Mathematical modeling and visualization of functional neuroimages

Rasmussen, Peter Mondrup; Hansen, Lars Kai; Madsen, Kristoffer Hougaard

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Mathematical modeling and visualization of functional neuroimages

Peter Mondrup Rasmussen

Kongens Lyngby 2011
IMM-PHD-2011-267

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This dissertation presents research results regarding mathematical modeling in the context of the analysis of functional neuroimages. Specifically, the research focuses on pattern-based analysis methods that recently have become popular analysis tools within the neuroimaging community. Such methods attempt to predict or *decode* experimentally defined cognitive states based on brain scans. The topics covered in the dissertation are divided into two broad parts:

The first part investigates the relative importance of model selection on the brain patterns extracted form analysis models. Typical neuroimaging data sets are characterized by relatively few data observations in a high dimensional space. The process of building models in such data sets often requires strong regularization. Often, the degree of model regularization is chosen in order to maximize prediction accuracy. We focus on the relative influence of model regularization parameter choices on the model generalization, the reliability of the spatial brain patterns extracted from the analysis model, and the ability of the model to identify relevant brain networks defining the underlying neural encoding of the experiment. We show that known parts of brain networks can be overlooked in pursuing maximization of prediction accuracy. This supports the view that the quality of spatial patterns extracted from models cannot be assessed purely by focusing on prediction accuracy. Our results instead suggest that model regularization parameters must be carefully selected, so that the model and its visualization enhance our ability to interpret brain function.

The second part concerns interpretation of nonlinear models and procedures for extraction of 'brain maps' from nonlinear kernel models. We assess the performance of the sensitivity map as means for extracting a *global* summary map from a trained model. Such summary maps provides the investigator with an overview of brain locations of importance to the model's predictions. The sensitivity map proves as a versatile technique for model visualization. Furthermore, we perform a preliminary investigation of the use of pre-image estimation for *localized* interpretation of nonlinear models. In the context of image denoising the pre-image analysis proves to enhance the reliability of brain patterns extracted from multivariate models of the neuroimaging data.

Resumé

Denne afhandling præsenterer forskningsresultater omhandlende matematisk modellering indenfor analyse af funktionelle hjernescanningsbilleder. Specifikt fokuserer afhandlingen på mønster-baserede analysemetoder, som nyligt er blevet populære indenfor hjerneforskning. Ved hjælp af sådanne modelleringsmetoderne forsøger forskere at prædiktere en eksperimentelt defineret mental tilstand ud fra hjernescanningsdata. Afhandlingen omhandler emner, der kan inddeles i to dele.

Første del undersøger hvorledes modelvalg indvirker på hjerneaktiveringsmønstre, som dannes på baggrund af analysemodeller. Typiske datasæt indenfor hjernescanningsforsøg indeholder relativt få observationer med en høj dimensionalitet. Modellering af sådanne datasæt kræver ofte en kraftig kompleksitetskontrol i modellerne. Ofte vælges graden af kompleksitetskontrol med henblik på at maksimere modellernes prædiktive nøjagtighed. Vi fokuserer på hvilken indvirkning valg af modelkompleksitet har på i) modellernes evne til at prædiktere nøjagtigt, ii) pålideligheden af hjerneaktiveringsmønstre, som dannes på baggrund af modellerne, og iii) modellernes evne til at identificere relevante strukturer i aktiveringsmønstre. Vi viser, at dele af velkendte aktiveringsmønstre kan blive overset, hvis der ensidigt fokuseres på maksimering af den prædiktive nøjagtighed. Disse observationer underbygger en anskuelse om, at kvaliteten af hjerneaktiveringsmønstre ikke kan vurderes på baggrund af en ensidig betragtning af den prædiktive nøjagtighed. Vores resultater indikerer, at modellers kompleksitet skal vælges nøjagtigt, således at modellerne og deres visualiseringer kan bidrage til en øget indsigt i den menneskelige hjernes funktion.

Anden del omhandler tolkningen af ikke-lineære modeller og procedurer til ekstraktion af hjerneaktiveringsmønstre fra ikke-lineære *kernel* modeller. Vi undersøger hvorvidt et *sensitivity map* er brugbart i forbindelse med global modelvisualisering. Globale visualiseringer danner et billede af hjerneaktiveringsmønstre og giver forskere mulighed for at få et samlet indblik i hvordan en prædiktiv model fungerer. *Sensitivity mappet* viser sig som en alsidig metode til modelvisualisering. Endvidere foretager vi en indledende undersøgelse af *pre-image* analyse som et værktøj til lokaliseret tolkning af ikke-linære modeller. I forbindelse med støjreduktion i hjernescanningsbilleder viser *pre-image* analyse sig som en effektiv metode, som muliggør en pålidelig identifikation af hjerneaktiveringsmønstre.

Preface

This dissertation was prepared at DTU Informatics, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering. The dissertation is based on work done over the period December 2008 - December 2011.

Lyngby, December 2011

Peter Mondrup Rasmussen

Acknowledgements

I am grateful to the volunteers who generously participated in the series of studies that provided data sets to the work in the project. I thank the three advisors of my Ph.D. project. Lars Kai Hansen, the Technical University of Denmark (DTU), who introduced me to this amazing research field that emerges when mathematical modeling and neuroscience interact. Torben Ellegaard Lund, the Danish National Research Foundation's Center of Functionally, Integrative Neuroscience (CFIN), who contributed with his great understanding on all aspect regarding neuroimaging ranging from planning of experiments to data analysis. Furthermore, I also thank Torben for initiating my collaboration with CFIN. Kristoffer Hougaard Madsen, the Danish Recearch Centre for Magnetic Resonance (DRCMR), who shared his impressive knowledge on machine learning within neuroimaging contexts. In particular I thank Kristoffer for providing the possibility of using the MR scanner facilities at DRCMR, Hvidovre Hospital, and for helping with data acquisition in late Friday nights. I would also here like to thank Stephen Charles Strother, the Rotman Research Institute (RRI), Baycrest, Toronto, Canada. Despite not serving as a formal supervisor of my project Stephen became highly involved during the project. I am grateful for his involvement and great contribution based on his impressive knowledge within the neuroimaging research field.

I would also like to thank my colleagues at both the Cognitive Systems group at DTU Informatics, and at CFIN for their company during the past three years. In particular, I wish to thank the people that have been involved in various parts of my project, including Morten Mørup, Carsten Stahlhut, Trine Julie Abrahamsen, all at DTU, and Lars Ribe at CFIN. I also wish to thank Ulla Nørhave at DTU for her kind assistance and patience regarding administrative aspects of my project. Additionally I wish to thank Kristjana Ýr Jónsdóttir, Kim Mouridsen, and Mikkel Wallentin, all at CFIN, for their involvement in different interesting projects that unfortunately did not succeed in becoming part of this thesis. I thank Leif Østergaard at CFIN for providing me with the opportunity of collaborating with researchers at CFIN. During the project I have also had great benefits from access to the cluster systems within the MINDLab core facilities. I also thank my collaborators at Center for Integrated Molecular Brain Imaging (CIMBI). In particular, I wish to thank Gitte Moos Knudsen, David Erritzøe, Vibe Gedsø Frøkjær, and Mette Haahr.

During my project I had a number of stays abroad. I wish to thank Stephen Charles Strother for organizing my stay at RRI. I thank all people at the Strother Lab for a warm welcome and for ensuring that I had a pleasant stay. In particular I thank Grigori Yourganov at RRI for his great hospitality. I also thank Grigori Yourganov, Nathan Churchill at RRI, and Tanya Schmah at University of Toronto, for continuing our collaboration after my stay. I participated in the 2009 University of California, Los Angeles (UCLA) NeuroImaging Training Program (NITP) Advanced Neuroimaging Summer School, and wish to thank the organizers Russell Poldrack at University of Texas at Austin, Mark Cohen at UCLA, and Susan Bookheimer at UCLA for inviting me to the summer school. I also participated in the 4th International Summer School in Biomedical Engineering, Leipzig, and wish to thank the organizers Jens Haueisen at Technischen Universität Ilmenau and Thomas Knösche at the Max Planck Institute for inviting me to the summer school.

I gratefully acknowledge the support of my financial sponsors. The PhD project was financed through a DTU Informatics scholarship. The Otto Mønsteds Foundation supported my participation in the Biosignals 2009 conference. The 4th International Summer School in Biomedical Engineering supported my participation with a travel grant. The Neuroimaging Training Program was supported by sponsorship of the NIH, through awards DA022768 an DA023422, and offered complementary admission to the summer school as well as free shared housing on the UCLA campus during the 2009 UCLA NITP Advanced Neuroimaging Summer School.

Finally, I wish to thank my friends, my whole family and my beloved Ulla for your support during my project and for filling me with energy.

Publications

During the project I have contributed to a number of scientific publications. The dissertation is based on the contributions highlighted with ✓.

Journal papers

[✓] Rasmussen, P. M., Abrahamsen, T. J., Madsen, K. H., Hansen, L. K., 2011. Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-image estimation. NeuroImage 60 (3), 1807-1818.

[✓] Rasmussen, P. M., Madsen, K. H., Churchill, N. W., Hansen, L. K., Strother, S. C., 2011. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45 (6), 2085-2100.

[✓] Rasmussen, P. M., Madsen, K. H., Lund, T. E., Hansen, L. K., 2011. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage 55 (3), 1120 - 1131.

[ ] Churchill, N. W., Yourganov, G., Spring, R., Rasmussen, P. M., Lee, W., Ween, J. E., Strother, S. C., 2012. PHYCAA: Data-driven measurement and removal of physiological noise in BOLD fMRI. NeuroImage 59 (2), 1299 - 1314.

[ ] Erritzøe, D., Frøkjær, V. G., Haugbøl, S., Marner, L., Svarer, C., Holst, K., Baaré, W. F. C., Rasmussen, P. M., Madsen, J., Paulson, O. B., Knudsen, G. M., 2009. Brain serotonin 2A receptor binding: Relations to body mass index, tobacco and alcohol use. NeuroImage 46 (1), 23 - 30.

[ ] Erritzøe, D., Frøkjær, V. G., Holst, K. K., Christoffersen, M., Johansen, S. S., Svarer, C., Madsen, J., Rasmussen, P. M., Ramsøy, T., Jernigan, T. L., Knudsen, G. M., 2011. In vivo imaging of cerebral serotonin transporter and serotonin(2A) receptor binding in 3,4- methylenedioxymethamphetamine (MDMA or "ecstasy") and hallucinogen users. Archives of General Psychiatry 68 (6), 562 - 76.

[ ] Haahr, M. E., Rasmussen, P. M., Madsen, K., Marner, L., Ratner, C. Gillings, N. Baaré W. F. C., Knudsen, G. M., 2012. Obesity is associated with high serotonin 4 receptor availability in the brain reward circuitry. NeuroImage 61 (4), 884-888.

[ ] Frøkjær, V. G., Erritzøe, D., Holst, K. K., Jensen, P. S., Rasmussen, P. M., Fisher, P. M., Baaré W., Madsen, K. S., Madsen, J., Svarer, C., Knudsen, G. M., 2012. Prefrontal serotonin transporter availability is positively associated with the cortisol awakening response. European Neuropsychopharmacology 2012.

[ ] Yourganov, G., Schmah, T., Small, S., Rasmussen, P. M., Strother, S. C., 2010. Functional connectivity metrics during stroke recovery. Archives Italiennes de Biologie 148 (3).

Conference papers

[✓] Rasmussen, P. M., Schmah, T., Madsen, K. H., Lund, T. E., Yourganov, G., Strother, S. C., Hansen, L. K., 2012. Visualization of nonlinear classification models in neuroimaging - signed sensitivity maps. In: Biosignals 2012 International Conference on Bio-inspired Systems and Signal Processing.

[ ] Frøkjær, V. G., Erritzøe, D., Jensen, P., Rasmussen, P. M., Holst, K. K., Madsen, J., Baaré, W., Knudsen, G. M., 2011. Frontal cortex serotonin transporter binding is positively associated with basal physiological stress reactivity in healthy volunteers. European Neuropsychopharmacology 21, (Suppl. 1), S74.

[ ] Bjerre, T., Henriksen, J., Nielsen, C. H., Rasmussen, P. M., Hansen, L. K., Madsen, K. H., 2009. Unified ICA-SPM analysis of fMRI experiments: Implementation of an ICA graphical user interface for the SPM pipeline. In: Biosignals 2009 International Conference on Bio-inspired Systems and Signal Processing.

[ ] Wilkowski, B., Szewczyk, M. M., Rasmussen, P. M., Hansen, L. K., Nielsen, F. A., 2010. Coordinate-based meta-analytic search for the SPM neuroimaging pipeline : The BredeQuery plugin for SPM5. In: Healthinf 2009, Proceedings of the Second International Conference on Health Informatics.

Poster presentations and conference abstracts

[✓] Rasmussen, P. M., Madsen, K. H., Lund, T. E., Hansen, L. K., 2010. Visualization of predictive models in neuroimaging - the sensitivity map. 16th Annual Meeting of the Organization for Human Brain Mapping.

[✓] Rasmussen, P. M., Madsen, K. H., Strother, S. C., Hansen, L. K., 2010. Sparseness in predictive models is associated with reduced pattern reproducibility. 16th Annual Meeting of the Organization for Human Brain Mapping.

Book chapters

[ ] Wilkowski, B., Szewczyk, M. M., Rasmussen, P. M., Hansen, L. K., Nielsen, F. A., 2010. BredeQuery: Coordinate-Based Meta-analytic Search of Neuroscientific Literature from the SPM Environment. Vol. 52 of Communications in Computer and Information Science. Springer-Verlag Berlin Heidelberg.

Contents

Abbreviations and notation

AFNI	Analysis of functional neuroimages (software package)
ANN	Artificial neural network
ANOVA	Analysis of variance
ARD	Automatic relevance determination
BOLD	Blood oxygenation level dependent
CB	Cerebellum
CBF	Cerebral blood flow
CBV	Cerebral blood volume
CCA	Canonical correlation analysis
$CMRO_2$	Cerebral metabolic rate of oxygen
CVA	Canonical variates analysis
DCM	Dynamic causal modeling
dHB	Deoxygenated hemoglobin
DWI	Diffusion-weighted imaging
EEG	Electroencephalography
ENET	Elastic net
EPI	Echo planar imaging

FDA	Fisher's discriminant analysis
FDR	False discovery rate
fMRI	Functional magnetic resonance imaging
FMRIB	Oxford centre for functional MRI of the brain
FSL	FMRIB Software Library (software package)
FSW	Feature space weighting
FWHM	Full width half maximum
GLM	General linear model
GMM	Gaussian mixture model
GNB	Gaussian Naïve Bayes
ICA	Independent component analysis
IF	Inferior-frontal
IRLS	Iteratively re-weighted least squares
KFDA	Kernel Fisher's discriminant analysis
KLR	Kernel logistic regression
KPCA	Kernel principal component analysis
KRR	Kernel ridge regression
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LogReg	Logistic regression
LP	Left precentral
MEG	Magnetoencephalography
MI	Mutual information
MNI	Montreal Neurological Institute
MNI152	Montreal Neurological Institute template
MO	Medial orbitofrontal
MR	Magnetic resonance

MRI	Magnetic resonance imaging
MT	Middle temporal
MVB	Multivariate Bayesian decoding
MVPA	Multi-voxel pattern analysis
NPAIRS	Nonparametric, prediction, activation, influence, reproducibility, re-sampling
OEF	Oxygen extraction fraction
OLS	Ordinary least squares
PBAIC	Pittbsurgh Brain Activity Interpretation Competition
PC	Principal component
PCA	Principal component analysis
PCC	Posterior cingulate cortex
PCR	Principal component regression
PDA	Penalized discriminant analysis
PET	Photon emission tomography
PHYCCA	Physiological correction using canonical autocorrelation analysis
PLS	Partial least squares
PreC	Precuneus
RETROICOR	Retrospective correction of physiological motion effects in fMRI
RFE	Recursive feature elimination
RKHS	Reproducing kernel Hilbert space
ROI	Region of interest
rSPI	Reproducible SPI
RVM	Relevance vector machine
S2	Secondary somatosensory cortex
SC	Subcortical regions
SMC	Sensorimotor cortex

SPECT	Single photon emission tomography
SPI	Statistical parametric image
SPL	Superior parietal lobes
SPM	Statistical parametric mapping (software package)
SVD	Singular value decomposition
SVM	Support vector machine
TR	Repetition time
TV	Total variation
VT	Ventral temporal
WB	Whole brain

Notation

x	Scalar
$\mathbf{x}$	Vector
$\mathbf{X}$	Matrix
$\mathcal{M}$	Three-dimensional brain scan volume
$\mathcal{X}$	Input domain (voxel space)
$\mathcal{F}$	Feature space
$\mathbf{I}_n$	Diagonal matrix of dimension $(n \times n)$ with ones in the diagonal
$\mathbf{1}_n$	Row vector of ones with n rows.
$\delta(\cdot)$	Dirac delta function
$\mathrm{var}(\cdot)$	Variance measure
$\mathrm{tr}(\cdot)$	Trace of a matrix
$(\cdot)^{-}$	Moore-Penrose generalised/pseudo inverse
$\mathrm{orth}(\cdot)$	Orthogonalization operation
$(\cdot)^{\top}$	Transpose

Chapter 1

Reading guide

The following provides an overview of the content of the dissertation. The dissertation is based on a series of publications, and the intention of the dissertation is to form a connection between these publications. Furthermore, the dissertation attempts to discuss the contribution of the Ph.D. project in a wider neuroimaging context. The dissertation is written to be self containing. Hence, there is a considerable overlap between results presented in this dissertation and the work reported as scientific publications during the Ph.D. project.

The dissertation is divided into six chapters. Following this introduction is Chapter 2 that provides a general motivation for our work. Hereafter, a brief review of different parts of the neuroimaging pipeline ranging from experimental design to interpretation of analysis results is provided. Finally, the main contributions of the Ph.D. project are outlined.
Chapter 3 provides background on mathematical modeling techniques used in the project.
Chapter 4 introduces the data sets used in the project.
Experimental results are presented in Chapter 5.
Finally, Chapter 6 concludes the dissertation and outlines future research perspectives.

Chapter 2

Dissertation background, context, and contribution

The central topic covered in this dissertation is *pattern-based analysis* of neuroimaging data. Our interest in research within this particular field is based on observations provided in the first section of this chapter. The second section discuss a series of important elements involved in the generation, analysis, and interpretation of neuroimaging data sets. These elements are discussed in the context of the dissertation focus. The last section outlines the contribution of the Ph.D. project.

Contents

2.1 Neuroimaging background

The neuroscientific research field concerns the study of the nervous system. Measuring various signals from the nervous system is often referred to as *neuroimaging*. Neuroimaging techniques provide means for mapping human brain function in time and space. The practical application of functional neuroimaging is broad. Applications range from basic research attempting to understand the physiology underlying the measured brain signals or to understand information processing in the healthy and the diseased brain, over brain-computer interfaces helping paralyzed people to communicate by measuring brain signals, to more controversial 'mind reading' applications e.g. attempting to use neuroimaging techniques as advanced polygraphs. A variety of measurement techniques offer unique opportunities to perform non-invasive measurements on the human brain. Emission tomography, e.g. photon emission tomography (PET) and single photon emission tomography (SPECT) allow e.g. receptor systems to be mapped by administering radionuclides to the human body. Electroencephalography (EEG) and magnetoencephalography (MEG) relies on measurements based on electromagnetic fields generated from ionic current flows caused by the neuron's electric behavior. Examples of other measurement techniques are ultrasound and optical imaging. The work in the present thesis focuses on measurements acquired with functional magnetic resonance imaging (fMRI). Each measurement modality has its own characteristics and ability to provide insight into different aspects of a particular brain system under study. Modalities can often be combined to provide more detailed descriptions. An introduction to the physics of medical imaging modalities is found in Bushberg et al. (2001).

fMRI is an active research field. A database query[1] searching for paper title, abstract, or keywords containing either *functional magnetic resonance imaging* or *fMRI* resulted in 4,026 publications in the calendar year 2010. Other queries were performed targeted at specific analysis procedures used in the analysis of data sets acquired with fMRI. Figure 2.1(A) is based on three early papers of some of the pioneers involved in the development of data analysis strategies, e.g. Friston et al. (1994, 1995b); Worsley and Friston (1995). These analysis strategies focus on characterizing regional specific effects of the experimental design in fMRI data by use of the so called mass-univariate analysis. Note that the number of citations seems to reach a plateau around the millennium followed by a slight decrease. This should not be taken as evidence that investigators have stopped using these analysis strategies. Possible other explanations are: i) these analysis strategies have been well established and investigators do not find it relevant to cite these methodological papers, ii) investigators now cite more recent papers. Figure 2.1(B) is based both early and more recent papers exploiting

[1] SciVerse Scopus citation database (www.scopus.com).

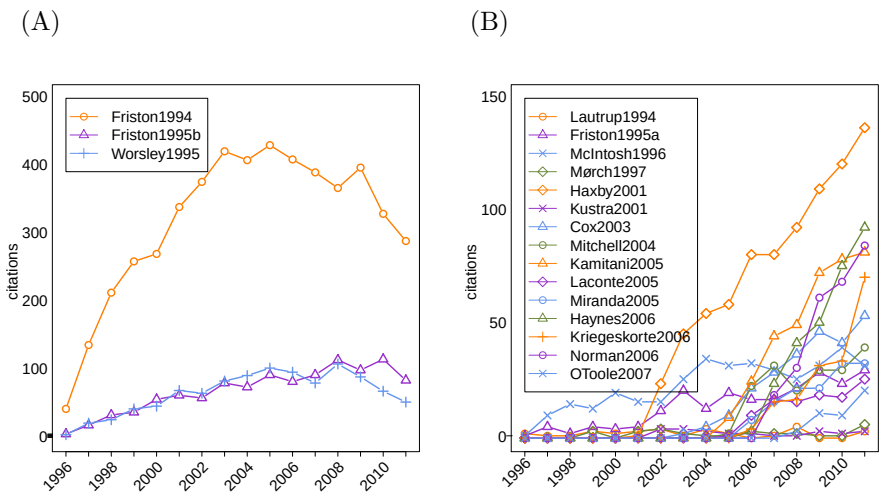

Figure 2.1: Literature search of the number of citations per year for a series of articles concerned with the analysis of functional magnetic resonance images. Statistics were retrieved using the SciVerse Scopus citation database as of late November 2011. **Panel A:** Some of the first publications on the mass-univariate analysis procedure. **Panel B:** Some early and more recent publications on the pattern-based analysis procedure.

analysis strategies that focus on the characterization of distributed brain patterns by use of the so called pattern-based analysis procedure. Recent papers exploiting pattern-based analysis by means of the support vector machines (LaConte et al., 2005; Mourão-Miranda et al., 2005) are seemingly becoming more popular than established pattern-based analysis procedures such as the partial least squares analysis (McIntosh et al., 1996). Haxby et al. (2001); Kamitani and Tong (2005); Kriegeskorte et al. (2006) presented evidence that pattern-based analysis provides insight into aspects of the acquired fMRI data that could not be detected with existing established and recognized analysis strategies. Haynes and Rees (2006); Norman et al. (2006); O'Toole et al. (2007) are review papers. Many papers formulate the pattern-based analysis as a classification problem. Lautrup et al. (1994); Mørch et al. (1997); Kustra and Strother (2001) are examples of early papers, explicitly formulating the data analysis as a classification problem, that are rarely cited.

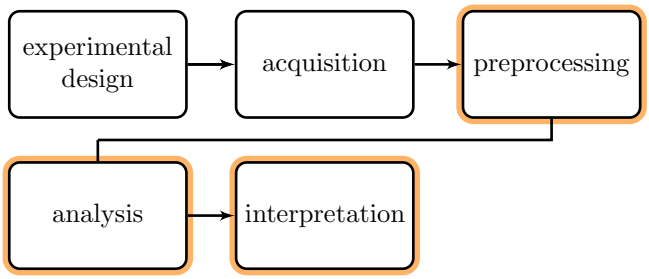

Figure 2.2: The neuroimaging pipeline. The diagram shows different parts of the neuroimaging pipeline. This dissertation primarily focuses on aspects regarding the preprocessing, data analysis, and the analysis interpretation.

2.2 The neuroimaging pipeline

The following provides an overview of elements in the *neuroimaging pipeline*. 'Pipeline' here refers to a series of choices and assumptions made by the investigator during a neuroimaging experiment. Figure 2.2 shows one way to organize the pipeline. The present dissertation focuses, in particular, on aspects regarding preprocessing, data analysis, and the interpretation of the analysis results. The following section reviews some of the assumptions and choices involved in each step of the pipeline.

2.2.1 Experimental design

The use of neuroimaging techniques is typically motivated by a desire to identify how information is processed in the human brain. Examples could be that we are interested in characterization of brain responses evoked by viewing images of faces or identification of brain areas that are involved in finger movement. The simplest type of experimental paradigms are based on the subtraction method where the basic logic is as follows: i) The participant is engaged in two experimental conditions defined as experimental blocks e.g. a condition of interest and a control condition. ii) The two conditions are expected to be characterized by the same cognitive or sensorimotor processes or signal structures except the process of interest. iii) Subtracting signals acquired under the two conditions gives a difference that is attributed the cognitive/sensorimotor process of interest. Typical blocks of stimulation have lengths $5 \sim 20$ seconds (Bandettini et al., 1992). Another experimental design is the event-related design aiming on detection of responses evoked by single trails, e.g. presentation of

a single sound (Buckner et al., 1996). Event-related designs allow for stimuli randomization[2] and may also be less affected to adaptation effects than block designs. A general introduction to experimental design in functional neuroimaging is found in e.g. Faro and Mohamed (2010). Other experiments acquire data under less controlled experimental settings. Examples are fMRI studies where participants were subjected to movie watching (Hasson et al., 2004) or story telling (Wallentin et al., 2011). In the Pittsburgh Brain Activity Interpretation Competition[3] (PBAIC) participants were engaged in navigating in a virtual reality environment. Investigations were provided with brain scans and subjects' behavioral measurements, and the objective was to build a pattern analysis systems that reliably could predict behavioral data from scans originating form a test run. Both block designs and event related designs are used in studies using pattern-based analysis techniques, e.g. Haxby et al. (2001); Kamitani and Tong (2005); Kriegeskorte et al. (2006).

2.2.2 Acquisition

The most common measured signal in fMRI is the blood oxygenation level dependent (BOLD) signal (Ogawa et al., 1992). The BOLD signal reflects the relative presence of oxygenated and deoxygenated hemoglobin. Following neural activity the local cerebral blood flow (CBF) will increase more than the increase in the cerebral metabolic rate of oxygen ($CMRO_2$) resulting in a decrease in the oxygen extraction fraction (OEF). The decrease in OEF will alter the relative levels of oxygenated and deoxygenated hemoglobin (dHB). Since these have different magnetic properties the change in the relative levels is measurable. Hence, a decrease in the content of dHB will lead to a BOLD signal increase. However, the measured BOLD signal does not measure the underlying neural activity directly. Instead the BOLD signal has a complex dependence on changes in CBF, $CMRO_2$, and the cerebral blood volume (CBV). Changes in CBF, $CMRO_2$, and CBV are also collectively referred to as the hemodynamic response to activation. A critical issue in interpreting fMRI experiments is an understanding of the relationship between the underlying neural activity, the hemodynamic response, and the measured BOLD signal (Buxton et al., 2004). The physiological relationship between neural activity, the hemodynamic response, and the BOLD signal is unclear, and different mathematical models, that primarily incorporates CBF, CBV, and $CMRO_2$ as dynamical variables, have been proposed, e.g. Buxton et al. (1998, 2004); Friston et al. (2000). Reviews on the interpretation of the BOLD signal are found in e.g. Dinesh (2005); Logothetis (2008).

[2]Block designs can also be randomized. However, experimentally evoked effects may be removed by a high-pass filter due to the slow nature of the design.

[3]http://pbc.lrdc.pitt.edu .

2.2.3 Organization of the data from an fMRI experiment

Typical neuroimaging data sets are characterized by a series of acquired variables. One way to organize the variables is to use a partitioning into *mesoscopic variables* and *macroscopic variables*.

Mesoscopic variables are here defined to be the voxels' time series acquired during the experiment. A single brain scan volume can be considered as organized into the tensor $\mathcal{M} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, where d_i are the number of voxels in the three dimensions. Without loss of generality we can organize $\mathcal{M}$ into the vector $\mathbf{m} \in \mathbb{R}^{P \times 1}$ where $P = d_1 d_2 d_3$. If N scans are acquired during the experiment we can further organize these into the *measurement matrix* $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_N]$,

Macroscopic variables refer to a series of other variables collected before, during, or after the scan acquisitions. Examples of such variables are information about the experimental paradigm, subject behavior during the experiment, physiological nuisance variables, and movement parameters estimated as part of preprocessing of the acquired scans. The macroscopic variables can be collected into the vector $\mathbf{g} \in \mathbb{R}^{R \times 1}$, with R being the number of collected variables. The mesoscopic variables can be further organized into the *design matrix* $\mathbf{D} = [\mathbf{g}_1, \ldots, \mathbf{g}_N]$ similar to the organization of the measurement matrix $\mathbf{M}$.

2.2.4 Preprocessing

Motion correction

Motion correction refers to the process of aligning individual scan volumes of a temporal sequence. This registration is important for a number of reasons (Friston et al., 1996): i) The signal changes, due to the hemodynamic response evoked by the experimental paradigm, can be small in comparison to signal differences originating from subject movement. A correlation between subject head motion and the experimental paradigm will result in spurious results, where signal changes due to motion are detected as 'activation'. ii) Movement-related signal will contribute to error variance of a statistical model. Hence, the test statistics will be smaller than if the movement related effects were removed. Motion correction is implemented in terms of a rigid body registration. Such registration involves a step that estimates transformation parameters, a step that apply the transformation parameters, and a re-slicing step where the registered image is written out. The rigid body realignment assumes that the movement of brain locations can be described by a linear transformation and that any motion happens between the acquisition of individual scans. While the first assumption

may be reasonable for most brain locations, the latter is hard to meet, since the process of scan acquisition is basically continuous over time. Some motion related artifacts are highly nonlinear (e.g. magnetization contamination as a result of spin excitation effects) (Friston et al., 1996). Such artifacts are not directly removed by rigid body realignment. Rigid body realignment is in general considered to be a standard preprocessing step. An investigation of the impact of motion correction on pattern-based analysis models is found in e.g. Zhang et al. (2009). Analysis results in Chen et al. (2006) provides an example of apparent task related effects which the authors interpret as originating from interaction between susceptibility and condition-related movement rather than 'real activation'. These effects were detected in a data set that was subjected to motion correction prior to the analysis. Hence, inspection of the brain pattern underlying a model's predictions becomes important.

Inter-modality or co-registration refers to the process of aligning scan volumes acquired with different modalities. Examples are registration of functional scans to a structural scan in order to localize effects based on anatomical information, or to perform a registration of functional scans to a standard template where the registration parameters have been estimated based on a structural scan. Co-registration is closely related to motion correction and also involves the estimation of parameters of a linear transformation model. The scans volumes subjected to motion correction will in general have a similar voxel intensity distribution, and a commonly used objective is to minimize the sum of squares difference between the volumes. Signal intensities may vary considerably across modalities and it may be advantageous to use other cost functions e.g. mutual information or normalized mutual information, see e.g. Friston et al. (2007) and references therein.

Stereotactic registration

Stereotactic registration refers to the process of the registration of scan volumes to a common template. Such registration enables the investigator to report findings in terms of coordinates defined within a standard coordinate system. Examples are the templates provided by the Montreal Neurological Institute (MNI). Stereotactic registration is also a commonly used procedure in multi-subject analyses in order to facilitate a voxel based analysis across subjects. Typically the stereotactic registration involves estimation of parameters of models that allow for non-linear warping. Examples of normalization procedures are 'low dimensional' warping based on a limited number of basis functions e.g. Ashburner and Friston (1999, 2005), 'high dimensional' warping based on flexible deformation methods e.g. Ashburner (2007), and surface based registration e.g. Van Essen (2004).

An alternative to the use of spatial normalization in multi-subject studies is parcellation methods. These methods attempt to derive groups of parcels that are coherent across subjects. Further information on such procedures is found in e.g. Thirion et al. (2006); Michel et al. (2011b) and references therein.

Spatial filtering

The use of spatial filtering or *smoothing* can be motivated by the following (Friston et al., 2007): i) By considering the signal to noise ratio it can be argued that the spatial filter should match the size of the signal to detect (matched filter theorem), ii) smoothing will render the error terms in the general linear model more normal making the statistical inference more valid, iii) smoothing may lead to better fulfillment of model assumptions when performing statistical thresholding according the random field theory, iv) smoothing may be necessary when the spatial registration procedure is insufficient to provide a good alignment of similar (functionally or anatomically) structures across subjects.

The above points i) and iv) are relevant in the context of pattern-based analysis. The impact of spatial smoothing has systematically been investigated within a resampling framework in LaConte et al. (2003). These authors demonstrated that spatial smoothing increased the performance of canonical variates analysis (CVA) (regularized by truncating a principal component analysis (PCA) basis) both in terms of prediction accuracy and the reproducibility of brain maps extracted from the model. The same behavior was found by Strother et al. (2004) demonstrating a rapid rise in both prediction accuracy and reproducibility for small amounts of smoothing from 1 to 2 pixel full width half maximum (FWHM) of a Gaussian in-plane filter. LaConte et al. (2005) provided a comparison of a CVA (regularized by truncating a PCA basis) and the support vector machine (SVM) (soft-margin) for different combinations of preprocessing steps, showing that i) for high levels of temporal filtering mainly the SVM showed decreased prediction accuracy with decreased level of spatial smoothing, ii) for low levels of temporal filtering both models, in particular the CVA, showed decreased prediction accuracy with decreased levels of spatial smoothing. Mourão-Miranda et al. (2005) reported empirical results on the performance of an SVM (hard-margin version) and linear discriminant analysis (LDA) (without regularization) concluding that by spatial smoothing i) the prediction accuracy was increased for both models and most prominently for LDA, ii) the spatial smoothing affected the brain maps obtained from both methods - in particular the map of the LDA. Recently, Op de Beeck (2010) reported results from grating experiments and object experiments, where the performance of pattern correlation analysis and SVM predictions were evaluated as a function of spatial smoothing. It was observed, that increasing the spatial smoothing increased the performance

of pattern correlation analysis, while the performance of the SVM was not affected by the spatial smoothing. Interestingly, the paper of Op de Beeck (2010) initiated some debate based on a following hypothesis that was presented in the paper: *"if multivariate analyses are picking up a small-scale functional organization, then it can be expected that smoothing will be detrimental to the ability to decode these fine-scale spatial signals"*. In a comment Kamitani and Sawahata (2010) argued that spatial smoothing not necessarily hurt the information residing in fine-scale patterns by the following argument: Spatial smoothing is essentially a linear transformation of the data. Hence, one can recover the original data from smoothed data by applying the inverse transformation. If some optimum (linear) discriminant function is constructed on the original data it is possible to construct a corresponding discriminant function in the smooth data. Classification according to the two discriminant functions will results in identical decisions. It is pointed out that whether it is advantageous to use smooth or unfiltered data depends on the model's ability to estimate a good decision function based on the given data. The key point is that absence of decreased performance when imposing spatial smoothing does not imply that information is represented only on larger scales. Information is preserved in the filtered data. Other contributions to the debate are found in Kriegeskorte et al. (2010); Shmuel et al. (2010). Indeed, the *searchlight* analysis strategy has be motivated as an analysis approach that exploits fine-grained spatial signal structures in data not subjected to spatial smoothing (Kriegeskorte et al., 2006; Raizada and Kriegeskorte, 2010). An interesting alternative to smoothing with a Gaussian filter kernel is an analysis procedure based on steerable filters acting as spatial basis functions (Friman et al., 2003). This analysis adapts filters to regions surrounding a center voxel and the localized analysis is performed throughout the entire brain (as in searchlight analyses).

Temporal filtering

The fMRI time series are characterized by structured noise that is not related to the experimental effects of interest. Such structures originates from a series of nuisance sources e.g. low frequency scanner drift, the cardiovascular system, and respiration, see e.g. Lund et al. (2006) and references therein. Typically, such effects are attempted reduced by applying a high-pass filter to the time series. Such filtering is implemented e.g. in terms of a set of orthogonal basis functions, defined by a set of discrete cosine basis functions up to a certain cut-off frequency as in the SPM software package (Friston et al., 1995b) or Legendre polynomials as in the AFNI software package (Cox, 1996). Another procedure for identification of basis functions is the retrospective correction of physiological motion effects in fMRI (RETROICOR) (Glover et al., 2000). RETROICOR constructs a set of basis functions based on physiological mea-

surements of cardiac and respiratory signals acquired during scan acquisition. These basis functions are included in the design matrix. The filtering process removes the modeled components using a least squares fit. This is equivalent to orthogonalizing the input time series with respect to the basis time series. Another important class of filters is decomposition methods, that attempt to learn the temporal structure of the noise directly from the fMRI time series. Examples are principal component analysis (PCA) (Bullmore et al., 1996; Hansen et al., 1999; Thomas et al., 2002) and independent component analysis (ICA) (McKeown et al., 1998, 2003). Such methods create a new set of variables as a linear combination of the original time series. The main idea is that some of the new variables can be regarded as relevant 'signal' components, while others are regarded as 'noise' components. Another example is the physiological correction using canonical autocorrelation analysis (PHYCCA) denoising method (Churchill et al., 2012b). This method identifies autocorrelated physiological noise sources with reproducible spatial structure, using canonical correlation analysis performed in a split-half resampling framework. Filtering by decomposition methods requires selection of a subset of the components, and it is a challenge to perform this selection in an unbiased way. For example, various classification schemes have been proposed in order to automatically identify signal and noise components (Thomas et al., 2002; De Martino et al., 2007; Tohka et al., 2008; Churchill et al., 2012b). Filtering by decomposition methods also amounts to an orthogonalization of the original time series with respect to the basis time series as identified by the decomposition. Systematic investigations of the impact of temporal filtering on predictive models are found in e.g. LaConte et al. (2003, 2005); Chen et al. (2006). Likewise, temporal filtering has been demonstrated to improve the performance of kernel regression models (Chu et al., 2011a). There seems to be converging evidence that temporal filtering improves model performance both with respect to prediction accuracy and the reproducibility of brain maps extracted from pattern-based analysis models.

Feature extraction and selection

The above preprocessing steps are quite standard in standard univariate analysis pipelines. Feature extraction and selection are often used as additional preprocessing steps when building pattern-based analysis models on neuroimaging data.

Feature extraction here refers to the process of generating a new set of variables based on the original variables (voxels). Such new variables can be constructed with decomposition methods, e.g. PCA (Bullmore et al., 1996) and ICA (McKeown et al., 1998). These methods construct a new set of variables as linear combinations of the original variables. An example of nonlinear feature extrac-

tion is kernel PCA (KPCA). KPCA has been successfully applied as a feature extraction stage in a computer-aided diagnosis system, that was build to distinguish Alzheimer's disease subjects from a control group (López et al., 2009). These authors used KPCA to extract nonlinear features from SPECT images and subsequently trained linear and nonlinear classifiers on the KPCA feature representation. Within analysis of fMRI, Thirion and Faugeras (2003) used KPCA to perform nonlinear dimensionality reduction prior to modeling, while Song et al. (2008) used KPCA and pre-image estimation to derive a nonlinear frequency analysis scheme for noise removal. Guo (2010) used KPCA as an feature extractor in analysis of fMRI data. A simple feature extraction procedure, instead of building models on the raw fMRI time series, is to derive features as temporal averages of scan blocks. Mourão-Miranda et al. (2006) investigated the impact of temporal averaging blocks on the performance of the SVM (soft-margin, build with default value of the regularization parameter, $C = 1$). It was observed that in comparison to models build on raw data i) temporal averaging led to an increased prediction accuracy, ii) the brain maps derived from the models build on temporal averaged data were more similar to maps obtained with a mass-univariate analysis (as evaluated by visual inspection of thresholded brain maps). Chen et al. (2006) investigated the impact of temporal averaging in combination with variation in other elements of the preprocessing pipeline (e.g. temporal filtering). It was found that the overall prediction accuracy increased with temporal averaging. Another related feature extraction procedure is to fit a general linear model to each voxel's time series and define features by either beta estimates or t-values. Investigations of the impact of such feature extraction procedures on the predictive performance of a variety of classifiers are found in Misaki et al. (2010); Mumford et al. (2011).

Feature selection here refers to the process of selecting a subset of the variables/features for further analysis. A comprehensive introduction of variable and feature selection is found in Guyon and Elisseeff (2003). De Martino et al. (2008) discuss and compare different feature selection procedures in the context of fMRI classification analysis. Perhaps the simplest method for feature selection is selection of voxels based on a priori defined region of interest (ROI). ROIs can be identified e.g. based on knowledge from prior studies in literature or on functional localizer scans. Another simple approach is to use some univariate selection criterion e.g. analysis of variance (ANOVA) based feature selection. Cox and Savoy (2003) used both ROI selection and ANOVA selection in classification analysis of patterns of fMRI activation evoked by various categories of objects presented as visual stimulation. They reported increased prediction accuracy with an increased number of voxels included reaching an asymptote in accuracy with ~ 100 voxels included in the models. ROI and ANOVA selection are examples of *filter* methods that select features as a preprocessing step. *Wrappers* are methods where feature selection is more integrated into the modeling process. Such methods use a predictive model to rank variables according

to their relative importance to the model. An example is recursive feature elimination (RFE) that recursively eliminates features according to their importance. Lautrup et al. (1994) performed RFE in a neural network in classification of PET scans. RFE has been used together with SVMs for classification of brain scans in e.g. (De Martino et al., 2008; Hanson and Halchenko, 2008). It was shown that RFE led to an increase in prediction accuracy when features were recursively eliminated. Finally, the process of feature selection is an integrated part of the model training procedure in *embedded* methods. A comprehensive discussion of these methods is found in Lal et al. (2006). Embedded methods perform feature selection by incorporating e.g. a sparsity enforcing term into the cost function that is subject to optimization during the training phase. An example within the Bayesian learning framework is automatic relevance determination (ARD) where feature selection is achieved by introducing a sparsity enforcing prior over the model's weights (MacKay, 1992, 1994). Yamashita et al. (2008) proposed a sparse logistic regression model using ARD as a model that automatically selects voxels relevant for classification of fMRI patterns. In data from a simple four quadrant visual paradigm they demonstrated, that the proposed model was able to either maintain or to increase prediction accuracy compared to a conventional logistic regression model (without feature selection). Additionally, the sparse model selected ~ 10 voxels out of approximately 6000 voxels. A variability in the selected variables across different splits of the data was reported. Furthermore, a procedure for stable feature selection was proposed. This procedure selects stable features according to the frequency at which individual features are included in the sparse model across different splits of the data. Examples of studies proposing models with the *elastic net* penalty for embedded feature selection are Grosenick et al. (2008); Carroll et al. (2009); Ryali et al. (2010), demonstrating capability of the models in maintaining or increasing prediction accuracy and to identify a subset of voxels, forming a distributed pattern, as being important to the predictive model. Recently Michel et al. (2011c) proposed total variation (TV) regularization for fMRI pattern classification. TV regularization performs feature selection (and fit the models weights) in such a way that voxels that are close in space will have similar weights in the predictive model.

