STATISTICAL ANALYSIS OF MEDICAL IMAGES WITH APPLICATIONS TO NEUROIMAGING

By
Rafal Kustra

THESIS SUBMITTED IN CONFORMITY WITH THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY GRADUATE DEPARTMENT OF PUBLIC HEALTH SCIENCES IN THE UNIVERSITY OF TORONTO TORONTO, ONTARIO AUGUST 2000

© Copyright by Rafal Kustra. 2000
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.
UNIVERSITY OF TORONTO
DEPARTMENT OF
PUBLIC HEALTH SCIENCES

The undersigned hereby certify that they have read and recommend to the School of Graduate Studies for acceptance a thesis entitled "Statistical Analysis of Medical Images with Applications to Neuroimaging" by Rafal Kustra in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 2000
External Examiner:  
__________________________
Keith Worsley. PhD

Research Supervisor:  
__________________________
Robert Tibshirani

Examining Committee:  
__________________________
Steven Strother. PhD

__________________________
James Stafford. PhD

__________________________
Randy McIntosh. PhD
UNIVERSITY OF TORONTO

Date: August 2000

Author: Rafal Kustra

Title: Statistical Analysis of Medical Images with Applications to Neuroimaging

Department: Public Health Sciences

Degree: Ph.D. Convocation: October Year: 2000

Permission is herewith granted to University of Toronto to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

______________________________
Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.
Contents

List of Tables ix

List of Figures x

Abstract xvi

Acknowledgments xvii

1 Introduction 1

1.1 Images as Data ................................................. 1
1.2 Motivation and the Setup for Statistical Image Analysis ........... 4
1.3 Functional Data ................................................ 6
1.4 Notation and Conventions ...................................... 9

2 Neuroimaging Data and Methods 12

2.1 Goals and Study Design in Neuroimaging .......................... 12
2.2 PET and fMRI Modalities ....................................... 15
  2.2.1 Positron Emission Tomography ............................. 15
  2.2.2 Functional Magnetic Resonance Imaging .................... 20
2.3 Literature Overview ............................................. 24
  2.3.1 Single Voxel Analysis: Statistical Parametric Mapping and Gaussian Random Fields ............................. 27
### 3 Penalized Linear Discriminant Analysis with Basis Expansion

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Classical Linear Discriminant Analysis</td>
<td>44</td>
</tr>
<tr>
<td>3.1.1 Discriminant Functions and MANOVA View of LDA</td>
<td>46</td>
</tr>
<tr>
<td>3.1.2 The Geometry of LDA in Two Class, 2D setting</td>
<td>47</td>
</tr>
<tr>
<td>3.2 LDA and Random Subject Effects</td>
<td>48</td>
</tr>
<tr>
<td>3.2.1 Simulation Study</td>
<td>51</td>
</tr>
<tr>
<td>3.3 Dimension Reduction in LDA using Smoothness Constraints and Penalization</td>
<td>57</td>
</tr>
<tr>
<td>3.3.1 Basis Expansion of Canonical Variates</td>
<td>58</td>
</tr>
<tr>
<td>3.3.2 Penalized Linear Discriminant Analysis</td>
<td>59</td>
</tr>
<tr>
<td>3.3.3 Penalized Discriminant Analysis and Statistical Parametric Mapping</td>
<td>61</td>
</tr>
<tr>
<td>3.3.4 PDA via Regression</td>
<td>64</td>
</tr>
<tr>
<td>3.3.5 Expressing the PDA algorithm in the N-dimensional space</td>
<td>66</td>
</tr>
<tr>
<td>3.3.6 Effective Degrees of Freedom</td>
<td>67</td>
</tr>
<tr>
<td>3.3.7 Prediction Error and its Estimates</td>
<td>68</td>
</tr>
<tr>
<td>3.4 A Note on Gaussian Assumption</td>
<td>73</td>
</tr>
<tr>
<td>3.5 Is Ridge Penalty Enough for the B-spline Basis?</td>
<td>74</td>
</tr>
</tbody>
</table>

### 4 Results with B-Spline and 3 dimensional Wavelets

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Wavelet Basis</td>
<td>79</td>
</tr>
<tr>
<td>4.1.1 Wavelets: Introduction</td>
<td>80</td>
</tr>
</tbody>
</table>
4.1.2 Orthogonal Wavelet Basis and Multiresolution Analysis ........ 81
4.1.3 Discrete Wavelet Transform ...................................... 85
4.1.4 3D Wavelet Basis .................................................. 86
4.1.5 Wavelet Thresholding ............................................... 88

4.2 Finger Opposition Data: Methods .................................... 90
4.2.1 Data and the Standard t-Test Analysis ............................. 90
4.2.2 Two-way Classification with TPS: Internal Optimization and Scan Normalization ................................................. 90
4.2.3 Two-Way Classification: External Optimization ................. 91
4.2.4 Deriving Time and Spatial projections .............................. 91

4.3 Finger Opposition Data: Results ..................................... 92
4.3.1 Two-way Classification and Internal Optimization ............... 92
4.3.2 External Optimization with Different Bases ....................... 94
4.3.3 Applying PDA to an Eight-Class Problem: State and Temporal Changes .............................................................. 102

4.4 Result with Wavelet Expansions ..................................... 105

5 Static Force fMRI Analysis ............................................. 112
5.1 Modeling the time series effects: Time-Smoothed PDA ............. 112
5.2 Introducing Between-Scan Smoothness within the Discriminant Framework ......................................................... 113
5.3 The O(N) Algorithm for Time-Smoothed Penalized Discriminant .... 115
5.3.1 Constructing the Second-Order B-spline Penalty Matrix ........ 117
5.4 Connections with Canonical Correlation Analysis and MANOVA .... 118
5.5 Penalized Discriminant Analysis of StaticForce data in B-splines and Wavelet domains .................................................. 123
5.6 Applying Time-Smoothed PDA model to the StaticForce data ....... 125
6 Conclusions and Extensions 128

6.1 Extending the predictive analysis ................................................. 129
6.2 Comparing the Results Across Non-Predictive Paradigms ................. 131
6.3 Wavelets and Basis Selection Techniques ....................................... 132
6.4 Inference and other issues ........................................................... 134

A Tensor Product B-Spline Basis 137

B Basis Expansion of Canonical Variates 139

C CCA via Regression 141

D Correspondence Between CCA and LDA variates 143

E Deriving Predictions in the n-Dimensional Space 145

F Ridge Regression With the Outer Product Matrix 146

G Centering the Design Matrix 149
List of Tables

1.1 Some common acronyms used throughout the thesis .......................... 11

3.1 Estimate of Effects in the four ANOVA models of the simulation results.
   The terms are input dimensionality. $P = \{5, 30\}$ as compared to $P = 10$. training set size $N = 50$ compared to $N = 100$. number of subjects. $Subjects=\{5, 10, \"N\"\}$ compared to 1 subject and ratio of variances of subject effect to error effect. $VarRatio=\{10, 100\}$ compared to $VarRatio=1$. There is also an interaction term between $P$ and $N$. .......................... 56
List of Figures

3.1 Demonstration of 2-class LDA in 2 dimensions. The light points (class 1) and darker points (class 2) show the 200 bivariate Gaussian observations generated from each class. The solid line is the true canonical variate (CV). The circles are class means, and the diamonds are the means projected onto the CV. The points marked with the cross represent the test point and its projections onto the mean-difference (broken) and CV lines. ........................................ 47

3.2 The orthonormal basis functions (columns of $U$) from the 2-D tensor-product B-spline expansion. There are $5 \times 5$ B-spline basis, which were diagonalized with SVD and displayed here in the order of decreasing eigenvalues. The penalties shown for each basis were calculated with $\lambda = 1$ .................................................. 77

4.1 Haar (left) and Daubechies Symmlet wavelet functions. The detail level grows from bottom up, and only some integer translates are drawn at each level ................................................................. 84
Internal optimization of the ridge tuning parameter (Sec. 4.3.1), expressed as a number of Effective Degrees of Freedom, for Penalized Discriminant Analysis in two-class problem. This example uses the data projected on the B25 tensor product B-spline basis set. The curves show the change of Prediction Error (both Misclassification (MC) rate and Squared Prediction Error (SPE)) as a function of Effective Degrees of Freedom (EDF). Both cross-validation (CV) and .632Bootstrap estimates are exhibited. The top left panel shows changes for un-normalized data, while the bottom panel deals with mean-normalized data. The top right panel shows CV and .632Bootstrap SPE curves for mean-normalized data which were obtained with a different random seed for every EDF; these portray the greater variability of CV estimates. Thin lines — CV estimates. Thick lines — .632+Bootstrap estimates. solid lines — estimates of SPE. dashed lines — estimates of MC rate.

Prediction Error (PE) curves as a function of the effective degrees of freedom (EDF) for all tensor-product B-spline (B15–B35) bases and unprojected raw (Braw) and smoothed (Bsmooth3.5) scans. The upper panel shows Squared Prediction Error (SPE=(1 − \( \hat{p}_C^2 \)), where \( \hat{p}_C \) is the estimated posterior probability for the true class), and the lower panel depicts Misclassification rate (MC rate) as a percentage of the total number of scans misclassified. The images from un-projected data are shown with larger-width curves. The markers show the minima for each curve. (The minimum for the B30 curve is not shown as it occurred beyond the figure frame).
4.4 Functionally activated $[^{15}\text{O}]$water PET voxels above the 93.4 percentile (white overlay) interleaved with registered grayscale MRI brain slices for Penalized Discriminant Analysis of: (A16-A40) unprojected raw data presmoothed with a $3 \times 3 \times 3$ voxel boxcar kernel (Bsmooth3): (B16-B40) unprojected raw data without presmoothing (Braw): (C16-C40) tensor product spline basis with 15 spline bases in each spatial dimension (B15): (D16-D40) tensor product spline basis with 35 spline bases in each spatial dimension (B35) (activation images A to D have decreasing squared prediction error (SPE) values as illustrated in Figure 4.3): (E16-E40) is a pooled standard deviation t-test image of scans presmoothed with a $3 \times 3 \times 3$ voxel boxcar kernel — the Bonferroni t-value ($t=4.65$) at the 93.4 percentile was used to define a conservative activation threshold with which to compare activation image peaks (white overlay) for a fixed number of voxels. PET and MRI slices are $128 \times 128$ with 3.1 mm pixels with center-to-center slice spacing of 3.4 mm (i.e., slice A23 and A26 are separated by $(26 - 23) \cdot 3.4 = 10.2$ mm) and are parallel to the AC-PC plane, which coincides with slice 24. Image left = brain left.
4.5 Scatter plot of pairs of activation image values for all Talairach brain voxels (1 point/voxel) for a single-voxel t-test image using a pooled standard deviation estimate, compared to penalized discriminant analysis (PDA) of a tensor product spline representation with 35 B-spline bases along each spatial dimension (B35). The dashed line depicts the principal axis from a principal component analysis of the scatter plot distribution. The circle highlights a group of voxels in the primary visual region that have moved from the 20th percentile in the t-test image to the 90th percentile in the PDA image. The solid vertical line depicts the Bonferroni t-value (t=4.65) at the 93.4 percentile of the t-test distribution of voxel values (white overlay: row E of Figure 4.4) and the solid horizontal line reflects the 93.4 percentile (value=0.0065) for the PDA distribution of voxel values (white overlay, row D of Figure 4.4).

4.6 Square Prediction Error (SPE) curves, in an 8-class problem, as a function of Equivalent Degrees of Freedom for 8 penalized discriminant models with different representations: 5 tensor-product B-spline projected datasets with varying numbers of basis functions (B15 to B35) and 3 unprojected raw (Braw) and smoothed (BSmooth3 and BSmooth5) datasets (thicker lines). The markers show the minima for each curve.

4.7 Projecting the data on first two canonical images obtained by Penalized Discriminant Analysis of the 8-way classification problem. The points are labeled according to the class (1: first baseline, 2: first active, 3: second baseline, etc), and the class means are shown in circles. This figure was obtained from tensor-product B-spline projected data using 35 B-spline basis in each dimension (B35) model with λ corresponding to the minimum Squared Prediction Error (SPE), but is similar across basis and λ's.
4.8 Top panel compares wavelet and B-spline results in the 2-class problem. Shown are Daubechies order 2 thresholded wavelet basis compared to raw and B-spline representations using .632+ Bootstrap estimate of squared prediction error. The bottom panel shows SPE curves for the two-class problem for various wavelet families. We compare Daubechies order 2 and 3 family and order 2 Coiflet system. For each family we investigate two thresholding strategies: simply “peeling off” one finest detail level in each dimension (32 × 32 × 16), and Donoho and Johnstone VisuShrink hard thresholding rule.

4.9 Visual comparison of Wavelet (top row) and B30 representations in the 2 class problem. First three slices show portions of the cerebellum. next two display the midbrain portions, and the last three slices depict the activation of the cortex. The grayscale image is the anatomical MRI scan in the Talairach space and the CV is overlayed on top of it using the hot-metal color coding. Both images were created using EDF that minimized the SPE: 74.6 for Wavelet vs 53.6 for B30.

4.10 Comparing the B-splines (B30) and wavelets (Daub2Thresh) Canonical Images using a corner-cube environment of Rehm et al. (1998) in the two-class setting. Except for the three major overlapping regions, the foci have been fit inside a ball of the same volume as that of corresponding focus. Blue foci correspond to B30 Canonical Image.
4.11 Squared Prediction Error for various wavelet families in the 8class problem.

Three wavelet functions are investigated: Daubechies order 2 and 4, and Coiflets order 2. For each family, we either remove top-scale wavelet coefficient level, resulting in $32 \times 32 \times 16$ wavelet coefficients or we apply D&J VisuShrink hard thresholding (Thresh). As a further dimension reduction technique, we investigate using all 7, or first 2 CV's to perform classification ($D7$ vs $D2$).

5.1 Projections on first 2 CVs from the PDA model applied to

5.2 Projections of and time-points (first row) and force level means onto the first four Canonical Images using the time-smooth PDA model with B25 Tensor Product B-spline basis and B-splines for the time axis. Force levels were: 1–baseline. 2–200g. 3–400g. 4–600g. 5–800g. 6–1000g. The time-structure penalty hyperparameter was set at $\lambda_Y = 10$.

5.3 Selected slices of the third canonical image resulting from applying the time-smoothed PDA model to the StaticForce data with B25 basis. EDF=50 and $\lambda_Y = 10$.

A.1 1D and 2D B-spline basis.
Abstract

We extend a classical multivariate technique: Linear Discriminant Analysis (LDA) and apply it in the analysis of PET and fMRI images of human brain function to discover regions of activation driven by the experimental stimuli. We re-examine and specialize some equivalences between LDA and: Canonical Correlation Analysis (CCA) and Multivariate ANOVA (MANOVA). Furthermore, efficient algorithms are derived to facilitate applying these multivariate models to extremely large image data. We deal with the ill-posed nature of the problem using spatial basis expansion and the penalization (with Penalized Discriminant Analysis (PDA) of Hastie et al. (1995)), and utilize efficient measures of predictive performance to optimize hyperparameters and validate the models in a robust fashion. We examine expanding the images into a 3D tensor-product B-spline and Wavelet basis and compare to the results obtained without expansion. Some parallels between our proposal and some of those currently popular in the neuroimage community are discussed. Another extension to PDA is derived and applied that allows one to model time series effects that exist in fMRI images. We conclude with many possible enhancements to the proposed paradigm.
Acknowledgments

This work has been a truly interdisciplinary effort and therefore owes immensely to a large number of people. I would like to thank the whole \VAPET team in Minneapolis for helping me in every way, fielding my questions and massaging the image data in some unorthodox and inconvenient ways. I would particularly like to thank Mr. Jon Anderson and Mr. Kirt Scharper for their data and system help.

I would also like to express my gratitude to Dr Nick Lange from Harvard Medical School. Nick has provided me with the many insights, data and some financial backing that has all been very important in completing this thesis. I truly hope that we will be able to work together more in the future.

No thesis work is possible without the guidance and encouragement of a supervisor. I have been most lucky with two such persons: prof. Rob Tibshirani whom I owe a huge debt of gratitude for introducing me into modern statistics, inspiring me to pursue excellence worthy of such a supervisor and instilling in me a belief in experiment and enthusiasm for unorthodox approaches to data. On the other hand I have enjoyed his unconditional trust and support in choosing the path I felt appropriate. There was so little hand-holding, even though I yearned for it at times; only recently have I been able to appreciate his restrain to lead me and his desire to let me discover my own way in the exciting and ever-growing, but definitely not simple, world of statistics.

This thesis also owes its existence to prof. Steven Strother, PI of the PET group in the \VA medical center in Minneapolis. Steve has given me enormous support and guidance
throughout our cooperation. He has taught me how to work with complicated data, how to interpret the results of analysis, how to write papers and fight for them later on, and has given me moral support all way long. His enthusiasm was crucial at times. I have also learned from him how to look beyond the beauty of mathematical models, how to accept the sometimes brutal truths spoken by the data which more often than not force us to abandon complication and retain simplicity.

I would also like to offer special thanks to all faculty in the Departments of Statistics and Biostatistics at University of Toronto, and in the Department of Statistic in Stanford University. I would especially like to thank prof. Paul Corey in Toronto for his support and encouragement, prof. James Stafford in Toronto for his support and enthusiasm, and prof. Trevor Hastie in Stanford for his suggestions and opinions. In particular, material in sections 3.5 and 5.2 has been suggested by Trevor. I also thank him for providing me with an excellent display of wavelet functions (figure 4.1). I am greatly indebted to my friends and colleagues, both in Toronto and in Stanford: Ilana Belitskaya, Annie Dupuis, Celia Greenwood, Carmen Mak, Mike Manno, Jorge Picazo, Bogdan Popescu, and Jiaming Sun: thank you!

Last, but certainly not the least, my love and deep appreciation goes to my family. I want to thank my wife, Kasia, who has been so patient with me, my brother and his family for love and support, and my parents who have always believed in me in the most difficult of times.
Chapter 1

Introduction

1.1 Images as Data

Data comes to us in various forms: increasingly, it is collected, stored and analyzed in the form of images. To me this is not very surprising: humans receive most of their information about the world surrounding us, as a visual input. It is only natural that we upgrade the position of this medium in the statistical analysis domain.

There are various reasons why it is only in recent years that the image data has become increasingly common. Most important are technological advances on many fronts. To store and process any data, be it sound, image, or categorical attributes, we need, as far as our current information processing techniques go, a numerical representation. The methodology for acquiring images and converting them to numbers, a process known as scanning or image digitizing, has become very accessible. Moreover, many medical instruments have been digitized and already store and process images in digital form. Other technological advances had to be made in the storage domain: image data is very byte-consuming. And here, again, much has been done in recent years, mainly to meet the demand of multimedia-savvy consumers. Hard disks of ever-increasing capacity and speed, and other forms of computer storage — most recently DVD that stores over an hour of good quality
video on a laser disk — have become very accessible.

As important, although less well known, have been the advances in image compression and processing. To appreciate their importance, one needs to understand the sheer size of image data after they have been converted to numerical representation or digitized. After digitizing, the image is viewed as a large number of indivisible components called pixels (picture elements), or voxels (volume elements) when the image is three dimensional. The more pixels there are the better resolution of the digitized image. To represent color information in each pixel, one usually needs three numbers for each of Red, Green and Blue colour channels. The number of available colors is governed by a maximum allowed value of each such number which is in turn connected to the number of bits allowed for each number. Today 24 bits per pixel (or 8 bits per number), so called “TrueColor”, giving \(2^{24}\) or almost seventeen million colours possible, is an industry standard. To give some idea about the storage requirement, a typical 640 by 480 TrueColor image needs 921,600 bytes, almost a megabyte, to store its 307,200 pixels.

It is obvious that storing images needs compression. The general compression algorithms are not optimal since they do not take into account the spatial structure of images. Better are methods that use the information about special meaning of the bytes and incorporate smoothness assumptions. Two such methods have become very popular: GIF and JPEG. Both are capable of reducing the size requirement many fold. Increasingly, wavelets are being tried for this task (Antonini et al., 1992).

The old saying: “picture speaks a thousand words” has a special meaning to statisticians and biostatisticians. After all, one of the main goals of statistics is to extract and summarize the information in the data, separating it from noise and error. If images indeed carry so much more information, then we should be particularly interested in adapting, extending and developing statistical methodology for analyzing them. Statistics, however, has been slow in joining other fields that have become heavily involved in image analysis, such
as the Machine Learning community. There are some notable exceptions, such as Duda and Hart (1973), who introduced some statistical methodology, mainly in the classification domain, to the pattern recognition community: Ripley (1981) who has summarized existing spatial statistics techniques; or Cressie (1993) who has provided a theoretical framework for mathematical image processing. These are just a few examples, of course, and many more could be found. The point is, however, that the statistical analysis of images is still not a well-established division in our field.

The Machine Learning community, which has already provided statisticians with many new exciting models and methods, such as Neural Networks (Hertz et al., 1991), Support Vector Machines (Vapnik, 1995), and very promising Boosting learning (Freund (1995). Freund and Schapire (1996), but see also Friedman et al. (In Press)) has been much quicker to approach image data. Most of the examples there, however, deal with building predictive models that use images as inputs: for instance character and zip-code recognition models. Work in the direction of statistical understanding and analyzing image data has been much less prominent.

There are various possible reasons why statisticians have been reluctant to consider image data. One is the difficulty in working with images: it requires considerable computing sophistication. Images are stored in many formats and easy to use libraries for reading, writing and transferring between them are not readily available. The problem of huge image sizes, where one image may be many times larger than a typical whole dataset used in statistics, requires much larger computational resources and special programming approaches. Another reason may be the total inadequacy of many common statistical tools. Let us take a simple example when the $n \times p$ input data matrix, $X$, consists of $n$ images, each with tens or hundreds of thousands ($p$) of pixels, and when each image is associated with one or more numerical responses, $y$. Now, if we think of running regression of $y$ onto $X$, where $n$ is many times smaller than $p$, we immediately see the difficulty with
such data. Most of the asymptotic results we use, increasingly questioned even with the "regular" data, are totally inadequate here. Working with images requires a new way of statistical thinking that questions and examines all the assumptions we have come to rely on in statistics.

1.2 Motivation and the Setup for Statistical Image Analysis

This work deals mostly with methods of analysis of medical images that have been acquired under various experimental conditions. Imaging techniques in medicine have been increasing in importance ever since the introduction of X-ray imaging. There are now many modalities that are routinely used to gather visual information in vivo about the workings of our organism. In this dissertation I center on the neuroimage data: scans of the living brain that represent patterns of neuronal activation. The techniques introduced, however, have, in my opinion, a much greater application. In this section I present a general experimental situation which I view as a basis for the methodology presented in the later chapters.

Let us imagine that we have $S$ subjects and that several images have been acquired from each of them. Suppose further that the images have been obtained under various conditions: either induced experimentally, or sampled from the population. For instance we may have "normal" and "sick" patients, or the patients may be asked to perform certain tasks. In the first case we have that each patient is uniquely assigned to one of the conditions, in the second case we have a blocked design with subjects as blocks. There could also be other variables collected in the patients, either for each image separately, or once per subject, but such variables are of secondary interest and would be used in the analysis to control
for confounding. Each observation could be presented as follows:

\[ \{ i_{sr}^{(k)}, x_{sr}^{(k)} \} \]

Here, and throughout the document, \( i \) will denote an image. The indexes \( s \) and \( r \) denote subjects and repetitive scans acquired from a given subject, respectively, while \( k \) refers to the condition or state under which the image was acquired. The image data may be supplemented by additional measurements, \( x \) as mentioned above.

It is assumed that the goal of the analysis is to estimate the "differences" among the images that were acquired under various conditions. I use the quotes on 'differences' to stress the fact that I do not necessarily mean algebraic differences, but any measure of disparity. We are thus required to summarize that part of the variability in the images that was induced by the conditions. The important goal of any descriptive analysis of experimental data is to provide some decomposition of that part of variability that is associated with the experimental setup. Furthermore, it is desired that the results be in the form of one or more images, with important measures (such as percentage of explained variability, assessment of the type of variability explained by each summary image, etc) attached. This is quite different from the predictive goal of many AI methods: we would not be content with a black box that takes the images as an input and predicts their condition, \( k \), for example.

One of the difficulties alluded to in previous sections is an atypical data setup. In the general framework introduced above, we have independent observations, \( \{ i, x \} \) that have huge apparent dimensionality equal to the number of pixels (plus whatever extra variables \( x \) we measured), while the number of such observations is many times smaller. This is referred to as an extremely ill posed problem (e.g., Lautrup et al., 1995). This is one of the reasons that the usual inferential statistics based on asymptotic results cannot be applied. The motivation of this work was to develop a framework that would be testable with as few distributional assumptions as possible. To this end, I utilize measures of predictive performance as a goodness of fit assessment that are robust across distributional
assumptions.

1.3 Functional Data

Image data may be thought of as a two and three dimensional extension of functional data: data that is realized from observing smooth functional processes discretized on a common lattice. Ramsey and Silverman (1997) have surveyed and extended common linear statistical methods for such data. They develop functional alternatives to Principal Component Regression, General Linear Models, including MANOVA, Canonical Correlation and Linear Discriminant Analysis, among others.

The starting point for all the models described in the book is the definition of the functional inner product. Since inner products are the workhorse of all the linear methodology in statistics, proving the right definition for functional data results in a functional equivalent of the model. The authors settle on the usual $L^2(\mathbb{R})$ Hilbert space inner product:

$$\langle x(t), y(t) \rangle = \int_{\mathcal{T}} x(t)y(t)dt$$  \hspace{1cm} (1.2)

where $\mathcal{T}$ is the domain of the data.

Of interest for this thesis is the functional approach to Canonical Correlation Analysis (CCA). The classical CCA may be stated as a maximization problem (see also Eq. 3.38):

$$\arg\max_{a,b} \frac{a^TS_{xy}b}{\sqrt{a^TS_{xx}a \cdot b^TS_{yy}b}}$$  \hspace{1cm} (1.3)

where $S_{..}$ are appropriate covariance and cross-covariance matrices for $x_i, y_i$, the $N$ observed pairs for which we seek to fit the CCA. If we now assume that what we observed are pairs of functions, $x_i(t), y_i(t)$ we may define variance operators, e.g.:

$$(V_{xy}f)(s) = \int_{\mathcal{T}} v_{xy}(s,t)f(t)dt$$  \hspace{1cm} (1.4)
where \( v_{xy} \) is a covariance kernel:

\[
v_{xy}(s, t) = N^{-1} \sum_{i=1}^{N} X_i(s)Y_i(t) \tag{1.5}
\]

With these definitions the functional CCA criterion is posed:

\[
\arg \max_{\zeta, \eta} \frac{\langle \zeta, V_{xy}\eta \rangle^2}{\langle \zeta, V_{xx}\zeta \rangle \cdot \langle \eta, V_{yy}\eta \rangle} \tag{1.6}
\]

where \( \zeta, \eta \) are functions belonging to Hilbert space defined by the inner product \((L(\mathbb{R}))^2 \) for the one defined in (1.2)). To obtain unique and interpretable solutions the inner products in the denominator are modified via penalization. Typically a second derivative penalty would be used. If one denotes the second-derivative differential operator by \( D^2 \), we can modify the criterion (1.6):

\[
\arg \max_{\zeta, \eta} \frac{\langle \zeta, V_{xy}\eta \rangle^2}{\langle \zeta, V_{xx} + \lambda_x \|D^2\zeta\|^2 \rangle \cdot \langle \eta, V_{yy} + \lambda_y \|D^2\eta\|^2 \rangle} \tag{1.7}
\]

which, under mild regularity and boundary conditions, is equivalent to:

\[
\arg \max_{\zeta, \eta} \frac{\langle \zeta, V_{xy}\eta \rangle^2}{\langle \zeta, (V_{xx} + \lambda_x D^4)\zeta \rangle \cdot \langle \eta, (V_{yy} + \lambda_y D^4)\eta \rangle} \tag{1.8}
\]

using the fourth-derivative operator, \( D^4 \).

Now, that the criterion is posed, what remains is to develop an algorithm to optimize it, given the observed data. One possibility pursued in Ramsey and Silverman (1997) is to use the basis expansion step which permits the use of standard tools for functional data. That is, given a system \( \{\phi_k(t)\}_{k=1}^{K} \) we assume:

\[
x_i(t) = \sum_k c_{ik}\phi_k(t); \quad y_i(t) = \sum_k d_{ik}\phi_k(t) \tag{1.9}
\]

(There is nothing in the following algebra that requires the use of the same system for \( x \) and \( y \); in fact for the linear discrimination only \( x \)'s are expanded). In place of operators \( V_{xx} \) etc, one has matrices expressing the covariance in the smooth basis domain:

\[
[V_{xx}]_{jk} = \sum_i c_{ij}c_{ik}; \quad [V_{yy}]_{jk} = \sum_i d_{ij}d_{ik}; \quad [V_{xy}]_{jk} = \sum_i c_{ij}d_{ik} \tag{1.10}
\]
The covariance matrix for the basis is simply $J_{jk} = \langle \phi_j(t), \phi_k(t) \rangle$. Similarly, the penalty matrix is $K_{jk} = \langle (D^2 \phi_j)(t), (D^2 \phi_k)(t) \rangle$. With these definitions we can show that the penalized criterion (1.7) becomes:

$$\arg\max_{\mathbf{a}, \mathbf{b}} \frac{\mathbf{a}^T J \mathbf{V}_{xy} J \mathbf{b}}{\mathbf{a}^T (J \mathbf{V}_{xx} J + \lambda_x \mathbf{K}) \mathbf{a} \cdot \mathbf{b}^T (J \mathbf{V}_{yy} J + \lambda_y \mathbf{K}) \mathbf{b}}$$  \hspace{1cm} (1.11)$$

where the left and right canonical variates now are:

$$\zeta(t) = \sum_k a_k \phi_k(t): \quad \eta(t) = \sum_k b_k \phi_k(t)$$  \hspace{1cm} (1.12)$$

To see that, let us look at one component of the denominator, $\langle \eta, \mathbf{V}_{yy} \eta \rangle$, without penalty at first.

The covariance kernel is:

$$v_{xx}(s, t) = N^{-1} \sum_i X_i(s)X_i(t)$$  \hspace{1cm} (1.13)$$

$$= N^{-1} \sum_i \sum_{\nu, \nu'} c_{i\nu} c_{i\nu'} \phi_\nu(s) \phi_{\nu'}(t)$$  \hspace{1cm} (1.14)$$

$$= N^{-1} \sum_{\nu, \nu'} \left[ \sum_i c_{i\nu} c_{i\nu'} \right] \phi_\nu(s) \phi_{\nu'}(t)$$  \hspace{1cm} (1.15)$$

$$= N^{-1} \sum_{\nu, \nu'} \left[ \mathbf{V}_{XX} \right]_{\nu\nu'} \phi_\nu(s) \phi_{\nu'}(t)$$  \hspace{1cm} (1.16)$$

Hence:

$$(v_{x} \eta)(t) = \int v_{xx}(s, t) \eta(t) dt$$  \hspace{1cm} (1.17)$$

$$= N^{-1} \sum_{\nu, \nu'} \int \left[ \mathbf{V}_{xx} \right]_{\nu\nu'} \phi_\nu(s) \phi_{\nu'}(t) \sum_\rho a_\rho \phi_\rho(t) dt$$  \hspace{1cm} (1.18)$$

$$= N^{-1} \sum_\nu \phi_\nu(s) \left[ \mathbf{V}_{xx} \right]_{\nu\nu'} \sum_\rho a_\rho \left[ \mathbf{J} \right]_{\nu\rho}$$  \hspace{1cm} (1.19)$$

$$= \phi^T(s) \mathbf{V}_{xx} J \mathbf{a}$$  \hspace{1cm} (1.20)$$

And similarly:

$$\langle \eta, v_{xx} \eta \rangle = \int \eta(s) v_{xx} \eta(t) dt$$  \hspace{1cm} (1.21)$$

$$= \mathbf{a}^T J \mathbf{V}_{xx} J \mathbf{a}$$  \hspace{1cm} (1.22)$$
Thus, with penalty term added, we get the two components of the denominator of (1.11). The numerator can be derived in an almost identical manner.

