GEOSTATISTICAL METHODS FOR SPATIO-TEMPORAL ANALYSIS OF FMRI DATA

by

Jun Ye

(Under the direction of Nicole A. Lazar)

Abstract

In this dissertation, I discuss and propose several geostatistical methods for functional Magnetic Resonance Imaging (fMRI) data. Geostatistics is a branch of applied statistics that focuses on providing quantitative descriptions of natural variables distributed in space or in time and space. Nowadays geostatistics is popular in many fields of science such as mining, environmental sciences, remote sensing and ecology. Functional Magnetic Resonance Imaging (fMRI) is a relatively new non-invasive technique for studying the workings of the active human brain. To date there has not been much work using geostatistical methods to analyze the brain in spite of the similarities of data types and questions of interest. Some recent exceptions are Spence et al. (2007), who used the variogram function to find neighbors of voxels of interest and Bowman (2007), who used the empirical variogram to define the spatial distance structure. My dissertation topic is applying geostatistical methods more broadly in fMRI data analysis.

There are three interrelated parts in geostatistics: Classification, Structural analysis, and Kriging. My research explores these three parts in detail as they apply to fMRI.

In clustering, I use geostatistical methods and sparse principal component analysis to analyze the fMRI data and establish a special clustering method for fMRI data time series; my results show that both techniques can effectively identify regions of similar activations.

A byproduct of my analysis is the finding that masking prior to clustering, as is commonly done in fMRI, may degrade the quality of the discovered clusters, and I offer an explanation for this phenomenon.

In structural analysis, I first introduce an alternative point of view of an axial image of the brain based on the empirical variograms during different time points, which gives a good understanding of how the brain reacts to the experimental task. I then deal with the variogram modeling of the same axial image, and use parametric and nonparametric hole effect models to look at the spatial character of the data. The models I use consider both physical and functional relations among the different parts of the brain, which distinguishes them from previous attempts to use variograms in fMRI. I show the effectiveness of the hole effect model compared with the regular monotonical model in describing the structure of the fMRI data.

In kriging, I choose filtered kriging as an alternative to spline smoothing to remove the measurement errors at the observed sites of the data, and maintain temporal consistency by controlling the noise to signal ratio of the smoothness – an idea borrowed from the smoothing function approach. This proposed new method incorporates combining both spatial and temporal information of the data into the smoothing procedure and can reduce the noise of the data in an intelligent way.

INDEX WORDS: Autocorrelation, Bessel function, cross-correlation, elastic net, functional neuroimaging, LASSO, hole effect, signal-to-noise ratio, smoothing ratio, thin-plate spline

GEOSTATISTICAL METHODS FOR SPATIO-TEMPORAL ANALYSIS OF FMRI DATA

by

Jun Ye

B.E., Soochow University, China, 1991

B.S., Nanjing University, China, 1996

M.S., Tennessee Technological University, 2004

M.S., The University of Georgia, 2007

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2008

GEOSTATISTICAL METHODS FOR SPATIO-TEMPORAL ANALYSIS OF FMRI DATA

by

Jun Ye

Approved:

Major Professor: Nicole A. Lazar

Committee: Gauri Datta
Yehua Li
Lynne Seymour
Andrew Sornborger

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

Dedication

To my parents, my mother-in-law, my wife and my son

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my dedicated advisor, Dr. Nicole A. Lazar. She has been such a great mentor by giving me inspiring guidance, indispensable suggestions, without which this dissertation would not have been possible. I am grateful for your constant encouragement and support accompanied me in all the time of research for and writing of this dissertation. Thank you for introducing me to the wonderful field of fMRI.

I want to take this opportunity to express my sincere gratitude to my committee members, Dr. Gauri Datta, Dr. Yehua Li, Dr. Lynne Seymour and Dr. Andrew Sornborger, for their patience, time, and suggestions in completing the dissertation. Not only have they provided guidance whenever needed, they have provided intellect.

I would like to extend my sincere thanks to fMRI studying group members, Dr. Jeongyoun Ahn, Dr. Abhyuday Mandal, Dr. Cheolwoo Park, Ana Moura Bargo, Ming-Hung Kao, for their valuable help and substantial consistent support throughout my study.

I wish to express my appreciation to Dr. Rebecca L. McNamee of the University of Pittsburgh for kindly providing the saccade data set and her conscientious assistance with physiological interpretations. I also would like to thank Dr. Nathan Yanasak, now at the Medical College of Georgia, for his assistance with the resting data acquisition.

Many thanks also go to my fellow colleagues and friends here for their constant support and for providing an atmosphere conducive to research. I will truly enjoy to work with these fine people again.

Finally, I want to thank my wife Ling Zhou and my son Andrew Ye, I could not have accomplished this dissertation without their love and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to fMRI

The brain is the most interesting but least understood organ in the human body. Even with the rapid development of neuro-imaging techniques in recent years, many problems still have not been solved. Especially, the detection of active brain areas is a challenging problem. *Functional brain mapping* is a useful method in analyzing the activations of the brain in research settings because it considers the dynamic brain changes over time. *Magnetic Resonance Imaging*(MRI) is a relatively new non-invasive technique to look at the structure of the brain. Functional brain mapping with Magnetic Resonance Imaging (MRI), usually called *functional Magnetic Resonance Imaging* (fMRI), uses *functional brain mapping* methods to study the workings of the active human brain by the MRI equipment; it has perhaps been the area of most rapid growth in image analysis (Huettel et al., 2004).

1.1.1 From MRI to fMRI

Compared with medical imaging techniques such as *Computer Tomography* (CT) or other X-ray based methods, MRI is a relatively non-invasive imaging technique which uses *Magnetic Resonance* (MR) to produce images of the inside of the human body (Noll, 2001). MRI is widely used to visualize the torn ligaments in the soft tissues; diagnose inflammations, infections or other irregularities that exist in organs; and so on. The basic imaging parameters in the pulse sequence of the MR scanner are time constants T_1 and T_2, where T_1 is the longitudinal relaxation time aligned in the direction of the magnet; and T_2 is the transverse relaxation time orthogonal to the direction of the magnet. Both T_1 and T_2 are tissue type

dependent and hence can be used to discriminate between tissues. A third relaxation time, T_2^*, measures the combined effect of magnet related and tissue related inhomogeneities. Image information are usually acquired by T_1, T_2 or T_2^*, or no relaxation time (Haacke et al., 1999; Huettel et al., 2004).

Functional MRI (fMRI) is a method to see the activations of the brain by the *blood oxygenation level dependent* (BOLD) contrast. It has been well-known that neuronal activity causes an increase metabolic demand, which changes the Cerebral Blood Flow (CBF) and the amount of deoxygenated hemoglobin in the related brain tissues. Since deoxygenated blood has a different magnetic susceptibility from oxygenated blood, the measured NMR signal is affected through the BOLD contrast effect. The changes in the ratio of oxygenated to deoxygenated blood are measured by the *hemodynamic response* (HDR) estimation, which is the change in MR signal on T_2^* images following neuronal activity (Huettel et al., 2004).

1.1.2 Experimental designs in fMRI

During an fMRI experiment, a sequence of magnetic resonance images is acquired while the subject performs specific cognitive tasks. Changes in the measured signal are used to identify and characterize the brain activity resulting from task performance. For comparing brain responses to different tasks during the experiment, there are two main approaches: one is the *blocked design*, the other is the *event-related design* (Lazar, 2008).

The *blocked design* is relatively simple and effective, and remains commonly used in fMRI (Jezzard et al., 2001). In the simplest case, the designed conditions are separated into two distinct states. Each condition is executed for a discrete period of time and the conditions alternate. The two designed conditions are usually chosen as one *control condition* and one *experimental condition*. The control condition is also called the *baseline condition* and the experimental condition is called the *task condition* (Huettel et al., 2004). For example, the analysis of eye movements (saccades) has long been used in neurology, since lesions in different brain structures may result in deficits in eye movement control, as may certain diseases

(Fischer and Everling, 1998). When subjects are instructed before a visual stimulus to look at the stimulus, this is called a *prosaccade*; when subjects are instructed to perform eye movements in the opposite direction from the location of a stimulus that appears in their peripheral vision, this is called an *antisaccade* (Hallett, 1978). Prosaccade and antisaccade experiments can be constructed as a blocked design in the analysis of deficits related to circumscribed brain lesions (Guitton et al., 1985), where prosaccade is the baseline condition and antisaccade is the task condition. The blocked design is simple, straightforward and it is adequate for many types of experiments in the fMRI research (Christidis and Reynolds, 2004). But it also has some disadvantages, e.g., it is predictable for the subject; it may be difficult to control the specific state for a long time in the block (e.g., the subject may not be always engaged in the task); it may not be appropriate for certain cognitive tasks (I will discuss it later) etc. (Christidis and Reynolds, 2004).

Event-related design associates brain processes with a sequence of discrete events, stimuli, or conditions which occur arbitrarily during MR scanning sessions (Rosen et al., 1998; Christidis and Reynolds, 2004). In the event-related design, stimuli from various experimental conditions (e.g., conditions A, B, C, and D) are presented individually in a randomized fashion, separated by an *inter-stimulus interval* (ISI) of a specified length (Christidis and Reynolds, 2004). Hence, the event-related design has more flexibility and randomization than the blocked design, making it especially useful in cognitive neuroscience (such as the field of memory research) (Rosen et al., 1998). For example, during an event-related study of different memory paradigms, subjects can demonstrate an ability to remember a specific stimulus (or not) through a particular action (Rosen et al., 1998). The event-related design is usually controlled by the ISI, which determines if the design is slow or rapid (Noll, 2001; Christidis and Reynolds, 2004). In the *slow event-related design*, the hemodynamic responses for different stimuli are not overlapped and each hemodynamic response results from one specific trial. In the *rapid event-related design*, the ISI is short between different stimuli and the hemodynamic responses for different stimuli are overlapped. The main dis-

advantage of the event-related design is that the design maybe too complicated particularly for the rapid event-related design, which brings difficulties in modeling the hemodynamic response (Christidis and Reynolds, 2004).

Given the advantages and disadvantages of the above designs, it is up to the researchers to choose particular design for their needs. It is possible to use both blocked designs and event-related designs in fMRI studies, where the blocked design is used for evaluating *state-related* effects and the event-related design is used for evaluating *item-related* effects (neural activity elicited by different individual items) (Otten et al., 2002).

1.1.3 Preprocessing steps

The data acquired from the MR scanner are raw spatial frequency space data, called $k-space$ *data*. Real brain images are reconstructed by Fourier transform. After the transformation, there are spatio-temporal variabilities across images, which include thermal noise, system noise, subject-related noise and task-related noise (Lazar, 2008). It is best, if possible, to remove these noise and artifacts at their sources, but some major sources are unavoidable (Kruggel et al., 1999). Hence, it is almost always necessary to do some data preprocessing prior to further statistical analysis.

To remove unwanted noise and artifacts, the most commonly used preprocessing steps include (Eddy et al., 1999; Jezzard et al., 2001; Huettel et al., 2004): *slice timing correction*, which shifts the different slices of data acquired from different time points to a fixed time point so that it looks as if all slices were scanned at the same time; *head motion correction*, which corrects the different brain images to the same position because the subject may move his (her) head during an experiment; *intensity normalization*, which rescales all images to the same intensity since the electrical or temperature effects of the MR scanner may change over time; *spatial filtering*, which uses a Gaussian filter as the signal of interest to blur each volume spatially, thereby maximizing the signal to noise ratio, where the filter width is used to control the degree of smoothness; *temporal filtering*, which removes the unwanted

components of a time series by high-pass filtering and low-pass filtering, where the high-pass filtering removes the slow drift-like trend from breathing, heartbeat etc., and the low-pass filtering removes the high frequency noise. Lab and research groups differ on whether, and how, these steps are implemented (e.g., not all researchers smooth data before analysis).

1.1.4 Description of fMRI data

An fMRI experiment yields a sequence of 3-dimensional images of the subject's brain. Each image comprises measurements of the MR signal over a grid of small regular volume elements called voxels. Voxel is an abbreviation for *volume element*, where volume is a three-dimensional figure measured in cubic units (Milot, 1998). A voxel is analogous to a pixel, which is an abbreviation for *picture element* in two dimensions. Voxel values contribute the intensity of the fMRI image (Wynn, 2000). Coordinates for brain maps are often defined using the following neurological convention for the three axes (Huettel et al., 2003). When the head is viewed from behind, the x direction is left to right, the y direction is front to back, and the z direction is top to bottom. An axial slice of an image consists of all (x, y) voxel locations for a fixed vertical location z (Figure 1.1). A coronal image consists of all (x, z) voxel locations for a fixed location y (Figure 1.2). A sagittal image consists of all (y, z) voxel locations for a fixed horizontal location x (Figure 1.3).

In this dissertation, I examine two data sets in detail, a saccade data set and a resting data set. The saccade data used here consist of 30 slices of size 64×64, taken over 156 time points, with images every 2.5 seconds, that is, $(x, y, z, t) = 64 \times 64 \times 30 \times 156$. A blocked design alternating anti-saccade and pro-saccade tasks was performed (Figure 1.4). The first two conditions (Anti 1; Pro 1) were set to allow for a 5 second delay in the hemodynamic response. The conditions during the 156 time points were thus: Anti, 1; Pro, 1; Pro, 12; Anti, 12; Pro, 12; ...; Anti, 12; Pro, 10. Hence there were 6 alternating blocks of size [Pro 12; Anti 12] in the design. Pre-processing steps performed on this data set included removal of spatial outliers, correction of head motion, outlier correction in image space, Gaussian filter

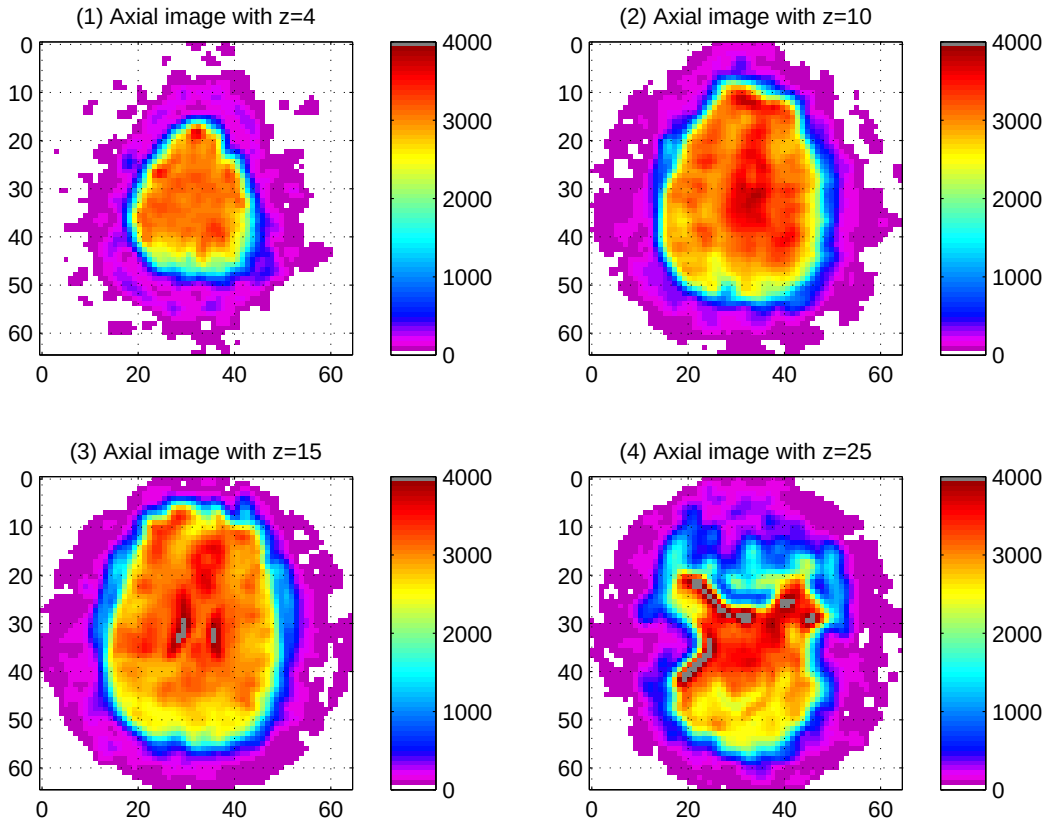

Figure 1.1: Axial view of the saccade data at time $t = 78$. (1) is axial image with $z = 4$, (2) is axial image with $z = 10$, (3) is axial image with $z = 15$, (4) is axial image with $z = 25$. z ranges from 1 to 30 in this data set.

smoothing with a radius of 2 voxels, removal of linear pixel-wise trends, and removal of linear drifts over time for each voxel.

The resting data contain three slices of size 64×64, taken over 1498 time points, with images every 2 seconds, that is, $(x, y, z, t) = 64 \times 64 \times 3 \times 1498$. This is a long range resting data set, i.e., no task was performed while the subject was in the scanner. The data were minimally preprocessed.

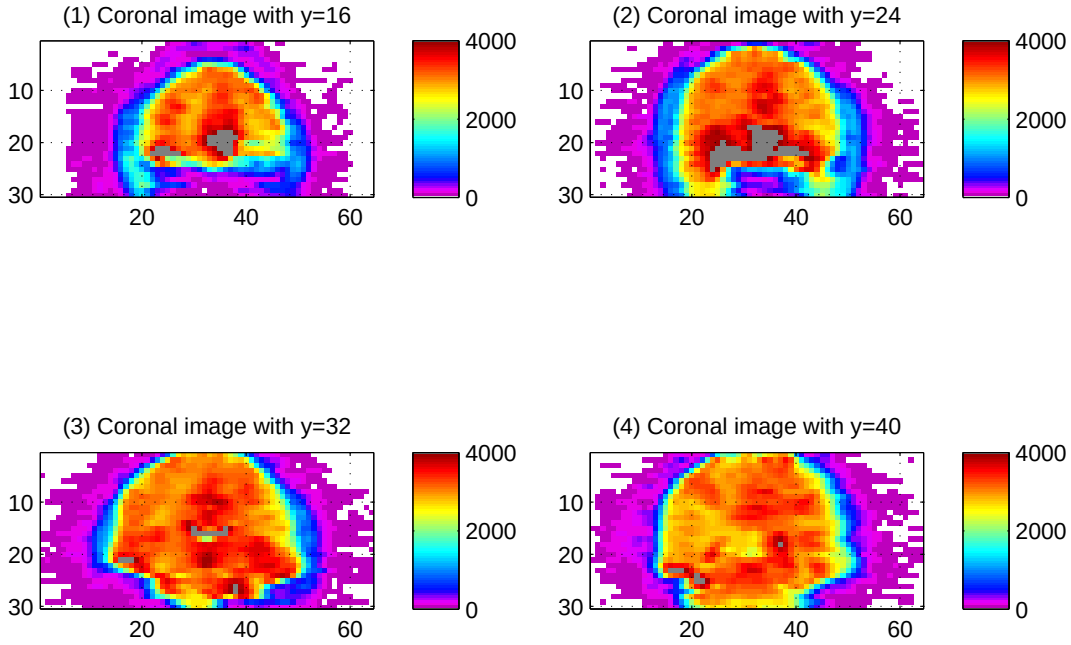

Figure 1.2: Coronal view of the saccade data at time $t = 78$. (1) is coronal image with $y = 16$, (2) is coronal image with $y = 24$, (3) is coronal image with $y = 32$, (4) is coronal image with $y = 40$. y ranges from 1 to 64 in this data set.

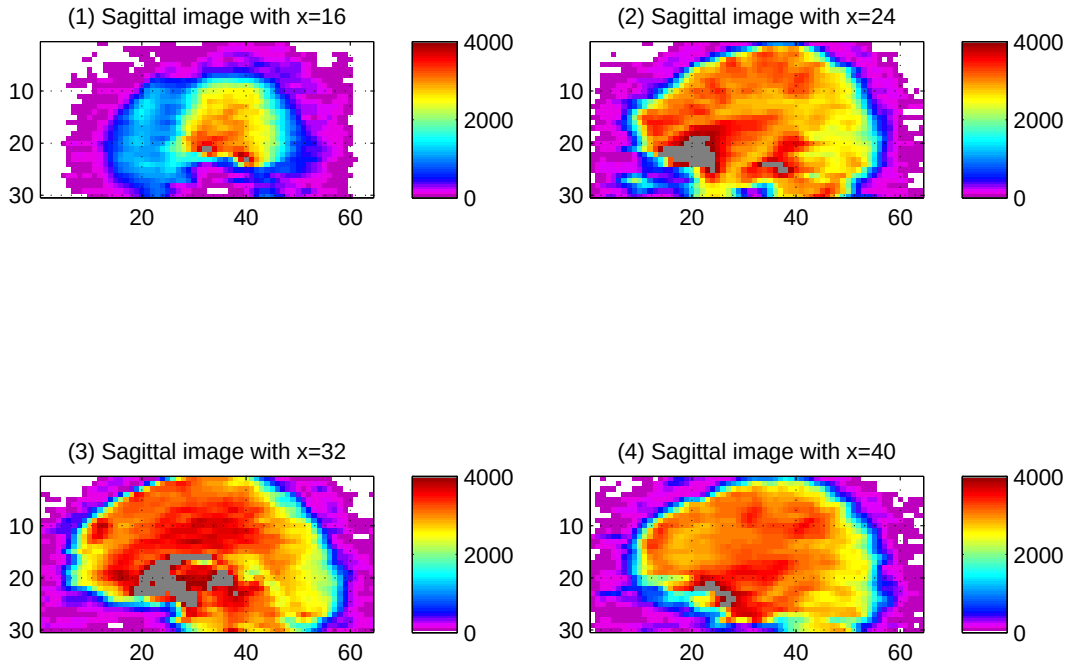

Figure 1.3: Sagittal view of the saccade data at time $t = 78$. (1) is sagittal image with $x = 16$, (2) is sagittal image with $x = 24$, (3) is sagittal image with $x = 32$, (4) is sagittal image with $x = 40$. x ranges from 1 to 64 in this data set.

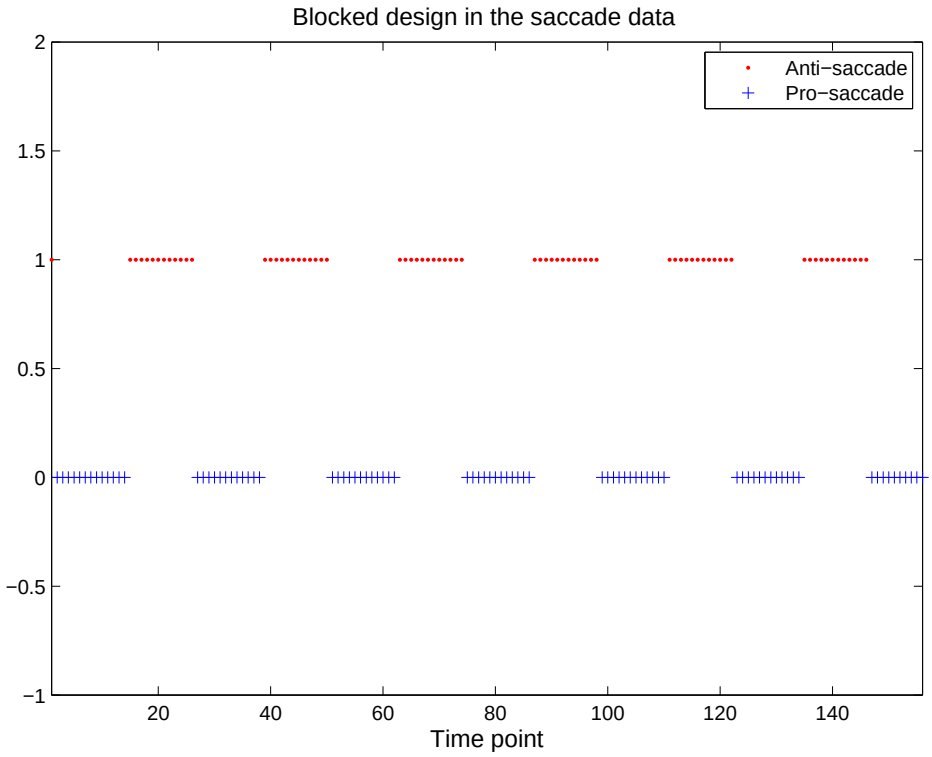

Figure 1.4: Blocked design in the saccade data, where prosaccade is the baseline condition and antisaccade is the task condition.

1.2 Statistical analysis in fMRI

According to Kosslyn (1999), there are two questions addressed in neuroimaging: one is when the particular structures and processes are involved in the activations over time; the other is how the brain implements the information processing. After the preprocessing steps, statistical analysis is carried out to answer these questions. But even after the data preprocessing, the change of true signal due to brain activity is small compared with artifacts and noise (Eddy et al., 1999; Noll, 2001), which brings a significant technical challenge to fMRI data (Noll, 2001). Hence carefully choosing statistical methods in the analysis is important. Currently, there are two approaches in the statistical analysis: *model-based* and *model-free* (Jezzard et al., 2001; Thirion, 2003; Huettel et al., 2004; Lazar, 2008.) .

1.2.1 Model-based approach

In the model-based approach, the data analysis depends heavily on the assumed model for the sampled data (Jezzard et al., 2001). Since each voxel time series is fitted by a model separately, this is considered to be a univariate approach (Thirion, 2003).

Basic analysis in model-based approach

In the model-based approach, the evaluation of activation status of a voxel (that is, whether it is considered active or not) is based on an hypothesis test, hence it is also called *hypothesis-driven analysis* in the literature (Huettel et. al., 2004). This method postulates that different voxels inside the brain have different responses to the stimulus. A parameterized function can be used to model the response and a statistical test assesses the level or shape of that response. Usually the analysis at each voxel can be used to build a map of statistic values for each voxel called a *statistical parametric map* (Huettel et al., 2004).

The *General Linear Model* (GLM) is a representative method (Friston et al., 1995). The idea is to build a linear model for the stimulus and to treat the data as a linear combination of dissociable factors. Under the assumption that the time series for each voxel is independent

of the others, the simplest form fits each voxel by a linear function $Y = XB + \epsilon$, where Y is the signal values at the n different time points ($n \times 1$ vector), B is the unknown model factor with m parameters ($m \times 1$ vector), X is the $n \times m$ design matrix of predictors, ϵ is the random error which is usually assumed to be normally distributed with mean 0 and variance σ^2. After estimating the model parameters by least squares, a $t-$statistic for the contrast $c^T\hat{B}/Var(c^T\hat{B})$ can be calculated, where c is the contrast parameter for the hypothesis test. Plotting all of these $t-$statistic back on a regular (e.g., 64×64) grid (for a single slice of data) results in the statistical parametric map mentioned in the previous paragraph. Statistical tests can be used to determine which voxels are likely to have had significant changes affected by the stimulus, e.g., which voxels are "active" and which are not (Thirion, 2003; Lazar, 2008).

Although the GLM method is easy to use, it makes some restrictive assumptions in its simplest form (Wang et al., 2003; Huettel et al., 2004; Lazar, 2008): voxels are independent without considering their spatial correlations; the error variances are kept the same at different time and locations; one model is globally fitted everywhere. Since these assumptions do not always hold in practice, it is necessary to use alternative methods to improve the validity of the GLM.

Temporal, spatial, and spatio-temporal approaches

Spatio-temporal correlation in fMRI data carries much information and should not be ignored. A great part of fMRI literature has concentrated on the partial improvement of the model-based approach by considering the spatio-temporal structure of the data.

Temporal approach For considering the temporal correlation of the model, two basic ways can be used (Lazar, 2008): one is analyzing the time domain directly; the other is analyzing the frequency domain from the transformed time series.

For the first way, there is a broad range of smoothing methods available from the classical time series analysis (Kruggel et al., 1999), such as moving average filter, finite impulse

response low-pass filter, autoregressive filter, Kalman filter and so on. The basic method in filtering (smoothing) is to prewhiten the correlated time series (Lazar, 2008). Because the correlation of the time series is considered now, the GLM model is change to $Y = XB + Z$, where Z is the error term with mean zero and variance $\sigma^2 KK^T$. This linear function is prewhitened as $K^T Y = K^T XB + K^T Z$, removing the correlation structure of the error term. Since KK^T is actually unknown, different filtering methods can be applied for the estimation.

For the second way, analysis is performed in the frequency domain (Lange and Zeger, 1997; Marchini and Ripley, 2000). The idea is to use the Fourier transform function

$$d_w(w_j) = \frac{1}{n} \sum_{k=1}^{n-1} w_k \exp(-i2\pi w_j \delta k)$$

where $w_j = j\delta/n$ are the Fourier frequencies, n is the length of the time series, δ is the sampling interval and $j = 0, 1, ..., n/2$. The model for the time series is transformed in the frequency domain $d_Y(w_j) = d_X(w_j)B + d_Z(w_j)$, where $d_Z(w_j)$ is uncorrelated for large n (Marchini and Ripley, 2000). This procedure is very useful for periodic stimulus designs (Thirion, 2003) and has the following advantages: the low frequencies of the noise are well separated (Lange and Zeger, 1997); the Fourier coefficients are almost independent for unbiased hypothesis testing (Marchini and Ripley, 2000).

Spatial approach In the spatial approach, Bayesian inference is gaining popularity in the fMRI literature (Lazar, 2008). Bayesian inference introduces priors to the statistical model and is made given the priors and the data (Thirion, 2003). These methods take advantage of the neighbors of the target voxels in the smoothing process. As an example, Hartvig and Jensen (2000) build a Bayesian spatial model based on mixtures. For the target voxel, the model is formulated through the marginal distributions with its neighbors showing active or non-active patterns, where the number of the neighbors is considered as 8 in a 3×3 grid (slice) of the voxel, or 26 in a $3 \times 3 \times 3$ grid (cube) of the voxel. For the neighbors with the non-active pattern, a prior with a pre-specified null distribution is used; for the

neighbors with the active pattern, three prior models with different marginal probabilities are presented in their paper which reflect different kinds of activations of the data.

Spatio-temporal approach Spatio-temporal models for fMRI jointly consider spatial information and temporal information. There are two main approaches here (Lazar, 2008): direct modeling and clustering of time series.

Direct modeling In this approach, all the model building currently considers space and time in the data to be separable. Descombes et al. (1998) and Kruggel et al. (1999) propose a *Markov Random Field* (MRF) model to restore (smooth) the signals. For the target voxel at one specific time point, they use a prior in their model which considers the information of its four closest spatial neighbors and two nearest temporal neighbors. In the spatial domain, anisotropic property of the neighbors is also considered. The posterior distribution of the spatio-temporal activation is implemented by the Markov Chain Monte Carlo (MCMC) simulation. Their method gives much better results than those obtained from Gaussian filtering, because it considers the spatio-temporal information of each voxel in smoothing rather than only blurring the voxel.

Katanoda et al. (2002) develope a separable spatio-temporal model based on multiple regression, where the covariance structure is a product of the spatial autocorrelation and the temporal autocorrelation. The time series is acquired in the frequency domain (Marchini and Ripley, 2000). In addition, the time series of six nearest neighbors are added to the regression for the target voxel (Katanoda et al., 2002). Generalized least squares method is used for parameter estimation. This method has the advantage of modeling the intrinsic spatio-temporal property together by the separable model (Wang et al., 2003).

Clustering of time series Clustering analysis is based on the assumption that the activated and non-activated voxels inside the brain have different spatio-temporal structures. Clustering itself is non-model based (Huettel et al., 2004; Keogh and Cordes, 2007). But

due to the high noise level in fMRI data, the results of clustering on the raw time series are often unsatisfactory and do not necessarily group data according to the similarity of their pattern of response to the stimulus (Goutte et al., 1999). Hence, many clustering techniques have focused instead on cross-correlation functions of the fMRI time series. Bandettini et al. (1993) consider the correlation between the data time series and a reference function from a "seed" voxel to characterize the temporal response of the brain to the paradigm. Goutte et al. (1999, 2001) suggest clustering voxels on the basis of the cross-correlation function between the fMRI activation and the experimental protocol signal. Yeo and Ou (2004) devise a "signal to protocol metric" by combining the cross-correlation of two fMRI signals and the Euclidean distance between voxels to do the clustering. Friman et al. (2002a) approach the problem by viewing the correlation between two time series as an objective function to be maximized with respect to some parameters, where one time series is from the target voxel and its eight neighbors; the other time series is from the hemodynamic function corresponding to the stimulus. Clustering on the cross-correlation function instead of the raw time series provides increased robustness and yields improved performance, but it depends heavily on the choice of reference function. I will explore clustering of time series in my dissertation.

Problems in model-based approach

The model-based approach provides a useful tool for fMRI data analysis, but it needs prior knowledge of the design and response, and strongly depends on the understanding of the anticipated hemodynamic response of the brain. For example, many experiments can not be addressed directly, or the model building may not be valid especially in the event-related design. Even if the hemodynamic response can be modeled "correctly", sometimes responses of the brain may not be directly corresponding to the task in the experimental design (Wang et al., 2003).

On the other hand, in the model-based approach, the hypothesis-driven analysis only can answer if the specific voxels with assumed responses to the stimuli are active or not, and can

not address the issues of connections between the active voxels and the main spatio-temporal patterns presenting in the image (Thirion, 2003).

1.2.2 Model-free approach

In the model-free approach, information from the fMRI data is extracted without reference to the experimental protocol; effects or components of interest are found from the intrinsic structure of the data directly. This approach does not depend on the estimation of any functions and is considered as an *data-driven analysis* (Huettel et al., 2004). In this approach, all voxels are considered simultaneously, hence it is also considered to be multivariate in nature (Thirion, 2003).

This approach uses the stratagem of the covariance paradigm (Sommer and Wichert, 2002) because it assumes that the different regions of the brain have different temporal covariance structure. Although this approach can not give the same types of answers as the hypothesis-driven approach, e.g., hypotheses on activation status of voxels, it can assess the patterns that appear in the data and how these patterns are spatio-temporally presented.

Examples here include principal component analysis (PCA) (Backfriender et al., 1996; Lai and Fang, 1999), independent component analysis (ICA) (McKeown et al., 1998 and 2003), and canonical correlation analysis (CCA) (Friman et al., 2002b) .

Principal component analysis *Principal component analysis* (PCA), also known as *empirical orthogonal function* (EOF), is a classical tool for analyzing large scale multivariate data. In PCA, the first component explains the largest proportion of the variance of the data, the second component explains the second largest proportion of the variance which is orthogonal to the first component, and so forth. Hence, PCA can be considered as an optimized method to simplify the data description. The purpose of data-driven analysis is to find the different areas inside the brain activated by the stimuli. Backfriender et al. (1996) propose an oblique rotated PCA method through *factor analysis* for fMRI. In their method, the rotated matrix is estimated by additional information about the expected factor images, where the

factor structures are assumed not to overlap each other (Jackson, 1991). To overcome the low density of the active signal, a local PCA in the selected *region of interest* (ROI) is calculated again after the regular PCA. To overcome the superposition of the active region with anatomical structure in fMRI, an artificial factor image of the brain anatomy is used. Hence, the final factor images can be extracted in several steps (Backfriender et al., 1996): at first, primary principal components which show the most prominent properties in the data are extracted, then a local PCA in the selected ROI is executed again, and finally, the local principal components are obliquely rotated. This method can well separate the activated region of the brain without prior knowledge about the experimental design.

Independent component analysis *Independent component analysis* (ICA) is a multivariate technique that has been widely applied in fMRI. ICA identifies statistically independent components in either the spatial or temporal dimensions underlying the observed data (Hyvarinen et al., 2001; Keogh and Cordes, 2007). The idea in ICA is similar to PCA, but PCA extracts orthogonal components based on the explained variance in the data, where the assumption is that the components are Gaussian and uncorrelated; on the other hand, ICA extracts independent components based on the sign of the fourth moment in the data, called *kurtosis* (Hyvarinen et al., 2001; Mckeown et al., 2003), where the assumption is that the components are non-Gaussian and independent. In fMRI, ICA can be performed either in the spatial domain, called spatial ICA (SICA) (Keogh and Cordes, 2007), or in the temporal, called temporal ICA (TICA) (Keogh and Cordes, 2007). Often, PCA is performed first as a dimension reduction step. SICA applied to fMRI produces a set of spatial components with different patterns of time series (Chen and Yao, 2004; Mckeown et al., 1998). In the TICA, the independent components are time series instead of groups of voxels (Calhoun et al., 2001, 2003). Also, both SICA and TICA can be combined in one analysis (Stone et al., 2002).