2.2.5 Supervised analysis I - Classical statistical modeling

Rationale behind mass-univariate analysis

Mass-univariate analysis approaches allow for classical inferences about regionally specific effects of the experimental design on the measured brain signals, e.g. (Friston et al., 1994). The assumed underlying model of brain function

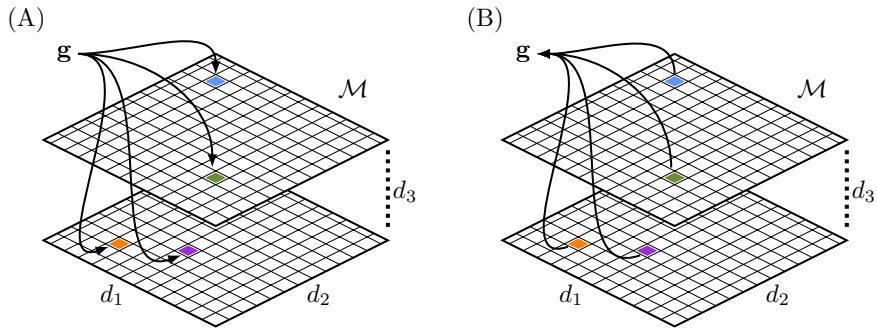

Figure 2.3: Supervised analysis. Conceptual difference between the 'conventional' mass-univariate analysis shown in (A) and the more 'recent' pattern-based analysis procedure shown in (B). The planes illustrate scan slices each composed of a number of voxels. In both procedures we are interested in the statistical relationship between the macroscopic variables in $\mathbf{g}$ and the mesoscopic variables in $\mathcal{M}$. The overall relationship can be modeled by the joint density $p(\mathbf{g}, \mathcal{M})$. Models in (A) attempts to explain the observed brain signals in terms of knowledge on the experimental paradigm. Hence, we are interested in modeling $p(\mathcal{M}|\mathbf{g})$. Models in (B) attempts to explain the experimental paradigm in terms of the measured brain 'responses' to stimulation. Hence, we are here interested in modeling $p(\mathbf{g}|\mathcal{M})$.

is functional specialization. For example, in block design experiments one may consider an activation study (two conditions - 'baseline' and 'active'), and the mass-univarite analysis approach focuses on identifying localized brain areas that are 'activated' according to the experimental design.

Mass-univariate analysis approach

The mass-univariate analysis procedure is the most prevalent analysis strategy within the neuroimaging community. Consider a data set from a neuroimaging experiment $\{(\mathbf{g}_n, \mathcal{M}_n)\}_{n=1}^{N}$, with N being the number of scans. The goal is to learn a so-called *encoding* model $f : \mathbf{g} \rightarrow \mathcal{M}$, that explains how information on the paradigm is encoded in the brain. Figure 2.3(A) provides an illustration of the direction of 'information flow' in univariate analyses. Within a probabilistic framework we are interested in modeling $p(\mathbf{m}|\mathbf{g})$, where $\mathbf{m}$ is the vectorized scan volume $\mathcal{M}$. The mass-univariate analysis assumes that the distribution

over the voxels factorizes such that

$$p(\mathbf{m}|\mathbf{g}) = \prod_{i=1}^{P} p(m_i|\mathbf{g}). \tag{2.1}$$

Hence, it is assumed that the voxels m are conditionally independent given $\mathbf{g}$. Specifically, the model in eq. (2.1) is implemented in terms of the general linear model (GLM), and parameters of the GLM, $\boldsymbol{\theta}_i$, are estimated in each and every voxel separately. Constructing test statistics, based on model parameters, allows for regionally specific hypotheses to be tested. While modeling of the brain signals takes place at the level of individual voxels, the multivariate nature of the signals is taken into account during a subsequent thresholding of statistical brain maps, e.g. by means of the random field theory. The mass-univariate analysis procedure is well implemented in a series of widely used software packages, e.g. in AFNI (Cox, 1996), FSL (Smith et al., 2004), and SPM (Friston et al., 2007).

2.2.6 Supervised analysis II - Machine learning in neuroimaging

Rationale behind pattern-based analysis

Recently, pattern-based analysis methods have gained much attention within the neuroimaging community. The ultimate goal of pattern-based analysis is to link patterns of brain 'activation' to some experimental defined cognitive state. The basic question that such methods often attempt to answer is :"Is it possible, based on the information in a activation pattern, to predict the cognitive state that a subject was engaged in during the experiment?". The idea of generation of predictions or classifications based on brain scans is not new (O'Toole et al., 2007). Early work includes classification of PET images. Clark et al. (1991) used LDA to successfully classify patients with Huntington's disease vs. healthy controls. Another example is the sub-profile scaling model that allows for quantification of disease progression or severity based on brain scans (Moeller and Strother, 1991). Pattern-based analysis was performed by use of artificial neural networks in the early 1990s. Lautrup et al. (1994) performed classification of whole brain PET scans (141,375 voxels), while Mørch et al. (1997) classified both fMRI and PET. Pattern-based analysis is also well established in terms of the partial least squares (PLS) analysis procedure (McIntosh et al., 1996; McIntosh and Lobaugh, 2004).

Over the past decade pattern-based analysis has been re-introduced within the neuroimaging community as e.g. *brain reading* (Cox and Savoy, 2003),

multi-voxel pattern analysis (MVPA) (Norman et al., 2006), *mental state decoding* (Haynes and Rees, 2006), and *information-based functional brain mapping* (Kriegeskorte et al., 2006) methods. In a study of object category representation in the ventral temporal (VT) cortex Haxby et al. (2001) investigated how information about categories of objects (e.g. faces, houses, and chairs) was represented in the VT cortex. By use of a simple pattern correlation classifier they were able to successfully predict the object category based on brain patterns of activation. Additionally, they performed an analysis excluding voxels that respond maximally to specific categories. Even with such voxels excluded from the classification analysis they were able to predict category label well above chance level. By adopting a pattern-based analysis procedure this study supported a view, that the information on faces and objects in the VT cortex is distributed and overlapping. Another study motivating the use of pattern-based analysis was reported by Kamitani and Tong (2005). They trained a classifier to predict stimulus orientation based on brain activation in the early visual cortex. Individual voxels were shown to provide poor response selectivity for different stimulus orientation. However, by integrating information across space, by use of pattern-based analysis, it was possible to correctly predict stimulus orientation by use of a linear classifier.

Figure 2.4 illustrates two scenarios, where a pattern-based analysis procedure could extract more information from a data set than could be done in a conventional mass-univariate analysis. In Figure 2.4(A) only voxel x_1 shows stimulus selectivity in terms of change in mean activation strength. Hence, it would be possible to detect the signal in x_1 using a mass-univariate analysis procedure. However, voxel x_1 and x_2 are correlated. By considering both voxels (i.e. multi-voxels analysis) it is fairly straightforward to infer which of the two experimental conditions a particular example belongs to. Hence, it could be concluded that x_1 and x_2 are informative with respect to the stimulus condition. A linear classifier will have the form $f(x_1, x_2) = w_1 x_1 + w_2 x_2 + b$, (orange examples assigned label 1, and green examples assigned label -1). Note that w_1 will be negative and w_2 will be positive according to the orientation of the weight vector. Often, linear models are visualized with a 'brain map' showing the weights of the weight vector corresponding to the individual voxels. However it is clear from Figure 2.4(A) that e.g. a positive value of w_2 not implies that voxel x_2 respond stronger on average to the orange condition than the green condition. Hence, claims about mean difference in activation strengths in individual voxels across conditions cannot be made based on the inspection of the sign of the model's weights. The logic from interpreting mass-univariate models in terms of mean differences does not apply directly to the interpretation of classification model. The positive value of w_2 indicates that increasing the signal in x_2 for a given value in x_1 will increase the likelihood of the example being classified as orange. Figure 2.4(B) shows an example where a mass-univariate analysis and also a linear pattern-based analysis method will fail to detect any relevant

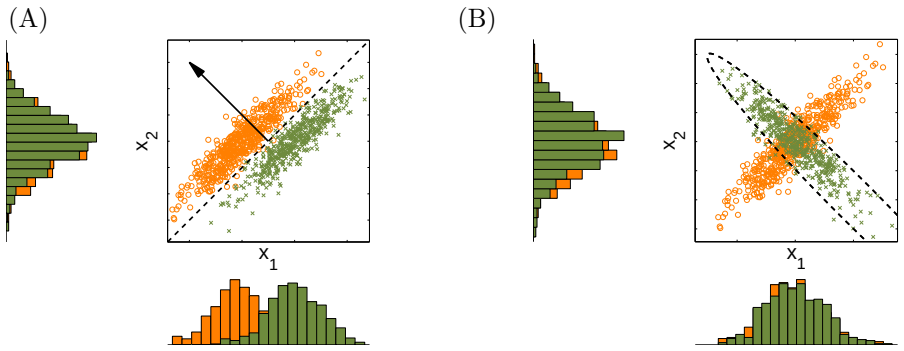

Figure 2.4: Illustration of the benefit of pattern-based analysis showing examples of experiments with two experimentally defined conditions (orange circles and green crosses) and measured signals in two voxels x_1 and x_2. **Panel A** shows an example where the conditions can be separated with a linear classifier indicated by the dashed line. The arrow illustrates the direction of the weight vector $\mathbf{w}$ of the linear classifier. An univariate analysis could also detect the signal in x_1 but may require a large number of observations due to the relatively large variation within conditions in comparison to the mean difference. **Panel B** shows an example of a situation where a nonlinear classification model is required to detect the underlying signal structure. Both univariate analysis and linear pattern-based analysis methods will fail to detect any relevant structure in the signals.

signal structure. The voxels show no stimulus selectivity in terms of mean activation differences. The example illustrates a scenario, where a change in brain covariance structure is evoked by the experimental paradigm. Signal detection is possible with a nonlinear method that allows for more complex signal structure to be modeled. For example, a quadratic discriminant analysis modeling the classes by distinct covariance matrices will be able to detect the underlying structure of the signals.

Pattern-based analysis approach

Pattern-based analysis can be formulated in terms of a *decoding* model $f : \mathcal{M} \rightarrow \mathbf{g}$. Hence, such a model considers the mesoscopic variables to be the causes and the macroscopic variables to be the consequences (Friston et al., 2008). This is in contrast to encoding models, that considers the macroscopic variables to be the causes and the mesoscopic variables to be the consequences. In decoding models we are interested in modeling the conditional distribution $p(\mathbf{g}|\mathbf{m})$. Note that within the neuroimaging community the process of learning $p(\mathbf{m}|\mathbf{g})$ is

considered a *classical* statistical analysis strategy while learning $p(\mathbf{g}|\mathbf{m})$ is often referred to as a *machine learning* approach to data analysis. Encoding models and decoding models can be related through Bayes' rule

$$p(\mathbf{g}|\mathbf{m}) = \frac{p(\mathbf{m}|\mathbf{g})\,p(\mathbf{g})}{p(\mathbf{m})}, \tag{2.2}$$

Note that $p(\mathbf{m})$ in general is extremely complex. See Kjems et al. (2002); Friston et al. (2008) for a further discussion of the relationship between encoding and decoding models.

Using the factorization in eq. (2.1) we can construct a simple decoding model by use of Bayes' rule

$$p(\mathbf{g}_k|\mathbf{m}) = \frac{\prod_{i=1}^{P} p(m_i|\mathbf{g}_k)\,p(\mathbf{g}_k)}{\sum_{k'=1}^{K}\prod_{i=1}^{P} p(m_i|\mathbf{g}_{k'})\,p(\mathbf{g}_{k'})}, \tag{2.3}$$

assuming that the macroscopic variable $\mathbf{g}_k$, $k \in \{1,\ldots,K\}$ encodes K discrete brain states. Further by assuming that the conditional distributions over the mesoscopic variables are Gaussian we obtain the Gaussian Naïve Bayes (GNB) classifier (Kjems et al., 2002).

A wide range of model types

Despite its simplicity the GNB model, introduced above, has proven good performance on a variety of pattern-based analysis tasks (Kjems et al., 2002; Mitchell et al., 2004; Chen et al., 2006; Pereira and Botvinick, 2011). Kjems et al. (2002) implemented a GNB model on top of a subspace identified with canonical variates analysis (CVA) for multi-class prediction in PET images. Mitchell et al. (2004) demonstrated that the GNB model was able e.g. to successfully predict stimulus type (visual / auditory) and in predicting the semantic category of words. Prediction was performed on both individual subject level and across multiple subjects. Chen et al. (2006) compared GNB models to the support vector machine (SVM) and logistic regression (LogReg) for a variety of combinations of data preprocessing choices. The GNB models appeared to be more sensitive (with respect to prediction accuracy) to the different preprocessing strategies in comparison to LogReg and SVM. However, the performance of GNB was at the same level as the LogReg and SVM for the preprocessing strategy that led to the highest prediction accuracies. In a recent study focusing on decoding the category of visual objects based on an event-related design, the GNB models were reported to perform significantly worse than e.g. the SVM

and LDA (Misaki et al., 2010). An important remark is that GNB models allow for computationally extremely fast analyses (Pereira and Botvinick, 2011). This issue may be important in the *searchlight* analysis procedure that builds local decoding models throughout the entire brain (Kriegeskorte et al., 2006; Pereira and Botvinick, 2011).

Another commonly used model is Fisher's discriminant analysis (FDA) (and the closely related CVA and LDA, see e.g. Izenman (2008)) that also has proven good classification performance. The use of CVA has been well established within the NPAIRS resampling framework for model evaluation, see e.g. Strother et al. (2002); Kjems et al. (2002); Strother et al. (2010). Carlson et al. (2003) used LDA in analysis of patterns characterizing categorical representation of objects (faces, chairs, houses). CVA was compared to the SVM for different preprocessing strategies in LaConte et al. (2005), where the predictive performance of CVA was reported to be more sensitive to specific preprocessing choices than the SVM. A recent study comparing a wide range range of classifiers reported LDA to provide at least the same classification accuracies as linear (and nonlinear) SVMs (Misaki et al., 2010).

The SVM is the most frequently adopted method for pattern-based analysis within the neuroimaging community. Cox and Savoy (2003) used the SVM for classifying distributed patterns of activation in the visual cortex, and Kamitani and Tong (2005) used the SVM for classification of oriented pattern based on signals in the early visual cortex. Frequently cited papers in studies using the SVM are LaConte et al. (2005) and Mourão-Miranda et al. (2005). LaConte et al. (2005) investigated the predictive performance of the SVM for ten different preprocessing choices in data from 16 subjects and proposed four different methods for extraction of activation maps from SVMs. Additionally, they reported the number of support vectors retained in the best performing model ranging from fractions $30 - 100\%$. The study of Mourão-Miranda et al. (2005) performed multi-subject classification in data from 16 subjects performing a face matching and a location matching task. In comparison to LDA, the SVM proved to have better predictive performance and to be less sensitive to whether spatial smoothing was used in the data preprocessing. It was proposed that the weight vector of a linear SVM represents a *discriminating volume*. A comparison between the discriminating volume and a statistical parametric image (SPI) from a conventional GLM analysis was performed. It was demonstrated that voxels belonging to the most discriminating regions, as identified by the SVM, tend to be closely related to the most significant voxels in the SPI. This was in contrast to the brain map derived from the LDA that provided less overlap with the SPI. The SVM was furthermore proposed as an analysis approach with clear benefits over e.g. LDA:

> *The brain state classification from fMRI data volumes corre-*

> *sponds to the classification of few points (scans) in a high-dimensional space (dimension = number of voxels). In this situation, there are many linear classifiers (i.e. hyperplanes) that separate the training data (Scholkopf and Smola, 2002), which heavily overfit and generalize badly. The SVM algorithm can solve this problem (Boser et al., 1992). It finds the optimal hyperplane, i.e. the separating hyperplane that generalizes better. This property makes the linear SVM an optimal tool to address the problem of finding a common brain network between subjects and use this information to classify data from a new subject. For all tests, the training error for the SVM (i.e. the error rate for classifying the training set) was zero, this means that the training data were linearly separable and the SVM algorithm found the optimal separating hyperplane. This reflects the fact that extensions of the SVM as nonlinear kernels or soft-margin SVM with slack variables are unnecessary here and would be counterproductive.* Mourão-Miranda et al. (2005).

The SVM has indeed proven good predictive performances in a long series of studies. Applications include classification of *IC-fingerprints* in order to automatically select relevant ICA components in fMRI data (De Martino et al., 2007; Tohka et al., 2008), prediction of speech content and speaker identity based on the brain activation pattern of a listener (Formisano et al., 2008a), prediction of object category based on patterns of brain activation evoked by visual stimuli (Cox and Savoy, 2003; Hanson and Halchenko, 2008), real-time brain state classification based on fMRI data LaConte et al. (2007), and classification of disease patterns based on structural scans (Golland et al., 2001, 2005; Klöppel et al., 2008; Koutsouleris et al., 2009).

A variety of other classifier types have successfully been applied in the pursue to extract information from patterns of brain activation, including artificial neural networks (Lautrup et al., 1994; Mørch et al., 1997; Hanson et al., 2004), logistic regression models (Chen et al., 2006; Yamashita et al., 2008; Rissman et al., 2010; Wolbers et al., 2011; Michel et al., 2011c), relevance vector machines (Lukic et al., 2007; Formisano et al., 2008b; Valente et al., 2011), kernel ridge regression (Chu et al., 2011a), and restricted Boltzmann machines (Schmah et al., 2008). The interested reader is referred to e.g. the study of Misaki et al. (2010) that compares six classification models based on an event-related experiment with visual object stimulation or Schmah et al. (2010) for a comparison of ten classification model applied to data from a longitudinal fMRI study of stroke recovery.

Model regularization

When building pattern-analysis models on neuroimaging data the issue of complexity control becomes important. Neuroimaging data sets are typically characterized by a high number of features/voxels (10K-100K), while a relatively small number of examples are available (100-1000). Hence, strong regularization in the models is often required to avoid over-fitting to the training data. To control model complexity one possibility is to select a few voxels or ROIs as input variables to the model as described in Section 2.2.4 on feature selection. Another approach to regularization is to derive an informative basis set based on the original features/voxels. A commonly used approach is to perform PCA and build the predictive model on a subspace defined in terms of a subset of the principal components (PC)s. The studies Strother et al. (2002); Kjems et al. (2002); LaConte et al. (2003); Strother et al. (2004) successfully implemented a CVA model on top of a PCA basis. The model complexity can be controlled by varying the number of retained PCs. These studies demonstrated a trade-off between prediction accuracy and model visualization 'reproducibility'. In general how model performance depends on the number of PCs retained seems to be significantly influenced by the specific preprocessing strategy used. There was a general tendency that prediction accuracy increased with the number of PCs retained (hence high model complexity) whereas model visualization reproducibility decreased when the model complexity increased. Carlson et al. (2003) successfully implemented LDA on top of a PCA basis in order to discriminate between object categories in fMRI data. Cox and Savoy (2003) compared LDA without regularization with SVM and reported poor performance of the LDA when the number of included features increased. The authors noted that this was not surprising since the estimate of the covariance matrix became increasingly singular as the number of features increased. Likewise, Mourão-Miranda et al. (2005) compared LDA without regularization to the SVM and concluded that the SVM provided the best predictive performance and proved best in identifying the most discriminating regions between brain states.

Another commonly used regularization method is the ridge regularization or ℓ_2 regularization. Kustra and Strother (2001) proposed a methodology based on penalized discriminant analysis (PDA) with a ridge penalty, i.e. regularization by adding a multiple of a diagonal matrix to the covariance matrix. Furthermore, they imposed spatial smoothness on the model structure by expanding the brain scans in terms of a tensor product B-spline basis. The model was successfully applied to a two class and an eight class classification problem in a multi-subject PET study. Ridge regularization has also been successfully applied together with logistic regression models e.g. Rissman et al. (2010); Schmah et al. (2010). A Bayesian version of logistic regression with a ridge penalty has been successfully applied in analysis of fMRi data in e.g. Yamashita et al.

(2008). Conventionally, the SVM is motivated as a margin maximizing method. However, the SVM can also be cast into a regularization framework, e.g. Hastie et al. (2009). Hence, the complexity control parameter of the SVM becomes similar to the (inverse) of the regularization parameter controlling the amount of ridge penalty. In neuroimaging contexts, it has been observed that model performance, as measured by prediction accuracy, is only degraded at low values of the SVM complexity parameter (LaConte et al., 2005; Marquand et al., 2010). These observations seem to support the use of the 'hard-margin' SVM, a special instance of the SVM with no regularization parameters that needs to be specified. The hard-margin SVM has been used in several pattern-based analyses of fMRI data, e.g. Mourão-Miranda et al. (2005); Wang et al. (2007), a 'default value' of the complexity parameter has been used in e.g. Mourão-Miranda et al. (2006); Wang (2009); Ecker et al. (2010); Marquand et al. (2010), while other studies optimize the regularization parameter in order to maximize prediction accuracy (Grosenick et al., 2008; Schmah et al., 2010; Michel et al., 2011c). The reason why selection of model regularization parameters is not a major concern of SVM users in the neuroimaging community may be explained by the fact that SVMs operated 'out of the box' show good generalization performance on present neuroimaging data sets. It has been argued by Yunqian Ma and Cherkassky (2005) that the SVM, in addition to the ridge penalty, also has an inherent regularization through the hinge loss function (margin regularization). This is an important property of the SVM. When other methods fails to identify an unique/stable solution due to ill-posed nature of the problem (e.g. inversion of a singular matrix) the SVM will be able to identify an unique solution through margin maximization - even if no regularization is imposed through the user controlled ridge penalty.

Ridge regularization provides uniqueness to the model fit and stabilize coefficient estimates. The solution will be dense, i.e. the model's coefficients will in general have values different form zero. Sparsity enforcing regularization e.g. the least absolute shrinkage and selection operator (LASSO) or the grouped LASSO methods will set coefficient corresponding to individual features (or predefined groups) to zero, see e.g. Tibshirani (1996); Meier et al. (2008) and references therein. Such regularization is also referred to as ℓ_1 regularization. However, the LASSO type regularization may not be appropriate in the analysis of neuroimaging data, since the procedure tends to select only a single of multiple correlated variables. Furthermore, LASSO selects at most N features within the number of features P exceed the number of scans available N. An attractive alternative is the elastic net (ENET) penalty, that uses a combination of ℓ_1 and ℓ_2 (Zou and Hastie, 2005). The ENET regularization was introduced as a method that does automatic variable selection and also selects groups of correlated variable to be included in the model. ENET has been successfully used in pattern-based analysis in a series of neuroimaging studies. Grosenick et al. (2008) used both the LASSO and ENET regularization with a PDA model.

They did a comparison of sparse PDA and conventional dense models; LogReg, LDA, and SVM. In prediction of purchase decisions based on fMRI data they demonstrated that both the LASSO and the ENET penalty lead to increased predictive performance. Additionally, it was argued that the sparse methods automatically selected a relevant set of model coefficients. Carroll et al. (2009) used ENET regularized regression in predicting a set of ratings based on the PBAIC fMRI experiment, where subjects were engaged in a virtual reality task. Model evaluation was performed not only by considering prediction accuracy. Additionally, the authors introduced procedures for quantifying the model's ability to provide 'interpretable' brain maps: i) a 'spatial distribution' metric estimating the spread of selected voxels throughout the brain, and ii) a 'robustness' metric measuring the overlap between selected voxels identified in models trained on independent splits of the data set. It was demonstrated that imposing increasing levels of ℓ_2 regularization, while fixing the ℓ_1 regularization, led to an increase in robustness without sacrificing prediction accuracy. Additionally, the spatial distribution metric was decreased with increasing ℓ_2 regularization. The authors suggested that correlated clusters from which variables are included were spatial proximal. An observation which they argued was consistent with neuroscientific intuition. Recently, Michel et al. (2011a) proposed an analysis method called Multiclass Sparse Bayesian Regression. This method performs a grouping of features (voxels) into several classes. The grouping of features into classes is controlled by a latent discrete variable, and features belonging to each class is then regularized differently (in contrast to the 'global' regularization in ridge regression).

A potential limitation of the regularization procedures discussed above is that they do not directly take into account the expected structure of the model coefficients. Examples of such underlying structure are signal structure defined by spatial distance measures, temporal similarity, prior knowledge on functionally or anatomically connectivity, see e.g. Thirion et al. (2006) and references therein. Friston et al. (2008) proposed to include structure into the regularization procedure by defining a prior over model coefficients within a Bayesian model comparison framework. It was suggested to impose structure on the coefficient covariance matrix. Effectively, this was implemented by defining the covariance matrix in terms of spatially smooth vectors, singular vectors, sparse vectors, or support vectors. Models build with different structure imposed could subsequently be compared within a Bayesian model comparison framework allowing for inference on the underlying signal structure. Cuingnet et al. (2010) proposed methodology for spatial regularizing the SVM. Specifically, they proposed to use Laplacian regularization in order to obtain more interpretable brain maps. By considering the notion of proximity between elements of a brain scan volume, prior knowledge is introduced by considering e.g. spatial proximity (voxels are close if they are close in space) or anatomical proximity (voxels are close if they belong to the same brain network as defined e.g. by a brain atlas or

fiber tracts). This methodology has been successfully applied in pattern-based analysis of stroke data in detection of difference between subjects with good and poor outcome based on diffusion-weighted imaging (DWI) data acquired at the acute stage (Cuingnet et al., 2011). Michel et al. (2011c) recently proposed a total variational (TV) regularization framework for pattern-based analysis in fMRI. TV is defined as the ℓ_1 norm of the image gradient and preserves edges. The use of this regularization was motivated by expectation on the underlying signal structure. Informative voxels, with respect to the macroscopic variable, were expected to be spatially distributed and that voxels selected by the model should be grouped into spatially connected clusters. In fMRI data from an experiment studying object representation the method proved successful in i) provide prediction accuracies comparable to that obtained with ENET regression, sparse multinomial logistic regression, and the SVM, ii) providing spatially coherent regions with similar weights, interpreted to be a simplified and still an informative set of selected voxels.

Linear and nonlinear models

Figure 2.4 provides examples of signal structures that can be detected by mass-univariate analysis and by linear and nonlinear pattern-based analysis. Aiming at improved modeling of effective connectivity, nonlinear modeling has been introduced within the dynamic causal modeling (DCM) framework (Stephan et al., 2008). In this context, nonlinear modeling allows for identification of models in which the connection between two brain regions is modulated by the activity in a third region. This argument extends to pattern-based analysis models, that is, nonlinear models allow for signal detection, where *changes in inter-regional interactions* are related to the experimental variable. Lautrup et al. (1994) used flexible artificial neural networks (ANNs) for classification of PET scans. A comparison of ANNs and linear models (FDA) was reported in Mørch et al. (1997). They performed pattern-based analysis in both PET and fMRI data sets, and compared the linear and nonlinear models by evaluating the prediction accuracy. By monitoring prediction accuracy as a function of available training examples (scans), they demonstrated crossing learning curves: Linear models performed best at small sample sizes, while the nonlinear models showed superior performance at larger sizes of the training set. A re-analysis of the data from the experiment of Haxby et al. (2001) has been performed by means of ANN analysis (Hanson et al., 2004). These authors used a noise perturbation approach to evaluate the contribution of individual voxels to the overall classification performance. By comparing maps obtained for each specific object category they reported considerable overlap between 'important' voxels across categories. Based on this finding they argued in favor of a combinatorial code in the VT lope. Cox and Savoy (2003) compared a linear SVM to a (non-

linear) polynomial SVM in the analysis of fMRI activations evoked by visual stimulation. They reported the linear models to be best performing as measured by prediction accuracy. Similarly, LaConte et al. (2003) compared SVMs build with linear and polynomial kernels, and reported best performance when using the linear models. Misaki et al. (2010) compared six different classifier types trained to predict based on response patterns recorded with fMRI in the early visual and inferior temporal cortex during an event related experiment. Overall the linear models were reported to perform the best as measured by prediction accuracy. In line with the observation by Mørch et al. (1997) the following explanations were suggested by Misaki et al. (2010): i) the true distribution's Bayes-optimal decision boundary were linear, ii) the data available was insufficient to build reliable nonlinear models, iii) a combination of i) and ii). In a recent study ten classification methods were compared in a longitudinal fMRI study of stroke recovery (Schmah et al., 2010). Three different two-class classification tasks were considered. The classes were heterogeneous in the sense that each class contained scans from four sub-classes. With respect to classification accuracy, the relative benefit of nonlinear methods compared to their linear counterparts varied over classifications tasks. In two classification tasks the nonlinear methods proved superior performance, while in one task there was no significant benefit from applying nonlinear methods. Recently, linear SVMs were compared to nonlinear SVMs based on the radial basis function kernel (Song et al., 2011). Model comparison was based on a fMRI data set, where the participants were subjected to visual stimulation by objects belonging to different object categories. The nonlinear models performed the best when build on relatively low number of voxels, while the linear models performed the best when a larger number of selected voxels were considered.

The use of nonlinear modeling within the neuroimaging community has been limited: i) There seems to be a limited benefit in using nonlinear models on present neuroimaging data sets, and ii) nonlinear models are considerable more difficult to interpret than basic linear models. Linear models are frequently interpreted or *visualized* by constructing brain maps showing the model's individual weights at corresponding brain locations, see e.g. McIntosh et al. (1996); Kjems et al. (2002); Mourão-Miranda et al. (2005); Hanson and Halchenko (2008). The sign of individual voxels in such a weight map reflects how a voxel's signal value should be changed in order to increase the likelihood of the scan being assigned to a particular class. Note that even a linear model becomes difficult to interpret based on the model's weights in multi-class settings (> 2 categories)(Kjems et al., 2002; Michel et al., 2011c). Kjems et al. (2002) introduced the *sensitivity analysis* as a generic technique for extraction of brain pattern maps, which can be applied to any model. The methodology builds on early work by Zurada et al. (1994, 1997). Kjems et al. (2002) extracted *sensitivity maps* from a multivariate Gaussian classifier build to discriminate brain states based on patterns of brain activation measured by PET imaging. Typically, a

single global summary map is extracted with the sensitivity map visualization procedure. Hanson et al. (2004) used a noise perturbation method in order to identify important voxels to a ANNs decisions. Specifically, Gaussian noise was added to each and every voxel individually, and it was observed whether the noise perturbation affected the classifier's decisions. Golland et al. (2005) proposed a localized interpretation approach for the (linear and nonlinear) SVM in the context of analysis of differences in anatomical shape between populations. They aimed for a representation of the differences between two classes captured by the classifier in the neighborhood of data examples. Specifically, these authors considered the decision function's sensitivity to changes in the input along different directions in the feature space. This procedure yields one visualization/brain image for individual data observation. The authors argued that generating maps corresponding to support vectors is of particular interest, since they are close to the separating boundary. Recently, Baehrens et al. (2010) proposed a general methodology for interpretation of trained classifiers by exploring *local explanation vectors* that are defined as class probability gradients. This procedure identifies features that are important for prediction at localized points in the data space. Hence, the methodology provides a means for explaining a classifier's individual decisions. LaConte et al. (2005) provides an insightful and comprehensive discussion of four visualization schemes for the SVM in the context of fMRI analysis. Specifically, a method termed feature space weighting (FSW) is proposed and analyzed. FSW comprise the following steps. First, an SVM is trained and a reduced data set is formed by removing scans corresponding to support vectors from the initial data set. Hereafter a summary map is generated by an univariate correlation analysis with the reference function (experimental/ macroscopic variable). Hence, the FSW visualization strategy focuses on data points that do not contribute to the decision function. The FSW scheme does however not provide a measure of the relative importance of voxels to the classifier.

2.2.7 Unsupervised analysis

In neuroimaging, unsupervised analysis typically concerns learning a model characterizing the mesoscopic variables $\mathbf{m}$. Hence, the macroscopic variables $\mathbf{g}$ are not directly used in the modeling. Sometimes, the unsupervised analysis strategy is introduced as 'model-free analysis', which seems to be referring to the fact that the model is not specified the same way as for example in the GLM analysis. In general, the result of the unsupervised analysis will depend on the structure imposed on the modeling procedure. PCA defines the data in terms of a new basis set composed by a series of orthogonal eigenimages (Bullmore et al., 1996; Hansen et al., 1999; Thomas et al., 2002). Another popular decomposition technique is ICA that attempts to identify spatially or temporally statistically

independent sources of variation (McKeown et al., 1998). ICA has successfully proven to allow for extraction of signal structures that are interpretable. E.g. the method may be able to separate the signals into components that are related to the experimental paradigm, respiration, heartbeat, and subjects motion (McKeown et al., 2003). ICA has in general proven to be useful in resting state experiments or in experiments where the a temporal model $\mathbf{g}$ of the data is not available (Calhoun et al., 2002).

2.2.8 Interpretation

The last part of the neuroimaging pipeline shown in Figure 2.2 is interpretation. Model interpretation is important in most scientific domains. In particular within the neuroimaging community there has been a long tradition in summarising experimental data by a statistical parametric image (SPI). Indeed spatial localization is the primary objective in many studies, that seeks to identify brain regions that 'respond' significantly to manipulations in a strictly controlled experimental variable. Within the past decade there has been an appreciation within the neuroimaging community of the usefulness of analysis methodology adopted from the research field of machine learning. E.g. the study of Kamitani and Tong (2005) exemplifies that pattern-based analysis allows for detection of signal structures that was far beyond the scale of what existing analysis methodology (mass-univariate analysis) is able to detect. It was demonstrated that reliable predictions of stimuli orientation could be formed based on the information present in the activity patterns present in the visual cortex. In many clinical applications reliable predictions are important, since the outcome of the pattern-based analysis potentially can assist a clinical diagnosis (Kippenhan et al., 1992; Klöppel et al., 2008; Ecker et al., 2010). In such applications high prediction accuracy is desirable.

Reporting prediction accuracy is often accompanied with a brain map extracted from the pattern-based model. Visualization of the model allows the investigator to interpret the model by identifying brain regions that seemingly drives the model's predictions. Hence, the classifier becomes more than a black-box saying 'yes' or 'no', since the model's visualization/interpretation could be related to existing knowledge. Or equally important, new scientific insight could be gained based on inspection of the model's visualization. Examples are the studies Haxby et al. (2001); Hanson et al. (2004) that made claims about object category representations in the VT lope based on a model's visualization.

Indeed, in many functional experiments the predictions are not relevant by themselves. Predictive performance provides a surrogate measure of the model's ability to explain the observed data, both in cases where the mapping is from

mesoscopic variables to macroscopic variables $f : \mathbf{m} \mapsto \mathcal{M}$ or the opposite $f : \mathcal{M} \mapsto \mathbf{m}$. After building a model, it is important to assess whether the model captured the statistical regularities of interest in the data. A natural procedure is to quantify performance in terms of the *generalizability* of the model (Mørch et al., 1997). Effectively, this is done by evaluating the prediction accuracy on a test set. We have more confidence in a model that correctly predicts the brain states, while it is hard to defend a model with poor generalization performance. Additionally, we are interested in the *interpretability* of the predictive model. Typically, such model interpretation is done on the basis of a brain map that reveals in which voxels the discriminative information resides (McIntosh et al., 1996; Kjems et al., 2002; LaConte et al., 2005; Mourão-Miranda et al., 2005). Hence, we could say that there is a hidden agenda in the use of classification models in analysis of neuroimaging data. That is, we are interested in how the discriminative information is encoded in the brain, rather than assignment of class labels to scans (since labels often are already known).

The final outcome of many mass-univariate analysis is a thresholded SPI. A statistical test is performed in individual voxels, and the resulting SPI is thresholded according to correction for multiple comparisons, e.g. Friston et al. (1994). Typically, such maps reveal a limited number of blobs of brain 'activation' distributed across the scan volume. Naturally, we may pursue the same characteristics in the model's visualization when using pattern-based analysis. A brain map extracted from e.g. a SVM will in general be dense. Each voxel will have a value different from zero. The sparse characteristics of the model's visualization can be achieved by i) thresholding the map according to some heuristics, see e.g. Kjems et al. (2002); LaConte et al. (2005); Mourão-Miranda et al. (2005) for examples, or ii) use resampling to obtain an empirical coefficient distribution in each voxel followed by some principled statistical thresholding procedure, see e.g. Cuingnet et al. (2011). Brain map sparsity can also be achieved inherently in the model building procedure. For example, feature elimination or other feature selection procedures automatically selects a subset of voxels available. Another approach is to build models using a sparsity enforcing regularization procedure. Such models will automatically set coefficients to be exactly zeros, yielding sparse characteristics of the resulting brain map extracted from the model (Grosenick et al., 2008; Yamashita et al., 2008; Carroll et al., 2009; Ryali et al., 2010; Michel et al., 2011c). Such methods have been introduced as providing more interpretable maps by automatically selecting 'relevant' brain regions.

In addition to the prediction accuracy metric (p) for model evaluation Strother et al. (1997) proposed a reproducibility metric (r) that measures the similarity between SPIs extracted from models trained on independent data samples. This approach has been formally established as the NPAIRS resampling framework (Strother et al., 2002). The NPAIRS framework has proven successful in

optimization of preprocessing pipelines in a series of studies and also in pattern-based analyses, see e.g. Strother et al. (2002, 2004, 2010); LaConte et al. (2003). The NPAIRS authors argued that both (p) and (r) play an important role with respect to model interpretation:

> *Simultaneously, Hansen and Strother, guided by the field of predictive learning in statistics (Hastie et al., 2001; Larsen and Hansen, 1997; Mjolsness and DeCoste, 2001), introduced the idea of using potentially unbiased cross-validation-based prediction metrics to measure data-analytic performance in functional neuroimaging (Hansen et al., 1999; Kjems et al., 2002; Kustra and Strother, 2001; Lautrup et al., 1995; Morch et al., 1997). Similar prediction metrics have recently been used by others (McKeown, 2000; Ngan et al., 2000). In addition, prediction metrics have been used to gain new insight into the debate over the spatially modular versus spatially distributed nature of human brain processing (Cox and Savoy, 2003; Haxby et al., 2001). We expect both prediction and reproducibility metrics to play an increasingly important role in the future optimization and interpretation of fMRI studies.* (Strother et al., 2004).

This viewpoint that prediction accuracy alone may be insufficient is shared by Carroll et al. (2009) who also argue in favor of building models that both predicts well and are robust:

> *Indisputably, prediction is an essential component of scientific modeling, and great effort should be put into maximizing it; however, as shown in this paper, equally predictive models can still be markedly different. In fMRI analysis, the core goal underlying predictive modeling is production of a model that can be interpreted to pinpoint all relevant voxel activity and exclude all irrelevant activity. Therefore, it is crucial to not lose sight of the interpretation of the resulting models in the quest to optimize prediction performance.*
> *... We have also shown that being preoccupied with prediction performance can be equally destructive. Models that function as highly predictive "black boxes" might be useful for neuro-engineering "mind reading" efforts, but for informing neuroscience, these models should also be reliable and valid.* Carroll et al. (2009) .

2.3 Project contribution

This section outlines the main contributions of the Ph.D. project in relation to the existing work presented in the previous sections.

2.3.1 Model sparsity and brain pattern interpretation

Interest is increasing in applying pattern-based analysis techniques to the analysis of functional neuroimaging data. Model interpretation is of great importance in the neuroimaging context, and is conventionally based on a 'brain map' extracted from the pattern-based analysis model. Prior studies have suggested that the support vector machine (SVM) may be capable in part to achieve an uncoupling between reproducibility[4] and prediction performance (Mourão-Miranda et al., 2005). Prior studies have observed quite stable predictive performance of the SVM for sufficiently high values of the SVM regularization parameter C (LaConte et al., 2005; Marquand et al., 2010). Other studies argue that regularization parameters not needs to be selected 'very precise' also relying on observation of relative stable predictive performance $\pm 3\%$ over a relatively large range of values for the regularization parameter Chu et al. (2011b).

We have studied the relative influence of model regularization parameter choices on the model generalization, the reliability of the spatial patterns extracted from the classification model, and the ability of the resulting model to identify relevant brain networks defining the underlying neural encoding of the experiment. The work was published in Rasmussen et al. (2012b).

- For a SVM, logistic regression (LogReg) and Fisher's discriminant analysis (FDA) we demonstrate that selection of model regularization parameters has a strong but consistent impact on the generalizability and both the reproducibility and interpretable sparsity of the models for both ℓ_2 and ℓ_1 regularization.

- In contrast to early studies comparing FDA (unregularized versions) and SVM we demonstrate similar performance in both prediction accuracy and brain pattern reproducibility of the SVM, LogReg, and FDA models. Our results suggest, that it may be more important carefully to tune model regularization than it is to select a specific classifier type.

- Importantly, we illustrate a trade-off between model spatial reproducibility and prediction accuracy for SVM, LogReg, and FDA models. Unlike as suggested in the literature, we observe that the SVM is not capable in uncoupling prediction accuracy and pattern reproducibility.

- We show that known parts of brain networks can be overlooked in pursuing maximization of classification accuracy alone with either ℓ_2 and/or ℓ_1 regularization. When performing feature selection using sparse models

[4] Reproducibility as defined in Strother et al. (2002).

(either via feature elimination or a sparsity enforcing prior) we find it useful also to report the prediction accuracy based on voxels excluded from the model. Such analysis could potentially enhance the interpretation of a sparse brain pattern.