The discriminant version of this is to use the class indicator matrix, $Y$ without any regularization, but to regularize the observations $x_i(t)$. Our approach has been simpler: as we show in Sec. 3.3.4 we only expand the canonical variate (here, $\zeta$) in the smooth B-spline and wavelet basis. As compared with the approach presented here, this does not necessitate fitting the basis to the data, to obtain coefficients $c_{ik}$, but merely projecting the data onto the basis. With least squares fitting, the difference is that we are using a non-orthogonal projection onto the space spanned by the basis functions; instead we orthogonally project observations $x_i(t)$ onto each basis separately. This means that it is not necessary to compute the basis covariance matrix, $J$, in our case. For orthogonal basis, like wavelets, the two approaches are identical. For non-orthogonal basis, like B-splines that we use, it would be a worthwhile experiment to compare our (simpler) model with that proposed by Ramsey and Silverman (1997).

1.4 Notation and Conventions

We will adhere to the typical statistical notation except for a few exceptions. Thus $\mathbf{x}$ will denote a dependent observation (column) vector, with one or more subscripts as required. For classes (or conditions) we will use superscripts in brackets. Thus $\mathbf{x}_{ij}^{(k)}$ will denote the $(ij)^{th}$ observation, with the meaning of subscripts explained at appropriate places, and obtained under the $k^{th}$ class or condition. In place of $\mathbf{x}$, we will use $\mathbf{i}$, if we are referring to the scan data. We intend to call the input images, scans, and usually reserve the word image to mean the result of some analysis, that lies in the same space as the input scans, and may thus be visualized in the same way. However, when the context makes the difference clear, we may sometimes use “scan” and “image” interchangeably for stylistic reasons.

We use **boldface** in formulas to distinguish vectors from scalars if there is a danger
of confusion. such as when both scalars and vectors appear. Upper case Roman (late alphabet: \(U, V, W, X, Y, Z\)) and Greek letters are used for matrices without boldface. In few places boldface upper case \(X, Y\) are used to denote random vectors. but, in general, we do not make notational distinctions between random variables (including vectors) and their realizations, unless it is necessary.

We have intended to limit and localize the use of acronyms. Some well-known ones (e.g., MANOVA) are used freely. Other acronyms used globally are explained in table 1.1.
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Stands for:</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
<td>Statistical technique, due initially to Fisher, for classifying multivariate observations into one of few populations by means of linear discriminant functions</td>
</tr>
<tr>
<td>PDA</td>
<td>Penalized Discriminant Analysis</td>
<td>Extension of LDA due to Hastie et al. (1995) that introduces general penalization of covariance matrix and provides an appealing algorithm with a penalized regression as a main component</td>
</tr>
<tr>
<td>CCA</td>
<td>Canonical Correlation Analysis</td>
<td>Another classical multivariate technique that, given observations with variables divided in two sets, finds &quot;left&quot; and &quot;right&quot; linear combination that exhibit maximum correlation. LDA can be seen as a special case of CCA</td>
</tr>
<tr>
<td>PET</td>
<td>Positron Emission Tomography</td>
<td>One of few tomographic imaging technique, especially useful for imaging of the brain. The PET camera picks up gamma rays emitted within the imaged organ by a previously injected radiotracer</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional Magnetic Resonance Imaging</td>
<td>Another imaging technique used by neuroscientists to study brain function. fMRI is a specialization of MRI that is able to measure relative concentrations of oxygenated blood</td>
</tr>
<tr>
<td>CV</td>
<td>Canonical Variate</td>
<td>The linear combination(s) that result in CCA and LDA. If the CV lies in the image space, or has been reconstructed using the B-spline or wavelet basis, we sometimes refer to it as a Canonical Image</td>
</tr>
<tr>
<td>SPE</td>
<td>Squared Prediction Error</td>
<td>The main measure of predictive performance that we use, defined as ( \text{SPE} = (1 - \hat{p}_C)^2 ), where ( \hat{p}_C ) is a posterior probability of the true class</td>
</tr>
<tr>
<td>EDF</td>
<td>Equivalent Degrees of Freedom</td>
<td>One way to normalize the ridge penalty hyperparameter by calculating the trace of the &quot;hat&quot; or projection matrix in ridge regression</td>
</tr>
</tbody>
</table>

Table 1.1: *Some common acronyms used throughout the thesis*
Chapter 2

Neuroimaging Data and Methods

2.1 Goals and Study Design in Neuroimaging

Neuroimaging is a relatively young discipline that attempts to study the workings of the brain and the central nervous system through imaging techniques (Frackowiak et al., 1997). The most important goal is to discover the functional organization of the brain: the networks of functionally connected structures in the brain specific to a given task or groups of tasks. It is postulated that the brain is organized in various, likely overlapping, networks that are connected by function rather than anatomically (Strother et al., 1995a, McIntosh et al., 1997). These networks, that can be observed as patterns of activation, come to life when the brain faces a specific challenge, and work together to deliver a response.

There are two opposing views of the brain organization. One depicts the brain as a monolithic black box of neurons. This view of massive parallelism has lead to the development of Artificial Neural Networks, which have become a very successful computational and modeling device, rather than a true model of the brain. The other view of precise localization of the areas within the brain has been supported throughout the century by a series of first anatomical, and then functional discoveries of specific regions in the brain starting with discoveries of language components in the brain by Broca, Wernicke and Lichtheim at
the end of the nineteenth century. The updated view lies, as it often happens in science, somewhere in between. One way to describe it (Strother et al., 1995a, McIntosh et al., 1997) is to consider functional networks of areas, where the networks are specific for the task. Therefore, on some level, we do have homogeneous areas in the brain, but it takes a system of these, not one, for the brain to process a task. The same regions are likely used in quite different situations when they will be connected in different networks. A similar view is espoused by the notions of functional segregation and functional integration (pp. 5, Frackowiak et al., 1997). The functional segregation concept refers to large number of spatially localized areas in the brain that work more or less independently. The integration idea refers to the global integration of these specialized areas in the face of the task. These two views are not exactly the same. If we assumed that some orthogonality needs to be imposed the functional networks’ concept would more easily correspond to the orthogonal networks of areas, while the functional integration/segregation description would be better served by the orthogonality among the specialized areas. It is not clear, however, that any orthogonality assumption is correct in describing the brain function.

To clarify some ideas I now provide an example using parts of the cortex responsible, to some degree, for our motor abilities. Most of the description here is taken from (Frackowiak et al., 1997, Ch. 11) and Noback et al. (1991). There are many parts of what is known as a motor cortex: primary motor cortex (MI), supplementary motor area (SMA), premotor cortex, also known as pre-SMA, secondary motor cortex and cingulate motor area (CMA). All of these correspond to the distinct Brodmann areas which are a meticulous division of a human brain done on the basis of the local properties of the brain cells (cytoarchitecture). It has been established that MI, pre-SMA and SMA contain multiple representations of the body, that is they are organized somatotopically. Another words, one can find mostly contiguous areas that correspond to all parts of our voluntary motor system, from toes to the tongue and facial muscles. The situation is quite complicated: the cingulate parts of the
motor cortex are still controversial, the pre-SMA and SMA areas seem to be composed of even smaller parts of some autonomy. and the function of secondary motor cortex remains largely unknown. There are also subcortical areas in the cerebellum and the ventral part of the thalamus, which contribute significantly to the functioning of the motor cortex. There are also parts of the somatosensory systems necessary for motor control.

Major research questions center on the functional significance and connectivity of all these, and other, related, systems. We would like to understand how the brain controls our musculature, how does it plan and execute movements, how is the working of the motor cortex of a fine pianist different from that of an average person. Can movements be divided into groups that correspond to distinct activity patterns? We are only beginning to tackle these questions mostly with neuroimaging techniques. The second half of the last, and all of the present century have provided us with a huge body of knowledge related to the anatomical arrangement of the brain. We thus know a fair amount about the major connections in the brain, but to enhance our understanding of this most important organ of ours, we must concentrate on its functional arrangement.

A common type of study design used in neuroimaging attempts to delineate the activation signal using two contrast states. The states are chosen in such a fashion that the difference in activation patterns will provide maximum information about a specific function of the brain. For example in our finger opposition (FOPP) PET data (Sec. 2.4.1) the baseline and activation states differ only in the presence of absence of paced finger movement; in particular the eyes are patched in both states and there is no auditory input except for the pacing signal in the active state. Similarly in the StaticForce fMRI (Sec. 2.4.2) experiment the subjects observe control lines during the baseline state to compensate for the visual stimulation due to the force level display in the active states.

In another kind of a study design one gradually varies a single parameter that relates to strength of a supposed neural signal; an example is the StaticForce dataset. This situation
is similar to the dose–response relationship that is commonly observed in pharmacological studies: here one would look for patterns of activity that change with the parameter. The change may be monotonic, linear or not, or may be abrupt at first (in a transition from baseline to active state) and not change much hereafter. Indeed, both kinds of activity may be related to two different neuronal patterns at the same time. For instance, Sadato et al. (1996) reports how bilateral primary motor area and contralateral ventral premotor cortex, among others, were equally activated during an active phase of finger movements of increasing complexity. On the other hand, the ipsilateral premotor area, also among others, has shown a linear increase with the movement complexity. We therefore observe at least two important networks of activation in this study: one associated with the movement itself, an executive network, and one responsible for processing and planning of movement which therefore has had an increased activity for complex tasks. Similar results have also been reported by Catalan et al. (1998).

2.2 PET and fMRI Modalities

The object of any neuroimage modality is to reveal the neuronal activity throughout the brain volume or part thereof. The two most commonly used whole–brain modalities are Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI). None of these actually record the neuronal spiking patterns: rather they go after a “proxy” measure whose correlation with the neuronal activity has been established.

2.2.1 Positron Emission Tomography

Positron Emission Tomography (PET) is a general imaging technique that is used for many purposes in medicine where an image of physiological function is required. PET modality is an improvement over other radioisotope–based modalities, such as single–photon emission
computed tomography (SPECT). The description in this section is based mainly on Ollinger and Fessler (1997).

PET works by counting the number of high energy (512 kV) photons emitted from the imaged organ. In summary, positrons are created when the injected radiotracer steadily decays: such decay produces a single positron which very shortly annihilates with an electron. The annihilation produces two 512 kV photons propagating in nearly opposite direction. The PET camera is able to detect single photons and synchronize two hits to establish that the two photons originated from the same annihilation. In this way discrete approximations to the line integrals of radiotracer density along many lines are computed and the 2D or 3D image of the activity obtained by the inverse Radon transform.

A PET camera has detectors made of crystals (usually bismuth-germanate) which convert a single high-energy 512 kV photon into about 2,500 light photons. These are then fed into Photo Multiplier Tubes (PMTs) which then change the light activity into electrical signals. Most of the scanners connect each block of small crystals (say, 7 × 8 array of crystals) into a block of fewer PMTs (say, 2 × 2 array of them). The crystals in the block differ slightly among each other which allows the camera to determine the one crystal in the block hit by a photon. The camera counts the number of events: a pair of photons hitting two crystals on opposite sides of the camera within a very short period of time, called the coincidence timing window, usually about 10 ns. These counts, after they have been processed by the inverse Radon transform, create the 3D image of organ function.

There are many simplifying assumptions and problems that decrease the signal-to-noise ratio of the data. First, it is assumed that the positron will annihilate an electron immediately after being emitted. It has been shown that the positron range is usually smaller than 1 mm, which is much smaller than the resolution of the scanner, and therefore ignored. Another assumption is that the annihilation will produce photons flying out in exactly opposite directions. It has been established that the divergence from collinearity
is on the order of one degree or less, and can also be ignored. The other problems are more serious and usually cannot be ignored. The first of these is attenuation: a decrease of photon's energy due to its interactions with body tissue and with outer shell electrons. The interaction with body tissue, a photoelectric interaction, while a big problem for SPECT, is negligible for PET due to the type of radiotracers used. The other type of interaction, Compton scatter, can be statistically corrected for in the image reconstruction process. This correction is possible because the attenuation experienced by the pair of photons is independent of the position of the annihilation event. To correct for attenuation one approximates the the probability of a single photon pair experiencing Compton scatter which depends on the total distance traveled, and then includes this correction in the image reconstruction step. The probability of a single photon, traveling along the line \( l \), not experiencing the Compton scatter is modeled by the following equation:

\[
P_l = \exp\left( - \int_l \mu(x) dx \right)
\]

where \( \mu(x) \) is the linear attenuation coefficient at position \( x \). This probability is usually approximated by obtaining two extra scans: transmission and blank. These are obtained by a line source of radiation rotating around the field of view of the camera with (transmission scan) and without (blank scan) the subject. The ratio of counted events for each possible line \( l \) approximates the the probability 2.1. The number of detected events in the regular emission scan is then corrected for the probability of scatter.

An additional way to correct for Compton scatter, so called scatter correction comes from the fact that the scattered photons have smaller energy. The energy can be measured, to some degree (to about 10% on most scanners with bismuth-germanate crystals) at the individual detectors and then the threshold established below which the events are not counted.

Compton scatter produces another undesired effect: it causes a deflection in the path of
the affected photon. Most of the time such a photon will not hit any detector; the unscattered complementary photon which will hit the detector is called a *single*. It is possible, given the large number of scattered photons, that two singles will hit the camera within the coincidence timing window, and therefore be erroneously counted as an event that occurred on the line joining them. Such undesired events are called *randoms* or *accidental coincidences*.

The last problem which needs to be corrected for is detector *deadtime*. This is due to the finite amount of time that the detector needs to process a hit: during this time the detector is not able to sense any other hits. The detector deadtime limits the maximum dose of the radiotracer that can be placed in the patient: the researcher will try to use the maximum dosage that will still not saturate the camera, but which provides enough events to make the discrete approximation, used by algorithms such as filtered backprojection, to line integrals, implicit in the Radon transform, viable.

The PET data can be collected in 2D or 3D acquisition modes. In 2D mode, the events are counted in slices physically determined by collimators: thin annular rings of tungsten called *septa*. This mode results in greater accuracy by decreasing the probability of scattered events and randoms, since many more of scattered photons originating in the 2D field of view will never hit the collimated detectors. However, 3D mode, in which the septa are retracted and all possible events are counted, has up to eight times increased sensitivity which leads to decreased image variance and/or lowered doses of radiotracer required. Until recently most of the PET data was collected in 2D mode, mostly due to lack of 3D reconstruction algorithms. However, 3D mode is now gaining a wide acceptance in neuroimage community.

The data collected by the PET camera are in the form of counts for each possible line in the field of view. To reconstruct the image of the radiotracer density inside the imaged organ one uses a computational approach to solve the inverse Radon transform,
called filtered-backprojection (FBP). After correcting for some or all artifacts, such as attenuation and scatter, one has, in the 2D scanning mode, the data in a form of photons emitted in a given line, indexed by depth and angle, and denoted by $\hat{M}_{\theta d}$. These counts constitute the input to the FBP. The goal is to estimate a 3D distribution function of an radioisotope, $\lambda(x, y, z)$ with values of line integrals available:

$$g_\theta(d) = \int_{\Omega(d, \theta)} \lambda(x, y, z) \, dx \, dy \, dz$$

FBP is a deterministic method that assumes we have observed perfect data. that is:

$$g_\theta(d) = \hat{M}_{\theta d}$$

over a discrete set of angles, $\theta$ and depths $d$. This is an instance of inverse problem and one specialized solution to this problem is the algorithm called filtered back-projection. The algorithm has been extended to full 3D reconstruction. There are also approaches that use distributional assumptions, mostly Poisson, but they have yet to gain widespread support.

**Functional Neuroimaging via PET**  
PET may be used for imaging many organs. In functional neuroimaging, where we want to obtain information about the neuronal activity, two possible radiotracers emerged: $[^{18}\text{F}]$fluoro-2-deoxy-D-glucose (FDG) and $[^{15}\text{O}]$water. FDG is a radiolabeled glucose and allows PET to show local glucose concentration in the brain. This in turn is able to show us the neuronal activity since increased neuronal activity is very quickly followed by the surge of glucose-rich blood. (Barinaga, 1997). Similarly, $[^{15}\text{O}]$water tracer allows PET to image the bloodflow in the brain. Blood is also thought to surge into active neural areas to meet the demand for oxygen needed for metabolism of glucose (there is, however, some controversy regarding the nature of metabolism of neurally active areas, see for example Buxton and Frank (1997) and Barinaga (1997)).
2.2.2 Functional Magnetic Resonance Imaging

We will first describe (non-functional) MRI, which is sometimes called anatomical MRI in the neuroimaging community, as it describes the static anatomy of an organ (the brain, for example) rather than the dynamic function. Most of the description here is based on the overview article by Wright (1997).

General Physics of MR Imaging

MR techniques are based on magnetic properties of atoms, called nuclear magnetic resonance, first observed in the forties. In medical imaging one uses, almost exclusively, the simplest atom: single proton hydrogen nucleus. A hydrogen atom may be thought of as a minuscule magnet with its two poles producing a local magnetization vector in a certain orientation. In the absence of any external magnetic field of significant strength, thermodynamic movement causes random distribution in the local magnetization vector directions which results in a net magnetization, \( M \), equal to zero.

MR machines used for human diagnostics apply a static field, \( B_0 \), of strength which is 5 orders of magnitude higher than the earth field (a typical MR machine has 1.5 Tesla (1.5T) static field which is about 20,000 larger than the earth field). The most visible effect of such a large static field is that it causes a small portion of hydrogen nuclei to align themselves in the direction of \( B_0 \), which we presume to be along the vertical axis, \( Z \), in a 3D reference frame. The process of alignment is not immediate, the net magnetization in the direction of \( B_0 \) has an exponential delay:

\[
M_z(t) = M_0(1 - \exp(-t/T_1))
\]  

(2.2)

\( M_z \) is a vertical magnetization at time \( t \) after the static field has been turned on, \( M_0 \) is the asymptote of this, and \( T_1 \) is a longitudinal relaxation time which is a property of a material studied. For example, at 1.5T, \( T_1 \) for grey matter in the brain is about 1000ms, while for white matter only about 650ms and even less for fat (260ms).
Assuming that there is a non-zero net magnetization component in the plane perpendicular to $\mathbf{B}_0$ field, i.e. in the $XY$ plane (which is not the case in a tissue without any magnetic influence: such component is introduced by the MR machine as described later), the strong static field, $\mathbf{B}_0$ causes the $XY$ magnetization component to rotate around the $Z$ axis as it tries to align itself with the static field. In the literature, this rotation is called precession. and its angular frequency is directly proportional to the strength of $\mathbf{B}_0$. This frequency is called the Larmor frequency. Any rotating magnetic dipole, such as a hydrogen proton, generates electrical current in the coil that is positioned perpendicular to the plane of rotation. This is the signal detected by the MR machine: the coils are made to resonate at the Larmor frequency of a proton to maximize the signal detected.

In general, any volume of a tissue that contains many protons will have not net magnetization in the $XY$, or transverse plane. This magnetization is induced in the MR machine with a RF pulse that rotates with the Larmor frequency in the transverse plane. Figuratively speaking, one may imagine the RF pulse as tipping over these dipoles that have aligned themselves with the static field. The RF pulse is applied long enough to tip the dipoles to the horizontal direction: with time the $XY$ magnetization component, while rotating around the $Z$ axis, will return to the thermal equilibrium condition where the only significant component is in the $Z$ direction. But an even larger component contributing to the decay of the $XY$ signal comes from the gradual lost of phase coherence among the precessing dipoles. After the RF pulse, the precessing dipoles will be in phase. Due to their heterogeneous physical environment they will precess with slightly different rates and as a consequence the phase coherency will be lost resulting in a diminishing signal. The associated exponential decay has a characteristic time constant, denoted by $T2^*$. To restore the phase coherency MR machines apply another magnetic pulse that induces a spin echo. Specifically, let the dipoles evolve for time $\tau$, when some phase discrepancy will be evident. Apply a short (as compared with $\tau$) magnetic pulse in a single direction (say $y$) in the
transverse plane (in practice one applies the pulse in the single direction of the transverse plane rotating with Larmor frequency). This pulse effectively "flips" the dipoles about the $y$ axis: those that were precessing faster and were "ahead" of $y$, now lag behind the same amount, and vice versa. The result of the refocusing pulse is that after time $2\tau$ the phase coherency will be restored, assuming that rate differences among dipoles do not change with time.

In practice, the precession frequencies of different dipoles do change with time, and, despite spin echo, the $XY$ signal will decline. The dynamics of this decline, taking into account the refocusing efforts, are modeled as an exponential decay with constant $T2$:

$$|M_{XY}(t)| = M_0 \exp(-t/T2)$$

(2.3)

As was the case for $T1$, the transverse relaxation time, $T2$, is tissue specific. For grey matter, white matter, and fat $T2$'s are: 106ms, 69ms and 60ms, (all at 1.5T) respectively. All the above considerations are captured in one equation discovered by Bloch in 1946, which describes the full dynamics of the magnetization field. $M(t) = (M_x(t), M_y(t), M_z(t))$:

$$\frac{dM(t)}{dt} = M(t) \times \gamma B(t) - \left(\frac{M_x(t)}{T2} + M_y(t)j\right) - \left(\frac{M_z(t) - M_0}{T1}\right)k$$

where $B$ is the total magnetic field applied. The first term describes the general precession dynamics ($\gamma$ is a gyromagnetic constant; for protons, $\gamma = 2\pi \cdot 42.6$ MHz/T) and the remaining terms deal with transverse decay (Eq. 2.3) and gradual alignment of dipoles with the static field (Eq. 2.2). After the signal declines to limiting levels, a short period is required for the system to return to the thermal equilibrium within the static field, before next the RF pulse is applied and the measurement process repeated. The total time between RF pulses is denoted by $TR$ and is usually in order of few seconds. The time between refocusing pulses is denoted by $TE= 2\tau$. 
Contrasts and Spatial Imaging in MRI

The MR signal measured, that comes from the precession dynamics is eventually used to produce the tissue images. Depending on the tissue under study different contrasts may be used: these are combinations of of $T1$ and $T2$ relaxation times. Depending on the time window during which the data is acquired, one weighs either of these more heavily in the resulting contrast. What is needed at this point is a way to spatially select regions for data acquisition to produce images. This is achieved in a few steps. Firstly, the “static” field $B_0$ in the $Z$ direction has a linear gradient. Since the precession frequency of protons in the transverse plane depends on the strength of the field, one may acquire the data in slices by applying RF pulses with different frequencies matched to the precession frequencies in thin slices along $Z$ axis. This ensures that the recorder signal comes mostly from the dipoles in a specific horizontal slice.

To locate the signal in the $XY$ plane, similar ideas are used. One introduces gradients in $X$ and $Y$ directions that also vary with time. The signal acquired up to time $t$, say, will be a 2D spatial Fourier transform of the total slice magnetization, sampled at the spatial frequency, $(k_x(t), k_y(t))$, the so-called $k$-space. The $k$-functions are integrals of the respective dynamic gradients over time, up to $t$. The image of the slice magnetization is reconstructed using inverse 2D Fourier transform.

Functional MRI

Functional MRI was developed by Ogawa et al. (1990a,b). The “functional” adjective refers to the novel utilization of MR technology for imaging physiological functions as opposed to static structures for which it was originally proposed.

Functional MRI takes advantage of the differing magnetic properties of hemoglobin that depend on whether it carries oxygen or not. Oxyhemoglobin is diamagnetic as are other tissues in the brain. Deoxyhemoglobin is paramagnetic and causes changes to the proton
molecules in the water within the blood and surrounding the blood vessel. This is called a Blood Oxygenation Level-Dependent (BOLD) contrast. The paramagnetic nature of deoxyhemoglobin is somehow “felt” by the water molecules “close-by” which amplifies the BOLD signal significantly. The change in magnetic susceptibility affects the distribution of Larmor frequencies of nearby photons, causing a much greater phase spread. This in turn causes a significant decrease in MR signal in affected areas and results in a magnetic contrast mostly dependent on the $T2^*$ time constant. Functional MRI produces images that show local concentrations of oxygenated hemoglobin in veins and capillaries. Since we believe that there is strong correlation between levels of oxidized blood and neuronal activation, BOLD images may be interpreted as images of the brain function.

2.3 Literature Overview

PET and, most recently, fMRI data have been analyzed by the a plethora of methods of ever-increasing complexity. The most important challenges seem to be:

**Huge input dimensionality:** each “observation” is an image composed of 30-500 thousand numbers. This leads to the “extremely-ill posed” situation (e.g., Morch et al., 1997) where the number of variables is much larger than the number of observations.

**Time series effects:** even if all we need is a pattern of neuronal activity which corresponds to the activity that we study, we know that we cannot obtain a truly repetitive experiment within a subject: the brain state changes throughout the experiment, as there is some learning, change in the environment, adaptation, and many other transient effects.

**Subject effects:** With multiple subject studies, needed to obtain results of some generality, we observe (Strother et al., 1995a) that the differences among brains produce effects which are much larger than the effect due to the stimulus under study.
**Spatial correlation:** The activity in nearby brain locations is correlated. This must be acknowledged in either the modeling stage or in hypothesis testing paradigms (e.g., Worsley et al., 1992) or both.

Many initial methods analyzed Regions of Interests (ROIs) which were a manually delineated using anatomical or other known regions in the brain (e.g., Clark et al., 1985). The average activity within each region was used as the input to any subsequent analysis. This resulted in a great reduction in input dimensionality as one would typically have a few dozen regions at most. This method has been mostly abandoned with the introduction of methodology to deal with extremely-ill posed problems and the corresponding software packages. The most fundamental criticism of ROI methodology is directed at the fact that manual and *ad hoc* ROI definitions imposes a strong, mostly anatomical, prior on the analysis.

Currently there are several groups of models used to estimate activation maps for PET and the closely related functional magnetic resonance imaging (fMRI) technique. Our categorization follows that recently proposed by Lange et al. (1999) in fMRI. First, consider techniques that explicitly incorporate the experimental state of the subject for each scan (e.g., baseline or activation) with possible additional explanatory variables such as neuropsychological performance measures, which correspond to the "interesting indicator/categorical variables" and "covariates" of the general linear model (GLM) approach of Friston et al. (1995). Simple subtraction of (possibly standardized) average images from two experimental states is the most widely used example of this approach (e.g., Fox and Mintum, 1989, Worsley et al., 1992), and ANOVA related methods for more than two states have been generalized by GLM. These methods attempt to find the activation pattern which is driven by (or which drives) the experimental or observed conditions: an imposed stimulus, motor task or abnormality.
The second category includes techniques such as principal component analysis (e.g. Friston et al., 1993, Strother et al., 1995b), which require no experimental brain-state information. In these methods, one attempts to explain the general variability with a set of independent or orthogonal components and then post hoc link some of these to the experimental conditions. The problem with these methods is that the variability is partitioned without any reference to the stimulus or experimental conditions, which are then "sought after" among the resulting components. The third wide category includes all the non-linear models such as neural networks (e.g., Kippenham et al., 1994, Lautrup et al., 1995, Morch et al., 1997) and very recently Volterra kernels within the GLM framework (Friston, 1998).

All three categories may be applied to two different spatial data representations that have evolved to deal with the highly ill-posed nature of the functional neuroimaging domain: namely the huge dimensionality of the input space, equal to the number of voxels in the image (e.g., 20 to 30 thousand in PET), compared to the available number of independent scans, which is typically only a few hundred. The most common spatial representation initially ignores this issue by analyzing individual voxels, or volumes of interest (VOI), as independent samples, generating a test statistic for each voxel (or VOI), and then post hoc allowing for simple local spatial correlations by using inferential tests based on random field theory to threshold the resulting statistical parametric maps (e.g., Worsley et al., 1992, Friston et al., 1995, Worsley et al., 1996). The second representation uses a data-driven basis, such as the one obtained from Singular Value Decomposition (SVD) of the input data matrix, to reduce the effective dimensionality of the modeling problem. This was introduced to PET for VOI measurements (Clark et al., 1985, Moeller et al., 1987, Moeller and Strother, 1991) and was then extended to voxel based $^{15}$O water studies (Lautrup et al., 1995, Strother et al., 1995a,b, Friston et al., 1996, Worsley et al., 1997).
2.3.1 Single Voxel Analysis: Statistical Parametric Mapping and Gaussian Random Fields

In this section I will summarize one popular method: Statistical Parametric Mapping (SPM) with Gaussian Random Field theory testing (Worsley et al., 1992. Friston et al., 1995. Worsley et al., 1996), a method that works separately with each voxel and then uses estimated spatial correlations to control the type I error in the multivoxel hypothesis testing.

With the simplest SPM setup, ones assumes that the scans come from two different conditions, say Baseline and Active (Worsley et al., 1992). A more general framework was provided in Worsley et al. (1996) where any (one) contrast could be used. Also in the newest incarnations SPM may come from any method that generates a 'Z', 't', 'F' or $\chi^2$ statistic at each voxel. We will keep to the simpler two condition situation but the extension to general contrasts is immediate. Let $i_{ijk}(x, y, z)$ denote the $k^{th}$ scan from subject $i$ ($i = 1, 2, \ldots, n$) under condition $j$ ($j \in \{A, B\}$). The normalized subject specific contrast images are formed:

$$\Delta_i^*(x, y, z) = \left(\bar{i}_{i,A}(x, y, z)/\bar{i}_{i,A}(\cdot, \cdot, \cdot) - \bar{i}_{i,B}(x, y, z)/\bar{i}_{i,B}(\cdot, \cdot, \cdot)\right)/\sqrt{2}$$

Thus the scans within each subject are averaged for each condition ($A, B$) and divided by the subject and condition specific constant $\bar{i}_{ij}(\cdot, \cdot, \cdot)$ which estimates the global blood flow. (Sometimes the scan-specific normalization by $\bar{i}_{ijk}(\cdot, \cdot, \cdot)$ is used before averaging over $k$ (Strother et al., 1995a. Appendix), to try to remove the scan global blood flow. Some normalization is necessary, especially for the PET measurements, as these are relative. The proper normalization methods are subject to some debate (Strother et al., 1995a. Appendix)). The $\sqrt{2}$ constant is used to keep the standard deviation of the difference the same as that of the original scans (Worsley et al., 1996), but it does not appear in the earlier versions of the method (Worsley et al., 1992). The contrast images are then averaged over
subjects to produce the mean difference image:

$$\Delta(x, y, z) = \sum_i \Delta_i^*(x, y, z)/\sqrt{n}$$  \hspace{1cm} (2.5)

where \( n \) is a number of subjects. Again, Worsley et al. (1992) uses \( n \) in place of \( \sqrt{n} \). Some estimate of standard deviation of \( \Delta \) is then used to normalize the mean-difference across voxels. There are many choices: the simplest, proposed in Worsley et al. (1992), is to calculate the subject-specific estimate for each voxel and then average over voxels:

$$\overline{S}^2 = \sum_{x,y,z} S^2(x, y, z)/V$$  \hspace{1cm} (2.6)

where \( V \) is number of voxels and:

$$S^2(x, y, z) = \sum_i (\Delta_i^*(x, y, z) - \Delta(x, y, z))^2/(n - 1)$$  \hspace{1cm} (2.7)

This assumes that the variance across voxels is the same. The other choices are not to pool across voxels, to pool over conditions using ANOVA or ANCOVA estimators (Friston et al., 1991, Worsley et al., 1996) or to combine the latter with the pooled estimator 2.6 (Worsley et al., 1996). The inherent dilemma is the low degrees of freedom available if no pooling across voxels is done (due to small number of subjects) and a strong assumption of homoscedasticity across voxels if such pooling is done.

