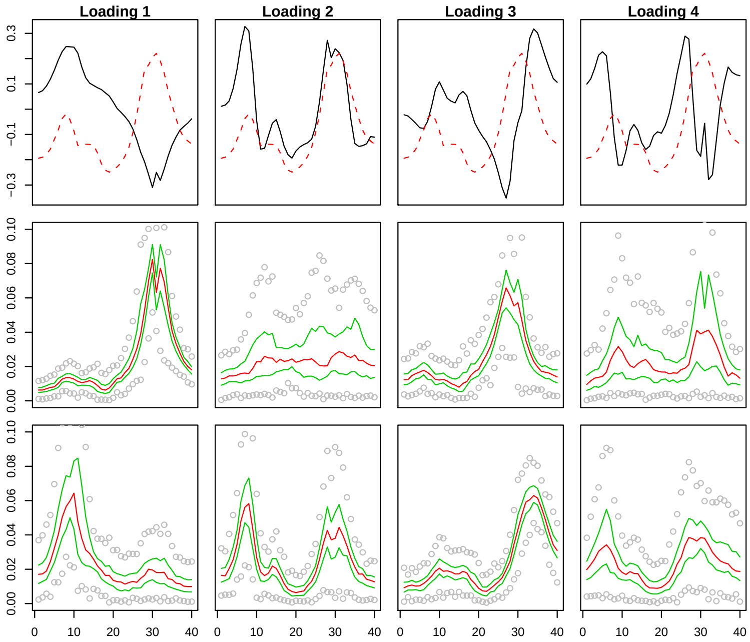


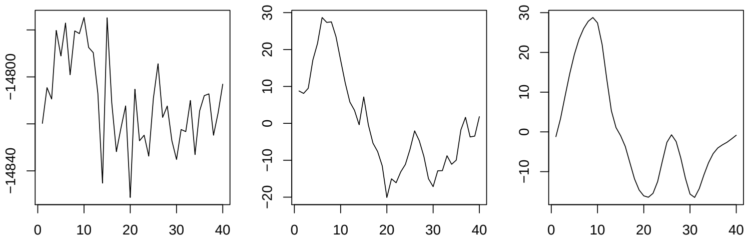
Figure 2.5.: First row: first four PLS loadings (the dashed lines sketch a prototypical spectrum with the three relevant peaks of Cho, Cr and Ci). Last two rows: median (red), hinges (green) and extreme points (circles) of the 5% of the training sample which score highest/lowest for the respective PLS loading.

nificantly better than the ratio rule (Cho+Cr)/Ci (q) (p = .0034). Comparing quantification based (SVM-rbf (q)) and PR methods (SVM-rbf (m)/GP-rbf (m)), the performance gain could still be considered significant (p = .0269/.0261). The performances of the linear PR approach PLS/LR (m) compared with SVM-rbf (q) and (Cho+Cr)/Ci (q), however, are statistically indistinguishable (p = .1531/.5966).

2.3.5. Interpretation of PR approaches.

PLS loadings obtained from the whole prostate dataset are presented in Fig. 2.5. The first row shows the L_2-normalized loadings along with a typical spectrum (dashed line). The last two rows show statistics (median, hinges and extreme points) of the upper/lower 5% of the training sample, sorted according to their PLS scores. This reveals spectral patterns which score high/low for the respective PLS loading and facilitates their interpretation.

Fig. 2.6 contrasts coefficient profiles obtained from three linear classifiers trained on the whole data set. Fig. 2.6a shows the coefficients obtained with unregularized LR



(a) Logistic Regression (b) LR with PLS (c) Logistic PSR

Figure 2.6.: Comparison of coefficient profiles learned with logistic regression models. The unregularized model (a) does not show a pattern whereas the PSR model (c) seems to oversmooth slightly. In contrast, the PLS model (b) shows a clear pattern and also preserves the details.

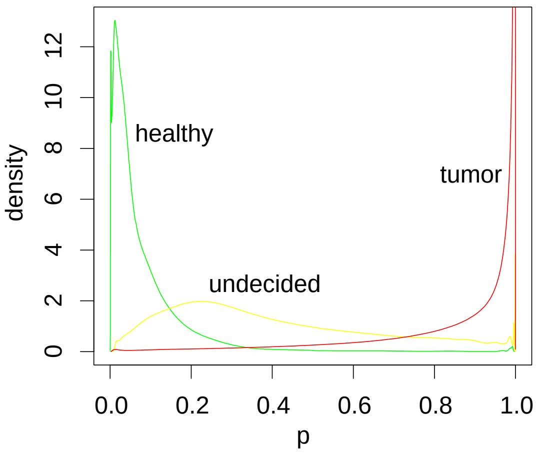
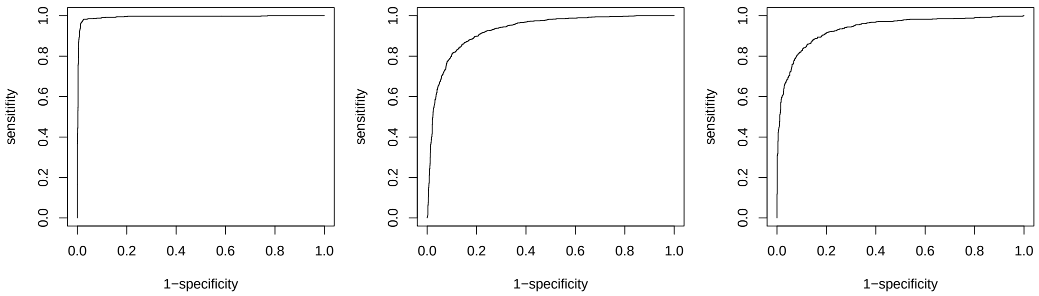


Figure 2.7.: Density estimates of the cross-validated tumor probability estimates for each of the classes in the training set (green=healthy, yellow=undecided, red=tumor).

on all 40 channels, Fig. 2.6b with LR on PLS scores and Fig. 2.6c shows the result for PSR.

2.3.6. Probability Estimates

Fig. 2.7 shows density estimates of the cross-validated probability estimates obtained with the binomial PLS method. The three classes are nicely separated and yield excellent ROC curves which are depicted in Figs. 2.9(a) to 2.9(c).



(a) ROC healthy/tumor (b) ROC healthy/undecided (c) ROC undecided/tumor

Figure 2.8.: Receiver Operator Characteristic curves for the binomial PLS classifier on the prostate tumor data.

In Fig. 2.9 the sigmoidal tumor probability estimate of the binomial PLS model is shown in its projection onto the first two PLS loadings (actually, four loadings have been used in the binomial PLS model, however, they cannot all be depicted simultaneously). Furthermore, some randomly drawn examples from the training set are displayed together with their 2σ-confidence intervals and their true class which is color-coded. It can be observed that the size of the confidence intervals depends not only on the predicted probability, but also on the quantity of proximate training examples which support the probability estimate.

2.3.7. CLARET

Only their simple availability and efficient accessibility will allow pattern recognition algorithms to be employed in clinical studies and routine use. Therefore, the CLARET tool (*C*SI-based *L*ocalization *A*nd *R*obust *E*stimation of *T*umor probability) has been developed for the diagnostic evaluation of MRSI data. CLARET implements pattern recognition methods as described above and allows for an automatic evaluation of MRSI volumes.

The evaluation of MRSI volumes with CLARET is designed for utmost user friendliness. After selecting an MRSI volume and a suitable MR image volume (usually T_2-weighted) from the DICOM data set, CLARET can be initiated to evaluate either individual slices or the whole loaded volume at once. The results are displayed in transparent probability maps superposed onto slices through the MRI volume (Fig. 2.10).

One can easily switch between tumor probability estimates and their 2σ confidence intervals. In addition, voxels which cannot be evaluated are masked out. In the

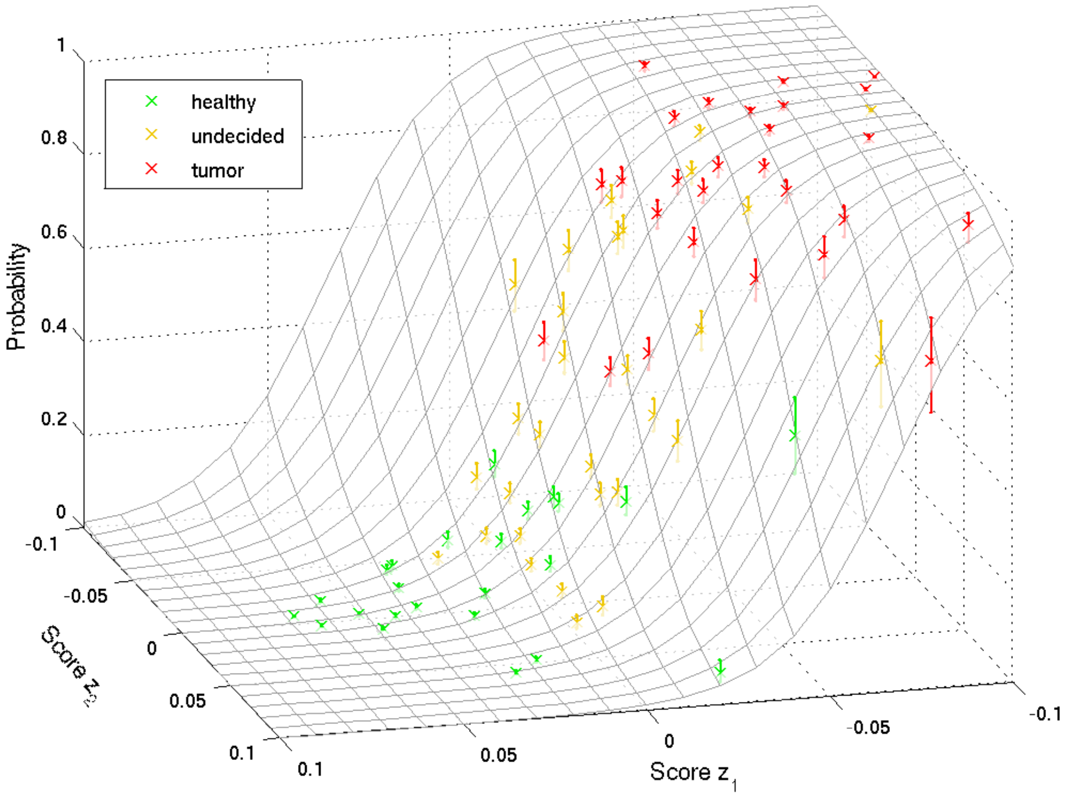
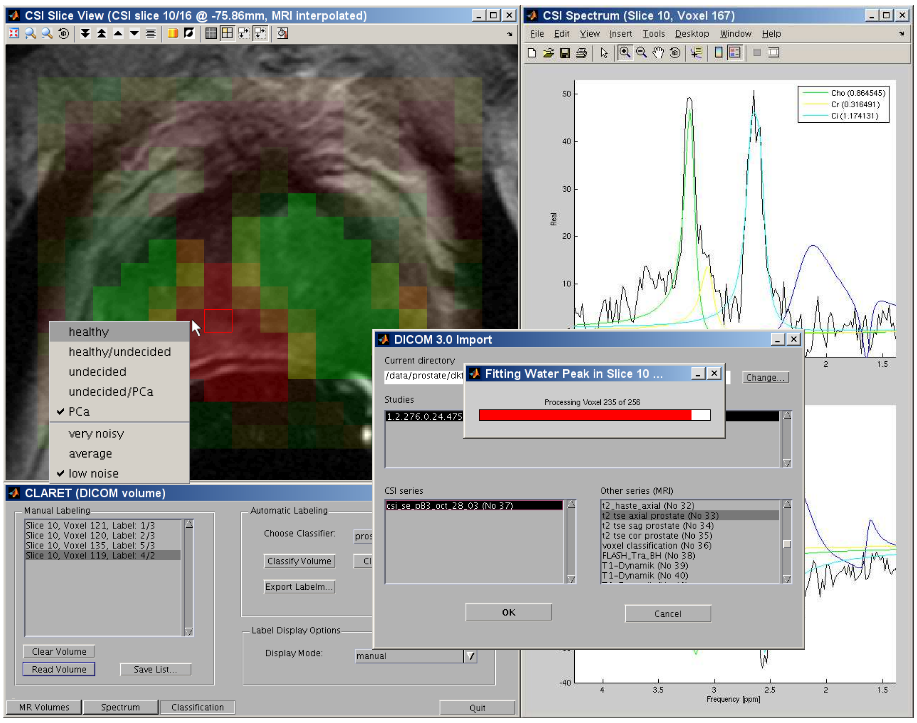


Figure 2.9.: Projection of training examples onto the first two PLS loadings and their tumor probability estimates with 2σ-confidence intervals. The true class labels are color-coded. The more training examples are located in the vicinity and the more extreme a probability estimate is, the smaller the confidence intervals get.

---

subsequent diagnosis the user can therefore concentrate only on regions marked suspicious in the probability map. In case of doubt, the original spectral signal is easily accessible and conspicuities in the T_2-image can also be scrutinized. Finally the extracted probability map can be stored in a file together with the analyzed MRI/MRSI volumes for later reference or it can be exported for use in the radiation planning software VIRTUOS [25]. If no user-interaction is needed at all, an automatic evaluation of MRSI data with CLARET can be initiated from within VIRTUOS directly (cf. Fig. 2.11).

CLARET has explicitly been designed for the application of pattern recognition methods. Therefore, it can also be used for the construction of training data sets. The automatic display of the respective spectral signals together with fitted model spectra and quantification results upon selection of an MRSI voxel allow for a semi-



Current directory: /data/prostate/dkf

Processing Voxel 235 of 256

Studies: 1.2.276.0.24.475

CSI series: csi_se_pB3_oct_28_03 (No 37)

Other series (MRI):
t2_haste_axial (No 32)
t2 tse axial prostate (No 33)
t2 tse sag prostate (No 34)
t2 tse cor prostate (No 35)
voxel classification (No 36)
FLASH_Tra_BH (No 38)
T1-Dynamik (No 39)
T1-Dynamik (No 40)

Slice 10, Voxel 121, Label: 1/3
Slice 10, Voxel 120, Label: 2/3
Slice 10, Voxel 135, Label: 5/3
Slice 10, Voxel 119, Label: 4/2

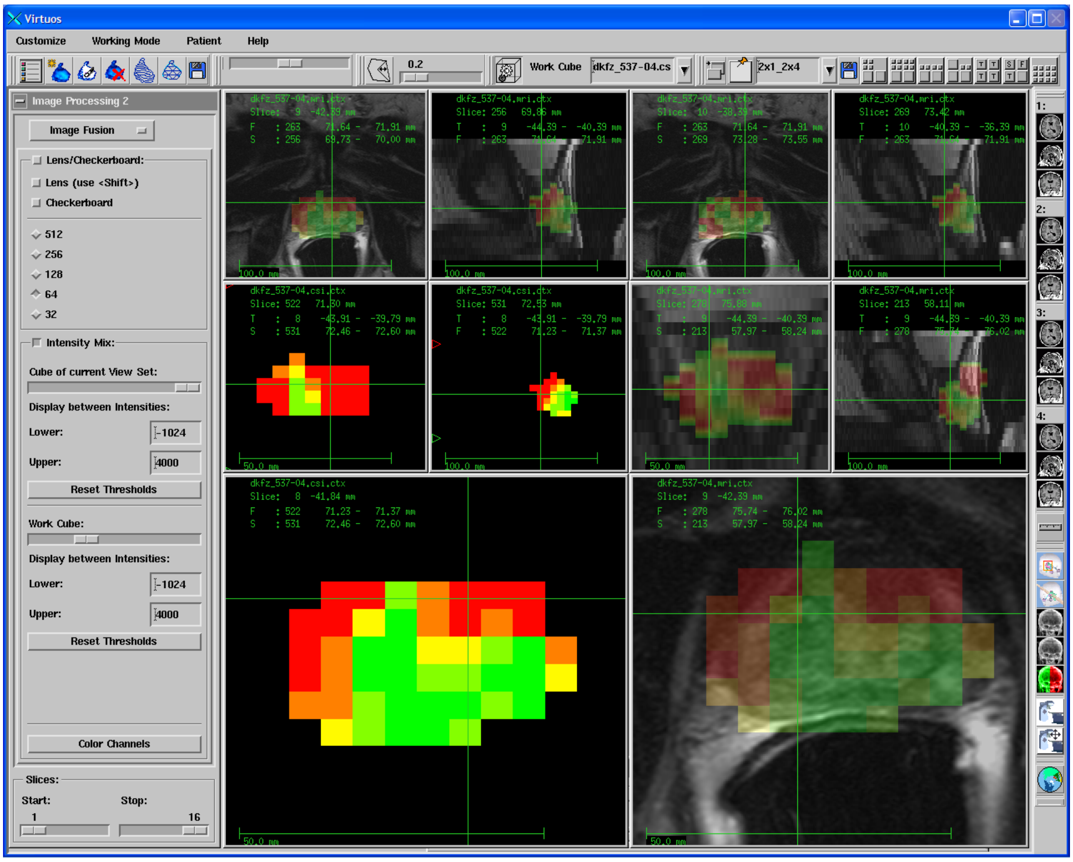
Choose Classifier: pros

Display Mode: manual

Figure 2.10.: The CLARET GUI can be used to evaluate MRSI efficiently. In routine use, the program automatically computes and displays tumor probability maps and confidence intervals on top of morphologic MR images. The program also allows for a point-and-click display of spectral raw data, it can perform quantification, and it may be used for the manual labeling or the semi-manual refinement of training data sets.

manual evaluation of the spectral data. The results from such a voxel-wise evaluation can easily be entered per mouse click in the probability map and stored as training data set. Since manual labels can also be entered after an automatic evaluation CLARET is also suitable for the correction of classification errors and is ready for active learning.

Fig. 2.12 shows the color-coded probability map obtained with CLARET using PLS/LR. Next to it, results from a histologic step-section examination are shown. It should be noted that the slice planes obtained from histologic examinations and MRSI are unlikely to coincide exactly. Also, since the histologic samples easily deform after radical prostatectomy, only qualitative comparisons are possible.



Work Cube dkfz_537-04.cs

2x1_2x4

Cube of current View Set:

Display between Intensities:

Lower: -1024

Upper: 4000

Work Cube:

Display between Intensities:

Lower: -1024

Upper: 4000

Slices:

Start: 1 Stop: 16

Figure 2.11.: Integration of CLARET in the software platform VIRTUOS (dkfz, Heidelberg). Tumor probability maps are generated automatically from MRSI and can, in addition to other imaging modalities, be used for radiation planning.

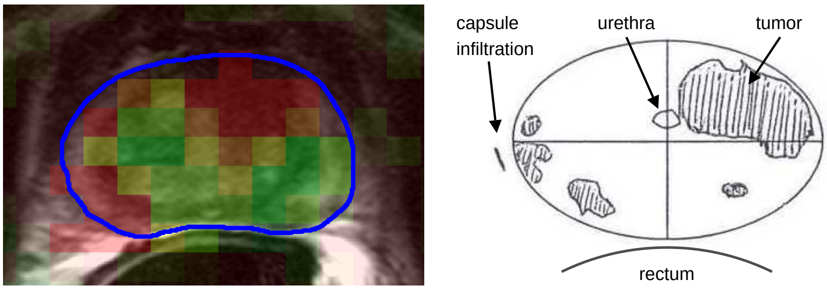


Figure 2.12.: Tumor probability map estimated with logistic regression based on partial least squares scores (PLS/LR) and histologic step-section result for the same slice. Up to minor deformations, the evaluated *in vivo* MRSI agrees very well with the histopathology.

2.4. Discussion

2.4.1. Spectral Preprocessing

For PR, two spectral representations have been employed in the present study, namely real and magnitude spectra. As opposed to real spectra, magnitude spectra are invariant w.r.t. zero-order phase shifts. The additional variation in the spectral pattern caused by phasing problems mainly degrades the performance of PCA/LR and ICA/LR (Fig. 2.4c). Although the difference to magnitude-based methods is smaller for other linear and nonlinear classifiers, magnitude spectra consistently yield better results. Improvements with real spectra might be obtainable when using more sophisticated automatic phasing algorithms, however, these might also be prone to similar robustness problems as quantification algorithms. It seems that the advantage obtained from omitting phase correction in magnitude spectra outweighs the disadvantage of increased line widths and peak overlap for prostate MRSI. Other studies present analogous results for brain MRS [61, 193, 123, 7].

It has also previously been found that some kind of normalization (L_1, L_2, L_∞) of the spectral patterns is important [7]. Experiments with L_1- and L_2-normalized prostate spectra did not yield very different results (not shown here). In contrast to [61, 193, 123] L_1-normalized spectra have been used because of the notable relationship to metabolite ratios. The L_1-norm can be regarded as an approximation to the integrated spectrum and corresponds to Cho + Cr + Ci in the prostate. Hence, linear combinations of the derived spectral features are similar to the ratio $r_2 = \mathrm{Ci}/(\mathrm{Cho} + \mathrm{Cr} + \mathrm{Ci})$ which is related to the usual ratio $r_1 = (\mathrm{Cho} + \mathrm{Cr})/\mathrm{Ci}$ by the monotonous transformation $r_2 = (r_1 + 1)^{-1}$. Therefore, r_1 and r_2 must have the same discriminating power which is confirmed by the results in Fig. 2.4a. Hence, L_1-normalization addresses the problem that absolute line intensities in MR spectra are unreliable and information is only contained in their ratio.

2.4.2. Quantification-based approaches

Despite only subtle mathematical differences in performing the quantification with VARPRO, AMARES or QUEST (with simulated metabolite templates), the classification results differ considerably (Fig. 2.4a). All quantification methods have been employed with the same number of Lorentzian shaped components but with slightly different constraints. Hence, the employed prior knowledge has considerable influence on the obtainable classification performance.

Implementation details of the employed algorithm also seem to matter. The superior performance of our VARPRO approach in comparison to AMARES might be surprising at first. However, deviating from the original VARPRO approach, which uses a modified Levenberg-Marquardt algorithm [204], an interior trust region algorithm [52] that appears to cope better with the variable projection functional has been used. The excellent performance of QUEST, on the other hand, might be due to its implicit baseline correction [160].

Most of the differences between quantification based approaches are due to choosing different quantification methods and not due to using different classifiers (Fig. 2.4a). However, none of the classifiers employed on quantified data could improve over the results obtained with the conventional (Cho + Cr)/Ci ratio. This indicates that the ratio rule is indeed a good approach for the discrimination of tissue classes in the prostate, provided that the quantification results are reliable. However, the latter is difficult to judge in the absence of ground truth.

2.4.3. Subspace Methods

The results listed in the second compartment of Fig. 2.4b show no significant difference between the tested subspace methods. In particular, identical performance is obtained with PCA and ICA. Given that the scores obtained from FastICA are necessarily a linear combination (scaling and rotation) of the PCA scores, this can be explained by noting that LR is invariant w.r.t. such feature transformations. But also NMF cannot improve the AUC. And although PLS and GPLS can increase the lower hinge in the discrimination of healthy and tumor tissue, these effects are not observed in the discrimination against voxels of the "undecided" class.

One reason for the use of subspace methods is that the basis of the constructed subspace is amenable to interpretation. Optimal subspaces along with the most important spectral patterns are automatically determined based on *in vivo* data. Therefore, not only protocol and metabolite dependent features of the signal but also the *in vivo* situation is considered. Furthermore, for PLS also the classification task at hand has an influence on the choice of the subspace. This distinguishes subspace from quantification approaches which either use theoretical models or metabolite templates derived from *in vitro* measurements.

The PLS loadings derived for the prostate data allow for a consistent interpretation (cf. Fig. 2.5). As published in various clinical studies on prostate spectroscopy (*e.g.* [173, 175, 149]), the ratio between Cho + Cr and Ci is the most important feature in discriminating cancerous from healthy tissue. Together with the L_1-normalization, the first loading clearly reflects this ratio. The second loading rewards high Cho-to-Cr ratios in the presence of a clear Ci peak. Spectra with small line widths and clear peaks for all metabolites get a high score on this component if Cho is elevated in comparison to Cr. This criterion is in accordance with medical studies, for example [149], where the Cho/Cr ratio has also been considered. The third loading reflects frequency shifts of the citrate peak and the fourth loading frequency shifts of the choline peak. Both these and higher order loadings are not relevant for tumor classification.

2.4.4. Linear and Nonlinear Classifiers

Fig. 2.6 demonstrates that subspace methods act as a regularizer which helps to overcome the problem of collinearities in spectral data. The unregularized model as applied to the highly correlated spectral channels yields a very rough coefficient profile with high offset (Fig. 2.6a). As opposed to that, the coefficient profile of the PLS model in Fig. 2.6b takes small values around zero and shows a clear pattern resulting from a linear combination only of the first four loadings. A similar profile is obtained for the logistic PSR model in Fig. 2.6c for which the coefficient profile has explicitly been modeled as a smooth spline function. Anyhow, the regularizing influence is not reflected in a clear performance gain (Fig. 2.4).

Increased performance is obtained when using nonlinear PR approaches. A SVM with linear kernel is a linear classifier and performs no better than its cognates. As evidenced by Fig. 2.4b, an improvement is obtained only when switching to the nonlinear RBF kernel. A significant improvement from using nonlinear classifiers can also be observed in Tab. 2.3. The performance of the RF and GP methods are very similar, indicating that some nonlinearity is indeed present in the prostate tumor classification task. However, an interpretation of the decision rules of a nonlinear classifier remains difficult. Significant differences between the nonlinear classifiers could not be observed.

Nonlinear classifiers can manifest their superiority only when applied to spectral patterns. In view of Tab. 2.3 and Fig. 2.4, if enough data is available and a nonlinear "black-box" method is acceptable, there remains little reason to use quantification for feature extraction.

Finally, diagnostic maps such as shown in Fig. 2.12 allow for a time-efficient evaluation of NMR spectroscopic images. In this example, the spectral patterns obtained with MRSI and the histopathological ground truth agree very well. Further clinical validation is certainly required and is attempted in [19].

2.4.5. CLARET

For the first time a software tool is available which generates pathophysiologic probability maps from MRSI data fully automatically. CLARET is currently employed in a prostate study at the German cancer research center Heidelberg (dkfz). The graphical user interface and integrated workflow allow for an efficient evaluation of MRSI. Direct import of DICOM data from the MR scanner and the subsequent fully automatic evaluation by means of powerful pattern recognition algorithms make its use simple. An application of CLARET for radiation therapy planning is enabled by the integration into the software platform VIRTUOS [25].

CLARET prototypically demonstrates the possibilities of a pattern recognition based MRSI evaluation. Here CLARET clearly contrasts with other MRSI evaluation tools such as jMRUI, LCModel or PRISMA which concentrate on the quantification of spectral data. Although these programs can also be used to compute and visualize color maps, only relative metabolite concentrations or ratios thereof are displayed. In contrast, CLARET is tailored towards the generation and visualization of pathophysiologic maps that report an explicit estimate for tumor probability in every voxel.

2.5. Summary

In this chapter, different single-voxel approaches to the automated estimation of tumor probability in ^{1}H NMR spectroscopic images of the prostate have been compared. The emphasis has been put on developing a fully automatic and reliable approach with optimal diagnostic results.

In particular, it was found that quantification based approaches heavily rely on an optimal choice of prior knowledge and on the algorithm used for quantification. In contrast, PR approaches do not require specific prior knowledge and can infer important spectral patterns from the *in vivo* training data automatically. The PR approach attempts to address the diagnostic question – healthy vs. tumorous tissue – directly and can therefore use the full statistical information contained in the raw spectral data.

Among the quantification based approaches, best results have been obtained with classifiers based on metabolite concentrations estimated with QUEST. However, the performance was not superior to the conceptually simple linear PR approaches based on magnitude spectra.

Several subspace methods proposed for spectral MR data before have been compared. In particular, PCA, ICA, NMF and PLS were used. Hardly any difference in performance could be observed between these. Still, methods especially designed for spectral data such as PLS and GPSR seemed to have slight advantages.

If a “black-box” approach is acceptable, superior performance can be obtained by using a suitable nonlinear classifier in conjunction with magnitude spectra.

Pattern recognition algorithms are not supported by available MRSI processing software. Using the CLARET tool introduced above, the proposed pattern recognition approaches can easily be employed by a physician, *e.g.* in clinical studies. It prototypically demonstrates the applicability of pattern recognition methods for clinical routine use.

Chapter 3.

Quantification Using Spatial Context

3.1. Introduction

Quantification based approaches might not be the best choice if one is only interested in the diagnostic classification of a voxel. In contrast, if one is interested in the pathophysiological process and the underlying biochemistry, a quantification is indispensable. The present chapter discusses an approach to perform quantification that exploits spatial prior knowledge in the form of a generalized Gaussian Markov random field (GGMRF) [35].

The general approach builds upon a nonlinear signal model $S_\theta(t_n)$ and a sequence of observed data points y_n for a certain voxel at discrete time points $\{t_n\}_{n=1}^N$. The parameters θ are usually estimated by minimizing the sum of squared residuals (SSR), i.e. by solving the nonlinear least squares (NLS) problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{n=1}^{N} (S_\theta(t_n) - y_n)^2 \tag{3.1}$$

Since this is a very general model that applies to many imaging modalities the approach has not only been applied to the quantification of MRSI but also to the quantification of Dynamic Contrast-Enhanced MR Images (DCE-MRI).

After the introduction of the GGMRF framework a specialized algorithm that efficiently exploits the sparse structure of the given optimization problem is proposed. Experiments and results on MRSI and DCE-MRI are presented and discussed in two subsequent subchapters. A summary repeating the most important points and proposing future enhancements will conclude the chapter.

3.2. GGMRF: A Generalized Gaussian Markov Random Field Prior

The generalized Gaussian Markov random field [35] is a Markov random field [27, 26, 213, 119] with particular compatibility functions (the logarithms of which are known as potentials). Every voxel in the region of interest (ROI) is represented by a site $s \in S$ which is associated with the vector-valued random variable θ_s. Like in the single-voxel case, the observation likelihood is Gaussian, *i.e.* $y_n^s \,|\, \theta_s \sim \mathcal{N}(S_{\theta_s}(t_n),\, \sigma^2)$. Adding the spatial GGMRF prior on the parameter maps θ yields a joint distribution over y and θ in form of the Gibbs distribution:

$$\Pr(\theta, y) = \frac{1}{Z} \prod_{s \sim t} \Psi(\theta_s, \theta_t) \prod_s \Phi(\theta_s, y_s) \tag{3.2}$$

where y and θ are vector variables obtained by stacking the site vector variables y_s and θ_s. Z is the global normalizer (partition function) and $s \sim t$ denotes pairs of neighboring sites according to the employed neighborhood system. The compatibility functions in $\Phi(\theta_s, y_s)$ and $\Psi(\theta_s, \theta_t)$ are defined by the potentials

$$\log \Phi(\theta_s, y_s) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (S_{\theta_s}(t_n) - y_n^s)^2 \tag{3.3}$$

$$\log \Psi(\theta_s, \theta_t) = -\frac{\alpha_{st}}{2} \, \|W(\theta_s - \theta_t)\|_p^p \tag{3.4}$$

where $1 < p \leq 2$ and $\alpha_{st} \geq 0$ are hyper-parameters determining smoothness properties of the sought parameter maps. W is a diagonal weighting matrix which accounts for the different scales and variability of the parameters in θ_s and can be used to adjust the smoothness of individual parameter maps.

The application of a GGMRF allows to vary continuously between a smoothing Gaussian MRF prior ($p = 2$) and an edge-preserving MRF ($p \to 1$) with properties comparable to a weighted median filter [35]. Furthermore, the GGMRF potential defined by (3.4) is convex and, as opposed to robust alternatives such as the Huber potential [91], it does not have a threshold parameter at which its behavior suddenly changes.

In the present work, only regular lattices will be of concern. In principle, a hierarchy of neighborhood systems can be defined on these, starting with four nearest neighbors (first order), to eight nearest neighbors (second order), and so on [23]. In this chapter, however, only first and second order neighborhoods will be considered. The graphical model of the described GGMRF with a first order neighborhood system is depicted in Fig. 3.1.

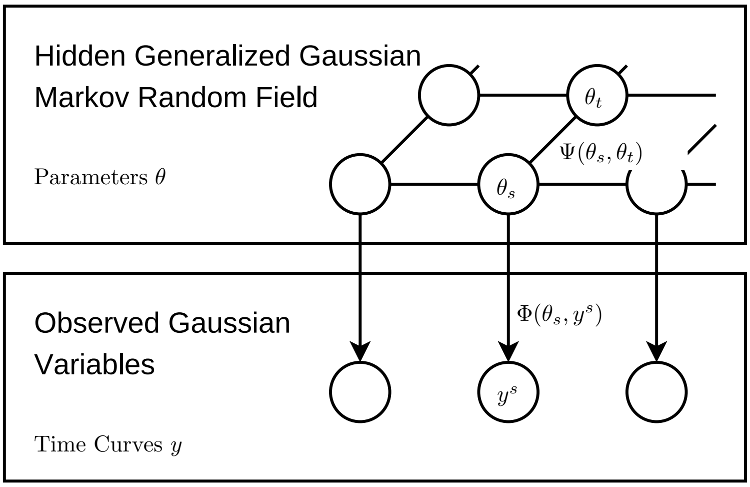

Hidden Generalized Gaussian Markov Random Field
Parameters θ
θ_t
$\Psi(\theta_s, \theta_t)$
θ_s
Observed Gaussian Variables
$\Phi(\theta_s, y^s)$
y^s
Time Curves y

Figure 3.1.: Graphical representation of the GGMRF lattice model with first order neighborhood system (4 neighbors) and nonlinear observation potentials $\Phi(\theta_s, y_s)$.

3.3. MAP Estimation with Block-ICM

Given an observed sequence of data points $\{y_n^s\}$, the maximum a posteriori (MAP) estimate $\hat{\theta}$ is found by minimizing the data term (SSR) and an additional spatial coupling term:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\sum_{s \in V} \sum_{n=1}^{N} (f_{\theta_s}(t_n) - y_n^s)^2 + \sigma^2 \sum_{s \sim t} \alpha_{st} \left\| W(\theta_s - \theta_t) \right\|_p^p \right] \tag{3.5}$$

For realistically sized images, the parameter vector θ gets extremely high-dimensional which makes the optimization problem (3.5) very hard to solve with standard NLS algorithms. However, the problem is sparse in the sense that most of the θ_s are not directly coupled. The MRF framework provides special algorithms which can exploit this sparsity such as the ICM (iterated conditional modes) algorithm [27].