- Our observations support the view that the quality of spatial patterns extracted from models cannot be assessed purely by focusing on prediction accuracy. Our results instead suggest that model regularization parameters must be carefully selected, so that the model and its visualization enhance our ability to interpret the brain.

2.3.2 Visualization of nonlinear kernel models

Kernel methods, e.g., SVMs, relevance vector machines (RVMs), or kernel ridge regression (KRR) are frequently used in pattern-based analysis. The practical use of nonlinear modeling within the neuroimaging community has been limited: i) there seems to be a limited benefit in using nonlinear models on present neuroimaging data sets (Cox and Savoy, 2003; LaConte et al., 2003; Misaki et al., 2010) but see also Stephan et al. (2008); Schmah et al. (2010); Song et al. (2011), and ii) nonlinear models are considerable more difficult to interpret than basic linear models (LaConte et al., 2005).

We have focused on visualization of nonlinear kernel models. Specifically, we investigated the sensitivity map (Zurada et al., 1994, 1997; Kjems et al., 2002) as a technique for generation of global summary maps of kernel classification models. The illustration of the performance of the sensitivity map was based on a fMRI data set from an experiment with visual stimuli. The work was published in Rasmussen et al. (2011) and Rasmussen et al. (2012b).

- We show that the performance of linear models is reduced for certain scan labelings/categorizations in the fMRI data set, while the nonlinear models provide more flexibility. Nonlinear models are capable in modeling data, where a considerable amount of heterogeneity within individual classes exists.

- We illustrate that the sensitivity map can be used to visualize nonlinear versions of LogReg, the FDA, and the SVM, and show that the sensitivity map is a versatile and computationally efficient tool for visualization of nonlinear kernel models in neuroimaging.

- Based on the original formulation of the sensitivity map visualization, we have developed further procedures for model visualization. Specifically,

we focus on the generation of maps that include sign information, unlike earlier versions of the sensitivity map. The sign information provides the investigator with a more detailed explanation of how signal changes in individual brain locations influence the classification.

- An important aspect of our analysis has been to assess the reliability/stability of the proposed model visualizations. The evaluation is performed within the NPAIRS framework (Strother et al., 2002), a data-driven split-half evaluation framework in which we build multivariate models of the data and base the evaluation on both brain state predictability and the reproducibility of brain maps extracted from multivariate models.

2.3.3 Nonlinear denoising using kernel principal component analysis

Kernel principal component analysis (KPCA) is a nonlinear generalization of PCA. The basic idea in KPCA is to map the data from voxel space to a feature space, and then perform PCA on the mapped data. Even though, the practical use of nonlinear kernel based preprocessing methods has been limited, recent years have seen an increased interest in applying KPCA as a preprocessing and analysis tool in the field of neuroimaging (Thirion and Faugeras, 2003; Song et al., 2008; López et al., 2009; Guo, 2010). The main challenge in denoising by KPCA is the mapping of denoised feature space points back into input space - also known to as the *pre-image problem*. The use of KPCA and pre-image estimation in neuroimaging was reported in Rasmussen et al. (2012a).

- We evaluate the performance of KPCA and the subsequent pre-image estimation as a tool for noise reduction in fMRI. Using the NPAIRS resampling framework (Strother et al., 2002) for this evaluation has been a key element in the analysis of our proposed procedures.
- We introduce manifold navigation for exploration of a nonlinear data manifold and illustrate how pre-image estimation can be used to generation brain maps in the continuum between experimentally defined brain states/classes. Our procedure extends the hyperplane navigation procedure proposed for linear models by Sato et al. (2008).

Chapter 3

Statistical modeling and model evaluation

This chapter describes the statistical modeling used in the project. Most of the methods are well established, and introduced here in order to assist the reader in interpreting the results presented in Chapter 5. The general linear model is introduced, since this is the most prevalent analysis approach at present. The multivariate Bayesian decoding model is also introduced, mainly since it is implemented in the widely used SPM software package. This is followed by an introduction of linear and nonlinear predictive models. Readers familiar with basic modeling techniques are encouraged to skip directly to the last sections on *global model visualization by sensitivity maps* and *denoising and localized visualization using kernel principal component analysis and pre-image estimation*.

Notation: In this chapter $\mathbf{y}$ will refer to the dependent variables in a model, while $\mathbf{X}$ will refer to independent/ explanatory variables. Note that either of these variables can contain macroscopic and mesoscopic variables as introduced in Section 2.2.3.

Contents

3.1 Univariate modeling

3.1.1 The general linear model

Using the conventional notation, we let the mesoscopic variables (e.g. voxel time series) in a single dimension (voxel) serve as dependent variables $\mathbf{y} \in \mathbb{R}^{N\times 1}$ and the macroscopic variables $\mathbf{D}$ (Section 2.2.3) serve as explanatory variables organized in the design matrix $\mathbf{X} = \mathbf{D}^\top \in \mathbb{R}^{N\times K}$. The general linear model (GLM) reads

$$\mathbf{y}_i = \mathbf{X}\mathbf{w}_i + \mathbf{e}_i, \tag{3.1}$$

where $\mathbf{w} \in \mathbb{R}^{K\times 1}$ contains model parameters/weights to be estimated, and $\mathbf{e} \in \mathbb{R}^{N\times 1}$ denotes the residual or error term of the GLM. The subscript i highlights, that the GLM models the signal at voxel-level. Also the error term is modeled at voxel-level. It is assumed that at voxel i $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$, where $\mathbf{\Sigma}_i$ models the covariance structure of the time series (noise autocorrelation). Estimation of the model weights at voxel i can be performed using generalized least squares estimation

$$\hat{\mathbf{w}}_i = \left(\mathbf{X}^\top {\mathbf{\Sigma}_i}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top {\mathbf{\Sigma}_i}^{-1}\mathbf{y}_i. \tag{3.2}$$

Note that this procedure requires the covariance structure, as modeled by $\mathbf{\Sigma}_i$ to be known. We refer to the literature for discussions on procedures for estimating this covariance structure, e.g. (Friston et al., 2007).

Statistical inference proceeds by definition of a contrast vector $\mathbf{c} \in \mathbb{R}^{K \times 1}$. For example, a t-statistics can be constructed by

$$t_i = \frac{\mathbf{c}^\top \hat{\mathbf{w}}_i}{\sqrt{\text{var}\left(\mathbf{c}^\top \hat{\mathbf{w}}_i\right)}}, \tag{3.3}$$

where the denominator is calculated as

$$\text{var}\left(\mathbf{c}^\top \hat{\mathbf{w}}_i\right) = \mathbf{c}^\top \left(\mathbf{X}^\top {\mathbf{\Sigma}_i}^{-1} \mathbf{X}\right)^{-1} \mathbf{c}, \tag{3.4}$$

see e.g. Friston et al. (2007). This test statistics will follow a Student's t distribution under the null-hypothesis. The Student's t distribution is governed by ν degrees of freedom, and the interested reader is referred to the literature for procedures to estimate ν, e.g. Friston et al. (2007). Calculating the t-statistics in eq. (3.3) for each $i \in [1, \ldots, P]$ leads to a *statistical parametric image* (SPI). By comparing the t-statistics to the null-distribution in each voxel, we can identify brain locations where we can reject the null-hypothesis at a certain level of significance.

The mass-univariate analysis procedure is well implemented in a series of widely used software packages, e.g. in AFNI (Cox, 1996), FSL (Smith et al., 2004), and SPM (Friston et al., 2007).

3.2 From univariate encoding models to multivariate decoding models

The software package SPM (Friston et al., 2007) provides a decoding model with a similar formulation as eq. (3.1). This scheme is named multivariate Bayesian (MVB) decoding. The MVB scheme defines the dependent variables $\mathbf{y}$ based on part of the design matrix containing macroscopic variables $\mathbf{D}$. The explanatory variables $\mathbf{X}$ are based on the mesoscopic variables $\mathbf{M}$, hence $\mathbf{X} \in \mathbb{R}^{N \times P}$ (Section 2.2.3). Using the above notation a decoding model can be formulated by

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \tag{3.5}$$

where $\mathbf{w} \in \mathbb{R}^{P \times 1}$ contains model weights to be estimated, and $\mathbf{e} \in \mathbb{R}^{N \times 1}$ denotes the residual or error term. As in the GLM in eq. (3.1) it is assumed that the residuals are characterized by some covariance structure $\mathbf{\Sigma}^{\mathbf{e}}$, thus $\mathbf{e} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}^{\mathbf{e}}\right)$.

Note that in contrast to the GLM that estimates model parameters at voxel-level $\mathbf{w}_i$, the MVB scheme estimates a single model weight vector $\mathbf{w}$.

Typically $P \gg N$ which means we cannot use the same strategy as in eq. (3.2) to estimate the model weights. The problem of estimating $\mathbf{w}$ is ill-posed in the sense, that there exists an infinite number of equally likely solutions. Hence, further assumptions are required. The approach in the MVB scheme is to introduce a prior over the model weights and assume $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{\mathbf{w}})$, where $\mathbf{\Sigma}^{\mathbf{w}}$ specify the covariance structure of the model weights. Estimation of model parameters in the MVB scheme proceeds by use of a variational inference scheme, and the reader is referred to Friston et al. (2008) for further details.

A key feature of the MVB scheme is model comparison within a Bayesian framework. By imposing a particular structure on the weight covariance $\mathbf{\Sigma}_i^{\mathbf{w}}$ the investigator can form a model hypothesis $\mathcal{M}_i$ and evaluate the model evidence $p(\mathbf{y}|\mathbf{X}, \mathbf{\Sigma}_i^{\mathbf{w}})$. By considering a series of covariance structures $\{\mathbf{\Sigma}_1^{\mathbf{w}}, \ldots, \mathbf{\Sigma}_M^{\mathbf{w}}\}$ we can choose among M hypotheses by choosing the covariance specification giving the highest model evidence. Examples of covariance structures available in the MVB scheme are structures that are modeled by sparse pattern representations, pattern representations defined by singular vectors of the data, spatial smooth vectors, or single observations (scans).

3.3 Decoding as predictive modeling

The GLM and MVB models above are formulated to relate the mesoscopic variables with the macroscopic variables. The interest is on performing inference on the model's mapping rather that predicting labels of new examples (Friston et al., 2008). Another (and more prevalent) approach to pattern-based analysis is to formulate the analysis in terms of a pattern recognition problem. In such settings we are interested in learning the model's parameters based on a training set in order to predict labels for 'new' examples with unknown labels. In the usual predictive modeling setup we consider a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where we let $\mathbf{x}_i \in \mathbb{R}^{P\times 1}$ be the i'th row of the matrix $\mathbf{M}$ holding the mesoscopic variables, and y_i be an associated target variable, e.g. an element of the design matrix $\mathbf{D}$ coding class membership or some behavioral variable. We then formulate a predictive (decoding) model as

$$\hat{y} = f(\mathbf{x}; \boldsymbol{\theta}), \tag{3.6}$$

where $\boldsymbol{\theta}$ are model parameters to be estimated[1]. For example, we can write a linear model as

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b, \tag{3.7}$$

with $\boldsymbol{\theta} = \{\mathbf{w}, b\}$ where $\mathbf{w} \in \mathbb{R}^{P \times 1}$ are model weights and b is the intercept. In a more general approach we can introduce a vector valued feature representation $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{F \times 1}$, where F is the dimensionality of the feature space. We then write the model eq. (3.6) as

$$\hat{y} = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) + b, \tag{3.8}$$

where $\mathbf{w} \in \mathbb{R}^{F \times 1}$ now.

If we let y be a continuous real variable the model eq. (3.6) will be a regression model. The model can also be considered as a classification model by using some coding of the classes e.g. $y \in \{-1, 1\}$ and classify an observation $\mathbf{x}$ based on the sign of $\hat{y}$.

3.3.1 Learning model parameters by model regularization

A common strategy to learn parameters of the predictive model eq. (3.6) is to use the loss and penalty formulation, e.g. Hastie et al. (2009)

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \lambda J(\mathbf{w}), \tag{3.9}$$

where the loss function $L(y_i, \hat{y}_i)$ measures the mismatch between the true target and the prediction, while $J(\cdot)$ is a penalty function on the model weights. $\lambda \geq 0$ is a regularization parameter controlling the balance between the loss-term and the penalty. There exist a wide range of penalty functions, for instance, the LASSO penalty $J(\mathbf{w}) = ||\mathbf{w}||_1$ (ℓ_1 penalty) (Tibshirani, 1996), the ridge penalty $J(\mathbf{w}) = ||\mathbf{w}||_2^2$ (ℓ_2 penalty) (Hoerl and Kennard, 1970), or the ENET penalty $J(\mathbf{w}) = \lambda_1||\mathbf{w}||_1 + \lambda_2||\mathbf{w}||_2^2$ where λ_i are regularization parameters controlling the balance between the LASSO and the ridge term (Zou and Hastie, 2005; Hastie et al., 2009).

[1] Note that the predictions depend on the variable $\mathbf{x}$, i.e. $\hat{y}(\mathbf{x})$. We use the notation $\hat{y}$ in order to keep the notation uncluttered.

3.4 Linear predictive models

The following reviews the basics behind a series of standard models all formulated with ℓ_2 regularization.

3.4.1 Ridge regression

Ridge regression is based on the quadratic loss function $L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ and the ℓ_2 penalty (Hoerl and Kennard, 1970). We can write the objective as

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2}(y_i - \hat{y}_i)^2 + \frac{\lambda}{2}||\mathbf{w}||^2. \tag{3.10}$$

Using matrix notation for the predictors $\mathbf{X} \in \mathbb{R}^{N \times P}$ (mesoscopic variables organized in rows) and targets $\mathbf{y} \in \mathbb{R}^{N \times 1}$ we can write the solution to eq. (3.10) as

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{3.11}$$

where $\Lambda \in \mathbb{R}^{P \times P}$ is a diagonal matrix with λ in the diagonal[2]. Note that we here assume that the data has been centered, and estimate the intercept b in eq. (3.7) by the mean of the targets $\mathbf{y}$. Alternatively we could augment $\mathbf{X}$ by a column with constant elements in order to model the intercept. Correspondingly, we will then augment $\mathbf{\Lambda}$ with a diagonal element set to 0 in order not to apply the penalty to the model intercept. As can been seen in eq. (3.11) we recover the ordinary least squares (OLS) estimator if $\lambda = 0$.

A further insight into the nature of the ridge penalty can be achieved by the following analysis (Hastie et al., 2009). Consider the singular value decomposition (SVD) of the predictor variables $\mathbf{X} = \mathbf{USV}^\top$ with $\mathbf{U} \in \mathbb{R}^{N \times P}$, $\mathbf{S} \in \mathbb{R}^{P \times P}$, and $\mathbf{V} \in \mathbb{R}^{P \times P}$. We rewrite the fitted response in eq. (3.7) in terms of the SVD

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\hat{\mathbf{w}} \\
&= \mathbf{X}\left(\mathbf{X}^\top \mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \mathbf{USV}^\top \left(\mathbf{VSU}^\top \mathbf{USV}^\top + \mathbf{\Lambda}\right)^{-1} \mathbf{VSU}^\top \mathbf{y} \\
&= \sum_{i=1}^{P} \mathbf{u}_i \frac{s_i^2}{s_i^2 + \lambda} \mathbf{u}_i^\top \mathbf{y}.
\end{aligned} \tag{3.12}$$

[2]Note that we in the text use matrices of size $P \times P$. However, effectively we operate on matrices of size $N \times N$ when $P \gg N$. See e.g. Lautrup et al. (1994); Mørch et al. (1997); Hastie and Tibshirani (2004).

The coordinates of $\mathbf{y}$ with respect to the orthogonal basis $\mathbf{U}$ are shrunk by factors $\frac{s_i^2}{s_i^2+\lambda}$, and the shrinkage is the largest for the coordinates of the basis vectors $\mathbf{u}$ with the smallest singular values. Note that the ridge regression can be compared to principal component regression (PCR) (Hastie et al., 2009). PCR truncates the basis representation and only retains the K basis vectors with the largest singular values. In PCR we have similar to eq. (3.12)

$$\hat{\mathbf{y}} = \sum_{i=1}^{K} \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y}. \tag{3.13}$$

In the linear model eq. (3.7) we generally may consider to have $(P+1)$ degrees of freedom. However, we may consider the regularization as a constraint on the model structure. Hence, by varying λ we effectively control the model flexibility or complexity. A measure of such model complexity is the *effective degrees of freedom* (Hastie et al., 2009) as defined by

$$\begin{aligned} \text{edf}(\lambda) &= \text{tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \mathbf{\Lambda}\right)^{-1}\mathbf{X}^\top\right) && (3.14)\\ &= \sum_{i=1}^{P} \frac{s_i^2}{s_i^2+\lambda}. && (3.15) \end{aligned}$$

It is seen that $\text{edf}\,(\lambda) \to (P)$ if $\lambda \to 0$ and $\text{df}\,(\lambda) \to 0$ if $\lambda \to \infty$. Furthermore, we have one additional degree of freedom if the model intercept b is included in the model.

3.4.2 Logistic regression

Consider a binary target variable with coding $y \in \{-1, 1\}$. Logistic regression is based on the assumption that the log likelihood ratio is linear in $\mathbf{x}$ so that the conditional probability for $y = 1$ can be written in terms of the sigmoid function $\sigma(\cdot)$

$$\begin{aligned} p\,(y=1|\mathbf{x},\mathbf{w},b) &= \frac{1}{1+\exp\,(-\mathbf{w}^\top\mathbf{x}-b)}\\ &= \sigma\left(\mathbf{w}^\top\mathbf{x}+b\right). && (3.16) \end{aligned}$$

Likewise we write $p\,(y=-1|\mathbf{x},\mathbf{w},b) = 1 - p\,(y=1|\mathbf{x},\mathbf{w},b) = \sigma\left(-\mathbf{w}^\top\mathbf{x}-b\right)$ so

$$p\,(y|\mathbf{x},\mathbf{w},b) = \sigma\left(y\left(\mathbf{w}^\top\mathbf{x}+b\right)\right). \tag{3.17}$$

Using the negative log likelihood function we can write the logistic loss function as $L\,(y,\hat{y}) = \log\,(1+\exp\,(-y\hat{y}))$, which leads to the loss and penalty formulation

eq. (3.9)

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{N} \log\left(1+\exp\left(-y\hat{y}\right)\right) + \frac{\lambda}{2}||\mathbf{w}||^2. \tag{3.18}$$

Unlike the minimization problem in ridge regression eq. (3.11) there exists no closed form solution to eq. (3.18) due to the nonlinearity of the sigmoid function. A common strategy is to use the Newton-Raphson iterative optimization scheme that leads to iteratively re-weighted least squares (IRLS) estimation

$$\hat{\mathbf{w}}^{\text{new}} = \hat{\mathbf{w}}^{\text{old}} - \mathbf{H}^{-1}\mathbf{g}, \tag{3.19}$$

where the gradient of the cost function with respect to the weights is

$$\begin{aligned} \mathbf{g} &= \sum_{i=1}^{N} \left(\sigma\left(y_n\hat{y}_n\right)-1\right) y_n\mathbf{x}_n + \lambda\mathbf{w} \\ &= \mathbf{X}^\top\mathbf{a} + \lambda\mathbf{w}, \end{aligned} \tag{3.20}$$

where $\mathbf{a} \in \mathbb{R}^{N\times 1}$ holds the elements $a_i = \left(\sigma\left(y_n\hat{y}_n\right)-1\right)y_n$. The Hessian matrix is given by

$$\begin{aligned} \mathbf{H} &= \sum_{i=1}^{N} \sigma\left(y_n\hat{y}_n\right)\left(\sigma\left(y_n\hat{y}_n\right)-1\right)\mathbf{x}_n\mathbf{x}_n^\top + \mathbf{\Lambda} \\ &= \mathbf{X}^\top\mathbf{W}\mathbf{X} + \mathbf{\Lambda} \end{aligned} \tag{3.21}$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with elements $w_{i,i} = \sigma\left(y_n\hat{y}_n\right)\left(\sigma\left(y_n\hat{y}_n\right)-1\right)$ along the diagonal, and $\mathbf{\Lambda}$ being a diagonal matrix holding λ in the diagonal. Hence, the update formula becomes

$$\hat{\mathbf{w}}^{\text{new}} = \left(\mathbf{X}^\top\mathbf{W}\mathbf{X} + \mathbf{\Lambda}\right)^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{z}, \tag{3.22}$$

where $\mathbf{z} \in \mathbb{R}^{N\times 1}$ holds the elements $z_i = \mathbf{x}_i^\top\hat{\mathbf{w}}^{\text{old}} - \frac{a_i}{w_{i,i}}$. For a comparison of different different numerical optimizers for logistic regression see e.g. Minka (2003). Logistic regression can naturally be generalized to multi-class scenarios (Bishop, 2006; Hastie et al., 2009).

As with the ridge regression model we can characterize the complexity of the regularized logistic regression model by estimating the effective degrees of freedom, see e.g. Park and Hastie (2008) and references therein. This can be approximated by

$$\text{edf}(\lambda) = \text{tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{W}\mathbf{X} + \mathbf{\Lambda}\right)^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{X}\right), \tag{3.23}$$

using $\mathbf{W}$ as obtained in the final step of the IRLS algorithm.

3.4.3 Support vector machines

Support vector machines (SVMs) are often introduced as margin maximizing classification models (Boser et al., 1992; Cortes and Vapnik, 1995). If we consider two classes that are linearly separable, we can define a separating hyperplane so that the training observations are perfectly separated. Indeed, we can define infinitely many of such hyperplanes. The SVM identifies the hyperplane that maximizes the margin. The margin is defined as the shortest distance from the hyperplane to the training data. In case of overlapping class distributions in the training data the SVM can be extended to allow training points being inside the margin or even on the wrong side of the hyperplane, i.e. training points will be misclassified.

For SVMs the optimization objective can also be written as based on the *hinge loss* function $L(y, \hat{y}) = [1 - y\hat{y}]_+$, which leads to the loss and penalty formulation eq. (3.9) as

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{N} [1 - y_n \hat{y}_n]_+ + \frac{\lambda}{2}||\mathbf{w}||^2, \tag{3.24}$$

see e.g. Hastie et al. (2004); Chapelle (2007). The task of learning the parameters of the SVM is often considered as a constrained quadratic programming problem. The interested reader is referred to the literature for details on model optimization, e.g. Chapelle (2007); Chang and Lin (2011). Specifically, the model estimation leads to a number of model coefficients $\gamma_i \geq 0$ where $i \in \{1, \ldots, N\}$, from which the models weight vector is estimated by

$$\hat{\mathbf{w}} = \sum_{i=1}^{N} \gamma_i y_i \mathbf{x}_i. \tag{3.25}$$

Importantly, a subset of the data points may have $\gamma_i = 0$. Specifically, only training points that are located on the margin, inside the margin, or on the wrong side of the separating hyperplane (decision boundary) will have $\gamma_i > 0$. Such points constitute the *support vectors*. This is in contrast to the solutions for ridge regression eq. (3.11) and logistic regression eq. (3.22), where all training points in general will have a non-zero contribution to $\hat{\mathbf{w}}$.

In scenarios where $P \gg N$ the training data can always be separated by a linear decision boundary. Hence, it may be appealing to search for the solution that maximizes the margin. Hastie et al. (2004) argued that selection of the regularization parameter λ can be critical in such scenarios. These authors provided a $P \gg N$ simulation showing that more regularized models can be closer to the Bayes optimum solution than the margin maximizing solution.

3.4.4 Fisher's discriminant analysis

Consider multi-class problem with C classes $y \in \{1, \ldots, C\}$, $C \geq 2$. Fisher's discriminant analysis (FDA) seeks to find optimal projection directions along which the ratio of the between-class scatter to the total scatter is maximized. In the multi-class classification problem the Fisher's discriminant is given by the matrix $\mathbf{W}$, a $C-1$ column matrix, that optimizes the objective

$$\underset{\mathbf{W}}{\operatorname{argmax}} \frac{|\mathbf{W}^\top \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^\top (\mathbf{S}_T + \boldsymbol{\Lambda}) \mathbf{W}|} \tag{3.26}$$

where $\mathbf{S}_B = \sum_{c=1}^{C} N_c (\mathbf{m}_c - \mathbf{m}) (\mathbf{m}_c - \mathbf{m})^\top$ is the between-class scatter matrix, and $\mathbf{S}_T = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^\top$ is the total scatter matrix, with N_c denoting the number of samples in class c and $\mathbf{m}_c$ and $\mathbf{m}$ class means and grand mean respectively. Note that we here consider regularized Fisher's discriminant, where $\boldsymbol{\Lambda} \in \mathbb{R}^{P \times P}$ is a diagonal matrix holding a regularization parameter λ in the diagonal (Zhang et al., 2010). FDA is often formulated as the solution to the following generalized eigenvalue problem

$$\mathbf{S}_B \mathbf{W} = (\mathbf{S}_T + \boldsymbol{\Lambda}) \mathbf{W} \boldsymbol{\Xi}, \tag{3.27}$$

where the eigenpairs $\{\mathbf{w}_i, \xi_i\}$ are hold in the columns of $\mathbf{W}$ and in the diagonal of $\boldsymbol{\Xi}$. Since the rank of $\mathbf{S}_B$ is $C-1$ at most, the dimensionality of the subspace as identified by FDA will correspondingly be $C-1$ at most (Bie et al., 2005; Zhang et al., 2010).

Regularized FDA can be shown to be closely related to ridge regression (Zhang et al., 2010). By use of specific coding schemes for the target variables y, the solutions are related such a way that the subspace (or weight vector) as defined by ridge regression (eq. (3.10)) onto target variables y and the solution to eq. (3.27) are related by an orthogonal transformation and a scaling. Hence, in classification settings we can directly interpret FDA in terms of the loss and penalty formulation eq. (3.9) based on the squared loss function and an ℓ_2 regularization term.

Estimating weights according to eq. (3.26) and (3.27) provides the model's weight vectors/ canonical variates (or FDA subspace) in the decision rule eq. (3.7) and (3.8). The projection of data observations onto the subspace is performed by

$$\mathbf{z}_\mathbf{x} = \mathbf{W}^\top \mathbf{x}. \tag{3.28}$$

However, also the bias coefficient b in the decision rule needs to be determined. One approach is to build a Bayes classifier on top of the basis set as identified

by FDA, see e.g. Kjems et al. (2002). The Bayes classifier is written in terms of Bayes' rule

$$P\left(c_j|\mathbf{z}_\mathbf{x}\right) = \frac{p\left(\mathbf{z}_\mathbf{x}|c_j\right)P\left(c_j\right)}{\sum_{j'=1}^{C}\left(\mathbf{z}_\mathbf{x}|c'_j\right)P\left(c'_j\right)}, \tag{3.29}$$

where $P\left(c_j|\mathbf{z}_\mathbf{x}\right)$ is the posterior probability for class c_j, $p\left(\mathbf{z}_\mathbf{x}|c_j\right)$ is the class conditional density, and $P\left(c_j\right)$ is the class prior probability. Under the assumption that the class conditional densities are well modeled by Gaussian densities we have

$$p\left(\mathbf{z}_\mathbf{x}|\boldsymbol{\mu}_c, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{C-1}{2}} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c||^2\right), \tag{3.30}$$

with $\boldsymbol{\mu}_c$ denoting the mean of the projections of members in class c and the variance σ^2 is assumed to be shared across classes. In cases where the class priors are equal the classifier as implemented by eq. (3.29) will be equal to a nearest mean classifier (assuming eq. (3.30)). Decisions are then based on

$$\underset{c\in\{1,\ldots,C\}}{\operatorname{argmax}}\ f_c\left(\mathbf{z}_\mathbf{x}\right) \tag{3.31}$$

with the classifier's c'th output channel given by

$$f_c\left(\mathbf{z}_\mathbf{x}\right) = -||\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c||^2. \tag{3.32}$$

3.5 Nonlinear predictive models - Kernel models

In Section (3.4) we outlined the modeling setup for a series of linear models. However, linear models have by definition limitations when faced with nonlinear problems. A solution is to consider new representations $\boldsymbol{\phi}\left(\mathbf{x}\right) \in \mathbb{R}^{F\times 1}$ based on the explanatory variables $\mathbf{x}$ as in eq. (3.8). For example, we could let interaction terms $x_i x_j$ be parts of the feature representation. Over the past decade the kernel methodology has proven quite useful in extending linear models to nonlinear models. In the following we will review some basics concepts of kernel based learning. The classical reference for theory of reproducing kernels and reproducing kernel Hilbert spaces is Aronszajn (1950). A general introduction to kernel based learning is found in Shawe-Taylor and Cristianini (2004). Interested readers are referred to these references for further/ more complete introductions to kernel based learning.

3.5.1 Basic kernel methodology

A central form in kernel based learning is the positive definite kernel function $k(\mathbf{x}, \mathbf{y})$. Similar to the eigenvalue decomposition in linear algebra we can define eigenfunctions $\phi(\cdot)$ and associated eigenvalues λ that fulfills

$$\int k(\mathbf{x}, \mathbf{y}) \phi(\mathbf{y}) \, \mathrm{d}\mathbf{y} = \lambda \phi(\mathbf{y}), \tag{3.33}$$

for all $\mathbf{x}$. It follows from Mercer's theorem that the positive definite kernel can be written in terms of M eigenpairs by

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{M} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y}) \quad M \leq \infty. \tag{3.34}$$

Associated with a kernel function there exists an unique reproducing kernel Hilbert space (RKHS) $\mathcal{F}$. The sequence of eigenfunctions $\{\phi_i(\cdot)\}_{i=1}^{M}$ creates an orthonormal basis in $\mathcal{F}$, so that any function $f \in \mathcal{F}$ can be written as $f(\mathbf{x}) = \sum_{i=1}^{M} a_i \phi_i(\mathbf{x})$, with $a_i \in \mathbb{R}$. The function $k(\mathbf{x}, \mathbf{y})$ is called the reproducing kernel for $\mathcal{F}$ and has the properties $f(\mathbf{y}) = \langle f(\cdot), k(\cdot, \mathbf{y}) \rangle_{\mathcal{F}}$ and $k(\mathbf{x}, \mathbf{y}) = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{y}) \rangle_{\mathcal{F}}$, where $\langle \cdot, \cdot \rangle$ denote the scalar product (Shawe-Taylor and Cristianini, 2004).

Rewriting eq. (3.34) as

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{M} \sqrt{\lambda_i} \phi_i(\mathbf{x}) \sqrt{\lambda_i} \phi_i(\mathbf{y}) \\ &= \boldsymbol{\phi}(\mathbf{x})^{\top} \boldsymbol{\phi}(\mathbf{y}). \end{aligned} \tag{3.35}$$

Hence, evaluating the kernel function corresponds to the dot product in some feature space $\mathcal{F}$, where the data is mapped by

$$\boldsymbol{\phi} : \mathbf{x} \to \left(\sqrt{\lambda_1} \phi_1(\mathbf{x}), \ldots, \sqrt{\lambda_M} \phi_M(\mathbf{x}) \right). \tag{3.36}$$

Kernel based learning algorithms relies on embedding data observations onto the feature space $\mathcal{F}$ and attempt to identify linear relations in the feature space. These methods calculate inner products in feature space by use of the kernel function rather than points' actual embeddings. Hence, modeling is often conducted based on the kernel representation rather than performing the mapping into $\mathcal{F}$ explicitly.

3.5.2 Learning kernel models by regularization

In section (3.3.1) the task of learning model parameters of predictive models was formulated as a loss and penalty objective. Similarly, many kernel based learning models can be based on the objective

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{N} L\left(y_i, f\left(\mathbf{x}_i\right)\right) + \frac{\lambda}{2}||f||_{\mathcal{F}}^2, \tag{3.37}$$

see e.g. Hastie et al. (2009). Now, according to the *representer theorem* we can write the solution to eq. (3.37) on the form

$$f\left(\mathbf{x}\right) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \tag{3.38}$$

with $a_i \in \mathbb{R}$, see e.g. Shawe-Taylor and Cristianini (2004) and references therein. According to the reproducing properties of the kernel, the squared norm can be written as $||f||_{\mathcal{F}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$. It then follows that eq. (3.37) can be written as

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i=1}^{N} L\left(y_i, \boldsymbol{\alpha}^\top \mathbf{k}_i\right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}. \tag{3.39}$$

where $\mathbf{k}_i$ is th the i'th column of the kernel matrix $\mathbf{K}$ with elements $\mathbf{K}_{i,j} = k\left(\mathbf{x}_i, \mathbf{x}_j\right)$. Hence. the problem of learning model parameters is reduced to a finite dimensional minimization problem. Model building proceeds by selecting a loss function $L\left(\cdot\right)$ and kernel function k. This procedure leads to a range of kernelized methods, e.g. kernel ridge regression (KRR) (Saunders et al., 1998), kernel logistic regression (KLR)(Cawley and Talbot, 2004), kernel Fisher's discriminant analysis (KFDA) (Mika et al., 1999a; Zhang et al., 2010), and kernel principal component analysis (KPCA) (Schölkopf et al., 1998; Mika et al., 1999b).

3.5.3 The kernel trick

In the previous section we considered kernel based learning from a function regularization approach. Alternatively we can be more explicit about the feature space mapping (Shawe-Taylor and Cristianini, 2004). Consider a nonlinear transformation of the input variables $\{\mathbf{x}_i\}_{i=1}^N$ into a feature space $\mathcal{F}$ by $\boldsymbol{\phi} : \mathbf{x}_i \rightarrow \boldsymbol{\phi}\left(\mathbf{x}_i\right) \in \mathcal{F}$. For example, assume that we are interested in building a ridge regression model in $\mathcal{F}$. Now, even in case of a high (possible infinite) dimensional feature space, we only have N observation available. Thus we restrict

the solution $\hat{\mathbf{w}}$ to be in the span of the mapped training points $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_{i=1}^N$, i.e. $\hat{\mathbf{w}} = \sum_{i=1}^N \alpha_i \boldsymbol{\phi}(\mathbf{x}_i)$ and seek the solution to

$$\begin{aligned} & \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2} \left(y_i - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i)\right)^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \\ = & \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2} \left(y_i - \sum_{i=1}^{N} \left(\alpha_i \boldsymbol{\phi}(\mathbf{x}_i)\right)^\top \boldsymbol{\phi}(\mathbf{x}_i)\right)^2 \\ + & \frac{\lambda}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_j) \\ = & \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{1}{2} \left(y_i - \boldsymbol{\alpha}^\top \mathbf{k}_i\right)^2 + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}. \end{aligned} \tag{3.40}$$

The procedure above is an example of use of the *kernel trick*. In this procedure one i) formulates the objective in terms of input points' feature mappings, ii) identifies where dot products in feature space appears, and ii) replace the dot products by kernel evaluations.

For binary versions of KRR, KLR, and the SVM the decision function eq. (3.8) becomes

$$f(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{k}_\mathbf{x} + b, \tag{3.41}$$

with the i'th element of $\mathbf{k}_\mathbf{x} \in \mathbb{R}^{N \times 1}$ corresponding to $k(\mathbf{x}_i, \mathbf{x})$. For the multi-class analysis with KFDA (eq. (3.26)) the projections of the data observation $\mathbf{x}$ onto the $C-1$ basis vectors (in feature space) can be written as

$$\mathbf{z}_\mathbf{x} = \mathbf{B}^\top \left(\mathbf{k}_\mathbf{x} - \frac{1}{N} \mathbf{K} \mathbf{1}_N\right), \tag{3.42}$$

with $\mathbf{B} \in \mathbb{R}^{N \times C-1}$ being a projection matrix (see Zhang et al. (2010) eq. 22 for a definition of $\mathbf{B}$). By constructing a Bayes classifier on top of these feature space projection we obtain a multi-class classifier as in eq. (3.29)

$$P(c_j | \mathbf{z}_\mathbf{x}) = \frac{p(\mathbf{z}_\mathbf{x} | c_j) P(c_j)}{\sum_{j'=1}^{C} \left(\mathbf{z}_\mathbf{x} | c'_j\right) P\left(c'_j\right)}. \tag{3.43}$$

Again we may assume that the class conditional densities of data points feature space projections are well modeled by Gaussian densities

$$p\left(\mathbf{z}_\mathbf{x} | \boldsymbol{\mu}_c, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{C-1}{2}} \exp\left(-\frac{1}{2\sigma^2} ||\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c||^2\right), \tag{3.44}$$

with $\boldsymbol{\mu}_c$ denoting the mean of the projection of members in class c and variance σ^2 is assumed to be shared across classes.

3.6 Global model visualization by sensitivity maps

In many applications it is relevant to interpret a trained predictive model. Often such interpretation is based on visualizing the model's weights w_i in the linear case eq. (3.7). Likewise, in the nonlinear model eq. (3.8) we could also inspect the model's weights. However, when using the kernel methodology we do not directly have access to the model's weight vector $\mathbf{w}$. Instead the model is parametrized by means of the $\boldsymbol{\alpha}$ coefficients in eq. (3.41) and $\mathbf{B}$ in eq. (3.42). In the following section we will develop a *sensitivity mapping* procedure that allows for extraction of visualizations from both linear and nonlinear kernel models. The visualization procedure is based on early work by Zurada et al. (1994, 1997) and more recent work by Kjems et al. (2002).

3.6.1 General sensitivity map definitions

Definition 3.1 Consider a given (vector valued) function $\mathbf{g}(\mathbf{x}) \in \mathbb{R}^{O\times 1}$ in a stochastic environment with a distribution over the inputs $\mathbf{x} \in \mathbb{R}^{P\times 1}$ given by the probability density function $p(\mathbf{x})$. Corresponding to the c'th element in $\mathbf{g}(\mathbf{x})$ we define a model visualization $\mathbf{s}_k^c$ by the expected value of the derivative of the function $g_c(\mathbf{x})$ with respect to its arguments

$$\mathbf{s}_k^c = \int_{\mathbf{x}\in\mathcal{I}} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x})\right]^k p(\mathbf{x})\, d\mathbf{x}, \quad k \in \{1,2\} \tag{3.45}$$

where $\mathcal{I}$ denote some region of integration and $\mathbf{s}_k^c \in \mathbb{R}^{P\times 1}$. For $k = 2$ the visualization $\mathbf{s}_2^c$ is referred to as a *sensitivity map*, and for $k = 1$ the visualization $\mathbf{s}_1^c$ is referred to as a *signed sensitivity map*.

In the following we consider Fisher's discriminant analysis (FDA) with a Bayes classifier build on top of the FDA basis as in eq. (3.29). Different choices for the visualization function $g_c(\mathbf{x})$ in eq. (3.45) exist. Among the possibilities are to use a classifier's output $g_c(\mathbf{x}) = f_c(\mathbf{z}_\mathbf{x})$ (Yourganov et al., 2010; Rasmussen et al., 2011), the probability function $g_c(\mathbf{x}) = P(c|\mathbf{z}_\mathbf{x})$ (Baehrens et al., 2010), or the logarithm of the probability function $g_c(\mathbf{x}) = \log[P(c|\mathbf{z}_\mathbf{x})]$ (Kjems et al., 2002).

3.6.2 Gradients

Linear models

In the following we calculate the gradient in eq. (3.45) for different choices of the visualization function $g_c(\mathbf{x})$. Note that we here neglect any global scaling factors to maintain notational simplicity[3].

I: $g_c(\mathbf{x}) = f_c(\mathbf{z_x})$

By use of the classifiers output for the c'th class eq. (3.32) we immediately calculate the gradient as

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = -\mathbf{W}(\mathbf{z_x} - \boldsymbol{\mu}_c). \tag{3.46}$$

II: $g_c(\mathbf{x}) = P(c|\mathbf{z_x})$

By use of the chain rule and the quotient rule we calculate the gradient as

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = -\mathbf{W}\left[(\mathbf{z_x} - \boldsymbol{\mu}_c) - \sum_{c'=1}^{C}(\mathbf{z_x} - \boldsymbol{\mu}_{c'})P(c'|\mathbf{z_x})\right]P(c|\mathbf{z_x}). \tag{3.47}$$

III: $g_c(\mathbf{x}) = \log[P(c|\mathbf{z_x})]$

By use of the chain rule and the quotient rule we calculate the gradient as

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = -\mathbf{W}\left[(\mathbf{z_x} - \boldsymbol{\mu}_c) - \sum_{c'=1}^{C}(\mathbf{z_x} - \boldsymbol{\mu}_{c'})P(c'|\mathbf{z_x})\right]. \tag{3.48}$$

Kernel models

We now calculate the gradient in eq. (3.45) for different choices of the visualization function $g_c(\mathbf{x})$ for kernel models. Here we let $\mathbf{z_x}$ denote the projection of the feature vector $\boldsymbol{\phi}(\mathbf{x})$ onto the FDA basis $\mathbf{W}$ as in eq. (3.42).

[3] By global scaling factors we here refer to factors that are constant over the input space $\mathcal{X}$ as well as across the elements x_i.

I: $g_c(\mathbf{x}) = f_c(\mathbf{z}_\mathbf{x})$

Since the classifiers output for the c'th class reads $f_c = -(\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c)^\top (\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c)$ we get by the chain rule

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = -\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}} \mathbf{B} (\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c), \tag{3.49}$$

with $\mathbf{B}$ being defined as in eq. (3.42).

II: $g_c(\mathbf{x}) = P(c|\mathbf{z}_\mathbf{x})$

By use of the chain rule and the quotient rule we calculate the gradient as

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = -\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}} \mathbf{B} \left[(\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c) - \sum_{c'=1}^{C} (\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_{c'}) P(c'|\mathbf{z}_\mathbf{x}) \right] P(c|\mathbf{z}_\mathbf{x}). \tag{3.50}$$

III: $g_c(\mathbf{x}) = \log[P(c|\mathbf{z}_\mathbf{x})]$

By use of the chain rule and the quotient rule we calculate the gradient as

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = -\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}} \mathbf{B} \left[(\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c) - \sum_{c'=1}^{C} (\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_{c'}) P(c'|\mathbf{z}_\mathbf{x}) \right]. \tag{3.51}$$

In eq. (3.49 - 3.51) we see, that the derivative of the kernel function is required in calculating the gradients. Note that the i'th element of $\mathbf{k}_\mathbf{x}$ is $k(\mathbf{x}_i, \mathbf{x})$ for $i \in \{1, \ldots, N\}$.