The statistical t-map is formed by dividing the mean difference image 2.5 by an estimate of noise, which can either be an image itself (i.e. voxel-specific) or a scalar, as described above. Using the estimate 2.6, for example, the t-map is:

$$t(x, y, z) = \Delta(x, y, z)/\overline{S}$$  \hspace{1cm} (2.8)

This gives a t-statistic for every voxel. The problem is now to determine which of the many thousand t-statistics are significant designating neuronal regions with significant change between the conditions. Typically in PET images there would be about 30,000–40,000 intercranial voxels which are used to form the t-map, and therefore that many
t-statistics. By using the unadjusted significance level, $\alpha$, we will seriously overestimate the over-all significance, or inflate the type I error because of the multiple testing problem. Since there are large spatial correlations existing in the images and therefore in the t-map, the simplest Bonferroni adjustment method, which is most effective when the tests are independent would be very conservative decreasing the power to very small levels. One solution proposed in Worsley et al. (1992), generalizing and making more rigorous the ideas in Friston et al. (1991), is based on the theory of maxima of Gaussian Random Fields (Adler and Hasofer. 1976, Hasofer and Adler. 1978). A three dimensional Gaussian Random Field with mean $\mu(x, y, z)$ and covariance $C((x_1, y_1, z_1), (x_2, y_2, z_2))$ is a continuous stochastic process, $G(x, y, z)$, such that for any finite $n$ and for any selection of points $\{(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)\}$ the joint distribution of $\{G(x_1, y_1, z_1), \ldots, G(x_n, y_n, z_n)\}$ is n-variate Gaussian with mean $\{\mu(x_1, y_1, z_1), \ldots, \mu(x_n, y_n, z_n)\}$ and the covariance matrix obtained by evaluating covariance function $C$ at the $n^2$ pairs of points.

The main idea is to derive the single threshold, $t_\alpha$, such that under the null hypothesis:

$$P(T_{\text{max}} > t_\alpha) = \alpha$$  \hspace{1cm} (2.9)

where $T_{\text{max}}$ is a maximum t value. Using the Gaussian Random Field theory, the convenient null hypothesis is that if there is no difference between conditions, the t-map will constitute a zero mean, Gaussian noise, with the (scalar multiple of) identity covariance function, $C$. Remarkably, using the notion of Euler characteristic number, one can approximately evaluate probability 2.9 for any Gaussian Random Field, $G$ (Adler and Hasofer, 1976, Worsley et al., 1992, Eq. 1):

$$P(T_{\text{max}} > t) \approx V |\Lambda|^{1/2}(2\pi)^{-2/3}(t^2 - 1)e^{-t^2/2}$$  \hspace{1cm} (2.10)

Here, $V$ is a volume of the image, in some units, and the $\Lambda$ is a $3 \times 3$ variance matrix of
partial derivatives of the field in each dimension $x, y, z$, in the same units as $V$:

$$\Lambda = \begin{bmatrix}
\text{Var} \left( \frac{\partial G}{\partial x} \right) & \text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial y} \right) & \text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial z} \right) \\
\text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial y} \right) & \text{Var} \left( \frac{\partial G}{\partial y} \right) & \text{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial z} \right) \\
\text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial z} \right) & \text{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial z} \right) & \text{Var} \left( \frac{\partial G}{\partial z} \right)
\end{bmatrix}$$

(2.11)

The matrix of partial derivatives (2.11) constitutes a way to specify the covariance structure for the continuous and homogenous random field. The diagonal entries tell us how the field varies in the three axial directions, and the off-diagonal entries give variability in the three diagonal directions.

If $\Lambda$ were known, the Eq. 2.10 could be (numerically) inverted to find the desired threshold, $t_\alpha$. The covariance matrix $\Lambda$ may be approximated using numerical differences, which is, however, a poor and unstable estimate. Another solution is proposed in Worsley et al. (1992), which uses the known properties of a smoother which is applied to the scans. Using the assumption that under the null hypothesis, the t-map is a white Gaussian noise field, one can derive an expression for the covariance matrix of the white noise convolved with a kernel smoother. This is then used in Eq. 2.10 to calculate $t_\alpha$ which is then used to threshold the t-map and select the "significant" voxels.

SPM with Gaussian Random Field theory for determining the threshold has been a great step forward in the analysis of neuroimages. It has also been successfully applied in other fields, such as astrophysics. The theory is as remarkable and practical as it is beautiful. There are, however, a number of assumptions going into the SPM method, some of which have been addressed in later papers, which could cause problems in interpreting the results. The normality assumption may only be viable if there are a large number of degrees of freedom going into estimating the t-map. Typically, this is only the case when the voxel-wise variance estimators are pooled across voxels. This however leads to the possibly over-simplistic assumption of homoscedasticity across the brain volume. The other possible problem with the SPM method is a specification of the null distribution:
that of the white noise Gaussian field convolved with the smoothing filter applied to the t-map. This seems to disregard any possibility of spatial smoothness present in the t-map before pre-smoothing is applied, when we know that the hemodynamic response, which is actually measured by PET and fMRI, has an extent of 3-5mm (Malonek and Grinvald, 1996) and the reconstruction techniques themselves impose spatial smoothness. When the null hypothesis is rejected, we therefore still do not know, even upholding the normality assumption, whether the breach came because of the non-zero mean, which is the desired result, or because of the misspecified covariance matrix, \( \Lambda \). Most likely it is both, which, at best, leads to the inconclusive answer, and at worst may point to the totally wrong regions.

With the lack of realistic simulation studies that would examine the robustness of the SPM method under \( \Lambda \) misspecification, it seems probable to me that errors in estimating \( \Lambda \) may easily lead to the rejection of the null without any support in the mean.

### 2.3.2 Scaled Subprofile Model: State–Driven Variance Decomposition with Global and Subject Effect Removal

Scaled Subprofile Model (SSM) of Moeller et al. (1987). Moeller and Strother (1991) has been developed to identify regional variation produced by a treatment or a stimulus allowing for heterogeneous covariance patterns and subject effects. It has been specially formulated to deal with high-dimensional PET datasets obtained using a small number of subjects, and to work with a minimal set of assumptions regarding subject, treatment and residual covariance patterns. It strives to partition the variability (similarly to ANOVA model) to dissociate the subject and treatment covariance patterns.

The two main equations of SSM are:

\[
\begin{align*}
    i_i &= s_i(\mu + \alpha_i) \\
    \alpha_i &= \sum_k \gamma_{ik} \phi_k + \text{error}
\end{align*}
\]
Initially, (Moeller et al., 1987, Moeller and Strother, 1991) the index $i$ was meant to denote subjects in studies of one scan per subject, and the method was implemented for pre-designated regions in the brain. SSM was later (e.g., Strother et al., 1995a) successfully applied in the situation where index $i$ denotes a combination of subjects, treatment (stimulus) and repetition effects, and is therefore unique for each scan, and the voxels are used in place of regions. In the above equation, each scan is decomposed into global and residual images ($\mu$ and $\alpha_i$) which are called Group Mean Profile, GMP, and Subject Residual Profiles, SRP, respectively, in Moeller and Strother (1991). Each scan has a scaling factor, $s_i$ associated with it. The GMPs are further decomposed into a set of orthogonal Group Invariant Subprofiles (GISs) here denoted by $\phi_k$. This is related to a previously mentioned approach, where each voxel is separately modeled with ANOVA or ANCOVA, and the residuals then grouped back together to form residual images, which are then decomposed using SVD or similar techniques. Strother et al. (1995a) has listed similarities and differences between these approaches.

The main part of SSM was the development of a procedure for estimating various parts of the model. Moeller and Strother (1991) provide a detailed description which we will briefly summarize here. Scaled Subprofile Model can be approximately expressed as a voxel-wise, two-way ANOVA on a log scale:

$$\ln i_{ij} \approx \ln s_i + \ln \mu_j + \left\{ \frac{\alpha_i}{\mu} \right\}_j$$  \hfill (2.14)

where the index $j$ refers to voxels, and the division in the residual term is made voxel-by-voxel. Small signal approximation $\ln(1 + x) \approx x$ for $x \ll 1$ is used to derive the ANOVA correspondence, where $x = \left\{ \frac{\alpha_i}{\mu} \right\}_j$, for each voxel $j$. The estimation procedure assumes model (2.14). It starts by removing two main effects from the log-transformed scans by double-centering the log-scan matrix, $V_{ij} = \ln i_{ij}$. The resulting matrix is then decomposed using Singular Value Decomposition. The left- and right-hand eigenvectors may be shown
to be $\gamma_{ik}$ and:

$$
\left[ \frac{\phi_k}{\alpha} \right]_j - \left[ \frac{\phi_k}{\alpha} \right]
$$

(2.15)

respectively. One can then use K-variate regression of $N$ average log-scans, $\ln \bar{I}_i$, onto $\gamma_{ik}$ to estimate the K offsets in (2.15) and $\ln s_i$, which come up as regression coefficients and residuals, respectively. In fact, the regression only identifies the part of $s$ which lies in an orthogonal complement of the subspace spanned by $\gamma_k$: without an assumption of orthogonality between the two, the SSM model is not identifiable. From the regression results one can estimate the remaining terms.

SSM provides an intuitively appealing, log-linear model for the PET scans collected from various subjects. The parameters have the following physical interpretations: $s_i$ are scan-specific multiplicative factors which are related to global scan effects, both physiological and methodological, e.g., subject dose. The global radioactivity levels are very hard to control at the experimental stage: they are a results of a complicated interrelationship between the dose of radiolabeled agent, weight and other physical characteristic of the patient and unknown physiological effects that affect the distribution of the agent within the brain. The mean pattern $\mu$ represents a hypothetical brain state that is common to all scans. This may include a coarse description of regional differences that is invariable across scans and subjects. The scan specific variations, $\alpha_i$ represent patterns superimposed onto this mean brain state pattern. The resulting image of the sum of the mean and scan-specific patterns is normalized via the global scaling factors, $s_i$.

The main result of SSM consists of a set of residual patterns, $\phi_k$ together with their weights, $\gamma_{ik}$, which are scan-specific. One may show (Eq. 4 Moeller and Strother, 1991) that the total variance of log-transformed scans may be approximately decomposed into a global term, an error term and the residual profiles terms which are independent. One may therefore represent the $i^{th}$ scan's contribution to the total variance by the sum of squared weights, $\sum_k \gamma_{ik}^2$ for this scan. Also, if the overall index $i$ is broken into $i_s$ for subjects, $i_c$
for conditions and \( i_r \) for repetitions, the following decomposition of variance results:

\[
\sum_{i_s, i_e, i_r} \gamma_{i_s, i_e, i_r, k}^2 = \sum_{i_s, i_e, i_r} (\gamma_{i_s, i_e, i_r, k} - \bar{\gamma}_{i_s, i_e, i_r, k})^2 + \sum_{i_s, i_e, i_r} (\bar{\gamma}_{i_s, i_e, i_r, k} - \bar{\gamma}_{i_s, i_e, i_r, k})^2 + \sum_{i_s, i_e, i_r} \bar{\gamma}_{i_s, i_e, i_r, k}^2 \]  

since \( \gamma_{\ldots, k} = 0 \) for each \( k \). What the three pieces represent are between-condition, repeat-trial and intersubject variance contributions for each Subject Residual Profile, \( \phi_k \), which are uncorrelated. This allows us to study the particular contribution of each SRP, and thus determines whether it is mostly associated with the subject variances or the study design.

### 2.3.3 Partial Least Squares

McIntosh et al. (1996) propose another interesting multivariate method for the analysis of neuroimages, called Partial Least Squares (PLS) (their method is not related to the well-known regression model under the same name, described in (e.g., Wald et al., 1984)). The authors motivate PLS as a unique approach that results in the spatial patterns which optimally explain the covariance between a set of scans and the "exogenous blocks". The latter can be formed by contrasts of interest, or may include external measures: such as behavioural, performance etc. PLS is related to both simple t-maps and, conceptually, to the Scaled Subprofile Model described in the previous section. It is also related to LDA and hence to our proposal.

Let, as before, \( X \) denote an \( N \times p \) matrix with \( N \) scans each with \( p \) voxels. Let \( Y \) be the \( N \times K \) "exogenous block" matrix, with \( K \) blocks: contrasts or external measures. For instance, to follow an example in the paper, we may have a multisubject PET data obtained under 3 conditions: 1 baseline and 2 active. \( Y \) may contain two columns: the first, comparing the baseline to the average of the other two, with \( \{2; -1; -1\} \) for a scan in condition 1, 2, 3, respectively; and the second comparing the two active conditions with a \( \{0; 1; -1\} \) contrast.

The basis for the PLS method is an SVD decomposition of a cross correlation matrix.
$S = X^*T Y^*$, a product of column-centered and column-normalized $X$ and $Y$. That is:

$$S = A D V^T$$

with $A$: a $p \times K$ matrix of orthogonal singular images, and $V$: a matrix of orthogonal profiles, and the diagonal matrix $D$ with singular values. The $\mathbf{a}_1, \mathbf{b}_1$ pair of the first column vectors of $A, B$ gives the best linear approximation to explaining the cross-correlation matrix $S$. The first singular value $d_1$ gives the strength of this association: when squared and divided by the sum of all squared singular values it is a proportion of the total variance explained. One may examine the image $\mathbf{a}_1$, together with the first profile, to deduce the features of the experiment and the associated spatial map, which contribute most to the overall variability. McIntosh et al. (1996) also introduce a third measure: subject scores obtained by projecting individual scans onto the singular images. Thus for each singular image one obtains $N$ subject scores (which could more aptly be called scan or image scores) which can be plot against the conditions or external measures to gain further insight into the feature represented by a particular singular image.

PLS approach includes a calibration method to validate the model and determine the number of significant singular image/profile pairs. McIntosh et al. (1996) show that the covariance between the $j^{th}$ subject score: $\delta_x = X \mathbf{a}_j$ and the similar $j^{th}$ left hand score: $\delta_y = Y \mathbf{b}_j$ is $d_j$, the $j^{th}$ singular value. Thus PLS is finding the singular pairs which successively maximize the covariances between the left and right hand scores. McIntosh et al. (1996) propose computing the regression of subject score on the rows of $X$, the contrast or external measure matrix. As a measure of validity they proposed $R^2$, the proportion of variance explained by the regression. To determine a significant cut-off point, PLS uses a permutation test: rows of $X$ are permuted and for each permutation an SVD applied and $R^2$ computed. The $R^2$ computed on the given, unpermuted data, is compared to the
distribution determined from the permutation test and, for a given significance value, its significance asserted or rejected. This way the whole PLS model and a number of significant singular pairs may be estimated.

PLS constitutes a complete descriptive technique for neuroimage analysis, in much the same way as a paradigm proposed for the PET data in this thesis (PLS does not include any extensions to deal with the temporal sequences such as ones found in fMRI data). It is multivariate in nature but does not attempt to model the spatial properties, such as spatial smoothness, in the data. These could be introduced in a way similar to that proposed in this thesis. The permutation test to determine the number of significant singular pairs represents a big step forward from other descriptive techniques, but it is not without problems. The scan data normalization, for example normalizing voxels to unity variance, introduces potentially big variability into the procedure, but it is not tested in the permutation stage. In general, we can think of LDA, and hence our procedure, as an extension of PLS, where the per-voxel normalization is replaced by full-scale covariance normalization: one that involves rotation, as well as, scaling. Since the full-rank covariance matrix cannot be estimated, we use penalization as shown in the next chapter.

2.4 Datasets Studied

The methodology described in this thesis was applied to two datasets. Here, we give their description, together with the study design, and explain the pre-processing steps applied.

2.4.1 Finger Opposition Task

The finger opposition dataset, which we will call FOPP, is the result of the $^{[15}O]$water PET study on 45 volunteers which were scanned between April of 94 and December of 97 in the PET research center of Veteran Affairs Medical Center in Minneapolis, MN. Due to the
limited axial size of the VA PET camera (10.8cm) it was not possible to cover the whole brain. Consequently, scans from 18 subjects were discarded as they did not adequately cover two particularly important areas: the motor area in the cortex (top of the head) and the cerebellum (at the bottom of the head).

Additionally, scans from 7 subjects had unacceptable between-scan movement. The unacceptable head movement was defined as that which causes a misalignment of more than one voxel (3.125 x 3.125 x 3.325mm³) between any [¹⁵O]water scan and the attenuation scan. The exact amount of movement was determined based on the results from the 6-parameter rigid body transformation (Woods et al., 1992). The idea of tracking movement based on the alignment transformation was described in Strother et al. (1994).

The final dataset used in this thesis consisted of the scans from 20 subjects: 7 males (ages: 42, 39, 30, 25, 44, 33, 37) and 13 females (ages: 45, 30, 25, 33, 53, 53, 56, 33, 41, 47, 24, 38, 27).

Each subject was scanned eight or ten times. Odd scans were taken under the baseline condition, with each study starting from scan 1, while the even scans were obtained while the subject performed a simple motor task. In both states, the subjects had their eyes covered with a patch and were lying relaxed. During the baseline state the ears were plugged with insert earphones, while during the active state an auditory pacing signal was delivered through the earphones. Each subject received one practice lesson before the study. The motor task consisted of sequentially touching, using the left-hand thumb, each of the remaining four fingers successively fore and back. The task started with the i.v. injection of the radioactive [¹⁵O]water bolus and a 90s image acquisition started when the radioactive bolus reached the brain (typically after 10-20s), as assessed by the total number of counts detected by the PET camera. The scans were acquired in the 3D mode and reconstructed using 3D filtered backprojection. The data was corrected for randoms, deadtime and attenuation, but not for scatter.
Scans for each subject were separately aligned to the first scan using the intramodality image ratio technique described in Woods et al. (1992). This process uses a linear transformation to correct for translation and rotation of the head between scans. The eight or ten subject scans were then averaged and the averages used to calculate the twelve subject-specific parameters for the inter-subject alignment algorithm (Woods et al., 1993). This algorithm transforms the subject scans to the common anatomical space of the brain in Talairach coordinates, by applying rotations and translations plus non-rigid body transformations such as shears, to the subject average volumes by comparing them with a simulated reference PET volume in Talairach coordinates' space. The intrasubject aligned volumes were smoothed with $3 \times 3 \times 3$ and $5 \times 5 \times 5$ boxcar smoothers, with simple boundary correction, and these together with the unsmoothed volumes were transformed into the common Talairach space using the subject-specific twelve parameters derived before. The $3 \times 3 \times 3$ smoothed volumes were used to derive the intracranial mask volume consisting of 1's for voxels inside the brain, and 0's outside. The mask was derived by thresholding each volume at the 45th percentile.

### 2.4.2 Static Force fMRI data

Seventeen volunteer subjects had been scanned with a static force paradigm. Each run begins with the baseline condition and alternates between active and baseline, as before. The active condition required the subject to apply a constant force to a small force transducer held between his/her thumb and index fingers. The actual force applied was displayed on a screen together with the “tolerance bars” according to the expected force level. There were 5 force levels: 200g, 400g, 600g, 800g, and 1000g, and the order of these was randomized for each subject and run. Each subject performed two runs. In each run there are 11 instances of alternating baseline and active conditions with 5 force levels. Each instance ran for 44 seconds. During the baseline conditions the subjects were resting and viewing control lines.
The data was collected using a 1.5T GE Sigma Scanner with a whole-brain echo-planar sequence (TR=4s, TE=70ms, tau offset=25ms). Each image volume consisted of thirty 5mm oblique axial slices with $64 \times 64$ voxels ($3.125 \times 3.125\text{mm}^2$) per slice. The postprocessing of volumes includes:

1. visual inspection and exclusion of images with obvious motion, artifact, poor positioning, and where the performance and neurophysiological measures indicate a failure to perform the task

2. semi-automated generation of brain masks for anatomical MRI and fMRI volumes

3. calculation of the 6-parameter within-subject rigid-body alignment matrices of masked fMRI scans to the first scan using AIR (Woods et al., 1998) — discard runs showing more than sub-voxel movement based on the maximum voxel movement in each volume after application of 6 parameter matrices to brainmasks (Strother et al., 1994)

4. aligning the within-subject fMRI scans and calculation of the subject-average aligned fMRI scan.

5. calculation of rigid-body alignment parameters for average fMRI to high-resolution anatomical MRI scan using AIR

6. visual inspection of the alignment between the average fMRI and MRI with and without the transformation and choosing the best

7. calculation of the between-subject 12 parameter affine alignment parameters of the high-resolution anatomical MRI to a high-resolution MRI template in Talairach space

8. formation and application of a single transformation matrix taking each masked fMRI scan to the Talairach space
9. detrending the time series using 4 cosine functions (voxel-by-voxel)

The data from three volunteers were eliminated during the first step of visual inspection, and six more were eliminated during further processing. In two of the remaining eight subjects the first run had poor neurophysiological performance measures with one missing force level. As a result the Static Force dataset contains only single (second) run from 8 subjects. Furthermore, first three scans and some transition scans (those occurring right before or after condition change) were dropped because of known artifacts and hemodynamic transition effects. The eight remaining subjects are 3 males and 5 females with an average age of 31 ± 6 years.
Chapter 3

Penalized Linear Discriminant Analysis with Basis Expansion

In this chapter we present the general framework for the analysis of neuroimages. We present our motivation, algebraic derivation and important details of the computational aspect which we developed to deal with the extremely ill-posed nature of the neuroimaging data.

The basis for our analysis has been Linear Discriminant Analysis (LDA). LDA and its almost-equivalent sister, Canonical Variate Analysis, which have been used in neuroimaging before (e.g., Azari et al., 1993, Rottenberg et al., 1996, Ardekani et al., 1998) with the experimental states defining the classes. These studies deal with the ill-posed nature of the problem by either defining a small number of Volumes of Interest (VOI) — anatomically homogeneous regions of the brain a’priori considered important to the stimulus studied – or by defining the SVD-derived basis. The resulting Canonical Variates — one less than the number of classes or states — are then interpreted as the neural activation patterns. With two classes the resulting single Canonical Variate can be viewed as an alternative to the methods that rely on images formed by subtracting the average of scans in each class, perhaps normalized and preprocessed by ANCOVA (Friston et al., 1991, Worsley et al.,
With more than two states, the LDA approach results in several Canonical Variates ordered in their contribution to explaining the between-state variance. Finally, the times-series effects may be studied by defining the classes to correspond to the temporal order of each scan.

Section 2.3 mentions a broad categorization of the methodology developed for neuroimaging analysis. There are methods that deal with stimulus-induced changes, general variance decomposition methods that do not take into account the state indicators, and methods that introduce non-linearity in various ways. There are also, existing quite independently of the three categories, two data representations used for analysis: single-voxel and SVD-derived basis.

In our approach, we acknowledge the multivariate, spatially correlated nature of the data and introduce a third representation by expanding the desired canonical activation image in a smooth basis. Then a penalized version of LDA, called PDA and developed in Hastie et al. (1995), is applied with smoothness constraints on the canonical variates. This is seen (Appendix B) as equivalent to projecting the input scans on each basis separately, carrying out the PDA analysis in the projected domain and reconstructing the canonical image from the resulting coefficients. A further ridge penalty on the within-class covariance matrix allows for data-dependent choice of the exact amount of smoothness required. Our method may also be seen as bridging the two categories: analysis of stimulus-induced changes and general variance-partitioning methods.

Even with a preliminary dimensionality reduction using SVD, or a smooth basis, the ill-posed nature of the functional neuroimaging problem precludes naive application of multivariate statistical methods such as a standard GLM or LDA, even though the data is clearly multivariate in nature. The problem of overfitting the data, here tied strongly to the "curse of dimensionality" (Bellman, 1961, Hastie and Tibshirani, 1990, pp. 83-84.), is especially acute in these data sets, and leads, in many cases, to singularities or saturated
models at best. With input dimensionality (number of voxels or basis elements) so high, even simple linear models become very flexible and powerful, with high overfitting potential. There are simply too many degrees of freedom available even with linear models, seemingly as many as the number of voxels, although we cannot obviously use more than the number of scans. The proper assessment of validity and generalizability of the modeling results is then of paramount importance. The classical goodness of fit techniques, based mostly on asymptotic results of some global measure of the residuals, are totally inadequate here since the asymptotic assumption of large \( N \), the number of observations, as compared to \( p \), the dimension of the space, are evidently not met. The need for optimal model selection techniques based on measures of model generalizability has been advocated by some (Kippenham et al., 1994, Lautrup et al., 1995, Strother et al., 1995b, 1997, 1998a, Morch et al., 1997, Morch, 1998, Hansen et al., 1999) but has been largely ignored in favour of typically asymptotic inferential tests of unknown generality (e.g., Friston, 1998).

Operationally there are two problems that we try to address with this approach: strategic dimensionality reduction of the input space, and proper assessment of model generalizability. The first problem is attacked in two ways: (1) we induce a smooth prior on the space of the resulting canonical image(s) in the form of a non-adaptive, smooth basis in which we expand the image(s) — in this thesis, we use tensor products of cubic B-splines and wavelets; (2) we regularize the model further with a simple ridge penalty that is a compromise between the model and the estimated spatial covariance and acts as an additional smoothness constraint by reducing the effective degrees of freedom.

By posing the estimation of activation maps as a classification problem we operate within the probabilistic framework of decision theory where we can address the issue of model generalizability with predictive performance measures. If we impose the need too operate within the predictive framework and require that a method must result in images that are interpretable in a given experimental context, we are led naturally to LDA-like
approaches. Our smoothness-constrained, penalized LDA is an extension and specialization of the general Penalized Discriminant Analysis (PDA) model proposed by Hastie et al. (1995) and also investigated by Nielsen et al. (1998) in the neuroimage domain. In section 3.3.4 we derive an efficient algorithm for fitting PDA suitable for this extremely ill-posed data.

We address the important problem of model generalizability and validity of the resulting activation maps using prediction error. In section 3.3.7 we propose two predictive performance measures: Misclassification rate (MC rate) and Squared Prediction Error (SPE) based on posterior probability estimates of class membership. To estimate these we use a state-of-the-art specialization of Bootstrap, the .632+Bootstrap (Efron and Tibshirani, 1997) (the name comes from the fact the the probability that any observation is included in the bootstrap sample is, in the limit, .632), which we compare with the more traditional cross-validation (CV). These are resampling techniques which give nearly unbiased estimates of prediction error and are largely free from distributional assumptions. We use these measures to: (1) compare results built with different numbers of B-spline basis functions, and results built without basis expansion but with different amounts of voxel presmoothing, (2) optimize the ridge parameter which fine-tunes the amount of smoothness in the result, and (3) compare simple preprocessing with and without mean scan normalization.

3.1 Classical Linear Discriminant Analysis

The linear discriminant function was introduced by Fisher (1936) for two classes as a sensible approach to discriminate between two sets of observations. Fisher sought to project the data on the line in a way that would maximize the separation between classes as measured by the between-class variance. The maximization problem must be normalized and it is intuitively appealing to carry it with respect to some measure of data variability.
With multivariate data, the pooled within-class covariance matrix is a good candidate. The problem then becomes one of finding a linear projection that maximizes the ratio of between-class to within-class variance.

The LDA method was later extended to multiple classes (Rao, 1948). LDA can also be derived, if all canonical variates are used for classification, as a maximum likelihood classification rule if the data in all classes is assumed to follow multivariate Gaussian distribution with a common covariance matrix (Mardia et al., 1979; Hastie et al., 1995). With a simple modification to account for prior class probabilities (usually estimated with the proportion of the observations in each class) the LDA method may also be seen as a plug-in Bayes estimator with the same assumptions.

Let us start by establishing some notation used in this section to introduce the classical LDA. Let \( \mathbf{x}_{i}^{(k)} \) be an \( i \)-th \( p \)-variate observation in class \( k = 1, \ldots, K \) in \( K \)-class classification problem. Let \( n_k \) be a number of observations in class \( k \) and \( N = \sum_k n_k \). Define the between-class and within-class covariance matrices by the usual MANOVA quantities:

\[
\Sigma_B = (K - 1)^{-1} \mathbf{B} \quad \text{where} \quad \mathbf{B} = \sum_{k=1}^{K} n_k (\mathbf{x}_{i}^{(k)} - \bar{x}^{(k)})(\mathbf{x}_{i}^{(k)} - \bar{x}^{(k)})^T \tag{3.1}
\]

\[
\Sigma_W = (N - K)^{-1} \mathbf{W} \quad \text{where} \quad \mathbf{W} = \sum_{i,k} (\mathbf{x}_{i}^{(k)} - \bar{x}^{(k)})(\mathbf{x}_{i}^{(k)} - \bar{x}^{(k)})^T \tag{3.2}
\]

Algebraically, LDA is a following optimization problem: find \( \mathbf{a}_h \) such that \( \mathbf{a}_h^T \mathbf{B} \mathbf{a}_h \) is maximized subject to \( \mathbf{a}_h^T \mathbf{W} \mathbf{a}_h = 1 \) and subject to \( \mathbf{a}_h^T \mathbf{W} \mathbf{a}_j = 0 \) for \( j = 1, \ldots, h - 1 \). This can be posed as a generalized eigenvalue problem:

\[
\arg\max_{\mathbf{a}_h} \frac{\mathbf{a}_h^T \mathbf{B} \mathbf{a}_h}{\mathbf{a}_h^T \mathbf{W} \mathbf{a}_h} \tag{3.3}
\]

subject to the aforementioned orthogonality constraints on \( \mathbf{a}_h \) with respect to \( \mathbf{W} \). The solution is the eigendecomposition of \( \mathbf{W}^{-1} \mathbf{B} \), with the first eigenvector defining the first canonical variate which explains most of the variability between classes. Geometrically, with two classes, LDA seeks to separate the classes in the \( p \)-dimensional space by the
straight line that is orthogonal to the line joining the two centroids of the data that has first been sphered using $W^{-1}$.