Canonical correlation analysis *Canonical correlation analysis* (CCA) is a multivariate analysis seeking linear combination weights that maximize the cross-correlation between two

random vectors; the resulting linear combinations are called canonical variates. Actually, PCA is a special case of CCA where the two random vectors in the CCA are the same. Friman et al. (2002b) propose a CCA method that can separately detect spatial and temporal properties by considering their autocorrelatuions themselves. In their method, for the temporal analysis, the two random vectors in CCA are referred to two time courses with different lags; for the spatial analysis, the two random vectors in CCA are referred to a target voxel and its four proximal neighbors. Their method shows advantages over ICA in terms of performance, i.e., it is robust for the small sample size in temporal analysis; it gives a natural order of the components to the features of interest etc.

Problems in model-free approach

Although the model-free approach makes fewer assumptions about the data and only considers the intrinsic structure directly, it also has some disadvantages. Firstly, since it is model-free, we can not assess significance of results (Thirion, 2003). Secondly, ICA and CCA methods usually need some methods, e.g., PCA, to reduce the dimension of the data first. Prior knowledge about the data is necessary because heavily dimension reduction may discard important information (Thirion, 2003). Thirdly, in the PCA method, variance partitioning is non-specific (Sommer and Wichert, 2002), i.e., the separation of different components is by their uncorrelated properties, not by their independent properties. Hence the results from PCA are always not satisfactory (Thirion, 2003). For example, the method by Backfriender et al. (1996) needs a sequence of extraction steps because of the high noise level of the data and the sensitivity of the oblique rotation procedure, which may make a simple case overcomplicated. Fourthly, the ICA method assumes statistical independence structure in the space or time domain of the data, and is non-compatible with spatial or temporal smoothing. But the assumption is violated when prior statistical smoothing is used for the data, revealing a non-consistency in the data analysis procedure (Thirion, 2003; Calhoun et al., 2001; Stone et al., 2002). Also, because of the intrinsically spatial or temporal correlations

of the data, SICA can not recover the data structure if there is strong spatial dependence, and TICA fails if there is strong temporal dependence (Lazar, 2008). Hence prior knowledge of the data structure is very important for ICA (Lazar, 2008). Finally, ICA does not provide an ordering of the components and they have to be inspected visually (Friman et al, 2002b). Since sometimes it is difficult to distinguish different components by their time courses directly, cross-correlation with the behavioral experiment is used instead for comparing components' time courses (McKeown et al., 2003; Petersen et al., 2000). But these alternative methods change the *data-driven* method back to *hypothesis-driven* and have some limitations in the extreme cases (Baumgartner et al., 2000).

1.3 Data smoothing in fMRI

Because the amplitude of the signal change inside the brain is small, brain activations are contaminated with larger number of un-activated voxels and noises (Eddy et al., 1999). Since studies rely on the detection of the true signal in the presence of substantial noise, the more we know the true structure of the brain, the better the chance of detection. Prior to the preprocessing step, there are some ways to increase the amplitude of the fMRI signal and decrease the noise, such as improving the experimental design or increasing the scanner field strength (Huettel et al., 2004). During the preprocessing step, Gaussian filtering in time and space smooths the data in some ways. But because of the limitations of these techniques, the existing methods may not be enough to remove considering statistical noise in the data (Lazar, 2008). instead, of course we can always expect some (statistical) noise to remain.

To improve the signal changes and reduce noise, I summarize several main data smoothing techniques commonly used in fMRI data analysis as follows:

The first technique is to average the data across trials, commonly used for event-related designs. This method assumes the signal of interest is identical over repeated stimulus presentations, and also requires the noise to be random with sufficient numbers of repetitions (Huettel et al., 2004). Hence, the new data are combined across multiple trials over time. This

method can be considered as an alternative smoothing method, with the voxel value at each location of the image equally weighted over trials. Averaging analyses substantially simplify computations and may capture the main features of the data. Although the experimental trials can be repeated more times, there are limits due to time and environment constraints, and the learning adaption of the subjects (Wang et al., 2003). Some noise or artifacts may be enhanced by averaging the trials over time (Huettel et al., 2004). Also, this approach may lose some important characteristics for a particularly activated region (Bowman, 2007).

The second technique is doing the low-pass Gaussian filtering (Huettel et al., 2004), either in space or time. This method uses a Gaussian filter as the signal of interest to maximize the signal to noise ratio, where the filter width is used to control the smoothness. For example, in spatial filtering, typical filter widths for fMRI are two or three voxels. The disadvantages are: the filter width is difficult to control because of the different noise levels of the data; the blurring only depends on the data originally from that voxel; the degree of smoothness is the same everywhere. Also, the degree of smoothness is limited since the filter width should be smaller than the size of the active region in spatial filtering (Jezzard et al., 2001). If not, the simple blurring will mask the difference between the active region and neighboring nonactive regions in the brain and change the nature of the spatial correlation among voxels (Descombes et al., 1998; Jezzard et al., 2001; Lazar, 2008).

The third technique is taking into account spatial and temporal correlation into smoothing during the model-based approach. As mentioned before, filtering (smoothing) methods in time series, either in the time domain directly or in the frequency domain, are well developed and have been widely used in fMRI. By contrast, spatial smoothing methods are quite limited. Current spatial smoothing in fMRI is usually using a fixed deterministic model to fit the spatial surface in fMRI as "close" as possible. The parameter estimation of the model is considered in two ways: one is least squares (Katanoda et al., 2002); the other is simulation, e.g., MCMC (Descombes et al., 1998). Although these methods show better results than in Gaussian filtering, they still have some limitations: it is difficult to

formulate a general spatial model to well fit the spatial surface (Hartvig and Jensen, 2000); the spatial model building is local, which means only a few proximal neighbors are chosen for smoothing. But it is possible that voxels that are not physical neighbors would still show high correlations in space (Bowman, 2007); when the statistical model inference relies on the simulation only, it is too computationally intensive even for a trivial task (Descombes et al., 1998).

1.4 Geostatistics and fMRI

1.4.1 Applications of geostatistics

Geostatistics is a branch of applied statistics that focuses on providing quantitative descriptions of natural variables distributed in space or in time and space. The term "geostatistics" was coined in a geographical context to denote statistical techniques that emphasize location within areal distributions (Cressie, 1993). The development of geostatistics in the 1960s resulted from the need for a methodology to evaluate the recoverable reserves in mining deposits (Goovaerts, 1997). Nowadays geostatistics is popular in many fields of science and industry where there is a need for evaluating spatially or spatio-temporally correlated data. The application of geostatistical techniques includes mining, environmental sciences, remote sensing and ecology (Wackernagel, 2003; Rossi et al., 1992). To date there has not been much work using geostatistical methods to analyze the brain in spite of the similarities of data types and questions of interest. Some recent exceptions are Spence et al. (2007), who used the Gaussian variogram function to find neighbors of voxels of interest and Bowman (2007), who used an alternative variogram model to define the physical and functional spatial distance structure inside the brain. My dissertation topic is applying geostatistical methods more broadly in fMRI data analysis.

Spatial dependence is particularly important in an analysis of spatially varying regions and temporal variables, yet many traditional statistical measures tend to ignore it. Based on random sampling, the classical estimators are linear sums of data, all of which carry the

same weight. But under spatial models, we can vary the weights attributable to the data and make the estimation more precise and efficient (Webster and Oliver 2001). The analysis of the brain is quite similar to the analysis of other spatial and temporal phenomena in natural science. For example, miners want to estimate the amounts of metal in ore bodies and the thickness of coal seams, and petroleum engineers want to know the positions and volumes of reservoirs. Usually, those deposits are not completely random but rather exhibit some form of structure in an average sense, reflecting the fact that regions close in space tend to have similar values (Chiles and Delfiner, 1999). fMRI analysis includes investigation of activations in different regions and at different times. Those activations are spatially and temporally dependent. Geostatistics brings to fMRI tools for the interpretation of spatial patterns of regions, of the numerous temporal components with which they interact, and of the joint spatial dependence between regions and time (Rossi et al., 1992).

1.4.2 Three parts in geostatistics

The purpose of my research is to provide a comprehensive and easily understood exploration of geostatistical approaches to fMRI data analysis. Geostatistical analysis usually consists of three different parts: (i) *structural analysis*, also called covariance and variogram analysis; (ii) smoothing, filtering or prediction, also called *kriging*; (iii) *classification*. These three parts are interrelated to each other. Structural analysis is also called structural classification in geostatistics; modeled structure can be used for kriging; kriging can also be used for regionalized classification. My research explores these three parts in detail as they apply to fMRI.

1.4.3 My contributions

I mainly bring three geostatistical ideas to fMRI data analysis. **Firstly**, I apply structural classification from geostatistics in fMRI. Due to the high noise of the raw data in fMRI, many clustering techniques focus on classifying the modeled cross-correlation functions instead of

the raw time series. This type of clustering is an example of a model-based approach, which depends heavily on having prior knowledge of the reference function. In geostatistics, use of autocovariance or variogram to characterize the spatial or temporal structure of the data is called *structural analysis*. When the autocovariance function is used in classification, it is called structural classification in geostatistics. This classification method has been widely used in ecology and remote sensing (Atkinson and Lewis, 2000), but has not been applied in fMRI. I will show that using the empirical covariance structure instead of the raw data or the cross-correlation function gives more robust results in classification, even without using any additional smoothing method. **Secondly**, I consider the use of real structure of the data to do the smoothing, which is called *filtered kriging* in geostatistics. This method is more reliable compared with Gaussian filtering or other smoothing methods in current fMRI data analysis, because it considers combining spatial information into the smoothing procedure, which can reduce the noise in an intelligent way. When carrying out the research in a specific region of the brain, two main factors are to be considered. One is the relation between a given voxel and other voxels in this specific region of the brain; the other is the relation between the region and other regions of the brain. I call the first relation the *physical region connection*, and the second relation the *functional region connection*. Although current spatial smoothing considers the spatial structure in the data, model building is local, which means only the physical region connection is considered. But in practice, functional region connections are important in fMRI as well. For example, different areas involved in language processing may present some functional relations inside the brain (Bowman, 2007). Geostatistics gives more choices in model selection and model inference in fMRI. When the variogram shows a cyclical pattern with a "down-hole", its structure is called *hole effect* (Journel and Huijbregts, 1978). The use of hole effect structure in variogram modeling will consider both the physical and functional relations in the data. **Thirdly**, I control the degree of smoothness in the spatial smoothing by introducing a *smoothing ratio*. When using filtered kriging to remove the measurement errors at the observed site for different time points, I creatively borrow the

idea of the smoothing ratio from spline method to control the degree of smoothing in filtered kriging. The smoothing ratio in the spline method does not have physical meaning and just provides a computational convenience. But when the smoothing ratio is used in filtered kriging, it has a clear statistical meaning and gives a good interpretation. This method considers both the spatial structure and the temporal properties of the fMRI data.

1.4.4 Organization of the dissertation

The remainder of the dissertation is organized as follows: In Chapter 2, I introduce basic concepts and definitions in geostatistics. Chapter 3 and Chapter 4 introduce geostatistical methods and sparse principal component analysis (Zou et al., 2006) in clustering and demonstrate their uses in fMRI time series. Chapter 5 uses hole effect variogram structure to model the spatial structure of the fMRI data, which considers both the physical and functional region connections inside the brain. Chapter 6 considers filtered kriging as a smoothing method and uses the smoothing ratio to control the degree of the smoothness. Chapter 7 summarizes the final conclusions and some directions for future research.

Chapter 2

Basic concepts and definitions in geostatistics

Before moving to an in-depth discussion of geostatistical analysis in fMRI, it is helpful to introduce some basic concepts and definitions in geostatistics. This chapter includes two parts: one deals with structural analysis, the other with kriging.

2.1 Structural analysis

Using the variogram or the covariance model to characterize the spatio-temporal structure of the interest variable is called structural analysis in geostatistics (Wackernagel, 2003; C-Olmo et al., 2000). The analysis treats a set of spatial data as a sample from the realization of a random process, and stresses the structural features.

2.1.1 Definitions

The variogram is a traditional geostatistical tool that provides quantification of the degree of directional spatial properties in the random variables (Rossi et al., 1992). The variogram gives a nice interpretation of the variance in a second-order stationary process (Schabenberger and Gotaway, 2005; Schabenberger and Pierce, 2002), and it only compares the average square difference between locations and need not to know the constant population mean. Based on these reasons, it is common to work with the variogram rather than the covariance function in geostatistical applications.

For the general spatial model $\{Z(s) : s \in D\}$, $s = (x, y)$ denotes the coordinates of the sample site (x, y); $Z(s)$ denotes the random variable of interest at spatial position $s = (x, y)$; D denotes the set of the region of interest. Considering a simple geostatistical model $Z(s) =$

$\mu + \epsilon(s)$, where μ is a constant population mean, $\epsilon(s)$ is the zero-mean random function at s, if we have

$$\begin{aligned} &E[\epsilon(s)] = 0,\ E[Z(s)] = \mu,\ s \in D; \\ &Var[\epsilon(s)] = \sigma^2,\ Var[Z(s)] = \sigma^2,\ s \in D; \\ &C(\mathbf{h}) = Cov[Z(s+\mathbf{h}), Z(s)],\ s+\mathbf{h}, s \in D, \end{aligned} \tag{2.1}$$

then the random function is called *second order stationary*. The separation $\mathbf{h}$ in the above covariance function is a vector with both distance and direction, usually called *lag*. The *variogram* is defined as an alternative measure of spatial dependence

$$\gamma(\mathbf{h}) = \tfrac{1}{2} Var[Z(s+\mathbf{h}) - Z(s)]. \tag{2.2}$$

A random field is *intrinsically stationary* if

$$\begin{aligned} &E[Z(s+\mathbf{h}) - Z(s)] = 0,\ s+\mathbf{h}, s \in D; \\ &\gamma(\mathbf{h}) = \tfrac{1}{2} Var[Z(s+\mathbf{h}) - Z(s)]. \end{aligned} \tag{2.3}$$

These two assumptions constitute Matheron's intrinsic hypothesis (Webster and Oliver, 2001; Wackernagel, 2003). Intrinsic stationarity allows for the possibility that $\sigma^2 = \infty$. Hence, all second order stationary random fields are intrinsic, but an intrinsic random field may not be second order stationary.

For second order stationary random fields, we have the relation $\gamma(\mathbf{h}) = C(0) - C(\mathbf{h})$, and the variogram can reach a limiting maximum value, i.e., $\gamma(\mathbf{h_0}) = C(0)$, which is called the *sill*. Here there is a lag $\mathbf{h_0}$ called the *range*, beyond which $Z(x)$ and $Z(x + \mathbf{h_0})$ are uncorrelated; $\gamma(\mathbf{h_0})$ reaches the variance $C(0) = \sigma^2$ of the process. Range is also known as *correlation range*, since it defines the average distance within which the random function remains correlated spatially. So, $C(\mathbf{h_0}) = 0$.

For $\gamma(\mathbf{h}) = C(0) - C(\mathbf{h})$, it is clear that in theory $\gamma(0) = 0$. But sometimes in practice, $\gamma(\mathbf{h}) \to \sigma_\epsilon^2$ as $\mathbf{h} \to 0$, where σ_ϵ^2 is called the *nugget effect* by Matheron (1962). Nugget effect is usually considered as a micro-scale or measurement error causing a discontinuity at the origin (Cressie, 1993).

2.1.2 Properties

Trend removal

The most generic geostatistical model is a decomposition of a random response variable into a mathematical structure describing variation and covariation among the responses, i.e., $Z(s) = \mu + \epsilon(s)$. We assume the mean is constant but this assumption is not always realistic. Hence we first perform trend surface analysis to remove any trend surface and create stationary data fields (Glover et al.,2006; Schabenberger and Gotway, 2005). Trend removal may introduce a bias, but this bias can be ignored when the sample size is large and the variogram is used instead of covariance in further analysis (Cressie, 1993), because the square of variogram bias goes to zero faster than does the variance and the variogram bias has smaller order than the corresponding covariance bias. Detail proofs are in Cressie (1993). Even if the bias can not be ignored, it is still considered less significant than the errors introduced by leaving the trend alone (Gringarten and Deutsch, 2001). According to Gringarten and Deutsch (2001), the reason is that residuals after trend removal are easy to consider stationarity. If a significant trend is presented in the data, the mean value is not independent of location in the variogram estimation.

For example, we can set up the model as

$$Y(s) = f(s) + \mu + \epsilon(s) = f(s) + Z(s),$$

where $f(s)$ is a trend function. In trend removal analysis, the order refers to the highest power of the variables, the rank refers to the dimensionality, the number of the variables. For two dimensions and a second-order trend in the model, we fit by a polynomial function:

$$f(x, y) = m_0 + m_1 x + m_2 y + m_3 x^2 + m_4 xy + m_5 y^2,$$

where $m_0, ..., m_5$ are unknown parameters estimated by least squares. After removing the trend $f(s)$, we render the residuals stationary and use $Z(s)$ to compute the variograms for further analysis, and add back the trend $f(s)$ at the end of the analysis. After trend removal,

we consider $Z(s)$ to be second order stationary, and assume that the sills and the nuggets are equal regardless of direction (Glover et al., 2006), which simplifies the subsequent work.

Unfortunately, we do not typically know the true order of the trend function in advance. Hence we have two choices: (i) assume that the model only has a first or second-order trend, and the error term has the high order structure; (ii) consider the model function to be of high order and the error term to be simple trend or even random structure. These are two different approaches to model fitting: the first one is called the *random function approach*, or kriging; the second one is called the *smooth function approach*, or the spline method (Watson, 1984). Actually, the final results of these two approaches are consistent, as I will discuss in detail later.

Positive definiteness

In the second order stationary case, if we consider a linear combination form $\sum_{i=1}^{n} \lambda_i Z(s_i)$, its variance

$$Var[\sum_{i=1}^{n} \lambda_i Z(s_i)] = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) \tag{2.4}$$

must be positive or zero, then the covariance matrix $C(s_i - s_j)$ should be non-negative whatever the λ_i chosen. This is called *positive definiteness*. The family of covariance in $\mathbf{R}^n$ satisfies the following properties (Chiles and Delfiner, 1999; Schabenberger and Gotway, 2005): (i) if $C(h)$ is a covariance in $\mathbf{R}^n$, then it is also valid in $\mathbf{R}^m$ for $m \leq n$; (ii) if $C_1(h)$ and $C_2(h)$ are two covariances in $\mathbf{R}^n$, then $aC_1(h) + bC_2(h)$ is a covariance in $\mathbf{R}^n$ for $a > 0, b > 0$; (iii) if $C_1(h)$ and $C_2(h)$ are two covariances in $\mathbf{R}^n$, then $C_1(h) \times C_2(h)$ is a covariance in $\mathbf{R}^n$. By these three properties, we can define different kinds of covariance functions.

If the variable is intrinsic only and the covariance does not exist, then

$$Var[\sum_{i=1}^{n} \lambda_i Z(s_i)] = C(0) \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \lambda_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma(s_i - s_j). \tag{2.5}$$

Since $C(0)$ is unknown now, we may eliminate it by making the weights sum to 0 (Webster and Oliver, 2001); then

$$Var[\sum_{i=1}^{n} \lambda_i Z(s_i)] = -\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(s_i - s_j). \tag{2.6}$$

Hence, the variogram must be *conditionally negative definite* with the condition that the weights sum to 0.

Anisotropy

For a specific spatial structure, if the variogram function is direction dependent, we call this *anisotropy*. If both the sill and the range of the variogram vary with direction, it is called *zonal anisotropy*. If only the range varies and the sill remains constant, it is called *geometric anisotropy*. As mentioned before, we have removed the trend and assume the residuals are second order stationary. Hence after trend removal we can only consider geometric anisotropy (Glover et al., 2006).

Geometric anisotropy can be corrected by a linear transformation of the coordinate system (Chiles and Delfiner, 1999). In $\mathbf{R}^2$, the axes rotation can be defined by the variogram

$$\gamma(\mathbf{h}) = \gamma_0(|\mathbf{Ah}|),$$

where $\gamma_0(\cdot)$ is an isotropic model, $\mathbf{A}$ is the transformation matrix from the initial space to the isotropic space, $\mathbf{h}$ is the initial distance vector.

Let θ and $\theta + \frac{\pi}{2}$ be the new coordinate system with axes parallel to the anisotropic directions with the maximum range and the minimum range; a_1 and a_2 be the ranges in the directions θ and $\theta + \frac{\pi}{2}$; h_1 and h_2 be the distances in the directions θ and $\theta + \frac{\pi}{2}$, then

$$|\mathbf{Ah}| = \begin{bmatrix} \frac{1}{a_1} & 0 \\ 0 & \frac{1}{a_2} \end{bmatrix} \begin{bmatrix} cos(\theta) & sin(\theta) \\ -sin(\theta) & cos(\theta) \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}. \tag{2.7}$$

The simplest case is when the anisotropy axes coincide with the coordinate axes.

2.1.3 Calculations of empirical covariance and variogram

Consider a spatial model $\{Z(s) : s \in D\}$, where D denotes the set of sample points $s_1, ..., s_n$ lying on a regular lattice. The empirical variogram $\hat{\gamma}(\mathbf{h})$ and the empirical autocovariance $\hat{C}(\mathbf{h})$ can be calculated as follows (Isaaks and Srivastava, 1989):

$$\begin{aligned} 2\hat{\gamma}(\mathbf{h}) &= \tfrac{1}{N(\mathbf{h})} \textstyle\sum_{(s_i,s_j)|\mathbf{h}_{s_i,s_j}=\mathbf{h}} [Z(s_i) - Z(s_j)]^2, \\ \hat{C}(\mathbf{h}) &= \tfrac{1}{N(\mathbf{h})} \textstyle\sum_{(s_i,s_j)|\mathbf{h}_{s_i,s_j}=\mathbf{h}} Z(s_i) \cdot Z(s_j) - \hat{\mu}_{-\mathbf{h}} \cdot \hat{\mu}_{+\mathbf{h}}, \end{aligned} \tag{2.8}$$

where $\hat{\mu}_{-\mathbf{h}}$ is the mean of all data values whose locations are $-\mathbf{h}$ away from some other data location:

$$\hat{\mu}_{-\mathbf{h}} = \tfrac{1}{N(\mathbf{h})} \textstyle\sum_{(s_i)|\mathbf{h}_{s_i,s_j}=\mathbf{h}} Z(s_i);$$

$\hat{\mu}_{+\mathbf{h}}$ is the mean of all data values whose locations are $+\mathbf{h}$ away from some other data location:

$$\hat{\mu}_{+\mathbf{h}} = \tfrac{1}{N(\mathbf{h})} \textstyle\sum_{(s_j)|\mathbf{h}_{s_i,s_j}=\mathbf{h}} Z(s_j).$$

The autocorrelation $\hat{\rho}(\mathbf{h})$ is the autocovariance standardized by the standard deviations:

$$\hat{\rho}(\mathbf{h}) = \tfrac{\hat{C}(\mathbf{h})}{\hat{\sigma}_{-\mathbf{h}} \cdot \hat{\sigma}_{+\mathbf{h}}}, \tag{2.9}$$

where $\hat{\sigma}^2_{-\mathbf{h}}$ is the variance of all data values whose locations are $-\mathbf{h}$ away from some other data location:

$$\hat{\sigma}^2_{-\mathbf{h}} = \tfrac{1}{N(\mathbf{h})} \textstyle\sum_{(s_i)|\mathbf{h}_{s_i,s_j}=\mathbf{h}} Z(s_i)^2 - \hat{\mu}^2_{-\mathbf{h}};$$

$\hat{\sigma}^2_{+\mathbf{h}}$ is the variance of all data values whose locations are $+\mathbf{h}$ away from some other data location:

$$\hat{\sigma}^2_{+\mathbf{h}} = \tfrac{1}{N(\mathbf{h})} \textstyle\sum_{(s_j)|\mathbf{h}_{s_i,s_j}=\mathbf{h}} Z(s_j)^2 - \hat{\mu}^2_{+\mathbf{h}}.$$

In the above calculations, the summation is over only the $N(\mathbf{h})$ pairs of observations which are separated by lag $\mathbf{h}$. Clearly, $N(\mathbf{h})$ decreases as $\mathbf{h}$ increases. Equally clearly, the larger $N(\mathbf{h})$, the greater the statistical reliability (Ross et al., 1992). Since a relatively large $N(\mathbf{h})$ is associated with a relatively small standard error (Morris, 1991), Journel and Huijbregts (1978) suggest the number of data pairs in each class should be at least 30, and this rule of thumb has been widely accepted.

2.1.4 Modeling the variogram

There are three reasons for modeling the empirical variogram (Webster and Oliver, 2001; Schabenberger and Gotaway, 2005): First, the empirical variogram only gives estimates at a finite set of lags, but what we need are estimates at arbitrary lags; second, the empirical variogram may not be valid since the corresponding covariance may not satisfy the property of positive definiteness; third, the empirical variogram may show meaningless fluctuations due to measurement errors. The reason is that the calculated variogram of a particular lag is only the estimated mean of that lag. Measurement error increases as the lag distance increases. Variogram modeling smooths the empirical variogram.

Methods for variogram fitting include *maximum likelihood* and *least squares*. Maximum likelihood estimation requires the assumption of Gaussian distribution of the data and the fit has to be based on all data; by contrast, least squares does not have the distribution assumption and one can restrict the lag distance for the data. I choose least squares here because of its flexibility property (Schabenberger and Gotaway, 2005; Schabenberger and Pierce, 2002). Since fitting models in this way is a form of non-linear regression, the Levenberg-Marquardt method (Marquardt, 1963) is used for model fitting (Webster and Oliver, 2001).

Because we do not know how much of the observed fluctuation in the empirical variogram is due to error and how much is structural, we follow three rules in the fitting process (Webster and Oliver, 2001; Ma and Jones, 2001): (i) ignore the point to point fluctuation and concentrate on general trends; (ii) estimate the variogram to be accurate at short lags, with less accuracy at longer lags; (iii) match the cyclic pattern of the variogram at least to the first peak or trough.

2.1.5 Examples of variogram models

The standard spherical model is defined as

$$\gamma_{sph}(h) = \begin{cases} \frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3} & \text{if } h \leq a \\ 1 & \text{if } h > a, \end{cases}$$

where a is the range. The standard nugget effect model is

$$\gamma_{nug}(h) = \begin{cases} 0 & \text{if } h = 0 \\ 1 & \text{if } h > 0. \end{cases}$$

Since the sum of two basic covariance models is still a valid covariance model, the spherical model with a nugget effect can be constructed as

$$\gamma(h) = \sigma_\epsilon^2 \gamma_{nug}(h) + (\sigma^2 - \sigma_\epsilon^2)\gamma_{sph}(h),$$

where σ_ϵ^2 is the nugget effect, and σ^2 is the sill of the new model. Figure 2.1 gives the graphs of the variogram model $8\gamma_{nug}(h) + 12\gamma_{sph}(h)$ in two different forms, where the nugget effect is 8, the sill is 20, and the range is 10. Graph (1) is the model in variogram form, which is a monotonically increasing function that reaches sill 20 after range 10. Graph (2) is the model in covariance form, which is a monotonically decreasing function that becomes 0 after range 10. The model in variogram form is negativly related with the model in covariance form, i.e., $\gamma(h) = 20 - C(h)$. The nugget effect is a discontinuous point at the origin in the variogram form, or a discontinuous point at 20 in the covariance form. The range 10 defines the maximum distance within which the points are correlated.

2.2 Kriging

2.2.1 General characteristics of kriging

In geostatistics interpolation or prediction is called *kriging*. For predicting the value of the random variable $Z(s_0)$ at an unsampled site s_0, from the data $Z(s_1), ..., Z(s_n)$ at sampled sites $s_1, ..., s_n$, kriging is given as the weighted sum of the values of the neighbors. The weights are calculated by minimizing the error variance of a given or assumed model of covariance for the data with regard to the spatial distribution of the observed data points. If the predicted variable is at one of the sampled sites, kriging gives the original value at this site, which is called interpolation.

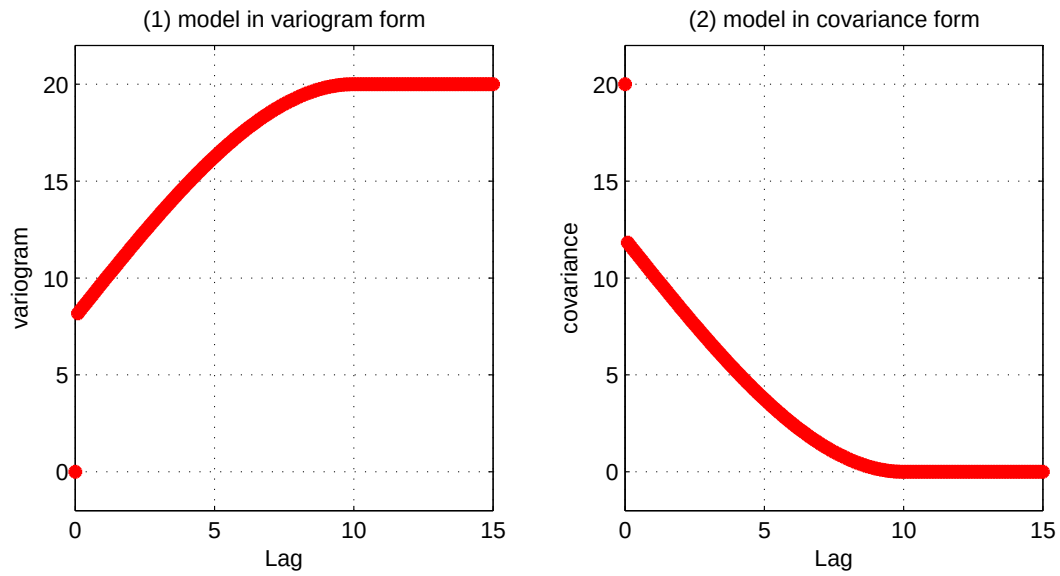

Figure 2.1: An example of spherical model with nugget effect 8, range 10 and sill 20. (1) is the model in variogram form, which is a monotonically increasing function and reaches sill 20 after range 10. (2) is the model in covariance form, which is a monotonically decreasing function and becomes 0 after range 10. The model in variogram form is negativly related with the model in covariance form, i.e., $\gamma(h) = 20 - C(h)$. The nugget effect is a discontinuous point at the origin in the variogram form, or a discontinuous point at 20 in the covariance form. The range defines the maximum distance within which the points are correlated.

In order to simplify the problem, it is common to restrict the estimation to linear predictors (Stein, 1999). An unknown value $Z(s_0)$ is therefore estimated by a weighted linear combination of n known values $Z(s_1), ..., Z(s_n)$. If this linear predictor minimizes the mean squared error among all linear predictors, it is called *best linear predictor* (Stein, 1999). If this predictor also has an unbiased constraint, it is called *best linear unbiased predictor*. Hence, kriging is a method to determine the best linear unbiased estimate of the already known points (Webster and Oliver, 2001).

Defining $C(0) = Var[Z(s_i)] = \sigma^2$ and $C(s_i, s_j) = Cov[Z(s_i), Z(s_j)]$, $i, j = 1, ..., n$, if $\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i)$, the random function $Z(s_0) - \sum_{i=1}^{n} w_i Z(s_i)$ is considered as an *intrinsic random function of order n* (n-IRF) by Dubrule (1983), with the properties:

$$\begin{aligned} &E[Z(s_0) - \textstyle\sum_{i=1}^{n} w_i Z(s_i)] = 0, \\ &Var[Z(s_0) - \textstyle\sum_{i=1}^{n} w_i Z(s_i)] = \sigma^2 - 2\sum_{i=1}^{n} w_i C(s_i, s_0) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C(s_i, s_j). \end{aligned} \tag{2.10}$$

These two functions construct the basic formulas in kriging. All krigings are special cases of the n-IRF.

2.2.2 Different krigings

Following the ideas and definitions in Cressie (1993) and Journel (1989), when $\hat{Z}(s_0)$ at an unsampled site s_0 is estimated from $Z(s_1), ..., Z(s_n)$ at sampled sites $s_1, ..., s_n$, $Z(s_i)$ is decomposed as a geostatistical model

$$Z(s_i) = \mu(s_i) + e(s_i),\ i = 0, ..., n,$$

where $e(s_i)$ is the zero-mean second-order or intrinsically stationary random function at spatial position s_i. The kriging is divided into three different classes depending on the prior knowledge of $\mu(s_i)$. When $\mu(s_i)$ is an already known constant, we have *simple kriging*; when $\mu(s_i)$ is an unknown constant, we have *ordinary kriging*; when $\mu(s_i)$ is presented as a trend, it is called *universal kriging*. The following are the formulas of the three types.

Simple kriging

$$Z(s_i) = \mu(s_i) + e(s_i), i = 0, 1, ..., n,$$

where $\mu(s_0), \mu(s_1), ..., \mu(s_n)$ are known constants.

The simple kriging is $\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i) + [\mu(s_0) - \sum_{i=1}^{n} w_i \mu(s_i)]$. When $E[Z(s_0) - \hat{Z}(s_0)]^2$ is minimized,

$$C(s_i, s_0) = \sum_{j=1}^{n} w_j C(s_i, s_j). \tag{2.11}$$

The kriging variance is

$$\sigma_{SK}^2 = E[Z(s_0) - \hat{Z}(s_0)]^2 = \sigma^2 - \sum_{i=1}^{n} w_i C(s_i, s_0). \tag{2.12}$$

Since $\gamma(s_i, s_j) = \sigma^2 - C(s_i, s_j)$, the above formulas also can be expressed as a form with variogram,

$$\gamma(s_i, s_0) = \sum_{j=1}^{n} w_j \gamma(s_i, s_j) + \sigma^2[1 - \sum_{j=1}^{n} w_j], \tag{2.13}$$

$$\sigma_{SK}^2 = \sum_{i=1}^{n} w_i \gamma(s_i, s_0) + \sigma^2[1 - \sum_{j=1}^{n} w_j]. \tag{2.14}$$

Ordinary kriging

$$Z(s_i) = \mu(s_i) + e(s_i), i = 0, 1, ..., n,$$

where $\mu(s_i) = \mu$ is an unknown constant.

The ordinary kriging is $\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i)$, subject to the linear constraint $\sum_{i=1}^{n} w_i = 1$. Let m be the Lagrange parameter, $E[Z(s_0) - \hat{Z}(s_0)]^2 - 2m[\sum_{i=1}^{n} \lambda_i - 1]$ is minimized by setting its partial derivatives with respect to w_i and m equal to zero. Hence,

$$\begin{cases} \sum_{i=1}^{n} w_i = 1, \\ C(s_i, s_0) = \sum_{j=1}^{n} w_j C(s_i, s_j) + m. \end{cases} \tag{2.15}$$

$$\sigma_{OK}^2 = E[Z(s_0) - \hat{Z}(s_0)]^2 = \sigma^2 - \sum_{i=1}^{n} w_i C(s_i, s_0) + m. \tag{2.16}$$

It also can be expressed as a form with variogram

$$\begin{cases} \sum_{i=1}^{n} w_i = 1, \\ \gamma(s_i, s_0) = \sum_{j=1}^{n} w_j \gamma(s_i, s_j) - m. \end{cases} \tag{2.17}$$

$$\sigma_{OK}^2 = \sum_{i=1}^{n} w_i \gamma(s_i, s_0) + m. \tag{2.18}$$

Universal kriging

$$Z(s_i) = \mu(s_i) + e(s_i), i = 0, 1, ..., n.$$

$\mu(s_i) = \sum_{l=0}^{L} a_l f_l(s_i)$ is an unknown trend, where $f_l(s_i)$ are known functions and $f_0(s_i) = 1$; a_l are unknown parameters.