Here, a generalized ICM algorithm is proposed which will be shown to converge faster than the standard ICM approach. As the algorithm considers collections of sites instead of single sites at each step, this approach is named *block-ICM*. Given an arbitrary subset of sites $\tilde{V} \subseteq S$, it follows from the Hammersley-Clifford theorem [213] that the posterior distribution $\Pr(\theta \,|\, y) = \Pr(\theta_V \,|\, y)$ can be factored as

$$\Pr(\theta_V \,|\, y) = \Pr(\theta_{\tilde{V}} \,|\, \theta_{\partial \tilde{V}}, y) \Pr(\theta_{V \setminus \tilde{V}} \,|\, y) \tag{3.6}$$

where $\partial \tilde{V} = \{s \,|\, t \sim s \,\wedge\, t \in \tilde{V} \,\wedge\, s \in V \setminus \tilde{V}\}$ is the border of $\tilde{V}$. Increasing the first factor $\Pr(\theta_{\tilde{V}} \,|\, \theta_{\partial \tilde{V}}, y)$ with respect to $\theta_{\tilde{V}}$ certainly cannot decrease $\Pr(\theta_V \,|\, y)$ since the second factor $\Pr(\theta_{V\setminus\tilde{V}}\,|\,y)$ does not depend on any of the variables in $\theta_{\tilde{V}}$. Hence, the MAP problem (3.5) can be solved iteratively by solving a series of smaller MAP problems over subsets of sites

$$\theta_{\tilde{V}}^{(k+1)} = \operatorname*{argmin}_{\theta_{\tilde{V}}} \Bigg[\sum_{s\in\tilde{V}} \sum_{n=1}^{N} (f_{\theta_s}(t_n) - y_n^s)^2 + \sigma^2 \sum_{\substack{s\sim t\\ s,t\in\tilde{V}}} \alpha_{st} \left\| W(\theta_s - \theta_t) \right\|_p^p + \sigma^2 \sum_{\substack{s\sim t\\ s\in\tilde{V}\\ t\in\partial\tilde{V}}} \alpha_{st} \left\| W\left(\theta_s - \theta_t^{(k)}\right) \right\|_p^p \Bigg] \tag{3.7}$$

The block-ICM algorithm can also be viewed as an iterative coordinate descent approach where the potentially intersecting subsets $\tilde{V}^{(k)}$ redefine generalized coordinates $\theta_{\tilde{V}}^{(k)}$ in every descent step. Also, it suffices to find a realization $\theta_{\tilde{V}}^{(k+1)}$ which decreases the objective (3.7) instead of finding the exact minimum in every descent step. The proposed procedure still converges to a local minimum.

Shape, size and update sequence of the subsets $\tilde{V}$ are design parameters of the block-ICM algorithm and should be chosen so as to trade off the problem size in each step against the number of sweeps required for convergence. If, *e.g.*, each of the subsets $\tilde{V}^{(k)}$ contains only one site s, the standard ICM algorithm is recovered [27] which is known to often converge rather slowly. If, contrariwise, only one (sub)set $\tilde{V} \equiv S$ is chosen, the complete MAP problem (3.5) which contains all variables is obtained. Hence, small subsets of sites should be chosen depending on the size of the local neighborhood and the strength of the mutual influence. Because of the locality of this influence, the size of the subsets does not have to be increased with growing lattices, yielding an algorithm which scales linearly with the number of sites. In the following experiments a fixed update schedule as sketched in Fig. 3.2 with different block sizes has been employed.

3.4. Quantification of Magnetic Resonance Spectroscopic Images

3.4.1. Introduction

Accurate quantification is a crucial prerequisite for the study of *in vivo* metabolisms by means of magnetic resonance spectroscopy (MRS). It may help the noninvasive diagnosis and characterization of pathophysiological changes and thus be an important tool for clinical research. The reliable quantification of MR spectroscopic signals

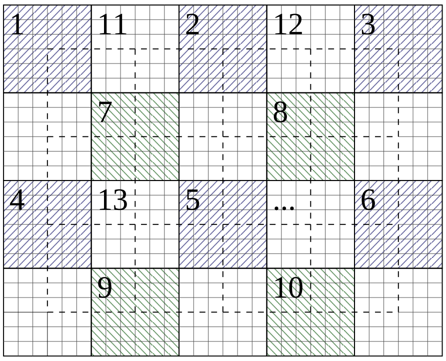


Figure 3.2.: Blocks and update schedule used for the block-ICM algorithm. In every odd sweep, square blocks of 6 × 6 sites are visited following the pattern indicated by the numbering. The even sweeps are performed in the same way but shifted by 3 sites (dashed squares).

depends on stable approaches to spectral fitting which is a challenging problem due to low signal-to-noise ratios (SNR), artifacts, overlapping peaks and baselines present in MRS data.

A common approach to improve the quality of spectral fits is the use of prior knowledge. For example, various forms of constraints on the model parameters can be considered using the AMARES algorithm [204]. Theoretical arguments that assure a decrease of the Cramér-Rao lower bound for hard (equality) constraints [45] as well as empirical evidence (*e.g.* [132]) have proven the usefulness of employing prior knowledge as much as possible.

Also algorithms such as LCModel [156], QUEST [160] and AQSES [185, 183] employ prior knowledge by using metabolite templates which implicitly impose hard parameter constraints on the signal components from each metabolite. Furthemore, all these algorithms provide some means for handling baselines stemming from macromolecules [185].

Although current endeavors in MR spectroscopy aim at recording *high resolution* spectroscopic images all MRS quantification algorithms work on a per-voxel basis. Spatial prior knowledge in form of *smoothness assumptions* for certain quantified parameters of the nonlinear signal model can be valuable and should be exploited.

In the present work the application of *spatial prior knowledge* for spectral fitting in MRS images is proposed. Using a Gaussian Markov random field prior, smoothness of selected parameter maps can be encouraged. Spatial prior knowledge is applied in addition to the commonly employed per-voxel prior knowledge. Although, the problem of baselines is not considered in the presented experiments, the proposed framework allows to add a nonparametric baseline model as proposed in [183].

3.4.2. Spectral Fitting of Magnetic Resonance Spectra

Most quantification algorithms fit a nonlinear signal model $S_\theta(t_n)$ to the observed spectral data y_n, either in time-domain [204, 160, 185] or in frequency-domain [156], by minimizing the sum of squared residuals

$$l(\theta) \quad = \quad \sum_{n=1}^{N}(S_\theta(t_n) - y_n)^2 \quad \text{with } t_n = n\Delta t \tag{3.8}$$

which is the maximum likelihood solution under the assumption of additive white Gaussian noise.

Similar to QUEST [160] and AQSES [185], the free induction decay (FID) signal is modeled as a linear combination of M possibly damped, phase- and frequency-shifted metabolite templates $T_m(t_n)$:

$$S_\theta(t_n) = \mathrm{e}^{j\phi_0} \sum_{m=1}^{M} T_m(t_n)\ a_m \mathrm{e}^{(j2\pi\Delta f_m - \Delta d_m)t_n} \tag{3.9}$$

with the imaginary unit $j = \sqrt{-1}$. Thus, the parameter vector θ contains a common phase correction ϕ_0 and for each of the M metabolites an amplitude a_m, frequency shift Δf_m and damping Δd_m. Both, the Δf_m and Δd_m are initialized to zero whereas initial guesses for a_m and ϕ_0 are obtained by linear least squares. Furthemore, each metabolite template follows the K-component Lorentzian model

$$T_m(t_n) \quad = \quad \sum_{k=1}^{K^{(m)}} a_k^{(m)} \mathrm{e}^{j\phi_k^{(m)}}\, \mathrm{e}^{(j2\pi f_k^{(m)} - d_k^{(m)})t_n} \tag{3.10}$$

where $f_k^{(m)}$ are the frequencies, $d_k^{(m)}$ the dampings, $\phi_k^{(m)}$ the phases and $a_k^{(m)}$ are the amplitudes of the metabolite components. In light of the signal model $S_\theta(t_n)$, only relative values are important here and encode available prior knowledge. The absolute values for $f_k^{(m)}$ and $d_k^{(m)}$ are chosen as to provide good initial guesses.

The restriction to metabolite templates of the form in Eq. (3.10) allows to use the AMARES algorithm [204] which is available and can be parameterized through the jMRUI software [139].

Spatial prior knowledge is incorporated using the Bayesian approach of specifying a *prior distribution* over the sought parameter maps [26, 49]. The maximum likelihood point estimate is then replaced by the mode of the posterior distribution. Here, a Gaussian random field is used to model a prior distribution that favors smooth parameter maps [10]. The optimization objective thus becomes

$$l(\Theta) \quad = \quad \sum_{s} \sum_{n=1}^{N} (S_{\theta_s}(t_n) - y_n^s)^2 + \sigma^2 \sum_{s \sim t} ||\theta_s - \theta_t||_W^2 \tag{3.11}$$

where Θ now contains the nonlinear parameters θ_s from all MRSI voxels which are indexed by s, σ is the standard deviation of the signal noise and $||a||_W^2 = a^T W a$ is a 2-norm weighted with the diagonal matrix W.

The first term in Eq. (3.11) just builds a sum over the squared residuals from all voxels s. Adding the second term which builds a sum over the squared distance $||\cdot||_W$ between all neighboring voxel pairs $s \sim t$ in the spectral image also encourages smooth solutions. Both σ and W determine the trade-off between fitting the individual signals against smoothing the parameter maps. Furthemore, W can be used to adjust the smoothness force of certain parameter maps individually, and even to turn off smoothing for some of the parameters by using a zero weight.

Without the second term in Eq. (3.11), the parameter vectors θ_s at individual voxels would be independent, the optimization problem would decouple and would yield exactly the same solution as with the single voxel approach in Eq. (3.8).

The optimization resulting from employing the proposed spatial prior is quite challenging and special algorithms that can exploit the sparse structure of the problem have to be used. To this end, a version of an iterated conditional modes (ICM) algorithm [36, 213] has been used, the block-ICM method as proposed in [10].

3.4.3. Experimental Setup

Brain MRSI Data

The proposed approach was applied to two different brain data sets the first of which was randomly selected from a clinical study with brain tumor patients. ^{1}H-MRSI data was acquired at the German Cancer Research Center (dkfz, Heidelberg, Germany) on a 1.5T Magnetom Symphony (Siemens Medical Solutions, Erlangen, Germany) with commercially available PRESS pulse sequences and a standard head coil. The MR spectra were obtained with a double spin-echo sequence with one pulse water signal suppression and long echo time (TR 1000ms, TE 135ms, 512 data

metabolite	$f_k^{(m)}$(Hz)	$d_k^{(m)}$(Hz)	$a_k^{(m)}$(a.u.)	$\phi_k^{(m)}$(rad)
Choline	-94.2	-12	1	0
Creatine	-107.5	-12	1	0
N-acetylaspartate	-171.2	-12	1	0

Table 3.1.: Metabolite templates for long echo brain spectra at 1.5T.

points, slice thickness 15mm, matrix size 16 × 16, FOV 160mm). Only 8 × 8 voxels within the PRESS-selected volume have been fitted.

The second ^{1}H-MRSI data was acquired from a healthy volunteer at the Institute for Biomedical Engineering (IBT, ETH Zürich, Switzerland). MR experiments were performed on a Philips Achieva 3T scanner (Philips Medical Systems, Best, The Netherlands) using a transmit-receive head coil with birdcage design. High-resolution PRESS-localized MRSI data using a in-plane resolution of 3.1mm (32×32 voxel, FOV 100mm) and a slice thickness of 10mm were acquired (TR 1300ms,TE 34ms, 1024 data points). To avoid foldover artifacts and therefore enable the reduced FOV, six saturation bands based on RF pulses with polynomial phase response (PPR, [180]) were applied according to the principle discussed in Henning *et al.* [88] to reach T_1 and B_1 insensitive outer volume suppression (OVS) prior to MRSI encoding. To avoid rephasing, amplitude and direction of spoiling gradients were periodically modulated following sinusoidal envelope functions in all three spatial dimensions that are shifted against each other [50, 82]. For water suppression two CHESS [83] pulses were applied prior to the OVS pulses.

Simulation Study

In comparing the single voxel and the GMRF approach it is important to analyze the bias-variance behavior of the two approaches. Since this requires knowledge of ground truth, a Monte Carlo study has been performed. To this end long echo brain MRSI at 1.5T has been simulated using the three metabolite resonances of choline (Cho), creatine (Cr) and N-acetylaspartate NAA with Lorentzian FID and the template parameters as provided in Tab. 3.1. Based on three frequency, damping, and amplitude maps and a phase shift map, noiseless MRSI data has been generated according to the signal model in Eq. (3.9) (dwell time $\Delta t = 1$ms, imaging frequency 64MHz). Then $R = 100$ versions have been simulated by adding isotropic white Gaussian noise (σ_n) to the generated MRSI data.

Two data sets have been generated with different parameter maps for the purpose of emphasizing different effects of a spatial prior:

- **Sharp edges.** First smooth random maps for all parameters have been generated. In the amplitude maps, sharp edges are introduced by adding a constant value to inverted parts of the amplitude images (cf. Fig. 3.3). This data set has been simulated with $N = 256$ data points and noise standard deviation of $\sigma_n = .232$.
- **Overlapping peaks.** Two spatially orthogonal, wedge-shaped amplitude maps for Cho and Cr (cf. Fig. 3.8) along with a smooth random amplitude map for NAA. All frequency and phase shifts have been set to zero, whereas all dampings have been increased by 10Hz (toward the models in Tab. 3.1). This data set has been simulated with $N = 512$ data points and noise standard deviation of $\sigma_n = .155$.

Using the known ground truth parameter ϑ and the R fit results $\hat{\vartheta}_i$ $(i = 1 \dots R)$ obtained for the repeatedly simulated MRSI data, the root-mean-squared error (RMSE) can be calculated in each voxel as

$$\text{RMSE} = \left(\frac{1}{R} \sum_{i=1}^{R} (\hat{\vartheta}_i - \vartheta)^2 \right)^{1/2} . \tag{3.12}$$

Note that this is an estimate for the root of the expected squared error of the estimator $\hat{\vartheta}$ under the data distribution, $\sqrt{\mathrm{E}[(\hat{\vartheta} - \vartheta)^2]}$.

More information can be gained from the decomposition of the RMSE into a bias and a standard deviation term which is provided by a *bias-variance decomposition* [85]:

$$\text{RMSE}^2 = \text{bias}^2 + \text{stdev}^2 \tag{3.13}$$

where

$$\text{bias} = \bar{\vartheta} - \vartheta \tag{3.14}$$

$$\text{stdev} = \left(\frac{1}{R} \sum_{i=1}^{N} (\bar{\vartheta} - \hat{\vartheta}_i)^2 \right)^{1/2} \tag{3.15}$$

with $\bar{\vartheta} = \frac{1}{R} \sum_{i=1}^{R} \hat{\vartheta}_i$.

3.4.4. Results

For the simulated data the noise variance σ^2 which is needed for the GMRF (cf. Eq. (3.11)) approach was known and thus did not need to be estimated. For real data, an estimate for the noise variance has been obtained from the residuals of manually verified single voxel fits.

The diagonal weighting matrix W has been determined once from the single voxel fits of a simulated MRS image ("sharp edges") based on a robust variogram estimate at pixel distance. Throughout the experiments $W_d = .2$ was used for the damping parameters, $W_f = 2$ for the frequencies and $W_{\phi_0} = 20/\pi$ for the phase map. Since it is not desirable to smooth the amplitude maps, W_a has been set to zero. The same weighting matrix has been used on simulated and real data.

Sharp Edges

Figure 3.3 presents results on the first simulated data set ("sharp edges"). The first row shows the ground truth amplitude images and the following rows show amplitude estimates obtained with different fitting approaches for the same realization of the simulated MRS image data. The second row shows results when using the spatial prior (sp, GMRF) whereas the third row shows results when using a single voxel method (sv, AMARES). The results in the last row have been obtained by smoothing the amplitude images obtained from AMARES with a Gaussian filter. Normalized convolution has been employed in order to avoid border effects and to make the linear filtering more similar to the GMRF approach. Furthermore, the Gaussian kernel has been chosen for each metabolite separately so as to minimize the squared error to the ground truth image. Note that such an optimization is certainly not possible given real data. Figure 3.4 shows results based on the same parameter estimates but in difference to the ground truth.

Compared with the single voxel approach, the spatial approach visibly improves the estimate of Cho and Cr. For NAA with its higher amplitudes, the improvements are not so clear but the GMRF prior certainly does not deteriorate the results. The smoothed single voxel solution is considerably better than the single voxel solution and, as judged by Fig. 3.3, also seems to be better than the GMRF solution. However, Fig. 3.4) reveals that posthoc smoothing introduces spatially correlated errors within the homogeneous areas and in particular at the edges which certainly results in estimation bias.

The Monte Carlo study using all $R = 100$ realizations of the "sharp edges" data reveals that the root-mean-squared error of the NAA amplitude estimate actually improves when using the spatial prior (cf. Fig. 3.5). Even more improvement is ob-

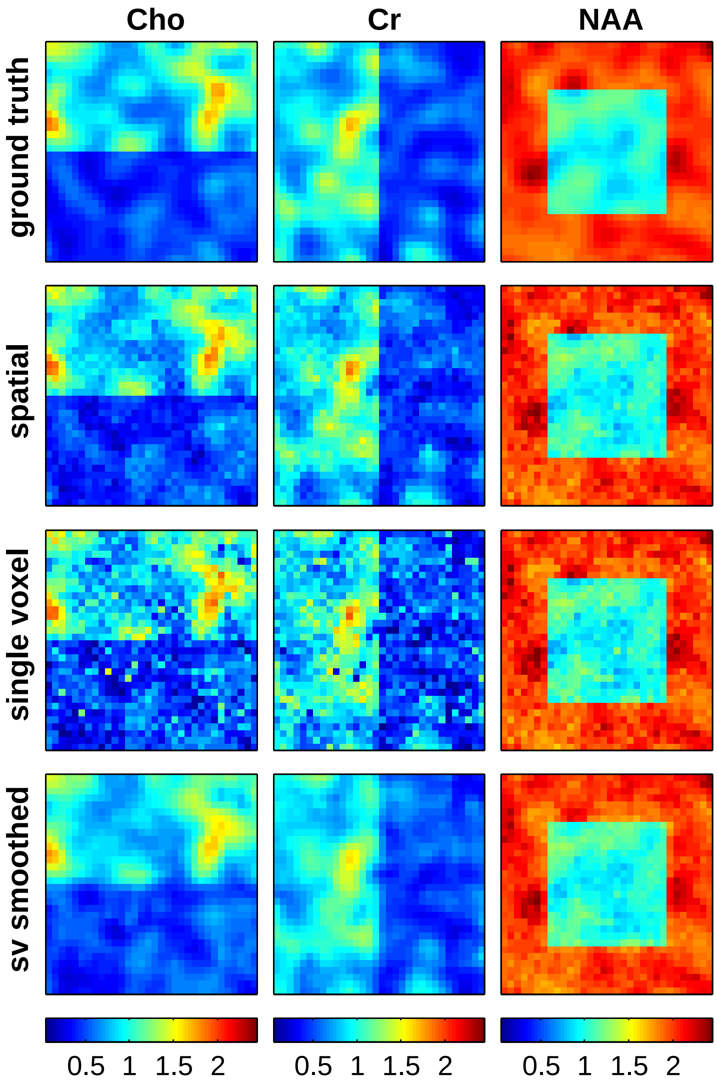


Figure 3.3.: Simulated MRSI data with sharp amplitude edges. From top to bottom: ground truth; GMRF estimate; AMARES estimate; optimally smoothed AMARES estimate. Since with the GMRF no smoothing is performed for the amplitudes, sharp edges are not smeared out. With posthoc smoothing (last row) the estimates of Cho and Cr can be improved at the cost of oversmoothing the edges (cf. Fig. 3.4). Within the homogeneous areas posthoc smoothing seems to provide better results than the other methods which suggests that some smoothing of the amplitudes with the GMRF could be useful. For NAA which has higher amplitudes all methods yield similar results for the example shown.

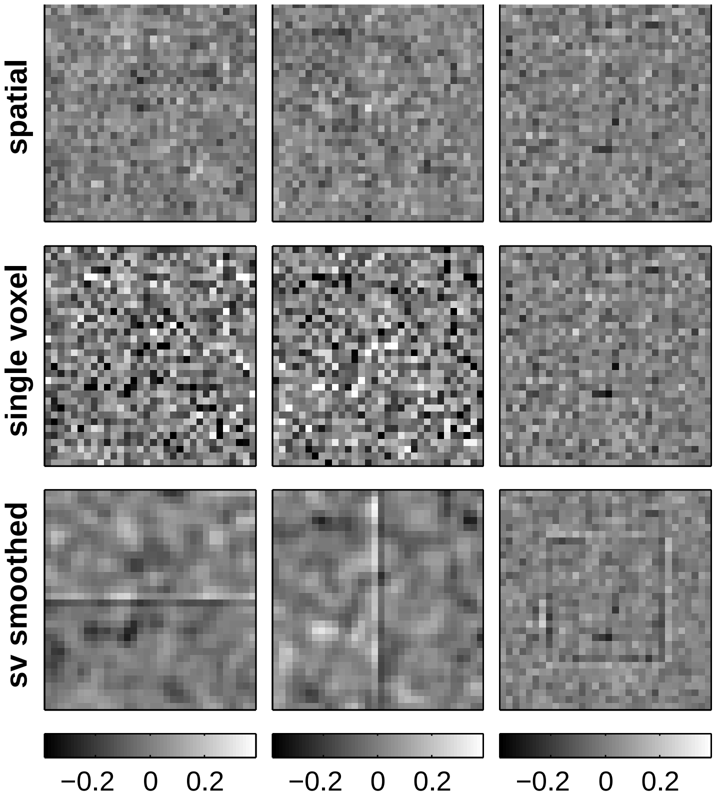
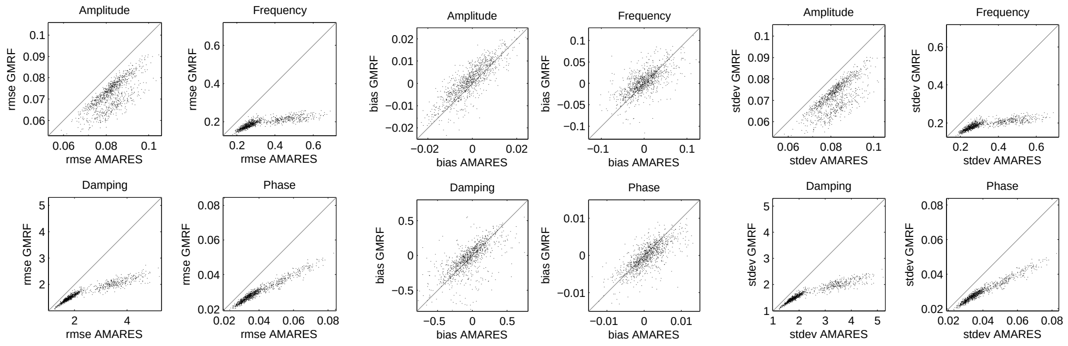


Figure 3.4.: Difference to ground truth images (cf. Fig. 3.3). From top to bottom: difference to GMRF estimate; difference to AMARES; difference to optimally smoothed AMARES. Unlike the GMRF, posthoc smoothing creates spatially correlated error maps and causes systematic errors at the edges.

tained for the remaining parameters. As apparent from the presented bias-variance decomposition of the RMSE in Fig. 3.5, the gain is mainly due to a reduction in standard deviation. Furthermore, the GMRF prior does not seem to introduce significantly more bias than AMARES for most voxels in the “sharp edges” data.

With the available ground truth the exact Cramér-Rao lower bound (CRLB) on the standard deviation of the parameter estimates can be calculated. A comparison of the estimated standard deviations with the corresponding CRLB for all metabolites in the “sharp edges” data is presented in Fig. 3.6. As expected, the unbiased single voxel estimates do not beat the CRLB since all points fall above the diagonal in the scatter plots of Fig. 3.7(a). Occasional points below the diagonal are due to the fact that the standard deviations obtained from the Monte Carlo study are only estimates based on $R = 100$ repetitions. Systematic improvements are obtained with the spatial prior as apparent from Fig. 3.7(b). In line with the observed clear



(a) root-mean-squared error

(b) bias

(c) standard deviation

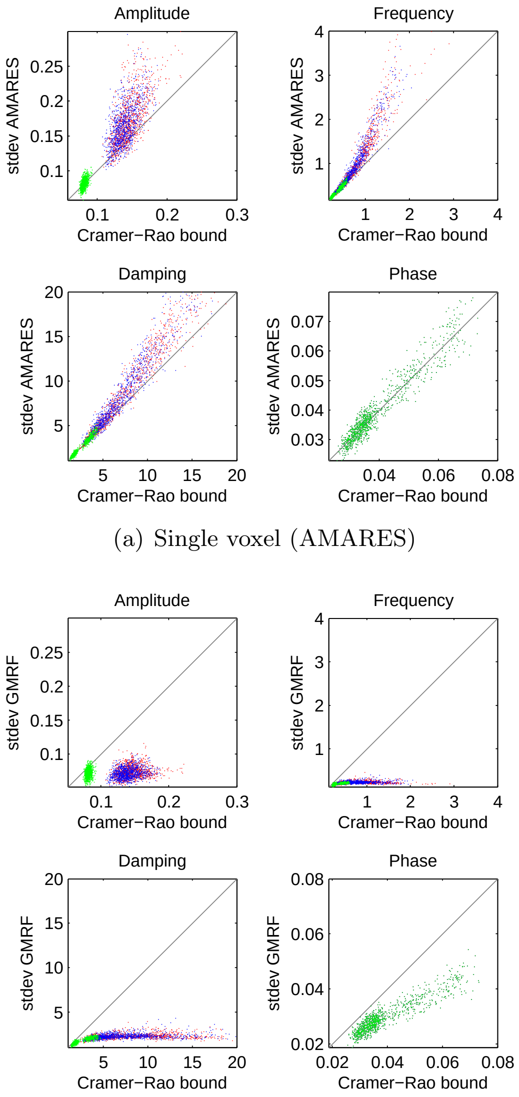
Figure 3.5.: Scatter plots of root-mean-squared error (RMSE), bias and standard deviation (stdev) of NAA amplitudes as estimated from $R = 100$ noisy realizations of the simulated "sharp edges" MRSI data. The GMRF approach yields lower RMSE than AMARES in all parameters. Since the bias is comparable, the gain must be entirely ascribed to a reduction in standard deviation.

improvement of the Cho and Cr difference images in Fig. 3.3, most gain in terms of standard deviation is obtained for these two metabolites (red and blue dots).

Overlapping Peaks

Concentrating on the overlapping peaks of Cho and Cr, the bias-variance analysis on the second simulated data set ("overlapping peaks") confirms that the spatial prior heavily affects the standard deviation of the amplitude estimates which make the main contribution to the RMSE. Surprisingly, however, the single voxel approach also shows more bias than the GMRF approach for small amplitudes at the left and upper border of the images. This effect is even more pronounced if one of the metabolites has a high and the other has a low amplitude at the same time. Then also the standard deviation increases significantly. The GMRF approach does not exhibit such behavior.

An analysis of these border voxels reveals that the observed effect is due to the fact that the single voxel approach sometimes explains both, the Cho and Cr resonance with only one component and sets the amplitude estimate for the smaller metabolite peak to zero as for the example shown in Fig. 3.9. Instead, the GMRF approach uses information from neighboring voxels to infer that an additional small peak is more likely.



(a) Single voxel (AMARES)

(b) Spatial prior (GMRF)

Figure 3.6.: Scatter plots of Cramér-Rao lower bounds (CRLB) against standard deviations (stdev) of the parameter estimates obtained with AMARES and the proposed GMRF approach. The results are based on 100 realizations of simulated MRSI images (32 × 32 voxels) containing the three metabolites Cho (red), Cr (blue) and NAA (green). While the single voxel approach does not beat the the CRLB, using spatial prior knowledge can reduce the estimation variance below the theoretical CRLB.

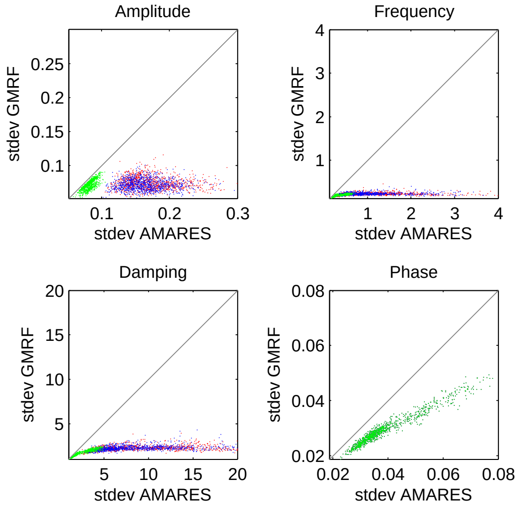


Figure 3.7.: Scatter plot comparing the standard deviations obtained with AMARES and the GMRF. The GMRF can clearly reduce the standard deviation for all model parameters including the amplitudes which are not explicitly regularized.

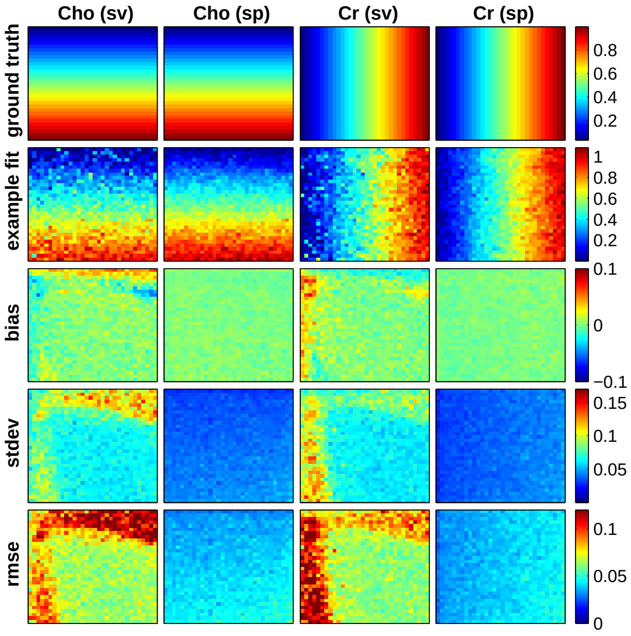
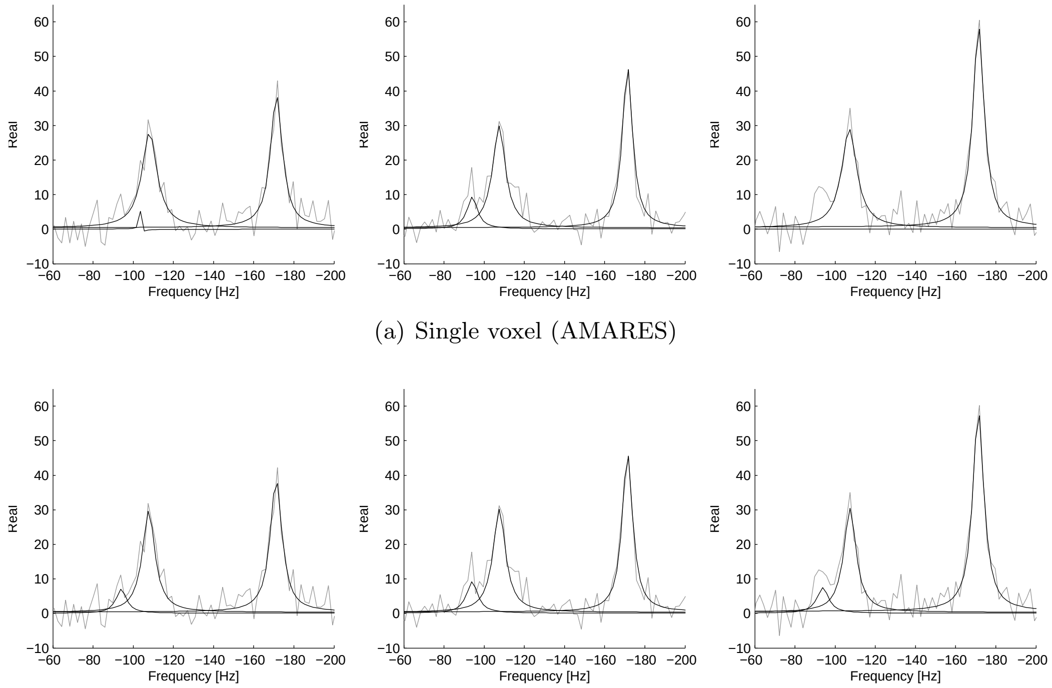


Figure 3.8.: Bias, standard deviation (stdev) and root-mean-squared error (RMSE) for amplitude estimates using the single voxel (sv, AMARES) and spatial (sp, GMRF) approaches when fitting two overlapping peaks: Choline (Cho) and Creatine (Cr).



(a) Single voxel (AMARES)

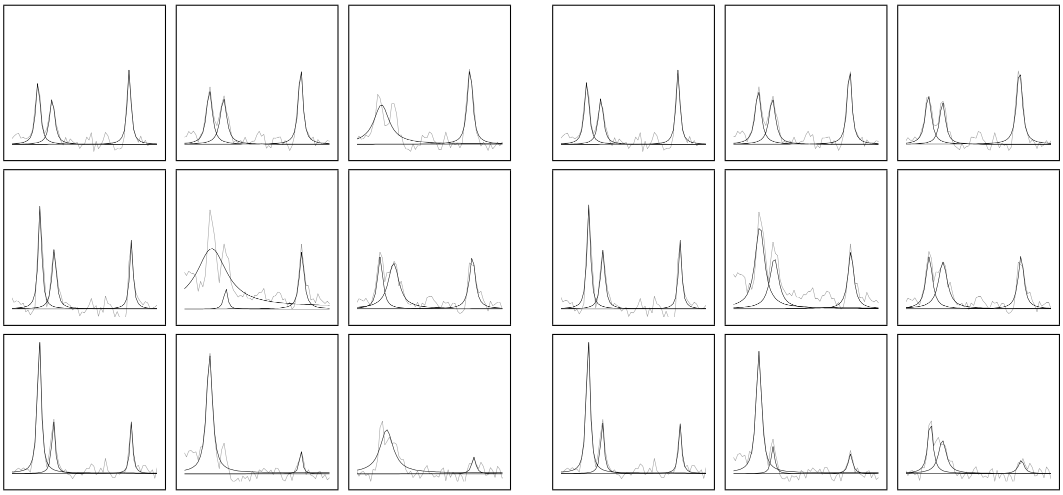
(b) Spatial prior (GMRF)

Figure 3.9.: Three adjacent voxels. For overlapping peaks one of the components could be sufficient to explain the observed signal. In this example, the Cho-component is only fitted with the GMRF approach which uses information from a local neighborhood.