### 3.1.1 Discriminant Functions and MANOVA View of LDA

With two classes, $k \in \{1, 2\}$. Anderson (1984) defines a discriminant function $W(x) = \mathbf{x}^T \delta$ as:

$$\arg\max_{\delta} \frac{\left[ \mathbf{E}(\mathbf{X}^T \delta \mid k = 2) - \mathbf{E}(\mathbf{X}^T \delta \mid k = 1) \right]^2}{\text{Var}(\mathbf{X}^T \delta)}$$

(3.4)

which, with appropriate plug-in estimators, leads to Eq. 3.3. From there we see that the two-class LDA seeks a univariate random variable, $\delta^T \mathbf{x}$, that maximizes the ratio of expected squared between class difference to its variance.

A generalization of that result may be viewed in MANOVA context. In section 12.5, Mardia et al. (1979) develops a test of dimensionality that leads directly to the LDA’s canonical variates. Briefly, with all LDA assumptions, let $r \leq \min(p, K - 1)$, be the proposed dimension of the hyperplane within which all $K$ class means lie. This test is one of possibilities to explore should the general (one way) MANOVA test of all means equal be rejected. Note that this test has no analogue in the univariate case: there one can only go after specific contrasts.

In general $K$ means span a $K - 1$ dimensional hyperplane, given that $p$ is at least $K$. One may wish to see whether the actual dimensionality of the problem is smaller, hence the test. The Likelihood Ratio version of this test entails a set of vectors, proportional to the canonical variates, that may be used to test for successively larger $r$: the first vector is used to test $r = 1$, the first and second to test whether $r \leq 2$ and so on. These vectors span the successively higher dimensional hyperplanes such that, for a given dimension $r$, each hyperplane is a Maximum Likelihood estimate of the hyperplane that contains the $K$ means under null hypothesis. Therefore, we expect that the first canonical image will exhibit the features that most distinguish the classes, normalized for the covariance structure.
Successive canonical images show the further features of the data that are uncorrelated with the previous ones.

### 3.1.2 The Geometry of LDA in Two Class, 2D setting

![Figure 3.1: Demonstration of 2-class LDA in 2 dimensions. The light points (class 1) and darker points (class 2) show the 200 bivariate Gaussian observations generated from each class. The solid line is the true canonical variate (CV). The circles are class means, and the diamonds are the means projected onto the CV. The points marked with the cross represent the test point and its projections onto the mean-difference (broken) and CV lines.](image)

Figure 3.1 shows a demonstration of the LDA with two classes and in two dimensions. The data has been generated using 2D Gaussian distribution with means (0.55, 0.45) for the first class and (0.25, 0.65) for the second class. The covariance matrix was chosen to
obtain non-circular shapes with an oblique angle. The test point (a cross at (0.38, 0.38))
that, given the shape of two Gaussians, quite clearly belongs to class 2, is actually closer
to the mean of class 1, when using the regular Euclidean distance. Using this distance
is equivalent to projecting onto the mean–difference line, shown in broken style in the
Figure 3.1, and carrying the Euclidean distance classification in one dimension. This line
represents the t-test image with the pooled estimate of standard deviation (equation 2.8).
The canonical variate line (solid) is corrected to reflect the non-circular shape of the data.
and is a compromise between the first principal component and the mean–difference line.
One projects the data and class means onto this line and then uses Euclidean distance to
classify.

3.2 LDA and Random Subject Effects

In all of the analysis in Chapter 4 we do not explicitly consider subject effects. In this
section we investigate how dangerous this is, and how much accuracy we lose to gain some
computational advantage offered by scan-space version of PDA. We will only look at the
classical LDA with two classes and in the “non-ill-posed” setting, i.e., with \( p < n \).

It turns out, that LDA is doing almost “the right thing” if the data are assumed
to come from a two-way replicated design with random subject and fixed class effects.
This illustrates the clear advantage of LDA over methods based on average difference
images, like t-maps. These, and other methods usually subtract subject effects, which is
equivalent to assuming a fixed-effect model. It seems much more plausible to regard subject
effects as random rather than fixed. Using random effects also lets us assess the predictive
performance of the model, something that is not possible with a fixed-effect structure. With
fixed effects, such as models that subtract the subject averages from the data, we cannot
extend the results beyond the population studied, and thus cannot use prediction error for
validation. Thus we will see, that in addition to the fact, that LDA is able to correct for
non-diagonal covariance matrix (which corresponds to the non-iid noise structure) it also "automatically" deals with random subject effects.

To show the workings of LDA under random subject effect model, assume that we have $S$ subjects indexed by $s$, and each subject had her/his observation obtained $r = 1, \ldots, R$ times in each class. The model for the observation $x_{rs}^{(k)}$ (in class $k$) is:

$$x_{rs}^{(k)} = \mu_k + \nu_s + \epsilon_{rs}$$

with:

$$\nu_s \sim \mathcal{N}(0, \Sigma_S) \quad \text{and} \quad \epsilon_{rs} \sim \mathcal{N}(0, \Sigma_e).$$

where $\Sigma_S$ and $\Sigma_e$ are covariance matrices for both effects. As usual, we assume that the subject terms $\nu_s$ are independent between subjects and independent of the iid sequence $\epsilon_{rs}$. Also note that both covariance matrices are the same in each class $k$, which is in the spirit of LDA.

The Gaussian view of LDA, assumes that we have a multinormal Gaussian distribution in each class with the same covariance matrix, and iid observations in each class. One then classifies by either maximum likelihood rule (Mardia et al., 1979, p. 301) or by Bayes rule. Ignoring prior probabilities, both rules assign observation $x$ to the class with maximum likelihood for $x$. (The Fisher LDA is equivalent when one uses all canonical variates and when a pooled within covariance matrix estimate is used for the common class covariance matrix).

In our case, although the observations from the same subject are no longer independent, the decision rule is similar to the LDA. We have:

$$x_{rs}^{(k)} \sim \mathcal{N}(\mu_k, \Sigma_S + \Sigma_e)$$

Thus the classification is based, as usual, on the Mahalanobis distance:

$$(x_{rs}^{(k)} - \bar{x}_k)'(\Sigma_S + \Sigma_e)^{-1}(x_{rs}^{(k)} - \bar{x}_k)$$
As compared to LDA, the only thing that changes is the covariance matrix. To examine how Fisher’s LDA is doing in this case, we need to look at whether the pooled within-class covariance matrix, used implicitly in LDA, is a good estimate of $\Sigma_S + \Sigma_e$.

We will start by working with the observations in a single class. Let $X$ be the $N \times p$ matrix of observations. We have that:

$$W = \frac{1}{N-1}(X - \bar{X})^T(X - \bar{X})$$

Let us look at the $(t_1, t_2)$ element of $W$:

$$W_{t_1,t_2} = \frac{1}{N-1} \sum_{rs} (x_{rs}(t_1) - \bar{x}(t_1)) \cdot (x_{rs}(t_2) - \bar{x}(t_2))$$

$$= \frac{1}{N-1} \left[ \sum_{rs} x_{rs}(t_1)x_{rs}(t_2) - N \bar{x}(t_1)\bar{x}(t_2) \right]$$

Now:

$$E x_{rs}(t_1)x_{rs}(t_2) = \mu(t_1)\mu(t_2) + \Sigma_S(t_1, t_2) + \Sigma_e(t_1, t_2)$$

and:

$$E \bar{x}(t_1)\bar{x}(t_2) = N^{-2} \sum_{r,s,r',s'} E x_{rs}(t_1)x_{r's'}(t_2)$$

$$= N^{-2} \sum_{r,s,r',s'} \mu(t_1)\mu(t_2)\zeta(r, s, r', s')$$

where:

$$\zeta(r, s, r', s') = \begin{cases} 
\Sigma_S(t_1, t_2) + \Sigma_e(t_1, t_2) & \text{when } (r, s) = (r', s') \\
\Sigma_S(t_1, t_2) & s = s', r \neq r' \\
1 & s \neq s', r \neq r' 
\end{cases}$$

There are $N$ pairs for the first case, $SR(R - 1) = N(R - 1)$ for the second case and $N^2 - N - N(R - 1)$ for the third. Thus:

$$E \bar{x}(t_1)\bar{x}(t_2) = N^{-2} \left[ N^2\mu(t_1)\mu(t_2) + N(R - 1)\Sigma_S(t_1, t_2) + N(\Sigma_S(t_1, t_2) + \Sigma_e(t_1, t_2)) \right]$$

$$= \mu(t_1)\mu(t_2) + R/N\Sigma_S(t_1, t_2) + N^{-1}\Sigma_e(t_1, t_2)$$
And thus we conclude, that:

\[
\begin{align*}
\mathbb{E} W(t_1, t_2) &= \frac{1}{N-1} 
\left[ N\mu(t_1)\mu(t_2) + N\Sigma_S(t_1, t_2) + N\Sigma_e(t_1, t_2) 
- N\left[ \mu(t_1)\mu(t_2) + RN^{-1}\Sigma_S(t_1, t_2) + N^{-1}\Sigma_e(t_1, t_2) \right] \right] \\
&= \frac{N-R}{N-1} \Sigma_S(t_1, t_2) + \Sigma_e(t_1, t_2)
\end{align*}
\]  

We can therefore see, that the usual pooled within-covariance estimator that the LDA uses will underestimate the combined covariance \(\Sigma_S + \Sigma_e\) that ML requires. The amount of under-estimate will depend on the relation between the subject effect and error term, and the number of replications per subject relative to the number of subjects. In our case, \(N = 20 \cdot 4 = 80\) and \(R = 4\) so the underestimate does not appear to be significant.

Asymptotically, if the number of scans for each subject is held fixed, the above estimate is unbiased. The result does not change when \(K\) classes are considered and the \(K\) estimates of the common covariance matrix in each class are pooled, since each class has the same covariance structure by assumption.

The amount of bias caused by the \(\frac{N-R}{N-1}\) factor in front of \(\Sigma_S\) will also depend on how different both covariance matrices are. For instance, if they are not different, then LDA results in an estimate which is correct up to a scalar multiple, and the same goes for the canonical variates. Thus if the action of the two matrices is concentrated on the first few eigenvectors which are similar for both \(\Sigma_S\) and \(\Sigma_e\), the bias would be minimal. On the other hand the bias will have a stronger effect on the canonical variates if the leading eigenvalues of subject covariance are large as compared to their \(\Sigma_e\) counterparts, and are associated with very different eigenvectors.

### 3.2.1 Simulation Study

To assess the effects of the biased covariance matrix estimate that LDA is implicitly using with subject effects, we have performed a simulation study. In the study we compare LDA using three estimators of the covariance matrix:
1. The usual pooled within-class covariance matrix, $W$.

2. The sum of common within-subject and error covariance matrices $\hat{\Sigma}_S + \hat{\Sigma}_E$.

3. The corrected within-class covariance matrix: $W + \frac{R}{N-1} \hat{\Sigma}_S$.

We used the estimators of $\Sigma_S$ and $\Sigma_E$ proposed in the MANOVA context by Anderson (1985), Anderson et al. (1986). These correct the usual within and between sum-of-square and products estimators to make them positive semidefinite. The authors concern themselves with one-way random effects MANOVA, but the results hold in our case of mixed-effects 2-way MANOVA. One starts by forming the usual within-subject and error sum-of-squares and products matrices:

$$SS_S = RK \sum_{s=1}^{S} (\bar{x}_s - \bar{x})(\bar{x}_s - \bar{x})^T$$

(3.17)

$$MSE_S = (S - 1)^{-1} SS_S$$

(3.18)

$$SS_E = \sum_{k,s,r} (x_{sr}^{(k)} - \bar{x}_s - \bar{x}^{(k)})(x_{sr}^{(k)} - \bar{x}_s - \bar{x}^{(k)})^T$$

(3.19)

$$MSE_E = (N - S - K + 1)^{-1} SS_E$$

(3.20)

The expectations are as follows:

$$EMSS_S = \Sigma_E + RK \Sigma_S$$

(3.21)

$$EMSS_E = \Sigma_E$$

(3.22)

This suggests an estimator for $\Sigma_S$:

$$\hat{\Sigma}_S = (RK)^{-1}(MSS_S - MSS_E)$$

(3.23)

which is not guaranteed to be positive definite. Anderson (1985), Anderson et al. (1986) have obtained modified estimators. These move that part of variance that would make (3.23) negative to the error variance. This is done by first simultaneously decomposing $MSS_S$ and $MSS_E$:

$$MSS_S = ZDZ^T, \quad MSS_E = ZZ^T$$

(3.24)
It may be achieved, for example, by eigendecomposing \( \Lambda S S^{-1/2}_E S S^{-1/2}_S \) and left-multiplying the resulting eigenvectors by \( S S^{-1/2}_E \). The estimator (3.23) now becomes:

\[
(RK)^{-1} Z (D - I_p) Z^T
\]  

(3.25)

where, as before, \( p \) is a dimension of covariance matrices. The idea is to remove this part of variance that would make the estimator negative. This is achieved by excluding these columns of \( Z \) with corresponding eigenvalues \( \nu < 1 \). Let \( p^* \) denote the number of eigenvalues \( \nu \geq 1 \). Let \( D^* \) be as \( D \) but with only the first \( p^* \) elements. Let \( Z^* \) be the corresponding eigenvector matrix with only the first \( p^* \) columns of \( Z \) included. Then we get the estimator of \( \Sigma_S \) which is guaranteed to be positive-semidefinite:

\[
\hat{\Sigma}_S = (RK)^{-1} Z^* (D^* - I_{p^*}) Z^{*T}
\]  

(3.26)

The variability removed from \( \Sigma_S \) is attributed to error variance which gives the modified estimator for \( \Sigma_E \):

\[
\hat{\Sigma}_E = (N - 1)^{-1} [(N - 1) \Lambda S S_E + (S - 1)(Z^{**} D^{**} Z^{*T} + Z^* Z^{*T})]
\]  

(3.27)

Here the \( ** \) subscripts refer to these parts of \( D, Z \) that were left out in Eq. 3.26. If there were none, then \( \hat{\Sigma}_E = \Lambda S S_E \) as usual. Anderson (1985). Anderson et al. (1986) show that these are maximum likelihood estimators under normality assumptions.

**Design of the Simulation Study**

We performed a study by simulating the data from multivariate Gaussian distribution with random subject and error effects. We only considered two classes with same covariance structure in each. Thus \( n \) observations would be generated according to the model in Eq. 3.5. Each observation is a realization of a 1D Gaussian process observed at \( p \) points on a line. The mean in class I and II was a discretized sinusoid and cosinusoid, respectively. Regardless of the input dimensionality, \( p \), we set the frequency of the sinusoid so that 4 periods would be covered. We chose this mean to reflect our interest in the functional data.
The crucial issue is specification of covariance matrices for subject and error terms. Again, to stay within the functional data framework we used covariance matrices of an isometric process. The covariance between two points only depends on the distance between them. Specifically, the covariance structure was:

\[
\text{Cov}(x_1, x_2) = \exp(-\alpha \cdot \text{dist}(x_1, x_2))
\]  

(3.28)

where "dist" was measured in number of voxels separating \(x_1, x_2\). Increasing \(\alpha\) leads to fast–dying correlations and thus to rough processes. We used \(\alpha = 5.0\) for the error term and \(\alpha = 0.05\) for the subject effect. This conforms to our intuition that subject effects have some large smoothness properties, while error should be mostly noise with only some spatial smoothness remaining.

These specify the correlation matrices. One of the parameters of the simulation was \(\text{VarRatio}\). The error term always had variance equal to 1.0 and the subject term variance was \(\text{VarRatio} = \{1.0, 10.0, 100.0\}\). Another parameter was input dimensionality, \(p\), with three choices: \(\{5, 10, 30\}\). We also varied the number of subjects, \(S = \{1, 5, 10, "N"\}\), where "\(N\)" meant number of observations in each class. \(N = \{50, 200\}\). Together, \(N\) and \(S\) determined the number of observations for a subject in each class, \(R = N/S\).

For each combination of \(\{\text{VarRatio}, N, p, S\}\) 50 training sets, each of size \(2N\) (two classes) and one test set of size \(2 \cdot 5,000 \ (2 \cdot 3,000 \text{ for } S = "N")\) due to computational limitations) were generated. Three LDA models, with three estimates of common covariance matrix, described at the beginning of this section (Sec. 3.2.1), were estimated on the training sets and applied to the test sets to obtain estimated posterior probabilities for each observation in the test set. We considered two estimates of prediction error: \(\text{Dev}\) and \(\text{SPE}\), defined as follows. If a test observation \(x_0\) came from class \(C = \{I, II\}\), and \(\{p_I(x_0), p_{II}(x_0)\}\) were two estimated posterior probabilities then:

\[
\text{Dev}(x_0) = -2 \log p_C \quad \text{and} \quad \text{SPE}(x_0) = (1 - p_C)^2
\]  

(3.29)
The test set estimated posterior probabilities were then used to calculate both prediction error estimators and these were averaged over all observations in the test set.

We analyzed the results of the simulation study using ANOVA model for 4-way factorial design with replications. The four factors were: \{VarRatio, N, p, S\} and there were N = \{50, 200\} replications. The tests and p-values are used mostly as guides since the normality assumption is likely questionable, especially for the Deviance response which exhibits many outliers. We analyze four responses:

\[
\begin{align*}
\text{Dev-ES} : & \quad \text{Dev}(W') - \text{Dev}(\Sigma_E + \Sigma_S) \\
\text{Dev-WS} : & \quad \text{Dev}(W') - \text{Dev}(W' + \frac{R - 1}{N - 1} \Sigma_S) \\
\text{SPE-ES} : & \quad \text{SPE}(W') - \text{SPE}(\Sigma_E + \Sigma_S) \\
\text{SPE-WS} : & \quad \text{SPE}(W') - \text{SPE}(W' + \frac{R - 1}{N - 1} \Sigma_S)
\end{align*}
\]

These four responses analyze the differences between the standard within-class and two modified estimators of the covariance matrix, on the Deviance and SPE scales. We used a simple additive ANOVA model of all four factors with one interaction between P and V terms. In many models that we tried, all terms were always significant, but as we mention above we were not overly concerned with p-values. More interesting are the effects’ estimates presented in Table 3.1. There are some clearly visible trends. As expected from the result in Eq. 3.16, increasing VarRatio leads to a better performance with modified estimators regardless of the PE metric used. The improvement is much more pronounced using the “W+S” estimator. Increasing the dimensionality of the data, P, also tends to favour modified estimators, although there is a surprising twist in the SPE-WS case. Similarly, the modified estimators work better with smaller training set sizes (N50) especially in high dimensions as the interaction term N50:P30 indicates. This suggests that the modified estimators benefit from lower signal-to-noise ratios. It is an apparently surprising finding, since in higher dimensions and with fewer observations one might have expected poor estimates of variance matrices. Since we need two such estimates for the modified
Table 3.1: Estimate of Effects in the four ANOVA models of the simulation results. The terms are input dimensionality: $P = \{5, 30\}$ as compared to $P = 10$, training set size $N = 50$ compared to $N = 100$, number of subjects, $\text{Subjects} = \{5, 10, "N"\}$ compared to 1 subject and ratio of variances of subject effect to error effect, $\text{VarRatio} = \{10, 100\}$ compared to $\text{VarRatio} = 1$. There is also an interaction term between $P$ and $N$.

LDA as opposed to one pooled within-class matrix, the modified method could be expected to suffer more under lower signal-to-noise ratios and in higher dimensions. One possible explanation is that the modified estimators use the available signals more efficiently. This hypothesis is partially supported by comparing “E+S” and “W+S” estimators: “E+S” could be expected to show larger improvement as it does not use the within matrix at all, and this is indeed the case. Another possibility is that the true covariance matrices had very simple structure dependent only on a single parameter $\alpha$. It may be possible that in that case more dimensions actually help in estimating these matrices.

A good case for modified LDA comes, not surprisingly, from number of subjects’ effect: Subjects. The baseline was Subjects = 1 when all methods are numerically equivalent. With $n$ subjects, there is understandably little effect of modified LDA, and in fact $W+S$ does worse than the baseline. But for 5 and 10 subjects both methods perform significantly better than the baseline, except that again there is a huge reversal in the SPE-WS column. For “E+S” estimator the effect for 5 subjects is itself enough to make this method better.
regardless of the state of other parameters, for both measures of Prediction Error. Together with higher VarRatio values, the subject effect makes the modified estimators, especially “E+S” very clear winners over classical LDA.

In general “E+S” modified estimator of the covariance matrix performs somewhat better than the pooled within covariance matrix in the context of classification, if there are strong subject effects present. The improvement seems to be more pronounced with higher signal-to-noise ratios, here indicated by either smaller training set sizes, higher dimensionality or both. The improvements are not large, however, and they are probably smaller when penalization is included. It may be worthwhile to develop a modified PDA method for PET/fMRI images that would take both subject and error covariance matrices into account.

3.3 Dimension Reduction in LDA using Smoothness Constraints and Penalization

As described in the preceding section, Linear Discriminant Analysis results in the set of orthogonal vectors in the data space, called canonical or discriminant variates, that best separate the class means with respect to the within-class covariance. The total number of discriminant variates is one less than the number of classes if the problem is of full rank. If all of the variates are used, then LDA can be derived as the Maximum Likelihood estimate of the optimal classification rule under the multivariate normality assumption with a common within-class covariance matrix (Hastie et al., 1995, Ripley, 1996, pp96).

Linear discriminant analysis is essentially equivalent to canonical correlation analysis, canonical variate analysis and optimal scoring, in that any one is sufficient to derive the others. In the context of images, where the classes are experimental states, the LDA can be used to obtain the variates (e.g., Azari et al., 1993, Rottenberg et al., 1996, Friston et al., 1996, Strother et al., 1996, Ardekani et al., 1998) in the voxel (or VOI) space, that can be
interpreted as activation images (or profiles). When LDA is used with scans as inputs, the canonical variates are usually called canonical images. With two classes, as in our baseline–activation analysis, the single canonical image may be interpreted as a pattern of the signal driving (or driven by) the activation state. Algebraically, the canonical image is just a mean difference pattern rescaled by the inverse of the estimated covariance matrix. With more than two classes, the principal canonical image gives this direction (in the scan space) that most separates the classes with respect to the within-class covariance structure. In that sense it is the image that carries the largest amount of information about the classes. Successive canonical images are chosen to extract most of the information about the classes in the orthogonal complement of the subspace spanned by the previous canonical images.

Naive application of LDA to the images will not work. Due to the ill-posed nature of the problem one will not be able to estimate the inverse of the within–class covariance matrix. Therefore, we need to constrain the problem and bring its dimensionality down. We achieve it in two ways: by constraining the roughness of canonical images and by penalizing the within–covariance matrix.

3.3.1 Basis Expansion of Canonical Variates

By constraining the problem through imposing spatial smoothness on the resulting canonical images we not only reduce the effective dimension of the problem and, thus potentially the variance of the result, but we also model some spatial smoothness which is known to exist in the scans. This is done by expressing the unknown canonical image(s) as a linear combination of the known basis functions of some smooth space.

If \( \mathcal{I}(\nu_1, \nu_2, \nu_3) \) is a canonical image, indexed by the location vector \((\nu_1, \nu_2, \nu_3)\), we require that:

\[
\mathcal{I}(\nu_1, \nu_2, \nu_3) = \sum_{j=1}^{J} \gamma_j B_j(\nu_1, \nu_2, \nu_3) \quad (3.34)
\]
Here, \( B_j(\nu_1, \nu_2, \nu_3) \) is a basis function in the voxel space. Many choices exist for a basis set. We have experimented with the tensor product B-splines (TPS) and wavelet bases.

Having constrained the spatial "roughness" of our canonical images, we need to estimate coefficients \( \gamma_j \). LDA works with the scores, \( \langle \beta, i \rangle \), which are (discrete) inner products between observed scan, \( i \), and the canonical image. Using Eq. 3.34, we have that:

\[
\langle \beta, i \rangle = \left( \sum_j \gamma_j B_j, i \right) = \sum_j \gamma_j \langle B_j, i \rangle
\]  

(3.35)

Thus to find the coefficients \( \gamma \) in the LDA framework, we need to project the scans \( i \) onto the basis set, and treat those as the input to the LDA. The resulting canonical variate will be a vector of coefficients \( \gamma \), which will let us reconstruct the canonical image via Eq. 3.34. Appendix B has more details proving that smoothness-constrained LDA (or PDA) leads to unconstrained LDA (PDA) with the projected data.

### 3.3.2 Penalized Linear Discriminant Analysis

By imposing the smoothness on canonical images, we already reduce the dimensionality of the problem: there will typically be fewer basis functions (\( B_j \)'s; see Eq 3.34) than voxels (we have a few thousand basis functions and about 30 thousand voxels remaining in \( i \)'s after masking is applied). The problem is still ill-posed, however, and further regularization is needed. We have worked with the PDA model developed by Hastie et al. (1995). They impose a penalty on the within-class covariance matrix (of the projected images data), which directly affects the canonical variates, here \( \gamma \), that result. Since we have already expanded the canonical image in a smooth basis, we use a simple ridge penalty, which adds a small value to the diagonal entries in the estimated within covariance matrix. This is equivalent to imposing a penalty on the sum of squared variates' coefficients, \( ||\gamma||^2 \). The PDA model with a ridge penalty is equivalent to the Canonical Ridge model of Vinod (1976), which is also being explored in functional neuroimaging by Nielsen et al. (1998).
The intuition behind penalization is as follows. While we constrain the image to lie in the smooth space, the smoothness constraint can still be overcome by large coefficients. If one specifies a large positive $\gamma_{j_1}$ in (3.34), for a basis function $j_1$ that is centered at some location, and a large negative coefficient $\gamma_{j_2}$ for a basis function $j_2$ that is centered at a nearby location, then the resulting canonical image will have a steep dip between locations, despite our efforts to impose smoothness. How steep a dip will depend on the type and number of basis functions and the size of the coefficients. Together these three control the effective smoothness of the canonical image. By penalizing the size of the coefficients, we control the amount of smoothness given the choice and number of the basis functions. There is a free tuning parameter, $\lambda$, that controls the importance of the penalty term relative to the criterion being minimized by LDA. This is yet another expression of the ubiquitous bias-variance tradeoff. (Friedman, 1994)

There are some advantages to basis expansion followed by penalization. If the basis uses smooth functions, like B-splines, projecting the scans on the basis is similar to smoothing them with the kernel of the shape of the basis function. The whole method, however, is very different from pre-smoothing the scans prior to analyzing them. With the basis-expansion idea we have the power to impose regionally different amount of smoothness or bandwidth: moreover this spatially varying bandwidth is utilized to maximize the discriminatory power of PDA. The regional smoothness is again determined by the individual basis functions, their placement and the size of the coefficients, $\gamma$. Thus, with $\gamma$ being small in some regions and large and variable in other regions, we are able to model a wide variety of possible canonical images that exhibit smoothness in some parts and roughness in the others. The ability to control the overall size of $\gamma$ through a ridge penalty gives us the ability to globally fine-tune the smoothness of the canonical image. Smoothing on some level is necessary: The images are re-constructed by a tomographic process which imposes spatial correlation (Pajevic et al., 1998), the registration techniques are not perfect (Kjems et al., 1999), and
the actual hemodynamic response of the brain, which PET and fMRI methods use as a proxy for neuronal activation, has spatial extent on the order of 3–5mm (Malonek and Grinvald, 1996).

The basis expansion idea is similar in spirit to Ruttimann et al. (1998). There the authors also use a specialized (wavelet) basis to induce a prior and reduce the dimensionality of the fMRI data. Their approach is geared more towards deriving inferential statistical tests on the obtained activation images, a task made easier by the orthogonality of the basis (which leads to the near-orthogonality of the coefficients), while our emphasis is on testing generalizability via prediction error, in a non-parametric way. A second important difference is in the input space to which the basis expansion is applied: we concentrate on the LDA approach, which works with the whole-brain spatial covariance, while Ruttimann et al. (1998) apply the expansion to the voxel-based pooled difference image, basically working with the first moment statistic.

To summarize our method: the data matrix is obtained by projecting each image, as in Eq. 3.35, using a chosen basis, \( B_j \). Then the PDA is applied to the projected data for some value of tuning parameter \( \lambda \), which results in the canonical variates matrix \( \Gamma \). Finally, canonical images are then reconstructed via Eq 3.34.

### 3.3.3 Penalized Discriminant Analysis and Statistical Parametric Mapping

Some interesting analogies to the voxel based methods that rely on (possibly scaled) images derived from the difference of class-averaged images, like SPM, can be established by considering the two–class problem. We have already indicated the basic difference in the geometry of both approaches in Section 3.1, on page 47. Disregarding for a moment the
basis projection step. the single canonical image from PDA is:

\[ c(\Sigma_{W'} + \lambda I)^{-1}(\mu_1 - \mu_2) \]  

(3.36)

where \( \Sigma_{W'} \) is a pooled within-class covariance matrix. \( \mu_1, \mu_2 \) are class-mean images, and \( c \) normalizes the image to length one. Therefore, the canonical image is a rotated, rescaled version of the simple class-mean difference image (MDI), where the rotation and rescaling attempts to equalize the variance and to decorrelate voxels, while the penalty term works in the opposite direction.

For very large values of \( \lambda \), when the penalized within-class covariance matrix becomes essentially (a constant multiple of) the identity, the canonical image is a scaled MDI. This assumes that variances across voxels are equal, and that voxels are uncorrelated, and thus resembles the voxel-wise t-map with pooled variance estimate. For moderate \( \lambda \), we can expect \( \Sigma_{W'} + \lambda I \) to be diagonally dominant with possibly different diagonal elements. The resulting image will be similar to diagonally scaled MDI where each voxel in the MDI is compared to its variance (now resembling the voxel-wise t-map with individual voxel variance estimates). For small \( \lambda \), we get close to fully estimating the within covariance matrix and rotating/scaling the MDI to account both for local variances, as well as covariances. With the internal optimization, we let Prediction Error decide how much information we have to move away from the unrealistic assumption of homoscedasticity and independence across voxels.

The tensor-product basis projection is helping the estimation by somewhat decorrelating the variables because it models part of the spatial covariance structure. We can expect the covariance matrix of the projected data to be more diagonally dominant than in the unprojected space. This in turn results in better estimates of the covariance matrix. We therefore expect PDA to be a more flexible method than SPM, with one hyperparameter \( \lambda \) that is able to control the tradeoffs between increased flexibility of full covariance normalization and necessity of simplifying assumptions of homoscedasticity or diagonal covariance.
matrix.

PDA also exhibits some similarity to the other well known method in Neuroimaging called Partial Least Squares (PLS) described in McIntosh et al. (1996) and in Section 2.3.3. (There is another well known algorithm, also called Partial Least Squares, widely used in the Chemometrics community. For some statistical description see, e.g., Wald et al. (1984) and note that PLS described here is a completely different algorithm). PLS starts with decomposing $X^T Y$ using SVD, where $X$ is, as in our case, the scan matrix, and $Y$ is an arbitrary design matrix. Furthermore, PLS provides an interesting paradigm for choosing a number of significant components that result from SVD. In light of the correspondence between CCA and LDA, which we prove in the next section, PLS may be seen as an unnormalized version of LDA: in LDA one looks at the singular value decomposition of:

$$
(X^T X)^{-1/2} X^T Y (Y^T Y)^{-1/2}
$$

(3.37) 
(see Eq.C.4), which is followed by rotating back the left-hand singular vectors. Another way to look at it is via the orthogonality constraint (Eq.3.11): PLS uses a Euclidean metric while LDA normalizes to unity variance using the within-class covariance matrix estimator. Assuming that the variance can be estimated effectively. LDA normalization is preferable as it puts all voxels on an equal footing. Of course, we cannot estimate the full covariance matrix of all voxels (or basis functions) and our recourse is to use penalization. One can again establish some analogies for different values of the ridge hyperparameter. Similarly as in the previous paragraph, we observe that for a small number of degrees of freedom we may expect our results to be similar to PLS ones, as the left-hand normalizing matrix will be close to a constant multiple of identity. The right-hand normalization simply reweights the observations by their class sizes.
3.3.4 PDA via Regression

In this section we will show how to obtain the Canonical Variates of the PDA model using two steps: penalized regression followed by the eigendecomposition of the regression results. Our proof is different from that given in Hastie et al. (1995) and relies only on matrix algebra. We also feel it is more appropriate for the neuroimaging community, as it hinges more closely on the current approaches widely used in this domain.