The universal kriging is $\hat{Z}(s_0) = \sum_{i=1}^{n} w_i [Z(s_i)]$, subject to the constraint $\sum_{i=1}^{n} w_i f_l(s_i) = f_l(s_0)$, for $l = 0, 1, ..., L$. Let $m_l, l = 0, 1, ..., L$ be the Lagrange parameters, $E[Z(s_0) - \hat{Z}(s_0)]^2$ is minimized with the constraint,

$$\begin{cases} \sum_{i=1}^{n} w_i f_l(s_i) = f_l(s_0) & \text{for } l = 0, 1, ..., L; \\ C(s_i, s_0) = \sum_{i=1}^{n} w_i C(s_i, s_j) + \sum_{l=0}^{L} m_l f_l(s_i) & \text{for } i = 1, ..., n. \end{cases} \tag{2.19}$$

$$\sigma_{UK}^2 = E[Z(s_0) - \hat{Z}(s_0)]^2 = \sigma^2 - \sum_{i=1}^{n} w_i C(s_i, s_0) + \sum_{l=0}^{L} m_l f_l(s_0). \tag{2.20}$$

It also can be expressed as a form with variogram

$$\begin{cases} \sum_{i=1}^{n} w_i f_l(s_i) = f_l(s_0) & \text{for } l = 0, 1, ..., L; \\ \gamma(s_i, s_0) = \sum_{j=1}^{n} w_j \gamma(s_i, s_j) - \sum_{l=0}^{L} m_l f_l(s_i) & \text{for } i = 1, ..., n. \end{cases} \tag{2.21}$$

$$\sigma_{UK}^2 = \sum_{i=1}^{n} w_i \gamma(s_i, s_0) + \sum_{l=0}^{L} m_l f_l(s_0). \tag{2.22}$$

2.2.3 Selection of neighborhoods

The selection of neighbors to be included in the estimation is important since this will affect the accuracy of prediction. In theory, for each prediction, the minimum mean square error is achieved when all sampled points are included. However, it is not necessary to use all data values in practice and a global neighborhood may result in a kriging matrix that is too large to be inverted numerically. Therefore, only a subset of all data in a spatial neighborhood around the predicted point is used. We limit our estimates to just the data within some predefined radius of our point of estimation, called the search radius. There are no strict rules for defining the neighborhood, rather it is usually chosen based on past experience. For example, Schabenberger and Gotway (2005) suggest to divide the neighborhood around the target point into quadrants or octants and use the nearest 2-3 points from each quadrant or octant. They think kriging with more than 25 points is often unnecessary. Restricting the number of neighbors for prediction has two advantages. One is saving on computing time; the other is that we assume stationarity only within this search radius, which is useful if the data in general are non-stationary (Shibli, 2003).

2.2.4 Discussions of different krigings

Actually, simple kriging is a special case of ordinary kriging, and ordinary kriging is a special case of universal kriging. The kriging variances are $\sigma^2_{SK} \leq \sigma^2_{OK} \leq \sigma^2_{UK}$ (Webster and Oliver, 2001). The ordinary kriging variance is the sum of the simple kriging variance plus the variance due to the estimation of the unknown mean (Wackernagel, 2003); the universal kriging variance is the sum of the ordinary kriging variance plus the variance due to the estimation of the unknown trend.

If the variogram has been assumed to be known, it is better to use universal kriging since it considers the trend automatically. But in practice, we have to estimate the variogram from the data before kriging. To guarantee stationarity over the domain of the study, the variogram only can be estimated under the already known trend. However, we usually do

not know the trend in advance. Hence, in the geostatistical analysis, we may estimate the trend first, then analyze the residuals of the trend by ordinary kriging and add back the trend after finishing the analysis (Cressie, 1993).

2.2.5 An example of ordinary kriging

Here is an example of ordinary kriging under the spherical variogram model. Assume there are four known points $Z(s_1) = 5, Z(s_2) = 10, Z(s_3) = 15, Z(s_4) = 20$ at locations $s_1 = (1, 5), s_2 = (5, 1), s_3 = (9, 5), s_4 = (5, 9)$ (Figure 2.2, graph(1)). The predicted value at a new location s_0 is

$$\hat{Z}(s_0) = w_1 Z(s_1) + w_2 Z(s_2) + w_3 Z(s_3) + w_4 Z(s_4)$$

with the constraint $w_1 + w_2 + w_3 + w_4 = 1$. If a spherical model with sill 1 and range 12 is assumed for this process, then the kriging results for the eight predicted points at different locations are shown in Table 2.1 and Figure 2.2 (right panel).

Note if the predicted location is at the observed location, e.g., $s_0^1 = s_1$ or $s_0^5 = s_3$, the kriging exactly interpolates the observed point and the kriging variance is 0. The kriging weight at the predicted location is 1, and the weights at the other three locations are 0. If the predicted location is in the middle of the four observed points e.g., $s_0^3 = (5, 5)$, the four observed points contribute the same weights and the predicted value is just the average of the four points. If the predicted location is in the middle of any two observed points, e.g., $s_0^7 = (3, 7)$ or $s_0^9 = (7, 7)$, the two proximal points contribute the same strong weights and the two distal points contribute weak weights.

Because the variogram defines the relationship between the variability of the data and the distance of locations, the study of the variogram range is very informative. It determines different contributions of the observed points to the predicted value. Since the spherical variogram is monotonically increasing, for predicting an unobserved point, the proximal points will contribute more weight and the distal points will contribute less weight. If we change the range in the variogram model, the predicted values and kriging variance are

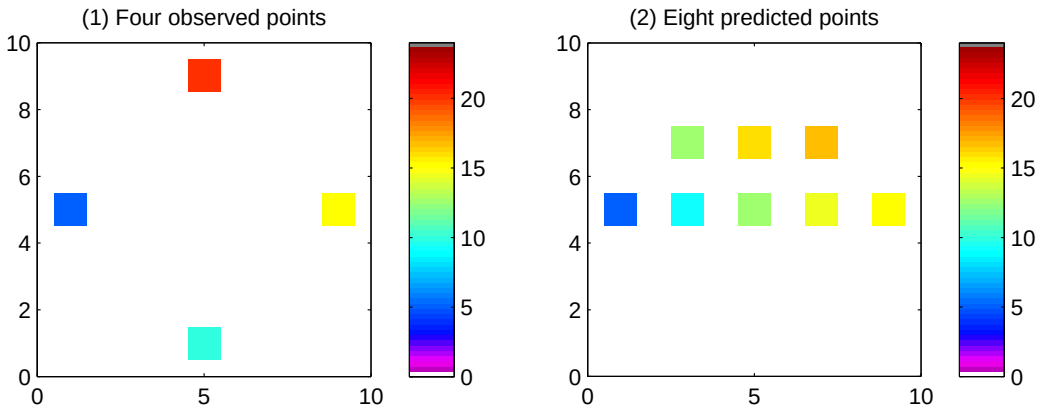

Figure 2.2: An example of ordinary kriging. (1) is the four observed points used in kriging. (2) is the eight predicted points by kriging.

changed at the unobserved locations. But just changing the sill does not change the predicted values and only the kriging variance is changed (Schabenberger and Pierce, 2002). I will discuss the property in later chapters.

s_0	$\hat{Z}(s_0)$	σ^2_{OK}	w_1	w_2	w_3	w_4
(1,5)	5.0000	0	1.0000	0	0	0
(3,5)	9.2023	0.3375	0.5798	0.1784	0.0635	0.1784
(5,5)	12.5000	0.4226	0.2500	0.2500	0.2500	0.2500
(7,5)	14.3654	0.3375	0.0635	0.1784	0.5798	0.1784
(9,5)	15.0000	0	0	0	1.0000	0
(3,7)	12.5000	0.3630	0.4674	0.0326	0.0326	0.4674
(5,7)	15.7977	0.3375	0.1784	0.0635	0.1784	0.5798
(7,7)	16.8477	0.3630	0.0326	0.0326	0.4674	0.4674

Table 2.1: Kriging prediction under the spherical variogram model with sill 1 and range 12. s_0 is the predicted location; $\hat{Z}(s_0)$ is the predicted value in kriging; σ^2_{OK} is the kriging variance; w_1, w_2, w_3, w_4 are the weights at observed locations $s_1 = (1, 5), s_2 = (5, 1), s_3 = (9, 5), s_4 = (5, 9)$ in kriging.

Chapter 3

Geostatistical analysis in clustering fMRI time series

3.1 Introduction

In recent years, characterizing brain activation by clustering the fMRI time series has gained popularity (for example, Goutte, et al., 1999). Since regions that react to the experimental task may be at several different physical locations, the goal in clustering time series is to partition the brain into clusters, where voxels within each cluster have similar temporal patterns. The underlying assumption is that voxels with similar temporal characteristics belong to the same functional regions of the brain (Baudelet and Gallez, 2003; Yeo and Ou, 2004).

Due to the high noise level in fMRI data, the results of clustering on the raw time series are often unsatisfactory and do not necessarily group data according to the similarity of their pattern of response to the stimulus (Goutte et al., 1999). Hence, many clustering techniques have focused instead on cross-correlation functions of the fMRI time series. Bandettini et al. (1993) consider the correlation between the data time series and a reference function from a "seed" voxel to characterize the temporal response of the brain to the paradigm. Goutte et al. (1999, 2001) suggest clustering voxels on the basis of the cross-correlation function between the fMRI activation and the experimental protocol signal. Yeo and Ou (2004) devise a "signal to protocol metric" by combining the cross-correlation of two fMRI signals and the Euclidean distance between voxels to do the clustering. Friman et al. (2002a) approach the problem by viewing the correlation between two time series as an objective function to be maximized with respect to some parameters. Clustering on the cross-correlation function instead of the raw time series provides increased robustness and yields improved performance. However,

methods based on cross-correlation depend heavily on the choice of reference function. Note that different types of reference functions are possible, and they result in correspondingly different analyses (Gibbons et al., 2004).

In this chapter, a special classification method for fMRI data is introduced. I first summarize the traditional classification methods and formalize them as three steps: feature extraction, choice of classification metric and choice of classification algorithm. In the feature extraction step, I propose the use of the autocorrelation function as the main feature and demonstrate its efficacy. In geostatistics, use of autocovariance (autocorrelation) or variogram to characterize the spatial or temporal structure of the data is called *structural analysis*. When the autocovariance (autocorrelation) function is used in classification, this is called *structural classification* in geostatistics. This classification method has been widely used in geostatistical areas, e.g., ecology and remote sensing (Atkinson and Lewis, 2000; Wackernagel, 2003; C-Olmo et al., 2000), but, to date, not in fMRI. The analysis treats a set of data as a sample from the realization of a random process, and stresses the structural features. The correlation structures for different functional regions of the brain are different, and can be used to inform classification. For example, in functional regions that react to the task, the time series exhibit periodic patterns, and the autocorrelation function may be expected to fluctuate periodically as well (Webster and Oliver, 2001). For irrelevant regions the autocorrelation is expected to be flat.

3.2 Methods

3.2.1 Clustering analysis

Three steps in classification

Classification is a procedure in which individual objects are placed into different classes based on their characteristics. When objects are assigned to one of several pre-defined classes, this is known as "supervised learning", "pattern recognition", or "discriminant analysis", depending

on the field. When neither the number of the classes nor the classes themselves are known in advance, this is known as "unsupervised learning" or "cluster analysis" (Gordon, 1999). In any classification problem, a good solution depends on three steps: feature extraction, the choice of the clustering metric, and the clustering algorithm (Friman et al., 2002a; Stanberry et al., 2003; Yeo and Ou, 2004).

Feature extraction step Feature extraction is a special form of data reduction that decreases the resources required to describe a large set of data accurately. There are many ways to extract the features from the object of interest. For example, principal component analysis and canonical correlation analysis are very commonly used for multivariate data. Since our primary interest is in time series data, we use the autocovariance (autocorrelation) structure as the main feature for clustering.

In geostatistical applications, it is common to work with the variogram, rather than the autocovariance (autocorrelation) function (Schabenberger and Gotway, 2005). The advantage of the former is that it only compares the average square difference between locations; this is quite useful in geology since one usually doesn't know the population mean and hence it is assumed to be constant. But it may be a disadvantage in other settings since the variogram does not account explicitly for mean and variance changes. Specifically, in fMRI data, the mean or variance may change over time. Therefore, I choose to work with the autocovariance (autocorrelation) instead of the variogram because of the former's robustness in the case that local means and variances vary (Rossi et al,. 1992, Radeloff et al., 2000).

Definitions Consider a time series $\{Z(t) : t \in D\}$, where t denotes time; $Z(t)$ denotes the random variable of interest at time t; D denotes the set of the time points of interest $t_1, t_2, ..., t_N$. We define the lag h as the separation between two different time points. The time series is second order stationary if:

$$E[Z(t_i)] = E[Z(t_i + h)] = \mu,$$

and

$$Cov[Z(t_i + h), Z(t_i)] = C(h),$$

for any $t_i + h, t_i \in D$, where μ is a constant and $C(\cdot)$ is a positive semi-definite function depending only on h. The autocorrelation function is defined as $\rho(h) = C(h)/C(0)$.

As mentioned before, different functional regions in the brain have different temporal behavior during the experiment, and their correlation structures are potentially also different. As a result, we can classify voxels by comparing their correlation structures. The reason to use correlation instead of covariance in clustering is that without the standardization, clustering may extract regions with large variances rather than similar temporal patterns (Goutte et al., 1999).

There are two main approaches for estimating the correlation function from the data (Atkinson and Lewis, 2000). The first approach is based on the nonparametric or empirical covariance estimator described below; the second approach is parametric, i.e. fitting a pre-existing covariance model to the empirical autocovariance and obtaining the model parameters. Usually the modeling approach is considered to be more efficient than the empirical one, as the covariance feature is usually described by a small number of parameters (usually three or four suffice). However the modeling approach relies on the restrictive assumption that the covariance belongs to a specific parametric family. No such parametric models have been proposed and proven to be valid for fMRI data. By contrast, the empirical approach is easy to use and very popular in geostatistics (Atkinson and Lewis, 2000). I consider the empirical approach here.

The empirical autocovariance $\hat{C}(h)$ can be calculated as follows (Isaaks and Srivastava, 1989):

$$\hat{C}(h) = \frac{1}{N(h)} \sum_{(t_i,t_j)|h_{t_i,t_j}=h} Z(t_i) \cdot Z(t_j) - \hat{\mu}_h^2, \tag{3.1}$$

where $\hat{\mu}_h = \frac{1}{N(h)} \sum_{(t_i)|h_{t_i,t_j}=h} Z(t_i)$; that is, the average of the $Z(\cdot)$ values for points within a lag h of the reference point. The autocorrelation $\hat{\rho}(h)$ is the autocovariance standardized

by the variance:

$$\hat{\rho}(h) = \hat{C}(h)/\hat{\sigma}_h^2, \tag{3.2}$$

where $\hat{\sigma}_h^2 = \frac{1}{N(h)} \sum_{(t_i)|h_{t_i,t_j}=h} Z(t_i)^2 - \hat{\mu}_h^2$; that is, the variance of the $Z(\cdot)$ values for points within a lag h of the reference point.

The definitions of autocovariance and autocorrelation can be extended to describing the relations between several variables. If we consider two different regionalized variables $Z_u(t_i)$ and $Z_v(t_i)$, where variables u and v both are second order stationary with respective means μ_u and μ_v, then for variable u we have

$$E[Z_u(t_i + h) - Z_u(t_i)] = 0,$$
$$C_{uu}(h) = E[\{Z_u(t_i) - \mu_u\}\{Z_u(t_i + h) - \mu_u\}],$$

and analogously for $C_{vv}(h)$. The cross-covariance and cross-correlation are defined as

$$C_{uv}(h) = E[\{Z_u(t_i + h) - \mu_u\}\{Z_v(t_i) - \mu_v\}],$$

and

$$\rho_{uv}(h) = C_{uv}(h)/\sqrt{C_u(0)C_v(0)}, \tag{3.3}$$

respectively. Cross-covariance and cross-correlation can be estimated by moment estimators similar to those for autocovariance (3.1) and autocorrelation (3.2) (see Goutte et al., 1999).

Classification metric step The classification metric is usually defined as the distance measure between two objects in multi-dimensional space or time. For all triples of objects (i, j, k), the distance measure d_{ij} between the i^{th} and j^{th} objects is said to be a metric if it satisfies: $d_{ij} \geq 0$ (non-negativity); $d_{ii} = 0$ (identity); $d_{ij} = d_{ji}$ (symmetry) and $d_{ij} \leq d_{ik} + d_{kj}$ (sub-additivity) (Gordon, 1999; Stanberry et al., 2003). In considering two objects of interest, there are two different types of measure. One measures their dissimilarity; the other measures their similarity. Both can be developed into metrics. Measures of dissimilarity include Euclidean distance, Mahalanobis distance, and the variogram. Measures of similarity include correlation and covariance.

Classification algorithm step There are two principal clustering approaches, hierarchical clustering and partitioning (Rencher, 2002). In hierarchical clustering, each object is first regarded as a separate cluster and clusters are then combined sequentially. The number of clusters is reduced at each step until only one cluster is left. The results are usually presented in a tree-like structure called a dendrogram, which shows the steps of the clustering procedure. In partitioning, objects are divided into k pre-defined clusters by some optimality criterion (Kaufman and Rousseeuw, 1990). The k means method (MacQueen, 1967) is probably the most widely used partitioning method.

Silhouette values

Silhouette values, first introduced by Rousseeuw (1987), are used to judge the results of a classification procedure. Each cluster is represented by a *silhouette*, showing which objects lie well within the cluster and which objects merely hold an intermediate position. The entire clustering procedure is displayed by plotting all silhouettes on a single diagram, allowing the user to compare the quality of the clusters (Kaufman and Rousseeuw, 1990). Silhouettes are constructed as follows: For each object i, let A be the cluster that has been assigned and $a(i)$ = average dissimilarity of i to all other objects in A. For any other cluster C different from A, define $d(i, C)$ = average dissimilarity of i to all objects in C. Let $b(i) = min_{C \neq A} d(i, C)$. $d(i, B) = b(i)$ is called the neighbor of object i; this is the second best choice for object i: If cluster A is discarded, cluster B is closest to i. The silhouette $s(i)$ is defined as

$$[b(i) - a(i)]/max[a(i), b(i)].$$

By definition, $s(i)$ is between -1 and 1. When $s(i)$ is close to 1, the "within" dissimilarity $a(i)$ is much smaller than the smallest "between" dissimilarity $b(i)$, thus object i is classified to the right cluster. When $s(i)$ is near 0, $a(i)$ and $b(i)$ are approximately equal and hence it is not clear whether i should have been assigned to A or B. Object i lies equally far away from both clusters and can be considered as an intermediate case. When the value is close

to -1, then $a(i)$ is much larger than $b(i)$, so i lies much closer to B than to A. Therefore, object i has probably been misclassified.

The average of $s(i)$ for all objects in a cluster is called the *average silhouette width* of that cluster. The average of the $s(i)$ for $i = 1, 2, ...$, is called the average silhouette width for the entire data set. The average silhouette width for the entire data set is used to select the number of clusters k, by choosing k so that the average silhouette width is highest (Kaufman and Rousseeuw, 1990).

3.3 Data Analysis

3.3.1 Classification steps for saccade data

I first perform clustering analysis to the saccade data, hoping to find the regions in the brain reacting to the task. I focus on the fourth axial slice of the brain, which for this subject is expected by the researcher who provided the data to have the most activity. I apply clustering based on autocorrelation, and compare the results with those based on cross-correlation (Goutte et al.,1999).

Data reduction The number of voxels in even a single slice of data still poses a challenge to many clustering techniques. Clustering methods usually do not work well for ill-balanced data. "Ill-balanced" means that the number of observations belonging to the different classes are widely disparate, e.g., if most observations belong to one class, then all observations might be put into a single cluster, even if there are different patterns in reality. In fMRI data the population of "activated" (i.e., stimulus-related) voxels is much smaller than the total population of voxels (Goutte et al., 1999; Fadili et al., 2000; Wang et al., 2007). Hence, a data reduction step is advisable before the three clustering steps to avoid all voxels being assigned to a single cluster.

Masking the brain is a popular dimension reduction technique. Once a specific slice or set of slices is chosen by the researcher, the data image is first processed to identify the location

of the brain, thereafter a suitable threshold method is used to remove all voxels outside the brain (see for example, Goutte et al., 1999, 2001; Stanberry et al., 2003; Gibbons et al., 2004; Bowman et al., 2004). For the saccade data, I will use both masked and unmasked data in clustering, and compare the results. For the slice I chose for the purpose of demonstration, there are 630 voxels (out of 4096) after masking.

Since the goal of the screening is to reduce the large amount of non-stimulus-related voxels that could seriously affect the robustness and sensitivity of the clustering results (Goutte et al., 1999; Fadili et al., 2000), I use a simple two-sample t test procedure comparing prosaccade to antisaccade to screen out the probable non-active voxels (Huettel et al., 2004). A suitably generous threshold ($|t| > 2$) is applied to create an image showing those regions of the brain with moderate to strong task-related activation. After thresholding, 345 of 4096 voxels are retained in the unmasked data and 167 of 630 voxels are retained in the masked data; only these are subject to clustering. Note the 167 voxels are included in the 345 voxels left in the unmasked data.

For the long range resting data, we only have masked data leaving us with 1096 voxels out of the original 4096. Since there is no contrast between different conditions, two-sample t test is not suitable here and all the 1096 voxels are used for clustering.

Feature extraction step For the 345 most active voxels in the unmasked data, I use the empirical autocorrelations at the 156 time points as the main feature (Marcotte, 1996). For the k-th voxel, $k = 1, 2, ..., 345$, define the time series $\{Z_k(t); t = 1, ..., 156\}$, with empirical autocovariance and autocorrelation $\hat{C}_k(h)$, $\hat{\rho}_k(h)$ as in equations (3.1) and (3.2). Every voxel is represented by the new vector $\hat{\boldsymbol{\rho}}_{\mathbf{k}} = (\hat{\rho}_k(0), \hat{\rho}_k(1), ..., \hat{\rho}_k(155))$ (Figure 3.1, graph (3)) instead of the original time series $Z_k(t)$ (Figure 3.1, graph (1)). I also calculate the cross-correlation (Figure 1, graph (4)) between each voxel $Z_k(t)$ and the stimulus series $Y(t)$ (Figure 3.1, graph (2)). As a comparison, in Section 3.3.1 I present results using the cross-correlation as the feature in clustering.

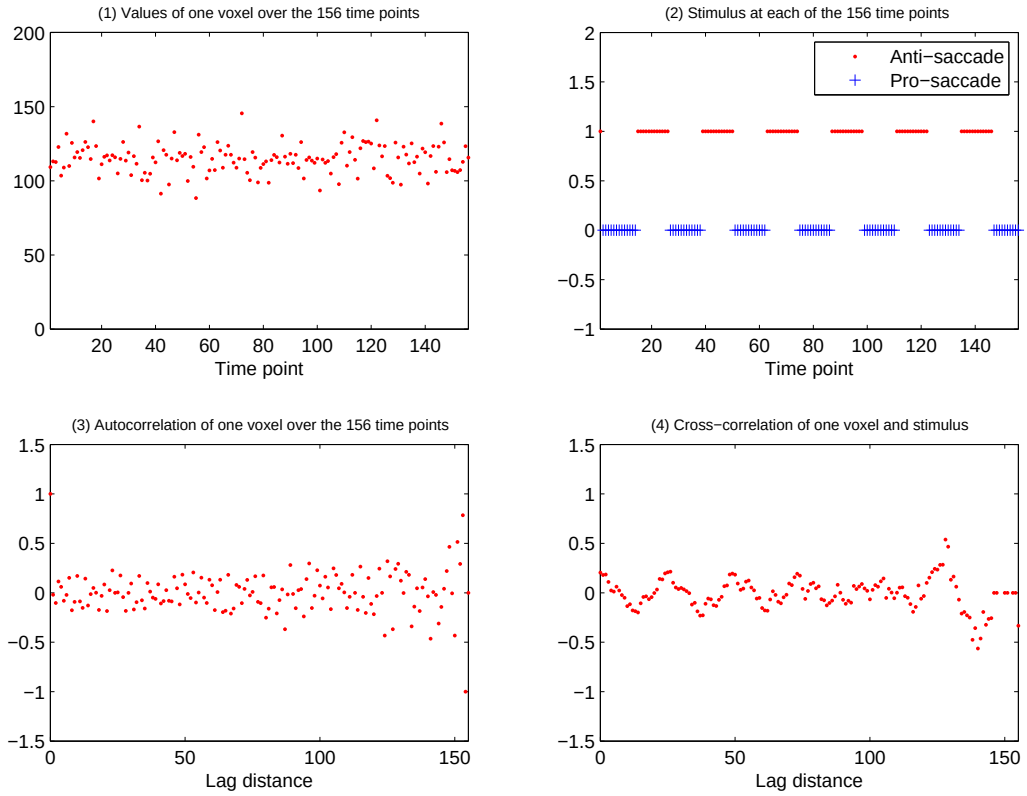

Figure 3.1: (1) time series of one voxel; (2) plot of stimulus time course; (3) autocorrelation of one voxel from lag distance 0 to 155. (4) cross-correlation of one voxel and the stimulus from lag distance 0 to 155. Graphs (3) and (4) clearly show that measurement error increases as the lag distance increases.

Due to a lack of data, $\hat{\rho}(h)$ becomes more variable when h is large. In my analysis, I thus consider h from 0 up to a maximum lag H. The results in the sequel are based on $H = 99$. However, the results are not very sensitive to the particular choice of H. I discuss this further in Section 3.4.

Clustering metric step Two metrics are commonly used at this step. One is the "generalized distance" (Goutte et al., 1999), also called the "functional distance" (Bowman, 2007). For two voxels $\rho_{\mathbf{m}}$ and $\rho_{\mathbf{n}}$ in R^{H+1}, $m, n = 1, ..., 345$ and $H = 99$, this distance is defined as

$$d_E^2(\hat{\rho}_{\mathbf{m}}, \hat{\rho}_{\mathbf{n}}) = (\hat{\rho}_{\mathbf{m}} - \hat{\rho}_{\mathbf{n}})^T \mathbf{M} (\hat{\rho}_{\mathbf{m}} - \hat{\rho}_{\mathbf{n}}),$$

where $\mathbf{M}$ is a $(H+1) \times (H+1)$ symmetric positive definite matrix. When $\mathbf{M}$ is the identity matrix, d_E^2 reduces to the Euclidean distance, called "Euclidean". The other metric is the square root of one minus the sample correlation between two voxels, called "correlation" (Rencher, 2002; Strauss et al., 1973), which is defined as

$$d_{corr}^2(\hat{\rho}_{\mathbf{m}}, \hat{\rho}_{\mathbf{n}}) = 1 - \frac{\sum_{h=0}^{H} (\hat{\rho}_m(h) - \hat{\mu}_m)(\hat{\rho}_n(h) - \hat{\mu}_n)}{\sqrt{\sum_{h=0}^{H} (\hat{\rho}_m(h) - \hat{\mu}_m)^2 \sum_{h=1}^{H} (\hat{\rho}_n(h) - \hat{\mu}_n)^2}}, \tag{3.4}$$

where $\hat{\mu}_m = \frac{1}{H} \sum_{h=0}^{H} \hat{\rho}_m(h)$, $\hat{\mu}_n = \frac{1}{H} \sum_{h=0}^{H} \hat{\rho}_n(h)$, $m, n = 1, ..., 345$, $H = 99$.

The "Euclidean" method can only compare the similarity of average profile levels among the voxels. We are also interested in the similarity of profile shapes, since voxels with similar patterns should belong to the same functional regions of the brain. The "correlation" metric is more appropriate for this task (Rencher, 1998 and 2002). The connections between the two metrics are as follows:

Given an $(n \times p)$ matrix $\mathbf{X}$, where n is the number of observations (time points) and p is the number of variables (locations). Define $\hat{\mu}_{\mathbf{j}}$ and $\hat{\sigma}_{\mathbf{j}}^{\mathbf{2}}$ as the mean and variance of the j^{th} vector $\mathbf{x_j} = (x_{1j}, ...x_{nj})^T$, where $j = 1, ..., p$; and define a new vector $\tilde{\mathbf{x}}_{\mathbf{j}} = (\tilde{x}_{1j}, ..., \tilde{x}_{nj})^T$, where $\tilde{x}_{ij} = (x_{ij} - \hat{\mu}_{\mathbf{j}})/(\sqrt{n}\hat{\sigma}_{\mathbf{j}})$, $i = 1, ..., n$. Then $\tilde{\mathbf{x}}_{\mathbf{j}}$ is normalized since $\tilde{\mathbf{x}}_{\mathbf{j}}^T \tilde{\mathbf{x}}_{\mathbf{j}} = \sum_{i=1}^{n} \tilde{x}_{ij}^2 = 1$. Note that $d^2(\tilde{\mathbf{x}}_{\mathbf{j}}, \tilde{\mathbf{x}}_{\mathbf{k}}) = (\tilde{\mathbf{x}}_{\mathbf{j}} - \tilde{\mathbf{x}}_{\mathbf{k}})^2 = 2(1 - \tilde{\mathbf{x}}_{\mathbf{j}}^T \tilde{\mathbf{x}}_{\mathbf{k}}) = 2d_{\text{corr}}^2(\mathbf{x_j}, \mathbf{x_k})$ for $j, k = 1, ..., n$. The square

of the "correlation" metric is an alternative square of the normalized "Euclidean" metric. Hence, the "correlation" metric can extract the similarity of profile shapes in clustering.

Clustering algorithm step The k means algorithm is common in neuroimaging applications because of its computational advantages: computations are fast, the algorithm does not require retention of all distances, and convergence occurs quickly (Bowman et al., 2004). For a given number of clusters k, it iteratively minimizes the within-class variance by assigning data to the nearest center and recalculating each center (Goutte et al. 2001). Hence, I choose k means algorithm here. To choose the number of clusters k, silhouette values (Rousseeuw, 1987) is used.

Results of autocorrelation method

In this subsection, I discuss clustering results using autocorrelation as the feature.

Unmasked data All 345 voxels that passed the initial t test screening are used for clustering. I use both average silhouette value and the index of Calinski and Harabasz (1974) to choose the number of clusters; according to both criteria, the appropriate number of clusters is four. After clustering, I mask the average brain image over time and discard all the voxels outside the brain. There are 167 voxels inside the brain of the 345 retained voxels (Figure 3.2).

Cluster 1 has 111 voxels, of which 88 are inside the brain (Figure 3.2, graphs (5) and (6)); empirical autocorrelations derived from the time points clearly exhibit strong peaks and troughs in the shape of waves, showing periods of time correlation (Figure 3.3, graph (4)), which likely result from variations of the blocks in the design paradigm (Radeloff et al., 2000; Chen, 2005). The signals of these voxels are quite strong (Figure 3.3, graph (3)), indicating the activation is probably due to "true" brain activity. The regions identified in cluster 1 (Figure 3.2, graphs (5) and (6)) are also confirmed to be active regions by another lab.

Cluster 2 has 80 voxels, of which 46 are inside the brain (Figure 3.2, graphs (7) and (8)). The empirical autocorrelations of these voxels also exhibit some peaks and troughs, but they do not match up with the stimulus sequence as well as those in cluster 1 (Figures 3.3, graph (6)). By graphs (7) and (8) in Figure 3.2, we can see the voxels in this cluster are around the brain. It seems the "activation" of this cluster is due to uncorrected head motion, according to the hypothesis of the researcher who supplied the data.

Cluster 3 has 62 voxels, the correlation is around zero. There are no obvious patterns in the signal sequence and the autocorrelation shows some overlapping cyclic patterns which do not correspond to the experimental design (Figure 3.3, graphs (7), (8)). As shown in graphs (9) and (10) in Figure 3.2, 33 voxels of this cluster are inside the brain and 29 voxels are outside the brain. I conclude this cluster contains noise voxels that failed to be screened out by the t test. Cluster 4 has 92 voxels, the mean of the correlations is almost zero (Figure 3.3, graph (10)). By looking at graphs (11) and (12) in Figure 3.2, all the voxels are outside the brain, hence are clearly noise.

Masked data Here the 167 retained voxels in the masked fMRI data are clustered directly. Since in unmasked data, all the voxels in cluster 4 are outside the brain, for comparing to the unmasked data fairly, I look here at the results when $k = 3$. Cluster 1 has 60 voxels and the empirical autocorrelations show strong periods of time correlations, indicating "activation" (Figure 3.4, graph (4)). Cluster 2 has 73 voxels and the empirical autocorrelations also show strong periods of time correlations (Figure 3.4, graph (6)), but its pattern is different from cluster 1. Since all the voxels are around the edges of the brain (Figure 3.5), cluster 2 indicates "head motion". Cluster 3 has 34 voxels and the empirical autocorrelation shows some overlapping cyclic patterns which do not corresponding to the experimental design (Figure 3.4, graph (8)).

The results are very similar to the unmasked method except the numbers of voxels in each cluster change (Figure 3.5). Both methods extract the 33 voxels of noise inside the brain (Figure 3.5, graph (9)), but the unmasked method prefers to attribute more voxels

to “activation” (Figure 3.5, graph (3)), whereas the masked method prefers to assign more voxels to “head motion” (Figure 3.5, graph (6)). By comparing the two “activation” clusters with the maps from another lab, the results from the unmasked method are closer to what has been found by other researchers. Because the saccade data has been preprocessed to correct for head motion, the number of voxels in the “head motion” cluster should be small compare to the number of voxels in the “activation” cluster. This is true for the unmasked data, but not for the masked data. Also, it is noted that the “head motion” cluster in masked data shows strong periods of time correlations (Figure 3.4, graph (6)), which means some voxels due to the “activation” may be misclassified to “head motion” cluster. Figure 3.6 shows the mean correlations of the 60 voxels in the masked method from Figure 3.5 graph (2), the 46 voxels in the unmasked method from Figure 3.5 graph (4), and the 28 voxels from Figure 3.5 graph (3). Clearly, the degree of similarity between graphs (1) and (3) is higher than that between graphs (2) and (3). Hence, the 28 voxels are closer to the “activation” cluster.

This phenomenon can be explained by the so-called “marginal effect”. In the calculation of a characteristic of a specific region, the lack of sufficient data outside this region may result in biases or other errors, which is called the “marginal effect” or “edge effect”. In order to reduce the impact of the marginal effect on a region, an easy solution is to calculate first on an extended region, then remove the extension after calculation to get the actual results. This idea can be referred to in our clustering analysis. By looking at the unmasked “head motion” cluster (Figure 3.2, graph (7)), we observe that almost all of the 80 voxels are around the edges of brain. The 34 voxels outside the brain have a strong “marginal effect” with the 46 voxels which are just inside the brain (Figure 3.2, graph (8)). Masking the brain before clustering removes all the voxels outside the brain, including the above 34 voxels, therefore the 46 voxels may group with some other voxels inside the brain (Figure 3.5, graphs (3) and (6)). Hence, to reduce the “marginal effect”, it is preferred to use unmasked

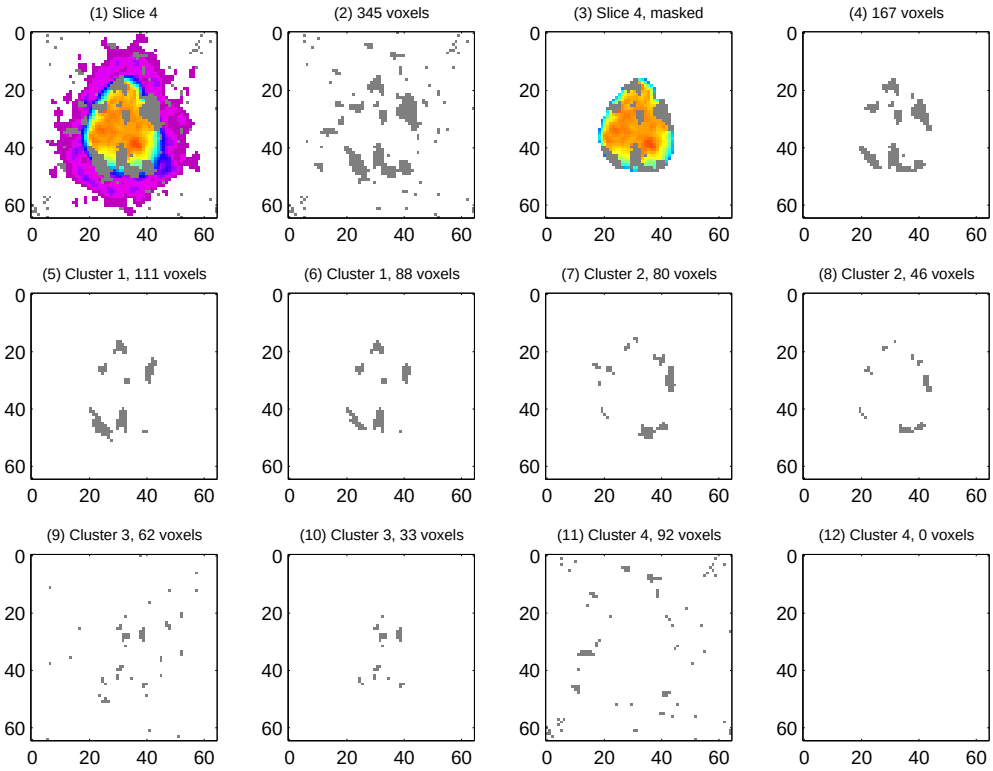

Figure 3.2: Maps of the brain for saccade data, autocorrelation method. (2) and (1) are maps of the 345 retained voxels and overlaid on the original brain. (4) and (3) are maps of the 167 voxels after masking and overlaid on the masked brain. (5) and (6) are maps of cluster 1 before and after masking; (7) and (8) are maps of cluster 2 before and after masking; (9) and (10) are maps of cluster 3 before and after masking; (11) and (12) are maps of cluster 4 before and after masking. There are 111, 80, 62, 92 voxels in the four clusters before masking, and 88, 46, 33, 0 voxels left after masking.