In the following we provide gradients for some 'popular' kernels. For the linear kernel $k(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i^\top \mathbf{x}$ we have

$$\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}} = \mathbf{X}^\top, \tag{3.52}$$

with individual observations $\mathbf{x}_n$ organized in the rows of $\mathbf{X}$. The gradient for the polynomial kernel $k(\mathbf{x}_i, \mathbf{x}) = \left(\mathbf{x}_i^\top \mathbf{x} + q\right)^2$ we have

$$\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{X}^\top \boldsymbol{\Gamma}, \tag{3.53}$$

where the matrix $\mathbf{\Gamma} \in \mathbb{R}^{N \times N}$ is a diagonal matrix holding elements $\mathbf{x}_i^\top \mathbf{x} + q$ in the diagonal.

Finally, the gradient of the Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\frac{1}{q}||\mathbf{x}_i - \mathbf{x}||^2\right)$ is

$$\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}} = \frac{2}{q}\mathbf{G}\mathbf{\Gamma}, \tag{3.54}$$

where the matrix $\mathbf{G} \in \mathbb{R}^{P \times N}$ holds the elements $G_{p,i} = x_p^i - x_d$ with x_p^i referring to the p'th element in training example $\mathbf{x}_i$. $\mathbf{\Gamma} \in \mathbb{R}^{N \times N}$ is a diagonal matrix holding elements $\mathbf{k}_\mathbf{x}$ in the diagonal.

Local support in sensitivity map estimation

In the following we consider the gradients based on kernel models eq. (3.49 - 3.51), and note that the intuition developed in the following also holds for the linear models eq. (3.46 - 3.48). In general we can consider the gradients for all three visualization functions $g_c(\mathbf{x})$ as based on the following decomposition.

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = \underbrace{-\frac{\partial \mathbf{k}_\mathbf{x}}{\partial \mathbf{x}}\mathbf{B}}_{\text{A}} \left[\underbrace{(\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_c)}_{\text{B}} - \underbrace{\sum_{c'=1}^{C} (\mathbf{z}_\mathbf{x} - \boldsymbol{\mu}_{c'}) P(c'|\mathbf{z}_\mathbf{x})}_{\text{C}} \right] \underbrace{P(c|\mathbf{z}_\mathbf{x})}_{\text{D}}. \tag{3.55}$$

We can interpret A as a factor common to all visualization functions and B, C, and D as weighting factors specific to each visualization function.

First consider the gradient as in eq. (3.49). We can write this gradient by $\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = \text{AB}$. When moving away from the class center $\boldsymbol{\mu}_c$ the factor B increases in magnitude. For a linear kernel this means that the magnitude of the gradient will increase (note that for the linear kernel A will be constant over the entire input space). For the e.g. Gaussian kernel the interaction between A and B is more complex, since also A varies over the input space

The gradient in eq. (3.51) can be written as $\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = \text{A(B-C)}$. Consider a scenario where the classifier is confident that an input point $\mathbf{x}$ belongs to class c, hence $P(c|\mathbf{z}_\mathbf{x}) \approx 1$. By this the factor (B-C) becomes small ((B-C) $\rightarrow 0$) and the magnitude of the gradient will correspondingly become small. If the classifier is confident that $P(c'|\mathbf{z}_\mathbf{x}) \approx 1$, $c' \neq c$, we can write (B-C) $\approx \boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_c$.

Finally, we can write the gradient in eq. (3.50) by $\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = A(B - C)D$. The situation is similar to the description above. Additionally, we have a weighting

with the posterior probability in the factor D. In result the magnitude of the gradient becomes small if $P(c|\mathbf{z}_\mathbf{x}) \approx 0$.

We constructed a simulation with a three class classification problem with $\mathbf{x} \in \mathbb{R}^{2\times 1}$ to illustrate the nature of the gradients based on the three different visualization functions. A KFDA classifier with a Gaussian kernel was used. Figure 3.1 and 3.2 show partitioning of the input space into three class regions along with the training examples (large markers). We considered the visualization function $g_c(\mathbf{x})$ corresponding to class 2 ($c = 2$) and calculated gradients for data observations belonging to class 1 in Figure 3.1 and class 2 in Figure 3.2. The gradients are plotted as black arrows. Their lengths are normalized in each plot such that the longest gradient has unit norm, since we are interested in the relative length variation across the gradients. Figure 3.1 illustrates that relatively many gradients have similar magnitude when using $g_c(\mathbf{x}) = f(\mathbf{z}_\mathbf{x})$ or $g_c(\mathbf{x}) = \log[P(c|\mathbf{z}_\mathbf{x})]$, while only a limited number of gradients have a significant magnitude when using $g_c(\mathbf{x}) = P(c|\mathbf{z}_\mathbf{x})$ as a visualization function. Figure 3.2 illustrates that using $g_c(\mathbf{x}) = \log[P(c|\mathbf{z}_\mathbf{x})]$ or $g_c(\mathbf{x}) = P(c|\mathbf{z}_\mathbf{x})$ leads to small magnitudes of the gradients when $p(c|\mathbf{z}_\mathbf{x}) \approx 1$. Additionally, using $g_c(\mathbf{x}) = P(c|\mathbf{z}_\mathbf{x})$ also leads to small magnitudes of the gradients when $p(c|\mathbf{z}_\mathbf{x}) \approx 0$.

3.6.3 Estimating sensitivity maps

Estimation of sensitivity maps requires integration over (part of the) input domain $\mathcal{X}$ according to the definition eq. (3.45). In general the density $p(\mathbf{x})$ is unknown, and we invoke the sampling distribution $p(\mathbf{x}) \approx \frac{1}{N_\mathcal{I}} \sum_{n\in\mathcal{I}} \delta(\mathbf{x} - \mathbf{x}_n)$. Hence, the integral is approximated by a finite sum over data observations

$$\mathbf{s}_k^c \approx \frac{1}{N_\mathcal{I}} \sum_{n\in\mathcal{I}} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \,|_{\mathbf{x}=\mathbf{x}_n} \right]^k, \quad k \in \{1, 2\} \tag{3.56}$$

where $\mathcal{I}$ is a set containing data observation indices, and $N_\mathcal{I}$ is the number of members in the set $\mathcal{I}$. In order to estimate sensitivity maps we must select the set $\mathcal{I}$ over which the summation in eq. (3.56) is done. Additionally, one or more output channels c must be selected. Many different choices and combinations exist, and in the following we will outline just a few.

I: Class specific map version 1

Let all observations of class c define the set $\mathcal{I}_c$. Output channel c is considered which leads to the map

$$\mathbf{s}_k^c = \frac{1}{N_{\mathcal{I}_c}} \sum_{n \in \mathcal{I}_c} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_n} \right]^k, \quad k \in \{1, 2\}. \tag{3.57}$$

We here use the notation $\mathbf{s}_k^c$ referring to that the output corresponding the c'th class is considered.

II: Class specific map version 2

Let all observations define the set $\mathcal{I}$. Output channel c is considered which leads to the map

$$\mathbf{s}_k^c = \frac{1}{N_{\mathcal{I}}} \sum_{n \in \mathcal{I}} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_n} \right]^k, \quad k \in \{1, 2\}. \tag{3.58}$$

Again the notation $\mathbf{s}_k^c$ refers to that the output corresponding the c'th class is considered.

III: Grand average map version 1

Let all observations of class c define the set $\mathcal{I}_c$. Furthermore all output channels are considered which leads to a grand average map $\mathbf{s}_k^{ga}$ as

$$\mathbf{s}_k^{ga} = \frac{1}{C} \sum_{c=1}^{C} \left\{ \frac{1}{N_{\mathcal{I}_c}} \sum_{n \in \mathcal{I}_c} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_n} \right]^k \right\}, \quad k \in \{1, 2\}. \tag{3.59}$$

IV: Grand average map version 2

Let all observations define the set $\mathcal{I}$. Furthermore all output channels are considered which leads to a grand average map $\mathbf{s}_k^{ga}$ as

$$\mathbf{s}_k^{ga} = \frac{1}{C} \sum_{c=1}^{C} \left\{ \frac{1}{N_{\mathcal{I}}} \sum_{n \in \mathcal{I}} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_n} \right]^k \right\}, \quad k \in \{1, 2\}. \tag{3.60}$$

V: Interclass contrast map

Let all observations in class c' define the set $\mathcal{I}_{c'}$. Furthermore output channel c is considered which leads to the map

$$\mathbf{s}_k^{c|c'} = \frac{1}{N_{\mathcal{I}_{c'}}} \sum_{n \in \mathcal{I}_{c'}} \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_n} \right]^k, \quad k \in \{1, 2\}. \tag{3.61}$$

We here use the notation $\mathbf{s}_k^{c|c'}$ referring to that the output corresponding the c'th class is considered, and that the map is based on data observations in class c'.

VI: Weighted maps

The contributions from the individual gradients are weighted by some factor w_n^m, where m indicates that we can have multiple weight factors per data observation. This procedure can be applied to all maps eq. (3.57 - 3.61). For example, applying the weighting factor to the interclass contrast map eq. (3.61) gives

$$\mathbf{s}_k^{c|c',m} = \frac{1}{N_{\mathcal{I}_{c'}}} \sum_{n \in \mathcal{I}_{c'}} w_n^m \left[\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_n} \right]^k, \quad k \in \{1, 2\}. \tag{3.62}$$

The motivation behind the introduction of the weight factor is as follows. Consider the maps eq. (3.57 - 3.61). Cancellation can occur when estimating the *signed sensitivity map* ($k = 1$) due the possible existence of sign differences in the gradients across the members of $\mathcal{I}$. Even when using linear classifiers such cancellation can occur if $C > 2$. One possible solution is to calculate the *sensitivity map* ($k = 2$) as a visualization, where the squaring remove such cancellation effects. However, it may be relevant to derive a visualization that contains sign information. By use of the weight factor we can hope to mitigate the problem of cancellation to some extent. For example, we could perform clustering of the data observations as defined by $\mathcal{I}$ such that observations with similar gradient $\partial g_c(\mathbf{x}) / \partial \mathbf{x}|_{\mathbf{x}=\mathbf{x}_n}$ are assigned to the same cluster. E.g. by modeling the data by M clusters we obtain $w_n^m = P(m|\mathbf{z}_{\mathbf{x}n})$ for $m \in 1, \ldots, M$ with $P(m|\mathbf{z}_{\mathbf{x}n})$ being the posterior probability of point $\mathbf{x}_n$ belonging to cluster m. The notation $\mathbf{s}_k^{c|c',m}$ refers to that the output corresponding the the c'th class is considered, and that the map is based on data observations in class c', where each gradient is weighed according to the weights w_n^m. Hence, the weighted maps can be seen as a refinement of the interclass contrast maps in (V), where we for each interclass contrast map obtain M maps by procedure (VI). Figure 3.3 illustrates a scenario, where some heterogeneity in gradient orientation within the same class

exists. No cancellation will occur when calculating the sensitivity map (squaring the gradients). However sign information is lost. Summation of gradients without squaring results in cancellation effects. The resulting map will fail to capture that the gradients have a considerable component along the second dimension. Generating weighted maps (multiple maps for each class) could lead to discovery of the relevance of the second dimension while maintaining sign information.

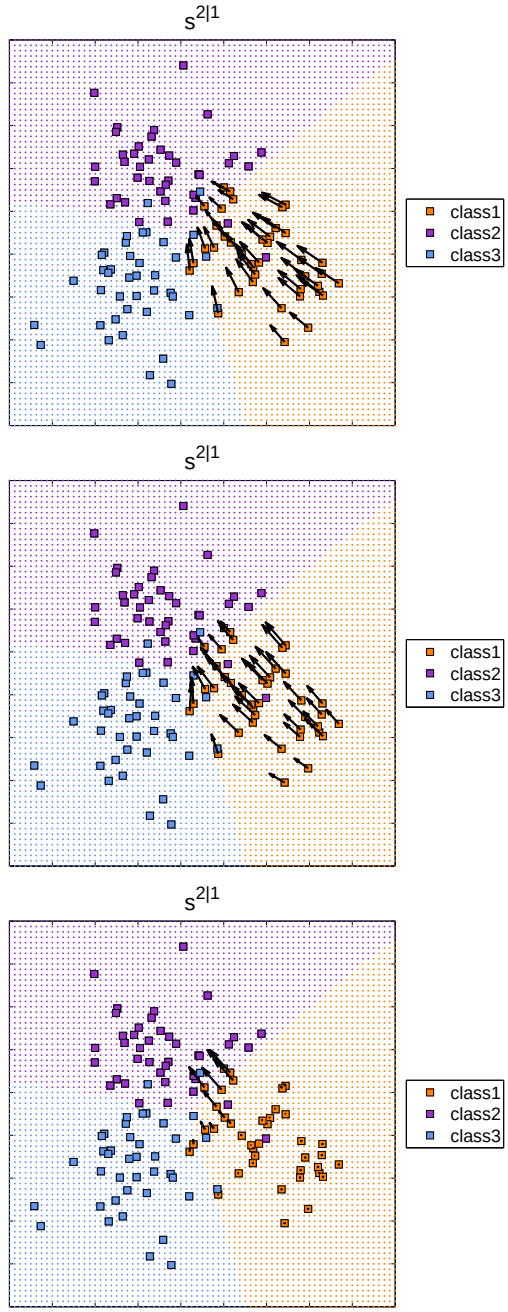

Figure 3.1: Illustration of gradients (arrows) based on three different visualization functions $g_c(\mathbf{x})$. From top to bottom: $g_c(\mathbf{x}) = f(\mathbf{z}_\mathbf{x})$, $g_c(\mathbf{x}) = \log[P(c|\mathbf{z}_\mathbf{x})]$, and $g_c(\mathbf{x}) = P(c|\mathbf{z}_\mathbf{x})$. The plots show how each point belonging to class 1 should be changed locally in order to increase the likelihood of the point being assigned to class 2.

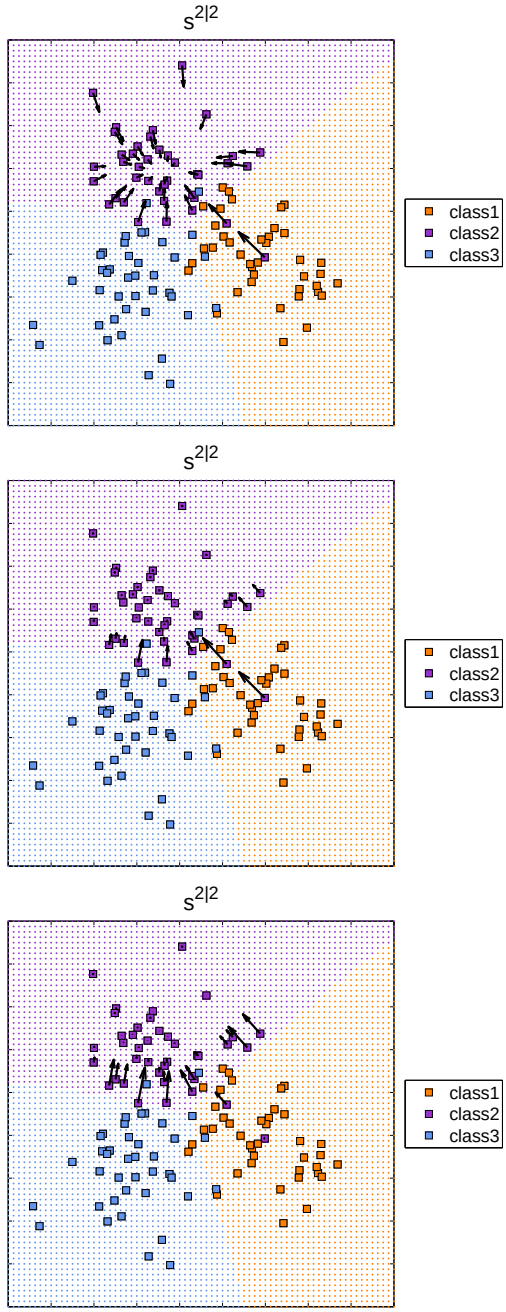

Figure 3.2: Illustration of gradients (arrows) based on three different visualization functions $g_c(\mathbf{x})$. From top to bottom: $g_c(\mathbf{x}) = f(\mathbf{z_x})$, $g_c(\mathbf{x}) = \log[P(c|\mathbf{z_x})]$, and $g_c(\mathbf{x}) = P(c|\mathbf{z_x})$. The plots show how each point belonging to class 1 should be changed locally in order to increase the likelihood of the point being assigned to class 1.

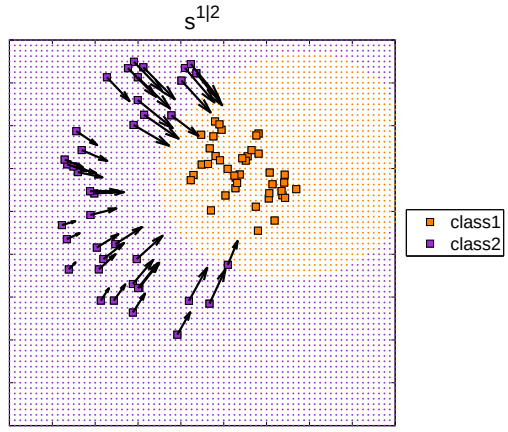

Figure 3.3: Illustration of scenario with considerable heterogeneity among the gradients' orientations within the same class. Summation of the gradients will lead to cancellation effect along the second dimension. Hence, a map based on the sum of the signed gradients will primarily identify the first dimension as important to the discriminative task. (Gradients are based on the visualization function $g_c(\mathbf{x}) = f(\mathbf{z}_\mathbf{x})$).

3.7 Denoising and localized visualization using kernel principal component analysis and pre-image estimation

Kernel principal component analysis (KPCA) is a nonlinear generalization of PCA and operates in a feature space $\mathcal{F}$, (Schölkopf et al., 1998; Mika et al., 1999b). Linear PCA is often used as an exploratory tool relying on the decomposition $\tilde{\mathbf{X}} = \mathbf{SV}$ of the centered data matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{N \times P}$, where $\mathbf{V} \in \mathbb{R}^{N \times P}$ holds orthonormal basis vectors in the rows, and $\mathbf{S} \in \mathbb{R}^{N \times N}$ holds the data observations' coordinates with respect to this new basis or the so called PC *scores* $\mathbf{S} = \tilde{\mathbf{X}}\mathbf{V}^\top$. Hence, the PC scores can be interpreted by inspecting a corresponding basis vectors, e.g. Bullmore et al. (1996); Hansen et al. (1999). Likewise, denoising can be performed by projecting a data observation onto a subspace spanned by a few basis vectors e.g. by truncating the basis $\mathbf{V}$ to only include the first K rows $\mathbf{V}_K$. The denoising operation is then $\tilde{\mathbf{X}}_{\text{denoised}} = \tilde{\mathbf{X}}\mathbf{V}_K^\top\mathbf{V}_K$. In KPCA it is straightforward to obtain the PC scores. However, is it complicated to i) going from a data point's projections in feature space $\mathcal{F}$ to the observation's representation in the input space $\mathcal{X}$, and ii) interpret the PC scores, since basis vectors do not directly have a representation in the input space. In the following we will briefly summarize KPCA as introduced by Schölkopf et al. (1998); Mika et al. (1999b). This is following by a description of the *pre-image* procedure that attempt to recover input space representations from data points feature space projections (Mika et al., 1999b). Finally, we propose an interpretation procedure for KPCA that is based on sensitivity mapping visualization.

3.7.1 Kernel principal component analysis

Let $\phi : \mathcal{X} \mapsto \mathcal{F}$ be a mapping from the D-dimensional input space, $\mathcal{X}$, to the feature space, $\mathcal{F}$. Now, let $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be N data points in $\mathcal{X}$ and $\{\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)\}$ be the corresponding points in $\mathcal{F}$.

KPCA is similar to PCA that estimates a set of orthogonal basis vectors that diagonalizes the covariance. However, rather than working in $\mathcal{X}$ KPCA operates in $\mathcal{F}$. Specifically, KPCA diagonalizes the covariance matrix in feature space

$$\mathbf{C} = \frac{1}{N} \sum_{n=1}^{N} \tilde{\boldsymbol{\phi}}(\mathbf{x}_n) \tilde{\boldsymbol{\phi}}(\mathbf{x}_n)^\top, \tag{3.63}$$

where it is assumed that the data has been centered (in feature space by $\tilde{\boldsymbol{\phi}}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}) - \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)$). By transforming the data by an orthogonal basis defined

as basis vectors in the rows of $\mathbf{V} \in \mathbb{R}^{N \times F}$ we achieve the new uncorrelated set of coordinates. Specifically, we seek eigenvalues $\lambda \geq 0$ and a corresponding set of eigenvectors $\mathbf{v}$ that satisfy the eigenvalue problem

$$\mathbf{C}\mathbf{v} = \lambda \mathbf{v}. \tag{3.64}$$

Now, it follows that the eigenvectors $\mathbf{v}$ lies in the span of the mapped data observations $\left\{\tilde{\boldsymbol{\phi}}(\mathbf{x}_1), \ldots, \tilde{\boldsymbol{\phi}}(\mathbf{x}_N)\right\}$, such that $\mathbf{v} = \sum_{n=1}^{N} \alpha_n \tilde{\boldsymbol{\phi}}(\mathbf{x}_n)$ with $\boldsymbol{\alpha} \in \mathbb{R}^{N \times 1}$ being a coefficient vector (Schölkopf et al., 1998; Mika et al., 1999b). By exploiting the kernel trick we can solve an equivalent eigenvalue problem

$$\tilde{\mathbf{K}}\boldsymbol{\alpha} = N\lambda\boldsymbol{\alpha}, \tag{3.65}$$

with $\tilde{\mathbf{K}}$ being the centered kernel matrix[4] The solutions $\boldsymbol{\alpha}^j$ are normalized by the requirement $\mathbf{v}^{j\top}\mathbf{v}^j = 1$, which translates into requiring $\lambda_j \boldsymbol{\alpha}^{j\top}\boldsymbol{\alpha}^j = 1$. Finally, a mapped data observation $\tilde{\boldsymbol{\phi}}(\mathbf{x})$ can be projected onto the basis vector $\mathbf{v}^j$ by the operation

$$\begin{aligned} \beta(\mathbf{x})_j &= \mathbf{v}^{j\top}\tilde{\boldsymbol{\phi}}(\mathbf{x}) \\ &= \sum_{n=1}^{N} \alpha_n^j \tilde{k}(\mathbf{x}_n, \mathbf{x}), \end{aligned} \tag{3.66}$$

with $\tilde{k}(\mathbf{x}_n, \mathbf{x})$ being the n'th element of the vector $\tilde{\mathbf{k}}_{\mathbf{x}}$ defined as the centralized version of the vector $\mathbf{k}_{\mathbf{x}}$.

As with PCA we may expect the underlying relevant structure of the data to be present in a subspace. Hence, we can retain $q < N$ components in order to perform KPCA denoising. Equivalent to PCA, the squared reconstruction error is minimal and the retained variance is maximal for KPCA. However, these properties hold in $\mathcal{F}$ not in $\mathcal{X}$. For a more thorough derivation of KPCA the reader is referred to Schölkopf et al. (1998); Mika et al. (1999b).

3.7.2 Pre-image estimation

Suppose that we are interested in a denoised version of an observation $\mathbf{x}_0$. A data observation's feature space projection can be obtained by means of eq. (3.66). The main challenge in denoising by KPCA is the mapping of denoised feature

[4] Centralizing the data in the feature space can be performed by the following operations (Schölkopf et al., 1998; Rosipal et al., 2001). Training data: $\tilde{\mathbf{K}} = \left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_n\mathbf{1}_n^\top\right)\mathbf{K}\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)$, with N being the number of training data observations. Test data: $\tilde{\mathbf{k}_{\mathbf{x}}}^\top = \left(\mathbf{k}_{\mathbf{x}}^\top - \frac{1}{N}\mathbf{1}_N^\top\mathbf{K}\right)\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)$.

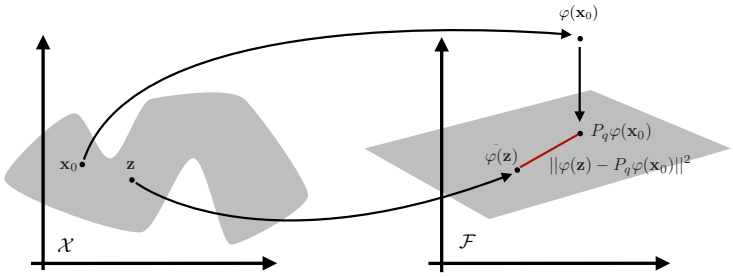

Figure 3.4: The pre-image problem in kernel principal component analysis denoising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace in feature space, $\mathcal{F}$. (This image was kindly provided by Trine Jule Abrahamsen, DTU Informatics).

space points back into input space. Assuming that the given feature space point lies in the span of $\left\{\tilde{\boldsymbol{\phi}}(\mathbf{x}_1), \ldots, \tilde{\boldsymbol{\phi}}(\mathbf{x}_N)\right\}$ implies that it can be represented as a linear combination of the training images. The pre-image problem consists of finding a point $\mathbf{z} \in \mathcal{X}$ such that $\tilde{\boldsymbol{\phi}}(\mathbf{z}) = P_q\tilde{\boldsymbol{\phi}}(\mathbf{x}_0)$, where P_q denote the projection onto a subspace. $\mathbf{z}$ is then called the pre-image of $P_q\tilde{\boldsymbol{\phi}}(\mathbf{x}_0)$.

Since a function has an inverse if and only if it is bijective, ϕ will not be invertible for most nonlinear kernel functions, and thus the pre-image problem is ill-posed (Burges, 1998; Schölkopf et al., 1998, 1999; Mika et al., 1999b; Kwok and Tsang, 2004). For many choices of kernels $\dim(\mathcal{F}) \gg \dim(\mathcal{X})$, and it follows that not all points in $\mathcal{F}$ or even the subspace spanned by $\{\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)\}$ is the image of any $\mathbf{x}$. Furthermore, whenever ϕ is not injective, uniqueness of a recovered pre-image is not guaranteed.

Since an exact pre-image often does not exist, various approaches to the nonlinear optimization problem of finding an approximate pre-image have been developed (Mika et al., 1999b; Kwok and Tsang, 2004; Abrahamsen and Hansen, 2011b). The original work by Mika et al. (1999b) proposed a fixed-point iterative approach by seeking a point in input space which maps into a point in feature space 'as close as possible' to $P_q\tilde{\boldsymbol{\phi}}(\mathbf{x}_0)$ (see Figure 3.4). Thus the pre-image estimate is defined as a point which minimizes the Euclidean distance between $\tilde{\boldsymbol{\phi}}(\mathbf{z})$ and $P_q\tilde{\boldsymbol{\phi}}(\mathbf{x}_0)$ with respect to $\mathbf{z}$

$$R(\mathbf{z}) = ||\tilde{\boldsymbol{\phi}}(\mathbf{z}) - P_q\tilde{\boldsymbol{\phi}}(\mathbf{x}_0)||^2. \tag{3.67}$$

For the Gaussian kernel Mika et al. (1999b) devised a fixed point iteration scheme to estimate pre-images based on minimization of the objective eq. (3.67). As any other iterative approach to nonlinear optimization problems, this method can suffer from convergence to local minima and sensitivity to the initialization.

Kwok and Tsang (2004) proposed a closed form solution to the pre-image problem. Their approach is based on the assumption that for any two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ there exists a simple relation between their Euclidean distance in input space and the distance between the corresponding ϕ-mapped images in feature space. The relation between the distance measures is obtained by exploiting the idea of multidimensional scaling, where a low dimensional distance preserving manifold is sought. Instead of using all the training points, only the k nearest neighbors in feature space are used for the pre-image estimation. The basic idea of the method by Kwok and Tsang (2004) is to estimate the pre-image by projection onto the subspace in input space spanned by the chosen neighbors.

3.7.3 Global visualization of kernel principal component analysis

The pre-image estimation provides reconstructions of *localized* feature space points in input space. Another relevant issue is to assess the importance of each input space dimension to the PC scores. Such interpretation can be achieved for linear PCA by inspecting the basis vectors in $\mathbf{V}$. Consider KPCA with a linear kernel. Since $\tilde{\boldsymbol{\phi}}(\mathbf{x}) = \mathbf{x} - \mathbf{m}$, with $\mathbf{m}$ defined as the observation mean vector, we can recover the j'th basis vector simply by

$$\begin{aligned} \mathbf{v}^j &= \sum_{n=1}^{N} \alpha_n^j \tilde{\boldsymbol{\phi}}(\mathbf{x}_n) \\ &= \sum_{n=1}^{N} \alpha_n^j (\mathbf{x}_n - \mathbf{m}). \end{aligned} \tag{3.68}$$

Our basic idea here is to apply the sensitivity mapping procedure in order to obtain similar interpretation of KPCA using other kernels than the linear kernel. Hence, we are interested in a model visualization of a KPCA residing in input space. To achieve this, the feature space projections $\beta(\mathbf{x})_j$ in eq. (3.66) is used as a visualization function in the general definition eq. (3.45). For example, consider a scenario where we are interested in the relative importance of input dimensions to data observations' embeddings along the j'th component of KPCA. The gradient of the visualizing function $g_c(\mathbf{x}) = \beta(\mathbf{x})_j$ reads

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) &= \frac{\partial}{\partial \mathbf{x}} \sum_{n=1}^{N} \alpha_n^j \tilde{k}(\mathbf{x}_n, \mathbf{x}) \\ &= \frac{\partial}{\partial \mathbf{x}} \sum_{n=1}^{N} \alpha_n^j \left[k(\mathbf{x}_n, \mathbf{x}) - \frac{1}{N} \sum_{n'=1}^{N} k(\mathbf{x}_n, \mathbf{x}) \right]. \end{aligned} \tag{3.69}$$

For example, using the linear kernel we immediately have the gradient as

$$\frac{\partial}{\partial \mathbf{x}} g_c(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n^j (\mathbf{x}_n - \mathbf{m}). \tag{3.70}$$

Hence, when using the gradient eq. (3.70) and the *signed sensitivity map* definition eq. (3.45) ($k = 1$) the visualization obtained with the sensitivity mapping procedure will be equivalent to the conventional visualization of linear PCA via visualization of basis vectors $\mathbf{v}$. In general similar visualizations can be derived from KPCA models using e.g. the Gaussian kernel with the gradient given in eq. (3.54).

3.8 Model evaluation

3.8.1 Prediction accuracy, model reproducibility, and NPAIRS resampling

Different metrics can be used in order to assess the 'quality' of a trained predictive model. A natural measure of a trained model's ability to provide meaningful predictions is the model's generalization error (or similarly the prediction accuracy) (Mørch et al., 1997; Hastie et al., 2009) as defined by

$$G_{\boldsymbol{\theta}} = \int e\left(y, \hat{y}|\boldsymbol{\theta}\right) p\left(\mathbf{x}, y\right) d\mathbf{x}dy, \tag{3.71}$$

where $e\left(y, \hat{y}|\boldsymbol{\theta}\right)$ is some 'error' measure[5] and $\boldsymbol{\theta}$ denote that the model is parametrized by a set of parameters. Examples of error measures $e\left(\cdot\right)$ are the squared error, deviance, or classification error. Typically, the joint distribution $p\left(\mathbf{x}, y\right)$ is unknown and we invoke the sampling distribution $p\left(\mathbf{x}, y\right) \approx \frac{1}{N}\sum_n \delta\left(\mathbf{x} - \mathbf{x}_n, y - y_n\right)$. The generalization error is then estimated over a finite number of samples. In ideal settings an independent data set will be available for performance evaluation. However, often only a limited sized data set $\mathcal{D}$ is available for model building, model selection, and performance evaluation. A strategy is to use $\mathcal{D}$ for both model building, selection, and performance assessment. This leads to the *re-substitution* error. When building flexible models the re-substitution error can be overly optimistic and underestimate the generalization error. To alleviate such bias different partitioning and resampling schemes has been proposed, see e.g (Molinaro et al., 2005). Such methods partition $\mathcal{D}$ into disjoint training and test sets $\mathcal{S}_{test}$ and $\mathcal{S}_{train}$. Furthermore, the training set can be partitioned into a training set and a validation set for model selection. Hence, the generalization error is estimated by

$$G_{\boldsymbol{\theta}_{\mathcal{S}_{train}}} = \frac{1}{N_{\mathcal{S}_{test}}} \sum_{i \in \mathcal{S}_{test}} e\left(y_i, \hat{y}_i|\boldsymbol{\theta}_{\mathcal{S}_{train}}\right), \tag{3.72}$$

where $\boldsymbol{\theta}_{\mathcal{S}_{train}}$ denote that the model parameters have been learned from the training set.

There exist a variety of strategies for splitting $\mathcal{D}$. One approach is n-fold cross-validation. This method assigns the observations in $\mathcal{D}$ to one of n partitions. $n-1$ partitions will serve as a training set and the last partition will serve as the test set. The generalization error is then assessed with each of the n partitions being the test set, and the resulting estimates of generalization error is

[5] Note that the predictions depend on the variable $\mathbf{x}$, i.e. $\hat{y}\left(\mathbf{x}\right)$.

averaged. The extreme case is leave-one-out cross-validation where the number of partitions equals the number of observations in $\mathcal{D}$. Resampling leads to estimation of the prediction error as

$$G = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{N_{\mathcal{S}_{test}^{(j)}}}\sum_{i\in\mathcal{S}_{test}^{(j)}} e\left(y_i, \hat{y}_i|\boldsymbol{\theta}_{\mathcal{S}_{train}^{(j)}}\right), \tag{3.73}$$

where $\mathcal{S}_{train}^{(j)}$ and $\mathcal{S}_{test}^{(j)}$ denote assignments of observations to the training and the test sets in the j'th partition. Another popular approach is the Bootstrap resampling scheme, where the training set is formed by sampling with replacement. The generalization error is then estimated based contribution from the re-substitution error and a test error estimate on samples not included in the training set, see e.g. Efron (1983); Efron and Tibshirani (1997). A general discussion of different resampling schemes is found in Hastie et al. (2009) while Molinaro et al. (2005) report on the performance of the different methods for estimating the generalization error.

Strother et al. (2002) proposed the NPAIRS (nonparametric prediction, activation, influence, and reproducibility resampling) framework for quantitative evaluation of functional neuroimaging experiments. This resampling framework use split-half resampling. The motivation behind NPAIRS is that models should not only be evaluated based on estimating the prediction accuracy/generalization error. It is equally important to assess the quality of models' visualizations (statistical parametric images (SPIs)), since they form the basis of experimental interpretation. In this split-half framework the data set $\mathcal{D}$ is split into two splits of equal size $\mathcal{S}_1$ and $\mathcal{S}_2$. A model is built on $\mathcal{S}_1$ and prediction accuracy is estimated from $\mathcal{S}_2$ and vice versa yielding two estimations of the prediction accuracy metric (p). Additionally, an SPI is extracted from each model denoted by $\mathbf{w}_1$ and $\mathbf{w}_2$. A *reproducibility* metric (r) is estimated based on a similarity measure of $\mathbf{w}_1$ and $\mathbf{w}_2$. Strother et al. (2002) proposed to measure the similarity based on the Pearson's product correlation coefficient. The split-half resampling procedure is repeated a number of times, each time with different combinations of the data observations assigned to $\mathcal{S}_1$ and $\mathcal{S}_2$.

In addition to estimates of prediction accuracy and visualization reproducibility the NPAIRS resampling scheme also provides an estimate of a reproducible SPI (rSPI): Each of the SPIs ($\mathbf{w}_i$'s, $i \in \{1, 2\}$) are normalized by their standard deviation and plotted against each other in a scatter plot, see Figure 3.5. Two axes are defined in the scatter plot. A signal axis is defined along the line of identity, and a uncorrelated noise axis is defined orthogonal to the signal axis. The projection of the normalized SPIs onto the signal and noise axes can be formed by $(\mathbf{w}_1 + \mathbf{w}_2)/\sqrt{2}$ and $(\mathbf{w}_1 - \mathbf{w}_2)/\sqrt{2}$ respectively. Finally, the rSPI is obtained by rescaling the projection onto the signal axis $\mathbf{s}$ with the standard

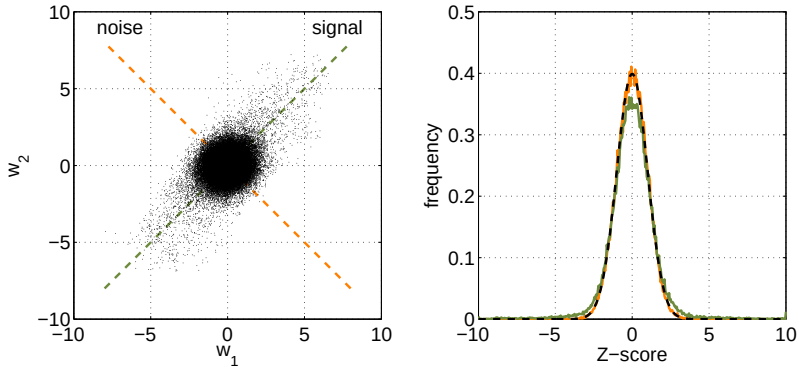

Figure 3.5: Example of a scatterplot used for generating reproducible model visualizations (rSPIs). A Fisher's linear discriminant analysis (FDA) model was trained to discriminate between (left) and (right) hand finger tapping in the finger tapping data set (Section 4.1). Two models were trained on two independent splits of the data set $\mathcal{S}_1$ and $\mathcal{S}_2$. The weight vector $\mathbf{w}$ in FDA was used as model visualization. The left plot shows a scatter plot based on weight vectors from each split-half (each scaled by its standard deviation). The lines indicate the signal axis and the uncorrelated noise axis. The right plot shows the distributions of the scatter cloud projected onto the two axes in the left plot (green is signal projection and orange is noise projection - each scaled by the standard deviation of the projection of the scatter cloud onto the noise axis). The dashed black line is the theoretical $\mathcal{N}(0,1)$ distribution.

deviation of the distribution of the scatter cloud projected onto the noise axis $\mathbf{n}$. This rescaled rSPI is referred to as a rSPI(Z), assuming that the distribution of $\mathbf{n}$ is approximated by a Gaussian $\mathcal{N}(0, 1-r)$ distribution, with r being the Pearson's product correlation coefficient. The interested reader is referred to Strother et al. (2002) for further details. An average reproducible visualization $\overline{\text{rSPI(Z)}}$ is finally obtained by averaging over the resampling repetitions[6].

[6]For model visualizations with only positive values (e.g. squared sensitivity maps) we scale the individual visualizations $\mathbf{w}$ to unit norm (Sigurdsson et al., 2004). In such cases the signal projection will only have positive values. The reproducible visualizations will then be denoted rSPI and $\overline{\text{rSPI}}$.

3.8.2 Statistical significance

The statistical significance of model performance can be assessed by means of permutation analysis (Golland and Fischl, 2003). By the permutation analysis we are interested in performing a hypothesis test, and possible reject the null hypothesis at a certain level of confidence α. Under the null hypothesis we assume that observations $\mathbf{x}$ and class labels y are independent, i.e. $p(\mathbf{x}, y) = p(\mathbf{x})\, p(y)$. We consider the generalization error eq. (3.72) as a test statistics. First, the test statistics is calculated with the correct labeling of data observation yielding t_0. Hereafter, the data observations are permuted and the test statistic t_n is recalculated. This is repeated M times yielding a distribution of t_n under the assumption that the data observations and labels are independent. Finally, t_0 can be compared to the empirical null distribution, and the null hypothesis may be rejected at level of significance α. The same procedure can be used to assess the statistical significance of the reproducibility metric. Further discussions on issues regarding the use of permutation analysis in the analysis of neuroimaging data set are found in Nichols and Holmes (2002); Pereira and Botvinick (2011).

Different approaches exist in order to assess the statistical significance of the models visualizations as summarized by the $\overline{\text{rSPI(Z)}}$ and $\overline{\text{rSPI}}$. One approach is to use the permutation analysis and consider each element in the $\overline{\text{rSPI(Z)}}$ and $\overline{\text{rSPI}}$ as a test statistics. First the reproducible visualization is constructed based on the correct labeling of data observations. Hereafter a permutation distribution is formed (for each element in the visualization) by permuting data observations and re-estimating the reproducible visualization yielding a null distribution for each element. Finally, the reproducible visualizations based on correct labeling is compared to the null distribution resulting in a p-value for each element in the model's visualization. If some null-distribution can be assumed we can obtain the set of p-values by comparing to a theoretical distribution. For example, we may compare the $\overline{\text{rSPI(Z)}}$ to the theoretical $\mathcal{N}(0, 1)$ distribution (Strother et al., 2002).

Chapter 4

Data sets

This chapter describes the fMRI data sets used in the analyses presented in the dissertation. The first data set origins from a multi-subject finger tapping experiment. This data set was used since the underlying brain network involved in finger tapping is relatively well characterized. Hence the models' ability to identify the underlying signal structure could be evaluated. The second data set origins from a multi-subject study, where we expected a lower signal to noise ratio in comparison to the finger tapping data set. The experimental design is an adaption of the Trail-Making Test. The third data set origins from a multi-subject experiment with visual checkerboard stimulation. While the evoked brain signals are expected to be confined to visual brain areas, the experimental design was deliberately constructed to allow for relatively complicated classification tasks to be formulated. The forth data set is a multi-subject data set containing several runs within each subject. This data set was included to investigate if our findings (based on multi-subject analysis using spatially filtered data) also generalized to data without spatial filtering. Additionally, this data set was expected to contain more subtle signal structure than e.g. the finger tapping data set. The four data sets were acquired from different subject groups at different centers/scanners. For compatibility with published results, we have maintained the centers' respective preprocessing pipelines.