Real data

The described effect can also be observed in real data. Figure 3.10 shows nine adjacent voxels from the 1.5T brain data set which shows clear Cho, Cr and NAA peaks. In four of the nine voxels the single voxel fit explains Cho and Cr with only one resonance component. By requiring that the damping (line width) of the resonance lines should not change rapidly in neighboring voxels the GMRF prior can resolve the two peaks in all voxels.

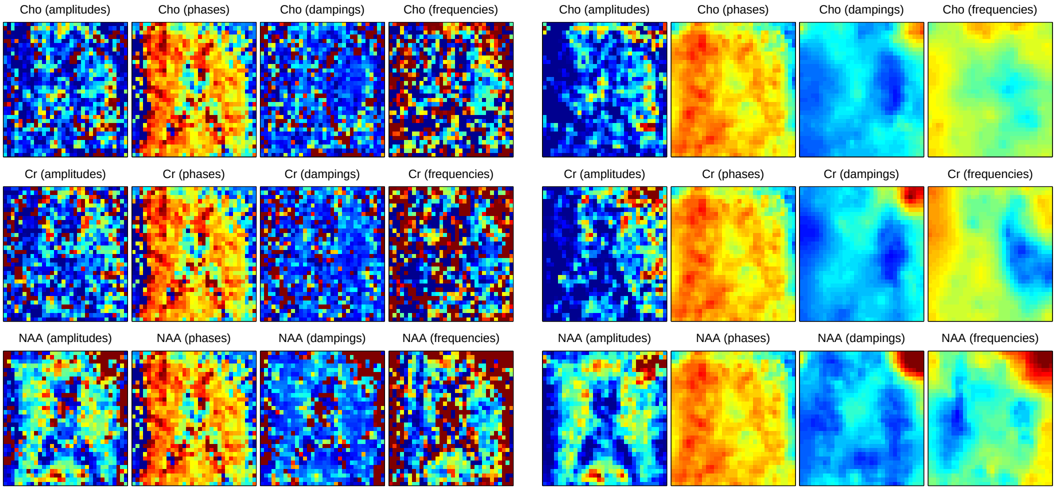
Finally results for the high-resolution data are presented. Figure 3.12(b) shows a comparison between parameter maps obtained for the 3T brain MRSI data with the single voxel approach AMARES and using the GMRF prior. Clearly the phase, damping and frequency maps in Fig. 3.12(b) are much smoother using the GMRF prior than the results obtained with a single voxel fit. The amplitude maps do not exhibit severe smoothing although many speckles from the single voxel approach do not occur with the GMRF prior, in particular for NAA.



(a) Single voxel (AMARES)

(b) Spatial prior (GMRF)

Figure 3.10.: Nine adjacent voxels from a patient with brain tumor and the corresponding AMARES and GMRF fits. The overlapping Cho and Cr peaks lead to erroneous single voxel fits in four of the nine voxels which are well captured when using spatial prior knowledge.

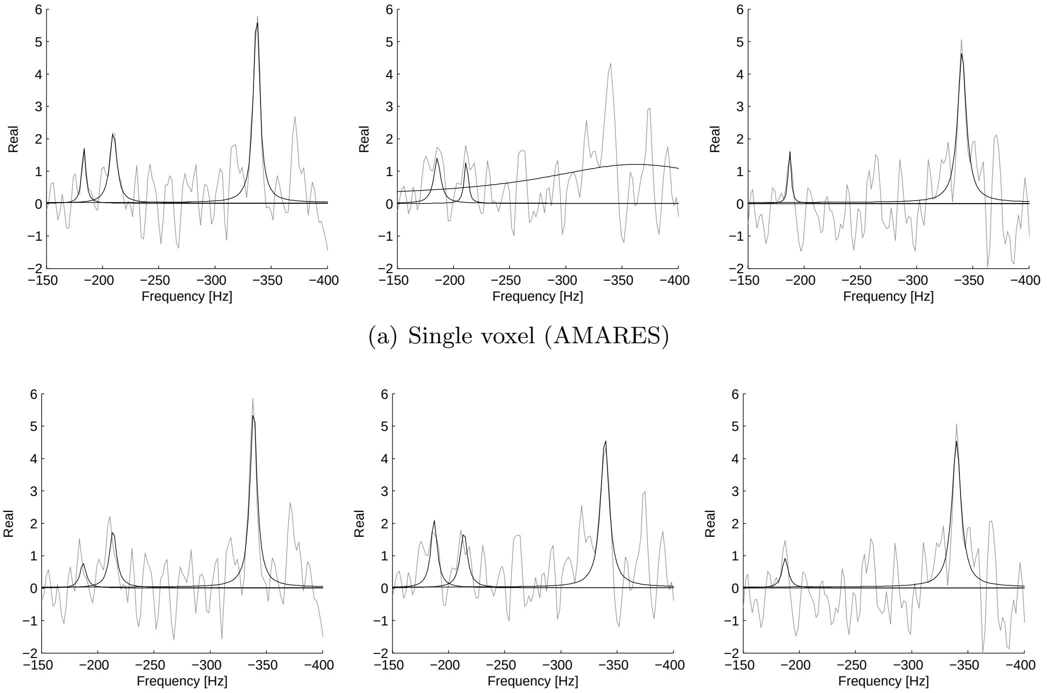


(a) Single voxel (AMARES)

(b) Spatial prior (GMRF)

Figure 3.11.: Metabolite maps from the brain of a healthy volunteer at 3T (32 × 32 voxels). Example spectra are shown in Fig. 3.12.

.



(a) Single voxel (AMARES)

(b) Spatial prior (GMRF)

Figure 3.12.: Spectra of three consecutive voxels from a MRSI brain scan at 3T with spectral line fits for Cho (−188.4Hz), Cr (−215Hz) and NAA (−342.4Hz). The top row shows results from AMARES whereas the bottom row shows the corresponding fits obtained with using spatial prior knowledge. The single voxel approach fails to fit the NAA peak in the middle voxel whereas the GMRF uses information from the neighboring spectra to stabilize the fit.

Spectra from three adjacent voxels are shown in Fig. 3.12 together with the AMARES and GMRF fits. The single voxel approach fails to properly fit the NAA peak in the middle voxel whereas the GMRF prior again prevents the sudden jump in line width (damping) and can provide a trustworthy fit. In contrast, the Cho and Cr peaks in the single voxel solution seem to better fit the observed signal and indicate some bias of the GMRF, especially for the left and right spectra in Fig. 3.12. However, since the two peaks nearly vanish in the signal noise the detection of a Cho or Cr resonance in the given data is questionable anyway. Also, as the amplitude maps for Cho and Cr do not show clear structures for either of the fit approaches in Fig. 3.11 it must be assumed that these metabolites cannot be estimated from the available 3T MRSI data.

3.4.5. Discussion

The simulation study has shown that by using spatial prior knowledge the estimation variance can be decreased significantly, even below the Cramér-Rao lower bound (CRLB) of the single voxel approach. However, since the single voxel CRLB assumes statistical independence between the FID parameters of each voxel, a spatial model such as the proposed GMRF prior is certainly contradictory and must lead to bias, in general. Since the CRLB provides a lower bound for *unbiased* estimators only it can be beaten using a *biased* estimate. Another biased estimator could be constructed with AMARES by using soft constraints. For example, if for the simulated "sharp edges" data set all frequency shifts would have been constrained to only vary ±1Hz, the standard deviation of the frequency estimates could not have exceeded 2Hz which would be better than the CRLB for some of the voxels (cf. Fig. 3.7(a)). However, if the true frequency lied outside this interval the estimate would certainly be biased. Thus, prior knowledge reduces estimation variance but increases bias if not appropriate.

Naturally the question arises whether spatial prior knowledge in form of the proposed GMRF is appropriate or not, *i.e.* whether it evokes bias or not. This question cannot be answered satisfactorily by a Monte Carlo study for which the data generating process is known. One can only try to put forward arguments that could justify that parameter maps are truly smooth. First, the employed MR sequences always cause a certain point spread which leads to FID signals that are spatially correlated. Second, parameters such as frequency shift and damping very much depend on the homogeneity of the applied B-field which is optimized in preceding shimming procedures. Ideally these parameters are therefore constant across the whole region of interest. Also, it has been observed that the zero-phase parameter usually varies only smoothly. Finally, the presented exemplary fits (cf. Fig. 3.10) seem much more plausible when using the proposed spatial prior than without it which is justified by comparison with spectra of neighboring voxels. For example, the spectral fit in the lower right corner of Fig. 3.10 only seems appropriate if the spectra in neighboring voxels are not known.

In the presented experimental study it was chosen not to impose a direct smoothness prior on the amplitude maps since this is the main parameter of interest in quantification for which spatial variation is expected and should not be smoothed. That even steep amplitude edges can be reconstructed without blurring has been demonstrated with the "sharp edges" data set (cf. Fig. 3.3). In contrast, posthoc smoothing of single voxel estimates lead to blurring but also gave improvements within the smooth parts of the test image. Hence, some spatial smoothing of the amplitude maps might also be desired and would argue for choosing a small positive weight W_a as well.

Further work is necessary to devise a procedure for fine-tuning the weighting matrix W that is better than the proposed variogram-based approach.

A strong advantage of the proposed GMRF approach is its ability to better resolve overlapping peaks than the single voxel approach. This suggests an application to short echo ^{1}H-MRSI in the brain. However, this requires appropriate baseline handling which will also be subject of future research.

3.4.6. Conclusion

An application of *spatial* prior knowledge in addition to commonly employed parameter constraints has been proposed in spectral fitting of magnetic resonance spectroscopic images. Using a Gaussian Markov random field that favors smooth maps for selected parameters such as phase, line width and frequency shift, the standard deviation of parameter estimates can be reduced below the Cramér-Rao lower bound that is obtained for the single voxel approach. An important advantage of using spatial prior knowledge is that overlapping peaks can reliably be resolved, also in cases where the single voxel approach fails to do so. The proposed method is particularly useful for high-resolution MRSI and allows to derive detailed metabolic images.

3.5. Quantification of Dynamic Contrast-Enhanced MR Images

3.5.1. Introduction

Kinetic parameter maps extracted from Dynamic Contrast-Enhanced MR Imaging (DCE-MRI) exhibit information about tissue perfusion and vascular permeability. In contrast to static imaging modalities such as T1- and T2-weighted MRI which mainly carry morphological information, DCE-MRI allows to derive physiologic information. It can therefore provide valuable evidence for clinical diagnostics, especially in cases where common modalities fail to distinguish the pathophysiologic process of interest. DCE-MRI has already proven to be useful in various clinical applications such as the examination of breast cancer [40], bone marrow [86, 148], brain [41, 209] and prostate tumors [223, 104, 178].

DCE-MRI can be acquired on common clinical MR scanners without any difficulty. However, it also requires some kind of postprocessing in order to image the diagnostic information, which can pose problems for low signal-to-noise ratios (SNR) and when unanticipated signals or artifacts are encountered. Various postprocessing strategies have been proposed which, in general, can be grouped into model-based and model-free approaches.

In model-based approaches, an appropriate pharmacokinetic model which describes the expected signal enhancement dynamics is derived [41, 197]. Its parameters are often associated with meaningful physiologic properties of the examined tissue and are determined with a nonlinear least squares (NLS) approach. Because of signal noise and local optima in the NLS objective these model fits can fail for individual voxels, resulting in speckled parameter maps. In [41], such voxels are identified by a SNR criterion and removed from the parameter maps, thus discarding possibly valuable information.

A different approach to cope with noise and unanticipated signal shapes is the application of model-free methods which do not make explicit prior assumptions about the data. For example, [199] only requires a labeled data set to train an artificial neural network without any physiologic model. As opposed to that, [141] proposes the application of an unsupervised clustering method: each signal-time curve is regarded as a feature vector and is projected onto a two-dimensional manifold by means of self-organizing maps. The resulting 2D coordinates can be color-coded and mapped on a diagnostic image. However, model-free approaches often lack physiologic interpretability and come with the flavor of "black-box" methods.

The approach taken here is model-based. In addition to a pharmacokinetic model it is assumes that the characteristics of the tissue vary gradually from voxel to voxel and, hence, that the parameter maps that best describe the physiologic properties of the tissue should exhibit some spatial smoothness. The parameter maps are modeled as a generalized Gaussian Markov random field (GGMRF) and the recorded DCE-MRI data is regarded as the noisy observation of a nonlinear transformation of the hidden parameter maps (the forward model). Hence, the parameter estimates at each voxel are supported by the estimates at its neighboring voxels which helps avoid spurious local optima of the NLS objective.

Markov random field (MRF) priors have been used for functional MRI (fMRI) data before (*e.g.* [60, 59, 150]). Related work can also be found in the application to fluorescene optical diffusion tomography [133] and to dynamic PET data [101]. Both works employ a GGMRF prior to improve the reconstruction of parameter images using an algorithm which the authors call parametric iterative coordinate descent (PICD) and which is similar to the conventional ICM algorithm.

3.5.2. Dynamic Contrast-Enhanced MR Imaging

DCE-MRI is used to track the diffusion of a paramagnetic contrast medium (CM) such as Gd-DTPA and study tissue perfusion and vascular permeability *in vivo* [41]. Therefore, DCE-MRI allows to detect pathologic tissue changes and can be used in clinical diagnostics [223, 104, 178, 40, 86].

During the intravenous injection of the CM, a sequence of several T1-weighted MR image volumes is recorded at intervals of a few seconds. Hence, a T1 signal-time curve is obtained for every voxel. An evaluation of this signal-time curve is usually based on a pharmacokinetic model whose parameters characterize the uptake and the washout of the CM from the underlying tissue [41].

Different pharmacokinetic models have been proposed in the literature (see [196, 197] and references therein). Here, only the widely employed two-compartment model by Brix *et al.* [41] is considered.

In [41], Brix *et al.* derive an explicit expression for the time-dependent contrast-enhanced MR signal $S_{CM}(t)$ from a system of first order ordinary differential equations (ODE). The time-intensity curves are thus described by

$$\frac{S_{CM}(t)}{S_0} = \begin{cases} 1 & t \leq t_0 \\ 1 + A\, C_{CM}(t - t_0) & t_0 < t \end{cases} \tag{3.16}$$

where S_0 is the T1 signal intensity obtained without CM and t_0 the lag time. The enhancement amplitude A depends on several tissue properties, the employed MR sequence and the infusion rate of the CM [41]. It is usually increased in tumors. The concentration of the CM evolves as

$$C_{CM}(t) = v\frac{\exp(k_{\text{el}}t') - 1}{\exp(k_{\text{el}}t)} - u\frac{\exp(k_{\text{ep}}t') - 1}{\exp(k_{\text{ep}}t)} \tag{3.17}$$

with $t' \equiv t$ for $t \leq \tau$, $t' \equiv \tau$ for $t > \tau$ and

$$u^{-1} = k_{\text{ep}}(k_{\text{ep}} - k_{\text{el}}) \tag{3.18}$$

$$v^{-1} = k_{\text{el}}(k_{\text{ep}} - k_{\text{el}}). \tag{3.19}$$

Here, k_{ep} (k_{21} in the original work, cf. [196]) describes the exchange rate between the two compartments (blood plasma and interstitium) which is often increased in tumors, and k_{el} is the first order elimination rate constant of the CM from the first compartment (plasma). The duration of the CM injection is described by τ. For convenience, all six model parameters are summarized in the parameter vector $\theta = (S_0, A, k_{\text{el}}, k_{\text{ep}}, t_0, \tau)$ in the following. Some examples of signal-time curves fitted by minimizing the single voxel SSR as in Eq. (3.1) are shown in Fig. 3.13.

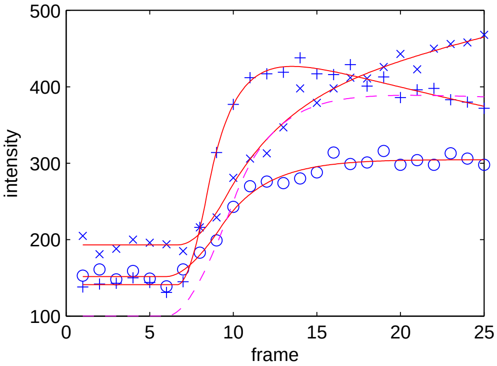


Figure 3.13.: Three examples for the employed pharmacokinetic model from [41] fitted to measured T1 signal-time curves. One frame was acquired every 11.25s. The dashed curve is obtained with the initialization parameter vector $\theta_0 = (100, .3, .4, .003, 10, 6)$.

The NLS problem (3.1) is usually solved voxel by voxel [41, 196]. Since a pharmacokinetic model such as (3.16) leads to a nonconvex NLS objective, multiple local optima may exist and an optimization routine may get trapped in one of these. The resulting parameter maps then contain single voxels or even regions of voxels with parameter estimates that are very different from their surroundings and that are deemed very unlikely to image a real physiologic process. Instead of identifying and masking these voxels in the parameter maps, a spatial smoothness prior in form of a generalized Gaussian Markov random field (GGMRF) [35] is employed. In this way, the NLS fit in each voxel is influenced by the data and the model fits in its local neighborhood, and is pushed towards a better solution.

3.5.3. Experimental Setup

DCE-MRI data from an ongoing prostate study has been collected at the German Cancer Research Center (dkfz, Heidelberg). The data was acquired on a clinical 1.5T scanner (Magnetom Symphony; Siemens Medical Solutions, Erlangen, Germany) with a disposable endorectal coil (MRInnervu; Medrad Inc., Indianola, PA, USA) using a T1-weighted FLASH sequence (TR/TE = 125ms/3.11ms) [223]. Series of 25 image volumes have been acquired with a temporal resolution of 11.25s. Each volume consisted of 16 slices with 256 × 160 voxels (.78mm × .78mm, slice thickness 3mm, 1mm gap). From the 36 patients which were available for the present study, ROIs of about 100 × 100 voxels have been selected in slices that contained tumorous tissue as determined from T1-MRI, MR spectroscopic imaging and histologic examinations (details in [4]).

For comparison, the parameters of the pharmacokinetic model (3.16) were determined using the voxel-wise approach first. The NLS fit was performed with an interior trust region method [52] as implemented in Matlab R14 (optimization toolbox). The same initialization $\theta_0 = (100, .3, .4, .003, 10, 6)$ was used in all voxels (cf. Fig. 3.13). A maximum number of 10000 iterations per voxel was allowed for in order to ensure proper convergence.

The proposed GGMRF approach was applied using a 2-norm ($p = 2$) which reduces the GGMRF to a Gaussian MRF (GMRF). A homogeneous parameterization was chosen and the influence of the diagonal coupling terms was reduced by choosing α_{st} 1.4 times smaller than for the horizontal/vertical coupling terms. For the block-ICM algorithm, the whole lattice was subdivided into two sets of blocks with 6×6 voxels, such that the second set had a horizontal and vertical displacement of three voxels (the dashed squares in Fig. 3.2). In every odd sweep, the blocks in the first set were visited in a doubly-quincunx pattern as indicated by the numbering in Fig. 3.2. In every even sweep, the same procedure was performed on the second, shifted set of blocks. A total of 12 sweeps were performed. To prevent premature convergence to a local minimum, the number of optimization steps in each block was restricted to 20 iterations in the first 10 sweeps. For the remaining sweeps, up to 2000 iterations were permitted. Initialization was done as in the voxel-wise approach and the same optimization algorithm was used.

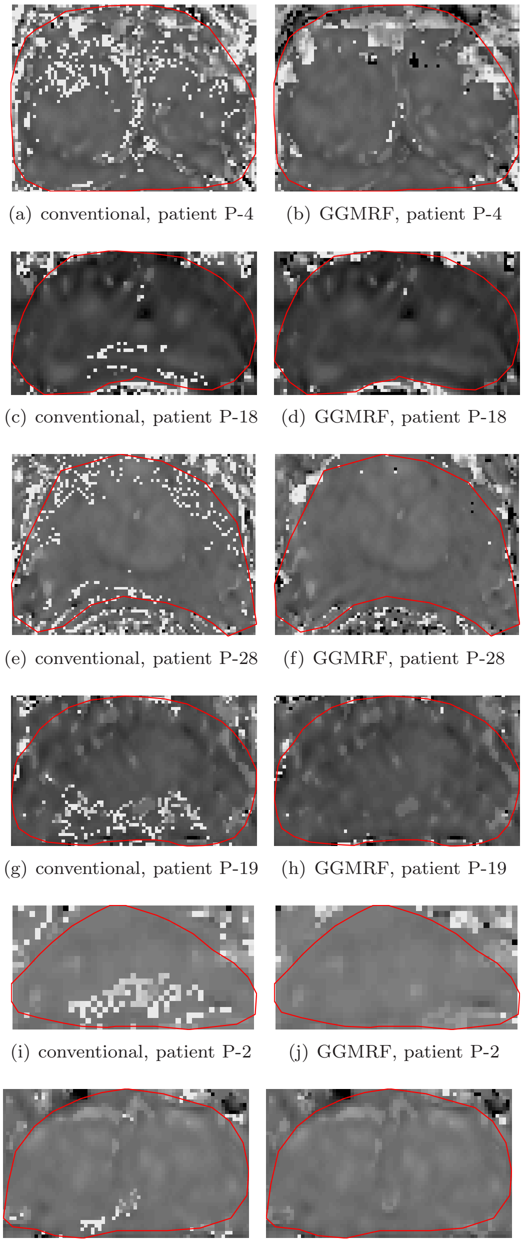
3.5.4. Results

The parameter maps for the important physiologic parameter k_{ep} shown in Fig. 3.14 allow for a qualitative comparison between the voxel-wise (left) and the GGMRF approach (right). Results from several patients are provided and all of them show white speckles in the voxel-wise approach. Since it is very unlikely that the k_{ep} value exhibits such erratic spatial changes, it can be assumed that all such speckles indicate failed parameter fits. Using the GGMRF, most of the speckles within the indicated boundary of the prostate can be removed.

Numerical Results for all 36 patients are provided in Tab. 3.2. In order to make results from time-curves with a different number of sampling points N comparable in terms of the residuals, the mean squared residuals (MSR) have been used instead of the SSR for the evaluation:

$$\mathrm{MSR} \quad = \quad \frac{1}{N}\mathrm{SSR} \tag{3.20}$$

For comparison, also the average difference in MSR (ΔMSR), *i.e.* averaged of the voxels, between the spatial GGMRF approach and the conventional voxel-wise approach



(a) conventional, patient P-4

(b) GGMRF, patient P-4

(c) conventional, patient P-18

(d) GGMRF, patient P-18

(e) conventional, patient P-28

(f) GGMRF, patient P-28

(g) conventional, patient P-19

(h) GGMRF, patient P-19

(i) conventional, patient P-2

(j) GGMRF, patient P-2

(k) conventional, patient P-3

(l) GGMRF, patient P-3

Figure 3.14.: Comparison of k_{ep}-maps for several patients (cf. Tab. 3.2). The GGMRF approach yields improved MSR for all shown patients but P-3. Nevertheless, the conventional k_{ep}-map of patient P-3 in (k) shows some speckles which are avoided with the GGMRF approach (l).

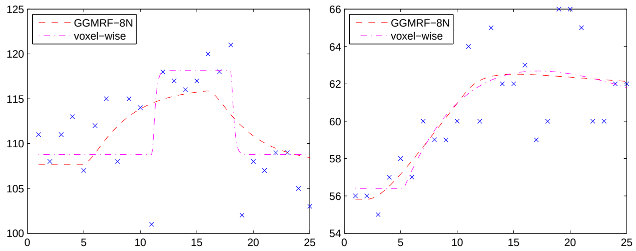


Figure 3.15.: Two examples for fits from patient P-7 which has the lowest signal-to-noise ratio in Tab. 3.2. In both cases the expected signal shape is hardly recognizable in the measured data (blue crosses).

was calculated. Furthermore, ΔMSR was splitted into its positive and negative contributions

$$\Delta\mathrm{MSR}^{-} = \frac{1}{|V|}\sum_{s\in V}\max(-\Delta\mathrm{MSR}(s),0) \tag{3.21}$$

$$\Delta\mathrm{MSR}^{+} = \frac{1}{|V|}\sum_{s\in V}\max(\Delta\mathrm{MSR}(s),0) \tag{3.22}$$

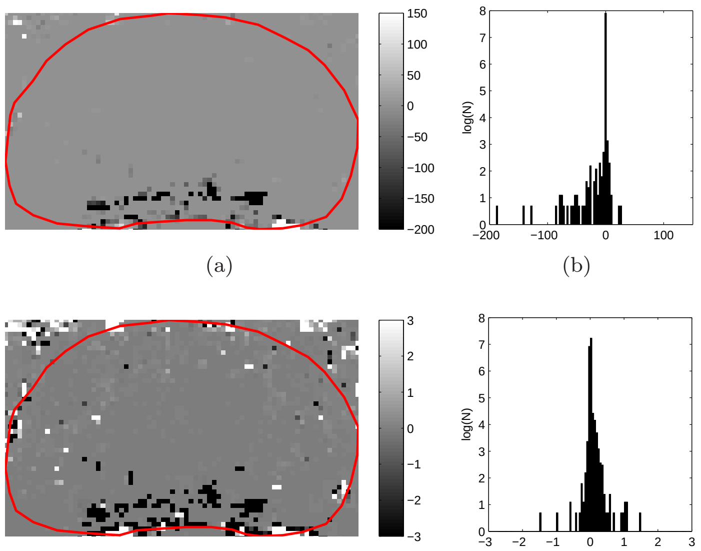
such that $\Delta\mathrm{MSR} = \Delta\mathrm{MSR}^{+} - \Delta\mathrm{MSR}^{-}$. $\Delta\mathrm{MSR}^{-}$ quantifies the MSR difference of the voxels for which the GGMRF approach performs better, and $\Delta\mathrm{MSR}^{+}$ summarizes the voxels in which the voxel-wise approach produces a lower MSR. The ratio $\Delta\mathrm{MSR}^{-}/\Delta\mathrm{MSR}^{+}$ thus allows for a comparison of the voxel-wise and the GGMRF approach; for ratios greater than one, the latter performs better. In this study, the GGMRF yields a better fit of the pharmacokinetic model, as measured by the MSR, in 34 out of 36 patients. The k_{ep}-parameter maps in Fig. 3.14 are from the patients printed in boldface and reflect the numerical results from Tab. 3.2 quite well.

Table 3.2 also provides a robust estimate of the signal-to-noise ratio (SNR). The lowest SNR is obtained for patient P-7 for which two exemplary fits are shown in Fig. 3.15. Both cases confirm that the measured image data is very noisy and show that the expected signal shape (cf. Fig. 3.13) can hardly be recognized. For example, in the left part of Fig. 3.15 it is difficult to judge which of the two very different fits should be favored by just looking at the isolated voxel-data.

A more detailed analysis of the resulting model fits was performed based on patient P-19. Figure 3.16(a) presents an image of the voxel-wise difference in MSR between the GGMRF and the conventional approach. The GGMRF yields smaller MSR in voxels that are dark in Fig. 3.16(a). Within the prostate, the GGMRF apparently never performs worse but in many voxels better than the conventional approach. However, if the same difference image is displayed on a smaller intensity range (Fig. 3.16(c)) the

Patient	SNR [dB]	ΔMSR$^-$	ΔMSR$^+$	ΔMSR	ratio
P-4	39.0	1410.241	2.186	-1408.055	644.981
P-30	33.8	1064.353	4.018	-1060.335	264.895
P-29	27.3	223.972	0.857	-223.115	261.351
P-14	51.7	35.729	0.162	-35.567	219.999
P-16	49.6	194.004	1.033	-192.972	187.885
P-23	50.9	101.946	0.967	-100.979	105.386
P-27	31.2	13.349	0.157	-13.192	85.142
P-15	36.8	21.289	0.266	-21.023	80.113
P-5	46.9	47.562	0.868	-46.694	54.785
P-18	45.4	42.064	0.829	-41.235	50.750
P-24	33.1	687.113	15.007	-672.106	45.786
P-26	22.4	81.634	1.887	-79.747	43.258
P-7	16.4	191.418	4.714	-186.704	40.603
P-35	38.0	194.517	5.400	-189.117	36.020
P-13	23.6	41.499	1.427	-40.072	29.085
P-31	26.8	34.853	1.426	-33.427	24.434
P-20	41.5	59.006	2.538	-56.468	23.251
P-28	41.9	69.005	3.597	-65.408	19.184
P-11	50.7	23.560	1.333	-22.227	17.668
P-12	41.8	61.956	5.112	-56.844	12.120
P-6	52.0	0.433	0.043	-0.390	10.017
P-36	45.1	4.046	0.472	-3.575	8.574
P-19	33.3	15.860	2.040	-13.820	7.776
P-33	33.9	6.387	0.913	-5.474	6.994
P-25	42.4	11.269	1.619	-9.650	6.959
P-9	29.9	0.632	0.095	-0.537	6.662
P-34	31.0	9.341	1.422	-7.919	6.569
P-2	30.3	4.117	0.802	-3.315	5.134
P-1	34.7	0.124	0.032	-0.092	3.889
P-10	28.9	6.589	1.869	-4.720	3.525
P-22	33.6	6.715	2.061	-4.654	3.258
P-21	40.8	1.763	0.655	-1.108	2.692
P-8	17.6	10.740	4.367	-6.373	2.459
P-32	32.7	21.184	12.499	-8.685	1.695
P-17	24.0	3.940	5.318	1.378	0.741
P-3	50.0	0.035	0.148	0.113	0.236

Table 3.2.: Results for all 36 patients ordered by the ratio $\Delta\text{MSR}^-/\Delta\text{MSR}^+$ (see text). Using the GGMRF prior improves the mean MSR in all but the last two cases. Parameter maps for bold patient identifiers are shown in Fig. 3.14.



(a) (b)

(c) (d)

Figure 3.16: Difference in mean squared residuals (MSR) for patient P-19. Darker pixels indicate sites for which the GGMRF model could find a better fit. (a) As for the k_{ep}-map in Fig. 3.15(h) the greatest benefit is obtained in the lower half of the image. (b) Histogram of the difference values within the prostate. (c) Contrast-enhanced version of (a). (d) Correspondingly zoomed histogram. Positive differences prevail in the displayed range, reflecting the bias introduced by the spatial prior.

contrast increases and a few white voxels emerge. In these voxels, the conventional approach yields lower MSR.

The two difference images in Fig. 3.16 indicate that the MSR difference is very big for voxels in which the GGMRF performs better and small in voxels for which the conventional approach is better. This observation is confirmed by the corresponding histograms. The histogram in Fig. 3.16(b) exhibits a lot of mass for the negative difference values which can be explained by the voxels with failed parameter fits. Then again, its zoomed version in Fig. 3.16(d) is clearly skewed towards positive values, an indication for bias.

Figure 3.16 thus allows to identify both such kinds of voxels and examine the respective model fits. For some of the black voxels in Fig. 3.16(a) in which the GGMRF yields significantly smaller MSR, these are compared in Fig. 3.17. Although the voxel-wise approach can find reasonable fits, especially for the first two examples, the solutions found using the GGMRF look much more convincing in all four cases. By contrast, Fig. 3.18 presents two examples from the encircled white voxels in Fig. 3.16(c). And, although the GGMRF fits in the right column of Fig. 3.18 pro-

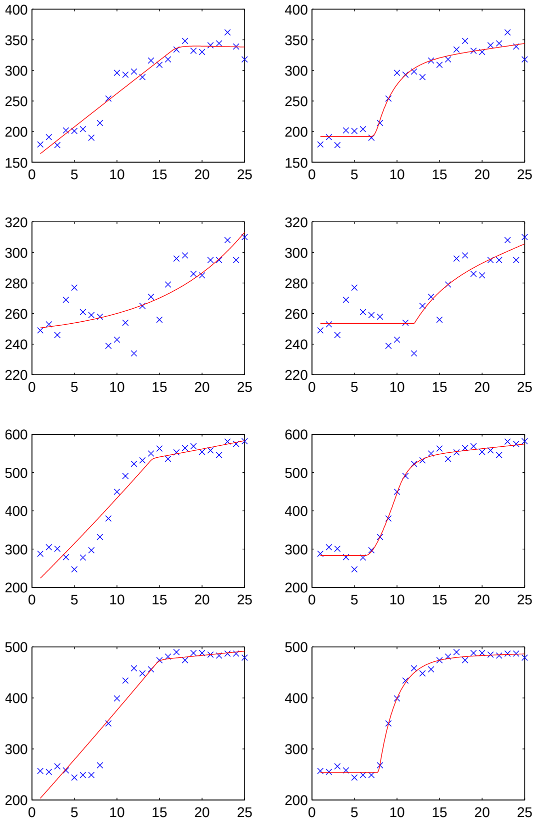


Figure 3.17.: Comparison of selected model fits (patient P-19). In these examples the GGMRF approach produced a lower MSR. The left column shows results from the voxel-wise approach whereas the right column shows the corresponding fit with the GGMRF prior.

duce slightly higher MSR than the voxel-wise fits in the left column, both solutions look reasonable. In fact, one might even prefer the solutions found with the GGMRF approach.

Finally, convergence results are provided in Fig. 3.19 that show the influence of using different block sizes in the proposed block-ICM algorithm. The special case of 1x1 blocks results in the conventional single-site ICM algorithm which clearly converges much slower than block-ICM using bigger blocks. When plotted over time (Fig. 3.20(a)) no difference in convergence speed is observed between the block sizes of 4x4, 6x6 and 9x9. A slight advantage of the biggest block size (9x9) becomes visible in Fig. 3.20(b) where the MAP objective is plotted against the number of evaluations of the model function $f_{\theta_s}(t)$.