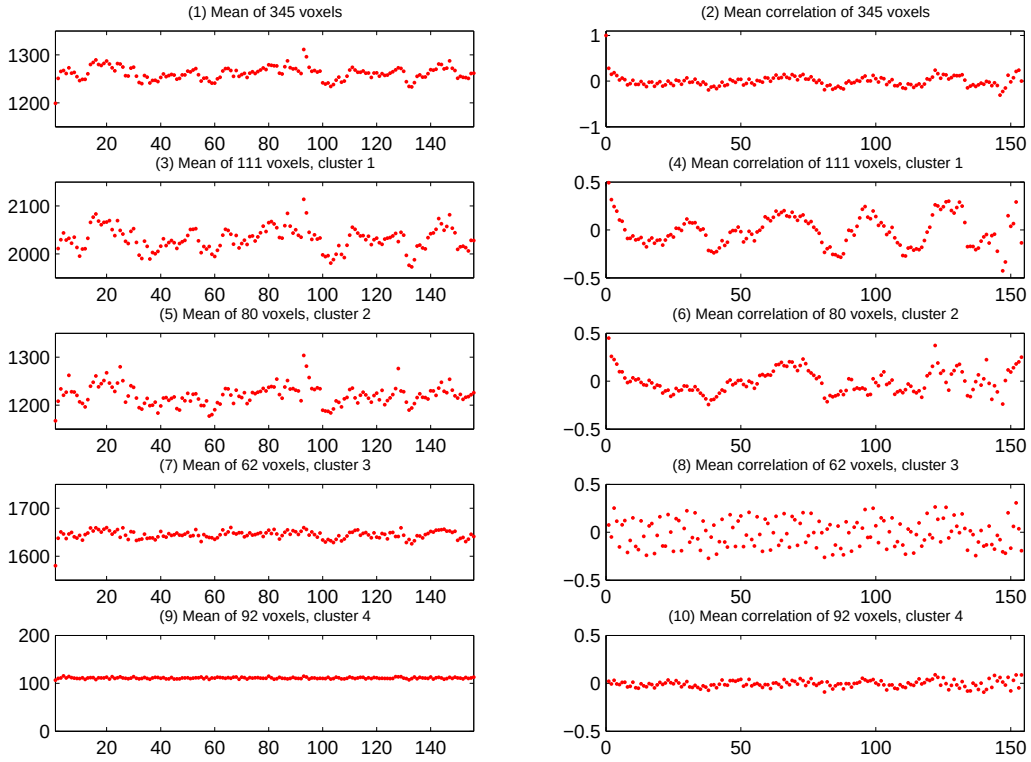

Figure 3.3: Time patterns for saccade data, unmasked method. (1) is the mean of the 345 retained voxels. (3), (5), (7), (9) are the means of the four clusters. (2) is the mean correlation of the 345 voxels. (4), (6), (8), (10) are the mean correlations of the four clusters. The mean correlation of cluster 1 shows clear peaks and troughs in the shape of waves; the mean correlation of cluster 2 shows weak peaks and troughs in the shape of waves; the mean correlations of cluster 3 shows some overlapping cyclic patterns which do not correspond to the experimental design, and cluster 4 shows no pattern.

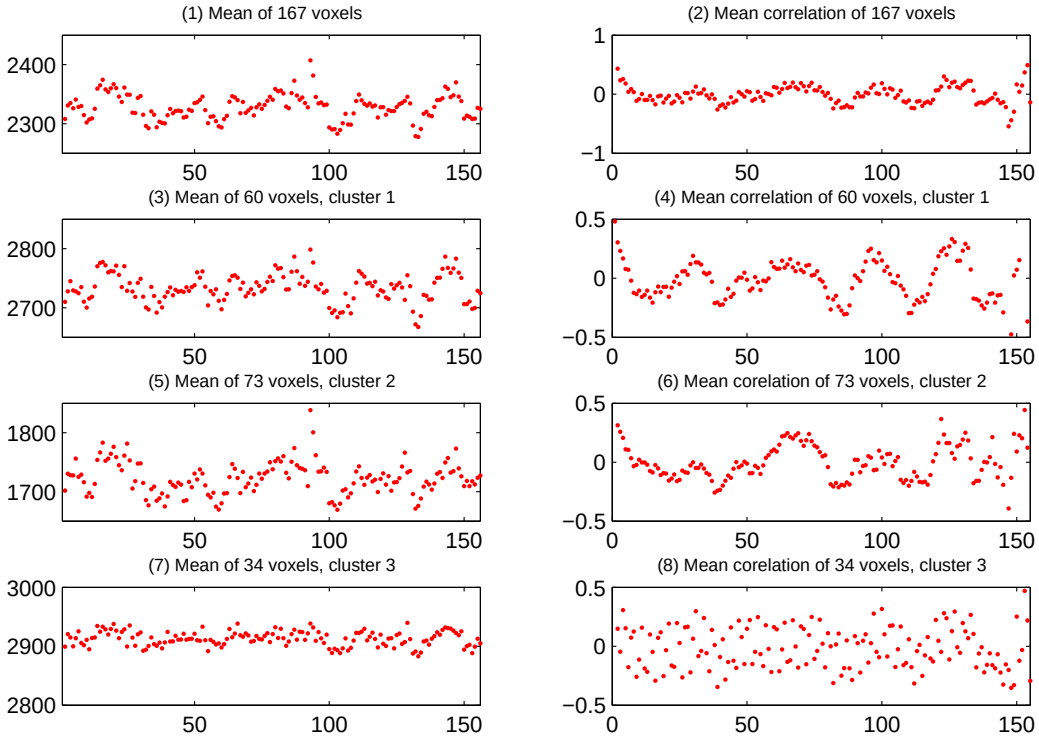

Figure 3.4: Time patterns for saccade data, masked method. (1) is the mean of the 167 retained voxels. (3), (5), (7) are the means of the three clusters. (2) is the mean correlation of the 167 voxels. (4), (6), (8) are the mean correlations of the three clusters. The mean correlation of cluster 1 shows clear peaks and troughs in the shape of waves; the mean correlation of cluster 2 also shows strong peaks and troughs in the shape of waves; the mean correlations of cluster 3 shows some overlapping cyclic patterns which do not correspond to the experimental design.

data in clustering especially when the data have head motion evident around the edges of the brain.

Results of cross-correlation method

I also use the cross-correlation between the 345 retained voxels and the stimulus as the main feature for clustering. Again for fair comparison, I look at results when $k = 4$. There are 72, 106, 109, 58 voxels in the four clusters and 50, 50, 66, 1 voxels inside the brain respectively. Using the maps in Figure 3.2 as our standard, the maps in Figure 3.7 show that clusters 1 and 3 are not well classified. The active voxels in the anterior and posterior of the brain are mixed up with noise inside the brain (Figure 3.7, graphs (6) and (10)). The intention of the cross-correlation method is to try to use the characteristics of delay and habituation between the stimulus and the response time series to create different partitions of the brain. The clustering metric is used to count the total number of relevant matches between the partitions. Perhaps differences in delay and habituation are not strong enough to be distinguished by a clustering method. Hence, in cross-correlation the clustering algorithm may just classify the voxels inside the brain by their voxel values, and not by their different properties. Because of this, researchers sometimes smooth the cross-correlation function (e.g., Goutte et al., 2001) to improve its precision and to get better results. This makes a simple case overly complicated, and furthermore is unnecessary, as shown by the autocorrelation results.

3.3.2 Classification for masked long range resting data

For the masked long range resting data set, there should be no task-related differences among the three slices since the subject performed no task. Hence I examine the first slice as an example. Since there is no task-related activation over time, the 2-sample t-test is not suitable here as a screening device and all 1096 voxels are used in clustering. This is a long range data with maximum lag distance 1498, and I pick an effective lag distance of $h = 1000$ to reduce

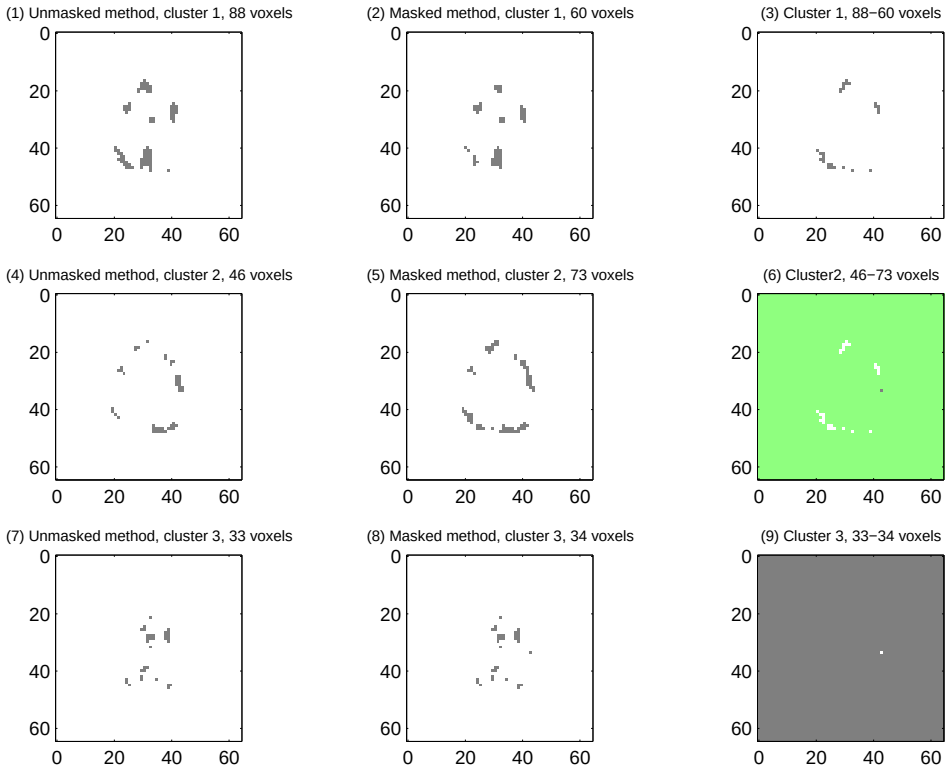

Figure 3.5: Comparison between unmasked and masked methods. The numbers of voxels in "activation", "head motion" and "noise" are 88, 46, 33 respectively in the unmasked method (graphs (1), (4), (7)), and 60, 73, 34 respectively in the masked method (graphs (2), (5), (8)). Both methods can extract the 33 voxels of noise (graph (9)), but the unmasked method prefers to assign more voxels to "activation" (graph (3)), while the masked method prefers to attribute more voxels to "head motion" (graph (6)).

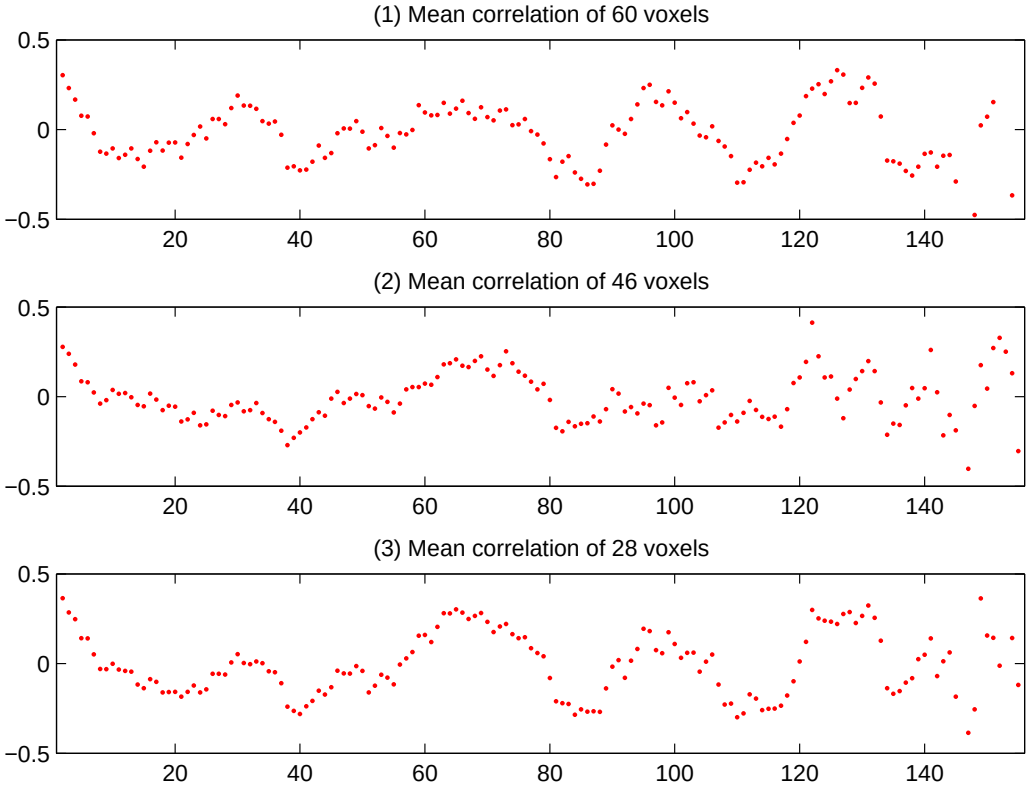

Figure 3.6: Time patterns for different clusters. (1) is the mean correlation of the 60 voxels in the masked method from Figure 3.5 graph (2). (2) is the mean correlation of the 46 voxels in the unmasked method from Figure 3.5 graph (4). (3) is the mean correlation of the 28 voxels from Figure 3.5 graph (3). Clearly, the degree of similarity between graphs (1) and (3) is higher than that between graphs (2) and (3).

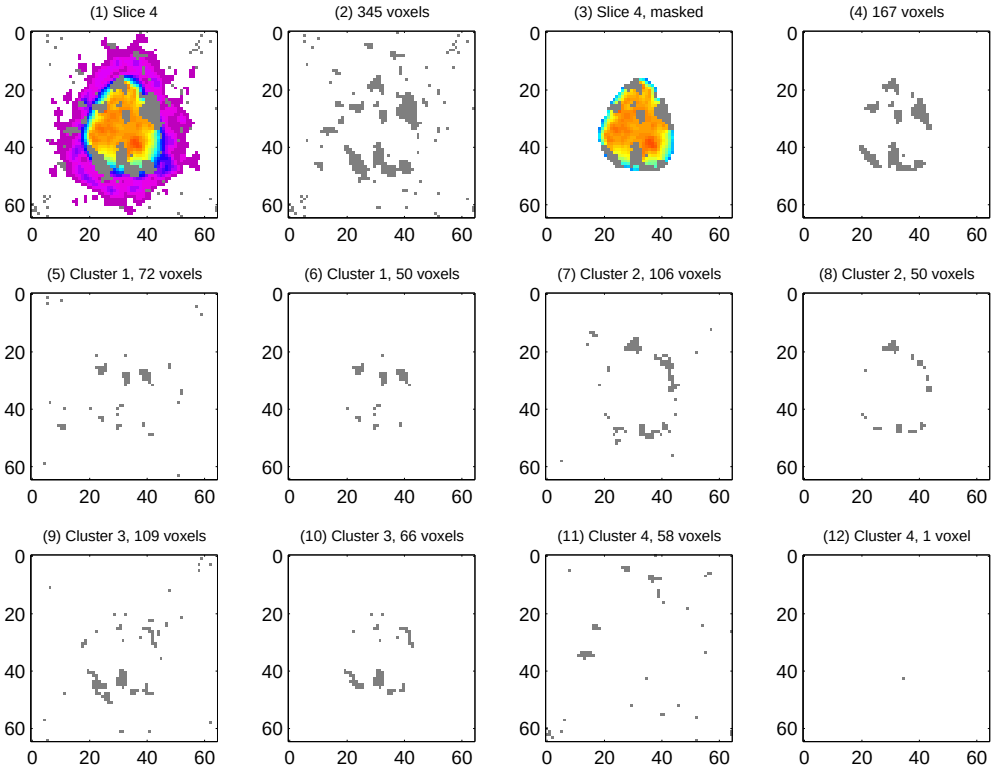

Figure 3.7: Maps of the brain for unmasked saccade data, cross-correlation method. (2) and (1) are maps of the 345 retained voxels and overlaid on the original brain. (4) and (3) are maps of the 167 voxels after masking and overlaid on the masked brain. (5) and (6) are maps of cluster 1 before and after masking; (7) and (8) are maps of cluster 2 before and after masking; (9) and (10) are maps of cluster 3 before and after masking; (11) and (12) are maps of cluster 4 before and after masking. There are 72, 106, 109, 58 voxels in the four clusters before masking, and 50, 50, 66, 1 voxels left after masking.

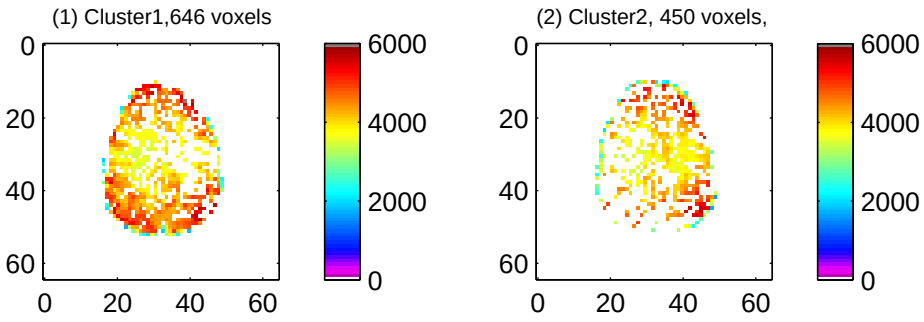

Figure 3.8: Maps of resting data. (1) is the map of cluster 1 with 646 voxels. (2) is the map of cluster 2 with 450 voxels.

the measurement error. I use the "correlation" metric again and the mean of silhouette values indicates that the best choice for the number of clusters is k=2. Cluster 1 has 646 voxels (Figure 3.8, graph (1)); cluster 2 has 450 voxels (Figure 3.8, graph (2)). There are no clear patterns in the two clusters. Neither cluster shows peaks or troughs in time (Figure 3.9), indicating that there are no task-related or systematic activations in the resting data. This is what we would expect.

3.4 Discussion and conclusion

3.4.1 Different comparisons

Comparison between autocorrelation and cross-correlation method Clustering on the cross-correlation function instead of the raw time series may provide increased robustness. This type of clustering is an example of an *hypothesis-driven analysis* (Huettel et al. 2004). A drawback of hypothesis-driven analysis is that it depends heavily on having prior knowledge of the reference function. Sometimes it is difficult to know this in advance, and sometimes such a reference waveform doesn't exist. For example, in an experiment that involves having the subject watch a film clip, there is not a clear "time course" with which to cross-correlate.

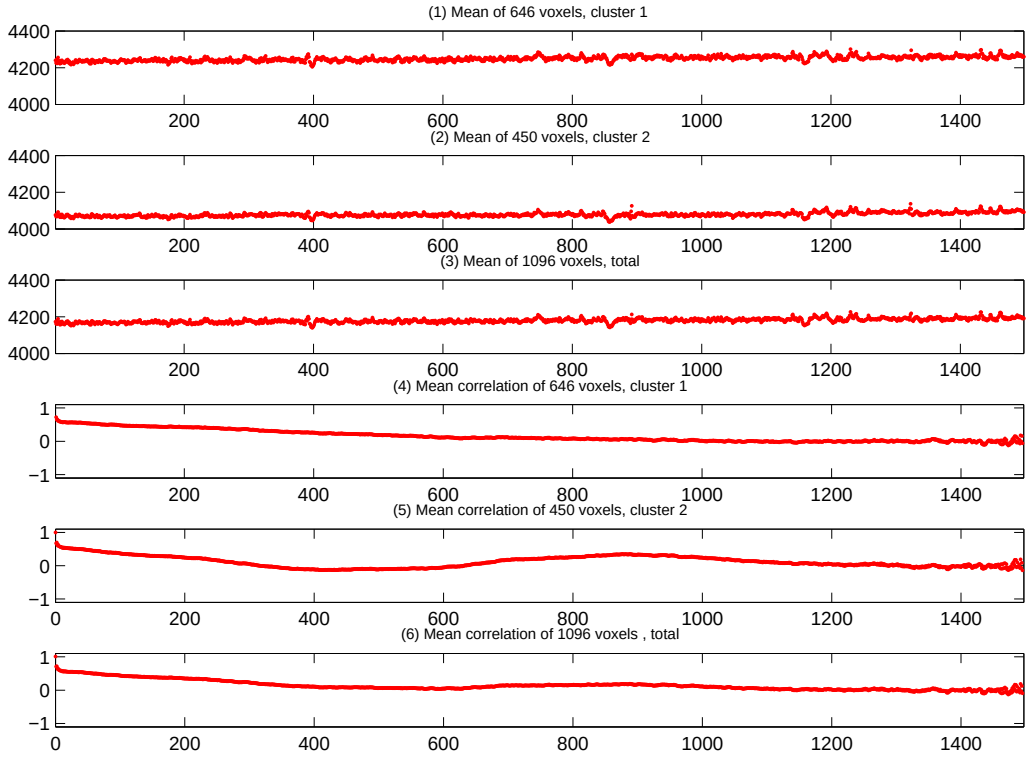

Figure 3.9: Time patterns of resting data. (1), (2) are the means of the two clusters. (3) is the mean of all 1096 voxels. (4), (5) are the mean covariances of the two clusters. (6) is the mean covariance of all 1096 voxels. Neither cluster exhibits obvious peaks and troughs in time, indicating that there are no systematic activations in the resting data.

Even in a study with a well-defined stimulus waveform, the subject may not follow the instructions correctly during the experiment and it is not clear how this will affect the cross-correlation analysis. Furthermore, we might not know the expected response to the task for a particular region because the study is exploratory. One can easily envision other scenarios where it might be difficult or undesirable to work with a predetermined reference function. The autocorrelation method does not rely on such an external standard, and hence can be considered a type of *data-driven analysis* (Huettel et al. 2004). It offers further improvement over the cross-correlation method, which itself improves on clustering of the raw time series. Using the autocorrelation function as a main feature in clustering makes fewer assumptions. In particular, as we have seen, we need not consider the hypothesized relation between the stimulus and the response time series (including concerns about lags in the onset of the hemodynamic response). My results indicate that this approach is superior in identifying active regions of the brain for task data, and can be used for resting data to examine functional similarities across the brain during ambient thought.

Comparison between saccade data and resting data The autocorrelations of the clusters in the saccade data exhibit peaks and troughs in the shape of waves in two of the clusters, showing periods of time correlations. This periodicity is clearer than that shown by the raw voxel time courses and reflects changes in the experimental condition. By contrast, the autocorrelations of clusters for the resting data do not show any clear patterns over time. These results demonstrate the good performance of autocorrelation in a variety of situations and its superiority over using the voxel time courses alone.

Comparison between unmasked method and masked method It is common to mask the brain as the first step in an analysis (Goutte et al., 1999, 2001; Stanberry et al., 2003; Gibbons et al., 2004; Bowman et al., 2004), since masking can eliminate a large number of "useless" voxels outside the brain and make the further analysis more convenient. But masking the brain before clustering may be problematic, as it ignores the effects of voxels

around the edges of the brain. These voxels may affect the clustering results especially for data with head motion.

3.4.2 Discussion of different techniques

2-sample t-test for dimension reduction Using the 2-sample t-test for dimension reduction is a popular technique in fMRI data analysis. It is simple, convenient and effective in extracting the most important voxels. But it also has some clear disadvantages. First, it requires a contrast between experimental conditions, so it does not work for resting data since there are no such conditions to compare. Nor is it truly suitable for event-related designs. Secondly, delay of the responses is not considered, and so the calculation of the two sample t statistic may not be accurate. Sparse principal component analysis (Zou et al., 2006) is a new technique which can combine dimension reduction and feature extraction together in clustering, and which may overcome the disadvantages of the 2-sample t-test. I explore this approach in the next chapter.

Use of empirical autocorrelation Using empirical autocorrelation in fMRI data analysis is attractive since it captures important characteristics of the voxels over time and does not require prior knowledge of the reference function. But there are some implementation issues. For example, when the method of moments is used to estimate the autocorrelation, the standard error of $\hat{\rho}(h)$ increases as lag h increases, because we have fewer data pairs for large h. Hence the autocorrelation needs to be truncated at an effective lag. On the other hand, we do not want to lose important structure by truncating too much. In practice, it is recommended to choose the number of data pairs to be no less than 30 or to truncate the lag distance at half of the maximum distance (Journel and Huijbregts, 1978; Bowman, 2007). I tried different effective lag values and found that the clustering results are not very sensitive, for a reasonable range of effective lags. Table 3.1 lists average silhouette values and the number of misclassified voxels inside the brain for different effective lag values. To keep

the balance between retaining more data information and reducing measurement error, an effective lag distance of 99 seems to be optimal for this data set.

Effective lag	91	93	95	97	99	101	103	105
Silhouette value	0.2747	0.2710	0.2623	0.2630	0.2682	0.2680	0.2628	0.2584
Misclassified voxels	4	3	3	6	1	3	2	3

Table 3.1: Average silhouette values and the number of misclassified voxels inside the brain at different lags. To keep the balance between retaining more data information and reducing measurement error, an effective lag distance of 99 is optimal.

Use of correlation instead of covariance As I mentioned before, the reason to use correlation instead of covariance in clustering is because in the absence of standardization, clustering will extract regions with large variances rather than similar temporal patterns (Goutte et al., 1999). But when I use "correlation" metric in clustering, the above standardization is not necessary here. Actually, since I have already used the covariance structure of the data in clustering instead of the raw data, the standardization should be performed for the covariance structure of the data now (Baudelet and Gallez, 2003).

3.4.3 Conclusions

In the analysis of brain imaging data, using the autocorrelation function offers an important advantage over existing cross-correlation approaches. Unlike conventional cross-correlation methods, the proposed method doesn't require prior knowledge about the reference function, and does characterize the important features of voxel changes in time. The analysis also provides evidence that masking the brain may affect the clustering results. Although many researchers often choose masking the brain as the first step to reduce dimension, I show that this is not necessarily effective and sometimes results in less convincing results, especially when the data have "head motion" evident around the edges of the brain.

Chapter 4

Geostatistical analysis and sparse principal component analysis in clustering fMRI time series

In the previous chapter, I used a data-driven function based on geostatistical ideas in clustering and got good results. But this method still needs a contrast between experimental conditions to do the dimension reduction before clustering. Here I use a technique called LASSO for the dimension reduction step. The proposed methods in this chapter will change the whole clustering process to be a data-driven approach.

4.1 Introduction

As discussed earlier, the model-free approach has advantages over the model-based approach, because it is a data-driven analysis by a model-free method, where the effects or components of interest are found from the intrinsic structure of the data directly (Jezzard et al., 2001; Huettel et al., 2004).

Principal component analysis (PCA) is a typical model-free approach. In PCA, the data are partitioned into uncorrelated components whose patterns vary over time along mutually orthogonal principal component axes ; each component can be considered as a cluster. But this method does not always work well, because PCA can not separate the data unambiguously into activation and noise, and components indicating activation may be contaminated with instrumental or physiological noise (Sommer and Wichert, 2002). The reason is because the amplitude of the signal change inside the brain is small and different clusters may be overlapped (Backfriender et al., 1996). To overcome these disadvantages, some researchers (e.g., Backfriender et al., 1996) consider performing *factor analysis* for a selected *region of*

interest (ROI) again in several additional steps, but their method depends heavily on the selection of the specific ROI and prior knowledge of the noise and artifacts.

Clustering based on autocorrelation structure offers an important advantage over the raw time series and cross-correlation approaches. This model-free approach captures the important characteristics of the voxels over time and can separate the activation and noise very well, as shown in the previous chapter. But it still has a weakness. Namely the lack of dimension reduction. Because clustering method usually does not work well for the *ill-balanced data* (Goutte et al., 1999; Fadili et al., 2000; Wang et al., 2007), where the population of "activated" (e.g., stimulus-related) voxels is much less than the total population of the voxels. Hence the above method needs a contrast between experimental conditions to remove a large portion of voxels that are almost impossible to be considered as "activated" before clustering, e.g., 2-sample t-test. But in some cases, such as in resting data or event-related data, there is no clear contrast for this dimension reduction step. Also, the delay of the responses is difficult to control, so the 2-sample t-test may not be accurate for dimension reduction.

Sparse principal component analysis (SPCA) (Zou et al., 2006) is a new technique which combines dimension reduction and feature extraction together in clustering, and thus can overcome the disadvantages of the 2-sample t-test. In SPCA, the focus is on the correlations among the time courses, which is very similar to the clustering method. The *sparse components* from SPCA can be regarded as different clusters. After the dimension reduction step in SPCA, I will show that SPCA and the geostatistical method give consistent results. But SPCA also has disadvantages as discussed below, and it is therefore necessary to do further clustering by the geostatistical method after SPCA.

Firstly, the SPCA outcomes need to be interpreted, which is similar to the Independent Component Analysis (ICA) method (Thirion, 2003). Since the informative content of the time courses is not considered in SPCA, only the different statistical structures, we still need to look at the autocovariance structure of the time course to determine which components

are of interest. Furthermore, we may have several components which show similar time patterns in SPCA; additional clustering may combine these components together and make different clusters more clear. Secondly, sometimes a few voxels in the SPCA outcomes need to be reassigned. In clustering analysis, any two clusters are mutually exclusive. But when the components from SPCA are not sparse enough, there are overlapped voxels across the different components (one voxel can belong to more than one component). Further clustering can reassign these voxels to the most probable clusters. In my analysis, I will show that using SPCA and geostatistical methods together greatly improves the efficiency of the clustering results. The result is a purely data-driven clustering procedure.

4.2 Methods

4.2.1 Sparse principal component analysis (SPCA)

Principal component analysis (PCA), also known as *empirical orthogonal functions*, is a classic tool for analyzing large scale multivariate data (Jolliffe, 1986). In PCA, the goal is to find a few components that explain a large proportion of the total sample variance of the original variables. The *principal components* can be regarded as the extracted features that maximally separate the individual observation vectors. However, one of the shortcomings of PCA is that each principal component is still a linear combination of all the original variables (Zou et al., 2006; Luss and d'Aspremont, 2006). *Sparse principal component analysis* (SPCA) (Zou et al., 2006) aims at producing easily interpreted components with only a few nonzero coefficients in the principal components, i.e., each new variable is a linear combination of a small subset of the original variables. Hence, SPCA has been a powerful tool in gene expression arrays selection (Zou et al., 2006) and medical shape modeling (Sjostrand et al., 2006).

Definitions in SPCA

Regularized ordinary least squares in regression For a linear regression model, given p predictors $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_p})$, the response $\mathbf{y}$ is predicted by

$$\hat{\mathbf{y}} = \hat{a}_1\mathbf{x}_1 + ... + \hat{a}_p\mathbf{x}_p = \mathbf{X}\hat{\mathbf{a}},$$

where $\mathbf{x}_1$ is defined as $\mathbf{1}$ when $\hat{a}_1$ is considered as the *intercept* (Hastie et al., 2001). *Ordinary least squares* is by far the most popular method in fitting the above regression model, i.e.,

$$\hat{\mathbf{a}}_{ols} = argmin_{\mathbf{a}}||\mathbf{y} - \mathbf{X}\mathbf{a}||^2,$$

where $||\cdot||$ represents the L_2 norm. This is the *best linear unbiased estimator*. However, if small bias is allowed, we can get better estimators by adding the L_2 constraint such that

$$\hat{\mathbf{a}}_{ridge} = argmin_{\mathbf{a}}||\mathbf{y} - \mathbf{X}\mathbf{a}||^2 + \lambda||\mathbf{a}||^2,$$

where $\lambda \geq 0$ is a parameter that gives the best compromise between goodness-of-fit and smoothness. This is called *ridge regression*. As a continuous shrinkage method, ridge regression achieves its better prediction performance through a bias-variance trade-off. Replacing the L_2 norm in the constraint with the L_1 norm gives

$$\hat{\mathbf{a}}_{LASSO} = argmin_{\mathbf{a}}||\mathbf{y} - \mathbf{X}\mathbf{a}||^2 + \lambda||\mathbf{a}||_1,$$

where $||\mathbf{a}||_1 = \sum_{i=1}^{p}|a_i|$. This method is called the *least absolute shrinkage and selection operator* (LASSO) by Tibshirani (1996); it not only shrinks the coefficients, but also enforces some of them to be exactly zero. This carries out a form of variable selection. LASSO has proven to be a very powerful regression and variable selection technique, but it still has a few limitations in some situations. If the number of variables p is greater than the number of observations n, LASSO chooses at most n variables in the analysis. If there is a group of strongly correlated predictors, LASSO tends to choose a single predictor and does not care which one is selected (Zou and Hastie, 2005; Tibshirani, 1996; Sjostrand et al., 2006). A new

method called *elastic net* (EN) regression is developed to overcome these limitations (Zou and Hastie, 2005), where

$$\hat{\mathbf{a}}_{EN} = argmin_{\mathbf{a}}||\mathbf{y} - \mathbf{X}\mathbf{a}||^2 + \lambda_1||\mathbf{a}||_1 + \lambda_2||\mathbf{a}||^2.$$

EN uses a combination of the L_1 norm and L_2 norm to do automatic variable selection and continuous shrinkage simultaneously, and it can select groups of correlated variables.

Regular principal component analysis The regular *principal component analysis* (PCA) takes an $(n \times p)$ matrix $\mathbf{X}$, where we assume we have measurements of some variable at locations $s_1, s_2, ..., s_p$ taken at time points $t_1, t_2, ..., t_n$. So n is the number of observations (time points) and p is the number of variables (locations). We form the covariance matrix of $\mathbf{X}$ by calculating $\mathbf{X}^T\mathbf{X}$ and solve the eigenvalue problem $\mathbf{X}^T\mathbf{X} = \mathbf{B}\mathbf{D}\mathbf{B}^T$ subject to $\mathbf{B}^T\mathbf{B} = \mathbf{I}_k$, where $d_i, i = 1, ..., k$ are the k positive *eigenvalues* and the columns of $\mathbf{B}$ are the *eigenvectors* (*loading vectors*). Let $\mathbf{Z} = \mathbf{X}\mathbf{B}$, $\mathbf{Z}$ will be much smaller than $\mathbf{X}$ if $k << p$ and the dimension of the problem thereby is reduced. This makes $\mathbf{Z}(n \times k)$ and $\mathbf{B}(p \times k)$. The columns of $\mathbf{Z}$ are the *principal components* (*expansion coefficients*), linear combinations of the variables in $\mathbf{X}$. The matrix $\mathbf{Z}$ is called the *score matrix* and the elements of $\mathbf{Z}$ are called *scores*, measuring the position of the observations on the derived axes in $\mathbf{B}$ (Sjostrand et al., 2006).