In the next section we will show how to “train” the PDA model, given $V$ scans (possibly projected) as input, and how to predict from this model with $O(V)$ computational effort, that is without dealing with large $p \times p$ matrices. This is of vital importance as we use resampling techniques to estimate the prediction error. Without this extension obtaining the hundreds of model estimates needed by the Bootstrap and cross-validation would be computationally prohibitive.

We start with an unpenalized version (LDA) and first show that the closely related method, Canonical Correlation Analysis (CCA), can be expressed as a multiresponse regression followed by an eigendecomposition. We then prove the relationship between CCA and LDA and, finally, introduce penalization and describe an extension to deal with the image data. Our proof differs from one in Hastie et al. (1995) in that we directly apply it to the CCA formula (Eqs. 3.40 and C.1).

CCA is a symmetric method that, given two sets of variables measured for each observation, $x, y$, seeks two linear combinations that exhibit maximal correlation. That is each observation is composed of $\{x_i, y_i\}$, where $x$ and $y$ are in general of different dimensions. One attempts to summarize the data by finding two linear combinations, $b, a$ of $x$ and $y$, respectively, such that:

$$
\frac{\text{Cov}(b^T x, a^T y)}{\sqrt{\text{Var}(b^T x) \cdot \text{Var}(a^T y)}}
$$

is maximized.
One extends the method by finding all such possible directions. \(a_k, b_k\) that successively maximize the correlation and are orthogonal to previously found pairs. Since:

\[
\text{Var}(a^T x) = a^T \text{Var}(x) a \quad \text{and} \quad \text{Cov}(b^T x, a^T y) = a^T \text{Cov}(x, y) b
\]  

(3.39)
the problem is to find matrices \(A, B\), with linear combinations in their columns, such that:

\[
N^{-1} B^T S_{xy} A
\]  

(3.40)

is maximized subject to:

\[
N^{-1} B^T S_{xx} B = I \quad \text{and} \quad N^{-1} A^T S_{yy} A = I.
\]  

(3.41)

where \(S_{xx}, S_{yy}\) and \(S_{xy}\) are respective covariance and cross-covariance matrices. This is a generalized SVD problem (Mardia et al., 1979, pp. 282), and one can show (Appendix C) that it may be solved, after suitable normalizations, via multiresponse regression of \(Y'\) onto \(X\), followed by the eigenanalysis of \(Y'^T Y'\) (where \(Y'\) are fitted values from the regression step).

**In our case, \(X\) denotes the data matrix** (whose rows are scans or projected scans), and \(Y\) is the \(N \times J\) class-indicator matrix, with 1's denoting the class of each scan. The classes are experimental conditions: here either two classes denoting the Active/Baseline states, or eight classes denoting the temporal order of tasks.

It is known (and we rederive in it Appendix D), that the Canonical Variates (CV's) associated with \(x\) are, up to a scaling factor, the same as the canonical variates that result from LDA (Hastie et al. (1995),(Mardia et al., 1979, Ex. 11.5.4)). In Appendix D we prove that:

\[
B_{LDA} = BD_{(1-c^2)^{-\frac{1}{2}}}
\]  

(3.42)

where \(D\) is a diagonal matrix. Thus we show how one obtains the canonical variates of LDA by rescaling \(B\).

As mentioned, the unpenalized version is unsuitable as it requires an inversion of a singular within-class covariance matrix, \(\Sigma_W\). To remedy that we apply penalization to \(\Sigma_W\).
or equivalently, to the total covariance matrix $S_{xx}$, which results in a penalized regression step:

$$\hat{y}_{\lambda,\Omega} = X\hat{\beta}_{\lambda,\Omega} = X(X^TX + \lambda\Omega)^{-1}X^TY$$

(3.43)

For any positive definite $\Omega$, this makes $X^TX + \lambda\Omega$ invertible. In this paper we use ridge penalty, $\Omega = I_n$, and then Eq. 3.43 defines a ridge regression solution.

### 3.3.5 Expressing the PDA algorithm in the $N$-dimensional space

Major effort has been spent in deriving efficient computational algorithms presented in this thesis. The importance of computational issues has increased greatly in statistics, often due to the resampling methods, that apply any given algorithm many times, due to popularity of simulations where the computer-generated data of large size is used to test the model, or simply due to ever increasing amounts of data the statisticians have to deal with. As mentioned in the introduction and throughout this chapter, images constitute a specially challenging form of the data due to their sizes. With a $128 \times 128 \times 48$ PET scan, we are dealing with 786,432 voxels, and if each is stored as a floating point number of single precision, each scan occupies over six megabytes of disk space. With dozens of images available in one data set, the computationally efficient methods are a must.

The PDA algorithm presented in Section 3.3.4 works in the $p$-dimensional space, where $p$ is the number of voxels or basis functions. The only place where $p$ dimensional quantities are needed is in the ridge regression step (Eqn. 3.43) and when the canonical variate ($\gamma$) or image (Eqn. 3.34) are constructed. In particular, in the ridge regression step, it appears that we need to form, and invert, a huge $p \times p$ matrix $X^TX + \lambda I$. In Appendix F we show that the fitted values $\hat{Y}$ of ridge regression step may be computed using only $N$-dimensional quantities, where $N$ is a number of scans. In order to do that, one needs to precompute the outer-product matrix, $XX^T$, an expensive step which needs $O(N^2p)$ operations but that is performed only once.
Since we use resampling methods to search for optimal $\lambda$ and thus need to run the above algorithm, with a given set of training inputs $X, Y$ many times. Additionally, since we then only need to compute the posterior probabilities (and hence, predicted class memberships), it pays to precompute $G = XX^T$ once and then apply cross-validation or bootstrap. Each bootstrap sample may then be obtained by selecting only those rows/columns of $G$ that correspond to the sample observations, and thus forming $G^*$. the bootstrap version of $G$.

This, after full-data $G$ is computed, lets us operate in the $N$-dimensional scan space for as long as we do not need to compute the canonical variate $B_{LDA}$. Since the posterior probability estimates can be obtained using only $Y$, fitted values $\hat{Y}$, right-hand eigenvectors $A$, and eigenvalues $D_e$, (Appendix E) we can perform the optimization of the ridge parameter $\lambda$ in the lower-dimensional space of the observations. Appendix F shows how ridge regression may be computed using only $N \times N$ matrix $G$ and matrix $Y$ of class indicators. Appendix G shows the algebraic trick of computing centered version of $G$, $\tilde{G} = \hat{X}\hat{X}^T$, where $\hat{X}$ had its columns means subtracted, using the uncentered $G$. In fact, since in the resampling methods we use the subset of the data as a training set, this Appendix shows how to center the partial $G^*$ corresponding to the subset chosen by resampling, and how to center the rows of remaining observations with the column means of training set used in fitting the PDA, all using the once-computed uncentered $G$.

### 3.3.6 Effective Degrees of Freedom

Linear statistical analysis defines a notion of *degrees of freedom* (d.f.). These specify a dimensionality of the space onto which we project the data, and in the case of iid Gaussian errors, the expected drop in Residual Sums of Squares if only noise variables are included in the model (e.g., Sec. 3.5, Hastie and Tibshirani, 1990).

Since we have expressed LDA (and PDA) with a regression as a building block we can carry over the notion of d.f. In the univariate, full-rank ($N > p$), linear regression case,
d.f. = \( p \), the number of variables. Then also:

\[
p = \text{trace}(X(X^TX)^{-1}X^T) = \text{trace}(H)
\] (3.44)

where \( H \) is a projection ("hat") matrix (i.e. \( \hat{Y} = HY \)). By analogy, in the ridge regression case we can define the effective degrees of freedom (EDF):

\[
\text{EDF} = \text{trace}(S(\lambda))
\] (3.45)

( Crave and Wahba, 1979. Hastie and Tibshirani, 1990), with \( S(\lambda) = X(X^TX + \lambda I)^{-1}X^T \). a penalized "projection" matrix obtained in the regression step of PDA. Please note that (3.45) is not the only possible definition for EDF: see [e.g.,][Hastie and Tibshirani (1990)] for other possibilities.

In our case, with \( n < p \), we can compute EDF using only matrix \( G \) of outer products.

A bit of algebra shows that:

\[
\text{EDF} = \text{trace}(S(\lambda)) = \sum_{j=1}^{N} \frac{\alpha_j}{\alpha_j + \lambda}
\] (3.46)

where \( \alpha_j \) are eigenvalues of \( G = XX^T \), same as non-zero eigenvalues of \( X^TX \). This shows that EDF combines the tuning parameter with the smoothness inherent in the basis representation, and is more informative than unscaled \( \lambda \). The EDF vary from 1 (for \( \lambda = \infty \) since the penalization is applied to the centered \( G \)) to \( N \), for \( \lambda \) near zero.

### 3.3.7 Prediction Error and its Estimates

It should be realized, that the need to impose constraints is more than just a numerical necessity. It affects the generalization ability of our model: i.e. whether the resulting activation map will be interpretable and significant, or whether it will be overshadowed by noise and peculiarities of the data at hand. That is directly related to the predictive performance of the model: if the model has not been constrained enough (and in the "right" way), then it will not be able to classify a new scan that was not used in training the model.
The generalizability of functional neuroimaging models has been addressed before (Kippenham et al., 1994, Lautrup et al., 1995, Morch et al., 1997, Morch, 1998, Strother et al., 1997, 1998a, Hansen et al., 1999). In particular, Morch (1998) contains a good introduction to generalization error, predictive performance and bias–variance trade-off issues in the context of neuroimaging. Even though prediction is not the main goal in analyzing PET images, one needs to be concerned about the generalizability of the activation maps (here, canonical images) derived. Prediction Error (PE) is a way to measure the generalizability of our modeling process. We use PE (or, rather, its estimate) as a function of EDF, in three ways: to choose the amount of smoothness, to assess the final usability of derived patterns, and to compare different data representations.

A Probabilistic Framework.

Linear Discriminant Analysis, while first established by Fisher as a sensible procedure regardless of distributional assumptions (Mardia et al., 1979), can be rederived within a probabilistic framework. If one assumes that the scans come from the multivariate Gaussian distribution, and assumes that these Gaussians have the same covariance structure among classes, then the LDA can be derived as a plug-in Bayes classifier for the data with the usual estimates of class-mean images and covariance matrix.

Specifically, for an image $i^{(k)}$ from class $k, k \in \{1, \ldots, K\}$, let $i^{(k)} \sim \mathcal{N}(\mu_k, \Sigma)$. In general the Bayes classifier would assign new image $i_0$ to that class $k_0$ which maximizes the posterior probability:

$$k_0 = \arg\max_k P(k \mid i_0) = \arg\max_k \frac{P(i_0 \mid k)P(k)}{P(i_0)} \quad (3.47)$$

where $P(i_0 \mid k)$ is a class–specific likelihood, here Gaussian, $P(k)$ is a prior probability of observing an image in class $j$, and $P(i_0)$ is a normalization constant.

Since the covariance matrix is assumed the same for all classes, the only class dependent component of multivariate Gaussian likelihood is the argument to the exponential function.
or Mahalanobis distance between \( i_0 \) and the mean of class \( k \), \( \mu_k \):

\[
D(i_0, \mu_k) = (i_0 - \mu_k)^T \Sigma^{-1}(i_0 - \mu_k)
\]  
(3.48)

It is an established fact (Hastie et al., 1995. Ripley, 1996. pp96). that the Mahalanobis distance is a Euclidean distance when the image and class means are projected onto all canonical variates. Therefore LDA results can be used to obtain both posterior probabilities and classification by:

\[
\hat{p}(k \mid i_0; \hat{\mu}_k, \hat{\Sigma}) = C \exp\left(-1/2 \| B_{LDA}^T(i_0 - \hat{\mu}_k) \|^2\right) \pi_k
\]  
(3.49)

where \( C \) normalizes the probabilities to add up to one. \( B_{LDA} \) is the matrix of canonical variates (in columns), which was derived via the route convenient for us. in section 3.3.4. Eq 3.42. and \( \pi_k \) are estimated prior probabilities for each class.

**Prediction Error Measures.**

We need to define a suitable measure of Prediction Error in the population. We have used two such measures: Misclassification rate (MC rate) and Squared Prediction Error (SPE). MC rate is a probability of misclassifying a new scan by the model fitted on the training data. It is a rough measure, with the discontinuous 0-1 penalty for misclassification. Thus the model which gives posterior probability of 49% to the correct class. will score the same error on this scan as a model which gives 1% posterior probability, assuming a 50% threshold is used as in the 2-way classification problem.

Another measure of PE with a more reasonable metric is Squared Prediction Error. \( SPE = (1 - \hat{p}_C)^2 \). where \( \hat{p}_C \) is the posterior probability estimated by the model for the correct class. One could also use deviance (minus twice the likelihood ratio), well known from the theory of Generalized Linear Models (McCullagh and Nelder, 1989). here simply \(-2 \log \hat{p}_C\). We experienced erratic behaviour of this measure. because posterior probabilities were often close to zero or one; deviance puts a very large penalty on cases where \( \hat{p}_C \approx 0 \). This issue has also been addressed by Hintz-Madsen et al. (1998).
**Resampling Estimates.**

The above are population parameters, conditional on the model and the training data. We need to derive their estimates. We use (5-fold) cross-validation (CV) and the bootstrap resampling techniques (Efron and Tibshirani, 1993). The 5-fold CV estimate is derived by first randomly dividing the data into 5 equal-sized parts. Then the model is trained on 4/5ths and used to obtain predictions for the remaining 1/5th of the data. This is repeated five times, for each of the five distinct training/validation set divisions. The prediction error is an average of errors accumulated over the five validation sets. The CV process mimics the situation where we have a set of independent observations on which to estimate the prediction error. However, using a five fold CV results in an estimate which is biased due to the diminished size (80%) of the CV-training set. Also, there is a variability associated with many possible ways to divide the data into five parts. The .632+ bootstrap estimate was designed to remedy that, and has been shown to outperform CV in simulation studies (Efron and Tibshirani, 1993, 1997).

The .632+ bootstrap procedure is a refinement of the “leave-one-out” bootstrap approach, which we now describe. One obtains $B$ bootstrap samples, with replacement, from the original data (in our case, $B = 50$). Then a model is built on each sample and tested on the observations that were (by chance) not included in the sample. The resulting prediction errors are averaged to give $\bar{PE}^{(1)}$, the leave-one-out Bootstrap PE estimate.

It should be noted, that we have used subjects, each with all his/her scans, as a sampling unit, for both CV and bootstrap resampling techniques. Otherwise large negative biases in the per-scan PE estimates will result due to the large between-subject variability in these data sets (Strother et al., 1995a,b).

The .632 correction (Efron and Tibshirani, 1993, Ch. 17) was derived to correct for the (positive) bias that results since each bootstrap sample contains only 63.2% of the original
sample, on average. The .632 estimate is:

$$
\overline{PE}^{(.632)} = 0.632\overline{PE}^{(1)} + 0.368\overline{PE}^{Tr}
$$

(3.50)

or a weighted average of the leave-one-out bootstrap estimate and the training error on all of the data. This will now underestimate the PE for models which highly overfit and have training errors close to zero. The '+' correction attempts to deal with that, by first estimating the no-information error rate. \( \gamma \) which is defined in the population as:

$$
\gamma = E_{F_{ind}} Q(y_0, r_{\mathbf{x}}(t_0))
$$

(3.51)

This means the following: assume distribution, \( F_{ind} \), of data points consisting of predictors and responses: \( \{t, y\} \), such that the marginal distributions of predictors and responses is the same as for the observed data, but the two are independent; that is there is no information in \( t_i \) about \( y_i \). Let \( r_{\mathbf{x}}(t_0) \) denote the prediction made by our model at point \( \{t_0, y_0\} \) from \( F_{ind} \), trained on the available data \( \mathbf{x} \). The point \( \{t_0, y_0\} \) is also independent of the training set \( \mathbf{x} \). The function \( Q(\cdot) \) is the prediction error measure. \( \text{SPE} \) or misclassification rate in our case. A possible empirical estimate of \( \gamma \), suggested by Efron and Tibshirani (1997), is:

$$
\hat{\gamma} = \sum_{i,i'=1}^{N} Q(y_i, r_{\mathbf{x}}(t_{i'}))
$$

(3.52)

which is an error rate computed for our data using all \( N^2 \) pairs of predictor and responses, that effectively mixes up the two and destroys their relationship.

For a misclassification rate, \( \hat{\gamma} = \hat{\pi}_1(1 - \hat{\rho}_1) + (1 - \hat{\pi}_1)\hat{\rho}_1 \), where \( \hat{\pi}_1 \) is a proportion of class 1 observations, and \( \hat{\rho}_1 \) is a proportion of observations predicted by \( r_{\mathbf{x}} \) to belong to class 1. The multiclass extension is \( \hat{\gamma} = \sum_{j=1}^{J} \hat{\pi}_j(1 - \hat{\rho}_j) \). For \( \text{SPE} \) Eq. 3.52 becomes:

$$
\hat{\gamma}_{\text{SPE}} = \sum_{i,i'} [1 - \hat{p}(j(i') \mid i_i)]^2
$$

(3.53)

where \( \hat{p}(j(i') \mid i_i) \) is an estimated posterior probability (as in (3.49)) of the class that a scan \( i' \) belongs to. One may calculate (3.53) in the following way. Class \( j \) will be a "correct"
class $n_j$ times for each observation (where $n_j$ is number of observation sin class $j$). Let $P$ denote the $N \times J$ matrix of estimated posterior probabilities. Then $\hat{\gamma}$ of equation 3.53 will be equal to the average of all row-sums in the scaled $P$, where each column $j$ is multiplied by $n_j$.

The no-information error rate is used to form the weight, $\hat{\omega}$, for the convex combination:

$$\overline{PE}^{(.632+)} = \hat{\omega}\overline{PE}^{(1)} + (1 - \hat{\omega})\overline{PE}^{Tr}$$

(3.54)

as a replacement for (3.50). The weight $\hat{\omega}$ is formed in the following way: first define the relative overfitting rate, $\hat{R}$:

$$\hat{R} = \frac{\overline{PE}^{(1)} - \overline{PE}^{Tr}}{\hat{\gamma} - \overline{PE}^{Tr}}$$

(3.55)

Relative overfitting rate measures the overfitting by the difference between the leave-one-out bootstrap estimate and the training error, relative to the "pure" overfitting as measured by the difference between no-information rate and training error. $\hat{R}$ varies between 0 — if there is no bias in the training error — and 1. if there is "full" overfitting: that is when $\overline{PE}^{(1)}$ equals to no-information error rate, $\hat{\gamma}$. The weight $\hat{\omega}$, defined as:

$$\hat{\omega} = \frac{.632}{1 - .368\hat{R}}$$

(3.56)

varies from .632, when $\hat{R} = 0$ to 1. With $\hat{\omega}$ so defined. Eq. 3.54 is seen providing some method-based adaptivity to Eq. 3.50.

### 3.4 A Note on Gaussian Assumption

The LDA procedure, as derived by Fisher, does not need to rely on Gaussian distributional assumption which is also true for its penalized and smoothness constrained version described here. The only place where the normality is used is in estimating posterior probability and thus in estimating Prediction Error measures developed in section 3.3.7. The
question of the validity of Gaussian assumption becomes then a question of the validity of PE estimates. One may conjecture that the departures from Normality would have detrimental effect on the predictive performance of our method, which would lead to larger prediction errors than one would obtain with the similar, but Gaussian distributed data. Perhaps more important, however, is the value of the ridge parameter $\lambda$ where the minimum PE happens. As this determines the final image we obtain from the analysis with a given basis set. It is quite possible that the departure from Normality changes the shape of PE curves (like these in Fig. 4.3). Again, we do not feel that the location of the minima would drastically change with the departure from normality (as this location is clearly independent to the monotonic transformations of the PE), but acknowledge a need for some robustness studies in that matter.

Finally, we have some consolation in that at least the PET data may not be very far from Gaussian. Each voxel is based on a linear combination of a large number of random photon counts (see section 2.2.1) and we thus hope the Gaussian approximation to Poisson will work. With smooth basis expansion these (reconstructed) counts are further smoothed with a large number of neighbouring ones which hopefully gives us a possibility that Central Limit Theorem may be applied.

### 3.5 Is Ridge Penalty Enough for the B-spline Basis?

Ridge penalty is very convenient for us to use computationally, but the question is whether it penalizes "the right thing". By that we usually mean high frequency components or higher derivatives. In one dimension B-splines are usually used with second order derivative penalty (e.g., Hastie and Tibshirani, 1990), which results in the natural cubic smoothing spline fit: similar penalty could be composed using 3 dimensional tensor product B-splines. O'Sullivan (1991) gives an very remarkable algorithm for composing an eigendecomposition of a discrete Laplacian penalty matrix, that penalizes square of the sum of the second
derivatives of the data. Also, there exists a wide literature on thin-plate splines (e.g., Green and Silverman, 1994) that are the most popular way to extend the cubic smoothing splines to higher dimensions. All these methods require that one handles \( p \times p \) matrices, which is computationally prohibitive in our case.

In this section we show, in a semi-formal way, that even though the ridge penalty is not optimal in the sense of penalizing second derivatives, it still behaves reasonably in the sense that higher frequency components are penalized more. The intuitive support was given in section 3.3.2.

Let us start by looking at the simpler regression problem. Let:

\[
y_i = f(x_i) + \varepsilon_i
\]

where \( \varepsilon \sim N(0, I_{n \times n}) \) and \( f(x) \) is a regression function to be estimated. If we want to constrain \( f(x) \) to be smooth, we can expand it into some basis of smooth functions. Let \( \{B_j(x)\}_{j=1}^p \) be such a basis. Then, if we denote the evaluated basis \( (n \times p) \) matrix by \( B \), we have:

\[
\hat{y} = B(B^TB + \lambda I_{p \times p})^{-1}B^Ty
\]

by fitting the ridge regression onto the smooth basis. In the context of B-splines, we would have matrix \( B \) as the B-spline matrix: i.e., \( p \) B-spline basis each evaluated at \( n \) design points, \( x_i \). If we wanted \( f \) to be a natural cubic spline, simple ridge penalty is not enough: in one dimension, we have to use \( p \times p \) penalty matrix:

\[
\Omega_{mn} = \int B'_m(x)B'_n(x)dx.
\]

We would like to avoid using more the complicated \( \Omega \), since:

- We operate in 3-D. Although there are extensions to higher dimensions (like thin-plate splines), we would like to use simpler tensor product basis, for which "proper" \( \Omega \) is not easy to calculate
• Simple ridge is much more feasible computationally in our case, as it lets us calculate fits in the $X$-dimensional scan space, as shown in Sec. 3.3.5.

A more formal, functional setting for the above problem is as follows: find function $f(\cdot)$ that minimizes the penalized regression problem:

$$\|y - f(t_i)\|^2 + \lambda \int (f''(t))^2 dt$$  \hspace{1cm} (3.60)

Remarkably, one can show (e.g., Wahba, 1990, Green and Silverman, 1994) that the minimizing function is a cubic smoothing spline with knots at distinct values $t_i$.

Let us diagonalize the regression equation (3.58), by decomposing $B = UD_\gamma V^T$, using Singular Value Decomposition. Here $U$ and $V$ are left and right orthonormal eigen-vectors, respectively, and $D$ has corresponding eigenvalues on its diagonal. The problem now becomes:

$$\tilde{y} = UD_\gamma V^T(VD_\gamma V^T + \lambda I)^{-1}VD_\gamma U^T y$$

$$= UD_\gamma V^T(V(D_\gamma^2 + \lambda I)V^T)^{-1}VD_\gamma U^T y$$

$$= UD_\gamma V^T(D_\gamma^2 + \lambda I)^{-1}V^T D_\gamma U^T y$$

$$= UD_\gamma^2/ (\lambda + \gamma_j^2) U^T y$$ \hspace{1cm} (3.61)

We have expressed the simple ridge regression problem (3.58) in the orthonormal basis $U$. The penalty associated with basis function $j$ is $(\lambda + \gamma_j^2)/\gamma_j^2$, showing that the basis associated with larger eigenvalues are penalized less.

The question now becomes: when arranged by decreasing eigenvalues, are the orthonormal basis functions increasing in “complexity”, thereby warranting higher penalties? The partial answer may be obtained by looking at figure (3.2). Here we have obtained the orthonormal basis for the B-spline problem in 2 dimensions. It is quite visible, that the “wiggly” basis are penalized more.

To go back to PDA with basis expansion, we look at the following problem: find function
were calculated with $A = I$ and displayed here in the order of decreasing eigenvalues. The penalies shown for each basis:

**Figure 3.2**: The orthonormal basis functions (columns of $U$) from the 2-D tensor-product B-spline expansion. There are $3 \times 3$ B-spline basis, which were diagonalized with SVD and B-spline expansion.
3(t) such that:
\[
\left\| y - \sum_j x(t_j)3(t_j) \right\|^2 + \lambda \int (3''(t))^2 dt
\] (3.62)
is minimized. This is a penalized regression problem and a solution for \(3(t)\) is again a cubic smoothing spline with knots at distinct values of \(t_j\). This problem is covered by a special case of theorem 1.3.1 in Wahba (1990), the so-called generalized smoothing spline problem. The details are explored in, for example, Hastie and Tibshirani (1993). The caveat for us is that one possible way to solve the problem is to expand the coefficient \(3(\cdot)\) in cubic B-spline basis, and apply the second order penalty matrix \(\Omega\) (Hastie and Tibshirani, 1993). Our approach of expanding the Canonical Image in B-spline basis has the same flavour and (for 1-D case) would result in the cubic smoothing spline if the right penalty matrix was used. We can use the heuristic argument of the previous paragraph to justify the use of the ridge penalty instead.
Chapter 4

Results with B-Spline and 3 dimensional Wavelets

4.1 Wavelet Basis

As it will be apparent from the results presented below, wavelets have shown themselves to be a possibly more efficient representation of the Canonical Variates in FOPP neuroimaging problem than B-splines. Fewer components basis are required to represent the signal and their predictive properties are superior to those of B-splines, in two class setting, although the results are surprisingly different in the eight-class problem. In this section we will introduce some properties of wavelets and provide partial justification for the choices we made when using wavelets for Canonical Variates bases.

In this section we will introduce wavelet bases, multiresolution analysis and the wavelet transform. This general discussion was adapted from three excellent books on wavelets:

Ogden (1997) This is the first, and a very readable, book on statistical analysis using wavelets.
Vidakovic (1999) This is a more comprehensive book on wavelets also designed for Statisticians. It offers a more complete theoretical framework and has a wider spectrum of auxiliary wavelet topics discussed.

Burrus et al. (1998) This is a book written for engineers in signal processing fields. It offers an excellent discussion on wavelet filters and their implementation, together with a good introduction to signal processing and filterbank theory.

4.1.1 Wavelets: Introduction

By a wavelet one usually means any family of functions that is composed from a single mother wavelet function, \( \psi(x) \) (Please note that we will use a customary wavelet notation here, which may conflict with previously introduced symbols). It is assumed that the mother wavelet satisfies an admissibility condition:

\[
C_\psi = \int \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \tag{4.1}
\]

where \( \Psi(\omega) \) is a Fourier transform of the mother wavelet. Loosely speaking, condition (4.1) says that the wavelet’s power must be concentrated in higher frequencies. Since the wavelet is in \( L^2 \), this effectively means that the wavelet must be a band-limited function. One easy consequence of the admissibility condition is that \( \Psi(0) = 0 \) which in turn implies that:

\[
\int \psi(x) dx = 0, \tag{4.2}
\]

that is, the mother wavelet must average to zero. It is also customary, to normalize the mother wavelet to unity norm:

\[
||\psi(x)||^2 = \int |\psi(x)|^2 dx = 1 \tag{4.3}
\]

Given a mother wavelet, one constructs the wavelet basis by \textit{diadic dilations} and integer translations:

\[
\psi_{j,k} = 2^{j/2} \psi(2^j x - k) \quad j, k \in \{0, \pm 1, \pm 2, \ldots \} \tag{4.4}
\]
The scaling factor $2^{j/2}$ keeps the unity norm. The translation index $k$ is easily understood for $j = 0$: it generates a sequence of mother wavelet translates, each moved to the right or left by an integer. The dilation index, $j$ rescales the $x$-axis, compressing or expanding the mother wavelet: it does it in the units of powers of 2.

Under mild conditions, the wavelet system is orthogonal:

$$\int \psi_{j_1,k_1}(x) \psi_{j_2,k_2}(x) dx = \delta_{j_1}^{j_2} \delta_{k_1}^{k_2}$$  \hspace{1cm} (4.5)

using a Kronecker $\delta$ symbol. There exist non-orthogonal wavelet systems: they are then usually bio-orthogonal. Bio-orthogonal systems have two sets of wavelets: one to project a function onto, to obtain the wavelet coefficients (the analysis wavelets) and one to reconstruct a function from, using the wavelet coefficients (synthesis wavelets). Bio-orthogonal systems maintain cross-orthogonality between the analysis and synthesis wavelets. The usual orthogonal system is a special case where the analysis and synthesis systems are the same. We will only concern ourselves with the orthogonal wavelet systems as they have statistical properties which are better understood.

### 4.1.2 Orthogonal Wavelet Basis and Multiresolution Analysis

We hinted above to the fact that the wavelets constitute an orthonormal basis for some functional spaces. Of these, the most important is $L^2(\mathbb{R})$, the space of all functions, $f(\cdot)$, with a finite $L^2$ norm:

$$\int f^2(x) dx < \infty$$  \hspace{1cm} (4.6)

Another well-known basis for $L^2(\mathbb{R})$ is a Fourier basis.

Given a wavelet basis for $L^2$ one can decompose any function $f(x) \in L^2$ into its wavelet coefficients:

$$d_{jk} = \int \psi_{jk}(x) f(x) dx$$  \hspace{1cm} (4.7)
and this mapping is one-to-one, i.e. the decomposition is reversible:

\[ f(x) = \sum_{j,k} d_{jk} \phi_{jk}(x) \]  
(4.8)

Equation 4.7 is called an \textit{analysis equation} and 4.8: a \textit{synthesis equation}.