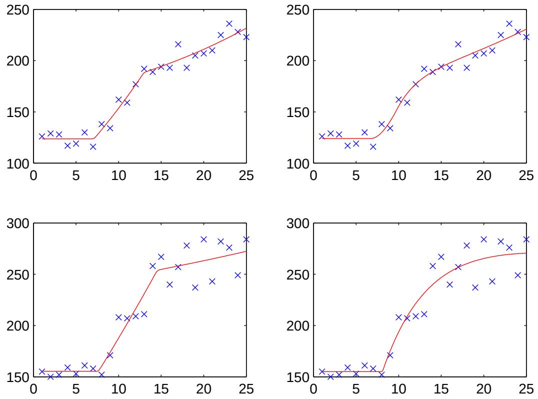


Figure 3.18.: Fits at the two voxels which are encircled in Fig. 3.16(c). The left column shows the voxel-wise and the right column the corresponding GGMRF fits.

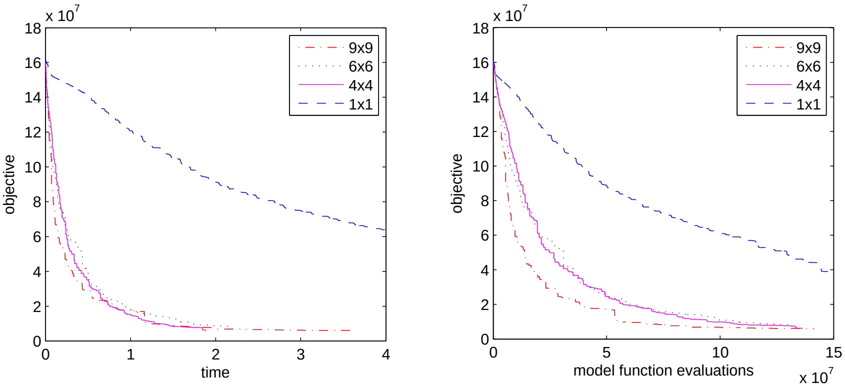


Figure 3.19.: Convergence behavior of ICM (1x1) and block-ICM for different choices of block sizes. The MAP objective is plotted over time (left) and over the number of model function evaluations (right). Block-ICM clearly outperforms the common ICM algorithm (1x1). Very similar performance is obtained for the block sizes 4x4, 6x6 and 9x9.

3.5.5. Discussion

At first sight, the decreased SSR observed in Tab. 3.2 and Fig. 3.16(a) might be surprising since the use of a smoothness prior should result in estimation bias and always yield greater SSR than an unbiased estimate. However, this relation must only hold true if the *global* optimum of the NLS objective (3.1) (voxel-wise) is compared with the global MAP optimum (3.5) (GGMRF). Hence, in all voxels for which the voxel-wise approach yields increased SSR, a local suboptimal solution has been found. Considering the white voxels in the contrast-enhanced SSR difference image in Fig. 3.16(c) and the histogram 3.16(d), estimation bias must indeed be assumed in voxels for which the conventional and the GGMRF approach converge to similar solutions.

Other approaches that support the convergence to better solutions of the NLS objective are conceivable. One could, for example, specify local constraints or priors on the model parameters. Compared to the GGMRF, this kind of prior knowledge is more explicit in that a very specific range of parameter values needs to be presumed. In contrast, no particular parameter value is preferred with the GGMRF since only differences between neighboring parameters are penalized. Still, this kind of local prior knowledge may be useful and is easily combined with the proposed GGMRF approach. Another approach to avoid local optima of the NLS objective is the use of initialization strategies that attempt to provide starting values close to the global optimum. However, such strategies need to be failsafe and consider the case of unexpected signal shapes since more harm than benefit could result from a poor choice of starting values. Certainly, sophisticated initialization strategies could also help the GGMRF approach to converge faster. In the present study, adaptive initialization strategies or local prior knowledge have not been employed in order to avoid confusion with the effects of the GGMRF prior.

In the presented experiments, the GGMRF prior was parameterized so that only weak smoothing effects have been obtained. The observed bias was very low and oversmoothing, which is often seen for Gaussian MRFs, did not pose a problem. Comparing the parameter maps in Fig. 3.14, faint structures are found in areas where both approaches yield good results. These are well preserved with the GGMRF prior which seems to be very adequate. One reason for that certainly is that due to the time constraints in the recording of DCE-MRI quite blurry image data is obtained. Another reason may be that the diffusion process in the prostate only yields rather smooth spatial concentration changes, anyway.

Overall, as shown in the examples in Figs. 3.17 and 3.18, the GGMRF prior usually yields curve fits that are more in line with the expected pharmacokinetic behavior even if the SSR is higher (Fig. 3.18). Thus, the effect of the GGMRF prior is twofold. Not only does it help avoid getting trapped in local optima, but it also reduces estimation variance.

Despite its favorable properties, using the GGMRF prior certainly makes parameter estimation computationally very demanding. Without a specialized optimization strategy such as ICM which can exploit the inherent sparseness of the MAP problem, the GGMRF prior would not be applicable for clinical purposes. Block-ICM can speed up convergence significantly and does not seem to be very sensitive to the choice of block size (as long as it does not reduce to conventional ICM). For the proposed GGMRF and on a second order neighborhood system, a block size of 6x6 appears to be a good choice in any case.

3.5.6. Conclusion

The application of spatial prior knowledge in form of a generalized Gaussian Markov random field prior has been proposed to improve the estimation of kinetic parameter maps from DCE-MRI. Since the nonlinear least squares problem used to fit a pharmacokinetic model in each voxel is nonconvex, the conventional approach is susceptible to getting trapped in local optima. It has been demonstrated that using the GGMRF prior can help avoid false optima and can yield parameter estimates with a lower mean square error. The noticeable speckles in the kinetic parameter maps resulting from failed parameter fits could largely be removed using the GGMRF prior. And, although a 2-norm has been used in the conducted experiments with prostate DCE-MRI, the GGMRF prior did not lead to substantial oversmoothing. The proposed block-ICM procedure demonstrated how the resulting, very high-dimensional optimization problem can be tackled efficiently, paving the way for a clinical application.

3.6. Summary

Usually quantification of vector-valued MR image data is performed by fitting a nonlinear signal model to the observed data in every voxel independently. This chapter has introduced an approach for including spatial prior knowledge about the model parameter maps.

Using a generalized Gaussian Markov random field (GGMRF) prior for the estimation of the parameter maps has several effects. First, the sum of squared residuals (SSR) can be decreased in many voxels as compared to a single voxel approach. This can only be explained by the failure of the single voxel approach to find the *global* optimum of the SSR objective whereas the spatial prior helps to avoid spurious local optima. Secondly, a smoothness prior on the parameters maps reduces estimation variance but potentially introduces bias. However, Monte Carlo studies have shown that the reduction in variance is much larger than the negligible amount of bias introduced, which results in a considerable decrease of root mean squared error. Finally, the resulting parameter maps for both dynamic contrast enhanced MR imaging (DCE-MRI) and magnetic resonance spectroscopic imaging (MRSI) simply look nicer and are more in line with the expected behavior of the mapped physiological process.

The optimization problem resulting from the GGMRF prior is over all parameters in the fitted image simultaneously, *i.e.* very big. A blocked version of the iterated conditional modes (ICM) algorithm has been proposed that can exploit the sparse structure of the problem in solving the optimization problem by dividing it into a number of smaller problems. An initial convergence analysis has shown significant speed-up as compared to conventional ICM.

Chapter 4.

Pattern Recognition Using Spatial Context

4.1. Introduction

Pattern recognition approaches are advisable if a labeled data set is available and one is mainly interested in obtaining a classification. In contrast to the previous chapter which proposed to incorporate spatial prior knowledge into quantification, the present chapter introduces and compares approaches that include spatial context into pattern recognition. In particular, a family of generative and discriminative random field models for simultaneous segmentation and classification are proposed and compared.

In MRSI, a trade-off between scan time, spatial resolution, spectral linewidth and signal-to-noise ratio (SNR) is made. Usually one tries to achieve "good" SNR at reasonable resolutions and scan times. At higher magnetic fields (3T), which are currently being introduced in standard clinical MR scanners, and using improved pulse sequences, these factors can be improved [46, 69, 81, 118]. In [118], Li *et al.* observed that, surprisingly, the SNR is not proportional to the voxel volume d^3 but only increases as d^2 due to a reduction of field inhomogeneities within small voxels. Gruber *et al.* exploit this effect by acquiring spectral images at very high resolutions, resulting in extremely noisy individual voxels, which are then averaged within manually segmented anatomical structures [81]. Thus, additional information can be gained from increasing imaging resolution when combined with sophisticated smoothing strategies in a postprocessing step.

Up to now virtually all proposed approaches concentrated on the processing of single spectra, neglecting the spatial nature of the acquired data (*e.g.* [7, 4, 32, 61, 193, 68, 70]). Given the classification (healthy/tumor+type) of every voxel the results have merely been display in so-called *nosologic images* [54]. Here, a classifier is to be learned using a labeled MRSI data set for an automated segmentation and classification of tumor regions using spatial probabilistic models [48].

In [113], Laudadio *et al.* describe an approach that classifies each voxel based on features extracted from a local neighborhood, thus by applying a *sliding window* method [64, 63][*] which is the first attempt to make use of spatial information in MRSI. However, the sliding window approach has two major disadvantages. First, spectra in the local neighborhood can have an adverse effect on classification performance, for example if a healthy voxel borders tumorous tissue. Secondly, the neighborhood is fixed and easily chosen too small so that long-range interactions that might support the local decision are neglected – a disadvantage in particular for very noisy spectral data.

These limitations can be overcome using global models such as Markov random fields (MRF) [28, 73, 64]. MRF-based models of various flavors have been proposed and used for the segmentation of scalar-valued as well as multi-spectral images (*e.g.* [31, 56, 215, 71, 47, 120, 181]). They can be used in both the unsupervised and the supervised setting. Usually, a *generative* approach is taken by modelling the data generating process of the observed image. Hence, a generative probabilistic model is derived by specifying a prior distribution $\Pr(\mathbf{y})$ on the class label map and the likelihood of the observed data given a label map $\Pr(\mathbf{x} \,|\, \mathbf{y})$. The prior combines information globally by favoring label maps that are smooth whereas the observation term models the local data attachment. For computational tractability the latter is often assumed to factorize over the sites, reflecting the assumption that the observed voxel data is independent given the label map.

Two major problems arise from this approach [64]. First, any correlation between labels has to be mitigated in the label map. If long-range interactions are to be modeled with a generative MRF the neighborhood system has to be enlarged resulting in increasingly difficult inference. Moreover, lacking further knowledge about unseen data, the label map prior is usually homogeneous and isotropic. Second, in order to accurately capture complex distributions, a powerful model for the observation term such as a mixture model should be chosen. This, however, results in a non-convex optimization problem for parameter estimation requiring good initialization schemes and/or optimization approaches that can escape local optima.

Both problems are addressed with conditional random fields (CRF) [111] which reflect the *discriminative* approach to probabilistic modelling [64]. Instead of expending modelling effort on the distribution of the observed data, the discriminative approach concentrates on the class label distribution which, in general, can be expected to be considerably less complex. In this sense, CRFs and MRFs relate to each other like

[*] although [64] is about *sequential* supervised learning, the described methods and their properties can be carried over to the spatial case.

logistic regression (LR) to linear discriminant analysis (LDA) which are the classical examples for comparing discriminative and generative modelling [166, 145].

With CRFs long-range interactions between observed data as well as class labels can be incorporated and data-dependent, spatially inhomogeneous and anisotropic class label distributions can be modeled in a sound way. The parameter estimation problem is convex and does not exhibit non-global optima. Nevertheless, parameter estimation remains computationally demanding due to the repeated evaluation of a log partition function.

Originally, CRFs have been introduced in the context of sequential supervised learning by Lafferty, McCallum and Pereira [111]. Since then, they have also been applied to computer vision tasks as different as character recognition [211], man-made structure detection and binary image denoising [110, 109, 108], image segmentation [87] and the demosaicking of color images [103].

In this work the application of conditional random fields to the automatic evaluation of MRSI data is examined. Generative and discriminative probabilistic models particularly suited for the processing of spectral images are proposed and described. A systematic comparison of analogous generative and discriminative models is conducted and the benefit of using spatial information is studied with regard to the trade-off between SNR and resolution in MRSI.

4.2. Generative and Discriminative Models

In the following the *generative* and *discriminative* approaches explored in this work are briefly described. Both have in common that the sought label map $\mathbf{y}$ is modeled as a Markov random field. The crucial difference lies in the way in which the observed image data $\mathbf{x}$ is incorporated.

4.2.1. Generative Approach

In the generative approach, models are derived by imitating the data generating process. A generative probabilistic model is constructed by specifying a prior distribution over class labels $\mathbf{y}$ and a class-conditional distribution $\mathrm{q}(\mathbf{x} \mid \mathbf{y})$, the observation model (cf. Fig. 4.1). In the generative approach these distributions should be carefully chosen since inappropriate model assumptions easily lead to severe bias. Therefore, powerful observation models with latent variables, such as mixture models, often perform best. In the following, generative models that are particularly suited for the simultaneous segmentation and classification of MRS images are proposed. Similar

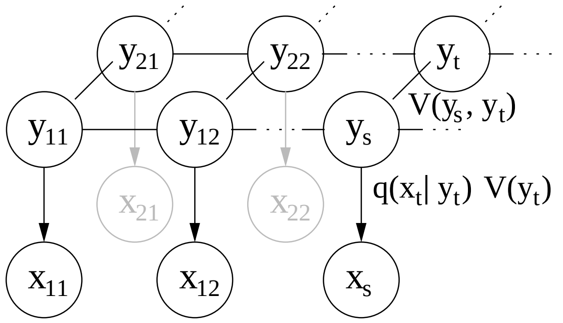


Figure 4.1.: A generative Markov random field (MRF) model. Given the spatially coupled label map $\mathbf{y}$, the class-conditional observation distributions $\mathrm{q}(x_s \,|\, y_s)$ are independent.

approaches to MRF-based segmentation of image data have been analyzed before (*e.g.* [71]).

Prior model

Let $G = (S, E)$ be an undirected graph where each site $s \in S$ represents a pixel in the label map and indexes an associated random variable Y_s that takes values in the discrete set $\mathcal{Y} = \{1 \dots L\}$, the set of class labels. The vector-valued random variable $\mathbf{Y} = (Y_s)$ takes values in the Cartesian product set $\mathcal{Y}^{|S|} = \mathcal{Y} \times \mathcal{Y} \cdots \times \mathcal{Y}$. In the conducted experiments, only regular lattice graphs with a first order neighborhood system (four neighbors) have been considered. The generalization to different neighborhood systems is straightforward. Each realization (label map) $\mathbf{y} \in \mathcal{Y}^{|S|}$ is assigned the probability

$$\mathrm{q}(\mathbf{y} \,|\, \theta) \quad = \quad \exp\left[\langle \theta, \Phi(\mathbf{y}) \rangle - A(\theta)\right] \tag{4.1}$$

with

$$\begin{aligned}
\langle \theta, \Phi(\mathbf{y}) \rangle &:= \sum_j \theta^n_j \phi^n_j(\mathbf{y}) + \sum_{j,k} \theta^e_{jk} \phi^e_{jk}(\mathbf{y}) && (4.2)\\
\phi^n_j(\mathbf{y}) &= \sum_s \delta_j(y_s) && (4.3)\\
\phi^e_{jk}(\mathbf{y}) &= \sum_{s \sim t} \delta_j(y_s) \delta_k(y_t) && (4.4)\\
A(\theta) &= \log \sum_{\mathbf{y} \in \mathcal{Y}^{|S|}} \exp \langle \theta, \Phi(\mathbf{y}) \rangle && (4.5)
\end{aligned}$$

where the Kronecker delta $\delta_j(y) = \delta(y - j)$ is 1 if $y = j$ and 0 otherwise, and $s \sim t$ denotes a neighboring pair of sites, *i.e.* $(s, t) \in E$. $\theta = (\theta_j^n, \theta_{jk}^e)$ summarizes the node parameters $\theta_n = (\theta_j^n)$ and the edge parameters $\theta_e = (\theta_{jk}^e)$, and $\Phi(\mathbf{y}) = (\phi_j^n(\mathbf{y}), \phi_{jk}^e(\mathbf{y}))$ the corresponding feature functions (sufficient statistics). $A(\theta)$ is a normalizing constant known as the log partition function.

That the described model is indeed a random field (cf. Fig. 4.1) can be seen by summarizing the terms in the energy function in Eq. (4.2) differently:

$$\begin{aligned} E(\mathbf{y}) &= \langle \theta, \Phi(\mathbf{y}) \rangle && (4.6) \\ &= \sum_s \sum_{j=1}^{L} \theta_j^n \delta_j(y_s) + \sum_{s\sim t} \sum_{j=1}^{L} \sum_{k=1}^{L} \theta_{jk}^e \delta_j(y_s) \delta_k(y_t) && (4.7) \\ &= \sum_s V_s(y_s) + \sum_{s\sim t} V_{st}(y_s, y_t) && (4.8) \end{aligned}$$

where the potentials $V_s(y_s)$ and $V_{st}(y_s, y_t)$ are defined as

$$\begin{aligned} V_s(y_s) &= \sum_j \theta_j^n \delta_j(y_s) && (4.9) \\ V_{st}(y_s, y_t) &= \sum_{j,k} \theta_{jk}^e \delta_j(y_s) \delta_k(y_t). && (4.10) \end{aligned}$$

For given parameters θ, the functions $V_s(y_s)$ can be defined by L-vectors and the $V_{st}(y_s, y_t)$ by $L \times L$ matrices. Roughly, these can be interpreted as unnormalized logarithmic probability tables.

The defined prior is homogeneous since θ_n and θ_e are both constant over the random field, and it is isotropic since θ_e is equal for horizontal and vertical edges. Furthermore, adding the linear symmetry constraint

$$\theta_{jk}^e - \theta_{kj}^e = 0 \qquad (4.11)$$

ensures that a spatial transition from class k to j is as likely as the reverse transition which is a reasonable assumption to make for the present purpose. The proposed prior model is different from the common Potts model [213, 119] in that potentials between different classes are not forced to be equal; the Potts model is obtained as a special case for $\theta_{jk} = -\beta\, \delta_j(k)$.

Observation models

The observation model is assumed to factorize over the sites. This conditional independence assumption ensures that the posterior $\mathrm{q}(\mathbf{y} \mid \mathbf{x}, \theta, \vartheta) \sim \mathrm{q}(\mathbf{y} \mid \theta)\, \mathrm{q}(\mathbf{x} \mid \mathbf{y}, \vartheta)$, which is needed for classification, remains a sparse random field with the same structure as the prior model. Thus,

$$\mathrm{q}(\mathbf{x} \mid \mathbf{y}, \vartheta) \quad = \quad \prod_{s \in S} \mathrm{q}(x_s \mid y_s, \vartheta). \tag{4.12}$$

Any generative single voxel based model $\mathrm{q}(x_s \mid y_s, \vartheta)$ with parameters ϑ can be used within this framework. For MRSI, linear discriminant analysis (LDA) and principal components analysis (PCA) have been shown to yield very good results in previous work [7, 4]. Corresponding generative models are the class-conditional Gaussian and the probabilistic principal component analyzer (PPCA) [194, 195, 76], respectively. Both can also be used for building mixtures, yielding more flexible models that can explain multimodal data distributions. Thus, the following observation models have been employed:

- LDA: d-dimensional Gaussians with class-specific mean and common full covariance matrix:

  $$\mathrm{q}(x_s \mid y_s, \vartheta) = \mathcal{N}_d(x_s \mid \mu_{y_s}, \Sigma) \tag{4.13}$$

- LDAi: Gaussians with class-specific mean and isotropic covariance:

  $$\mathrm{q}(x_s \mid y_s, \vartheta) = \mathcal{N}_d(x_s \mid \mu_{y_s}, \sigma^2 \mathbf{I}). \tag{4.14}$$

- PPCA-p: each class is modeled as a principal component analyzer of latent dimension p, *i.e.* Gaussians with constrained covariance matrices:

  $$\mathrm{q}(x_s \mid y_s, \vartheta) = \mathcal{N}_d(x_s \mid \mu_{y_s}, W_{y_s} W_{y_s}^T + \sigma_{y_s}^2 \mathbf{I}), \tag{4.15}$$

  where W_{y_s} is $d \times p$. Note that originally the PPCA-p model is explained as a model with p latent variables [194, 29, 30]. It is shown in that for the maximum likelihood estimate, the columns of W_{y_s} span the p-dimensional subspace that is also found with common PCA.

- K-MoG: the data distribution for each class is modeled as a mixture of K isotropic Gaussians:

  $$\mathrm{q}(x_s \mid y_s, \vartheta) = \sum_{k=1}^{K} \pi_k\, \mathcal{N}_d(x_s \mid \mu_{k y_s}, \sigma_{k y_s}^2 \mathbf{I}). \tag{4.16}$$

  MoGs can be explained using a model with K latent variables [30].

- K-MoPPCA-p: the data distribution for each class is modeled as a mixture of K principal component analyzers:

$$\mathrm{q}(x_s \mid y_s, \vartheta) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}_d(x_s \mid \mu_{ky_s}, W_{y_s} W_{y_s}^T + \sigma_{ky_s}^2 \mathbf{I}). \tag{4.17}$$

The prior model parameters θ of as well as the observation model parameters ϑ are estimated from observed data. Since there is considerable overlap with the estimation techniques for the discriminative approach a description of parameter estimation is deferred to section 4.3. We now proceed with the introduction of the discriminative model.

4.2.2. Discriminative Approach

The discriminative counterpart to the generative MRF model described previously is the conditional random field (CRF). Instead of modelling the data generating process, CRFs model the class distribution given observed data $\mathrm{p}(\mathbf{y} \mid \mathbf{x}, \theta)$ directly.

The definition of CRFs is quite generic and inconspicuous in that the only requirement is that $\mathrm{p}(\mathbf{y} \mid \mathbf{x}, \theta)$ constitutes a random field, *i.e.* that it factorizes according to an undirected graph [111]. The full power of CRFs arises from the possibility to use arbitrary feature functions $\phi(\mathbf{x})$ which can capture non-local correlations in the specification of $\mathrm{p}(\mathbf{y} \mid \mathbf{x}, \theta)$. Hence, the strong conditional independence assumption of the generative approach is avoided. This means that arbitrary features extracted from the observed image data, such as edge or texture information, can be incorporated in a sound probabilistic model.

Here, a CRF model suitable for simultaneous multi-class classification and segmentation is described. The formulation is different from [110] in that the site-potentials are not locally normalized. Considering only pairwise interactions, the conditional random field is defined by

$$\mathrm{p}(\mathbf{y} \mid \mathbf{x}, \theta) \quad = \quad \exp\left[\langle \theta, \Phi_{\mathbf{x}}(\mathbf{y}) \rangle - A(\theta, \mathbf{x})\right] \tag{4.18}$$

with

$$\begin{aligned}
\langle \theta, \Phi_{\mathbf{x}}(\mathbf{y}) \rangle &:= \sum_{j=1}^{L} \sum_{i=1}^{F_n} \theta^n_{ij} \phi^n_{ij}(\mathbf{y}, \mathbf{x}) + \sum_{j=1}^{L} \sum_{k=1}^{L} \sum_{i=1}^{F_e} \theta^e_{ijk} \phi^e_{ijk}(\mathbf{y}, \mathbf{x}) && (4.19)\\
\phi^n_{ij}(\mathbf{y}, \mathbf{x}) &= \sum_{s} \varphi^s_i(\mathbf{x}) \delta_j(y_s) && (4.20)\\
\phi^e_{ijk}(\mathbf{y}, \mathbf{x}) &= \sum_{s \sim t} \varphi^{st}_i(\mathbf{x}) \delta_j(y_s) \delta_k(y_t) && (4.21)\\
A(\theta, x) &= \log \sum_{\mathbf{y} \in \mathcal{Y}^{|S|}} \exp \langle \theta, \Phi_{\mathbf{x}}(\mathbf{y}) \rangle && (4.22)
\end{aligned}$$

where $\theta = (\theta^n_{ij}, \theta^e_{ijk})$ again summarizes the model parameters and the corresponding vector of sufficient statistics is $\Phi_{\mathbf{x}}(\mathbf{y}) = (\phi^n_{ij}(\mathbf{y}, \mathbf{x}), \phi^e_{ijk}(\mathbf{y}, \mathbf{x}))$. An explanation of the role of the F_n (F_e) node (edge) feature functions is provided later. Note that the summation index i in the above formulas is used for explicating dot products whereas j and k are used for class labels.

Again, the random field (cf. Fig. 4.2) can be expressed like in Eq. (4.8), leading to the potential functions

$$\begin{aligned}
V_s(y_s, \mathbf{x}) &= \sum_{i,j} \theta^n_{ij} \varphi^s_i(\mathbf{x}) \delta_j(y_s) && (4.23)\\
V_{st}(y_s, y_t, \mathbf{x}) &= \sum_{i,j,k} \theta^e_{ijk} \varphi^{st}_i(\mathbf{x}) \delta_j(y_s) \delta_k(y_t), && (4.24)
\end{aligned}$$

which, for given parameters θ and observation $\mathbf{x}$, again evaluate to a collection of vectors $V_s(y_s, \mathbf{x})$ and matrices $V_{st}(y_s, y_t, \mathbf{x})$ where each entry is obtained as the dot product between a localized feature function ($\varphi^s(\mathbf{x})$ or $\varphi^{st}(\mathbf{x})$) and a global parameter (θ^n_j or θ^e_{jk}).

In principle, the F_n real-valued node feature functions $\varphi^s(\mathbf{x}) = (\varphi^s_i(\mathbf{x}))$ and the F_e edge feature functions $\varphi^{st}(\mathbf{x}) = (\varphi^{st}_i(\mathbf{x}))$ could depend on the whole spectral image $\mathbf{x}$. In the present work, however, features are only calculated from a local patch around the site s and the edge $s \sim t$. Note that it is understood that one feature function, *i.e.* $\varphi^s_0(\mathbf{x})$ and $\varphi^{st}_0(\mathbf{x})$, is always the constant value 1. This so-called *bias feature* results in corresponding *bias parameters* that are always added in the dot product even if all feature values are zero.

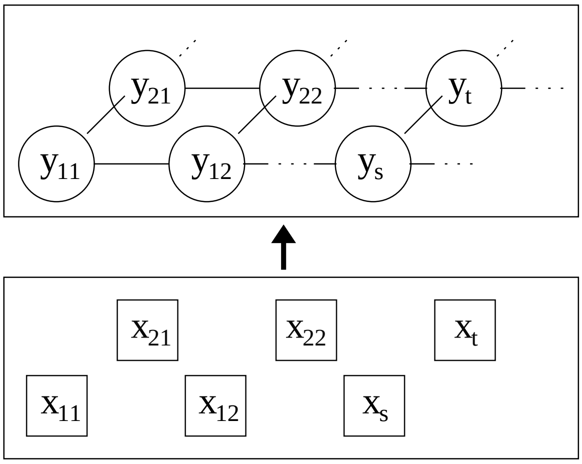


Figure 4.2.: The conditional random field (CRF) is a discriminative model. Given the observations $\mathbf{x}$ the random field over $\mathbf{y}$ has the same graphical structure as the prior model used in the generative approach, however the pair-potentials $V_{st}(y_s, y_t, \mathbf{x})$ evaluate to different values for each edge leading to an inhomogeneous random field. Using squares instead of circles for the observed variables $\mathbf{x}$ indicates that those do not have to be random variables, *i.e.* a distribution over $\mathbf{x}$ is not required.

For example, in a binary classification problem ($\mathcal{Y} = \{1 \dots 2\}$) the potential functions could be written as:

$$V_s(y_s, \mathbf{x}) = \begin{cases} \langle \theta_1^n, \varphi^s(\mathbf{x}) \rangle & \text{if } y_s = 1 \\ \langle \theta_2^n, \varphi^s(\mathbf{x}) \rangle & \text{if } y_s = 2 \end{cases} \tag{4.25}$$

$$V_{st}(y_s, y_t, \mathbf{x}) = \begin{cases} \left\langle \theta_{11}^e, \varphi^{st}(\mathbf{x}) \right\rangle & \text{if } y_s = 1 \text{ and } y_t = 1 \\ \left\langle \theta_{12}^e, \varphi^{st}(\mathbf{x}) \right\rangle & \text{if } y_s = 1 \text{ and } y_t = 2 \\ \left\langle \theta_{21}^e, \varphi^{st}(\mathbf{x}) \right\rangle & \text{if } y_s = 2 \text{ and } y_t = 1 \\ \left\langle \theta_{22}^e, \varphi^{st}(\mathbf{x}) \right\rangle & \text{if } y_s = 2 \text{ and } y_t = 2 \end{cases} \tag{4.26}$$

where the vector θ_{11}^e, for example, summarizes the parameter values θ_{i11}^e for all i of which there are as many as corresponding edge feature functions $\varphi^{st}(\mathbf{x})$. The parameter vectors θ_{12}^e, θ_{21}^e, θ_{22}^e, θ_1^n and θ_2^n are defined in a similar way.

In some sense, the single-site feature function $\varphi_i^s(\mathbf{x})$ takes over the role of the observation model $\mathrm{q}(x_s \,|\, y_s)$ in the generative approach. In fact, if the pairwise interaction terms in the MRF and CRF were removed and using the LDA observation model as well as the identity feature function $\varphi^s(\mathbf{x}) = x_s$, one would just end up with usual LDA versus logistic regression (cf. [85, pp103ff]). If instead, *quadratic* discriminant analysis (QDA) was used, a corresponding discriminative model would be logistic regression with a more complex feature function which calculates all pairwise products, *i.e.* $\varphi^s(\mathbf{x}) = \text{vec}\left[(1x_s^T)(1x_s^T)^T\right]$. A linear threshold in the augmented feature space generated by this feature function thus allows for quadratic boundaries in the original domain of the data x_s like with QDA. The proposed CRF allows for even more complexity in that single-site feature functions $\varphi^s(\mathbf{x})$ are admissible which compute features not only from the local x_s but from a local neighborhood or even the complete image $\mathbf{x}$. Correctly capturing such dependencies in a generative approach would lead to intractable models even without a direct interaction between class labels.

Similarly, the edge feature functions can be chosen so as to increase modelling power. Since high interaction potentials penalize the corresponding configurations of the random field, such realizations a less likely. Using a trainable model also for the edge potentials allows for data-adaptive penalties. Given a suitably chosen set of feature functions such as the channel-wise directional derivatives for spectral images and a set of training examples, this data-dependence as well as the overall coupling strength are automatically inferred and optimized for the given classification task.

In general, the idea of feature functions in combination with dot products is very popular in machine learning. It is also the basic principle behind *kernel methods* [179] such as support vector machines and Gaussian processes. Here, only finite-dimensional feature functions are considered and, unlike for kernel methods, parameter estimation is performed in the primal parameter space as opposed to the Lagrangian dual space. This has the advantage that the learned parameter vectors admit interpretation.

Note that the CRF (Eq. (4.18)) takes a form very similar to the MRF prior in Eq. (4.1). In fact, the MRF prior is obtained as a special case of the CRF by choosing $\varphi_i^s(\mathbf{x}) = \varphi_i^{st}(\mathbf{x}) \equiv 1$, *i.e.* by only using the bias feature. Provided that the edge feature functions $\varphi_i^{st}(\mathbf{x})$ are equal for horizontal and vertical edges, the conditional random field can again be called homogeneous and isotropic since the parameters θ_{ij} and θ_{ijk} are independent of location and direction. However, since the feature functions usually evaluate to different values at each location, the random field $\mathbf{Y} \mid \mathbf{X}$ has anisotropic and inhomogeneous potentials.

Like for the MRF prior, one might want to enforce the symmetry constraint

$$V_{st}(y_s, y_t, \mathbf{x}) - V_{ts}(y_t, y_s, \mathbf{x}) \quad = \quad 0. \tag{4.27}$$

Depending on the features $\varphi_i^{st}(\mathbf{x})$, however, this does not directly translate into a symmetry constraint on θ_{ijk} as for the MRF. However, given a certain symmetry of $\varphi_i^{st}(\mathbf{x})$, Eq. (4.27) can be enforced by constraints on θ_{ijk}. In particular,

$$\begin{array}{rclcrcll} \varphi_i^{st}(\mathbf{x}) - \varphi_i^{ts}(\mathbf{x}) & = & 0 & \Rightarrow & \theta_{ijk} - \theta_{ikj} & = & 0 & \text{(symmetric)} \\ \varphi_i^{st}(\mathbf{x}) + \varphi_i^{ts}(\mathbf{x}) & = & 0 & \Rightarrow & \theta_{ijk} + \theta_{ikj} & = & 0 & \text{(antisymmetric)} \end{array} \tag{4.28}$$

Hence, the symmetric constraint is appropriate for the bias whereas the antisymmetric constraint should be used for spatial derivative features.

Note that the distribution described by Eq. (4.18) is an exponential family and therefore exhibits many favorable properties [208, 24]. For example, maximum likelihood (ML) parameter estimation yields a convex optimization problem and, therefore, problems due to local optima are avoided (cf. section 4.3.2).

4.3. Parameter Estimation

Given independent identically distributed (iid) training data, *i.e.* a set of hand-segmented spectral images, parameter estimates with optimal generalization performance are sought. Usually, a training data set with observed realizations is available. Here, the standard approach is extended by also allowing for *soft evidence* such that beliefs over class labels [151] can be specified. For example, an expert's statement "tumor with 20% probability and healthy with 80%" can directly be translated into the belief $y_s = (.2, .8)$. Hard labels are included in this framework as the extreme beliefs $y_s = (1, 0)$ and $y_s = (0, 1)$. Maximum likelihood (ML) estimation with soft evidence is most conveniently stated using expectations with respect to *empirical distributions* $\mathrm{E}_{\tilde{\mathrm{p}}}[\cdot]$ which are defined by the observed data set $\mathcal{D}$ (cf. appendix D).