Regularized principal component regression Sparse principal component analysis focuses on a method for computing eigenvectors by a variable selection method in a "self-contained" *principal component regression* (Zou et al., 2006), in which the data set is regressed on the principal components of itself. If $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$, the ordinary least squares principal component regression,

$$\hat{\mathbf{B}} = argmin_{\mathbf{B}} \sum_{i=1}^{n} ||\mathbf{x}_i^T - \mathbf{B}\mathbf{B}^T\mathbf{x}_i^T||^2,$$

subject to $\mathbf{B}^T\mathbf{B} = \mathbf{I}_k$, where k is the number of components. Since $\mathbf{Z} = \mathbf{XB}$, the above formula can also be expressed as

$$\hat{\mathbf{b}}_j = argmin_{\mathbf{b}}||\mathbf{z}_j - \mathbf{X}\mathbf{b}_j||^2,$$

where $j = 1, ..., k$, $\mathbf{z}_j$ is the j^{th} regular principal component and $\mathbf{b}_j$ is the j^{th} sparse eigenvector. By using ridge regression to introduce some bias in the eigenvectors, large variances due to multi-collinearity of the eigenvectors are reduced. To obtain sparse eigenvectors, the least absolute shrinkage and selection operator (LASSO) method is used together (Tibshirani, 1996). Hence, the *regularized principal component regression*, called SPCA by Zou et al. (2006), becomes

$$\hat{\mathbf{b}}_j = argmin_{\mathbf{b}}||\mathbf{z}_j - \mathbf{X}\mathbf{b}_j||^2 + \lambda||\mathbf{b}_j||^2 + \lambda_1||\mathbf{b}_j||_1,$$

The weakness of this approach is that the results are heavily guided by regular PCA (Sjostrand et al., 2006), thus Zou et al. (2006) proposed an alternative *SPCA criterion*:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = argmin_{\mathbf{AB}} \sum_{i=1}^{n} ||\mathbf{x}_i^T - \mathbf{A}\mathbf{B}^T\mathbf{x}_i^T||^2 + \lambda \sum_{j=1}^{k} ||\mathbf{b}_j||^2 + \lambda_1 \sum_{j=1}^{k} ||\mathbf{b_j}||_1,$$

subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$. This new criterion effectively transforms the PCA problem to a regression-type problem (Zou et al., 2006). According to Sjostrand et al. (2006), this expression can be clarified as follows: firstly, $\mathbf{B}^T\mathbf{x}_i^T$ projects the variables of observation i onto the loading vectors of $\mathbf{B}$; secondly, $\mathbf{A}\mathbf{B}^T\mathbf{x}_i^T$ transforms the scores of $\mathbf{B}^T\mathbf{x}_i^T$ back to the original space by the orthogonal constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$. For proof of the *SPCA criterion* see Zou et al. (2006).

4.2.2 Cross-validation

Cross-validation, also known as *holdout*, is probably the simplest and most widely used technique to compare estimated and true values using only the already known data set (Hastie et al., 2001; Isaaks and Srivastava, 1989). A cross-validation study can help to choose the tuning parameters in SPCA. K-fold cross-validation is commonly used. It has the advantage

that all observations in the data set are eventually used for both training and testing. The data with sample size n are split into K roughly equal sized parts (folds), each containing roughly n/K observations. For the k^{th} part, the model is fitted to the other $K-1$ parts of the data, called the *training data*, and the prediction accuracy is measured on the k^{th} part, called the *testing data*. We do this for $k = 1, 2, ..., K$ and combine the K estimates of prediction error. When $K = n$, the procedure is known as *leave-one-out cross-validation*.

The choice of K depends on the objective and the data set. Typical choices of K are 5 and 10 (Hastie et al., 2001). When $K = n$, cross-validation is approximately unbiased but has high variance, in addition to being computationally intensive. By Hastie et al. (2001), when the number of time points n is greater than 125, 5-fold cross-validation has lower variance and does not suffer from much bias. Since $n = 156$ for the saccade data and $n = 1498$ for the resting data, I only consider $K = 5$ in my analysis. In our setting, for saccade data, the 5 folds are $Z(t_1), ..., Z(t_{31})$; $Z(t_{32}), ..., Z(t_{62})$; ...; $Z(t_{125}), ..., Z(t_{156})$; for resting data, the 5 folds are $Z(t_1), ..., Z(t_{300})$; $Z(t_{301}), ..., Z(t_{600})$; ...; $Z(t_{1201}, ..., Z(t_{1498})$. In cross-validation, the data in each fold are removed in turn and new values are predicted at those time points by the other observations in other folds using SPCA method. Let $\hat{Z}(t_1), ..., \hat{Z}(t_n)$ be the estimates, the Mean Squared Error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}[Z(t_i) - \hat{Z}(t_i)]^2$$

is minimized in the cross-validation.

4.2.3 Steps in SPCA

Step one: Data preprocessing

The necessary preprocessing step in SPCA is to normalize or standardize the raw data in time, because in the absence of normalization, SPCA extracts variables with large variances rather than variables with similar patterns (Rencher, 2002; Baudelet and Gallez, 2003; Goutte et al., 2001; Zou et al., 2006). Given an $(n \times p)$ matrix $\mathbf{X}$, where n is the number

of observations (time points) and p is the number of variables (locations), define $\hat{\mu}_{\mathbf{j}}$ and $\hat{\sigma}^{\mathbf{2}}_{\mathbf{j}}$ as the mean and variance of the j^{th} vector $\mathbf{x_j} = (x_{1j}, ...x_{nj})^T$, where $j = 1, ..., p$; and define a new vector $\tilde{\mathbf{x}}_{\mathbf{j}} = (\tilde{x}_{1j}, ..., \tilde{x}_{nj})^T$, where $\tilde{x}_{ij} = (x_{ij} - \hat{\mu}_{\mathbf{j}})/(\sqrt{n}\hat{\sigma}_{\mathbf{j}})$, $i = 1, ..., n$. Then $\tilde{\mathbf{x}}_{\mathbf{j}}$ is normalized since $\tilde{\mathbf{x}}^{\mathbf{T}}_{\mathbf{j}}\tilde{\mathbf{x}}_{\mathbf{j}} = \sum_{i=1}^{n} \tilde{x}^2_{ij} = 1$. Zou et al. (2006) prove that normalization can ensure the reconstruction of principal components in *principal component regression*. If normalized fitted coefficients are used, the scaling factor does not affect the eigenvectors. This is also explained by Rencher (1998, 2002), because the principal components from the covariance structure are not scale invariant, but the principal components from the correlation structure are scale invariant. Normalization is quite useful for cluster analysis. By Rencher (1998, 2002), we need to be concerned with the units in which variables are measured. If the variances differ widely or if the measurement units are not commensurate, the components just from centered data will be dominated by the variables with large variance. The other variables will contribute very little. However we want to extract variables with similar patterns rather than variables with large variances. In this case, the principal components from normalized data will be more interpretable in the sense that all the variables can contribute evenly (Rencher, 1998 and 2002; Baudelet and Gallez, 2003; Goutte et al., 2001).

Step two: Choose the number of components

In principal component analysis, the number of principal components is often determined by the *scree graph* (Cattell, 1966). To apply the scree method, plot the value of each successive eigenvalue against the rank order and look for a natural break between the "large" eigenvalues and the "small" eigenvalues. The recommendation is to retain those eigenvalues in the steep curve before the first one on the straight line (Rencher, 2002). The plot from this point on is mere "scree", which means "rubbish at the foot of a steep seashore" (Tatsuoka and Lohnes, 1988). The smaller eigenvalues tend to lie along a straight line and just represent random variation.

Step three: Determine the constrained parameter in L_1 norm term

For SPCA calculation, I use the Matlab toolbox of Sjostrand (2005), which is equivalent to Zou et al.'s (2006) R source code. In this toolbox, the L_1 norm term is formulated in equivalent forms where the penalty function enters the equation as a separate constraint:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = argmin_{\mathbf{AB}} \sum_{i=1}^{n} ||\mathbf{x}_i^T - \mathbf{A}\mathbf{B}_i^T\mathbf{x}_i^T||^2 + \lambda \sum_{j=1}^{k} ||\mathbf{b}_j||^2$$

subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$ and $\sum_{j=1}^{k} ||b_j||_1 \leq t$ for some t. Zou et al. (2006) recommend to use cross-validation to find the best constrained value t. But the case is different here. Because fMRI data is ill-balanced, i.e., the activated region only represents a very small proportion of the brain and a large amount of voxels are non "activated" (stimulus-related) (Fadili et al., 2000). Here I want the components from SPCA to be sparse enough to get the most significant patterns in the brain, i.e., dimension reduction. Hence, for my case, I have to use prior knowledge to determine the desired number of nonzero variables (voxels). In my opinion, the choice of constrained parameter is not very strictly, since the goal of the screening is to reduce the large amount of non-stimulus-related voxels that could seriously affect the robustness and sensitivity of the clustering results (Fadili et al., 2000). Actually, we will see below that the explained variances in the components are very small when the number of voxels is greatly reduced. Hence this step is almost equivalent to the dimension reduction step in regular clustering.

Step four: Choose the tuning parameter λ in L_2 norm term

The effect of the parameter λ in the L_2 norm term is generally small (Zou et al., 2006; Sjostrand et al., 2006). But the case is different for us now. Since we have reduced the variables (voxels) to a small number which only explains very little of the total variance, a minor change in λ will greatly affect the results of SPCA. Hence I use 5-fold cross-validation to find the best value for λ.

4.2.4 Steps in further clustering after SPCA

As mentioned in the introduction, since nothing tells us which components are of interest in SPCA, further clustering will typically be needed to discover important clusters. The aim of any clustering method in fMRI is to create a partition of the entire data set into distinct regions (Stanberry et al., 2003), where each region is represented by voxels exhibiting similar temporal behavior, but regions are dissimilar to each other. Hence, in the further clustering after SPCA, I use the autocovariance of time as the main feature.

Step one: Define the new data

For the normalized $(n \times p)$ data matrix $\mathbf{X}$, I formalize the new data after SPCA:

$$\mathbf{Y} = \mathbf{X}\mathbf{b}_1\mathbf{b}_1^T + \mathbf{X}\mathbf{b}_2\mathbf{b}_2^T + ... + \mathbf{X}\mathbf{b}_k\mathbf{b}_k^T.$$

Although the new data $\mathbf{Y}$ is still a $(n \times p)$ matrix, most of the p variables have been forced to be zeros. Hence the dimension of the data is in fact greatly reduced.

Step two: Calculate the autocovariance

For the new $(n \times p)$ data matrix $\mathbf{Y}$, I calculate the empirical autocovariances of the non-zero variables in p at the n time points. For reducing estimation error, I choose effective lag distance of $n - 30$ and discard the number of pairs less than 30 (Journel and Huibregts, 1978). Matlab code from Marcotte (1996) is used for these calculations.

Steps three, four and five

Steps three, four and five are: define the classification metric, use k-means algorithm in the clustering, and determine the final results by silhouette values. I have introduced these steps in Chapter 3.

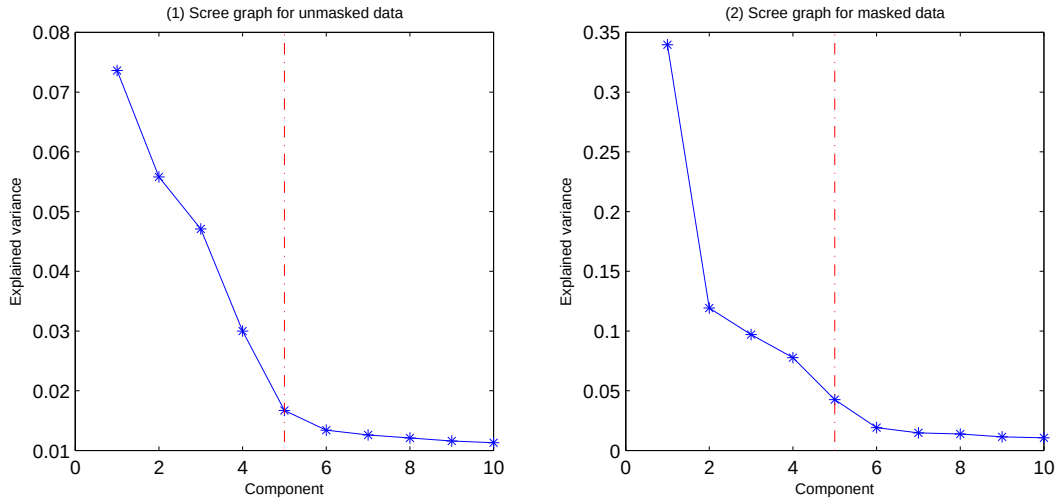

Figure 4.1: Scree graphs for unmasked and masked saccade data. The turning points between the steep curve and the straight line are 5 in both graphs. Hence 5 components are chosen.

4.3 Data analysis

4.3.1 Sparse principal component analysis for saccade data

As in the previous chapter, and for purposes of comparison, I consider here as well the 4^{th} sice of the saccade data.

Unmasked data

For the unmasked saccade data, all 4096 voxels are used in SPCA. The scree graph for the PCA is shown in Figure 4.1, graph (1), Because the turning point between the steep curve

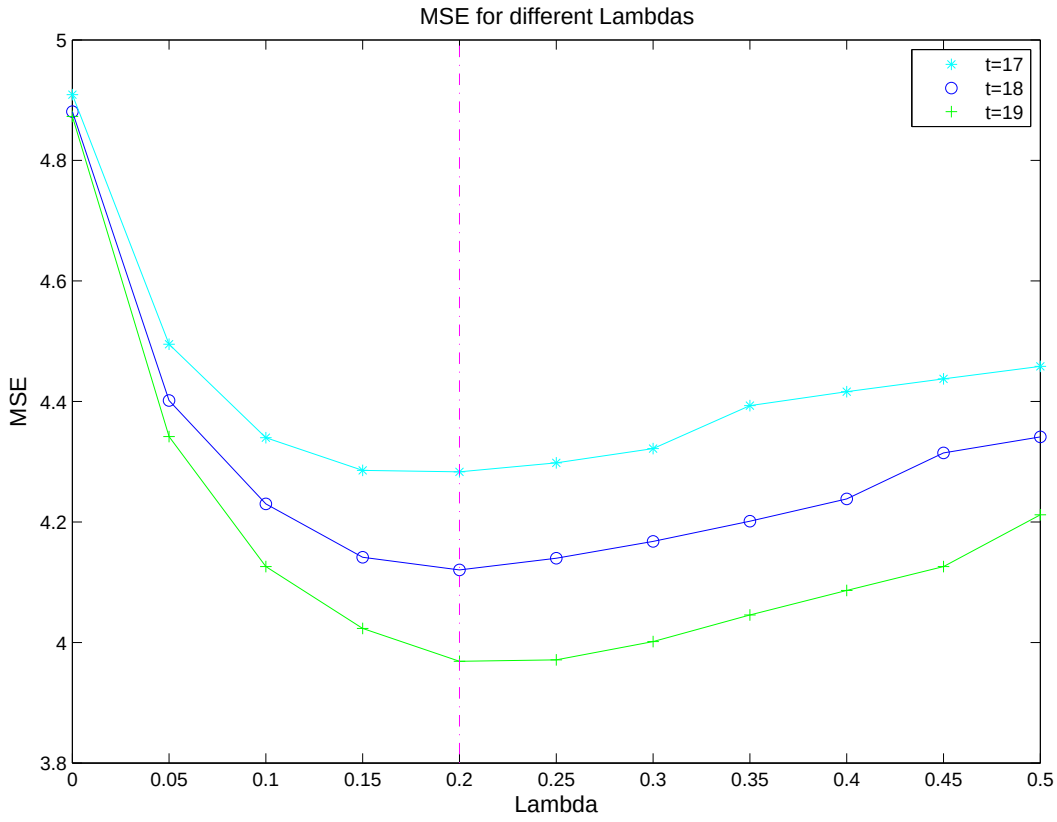

Figure 4.2: Plots of MSE from different constraint parameters t and tuning parameters λ for unmasked saccade data. Minimum MSEs are all at $\lambda = 0.2$. These plots also indicate that the MSE monotonically decreases as the constraint parameter t increases. I consider $t = 18$ and $\lambda = 0.2$ as an example.

and the straight line is 5, 5 components are chosen. Figure 4.2 shows the MSE of SPCA for constrained parameter $t = 17, 18, 19$ with different choices of the tuning parameter λ. These plots at different ts are not crossed, indicating that the choice of t is just used for dimension reduction. As mentioned before, because the goal of the screening is to reduce the large amount of non-stimulus-related voxels that could seriously affect the robustness and sensitivity of the clustering results (Fadili et al., 2000). The choice of t is not very strictly and depends on prior knowledge of the data or the needs of the researcher. In order to compare with results from the previous chapter, I consider $t = 18$ as an example here, which reduced the total number of voxels to fewer than 400. The tuning parameter in L_2 norm term is chosen as 0.2 by 5-fold cross-validation where it has the minimum MSE (Figure 4.2).

Figure 4.3 shows the clustering results from the SPCA method. The number of the first 5 non zero loadings are 102, 86, 48, 59, 46 voxels respectively, and the 4^{th} and 5^{th} loadings have one overlapped voxel. Hence there are 340 voxels in total (Figure 4.3). These 5 components explain only 2.97% of the total variance, indicating the dimension of the data is greatly reduced. The graphs (1) and (3) in Figure 4.3 show apparently "true" brain activation, "head motion" patterns in the first two components, consistent with the results from the previous chapter. I also do further clustering to double check the results. The graphs of silhouette values are shown in Figure 4.4. All the silhouette values using the "correlation" metric are greater than 0.80, which means strong structures have been found (Kaufman and Rousseeuw, 1990). Graph (5) ($k = 6$) in Figure 4.4 has the highest mean of silhouette values 1.0000, where the numbers of voxels in the 6 clusters are 102, 86, 48, 58, 45, 1. Cluster 6 extracts the overlapped voxel from components 4 and 5 in the SPCA analysis. Figure 4.4 (Graph (4)) also shows the second best structure is 5 clusters, where the numbers of voxels are 102, 86, 49, 58, 45 respectively. Now the overlapped voxel is reassigned to the most probable cluster. Hence the best number of clusters could be 5 or 6. These results are consistent with those from SPCA. For k=6, cluster 1 has 102 voxels, we can see very strong peaks and troughs in the shape of waves, showing periods of time correlations corresponding to the experimental

design (Figure 4.5, graph (2)), cluster 2 has 86 voxels, we also can see some peaks and troughs in the shape of waves, showing weak periods of time correlations corresponding to the experimental design (Figure 4.5, graph (3)). Clusters 3, 4 and 5 show some overlapping cyclic patterns, which do not correspond to the experimental design (Figure 4.5, graphs (4), (5), (6)). Cluster 6 shows no pattern. (Figure 4.5, graph (7)). Figure 4.6 shows the means of the original time courses of the six clusters. Note that the time patterns in the original courses are not as clear as in the autocovariances. Hence the use of the autocovariances of the time courses helps in the interpretation of the SPCA outcomes.

Based on the time patterns of the clusters, k=3 would also be a reasonable choice (Figure 4.4, graph (3)) since it has the third largest mean silhouette value. The clusters of "true" brain activation and "head motion" would be the same as before, and the other four "noise" clusters would be combined to form one. Hence, doing the clustering again provides a chance to double check the results from SPCA and enhance interpretation with a scientific basis. Figure 4.7 shows the final results of the unmasked method with three clusters. After clustering, I mask the average brain images over time and discard all the voxels outside the brain, leaving 236 voxels inside the brain of the 340 retained voxels; 83 for "activation", 45 for "head motion", and 108 for unknown "noise".

MSE	$\lambda = 10^{-6}$	$\lambda = 0.05$	$\lambda = 0.10$	$\lambda = 0.15$	$\lambda = 0.20$	$\lambda = 0.25$	$\lambda = 0.30$
$t = 11$	0.00205	0.00179	0.00176	0.00170**	0.00173	0.00177	0.00193
$t = 12$	0.00178	0.00127**	0.00133	0.00143	0.00147	0.00147	0.00153

Table 4.1: MSE from different constraint parameter t and tuning parameter λ for masked saccade data,"**" indicates the minimum MSE.

Masked data For the masked saccade data, only 630 retained voxels inside the brain are used in SPCA. The scree graph is shown in Figure 4.1, graph (2). Because the turning point between the steep curve and the straight line is five, five components are again chosen.

For comparing to previous results, I consider two choices in the dimension reduction step, i.e., $t = 11$ and $t = 12$. When $t = 11$, the number of voxels is reduced to around 250, which is used to compare with the masked data in the previous chapter; when $t = 12$, the number of voxels is reduced to around 400, which is used to compare with the unmasked data. In method one, I control the constrained parameter t in the L_1 norm term as 11, where the best choice for the tuning parameter λ in the L_2 norm term is 0.15 (Table 4.1). In SPCA, the numbers of the 5 non zero loadings are 59, 49, 47, 56 and 46 voxels (Figure 4.8). There are no overlapped voxels, and the 257 retained voxels explain 14.6% of the total variance. Further clustering shows the mean silhouette value equals to 1 when $k = 5$, where the results are exactly the same as those in SPCA. Similarly, the brain activation cluster has 46 voxels; the "head motion" cluster has 59 voxels, and the three noise clusters have 152 voxels in total.

In method two, I control the constrained parameter t in the L_1 norm term as 12, where the best choice for the tuning parameter λ in the L_2 norm term is 0.05 (Table 4.1). It is seen that the maps of the 5 components are very similar to those with $t = 11$ (Figure 4.8), except the numbers of the 5 non zero loadings increase to 90, 84, 70, 82 and 95 voxels respectively (Figure 4.9). The total number of voxels is 406 and they explain 20.29% of the variance; notice that the percentage of explained variance increases as the number of voxels increases. There are 15 overlapped voxels among the 5 components and 40% of them are overlapped between the 1^{st} component and the 5^{th} component. The 1^{st} component indicates "head motion" and the 5^{th} component indicates brain activation, hence the overlap reveals a confusion between "head motion"and brain activation. Further clustering reassigns the overlapped voxels to the most probable clusters. After reassigning, the number of voxels corresponding to the previous 5 components became 87, 83, 69, 79, 88 without overlapping.

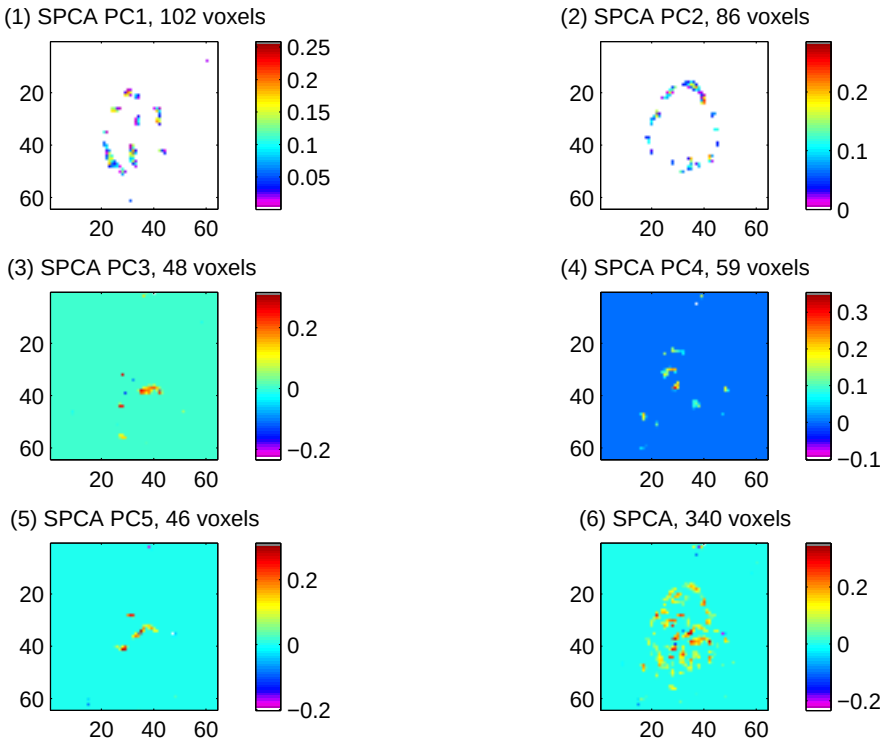

Figure 4.3: Maps of five SPCA components for unmasked saccade data, with constrained parameter t=18 and tuning parameter $\lambda = 0.2$. The first 5 components only retain 340 voxels. The numbers of the 5 non zero loadings are 102, 86, 48, 59, 46 voxels respectively, and the 4^{th} and 5^{th} loadings have one overlapped voxel.

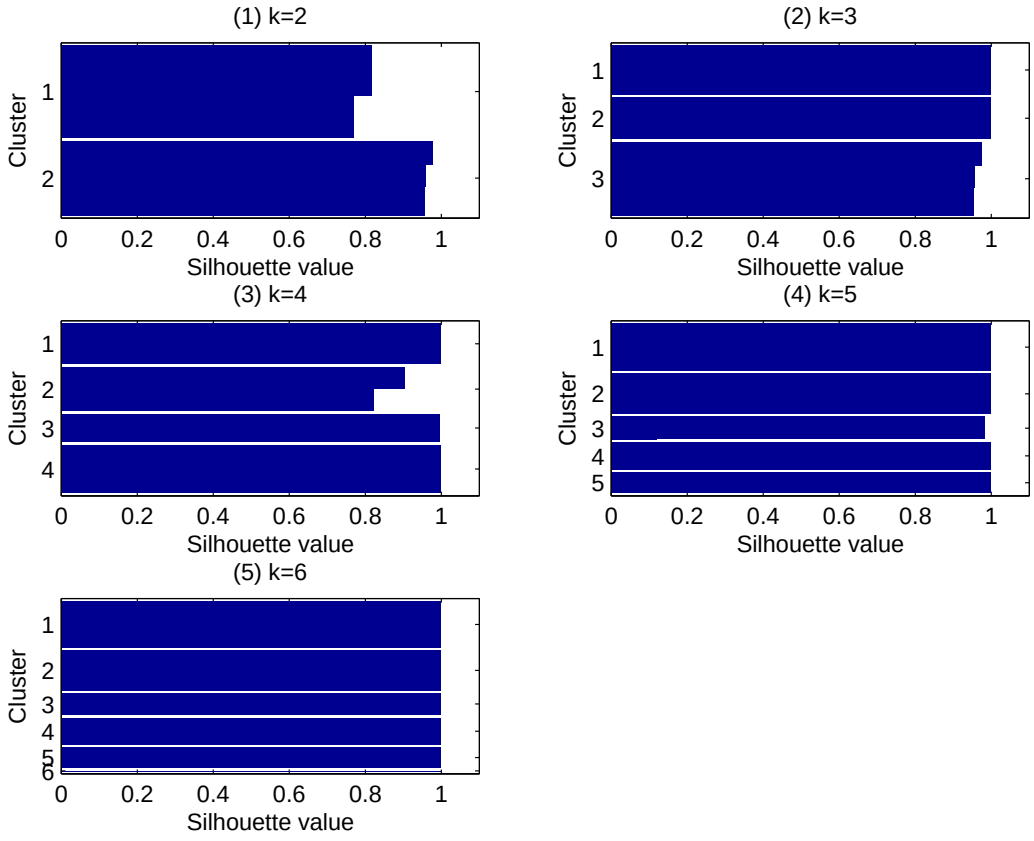

Figure 4.4: Silhouette values for unmasked saccade data with different numbers of clusters using "correlation" metric. (1), (2), (3), (4), (5) are the silhouette values in "correlation" metric with $k = 2, 3, 4, 5, 6$. The mean silhouette values are 0.8707, 0.9830, 0.9604 0.9952, 1.0000 respectively. The selected number of clusters k could be 3, 5 or 6. See text for explanation. Results are consistent with those from SPCA.

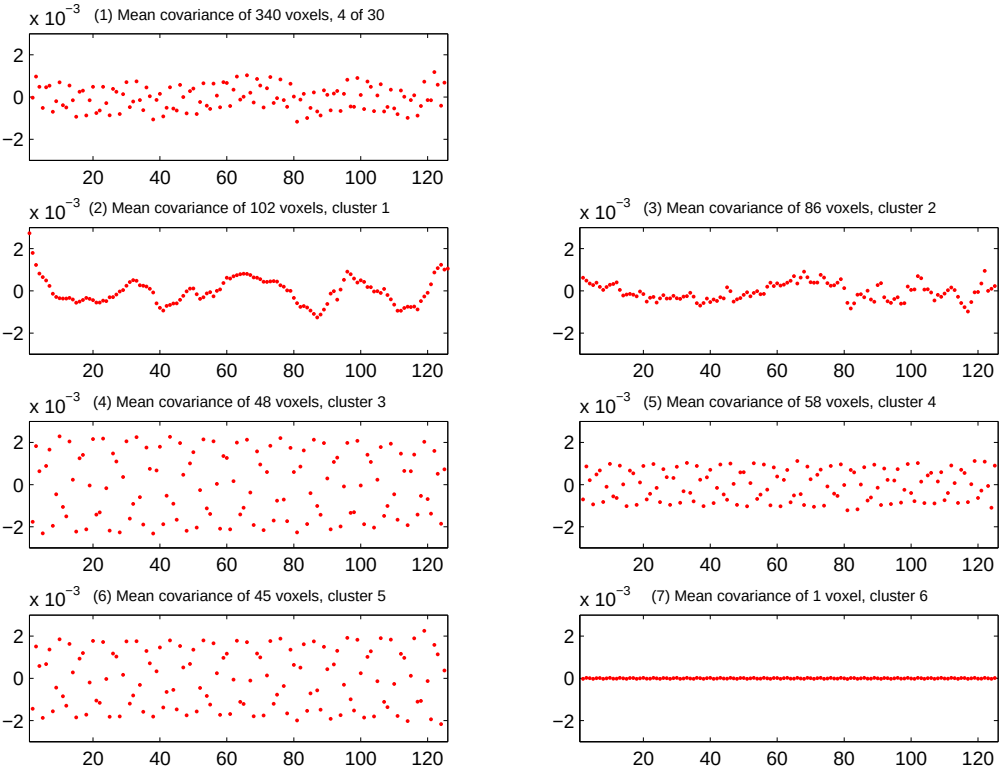

Figure 4.5: Time patterns for unmasked saccade data. (1) is the mean covariance of the 340 retained voxels. (2), (3), (4), (5), (6), (7) are the mean covariances of the six clusters. Cluster 1 shows strong peaks and troughs in the shape of waves corresponding to the experimental design. Cluster 2 shows weak peaks and troughs in the shape of waves corresponding to the experimental design. Clusters 3, 4 and 5 show overlapping cyclic patterns which do not correspond to the experimental design. Cluster 6 shows no pattern.

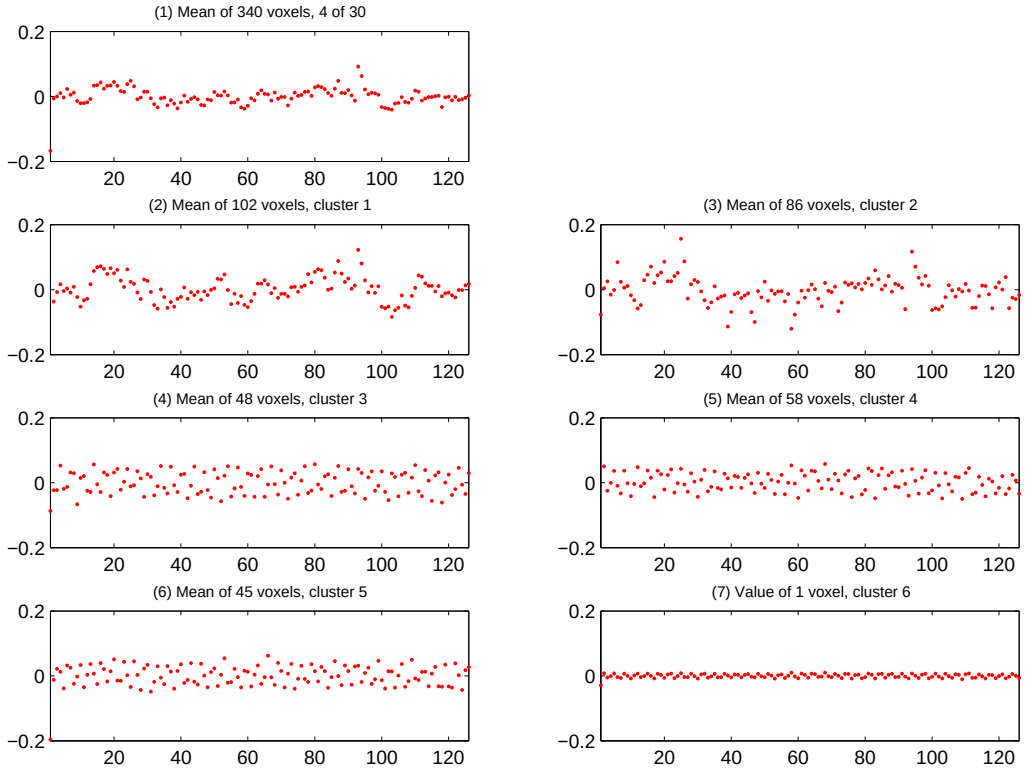

Figure 4.6: Time patterns for unmasked saccade data. (1) is the mean of the 340 retained voxels. (2), (3), (4), (5), (6), (7) are the means of the six clusters. The time patterns by their original time courses are not as clear as those by their autocovariances in Figure 4.5.

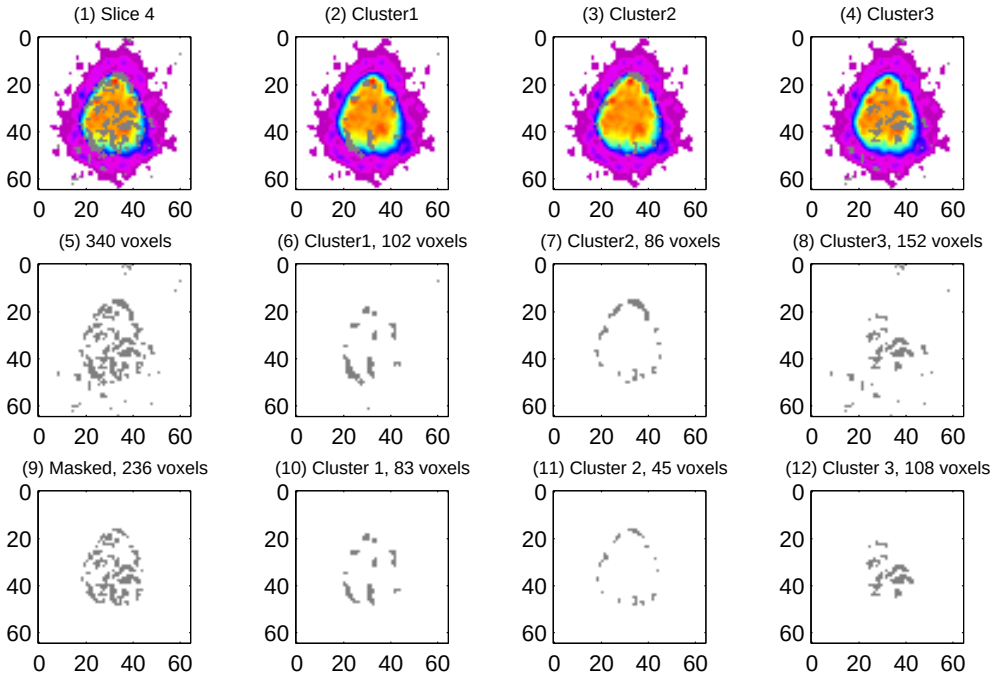

Figure 4.7: Maps of the brain for saccade data, unmasked method. (5) and (1) are maps of 340 voxels and overlaid on the original brain. (6) and (2) are maps of cluster 1 and overlaid on the brain. (7) and (3) are maps of cluster 2 and overlaid on the brain. (8) and (4) are maps of cluster 3 and overlaid on the brain. (9) is the map of 236 voxels after masking. (10), (11), (12) are the maps of the clusters after masking. There are 102, 86, 152 voxels in the three clusters before masking and 83, 45, 108 voxels left after masking.

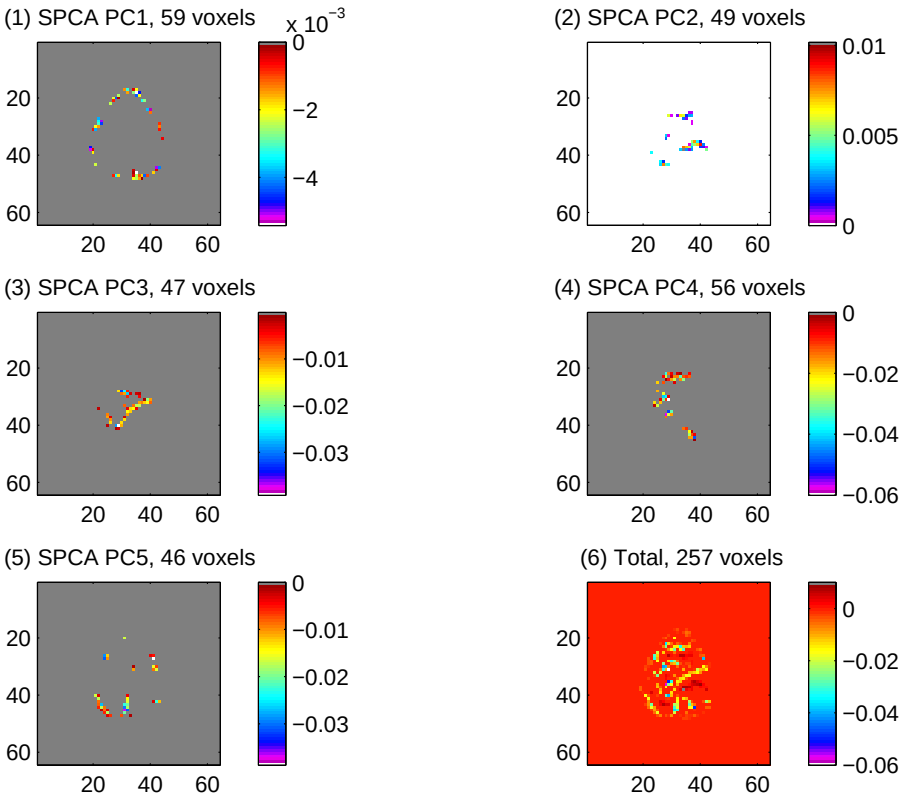

Figure 4.8: Maps of the five SPCA components for masked saccade data in method one, where constrained parameter t=11 and tuning parameter $\lambda = 0.15$. The first 5 components retain 257 voxels of the original 630. The number of the 5 non zero loadings are 59, 49, 47, 56, 46 voxels respectively, and there are no overlapped voxels among the components.