Contents

4.1 Finger tapping experiment

The finger tapping data set origins from a multi-subject study. The experimental paradigm consisted of two paced motor conditions in the following sequence: (right) right hand finger tapping, (left) left hand finger tapping. Pacing was provided by means of a red (left condition) or green (right condition) circle flashing at 1 Hz presented at the center of a screen. Each condition was presented for 20 s followed by 9.88 s of rest with no finger tapping. The stimulation cycle was repeated 10 times in the experimental run, and 240 scan volumes were acquired in total. One experimental run per subject was conducted. The work in the present dissertation is based on up to 28 subjects from this study. The fMRI data were acquired on a 3T MR scanner (Magnetom Trio; Siemens AG, Erlangen, Germany) using a standard 1 channel birdcage transmit/receive head coil. The data set consists of functional images acquired with a repetition time (TR) of 2490 ms and structural scans for the individual subjects. Preprocessing of the fMRI time series data included the following steps for each subject: (1) rigid body realignment, (2) co-registration of the functional images to the structural scan, (3) spatial normalization of the structural scan to the MNI152 template (Montreal Neurological Institute template), (4) reslicing of images into MNI space at 3 mm isotropic voxels, (6) spatial smoothing of spatial normalized images using an isotropic 6 mm FWHM Gaussian filter, (7) low frequency components were removed from the time series with a set of discrete cosine basis functions up to a cut-off period of 128 seconds, (8) the mean resting-state volume was subtracted, based on the last two images of each rest period. Additionally, the scans were masked with a rough whole-brain mask (57,988 voxels). For the classification analysis we extracted scans from the (right) and (left) epochs, discarding two transition scans at the start of each block, which gave 120 scans in total, per subject. Further information on acquisition and preprocessing is found in Rasmussen et al. (2012b).

4.2 Trail-Making Test experiment

The Trail-Making Test data set origins from a multi-subject study. The data set is also referred to as the *trailsAB* data set in this dissertation. The experimental paradigm is an adaptation of the Trail-Making Test (AITB, 1944; Bowie and Harvey, 2006), designed for the fMRI environment (Tam et al., 2011). Task blocks alternately consisted of (Trails A), where numbers 1-14 were pseudorandomly distributed on a viewing screen, and (Trails B), where numbers 1-7 and letters A-G were shown. Subjects drew a line connecting items in sequence (1-2-3-... or 1-A-2-B-...) as quickly as possible while maintaining accuracy, over

a 20 s block using an fMRI-compatible writing tablet and stylus (Tam et al., 2011). After each task block, a 20 s Baseline block was shown, in which subjects drew a line from the center of the screen to a random dot on a circle and back, every 2 s. A 4-block, 40-scan epoch of Trails A-Baseline-Trails B-Baseline was performed two times per run, and 80 scans were acquired in total per run. Two experimental runs per subjects were conducted. Data from 14 subjects and only the second run was used in the analyses. The fMRI data were acquired on a 3T MR scanner (Magnetom Trio; Siemens AG, Erlangen, Germany) using a 12 channel birdcage transmit/receive head coil. The data set consists of functional images acquired acquired with a TR of 2000 ms and structural scans for the individual subjects. Preprocessing of the fMRI time series data included the following steps for each subject: (1) rigid body realignment, (2) in-plane spatial smoothing with a 6 mm FWHM Gaussian kernel, (3) temporal filtering using 0-3rd-order Legendre polynomials, (4) spatial normalization of the structural scan to a study specific template based on individual subjects' structural scans registered to the MNI152 template, (5) reslicing of images into MNI space at $3.125 \times 3.125 \times 5$ mm voxels, (6) the scans were masked with a rough whole-brain mask (35,132 voxels). For the classification analysis we extracted eight scans from the Trails A and Trails B epochs, discarding two transition scans at the start of each block, which gave 32 scans total, per subject. Further information on acquisition and preprocessing is found in Rasmussen et al. (2012b). For other analyses of this data set see Churchill et al. (2012a).

4.3 Xor experiment

The xor data set origins from a multi-subject study. Six subjects were enrolled after informed consent as approved by the local Ethics Committee. In the experimental paradigm the participants were subjected to four conditions presented on a screen in the following sequence: (no) no visual stimulation, (left) reversing checkerboard on the left half of the screen, (right) reversing checkerboard on the right half of the screen, (both) reversing checkerboard on both halves of the screen. In order to maintain attention the participants were instructed to keep focus on a small circle presented in the center of the screen during the experiment, and to respond with a right hand button press to a change in the color of the circle. Each condition was presented for 15 s followed by 5.04 s of rest with no visual stimulation. The stimulation cycle was repeated 12 times in the experimental run, and 576 scan volumes were acquired in total. One experimental run per subject was conducted. The fMRI data were acquired on a 3T MR scanner (Magnetom Trio; Siemens AG, Erlangen, Germany) using an 8 channel birdcage transmit/receive head coil. The data set consists of functional images acquired with a TR of 1670 ms and structural scans for the

individual subjects. Preprocessing of the fMRI time series data included the following steps for each subject: (1) rigid body realignment, (2) co-registration of the functional images to the structural scan, (3) spatial normalization of the mean echo planar imaging (EPI) image to the EPI template in SPM8, (4) reslicing of images into MNI space at 2 mm isotropic voxels, (5) spatial smoothing of spatial normalized images using an isotropic 8 mm FWHM Gaussian filter, (6) low frequency components were removed from the time series with a set of discrete cosine basis functions up to a cut-off period of 128 seconds, (7) standardization of the individual voxels time series, (8) the scans were masked with a rough whole-brain mask (75,257 voxels). For the classification analysis we extracted scan 7-11 in each epoch, and the remaining volumes were discarded to avoid contaminating effects of the hemodynamic BOLD signal. Finally, the scans extracted from each block were averaged, which gave 48 scans in total, per subject. Further information on acquisition and preprocessing is found in Rasmussen et al. (2011).

4.4 Object recognition experiment

This data set originates from the experiment of Haxby et al. (2001) on face and object representation in the human ventral temporal cortex[1]. In the experimental paradigm the subjects were viewing gray scale images of eight object categories {bottle, cat, chair, face, house, scissors, scrambled, shoe} while performing a one-back repetition detection task. Stimuli were grouped into 24 seconds blocks separated by rest periods in each experimental run. 12 experimental runs per subjects were conducted. The data set contains data from six subjects. The fMRI data were acquired on a 3T MR scanner (General Electric, Milwaukee, USA). The data set consists of functional images acquired with a TR of 2500 ms and structural scans for the individual subjects. Preprocessing of the fMRI time series data comprised the following steps for each subject: (1) The functional images were scull-stripped, (2) correction for rigid-body movement, (3) different versions of the data set were created by spatially smoothing with $\{0, 3, 6, 9, 12, 15\}$ mm FWHM isotropic Gaussian filters, (4) the time series were linearly de-trended and standardized within each run, (5) the scans were masked with subject specific masks (`mask_vt.nii`) provided with the data set (307-675 voxels, voxel size ($3.5 \times 3.75 \times 3.75$ mm)). For the analysis we used scan from the eight conditions, which gave 864 scans in total, per subject[2] Further details on the experiment and acquisition are found in Haxby et al. (2001).

[1] The data was obtained from the PyMVPA web site http://www.pymvpa.org. The authors of Haxby et al. (2001) hold the copyright of the dataset and it is available under the terms of the Creative Commons Attribution-Share Alike 3.0 license

[2] Only 792 scans were available for subject 5.

Further information on preprocessing is found in Rasmussen et al. (2012a).

Chapter 5

Experimental results

This chapter presents experimental results. The first section concerns an investigation of the relative influence of model regularization parameter choices on both the model generalization, the reliability of the spatial patterns extracted from the classification model, and the ability of the resulting model to identify relevant brain networks defining the underlying neural encoding of the experiment. This section summarizes results reported in Rasmussen et al. (2012b) based on analysis of the trailAB data set and the finger tapping data set. Additional results based on the object recognition data set are presented. The next section concerns visualization of nonlinear kernel models by sensitivity maps. This section summarizes results reported in Rasmussen et al. (2011) and Rasmussen et al. (2012c). Additionally, it is shown how the sensitivity map can provide a global visualization of a kernel principal component analysis (KPCA) model. The final section concerns image denoising using KPCA and pre-image estimation. These results have been reported in Rasmussen et al. (2012a).

Contents

5.1 Discovery of brain networks

An investigation of the relative influence of model regularization parameter choices was performed in Rasmussen et al. (2012b). Specifically, we focused how selection of model regularization parameter affected classification models' ability to: i) Provide good generalization (high prediction accuracy on a test set), ii) provide a high degree of reliability/reproducibility of the spatial patterns extracted from the models, and iii) identify relevant brain networks defining the underlying neural encoding of the experiment.

5.1.1 Analysis setup

The analysis was based on 14 subjects from the finger tapping data set, the Trail-Making Test (also referred to as trailsAB) data set, and the object recognition data set. We formulated binary classification tasks as (left) vs. (right) in the finger tapping data set, (Trails A) vs. (Trails B) in the trailsAB data set, and (bottle) vs. (face) and (face) vs. (house) in the object recognition data set[1].

The underlying brain network expected to support the classifiers decisions in the finger tapping data set is relatively well known, see e.g. Moritz et al. (2000a,b); Kustra and Strother (2001); Riecker et al. (2003); Eickhoff et al. (2005); Witt et al. (2008). To investigate to what extent different brain regions contain discriminative information we performed localized analyses of the finger tapping data set. First, we performed a region based analysis. The regions of interest (ROIs) were based on the Harvard-Oxford cortical and subcortical structural atlases and the Probabilistic cerebellar atlas included in the FSL 4.1 software package (Smith et al., 2004). The ROIs were sensorimotor cortex (SMC), cerebellum (CB), secondary somatosensory cortex (S2), and subcortical regions (SC). We also considered a whole brain (WB) region in the analysis. Figure 5.1 shows the ROIs projected onto brain slices. ROI identification was based on prior knowledge from a series of experiments involving finger tapping tasks (Moritz et al., 2000a,b; Kustra and Strother, 2001; Riecker et al., 2003; Eickhoff et al., 2005; Witt et al., 2008). The ROIs were defined according to Table 1 in the supplementary materials of Rasmussen et al. (2012b). Classification of brain scans was performed by means of an SVM. The regularization parameter C of the SVM was selected in order to maximize prediction accuracy by nested cross-validation on the training set. Secondly, to quantify the local information content throughout the entire brain we employed the searchlight

[1] In the analysis of the object recognition data we considered binary classification tasks in order to use exactly the same modeling and model visualization framework as in the analysis of the finger tapping data set and in the trailsAB data set.

method Kriegeskorte et al. (2006): For each voxel in the brain, we defined a spherical cluster with a radius of two voxels (6 mm). A local classifier was trained based on information from the voxels within the cluster (33 voxels), and the trained classifier was used to assign labels to scans in a test set. For classification we used the Gaussian Naïve Bayes (GNB) classifier, e.g. Hastie et al. (2009); Pereira and Botvinick (2011). The classifier was trained with the searchlight cluster centered on each voxel in the brain volume, giving a map of prediction accuracies for individual spatial positions. In both analyses the models were trained on seven subjects and tested on seven subjects. To estimate the prediction accuracy the training/test procedure was repeated 50 times, where subjects, in each resampling run, were randomly assigned to the two partitions.

Logistic regression (LogReg), Fisher's discriminant analysis (FDA), and support vector machine (SVM) models were used in a whole-brain classification analysis. Specifically, the objective of the analysis was to investigate how selection of the regularization parameter λ affects model performance. Hence λ was varied over the range $\lambda \in \{2^{-40}, 2^{-39}, \ldots, 2^{40}\}$ (relative to the mean of the non-zero squared singular values of the data matrix). Only linear models were considered in order to simplify the analysis. By using linear models a visualization can be directly derived from the trained model in terms of the weight vector as in eq. (3.7). In both the finger tapping data set and the trailsAB data set we considered subjects as the basic resampling unit. In the object recognition data set the runs within the individual subjects were considered as resampling units. Additionally, in the object recognition data set the impact of spatial smoothing on classifier performance was assessed by analyzing versions of the data set subjected to various degrees of smoothing ($\{0, 3, 6, 9, 12, 15\}$ mm FWHM). The split-half NPAIRS resampling strategy was used in order to evaluate the models both in terms of prediction accuracy and pattern reproducibility. Furthermore, reproducible brain images ($\overline{\text{rSPI(Z)}}$s) were generated as described in Section 3.8. 50 NPAIRS resampling splits were performed.

Additionally, we performed analyses where sparsity in the voxel dimension was imposed directly on the models of the finger tapping data set. This was done in order to explore these methods impact on visualization of the known components of the motor network. We considered two strategies for obtaining sparsity - LogReg with the ENET penalty and SVM based recursive feature elimination (RFE) (see the description of feature selection in Section 2.2.4). Resampling was again performed within a split-half framework, and models were evaluated in terms of test error averaged over two splits, with pattern reproducibility measured using the two metrics of mutual information (MI) between model weights, and overlap between voxels retained in the two models. The rSPI(Z) mapping procedure was not applied to the sparsity enforcing models. To assess the consensus in voxel selection across splits of the data, we recorded the frequency across the resampling splits at which each voxel was included in the models. Fur-

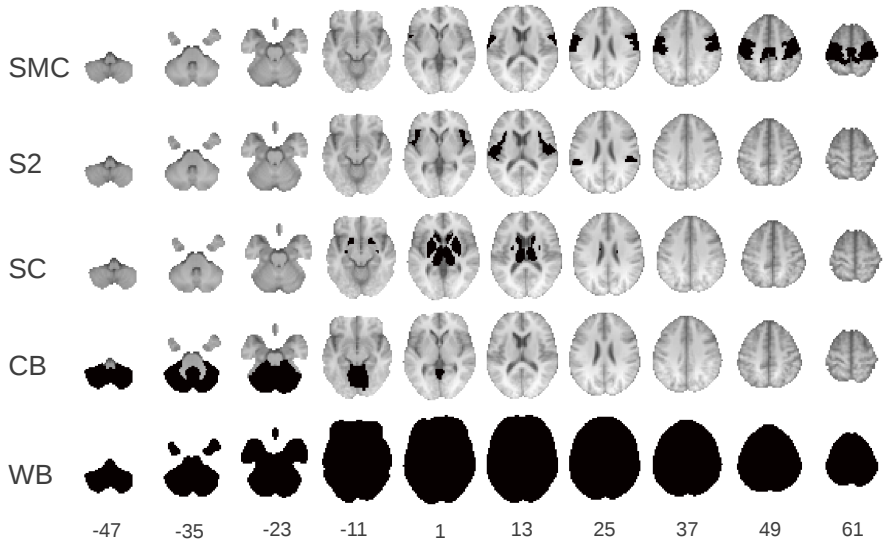

Figure 5.1: Visualization of the different regions of interest (ROIs) used in the regional analysis of the finger tapping data set. The ROIs are projected onto an average anatomical scan of the 14 subjects used in the analysis. Voxels defining the ROIs are marked with black color. The numbers below the last row of brain slices denote z coordinates in MNI space.

Region	SMC	S2	SC	CB	WB
Prediction accuracy	99.0 ***	78.5 ***	80.8 ***	98.4 ***	98.5 ***

Table 5.1: Region of interest analysis of the finger tapping data set. Split-half prediction accuracies for five brain regions. Classification was performed with an SVM. Results are based on 50 resampling splits. Statistical significance is based on a permutation test with 5000 permutations. Significance code ***: $p < 0.001$.

ther details on calculation of performance metrics and specific implementations of LogReg with ENET penalty and SVM based RFE are found in Rasmussen et al. (2012b).

5.1.2 Results

Table 5.1 provides the results of the ROI based classification analysis of the finger tapping data set. The classifiers trained on data from the WB, CB, and

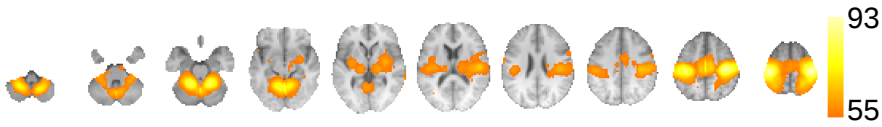

Figure 5.2: Searchlight analysis of the finger tapping data set. Accuracy map shown on subjects average anatomical scan. The map is thresholded according to $p < 0.05$ FDR correction, based on a nonparametric permutation test with 5000 permutations. The accuracy map is the mean of 50 resampling splits.

SMC regions provided high prediction accuracy, while the SC and S2 regions provided intermediate accuracies. Note that all regions provided prediction accuracies well above chance level (50%). Figure 5.2 shows the results of the searchlight analysis. A total of 12911 searchlight center voxels provided a significant prediction accuracy. Statistical thresholding was performed using the false discovery rate (FDR) control for multiple comparisons (Benjamini and Hochberg, 1995). The SC and S2 regions provided low (but still significantly different from chance level) to intermediate prediction accuracies, while CB and SMC regions provided high prediction accuracies. Note the vertical line located around SMA. Here the searchlight sphere covered voxels in both hemispheres resulting in relatively high prediction accuracy.

Figure 5.3 shows the performance of the three classifiers (SVM, LogReg, FDA) for the whole brain classification analysis of the trailsAB data set over a range of values of the regularization parameter λ. Figure 5.4 reports corresponding results, where the block labeling was permuted within each subject (see Section 3.8.2 for a description of the permutation test). As seen in Figure 5.3 all classifiers showed a transition in prediction accuracy from best accuracy at light regularization to a decreased accuracy at stronger regularization. Around $\lambda = 2^8$ we observe maximum accuracy for all classifiers. The SVM showed a somewhat steeper transition from high to low accuracy compared to the other models. All classifiers showed a transition from low reproducibility at light regularization to high reproducibility at strong regularization. For the SVM, we also plot the number of support vectors retained in the model, which tend to increase with increasing reproducibility. In addition, to obtain good reproducibility (e.g., > 0.3) we need to retain the majority of data points, with > 200 support vectors from our 224 input scans in a split-half subsample after dropping transitions. Note that the 'hard-margin' SVM corresponds to the limit with low regularization (high C since $C = 1/\lambda$), producing the least reproducible model. In the finger tapping data set we observed the same behavior in the performance metrics, but with higher values of prediction accuracy and reproducibility. Figure 5.5 shows performance of the LogReg classifier as a function of the model's effec-

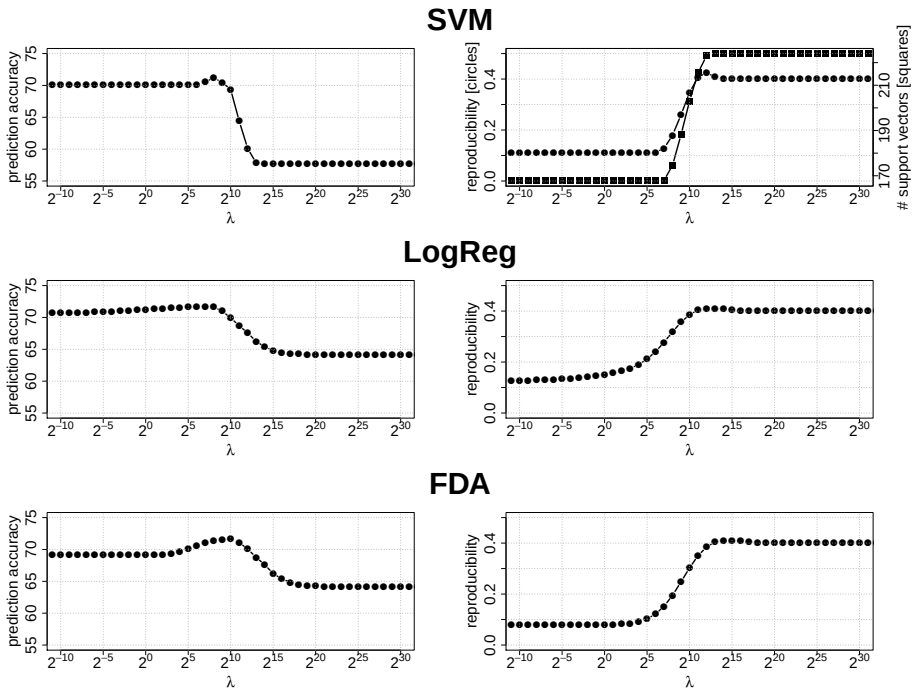

Figure 5.3: Performance of the SVM, LogReg, and FDA classifiers on the trailsAB data set, over a range of values for the regularization parameter λ. Large values of λ corresponds to strong regularization, while small values of λ corresponds to light regularization (the regularization parameter λ is here reported relative to the mean of the non-zero squared singular values of the data matrix). For the SVM the conventionally used complexity parameter C is given by $C = 1/\lambda$. Note that a 'hard margin' SVM corresponds to an SVM with light regularization. Prediction accuracies are reported in the left panels, and pattern reproducibilities in terms of the Pearson's product correlation coefficient are reported in the right panel. The number of support vectors in the SVM are also plotted over the regularization range. The curves are based on averages of 50 NPAIRS resampling splits.

tive degrees of freedom as estimated according to eq. 3.23. The reproducibility metrics supports a model of relatively low model complexity, while prediction accuracy supports more complex models.

Figure 5.6 shows pr-curves based on prediction accuracies and pattern reproducibilities for the three models in the finger tapping data set and in the trailsAB data set. A point on the curves corresponds to a particular value of λ. In the

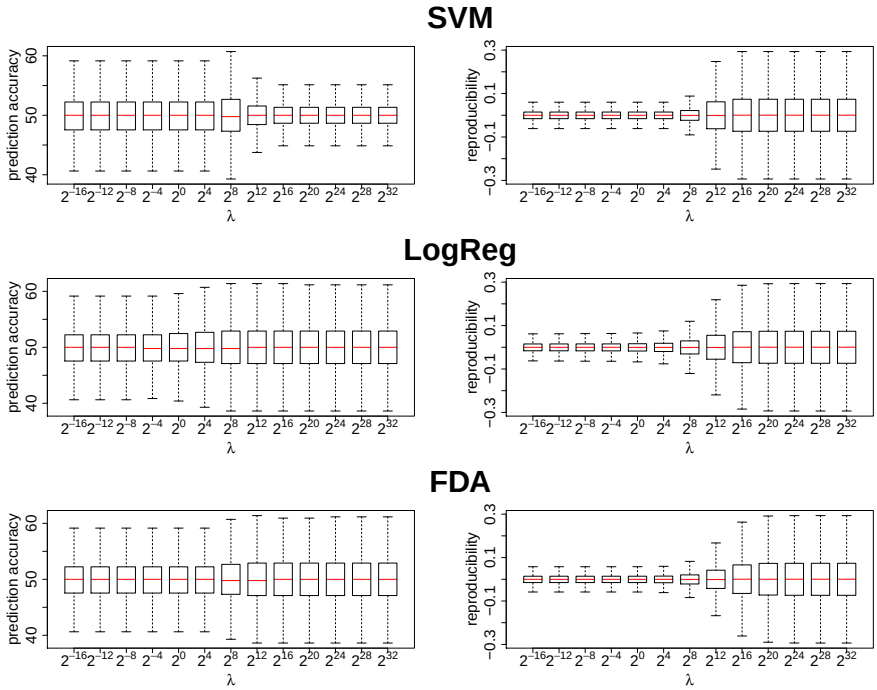

Figure 5.4: Performance of the SVM, LogReg, and FDA classifiers on the trailsAB data set, over a range of values for the regularization parameter λ in permuted data. Large values of λ corresponds to strong regularization, while small values of λ corresponds to light regularization. For the SVM the conventionally used complexity parameter C is given by $C = 1/\lambda$. Prediction accuracies are reported in the left panels, and pattern reproducibilities in terms of the Pearson's product correlation coefficient are reported in the right panel. The plots are based on permutation of block labels within each subject. 5000 permutations were performed.

finger tapping data set, we observe the best reproducibilities at a high degree of regularization and the best prediction accuracy with decreasing regularization. The curves follow the same paths, with SVM having a somewhat lower accuracy with strong regularization, and FDA has its maximum prediction at lower reproducibility than the other models with weak regularization. Note that model performance in terms of test error is $\geq 90\%$ well above chance level for all models. In the trailsAB data set we observe that the FDA and LogReg classifiers follow the same path, while that of the SVM has lower prediction and/or reproducibility except for weak and strong regularization. In terms of prediction accuracy and pattern reproducibility we observe the best combined performance

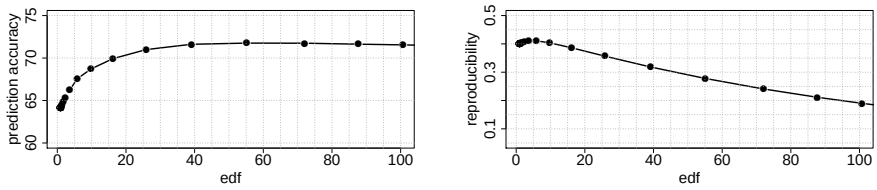

Figure 5.5: Performance of the LogReg classifier on the trailsAB data set plotted against an estimate of the model's effective degrees of freedom (edf).

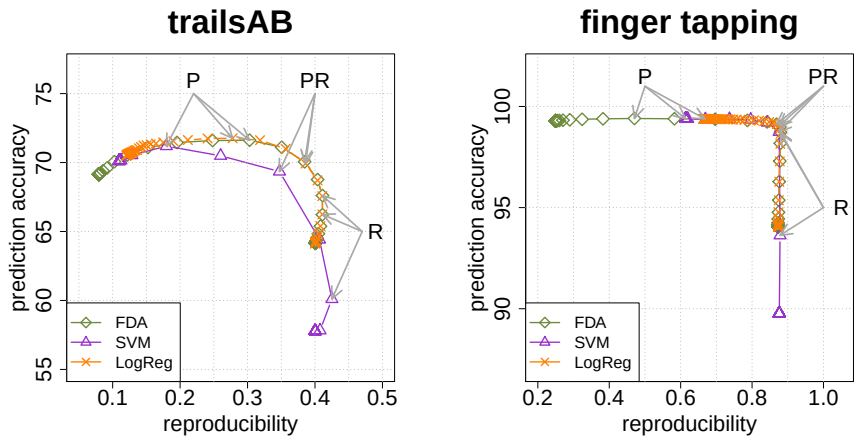

Figure 5.6: Prediction/reproducibility curves (pr-curves) for the three different classifiers. Left is the model performance on the trailsAB data set and right is the performance on the finger tapping data set. The curves are constructed by changing the regularization parameter in the models. The curves show the mean of 50 NPAIRS resampling splits. Based on the curves we selected three models within each classifier type indicated by arrows in the plots. For each classifier type, P, PR, and R correspond to optimization of prediction accuracy, joint optimization of prediction accuracy and reproducibility, and optimization of reproducibility respectively.

at intermediate levels of regularization for all classifiers. Importantly, there is a trade-off between prediction accuracy and pattern reproducibility. In both data sets and all models, there are relatively low gains in prediction accuracy and large losses in reproducibility when moving from PR to P on the pr-curves.

Figure 5.7 shows $\overline{\text{rSPI(Z)}}$s for the finger tapping data set, for the 9 different

model choices (3 classifier types and 3 selected models per classifier). Positive voxel values represent voxels in which a signal increase will drive the classifier towards a (right) classification. Negative voxel values represent voxels in which a signal increase will drive the classifier towards a (left) classification. The most reproducing voxels (with the highest Z-scores) are primarily expressed in the contralateral sensorimotor cortex (SMC), supplementary motor area (SMA) and in the ipsilateral regions in superior and inferior cerebellum (CB). Voxels with intermediate Z-scores are expressed in the contralateral second somatosensory area (S2), thalamus and putamen (SC). These activations are consistent with many prior studies, for example using multivariate classification in PET (Kustra and Strother, 2001), and a voxel-wise coordinate-based meta analysis of fMRI and PET studies (Witt et al., 2008). In general, there is a strong consistency between $\overline{\text{rSPI(Z)}}$s of the R and PR models across all three classifiers. For all three classifiers, and in particular for FDA, choosing prediction as an optimization criterion has an impact of the $\overline{\text{rSPI(Z)}}$s. The subcortical regions, some anterior cerebellar regions, and S2 are expressed with decreased Z-scores in the P models, and also the spatial extent of voxels with high Z-scores in SMC, SMA, and cerebellar regions is reduced.

The $\overline{\text{rSPI(Z)}}$s for the trailsAB task are presented in Figure 5.8. Voxels that drive the classifier towards a (Task B) classification (i.e., positive Z-scores) are primarily expressed in the precuneus (PreC) and superior parietal lobes (SPL), as well as the SMA and left precentral (LP) gyrus. More ventrally, reproducible signal is also shown in the left inferior-frontal (IF) and postcentral gyri, and the right cerebellar (CB) lobe. Reproducible voxels that drive the classifier towards a (Task A) classification (negative Z-scores) are also observed in the posterior cingulate cortex (PCC) and middle temporal (MT) lobe (predominantly right-side). Reproducible signal is also observed in the superior frontal gyrus and medial orbitofrontal (MO) gyrus, along with potentially artifactual signal near the brainstem. These activations have also been previously observed for multivariate analysis of the trailsAB task (Churchill et al., 2012a). Of the three classifiers, only the LogReg $\overline{\text{rSPI(Z)}}$ appears somewhat less sensitive to the chosen optimization criterion. For SVM, prediction has a marked influence on the $\overline{\text{rSPI(Z)}}$, as both P and PR optimization tend to primarily reinforce dorsal activation in the SMA, PreC and SPL, along with CB, while other loci show reduced Z-scores. The Z-scores of voxels that drive the classifier towards a (Task A) classification are generally reduced, indicating that these regions contribute to a less predictive, but more reproducible model. For the FDA model, optimization on P again reinforces dorsal regions of activation. However, PR and R optimization methods now produce relatively similar $\overline{\text{rSPI(Z)}}$.

Figure 5.9 provides a quantitative comparison of the average $\overline{\text{rSPI(Z)}}$s for all models in both data sets: finger tapping above the diagonal and trailsAB below. The R and PR models' spatial patterns are very similar across the classifiers,

whereas for the P models we observe less agreement. In addition, within each model the P and R models at different ends of the pr-curves in Figure 5.6 are most dissimilar. Note that the correlation between the average rSPI(Z)s generally are high - the lowest value is found for SVM P vs. SVM R with 0.79 in finger tapping. Similarities across and within models are generally higher than the reproducibilities between splits within each model (see Figure 5.6). These map similarities also appear consistent with the relative spatial similarities of the un-thresholded maps in Figure 5.7 and 5.8.

To identify significant activations in the rSPI(Z)s obtained from the models, the rSPI(Z)s were thresholded as follows: i) for each voxel we computed a p-value based on the voxels Z-score in the rSPI(Z) (by use of the theoretical $\mathcal{N}(0,1)$ distribution, see Section 3.8.2), ii) the rSPI(Z) was then thresholded using a statistical threshold of $p < 0.05$ FDR correction. Hence, for a particular classification model with a particular regularization parameter value we obtain a brain map that is sparse due to the statistical thresholding. For the finger tapping data set we then counted the number of voxels in the thresholded map that were included within each of the ROIs shown in Figure 5.1. This was repeated for all regularization parameter values and all classification models. A second analysis was performed for both the finger tapping data set and the trailsAB data set. Here we considered (for each classifier type) a single ROI which was defined by the network of voxels in the rSPI(Z) of the pr-maximizing model that survived thresholding according to $p < 0.05$ (FDR correction). As in the first ROI analysis we thresholded rSPI(Z)s corresponding to all regularization parameter values and all classification models according to $p < 0.05$ (FDR correction) and counted the number of significant voxels within the ROI defined from the pr-maximizing model. Here we also recorded the number of significant voxels outside the ROI. Figure 5.10 shows the analysis of the signal detection for FDA as a function of the regularization parameter λ in both data sets (similar results were obtained for SVM and LogReg). The pr-maximizing points were found at $\lambda = 2^{14}$ (finger tapping) and $\lambda = 2^{12}$ (trailsAB). Panel (A) shows the prediction accuracy and pattern reproducibility as a function of the value of λ in the finger tapping data set (a corresponding plot for the trailsAB data set is provided in Figure 5.3). As seen in Figure 5.10 there is a transition in prediction accuracy from high to low when λ increases, whereas pattern reproducibility increases with increasing λ. Panel (B) shows the number of voxels in the thresholded rSPI(Z)s (FDR correction) within the four known motor network regions defined from brain atlases. In general the number of voxels identified within the four regions increases with λ and reaches a plateau at large values of the regularization parameter. A peak in the number of voxels detected for the SC region is observed around the pr-maximizing point. With low values of λ we observe a dramatic decrease in the number of identified voxels within the regions. In particular the models with low λ (and high prediction accuracy) completely fail to detect voxels in the SC and S2 regions. This effect is also directly observed from the average rSPI(Z)

corresponding to the P model in Figure 5.7. Figure 5.10 panel (C) shows the number of voxels in the thresholded $\overline{\text{rSPI(Z)}}$s within a region defined by the thresholded $\overline{\text{rSPI(Z)}}$ corresponding to the pr-maximizing model in Figure 5.6. When λ decreases from the pr-maximizing value we observe a rapid decrease in the relative voxel detection. With increasing λ values from the pr-point there is a slight increase in the number of significant voxels outside the region defined by the pr-maximizing model. Panel (D) shows a similar but stronger trend for the trailsAB data set. Only relatively few voxels are detected for low λ values. When λ increases there is an increase in the number of significant voxels outside the region based on the pr-maximizing model. The maximum in the total number of suprathreshold voxels (sum of the two curves in panel (D)) is found at 2^{14} which corresponds to the R point on the pr-curve in Figure 5.6.

Figure 5.11 - 5.14 show pr-curves based on prediction accuracies and pattern reproducibilities for the three models (FDA, SVM, LogReg) for two subjects in the object recognition data set (similar results were obtained for the other subjects). The different plots show pr-curves for various levels of spatial smoothing. The relationship between regularization and the performance metrics was as observed in the finger tapping data set and in the trailsAB data set. I.e. low reproducibility at low levels of regularization, and increasing reproducibility at increasing regularization. Hence, when regularization increases, one moves from left towards right on the pr-curves. In general the pr-curves for each classifier type tend to follow the same path. At increasing degrees of smoothing and low levels of regularization the FDA and LogReg models show decreased reproducibility in comparison to the SVM. However, note that the FDA and LogReg models have at least the same performance as the SVM at increased levels of regularization. For subject 1 and classification task (bottle) vs. (face) (Figure 5.11) the reproducibility increases with an increasing degree of smoothing with best performance at 6 mm FWHM. Both prediction accuracy and reproducibility decrease with larger degrees of smoothing (9-15 mm FWHM). The best combined performance is observed at 6 mm FWHM. For subject 2 and classification task (bottle) vs. (face) (Figure 5.13) both the prediction accuracy and the reproducibility increase with an increasing degree of smoothing with best combined performance at 6 mm FWHM. Both performance measures decrease with larger degrees of smoothing (12-15 mm FWHM). For subject 1 and classification task (face) vs. (house) (Figure 5.12) we observe relatively high performances with respect to prediction accuracy and also reproducibility. The reproducibility increases with increasing degrees of smoothing, whereas prediction accuracy decreases with larger degrees of smoothing. The best combined performance is observed at 6 mm FWHM. For subject 2 and classification task (face) vs. (house) (Figure 5.14) we observe relatively high performance with respect to prediction accuracy for all levels of smoothing. The reproducibility metric increases with increasing degrees of smoothing without decreases in prediction accuracy as observed of the classification task (bottle) vs. (face) (Figure

5.13). The best combined performance is observed at 9 mm FWHM. Figures 5.15 and 5.16 show consensus analyses of the reproducible brain maps rSPI(Z)s extracted from models build on data of subject 1. Figure 5.15 is based on data without spatial smoothing, and Figure 5.16 is based on data smoothed with 6 mm FWHM. As for the trailsAB data set and the finger tapping data set (Figure 5.9) we observe a large degree of consensus across the classifiers. I.e. there is a strong correlation between brain maps extracted from models selected to maximize e.g. prediction accuracy and pattern reproducibility jointly (PR). Note that the models showing the least degree of consensus, across classifier type, are models chosen to maximize prediction (P). This is consistent with the observation from the two other data sets in Figure 5.9. The significance of the observations in Figures 5.11 - 5.16 is that the models' behavior observed in the trailsAB data set and in the finger tapping data set translates to the object recognition data set. Hence, the results generalizes to a data set with potentially more subtle pattern differences supporting discriminative information to the classifiers. Additionally, the shapes of the pr-curves, as observed in the finger tapping data set and the trailsAB data set, are also observed in versions of the object recognition data without using spatial smoothing as part of the preprocessing chain[2].

Figure 5.17 provides results of the analysis where sparsity is enforced on the model structure in the analysis of the finger tapping data set. The top panels (I)-(IV) show model performance over the parameter grid for LogReg with the ENET penalty. Multiple regularization parameter combinations lead to the same maximum model performance in terms of prediction accuracy (99.7%), but many of these tend to have low reliability demonstrated by low overlap (II) and mutual information (III). For further analysis we traced out the behavior of the three model performance metrics for a fixed λ_2 value and plotted these against the average number of voxels retained in the model. The extracted models are marked with rectangles in (I)-(IV). The λ_2 value was chosen so that the path contains the maximum prediction accuracy (top panel (I)) and also provides a relative high degree of overlap (top panel (II)). Note that it is not possible to simultaneously obtain maximum prediction accuracy, maximum overlap and even moderate levels of mutual information. From the plots in the middle row we observe maximum average prediction accuracy with 824 voxels retained in the model on average, maximum in overlap (corrected) found at 824 voxels, and maximum mutual information was with all voxels included in the model. *Corrected* overlap means that the overlap is corrected for the overlap that one would expect at chance (see Rasmussen et al. (2012b) for a description of this correction). For the SVM with RFE (Figure 5.17 bottom panel) maximum in prediction accuracy (99.7%) was with 1043 voxels retained, maximum in overlap

[2]Note that some degree of smoothing is introduced by the re-slicing step in the motion correction as part of data preprocessing (Haxby et al., 2001; Kamitani and Sawahata, 2010).

(corrected) was observed with 2039 voxels retained in the model, and maximum mutual information was with all voxels included in the model. Based on the curves for prediction accuracy we extracted maps from the models with best prediction performance (A, C, marked with crosses in the plots of prediction accuracy), and also a sub-optimal model (with respect to prediction accuracy) for the LogReg ENET model (marked with B). The brain maps in Figure 5.17 show the relative fraction of times each voxel was selected by the classifiers. These are based on 100 models (50 resampling splits with two models from each split). The brain slices in the top panel corresponding to the (A) model show that voxels in SMC and cerebellum are selected with high consistency but nothing else, thereby ignoring large sections of the known motor network subserving finger tapping. The brain slices in the top panel corresponding to the sub-optimal model (B) provides a brain pattern that is more similar to Figure 5.7 and also selects voxels in SMA, subcortical regions, and S2 with high consistency. The brain slices corresponding to the (C) model of SVM RFE (bottom panel) appears to be intermediate between A and B with voxels in SMC and cerebellum with high consistency, and a few additional, weakly consistent areas, e.g. subcortical regions.

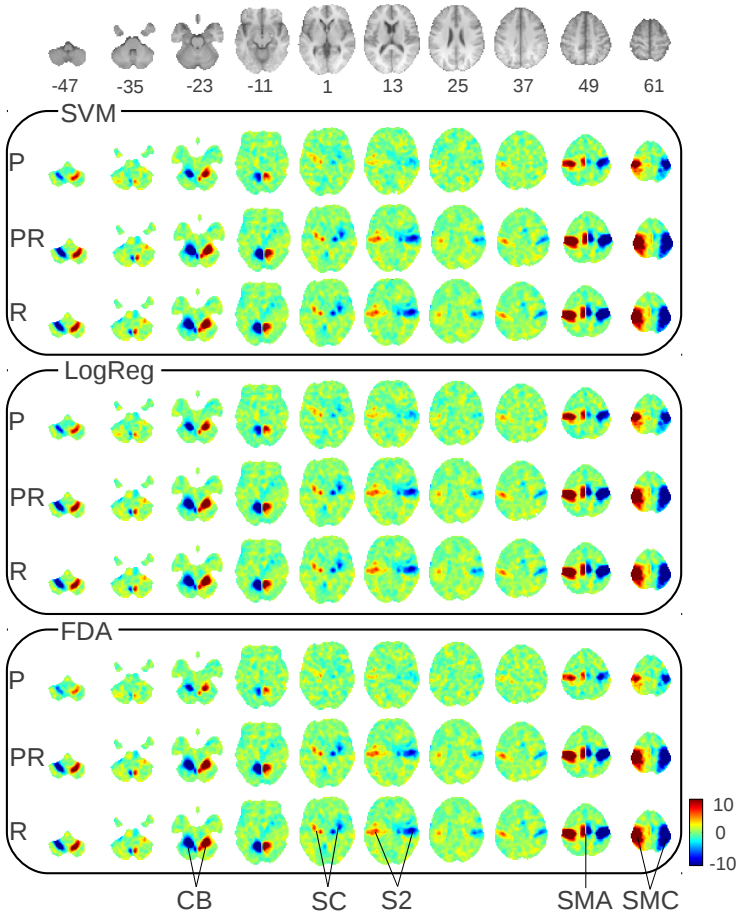

Figure 5.7: Finger tapping data set. Classification of (left) against (right). Results are based on 50 NPAIRS resampling splits. For each classifier type, P, PR, and R correspond to optimization of prediction accuracy, joint optimization of prediction accuracy and reproducibility, and optimization of reproducibility respectively. Shown are the Z-score reproducible $\overline{\text{rSPI(Z)}}$s. The top row shows an average anatomical scan of the 14 subjects included in the analysis (masked to only show voxels included in the analysis). Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention (right side of a brain slice is the right side of the brain). The lines drawn on the bottom row of slices indicate relevant brain regions: cerebellum (CB), subcortical areas -including caudate, thalamus, putamen (SC), second somatosensory cortex (S2), supplementary motor area (SMA), and sensorimotor cortex (SMC). Note that the regions extend throughout several slices (not marked).