4.3.1. Generative Approach

An important difference between generative and discriminative models is that the former are trained by *joint likelihood* whereas the latter are trained by conditional likelihood. Thus, ML parameters for generative models are found by maximizing

$$\begin{aligned} l_j(\theta, \vartheta) &= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{x},\mathbf{y})}[\log \mathrm{q}(\mathbf{x}, \mathbf{y} \,|\, \theta, \vartheta)] && (4.29)\\ &= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{x},\mathbf{y})}[\log \mathrm{q}(\mathbf{x} \,|\, \mathbf{y}, \vartheta)] + \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{y})}[\log \mathrm{q}(\mathbf{y} \,|\, \theta)]. && (4.30) \end{aligned}$$

Hence, joint likelihood for generative models decomposes into two terms which allow to perform training of the observation model $\mathrm{q}(\mathbf{x} \,|\, \mathbf{y}, \vartheta)$ and the prior model $\mathrm{q}(\mathbf{y} \,|\, \theta)$ independently. For hard labels (cf. appendix D), this amounts to maximizing

$$l_j(\theta, \vartheta) = \frac{1}{N} \sum_{i=1}^{N} \left(\log \mathrm{q}(\mathbf{x}_i \,|\, \mathbf{y}_i, \vartheta) + \log \mathrm{q}(\mathbf{y}_i \,|\, \theta) \right). \qquad (4.31)$$

Since the prior model in the generative approach is just a special case of the CRF model (compare Eq. (4.1) to Eq. (4.18)), training of the prior model can be performed with the same algorithms which are described in the following section.

Training the observation model in the generative approach is no harder than training a single-voxel model due to the decomposition of the joint likelihood objective. For the LDA and LDAi models this is done by calculating the (weighted) mean for each class followed by the estimation of a full, respectively isotropic covariance matrix from the pooled data. For the remaining models, PPCA, MoG and MoPPCA, maximum (soft-) likelihood parameters have been obtained with the expectation maximization (EM) algorithm [194].

4.3.2. Discriminative Approach

Unlike generative models, discriminative models are trained by *conditional likelihood*, *i.e.* by maximizing

$$\begin{aligned} l_c(\theta) &= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{x},\mathbf{y})}[\log \mathrm{p}(\mathbf{y} \,|\, \mathbf{x}, \theta)] && (4.32)\\ &= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{x})}[\mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{y} \,|\, \mathbf{x})}\,[\log \mathrm{p}(\mathbf{y} \,|\, \mathbf{x}, \theta)]]\,, && (4.33) \end{aligned}$$

which for hard labels becomes

$$l_c(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log \mathrm{p}(\mathbf{y}_i \,|\, \mathbf{x}_i, \theta). \tag{4.34}$$

Due to the high correlation and collinearities intrinsic to spectral data, plain maximum likelihood usually fails to provide good parameter estimates in the discriminative case. Therefore, a Bayesian approach is adopted by introducing a zero-mean Gaussian prior over the parameter vector and using the maximum a posteriori (MAP) estimate instead. Thus, the objective function to maximize becomes

$$l_c^{MAP}(\theta) = l_c(\theta) - \frac{\theta_n^T L_n \theta_n}{2\sigma_n^2} - \frac{\theta_e^T L_e \theta_e}{2\sigma_e^2} \tag{4.35}$$

where the L_n and L_e are precision matrices which can be chosen so as to favor small and smooth parameter coefficient profiles in the style of a generalized Tikhonov regularizer. For example, in the case of spectral data, a good regularizer is obtained from the positive semidefinite matrix $L = D_k^T D_k$ where D_k is the k-th order difference operator. The overall strength of the edge and node priors can be adjusted individually using the parameters σ_n^2 and σ_e^2 which reduce to Gaussian variances for $L_n = L_e = \mathbf{I}$.

Since the described random field models describe exponential families, their log partition functions are convex in θ [24]. Furthermore, the gradient of the log partition function equals the expected value of the sufficient statistics (cf. appendix C, Lemma 1), leading to

$$\begin{aligned}
\nabla l_c(\theta) &= \frac{\partial}{\partial\theta} l_c(\theta) && (4.36)\\
&= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{y},\mathbf{x})}\left[\frac{\partial}{\partial\theta}\langle\theta, \Phi_{\mathbf{x}}(\mathbf{y})\rangle\right] - \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{y},\mathbf{x})}\left[\frac{\partial}{\partial\theta}A(\theta)\right] && (4.37)\\
&= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{y},\mathbf{x})}[\Phi_{\mathbf{x}}(\mathbf{y})] - \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{x})}\left[\mathrm{E}_\theta[\Phi_{\mathbf{x}}(\mathbf{y})]\right] && (4.38)\\
&= \mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{x})}\left[\mathrm{E}_{\tilde{\mathrm{p}}(\mathbf{y}\,|\,\mathbf{x})}[\Phi_{\mathbf{x}}(\mathbf{y})] - \mathrm{E}_\theta[\Phi_{\mathbf{x}}(\mathbf{y})]\right]. && (4.39)
\end{aligned}$$

Hence, the gradient is proportional to the expected difference between the expectation of the sufficient statistics under the empirical distribution and its expectation under the model parameterized by θ. The first term is just a summation over the sample, which is easily calculated and has to be done only once since it does not depend on θ. Problems arise from the calculation of the second term, the expectation $\mathrm{E}_\theta[\Phi_{\mathbf{x}}(\mathbf{y})]$, and the log partition function $A(\theta, \mathbf{x})$ (Eq. (4.34)) which both involve summations over a number of states that grows exponentially with the number of sites $|S|$. Various algorithms have been proposed to approach this problem, among which are Monte Carlo methods [213, 162], variational and mean field methods [75] and message passing methods [217, 107, 207, 151].

For efficiency and simplicity, asynchronous belief propagation (BP) has been used, *i.e.* BP with a random message passing schedule [218]. In its sum-product version, BP can be used to obtain approximate marginals for a given random field (or more general a factor graph) with unknown partition function [107], *i.e.* it attempts to calculate

$$Z \cdot \mathrm{p}(y_s \,|\, \mathbf{x}, \theta) = \sum_{y_1}\sum_{y_2}\dots\sum_{y_{s-1}}\sum_{y_{s+1}}\dots\sum_{y_{|S|}} \exp\left[\langle\theta, \Phi_{\mathbf{x}}(\mathbf{y})\rangle\right] \qquad (4.40)$$

where Z is now easily calculated as normalizer for $\mathrm{p}(y_s \,|\, \mathbf{x}, \theta)$. A similar formula holds for the marginals over pairs of sites (y_s, y_t). With a single run of BP *all* site and pairwise marginals are obtained. This works by repeatedly passing messages from a node to its neighbors until *local consistency* is obtained. For tree-structured graphs, local consistency implies *global consistency* and BP is known to be exact then. Details and more references can be found, for example, in Yedidia *et al.* [218, 217] who also showed that the stationary points of BP are also stationary points of the

Bethe free energy. Thus, different algorithms could be used to find local optima of the Bethe free energy which might give better performance. However, the choice and influence of using different algorithms for obtaining approximate marginals has not been examined in the present work.

Upon convergence of BP, the beliefs $\mathrm{b}_{st}(y_s, y_t)$ and $\mathrm{b}_s(y_s)$ can be used to replace the exact marginals of $\mathrm{p}(\mathbf{y} \,|\, \mathbf{x}, \theta)$. Using the Bethe approximation of the loglikelihood, the calculation of the log partition function for Eq. (4.34) is avoided since

$$\log \mathrm{p}(\mathbf{y} \,|\, \mathbf{x}, \theta) \approx \sum_s (1 - d_s) \log \mathrm{b}_s(y_s) + \sum_{s \sim t} \log \mathrm{b}_{st}(y_s, y_t) \tag{4.41}$$

where d_s denotes the number of sites connected to site s [218, 217]. Furthermore, the same set of beliefs can be used for an approximation of the expected sufficient statistics. In particular,

$$\mathrm{E}_\theta[\phi^n_{ij}(\mathbf{y}, \mathbf{x})] \approx \sum_s \sum_{y_s} \mathrm{b}_s(y_s) \varphi^s_i(\mathbf{x}) \delta_j(y_s) \tag{4.42}$$

$$\mathrm{E}_\theta[\phi^e_{ijk}(\mathbf{y}, \mathbf{x})] \approx \sum_{s \sim t} \sum_{y_s, y_t} \mathrm{b}_{st}(y_s, y_t) \varphi^{st}_i(\mathbf{x}) \delta_j(y_s) \delta_k(y_t). \tag{4.43}$$

Stacked like $\Phi_{\mathbf{x}}(\mathbf{y}) = (\phi^n_{ij}(\mathbf{y}, \mathbf{x}), \phi^e_{ijk}(\mathbf{y}, \mathbf{x}))$, these expectations allow the calculation of the gradient in Eq. (4.39).

Using these approximations, $l^{MAP}_c(\theta)$ and its gradient are ready for use in any gradient-based optimization method and summarized in Fig. 4.3. Here, the limited memory BFGS [44] method has been used, a quasi-Newton method which has been designed to cope with large numbers of variables. This is particularly useful for CRFs since the discriminative approach draws its power and flexibility from the definition of a large number of feature functions resulting in a correspondingly large number of parameters.

4.4. Segmentation and Classification

Given a set of trained parameters the posterior class label distributions $\mathrm{p}(\mathbf{y} \,|\, \mathbf{x})$ (CRF) and $\mathrm{q}(\mathbf{y} \,|\, \mathbf{x}) \propto \mathrm{q}(\mathbf{x} \,|\, \mathbf{y})\,\mathrm{q}(\mathbf{y})$ (MRF) could in principle be calculated for any previously unseen image $\mathbf{x}$. Algorithmic problems arise again from the partition function. From a decision-theoretic point of view this posterior contains the max-

```
Data: current parameters $\theta^{(k)} = (\theta_n^{(k)}, \theta_e^{(k)})$
Result: objective $f$, gradient $g$
$f = 0$;
$g = 0$;
for all training images do
    calculate beliefs $\mathrm{b}_s(y_s)$ and $\mathrm{b}_{st}(y_s, y_t)$ using current $\theta^{(k)}$;
    $f \;=\; f + \sum_s (d_s - 1) \sum_{y_s} w_s(y_s) \log \mathrm{b}_s(y_s)$
          $- \sum_{s \sim t} \sum_{y_s} \sum_{y_t} w_s(y_s) w_s(y_t) \log \mathrm{b}_{st}(y_s, y_t)$
    $g_{ij}^n \;=\; \sum_s \sum_{y_s} (\mathrm{b}_s(y_s) - w_s(y_s)) \varphi_i^s(\mathbf{x}) \delta_j(y_s)$
    $g_{ijk}^e \;=\; \sum_{s \sim t} \sum_{y_s, y_t} (\mathrm{b}_{st}(y_s, y_t) - w_s(y_s) w_t(y_t)) \varphi_i^{st}(\mathbf{x}) \delta_j(y_s) \delta_k(y_t)$
    $g \;=\; g + \left(g_{ij}^n, g_{ijk}^e\right)$
end
$f \;=\; f + \frac{\theta_n^T L_n \theta_n}{2\sigma_n^2} + \frac{\theta_e^T L_e \theta_e}{2\sigma_e^2}$
$g^n \;=\; \sigma_n^{-2} L_n \theta_n^{(k)}$
$g^e \;=\; \sigma_e^{-2} L_e \theta_e^{(k)}$
$g \;=\; g + (g^n, g^e)$
```

Figure 4.3.: Pseudo-code for calculating the objective/gradient to be minimized for CRF training with soft evidence ($\mathcal{D} = \{w_j(x)\}_{j=1}^N$, cf. appendix D), *i.e.* the conditional likelihood. Note that the same code can be used for training the prior model of the MRF by using the bias feature only (cf. section 4.2.2).

imum amount of information and allows for a Bayes-optimal decision w.r.t. to any associated cost function $C(\hat{\mathbf{y}}, \mathbf{y})$, *i.e.* an estimate $\hat{\mathbf{y}}$ that minimizes the Bayes-risk

$$R(\hat{\mathbf{y}}) \;=\; \int C(\hat{\mathbf{y}}, \mathbf{y})\, \mathrm{p}(\mathbf{y} \,|\, \mathbf{x})\, \mathrm{d}\mathbf{y}. \tag{4.44}$$

Using the *zero-one loss function*

$$C(\hat{\mathbf{y}}, \mathbf{y}) \;=\; 1 - \delta(\hat{\mathbf{y}} - \mathbf{y}), \tag{4.45}$$

it can be shown that the Bayes estimate is the Maximum A Posteriori (MAP) estimate, which is the mode of the posterior distribution [213, p.27], [119], *i.e.*

$$\hat{\mathbf{y}}^{\mathrm{MAP}} \;=\; \underset{\mathbf{y}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{y} \,|\, \mathbf{x}) \;=\; \underset{\mathbf{y}}{\operatorname{argmax}} \log \mathrm{p}(\mathbf{y} \,|\, \mathbf{x}). \tag{4.46}$$

The maximization of such an energy of a large number of discrete variables is a well-studied standard problem in discrete optimization. Various algorithms can be used under appropriate conditions [121, 79, 38, 93, 105, 37].

The zero-one loss function, however, might be too restrictive for images since it is only 0 for an estimated label map that exactly agrees with the ground truth. An alternative is the *error rate loss function*

$$C(\hat{\mathbf{y}}, \mathbf{y}) \quad = \quad \frac{1}{|S|} \sum_{s \in S} (1 - \delta(\hat{y}_s - y_s)) \tag{4.47}$$

which quantifies the number of erroneously classified voxels. In addition, this loss function is better suited for a comparison with single voxel approaches. The Bayes estimate for the error rate is the Marginal Posterior Mode Estimate (MPME) [213, p.28], *i.e.*

$$\hat{y}_s^{\mathrm{MPME}} \quad = \quad \operatorname*{argmax}_{y_s} \mathrm{p}(y_s \,|\, \mathbf{x}). \tag{4.48}$$

The marginals $\mathrm{p}(y_s \,|\, \mathbf{x})$ (cf. Eq. (4.40)) can be determined with the same algorithms as in section 4.3.2 and again asynchronous belief propagation was used.

A further advantage of using marginals is that instead of displaying the MPME, the marginal probabilities can be displayed in probability maps which convey more information than the MPME point estimates.

Finally, since for given data $\mathbf{x}$ both CRF and MRF are represented by the same factor graph [107], MAP as well as MPME estimation does not differ at all between the discriminative and the generative approach. Thus, although MRFs are less difficult to train than CRFs (due to the decomposition of the likelihood function), the prediction of class labels given unseen data poses exactly the same computational burden. The cost for powerful modelling and more flexibility with CRFs is entirely paid during the training phase.

4.5. Experimental

4.5.1. Simulated Data

Artificial MRSI data was generated for the purpose of a Monte Carlo study using the common FID model, a superposition of K exponentially decaying complex sinusoidals (see also appendix A):

$$s(t) \quad = \quad \exp\left(j\phi_0\right) \sum_{k=1}^{K} a_k \exp\left(j\phi_k + j\, 2\pi f_k t - d_k t\right) \tag{4.49}$$

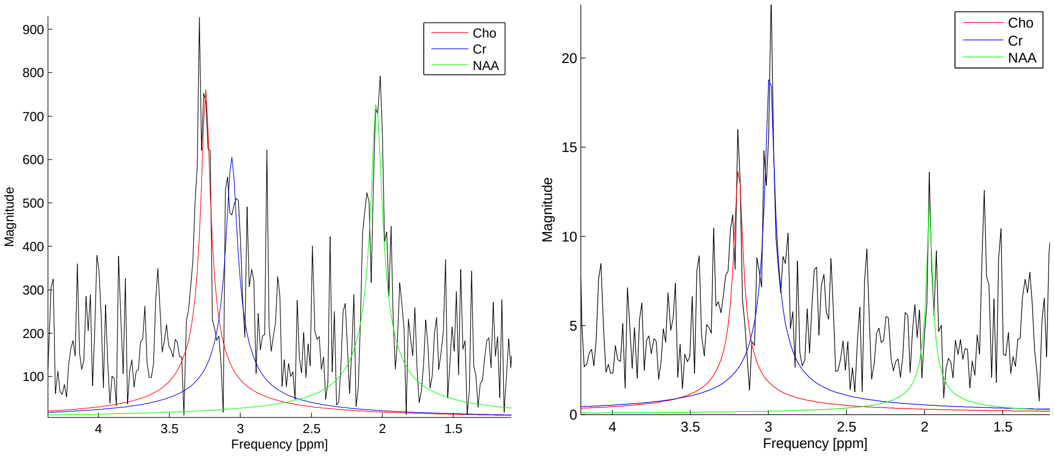
with the imaginary unit j and where a_k denotes the amplitude of the k^{th} component, f_k the frequency, ϕ_k the phase shift, and d_k the damping which determines the (Lorentzian) line width.

In order to obtain realistic tumor sizes and shapes, ground truth was generated based on available recordings from three patients. Ground truth is required to train the different classifiers and allows to calculate various error measures for their evaluation. Plausible high-resolution concentration maps of N-acetylaspartate (NAA), choline (Cho) and creatine (Cr) were created manually. The concentration of NAA was decreased in the tumor whereas the Cho concentration was increased. Altough Cr varied spatially it was not altered specifically within the tumor (cf. Fig. 4.4).

Downsampled concentration maps at various resolutions (8×8 up to 64×64) were calculated from the manually created high-resolution maps by averaging the concentration values within each low-resolution voxel. Spatially resolved FIDs were then generated using the FID model in Eq. (4.49) where the amplitudes were chosen proportional to the concentration of the respective metabolites. NAA was modeled with a frequency shift of 171.2Hz, Cr with 107.5Hz and Cho with 94.2Hz. A damping of $d_k = 12$Hz and no phase shift ($\phi_0 = \phi_k = 0$) was applied to all the resonances. Finally, complex-valued white isotropic Gaussian noise with mean zero was added to yield MRSI data with specific and known SNR. In Fig. 4.4 a real data example is contrasted with a simulated noisy spectrum (cf. section 4.5.1).

4.5.2. Feature Extraction

From the complex-valued FID signals, suitable feature vectors were extracted with the following steps (for details see [4, 7] or [222]). Fist, residual water and nuisance peaks were removed with time-domain selective filtering by means of a Hankel Singular Value Decomposition (HSVD) (cf. appendix A). Apodization with an exponential



(a) patient data

(b) simulated data

Figure 4.4.: Example for MRSI magnitude spectra from a patient with brain tumor and from simulation.

function and removal of the first few sampling points of the FID suppressed noise and baselines. Then, the filtered FID signal was Fourier transformed and magnitude spectra were calculated. Using magnitude spectra instead of real spectra avoids problems due to the unknown null-phase which is difficult to reliably estimate from noisy MRSI data [4]. Then, the spectra were truncated to the frequency band that contains the interesting metabolite resonances. Finally, the spectral pattern was L_1-normalized. The latter is necessary since MRSI signal amplitudes may vary not only due to metabolite concentration changes but also due to the spatially varying sensitivity of the receiving coils.

For the CRF, also edge features ϕ^e are used. Since a linear combination of the edge features determines the pairwise coupling strength of the labels, these are designed to contain edge information. The response of a derivative filter appears to be suitable. Here, Sobel-like filters are used [95, 96], *i.e.* each channel of the spectral image first is smoothed with the Binomial filter [1 2 1] in one direction and then convolved with the first order difference filter [1 -1] in the orthogonal direction.

4.6. Results

Training was performed based on six simulated spectral images with 32 × 32 voxels and 16% noise level (SNR = 1/.16). For the CRF, a parameter prior was derived

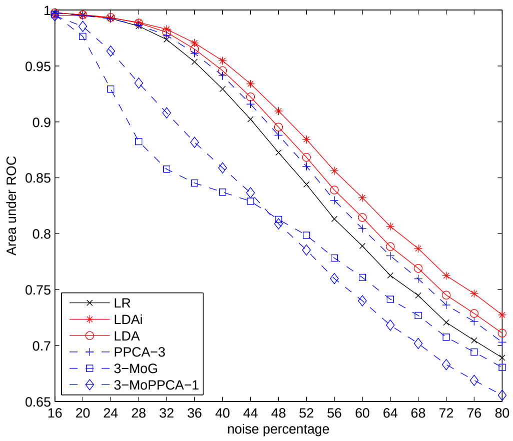


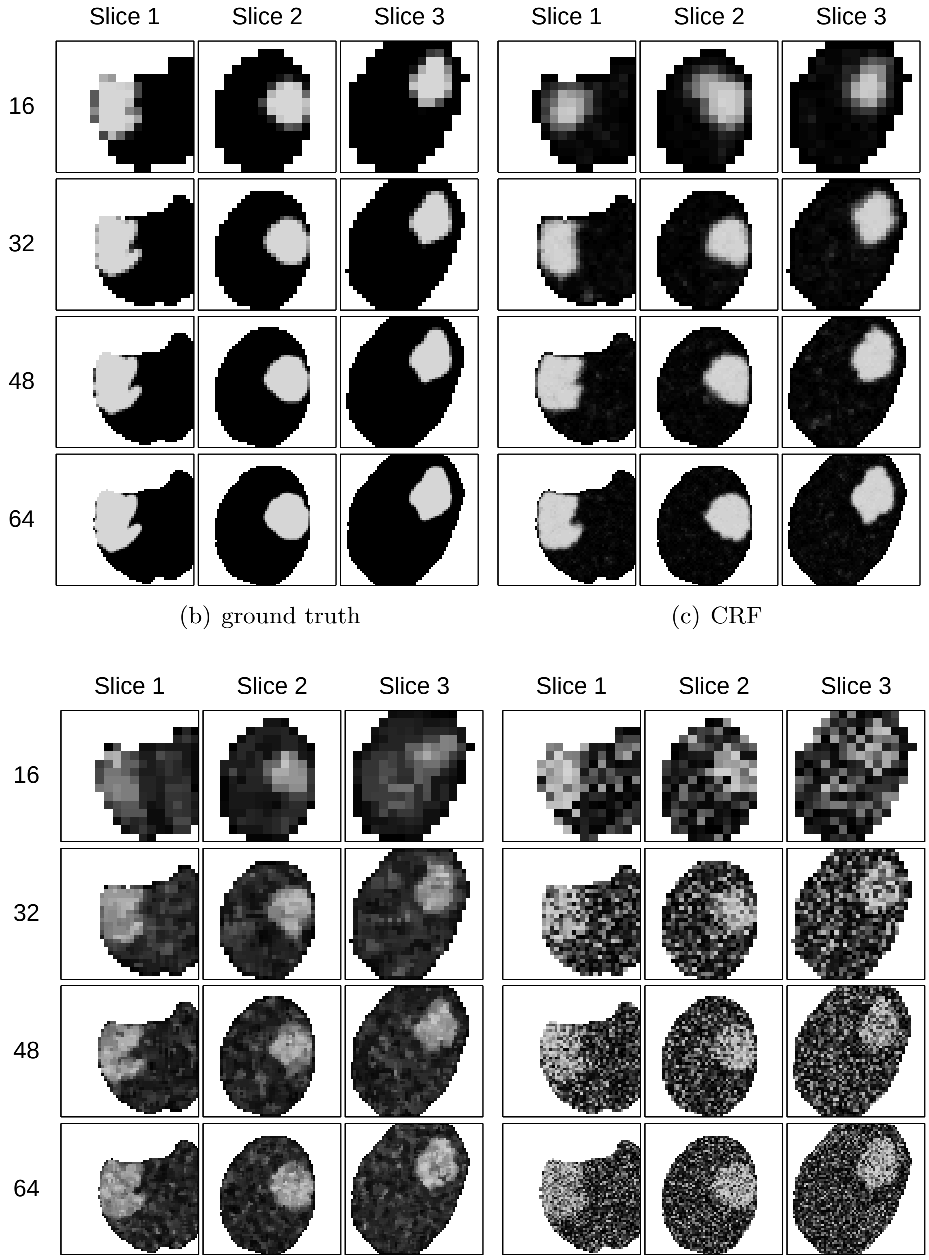
Figure 4.5.: Generative versus discriminative methods in a single voxel approach: mean area under the receiver operator characteristic.

from the second order difference operator and the hyperparameters σ_n and σ_e were determined so as to maximize the loglikelihood on an independent validation set. For the MRF, the global coupling strength of the homogeneous prior model was determined in the same way. Training was repeated 16 times with fixed hyperparameters on independent data sets yielding 16 different parameter estimates for each model.

For testing, 18 spectral images with varying resolutions and SNR were used. From the estimated marginal posteriors the marginal posterior mode estimate (MPME) and the threshold-independent area under the receiver operator characteristics (AUC) were determined for each image.

In Fig. 4.5, the AUCs for the generative and discriminative approaches without spatial coupling are compared. Training the CRF without spatial coupling is equivalent to using logistic regression (LR) which represents the discriminative approach. For the generative approach five different observation models were evaluated (cf. section 4.2.1). Three latent variables have been used for the respective observation models since this corresponds to the number of metabolite resonances present in the spectral patterns. According to Fig. 4.5 all models exhibit a similar performance loss with increasing noise level and only the mixture models show a slight disadvantage.

Comparing the ground truth of the simulated MRSI slices and the marginal posterior maps obtained with LR and the CRF in Fig. 4.6 suggests that using spatial context improves the estimate at all resolutions. Although posthoc median filtering can improve the results from logistic regression considerably, the CRF clearly recovers the ground truth more accurately. A further analysis of posthoc smoothing has therefore not been pursued.



(b) ground truth

(c) CRF

(e) LR median filtered

(f) LR

Figure 4.6.: Ground truth and estimated tumor probabilities (marginals of the posterior distribution) from logistic regression, median filtered logistic regression (3 × 3 mask) and CRF for three simulated MRSI slices at resolutions 16, 32, 48 and 64 with 48% noise.

The advantage of the CRF over LR is confirmed by the numerical results presented in Fig. 4.8(a). In addition to the median AUC also the .45, .35 and .25 quantiles are reported which provide an indication for the spread of the 16 × 18 AUC values. Fig. 4.8(a) also shows that, as expected, the performance of LR remains unchanged for different resolutions. Unlike that, the CRF clearly profits if a resolution of about 20 × 20 voxels is exceeded.

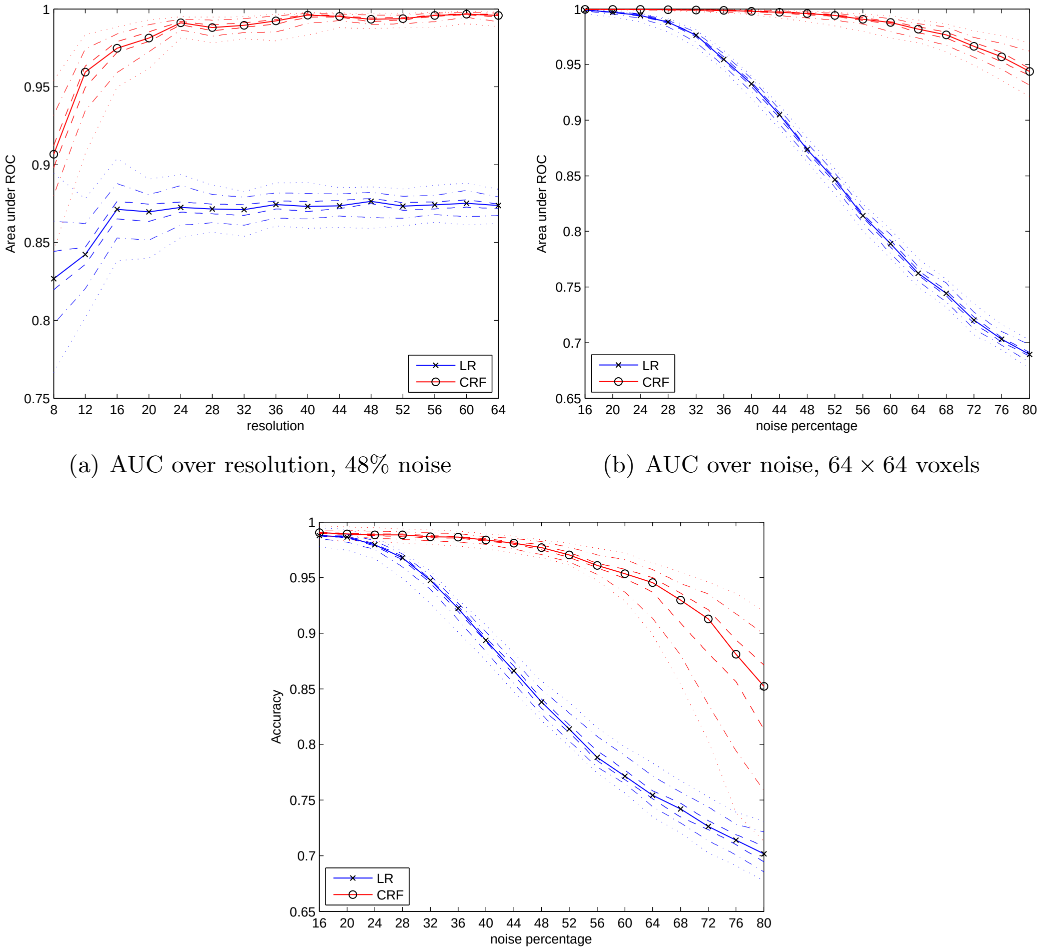
According to Fig. 4.8(b), which shows the same statistics for increasingly noisy test images, the CRF maintains an excellent AUC even at higher noise levels. However, beyond a level of about 60% noise the lower quartile of the AUC distribution starts to degrade rapidly which indicates that the CRF severely breaks down for some images. A similar behavior is observed for the accuracy in Fig. 4.8(c), though it seems to be more sensitive in that the lower quantile starts to break down earlier and faster. Still a clearly significant gain of up to about 15% accuracy could be reached on the simulated data.

With the generative approach the same principal behavior is observed and equivalent results are provided in Fig. 4.8. Using the MRF improves the performance for all observation models up to a certain noise level. Beyond that level the AUC degrades rapidly and the MRF prior even harms.

Figure 4.9 compares the performance of the CRF and the different MRFs. On the average, the CRF can achieve a slightly higher AUC for all noise levels. With respect to the mean error, however, the CRF hardly shows advantages. Its error rate is significantly lower than that of all other approaches only between the resolutions of 44 and 64.

A more realistic trade-off between resolution and noise was simulated by adjusting both simultaneously under the assumption of constant scan times. Figure 4.10 shows the performance of single voxel LR versus spatial CRF when the noise level is increased with the resolution such that the ratio noise/resolution^2 remains constant (cf. [118]). For resolutions between 32×32 voxels (36% noise) to 40×40 voxels (56% noise) the CRF shows a clear advantage over LR.

Figure 4.11 presents the edge weights θ^e (without the bias parameter) that have been learned from the training data. Note that an antisymmetric constraint on the parameters corresponding to the directional derivative features and a symmetric constraint on the bias parameters has been used. The edge parameter vectors can be interpreted and used to identify those spectral channels that determine spectral edges as defined by the label map.



(a) AUC over resolution, 48% noise

(b) AUC over noise, 64 × 64 voxels

(c) Accuracy over noise, 64 × 64 voxels

Figure 4.7.: Discriminative approach: Accuracy and area under curve (AUC) for single voxel logistic regression (LR) and conditional random field (CRF). In addition to the median, the .45, .35 and .25 quantiles are shown to indicate the spread of the distribution.

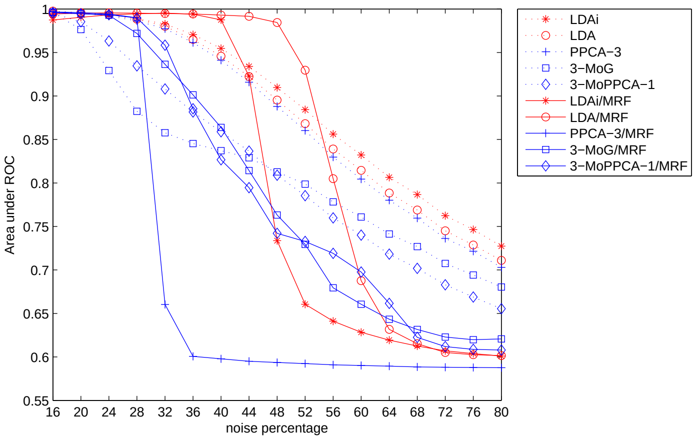
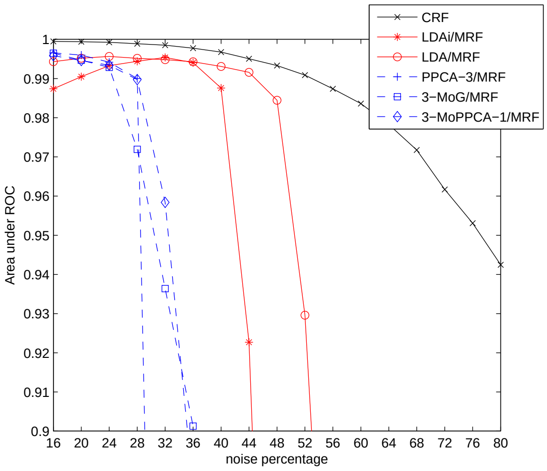


Figure 4.8.: Generative approach: mean area under ROC comparing models with and without spatial MRF prior.

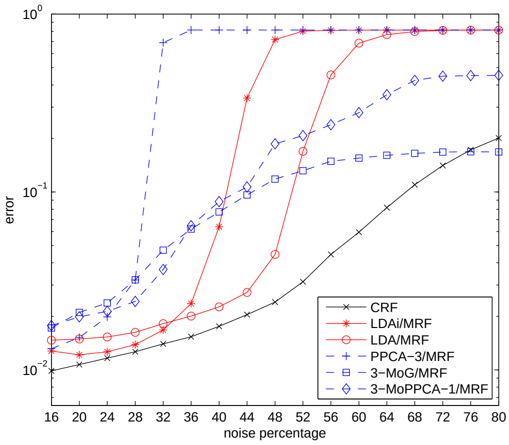
4.7. Discussion

At first sight, using simulated data for the comparison of generative and discriminative models seems to be a bad idea. Since the data generating process is known, the generative model should always perform better. Despite the nonlinear feature extraction steps that clearly result in non-Gaussian noise in the feature vector, the LDA model with isotropic covariance (LDAi) resembles the true distribution very well. This is also reflected in the similar performance of LR, LDAi, LDA and PPCA-3 in Fig. 4.5 where the generative models tend to show better performance. The potential of representing multi-modal distributions does not help the mixture models (3-MoG, 3-MoPPCA-1) to superior performance. On the contrary, the MoG shows a quite unstable and atypical behavior as compared to the other methods. A reason could be that the MoG is a generative model that does not resemble at all the true data generating process.