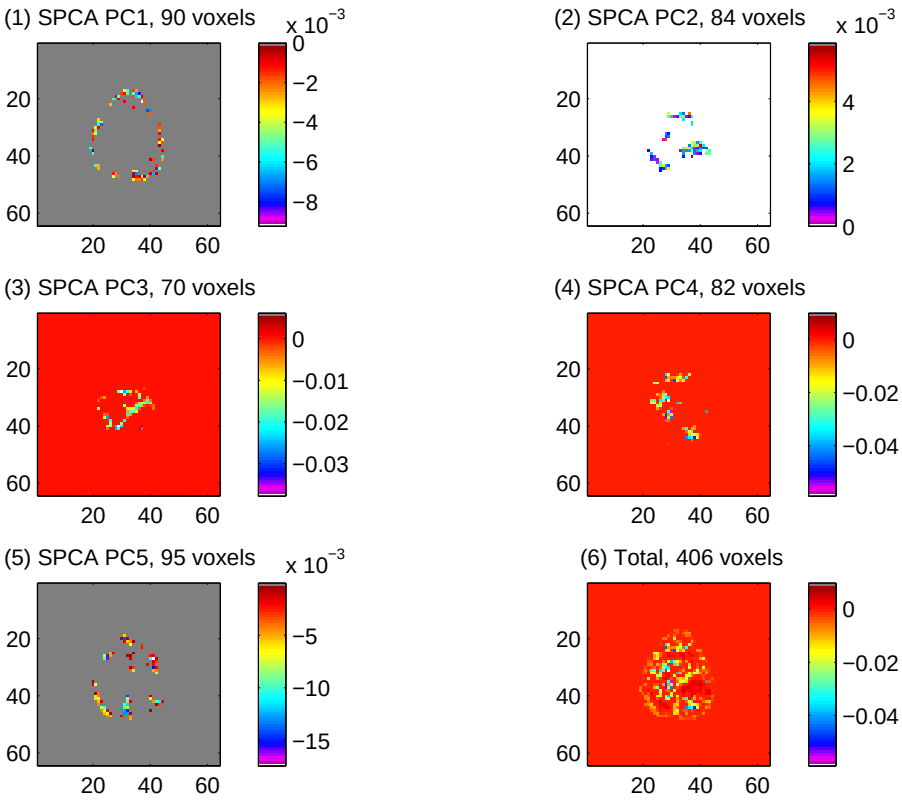

Figure 4.9: Maps of the five SPCA components for masked saccade data in method two, where constrained parameter t=12 and tuning parameter $\lambda = 0.05$. The first 5 components retain 406 voxels of the original 630. The number of the 5 non zero loadings are 90, 84, 70, 82, 95 voxels respectively. There are 15 overlapped voxels among the 5 components and 40% of them are overlapped between the 1^{st} component and the 5^{th} component.

4.3.2 Sparse principal component analysis for masked long range resting data

For comparison with previous results, I again consider the first slice of the resting data.

Similarly, the scree graph for the PCA is executed at first. By looking at the scree graph, I find the turning point between the steep curve and the straight line is not as clear as before, with 2 and 6 both being plausible values (Figure 4.10). Hence, I consider 6 components in method one, and choose 3 components in method two for comparison.

The reason to choose 3 components instead of 2 components in method two is: if the most probable answer for the number of cluster is 2, the further clustering in method one will reassign 6 components to 3 clusters, 2 of them are almost the same as the first 2 components in method two, and the other 4 of them are combined into the 3^{rd} component. Hence the two methods will get consistent results.

Method one In method one, the constrained parameter in the L_1 norm term determines the effective voxels in the brain; I choose $t = 16$ as an example. The tuning parameter λ in the L_2 norm term is determined by 5-fold cross-validation. By Table 4.2, for $t = 16$, $k = 6$, MSE is minimized when $\lambda = 0.05$.

The first 6 components have 718 voxels and explain 13.79% of the total variance. The number of the 6 non zero loadings are 138, 77, 152, 263, 130 and 102 voxels, and there are 144 loadings overlapped among the 6 components. The graphs of SPCA do not show any clear patterns, perhaps due to the overlapped voxels. It is necessary to do further clustering, and the means of silhouette values in the "correlation" metric for k=2, 3, 4, 5, 6, 7 are shown in Figure 4.12. The best choice is $k = 3$ since it has the largest mean of silhouette values (Figure 4.12, graph (2)). Maps of the resulting three clusters are shown in Figure 4.11 (Graphs (7), (8), (9)). Cluster 1 has 126 voxels, and is very similar to component 3; cluster 2 has 514 voxels, and is very similar to the combination of components 1, 4, 5, 6; cluster 3 has 78 voxels, and is very similar to component 2.

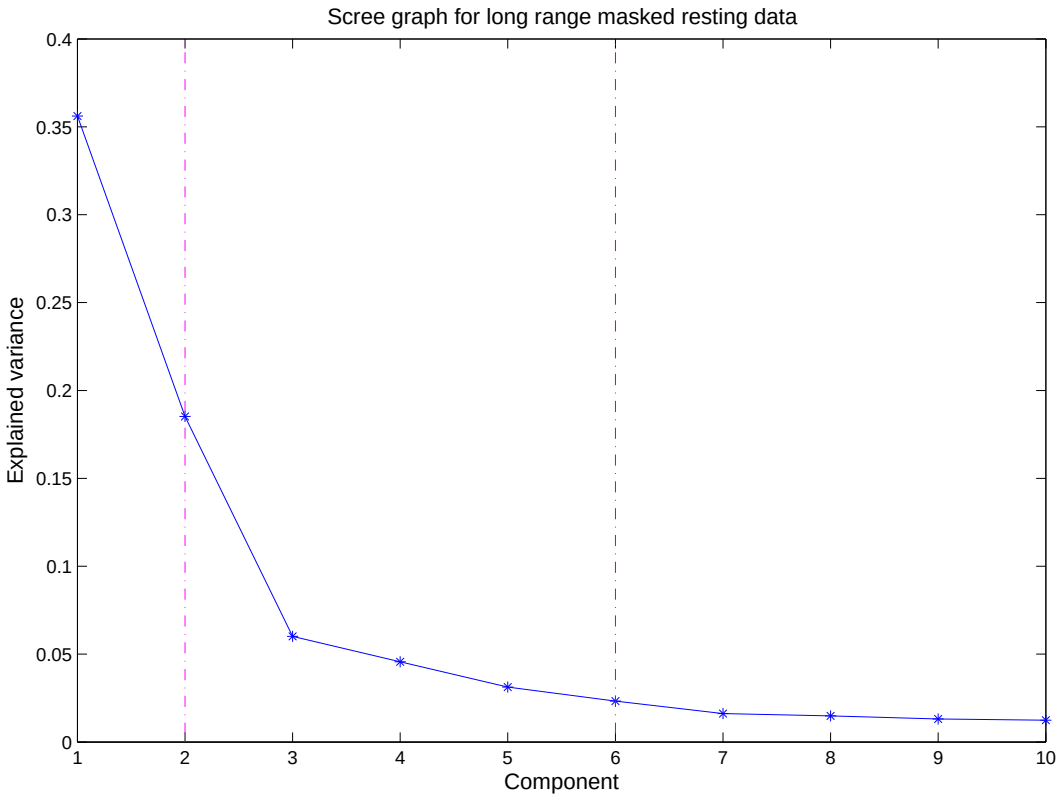

Figure 4.10: Scree graph for the resting data, the turning point between the steep curve and the straight line could be 2 or 6. I consider 6 components in method one; and choose 3 components in method two for comparison.

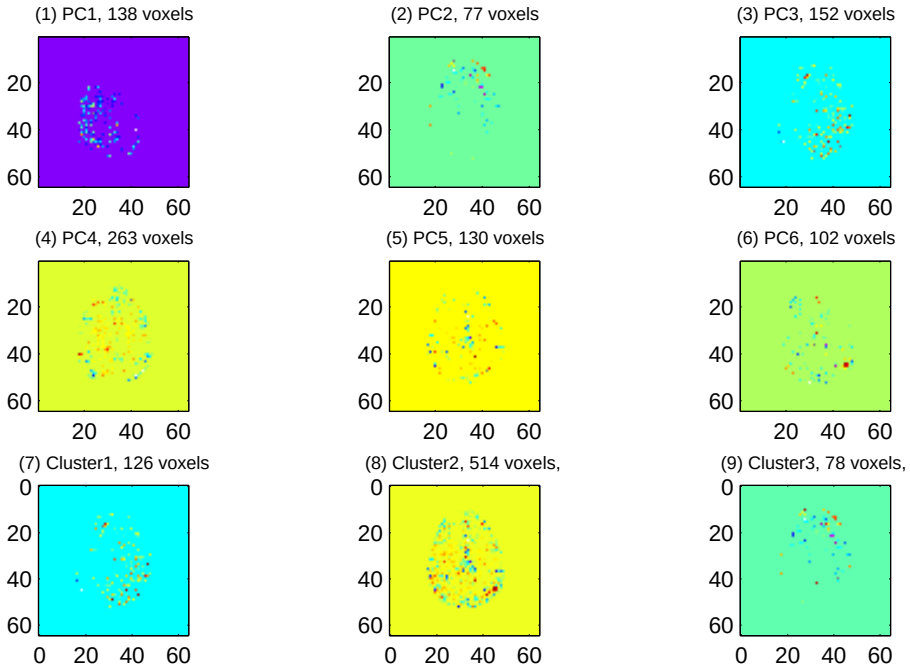

Figure 4.11: Maps of the brain in method one for the resting data. (1)-(6) are the results of SPCA, where the number of components k=6, constrained parameter t=16 and tuning parameter $\lambda = 0.05$. The first 6 components have 718 voxels in total. The number of the 6 non zero loadings are 138, 77, 152, 263, 130 and 102 voxels respectively, and there are 144 loadings overlapped among the 6 components. (7), (8), (9) are the results after further clustering. There are no overlapped voxels among the three clusters. (7) is cluster 1 with 126 voxels, which is very similar to component 3; (8) is cluster 2 with 514 voxels, which is very similar to the combination of components 1, 4, 5, 6; (9) is cluster 3 with 78 voxels, which is very similar to component 2.

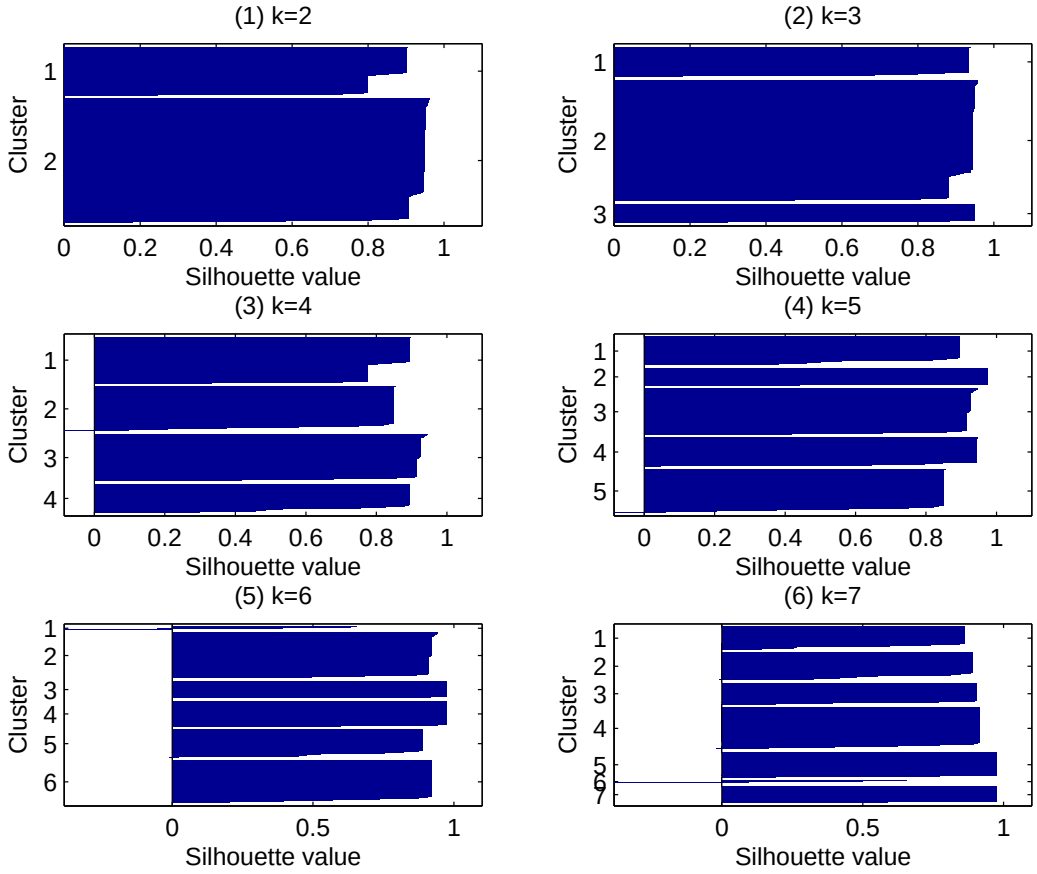

Figure 4.12: Silhouette values for the resting data. For $k = 2, 3, 4, 5, 6, 7$, the means of the silhouette values in the "correlation" metric are 0.9076, 0.9165, 0.8402, 0.8641, 0.8766, 0.8467 respectively. The best choice is $k = 3$.

Method two Since we only see some weak time patterns in cluster 1 in the method one (Fig 4.13), I think the best choice for the number of components might be 2. In method two, I choose the number of components to be $k = 3$ and do the same analysis as before for comparison. The parameters are chosen as $k = 3$, $t = 16$ and $\lambda = 0.05$ (Table 4.2). The results are shown in Figure 4.14. The first 3 components have 149 voxels and explain 5.66% of the total variance. The number of the 3 non zero loadings are 62, 45, 42, voxels, respectively and there are no loadings overlapped. I also combine the 3 components and again do further clustering. Results are exactly the same as in SPCA with mean of silhouette values equal to 1. The different temporal patterns of the three clusters are shown in Figure 4.15. Compared to the results in method one (Figure 4.13), we can see the time patterns of the three clusters are very similar. Although what gets called "cluster 1", "cluster 2", "cluster 3" differs from method to method. Hence, after further clustering, I conclude that the number of components for the resting data might be 2.

MSE	$\lambda = 10^{-6}$	$\lambda = 0.05$	$\lambda = 0.10$	$\lambda = 0.15$	$\lambda = 0.20$	$\lambda = 0.25$	$\lambda = 0.30$
$k = 6$	0.6907	0.6606**	0.6702	0.7001	0.7280	0.7488	0.7648
$k = 3$	0.9206	0.9163**	0.9211	0.9236	0.9275	0.9306	0.9330

Table 4.2: MSE from different number of components k and tuning parameter λ with constraint parameter t=16 for resting data,"**" indicates the minimum MSE.

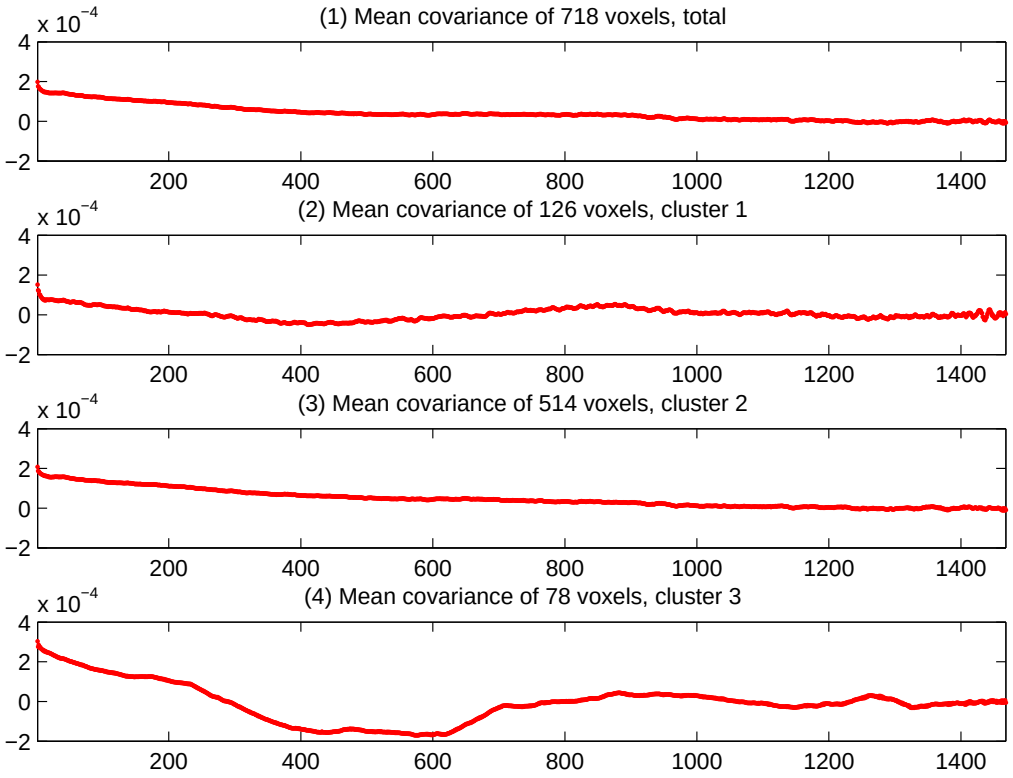

Figure 4.13: (1) is the mean covariance of the 718 voxels in method one for resting data. (2), (3), (4) are the mean covariances of the three clusters.

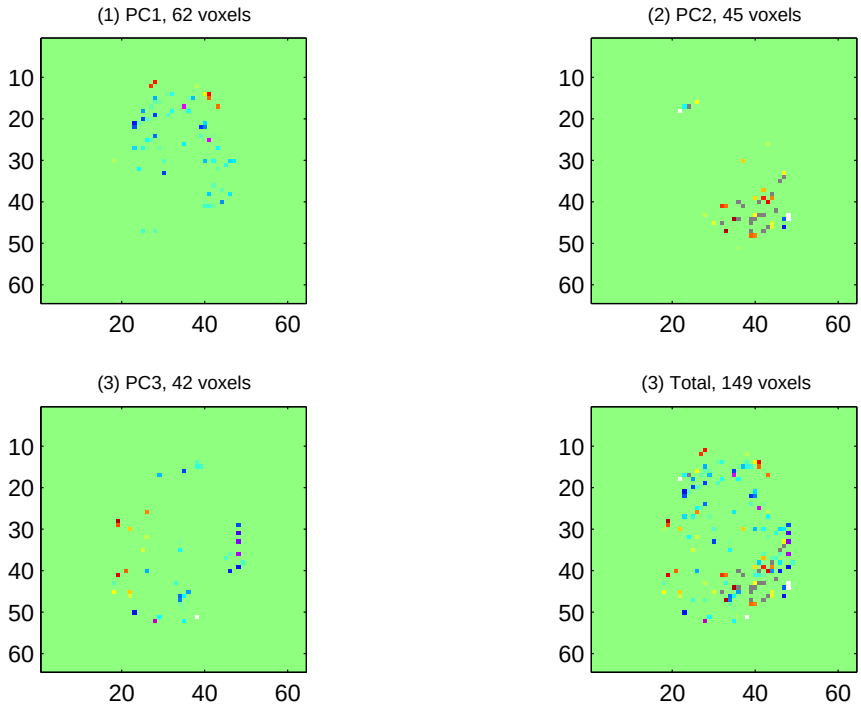

Figure 4.14: Maps of the brain in method two for resting data, where the number of components k=3, constrained parameter t=16 and tuning parameter $\lambda = 0.05$. The first 3 components have 149 voxels and explain 5.66% of the variance. The numbers of the 3 non zero loadings are 62, 45, 42 voxels, and there is no overlap.

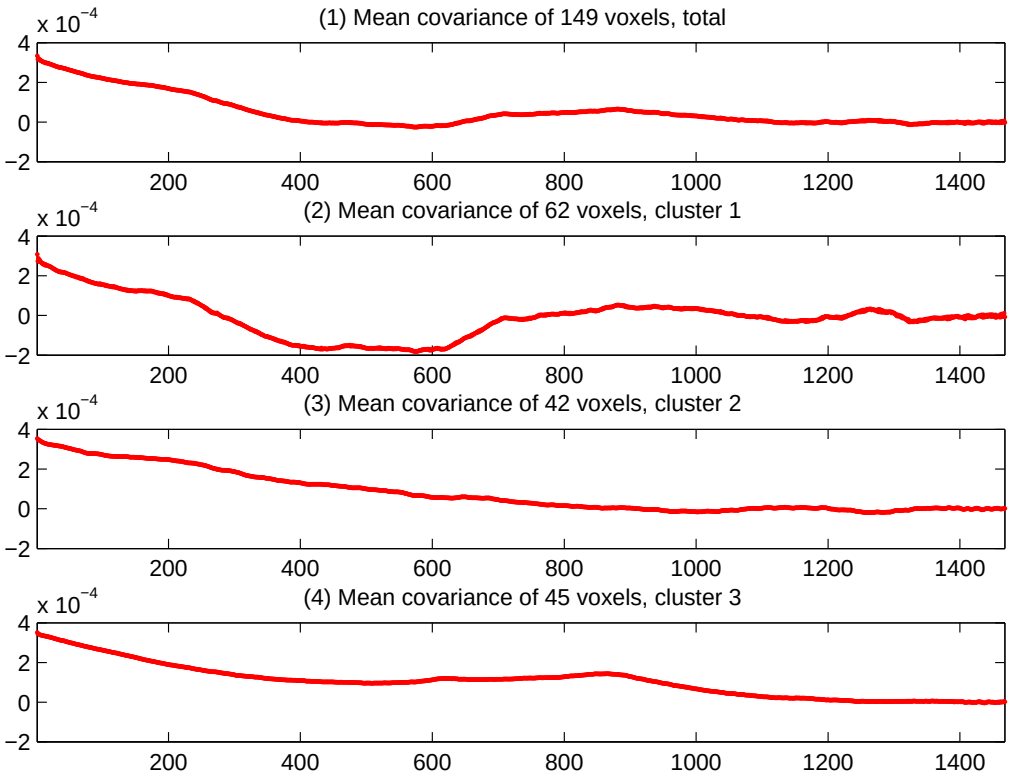

Figure 4.15: (1) is the mean covariance of the 149 voxels in method two for resting data. (2), (3), (4) are the mean covariances of the three clusters. Compared to the results in method one (Figure 4.13), we can see the time patterns of the three clusters are very similar. Although what gets called "cluster 1", "cluster 2", "cluster 3" differs from method to method.

Clusters	"Activation"	"Head motion"	"Noise"	Total
Method A	83 voxels	45 voxels	108 voxels	236 voxels
Method B	46 voxels	59 voxels	152 voxels	257 voxels
Method C	86 voxels	87 voxels	231 voxels	406 voxels
Overlaps in A and B	42 voxels	36 voxels	63 voxels	144 voxels
Overlaps in A and C	62 voxels	42 voxels	98 voxels	208 voxels
Overlaps in B and C	45 voxels	58 voxels	134 voxels	241 voxels

Table 4.3: Classified voxels inside the brain in different methods for saccade data. Method A indicates the unmasked method which uses the original data to do the analysis directly. Method B indicates the masked method one which uses the masked data with constrained parameter $t = 11$. Method C indicates the masked method two which uses the masked data with constrained parameter $t = 12$.

4.4 Discussion and conclusion

4.4.1 Discussion

Comparisons of different methods

By the results from the saccade data and the resting data, we can see that when components of SPCA are sparse enough, the results in further clustering are exactly the same as those in SPCA, showing the consistency of SPCA and geostatistical analysis. The reason lies in that both methods consider the correlations among the time courses. When SPCA components are not sparse, further clustering by geostatistical analysis can reassign the overlapped voxels to the most probable clusters. The analysis in the resting data also shows that when the number of components is mis-determined in SPCA, further clustering by geostatistical analysis can well interpret the results from SPCA and find the most probable cluster allocation.

Masking the brain

Table 4.3 and Figure 4.16 show the results of the three different approaches for the saccade data. The unmasked method uses the original data for the analysis and masks the clusters at the end. Masked methods one and two use the same masked data to do the analysis directly, but with different constrained parameters t chosen for dimension reduction. By comparing the "activation" graphic results with the maps from the previous chapter, it seems that results from the unmasked method are closer to what has been found by other researchers.

Unmasked method and masked method one have a similar number of voxels inside the brain, but the numbers of voxels in the clusters are different (Table 4.3). In masked method one, we can see the number of voxels in "activation" cluster decreases, and the numbers of voxels in "head motion" and "noise" increase compared to the unmasked method. Masking the data at first does not improve the results of clustering. To double check this conclusion, I also use masked method two with dimension reduction which increases the number of voxels in "activation" and "head motion". Although masked method two has a similar number of voxels in "activation" as that in unmasked method, there are 62 overlapped voxels only and more noise is included (Figure 4.16, graph (3)). This phenomenon has been explained by the "marginal effect" in the previous chapter.

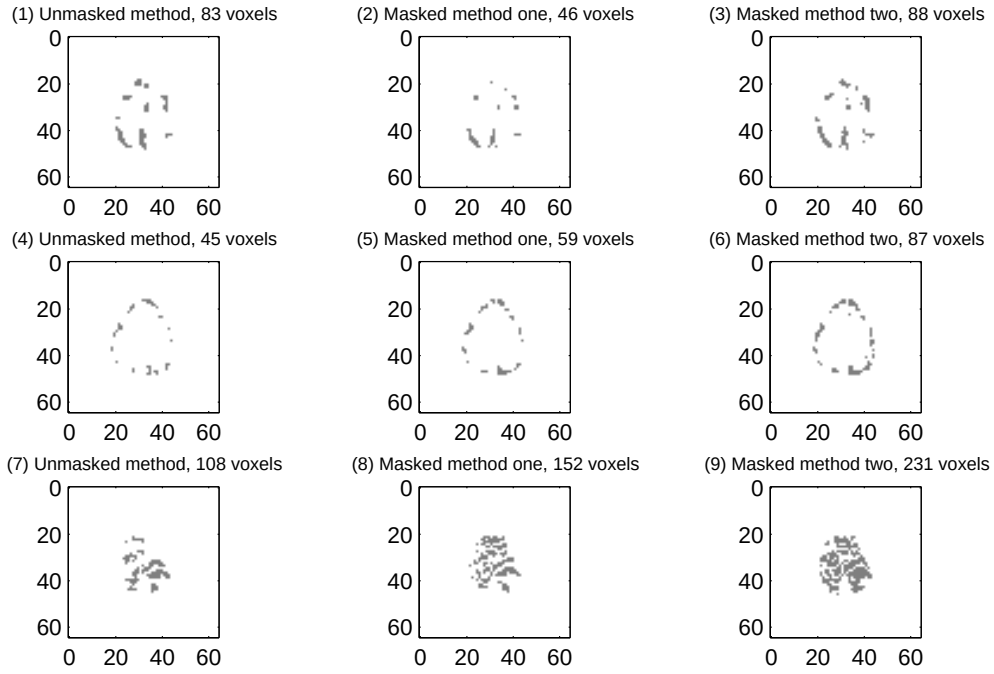

Figure 4.16: Classified voxels inside the brain in different methods for saccade data. (1), (4), (7) are "activation", "head motion", "noise" in unmasked method. (2), (5), (8) are "activation", "head motion", "noise" in masked method one. (3), (6), (9) are "activation", "head motion", "noise" in masked method two. There are 42 overlapped voxels between (1) and (2); 62 overlapped voxels between (1) and (3); 45 overlapped voxels between (2) and (3). There are 36 overlapped voxels between (4) and (5); 42 overlapped voxels between (4) and (6); 58 overlapped voxels between (5) and (6). There are 63 overlapped voxels between (7) and (8); 98 overlapped voxels between (7) and (9); 134 overlapped voxels between (8) and (9).

4.4.2 Conclusion

We present and analyze two clustering methods and demonstrate their uses in fMRI data analysis. The results of SPCA, which combines dimension reduction, feature extraction, clustering together are encouraging. It can be used in resting data for dimension reduction and overcome the limitation of the traditional 2-sample t-test. But SPCA still has a few small disadvantages: First, the explained variance in each component is so small that we can not determine the importance of the components by explained variance alone. Hence the component needs to be interpreted by its autocovariance structure over time. Second, although the different components from SPCA are sparse enough, sometimes they are not mutually exclusive. The overlapped voxels across the different components need to be reassigned by further clustering. Third, we have to decide the number of components, constrained parameter and tuning parameter in advance. Sometimes it is hard to make the decision and the first decision might not be the best choice. Hence, it is necessary to consider further clustering after SPCA. The use of autocovariance in the further clustering is also attractive since it captures important characteristics of the voxels over time. Unlike conventional "correlation" methods, the proposed two *data-driven* methods don't require prior knowledge about any reference functions, can effectively identify regions of similar activations and provide consistent results. The analysis also provides evidence that masking the brain may affect the clustering results. Although many researcher often choose masking the brain as the first step to reduce the dimension, I have shown that this is not necessarily effective and sometimes results in less convincing inference, especially when the data have "head motion" evident around the edges of brain. This conclusion is consistent with that in the previous chapter.

Chapter 5

Structural analysis in fMRI data

Having focused until now on the temporal properties of fMRI, I change my view in this chapter to spatial analysis of the fMRI data. I examine the specific characteristics of fMRI data and try to find a suitable variogram model to describe the true spatial structure of the data.

5.1 Introduction

In detecting activation inside the brain, understanding the structure of the data is a necessary first step. It is particularly important to consider the spatial structure of the data in an analysis of spatially varying regions inside the brain. Recently, some researchers have begun to apply geostatistical ideas in analyzing the spatial structure of the brain. Spence et al. (2007) use a Gaussian variogram model to find neighbors of voxels of interest, and show such spatial analyses can identify regions of the brain that exhibit statistically significant group differences. But their choice of variogram model is quite arbitrary. They select the Gaussian model from a limited set of monotonic increasing variogram models because it has a positive nugget effect. Also, the selection only considers the physical neighbors in the voxels of interest but different regions in the brain may be functionally related, which means the spatial correlations between voxels do not necessarily decrease with their increasing physical distance. Bowman (2007) notices these functional relations inside the brain. He considers an alternative monotonically increasing variogram model, which changes the lag distance between two voxels from the location distance to the voxel value distance (the difference in the measured signal between two voxels). But his fitting still has some problems. The usual

approach in spatial statistics is to ignore an error in location and to assume that error in voxel value is additive or perhaps multiplicative (Stein, 1999). Voxel value distance is not as precise as location distance since it contains measurement error. Hence Bowman uses empirical values from unspecified auxiliary data. Geostatistics gives more choices in model selection and model inference for fMRI. When a variogram shows a cyclical pattern with a "down-hole", its structure is called *hole effect* (Journel and Huijbregts, 1978). Here I will instead use the structural classification idea in geostatistics to analyze the characteristics of the brain by looking at empirical variograms of fMRI data, and use a variogram model with hole effect structure to fit the spatial structure of the brain. This method will consider both the physical and functional relations inside the brain and does not need any auxiliary data. Hole effect structural analysis is a relatively unexplored area in geostatistics, because in the mining industry, the fluctuations in the empirical variogram are usually considered as more or less random noise (Webster and Oliver, 2001). But in fMRI data, since we know that the fluctuations in the variogram may be due to correlations among different regions of the brain, the hole effect structure should not be ignored.

5.2 Concepts and methods

5.2.1 Assumptions in analysis

Structural analysis, i.e., covariance and variogram analysis, is the first and indispensable step in geostatistics, either for kriging or for classification. It is a procedure for characterizing the structural information of the *regionalized variables* (Journel and Huijbregts, 1978).

For a specific slice of the fMRI image, the measured signals are obtained over a regular grid of voxels, i.e. $\{z(s_i) : s_i \in D\}$, where $z(s_i)$ is a particular voxel value at location s_i, D is the set of the region of interest. Voxel values contribute the intensity of the fMRI image (Wynn, 2000).

The direct study of $\{z(s_i) : s_i \in D\}$ is a *nonparametric approach*, using a fixed deterministic function to fit the spatial surface as "closely" as possible, but it does not take account

of the spatial relations between different $z(s_i)$s (Journel and Huijbregts, 1978) because the structure of the fitting model is fixed.

In geostatistics, the set of $z(s_i)$s are defined as *regionalized variables* and $\{z(s_i) : s_i \in D\}$ is considered as a particular realization of a random process $\{Z(s_i) : s_i \in D\}$ (Cressie, 1993; Webster and Oliver, 2001); this is called the *random function approach*. In this approach, two random variables $Z(s_i)$ and $Z(s_j)$ tend to be correlated and their relation depends on the nature of the variable considered and the vector distance between locations s_i and s_j (Journel and Huijbregts, 1978). Hence geostatistics theory considers that the variabilities of all *regionalized variables* have a particular structure (Journel and Huijbregts, 1978). The exact data values are only samples of a realization $\{z(s_i) : s_i \in D\}$ from a random function $\{Z(s_i) : s_i \in D\}$. The advantage of this approach is that we only need to characterize the main features of the random function and not the particular realizations $z(s_i)$ (Wackernagel, 2003). Under the second order stationarity assumption, we can use covariance or variogram to describe quantitatively how the spatial variability of the regionalized voxel values can be characterized in the brain.

5.2.2 Two types of distances

By the assumption of second order stationarity, the variogram defines the relationship between the variability and the lag distance. Hence variance is a function of lag alone. Understanding the characteristics of different variabilities in fMRI data is very important for further analysis, e.g., kriging.

Nonlinear distance

In the mining and petroleum industries, as the magnitude of the lag separation vector increases, the variogram increases too, showing some linear properties. This is reasonable since regions close in space tend to have similar values. This is generally observed in nature

and we call it the *physical distance*. In fMRI data analysis, a similar distance exists. The difference is that the spatial correlation between proximal voxels is stronger than that in many other data types, therefore the variogram approaches the origin with a somewhat reversed curvature, showing quadratic patterns. I call it the *nonlinear distance*. This property has been indicated by Spence et al. (2007).

Functional distance

Sometimes, the variogram may seem to fluctuate more or less periodically, rather than increasing monotonically (Webster and Oliver, 2001) and cyclicity is observed as a "downhole" variogram. This is called *hole effect* in geostatistics (Gringarten and Deutsch, 2001; Journel and Huijbregts, 1978). Hole effect structure is usually considered to be artificial and is ignored in most applications because near things tend to be more related than distant things in nature (Bowman, 2007; Tobler, 1970). But in fMRI data analysis, different regions of the brain may be functionally related even though they are not neighbors; we call this phenomena the *functional distance*. Ignoring these non-monotonic structures may result in unrealistic heterogeneity models that do not produce the observed patterns of variability (Pyrcz and Deutsch, 2007). If the model has hole effect structure, then the sill represents the global variance. When the variogram exceeds the sill, the correlation is negative between locations separated by lag h (Pyrcz and Deutsch, 2007), i.e., $\gamma(h) = C(0)[1 - \rho(h)]$ for $\rho(h) \in [-1, 1]$. Hence, the hole effect model contains both positive and negative correlations.

5.2.3 Variogram models

After computing an empirical variogram, we need to fit a model to the variogram. There are many variogram models for us to choose, such as Gaussian, spherical, and exponential (Webster and Oliver 2001; Goovaerts, 1997). Current geostatistical practice in selecting a model is often rather subjective, relying on empirical guidelines (Gorsich and Genton, 2000). By the special properties of fMRI data, I will jointly consider the nonlinear distance and functional

distance in variogram modeling. I also consider two approaches here, one parametric, the other nonparametric.