One special property that distinguishes the wavelets from other basis is the \textit{Multi Resolution Analysis (MRA)} property, which we will now discuss. Let us imagine that the (infinite dimensional) space \( L^2(\mathbb{R}) \) has the following decomposition:

\[ \ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots \]  
(4.9)

such that:

\[ \bigcap_j V_j = \{0\}; \quad \bigcup_j V_j = L^2(\mathbb{R}); \quad f(2^j x) \in V_j \iff f(x) \in V_0 \]  
(4.10)

Here, the \( V_j \) are subspaces of \( L^2 \) which contain functions of increasing detail, as we will see. The closure of their union, indicated by the overbar, is the whole \( L^2 \), but their intersection is null. The last condition says that for each function \( f(x) \) that is in \( V_j \) there is a unique function \( f_2(x) = f(2x) \) in \( V_{j+1} \) that changes twice as fast, or with twice as much detail.

We suppose that there is an orthonormal basis for \( V_0 \) consisting of integer translations of a \textit{scaling function} or a \textit{father wavelet} \( \phi(x) \):

\[ f(x) \in V_0 \iff f(x) = \sum_k \langle f(x), \phi(x-k) \rangle \phi(x-k) \]  
(4.11)

The diadic dilations, \( \{2^{j/2} \phi(2^j x-k)\}_{k \in \mathbb{N}} \) of the basis for \( V_0 \) become the basis for \( V_j \). Further, since the subspaces contain each other, we decompose the subspace \( V_{j+1} \):

\[ V_{j+1} = V_j \oplus W_j \]  
(4.12)

i.e., into the direct sum of the previous-level subspace, \( V_j \) and the \textit{detail} space, \( W_j \). There exist a canonical ortho-basis for \( W_j \) composed of integer translations of the dilated mother wavelet function, \( \psi(\cdot) \). Since:

\[ V_j = V_{j-1} \oplus W_{j-1} = V_{j-2} \oplus W_{j-2} \oplus W_{j-1} = \bigoplus_{l \leq j} W_l \]  
(4.13)
it is not surprising that the wavelet system, associated with a particular MRA, constitutes an ortho-basis for $L^2(\mathbb{R})$.

The most famous, the simplest and the least practically usable wavelet system is a Haar basis. The Haar scaling function is:

$$\phi(x) = 1\{x \in [0, 1]\}. \quad (4.14)$$

that is, a unity constant function between 0 and 1. It is intuitively clear that by scaling this function down to cover smaller and smaller intervals, and translating it, one can, in the limit, represent any reasonable (e.g., one in $L^2(\mathbb{R})$) function.

The mother wavelet associated with the Haar scaling function is:

$$\psi(x) = \begin{cases} 
1, & 0 \leq x < \frac{1}{2} \\
-1, & \frac{1}{2} \leq x < 1 \\
0, & \text{otherwise}
\end{cases} \quad (4.15)$$

This is a Haar basis for $W_0$: zero-level detail subspace. Projecting $f(x)$ onto the integer translates of $\psi(\cdot)$, one obtains the local difference between $f(\cdot)$ represented with $\phi_{1,k}(\cdot)$ and with $\phi_{0,k}(\cdot)$, i.e., between $f(\cdot)$ represented with first-order detail and $f(\cdot)$ represented with zero-order detail. It is easy to check that different translates and dilates of $\psi(\cdot)$ are orthonormal: more other, $\psi_{j_0,\cdot}$ are orthogonal to $\phi_{j,k}$, for any $k$ and $j \leq j_0$, as we would expect from relation (4.12).

Figure 4.1 shows some examples of wavelet functions. At each level there are twice as many wavelets as on the previous level but each of them gets more "squeezed" which enables it to uncover more detail in the signal. Another feature of many wavelets is their "spikiness" which is a result of requiring compact support and orthogonality. Higher order wavelets get visibly smoother at the expense of longer filter lengths (next section).
Figure 4.1: Haar (left) and Daubechies Symmlet wavelet functions. The detail level grows from bottom up, and only some integer translates are drawn at each level.
4.1.3 Discrete Wavelet Transform

In practice, one is interested in obtaining the wavelet representation of a function, just as we are interested in Fourier (or frequency domain) representations computed with Fourier Transforms. Given a function $f(x)$, one wants to calculate the lowest-level scaling coefficients $c_{j_0,k}$ and the detail coefficients $d_{j,k}$, where:

$$
c_{0,k} = \int \phi(x-k)f(x)dx \quad d_{j,k} = \int \wp_{j,k}(x)f(x)dx
$$

(4.16)

The arbitrary level $j_0$, for which we calculate the scaling coefficients, represents the coarsest scale we are interested in for a function $f(\cdot)$ under study. In practice, one does not have a function but a sample of it obtained with a given sampling rate (we usually assume that the function $f(\cdot)$ has been sampled uniformly over the $x$-axis). We then assume that the function $f(\cdot)$ is piecewise constant over the sampling subintervals: $f(x) = f_i$ for $x \in \Delta_i$. A given sampling rate determines the highest detail level, $J$, we can possibly calculate. We can then approximately assume that what we have is the projection of function $f(\cdot)$ onto $V_J$, or that:

$$
c_{J,k} \equiv f_k
$$

(4.17)

This is the starting point for the Discrete Wavelet Transform (DWT) which is used extensively in practice. What remains to be shown is how to obtain detail and coarser level scaling coefficients $d_{j,k}, c_{j,k}, j = j_0, \ldots, J$. Since $V_0 \subset V_1$ we can represent the zero level scaling function using the first level ones:

$$
\phi(x) = \sum_{k \in \mathbb{Z}} h_k \sqrt{2} \phi(2x-k)
$$

(4.18)

This is a so-called scaling equation and is fundamental in constructing wavelets. The filter $\{h_k\}$ is of finite length when the support of $\phi(x)$ is finite, and is then an example of a Finite Impulse Response filter. Similarly, since $W_0 \subset V_1$, we have:

$$
\wp(x) = \sum_{k \in \mathbb{Z}} g_k \sqrt{2} \phi(2x-k)
$$

(4.19)
An important theorem, implied by the orthogonality of wavelet and scaling functions at the same level, states that:

\[ g_k = (-1)^k h_{1-k} \]  

Now, the scaled and translated version of the scaling equation is:

\[ \phi(2^j x - k) = \sum_{m \in \mathbb{Z}} h_m \sqrt{2} \phi(2(2^j x - k) - m) \]  

\[ = \sum_{m \in \mathbb{Z}} h_{m-2k} \sqrt{2} \phi(2^{j+1} x - m) \]

A similar relationship holds for the wavelet coefficients' equation (4.19). By writing down the definition of the wavelet and scaling coefficients one can use the above results to obtain the two fundamental equations of the DWT:

\[ c_{j,k} = \sum_{m \in \mathbb{Z}} h_{m-2k} c_{j+1,m} \]  

\[ d_{j,k} = \sum_{m \in \mathbb{Z}} g_{m-2k} c_{j+1,m} \]

These equations show how to calculate the lower-level scaling and wavelet coefficients from higher level scaling ones. Given the starting values \( c_{J,k} \) from (4.17) one proceeds to calculate the wavelet coefficients for levels \( J, J-1, \ldots, j_0 \) and the final scaling coefficients for level \( j_0 \). There are similar equations for going up the scale: calculating \( c_{j+1,k} \) from pairs of coefficients \( c_{j,k} \) and \( d_{j,k} \). These equations are used in the synthesis stage, while the equations (4.23, 4.24) in the analysis stage.

### 4.1.4 3D Wavelet Basis

The discussion so far was centered on the one dimensional wavelet basis. In order to use wavelets for analyzing PET and fMRI scans we need to construct the 3D basis. The method of choice is, as in B-spline case, tensor product of one dimensional wavelet functions. One has to be careful, however, to obtain an orthogonal basis with an appropriate MRA
decomposition. To this end one first generalizes MRA. Let $D$ denote the dimension of the domain $\mathbb{R}^D$. Then we have $D$ univariate MRA decompositions:

$$\ldots \subset V_{-2, (d)} \subset V_{-1, (d)} \subset V_{0, (d)} \subset V_{1, (d)} \subset V_{2, (d)} \subset \ldots$$

(4.25)

for $d = 1, 2, \ldots, D$. We are interested in the $D$-dimensional MRA:

$$\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots$$

(4.26)

where:

$$V_j = \bigotimes_{d=1}^D V_{j, (d)} \subset L^2(\mathbb{R}^D)$$

(4.27)

i.e., each $D$-variate MRA subspace is a tensor product of the corresponding univariate ones. By a tensor product space we mean that its basis consists of tensor product of univariate scaling functions that form the basis of $V_j$:

$$2^{j/2} \phi(2^j (x_1, \ldots, x_D) - k) = 2^{jD/2} \prod_{d=1}^D \phi_d(2^j x_d - k_d)$$

(4.28)

for any $k \in \mathbb{Z}^D$. To obtain the multidimensional wavelets we start with expressing $V_{j+1}$ as a direct sum between previous level $V_j$ and a detail space, $W_j$. To be concrete let us use $D = 3$:

$$V_{j+1} = V_{j+1, (1)} \otimes V_{j+1, (2)} \otimes V_{j+1, (3)}$$

$$= (V_{j, (1)} \otimes W_{j, (1)}) \otimes (V_{j, (2)} \otimes W_{j, (2)}) \otimes (V_{j, (3)} \otimes W_{j, (3)})$$

$$= V_j \oplus (V_{j, (1)} \otimes V_{j, (2)} \otimes W_{j, (3)}) \oplus (V_{j, (1)} \otimes W_{j, (2)} \otimes V_{j, (3)})$$

$$\oplus (V_{j, (1)} \otimes W_{j, (2)} \otimes W_{j, (3)}) \oplus (V_{j, (2)} \otimes V_{j, (3)})$$

$$\oplus (W_{j, (1)} \otimes V_{j, (2)} \otimes V_{j, (3)})$$

$$= V_j \oplus W_j^1 \oplus W_j^2 \oplus W_j^3 \oplus W_j^4 \oplus W_j^5 \oplus W_j^6 \oplus W_j^7$$

(4.29)
The detail spaces $W^j_s$ will emphasize local features in various canonical directions. If we imagine the directions in the space ordered as run horizontally, vertically and "into the page", then spaces $W^1_s, W^2_s, W^3_s$ will be "turned on" by features in the "depth", vertical and horizontal features, and the remaining 4 will pick up various diagonal directions. The space for $W^4_s$ is spanned by 3-D integer translations of:

$$2^{3j/2} \psi^s(x - k) = \prod_{d=1}^{3} \zeta_{d(s)}(2^j x_d - k_d)$$

(4.30)

where:

$$\zeta_i(x) = \begin{cases} 
2^{j/2} \phi(x) & \text{if } i = 0 \\
2^{j/2} \psi(x) & \text{if } i = 1 
\end{cases}$$

(4.31)

and $d(s)$ denotes the $d^{th}$ digit in the binary expansion of $s$.

### 4.1.5 Wavelet Thresholding

Donoho and Johnstone (1994, 1995) have developed thresholding rules for denoising signals using wavelets. These results are optimal in the minimax, or "worst case scenario" sense. Specifically, one assumes a typical sequence of noisy observations:

$$y(t_i) = f(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n$$

(4.32)

where $f(\cdot)$ belongs to some functional class and $\epsilon_i$ are i.i.d. sequence of standard Gaussian variable. Donoho and Johnstone develop a series of minimax estimators of $f(\cdot)$ using hard and soft wavelet coefficient thresholding. We will only describe here a hard universal thresholding rule which they termed VisuShrink.

DWT is an orthogonal transform which may be represented by a matrix multiplication. If one denotes a sequence of wavelet coefficients of $f(\cdot)$ by $\theta$ then the DTW of the noisy signal $y(t_i)$ is:

$$w_{jk} = \mathcal{W}y(t_i) = \theta_{jk} + \epsilon_{jk}$$

(4.33)
where the transformed noise coefficients are still i.i.d. standard normal because of the orthogonality of $W$. The idea of thresholding is to only keep these coefficients that carry the signal, i.e. that are "big enough". The idea hinges on the fact that for a wide classes of signals wavelets provide a sparse representation, that is the wavelet expansion of these signals have many zero or near-zero coefficients. The question is how to determine which wavelet coefficient of a noisy observation do not carry any signal. Many solutions have been proposed. but Donoho and Johnstone (1994) proved that a particularly simple rule has near-optimal MSE in the minimax sense. This rule is: replace $w_{jk}$ by:

$$
\hat{w}_{jk} = \begin{cases} 
  w_{jk} & w_{jk} \geq \sqrt{2 \log n} \\
  0 & w_{jk} < \sqrt{2 \log n}
\end{cases}
$$

(4.34)

In practice the threshold becomes $\hat{\sigma} \sqrt{2 \log n}$, where $\sigma$ is an estimate of the homoscedastic noise level. Donoho and Johnstone (1995) propose to use the median of the finest level coefficients divided by 0.6745 as an estimate of $\sigma$: the constant is derived from the Gaussian case. Others have proposed a similar Median Absolute Deviation of all wavelet coefficients, in place of simple median at the highest level. Everyone agrees that because of the sparcity properties of wavelets a robust estimator of variance should be used.

The above results work for the i.i.d. case. Remarkably, Johnstone and Silverman (1997) show that if the noise is correlated, all results of Donoho and Johnstone are still valid, provided the thresholding is done separately on each level. That is both the threshold and the variance are estimated separately for each detail level. In the case of images. we do that separately for each combination of level and direction, i.e. separately for each $W^\alpha_j$ in the Eq. 4.29.
4.2 Finger Opposition Data: Methods

4.2.1 Data and the Standard t-Test Analysis

We apply the Penalized Discriminant Analysis using a simple ridge penalty and tensor-product B-spline (TPS) basis to the FOPP data described in Section 2.4.1.

After processing by the $3 \times 3 \times 3$ box-car smoother and scan-mean normalization (i.e., dividing each voxel by the mean of all voxels within the brain mask for that scan) a pooled standard deviation estimate was calculated (Worsley et al., 1992) and an activation t-test value obtained for each voxel, as described in Strother et al. (1995a). Note, that such a pooled t-test activation image has been shown to outperform single-voxel t-test images with alternate preprocessing schemes (Strother et al., 1998a). making the pooled activation image a good reference pattern for the various PDA canonical images presented in this paper.

4.2.2 Two-way Classification with TPS: Internal Optimization and Scan Normalization

Each scan was assigned a label, from the \{Active, Baseline\} set, according to its experimental condition. The tensor-product (cubic) B-spline (TPS) basis, composed of 25 B-spline basis in each dimension, (defined as B25 in the next section) was set-up in the smallest 3D box that circumscribed the logical AND of all subject masks for raw (i.e. unsmoothed) scans with and without scan-mean normalization.

Five-fold cross-validation and leave-one-out bootstrap were used to estimate PE measures for a grid of values of the ridge tuning parameter, $\lambda$. The sampling units for both resampling methods were the subjects; thus all scans from a subject were included if a subject was sampled. Fifty bootstrap samples were used, for each value of $\lambda$, to derive a .632+ bootstrap estimate of both MC rate and SPE. We call this process of searching for
the optimal value of a ridge parameter an Internal Optimization.

4.2.3 Two-Way Classification: External Optimization

The scans were assigned class labels, as before. Different parameterizations of the data were used, which we denote by \( B15, B20, B25, B30, B35, \) Braw, BSmooth3, BSmooth5. We will use the word "basis" interchangeably with "representation", in reference to any of these. \( B?? \) are tensor-product spline projected images. The two digits denote the number of B-spline basis in each dimension. That results in a wide range of input dimensionality (i.e., total number of basis functions with support within the AND mask): between 2.500 and 25.000. Braw denotes the un-projected data, with voxels within the AND of all masks. This results in 28.500 voxels used. BSmoothk is similar, but from the data that has first been smoothed with a \( k \times k \times k \) box-car, (ie 3D square kernel) smoother: typically, \( k = 3 \) would be used (e.g., Strother et al., 1995a).

PDA was applied to the mean-normalized data for different basis choices. For each basis, the bootstrap analysis over the same grid of \( \lambda \) was done, with \( B=50 \) bootstrap samples, as described above. A number of bootstrap samples, 50, is admitedly low, but seems (cf. fig. 4.2 a good compromise given the huge computational burden associated with resampling. We compare the error curves (for both MC rate, or SPE) and the position of the minima on the SPE/EDF plane.

4.2.4 Deriving Time and Spatial projections

We can use the same model to obtain the activation maps of the within-subject temporal changes throughout the experiment. This is done by defining an 8-way classification problem, where classes denote the order the scan was taken (for subjects with 10 scans, the 9th scan was pooled with the 7th, and the 10th with the 8th). No information about the experimental-state or about the temporal ordering is available to the model. Out of 7
canonical variates that result, we look for those that represent time and state changes. We project the labeled data and class means onto the canonical variates and choose two that are most appropriate.

4.3 Finger Opposition Data: Results

4.3.1 Two-way Classification and Internal Optimization

Figure 4.2 shows the effect of the ridge tuning parameter (expressed as EDF) on the prediction error. Both CV and .632+ bootstrap estimates of PE measures are shown. We note that the MC rate has an erratic behaviour due to its discontinuous definition. Examining the SPE curves on the top-left and the bottom panel we note that the minima for CV estimates (thin lines) occur earlier and rise faster with increasing EDF than those for .632+Bootstrap (thick lines). This is likely since CV estimates are not corrected for the smaller training set size. This would cause the increased variability of the canonical variates, that comes with higher EDF, to have more pronounced and quicker effect than with the full training set.

In the top-right panel we see that CV estimates exhibit higher variance as compared to their .632+Bootstrap counterparts. To obtain this plot we did not set the random seed to the same value for every λ in the resampling exercise, as was done for every other plot in the paper. This allows us to exhibit between sample variability of both resampling methods that contributes to the variance of the estimate of the prediction error. Another problem with CV is the open question of how many folds to use. On one extreme we have n-fold cross validation, which would have negligible training set bias but high variance of the PE estimates in each fold. On the other and there is a two-fold cross-validation that has lower variance for each fold but high bias due to much smaller training set size and potentially large variance contribution that comes from many ways to divide the training
Figure 4.2: Internal optimization of the ridge tuning parameter (Sec. 4.3.1), expressed as a number of Effective Degrees of Freedom, for Penalized Discriminant Analysis in two-class problem. This example uses the data projected on the B25, tensor product B-spline basis set. The curves show the change of Prediction Error (both Misclassification (MC) rate and Squared Prediction Error (SPE)) as a function of Effective Degrees of Freedom (EDF). Both cross-validation (CV) and .632Bootstrap estimates are exhibited. The top left panel shows changes for un-normalized data, while the bottom panel deals with mean-normalized data. The top right panel shows CV and .632Bootstrap SPE curves for mean-normalized data which were obtained with a different random seed for every EDF; these portray the greater variability of CV estimates. Thin lines — CV estimates, Thick lines — .632+Bootstrap estimates, solid lines — estimates of SPE, dashed lines — estimates of MC rate.
set in two. Leave-one-out bootstrap may be seen (Efron, 1983) as approximating two-fold cross validation but with many two-fold splits and the .632+ correction deals with the bias. We have thus chosen the bootstrap estimator for reporting the rest of the results in this thesis.

The bottom panel shows the curves for the mean-normalized data, where each scan has been divided by its brain-voxel mean. The errors are lower, indicating that a large source of variance has been removed. The MC rate (as estimated by bootstrap) drops down to just below 13%, which is a large improvement over the no-information rate of 50%. More noticeably, minimum SPE is around .104, as compared to 0.25 for no-information value. The SPE minimum is at around 50 EDF, for the normalized data, and around 40 EDF for un-normalized. This suggests, that after optimizing for the degree of smoothness, the model is able to extract more information from normalized data. This observation is consistent with our intuition: mean-normalization removes a large source of variance increasing the signal-to-noise ratio and allowing more generalizable structure to be found in the data.

4.3.2 External Optimization with Different Bases

We have investigated the influence of image representation on the resulting canonical images and the prediction error. Figure 4.3 shows PE curves (SPE: top panel and MC rate: bottom panel) across EDF for all representations. We will concentrate on the SPE curves first. All TPS-projected data behave very similarly, except for the B15 basis. The B15 representation is likely too coarse to capture all but the major, spatially extensive components of the underlying activation pattern. Its minimum occurs quite early on the EDF scale (EDF=36.6), and its SPE rises sharply as EDF increases, indicating that there is less useful structure to be extracted in this representation. The Braw SPE curve shows the worst performance for small EDF, and for larger EDF increases faster than all but the B15 projected basis. This representation seems most sensitive to the choice of the tuning
Figure 4.3: Prediction Error (PE) curves as a function of the effective degrees of freedom (EDF) for all tensor-product B-spline (B15–B35) bases and unprojected raw (Braw) and smoothed (BSmooth3,5) scans. The upper panel shows Squared Prediction Error ($\text{SPE}=(1-\hat{p}_C^2)$, where $\hat{p}_C$ is the estimated posterior probability for the true class), and the lower panel depicts Misclassification rate (MC rate) as a percentage of the total number of scans misclassified. The images from un-projected data are shown with larger-width curves. The markers show the minima for each curve. (The minimum for the B30 curve is not shown as it occurred beyond the figure frame).
parameter. The other two unprojected representations behave similarly to the projected ones, except that the two minima are farther apart and their SPEs are larger than those of the projected basis for most of the EDF range.

Examining SPE plots at around 20 EDF, we note that the SPE curves of projected bases and Braw arrange themselves in order of increasing smoothness and decreasing SPE. This places BSmooth3 between B30 and B35, and BSmooth5 between B20 and B25. At these low degrees of freedom, the canonical image is very flat, except for some extended bumps occurring at the most predictive spatial locations. The canonical images are then driven by large features, rather than by small regional changes, and smoother representations are understandably better in those circumstances. Therefore, it makes sense that the curves order themselves according to the degree of smoothness of the representation.

We also examined the placement of minima for each representation in the SPE–EDF plane. The unprojected bases exhibit lowering of the prediction error with rising EDF, as one moves from smoothest (BSmooth5) to the roughest (Braw) representations. The projected basis exhibit a similar pattern but the trend seems to be heading for larger SPE with larger EDF for “rounder” representations, while all along maintaining a lower SPE than the unprojected bases. One explanation is that the projected basis are somewhat more efficient for this problem allowing the canonical images to contain more structure (higher EDF) while controlling the SPE level. Their utility is, however, limited and the SPE starts to rise with very large numbers of basis functions and an associated higher EDF. Also note, that the EDF-at-min-SPE is again ordered with respect to smoothness of representation, for Braw (28,500 voxels) and projected basis (≤ 25,000 basis functions), and that again the smoothed, unprojected representations fall somewhere in between B15 and B35.

The MC rate curves in the bottom panel of Fig. 4.3 again demonstrate the markedly different behaviour of the B15 representation. Even though this time it achieves second-lowest error at its minimum, it performs significantly worse for the higher EDF values.
as before. Except for B15, all curves from un-projected data have worse MC rates than TPS projected data, for most of the EDF spectrum. Braw exhibits performance that is visibly worse than the other unprojected curves, reversing the ranking of the minima of the SPE plots. All curves, besides B15, flatten-out beyond EDF = 70. We also note, that the EDF-at-min-MC are much higher, but less well determined than their SPE counterparts.

In both cases, the SPE measure of PE seems much informative than the MC rate curves. The minima in the SPE curves are more pronounced, and the curves themselves smoother. We also believe, that the measure which takes into account how much the model’s predictions are right or wrong, is more informative to our goal of assessing the model generalizability. The MC rate plots are highly variable, due to the discontinuous (0/1) error structure. That they flatten out for higher EDF, with their minima occurring much further to the right than in the SPE plots is likely due to the fact that the two-way classification problem is driven mostly by a few strong regions of activation which are quite clear indicators of the active state: mostly the right sensory motor area and left cerebellum (see row C, Fig. 4.4). Once enough weight is put on these regions, (achieved with high EDF) the classifier can perform well, on average, regardless of the rest of the image. There will be some scans, however, where the activation maps needed to perform classification are different. In those cases canonical images composed using high EDF, will perform very badly. While such cases will be penalized relatively mildly in terms of misclassification (they will score a penalty of 1 regardless of how “badly” they misclassify), their posterior probabilities will be very much off, and their SPE penalties much stiffer.

We prefer the SPE measure of the prediction error. The MC rate is, however, better established in testing pattern recognition models and has a somewhat more direct interpretation. We report both while concentrating our interpretation on SPE-based results. In all that follows we will use mean-normalized data and .632+Bootstrap estimates.

However, as evidenced by Fig. 4.4, the SPE differences at the minima are only part
Figure 4.4: Functionally activated [$^{15}$O]water PET voxels above the 93.4 percentile (white overlay) interleaved with registered grayscale MRI brain slices for Penalized Discriminant Analysis of: (A16-A40) unprojected raw data presmoothed with a $3 \times 3 \times 3$ voxel boxcar kernel (BSmooth3); (B16-B40) unprojected raw data without presmoothing (Braw); (C16-C40) tensor product spline basis with 15 spline bases in each spatial dimension (B15); (D16-D40) tensor product spline basis with 35 spline bases in each spatial dimension (B35) (activation images A to D have decreasing squared prediction error (SPE) values as illustrated in Figure 4.3): (E16-E40) is a pooled standard deviation t-test image of scans presmoothed with a $3 \times 3 \times 3$ voxel boxcar kernel — the Bonferroni t-value ($t=4.65$) at the 93.4 percentile was used to define a conservative activation threshold with which to compare activation image peaks (white overlay) for a fixed number of voxels. PET and MRI slices are $128 \times 128$ with 3.1 mm pixels with center-to-center slice spacing of 3.4 mm (i.e., slice A23 and A26 are separated by $(26 - 23) \cdot 3.4 = 10.2$ mm) and are parallel to the AC-PC plane, which coincides with slice 24. Image left = brain left.
of the story. This figure shows the top 6.6 percent activation in nine chosen slices from the canonical image obtained using: A — BSmooth3, B — Braw, C — B15, D — B35 representations, and E — a t-test image. The 6.6% threshold is defined by the Bonferroni t-value of image E and was selected to compare equal number of voxels across the five activation images. The ridge parameter, λ, was set to optimize the SPE for each basis. The first four PDA images are arranged in order of decreasing SPE. Of particular interest are Braw and B35 images. These two representations are somewhat analogous to each other: Braw and B35 images contain the most structure in their unprojected and projected groups, respectively), their respective minimal SPEs are not far from the group minima, and they both represent the roughest representations in each group. The figure shows that the B35 representation results in a less noisy and fragmented image which is more visually appealing. BSmooth3 regains more smoothness as compared to Braw, making it more appealing, but it does so at the expense of predictive performance. We also note significant interesting differences on the t-test image which seems to be missing potentially important structures: contralateral midbrain tegmentum in slice E23 which contains key parts of the motor system such as the substantia nigra, the ipsilateral auditory area in slice E30, and while ipsilateral parietal and premotor regions are seen on slices A36 and B36. D36 shows only the parietal area and E36 the premotor area.

Figure 4.5 shows the scatter plot (one point per Talairach voxel) comparing the t-test and the B35 images. There is a non-linear trend upwards for the significant voxels, in the upper left part of the plot, which shows that the mutually significant activation regions are more pronounced in the B35 canonical image. The circle shows a small cluster of voxels in the primary visual cortex that have been elevated by the PDA from the 20th percentile in the t-test image to the 90th percentile in the B35 image — these voxels are visible in the primary visual cortex in the slices A26, B26 and D26 in Fig. 4.4.

In this finger-opposition data set a single-voxel t-test using pooled standard deviation
Single-voxel t-test (pooled standard deviation)

Figure 4.5: Scatter plot of pairs of activation image values for all Talairach brain voxels (1 point/voxel) for a single-voxel t-test image using a pooled standard deviation estimate, compared to penalized discriminant analysis (PDA) of a tensor product spline representation with 35 B-spline bases along each spatial dimension (B35). The dashed line depicts the principal axis from a principal component analysis of the scatter plot distribution. The circle highlights a group of voxels in the primary visual region that have moved from the 20th percentile in the t-test image to the 90th percentile in the PDA image. The solid vertical line depicts the Bonferroni t-value (t=4.65) at the 93.4 percentile of the t-test distribution of voxel values (white overlay, row E of Figure 4.4) and the solid horizontal line reflects the 93.4 percentile (value=0.0065) for the PDA distribution of voxel values (white overlay, row D of Figure 4.4).
(SD) estimates has been shown to predict population based activation image patterns significantly better than for single-voxel t-tests using individual voxel SD estimates, and to perform at the same level as a canonical variate analysis built with the SVD basis (Strother et al., 1998a). Therefore, it is not surprising that the BSmooth3 and t-test activation overlays in Figure 4.4 are quite similar, probably reflecting the fact that for this simple two-state analysis the ridge penalty is relatively large (see section 3.3.3). However, there are several important differences between the PDA solutions and the pooled t-test result. Figure 4.5 demonstrates that the PDA result has nonlinearily enhanced the most significant voxels relative to the corresponding t-test values and "noise" values around zero. At least one area (the primary visual voxels within the circle in Figure 4.5) has been completely reordered relative to other activated regions so that it is now potentially active while in the t-test result it was negative and not distinguishable from noise. In addition, in Figure 4.4 there are "activated areas" which are plausible given this motor task paced by auditory cues, that appear in slices 19, 23, 30 and 36 of the PDA results in rows A, B and D, but not in the t-test result. The key point here is not that the PDA results are right and the t-test wrong, but that the distribution of potentially activated peaks agree for many expected areas and there is a hint that the tunable PDA results may be more sensitive and able to identify areas that could change the neuroscientific interpretation of the brain response to the task. In addition, the PDA framework is much more flexible, as shown by the eight-class results, and internally optimized through prediction error estimates so that we do not need to put as much faith in the validity of distributional assumptions within an inferential testing framework.

The Effective Degrees of Freedom (EDF) provides us with the way of calibrating the amount of information extracted from the data, analogous to the dimension of the \( \beta \) space in linear regression. It seems, from Fig. 4.3, that it is both the SPE and the EDF-at-min-SPE that are of interest. Ideally, we would like to extract as much structure from the
data as possible (since we believe that the brain function is anything but simple) while maintaining low levels of SPE and thus high generality of canonical images. In that sense, the most appealing results are Braw and B35 representations which have the highest EDF–at–min–SPE and a small difference in their min SPE. Figure 4.4 shows that of those two, the projected B35 representation is much more appealing visually. By examining the trend of the minima in Fig 4.3 we note that likely nothing more can be gained in the unprojected representations (the smoothing tends to worsen the results and with Braw we are at the end of the roughness scale), while we can hope to achieve better results by other choices of bases to project the data onto.

### 4.3.3 Applying PDA to an Eight–Class Problem: State and Temporal Changes

The SPE (Fig. 4.6) curves show some differences when compared to the 2–way classification problem. There is an increase in the EDF–at–min–SPE, hinting at more extracted structure in this more complicated, multi–class setting. The Braw representation, which again has the highest EDF, also has higher SPE (0.555) than BSmooth3, the winner in the unprojected group (SPE=0.550). B35 has the lowest SPE (0.545) and second lowest EDF–at–min–SPE (62.1). By chance, one would expect 0.757 for the no–information SPE.