As soon as spatial context is considered (Figs. 4.8 and 4.10(a)), the relative performance changes. Most notably, the performance of the PPCA-3/MRF model breaks down first and, different from the single voxel case, is not superior to the mixture models. Furthermore, the LDA/MRF model with full covariance matrix now performs best. This reflects the fact that the true distribution, considering spatial correlations, is not very well captured by the isotropic and homogeneous MRF prior which can partially be compensated by more complex observation models. Unlike that, the CRF better models the spatial label distribution and thus its performance gains most from including spatial context.

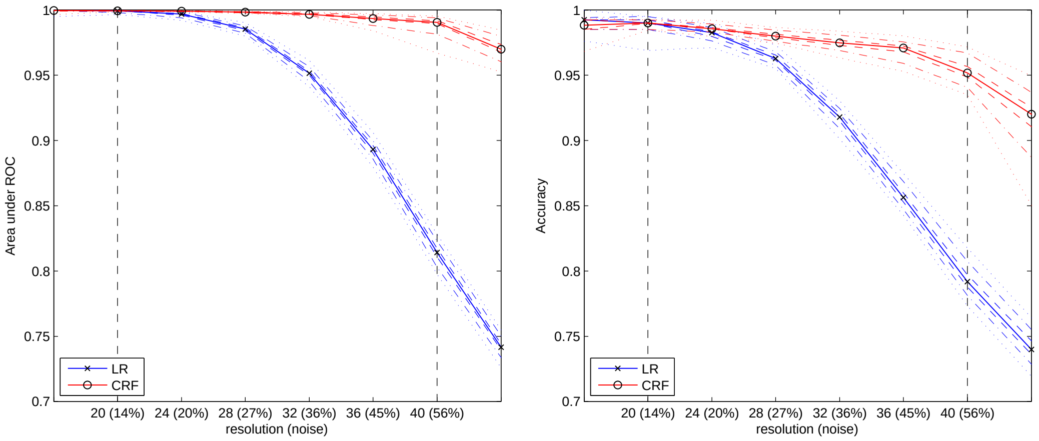


(a) Mean AUC



(b) Mean Error

Figure 4.9.: Comparison of AUC and error obtained with LR and CRF for increasingly noisy signals.



(a) AUC

(b) Accuracy

Figure 4.10.: Performance of LR and CRF when the noise level increases as the square of the resolution (noise/resolution^2 $= 56\%/40^2$).

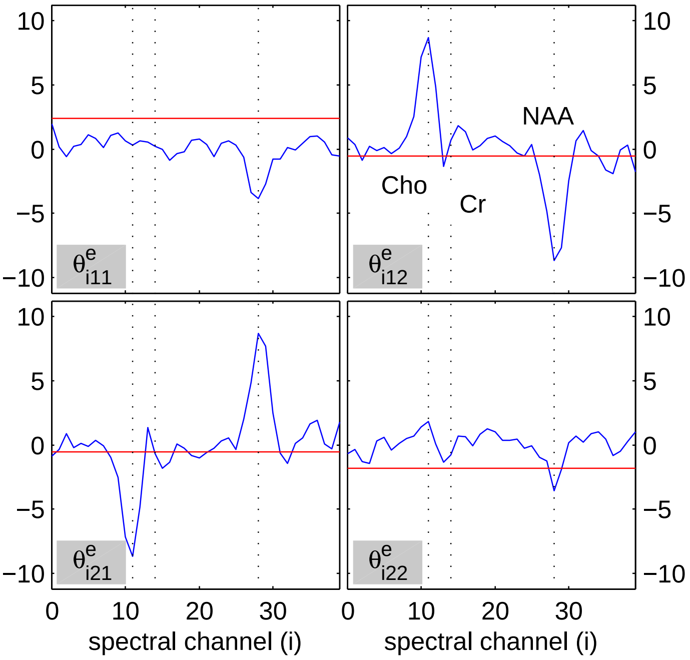


Figure 4.11.: Edge weights θ^e without bias learned with the CRF. Since the weights correspond to spectral channels (blue curves), spectral regions can be identified that determine an edge and make a spatial label change more probable. The horizontal red lines show the constant bias.

Although locally adaptive priors have been proposed (*e.g.* [56]), such modifications lead to solutions not modeled by the original MRF since the local weights either have to depend on the observed data or have to be included in the original prior. Unlike that, the proposed CRF constitutes a sound probabilistic model for incorporating data-dependent edge weights.

Figures 4.7 and 4.10 show that CRFs can indeed be used to increase the acceptable signal noise and thus record MRSI at higher resolutions. However, beyond a certain noise level the results obtained with a random field prior have to be regarded with care since the variance of the predictions gets very big. For example, for the second slice at resolution 16 in Fig. 4.6, the CRF erroneously adds a small additional region to the top left of the tumor region. However, in order to examine the break-down behavior of the different methods, the simulated noise levels have been extremely high. In real applications, an optimal noise level has to be found that still allows for reliable evaluation using the proposed approach.

In general, it is not clear how an "edge" should be defined for spectral images, in particular, it is not clear how to combine the derivative information from different spectral channels. Using the CRF, such information can be learned from labeled images which implicitly define what the notion of an edge actually means. For the simulated MRSI data, an edge is identified in places where the magnitudes of the Cho and NAA resonances change a lot between neighboring pixels as apparent from Fig. 4.11. Within the homogeneous regions the spatial derivatives can be expected to be close to zero, resulting in a potential matrix that is only determined by the bias parameters. At transitions from tumor (class 1) to healthy tissue (class 2) the Cho resonance decreases whereas the NAA resonance increases. A dot product of the resulting directional derivatives with θ^e_{i12} results in a negative value which is subtracted from the bias and results in a potential matrix that favors a transition from the first to the second class and makes the opposite transition, *i.e.* class two to one (θ^e_{i21}), less probable. Hence, the smoothing strength at spectral edges is reduced as desired. In light of the flat curves obtained for the within class transitions (θ^e_{i11} and θ^e_{i22}) it might be reasonable to explicitly constrain these parameters to zero during CRF training.

4.8. Summary

In the present chapter, pattern recognition approaches that include spatial context for the classification of individual voxels have been introduced and compared. Using probabilistic graphical models, a family of generative and analogous discriminative approaches for the combined segmentation and classification of spectral images have been proposed.

In particular, a family of generative models has been constructed by defining a discrete-valued homogeneous Markov random field prior of the lattice-structured label map together with five different observation models of different complexity and particularly suited for spectral data. Three observation models have been conditionally Gaussian with differently constrained covariance matrices (full, isotropic and probabilistic principal components analyzer (PPCA)) and two observation models have been conditional mixture models (isotropic Gaussian and PPCA).

For the analogous discriminative approach, a conditional random field (CRF) has been defined. Unlike the generative models, the CRF allows for an adaptive coupling strength between individual voxels in the label map. This allows to learn which spectral channels define an edge from labeled training data and prevent label smoothing across such edges. Unlike with the generative models, feature information can be extracted from the whole image and used with the CRF in a sound way. Since the CRF model has been designed as an exponential family in natural parameters, the parameter estimation problem is convex. Furthermore, with a single run of belief propagation, marginals are obtained that can be used to approximate the likelihood function (Bethe) as well as to calculate its gradient. Finally, the resulting parameter vectors allow for interpretation and can be used to define a weighing of the spectral channels of a directional derivative that indicate a spectral edge.

Since CRFs incorporate information from a local neighborhood and thus perform a kind of local averaging, data with more noise can be processed. Using a CRF could improve classification accuracy by up to 15% as compared to the single voxel approach (LR) and even more improvement is obtained for the threshold-independent area under the receiver operator characteristic (AUC). This in turn allows to use MRSI at higher resolutions.

Chapter 5.

Conclusion

IMAGING modalities that allow conclusions to be drawn about physiological processes increasingly gain importance for clinical purposes. Two such modalities from the family of nuclear magnetic resonance imaging (MRI) techniques have been examined in the present work, namely dynamic contrast-enhanced MRI (DCE-MRI) which reveals permeability and perfusion properties of the depicted tissue, and magnetic resonance spectroscopic imaging (MRSI) which is a metabolic imaging technology and can thus be used to map the concentration of certain biomolecules.

In order efficiently access their diagnostic information, both modalities require the processing of vector-valued image data. In the present work two approaches to such a processing have been identified. Given a set of evaluated example images, the *pattern recognition* approach tries to imitate the decision rules applied by an expert (physician), thus it is data-oriented. In contrast, the *quantification* approach is model-oriented in that the data is evaluated based on a physical model which usually requires the fitting of a nonlinear function to the observed data. Clearly, the quantification approach incorporates more specific prior knowledge and does not require a training data set whereas the pattern recognition approach is more powerful when it comes to effects in the data that cannot be well modeled but are empirically considered by an expert.

Both approaches are usually applied in a voxelwise fashion. A comparison of voxelwise quantification and pattern recognition approaches on a clinical problem, namely the estimation of tumor probability in prostate MRSI, has been presented in chapter 2. An extensive collection of linear and nonlinear classifiers as well as common quantification algorithms have been systematically evaluated.

The present thesis has proposed methods to enhance the voxelwise evaluation with considering spatial context. In chapter 3, a Generalized Gaussian Markov Random Field (GGMRF) prior over parameter maps has been proposed for quantification. Its application to DCE-MRI as well as MRSI has been examined and found to provide better solutions. A bias-variance decomposition using simulated MRSI data has shown that the GGMRF prior significantly reduces variance and hardly introduces bias. This results in reduced mean squared error (MSE) for the parameter estimates. Using the GGMRF resulted in a lower sum of squared residuals (SSR) for the DCE-MRI data, indicating that the prior also helps the optimization routine to avoid local optima of the model fit function. An efficient optimization strategy has been proposed that is a blocked version of the iterated conditional modes algorithm (Block-ICM) which converges significantly faster than conventional ICM.

Incorporating spatial context using random fields for the purpose of pattern recognition has been examined in chapter 4. There, an analogous family of generative and discriminative models for joint segmentation and classification of spectral images have been proposed. It has been shown that including spatial context can significantly improve classification accuracy and the area under the receiver operator characteristic (AUC) up to a certain noise level. Using the discriminative approach, random fields that perform anisotropic and inhomogeneous label smoothing can be learned using a sound probabilistic model. Furthermore, the resulting weight vectors allow for interpretation and can be used to identify spectral channels that identify a spectral edge.

Appendix A.

Quantification of Magnetic Resonance Spectra

In its most general form, the free induction decay (FID) is modeled as the sum over exponentially decaying sinusoidal complex-valued time-signal components:

$$S_\theta(t_n) = \sum_{k=1}^{K} a_k \mathrm{e}^{j\phi_k} \mathrm{e}^{(j2\pi f_k - d_k - g_k t_n)t_n} \quad \text{with } t_n = n\Delta t + t_0 \tag{A.1}$$

where $j = \sqrt{-1}$, a_k denotes the k^{th} amplitude, ϕ_k the phase, d_k determines the Lorentzian line width and g_k the width of the Gaussian part. Δt is the sampling rate and t_0 an offset. This so-called *Voigt* model [126, 186] comprises the *Lorentz* (for $g_k = 0$) as well as the *Gauss* ($d_k = 0$) models. While the Lorentz model follows from basic NMR physics, field inhomogeneities and partial volume effects can lead to signals with more Gaussian shape (as a superposition of many slightly shifted Lorentz lines).

Quantification denotes the process of determining the most likely FID signal parameters $\theta = (t_0, a_k, f_k, d_k, g_k, \phi_k)$ given an observed sampled FID signal y_n. In the case of additive isotropic zero mean white Gaussian noise the sum of squared residuals is to be minimized:

$$\begin{aligned} \mathrm{SSR}(\theta) &= \sum_{n=1}^{N} (S_\theta(t_n) - y_n)^2, && \text{(A.2)} \\ &= ||\mathbf{s} - \mathbf{y}||^2 && \text{(A.3)} \end{aligned}$$

where the components of the vector $\mathbf{s} = (s_n)$ are $s_n = S_\theta(t_n)$.

Often *prior knowledge* about some of the parameters is available, *e.g.* certain metabolites such as citrate (Ci) consist of several components with known relative phase and amplitudes. Such prior knowledge can be used to reduce the number of sought parameters. Depending on the kind of prior knowledge to be employed, different algorithms with different properties can be used for quantification.

An abundance of methods have been proposed for the quantification of MR spectra. Here, only the most important time-domain methods are briefly reviewed (more details can be found, *e.g.*, in [206, 131, 203]). All algorithms described in the following are available in the jMRUI tool (*J*ava *M*agnetic *R*esonance *U*ser *I*nterface, `http://www.mrui.uab.es/`) [139, 140].

A.1. Hankel Singular Value Decomposition

If the Lorentz model is assumed ($g_k = 0$) an approach based on Singular Value Decomposition (SVD) can be used that allows for computationally efficient quantification [154, 33, 112].

For the Hankel Singular Value Decomposition (HSVD), the noiseless signal $s_n = S_\theta(t_n)$ ($n = 0, \ldots, N-1$; $t_0 = 0$) is first arranged in a $L \times M$ Hankel matrix as

$$\mathbf{H} = \begin{bmatrix} s_0 & s_1 & \cdots & s_{M-1} \\ s_1 & s_2 & \cdots & s_M \\ \vdots & \vdots & \vdots & \vdots \\ s_{L-1} & s_L & \cdots & s_{N-1} \end{bmatrix} \tag{A.4}$$

where $L = N - M + 1$, $M > K$ and $L > K$. From Eq. (A.1) it is easily verified that the following Vandermonde decomposition holds

$$\mathbf{H} = \begin{bmatrix} 1 & \cdots & 1 \\ z_1^1 & \cdots & z_K^1 \\ \vdots & \vdots & \vdots \\ z_1^{L-1} & \cdots & z_K^{L-1} \end{bmatrix} \begin{bmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_K \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ z_1^1 & \cdots & z_K^1 \\ \vdots & \vdots & \vdots \\ z_1^{M-1} & \cdots & z_K^{M-1} \end{bmatrix}^T \tag{A.5}$$

$$= \zeta_{LK}\, \mathbf{C}\, \zeta_{MK}^T \tag{A.6}$$

with the poles $z_k = \exp\left[(j2\pi f_k - d_k)\Delta t\right]$ and the (complex) amplitudes $c_k = a_k \exp[j\phi_k]$. Hence, all sought parameters are easily computed if the Vandermonde decomposition of the Hankel matrix H is known. However, this cannot be computed directly. Instead, an indirect way can be derived to determine the poles z_k first and from these the amplitudes c_k. Figure A.1 shows an example for an FID with four Lorentzian shaped components.

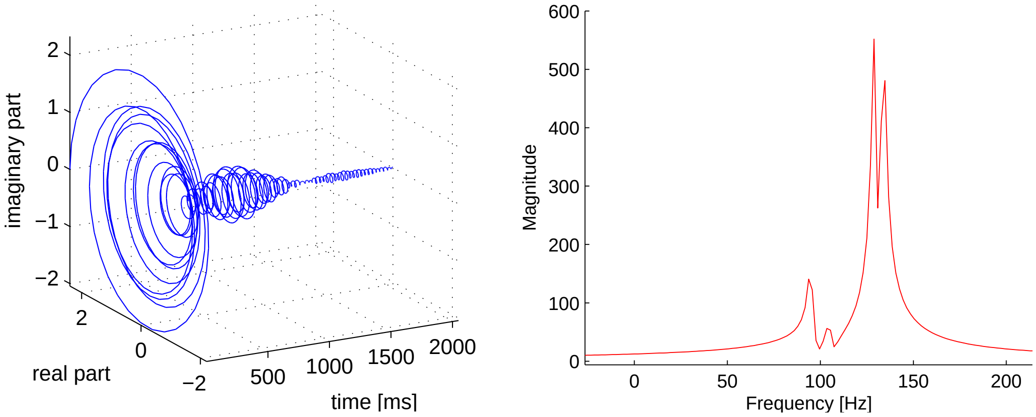


Figure A.1.: Simulated free induction decay (FID) at 1.5T with four components at the frequencies of Choline (95Hz), Creatine (105Hz) and Citrate (129.1Hz & 133.7Hz). Left: complex-valued time-domain signal. Right: corresponding magnitude spectrum.

With $\mathbf{Z} = \mathrm{diag}(z_k)$, a diagonal matrix with the signal poles, it is easily seen that ζ_{LK} (and also ζ_{MK}) are shift-invariant in the sense that

$$\zeta^{\uparrow}_{LK} = \zeta^{\downarrow}_{LK}\mathbf{Z} \tag{A.7}$$

where the up (down) arrow denotes the removal of the top (bottom) row of ζ_{LK}.

The above Hankel matrix H can certainly also be decomposed by SVD, which allows a factorization as

$$\begin{aligned} \mathbf{H} &= \mathbf{U}\Sigma\mathbf{V}^* && \text{(A.8)} \\ &= \begin{bmatrix} \mathbf{U}_K & \mathbf{U}_0 \end{bmatrix} \begin{bmatrix} \Sigma_K & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_K & \mathbf{V}_0 \end{bmatrix}^* && \text{(A.9)} \end{aligned}$$

where $\cdot^*$ denotes Hermitian conjugation. The decomposition in the second line is possible since it is known that $\mathbf{H}$, for a noise-less signal, has at most rank K.

Comparing the Vandermonde decomposition in Eq. (A.6) with the SVD decomposition in Eq. (A.9) reveals that $\mathbf{U}_K$ and ζ_{LK} must span the same column space and, hence, are equal up to a multiplication by an invertible matrix $\mathbf{T} \in \mathbb{C}^{K\times K}$:

$$\mathbf{U}_K = \zeta_{LK}\mathbf{T}. \tag{A.10}$$

Thus, from the shift-invariance property in Eq. (A.7) one obtains that

$$\mathbf{U}_K^{\uparrow} \quad = \quad \mathbf{U}_K^{\downarrow}\mathbf{T}^{-1}\mathbf{Z}\mathbf{T} \tag{A.11}$$

which allows to compute ($\mathbf{T}^{-1}\mathbf{Z}\mathbf{T}$) from the left singular vectors $\mathbf{U}_K$. In the case of noisy signals, this solution is found in a least squares sense, yielding the HSVD method if the usual Moore-Penrose pseudoinverse [154] and the HTLS method if a total least squares approach [202] is used. The poles z_k are obtained as the eigenvalues of ($\mathbf{T}^{-1}\mathbf{Z}\mathbf{T}$) which again allow the computation of c_k by least squares. The described HSVD algorithm is implemented in only a few lines of Matlab code (cf. Fig. A.2).

With some exceptions (*e.g.* [187, 114]), SVD-based quantification does not allow the application of prior knowledge. Therefore, it is usually referred to as "black-box" approach. It is often used to provide starting values for iterative algorithms or to remove unwanted signal components with poles outside interesting spectral regions, in particular residual water components.

A.2. VARPRO, AMARES and QUEST

VARPRO (*var*iable *pro*jection), AMARES (*a*dvanced *m*ethod for *a*ccurate, *r*obust, and *e*fficient *s*pectral fitting) and QUEST (*qu*antitation based on *qu*antum *est*imation) are quantification methods which are based on iterative nonlinear optimization algorithms that can incorporate various types of constraints. All methods minimize the SSR from Eq. (A.2) but with slightly different forms of the model function $S_\theta(t_n)$.

VARPRO uses a modified Levenberg-Marquart algorithm to minimize the SSR w.r.t. to its nonlinear variables while the variables that have a linear influence are solved by linear least squares. VARPRO uses the signal model

$$S_\theta(t_n) = \sum_{k=1}^{K} c_k \gamma_{\theta_k^{\mathrm{nl}}}(t_n) \quad \text{with } t_n = n\Delta t + t_0 \tag{A.12}$$

where c_k is the complex amplitude of the k^{th} component function $\gamma_{\theta_k}(t_n)$ which contains all the nonlinear parameters, *i.e.* $\theta_k^{\mathrm{nl}} = (t_0, f_k, d_k, g_k)$. With $\mathbf{s} = (S_\theta(t_n))$ Eq. (A.12) can be written as the linear system of equations

$$\mathbf{s} \quad = \quad \Gamma(\theta^{\mathrm{nl}})\mathbf{c} \quad = \quad \Gamma\mathbf{c}. \tag{A.13}$$

```
function [zk, ck] = hsvd(signal, K)
% signal: complex FID signal
% K: number of components in signal
% zk: complex signal poles
% ck: complex amplitudes

% number of sampling points
N = length(signal);

% number of columns/rows in Hankel matrix
M = round(3*N/4+1)
L = N-M+1;

% create the Hankel matrix H from the data
H = hankel(signal(1:L),signal(L:end));

% compute singular values (normal equations approach)
[U,S] = eig(H*H'); % NOTE: ensure that the EVs are sorted!

% use left singular vectors (U) to calculate the
% signal poles, only K singular vectors are used
U  = U(:,L-K+1:L);
Ut = U(2:L,:);
Ub = U(1:L-1,:);

% compute signal poles
zk = eig((Ub'*Ub)\Ub'*Ut);

% create Vandermonde matrix from signal poles
n = 0:(N-1);
A = exp(log(zk)*n).';

% compute amplitudes by least squares
ck = A\signal(:);
```

Figure A.2.: Matlab source code of the HSVD algorithm which can be used for the quantification of MRS data.

For given $\Gamma(\theta^{\mathrm{nl}})$ and observed signal $\mathbf{y} = (y_n)$ the SSR is minimized w.r.t. $\mathbf{c}$ by

$$\hat{\mathbf{c}} \;=\; (\Gamma^*\Gamma)^{-1}\Gamma^*\mathbf{y}. \qquad \text{(A.14)}$$

The remaining nonlinear variables are found by minimizing the *variable projection functional*

$$\mathrm{SSR}(\theta^{\mathrm{nl}}) \;=\; ||\mathbf{y} - \Gamma\hat{\mathbf{c}}||^2 \;=\; \left|\left|\mathbf{P}(\theta^{\mathrm{nl}})\,\mathbf{y}\right|\right|^2 \qquad \text{(A.15)}$$

with the projection matrix $\mathbf{P}(\theta^{\mathrm{nl}}) = \mathbf{I} - \Gamma(\Gamma^*\Gamma)^{-1}\Gamma^*$.

AMARES minimizes the SSR w.r.t. to all variables simultaneously using the NL2SOL algorithm [99, 204]. The implementation of AMARES allows for various equality and inequality constraints between the parameters which allows more flexibility than with the available implementation of VARPRO [140, 139].

QUEST uses a Levenberg-Marquart algorithm to minimize the SSR using a signal model that allows to specify prior knowledge and starting values on the components more implicitly or even using measured *in vitro* metabolite templates [160, 159]. As opposed to Eq. (A.1) the FID is now modeled as sum over M distorted metabolite templates $T_m(t_n)$ rather than components:

$$S_\theta(t_n) = \mathrm{e}^{j\phi_0} \sum_{m=1}^{M} T_m(t_n)\; a_m \mathrm{e}^{j\Delta\phi_m} \mathrm{e}^{(j2\pi\Delta f_m + \Delta d_m)t_n} \quad \text{with } t_n = n\Delta t + t_0 \qquad \text{(A.16)}$$

The parameters in this model comprise the null-phase, the time-lag, the metabolite amplitudes and distortion parameters, *i.e.* $\theta = (\phi_0, t_0, a_m, \Delta\phi_m, \Delta f_m, \Delta d_m)$. All parameters but a_m are naturally initialized to 0. An initial guess for the amplitudes a_m is then obtained by least squares. If *in vitro* measurements for the metabolite templates $T_m(t_n)$ are not available or desired, the templates can be generated by simulation.

One way to obtain metabolite templates certainly is to use the signal model from Eq. (A.1) which, when used with QUEST, leads to the same results as AMARES with corresponding constraints. Thus, much of the available prior knowledge is hidden in the metabolite templates and does not have to be specified in the form of constraints which may become tedious for spectra with many metabolites. From a mathematical point of view there is not much of a difference.

Another advantage of QUEST over AMARES is that the method can also handle baselines in the spectra stemming from macromolecules. Recent enhancements of AMARES use a semiparametric signal model by adding a spline term which allows for simultaneous quantification and baseline removal [183, 185]. In using metabolite templates, QUEST follows ideas first introduced with LCModel [156, 155].

Appendix B.

Statistical Subspace Methods

The term *statistical subspace methods* summarizes a family of methods that perform dimensionality reduction by constructing a subspace in a given feature space that captures all or most of the *relevant information.* Although linear in nature, these methods can also be enhanced for nonlinear dimensionality reduction by using the *kernel trick* [179] which corresponds to an application of the linear method to an implicitly inflated feature space. For conciseness only the linear theory is reviewed which is also sufficient and appropriate for the spectral data encountered in the present thesis.

The idea behind subspace methods is to construct a compressed representation of the given mean-centered data set which is optimal with respect to some optimality criterion [72, 188, 34, 53]. More formally, K uncorrelated latent variables $z_k(x) = \alpha_k^T x$ (the score variables, $k = 1 \dots K$) are sought such that

$$\alpha_k = \underset{\substack{\mathrm{corr}(\alpha^T x, \alpha_j^T x)=0,\, j<k \\ ||\alpha||=1}}{\mathrm{argmax}} T(\alpha) \tag{B.1}$$

where $\mathrm{corr}(\alpha^T x, \alpha_j^T x)$ is the correlation w.r.t. the empirical distribution, *i.e.*

$$\mathrm{corr}(\alpha^T x, \alpha_j^T x) = \alpha^T \mathbf{S}_x \alpha_j \tag{B.2}$$

with the sample covariance $\mathbf{S}_x = N^{-1}\mathbf{X}^T\mathbf{X}$. If $K = P$ components are constructed, all subspace methods yield a set of basis vectors which are orthogonal in $\mathbf{S}_x$ and span $\mathbb{R}^P$. The crucial difference between various subspace methods comes from the fact that only the first $K << P$ loadings are used. The low-dimensional projection of x thus neglects different aspects of its variations depending on how $T(\alpha)$ is chosen.

Prominent representatives of the class of subspace methods are principal component analysis (PCA), partial least squares (PLS) and ordinary least squares (OLS) regression, which are obtained for specific choices of $T(\alpha)$:

$$\begin{array}{lcllll} T_{\mathrm{OLS}}(\alpha) & = & \mathrm{corr}^2(\alpha^T x, y) & \propto & (\alpha^T \mathbf{S}_x \alpha)^{-1}(\alpha^T \mathbf{X}^T \mathbf{y})^2 \\ T_{\mathrm{PCA}}(\alpha) & = & \mathrm{var}(\alpha^T x) & \propto & \alpha^T \mathbf{S}_x \alpha \\ T_{\mathrm{PLS}}(\alpha) & = & \mathrm{corr}^2(\alpha^T x, y)\ \mathrm{var}(\alpha^T x) & \propto & (\alpha^T \mathbf{X}^T \mathbf{y})^2 \end{array} \tag{B.3}$$

where $\propto$ means equal up to a constant factor. Hence, PLS can be viewed as an intermediate approach between OLS and PCA. As opposed to PCA, PLS also considers the target labels y when determining the optimal directions α_k. Therefore, the derived latent variables in PLS are tuned towards discriminating the labels and downweight variations in the pattern which do not convey label information. The advantage over OLS is that PLS is less vulnerable to overfitting when given many correlated features such as a spectral pattern. Due to these advantages, PLS has been used extensively in chemometric applications [72]. From a theoretical point of view it therefore seems advantageous to prefer PLS over PCA if labels are available and a classification is the aim.

Subspace methods deliver (projection) *directions* α_k and *loadings* (coordinate vectors in the original feature space) together with an importance ordering. Thus PCA reveals the dominant spectral patterns of maximum variance ($\mathbf{S}_x \alpha_k \propto \alpha_k$, the principal components) whereas PLS reveals the most important patterns regarding the given classification task. Visualizing these patterns allows to understand the decision process of the trained classifier much better than just looking at the coefficient profile obtained from a linear model.

The relationship between loadings and directions is derived as follows. Using the $P \times K$ matrix $\mathbf{R} = (\alpha_k)$ with directions, the $P \times K$ matrix $\mathbf{L}$ with loadings, and the $N \times K$ score matrix $\mathbf{Z} = (z_k)$, the examples collected in the $N \times P$ design matrix $\mathbf{X}$ are to be approximated in a K-dimensional subspace ($K << P$) as

$$\mathbf{X} \approx \tilde{\mathbf{X}} \quad = \quad \underbrace{\mathbf{Z}}_{\text{scores}} \quad \underbrace{\mathbf{L}^T}_{\text{loadings}} \ . \tag{B.4}$$

In general, the loadings are not orthogonal (only for PCA they are) and the scores are obtained in the least squares sense, *i.e.*

$$\mathbf{Z} \quad = \quad \mathbf{X} \underbrace{\mathbf{L}(\mathbf{L}^T \mathbf{L})^{-1}}_{\text{directions } \mathbf{R}} . \tag{B.5}$$

From this we have that

$$\begin{aligned}\tilde{\mathbf{X}}\mathbf{R} &= \mathbf{Z}\mathbf{L}^T\mathbf{R} &\text{(B.6)}\\ &= \mathbf{X}\mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1} &\text{(B.7)}\\ &= \mathbf{X}\mathbf{R}. &\text{(B.8)}\end{aligned}$$

Since the latent variables are supposed to be uncorrelated $(\mathrm{corr}(\alpha_k^T x, \alpha_j^T x) = 0$ for $j \neq k)$ and because of the normalization constraint $||\alpha|| = 1$ we have that

$$\mathbf{Z}^T\mathbf{Z} \;=\; \mathbf{R}^T\mathbf{S}_x\mathbf{R} \;=\; \mathbf{D} \qquad \text{(B.9)}$$

where $\mathbf{D}$ is an (invertible) diagonal matrix determined by

$$(\mathbf{R}^T\mathbf{R})_{ii} = 1. \qquad \text{(B.10)}$$

We then have

$$\begin{aligned} \tilde{\mathbf{X}} &= \mathbf{Z}\mathbf{L}^T\\ \Rightarrow\; \mathbf{Z}^T\tilde{\mathbf{X}} &= \mathbf{Z}^T\mathbf{Z}\mathbf{L}^T\\ \Rightarrow\; \mathbf{L} &= \tilde{\mathbf{X}}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1} &\text{(B.11)}\end{aligned}$$

and with Eq. (B.5) obtain

$$\begin{aligned}\mathbf{L} &= \tilde{\mathbf{X}}^T\mathbf{X}\mathbf{R}(\mathbf{R}^T\mathbf{X}^T\mathbf{X}\mathbf{R})^{-1} &\text{(B.12)}\\ &= \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{R}\mathbf{D}^{-1} &\text{(B.13)}\\ &= \tilde{\mathbf{S}}_x\mathbf{R}\mathbf{D}^{-1}. &\text{(B.14)}\\ \mathbf{R}^T\mathbf{L} &= \mathbf{I} &\text{(B.15)}\end{aligned}$$

Hence, the scaled loadings are obtained by multiplying the directions ith the approximate empirical covariance matrix $\tilde{\mathbf{S}}_x$. Since the loadings define coordinate vectors in the original high-dimensional space and since the observed data points in this space are explained as linear combinations of the loadings, these can be regarded as “components” that make up the observed patterns and are thus displayed for interpretation.

An exception within the subspace framework is OLS for which only the first loading gets non-zero weight in the subsequent linear regression [188]. Therefore, the loading and the coefficient profile are identical for OLS. The subspace view does not yield additional interpretability in this case.

Appendix C.

Graphical Models, Exponential Families and Convex Analysis

A particularly useful subclass of the distributions that can be defined as graphical models are exponential families. Many well-known distributions are known to be exponential families such as the Binomial, Multinomial, Poisson, Geometric, Laplace, Beta, Gamma, Exponential and the Gaussian distribution (cf. C.1). The definition and some useful properties of exponential families are summarized in the following, mostly based on [208] and [24].

C.1. Definitions

Given a graph $G = (V, E)$, for each site $s \in V$ let x_s be a random variable taking values in some ν_s-measurable sample space $\mathcal{X}_s$ which may be discrete or continuous (*e.g.* $\mathcal{X}_s = \mathbb{R}$ with the Lebesgue measure). The random vector $x = (x_s)$ then takes values in the Cartesian product space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$ endowed with the corresponding product measure ν where $N = |V|$.

Furthermore, let $\mathcal{C} = \{c \colon c \subseteq V\}$ be a collection of cliques on the graph G. Let $t(x) = (t_c(x_c))$ define a corresponding collection of Borel measureable functions $t_c(x_c) \colon \mathcal{X}_c \to \mathbb{R}^{n_c}$ (the *sufficient statistics*) and let $\theta = (\theta_c) \in \mathbb{R}^n$ be a vector of parameters (the *canonical* or *exponential parameters*). An induced dot product $\langle \theta, t(x) \rangle$ can thus be obtained as $\sum_{c \in \mathcal{C}} \langle \theta_c, t_c(x_c) \rangle$ where for the purpose of this review only the usual Euclidean dot product is used[*].

[*] Replacing the dot product with a positive definite kernel can lead to very interesting non-parametric models [117].

Table C.1.: Some univariate exponential family distributions.

Family	$\mathcal{X}$	ν^1	$t(x)$	$A(\theta)$	Θ
Bernoulli	$\{0,1\}$	C	x	$\ln[1+\exp(\theta)]$	$\mathbb{R}$
Gaussian	$\mathbb{R}$	L	(x,x^2)	$\frac{1}{2}\ln\frac{\pi}{-\theta_2}-\frac{\theta_1^2}{4\theta_2}$	$\{\theta\in\mathbb{R}^2\vert\theta_2<0\}$
Exponential	$(0,+\infty)$	L	$-x$	$-\ln\theta$	$(0,+\infty)$
Poisson	$\{0,1,2,\ldots\}$	C^2	x	$\exp\theta$	$\mathbb{R}$

Also: Geometric, Laplace, Beta, Gamma, Wishart, Dirichlet, von Mises-Fisher, ...