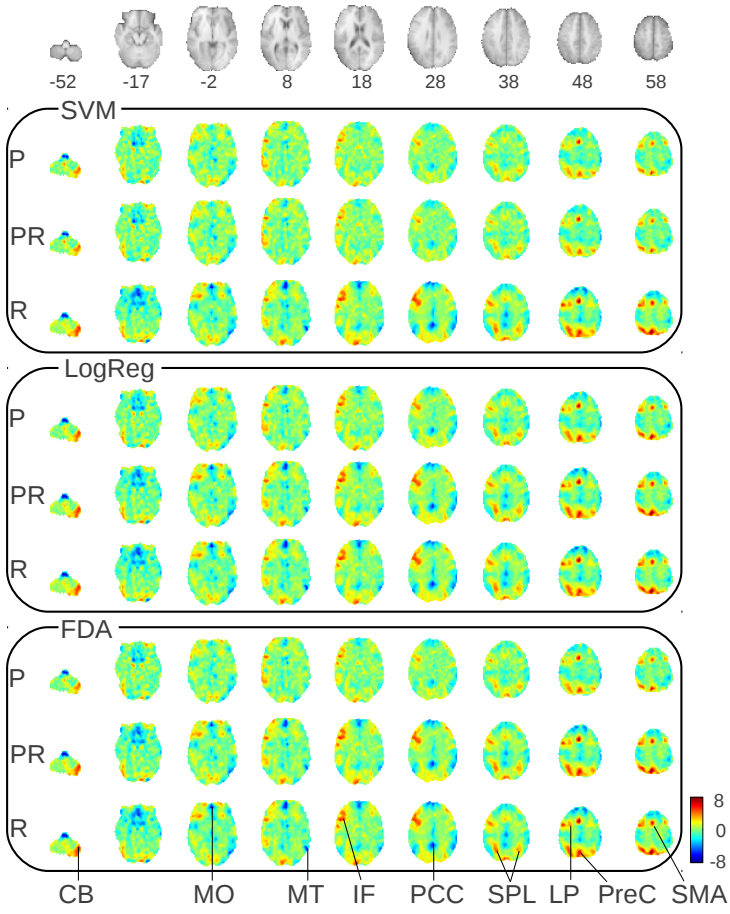

Figure 5.8: TrailsAB data set. Classification of (Trials A) against (Trails B). Results are based on 50 NPAIRS resampling splits. For each classifier type, P, PR, and R correspond to optimization of prediction accuracy, joint optimization of prediction accuracy and reproducibility, and optimization of reproducibility respectively. Shown are the Z-score reproducible $\overline{\text{rSPI(Z)}}$s. The top row shows an average anatomical scan of the 14 subjects included in the analysis. Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention (right side of a brain slice is the right side of the brain). The lines drawn on the bottom row of slices indicate the regions: cerebellum (CB), medial orbitofrontal gyrus (MO), middle temporal lope (MT), inferior-frontal gyrus (IF), posterior cingulate cortex (PCC), superior parietal lobes (SPL), left precentral gyrus (LP), precuneus (PreC), supplementary motor area (SMA).

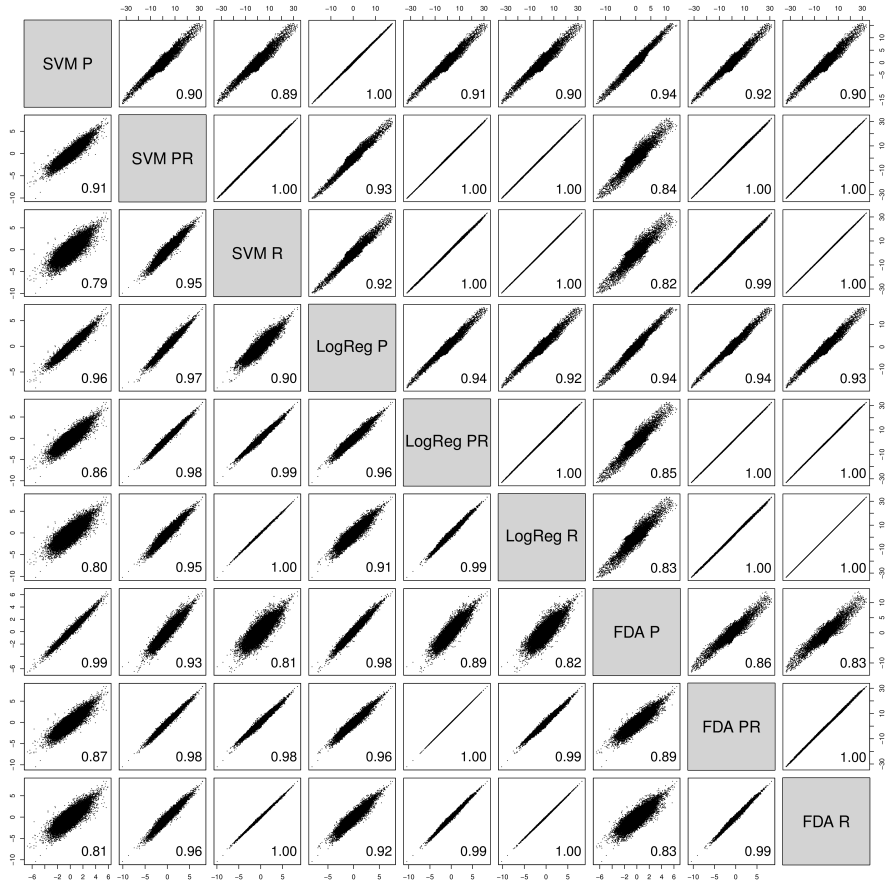

Figure 5.9: Consensus analysis of average reproducible brain maps ($\overline{\text{rSPI(Z)}}$s) across classifier types and models. For each classifier type, P, PR, and R correspond to optimization of prediction accuracy, joint optimization of prediction accuracy and reproducibility, and optimization of reproducibility respectively. Each point in the plots corresponds to a voxel. Upper-diagonal plots are the finger tapping data set, while plots below the diagonal are based on the trailsAB data set. The Pearson's product correlation coefficient in each plot summarizes the scatter cloud.

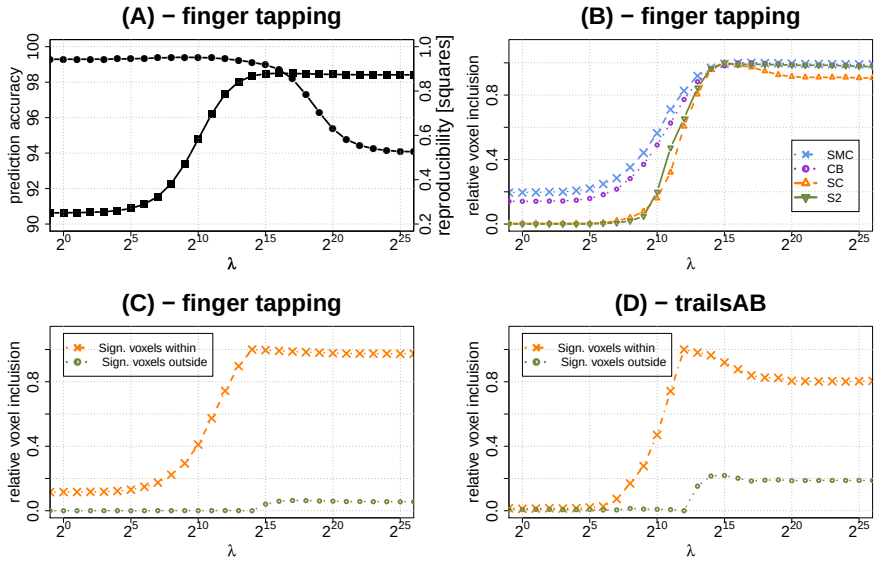

Figure 5.10: Analysis of signal detection by Fisher's discriminant analysis (FDA) as a function of the regularization parameter λ in the finger tapping (panel A-C) and trailsAB data sets (panel D). **Panel (A);** The prediction accuracy and pattern reproducibility in FDA as a function of the value of λ in the finger tapping data set (a corresponding plot for the trailsAB data set is provided in Figure 5.3). **Panel (B);** The number of voxels in the thresholded $\overline{\text{rSPI(Z)}}$s (FDR correction) within four regions defined from brain atlases. **Panel (C-D);** The number of voxels in the thresholded $\overline{\text{rSPI(Z)}}$s within a region defined by the thresholded $\overline{\text{rSPI(Z)}}$ corresponding to the pr-maximizing model in Figure 5.6. The number of suprathreshold voxels outside the region defined by the pr-maximizing model is also plotted. The number of voxels in panel (B-D) is relative to the maximum number of voxels included across the entire λ range.

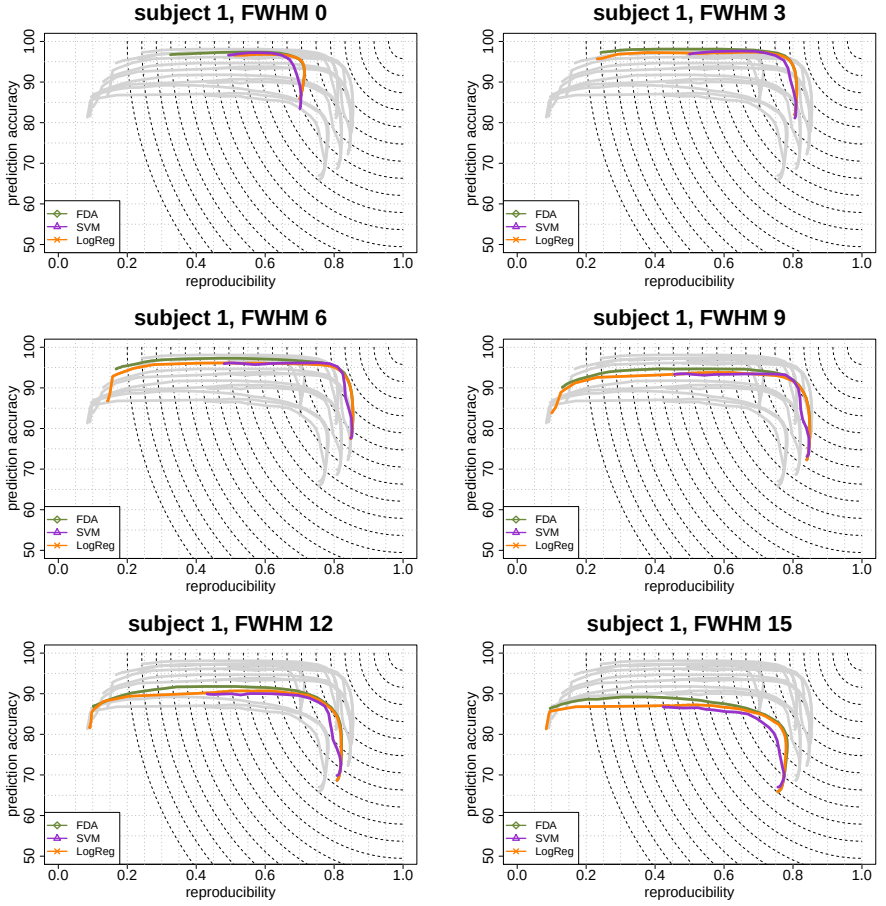

Figure 5.11: Object recognition data set, classification task (bottle) vs. (face), subject 1. Prediction/reproducibility curves (pr-curves) for the three classifiers for various degrees of spatial filtering. The width of the Gaussian smoothing kernel was varied as $\{0, 3, 6, 9, 12, 15\}$ mm full width half maximum (FWHM). The curves were constructed by changing the regularization parameter in the models and show the mean of 50 NPAIRS resampling splits. The gray curves show pr-curves for all classifiers at all smoothing levels, while the colored curves highlight the pr-curves at particular degrees of smoothing. Isolines indicate distances to the point $(p, r) = (100, 1)$.

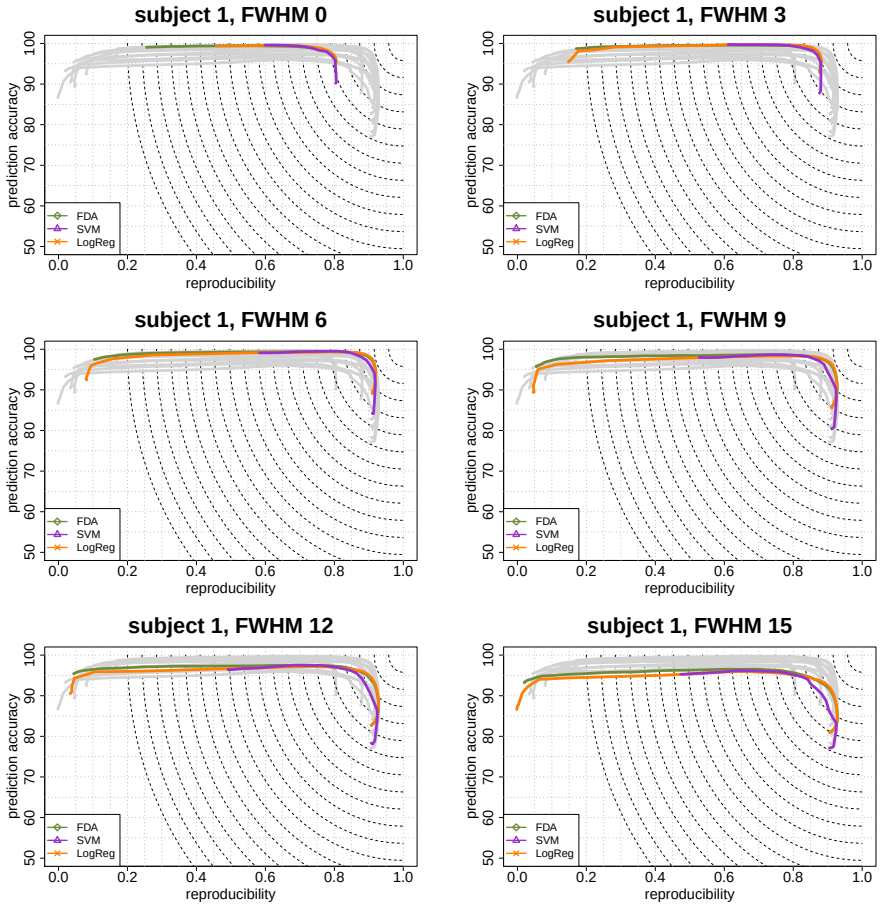

Figure 5.12: Object recognition data set, classification task (face) vs. (house), subject 1. Prediction/reproducibility curves (pr-curves) for the three classifiers for various degrees of spatial filtering. The width of the Gaussian smoothing kernel was varied as $\{0, 3, 6, 9, 12, 15\}$ mm full width half maximum (FWHM). The curves were constructed by changing the regularization parameter in the models and show the mean of 50 NPAIRS resampling splits. The gray curves show pr-curves for all classifiers at all smoothing levels, while the colored curves highlight the pr-curves at particular degrees of smoothing. Isolines indicate distances to the point $(p, r) = (100, 1)$.

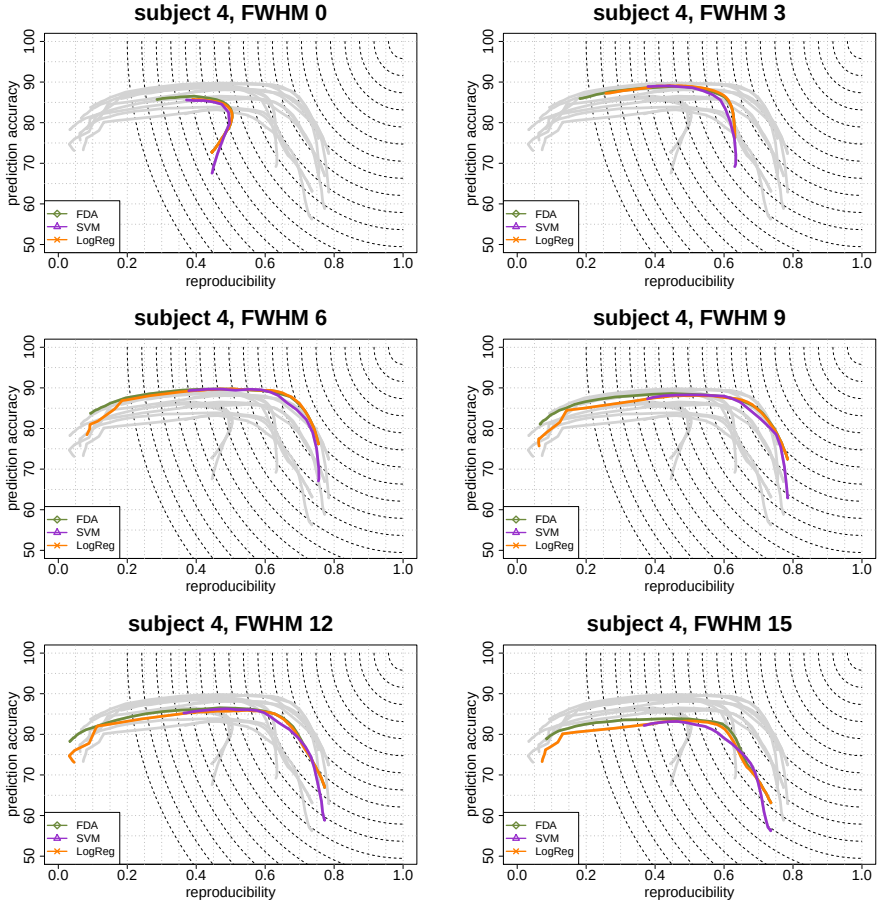

Figure 5.13: Object recognition data set, classification task (bottle) vs. (face), subject 4. Prediction/reproducibility curves (pr-curves) for the three classifiers for various degrees of spatial filtering. The width of the Gaussian smoothing kernel was varied as $\{0, 3, 6, 9, 12, 15\}$ mm full width half maximum (FWHM). The curves were constructed by changing the regularization parameter in the models and show the mean of 50 NPAIRS resampling splits. The gray curves show pr-curves for all classifiers at all smoothing levels, while the colored curves highlight the pr-curves at particular degrees of smoothing. Isolines indicate distances to the point $(p, r) = (100, 1)$.

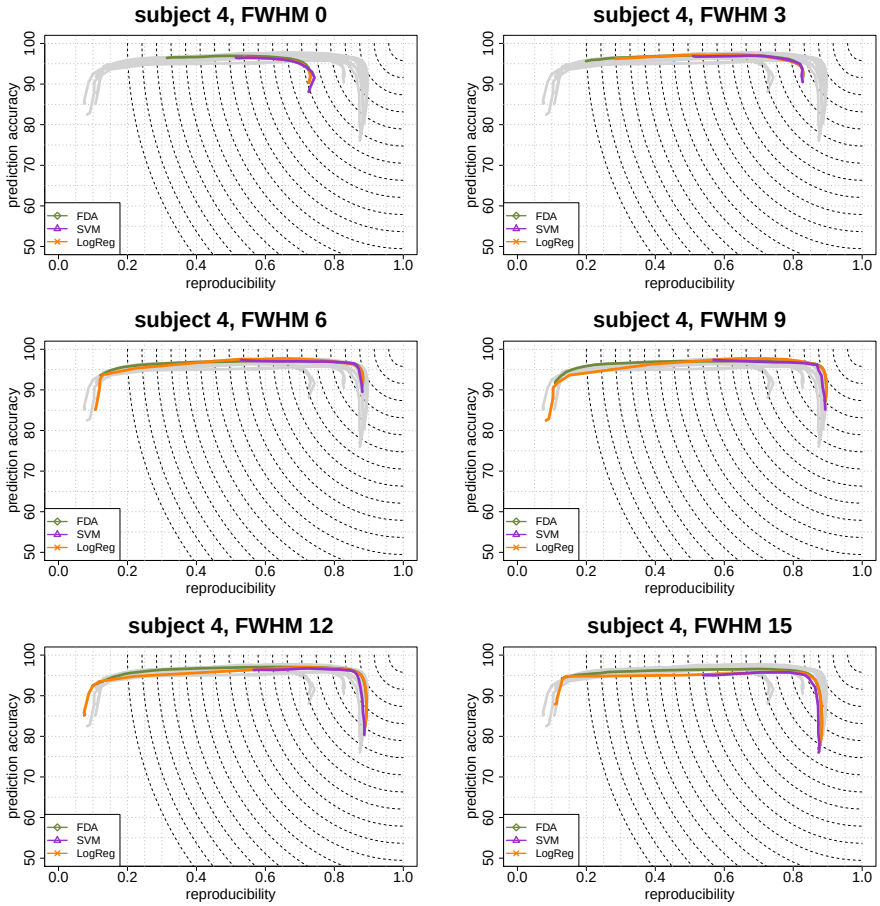

Figure 5.14: Object recognition data set, classification task (face) vs. (house), subject 4. Prediction/reproducibility curves (pr-curves) for the three classifiers for various degrees of spatial filtering. The width of the Gaussian smoothing kernel was varied as $\{0, 3, 6, 9, 12, 15\}$ mm full width half maximum (FWHM). The curves were constructed by changing the regularization parameter in the models and show the mean of 50 NPAIRS resampling splits. The gray curves show pr-curves for all classifiers at all smoothing levels, while the colored curves highlight the pr-curves at particular degrees of smoothing. Isolines indicate distances to the point $(p, r) = (100, 1)$.

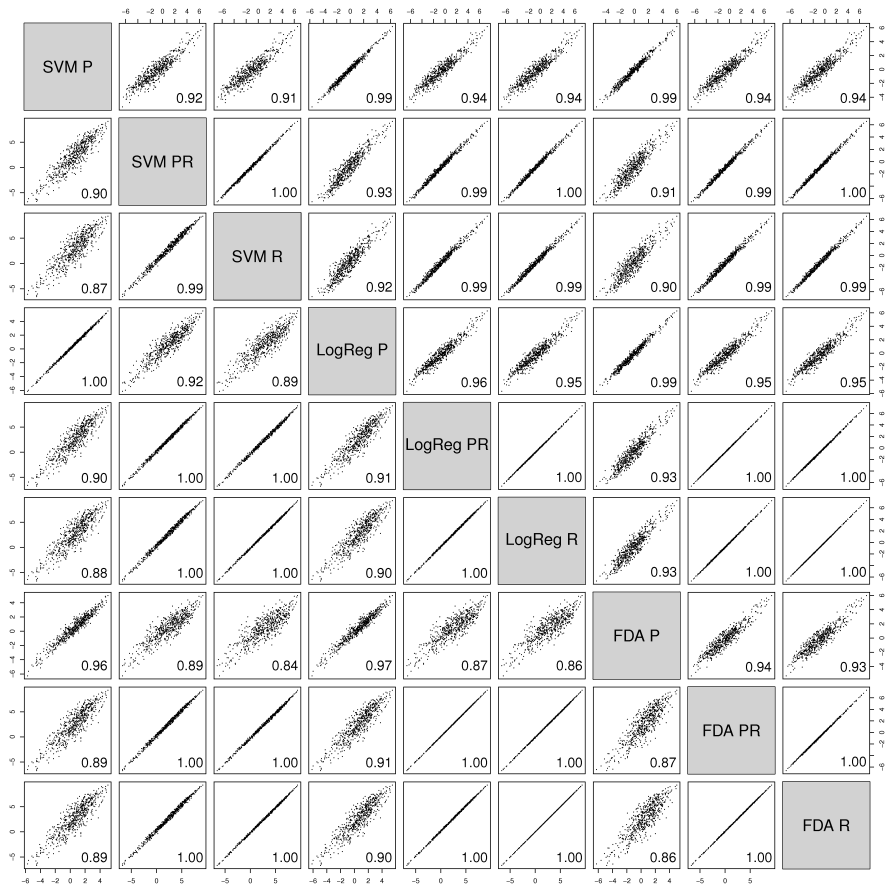

Figure 5.15: Object recognition data set, subject 1, no spatial filtering of the data. Consensus analysis of average reproducible brain maps ($\overline{\text{rSPI(Z)}}$s) across classifier types and models. For each classifier type, P, PR, and R correspond to optimization of prediction accuracy, joint optimization of prediction accuracy and reproducibility, and optimization of reproducibility respectively. Each point in the plots corresponds to a voxel. Upper-diagonal plots are the (bottle) vs. (face) classification task, while plots below the diagonal are based on the (face) vs. (house) classification task. The Pearson's product correlation coefficient in each plot summarizes the scatter cloud.

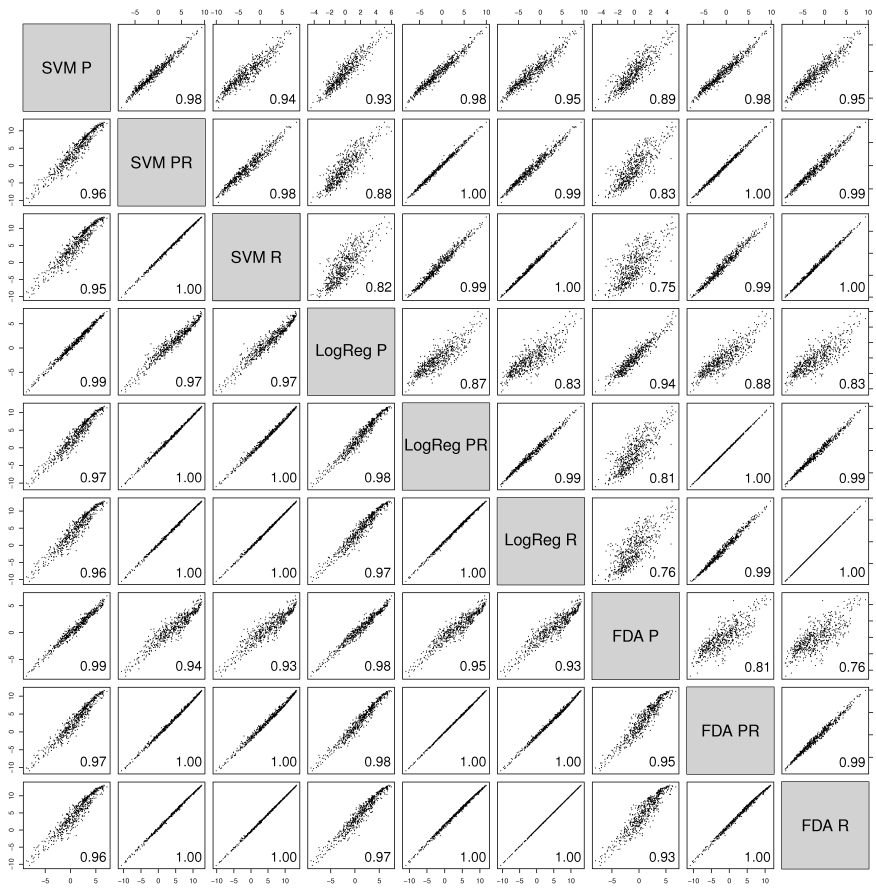

Figure 5.16: Object recognition data set, subject 1, spatial filtering of the data with a 6 mm full width half maximum Gaussian kernel. Consensus analysis of average reproducible brain maps ($\overline{\text{rSPI(Z)}}$s) across classifier types and models. For each classifier type, P, PR, and R correspond to optimization of prediction accuracy, joint optimization of prediction accuracy and reproducibility, and optimization of reproducibility respectively. Each point in the plots corresponds to a voxel. Upper-diagonal plots are the (bottle) vs. (face) classification task, while plots below the diagonal are based on the (face) vs. (house) classification task. The Pearson's product correlation coefficient in each plot summarizes the scatter cloud.

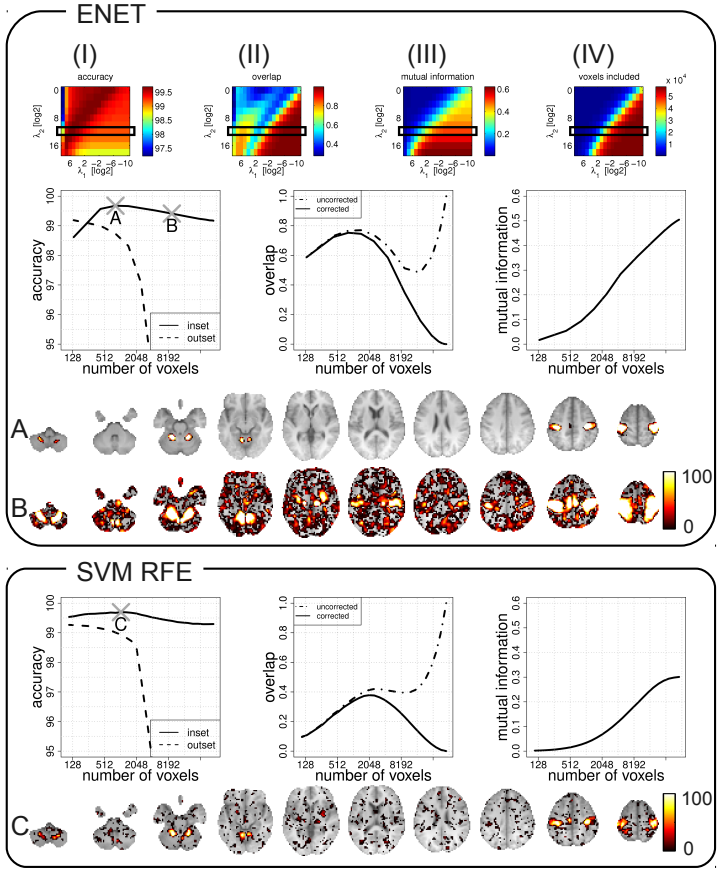

Figure 5.17: Finger tapping data set. Effect of imposing sparsity on the model structure with the ENET penalty in LogReg and SVM based RFE. Models are evaluated with split-half resampling with 50 resampling splits. As model performance metrics we used prediction accuracy on the voxels retained in the model (inset), prediction accuracy on the voxels excluded from the model (outset), mutual information, and overlap between non-zero voxels. *Corrected* overlap means that the overlap is corrected for the overlap that one would expect at chance (see Rasmussen et al. (2012b) for a description of this correction). **Top panel;** The top row shows performance of the LogReg model over the regularization parameter grid. The middle row shows model performance as a function of voxels included in the models for a fixed λ_2 marked by rectangles in the top row. The crosses A and B indicate an optimal and sub-optimal model with respect to prediction accuracy respectively. The brains slices show for model A and B the relative percentages how often each voxel was included in the LogReg model across the resampling splits. **Bottom panel;** The top row shows the performance metrics for the SVM as a function of voxels included in the model. The cross C in the plot of prediction accuracy marks the accuracy maximizing model. The relative voxels inclusion for model C is visualized in the bottom row.

5.2 Global model visualization by sensitivity maps

Investigations of the applicability of the sensitivity mapping procedure for visualization of nonlinear kernel models were reported in Rasmussen et al. (2011) and Rasmussen et al. (2012c). The analyses were based on the xor data set. Additionally, we here present a procedure for visualization of KPCA build by use of a nonlinear kernel.

5.2.1 Analysis setup

Three classification tasks were formulated based on the scan block labeling as follows.

- **Classification task I**: Scans for condition (left) were assigned to class -1 while (right) was assigned to class 1. We expected this classification task to be relatively simple for linear (and also nonlinear) methods to solve.
- **Classification task II**: Scans were grouped so that (no, both) were assigned to class -1 and (left, right) were assigned to class 1. This task was deliberately formulated to be harder for linear methods to solve and possible relatively easy for nonlinear methods to solve.
- **Classification task III**: Here we considered a four class classification task by partitioning the scans according to the block labels: no, left, right, and both. This classification task was similarly to Classification task I expected to be relatively easy for a linear classifier to solve.

Subjects were considered as the basic resampling unit, hence we build the models on a subset of the subject and test the models on the out-of-sample subjects. The split-half NPAIRS resampling strategy was used in order to evaluate the models both in terms of prediction accuracy and pattern reproducibility as described in Section 3.8. I.e. each split-half consisted of scans from three subjects. 10 NPAIRS resampling splits was performed (all possible combinations of subjects).

To underline the generality of the nonlinear modeling and the sensitivity mapping visualization procedure for model visualization we used the support vector machine (SVM), kernel logistic regression (KLR), and kernel Fisher's discriminant analysis (KFDA) models for the binary Classification tasks I and II. An illustration of signed sensitivity maps was based on the KFDA model in Classification tasks II and III. All models have a regularization parameter that needs to be selected (λ for KLR and KFDA and $C = 1/\lambda$ for the SVM). Additionally, use

of the Gaussian kernel also requires selection of the parameter q that controls the kernel width. Selection of model parameters was performed within the NPAIRS resampling framework by measuring prediction accuracy (p) and visualization reproducibility (r) for each parameter combination. Model selection was based on minimization of the Euclidean distance from the point $(p/100\%, r)$ to (1,1). In the following we refer to such models as pr-maximizing models. Additionally, we derived average reproducible brain images ($\overline{\text{rSPI}}$s) as described in Section 3.8.

Visualization of binary classifiers with sensitivity maps

Both the linear kernel and the Gaussian kernel were considered. The visualization of models build with a linear kernel was based on the squared weight map. The grand average sensitivity map $\mathbf{s}_2^{\text{gm}}$ defined as in eq. (3.60) served as a visualization of models build with the Gaussian kernel in order to visualize the classifiers based on a single map. The classifier output eq. (3.41) was used as the visualization function $g_c(\mathbf{x})$ in eq. (3.45). The map was based on squaring the individual sensitivity contributions ($k = 2$ in the definition eq. (3.45)). Squaring was done in order to avoid possible cancellation effects.

Visualization of classifiers with signed sensitivity maps

An illustration of the use of signed sensitivity maps ($k = 1$ in the definition eq. (3.45)) was done as follows. The classifiers outputs eq. (3.32) were used as visualization functions. In classification task III (four class task) we derived an overall visualization of the trained classifier by means of the grand mean sensitivity map $\mathbf{s}_2^{\text{gm}}$ eq. (3.60) based on squaring of individual sensitivity contributions ($k = 2$ in eq. (3.60)). Additionally, signed interclass contrast sensitivity maps $\mathbf{s}_1^{\text{c}|\text{c}'}$ were derived as in eq. (3.61) with $k = 1$ (no squaring of individual sensitivities) in order to interpret the classifier in terms of brain maps with sign information. In Classification task II we expected a relatively large heterogeneity between single sensitivities of observations within the same class. Note that the scans were grouped as (no, both) and (left, right) . First, interclass contrast sensitivity maps $\mathbf{s}_2^{\text{c}|\text{c}'}$ were constructed with $k = 2$ in eq. (3.61). Hence, the individual sensitivity contributions were squared in order to avoid cancellation effects. Second, we constructed signed interclass contrast sensitivity maps by adopting the weighted mapping procedure in eq. (3.62). Specifically, each of the interclass contrast maps $\mathbf{s}_2^{\text{c}|\text{c}'}$ was refined as follows: (1) Single sensitivities $\mathbf{s} = \partial g(\mathbf{x})/\partial \mathbf{x}|_{\mathbf{x}=\mathbf{x}_n}$ were calculated for all $n \in \mathcal{I}_{c'}$. (2) A feature vector $\mathbf{f}$ was constructed for each observation by stacking the single sensitivity $\mathbf{s}$ and the observation $\mathbf{x}$ itself $\mathbf{f} = [\mathbf{x}; \mathbf{s}]$ ($\mathbf{x}$ and $\mathbf{s}$ were both scaled to unit norm to put them on the somewhat same scale), so that $\mathbf{f}$ was a 2P-dimensional vector. (3) Principal component analysis (PCA) was performed on the feature vectors $\mathbf{f}$, and the feature vectors were projected onto the PCA subspace spanned by the

first two components[3]. (4) Based on the low dimensional feature representation, we build a Gaussian mixture model (GMM). To estimate the number of components/clusters $M \in \{1, \ldots, 6\}$ we used nested cross validation where the GMM was trained on a subset of observations, and model generalizability was estimated by evaluating the GMM likelihood on the left out observations. (5) Steps (1-4) were repeated for all 10 NPAIRS splits and the number of components M was chosen according to maximization of the mean likelihood across the 10 splits. (6) Finally, a second pass through the NPAIRS resampling procedure was performed, where M was fixed across all split-halves and resampling runs in order to obtain the same number of clusters across all models build on individual splits of the data. The labels of clusters identified in individual split-halfs must be aligned across splits in order to derive $\overline{\text{rSPI}}$s. A simple reference filtering procedure was used, there the cluster's labels of a particular split was permuted in order to maximize the correlation between maps across splits. The outcome of the steps (1-6) was weighting factors w_n^m in eq. (3.62). These weighing factors were defined as the posterior probability of a particular observation $\mathbf{x}_n$ belonging to component m as $w_n^m = P\left(m|\mathbf{z}_{\mathbf{x}n}\right)$ for $m \in \{1, \ldots, M\}$. Hence, each of interclass contrast maps $\mathbf{s}_2^{\text{c}|\text{c}'}$ is refined into M maps $\mathbf{s}_1^{\text{c}|\text{c}',\text{cluster m}}$ with sign information.

Visualization of KPCA by sensitivity maps

In addition to visualization of supervised learning models, the sensitivity mapping visualization procedure was used to visualize a trained kernel principal component analysis (KPCA) model. Specifically, we here considered the projection function $\beta\left(\mathbf{x}\right)_j$ of KPCA eq. (3.66) as a visualization function $g_c\left(\mathbf{x}\right)$ in the sensitivity map definition in eq. (3.45). Individual sensitivities were squared ($k = 2$). Details on the visualization procedure are found in Section 3.7.3.

5.2.2 Results

Visualization of binary classifiers with sensitivity maps

Figure 5.18 shows model performance in Classification tasks I and II for the SVM with a Gaussian kernel as measured by prediction accuracy and pattern reproducibility over a range of parameter values. The width of the Gaussian kernel is reported relative to the average input-space distance measure of the Gaussian kernel to the nearest 25% points across all data points. For Classifi-

[3]The dimensionality of the PCA subspace was chosen heuristically. More principled approaches to the clustering is a topic for future research.

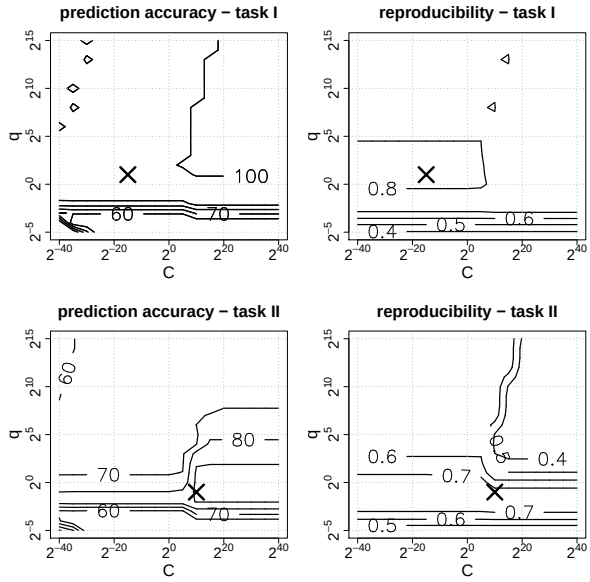

Figure 5.18: Parameter optimization grid for the support vector machine (SVM) with a Gaussian kernel. C is the 'complexity'/regularization parameter of the SVM, and q is the kernel width. Models were optimized against both prediction accuracy (p) and pattern reproducibility (4). The plots are based on mean values of 10 NPAIRS splits. The top row shows model performance in classification task I and the bottom row shows model performance in classification task II. The crosses indicate models selected according to minimization of the Euclidean distance from the point (p/100%,r) to (1,1).

cation task I (top row) we observe that for a fixed width of the Gaussian kernel q there is a tendency of an increased prediction accuracy at large values of C. Conversely, there is a tendency to increased reproducibility at lower values of C. Both metrics tends to increase with increasing width of the Gaussian kernel. Hence, there is a preference towards more linear models. For Classification task II (bottom row) we observe preference towards a relatively small kernel width as compared to that in Classification task I. For a particular value of the Gaussian kernel there is a tendency to increased prediction accuracy with large values of C and increased reproducibility with decreasing values of C.

Table 5.2 summarizes model performance for the three different classifiers with linear kernels and Gaussian kernels. For Classification task I it is observed that all models provide good performance both in terms of prediction accuracy and pattern reproducibility. For Classification task II there is a major reduction in

	Task I		Task II	
	prediction	reproducibility	prediction	reproducibility
SVM linear	100.0 ***	0.79 ***	67.8 ***	0.57 **
SVM Gaussian[a]	100.0 ***	0.79 ***	-	-
SVM Gaussian	99.7 ***	0.81 ***	92.2 ***	0.75 ***
KLR linear	100.0 ***	0.80 ***	57.4 **	0.57 **
KLR Gaussian[a]	100.0 ***	0.80 ***	-	-
KLR Gaussian	100.0 ***	0.81 ***	92.0 ***	0.76 ***
KFDA linear	100.0 ***	0.80 ***	57.4 **	0.57 **
KFDA Gaussian[a]	100.0 ***	0.80 ***	-	-
KFDA Gaussian	99.9 ***	0.81 ***	92.3 ***	0.75 ***

Table 5.2: Results for classification task I and II. For all models performances are reported for parameter settings that optimize both prediction accuracy (p) and reproducibility (r) jointly. The width of the Gaussian kernel was fixed at $q = 2^{15}$ in the models marked with [a]. The table reports mean values 10 NPAIRS splits. Significance codes; **: $p < 0.01$, ***:$p < 0.001$. Statistical significance is based on a nonparametric permutation test (5000 permutations).

both prediction accuracy and reproducibility for the linear models in comparison to the models' performances in Classification task I. Although also showing decreased performance in comparison to performance in Classification task II the nonlinear models were capable in maintaining relatively good performance with respect to both prediction accuracy and pattern reproducibility.

Figure 5.19 shows $\overline{\text{rSPI}}$s based on sensitivity maps derived from a trained SVM. Figure 5.20 shows the same maps but thresholded according to $p < 0.05$ FDR correction. The maps are calculated as in eq. (3.60) with squared individual sensitivities in order to avoid potential cancellation effects. Hence the maps contain only positive values. Figure 5.19 and 5.20 panel A-C show the $\overline{\text{rSPI}}$s based on models with a linear kernel, a Gaussian kernel with a large kernel width ($q = 2^{15}$), and a kernel width estimated according to pr-optimization respectively. The maps, hence the different models, tend to identify the same voxels as important to the classifiers decisions. In Classification task I both the linear and the nonlinear models are capable of using information in the visual cortex, for solving the classification task. In Classification task II, Figure 5.19 and 5.20 panel D-E, there is a large degree of discrepancy in which voxels that supports discriminative information to the models. The nonlinear model (panels D) tends to use the same voxels as all classifiers in Classification task I (panel A-C), while the linear model use information that are not in the primary visual areas. This is in particular seen in the thresholded map Figure 5.20(D). Interestingly, the linear model appears to identify relatively large regions of

voxels with high consistency as indicated by the relatively high value in the $\overline{\text{rSPI}}$ Figure 5.19(D). Also relatively large regions survive the statistical thresholding in 5.20(D).

Figure 5.21 provides a consensus analysis between the $\overline{\text{rSPI}}$s of all models in Classification tasks I and II. Model parameters were selected according to pr-optimization. First, we observe that within each kernel type and classification task there is a large degree of consensus across classifiers. This observation is consistent with the results reported in Section 5.1 that also show great similarities across classifier types. In Classification task I there is also a strong consensus between the pr-optimized linear and nonlinear models. For Classification task II there is less consensus between the linear classifiers and the classifiers of Classification task I. This is in contrast to the nonlinear models in Classification task II, that show larger similarities with all models in classification task I.