More importantly, Figure 4.7 shows that the model is able to extract two components, which we a priori consider important: state and temporal effects. The class centroids projected on the first canonical image arrange themselves largely in the temporal order in each of the two states. There is also a large jump between first active/baseline scan (class 1 and 2) and the second of these scans. This is intuitively appealing as the subjects are probably still learning (in the case of the first active scan) or reflecting on the tasks to be performed and generally adjusting to the situation (in the case of the first baseline scan). The first canonical image clearly separates the baseline (odd class numbers) and
Figure 4.6: Square Prediction Error (SPE) curves, in an 8-class problem, as a function of Equivalent Degrees of Freedom for 8 penalized discriminant models with different representations: 5 tensor-product B-spline projected datasets with varying numbers of basis functions (B15 to B35) and 3 unprojected raw (Braw) and smoothed (BSmooth3 and BSmooth5) datasets (thicker lines). The markers show the minima for each curve.
active scans. It is worth repeating, that the model had no knowledge about both states and temporal ordering: its task was simply to differentiate among 8 unordered classes. These two components come up as the first two canonical variates and thus account for the majority of the between-class variance (62%). The figure also shows: (1) potential interaction between the two experimental states and the temporal process, as the means in the two groups arrange themselves on lines with different slopes, and (2) a first scan effect in both states, scans 1 and 2.

By examining the SPE curves we note that the EDF–at-min–SPE are somewhat higher (62.1 vs 56.8, for B35 and 75.0 vs 68.8 for Braw) suggesting that more information is extracted from the data, when the temporal structure of the problem is not included in the within-class covariance. This improvement is consistent with our observation that the temporal structures seem different for two experimental states, potentially violating the common within-state covariance assumption for the two-class analysis. This setting also
demonstrates that the projected representations are potentially even more useful in more sophisticated situations: the lowest SPE for unprojected data is achieved with BSmooth3, and for projected data with B35 somewhat reversing the trend found in the two-class problem. This shows, even more clearly than in the two-way problem, that to extract more of the *generalizable* structure we need to impose some constraints. We have attempted to compare the first canonical image from this PDA (corresponding to the experimental state classification) to the canonical image obtained in the two-class paradigm, but one problem is the arbitrary rotation allowed in the space of the first two canonical images. As a possible future work we may develop a PDA model where the first Canonical Variate is specified from the two-way analysis, and apply PDA to the eight-class problem to discover secondary structures.

### 4.4 Result with Wavelet Expansions

The 182 scans have been preprocessed with a Discrete Wavelet Transform, which is equivalent to projecting them on a wavelet basis. Two families were used: Daubechies orthogonal wavelets and Coiflets, which is also an orthogonal family. We have investigated order 2 and 3 Daubechies wavelets (*Daub2* and *Daub3*) and order 2 Coiflets with 6 coefficients (*CoifN6*). Daubechies wavelets are perhaps the most famous wavelet family. They were constructed to be as symmetric as possible (the only fully symmetric, orthogonal wavelet with compact support is a Haar wavelet). The order number refers to the highest moment of the wavelet function that is equal to zero: Daubechies wavelet of order 2 has mean and second moment equal to zero. This is directly related to the smoothness of the wavelet. Coiflets have additional zero-moment requirements on the scaling function (Burrus et al., 1998).

The Donoho & Johnstone thresholding bring about a tremendous dimension reduction. The thresholded *Daub2*, *Daub3* and *CoifN6* representations result in 1884, 4187 and 3850 wavelet functions respectively. This has to be contrasted with 9670 for *B25* and 25,071
Given the much better prediction results achieved by wavelets, this great reduction of dimensionality seems to have successfully decreased the variance of projections.

We first look at the 2-class problem that extracts the single baseline-activation image. The top panel in figure 4.8 compares one particular wavelet representation with B-splines and unprojected results. The improvement is dramatic and is much larger than the difference between B-splines and unprojected representations. We offer one possible explanation for this improvement. Wavelets combined with thresholding offer a great reduction in dimensionality, without lowering the discriminatory capability of the PDA. Reduction of dimensionality has the effect of lowering the variability of the results: that is of the estimated canonical variates, hence projection and hence posterior probability estimates. It seems that this reduction in variance is much greater than the associated increase in bias of these estimates. In fact, due to the scale-space tiling property of wavelets, even the thresholded wavelet basis may have better resolution than the much higher dimensional B35 spline basis. The thresholding attempts to keep the high resolution wavelets only where they seem to be needed to estimate the brain function well, and reduces the resolution elsewhere.

In the bottom panel we compare the three wavelet families with and without thresholding. The thresholding helps the classification problem somewhat, particularly for the two Daubechies families. The Daub2 is a winner with $\text{SPE} = 0.075$ at $\text{EDF} = 74.6$. As this is the coarsest wavelet family it indicates that the discrimination problem hinges on a well defined, sharp structure which is best picked up by the low-order Daub2 wavelet. The D&J thresholding gives small, but consistent improvement.

Figure 4.9 shows a few slices of the Canonical Eigenimage that results from applying PDA with ridge hyperparameter optimization using two representations: Daubechies or order 2, and B30. The wavelet representation has a much sharper focus on the activated areas than B-spline, with much smaller "bleed-over" from neighbouring pixels or slices. On
Figure 4.8: Top panel compares wavelet and B-spline results in the 2-class problem. Shown are Daubechies order 2 thresholded wavelet basis compared to raw and B-spline representations using .632+ Bootstrap estimate of squared prediction error. The bottom panel shows SPE curves for the two-class problem for various wavelet families. We compare Daubechies order 2 and 3 family and order 2 Coiflet system. For each family we investigate two thresholding strategies: simply "peeling off" one finest detail level in each dimension (32 x 32 x 16), and Donoho and Johnstone VisuShrink hard thresholding rule.
Figure 4.9: Visual comparison of Wavelet (top row) and B30 representations in the 2 class problem. First three slices show portions of the cerebellum, next two display the midbrain portions, and the last three slices depict the activation of the cortex. The grayscale image is the anatomical MRI scan in the Talairach space and the CV is overlaid on top of it using the hot-metal color coding. Both images were created using EDF that minimized the SPE: 74.6 for Wavelet vs 53.6 for B30.

the other hand, it still has fewer spikes and speckles than the unprojected representations such as Braw or t-map (not shown) which improves interpretability.

Figure 4.10 compares the Wavelet and B-spline results using a corner-cube environment (CCE) (Rehm et al., 1998). CCE finds several connected areas with high average activation (here: above 99 percentile) and with pre-set minimal volume, for images being compared. These areas, called CCE foci, are then displayed using stems and projections onto the walls of the 3D volume. The figure shows clearly that the B-spline results are smoother and more spread than wavelets. Also, the wavelet PDA shows some smaller regions of activation that are either absent or have much smaller activation levels in the B-spline volume. This is due to the CCE algorithm: except for the major centers of activation (motor and auditory cortices) the relative levels will be lower in the B-spline CV due to imposed smoothness, which causes them to be larger but suppresses the peaks, and thus prevents them from being picked up by CCE.
Figure 4.10: Comparing the B-splines (B30) and wavelets (Daub2Thresh) Canonical Images using a corner-cube environment of Rehm et al. (1998) in the two-class setting. Except for the three major overlapping regions, the foci have been fit inside a ball of the same volume as that of corresponding focus. Blue foci correspond to B30 Canonical Image.
Figure 4.11: Squared Prediction Error for various wavelet families in the 8-class problem. Three wavelet functions are investigated: Daubechies order 2 and 4, and Coiflets order 2. For each family, we either remove top-scale wavelet coefficient level, resulting in $32 \times 32 \times 16$ wavelet coefficients or we apply D&J VisuShrink hard thresholding (Thresh). As a further dimension reduction technique, we investigate using all 7, or first 2 CV's to perform classification ($D7$ vs $D2$. )
In the 8 class problem that decomposes a full covariance structure associated with both time and experimental design wavelets perform surprisingly badly. Figure 4.11 shows the SPE curves for the same wavelet families and thresholding rules that we used in the 2-class case. We have also investigated restricting the dimensionality of the PDA model from full 7 to 2, since there is an a’priori belief that only baseline-activation and time structures are important for this problem. That is we only predict using the first two canonical variates. While this rank reduction helps somewhat, the results are still significantly worse from those of B-splines and raw representations: the lowest SPE for wavelet families is achieved by Daub2 family, D&J thresholded and restricted to two Canonical Variates (Wave64Daub4ThreshD2). It achieves SPE=0.672 at EDF=61.7 which we compare with 0.545 at EDF=62.1 for B35 representation.

Figure 4.11 shows that dimensionality reduction greatly improves the prediction for wavelet families, while, as was the case in the 2-class setting, the thresholding strategy seems less important. This suggests that the errors may be driven by variance, which is reduced when only 2 CV’s are used. We suspect that the common covariance assumption may be grossly violated in the wavelet domain when all eight classes are used. Some support for this assertion comes from observing the curves generated by using all 7 CV’s. They achieve their minima at very low EDF, as compared to B-spline representations, and rise sharply afterwards. Since large ridge penalty (and hence small EDF) works to counteract the effect of unequal covariance matrices, these would indicate that the common within-covariance matrix assumption may be badly violated.
Chapter 5

Static Force fMRI Analysis

In Section 2.4.2 we described the static force data. In this chapter we will extend the methodology developed for FOPP task. Our goal is to remain in the same paradigm as before: develop a descriptive tool to offer several views of the data, as driven by the experimental setup, but taking into the consideration the residual covariance structure. That is, we would like to look at the data through the canonical variates, which describe the experimental “gradient”: where do the experimental conditions really make the difference. We feel that even though the task in front of us is not a classification task, the Discriminant approach to the data is still suitable, as it disassembles the between-conditions covariance structure into orthogonal pieces of decreasing influence.

5.1 Modeling the time series effects: Time–Smoothed PDA

The main challenge of this data, as contrasted with the PET FOPP data, is the existence of time series effects on a much finer scale than before. In the PET data, we dealt with the time series by extending the class structure (our eight-way PDA analysis), and therefore
allowing the time effects to be arbitrary. The staticForce data contains the 8 subjects' time series, each of length 91, where each image is taken in 4s intervals. The force levels, which are the experimental conditions here, are super-imposed on this time-series, and there are about 8 scans during each instance of a force condition (about 11 for baseline). It is reasonable to suspect, that some part of the variability in the scans is due to the time-dependent changes independent of the conditions. These could be related to time drifts in the MRI machine (the simple linear drift that almost always accompanies the fMRI series has been removed in the preprocessing stage, but there may be "higher order" changes) to the hemodynamic processes in the brain that occur during a given instance of a condition and may possess a systematic structure, and, as before, to the long-term brain processes like adaptation, over-learning and fatigue.

It seems intuitively clear, that there should be some continuity in the time-series of scans, since they are taken in every 4s. This means that consecutive scans within the same condition instance should change in a smooth way, apart from noise. We may force smoothness onto the result similarly as we forced spatial smoothness onto the canonical variates. The difference here is that we are forcing smoothness between the scans that constitute the observations, rather than within the result.

5.2 Introducing Between–Scan Smoothness within the Discriminant Framework

The LDA algorithm may be cast in terms of the orthogonal projection operator, $P_Y$ that projects the scans onto the structure of $Y$. In the typical LDA, $Y$ is just a $N \times J$ indicator matrix, and then:

$$P_Y = Y(Y^TY)^{-1}Y^T$$  \hspace{1cm} (5.1)
projects any data onto the class structure. For example, if, as before, $X$ denotes the $N \times p$ scan matrix, $P_Y X$ is the $N \times p$ matrix that has scan (row) $i$ replaced by the class-average of all scans in the class that the observation $i$ belongs to. The between- and within-class covariance matrices are (see also Eqn. D.1) $X^T P_Y X$ and $X^T(I_N - P_Y)X$, respectively. The idea here is to work with the $P_Y$ operator forcing smoothness between scans in the time series.

We can develop the idea intuitively, as follows. Initially, we could partition the time axis into non-overlapping bins, of say three scans each (about 12s). We could designate each bin as a separate class. We would then have about 30 classes, just from the time structure alone (We will introduce the combination of the time and force levels effects later). If we assume that the scans are in temporal order, then the first four rows of the $Y$ matrix would look as follows:

$$
Y = \begin{bmatrix}
1 & 0 & \ldots \\
1 & 0 & \ldots \\
1 & 0 & \ldots \\
0 & 1 & \ldots \\
\vdots & \vdots & \ddots 
\end{bmatrix}
$$

Our proposal is to replace the rigid 0/1 design above, which corresponds to square kernels on the time axis, with smoother kernels. If we pick a smooth basis like B-splines, we can achieve the desired effect by setting up the response matrix $Y$ to be an $N \times J_2$ matrix of $J_2$ B-spline basis evaluated at the $N$ time points. In fact any smooth kernel-shaped function could be used, and our intuition suggests that very similar results would then be obtained. B-splines have the advantage of compact support which, together with the banded penalty matrix, leads to efficient numerical implementations.

We still need to include the force levels, which is the main experimental design effect. Our proposal is to create the $Y$ matrix that combines the force level and the time structures
in a natural way. We could also impose smoothness onto the force structure but we chose not to, and allow arbitrary force level effects. This is feasible, since there are only 6 different force levels, and lets us assess the relationship between force levels and the brain response visually which could then be followed by a more formal investigation in a hypothesis testing framework.

To complete the story we propose to penalize the time–axis parameterization. It is natural to regularize the B-spline basis by the second–order penalty matrix, which penalizes the second–derivative of the resulting function to control its “wiggliness” (Hastie and Tibshirani, 1990. Green and Silverman. 1994. e.g.,). Therefore the complete proposal is to set-up the response matrix \( Y \) as:

\[
Y = [Y_f, Y_B] = [f_1, \ldots, f_{J_1}, B_1(t), \ldots, B_{J_2}(t)]
\]  

(5.2)

where \( f_j \) are indicator columns for force levels, and \( B_j(t) \) are \( J_2 \) B-spline basis functions evaluated on the \( N \) time points. Then the projection matrix is constructed:

\[
P_Y = Y(Y^T Y + \lambda_Y \Omega)^{-1} Y^T
\]

Here \( \Omega \) is a penalty matrix for the B-splines, with rows and columns, that correspond the force level basis, zeroed out, and \( \lambda_Y \) is another free hyperparameter that controls the exact amount of smoothness in the time domain. We call the resulting model a time–smoothed PDA.

### 5.3 The O(N) Algorithm for Time–Smoothed Penalized Discriminant

As presented above, the algebra associated with the time–smoothed PDA would entail computing \( p \times p \) matrices and \( p \)-vectors, where \( p \) is a number of voxels (or image basis functions) and is much larger than \( N \). We need to modify the algorithm presented in
Section 3.3.4. which was expressed in terms of the $N \times N$ matrix of inner products. Our approach is to construct and disassemble the $P_Y$ projection operator, which then lets us use the usual PDA algorithm.

Specifically, we start by creating $Y$ as in Eqn. 5.2. Then we compute $S = Y^T Y + \lambda_Y \Omega$ and the Singular Value decomposition of it:

$$S = U_Y D_Y V_Y^T$$
(5.3)

We then compute the normalized response matrix:

$$Y_n = Y S^{-1/2} = Y U_Y D_Y^{-1/2} V_Y^T$$
(5.4)

Since $D_Y$ is a diagonal matrix the inversion above is trivial. One can now easily show that running our algorithm from section 3.3.4. with a response matrix $Y_n$ from Eq. 5.4 is equivalent to eigen-analyzing $\Sigma^{-1}_A \Sigma_B$ (with both covariance matrices defined through the projection operator, $P_Y$ from Eq. 5.1) which is the holy grail of LDA.

There are many ways to display the results. Obviously, we will want to look at the Canonical Variates, but it is also important to understand what do the CV’s represent. The fastest way to assess that in the usual LDA, is to project the class means on a pair of CV’s and display the projections. In our case this corresponds to projecting force levels’ and time means. If $Y_f$ and $Y_t$ represent the indicator matrices for the force levels and the time points, respectively, we need (in the notation of Eqn. C.8):

$$M_x B_{LDA} = (Y_x^T Y_x)^{-1} Y_x^T X B_{LDA}$$
(5.5)

$$= \sqrt{(n)} (Y_x^T Y_x)^{-1} Y_x^T X \beta A^* D_{(c \sqrt{1-c^2})^{-1}}$$
(5.6)

$$= (Y_x^T Y_x)^{-1} Y_x^T A^* D_{(c \sqrt{1-c^2})^{-1}}$$
(5.7)

where $M_x$ is a matrix of time/force level means, and subscript $x$ stands for either time $t$ or force level $f$ structure. This also shows that we can examine the projected means in the $N$-space, without computing the expensive Canonical Variates.
5.3.1 Constructing the Second-Order B-spline Penalty Matrix

We point out in section 3.5 that it is desirable to use a “proper” second-derivative penalty matrix, \( \Omega \), to penalize the B-splines. Using such a penalty was computationally inconvenient for us in the case of B-spline expansion of Canonical Variates, but is completely feasible in the current case. To obtain the cubic smoothing spline representation of the time structure, we need a 91 B-spline basis with knots at unique time points and second-derivative \( \Omega \). Here we describe a computational trick that lets us avoid explicit construction of the B-spline basis matrix and \( \Omega \) by using an existing Splus smoothing spline function, `smooth.spline`. For a given \( \lambda \), `smooth.spline` delivers, among other things, predicted values, \( \hat{y} \):

\[
\hat{y} = B(B^T B + \lambda \Omega)^{-1} B^T y
\]  

(5.8)

where \( B \) is an \( n \times n \) matrix of \( n \) B-splines evaluated at the \( n \) design points (assuming all design points are unique). We evaluate `smooth.spline` \( n \) times, with \( n \) canonical basis vectors for \( y \) (i.e. at the \( k^{th} \) evaluation \( y \) is a vector of all zeros and a single unity at the \( k^{th} \) place), and with \( x \) which is a sequence of \( n \) design points. In our case, \( x \) is a vector with 91 time points for the fMRI sequence, \( x = [0.4.8...360] \). For each call to `smooth.spline` we get \( \hat{y} \) which is a row of the hat matrix; and thus after \( n \) evaluations we can reconstruct:

\[
H = B(B^T B + \lambda \Omega)^{-1} B
\]  

(5.9)

Now, B-splines are just one possible (and numerically efficient) basis to obtain a solution to the smoothing spline problem, (Eq. 3.60), but any other full-rank basis system will give the same fitted values. In particular, we can change \( B \) to the unity matrix and obtain the solution in the Natural cubic splines basis (eq. (2.10) Hastie and Tibshirani, 1990). That is, we would obtain:

\[
H = (I + \lambda K)^{-1}
\]  

(5.10)
where $K$ is a penalty matrix for the natural cubic spline basis. For $\lambda = 1$, we can compute $K$ from $H$ by eigendecomposing it:

$$H = UD\gamma U^T$$

(5.11)

then inverting and subtracting 1 from eigenvalues $\gamma$ and reconstructing $K$ from these and the eigenvectors:

$$K = UD\gamma^{-1}U^T$$

(5.12)

We can use $K$ with time-structure response $Y_B$ from Eq. 5.2 being just an indicator matrix.

5.4 Connections with Canonical Correlation Analysis and MANOVA

As we have shown in Section 3.3.4 the LDA is basically equivalent to Canonical Correlation Analysis (CCA) when the class-indicator matrix is used for $Y$. This connection is even more appealing in the presently proposed model. CCA does not put any requirements on the right-hand variables. Thus we may choose any representation for $Y$, for instance the structure shown in Eqn. 5.2. It is up to a researcher to make sure that the representation is sensible from the interpretation point of view. In the present context, we are seeking the canonical correlations of scans with both the force level and the time structure. In addition, it makes sense to parameterize the time axis using smooth basis functions to model part of the interscan correlations that exist due to proximity in time.

The penalization scheme proposed here is also appealing in the CCA context. The PDA regularization penalizes the left-hand side of the CCA equation, or modifies the norm for left-hand side Canonical Variates that correspond to the scan data. The penalization of time-axis B-spline basis, does the same to the right-hand side of the symmetric CCA equation. It is, again, up to a researcher to make sure that the penalization scheme is reasonable.
from the analytic point of view. The model proposed here is similar to the one described in Chap. 12 of Ramsey and Silverman (1997) which we summarized in Sec. 1.3: now both parts of the criterion (1.6) are treated as functional and regularized. One difference is that we have a mixed response (or $Y$) structure: fixed force levels and smooth time which we deal with using an additive model.

In Section 3.1.1 we hinted at the connection between classical LDA and MANOVA. Here we will show that the proposed time-smoothed PDA model has a similar connection with an appropriately parameterized MANOVA.

In Section 12.5 (Mardia et al., 1979) shows that the test of dimensionality in one-way MANOVA leads to similar results as the LDA. Specifically, if we assume the model:

$$i_i = \mu_{j(i)} + \epsilon_i$$  \hspace{1cm} (5.13)

for the scans $i_i$, $i = 1, \ldots, N$ that are in $J$ classes, and with the usual assumption of i.i.d. Gaussian errors $\epsilon \sim \mathcal{N}(0, \Sigma_{W'})$, we can first test the null hypothesis of equality between the class means $\mu_j$. If we reject the null then we have at least two options. We can test for specific contrasts, as in usual ANOVA, but we can also perform a more general test of dimensionality. That is, we can test whether the $J$ class means (which lie in the $p$ dimensional space of $i_i$'s) span the $r$ dimensional hyperplane, with $r < J - 1$). The GLR test results in the sum of the first $r$ eigenvalues of $\Sigma_{W'}^{-1}\Sigma_B$, which is the decomposition that also gives LDA results. Also, one can show (Mardia et al., 1979, Sec.5.4), that the estimated hyperplane for the class means can be parameterized in terms of the eigenvectors of $\Sigma_{W'}^{-1}\Sigma_B$, which are (up to a scale factor) same as CV's from LDA. Similar connection is proven by Hastie and Tibshirani (1995) who use these results to derive the EM algorithm for reduced-rank mixture discrimination.

We will now show that similar results hold for the model proposed in this chapter. If $j(i)$ denotes the class (force level) of scan $i_i$ and $t_i$ its time, then we can propose a 2-way
The Residual Sum of Squares (RSS) for this model has the form:

$$RSS = \sum_i (i_i - \alpha_{j(i)} - \beta(t_i))^T \Sigma_{jj}^{-1} (i_i - \alpha_{j(i)} - \beta(t_i))$$  \hspace{1cm} (5.15)

Let us perform a change of basis to orthogonalize RSS. This involves left-multiplying scans $i_i$ and factors $\alpha, \beta$ with $\Sigma_{jj}^{-1/2}$. I will retain the same symbols for all of these to avoid trivial notational changes. Let us now assume that the effects span the $r$-dimensional hyperplane with an orthogonal basis, that will turn out to be the Canonical Variates:

$$\Phi = [\phi_1, \ldots, \phi_r]$$  \hspace{1cm} (5.16)

That is we assert that:

$$\alpha_j = \sum_{k=1}^{r} \gamma_{jk} \phi_k; \quad \beta(t) = \sum_{k=1}^{r} \gamma_{k}^*(t) \phi_k$$  \hspace{1cm} (5.17)

We now parameterize the time effects to achieve the smoothness. We choose a basis for the time axis with $J_2$ components $B_l(t), l = 1, \ldots, J_2$ and use that to parameterize the coefficients $\gamma_{k}^*(t)$:

$$\gamma_{k}^*(t) = \sum_l \lambda_{lk} B_l(t)$$  \hspace{1cm} (5.18)

and then to get the time effects in this basis:

$$\beta(t) = \sum_k \sum_l \lambda_{lk} B_l(t) \phi_k$$  \hspace{1cm} (5.19)

The RSS now becomes:

$$RSS = \sum_i \left\| i_i - \sum_k \gamma_{jk} \phi_k - \sum_k \sum_l \lambda_{lk} B_l(t) \phi_k \right\|^2$$  \hspace{1cm} (5.20)
Let us write the RSS in a matrix form. In addition to CV matrix $\Phi$ that we defined in Eqn. 5.16. and the response vectors $y_i$ that are the rows of the matrix $Y$ from Eqn. 5.2. we define the $(J_1 + J_2) \times r$ matrix of effects' coefficients. $\eta$:

$$\eta = \begin{bmatrix} \Gamma \\ \Lambda \end{bmatrix}; \quad \Gamma = \{ \gamma_{jk} \}_{j=1:k=1}^{J_1:r}; \quad \Lambda = \{ \lambda_{lk} \}_{l=1:k=1}^{J_2:r}$$ (5.21)

The RSS can be written us:

$$RSS = \sum_i \| i_i - \Phi \eta^T y_i \|$$ (5.22)

Let $\Phi_C = [\Phi \Phi_\bot]$ be a $p \times p$ matrix with columns forming the orthonormal basis for $\mathbb{R}^p$. The first $r$ columns are the canonical variates $\phi_k$, as defined above, and the remaining $p - r$ columns are the orthonormal basis for the orthogonal complement of the CV space. We then have:

$$RSS = \sum_i \| i_i - \Phi \eta^T y_i \|^2_{\Phi_C \Phi_C^T}$$ (5.23)

$$= \sum_i \| [\Phi \Phi_\bot]^T (i_i - \Phi \eta^T y_i) \|^2$$ (5.24)

$$= \sum_i \| \Phi^T i_i - \eta^T y_i \|^2 + \| \Phi^T y_i \|^2$$ (5.25)

For any choice of orthonormal CVs, $\Phi$, we can minimize the RSS with respect to coefficients, $\eta$. Since the second term above does not depend on $\eta$, the result is just a regression of CV-projected scans $\Phi^T i_i$ onto $Y$ and thus the minimizing solution is:

$$\hat{\eta} = \arg\min_{\eta} RSS = (Y^T Y)^{-1} Y^T X \Phi$$ (5.26)

where $X$ is an $N \times p$ scan matrix.

To find the canonical variates, $\phi_k$, we just consider a case of the one CV, i.e., $r = 1$. Since the CV's are orthogonal, we can do the minimization separately, which simplifies the notation which would otherwise require traces of matrices. With just a single $\phi$, the
partially minimized RSS becomes:

\[
\text{RSS}_{r=1} = \|X\phi_1 - Y\tilde{\eta}\|^2 + \sum_i \|\Phi_i^T i_i\|^2 
\]

(5.27)

\[
= \sum_i ||i_i||^2 - 2\phi_i^T X^T Y\tilde{\eta} + \tilde{\eta}^T Y^T Y\tilde{\eta} 
\]

(5.28)

\[
= \sum_i ||i_i||^2 - \phi_i^T X^T Y(Y^T Y)^{-1} Y^T X\phi_i 
\]

(5.29)

\[
= c - \phi_i^T R\phi_i 
\]

(5.30)

The minimizing \(\phi_1\) can now be easily seen to be the first eigenvector of \(R = X^T Y(Y^T Y)^{-1} Y^T X\).

and in light of the orthogonality of \(\phi_k\)’s, the full solution to the minimization problem is \(\{\tilde{\eta}, \Phi\}\), where \(\Phi\) has first \(r\) eigenvectors of \(R\) in its columns. To finish the presentation we have to remember that the minimization was carried out in the rotated system. Therefore to project the unrotated scan \(i_i\) onto the hyperplane, we need to change its basis before projecting it onto \(\phi_k\)’s. Thus the final estimate of the hyperplane’s basis are the first \(r\) rotated eigenvectors of \(R\), or \(\Sigma_w^{-1/2}\phi_k, k = 1, \ldots, r\).

What we have computed are the MLE estimates of the successively higher-dimensional hyperplanes that are hypothesized to contain the force level and time effects. The proposed model was the 2 way MANOVA with B-spline parameterization of the time effect. This result also forms the basis for the GLR test of the dimensionality (as in (Mardia et al., 1979, Sec.12.5)) for the 2 way MANOVA model with the proposed parameterization of the time effects. To see that \(\phi\) are essentially the unrotated CV’s we note that they are the eigenvectors of \(R\):

\[
R = \Sigma_B = \Sigma_w^{-1/2}\Sigma_B\Sigma_w^{-1/2} 
\]

(5.31)

By Theorem A.9.2 of Mardia et al. (1979) we know that the \(\Sigma_w^{-1/2}\phi\) are then the eigenvectors of \(\Sigma_w^{-1}\Sigma_B\), which gives the LDA decomposition.

This connection between an appropriately parameterized 2-way MANOVA model and our proposal may be used to obtain further insights. It is now clearer that the penalization
of the B-spline basis reduces the effective dimensionality of coefficients $\eta$ and thus regularizes the time effects $\beta(t)$. This is on top of the crude regularization that is provided by limiting the canonical dimensionality. The B-spline penalization prohibits excessive variation of the time effects and thus forces the estimation procedure to explain the variability in other, hopefully more suitable, ways.

5.5 Penalized Discriminant Analysis of StaticForce data in B-splines and Wavelet domains

We have applied the PDA model with ridge regression to the StaticForce fMRI data described in Sec. 2.4.2. We use a 6 class structure: baseline and 5 force levels. There are 91 scans for each subject: 46 of them are in 6 baseline instances, and 45 in 5 active classes. With 8 subjects we have a total of 728 scans.

![Projections on first 2 CVs from the PDA model applied to StaticForce data](image)

**Figure 5.1:** Projections on first 2 CVs from the PDA model applied to StaticForce data

The figure 5.1 shows projections on the first two CVs for the StaticForce data using second-order Daubechies wavelets with thresholding. The right panel shows the results of the model that was fitted with $\text{EDF}=200$. The projections indicate that the PDA model extracts a reasonable structure: the first CV divides baseline and active scans and the
second CV corresponds to force levels: apart from class 6 (1000g) the scores on this CV increase with a force level. The static force experiment with 1000g is seen as somewhat different from the other ones: it is apparently quite hard to maintain the force of this level for 45s through the experiment. It is reasonable to expect that different brain structures will be involved.

We have performed an extensive predictive analysis study using 3 thresholded wavelet families, as described in section 4.4 and B25 B-spline basis. We also used reduced rank discrimination with 2, instead of 5 canonical variates used for prediction. In all these cases, both SPE and misclassification rate achieve their minima at very low degrees of freedom. Reduced rank helps keep the errors from increasing rapidly for higher EDF, but is otherwise not better than the full rank model. The minima obtained are invariably around the base rates for this data: rates that would be achieved by the model that predicts based on the prior probabilities. The base misclassification rate is 368/728 or 50.55%. Similarly, the base SPE is $0.5055(1 - 0.5055)^2 + 5 \times 72/728(1 - 72/729)^2 = 0.525$, since for each of five classes there are 72 scans (out of 728). The left panel of figure 5.1 shows the prediction for the PDA model fitted with EDF=8. It shows that the model is doing exactly what we suspected: predicting the a'priori most probable class regardless of the scan characteristics. That the minimum error occurs at these low degrees of freedom, suggests that PDA is not able to effectively predict the class of each particular scan. The predictive failure of PDA does not completely disqualify it from analyzing the data. As we saw in fig. 5.1. PDA extracts two reasonable components: it is their generalizability over subjects that is in question. Also there is an important time axis here which is completely ignored in the current analysis and which may constitute a much stronger effect than the condition under which the scan was taken.
5.6 Applying Time–Smoothed PDA model to the StaticForce data

We apply the time–smoothed PDA model to the StaticForce data. We use B25: tensor-product of 25 B-spline basis functions in each dimension, for image representation, and 91 B-splines basis function for time axis with knots at the unique data points. We use the “proper” B-spline penalization with the second-order penalty matrix.