Basic models

Three different covariance and variogram models are used.

Nugget effect model The nugget effect model can be thought of as white noise, which is usually due to micro-scale variation or measurement error (Schabenberger and Gotway, 2005; Schabenberger and Pierce, 2002). All random variables $Z(s_i), i = 1, ..., N$ have the same mean and variance, and are without cross-correlations. The nugget effect can be considered as a discontinuity of the covariance function at the origin. Since the sum of valid covariance functions is itself a valid function, the nugget effect model is usually considered as a nested structure in other models. The nugget effect model can be expressed in a covariance form or a variogram form:

$$C(h) = \begin{cases} \sigma_\epsilon^2 & \text{if } h = 0, \\ 0 & \text{if } h > 0. \end{cases}$$

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0, \\ \sigma_\epsilon^2 & \text{if } h > 0. \end{cases}$$

Gaussian-type model Wackernagel (2003) defines the Gaussian-type model in a covariance form or a variogram form:

$$C(h) = \sigma^2 exp(-\tfrac{h^c}{a^c}),$$

$$\gamma(h) = \sigma^2[1 - exp(-\tfrac{h^c}{a^c})],$$

where $1 \leq c \leq 2$. This model reaches its maximum asymptotically. An *effective range* is defined as $\sqrt[c]{3}a$, which is the distance at which $\gamma(\cdot)$ equals 95% of the sill.

As $c \to 1$, the model is more a linear at small lags and the spatial correlation between nearby points is weaker. As $c \to 2$, the model approaches the origin with a more quadratic

shape and appears more sigmoid, and the spatial correlation between nearby points is higher. The two extreme cases are: when $c = 1$, it becomes a exponential model; when $c = 2$, it becomes a Gaussian model.

The pure Gaussian model is unstable (Webster and Oliver, 2001), because the quadratic behavior at the origin in the variogram will generate extreme values at the borders of the estimated map in kriging (Wackernagel, 2003). This is unrealistic in practice (Webster and Oliver, 2001). Hence, a nugget effect model is usually added as a nested structure in the Gaussian model to present a discontinuity at the origin, which avoids the extreme extrapolation properties and makes the kriging results more stable (Webster and Oliver, 2001; Wackernagel, 2003).

Bessel model A variogram may seem to fluctuate more or less periodically, rather than increase monotonically. This type of covariance function is defined as (Schabenberger and Gotaway, 2005)

$$C(h) = 2^{\nu}\Gamma(\tfrac{d}{2})(bh)^{-\nu}J_{\nu}(bh),$$

where b is the number of sign changes, d is the dimension of the data and $\nu = \frac{d}{2} - 1$. $J_{\nu}(\cdot)$ is the Bessel function of the first kind of order ν. Note $C(h) = cos(bh)$ when $d = 1$; $C(h) = J_0(bh)$ when $d = 2$; $C(h) = (\frac{1}{bh})sin(bh)$ when $d = 3$; $C(h) = exp[-(bh)^2]$ when $d \to \infty$. The periodic function has weaker hole effect structure as the dimension d increases, and it becomes a Gaussian function when $d \to \infty$.

In my problem, I mainly look at the function in $\mathbf{R}^2$, $C(h) = J_0(bh)$, which can be expressed as

$$J_0(bh) = \textstyle\sum_{k=0}^{\infty} \frac{(-1)^k}{k!^2}(bh)^{2k}.$$

Figure 5.1 gives examples of Bessel function $J_0(bh)$ with $h \in [0, 3.8]$ and $b = 2, 3, 4, 5$.

The Bessel function used in my analysis is

$$C(h) = \sigma^2 J_0(bh),$$
$$\gamma(h) = \sigma^2[1 - J_0(bh)],$$

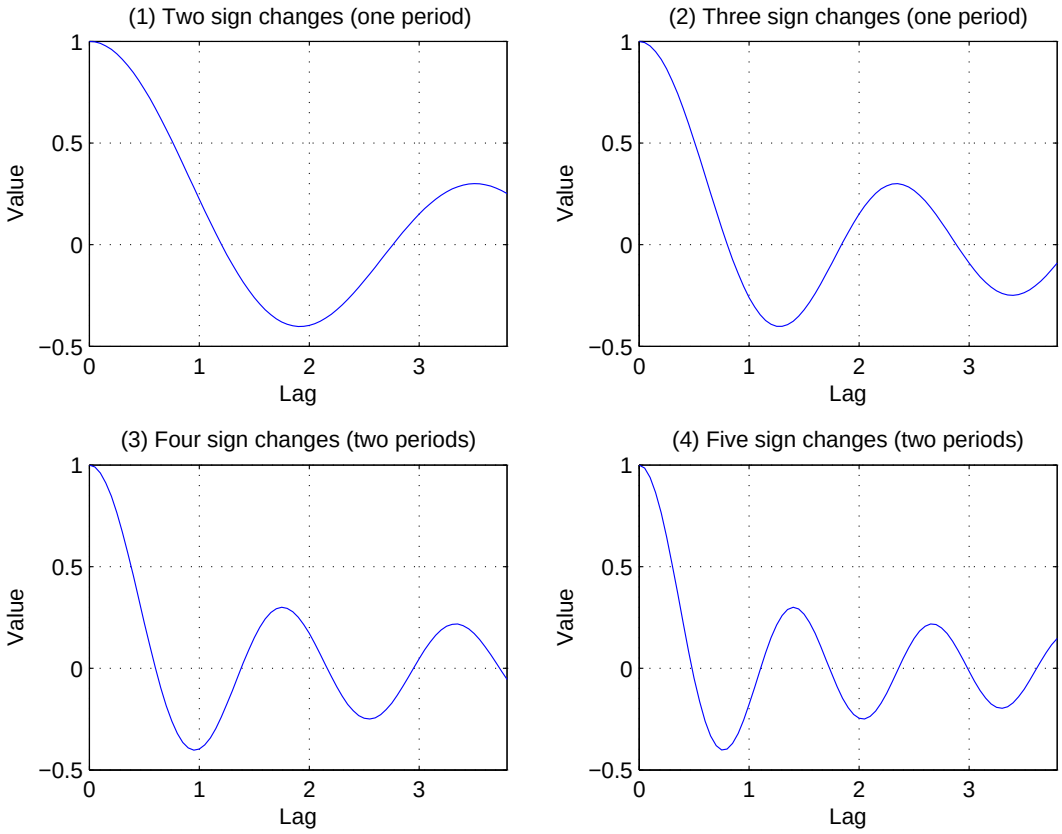

Figure 5.1: Examples of Bessel function $J_0(bh)$ with $h \in [0, 3.8]$ and $b = 2, 3, 4, 5$.

where σ^2 is the sill. Note that the range is not defined in the hole effect structure model (Ecker and Gelfand, 1997).

Parametric approach

Bessel Gaussian model Since the multiplication of two basic covariance models is still a valid covariance model, here I construct a variogram model based on a combined Gaussian-Bessel covariance model in $\mathbf{R}^2$:

$$C(h) = \sigma^2 exp(-\frac{h^c}{a^c}) J_0(bh),$$

$$\gamma(h) = \sigma^2[1 - exp(-\frac{h^c}{a^c}) J_0(bh)],$$

where $1 \leq c \leq 2$. When $b = 0$, it is a Gaussian-type model, i.e.,$\gamma(h) = \sigma^2[1 - exp(-\frac{h^c}{a^c})]$; when $a \to \infty$, it is a hole effect model, i.e., $\gamma(h) = \sigma^2[1 - J_0(bh)]$.

A nested variogram structure When the nugget effect model is thought of as a nested structure, the above constructed variogram model can be defined as

$$\gamma(h) = \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - exp(-\frac{h^c}{a^c}) J_0(bh)].$$

This alternative hole effect model considers both functional distance and nonlinear distance in the fMRI data, where σ_ϵ^2 is the nugget effect, σ^2 is the sill, $\sqrt[c]{3}a$ is the range, b is the number of sign changes.

Nonparametric approach

To increase the flexibility in modeling the variogram without violating the condition of positive definiteness of the covariance, we may choose a linear combination of valid covariance functions (Schabenberger and Gotway, 2005). This method is one kind of *non-parametric approach* (Ecker and Gelfand, 1997). The valid covariance functions are called *basis functions* or *step functions* (Shapiro and Botha, 1991). Since our data are two dimensional, $J_0(\cdot)$ can

be considered as the basis function (Ecker and Gelfand, 1997; Shapiro and Botha, 1991). The new variogram model is defined as

$$\gamma(h) = \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - \textstyle\sum_{k=1}^{p} w_k J_0(b_k h)],$$

where σ_ϵ^2 is the nugget effect, σ^2 is the sill, p is the number of basis functions, w_k is the weight assigned to the k^{th} function, b_k is the number of sign changes with geometric anisotropy property at the k^{th} function. To fit a variogram nonparametrically, the more basis functions we use, the better fitting for the data structure. Some researchers consider using hundreds of basis functions in the model fitting, e.g., Cherry et al. (1996) use 200 basis functions. But this is too computationally intensive for large data sets. Ecker and Gelfand (1997) suggest that at most five Bessel basis functions are enough based on their experience, because too many basis functions will result in over-fitting, i.e., a large number of weights of the basis functions will tend to be zero in the least squares method (Schabenberger and Pierce, 2002). To simplify model fitting again, Ecker and Gelfand (1997) also recommend to choose equal weights in the basis functions because the choice of the number of sign changes b_k is more sensitive to the choice of the weight w_k. To determine the appropriate b_k's, it is recommended to fit the empirical variogram with various numbers of sign changes and select the best one by least squares (Schabenberger and Pierce, 2002).

5.2.4 Variogram model selection

To choose a best model, one way is to use different kinds of variogram models for kriging and see how well they perform by some preset criteria (Webster and Oliver, 2001). *Cross-validation* is probably the simplest and most widely used technique to compare estimated and true values using only the already known data set (Hastie et al., 2001; Isaaks and Srivastava, 1989). A cross-validation study can help to choose between different variogram models by the different kriging results. K-fold cross-validation is commonly used.

In our setting, for data $Z(s_1), ..., Z(s_N)$, $K-$fold cross-validation splits the observations into K folds, where the k^{th} fold has data $Z(s_1^k), ..., Z(s_{N/K}^k)$ for $k = 1, ..., K$. In cross-validation, the data in each fold are removed in turn and new values are predicted at those locations by the observations in other folds using kriging. Let $\hat{Z}(s_1^k), ..., \hat{Z}(s_{N/K}^k)$ be the kriging values predicted by the (functional and physical) neighbors when the true observations are removed at locations $s_1^k, ..., s_{N/K}^k$, and $\hat{\sigma}^2(s_1^k), ..., \hat{\sigma}^2(s_{N/K}^k)$ be the kriging variance predicted by the neighbors. There are two criteria commonly used in geostatistics (Webster and Oliver, 2001): One is Mean Squared Error (MSE),

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N/K}[Z(s_i^k) - \hat{Z}(s_i^k)]^2.$$

For a selected variogram, we want the estimated MSE to be as small as possible. The other is Mean Squared Deviation Ratio (MSDR),

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N/K}[Z(s_i^k) - \hat{Z}(s_i^k)]^2/\hat{\sigma}^2(s_i^k).$$

MSDR can be considered as a weighted MSE (Ripley, 1981). Dividing the error allows one to compare the magnitudes of both the actual and the predicted error in the cross-validation (Davis, 1987). When MSDR >1, kriging variance underestimates the true estimation variances; when MSDR <1, kriging variance exaggerates the true estimation variances; when MSDR $\cong$1, the actual estimated error is equal on average to the error predicted by the model. For a selected variogram, we want the estimated MSDR to be as close to 1 as possible. The choice of criterion depends on the needs of the researcher (Davis, 1987). I will discuss this in more detail later.

5.2.5 Non-parametric marginal model analysis

Non-parametric marginal model analysis (Brunner et al., 2002) is used to test the group difference, where the null hypothesis is that there is no difference between two groups. The different groups in my analysis can be considered as different clusters in space or different

variogram models. This method is primarily used for longitudinal data where the observations of an experimental unit (subject) are repeated in time and called repeated measures analysis. Akritas and Arnold (1994) extend this method to general cases where the repeated measurements on an experimental unit can be time, space or other different occasions. In an experimental design with $i = 1, ..., a$ groups, each group has $k = 1, ..., n_i$ subjects, each subject is examined at $s = 1, ..., t$ occasions, i.e., a vector

$$\mathbf{X}_{ik} = (X_{ik1}, ..., X_{ikt})^T$$

is k^{th} subject in i^{th} group with t occasions. Usually different subjects are assumed to be independent and different occasions are assumed to be dependent. Under the assumption that the marginal distributions of $\mathbf{X}_{ik}$'s are identical, the hypotheses can be formulated in terms of either the distribution functions or the corresponding relative effects (expectations) (Brunner et al., 2002). The advantage of this method is the covariance structure of the different occasions need not to be known in the analysis, which uses the dependencies from the data directly.

5.3 Data analysis

5.3.1 Preprocessing steps

As in the previous chapter, the fourth slice of the saccade data is chosen here for demonstration purposes. Note the data for the analysis is $(x, y, t) = (64, 64, 156)$ for $z = 4$.

Although the data have been processed initially for statistical analysis, it is still necessary to adjust the data again for structural analysis. The following are three preprocessing steps.

Step one: Masking the brain

The first step is masking the brain. As already discussed in earlier chapters, this is a popular data reduction technique because it is only the brain that is of interest and not the

surrounding area. Therefore, it is often (if not always) helpful to remove the non-brain structures prior to further structural analysis (Jezzard et al., 2001).

In masking, the data image is first processed to identify the location of the brain, thereafter a suitable threshold method is used to remove all the voxels outside the brain (Goutte et al., 1999, 2002; Stanberry et al., 2003; Gibbons et al., 2004). For the slice I choose, there are 630 voxels out of 4096 left after masking (Figure 5.2, graph (2)).

Step two: Subtracting the mean of time image

By Huettel et al. (2004), the *raw signal* is not important in image processing and only the *functional signal* is of interest to the researcher. Functional signal, sometimes called *dynamic contrast*, is the intensity of the signal changes (called *contrast*) between voxels over time. Here we use the *contrast* between the raw image at each of the 156 time points and the mean of time image to do the analysis.

The second step is therefore subtracting the mean of time image at each of the 156 time points, which is equivalent to centering by time. Before centering the data, the images at the different time points are dominated by the mean of time image, and signal changes are so small that we can not see the changes clearly over time (Figure 5.2, graphs (3) and (5)). Since what we want is to see the true activations of the brain responding to the eye movement experiment, it is necessary to remove the background and to concentrate only on the changes over time (Figure 5.2, graphs (4) and (6)). Figure 5.2 gives the second preprocessing step at time points 70 and 71. The same preprocessing step is done at each time point.

Step three: Removing the trend

As mentioned before, to guarantee the surface to be stationary in the variogram estimation, trend removal is also necessary. I assume the data only has a first order trend, i.e., $f(x, y) = m_0 + m_1 x + m_2 y$. The third step is thus doing the trend surface fitting to remove the first order trend (Glover et al., 2006) (Figure 5.3). We may add back the trend after kriging if

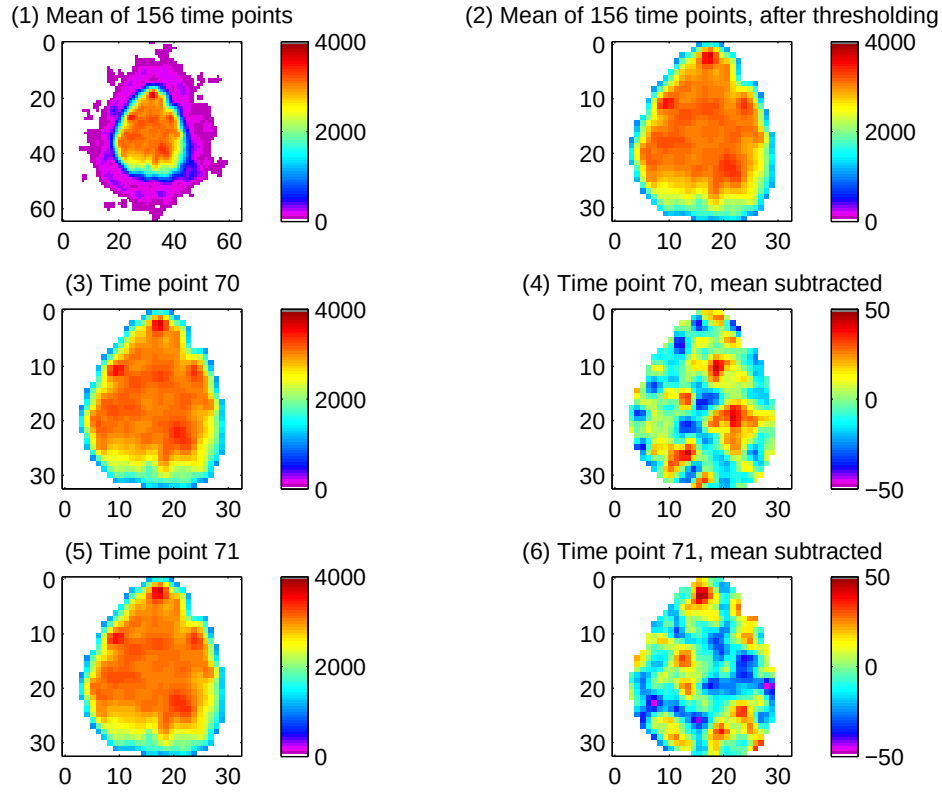

Figure 5.2: (1) is the mean of time image at the fourth slice. (2) is the masked mean of time image. (3) is the masked map at time point 70. (4) is the mean subtracted map at time point 70. (5) is the masked map at time point 71. (6) is the mean subtracted map at time point 71. From (1) to (2), is the first preprocessing step in structural analysis. From (3) to (4) and from (5) to (6) are the second preprocessing steps in structural analysis at time points 70 and 71 respectively. The signal changes are small from (3) to (5) but the changes are clearer from (4) to (5) after subtracting the mean of time image. Similar preprocessing step is done at every time points ranging from 1 to 156.

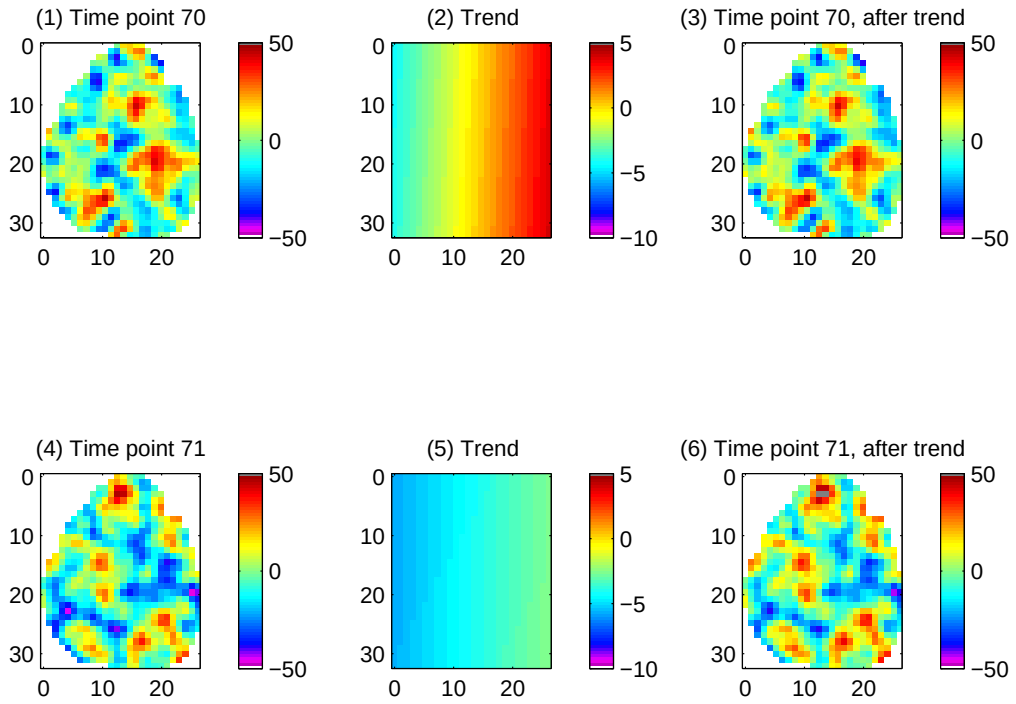

Figure 5.3: (1) is the mean subtracted map at time point 70. (2) is the first order trend map at time point 70. (3) is the trend removed map at time point 70. (4) is the mean subtracted map at time point 71. (5) is the first order trend map at time point 71. (6) is the trend removed map at time point 71. From (1) to (3) and from (4) to (6) is the third preprocessing step in structural analysis at time points 70 and 71 respectively.

necessary. Figure 5.3 gives the third preprocessing step at time points 70 and 71. The same preprocessing step is carried out at every time point.

5.3.2 Empirical analysis

The empirical variograms over the 156 time points for the trend removed data are calculated (Marcotte, 1996). To reduce the measurement error in the calculation (Journel and Huijbregts, 1978), lag distance of the variogram is chosen as 19, where the number of pairs of observations in the x direction is ≥ 94, and the number of pairs of observations in the y direction is ≥ 176.

Figure 5.4 shows the graphs of the empirical variograms at time point 70 in different directions. "$\circ$" indicates the variogram values in the x direction, "$+$" indicates the values in the y direction, "$*$" indicates the values in the 45 degree of x and y directions. "$\times$" indicates the values in the 135 degree of x and y directions.

I look at the variograms at different time points, and find that most of them have similar values and patterns as demonstrated in Figure 5.4 for time point 70. The variograms in the x and the y directions largely overlap, showing the second order stationarity in these two directions. The variograms also show "waves" as the lag distance increases, indicating functional relationships. There are some exceptions, i.e., time points at which the variograms are not stable, showing significant differences with the others. This makes sense from a geostatistical perspective, because spatial information can be used to provide data on structure; different variogram structure reflects different activation patterns in the brain. Hence the structural classification method is used again here.

In order to find those significantly different time points, an omnidirectional classification among the 156 time points is performed by averaging the x and y directional variograms. Clusters are characterized by their spatial patterns, i.e. variograms. In each cluster, the variograms have different behaviors due to variations in structure. The steps in the clustering are: (1) average the variograms in the x and y directions; (2) define the "Euclidean" metric

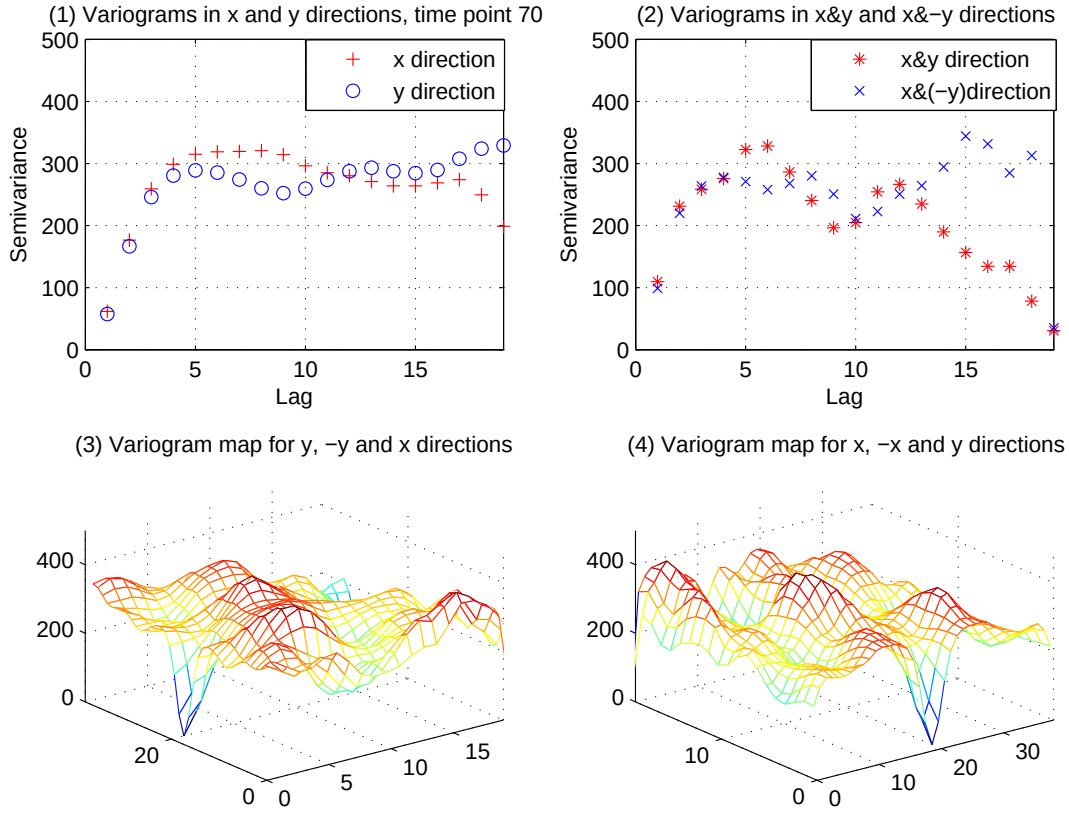

Figure 5.4: (1) are the empirical variograms in the x and y directions. (2) are the empirical variograms in the 45 degree and 135 degree of x, y directions. (3) is the empirical variogram map in all y, $-y$ and x directions. (4) is the empirical variogram map in all x, $-x$ and y directions.

since we want to compare the similarity of average profile levels in the variogram; (3) use the k-means algorithm to do the clustering for different k; (4) determine the final results by silhouette values.

The first choice for the number of clusters is two, since it has the largest mean of the silhouette values. Cluster 1 has 2 time points, which are 1 and 93; cluster 2 contains all other points. When I look at the time course of the 340 voxels from SPCA in the previous chapter (Figure 5.5, graph (1)), I find there are large upward movements from time point 1 to 2 (Figure 5.5, graph (2)) and from time point 92 to 93 (Figure 5.5, graph (3)), these discontinuities are not task related and might be from the uncorrected head motion (Jezzard et al., 2001) because motions over time are visible as largely vertical movements on the plots (Huettel et al., 2004). Hence, I ignore these two time points and see the other options. The second choice of the number of the clusters is three. This process results in three clusters. Cluster 1 has time points 1 and 93; cluster 2 has 28 time points, which are 15, 16, 17, 25, 33, 39, 40, 46, 47, 58, 59, 66, 81, 83, 87, 94, 95, 103, 104, 112, 113, 127, 128, 132, 133, 143, 144, 147; cluster 3 contains all other points. Each cluster presents a different spatial variability profile.

A non-parametric marginal model analysis (Brunner et al., 2002) is used to test if the groups of voxels in clusters 2 and 3 are significantly different, where each group as well as the different lags of the variograms is subject to a structure. The null hypothesis is that there is no difference between the two groups. Since the $p-$value is < 0.0001, the null hypothesis is rejected. Hence I conclude there are 28 time points with significantly different variograms from the rest (Table 5.1). Among the 28 time points, 16 are in anti-saccade time periods, and the important changes usually happen near the beginning of the task, showing the brain has strong activations when the pro-saccade task is switched to the anti-saccade task; 12 of them are in pro-saccade time periods, and the important changes usually happen at the middle of the task. I also use the non-parametric marginal model analysis (Brunner et al., 2002) to test if the 16 time points in the anti-saccade condition are significantly different

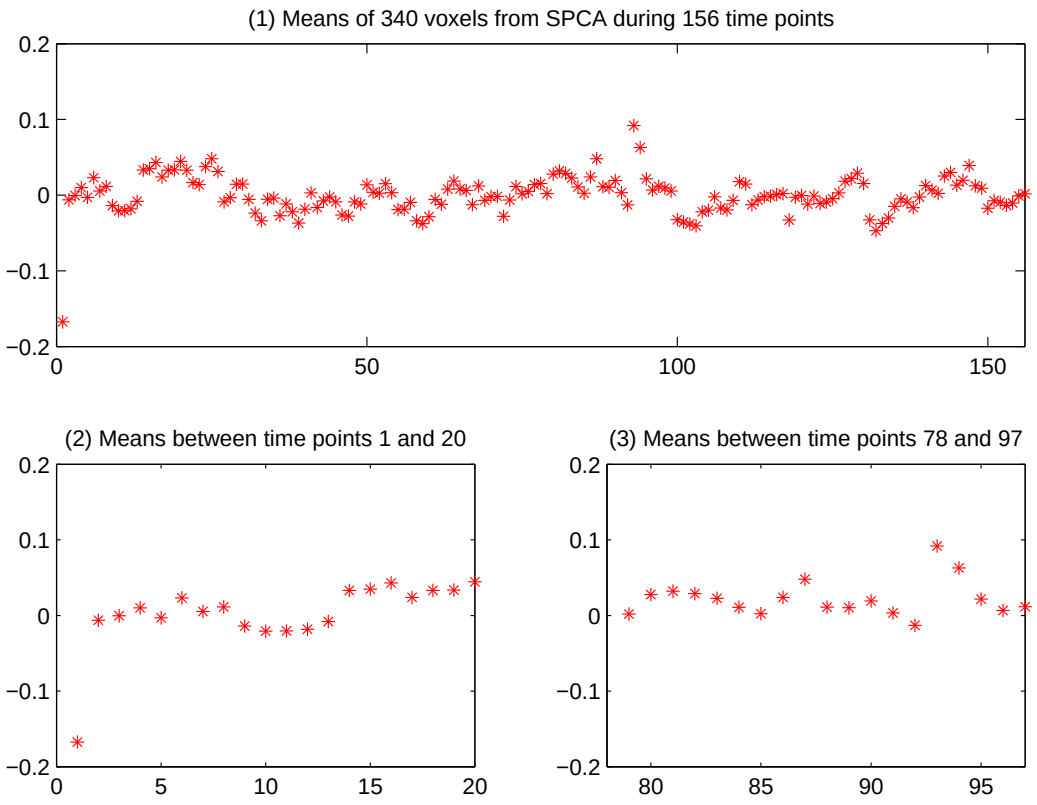

Figure 5.5: (1) is the means of 340 voxels from SPCA during 156 time points. (2) is the means between time points 1 and 20, where there is a large upward movement from time point 1 and 2. (3) is the means between time points 78 and 97, where there is a large upward movement from time point 92 to 93. These discontinuities are considered as uncorrected motion which is often visible as largely vertical movements on the time course plot.

from the 12 time points in the pro-saccade condition. The null hypothesis is not rejected. The differences in the two tasks are not significant. In my opinion, the reason is the active regions corresponding to the eye movement task are similar, so the variogram structure does not have big changes.

Anti-saccade	Significant time points	Pro-saccade	Significant time points
1	-	2-14	-
15-26	15, 16, 17, 25	27-38	33
39-50	39, 40, 46, 47	51-62	58, 59
63-74	66	75-86	81, 83
87-98	87, 94, 95	99-110	103, 104
111-122	112, 113	123-134	127, 128, 132, 133
135-146	143, 144	146-156	147

Table 5.1: Number of time points with significantly different variograms. There are 16 time points in anti-saccade time periods, and the important changes usually happen near the beginning of the anti-saccade task. There are 12 time points in pro-saccade time periods, and the important changes usually happen at the middle of the pro-saccade task. Among the 28 time points, variogram profile levels during anti-saccade are higher than those in pro-saccade, but these differences are not significant.

5.3.3 Variogram modeling

As mentioned before, it is necessary to model the empirical variogram if we want to do some further analysis. I consider modeling the variogram during time periods 61-90 for demonstration purpose. Since the trend has been removed from the data, I assume each time point is second order stationary (Glover et al., 2006). I also assume the variogram is geometric anisotropic, which means the variogram model is the same in both x and y directions but the parameters of the model maybe different.

Parametric approach

First, I model the 30 time points by a Gaussian-type model

$$\gamma(h) = \sigma_\epsilon^2 + \sigma^2[1 - exp(-\frac{h^c}{a^c})],$$

with $c = 1, 1.5, 2$ in both x and y directions, called G-100, G-150 and G-200 respectively. For comparison, I also model those time points with hole effect structures by the Bessel Gaussian model

$$\gamma(h) = \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - exp(-\frac{h^c}{a^c})J_0(bh)],$$

with $c = 1, 1.5, 2$ in both x and y directions, called BG-100, BG-150, BG-200 model. I assume the number of sign changes b is the same in both directions. Hence in the Bessel Gaussian model, the geometric anisotropy property is the same as that in the Gaussian-type model. The only difference is that the Bessel Gaussian model has an additional parameter b to control the hole effect structure.

All the 30 time points can be well fit by the Gaussian-type model (for example, see Figures 5.6, 5.7, 5.8, 5.9 and 5.10)). But for the Bessel Gaussian modeol, only 10 time points are well fitted by the BG-100, BG-150 and BG-200 models: 61, 64, 65, 70, 71, 76, 80, 83, 85, 89 (e.g., Figures 5.6 and 5.7). 4 time points can be fitted by the BG-100 or BG-150 models: 62, 75, 86, 90 (Figure 5.8). In these four points, as $c \to 2.00$, the number of signs $b \to 0$, so the BG-200 model is equivalent to the Gaussian-type model, where the hole effect structure disappears.

The other 16 points are not well fit by the Bessel Gaussian model. In my opinion, this has two reasons: one is that the assumption that the number of sign changes is the same in all directions is violated (time points 66, 67, 69, 72, 73, 77, 78, 79, 81, 82, 84, 87, 88). Under this situation, the range parameter $a \to \infty$ in the Bessel Gaussian model and the model becomes a pure Bessel model. The other is that the hole effect structure is weak, so the sign change parameter $b \to 0$ (time points 63, 68, 74).

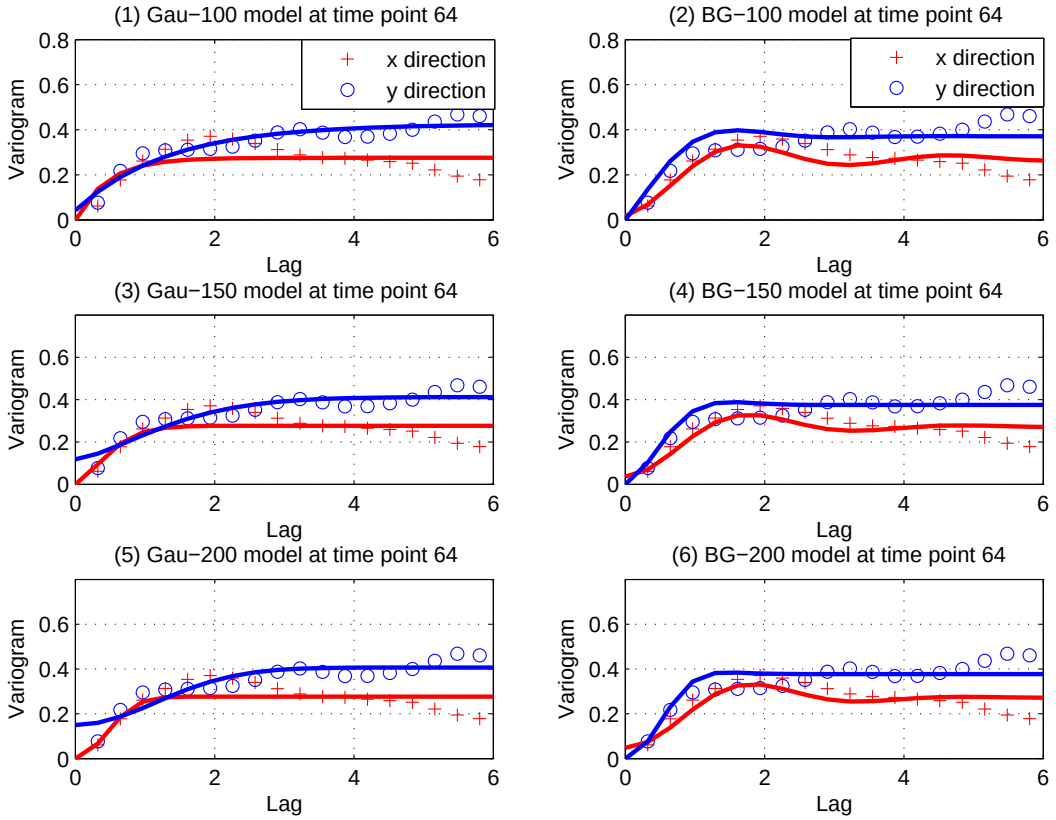

Figure 5.6: (1), (3), (5) are time point 64 in Gau-100, Gau-150, Gau-200 model fitting. (2), (4), (6) are time point 64 in BG-100, BG-150, BG-200 model fitting.