Visualization of classifiers with signed sensitivity maps

Figure 5.22 shows the results of Classification task III (four class classification task) obtained with the KFDA classifier. The average prediction accuracy was 92.3%. Figure 5.22(A) shows the $\overline{\text{rSPI}}$ based on the grand average maps $\mathbf{s}_2^{\text{ga}}$ derived from the model as in eq. (3.60) using squared individual sensitivities. The average reproducibility was 0.82. The map is thresholded according to $p < 0.05$ FDR correction. Primarily, voxels in the visual cortex are identified with consistency across the resampling splits as important to the classifier's decisions. The map has great similarities with the maps derived on models build in Classification tasks I and II in Figure 5.20. Figure 5.22(B) shows examples of signed interclass contrast sensitivity maps, see eq. (3.61). The maps are masked to show the same voxels as in Figure 5.22(A). The notation e.g. $\mathbf{s}_1^{\text{left}|\text{no}}$ means that the map indicates how scans belonging to the (no) class should be changed in order to move the scans towards regions of the input space where scans are being classified as (left). According to the map $\mathbf{s}_1^{\text{left}|\text{no}}$ a signal increase primarily in the right visual cortex will make the scans belonging to the (no) condition move towards the (left) class. Likewise, the map $\mathbf{s}_1^{\text{left}|\text{right}}$ indicates that lowering the signal in the left visual cortex and increasing signal in the right visual cortex will make scans belonging to the (right) condition move towards being classified as belonging to the (left) class. Finally, lowering the signal primarily in the left visual cortex will make scans belonging to the (both) condition move towards being classified as belonging to the (left) class as seen in the map $\mathbf{s}_1^{\text{left}|\text{both}}$. The signed interclass contrast maps have rather high reproducibilities indicating that possible cancellation effects may not be pathological.

Figure 5.23 shows both sensitivity maps and signed sensitivity maps derived form the KFDA model in classification task II: (no, both) vs. (left, right).

Figure 5.23(A) shows the grand average sensitivity map $\mathbf{s}_2^{\text{ga}}$ thresholded according to $p < 0.05$ FDR correction. The map has great similarities with that of the SVM in Figure 5.20. This similarity is also evident in the consensus analysis in Figure 5.21. Figure 5.23(B) shows interclass contrast sensitivity maps $\mathbf{s}_2^{\text{c}|\text{c}'}$ based on squared individual sensitivities. These maps also had a relatively high reproducibility ~ 0.74 across the NPAIRS splits. These map reproducibilities were reduced to ~ 0.22 if the interclass contrast maps were based on *signed* individual sensitivities (as in Figure 5.22). This reduction in reproducibility may be explained by the presence of cancellation effects. Figure 5.23(C) shows signed interclass contrast sensitivity maps. The GMM clustering procedure provided evidence towards presence of two clusters in each of the groups (no, both) and (left, right). Hence, the interclass contrast maps in Figure 5.23(B) were each decomposed into two maps with sign information. For example $\mathbf{s}_1^{\text{(no, both)}|\text{(left, right)},\text{cluster 1}}$ denotes that the signed sensitivity map is based on the output class (no, both), and that the sum in eq. (3.62) is calculated over the members of class (left, right). Furthermore, the contributions of the individual observations $\mathbf{x}_n$ to the signed sensitivity maps were weighted by the weighting factor w_n^1 being the posterior probability for observation $\mathbf{x}_n$ belonging to component 1 in the GMM. For the map $\mathbf{s}_1^{\text{(no, both)}|\text{(left, right)},\text{cluster 1}}$ we found that members of the condition (left) had an average weighting factor of ~ 0 in the sum while the members of the condition (right) and an average weighting factor of 0.9854. Hence, members of the condition (right) contribute the most to this map. Likewise, the members of condition (left) contributed the most to the map $\mathbf{s}_1^{\text{(no, both)}|\text{(left ,right)},\text{cluster 2}}$. For the map $\mathbf{s}_1^{\text{(left, right)}|\text{(no, both)},\text{cluster 1}}$ we found that the members of condition (no) had an average weighting factor of 0.0125 in the sum while members of the condition (both) and an average weighting factor of 1.00. Hence members of the condition (both) contribute the most to the map $\mathbf{s}_1^{\text{(left, right)}|\text{(no, both)},\text{cluster 1}}$. Likewise, the members of the condition (no) contribute the most to the map $\mathbf{s}_1^{\text{(left, right)}|\text{(no, both)},\text{cluster 2}}$. Note that the map reproducibilities are intermediate ~ 0.50 indicating that the cancellation effects have been mitigated to some extend by the weighting procedure.

Visualization of KPCA by sensitivity maps

Figure 5.24 shows an example of KPCA based on a single NPAIRS split in Classification task II. The width of the Gaussian kernel was the same as identified by pr-maximization in the analysis of Classification task II ($q = 0.5$), see Figure 5.18. Data from three subjects were used to estimate the KPCA basis, and data from the remaining three subjects served as test data. Figure 5.24 shows data observations' projections onto principal components $j \in \{1, 2, 5\}$. The shown projections were chosen to show a subspace in which the classes as defined in Classification task II (no, both) vs. (left, right) appear to be fairly linearly separable. Note that test points' projection appears to be more condensed than

training points' projections. This phenomenon has been referred to as *variance inflation*, see Abrahamsen and Hansen (2011a) for further information. Figure 5.25 provides a global interpretation of the data observations embeddings in Figure 5.24 by a sensitivity map estimated according to Section 3.7.3. The sensitivity map highlights, that changing the signal primarily in the visual cortex will lead to a change in the data observations' projections onto the three shown KCPA axes.

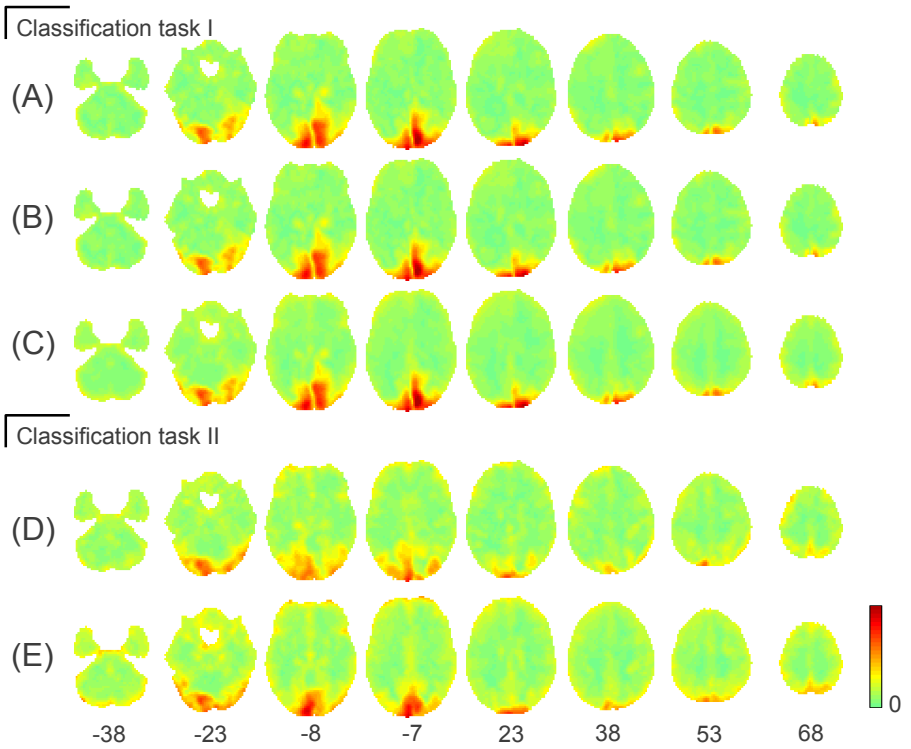

Figure 5.19: Interpretation of a trained support vector machine (SVM) by the sensitivity mapping visualization strategy. Classification task I is the (left) vs. (right) problem and Classification task II is the (no, both) vs (left, right) problem. The rows show reproducible brain images $\overline{\text{rSPI}}$s as estimated within the NPAIRS resampling framework. 10 resampling splits. **Panel A,D** show the $\overline{\text{rSPI}}$s based on an SVM with a linear kernel, where the model visualization was based on the squared model weights w^2. **Panel B,C,E** show the $\overline{\text{rSPI}}$s based on an SVM with a Gaussian kernel. The width of the Gaussian kernel was fixed at $q = 2^{15}$ in Panel B (approach linear classifier). The width of the Gaussian kernel in Panel C and E was chosen according to pr-maximization. Model visualization was based on the sensitivity map for SVMs with the Gaussian kernel. The sensitivity map express the relative importance of each voxel to the classifiers' decisions. The color bar indicates each voxels' value in the $\overline{\text{rSPI}}$ expressing the consistency in the sensitivity measure for individual voxels across the resampling splits. Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention (right side of a brain slice is the right side of the brain).

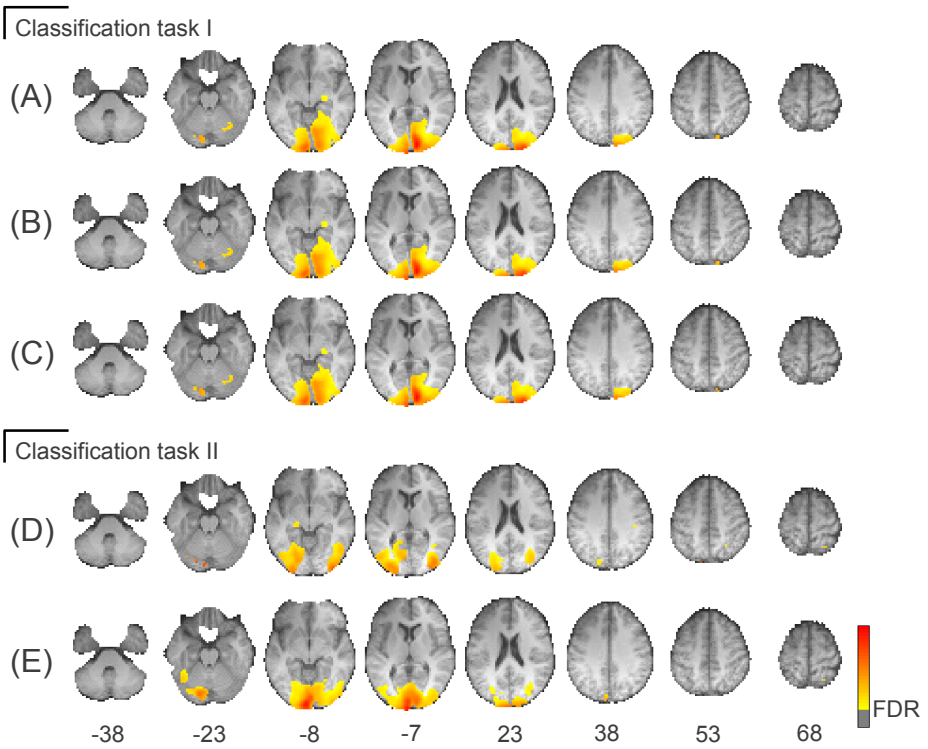

Figure 5.20: Same analysis as in Figure 5.19 and with reproducible brain images $\overline{\text{rSPI}}$ thresholded according to a non-parametric permutation analysis and correction for multiple comparisons using false discovery rate (FDR). A null distribution was build by permuting scan blocks' labels and retraining the classifiers. The $\overline{\text{rSPI}}$s are thresholded according to $p < 0.05$ FDR correction. 5000 permutations were conducted. The thresholded maps are projected onto an average anatomical scan of the six subjects.

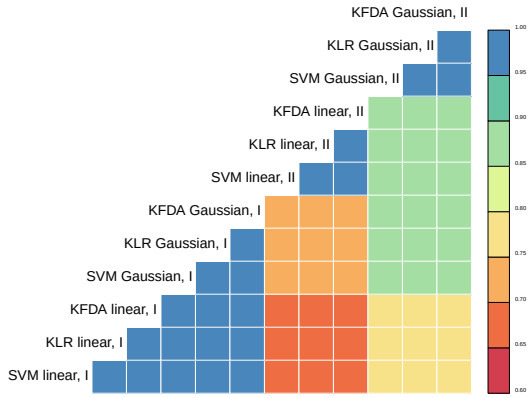

Figure 5.21: Across classifier consensus analysis. For each classifier we obtained an $\overline{\text{rSPI}}$ based on 10 NPAIRS splits. The plot shows the correlation of these brain maps across classification tasks and classification models. Models with code I and II are build on classification task I and II respectively.

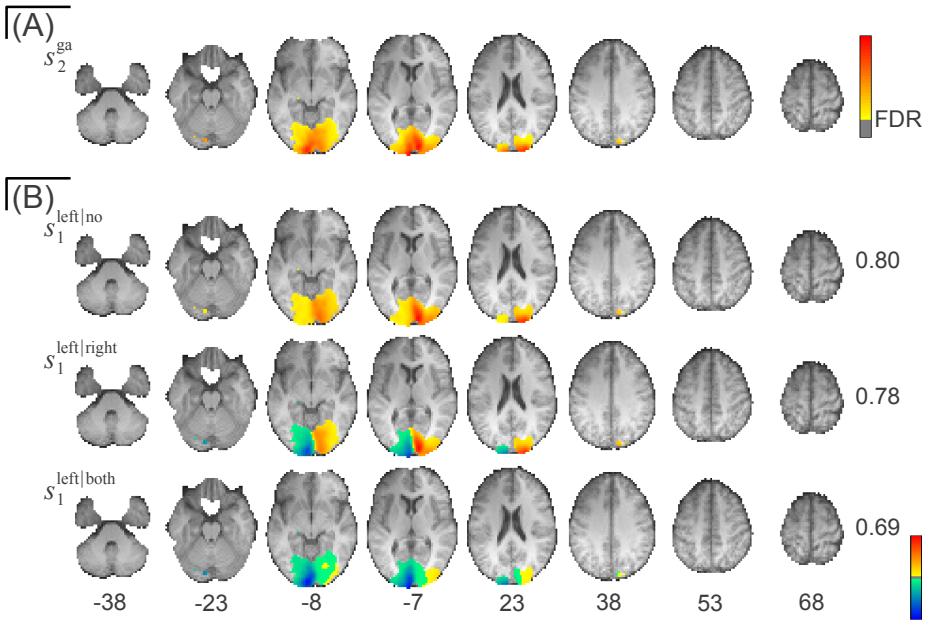

Figure 5.22: Interpretation of a trained classifier with *signed sensitivity maps*. The maps are extracted from a kernel Fisher's discriminant analysis (KFDA) with a Gaussian kernel in the four class Classification task III with classes (no, left, right, both). **Panel A** shows the reproducible brain image $\overline{\text{rSPI}}$ based on the grand average sensitivity map eq. (3.59) providing a model visualization by a single brain map. The average reproducibility of the sensitivity map was 0.82 as measured within the NPAIRS resampling framework. 10 resampling splits. The $\overline{\text{rSPI}}$ is thresholded at $p < 0.05$ according to false discovery rate (FDR) correction for multiple comparisons. **Panel B** shows $\overline{\text{rSPI}}$s based on the signed interclass contrast sensitivity maps eq. (3.61). Images in Panel B are masked with the same mask as in Panel A. Warm colors and cold colors are positive and negative values respectively. Numbers right to the slices report map reproducibilities.

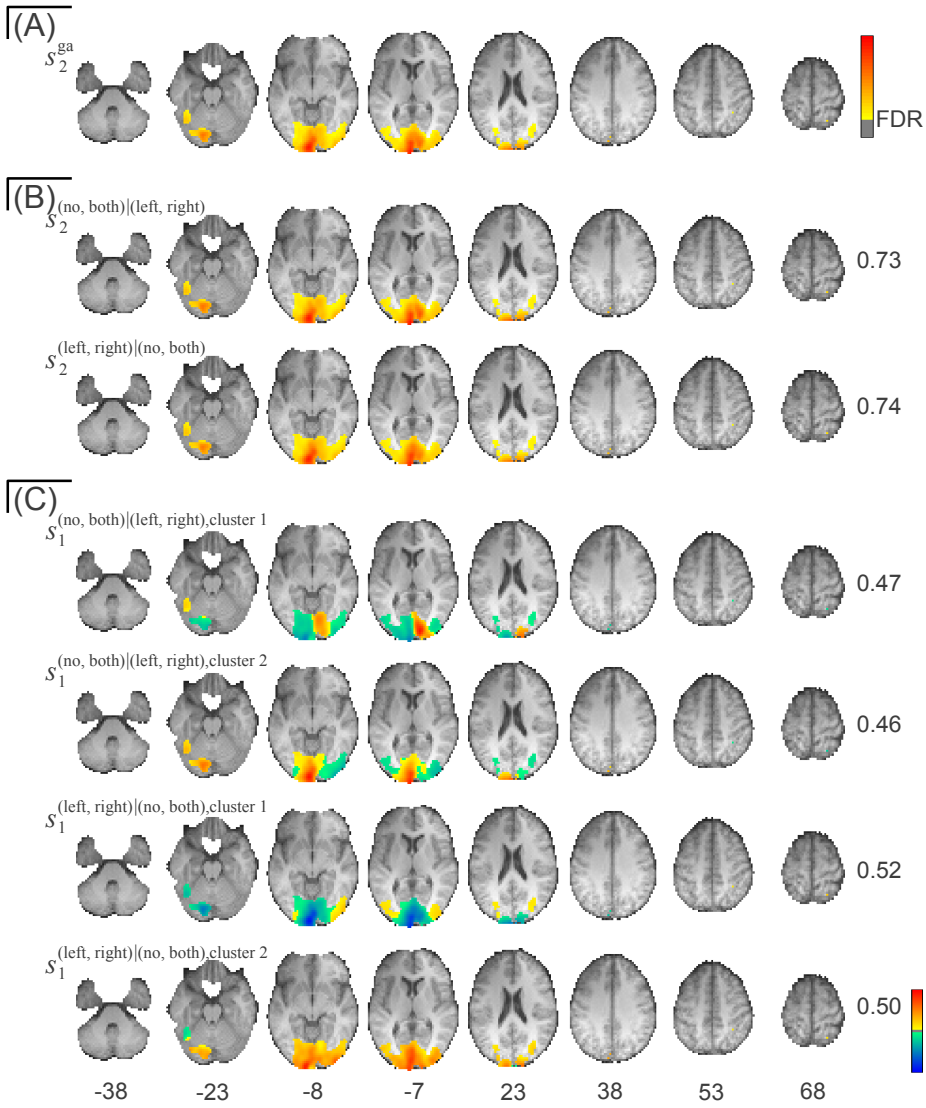

Figure 5.23: Interpretation of a trained classifier with *signed sensitivity maps*. The maps are extracted from a kernel Fisher's discriminant analysis (KFDA) with a Gaussian kernel in Classification task II, (no, both) vs. (left, right). **Panel A** shows the reproducible brain image $\overline{\text{rSPI}}$ based on the grand average sensitivity map eq. (3.59) providing a model visualization by a single brain map. The average reproducibility of the sensitivity map was 0.75 as measured within the NPAIRS resampling framework. 10 resampling splits. The $\overline{\text{rSPI}}$ is thresholded at $p < 0.05$ according to false discovery rate (FDR) correction for multiple comparisons. **Panel B** shows $\overline{\text{rSPI}}$s based on the interclass contrast sensitivity maps eq. (3.61) (with squared sensitivities). **Panel C** shows $\overline{\text{rSPI}}$s based on the weighted sensitivity map eq. (3.62). Each class is characterized by two clusters. Warm colors and cold colors are positive and negative values respectively. Numbers right to the slices report map reproducibilities.

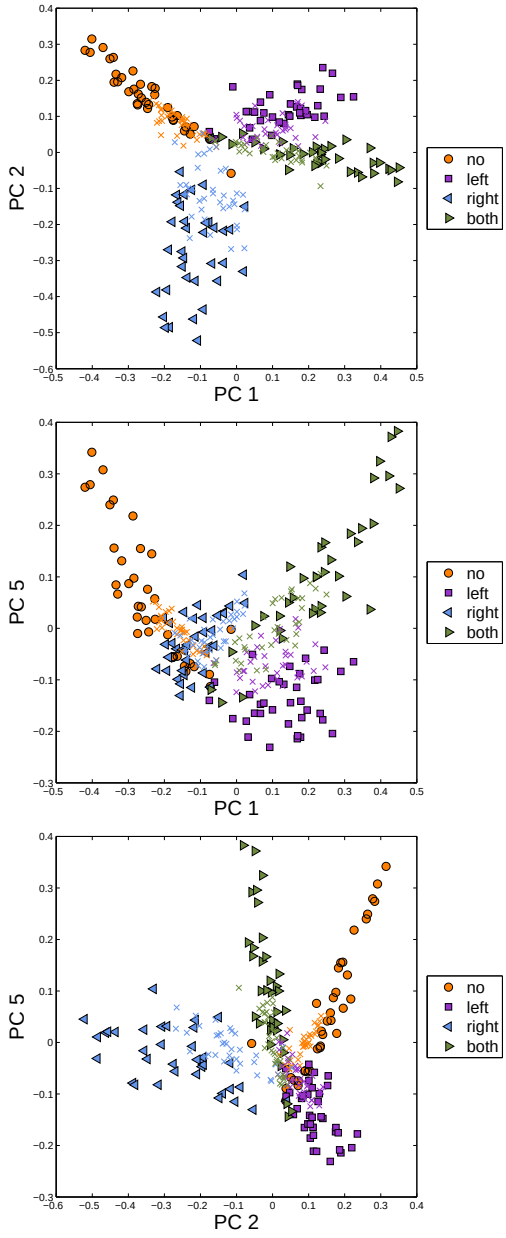

Figure 5.24: Kernel principal component analysis (KPCA) of the xor data set. The Gaussian kernel was used with the same width parameter as used in Classification task II ($q = 0.5$) (Figure 5.18 and Table 5.2). The plots show data points' projections onto principal components $j \in \{1, 2, 5\}$. The KPCA basis was estimated from three subjects, and the remaining three subjects was used as 'test data'. Filled markers are training points while the crosses mark test points.

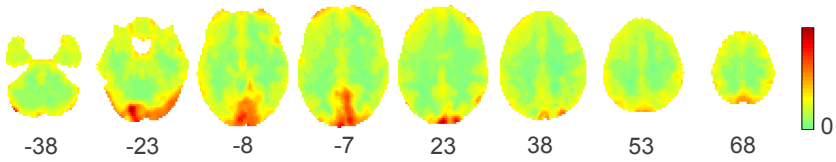

Figure 5.25: Global interpretation of a trained kernel principal component analysis (KPCA) model by the sensitivity mapping strategy. The visualization is based on feature space projections $\beta(\mathbf{x})_j$ acting as visualization functions in eq. (3.1). The sensitivity map is calculated over principal component $j \in \{1, 2, 5\}$ as shown in Figure 5.24.

5.3 Image denoising by kernel principal component analysis and pre-image estimation

An investigation of the applicability of kernel principal component analysis (KPCA) and pre-image estimation for image denoising was based on analysis of the finger tapping data set (28 subjects included in the analysis) and the object recognition data set (six subjects included in the analysis). The analysis of the object recognition data set was performed on the version of the data set without spatial smoothing as a preprocessing step (see Section 4.4). In Section 5.1 we observed that spatial smoothing could lead to increases in both prediction accuracy and pattern reproducibility. Our motivations for working on data without spatial smoothing here are: i) to keep the preprocessing minimal, ii) it may seem controversial to smooth data where a fine-grained signal structure is expected, iii) spatial smoothing was not applied in previous analyses of the object recognition data set, e.g. Haxby et al. (2001); Hanson et al. (2004). Note that, as any other preprocessing step, the KPCA denoising will interact with all other elements of the data processing pipeline (Strother et al., 2002). A comprehensive investigation of such interactions is a future research topic.

Results presented in this section have been reported in Rasmussen et al. (2012a)[4].

5.3.1 Analysis setup

Image denoising

Image denoising was performed by means of KPCA using a Gaussian kernel and pre-image estimation as described in Section 3.7.1 and 3.7.2. Pre-image estimation was based on the methods of Mika et al. (1999b) and Kwok and Tsang (2004) referred to as *Mika's method* and *Kwok and Tsang's method* respectively. Denoising requires selection of the width of the Gaussian kernel and the KPCA subspace size. For a particular parameter combination of the width, σ, of the Gaussian kernel and the dimensionality of the KPCA subspace, q, a denoised version, $\mathbf{Z}^{(\sigma,q)}$, of the original data observations $\mathbf{X}$ was determined. In both data sets we performed image denoising at the subject level. A KPCA basis was estimated from all scans of a particular subject, and the images were subsequently denoised by projecting the images onto the KPCA basis followed by pre-image estimation. For the finger tapping data set we created denoised versions of the

[4]Note that the notation in this section slightly differs from the previous two sections in order to comply with the notation used in Rasmussen et al. (2012a). In this section a prediction accuracy of 1 means perfect prediction (100% correct). Note also that σ and q denote kernel width and subspace dimensionality respectively.

original data set by varying the parameters over the grid $\sigma \in [2^{-3}, 2^{-2}, ..., 2^{10}]$ and $q \in [2, 4, 8, 16, 32, 50, 75, ..., 175, 240]$. For the object recognition data set we explored the grid $\sigma \in [2^{-3}, 2^{-2}, ..., 2^{10}]$ and $q \in [10, 20, ..., 250]$. The kernel width was scaled relative to the average input-space distance measure of the Gaussian kernel to the nearest 25% points across all data points. For Kwok and Tsang's reconstruction method we initially considered $k = \{5, 10, 15, 20, 50\}$ neighbors in the finger tapping data set and found no major impact on the model performance. In the finger tapping data set we report results by use of ten neighbors as suggested in Kwok and Tsang (2004). In the object recognition data set we found that a relatively large number of nearest neighbors was required to achieve stable model performance (see the Supplementary Materials of Rasmussen et al. (2012a) Figure 1). Hence, we report results based on 500 nearest neighbors for the object recognition data set. A relative change below 10^{-9} was used as a convergence criterion for Mika's method.

Evaluation of impact of image denoising

The impact of image denoising was assessed by means of multivariate classification model evaluated within the NPAIRS resampling framework (Section 3.8). Specifically, a Fisher's discriminant analysis (FDA) model (linear version) was trained to predict scan labels. Model evaluation was based on evaluation of the model's prediction accuracy and the reproducibility of the visualizations extracted from the model. In the finger tapping data set the model was trained to discriminate between the conditions (left) vs. (right). In the object recognition data set the model was trained to discriminate between the eight object categories as in Hanson et al. (2004). Details on specific resampling procedures are provided in the following.

Resampling details - Finger tapping data set

We split the finger tapping data set into a training set of 10 subjects and a test set of 18 subjects. Selection of the denoising parameters (σ and q) was based on the training set. The training set was repeatedly split into two disjoint sets, each with five subjects, and model performance was evaluated using the NPAIRS resampling scheme. To evaluate the reproducibility we used the weight vector/single canonical variate. In the FDA model 20 NPAIRS resampling splits were performed, and the average minimum distance on the pr-curve to the point (1,1) was obtained across the entire parameter grid. The test set was then denoised using the parameter combination giving the minimum distance. The impact of image denoising was then evaluated by constructing pr-curves based on analysis of the raw test data and denoised test data within the NPAIRS resampling framework. 20 NPAIRS splits was performed, where nine subjects

were randomly assigned to each of the split-halves.

Resampling details - Object recognition data set

In the object recognition data set we performed the evaluation of image denoising at the subject level. For a particular subject the data was split into a training and a test set - each with six runs. As with the finger tapping data set the selection of denoising parameters was based on the training set. The training set was repeatedly split into two disjoint sets, each with three runs, and model performance was evaluated using the NPAIRS resampling scheme. With eight classes we obtain seven canonical variates in the FDA. To evaluate the reproducibility we considered the first canonical variate (Chen et al., 2006). When training FDA models on different data samples the canonical variates of the FDA models are defined up to a sign and permutation ambiguity. To align canonical variates across splits we used the reference set filtering described in Strother et al. (2002). In the reference filtering procedure we initially fit a model to the entire data set and extract a set of canonical variates from this model. This set is considered as a reference set. When performing the resampling splits, we permute and flip signs of the split's individual canonical variates in order to maximize the correlation with the reference set. 10 NPAIRS resampling splits were performed (all possible combinations of runs), and the average minimum distance on the pr-curves to the point (1,1) was obtained across the entire parameter grid. Denoising parameters were selected according to minimization of the distance to (1,1) metric. The impact of image denoising was then evaluated by constructing pr-curves based on analysis of the raw test data and denoised test data (six runs) within the NPAIRS resampling framework. The entire evaluation procedure was repeated 10 times, with different runs randomly assigned to the training and test sets in each repetition. In addition to the model visualization via the canonical variates, the trained models were also visualized by means of a grand average sensitivity map, eq. (3.59).

5.3.2 Results

Image denoising in the finger tapping data set

Figure 5.26 shows results of the classification analysis of the effect of image denoising in the finger tapping data set, where image denoising was based on Mika's method. Figure 5.26(A) shows model performance, as measured by the minimum distance from the pr-curve to the point (1,1), based on the 10 subjects in the training data set. The distance first decrease with an increased number of retained components in the KPCA subspace and then tend to increase with

a high number of components retained. Least distances are observed with 16-32 components retained. For a fixed number of components there is a general tendency to decreased distance with increasing width of the Gaussian kernel. Figure 5.26(B) shows pr-curves based on analysis of the 18 subjects in the test data set. In general we observe high accuracies and reproducibilities, and the models used on denoised data are characterized by an increased reproducibility compared to models build on the raw data. For the raw data the minimum distance was 0.100 and the corresponding prediction accuracy and reproducibility was 0.994 and 0.900 respectively. For the denoised data based on the Mika's method the minimum distance was 0.0896 and the corresponding prediction accuracy and reproducibility was 0.994 and 0.911 respectively, thus maintaining the prediction accuracy and increasing the reproducibility in comparison to the models build on the raw data set. Denoisning did not result in increased prediction accuracy ($p = 0.45$), while the denoising lead to a significant increase in reproducibility ($p < 0.001$) as assessed with a nonparametric permutation test. See Rasmussen et al. (2012a) for details on the permutation analysis. For the denoised data based on Kwok and Tsang's method the minimum distance was 0.0942 and the corresponding prediction accuracy and reproducibility was 0.987 and 0.910 respectively, hence a decrease in prediction accuracy ($p < 0.001$) and an increased reproducibility ($p < 0.001$) relative to the raw data set.

Figure 5.27 shows the effect of image denoising on spatial brain maps using Mika's method for pre-image estimation. Figure 5.27(A) is based on the FDA classification models trained within the NPAIRS framework. The maps were thresholded according to correction for multiple comparisons by means of the FDR procedure using the theoretical $\mathcal{N}(0,1)$ distribution to obtain p-values for the $\overline{\text{rSPI(Z)}}$s, see Section 3.8.2. Cerebellar regions (slice -40 to -11), subcortical regions (slice 1), secondary supplementary motor area (S2)(slice 13) and sensorimotor cortex (SMC) and supplementary motor areas (SMA) (slice 37-61) are consistently identified as important by models build on both raw and denoised data. In general we observe highest Z-scores in the $\overline{\text{rSPI(Z)}}$ based on models build on the denoised data. At edges of the superthreshold regions, primarily in cerebellum, we observe a small decrease in Z-score values of the $\overline{\text{rSPI(Z)}}$. The intersection mask between the FDR thresholded maps comprised 7291 voxels. In the intersection mask 6658 voxels showed an increase in the $\overline{\text{rSPI(Z)}}$ value due to image denoising. Additionally, 701 and 91 voxels were uniquely identified in the maps corresponding to denoised and raw data respectively. Figure 5.27(B-C) show that these unique voxels primarily appear on edges of the regions identified in Figure 5.27(A).

Image denoising in the object recognition data set

Figure 5.28 depicts model performance measured in terms of minimum distance from the pr-curve to (1,1) across the denoising parameter grid for image denois-

ing with Mika's method. In general we observe a preference towards a relatively low number of retained components in the KPCA subspace. For a fixed number of principal components the distance tend first to decrease with the width of the Gaussian kernel and again slightly increase at large kernel widths, suggesting that the signal manifold may be nonlinear. In general the maximum performance was observed at an intermediate kernel width. Corresponding plots for all six subjects are provided in the supplementary materials of Rasmussen et al. (2012a) Figure 2.

Figure 5.30 demonstrates the impact of image denoising for all six subjects in the object recognition data set. Figure 5.29 shows the corresponding pr-curves. For both reconstruction methods, (Figure 5.30(A) based on Mika's method and Figure 5.30(B) based on Kwok and Tsang's method), we observe an increase in model performance across all subjects, i.e. decrease in the minimum distance from the pr-curve to (1,1). In the prediction plots (column 2) the image denoising tend to induce both slightly increases and decreases in prediction accuracy for most subjects, whereas a more dramatic decrease is observed subject 4. The reproducibility plots (column 3) show a prominent increase in reproducibilities in most subjects. We also observe an increase in the reproducibility of the sensitivity map for all subjects (column 4). Note that model selection was based on the minimum distance from the pr-curve to (1,1). Hence, the decrease in prediction accuracy for subject 4 (Figure 5.30 column two), is fully compensated by the increased reproducibility (Figure 5.30 column three) leading to a general decrease in distance (Figure 5.30 column one and Figure 5.29). As a statistical test of the impact of image denoising we used a nonparametric Wilcoxon Signed Rank Test. For all measures except prediction accuracy we could reject the null-hypothesis, that the median difference between pairs of preprocessing methods was zero, at significance level 0.05. Hence, denoising lead to changes in the minimum distance from the pr-curve to (1,1), reproducibility of the FDA basis, and reproducibility of the sensitivity map, while prediction accuracy was not significantly affected.

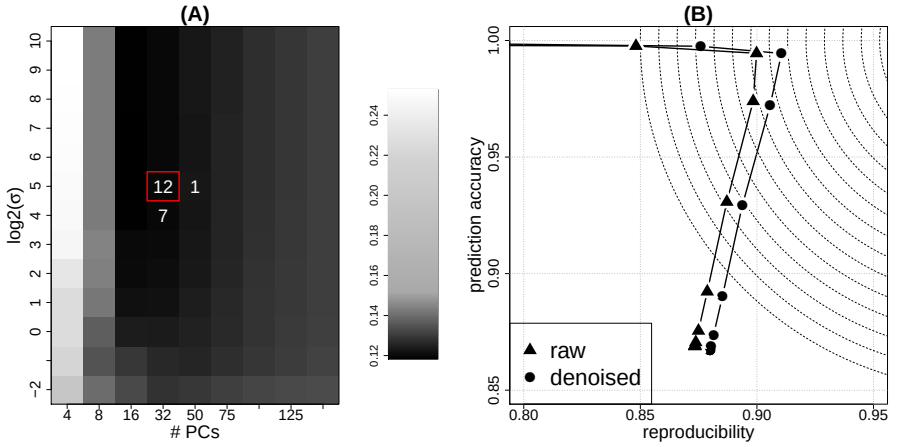

Figure 5.26: Effect of denoising in the finger tapping data set with Mika's estimation method. **Panel (A);** Model performance across part of the explored parameter grid (kernel width and kernel principal component analysis (KPCA) subspace dimensionality) based on 10 subjects. Denoising was performed at the subject level. The model performance was measured as the minimum distance on the pr-curve to the point (1,1). Resampling was performed within the NPAIRS resampling framework. The grid show the average distance across 20 NPAIRS resampling splits. The white numbers indicate the frequency at which a particular parameter combination had the lowest distance on the pr-curve to the point (1,1) across the splits. **Panel (B);** Model performance based on denoised and raw data from 18 test subjects (different from subjects used in Panel (A)). Denoising parameters were selected according to the red square in Panel (A). The pr-curves show model performance in terms of prediction accuracy and pattern reproducibility, where the pr-curves are traced out by varying the regularization parameter in the FDA classification model. The isolines indicate distances to the point (1,1). Denoisning did not result in increased prediction accuracy ($p = 0.45$), while the denoising lead to a significant increase in reproducibility $p < 0.001$ (nonparametric permutation test).

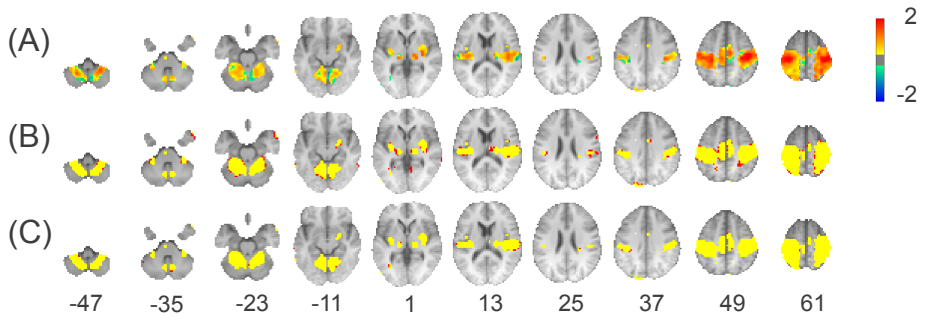

Figure 5.27: Spatial maps showing the effect of denoising in the finger tapping data set. The evaluation was based on Fisher's discriminant analysis within the NPAIRS resampling framework. **Panel (A);** Average rSPI(Z)s from models build on raw and denoised data were thresholded according to $p < 0.05$ FDR correction for multiple comparisons. Voxels shown are in the intersection mask of the two thresholded rSPI(Z)s. Voxel coloring indicate sign and magnitude of the difference between the absolute value rSPI(Z)s. Warm colors correspond to higher Z-scores in the map based on denoised data, and cold colors correspond to higher Z-scores in the map based on raw data. **Panel (B-C);** Binary masks showing voxels surviving thresholding according to FDR correction. Color coding: yellow is an intersection mask (same voxels as in panel (A)), red in panel (B) are unique to the rSPI(Z) based on denoised data, and red in panel (C) are unique to the rSPI(Z) based on the raw data. Numbers under the slices denote z coordinates in MNI space. Slices are displayed according to neurological convention (right side of a brain slice is the right side of the brain).

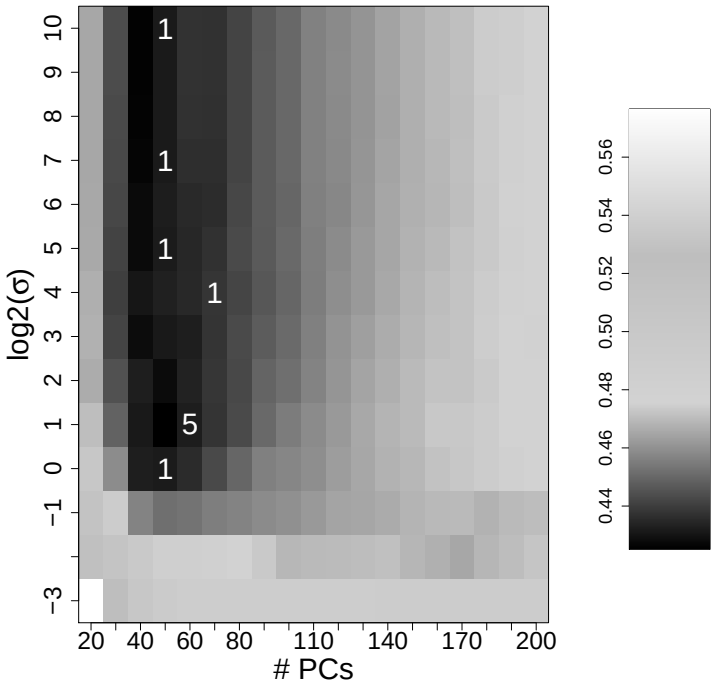

Figure 5.28: Effect of denoising in the object recognition data set - impact of denoising parameters. Denoising was performed, with Mika's method, at the subject level for each combination of the kernel width and the number of principal components. For each parameter combination an evaluation of the impact of image denoising was performed within the NPAIRS resampling framework, and the distances between the pr-maximizing point on the pr-curve to (1,1) was measured and used as a model performance metric. The distance metric was based on prediction accuracy and reproducibility of the first canonical variate in a Fisher's discriminant analysis model. Selection of denoising parameters was based on six randomly selected runs. The remaining six runs served as a test set for the evaluation of denoising in Figure 5.30. The plot shows the average distance metric across 10 resamping iterations (with 10 nested NPAIRS resamping splits within each iteration). The white numbers indicate the frequency at which a particular parameter combination had the lowest distance on the pr-curve to the point (1,1) across the 10 resampling iterations.

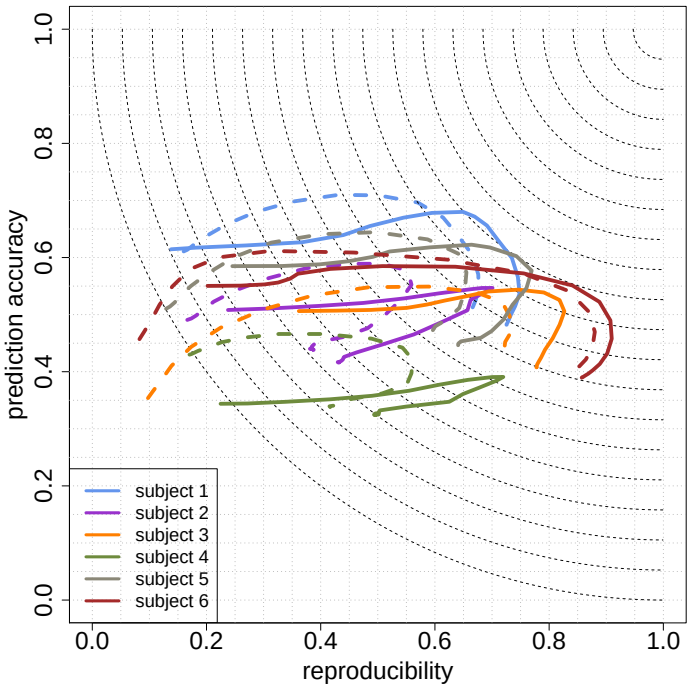

Figure 5.29: Denoising of the object recognition data set by kernel principal component analysis and pre-image estimation. Denoising parameters were selected according to minimization of the distances between the pr-maximizing point on the pr-curve to (1,1) was measured and used as a model performance metric. The distance metric was based on prediction accuracy and reproducibility of the first canonical variate in the Fisher's discriminant analysis (FDA) model. Selection of denoising parameters was based on six randomly selected runs. The remaining six runs served as a test set for the evaluation of denoising. The curves are based on averages of test set curves over 10 resampling iterations. Dashed curves are based on 'raw' data while dense curves are based on denoised data. The pr-curves are traced out by varying the regularization parameter in the FDA classification model.