![Time and Force Level Projections: B25 with EDF=50, lambda_Y=10](image)

Figure 5.2: Projections of and time-points (first row) and force level means onto the first four Canonical Images using the time–smooth PDA model with B25 Tensor Product B-spline basis and B-splines for the time axis. Force levels were: 1–baseline, 2–200g, 3–400g, 4–600g, 5–800g, 6–1000g. The time-structure penalty hyperparameter was set at \( \lambda_Y = 10 \).
Figure 5.2 shows the projections of the average of all scans at a given time point (first row) and a given force level (second row) onto the first three Canonical Images. These account for about 85% of the total variance. The hyperparameters were set at EDF of about 50, and $\lambda_Y = 10$. These were not optimized.

The first CV accounts for almost 68% of the variance. It clearly separates the baseline state from all the others. The corresponding time projection shows a possible quadratic time relationship for baseline states: it starts higher for first baseline state then decreases and increases back for last baseline. It may correspond to some kind of “anticipation” however the effect is much weaker than the baseline-activation effect and thus hard to interpret. In addition to this activation effect, the force levels are ordered on this CV which may provide some insight into how the force level is modeled within the brain. The second CV (about 10% of the variance explained) shows a curious time trend which is quite linear for most of the time interval. This may be related to a number of things (including MRI machine trend) and requires further scrutiny. The corresponding force level effect is also strong and is mostly geared towards distinguishing the third force level. The third CV (about 7% of variance explained) has a rather noisy time structure, with some periodic behaviour, mostly visible in the early time and related to the baseline-active changes. The corresponding force level display hints of structure in the brain which is associated with the strength of the force exerted. The baseline condition is an exception having the same score as force level 4. This may indicate that the baseline is quite distinct from zero-level force that it is supposed to model and should not be considered together with other force levels. The force level ordering on this CV suggest that it may be the most interesting to look at when searching for the answers on the relationship between the amount of force applied and the controlling brain structures.

We also note some relationships between the CVs discovered here and in the PDA modeling in the previous section. The first CVs of both models are clearly quite similar.
The time-smoothed PDA CV described here has a stronger association with the force levels in addition to modeling the baseline-activation changes. The second CV of the time-smoothed PDA model seems to be a novel discovery as it is strongly related to the time axis. The third time-smoothed CV is similar to the second CV of the PDA model. The difference is that it does put the 1000g force level in the right order with respect to the other forces. It may be occurring because of the explicit modeling of the time axis: this force level does not occur as a first active state in any of the 8 subjects. Thus its unexpected score on the second CV in the PDA model may be a result of the confounding of time-order effect.

In general, we believe that the time-smoothed PDA model is potentially very useful for modeling fMRI data. It provides a decomposition of the covariance matrix along the variance components induced by the experimental setup but it also takes into the account the strong time series effects. Currently we lack the criterion for optimizing the hyperparameters and assessing goodness-of-fit, since the classification performance is not longer useful in this paradigm, but we mention a possibility in the next chapter.
Chapter 6

Conclusions and Extensions

The presented paradigm provides a flexible option for constructing summary images from both PET and fMRI studies. It takes into account different experimental setups and is flexible enough to accommodate two smoothness sources known to exist in the data: spatial and temporal (fMRI). For PET studies, the predictive analysis constitutes a validating technique that gives a researcher a degree of confidence in the resulting images and allows him/her to make choices (e.g., among different bases or in number of degrees of freedom). Our method has some advantages over others proposed in the literature:

- It deals with the full 3D (4D, with the temporal extension) data in a cohesive way, without a need to delineate the regions of interest or perform voxel-based analysis.

- It acknowledges existing spatial and temporal smoothness in a simple way via basis expansion, which has an added benefit of reducing dimensionality and thus possibly variance.

- Using a fixed basis and regularization it avoids the SVD basis which are wholly data and variance driven and thus does not take into the account the spatial nature of scans. When using an SVD basis one also faces a task of choosing a subset of them, which is an exponential complexity task, which we avoid by regularization with a
single hyperparameter

- It provides a simple predictive framework for assessing the goodness-of-fit. While prediction is not a goal per se in neuroimaging studies (although it may become one as the diagnostic value of the PET/fMRI brain scans increases), the Prediction Error provides a simple one-number summary of the effectiveness of the resulting image(s).

- We develop an associated, computationally appealing, algorithm that avoids constructing huge covariance matrices.

### 6.1 Extending the predictive analysis

We use prediction error estimates as a way to both choose a basis and hyperparameters and to validate a resulting image. We would like to extend this paradigm to the Two-Way PDA model (Section 5.2). One possibility is to use the MANOVA connection: if we think of Canonical Variates (CVs) as a basis for the scan space, we can use bootstrap to estimate the Mean Square Error (MSE). Specifically, MANOVA tells us that if our model is correct, each scan is composed of a linear combination of Canonical Variates and an error term (Eqs. 5.14 and 5.17). We propose to validate the process by estimating the true MSE:

$$E_0E_{\mathbf{X}}\left\| i_0 - \hat{\alpha} - \hat{\beta} \right\|_{\mathbb{W}^{-1}}^2$$  \hspace{1cm} (6.1)

Here, the double expectations taken over the distribution of the training sets, \( \mathbf{X} \), and then over the distribution of independent test scans, \( i_0 \). We use the within-covariance rather than the Euclidean norm to orthogonalize the Canonical Variates, as in Section 5.4. This way, -2MSE is (up to an additive constant) a log-likelihood of a new test scan, \( i_0 \).

To proceed, let us first orthogonalize the system, as before (Sec. 5.4). What we have now are the estimated orthogonal canonical variates, \( \phi_k \), and we can write the norm in the
new basis system as:

\[
\sum_k \tilde{\gamma}_{0k} \phi_k^* ~ (6.2)
\]

for some canonical coefficients \( \tilde{\gamma}_{0k} \) which combine both the time and force level structure. The starred quantities refer to the rotated quantities, for example \( i_0^* = \Sigma_{W}^{-1/2} i_0 \).

We propose to estimate the MSE using the best linear coefficients, \( \tilde{\gamma}_{0k} \) for a given test scan, \( i_0 \), and a \( k \)th CV, \( \phi_k \). Our motivation for this is that we are interested in how well the estimated C\( V \)s represent the data, and \( \tilde{\gamma}_{k0} \) are a nuisance parameters in this context. By “best” we understand as resulting in the smallest MSE. It is trivial to show, since Canonical Variates are orthogonal in the rotated basis, that the minimizing coefficients, \( \tilde{\gamma}_{0k} \) are just projections of a test scan onto each CV. Since:

\[
\tilde{\gamma}_{0k} = i_0^* \phi_k^* = i_0^* \Sigma_{W}^{-1/2} \phi_k^* = i_0^* \phi_k \quad (6.3)
\]

and therefore we can project the test scan onto the unrotated Canonical Variate to calculate the coefficients. Using Eqn. E.2 we see that \( \tilde{\gamma}_{0k} \) may be calculated without actually resorting to the projection which is an expensive \( O(p) \) operation.

The MSE then is a double expectation of:

\[
\frac{1}{n_W} \sum_{i_0} \sum_k \tilde{\gamma}_{0k}^2 = \frac{1}{n_W} \sum_{i_0} \tilde{\gamma}_{0k}^2 = \frac{1}{n_W} \sum_{i_0} \tilde{\gamma}_{0k}^2 \quad (6.4)
\]

It should be possible to compute the first summand using only outer-product matrix \( G \) and the model fitted to the training set, or \( O(\mathcal{N}) \) quantities, without resorting to the expensive operation on the actual scans.

To estimate MSE (Eq. 6.1) we can use .632+Bootstrap or cross-validation. Given the estimates of \( \Sigma_{W}, \phi_k \) obtained using the bootstrap set, we would apply them to compute the MSE for the scans in the validation set, and average as before. Even in the “regular” PDA models, this could be a more appealing alternative to the prediction error that we use currently.
6.2 Comparing the Results Across Non-Predictive Paradigm

It would be of interest to compare the results of our model (Smooth Canonical Images) with these of other methods currently in use, such as t-maps, ANOVA/ANCOVA preprocessed PCA and Scaled Subprofile Model. For two classes, one possible way to compare these would be the ROC analysis. ROC curves are a measure of a classification model predictive power when we do not want to assume any thresholds for determining the class. The area under ROC curve represents the total amount of information about the class in the result, under linear model. One powerful feature of ROC analysis is that it is invariant under monotonic transformation of an image.

The preferred paradigm would be to perform a bootstrap study: for each bootstrap sample compute the summary image (Canonical Image, t-map, first Eigen-Image or Group Invariant Subprofile) and project each test scan onto it saving the score and the true class. At the end compute the ROC curves and areas under it for each model. Similar paradigm was proposed and tested on a set of simulated data, by Lange et al. (1999), and it has been warmly received by the community.

Different approach, that works for more than two classes was developed by Strother et al. (1998b), and called NPAIRS. It involves assessing the variability of the resulting image using pairwise permutation studies. Briefly, one performs a large number of experiments in which the data is split randomly in two halves. One obtains the summary image for each half and computes the correlation coefficient between them. The coefficients are averaged over many random samples to give a total variability measure of the image. One problem with NPAIRS is that it does not take into account the "bias" in the result: by bias I mean some measure of the relevancy of the resulting image: a useless method that always returns the same image would score perfectly in this system. However, for any "reasonable" method, especially if one that has been internally optimized using, for example prediction error, NPAIRS gives a useful indication of the total variability.
6.3 Wavelets and Basis Selection Techniques

We have experimented with two kinds of basis: Tensor–Product B-spline and wavelets. There are many other possibilities of course, and even within these two meta-families a great many more things may be explored.

Our current approach with B-spline basis is to delete those basis that fall outside of the overall mask. We have not conducted systematic experiments to check whether the overall mask should be an AND or an OR of all masks, or perhaps something in between. More generally, some basis selection (a’la wavelet denoising, perhaps) may be useful. Our approach has been to shift the burden of basis selection onto the ridge penalty. However this may be an over-simplistic strategy and some combination may be desirable. One possibility would be to use a sum-of-absolute-values penalty, like the LASSO strategy of Tibshirani (1995). This offers a compromise between shrinking and basis selection and has been successful in many situations when compared to classical shrinkage of ridge regression.

On the other hand, we have mentioned before that it would be desirable to replace the ridge penalty with the second-derivative one to perhaps obtain a thin-plate spline solution. It may be possible to combine both strategies: a LASSO–like penalty for shrinkage and basis selection with thin–plate spline second derivative penalty. One major obstacle is to implement this in a computationally appealing way. that would, similarly to the algorithms presented in this thesis, avoid constructing covariance basis in the voxel space.

There exists a more systematic approach to selecting basis from many families. Wavelet packets and the associated Best-Basis Pursuit algorithms (e.g., Vidakovic, 1999) start with overcomplete dictionaries which contain redundant basis from the one family or multiple families. Best–Basis pursuit was developed for signal and image denoising, but the idea has been extended to multiple images and LDA by Coifman and Saito (1994, 1996), which describe the Local Discriminant Basis (LDB). LDB searches a large redundant basis dictionary in a rapid way picking these basis that have high discriminatory power. Crucial
to fast implementation is the additivity of the discriminatory measure. The examples are Kullback-Liebler divergence and Hellinger distance. We definitely feel that the area of basis selection, especially with wavelet basis, warrants more exploration.

We feel strongly that working in the wavelet domain has great potential in neuroimaging. It has a potential for great dimensionality reduction without affecting the results in major way. Indeed, if done correctly, one may obtain better results, as we saw in two-class PDA analysis for the FOPP data, due to decrease in variance. Our approach to basis selection based on image-wise thresholding is simple to implement but has large potential drawbacks. First, it does not pool information across scans. One possibility would be to perform a robust version of ANOVA analysis, using medians and absolute distances instead of means and square metric, to estimate the level/channel–dependent noise across scans. This would be in direct analogy to the current MAD estimator but would take variability across scans, as well as potentially between subjects and conditions, into account. It is quite possible that with the current strategy we may be deleting basis important for signal discrimination. It is also possible that in some parts of a scan, the variability is high, but that part still has some discriminatory power. Possibly more likely is the reverse scenario: there are low–variability regions in the scans with low discriminatory power, which currently survive the thresholding, perhaps at the expense of other regions, only because we do not take the discrimination problem into account when constructing the wavelet basis. One problem when extending the thresholding strategies to account for subject and condition effects is that it will require a different approach for validating the image: using fixed subject effects and conditions to select wavelet basis would currently require that this step be performed for every bootstrap sample, which would be computationally prohibitive.

Wavelets also offer a possibility for more intelligent penalization schemes. Each wavelet has a position and the scale associated with it, and thus we may use any prior information to differentiate penalties for different spots in the brain on different scales. For example,
we may penalize places with white matter or ventricles more, as they are not likely to participate in the brain function. The smoother representation is already penalized less: since there are twice as many wavelet coefficients at the next higher level, and since each coefficient receives the same penalty, collectively the ridge penalty “favours” smoother results: this could be enforced with location-based penalties mentioned above. Since all these schemes result in a diagonal penalty matrix they are easy to implement in the current algorithmic paradigm.

6.4 Inference and other issues

We have not done much work on region-specific inference. LDA and other multivariate techniques, as applied to images, are spatially global in nature and have mainly a descriptive appeal. We use prediction error to validate the procedure, but we have not made any attempts to designate specific regions as significantly activated.

A simple approach would be to assume normality, construct a T-map from a Canonical Variate and threshold using the Bonferroni correction. This may be much appealing here than it was in the case of t-maps constructed with Statistical Parametric Mapping (Section 2.3.1), since the basis coefficients, γ, that result from PDA and projected images are potentially a lot less correlated than the original scans. For one, some spatial correlation has been removed via basis expansion step, and PDA decorrelates Canonical Variates further by working in the rotated space. Similarly, it may be possible to utilize Gaussian Random Field thresholding of Worsley et al. (1992) on the reconstructed Canonical Image: one would assume that under the null hypothesis the canonical variate resulting from PDA applied to the projected data is a zero-mean field, as before. Then the covariance matrix (2.11) of the canonical image, that results from applying Eqn. 3.34, would be possible to calculate using the properties of the basis used. This may be more appealing than the SPM approach since the homoscedasticity is more tenable in the CV space and the Prediction
Error (or MSE) would give us some non-parametric confidence.

Another issue is that of canonical dimension reduction: choosing a number of significant canonical variates. For example, in the 8-class FOPP problem, we felt that first 2 canonical variates retain most of the structure associated with the problem. The easiest approach would be to extend the prediction error selection to choose the canonical dimension. Some asymptotic results may be tenable for this problem, however, since we are working with the summary data. A related issue is that of allowable rotation of canonical variates. For example, if only the first two CV's are designated as significant, what we really obtain is a two-dimensional view of the between-class covariance. It is quite possible that some rotation of the CV's would result in more appealing structures. Similar issues are present in the principal component analysis literature and a number of automatic rotation procedures (such as VARIMAX) have been developed.

LDA and PDA depend heavily on the ability to estimate the covariance matrix. Since a full-rank covariance matrix cannot be estimated we "cheat" a little by penalizing it, which effectively adds some volume in each direction. More importantly, LDA/PDA use pooled estimates of covariance over all classes, assuming the same shape. The alternative of estimating separate covariance matrices for each class, called Quadratic Discriminant Analysis is clearly untenable in the present case. Friedman (1989) offers one intermediate solution, termed Regularized Discriminant Analysis: first shrink the covariance matrix for each class towards a circular one (via ridge penalty), and separately ridge-penalize the average covariance matrix. Two hyperparameters result, which may be estimated with cross-validation or bootstrap, as in our case. Another possibility is the Mixed Discriminant Analysis (MDA) proposal of Hastie and Tibshirani (1995). There each class (or all of the data) is modeled by a mixture of Gaussians with a resulting mixture of covariance matrices modeling the common covariance structure. Each mixture-covariance matrix is penalized with a global
hyperparameter, which helps keep the degrees of freedom low. This proposal has a potential for modeling different shapes for each class using different mean and covariances for class-specific Gaussian components. MDA algorithm involves Expectation–Maximization (EM) iterations of basic PDA algorithm, and is thus computationally appealing in our case, since we can run the analysis in the $O(N)$ time.
Appendix A

Tensor Product B-Spline Basis

Tensor products provide a general way to extend a one-dimensional basis to more dimensions. See, for instance, Green and Silverman (1994), which deals with cubic spline bases or Ogden (1997), for an example of a tensor product basis in the wavelet domain. On the modeling side, Friedman (1991) develops a powerful and adaptive model using first-order tensor product B-spline basis with backward elimination.

B-splines, discussed at length by de Boor (1978), were developed as a numerically-efficient basis for polynomial splines. If $B_j$ denotes the $j^\text{th}$ B-spline basis, we compose a 3-D basis by multiplying the unidimensional ones:

$$B^{(3D)}_{j_1,j_2,j_3}(x,y,z) = B_{j_1}(x)B_{j_2}(y)B_{j_3}(z) \quad (A.1)$$

Thus the basis in 3D involves all possible products of the unidimensional bases. Figure A.1 shows the B-spline basis in one and two dimensions. One notable feature of B-splines is their compact support, which results in banded design and penalty matrices leading to efficient algorithms.
Figure A.1: 1D and 2D B-spline basis.
Appendix B

Basis Expansion of Canonical Variates

In this appendix we show that the basis expansion of canonical variates leads to LDA or PDA with projected data.

As in section 3.3, let a resulting canonical image be constrained as

\[ 3(x, y, z) = \sum_{j=1}^{\gamma_j} B_j(x, y, z) = B\Gamma \]  \hspace{1cm} (B.1)

where \( B \) is a basis matrix with one basis in each column, evaluated over the voxels (rows).

PDA can be expressed as an optimization problem: for two classes it finds \( \beta \) that maximizes \( \beta^T \Sigma_{\text{BET}} \beta \) subject to \( \beta^T \Sigma_{\text{W}} \beta = 1 \). For more than two classes one successively maximizes the criterion subject to the orthogonality with metric \( \Sigma_{\text{W}} \), which does not affect the following result.

If we add the constraint on \( \beta \) the criterion and condition become \( \Gamma^T B^T \Sigma_{\text{BET}} B \Gamma \) and \( \Gamma^T B^T \Sigma_{\text{W}} B \Gamma \), respectively. The between and within covariance matrices are:

\[ \Sigma_{\text{BET}} = N^{-1}(P_Y X)^T(P_Y X) \]  \hspace{1cm} (B.2)
\[ \Sigma_{\text{W}} = N^{-1}\{[(I - P_Y)X]^T((I - P_Y)X) + \lambda \Omega\} \]  \hspace{1cm} (B.3)
is an orthogonal projection operator on the column space of \( Y \) \( (P_1 = Y(Y^TY)^{-1}Y^T) \); for LDA, \( \lambda = 0 \), and for PDA \( \Omega \) is a chosen penalty matrix in the original space. It is clear now, that the PDA problem with smoothness basis constraint is an unconstrained PDA problem in projected data matrix \( XB \) and a modified penalty, \( \Omega^* \). Our choice, partly for computational expediency, and partly due to the limited knowledge of the true nature of the data in relation to the TPS basis, has been to set \( \Omega^* = I \).
Appendix C

CCA via Regression

Here we derive the (unpenalized) CCA algorithm via regression (see also Hastie et al., 1995). Let \( N \) be the number of observations (i.e., scans) with \( p \) variables as inputs (here, voxels or basis functions). We assume here, for the unpenalized version, that \( N > p \). Let \( X \) be the \( N \times p \) data matrix, and let \( Y \) be a \( N \times J \) class-indicator matrix, with \( J = \) number of classes. We can obtain the solution to the CCA problem from the regular SVD of:

\[
K = S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} = D_c A^T D_c \tag{C.1}
\]

where \( c_j \) are singular values, and \( D_c = \text{diag}(c) \). Anticipating the LDA problem (Appendix D), we are interested in left canonical variates \( B \), which we will refer to as CV's. These are the rescaled left eigenvectors of \( K \) (Eq.3.40), or:

\[
B = S_{xx}^{-1/2} B^* \tag{C.2}
\]

because of the normalization requirement of CCA. If \( X \) and \( Y \) have been centered then we can use the sample estimates:

\[
\hat{S}_{xx} = \frac{1}{N} X^T X, \quad \hat{S}_{yy} = \frac{1}{N} Y^T Y, \quad \hat{S}_{xy} = \frac{1}{N} X^T Y \tag{C.3}
\]

Thus the sample version of Eq. C.1 becomes:

\[
K = \sqrt{N} (X^T X)^{-1/2} \frac{1}{N} X^T Y \sqrt{N} (Y^T Y)^{-1/2} = (X^T X)^{-1/2} X^T Y (Y^T Y)^{-1/2} \tag{C.4}
\]
Now $K^TK$, whose eigenvectors are the right eigenvectors of $K$, is:

$$K^TK = (Y^TY)^{-1/2}Y^TXX^TY(Y^TY)^{-1/2}$$ (C.5)

If we choose orthogonal contrasts for classes (i.e., normalize $Y$ by $(Y^TY)^{-1/2}$) we obtain:

$$K^TK = Y^TXX^TY = Y^T\hat{Y}$$ (C.6)

where $\hat{Y} = X(X^TX)^{-1}X^TY$. The two-steps mentioned above, are now clearly visible: run a multi-response regression of (centered) data matrix $X$ on (centered and orthonormalized) group-indicator matrix, $Y$, (i.e.: $\hat{Y} = X\hat{\beta}$) and derive the right-hand eigenvectors of $K$ (Eq. C.1). by eigenanalysis of $Y^TY$:

$$K^TK = A^Tc^2A^T$$ (C.7)

Then obtain the left-hand eigenvectors using Eqs C.1, C.2 and C.4:

$$B = \sqrt{N}(X^TX)^{-1/2}K^T \hat{\beta}^TD^{-1}c$$

$$= \sqrt{N}(X^TX)^{-1/2}(X^TX)^{-1/2}X^TY(Y^TY)^{-1}A^T \hat{\beta}^TD^{-1}$$

$$= \sqrt{N}\hat{\beta}A^Tc$$ (C.8)

where $\hat{\beta}$ is a matrix of coefficients from the regression step.
Appendix D

Correspondence Between CCA and LDA variates

In this section we will derive the exact amount of rescaling needed to convert the CCA variates, \( B \), in the notation of Eqs. C.1, C.2 to LDA canonical variates, \( B_{LDA} \), proving Eq. 3.42.

LDA is a generalized eigenvalue problem: find \( B_{LDA} \) that successively maximizes \( B^T \Sigma_{BET} B \) subject to \( B^T \Sigma_w B = 1 \), where \( \Sigma_{BET}, \Sigma_w \) are between- and within-class covariance matrices, as in the previous appendix. If \( X \) has been centered, then, for LDA:

\[
\Sigma_{BET} = \frac{1}{N} X^T P_Y X, \quad \Sigma_w = \frac{1}{N} X^T (I - P_Y) X
\]  

Since \( B^* \) are left eigenvectors of \( K \), we have that:

\[
D_c = B^* K K^T B^*
\]  

\[
= B^* (X^T X)^{-1/2} X^T P_Y X (X^T X)^{-1/2} B^*
\]  

\[
= \frac{1}{\sqrt{N}} B^T X P_Y X B \frac{1}{\sqrt{N}}
\]  

\[
= B^T \Sigma_{BET} B
\]
which shows that $B$ diagonalizes $\Sigma_{BET}$. Also, since $\Sigma_W = X^T X - \Sigma_{BET}$:

$$B^T \Sigma_W B = B^* T B^* - D_{c^2} \quad \text{(D.6)}$$

$$= D_{1-c^2} \quad \text{(D.7)}$$

and thus we need to rescale: $B_{LDA} = BD_{(1-c^2)^{-1/2}}$ to meet the LDA constraint.
Appendix E

Deriving Predictions in the n-Dimensional Space

In this section we show how to derive posterior probability estimates using the fitted values and eigendecomposition step results.

From equations 3.47-3.49 we note that we need $x_0^T B_{LDA}$ and $\hat{\mu}_j^T B_{LDA}$ to obtain the estimates. Now, using Eqs C.8 and 3.42, we have that:

$$
\begin{align*}
x_0^T B_{LDA} &= \sqrt{N} x_0^T \hat{\beta} A^* D_{[c^2(1-c^2)]^{-1/2}} \\
&= \sqrt{N} \hat{y}_0 A^* D_{[c^2(1-c^2)]^{-1/2}}
\end{align*}
$$

(E.1)

(E.2)

where $\hat{y}_0$ is a vector of fitted values for predictor $x_0$.

The $K \times p$ matrix of class centroids ($\hat{\mu}_k$’s) may be obtained by $(Y^T Y)^{-1} Y^T \hat{X}$, where $Y$ is an $N \times K$ class-indicator matrix. Therefore the required $K$ quantities, $\hat{\mu}_k^T B_{LDA}$, are $(Y^T Y)^{-1} Y^T \hat{X} B_{LDA}$ and may be calculated, similarly as in Eq. E.2, using rescaled fitted values $\hat{Y}$ and $A^*$. By using Eq. E.2, all $K$ posterior probabilities are obtained for $x_0$. 

145
Appendix F

Ridge Regression With the Outer Product Matrix

Any penalized regression, can be expressed using only \(N(N-1)\) dot products of observations in \(p(=\text{number of columns})\) dimensional space, or using an outer-product matrix. \(G = XX^T\).

For our purposes, \(X\) is an image matrix, one image per row. We will work with ridge regression, but any penalized regression can be brought into ridge form by suitable change of basis.

Ridge regression is the solution of the following problem:

\[
\text{argmin}_{\beta} \quad (y - X\beta)^T(y - X\beta) + \lambda \beta^T\beta
\]  

(F.1)

By taking derivatives wrt \(\beta\) we have that:

\[
-X^T(y - X\beta) + \lambda \beta = 0 \\
-X^Ty + X^TX\beta + \lambda \beta = 0 \\
-XX^Ty + XX^TX\beta + \lambda X\beta = 0 \\
(XX^T + \lambda I)X\beta = XX^Ty \\
(XX^T + \lambda I)\hat{\beta} = XX^Ty
\]  

(F.2)
Thus the fitted values are

$$\hat{y} = X\beta = (G + \lambda I)^{-1}Gy = G(G + \lambda I)^{-1}y.$$  \hspace{1cm} (F.3)

and the predicted values at new design points, $X^*$.

$$\hat{y}^* = X^*\beta = G^*(G + \lambda I)^{-1}y$$  \hspace{1cm} (F.4)

where $G^*$ is an $N_0 \times N$ matrix of dot-products between $N_0$ new images and $N$ training images.

Another derivation looks at the projection matrix $S_\lambda = X(X^TX + \lambda I)^{-1}X^T$. Start with Singular Value Decomposition of $X$:

$$X = UD\sqrt{V}$$  \hspace{1cm} (F.5)

In our case, the images will usually span the $N$ dimensional subspace of the $p$ dimensional voxel space. To keep things general, let’s assume that the images span a $k \leq N$ dimensional space, i.e., $U$ is a $N \times k$, $D$ is a diagonal with $k$ strictly positive entries, and $V$ is $p \times k$. The matrices $U, V$ are column-orthonormal, i.e.: $I_k = U^TU = V^TV$.

Now,

$$S_\lambda = X(X^TX + \lambda I_p)^{-1}X^T$$  \hspace{1cm} (F.6)

$$= UDV^{-T}V(D^2 + \lambda I_p)^{-1}V^TDU^{-T}$$  \hspace{1cm} (F.7)

$$= UD^2(D^2 + \lambda I_k)^{-1}U^T$$  \hspace{1cm} (F.8)

$$= UD^2U^TU(D^2 + \lambda I_k)^{-1}U^T$$  \hspace{1cm} (F.9)

$$= XX^T(XX^T + \lambda I_N)^{-1}$$  \hspace{1cm} (F.10)

$$= G(G + \lambda I_N)^{-1}$$  \hspace{1cm} (F.11)

Lines F.7 and F.10 come from the fact that $\{V; (D^2 + \lambda I_p)\}$ and $\{U; (D^2 + \lambda I_N)\}$ are eigensolutions of $X^TX + \lambda I_p$ and $XX^T + \lambda I_N$, respectively, and that both of these matrices are invertible.
The PDA algorithm is composed of a penalized multireponse regression of an image matrix on group-indicator matrix, $Y$, followed by the eigenanalysis of $Y^T\hat{Y}$. Thus, if all we need is posterior probabilities at any (new) image, $x_0$ we can operate entirely in the space of observations. (much smaller than the space of predictors). once $G$ and $G^*$ are precomputed. We need one more step to deal with centering of $X$ matrix using only the outer-product matrix $G$. 
Appendix G

Centering the Design Matrix

For PDA, we need to center the matrix $X$ first, before computing $G$. However, to run the resampling validation studies, the training set will change for each bootstrap (or CV) iteration. plus we need to center the validation examples by the training set means. This would require precomputing the outer-product matrix for each bootstrap iteration separately, defying the computational advantage of this operation. We therefore need to find the way to compute the centered version of $G$ and a way to center validation set matrix, for any selection of training set examples, given an uncentered $G$ computed using all uncentered $N$ observations.

Let $G_{\text{All}} = XX^T$ be the outer-product matrix of all un-centered data. Any given bootstrap/CV sample specifies a subset of rows of $X$, as a training set, with the rest being a validation set. From there, one obtains $G$ and $G^*$ to get predictions (Eq. F.4). If $G_{\text{All}}$ is re-arranged, so that first $N_1$ columns/rows correspond to the training images, and last $N_0$ to the validation images, then $G$ and $G^*$ are $N_1 \times N_1$ upper-left, and $N_0 \times N_1$ lower-left submatrices of $G_{\text{All}}$, respectively.

The centering operator associated with any $N \times p$ matrix $X$ is:

$$
\hat{X} = X - \frac{1}{N} 1_N 1_N^T X
$$

(G.1)
where $1_N$ denotes a column $n$-vector of ones. We want $\hat{G} = \hat{X} \hat{X}^T$ in terms of $G$. We have, explicitly:

$$\hat{G} = (X - \frac{1}{N} 1_N 1_N') (X - \frac{1}{N} 1_N 1_N')'$$

$$= XX^T - \frac{1}{N} XX^T 1_N 1_N' - \frac{1}{N} 1_N 1_N' XX^T + \frac{1}{N^2} 1_N 1_N' XX^T 1_N 1_N'$$

$$= G - \frac{1}{N} G 1_N 1_N' - \frac{1}{N} 1_N 1_N' G + \frac{1}{N^2} 1_N 1_N' G 1_N 1_N'$$

$$= G - \Delta G$$

where $(\Delta G)_{ik} = \bar{g}_i + \bar{g}_k - \bar{g}$, and $\bar{g}_i$, $\bar{g}$ denote column (row) mean and over-all mean of $G$, respectively. For $G^*$, with $N_0$ validation points, we proceed similarly, using the column means of $G$:

$$\hat{G}^* = (X^* - \frac{1}{N} 1_N 1_N')(X^* - \frac{1}{N} 1_N 1_N')'$$

$$= G^* - \Delta G^*$$

where, similarly as before, $(\Delta G^*)_{ik} = \bar{g}^*_i + \bar{g}^*_k - \bar{g}$. and $\bar{g}^*_i$ is a mean of $i^{th}$ row of $G^*$. 