[1] dominating measure: C=Counting, L=Lebesgue

[2] with $\nu = \delta_{\mathcal{X}}(x)/x!$

An exponential family distribution that is dominated by the measure ν and factors according to the graph $G = (V, E)$ is then defined as

$$p_\theta(x) \quad := \quad \frac{\mathrm{dP}_\theta}{\mathrm{d}\nu}(x) \quad = \quad \exp[\langle\theta, t(x)\rangle - A(\theta)] \tag{C.1}$$

$$= \quad \exp[-A(\theta)] \prod_{c\in\mathcal{C}} \exp[\langle\theta_c, t_c(x_c)\rangle] \tag{C.2}$$

with the log partition function

$$A(\theta) \quad = \quad \ln \int_{\mathcal{X}} \exp\langle\theta, t(x)\rangle \;\mathrm{d}\nu \tag{C.3}$$

Eq. (C.2) also shows that the functions $t_c(x_c)$ are *sufficient statistics* for the exponential family by the Fisher-Neyman factorization theorem [152, Thm 1.2.10]. Finally, the *natural parameter space* Θ is defined as the set of parameters for which the integral in (C.3) exists, *i.e.*

$$\Theta \quad = \quad \left\{ \theta\in\mathbb{R}^n \colon \int_{\mathcal{X}} \exp\langle\theta, t(x)\rangle \;\mathrm{d}\nu < \infty \right\} \tag{C.4}$$

Some univariate examples for exponential families are listed in Table C.1.

C.2. The Log Partition Function

Several important inference problems in graphical models exist, such as

- parameter estimation by maximizing the joint, conditional or marginal likelihood of observed data
- most probable explanation (MPE), maximum a posteriori (MAP) or mode estimation
- calculation of marginal and conditional distributions.

Although seemingly very different, all these operations lead to similar problems. In the following sections some of them will be considered in the context of exponential families which will highlight the crucial role of the log partion function.

Unfortunately, the log partition function often causes computational problems. A straightforward evaluation of the integral (respectively sum) in Eq. (C.3) is a daunting task if the number of sites N and therefore the collection of random variables grows big. For example, the number of summands in Eq. (C.3) for a random field with N binary nodes (Ising model) would be 2^N. In the case of a Gaussian random field the challenge lies in the calculation of the determinant of a huge matrix which, in general, comes at computational costs of $O(N^3)$.

A particularly useful property of the log partition function in exponential families is that it is also a *cumulant generating function*, *i.e.*

Lemma 1. *If $\theta \in \operatorname{int}\Theta$ then the statistic $t = t(x)$ has moments of all orders w.r.t. $p_\theta(x)$ and the i^{th}derivative of the log partition function $A(\theta)$ equals the i^{th}order cumulant of t. In particular, the first two derivatives coincide with the mean and the covariance of the sufficient statistic t:*

$$\frac{\partial A}{\partial \theta}(\theta) = \mathrm{E}_\theta[t] \tag{C.5}$$

$$\frac{\partial^2 A}{\partial \theta \partial \theta^T}(\theta) = \mathrm{E}_\theta[tt^T] - \mathrm{E}_\theta[t]\,\mathrm{E}_\theta[t^T] \tag{C.6}$$

Furthermore,

$$\lim_{n\to\infty} \left|\left| \frac{\partial A}{\partial \theta}(\theta^n) \right|\right| = +\infty \tag{C.7}$$

for any sequence $\{\theta^n\} \subset \operatorname{int}\Theta$ converging to a point on the boundary of Θ.

Proof. A proof can be found in [24, Thm 8.1]. □

Note that since the Hessian of $A(\theta)$ is equal to the covariance of t it must be positive (semi-)definite (Eq. (C.6)). This is sufficient to establish the convexity of $A(\theta)$ and also implies the convexity of the natural parameter space (*e.g.* [208, Cor 1], [163] or [152, Thm 1.6.5]). Together with Eq. (C.7) lemma 1 states that $A(\theta)$ is *essentially smooth* or *steep* [163, 24, 42, 153].

C.3. Maximum Likelihood Estimation

In parametric maximum likelihood (ML) estimation a set of parameters $\hat{\theta}$ is sought that maximizes the (joint) likelihood of an observation $o = \{x^i\}_{i=1}^N$. Assuming independent and identically distributed (iid) examples x^i, the distribution from which the observed instance o has been drawn is

$$\begin{aligned} p_\theta(\mathcal{O} = o) &= \prod_{i=1}^{N} p_\theta(x^i) && \text{(C.8)} \\ &= \exp[-NA(\theta)] \prod_{c \in \mathcal{C}} \exp\left[N\langle\theta_c, \tilde{\eta}_c\rangle\right] && \text{(C.9)} \end{aligned}$$

with the empirical mean $\tilde{\eta}_c = \frac{1}{N}\sum_{i=1}^N t_c(x_c^i) = \mathrm{E}_{\tilde{p}(x)}[t_c(x_c^i)]$. Note that the assumption of iid examples is not very limiting here. In fact, certain graphical models allow parameter estimation from only "one" observation ($N = 1$) in that clique parameters are tied, *e.g.* $\theta_c \equiv \theta_0\ \forall c \in \mathcal{C}$. Examples of such models are graphical chain models corresponding to Kalman smoothers/filters or hidden Markov models with time-invariant Markov kernel and observation distribution. In image processing (multidimensional) lattices with pairwise interactions form a natural generalization of such chain models.

The maximum of the joint likelihood (C.9) can be found by minimizing the negative loglikelihood which yields

$$\begin{aligned} \hat{\theta} &= \operatorname*{argmin}_{\theta \in \Theta} \left[A(\theta) - \sum_c \langle\theta_c, \tilde{\eta}_c\rangle\right] && \text{(C.10)} \\ &= \operatorname*{argmin}_{\theta \in \Theta} \left[A(\theta) - \langle\theta, \tilde{\eta}\rangle\right] && \text{(C.11)} \end{aligned}$$

where $\tilde{\eta} = (\tilde{\eta}_c)$ in analogy to the vectorization of $t(x)$ and θ. In light of Lemma 1 this is certainly a convex optimization problem which requires that

$$\frac{\partial}{\partial\theta} A(\hat{\theta}) \quad = \quad \mathrm{E}_{\hat{\theta}}[t(x)] \quad = \quad \tilde{\eta} \qquad \text{(C.12)}$$

The interpretation of this result is very intuitive as it requires that the empirical mean equals the mean of the maximum likelihood model. Note that despite this simplicity it still remains difficult to actually calculate $\hat{\theta}$ since the evaluation of the objective function (C.11) requires the calculation of $A(\theta)$. Furthermore, its gradient and Hessian require the calculation of marginals of $p_\theta(x)$ (*e.g.* for $\mathrm{E}_{\hat{\theta}}[t(x)]$) which is similarly difficult as the calculation of $A(\theta)$.

C.4. Exponential and Moment Parameters

Having identified $t(x)$ as sufficient statistic allows to parameterize any distribution of the exponential family in two ways. It is certainly determined by the *exponential parameters* θ but it can equally well be specified by fixing the *moment parameters* $\eta = \mathrm{E}[t(x)]$. The dual parameterization of exponential families is also the crucial ingredient that distinguishes *information geometry* from common differential geometry [21]. But before examining the relationship between these two parameterizations some further definitions are required.

Definition 1. *Given the sufficient statistic $t(x)$ and the measure ν, $\mathcal{M}$ is defined as the set of expectations of the sufficient statistic $t(x)$ under any probability measure* P *that is dominated by ν:*

$$\mathcal{M} \quad = \quad \{\eta \in \mathbb{R}^n \colon \exists \mathrm{P} \prec \nu \text{ s.t. } \mathrm{E}_{\mathrm{P}}[t(x)] = \eta\}\,. \tag{C.13}$$

$\mathcal{M}$ is a convex set. Furthermore, it can be shown that any $\eta \in \operatorname{ri} \mathcal{M}$ can be obtained under the exponential family P_θ generated by ν and $t(x)$ for some $\theta \in \Theta$ (see [208, Thm 1]).

The relationship between moment and exponential parameterizations can then be characterized as *conjugate duality* known from convex analysis [163, 24, 208]. The conjugate dual of the log partition function is defined as

$$A^*(\eta) \quad := \quad \sup_{\theta \in \Theta} [\langle \eta, \theta \rangle - A(\theta)]. \tag{C.14}$$

For $\eta \in \operatorname{ri} \mathcal{M}$ it is straightforward to verify that the conjugate dual evaluates to the negative differential entropy $\mathrm{H}(\theta(\eta)) := -\mathrm{E}_\theta[\ln p_{\theta(\eta)}(x)]$. In general, it can be determined as

$$A^*(\eta) \quad = \quad \begin{cases} -\mathrm{H}(\theta(\eta)) & \eta \in \operatorname{ri} \mathcal{M} \\ -\lim_{n\to\infty} \mathrm{H}(\theta(\eta^n)) & \eta \in \operatorname{bd} \mathcal{M}, (\eta^n) \to \eta \\ +\infty & \eta \notin \operatorname{cl} \mathcal{M} \end{cases} \tag{C.15}$$

where $\{\eta^n\} \subset \mathcal{M}$ is a sequence converging to η and $\theta(\eta) \in \Theta$ is an exponential parameter that fulfills $\mathrm{E}_\theta[t(x)] = \eta$, *i.e.* the pair (θ, η) is *dually coupled* [208, Thm 2]. In fact, this duality describes the well-known equivalence between the "Maximum Entropy Constraint Distribution" and the "Maximum Likelihood Gibbs Distribution" for exponential families (*e.g.* [57]).

Similarly to lemma 1, the derivatives of the conjugate dual $A^*(\eta)$ take a particularly nice form.

Lemma 2. *For $\eta \in \operatorname{ri}\mathcal{M}$ and $\eta = \mathrm{E}_\theta[t]$ the derivatives of $A^*(\eta)$ are*

$$\frac{\partial A^*}{\partial \eta}(\eta) = \theta \tag{C.16}$$

$$\frac{\partial^2 A^*}{\partial \eta \partial \eta^T}(\eta) = (\mathrm{E}_\theta[tt^T] - \mathrm{E}_\theta[t]\,\mathrm{E}_\theta[t^T])^{-1} \tag{C.17}$$

Proof. A proof can be found in [24]. □

Conversely, the conjugate dual of $A^*(\eta)$ is again the log partition function [208, 213] which leads to

$$A(\theta) = \sup_{\eta \in \mathcal{M}} [\langle \theta, \eta \rangle - A^*(\eta)]. \tag{C.18}$$

Furthermore, Fenchel's inequality [163] for the dual pair (A^*, A) yields

$$A^*(\eta) + A(\theta) - \langle \eta, \theta \rangle \geq 0 \qquad \text{for any } (\theta, \eta) \in \mathbb{R}^n \times \mathbb{R}^n \tag{C.19}$$

which is also known as *Gibbs variational principle* [213, p.60]. It holds with equality if and only if $(\theta,\ \eta)$ is dually coupled and is basis for variational representations of the log partition function such as *mean field* approximations [100, 94].

C.5. Kullback-Leibler Divergence

In general, the *Kullback-Leibler* (KL) divergence or *relative entropy* of two distributions is defined as

$$\mathrm{D}(q\,||\,p) = \int_{\mathcal{X}} q(x) \ln \frac{q(x)}{p(x)}\, \mathrm{d}\nu = \mathrm{E}_q[\ln q] - \mathrm{E}_q[\ln p] \tag{C.20}$$

For two distributions $p_{\theta_1}(x)$ and $p_{\theta_2}(x)$ from the same exponential family it takes the convenient form:

$$\begin{aligned} \mathrm{D}(\theta_1 \,||\, \theta_2) &:= \mathrm{D}(p_{\theta_1}(x) \,||\, p_{\theta_2}(x)) && \text{(C.21)} \\ &= \mathrm{E}_{\theta_1}[\ln p_{\theta_1}(x)] - \mathrm{E}_{\theta_1}[\ln p_{\theta_2}(x)] && \text{(C.22)} \\ &= A(\theta_2) - A(\theta_1) + \langle \theta_1 - \theta_2, \eta_1 \rangle && \text{(C.23)} \end{aligned}$$

with $\eta_1 = \mathrm{E}_{\theta_1}[t(x)] = \frac{\partial A}{\partial \theta}(\theta_1)$. In terms of convex analysis the KL divergence can thus be identified as a *Bregman distance* [153, 163].

Using Eq. (C.19) with the dually coupled parameter pairs (θ_1, η_1) and (θ_2, η_2), a mixed and a dual form of the KL divergence are obtained as

$$\begin{aligned} \mathrm{D}(\theta_1 \,||\, \theta_2) &= \mathrm{D}(\eta_1 \,||\, \theta_2) = A(\theta_2) + A^*(\eta_1) - \langle \eta_1, \theta_2 \rangle && \text{(C.24)} \\ &= \mathrm{D}(\eta_1 \,||\, \eta_2) = A^*(\eta_1) - A^*(\eta_2) + \langle \eta_2 - \eta_1, \theta_2 \rangle && \text{(C.25)} \end{aligned}$$

Appendix D.

Empirical Distributions and Soft Evidence

A generalized concept of *empirical distributions* is used to find the most likely point estimate for the parameters of a given model in a computationally efficient manner. After a short review of an alternative representation of likelihood as an expectation w.r.t. to the empirical distribution [80, 57] the *soft empirical distribution function* is introduced which allows an analogous formulation using the notion of *soft evidence* [151, 182, 58].

Definition 2 (Empirical distribution function, [80])**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ *be a probability space and let* $X_j : \Omega \to \mathbb{R}$ *be* N *independent and identically distributed (iid) random variables with realizations* x_j*. The* empirical distribution function F_N *is a cumulative probability distribution function defined by*

$$F_N(x) = \frac{1}{N} \sum_{j=1}^{N} I_j(x), \tag{D.1}$$

where $I_j(x)$ *is an indicator function for* $\{x_j \leq x\}$*.*

Using this definition, one can write

$$\begin{aligned} \tilde{\mathrm{p}}(\bar{x}) &:= \frac{\mathrm{d}F_N}{\mathrm{d}x}(\bar{x}) \\ &= \frac{1}{N} \sum_{j=1}^{N} \delta(\bar{x} - x_j) \end{aligned} \tag{D.2}$$

where $\tilde{\mathrm{p}}(\bar{x})$ is called *empirical probability* (*empirical density*) for discrete (continuous) random variables and δ is the Kronecker delta (Dirac distribution). Thus, the loglikelihood of the realizations x_j based on the training set $\mathcal{D} = \{\delta(x - x_j)\}_{j=1}^N$ can be written as

$$\begin{aligned} l(\theta; \mathcal{D}) &= N \cdot \mathrm{E}_{\tilde{\mathrm{p}}(x)}[\log \mathrm{p}(x \,|\, \theta)] && \text{(D.3)} \\ &= N \int\limits_{\mathbb{R}} \tilde{\mathrm{p}}(x) \log \mathrm{p}(x \,|\, \theta)\, \mathrm{d}x && \text{(D.4)} \\ &= \sum_{j=1}^{N} \log \mathrm{p}(x_j \,|\, \theta) && \text{(D.5)} \end{aligned}$$

Note that the maximum likelihood estimate of θ also minimizes the Kullback-Leibler divergence $\mathrm{D}(\tilde{\mathrm{p}} \,||\, \mathrm{p}_\theta) = \mathrm{E}_{\tilde{\mathrm{p}}(x)}[\log \tilde{\mathrm{p}}(x)] - \mathrm{E}_{\tilde{\mathrm{p}}(x)}[\log \mathrm{p}(x \,|\, \theta)]$ since the first term (empirical entropy) does not depend on the model parameters.

We now proceed with the notion of *soft evidence* and introduce the *soft empirical distribution function* in analogy to definition 2. Given a random variable $X\colon \Omega \mapsto \mathbb{R}$, soft evidence on X can be incorporated by introducing a virtual random variable E and specifying the likelihood function $\bar{w}(x) = \Pr(E = 1 \,|\, X = x)$. Introducing evidence in this way does not correspond to the usual concept of observing the realization of a random variable but instead consists in defining and adding a conditional probability distribution to the previously defined model. Hence, soft evidence should reflect degrees of belief that stem from external knowledge about the random variable X, *i.e.* from sources not captured by the probability model [151, 182].

Definition 3 (Soft empirical distribution function). *Let $X_j\colon \Omega \mapsto \mathbb{R}$ be iid random variables and $\mathcal{D} = \{w_j(x)\}_{j=1}^N$ a training set with soft evidence on X_j. Without loss of generality, it is required that the $w_j(x)$ integrate to one,* i.e. $w_j(x) := \bar{w}_j(x)(\int_{\mathbb{R}} \bar{w}_j(x)\, \mathrm{d}x)^{-1}$. *Then, the* soft empirical distribution function $\tilde{F}_{\mathcal{D}}$ *based on the training set $\mathcal{D}$ is defined as*

$$F_{\mathcal{D}}(x) = \frac{1}{N} \sum_{j=1}^{N} W_j(x). \tag{D.6}$$

where $W_j(\bar{x}) = \int_{-\infty}^{\bar{x}} w_j(x)\, \mathrm{d}x$.

Given a parametric model $\mathrm{p}(x \,|\, \theta)$, a uniform prior on θ and soft evidence $\mathcal{D}$, the mode of the posterior parameter distribution $\mathrm{p}(\theta \,|\, \mathcal{D})$ is found by maximizing the *soft loglikelihood*

$$l^s(\theta; \mathcal{D}) = \sum_{j=1}^{N} \log \int_{\mathbb{R}} \bar{w}_j(x)\, \mathrm{p}(x \,|\, \theta)\, \mathrm{d}x \tag{D.7}$$

$$= \sum_{j=1}^{N} \log \int_{\mathbb{R}} w_j(x)\, \mathrm{p}(x \,|\, \theta)\, \mathrm{d}x + c. \tag{D.8}$$

The maximizer $\hat{\theta}$ can be regarded as the maximum likelihood estimate of θ given the soft evidence specified by $\bar{w}_j(x)$. In fact, for the special case of hard labels ($\bar{w}_j(x) = \delta(x - x_j)$), the soft loglikelihood $l^s(\theta; \mathcal{D})$ directly turns into loglikelihood (cf. Eqn. (D.5)). However, in the case of X being a high-dimensional random field, the marginalization required for the evaluation of $l^s(\theta; \mathcal{D})$ makes its computation considerably more difficult. Alternatively, the empirical expectation $\mathrm{E}_{\tilde{\mathrm{p}}(x)}[\log \mathrm{p}(x \,|\, \theta)]$ can be maximized which is much simpler to compute but only provides a lower bound on $l^s(\theta; \mathcal{D})$ in general.

Proposition 1. *Let* $\mathrm{p}(x \,|\, \theta)$ *be a parametric model and* $\mathcal{D} = \{w_j(x)\}_{j=1}^{N}$ *a training set with soft evidence. Then, the expectation* $\mathrm{E}_{\tilde{\mathrm{p}}(x)}[\log \mathrm{p}(x \,|\, \theta)]$ *w.r.t. the empirical distribution (density)* $\tilde{\mathrm{p}}(\bar{x}) := \frac{\mathrm{d}F_{\mathcal{D}}}{\mathrm{d}x}(\bar{x})$ *provides a lower bound on the soft loglikelihood* $l^s(\theta; \mathcal{D})$*. The bound is tight for hard evidence (*$w_j(x) = \delta(x - x_j)$*).*

Proof. The claim is a direct consequence of Jensen's inequality. Since

$$\log \int_{\mathbb{R}} w_j(x)\, \mathrm{p}(x \,|\, \theta)\, \mathrm{d}x \geq \int_{\mathbb{R}} w_j(x) \log \mathrm{p}(x \,|\, \theta)\, \mathrm{d}x,$$

we have

$$\begin{aligned} l^s(\theta; \mathcal{D}) &\geq \sum_{j=1}^{N} \int_{\mathbb{R}} w_j(x) \log \mathrm{p}(x \,|\, \theta)\, \mathrm{d}x \\ &= N\ \mathrm{E}_{\tilde{\mathrm{p}}(x)}[\log \mathrm{p}(x \,|\, \theta)]. \end{aligned}$$

For $w_j(x) = \delta(x - x_j)$ (hard evidence) equality holds. □

List of Symbols and Expressions

Acronyms

^{1}H	Proton
AMARES	Advanced Method for Accurate, Robust, and Efficient Spectral fitting
AUC	Area Under Curve
CCA	Canonical Correlation Analysis
Ci	Citrate
CLARET	CSI-based Localization And Robust Estimation of Tumor probability
Cr	Creatine
CRF	Conditional Random Field
CRLB	Cramér-Rao lower bound
CV	Cross-validation
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging
FID	Free Induction Decay
FOV	Field Of View
GGMRF	Generalized Gaussian Markov Random Field
GP	Gaussian Process
GPLS	Generalized Partial Least Squares
HSVD	Hankel Singular Value Decomposition
ICA	Independent Components Analysis
ICM	Iterated Conditional Modes
KL	Kullback-Leibler
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MAP	Maximum A Posteriori
ML	Maximum Likelihood

NMF	Nonnegative Matrix Factorization
MoG	Mixture of Gaussians
MPME	Marginal Posterior Mode Estimate
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
MRSI	Magnetic Resonance Spectroscopic Imaging
MRS	Magnetic Resonance Spectroscopy
MSR	Mean Squared Residuals
NAA	N-acetylaspartate
NLS	Nonlinear Least Squares
NMR	Nuclear Magnetic Resonance
OLS	Ordinary Least Squares
PCA	Principal Components Analysis
PLS	Partial Least Squares
PPCA	Probabilistic Principal Components Analysis
PRESS	Point Resolved Spectroscopy
PSR	P-spline Signal Regression
QUEST	QUantitation based on QUantum ESTimation
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Squared Error
ROC	Receiver Operator Characteristics
Cho	Choline
SNR	Signal-to-noise Ratio
SSR	Sum of Squared Residuals
SVD	Singular Value Decomposition
SVM	Support Vector Machine
VARPRO	VARiable PROjection

General notation

$\langle \mathbf{a}, \mathbf{b} \rangle$	Standard dot product between vectors **a** and **b**
$\langle\langle \mathbf{A}, \mathbf{B} \rangle\rangle$	Dot product between matrices **A** and **B**

bd	Border
cl	Closure
$\operatorname{diag}\mathbf{a}$	A diagonal matrix with vector **a** on its diagonal.
$\mathcal{N}_d(\mathbf{x} \mid \mu, \Sigma)$	d-variate normal distribution with mean μ and covariance matrix Σ
ri	Relative Interior
(a_i)	denotes the vector **a** with components a_i.
$\operatorname{tr}\mathbf{A}$	The trace of matrix **A**.
$\mathbf{X}$	The $N \times P$ design matrix (N observations, P features).
$\mathbf{y}$	An N-vector of responses (N observations).

Greek Symbols

η	Moment parameters of an exponential family
ϕ, φ, Φ	Feature functions
σ	Standard Deviation of a Normal Distribution
θ	Parameters of a probability distribution, exponential parameters for exponential families

Latin Symbols

$\mathbb{C}$	Complex Numbers
$\mathbb{N}$	Natural Numbers
$\mathbb{R}$	Real Numbers
$\mathbb{R}^n$	n-dimensional Vector Space over $\mathbb{R}$

List of Figures

List of Tables

Bibliography

Own Contributions

Peer-Reviewed Publications

[1] Kelm BM, Menze BH, Henning A, Weber MA, Hamprecht FA. *Using spatial prior knowledge in the spectral fitting of magnetic resonance spectroscopic images.* Magn Reson Med (to be submitted)

[2] Kelm BM, Menze BH, Nix O, Zechmann C, Hamprecht FA. *Estimating kinetic parameter maps from dynamic contrast-enhanced MRI using a generalized Gaussian MRF with Block-ICM.* IEEE Trans Med Imaging (under revision)

[3] Kelm BM, Menze BH, Weinman J, Henning A, Görlitz L, Hamprecht FA. *Trading resolution against noise in NMR spectroscopic images with conditional random fields.* IEEE Trans Med Imaging (to be submitted)

[4] Kelm BM, Menze BH, Zechmann CM, Baudendistel KT, Hamprecht FA. *Automated estimation of tumor probability in prostate MRSI: Pattern recognition vs. quantification.* Magn Reson Med 2007. 57(1): 150 – 159. doi: 10.1002/mrm.21112

[5] Menze BH, Kelm BM, Masuch R, Lichy MP, Himmelreich U, Petrich W, Hamprecht FA. *Feature selection in spectral data using the multivariate Gini importance.* (to be submitted)

[6] Menze BH, Kelm BM, Weber MA, Bachert P, Hamprecht FA. *Mimicking the human expert: Pattern recognition for an automated assessment of data quality in MRSI.* Magn Reson Med (under revision)

[7] Menze BH, Lichy MP, Bachert P, Kelm BM, Schlemmer HP, Hamprecht FA. *Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors.* NMR Biomed 2006. 19(5): 599 – 609. doi:10.1002/nbm.1041

Conference Proceedings

[8] Kelm BM, Menze BH, Mueller N, Hamprecht FA. *Estimation of pharmacokinetic parameters using spatial prior knowledge.* In: Proc of the 23rd Annual Scientific Meeting of the ESMRMB. Warsaw/Poland, 2006 EPOS highlights session

[9] Kelm BM, Menze BH, Neff T, Zechmann CM, Hamprecht FA. *CLARET: a tool for fully automated evaluation of MRSI with pattern recognition methods.* In: H Handels, et al (eds.), Bildverarbeitung für die Medizin 2006 - Algorithmen, Systeme, Anwendungen, Informatik Aktuell. Springer, Hamburg/Germany, 2006 51–55

[10] Kelm BM, Müller N, Menze BH, Hamprecht FA. *Bayesian estimation of smooth parameter maps for dynamic contrast-enhanced MR images with block-ICM.* In: Proc Computer Vision and Pattern Recognition Workshop (MMBIA). New York/USA, 2006 96–103. doi:10.1109/CVPRW.2006.41

[11] Kelm BM, Pal C, McCallum A. *Combining generative and discriminative methods for pixel classification with multi-conditional learning.* In: Proc International Conference on Pattern Recognition, vol. 2. Hong Kong, 2006 828–832. doi:10.1109/ICPR.2006.384

[12] Kelm M, Menze B, Hamprecht F. *Automatische Lokalisation von Tumoren in 1H-NMR-spektroskopischen in vivo Aufnahmen.* In: GMA-Kongress: Automation als interdisziplinäre Herausforderung, vol. 1883 of *VDI-Berichte*. VDI/VDE - Gesellschaft Mess- und Automatisierungtechnik, Baden-Baden/Germany, 2005 457–466

[13] Menze BH, Kelm BM, Hamprecht FA. *Automated separation of low quality and artifact spectra by pattern recognition in the processing of MR spectral images.* In: Proc of the 14th ISMRM Scientific Meeting and Exhibition. Seattle/USA, 2006

[14] Menze BH, Kelm BM, Hamprecht FA. *From eigenspots to fisherspots - latent spaces in the nonlinear detection of spot patterns in a highly variable background.* In: HJ Lenz, R Decker (eds.), Advances in Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization, vol. 33. Springer, 2007 (in press)

[15] Menze BH, Kelm BM, Heck D, Lichy MP, Hamprecht FA. *Machine-based rejection of low quality spectra and estimation of brain tumor probabilities from*

magnetic resonance spectroscopic images. In: H Handels, et al (eds.), Bildverarbeitung für die Medizin, Informatik Aktuell. Hamburg/Germany, 2006 31–36

[16] Menze BH, Kelm BM, Lichy MP, Bachert P, Schlemmer H, Hamprecht FA. *Optimal processing in the automatic detection and localization of brain tumors using MRSI.* In: Proc of the 13th ISMRM Scientific Meeting and Exhibition. Miami/USA, 2005

[17] Pal C, Kelm M, Wang X, Druck G, McCallum A. *On discriminative and semi-supervised dimensionality reduction.* In: Proc of the NIPS Workshop - Novel Applications of Dimensionality Reduction. Vancouver/Canada, 2006

[18] Pal C, Wang X, Kelm M, McCallum A. *Multi-conditional learning for joint probability models with latent variables.* In: Proc of the NIPS Workshop - Advances in Structured Learning for Text and Speech Processing. Vancouver/-Canada, 2005

[19] Zechmann CM, Kelm BM, Zamecnik P, Ikinger U, Waldherr R, Röll S, Delorme S, Hamprecht FA, Bachert P. *Can man still beat the machine? automated vs. manual pattern recognition of 3D MRSI data of prostate cancer patients.* In: Proc of the 16th ISMRM Scientific Meeting and Exhibition. Berlin/Germany, 2007

Book Chapters

[20] Carlsohn MF, Menze BH, Kelm BM, Hamprecht FA, Kercek A, Leitner R, Polder G. Color Image Processing: Methods and Applications, chap. Spectral Imaging and Applications, 393–419. Image Processing. CRC Press, 2006. Vol 7

Other References

[21] Amari SI, Nagaoka H. Methods of Information Geometry, vol. 191. American Mathematical Society, 2001

[22] Bachert P. *MR-Spektroskopie.* In: W Schlege, J Bille (eds.), Medizinische Physik 2, 297–314. Springer, 2002

[23] Balram N, Moura JMF. *Noncausal Gauss-Markov random fields: Parameter structure and estimation.* IEEE Trans Inform Theory 1993. 39(4): 1333–1355

[24] Barndorff-Nielsen O. Information and Exponential Families in Statistical Theory. Wiley Series in Probability and Mathmatical Statistics. John Wiley & Sons, 1978

[25] Bendl R, Pross J, Hoess A, Keller M, Preiser K, Schlegel W. *VIRTUOS - a program for virtual radiotherapy simulation and verification.* In: Proc Intl Conf on The Use of Computers in Radiation Therapy. A. R. Hounsell u. a. Manchester: North Western Med. Physics Dept., 1994 226–227

[26] Besag J, Green P, Higdon D, Mengersen K. *Bayesian computation and stochastic systems.* Stat Sci 1995. 10(1): 3–66

[27] Besag JE. *Spatial interaction and the statistical analysis of lattice systems (with discussion).* J Roy Statist Soc Ser B 1974. 36: 192–236

[28] Besag JE. *On the statistical analysis of dirty pictures.* J Roy Statist Soc Ser B 1986. 48(3): 259–302

[29] Bishop CM. *Latent variabel models.* In: MI Jordan (ed.), Learning in Graphical Models, 371–403. MIT Press, 1999

[30] Bishop CM. Pattern Recognition and Machine Learning. Springer, 2006

[31] Blake A, Rother C, Brown M, Perez P, Toor P. *Interactive image segmentation using an adaptive GMMRF model.* In: Lecture Notes in Computer Science, vol. 3021. 2004 428 – 441

[32] Bonekamp D, Horská A, Jacobs MA, Arslanoglu A, Barker PB. *Fast method for brain image segmentation: application to proton magnetic resonance spectroscopic imaging.* Magn Reson Med 2005. 54(5): 1268–1272. doi:10.1002/mrm.20657

[33] van den Boogart A, van Ormondt D, Pijnappel WWF, de Beer R, Ala-Korpela M. *HLSVD water filtering.* In: JG McWhirter (ed.), Mathematics in Signal Processing III, 175–195. Clarendon Press, Oxford, 1994

[34] Borga M, Landelius T, Knutsson H. *A unified approach to PCA, PLS, MLR and CCA.* Report LiTH-ISY-R-1992, ISY, SE-581 83 Linköping, Sweden, 1997

[35] Bouman C, Sauer K. *A generalized Gaussian image model for edge-preserving MAP estimation.* IEEE Trans Image Processing 1993. 2(3): 296–310

[36] Bouman CA, Sauer K. *A unified approach to statistical tomography using coordinate descent optimization.* IEEE Trans Image Processing 1996. 5(3): 480–492

[37] Boykov Y, Kolmogorov V. *An experimental comparison of Min-Cut/Max-Flow algorithms for energy minimization in vision.* IEEE Trans Pattern Anal Machine Intell 2004. 26(9): 1124–1137

[38] Boykov Y, Veksler O, Zabih R. *Efficient approximate energy minimization via graph cuts.* IEEE Trans Pattern Anal Machine Intell 2001. 20(12): 1222–1239

[39] Breiman L. *Random forests.* Mach Learn 2001. 45(1): 5–32

[40] Brix G, Henze M, Knopp M, Lucht R, Doll J, Junkermann H, Hawighorst H, Haberkorn U. *Comparison of pharmacokinetic MRI and [18F] fluorodeoxyglucose PET in the diagnosis of breast cancer: initial experience.* Eur Radiol 2001. 11(10): 2058–70. doi:10.1007/s003300100944

[41] Brix G, Semmler W, Port R, Schad L, Layer G, Lorenz W. *Pharmacokinetic parameters in CNS Gd-DTPA enhanced MR imaging.* J Comput Assist Tomogr 1991. 15(4): 621–8

[42] Brown LD. Fundamentals of statistical exponential families: with applications in statistical decision theory. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986

[43] Butzen J, Prost R, Chetty V, Donahue K, Neppl R, Bowen W, Li SJ, Haughton V, Mark L, Kim T, Mueller W, Meyer G, Krouwer H, Rand S. *Discrimination between neoplastic and nonneoplastic brain lesions by use of proton MR spectroscopy: the limits of accuracy with a logistic regression model.* Am J Neuroradiol 2000. 21(7): 1213–1219

[44] Byrd RH, Nocedal J, Schnabel RB. *Representations of quasi-newton matrices and their use in limited memory methods.* Math Program 1994. 63: 129–156

[45] Cavassila S, Deval S, Huegen C, van Ormondt D, Graveron-Demilly D. *Cramér-Rao bounds: an evaluation tool for quantitation.* NMR Biomed 2001. 14: 278–283

[46] Chen AP, Cunningham CH, Kurhanewicz J, Xu D, Hurd RE, Pauly JM, Carvajal L, Karpodinis K, Vigneron DB. *High-resolution 3D MR spectroscopic imaging of the prostate at 3 T with the MLEV-PRESS sequence.* J Magn Reson Imag 2006. 24(7): 825–832. doi:10.1016/j.mri.2006.03.002

[47] Cheng H, Bouman C. *Multiscale Bayesian segmentation using a trainable context model.* IEEE Trans Image Processing 2001. 10(4): 511–525

[48] Chilès JP, Delfiner P. Geostatistics: Modeling Spatial Uncertainty. Wiley series in probability and statistics. Wiley, New York, 1999

[49] Chou PB, Brown CM. *The theory and practice of Bayesian image labelling.* Intl J Comp Vision 1990. 4: 185–210