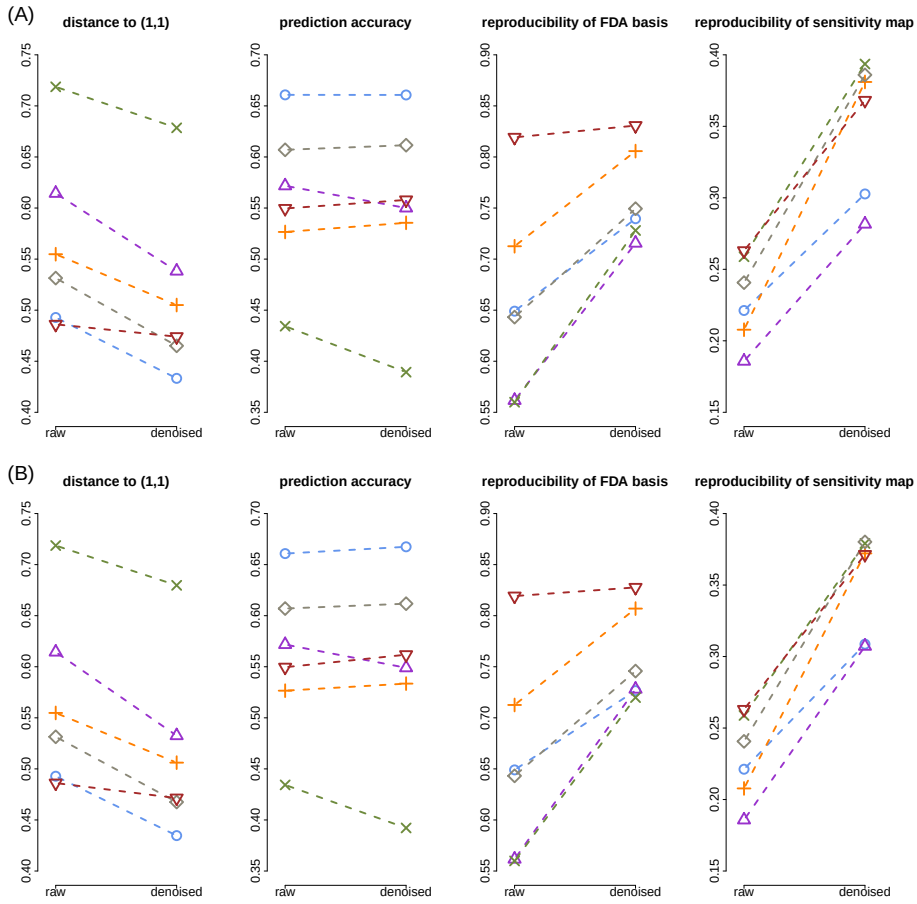

Figure 5.30: Effect of denoising in the object recognition data set - changes in model performance at the subject level. Panel (A) is based on Mika's image reconstruction method and panel (B) is based on Kwok and Tsang's method. Comparisons are based on pr-maximizing models with denoising parameters selected across the denoising parameter grid (on a training set) (see Figure 5.28), and pr-maximizing models build on the raw data. The first column shows model performance measured as the minimum distance from the pr-maximizing point on the pr-curve to (1,1). The second column shows prediction accuracy, the third column shows pattern reproducibility - both measured at the pr-maximizing point, and the fourth column shows the reproducibility of the corresponding sensitivity map. The symbols $\{\circ, \triangle, +, \times, \diamond, \triangledown\}$ correspond to subject 1-6.

Chapter 6

Conclusion and outlook

The main focus of the research presented in this dissertation has been on pattern-based analysis approaches in neuroimaging. Until recently the main approach to data analysis within the neuroimaging community has been the mass-univariate analysis (Friston et al., 1995b). This being despite the fact that the principles and methodologies underlying pattern-based analysis are not novel, e.g. Moeller and Strother (1991); Lautrup et al. (1994); Friston et al. (1995a); McIntosh et al. (1996); Mørch et al. (1997). Within the past decade there has been an appreciation within the neuroimaging community of the usefulness of pattern-based analysis procedures. The study of Kamitani and Tong (2005) provides an excellent example of pattern-based analysis allowing for detection of signal structures that conventional mass-univariate analyses may fail to identify.

Importance of selecting regularization parameters

Usually, the pattern-based analysis is implemented in terms of classification analysis. One of the most frequently adopted classifiers within the neuroimaging community is the support vector machine (SVM). The SVM has proven to be useful in providing good decoding performances (as measured by prediction accuracy) in a long series of recent papers. Many researchers motivate their preference for selecting the SVM based on results of early papers introducing the SVM to the neuroimaging community, e.g. LaConte et al. (2003); Mourão-Miranda et al. (2005). LaConte et al. (2005) reported consistent high predictive performance at high levels of the SVM regularization parameter C, while Mourão-Miranda et al. (2005) argued for the use of the 'hard-margin' SVM. A considerable number of recent papers build on observations from these studies and use the hard-margin SVM, a 'default' value or a 'high' value of C. By means of the NPAIRS resampling framework we have studied the relative influence of model regularization parameter choices on the model generalization, the reliability of the spatial patterns extracted from the classification model, and the ability of the resulting model to identify relevant brain networks defining the underlying neural encoding of the experiment. We observed the following important behaviors:

- Model reproducibility, as measured within the NPAIRS resampling framework, may vary considerably as a function of model regularization.
- While prediction accuracy may appear quite stable over a range of values of the regularization parameter, the reproducibility of the extracted pattern may vary.
- The hard-margin SVM may neither be optimal with respect to prediction accuracy nor with respect to the reproducibility of the extracted patterns.
- For the SVM, logistic regression (LogReg), and Fisher's discriminant analysis (FDA) we found a large degree of consensus between patterns extracted from the models.
- It may be more important carefully to select model regularization parameters than to select a particular model type. In particular we observe that FDA and LogReg have at least the same performance as the SVM. Note that the the first studies comparing SVM and FDA (LDA) considered unregularized versions of FDA.

We have used the NPAIRS resampling framework in order to assess model performance. In this framework we measure performance within a prediction (p) / reproducibility (r) space. By varying the model complexity one moves along a pr-curve. In general, moving along the pr-curve may allow the investigator to explore brain patterns within a hierarchy of brain↔behavior coupling, see also Strother et al. (2004). That is, there is no single classification model that optimally links task states and brain responses. By varying the regularization parameter we obtain a *continuum* of models, that may each provide information about a particular aspect of the brains response in terms of the modeled activation pattern. By selecting a particular point along the pr-curve the investigator can focus on a particular aspect of the underlying signal structure. For example, selecting the point optimizing prediction allows for identification of the subset of voxels that provides the best predictions of the scan labels. Selection of more reproducible models may allow for a more complete identification of the underlying brain network as our results suggest. This issue has indeed been discussed in a series of studies, e.g. Strother et al. (2002, 2004); Kjems et al. (2002); LaConte et al. (2003); Yourganov et al. (2010). Note that using an SVM with a 'default' regularization parameter value will in general result in a model located at an arbitrary location along the pr-curve. In such cases it becomes less clear which aspects of the underlying signal the investigator seeks to capture by the model.

An important approach to modeling, that we have not addressed in this dissertation, is the Bayesian approach. For example, consider logistic regression, e.g. Yamashita et al. (2008). Within a Bayesian framework we specify a prior

distribution, e.g. a normal distribution, over the model's weights. The prior is governed by a hyperparameter effectively controlling the regularization strength. Again we can specify a hyperprior, e.g. a gamma distribution, over the hyperparameter. The gamma distribution is parametrized by two parameters (hyper-hyperparameters) that need to be selected. In our research we experience, that by varying these hyper-hyperparameters we move along the pr-curve. Hence, instead of selecting the hyperparameter (regularization parameter) one needs to select the hyper-hyperparameters - a task that is indeed not trivial. Procedures for specifying such Bayesian hyperparameters is a topic for future research. One possible strategy is to use the NPAIRS resamping framework to facilitate the selection. Jacobsen et al. (2008) performed an evaluation of Bayesian models within the NPAIRS resampling framework. It could be interesting to perform a formal investigation of e.g. Bayesian logistic regression, logistic regression with automatic relevance determination (Yamashita et al., 2008) or the Multiclass Sparse Bayesian Regression (Michel et al., 2011a) within the NPAIRS resampling framework in order to get insight into the impact of hyper-hyperparameter choices on the model's ability to identify relevant brain networks, for example, in the finger tapping data set.

In our analysis we have studied the impact of selecting model regularization parameter while holding all other components in the neuroimaging pipeline constant (with exception of the smoothness investigation in the object recognition data set). It is important to note, that different components/choices regarding the pipeline interact - a fact that has been highlighted several times, e.g. Strother et al. (2002, 2004). In several data sets presented in this dissertation, characterized by, for example different experimental designs and preprocessing strategies, we have observed a strong and consistent dependence of regularization on model performance. Our results and conclusions may therefore generalize to other settings of the neuroimaging pipeline.

Visualization of nonlinear kernel models by sensitivity maps

Model visualization is an important aspect in the analysis of neuroimaging data sets. Often the generation of model visualizations or 'brain maps' is the ultimate goal of neuroimaging analyses. Based on such brain maps the investigator seeks to formulate claims about how information is represented in the brain. Such brain maps can be directly derived from linear models by visualizing the model's weights. It is not equally straightforward to visualize nonlinear models. Earlier studies have proposed the *sensitivity map* as a potential visualization strategy for kernel based methods (Kjems et al., 2002; LaConte et al., 2005). In Rasmussen et al. (2011) we investigated the sensitivity map as a technique for generation of global summary maps for kernel classification models. The sensitivity map visualization strategy proved to be a versatile and computationally efficient tool for such model visualization. The work on visualization of

nonlinear models were further extended in Rasmussen et al. (2012c) that investigated procedures for deriving model visualizations containing sign information. An important aspect of our analysis of the sensitivity maps as a visualization technique has been to assess the reliability/stability of the model's visualization as measured within the NPAIRS resampling framework. The sensitivity map proved as a reliable model visualization. A natural future application is to apply the nonlinear modeling and visualization framework in data sets containing more subtle and possible nonlinear effects.

Model sparsity and brain pattern interpretation

Model visualization is closely linked to the interpretation of neuroimaging experiments. Traditionally, the neuroimaging community has reported sparse statistical parametric images, where the sparse nature of the spatial maps origins from the statistical testing. Statistical tests are performed at the voxel-level. The resulting statistical parametric image is subsequently thresholded in order to control e.g. the family-wise error rate. Building pattern-based analysis models of neuroimaging data sets generally results in an estimate of the prediction accuracy. Inspecting prediction accuracies allows the investigator to assess whether the model succeed in capturing the relevant underlying signal structure, i.e. is capable of performing the mapping from brain scans to scan labels. Additionally, it is often relevant to identify the brain locations that supports discriminative information to the models. For example, it may be expected that the information is encoded in a distributed pattern of localized clusters. Pattern-based analysis analysis that enforce spatial sparsity has been introduced to the neuroimaging community as *interpretable* models, implying that dense predictive models are difficult to interpret. We find that this distinction between sparse and dense models is overly simplified. Sparse models are not necessarily more interpretable than dense models. Consider a data set with scans of 10^5 voxels. Of these voxels a subset $\mathcal{A}$ is closely coupled to the experimental task. Another subset $\mathcal{B}$ shows an intermediate level of task coupling while the remaining voxels $\mathcal{C}$ are uncoupled to the experimental task (and hence irrelevant in a decoding context). A sparse linear model optimized to maximize prediction accuracy may primarily identify voxels in $\mathcal{A}$ as having non-zero weights. However, it is important to note that such a model still is parametrized by 10^5 weights. Most of these weights are set to zero, which in itself is a strong statement. Brain maps based on sparse models may often be presented with lack of quantification of the stability or significance of the sparse pattern identified by the models. Hence, there remains several open questions to be answered. What is the statistical significance of the sparse brain pattern? Is the brain pattern stable across resampling splits? What characterizes the voxels with non-zero weights? Do voxels with a weight set to zero have no discriminative information? It may be challenging and difficult to interpret a dense model, and we may prefer a model identifying a subset of voxels. However, cautions must be

taken if claims about information representation is based on the sparse model, e.g. the voxels in $\mathcal{B}$ may be part of the underlying brain network. However, such voxels are not identified as important by the model since we formulate the optimization objective in order to maximize the prediction accuracy. Sparse models are not necessary interpretable *per se*. Other types of sparsity are seen in models that are sparse in the observation dimension (e.g. the SVM) and in sparse feature representation as identified by e.g. a PCA subspace. In Rasmussen et al. (2012b) we found that maximizing prediction accuracy lead to models in which the weight vector depended on relatively few training observations. However, more observations' support to the weight vector was required in order to increase reproducibility and enhance the model's ability to detect the relevant underlying brain networks. Models supported by relatively few data observation tended also to produce sparse brain patterns following a statistical thresholding procedure. In many general machine learning applications sparsity (in the observation dimension) is a desirable model property, since it can speed up processing in e.g. digit recognition systems. It is questionable if such sparsity is of equal desire in the analysis of neuroimaging data, if the purpose of the modeling is to learn the underlying information representation in the brain. Our results in Rasmussen et al. (2012b) suggest that solutions based on a relatively high fraction of the training observations will produce model visualizations that are more stable than visualizations based on only few data observations.

Nonlinear denoising and analysis with kernel principal component analysis and pre-image estimation

In Rasmussen et al. (2012a) we investigated the use of kernel principal component analysis (KPCA) and pre-image estimation as a means for image denoising as part of the image preprocessing pipeline. We based the investigation on two fMRI data sets, and evaluated the proposed method within the NPAIRS resampling framework. The proposed denoising strategy proved primarily to increase pattern reproducibility as measured within the NPAIRS resampling framework. Important future research topics include procedures for identification of signal/noise components from the KPCA feature representation. Another natural extension is to develop a nonlinear generalization of the multivariate, data-driven method for the characterization and removal of physiological noise in fMRI data, (PHYCAA) proposed by Churchill et al. (2012b). This method uses canonical correlation analysis (CCA) to identify noise structures in fMRI data. Pre-image estimation could be used directly in conjunction with kernel CCA in a similar *nonlinear* denoising scheme. In Rasmussen et al. (2012a) we further proposed a manifold navigation procedure for exploration of a nonlinear data manifold as an extension to existing technology applicable to linear models (Sato et al., 2008). The proposed method can be used to generate brain maps in the continuum between experimentally defined brain states. We provided an illustration that the method is capable of exploring a nonlinear manifold by con-

structing interpolated images via pre-image estimation. However, it is important to emphasize that we provide an illustration and nothing more. The nonlinearity was mainly observed for extrapolated data points. In future research it is relevant to acquire an fMRI data set, where interpolated/extrapolated data points can be compared with actual brain scans residing in the continuum between the brain states. Hence, an assessment whether the method is capable of predicting novel stimuli that are not in the training set should be performed (Raizada and Kriegeskorte, 2010).

Concluding remarks

The work presented in this dissertations is motivated by two overall goals. Firstly, we attempted to obtain further insight into commonly applied pattern-based analysis models' ability to identify relevant signal structures in neuroimaging data sets. In our research we used the NPAIRS resampling framework (Strother et al., 2002) to evaluate the model visualization reproducibility as means for model evaluation in addition to prediction accuracy. We hope that our results highlights that there are open issues to be addressed - even when one considers models that may be considered to be well established. How should the models and models' parameters be selected in order to maximize the scientific outcome of the analyses? How do we formulate an objective that enhances scientific discovery rather than prediction accuracy? We hope that our results will stimulate investigators in continuing pursuing these research topics in future research. Secondly, we performed an investigation of the applicability of nonlinear methods within the context of the preprocessing and analysis of neuroimaging data sets. We hope that our illustrations and analyses convinces the reader that nonlinear methods provide means for enhanced signal detection in cases where linear modeling may be too restrictive. We look forward to future applications of nonlinear learning within the analysis of functional neuroimages. Finally, we here formulate two ultimate goals to be addressed by pattern-based analyses:

- Given a brain scan, can we, based on the activation pattern, infer which of multiple brain states a subject was engaged in when the scan was acquired?
- What is the most complete and reliable spatial pattern reflecting the underlying neural encoding of the experiment defining the multiple brain states?

We advocate that both goals should be pursued in order to maximize the neuroscientific outcome of the analysis of functional neuroimages.

Bibliography

Abrahamsen, T. J., Hansen, L. K., 2011a. A Cure for Variance Inflation in High Dimensional Kernel Principal Component Analysis. Journal of Machine Learning Research 12, 2027–2044.

Abrahamsen, T. J., Hansen, L. K., 2011b. Regularized pre-image estimation for kernel PCA de-noising input space regularization and sparse reconstruction. Journal of Signal Processing Systems 65 (3), 403–412.

AITB, 1944. Army individual test battery. Manual of directions and scoring. Washington, DC: War Department, Adjutant General's Office.

Aronszajn, N., 1950. Theory of reproducing kernels. Transactions of the American Mathematical Society 68, 337–404.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113.

Ashburner, J., Friston, K. J., 1999. Nonlinear spatial normalization using basis functions. Human Brain Mapping 7 (4), 254–266.

Ashburner, J., Friston, K. J., 2005. Unified segmentation. NeuroImage 26 (3), 839–851.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanab, M., Hansen, K., Müller, K.-R., 2010. How to explain individual classification decisions. Journal of Machine Learning Research 11, 1803–1831.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., Hyde, J. S., 1992. Time course epi of human brain function during task activation. Magnetic Resonance in Medicine 25 (2), 390–397.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289–300.

Bie, T. D., Cristianini, N., Rosipal, R., 2005. Eigenproblems in Pattern Recognition. In: Handbook of Geometric Computing : Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics. Springer-Verlag, Heidelberg.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, 144–152.

Bowie, C. R., Harvey, P. D., 2006. Administration and interpretation of the trail making test. Nat. Protocols 1, 2277–2281.

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., Rosen, B. R., 1996. Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. Proceedings of the National Academy of Sciences 93 (25), 14878–14883.

Bullmore, E. T., Rabe-Hesketh, S., Morris, R. G., Williams, S. C. R., Gregory, L., Gray, J. A., Brammer, M. J., 1996. Functional magnetic resonance image analysis of a large-scale neurocognitive network. NeuroImage 4 (1), 16–3.

Burges, C. J. C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167.

Bushberg, J. T., Seibert, J. A., Leidholdt, E. M., Boone, J. M., 2001. The Essential Physics of Medical Imaging (2nd Edition), 2nd Edition. Lippincott Williams & Wilkins.

Buxton, R. B., Uludağ, K., Dubowitz, D. J., Liu, T. T., 2004. Modeling the hemodynamic response to brain activation. NeuroImage 23, Supplement 1 (0), S220–S233.

Buxton, R. B., Wong, E. C., Frank, L. R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magnetic resonance in medicine 39 (6), 855–864.

Calhoun, V., Pekar, J., McGinty, V., Adali, T., Watson, T., Pearlson, G., 2002. Different activation dynamics in multiple neural systems during simulated driving. Human Brain Mapping 16 (3), 158–167.

Carlson, T. A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. Journal of Cognitive Neuroscience 15 (5), 704–717.

Carroll, M. K., Cecchi, G. A., Rish, I., Garg, R., Rao, A. R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44 (1), 112–122.

Cawley, G. C., Talbot, N. L. C., 2004. Sparse Bayesian Kernel Logistic Regression. Neural Networks, 133–138.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chapelle, O., 2007. Training a support vector machine in the primal. Neural Computation 19 (5), 1155–1178.

Chen, X., Pereira, F., Lee, W., Strother, S., Mitchell, T., 2006. Exploring predictive and reproducible modeling with the single-subject FIAC dataset. Human Brain Mapping 27 (5), 452–461.

Chu, C., Ni, Y., Tan, G., Saunders, C. J., Ashburner, J., 2011a. Kernel regression for fMRI pattern prediction. NeuroImage 56 (2), 662–673.

Chu, C. J., Chiu, Y.-C., Kriegeskorte, N.and Tan, G. A. J., 2011b. Utilizing temporal information in fmri decoding: Classifier using kernel regression methods. NeuroImage 58 (2), 560–571.

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J. E., Graham, S. J., Strother, S. C., 2012a. Optimizing preprocessing and analysis pipelines for single-subject fMRI: 1. Standard temporal motion and physiological noise correction methods. Human Brain Mapping 59 (2), 1299–1314.

Churchill, N. W., Yourganov, G., Spring, R., Rasmussen, P. M., Lee, W., Ween, J. E., Strother, S. C., 2012b. PHYCAA: Data-driven measurement and removal of physiological noise in BOLD fMRI. NeuroImage 59 (2), 1299–1314.

Clark, C. M., Ammann, W., Martin, W. R. W., Ty, P., Hayden, M. R., 1991. The FDG/PET methodology for early detection of disease onset: A statistical model. Journal of Cerebral Blood Flow & Metabolism 11, A96–A102.

Cortes, C., Vapnik, V., 1995. Support-vector networks. In: Machine Learning. pp. 273–297.

Cox, D. D., Savoy, R. L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19 (2), 261–270.

Cox, R. W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.

Cuingnet, R., Chupin, M., Benali, H., Colliot, O., 2010. Spatial and anatomical regularization of svm for brain image analysis. Advances in Neural Information Processing Systems 23, 460–468.

Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., Samson, Y., Colliot, O., 2011. Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. Medical Image Analysis 15 (5), 729–737.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Salle, F. D., Goebel, R., Formisano, E., 2007. Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. NeuroImage 34 (1), 177–194.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43 (1), 44–58.

Dinesh, G. N., 2005. About being BOLD. Brain Research Reviews 50 (2), 229–243.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourão-Miranda, J., Marquand, A., Daly, E. M., Brammer, M. J., Murphy, C., Murphy, D. G., 2010. Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. NeuroImage 49 (1), 44–56.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association 78 (382), 316–331.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: The .632+ bootstrap method. Journal of the American Statistical Association 92 (438), 548–560.

Eickhoff, S. B., Amunts, K., Mohlberg, H., Zilles, K., 2005. The human parietal operculum. ii. stereotaxic maps and correlation with functional imaging results. Cerebral Cortex 16 (2), 268–279.

Faro, S., Mohamed, F., 2010. BOLD FMRI: A Guide to Functional Imaging for Neuroscientists. Springer.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008a. "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. Science 322, 970–973.

Formisano, E., De Martino, F., Valente, G., 2008b. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magnetic Resonance Imaging 26, 921–934.

Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2003. Adaptive analysis of fMRI data. NeuroImage 19 (3), 837–845.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W., 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. NeuroImage 39 (1), 181–205.

Friston, K., Frith, C., Frackowiak, R., Turner, R., 1995a. Characterizing Dynamic Brain Responses with fMRI: A Multivariate Approach. NeuroImage 2 (2, Part A), 166–172.

Friston, K., Holmes, A., Poline, J.-B., Grasby, P., Williams, S., Frackowiak, R., Turner, R., 1995b. Analysis of fMRI Time-Series Revisited. NeuroImage 2 (1), 45–53.

Friston, K., Mechelli, A., Turner, R., Price, C., 2000. Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels, and Other Hemodynamics. NeuroImage 12 (4), 466–477.

Friston, K. J., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time-series. Human Brain Mapping 1 (2), 153–171.

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., Turner, R., 1996. Movement-related effects in fMRI time-series. Magnetic Resonance in Medicine 35 (3), 346–355.

Glover, G. H., Li, T. Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. 44 (1), 162–167.

Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies. Information processing in medical imaging proceedings of the conference 18, 330–341.

Golland, P., Grimson, W., Shenton, M., Kikinis, R., 2001. Deformation analysis for shape based classification. Inf Process Med Imaging. IPMI 17, 517–530.

Golland, P., Grimson, W. E. L., Shenton, M. E., Kikinis, R., 2005. Detection and analysis of statistical differences in anatomical shape. Medical Image Analysis 9, 69–86.

Grosenick, L., Greer, S., Knutson, B., 2008. Interpretable classifiers for fmri improve prediction of purchases. Neural Systems and Rehabilitation Engineering, IEEE Transactions on 16 (6), 539–548.

Guo, Y., 2010. A weighted cluster kernel PCA prediction model for multi-subject brain imaging data. Statistics And Its Interface 3 (1), 103–111.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3 (7/8), 1157–1182.

Hansen, L. K., Larsen, J., Nielsen, F. A., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., Paulson, O. B., May 1999. Generalizable patterns in neuroimaging: How many principal components? NeuroImage 9 (5), 534–544.

Hanson, S. J., Halchenko, Y. O., 2008. Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. Neural computation 20 (2), 486–503.

Hanson, S. J., Matsuka, T., Haxby, J. V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? NeuroImage 23 (1), 156–166.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303 (5664), 1634–1640.

Hastie, T., Rosset, S., Tibshirani, R., Zhu, J., 2004. The entire regularization path for the support vector machine. Journal of Machine Learning Research 5 (1), 1391–1415.

Hastie, T., Tibshirani, R., 2004. Efficient quadratic regularization for expression arrays. Biostatistics 5 (3), 329–340.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference and prediction. Springer.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., Pietrini, P., Sep. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293 (5539), 2425–2430.

Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7, 523–534.

Hoerl, A. E., Kennard, R. W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12 (1), 55–67.

Izenman, A. J., 2008. Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. Springer-Verlag.

Jacobsen, D. J., Hansen, L. K., Madsen, K. H., 2008. Bayesian model comparison in nonlinear BOLD fMRI hemodynamics. Neural Computation 20 (3), 738–755.

Kamitani, Y., Sawahata, Y., 2010. Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. NeuroImage 49 (3), 1949–1952.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nature Neuroscience 8, 679–685.

Kippenhan, J. S., Barker, W. W., Pascal, S., Nagel, J., Duara, R., 1992. Evaluation of a neural-network classifier for pet scans of normal and alzheimer's disease subjects. Journal of Nuclear Medicine 33 (8), 1459–1467.

Kjems, U., Hansen, L., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., Strother, S., 2002. The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. NeuroImage 15 (4), 772–786.

Klöppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., Rohrer, J., Fox, N., Jack, C., Ashburner, J., Frackowiak, R., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., Gaser, C., 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Archives of general psychiatry 66 (7), 700–712.

Kriegeskorte, N., Cusack, R., Bandettini, P., 2010. How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex spatiotemporal filter? NeuroImage 49 (3), 1965–1976.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103 (10), 3863–3868.

Kustra, R., Strother, S. C., 2001. Penalized discriminant analysis of [15-O]-water PET brain images with prediction error selection of smoothness and regularization hyperparameters. IEEE Transactions on Medical Imaging 20 (5), 376–387.

Kwok, J. T.-Y., Tsang, I. W.-H., 2004. The pre-image problem in kernel methods. IEEE transactions on neural networks 15 (6), 1517–1525.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L. K., Yacoub, E., Hu, X., Rottenberg, D., Strother, S., 2003. The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics. NeuroImage 18 (1), 10–27.

LaConte, S., Peltier, S., Hu, X., 2007. Real-time fMRI using brain-state classification. Human Brain Mapping 28 (10), 1033–1044.

LaConte, S., Strother, S. C., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26 (2), 317–329.

Lal, T., Chapelle, O., Weston, J., Elisseeff, A., 2006. Embedded methods. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L. (Eds.), Feature Extraction. Vol. 207 of Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, pp. 137–165.

Lautrup, B., Hansen, L., Law, I., Mørch, N., Svarer, C., Strother, S., 1994. Massive weight sharing: A cure for extremely ill-posed problems. Proceedings of the Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks. World Scientific, Ulich, Germany, 137–148.

Logothetis, N. K., 2008. What we can do and what we cannot do with fMRI. Nature 453 (7197), 869–878.

López, M., Ramírez, J., Górriz, J., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R., 2009. SVM-based CAD system for early detection of the Alzheimer's disease using kernel PCA and LDA. Neuroscience Letters 464 (3), 233–238.

Lukic, A., Wernick, M., Tzikas, D., Chen, X., Likas, A., Galatsanos, N., Yang, Y., Zhao, E., Strother, S., 2007. Bayesian kernel methods for analysis of functional neuroimages. IEEE Transactions on Medical Imaging 26 (12), 1613–1624.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., Nichols, T. E., 2006. Non-white noise in fMRI: Does modelling have an impact? NeuroImage 29 (1), 54–66.

MacKay, D. J. C., 1992. Bayesian interpolation. Neural Computation 4 (3), 415–447.

MacKay, D. J. C., 1994. Bayesian non-linear modelling for the prediction competition. In ASHRAE Transactions V1.00 (2), 1053–1062.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourão-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. NeuroImage 49 (3), 2178–2189.

McIntosh, A., Bookstein, F., Haxby, J., Grady, C., 1996. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage 3 (3), 143–157.

McIntosh, A. R., Lobaugh, N. J., 2004. Partial least squares analysis of neuroimaging data: applications and advances. NeuroImage 23, Supplement 1, S250–S263.

McKeown, M. J., Hansen, L. K., Sejnowsk, T. J., 2003. Independent component analysis of functional MRI: what is signal and what is noise? Current Opinion in Neurobiology 13 (5), 620–629.

McKeown, M. J., Jung, T.-P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T.-W., Sejnowski, T. J., 1998. Spatially independent activity patterns in functional MRI data during the stroop color-naming task. Proceedings of The National Academy of Sciences 95, 803–810.

Meier, L., Van De Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (1), 53–71.

Michel, V., Eger, E., Keribin, C., Thirion, B., 2011a. Multiclass Sparse Bayesian Regression for fMRI-Based Prediction. International Journal of Biomedical Imaging 2011, doi:10.1155/2011/350838.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B., 2011b. A supervised clustering approach for fMRI-based inference of brain states. Pattern Recognition doi:10.1016/j.patcog.2011.04.006.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011c. Total Variation Regularization for fMRI-Based Prediction of Behavior. Medical Imaging, IEEE Transactions on 30 (7), 1328–1340.

Mika, S., Rätsch, G., Schölkopf, B., Smola, A., Weston, J., Müller, K.-R., 1999a. Invariant feature extraction and classification in kernel spaces. Advances in Neural Information Processing Systems 12, 526–532.

Mika, S., Schölkopf, B., Smola, A., Müller, K., Scholz, M., Rätsch, G., 1999b. Kernel PCA and de-noising in feature spaces. Advances in neural information processing systems 11 (1), 536–542.

Minka, T., 2003. A comparison of numerical optimizers for logistic regression. Tech. rep., Available from research.microsoft.com/˜minka/papers/logreg/.

Misaki, M., Kim, Y., Bandettini, P. A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53 (1), 103–118.

Mitchell, T., Hutchinson, R., Niculescu, S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. Machine Learning 57 (1-2), 145–175.

Moeller, J. R., Strother, S. C., 1991. A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. Journal of Cerebral Blood Flow & Metabolism 11, A121–A135.

Molinaro, A. M., Simon, R., Pfeiffer, R. M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21 (15), 3301–3307.

Mørch, N., Hansen, L., Strother, S., Svarer, C., Rottenberg, D., Lautrup, B., Savoy, R., Paulson, O., 1997. Nonlinear versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover. In: Proceedings of the 15th International Conference on Information Processing in Medical Imaging, 1997. Vol. Springer Lecture Notes in Computer Science 1230. pp. 259–270.

Moritz, C. H., Haughton, V. M., Cordes, D., Quigley, M., Meyerand, M. E., 2000a. Whole-brain Functional MR Imaging Activation from a Finger-tapping Task Examined with Independent Component Analysis. American Journal of Neuroradiology 21 (9), 1629–1635.

Moritz, C. H., Meyerand, M. E., Cordes, D., Haughton, V. M., 2000b. Functional MR Imaging Activation after Finger Tapping Has a Shorter Duration in the Basal Ganglia Than in the Sensorimotor Cortex. American Journal of Neuroradiology 21 (7), 1228–1234.

Mourão-Miranda, J., Bokde, A. L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data. NeuroImage 28 (4), 980–995, special Section: Social Cognitive Neuroscience.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on svm analysis of single-subject and multi-subject fmri data. NeuroImage 33 (4), 1055–1065.

Mumford, J. A., Turner, B. O., Ashby, F. G., Poldrack, R. A., 2011. Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. NeuroImage, doi:10.1016/j.neuroimage.2011.08.076.

Nichols, T. E., Holmes, A. P., 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping 15 (1), 1–25.

Norman, K. A., Polyn, S. M., Detre, G. J., Haxby, J. V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10 (9), 424–430.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., Ugurbil, K., 1992. Intrinsic signal changes accompanying sensory stimulation:

functional brain mapping with magnetic resonance imaging. Proceedings of the National Academy of Sciences 89 (13), 5951–5955.

Op de Beeck, H. P., 2010. Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? NeuroImage 49 (3), 1943–1948.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., Parent, M. A., Nov. 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. Journal of cognitive neuroscience 19 (11), 1735–1752.

Park, M. Y., Hastie, T., 2008. Penalized logistic regression for detecting gene interactions. Biostatistics 9 (1), 30–50.

Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: A comparative study. NeuroImage 56 (2), 476–96.

Raizada, R. D. S., Kriegeskorte, N., March 2010. Pattern-information fmri: New questions which it opens up and challenges which face it. Int. J. Imaging Syst. Technol. 20, 31–41.

Rasmussen, P. M., Abrahamsen, T. J., Madsen, K. H., Hansen, L. K., 2012a. Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-image estimation. NeuroImage 60 (3), 1807–1818.

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Chruchill, N. W., Strother, S. C., 2012b. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45 (6), 2085–2100.

Rasmussen, P. M., Madsen, K. H., Lund, T. E., Hansen, L. K., 2011. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage 55 (3), 1120–1131.

Rasmussen, P. M., Schmah, T., Madsen, K. H., Lund, T. E., Yourganov, G., Strother, S. C., Hansen, L. K., 2012c. Visualization of nonlinear classification models in neuroimaging - signed sensitivity maps. In: Biosignals 2012 International Conference on Bio-inspired Systems and Signal Processing.

Riecker, A., Wildgruber, D., Mathiak, K., Grodd, W., Ackermann, H., 2003. Parametric analysis of rate-dependent hemodynamic response functions of cortical and subcortical brain structures during auditorily cued finger tapping: a fMRI study. NeuroImage 18 (3), 731–739.

Rissman, J., Greely, H. T., Wagner, A. D., 2010. Detecting individual memories through the neural decoding of memory states and past experience. Proceedings of the National Academy of Sciences doi:10.1073/pnas.1001028107.

Rosipal, R., Be, P. P., Trejo, L. J., Cristianini, N., Shawe-taylor, J., Williamson, B., 2001. Kernel partial least squares regression in reproducing kernel hilbert space. Journal of Machine Learning Research 2, 97–123.

Ryali, S., Supekar, K., Abrams, D. A., Menon, V., 2010. Sparse logistic regression for whole-brain classification of fMRI data. NeuroImage 51 (2), 752–764.

Sato, J. R., Thomaz, C. E., Cardoso, E. F., Fujita, A., da Graa Morais Martin, M., Jr., E. A., 2008. Hyperplane navigation: A method to set individual scores in fmri group datasets. NeuroImage 42 (4), 1473–1480.

Saunders, C., Gammerman, A., Vovk, V., 1998. Ridge regression learning algorithm in dual variables. In: In Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, pp. 515–521.

Schmah, T., Hinton, G. E., Zemel, R. S., Small, S. L., Strother, S. C., 2008. Generative versus discriminative training of RBMs for classification of fMRI images. In: NIPS. pp. 1409–1416.

Schmah, T., Yourganov, G., Zemel, R., Hinton, G., Small, S., Strother, S., 2010. A Comparison of Classification Methods for Longitudinal fMRI Studies. Neural Computation 22, 2729–2762.

Schölkopf, B., Mika, S., Burges, C. J. C., Knirsch, P., Müller, K. R., Rätsch, G., Smola, A. J., 1999. Input space versus feature space in kernel-based methods. IEEE Transactions On Neural Networks 10 (5), 1000–1017.

Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10 (5), 1299–1319.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Shmuel, A., Chaimow, D., Raddatz, G., Ugurbil, K., Yacoub, E., 2010. Mechanisms underlying decoding at 7 t: Ocular dominance columns, broad structures, and macroscopic blood vessels in v1 convey information on the stimulated eye. NeuroImage 49 (3), 1957–1964.

Sigurdsson, S., Philipsen, P., Hansen, L., Larsen, J., Gniadecka, M., Wulf, H., 2004. Detection of skin cancer by classification of Raman spectra. IEEE Transactions on Biomedical Engineering 51 (10), 1784–1793.

Smith, S., Jenkinson, M., Woolrich, M., Beckmann, C., Behrens, T., Johansen-Berg, H., Bannister, P., De Luca, M., Drobnjak, I., Flitney, D., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J., Matthews, P., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 26 (S1), 208–219.

Song, S., Zhan, Z., Long, Z., Zhang, J., Yao, L., 02 2011. Comparative Study of SVM Methods Combined with Voxel Selection for Object Category Classification on fMRI Data. PLoS ONE 6 (2), e17191.

Song, X., Ji, T., Wyrwicz, A. M., 2008. Baseline drift and physiological noise removal in high field fMRI data using kernel PCA. In: ICASSP. pp. 441–444.

Stephan, K., Kasper, L., Harrison, L., Daunizeau, J., den Ouden, H., Breakspear, M., Friston, K., 2008. Nonlinear dynamic causal models for fMRI. NeuroImage 42 (2), 649–662.

Strother, S., Anderson, J., Hansen, L., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework. NeuroImage 15 (4), 747–771.

Strother, S., Conte, S. L., Hansen, L. K., Anderson, J., Zhang, J., Pulapura, S., Rottenberg, D., 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. NeuroImage 23, Supplement 1, S196–S207.

Strother, S., Oder, A., Spring, R., Grady, C., 2010. The NPAIRS Computational Statistics Framework for Data Analysis in Neuroimaging. Proc. 19th Int. Conf. on Computational Statistics, Paris, 111–120.

Strother, S. C., Lange, N., Anderson, J. R., Schaper, K. A., Rehm, K., Hansen, L. K., Rottenberg, D. A., 1997. Activation pattern reproducibility: Measuring the effects of group size and data analysis models. Human Brain Mapping 5, 312–316.

Tam, F., Churchill, N. W., Strother, S. C., Graham, S. J., 2011. A new tablet for writing and drawing during functional MRI. Human Brain Mapping 32 (8), 240–248.

Thirion, B., Faugeras, O., 2003. Dynamical components analysis of fMRI data through kernel PCA. NeuroImage 20 (1), 34–49.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.-B., 2006. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. Human Brain Mapping 27 (8), 678–693.

Thomas, C. G., Harshman, R. A., Menon, R. S., Nov. 2002. Noise reduction in BOLD-based fMRI using component analysis. Neuroimage 17 (3), 1521–1537.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58 (1), 267–288.

Tohka, J., Foerde, K., Aron, A. R., Tom, S. M., Toga, A. W., Poldrack, R. A., 2008. Automatic independent component labeling for artifact removal in fMRI. NeuroImage 39 (3), 1227–1245.

Valente, G., De Martino, F., Esposito, F., Goebel, R., Formisano, E., 2011. Predicting subject-driven actions and sensory experience in a virtual world with relevance vector machine regression of fmri data. NeuroImage 56 (2), 651–661.

Van Essen, D. C., 2004. Surface-based approaches to spatial localization and registration in primate cerebral cortex. NeuroImage 23, Supplement 1 (0), S97–S107, mathematics in Brain Imaging.

Wallentin, M., Nielsen, A. H., Vuust, P., Dohn, A., Roepstorff, A., Lund, T. E., 2011. BOLD response to motion verbs in left posterior middle temporal gyrus during story comprehension. Brain and Language 119 (3), 221–225.

Wang, Z., 2009. A hybrid SVM-GLM approach for fMRI data analysis. NeuroImage 46 (3), 608–615.

Wang, Z., Childress, A., Wang, J., Detre, J., 2007. Support vector machine learning-based fMRI data group analysis. NeuroImage 36 (4), 1139–1151.

Witt, S. T., Laird, A. R., Meyerand, M. E., Aug. 2008. Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. NeuroImage 42 (1), 343–356.

Wolbers, T., Zahorik, P., Giudice, N. A., 2011. Decoding the direction of auditory motion in blind humans. NeuroImage 56 (2), 681–687.

Worsley, K., Friston, K., 1995. Analysis of fMRI Time-Series Revisited-Again. NeuroImage 2 (3), 173–181.

Yamashita, O., aki Sato, M., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 42 (4), 1414–1429.

Yourganov, G., Schmah, T., Small, S. L., Rasmussen, P. M., Strother, S. C., 2010. Functional connectivity metrics during stroke recovery. Archives Italiennes de Biologie 148 (3), 259–270.

Yunqian Ma, Cherkassky, V., 2005. Characterization of data complexity for svm methods. IEEE International Joint Conference on Neural Networks 2, 919–924.

Zhang, J., Anderson, J. R., Liang, L., Pulapura, S. K., Gatewood, L., Rottenberg, D. A., Strother, S. C., 2009. Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. Magnetic Resonance Imaging 27 (2), 264–278.

Zhang, Z., Dai, G., Xu, C., Jordan, M., 2010. Regularized Discriminant Analysis, Ridge Regression and Beyond. Journal of Machine Learning Research 11, 2199–2228.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320.

Zurada, J., Malinowski, A., Cloete, I., 1994. Sensitivity analysis for minimization of input data dimension for feedforward neural network. 1994 IEEE International Symposium on Circuits and Systems, 1994. ISCAS'94. 6, 447–450.

Zurada, J., Malinowski, A., Usui, S., 1997. Perturbation method for deleting redundant inputs of perceptron networks. Neurocomputing 14 (2), 177–193.