[50] Chu A, Alger JR, Moore GJ, Posse S. *Proton Echo-Planar Spectroscopic Imaging with highly effective outer volume suppression using combined presaturation and spatially selective echo dephasing.* Magn Reson Med 2003. 49: 817–821

[51] Coackley FV, Teh HS, Qayyum A, Swanson MG, Lu Y, Roach III M, Pickett B, Shinohara K, Vigneron DB, Kurhanewicz J. *Endorectal MR imaging and MR spectroscopic imaging for locally recurrent prostate cancer after external beam radiation therapy: Preliminary experience.* Radiology 2004. 233: 441–448

[52] Coleman T, Li Y. *An interior trust region approach for nonlinear minimization subject to bounds.* SIAM J Opt 1996. 6: 418–445

[53] De Bie T, Cristianini N, Rosipal R. *Eigenproblems in pattern recognition.* In: E Bayro-Corrochano (ed.), Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics. Springer, Heidelberg, 2004

[54] De Edelenyi FS, Rubin C, Estéve F, Grand S, Décorps M, Lefournier V, Le Bas JF, Rémy C. *A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images.* Nature Med 2000. 6(11): 1287–1289

[55] Dean BL, Drayer BP, Bird CR, Flom RA, Hodak JA, Coons SW, Carey RG. *Gliomas: classification with MR imaging.* Radiology 1990. 174: 411–415

[56] D'Elia C, Poggi G, Scarpa G. *A tree-structured Markov random field model for Bayesian image segmentation.* IEEE Trans Image Processing 2003. 12(10): 1259–1273

[57] Della Pietra S, Della Pietra V, Lafferty J. *Inducing features of random fields.* IEEE Trans Pattern Anal Machine Intell 1997. 19(4): 380–393. doi:http://dx.doi.org/10.1109/34.588021

[58] Dempster AP. *A generalization of Bayesian inference.* J Roy Statist Soc Ser B 1968. 30(2): 205–247

[59] Descombes X, Kruggel F, von Cramon DY. *fMRI signal restoration using a spatio-temporal Markov Random Field preserving transitions.* Neuroimage 1998. 8(4): 340–349. doi:10.1006/nimg.1998.0372

[60] Descombes X, Kruggel F, von Cramon DY. *Spatio-temporal fMRI analysis using Markov random fields.* IEEE Trans Med Imag 1998. 17(6): 1028–1039

[61] Devos A, Lukas L, Suykens JAK, Vanhamme L, Tate A, Howe F, Majós C, Moreno-Torres A, van der Graaf M, Arús C, Van Huffel S. *Classification of brain tumours using short echo time ^1H MR spectra.* J Magn Reson 2004. 170(1): 164–175. doi:10.1016/j.jmr.2004.06.010

[62] Devos A, Simonetti A, van der Graaf M, Lukas L, Suykens JAK, Vanhamme L, Buydens LMC, Heerschap A, Huffel SV. *The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification.* J Magn Reson 2005. 173(2): 218–28. doi: 10.1016/j.jmr.2004.12.007

[63] Dietterich TG. *Machine learning for sequential data: A review.* In: T Caelli (ed.), Lecture Notes in Computer Science. Springer-Verlag, 2002

[64] Dietterich TG. Structural, Syntactic, and Statistical Pattern Recognition., vol. 2396 of *Lecture Notes in Computer Science*, chap. Machine Learning for Sequential Data: A Review, 15–30. Springer, 2002

[65] Dietterich TG, Bakiri G. *Solving multiclass learning problems via error-correcting output codes.* J Artif Intell Res 1995. 2: 263–286

[66] Dössel O. Bildgebende Verfahren in der Medizin: von der Technik zur medizinischen Anwendung. Springer, 1999

[67] Duda RO, Hart PE, Stork DG. Pattern Classification. Wiley, New York, 2000

[68] Dumas M, Canlet C, Andrè F, Vercauteren J, Paris A. *Metabonomic assessment of physiological disruptions using ^1H-$^{(13)}C$ HMBC-NMR spectroscopy combined with pattern recognition procedures performed on filtered variables.* Anal Chem 2002. 74: 2261–2273

[69] Dydak U, Meier D, Lamerichs R, Boesiger P. *Trading spectral separation at 3T for acquisition speed in multi spin-echo spectroscopic imaging.* Am J Neuroradiol 2006. 27(7): 1441–1446

[70] El-Deredy W. *Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review.* NMR Biomed 1997. 10: 99–124

[71] Farag AA, El-Baz A, Gimel'farb GL. *Precise segmentation of multimodal images.* IEEE Trans Image Processing 2006. 15(4): 952–968

[72] Frank IE, Friedman JH. *A statistical view of some Chemometrics regression tools.* Technom 1993. 35(2): 109–135

[73] Geman S, Geman D. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.* IEEE Trans Pattern Anal Machine Intell 1984. 6: 721–741

[74] Gestel TV, Suykens JAK, Lanckriet G, Lambrechts A, Moor BD, Vandewalle J. *Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis.* Neural Comp 2002. 14(5): 1115–47. doi:10.1162/089976602753633411

[75] Ghahramani Z, Beal M. *Graphical models and variational methods.* In: M Opper, D Saad (eds.), Advanced Mean Field Methods – Theory and Practice. MIT Press, 2001

[76] Ghahramani Z, Hinton GE. *The EM algorithm for mixtures of factor analyzers.* Tech. rep., Department of Computer Science, University of Toronto, 1997

[77] Granlund GH, Knutsson H. Signal processing for computer vision. Kluwer Academic, Dordrecht, NL, 1995

[78] Gray HF, Maxwell RJ, Martinez-Pérez I, Arús C, Cerdán S. *Genetic programming for classification and feature selection: analysis of ^1H nuclear magnetic resonance spectra from human brain tumour biopsies.* NMR Biomed 1998. 11: 217–224

[79] Greig DM, Porteous BT, Seheult AH. *Exact maximum a posteriori estimation for binary images.* J Roy Statist Soc Ser B 1989. 51(2): 271–279

[80] Grimmet G, Stirzaker D. Probability and Random Processes. Oxford University Press, 3rd ed., 2001

[81] Gruber S, Mlynárik V, Moser E. *High-resolution 3D proton spectroscopic imaging of the human brain at 3 T: SNR issues and application for anatomy-matched voxel sizes.* Magn Reson Med 2003. 49(2): 299–306. doi:10.1002/mrm.10377

[82] Günther M, Feinberg DA. *Simultaneous spin echo refocusing.* Magn Reson Med 2005. 54: 513–523

[83] Haase A, Frahm J, Hanicke W, Matthaei D. *1H NMR chemical shift selective (CHESS) imaging.* Phys Med Biol 1985. 30(4): 341–344

[84] Hagberg G. *From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods.* NMR Biomed 1998. 11: 148–156

[85] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, 2001

[86] Hawighorst H, Libicher M, Knopp M, Moehler T, Kauffmann G, Kaick G. *Evaluation of angiogenesis and perfusion of bone marrow lesions: role of semi-quantitative and quantitative dynamic MRI.* J Magn Reson Imag 1999. 10(3): 286–94

[87] He X, Zemel R, Carreira-Perpinan M. *Multiscale conditional random fields for image labelling.* In: Proc IEEE Conf on Comp Vision Pat Rec, vol. 2. 2004 695–702. doi:10.1109/CVPR.2004.1315232

[88] Henning A, Schär M, Schulte RF, Wilm B, Pruessmann KP, Boesiger P. *High field MRSI localization based on T1 and B1 insensitive highly selective outer volume suppression.* Magn Reson Med 2007. Submitted

[89] Horn RA, Johnson CR. Matrix Analysis. Cambridge University Press, 1985

[90] Horn RA, Johnson CR. Topics in Matrix Analysis. Cambridge University Press, 1991

[91] Huber PJ. Robust Statistics. New York: John Wiley and Sons Inc., 1981

[92] Hyvärinen A, Oja E. *Independent component analysis: algorithms and applications.* Neural Net 2000. 13(4-5): 411–30

[93] Ishikawa H. *Exact optimization for markov random fields with convex priors.* IEEE Trans Pattern Anal Machine Intell 2003. 25(10): 1333–1336

[94] Jaakkola T. *Tutorial on variational approximation methods.* In: Advanced mean field methods: theory and practice. MIT Press, 2000

[95] Jähne B. Digital image processing. Springer, Berlin, 5 ed., 2002

[96] Jain AK. Fundamentals of digital image processing. Prentice-Hall, Englewood Cliffs, NJ, 1989

[97] Jebara T. Discriminative, Generative and Imitative Learning. Ph.D. thesis, Media Laboratory, Massachusetts Institute of Technology, 2001

[98] Jebara T, Pentland A. *Maximum conditional likelihood via bound maximization and the CEM algorithm*, 1998

[99] John E Dennis J, Gay DM, Welsch RE. *NL2SOL – an adaptive nonlinear least-squares algorithm.* ACM Trans Math Softw 1981. 7(3): 369–383. doi: http://doi.acm.org/10.1145/355958.355966

[100] Jordan MI, Ghahramani Z, Jaakkola T, Saul LK. *An introduction to variational methods for graphical models.* Mach Learn 1999. 37(2): 183–233

[101] Kamasak ME, Bouman CA, Morris ED, Sauer K. *Direct reconstruction of kinetic parameter images from dynamic PET data.* IEEE Trans Med Imag 2005. 24(5): 636–650

[102] Karatzoglou A, Smola A, Hornik K, Zeileis A. *kernlab - An S4 package for kernel methods in R.* Tech. Rep. 9, Dept. Stat. Math., Wien, 2004

[103] Kelm BM. Demosaicking of Color Images by Means of Conditional Random Fields. Master's thesis, Oregon State University, 2003

[104] Kiessling F, Lichy M, Grobholz R, Heilmann M, Farhan N, Michel MS, Trojan L, Ederle J, Abel U, Kauczor HU, Semmler W, Delorme S. *Simple models improve the discrimination of prostate cancers from the peripheral gland by T1-weighted dynamic MRI.* Eur Radiol 2004. 14(10): 1793–801. doi:10.1007/s00330-004-2386-1

[105] Kolmogorov V, Zabih R. *What energy functions can be minimized via graph cuts?* IEEE Trans Pattern Anal Machine Intell 2004. 26(2): 147–159

[106] Kreis R. *Issues of spectral quality in clinical 1H-magnetic resonance spectroscopy and a gallery of artifacts.* NMR Biomed 2004. 17(6): 361–381. doi:10.1002/nbm.891

[107] Kschischang FR, Frey BJ, Loeliger HA. *Factor graphs and the sum-product algorithm.* IEEE Trans Inform Theory 2001. 47(2): 498–519

[108] Kumar S, Hebert M. *Discriminative fields for modeling spatial dependencies in natural images.* In: Proc Neural Inf Proc Systems. 2003

[109] Kumar S, Hebert M. *Discriminative random fields: a discriminative framework for contextual interaction in classification.* In: Proc IEEE Conf on Comp Vision Pat Rec. 2003 1150–1157

[110] Kumar S, Hebert M. *Discriminative random fields.* Intl J Comp Vision 2006. 68(2): 179–201. doi:10.1007/s11263-006-7007-9

[111] Lafferty J, McCallum A, Pereira F. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In: Proc Intl Conf on Mach Learn. 2001 282–289

[112] Laudadio T, Mastronardi N, Vanhamme L, Hecke PV, Huffel SV. *Improved Lanczos algorithms for blackbox MRS data quantitation.* J Magn Reson 2002. 157(2): 292–297

[113] Laudadio T, Pels P, Lathauwer LD, Hecke PV, Huffel SV. *Tissue segmentation and classification of MRSI data using canonical correlation analysis.* Magn Reson Med 2005. 54(6): 1519–1529. doi:10.1002/mrm.20710

[114] Laudadio T, Selén Y, Vanhamme L, Stoica P, Hecke PV, Huffel SV. *Subspace-based MRS data quantitation of multiplets using prior knowledge.* J Magn Reson 2004. 168(1): 53–65. doi:10.1016/j.jmr.2004.01.015

[115] Lauterbur PC. *Image formation by induced local interactions: Examples employing nuclear magnetic resonance.* Nature 1973. 242: 190–191

[116] Lauterbur PC. *Nobel Lecture. All science is interdisciplinary–from magnetic moments to molecules to men.* Biosci Rep 2004. 24(3): 165–178

[117] Le QV, Smola AJ, Canu S. *Heteroscedastic Gaussian process regression.* In: Proc Intl Conf on Mach Learn. 2005 489 – 496

[118] Li BS, Regal J, Gonen O. *SNR versus resolution in 3D 1H MRS of the human brain at high magnetic fields.* Magn Reson Med 2001. 46(6): 1049–1053

[119] Li SZ. Markov random field modeling in image analysis. Computer Science Workbench. Springer, Tokyo, 2nd ed., 2001

[120] Liang Z, MacFall JR, Harrington DP. *Parameter estimation and tissue segmentation from multispectral MR images.* IEEE Trans Med Imag 1994. 13(3): 441–449

[121] Liers F, Jünger M, Reinelt G, Rinaldi G. New Optimization Algorithms in Physics, chap. Computing Exact Ground States of Hard Ising Spin Glass Problems by Branch-and-cut, 47–70. Wiley, 2004

[122] Lisboa PJG, Kirby SPJ, Vellido A, Lee YYB, El-Deredy W. *Assessment of statistical and neural networks methods in NMR spectral classification.* NMR Biomed 1998. 11: 225–234

[123] Lukas L, Devos A, Suykens JAK, Vanhamme L, Howe F, Majós C, Moreno-Torres A, der Graaf MV, Tate A, Arús C, Huffel SV. *Brain tumor classification based on long echo proton MRS signals.* Artif Intell Med 2004. 31(1): 73–89. doi:10.1016/j.artmed.2004.01.001

[124] Lukas L, Devos A, Suykens JAK, Vanhamme L, Van Huffel S, Tate AR, Majós C, Arús C. *The use of LSSVM in the classification of brain tumors based on ^{1}H-MR spectroscopy signals.* In: Proc IEEE Workshop on Medical Applications of Signal Processing. 2002 15/1–15/5

[125] Mansfield P, Maudsley AA. *Medical imaging by NMR.* Br J Radiol 1977. 50(591): 188–194

[126] Marshall I, Higinbotham J, Bruce S, Freise A. *Use of Voigt lineshape for quantification of in vivo 1H spectra.* Magn Reson Med 1997. 37(5): 651–657

[127] Marx BD. *Iteratively reweighted partial least squares estimation for generalized linear regression.* Technom 1996. 38(4): 374–381

[128] Marx BD, Eilers PHC. *Generalized linear regression for sampled signals or curves: A p-spline approach.* Technom 1999. 41(1): 1–13

[129] McGill R, Tukey JW, Larsen WA. *Variations of box plots.* Am Stat 1978. 32(1): 12–16

[130] McIntosh AR, Bookstin FL, Haxby JV, Grady CL. *Spatial pattern analysis of functional brain images using partial least squares.* Neuroimage 1996. 3: 143–157

[131] Mierisová S, Ala-Korpela M. *MR spectroscopy quantitation: a review of frequency domain methods.* NMR Biomed 2001. 14: 247–259

[132] Mierisová S, van den Boogaart A, Tkác I, Hecke PV, Vanhamme L, Liptaj T. *New approach for quantitation of short echo time in vivo 1H MR spectra of brain using AMARES.* NMR Biomed 1998. 11(1): 32–39

[133] Milstein AB, Webb KJ, Bouman CA. *Estimation of kinetic model parameters in fluorescence optical diffusion tomography.* J Opt Soc Am A 2005. 22(7): 1357–1368

[134] Minka T. *Discriminative models, not discriminative training.* Tech. rep., Microsoft Research Cambridge, 2005

[135] Mitchell DG, Cohen MS. MRI Principles. Elsevier, 2nd ed., 2004

[136] Moon TK, Stirling WC. Mathematical methods and algorithms. Prentice Hall, Upper Saddle River, NJ, 2000

[137] Moré JJ. The Levenberg-Marquardt Algorithm: Implementation and Theory., 105–116. Lecture Notes in Mathematics 630. Springer, 1977

[138] Müller M. Generalized Linear Models, vol. 1 of *Handbook of Computational Statistics.* Springer-Verlag, Heidelberg, 2004

[139] Naressi A, Couturier C, Castang I, de Beer R, Graveron-Demilly D. *Java-based graphical user interface for MRUI, a software package for quantitation of in vivo/medical magnetic resonance spectroscopy signals.* Comp Bio Med 2001. 31: 269–86

[140] Naressi A, Couturier C, Devos JM, Janssen M, Mangeat C, de Beer R, Graveron-Demilly D. *jMRUI, MRUI for Java.* Magn Reson Mat in Phys, Biol and Med 2001. 12(2-3): 141–152

[141] Nattkemper TW, Wismüller A. *Tumor feature visualization with unsupervised learning.* Med Image Anal 2005. 9(4): 344–51. doi:10.1016/j.media.2005.01.004

[142] Neal R, Hinton G. *A view of the EM algorithm that justifies incremental, sparse, and other variants.* In: MI Jordan (ed.), Learning in Graphical Models. Kluwer, 1998 355 – 368

[143] Neff T. Verbesserung der Zielvolumendefinition in der Strahlentherapieplanung durch den Einsatz der biologischen Bildgebung. Ph.D. thesis, University of Mannheim, 2005

[144] Nelson SJ. *Multivoxel magnetic resonance spectroscopy of brain tumors.* Molec Cancer Therap 2003. 2: 497–507

[145] Ng AY, Jordan M. *On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes.* In: Neural Information Processing Systems, 14. 2002 657 – 664

[146] Nikulin AE, Dolenko B, Bezabeh T, Somorjai RL. *Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra.* NMR Biomed 1998. 11: 209–216

[147] Nishimori H. Statistical Physics of Spin Glasses and Information Processing. Oxford University Press, 2001

[148] Nosàs-Garcia S, Moehler T, Wasser K, Kiessling F, Bartl R, Zuna I, Hillengass J, Goldschmidt H, Kauczor HU, Delorme S. *Dynamic contrast-enhanced MRI for assessing the disease activity of multiple myeloma: a comparative study with histology and clinical markers.* J Magn Reson Imag 2005. 22(1): 154–162. doi:10.1002/jmri.20349

[149] Noworolski SM, Henry RG, Vigneron DB, Kurhanewicz J. *Dynamic contrast-enhanced MRI in normal and abnormal prostate tissue as defined by biopsy, MRI, and 3D MRSI.* Magn Reson Med 2005. 53: 249–255

[150] Ou W, Golland P. *From spatial regularization to anatomical priors in fMRI analysis.* In: Proc IPMI, no. 3565 in LNCS. 2005 88–100

[151] Pearl J. Probabilistic Reasoning in Intelligent Systems. CA. Morgan Kaufmann, 1988

[152] Pfanzagl J. Parametric Statistical Theory. Walter de Gruyter, 1994

[153] Pietra S, Pietra V, Lafferty J. *Duality and auxiliary functions for Bregman distances.* Tech. Rep. CMU-CS-01-109R, School of Computer Science, Carnegie Mellon University, 2002

[154] Pijnappel WWF, van den Boogart A, de Beer R, van Ormondt D. *SVD-based quantification of magnetic resonance signals.* J Magn Reson 1992. 97: 122–134

[155] Provencher SW. *Estimation of metabolite concentrations from localized* in vivo *proton NMR spectra.* Magn Reson Med 1993. 30: 672–679

[156] Provencher SW. *Automatic quantitation of localized* in vivo *^{1}H spectra with LCModel.* NMR Biomed 2001. 14e: 260–264

[157] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0

[158] Rabiner LR. *A tutorial on Hidden Markov Models and selected applications in speech recognition.* Proc IEEE 1989. 77(2): 257–286

[159] Ratiney H, Mitri F, Coenradie Y, Cavassila S, Van Ormondt D, Graveron-Demilly D. *Time-domain quantitation based on a metabolite basis set in magnetic resonance spectroscopy.* In: Proceedings of ProRISC-IEEE. 2002 432 – 437

[160] Ratiney H, Sdika M, Coenradie Y, Cavassila S, van Ormondt D, Graveron-Demilly D. *Time-domain semi-parametric estimation based on a metabolite basis set.* NMR Biomed 2005. 18(1): 1–13. doi:10.1002/nbm.895

[161] Ripley BD. Pattern Recognition and Neural Networks. Cambridge Univ. Press, 1997

[162] Robert CP, Casella G. Monte Carlo Statistical Methods. Springer Texts in Statistics. Springer, 2nd ed., 2004

[163] Rockafellar RT. Convex Analysis. Princeton University Press, 1970

[164] Roda JM, Pascual JM, Carceller F, Gonzàlez-Llanos F, Pérez-Higueras A, Solivera J, Barrios L, Cerdán S. *Nonhistological diagnosis of human cerebral tumours by ^{1}H magnetic resonance spectroscopy and amino acid analysis.* Clin Cancer Res 2000. 6: 3983–3993

[165] Roweis S, Ghahramani Z. *A unifying review of linear Gaussian models.* Tech. rep., Gatsby Computational Neuroscience Unit, University College London, 6 King's College Road, Toronto M5S 3H5, Canada, 1997

[166] Rubinstein YD, Hastie T. *Discriminative vs informative learning.* In: Knowledge Discovery and Data Mining. 1997 49–53

[167] Saindane AM, Cha S, Law M, Xue X, Knopp EA, Zagzag D. *Proton MR spectroscopy of tumefactive demyelinating lesions.* Am J Neuroradiol 2002. 23: 1378–1386

[168] Sajda P, Du S, Brown TR, Stoyanova R, Shungu DC, Mao X, Parra LC. *Non-negative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain.* IEEE Trans Med Imag 2004. 23(12): 1453–65

[169] Salakhutdinov R, Roweis S, Ghahramani Z. *Optimization with EM and expectation-conjugate-gradient.* In: Proc Intl Conf on Mach Learn. 2003 672–679

[170] Salibi N, Brown MA. Clinical MR Spectroscopy: First Principles. Wiley, 1998

[171] Salvan AM, Confort-Gouny S, Chabrol B, Cozzone PJ, Vion-Dury J. *Brain metabolic impairment in non-cerebral and cerebral forms of X-linked adrenoleukodystrophy by proton MRS: identification of metabolic patterns by discriminant analysis.* Magn Reson Med 1999. 41(6): 1119–1126

[172] Scharr H. Optimale Operatoren in der digitalen Bildverarbeitung. Ph.D. thesis, University of Heidelberg, 2000

[173] Scheenen T, Weiland E, Futterer J, van Hecke P, Bachert P, Villeirs G, Lu J, Lichy M, Holshouser B, Roell S, Barentsz J, Heerschap A. *Preliminary results of IMAPS: An international multi-centre assessment of prostate MR spectroscpoy.* In: Proc Intl Soc Magn Reson Med, 13. Springer, 2005

[174] Scheenen TWJ, Klomp DWJ, Röll SA, Fütterer JJ, Barentsz JO, Heerschap A. *Fast acquisition-weighted three-dimensional proton MR spectroscopic imaging of the human prostate.* Magn Reson Med 2004. 52(1): 80–8. doi:10.1002/mrm.20103

[175] Scheidler J, Hricak H, Vigneron DB, Yu KK, Sokolov DL, Huang LR, Zaloudek CJ, Nelson SJ, Carroll PR, Kurhanewicz J. *Prostate cancer: Localization with three-dimensional proton MR spectrosopic imaging – clinicopathologic study.* Radiology 1999. 213: 473–480

[176] Schlemmer HP, Bachert P, Hence M, Buslei R, Herfarth KK, Debus J, Kaick G. *Differentiation of radiation necrosis from tumor progression using proton magnetic resonance spectroscopy.* Neuroradiol 2002. 44: 216–222

[177] Schlemmer HP, Bachert P, Herfarth KK, Zuna I, Debus J, van Kaick G. *Proton MR spectroscopic evaluation of suspicious brain lesions after stereotactic radiotherapy.* Am J Neuroradiol 2001. 22: 1316–1324

[178] Schlemmer HP, Merkle J, Grobholz R, Jaeger T, Michel MS, Werner A, Rabe J, van Kaick G. *Can pre-operative contrast-enhanced dynamic MR imaging for prostate cancer predict microvessel density in prostatectomy specimens?* Eur Radiol 2004. 14(2): 309–17. doi:10.1007/s00330-003-2025-2

[179] Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2001

[180] Schulte RF, Henning A, Tsao J, Boesiger P, Pruessmann KP. *Design of broadband RF pulses with polynomial phase response.* J Magn Reson 2007. In press

[181] Schwartzkopf W, Bovik A, Evans B. *Maximum-likelihood techniques for joint segmentation-classification of multispectral chromosome images.* IEEE Trans Med Imag 2005. 24(12): 1593– 1610

[182] Shafer G. *Jeffrey's rule of conditioning.* Phil Sci 1981. 48(3): 337–362

[183] Sima DM, Van Huffel S. *Regularized semiparametric model identification with application to nuclear magnetic resonance signal quantification with unknown macromolecular base-line.* J Royal Stat Soc: Series B 2006. 68(3): 383. doi: 10.1111/j.1467-9868.2006.00550.x

[184] Simonetti AW, Melssen WJ, de Edelenyi FS, van Asten JJA, Heerschap A, Buydens LMC. *Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification.* NMR Biomed 2005. 18(1): 34–43. doi:10.1002/nbm.919

[185] Simonetti AW, Poullet JB, Sima DM, De Neuter B, Vanhamme L, Lemmerling P, Van Huffel S. *An open source short echo time MR quantitation software solution: AQSES.* Tech. Rep. ESAT-SISTA/TR 2005-168, Katholieke Universiteit Leuven, 2006

[186] Slotboom J, Boesch C, Kreis R. *Versatile frequency domain fitting using time domain models and prior knowledge.* Magn Reson Med 1998. 39(6): 899–911

[187] Stoica P, Selén Y, Sandgren N, Huffel SV. *Using prior knowledge in SVD-based parameter estimation for magnetic resonance spectroscopy–the ATP example.* IEEE Trans Biomed Eng 2004. 51(9): 1568–1578

[188] Stone M, Brooks RJ. *Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression.* J Roy Statist Soc Ser B 1990. 52(2): 237–269

[189] Stoyanova R, Brown TR. *NMR spectral quantitation by principal component analysis.* NMR Biomed 2001. 14: 271–277

[190] Swindle P, McCredie S, Russell P, Himmelreich U, Khadra M, Lean C, Mountford C. *Pathologic characterization of human prostate tissue with proton MR spectroscopy.* Radiology 2003. 228(1): 144–151

[191] Tate AR, Foxall PD, Holmes E, Moka D, Spraul M, Nicholson JK, Lindon JC. *Distinction between normal and renal cell carcinoma kidney cortical biopsy samples using pattern recognition of ^1H magic angle spinning (MAS) NMR spectra.* NMR Biomed 2000. 13: 64–71

[192] Tate AR, Griffiths JR, Martinez-Pérez I, Moreno A, Barba I, Cabañas ME, Watson D, Alonso J, Bartumeus F, Isamat F, Ferrer I, Vila F, Ferrer E, Capdevila A, Arús C. *Towards a method for automated classification of ^1H MRS spectra from brain tumors.* NMR Biomed 1998. 11: 177–191

[193] Tate AR, Majós C, Moreno A, Howe FA, Griffiths JR, Arús C. *Automated classification of short echo time in in vivo ^1H brain tumor spectra: a multicenter study.* Magn Reson Med 2003. 49(1): 29–36. doi:10.1002/mrm.10315

[194] Tipping ME, Bishop CM. *Mixtures of probabilistic principal component analysers.* Neural Comp 1999. 11(2): 443–482

[195] Tipping ME, Bishop CM. *Probabilistic principal component analysis.* J Roy Statist Soc Ser B 1999. 61: 611–622

[196] Tofts P, Brix G, Buckley D, Evelhoch J, Henderson E, Knopp M, Larsson H, Lee T, Mayr N, Parker G, Port R, Taylor J, Weisskoff R. *Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols.* J Magn Reson Imag 1999. 10(3): 223–32

[197] Tofts PS. *Modeling tracer kinetics in dynamic Gd-DTPA MR imaging.* J Magn Reson Imag 1997. 7(1): 91–101

[198] Tosi MR, Fini G, Tinti A, Reggiani A, Tugnoli V. *Molecular characterisation of human healthy and neoplastic cerebral and renal tissues by in vitro ^1H NMR spectroscopy (review).* Intl J Molec Med 2002. 9: 299–310

[199] Twellmann T, Lichte O, Nattkemper TW. *An adaptive tissue characterization network for model-free visualization of dynamic contrast-enhanced magnetic resonance image data.* IEEE Trans Med Imag 2005. 24(10): 1256–66

[200] Tzika AA, Zurakowski D, Young Poussaint T, Goumnerova L, Astrakas LG, Barnes PD, Anthony DC, Billet AL, Tarbell NJ, Scott RM, Black PM. *Proton magnetic spectroscopic imaging of the child's brain. the response of tumors to treatment.* Neuroradiol 2001. 43

[201] Ulusoy I, Bishop CM. *Generative versus discriminative methods for object recognition.* In: IEEE Conf on Comp Vision Pat Rec, vol. 2. 2005 258– 265. doi:10.1109/CVPR.2005.167

[202] Van Huffel S, Chen H, Decanniere C, Van Hecke P. *Algorithm for time-domain NMR data fitting based on total least squares.* J Magn Reson 1994. 110: 228–237

[203] Van Huffel S, Laudadio T. *Magnetic Resonance Spectroscopic Imaging: a survey of quantification and classification algorithms.* Tech. Rep. 05-249, K. U. Leuven (ESAT-SISTA), Leuven (Belgium), 2005

[204] Vanhamme L, van den Boogaart A, Van Huffel S. *Improved method for accurate and efficient quantification of MRS data with use of prior knowledge.* J Magn Reson 1997. 129(1): 35–43

[205] Vanhamme L, Lennerling P, Van Huffel S. *Comments on "Confidence Images for MR Spectroscopic Imaging" by Karl Young, Dennis Khetselius, Brian J. Soher, and Andrew A. Maudsley (Magn Reson Med 2000; 44:537-545).* Magn Reson Med 2001. 46: 1254–1255

[206] Vanhamme L, Sundin T, Van Hecke P, Van Huffel S. *MR spectroscopy quantitation: a review of time-domain methods.* NMR Biomed 2001. 14: 233–246

[207] Wainwright MJ, Jaakkola T, Willsky AS. *A new class of upper bounds on the log partition function.* IEEE Trans Inform Theory 2005. 51: 2313–2335

[208] Wainwright MJ, Jordan MI. *Graphical models, exponential families, and variational inference.* Tech. Rep. 649, Department of Statistics, University of California, Berkeley, 2003

[209] Weber MA, Zoubaa S, Schlieter M, Jüttler E, Huttner HB, Geletneky K, Ittrich C, Lichy MP, Kroll A, Debus J, Giesel FL, Hartmann M, Essig M. *Diagnostic*

performance of spectroscopic and perfusion MRI for distinction of brain tumors. Neurology 2006. 66(12): 1899–1906. doi:10.1212/01.wnl.0000219767.49705.9c

[210] Weber-Fahr W, Ende G, Braus DF, Bachert P, Soher BJ, Henn FA, Buechel C. *A fully automated method for tissue segmentation and CSF-correction of proton MRSI metabolites corroborates abnormal hippocampal NAA in schizophrenia.* Neuroimage 2002. 16: 49–60

[211] Weinman JJ, Learned-Miller E. *Improving recognition of novel input with similarity.* In: Proc IEEE Conf on Comp Vision Pat Rec. New York, 2006 308–315. doi:10.1109/CVPR.2006.151

[212] Wilhelm T, Aisenbrey C, Bachert P. *Schnelle ^{1}H-spektroskopische Bildgebung am Gehirn des Menschen*

[213] Winkler G. Image Analysis, Random Fields and Dynamic Monte Carlo Methods, vol. 27 of *Applications of Mathematics.* Springer, 2nd ed., 2003

[214] Wold S, Ruhe A, Wold H, Dunn WJ. *The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses.* SIAM J Sci Stat Comput 1984. 5(3): 735–743

[215] Woolrich MW, Behrens TEJ, Beckmann CF, Smith SM. *Mixture models with adaptive spatial regularization for segmentation with an application to fmri data.* IEEE Trans Med Imag 2005. 24(1): 1–11

[216] Ye JC, Webb KJ, Bouman CA, Millane RP. *Optical diffusion tomography using iterative coordinate descent optimization in a Bayesian framework.* J Opt Soc Am A 1999. 16: 2400–2412

[217] Yedidia J, Freeman W, Weiss Y. *Constructing free-energy approximations and generalized belief propagation algorithms.* IEEE Trans Inform Theory 2005. 51(7): 2282–2312

[218] Yedidia JS, Freeman WT, Weiss Y. *Understanding belief propagation and its generalizations.* Tech. Rep. TR-2001-22, Mitsubishi Electric Research Laboratories, 2002

[219] Young K, Khetselius D, Soher BJ, Maudsley AA. *Confidence images for MR spectroscopic imaging.* Magn Reson Med 2000. 44: 537–545

[220] Zakian KL, Eberhardt S, Hricak H, Shukla-Dave A, Kleinman S, Muruganandham M, Sircar K, Kattan MW, Reuter VE, Scardino PT, Koutcher JA. *Transition zone prostate cancer: Metabolic characteristics at ^{1}H MR spectroscopic imaging – initial results.* Radiology 2003. 229(1): 241–151

[221] Zakian KL, Sircar K, Hricak H, Chen HN, Shukla-Dave A, Eberhardt S, Muruganandham M, Ebora L, Kattan MW, Reuter VE, Scardino PT, Koutcher JA. *Correlation of proton MR spectroscopic imaging with gleason score based on step-section pathologic analysis after radical prostatectomy.* Radiology 2005. 234: 804–814

[222] Zandt Hi, van der Graaf M, Heerschap A. *Common processing of* in vivo *MR spectra.* NMR Biomed 2001. 14: 224–232

[223] Zechmann C, Baudendistel K, Aftab K, Trojan L, Michel MS, Kauczor HU, Delorme S. *Dynamic contrast-enhanced T1-weighted MRI combined with MR spectroscopic imaging in patients with prostate cancer - initial experience.* In: Proc Intl Soc Mag Reson Med, vol. 13. 2005