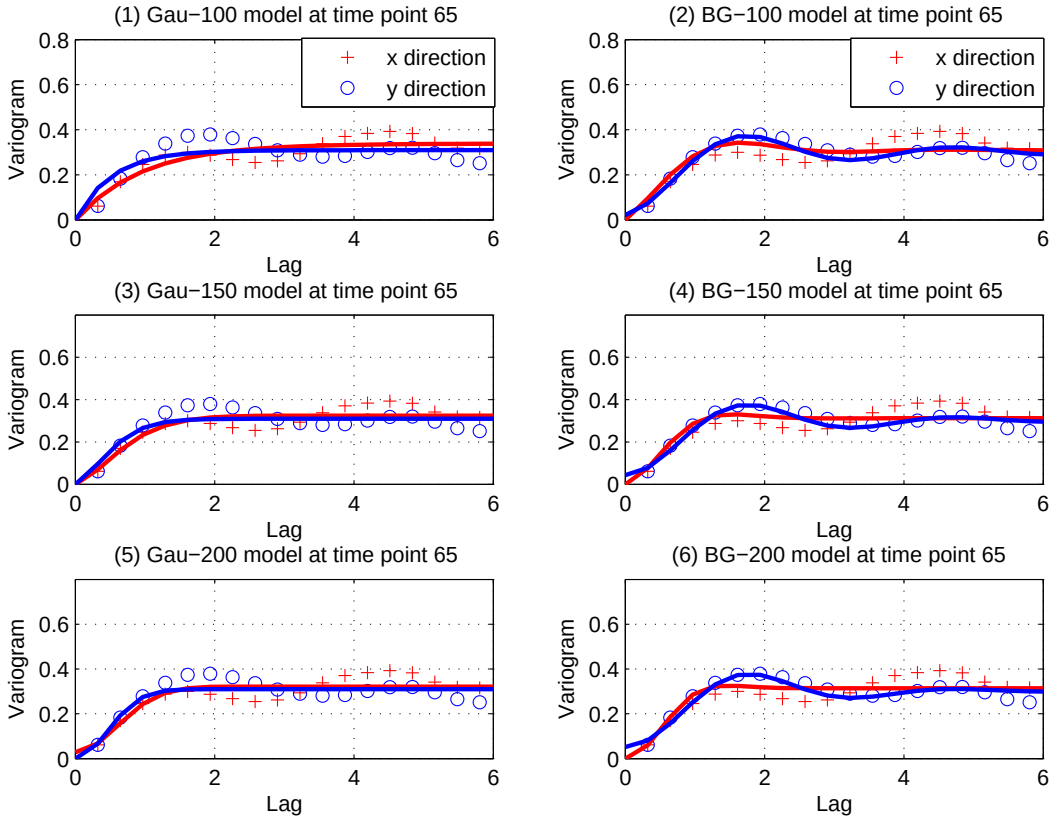

Figure 5.7: (1), (3), (5) are time point 65 in Gau-100, Gau-150, Gau-200 model fitting. (2), (4), (6) are time point 65 in BG-100, BG-150, BG-200 model fitting.

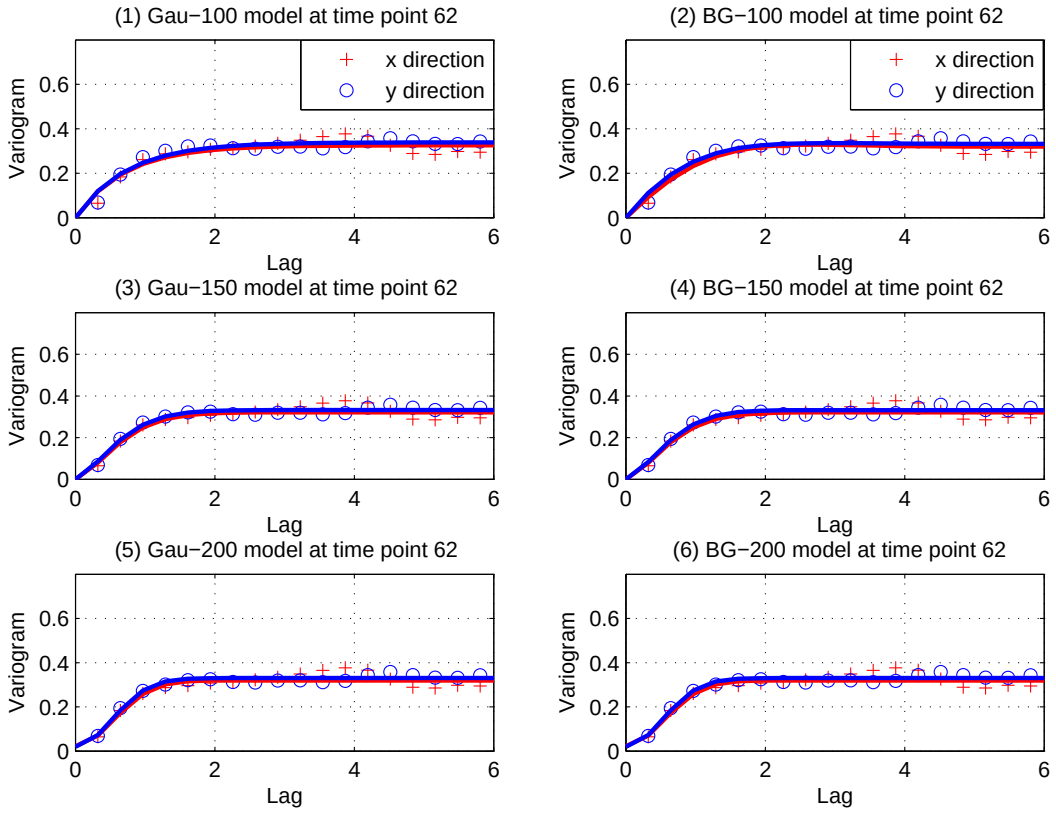

Figure 5.8: (1), (3), (5) are time point 62 in Gau-100, Gau-150, Gau-200 model fitting. (2), (4), (6) are time point 62 in BG-100, BG-150, BG-200 model fitting.

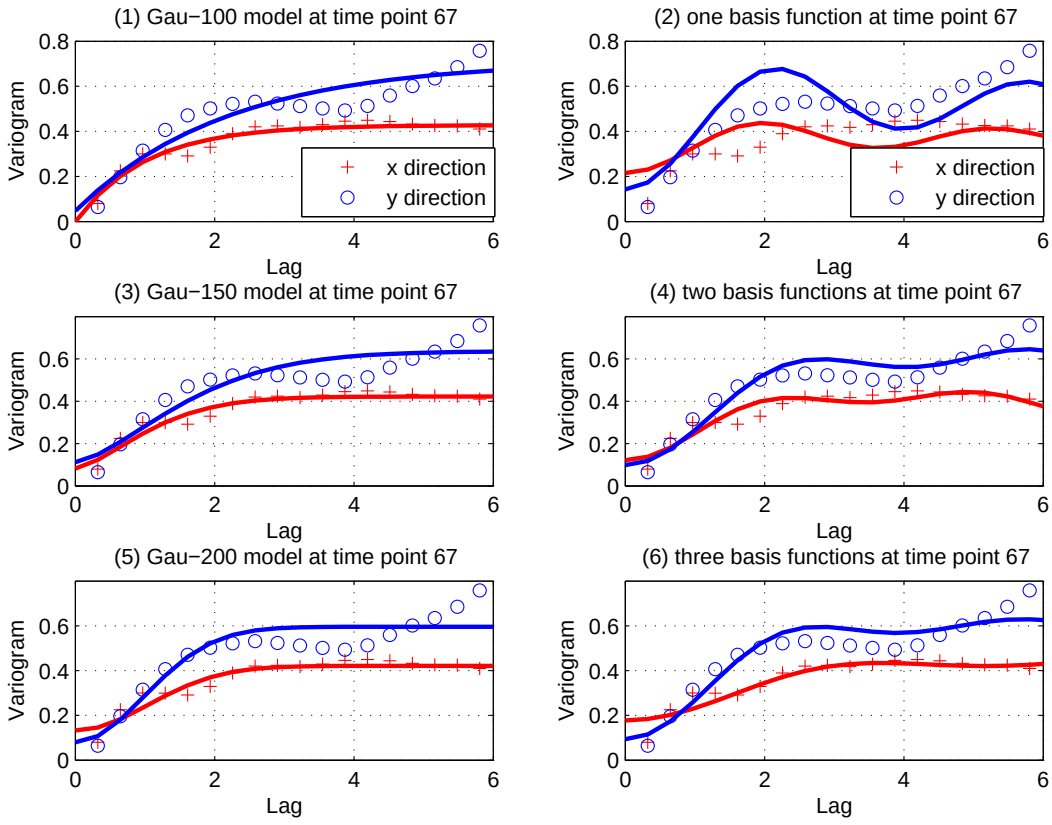

Figure 5.9: (1), (3), (5) are time point 67 in Gau-100, Gau-150, Gau-200 model fitting. (2), (4), (6) are time point 67 in one basis function, two basis functions, three basis functions model fitting.

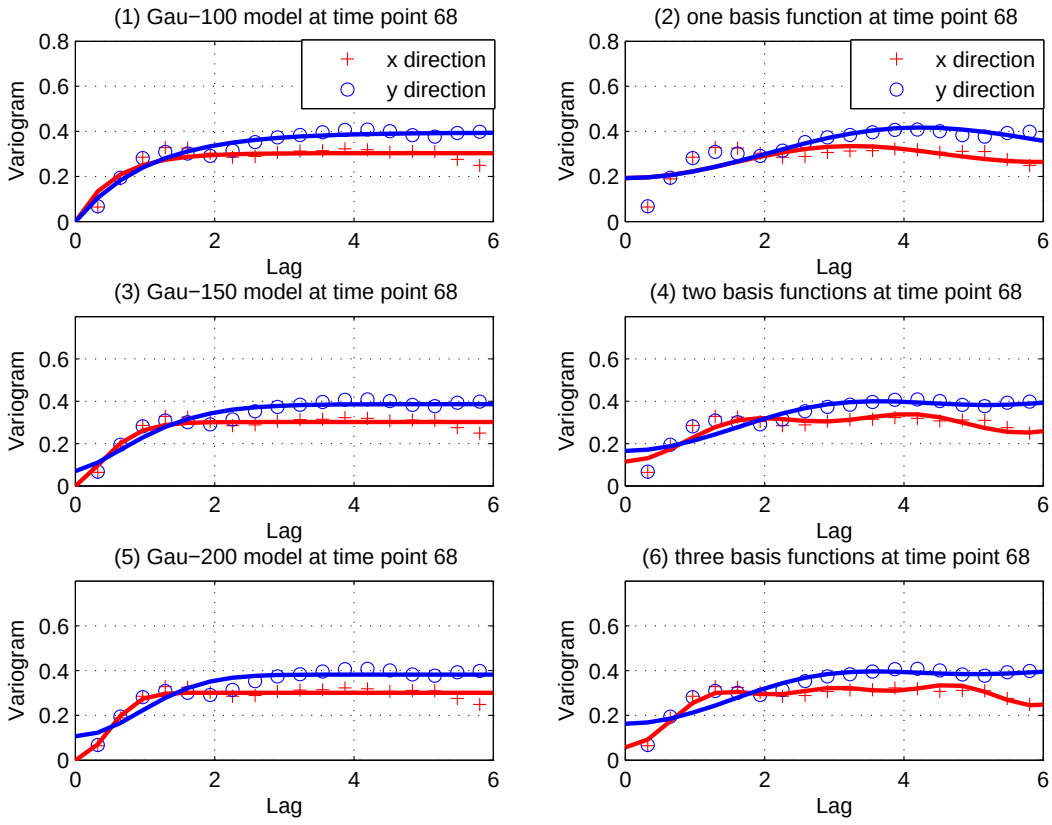

Figure 5.10: (1), (3), (5) are time point 68 in Gau-100, Gau-150, Gau-200 model fitting. (2), (4), (6) are time point 68 in one basis function, two basis functions, three basis functions model fitting.

Nonparametric approach

For the 16 time points that can not be well fitted by a Bessel Gaussian model (Figures 5.9 and 5.10), I model them by a nonparametric method, where the pure Bessel model is considered as a basis function, and the number of basis functions is chosen to be one, two or three. Because the choice of change of signs is more sensitive than the choice of weight, I consider each basis function as being equally weighted and the geometric anisotropy property happens in the change of signs. Now the function is defined as

$$\gamma(h) = \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - \textstyle\sum_{k=1}^{p} \frac{1}{p} J_0(kbh)],$$

where $p = 1, 2, 3$. Hence, the three fitting functions are:

$$\begin{aligned}
\gamma(h) &= \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - J_0(bh)], \\
\gamma(h) &= \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - (\tfrac{1}{2} J_0(bh) + \tfrac{1}{2} J_0(2bh))], \\
\gamma(h) &= \sigma_\epsilon^2 + (\sigma^2 - \sigma_\epsilon^2)[1 - (\tfrac{1}{3} J_0(bh) + \tfrac{1}{3} J_0(2bh) + \tfrac{1}{3} J_0(3bh))].
\end{aligned}$$

5.3.4 Variogram model selection

The best variogram model is chosen by cross-validation in kriging. According to Hastie et al. (2001), when the sample size N is greater than 125, 5-fold cross-validation has lower variance and does not suffer from much bias. Since there are 630 voxels considered in the analysis, I use $K = 5$ for model selection. Figure 5.11 shows the 5 cross-validation folds. In each fold, 126 voxels are removed in turn for estimation, as indicated by the dark squares.

For model selection, Mean Squared Deviation Ratio (MSDR) and Mean Squared Error (MSE) are both calculated for each time point.

Nonlinear distance

As shown in Table 5.2, most of the MSDR values are less than 1, indicating the kriging variance overestimated the true estimation variance. MSDR increases as the parameter c increases from 1 to 2, indicating the model fit gets better as c increases. Most time points

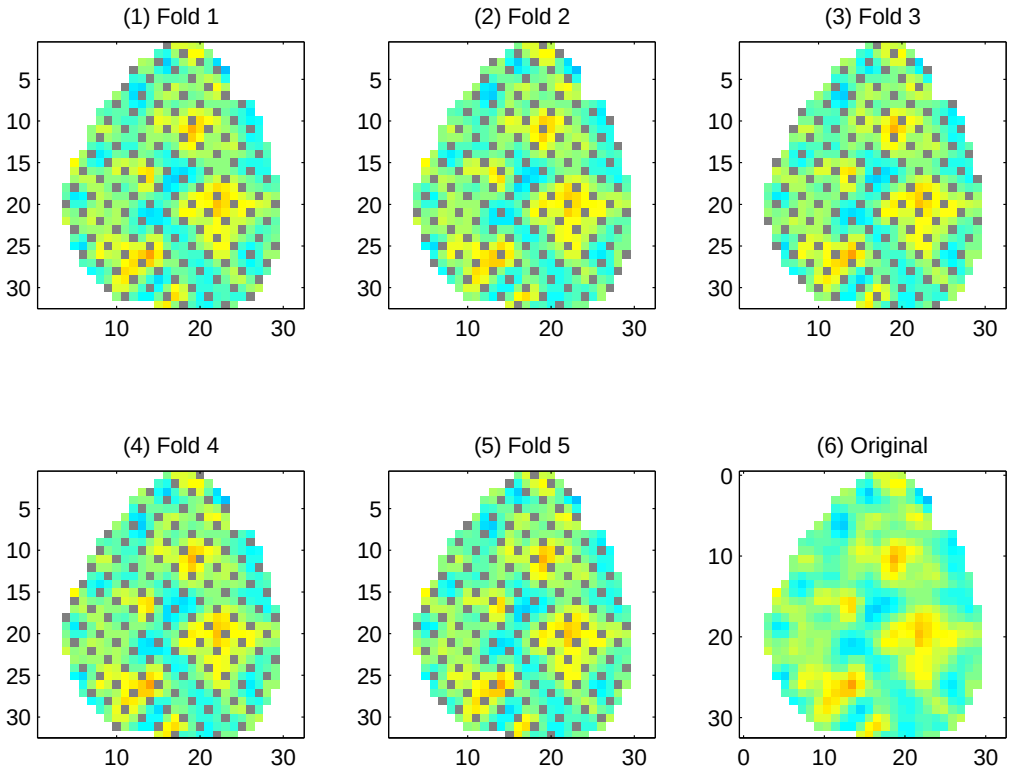

Figure 5.11: Maps for 5-fold cross-validation. (1), (2), (3), (4), (5) are the five folds where the voxels denoted by dark squares (126 voxels) are removed in each fold for estimation. (6) is the original map with 630 voxels.

are fitted best by the G-200 model, showing the nonlinear distance in the neighbors of the voxels. Hence proximal voxels in the brain have stronger relations.

Functional distance

By looking at the 14 time points fitted by the Bessel Gaussian model Table 5.2), we see the MSDR is closer to 1 than for the Gaussian-type model. Hence, if we consider the functional distance in the model, the estimation will more approach the true model. I also use the nonparametric marginal model analysis (Brunner et al., 2002) to test the differences in MSDR between the Gaussian-type model group and the Bessel Gaussian model group. The p-value is 0.0096, indicating the difference of the two groups is significant. For these 14 time points, using hole effect structure can significantly improve the model fitting. Table 5.3 also show that the MSE values slightly increase if we use Bessel Gaussian model instead of Gaussian-type model. I use the same method to test differences of MSE between the Gaussian-type model group and the Bessel Gaussian model group. The p-value now is 0.6587, indicating the difference of the two group is not significant. Overall, jointly considering the functional distance will significantly improve the model fitting in these 14 time points.

For the 16 time points fitted by one, two, three basis functions, we also see the MSDR is closer to 1 than for the Gaussian-type model (Table 5.4), but the MSEs are also larger than for the Gaussian-type model (Table 5.5). I also do the same hypothesis tests as above, both p-values are < 0.0001, indicating significant differences between these two approaches. The reason is: even though the nonparametric approach fits the data very well in general, it produces a larger nugget effect than that in the Gaussian-type model fitting (Figures 5.9 and 5.10), which means the closest neighbors are not fit well. Hence, we have two choices for these 16 time points, either getting better model fitting but with significantly larger errors, or having smaller errors but getting worse model fitting.

According to Davis (1987), a model chosen to be the best by the cross-validation method in kriging, is only under some specific conditions, e.g., the choice of discrepancy measure. No

models can be chosen to be the best under universal discrepancy measures. Although MSE is a more popular measure in the literature, MSDR gives a way of assessing the adequacy of the model and of its prediction. I consider MSDR as the main criterion in my variogram model selection, because what I want is to find a valid model that is accurate for the variogram in general trends. Hence MSDR is a good measure (Davis, 1987; Webster and Oliver, 2001). From the above results, to fit those time points with fluctuations in the variogram, the variogram model with hole effect structure gives much better kriging results than monotonically increasing models.

5.4 Conclusions and discussion

By looking at the empirical variograms across the 156 time points, we find their structures reflect the activations of the brain by the saccade task. When there are no task related activations in the brain, the variograms show strong stationarity, and the sills of the variograms are low. When there are task-related activations in the brain, the structure of the brain shows significant differences. The variograms have stronger anisotropic properties in the x and y directions and the sills of the variograms are higher. For the anti-saccade periods, those significant time points are usually at the beginning of the epoch. For the pro-saccade periods, those significant time points are usually in the middle of the task. The variogram structures tend to be somewhat stronger in anti-saccade time periods than those in pro-saccade, but the differences are not significant.

Spence et al. (2007) use the Gaussian model for variogram fitting. This is intuitively obvious from my results that the proximal voxels in the brain have stronger relations. So I confirm the validity of their results.

In geostatistics, the variogram model is monotonically increasing and "hole effect" structure is usually ignored. For fMRI data analysis, I consider both functional distance and physical distance. This is a major difference between geostatistical data and fMRI data. In

geostatistics, physical distance is most relevant since near regions are more related than distant regions, and distant regions will tend to not be related. But in fMRI, different regions may be functionally related even though they are not neighbors. This claim is similar to that in Bowman (2007), but my approach is different. The advantage of my method is that it does not need auxiliary data any more, and considers the original data only. I use the hole effect variogram model in model fitting, because the presence of such an effect provides valuable information concerning spatial variability, indicating a form of periodicity (Pyrcz and Deutsch, 2007).

In the analysis of brain imaging data, using the hole effect variogram model offers an important advantage over the current monotonic variogram, whether a parametric method or nonparametric method is used. This provides a good example of applying geostatistical ideas in a new area. But my hole effect model still has two disadvantages: One is that it may produce a larger nugget effect in the model fitting stage, especially in the nonparametric approach, which means if we fit the functional relations very well, we have to sacrifice some physical relations for the target voxel. The other is that the regular hole effect model in geostatistics is with dampening, which means the functional relations are always weaker than the physical relations. But for a target voxel in fMRI data, its distant voxels may exhibit higher correlations than some closer neighbors (Bowman, 2007). Hence the regular hole effect model in geostatistics may not be suitable. How to well define a suitable hole effect variogram model and keep the balance between the functional relations and physical relations in the model fitting will be considered in my future work.

MSDR	Gau-100	Gau-150	Gau-200	BG-100	BG-150	BG-200
61	0.1623	0.1622	0.5558	0.3731	0.5609	0.6141
62	0.2038	0.1905	0.4340	0.2390	0.1957	-
63	0.2414	0.3821	0.6499	-	-	-
64	0.3442	0.4448	0.6301	0.4668	0.6788	1.0026
65	0.2356	0.1928	0.4863	0.5293	0.7675	0.9634
66	0.5023	0.7017	0.9157	-	-	-
67	0.3824	0.3794	0.6036	-	-	-
68	0.2493	0.3406	0.5928	-	-	-
69	0.8143	1.1373	1.4314	-	-	-
70	0.1939	0.1482	0.4286	0.2234	0.3474	0.6226
71	0.1649	0.1678	0.5540	0.4925	0.7397	0.8088
72	0.2553	0.3342	0.5304	-	-	-
73	0.4075	0.6141	0.9003	-	-	-
74	0.2166	0.2648	0.5312	-	-	-
75	0.3120	0.4904	0.8536	0.2438	0.3088	-
76	0.1766	0.1160	0.3457	0.2250	0.1704	0.4083
77	0.8709	0.3477	0.4978	-	-	-
78	0.2349	0.2540	0.4855	-	-	-
79	0.6453	0.5982	0.9221	-	-	-
80	0.2405	0.3073	0.5337	0.2179	0.3175	0.6621
81	0.3252	0.4411	0.7044	-	-	-
82	0.4240	0.2898	0.3791	-	-	-
83	0.1954	0.1831	0.6250	0.2111	0.2044	0.9950
84	0.4911	0.6099	0.7161	-	-	-
85	0.2792	0.4200	0.6602	0.1962	0.3262	0.6249
86	0.2607	0.1908	0.3377	0.2986	0.1889	-
87	0.2180	0.3365	0.5484	-	-	-
88	0.3140	0.4890	0.7676	-	-	-
89	0.2130	0.1890	0.4811	0.3341	0.5254	0.6298
90	0.2298	0.2028	0.4486	0.2251	0.2246	-

Table 5.2: MSDR for different time points (61-90) in the Gaussian-type model, and MSDR for 14 time points in the Bessel Gaussian model. Compared with the Gaussian-type model, the Bessel Gaussian model attains better MSDR values. The differences between the two models are significant.

MSE	G-100	G-150	G-200	BG-100	BG-150	BG-200
61	22.2578	10.3682	13.5639	24.9593	39.2324	42.1805
62	23.1404	9.8625	14.5760	21.0463	9.5371	-
63	18.2641	33.0493	66.6354	-	-	-
64	40.6288	47.0910	59.6362	25.5761	27.4002	33.4274
65	24.7432	9.8289	12.0064	26.2968	31.4134	33.7669
66	57.6366	95.4935	132.3619	-	-	-
67	45.5645	54.7449	74.2879	-	-	-
68	25.9159	28.0734	41.9934	-	-	-
69	76.3764	125.3946	171.2664	-	-	-
70	22.3660	8.4836	1.5125	13.3760	13.9182	20.7360
71	16.8173	7.0828	11.4915	32.2751	43.8742	45.6591
72	31.7380	31.3293	35.8695	-	-	-
73	29.4400	45.2649	66.5082	-	-	-
74	19.2579	13.0700	26.1599	-	-	-
75	20.9743	24.5880	40.5677	14.1453	6.9437	-
76	22.1729	7.6267	1.4703	16.5448	6.4087	6.2182
77	106.6823	27.4747	25.1347	-	-	-
78	29.8707	17.0820	27.0628	-	-	-
79	67.3441	90.0690	123.8678	-	-	-
80	29.9014	21.4519	36.2870	18.6205	18.3886	30.4802
81	43.7206	36.8092	68.7382	-	-	-
82	93.1294	30.3147	33.5358	-	-	-
83	37.9755	17.0388	15.1281	34.8301	12.2169	15.3086
84	39.7562	51.8816	64.7611	-	-	-
85	27.7348	33.4896	48.6876	13.8412	16.0320	25.6907
86	39.9071	15.3667	6.6063	29.3797	12.0022	-
87	39.1465	45.5264	72.7195	-	-	-
88	41.2146	59.9628	92.9210	-	-	-
89	20.3588	7.9737	10.1546	17.9671	23.5594	18.4348
90	18.9660	7.7225	12.2218	13.9418	6.2454	-

Table 5.3: MSE for different time points (61-90) in the Gaussian-type model, and MSE for 14 time points in the Bessel Gaussian model. Compared with the Bessel Gaussian model, the Gaussian-type models have smaller MSE values. The differences between the two models are not significant.

MSDR	Gau-100	Gau-150	Gau-200	One basis	Two basis	Third basis
63	0.2414	0.3821	0.6499	0.6213	0.7927	1.0849
66	0.5023	0.7017	0.9157	0.6949	0.7530	0.7548
67	0.3824	0.3794	0.6036	0.4511	0.7567	0.6977
68	0.2493	0.3406	0.5928	0.6372	0.6251	0.6885
69	0.8143	1.1373	1.4314	0.7967	0.8368	0.9723
72	0.2553	0.3342	0.5304	0.4793	0.6100	0.5860
73	0.4075	0.6141	0.9003	0.7940	0.8325	0.9379
74	0.2166	0.2648	0.5312	0.6878	0.6440	0.7119
77	0.8709	0.3477	0.4978	0.7630	1.0908	1.5687
78	0.2349	0.2540	0.4855	0.6385	0.6026	0.5960
79	0.6453	0.5982	0.9221	0.8324	0.8233	0.9303
81	0.3252	0.4411	0.7044	0.6339	0.7335	0.8919
82	0.4240	0.2898	0.3791	0.5746	0.7317	0.7205
84	0.4911	0.6099	0.7161	0.6669	1.0138	0.7020
87	0.2180	0.3365	0.5484	0.6682	0.6738	0.7056
88	0.3140	0.4890	0.7676	0.3846	0.6275	0.5564

Table 5.4: Comparisons of MSDR for 16 time points under the Gaussian-type method and the nonparametric method. The nonparametric approach attains better MSDR values. These differences are significant by non-parametric marginal model analysis.

MSE	Gau-100	Gau-150	Gau-200	One basis	Two basis	Third basis
63	18.2641	33.0493	66.6354	89.3964	67.0004	30.0501
66	57.6366	95.4935	132.3619	187.8435	174.5950	171.6620
67	45.5645	54.7449	74.2879	91.2558	93.9873	104.6891
68	25.9159	28.0734	41.9934	131.9907	97.0111	86.0126
69	76.3764	125.3946	171.2664	192.8486	172.9570	160.1111
72	31.7380	31.3293	35.8695	75.8956	100.7255	57.3930
73	29.4400	45.2649	66.5082	111.8570	100.2815	85.3291
74	19.2579	13.0700	26.1599	122.5606	89.3827	51.1313
77	106.6823	27.4747	25.1347	84.0561	66.6550	52.7147
78	29.8707	17.0820	27.0628	130.5393	97.4070	92.4328
79	67.3441	90.0690	123.8678	151.6380	98.1363	145.5408
81	43.7206	36.8092	68.7382	170.4746	112.8929	95.8092
82	93.1294	30.3147	33.5358	146.5626	161.1876	154.8976
84	39.7562	51.8816	64.7611	119.2967	65.9191	71.1864
87	39.1465	45.5264	72.7195	201.0792	138.0778	101.9978
88	41.2146	59.9628	92.9210	84.3343	136.8792	84.2873

Table 5.5: Comparisons of MSE for 16 time points under the Gaussian-type method and the nonparametric method. The nonparametric method tends to larger MSE values than the Gaussian-type method. These differences are significant by non-parametric marginal model analysis.

Chapter 6

Spatio-temporal smoothing in fMRI

In the previous chapter, I found suitable models to describe the spatial structure of the fMRI data. The purpose of this chapter is to use the selected variogram model to do the smoothing in fMRI. By considering both the spatial and temporal properties of the data, the proposed smoothing method will greatly reduce the noise of the data in an intelligent way.

6.1 Introduction

In fMRI data analysis, *spatial filtering*, sometimes called *Gaussian filtering*, is commonly used as a preprocessing step to increase the signal to noise ratio in the original data (Huettel et al., 2004). The width of this filter determines the extent of the smoothing that takes place, i.e., the wider the filter, the smoother the data. It is usually required that the extent of the smoothing is not larger than the size of the activated region (Jezzard et al., 2001). Over-smoothing will mask the differences between different regions (Spence et al., 2007; Lazar, 2008), and result in failure to detect activation inside the brain. Based on this, the typical filter width for fMRI data is about 2 or 3 voxels, e.g., the saccade data used in our analysis was preprocessed by Gaussian filtering with a radius of 2 voxels. But because of the limitations of the Gaussian filtering, there is still considerable variability (statistical noise) in the data (Lazar, 2008). Rather than doing simple blurring by Gaussian filtering, we may consider incorporating spatial information into the smoothing procedure to reduce noise in an intelligent way. For each target voxel, this method can selectively choose different weights for its neighbors (Jezzard et al., 2001). Here I consider *filtered kriging* as an example of such

a smoothing method because it has an initial consideration of the spatial structure of the data, i.e., the modeled variogram of the data.

In time series there are three distinct types of prediction problems (Chiles and Delfiner, 1999; Schabenberger and Gotway, 2005): *forecasting*, which means prediction of future data; *filtering*, which means prediction of current data; *smoothing*, which means prediction of past data. In spatial statistics, we can use forecasting to predict data at new locations, but we can not distinguish past and current data, therefore filtering and smoothing refer to the same operation. As mentioned before, kriging can be used for prediction at new locations. When the predicted variable is at one of the sampled sites, kriging gives the original data value, which is called *interpolation*. This is under the assumption that there is no measurement error at this sampled site. Often in practice data may contain measurement error. If we want to remove the measurement error in kriging, we can consider kriging as a smoothing method, and this is called *filtered kriging* (Cressie, 1993).

6.2 Concepts and methods

6.2.1 Kriging and filtered kriging

In kriging, it is assumed that there is no measurement error in the process $\{Z(s), s \in D\}$. This has two meanings. The first meaning is to assume the random variable $Z(s)$ is observed at the exact location s and measurement errors in location s are ignored (Stein, 1999), which makes sense and this assumption is always kept in kriging. The second meaning is to assume either that the random variable $Z(s)$ is continuous without error, or that the discontinuous nugget effect structure in $Z(s)$ is because of micro-scale variation (Cressie, 1993). Under this assumption, kriging gives exact interpolation at the observed locations since it is the best linear unbiased estimation. But sometimes this is not realistic. The nugget effect can also be considered as a measurement error at the observed location (Dubrule, 1983). Stein (1999) presumes all the discontinuous nugget effect structure is due to measurement error, and it can not be ignored even if it is very small. Under this circumstance kriging has to

filter the measurement errors at the observed location instead of interpolating the noisy data, and the nugget effect can be considered as a smoothing parameter to control the degree of smoothness of the data (Billings et al., 2002a), which is called *filtered kriging* (Chiles and Delfiner, 1999; Cressie, 1993).

Since simple kriging and ordinary kriging are special cases of universal kriging, I only discuss universal kriging as an example here. Considering a geostatistical model $Z(s) = \mu(s)+e(s)$, $\mu(s_i) = \sum_{l=0}^{L} a_l f_l(s_i)$ is an unknown trend, where $f_l(s_i)$ are known functions and $f_0(s_i) = 1$; a_l are unknown parameters; $e(s)$ is the zero-mean random function at s. If the process contains a nugget effect, called $\epsilon(s)$, then the model becomes $Z(s) = \mu(s)+e(s)+\epsilon(s)$. Assume $S(s) = \mu(s)+e(s)$, then we have $Z(s) = S(s)+\epsilon(s)$. The nugget effect term has the following properties: $E(\epsilon) = 0$ (nonsystematic), $Cov(\epsilon S) = 0$ (additive), $Cov(\epsilon_i\epsilon_j) = 0$ for $i \neq j$ (mutually independent). If the nugget effect is considered as a micro-scale variation, the prediction is for $Z(s)$, which is called kriging, as I discussed in chapter 2. If the nugget effect is considered as a measurement error, the prediction is for a noiseless process $S(s)$, which is called *filtered kriging*. The following gives basic formulas in regular kriging and filtered kriging, where regular kriging is for $Z(s)$ and filtered kriging is for $S(s)$.

Regular universal kriging

The model is defined as

$$Z(s_i) = S(s_i) + \epsilon(s_i), i = 0, 1, 2..., n. \tag{6.1}$$

It is noted that $Var[S(s_i)] = \sigma^2$, $Var[\epsilon_i] = \sigma_\epsilon^2$, and $Var[Z(s_i)] = \sigma^2 + \sigma_\epsilon^2$. The universal kriging at site s_0 is predicted by $\hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i)$, when the estimation $E[Z(s_0)-\hat{Z}(s_0)]^2$ is minimized subject to the constraint $\sum_{i=1}^{n} w_i f_l(s_i) = f_l(s_0)$, for $l = 0, 1, ..., L$. Let $m_l, l = 0, 1, ..., L$ be the Lagrange parameters, the Lagrange formalism is defined as

$$g(w_i, i = 1, ..., n; m_l, l = 0, ..., L) = E[Z(s_0) - \hat{Z}(s_0)]^2 + 2\sum_{l=0}^{L} m_l[\sum_{i=1}^{n} w_i f_l(s_i) - f_l(s_0)].$$

Define $C_Z(s_i, s_j) = Cov[Z(s_i), Z(s_j)]$. By setting the partial derivatives of $g(w_i, i = 1, ..., n; m_l, l = 0, ..., L)$ with respect to w_i and m_l, then

$$\frac{\partial g}{\partial m_l} = 0 \Longrightarrow \sum_{i=1}^{n} w_i f_l(s_i) = f_l(s_0),\ l = 0, 1, ..., L;$$
$$\frac{\partial g}{\partial w_i} = 0 \Longrightarrow C_Z(s_i, s_0) = \sum_{j=1}^{n} w_j C_Z(s_i, s_j) + \sum_{l=0}^{L} m_l f_l(s_i),\ i = 1, ..., n.$$

Hence, the variance of universal kriging is

$$\sigma_k^2 = E[Z(s_0) - \hat{Z}(s_0)]^2 = \sigma^2 + \sigma_\epsilon^2 - \sum_{i=1}^{n} w_i C_Z(s_i, s_0) + \sum_{l=0}^{L} m_l f_l(s_0).$$

It also can be expressed as a form with variogram

$$\begin{cases} \sum_{i=1}^{n} w_i f_l(s_i) = f_l(s_0),\ l = 0, 1, ..., L; \\ \gamma_Z(s_i, s_0) = \sum_{j=1}^{n} w_j \gamma_Z(s_i, s_j) - \sum_{l=0}^{L} m_l f_l(s_i),\ i = 1, ..., n. \end{cases}$$

The variance of universal kriging is

$$\sigma_k^2 = \sum_{i=1}^{n} w_i \gamma_Z(s_i, s_0) + \sum_{l=0}^{L} m_l f_l(s_0).$$

Filtered universal kriging

The model is defined as

$$Z(s_i) = S(s_i) + \epsilon(s_i), i = 0, 1, ..., n.$$

The filtered universal kriging at s_0 is $\hat{S}(s_0) = \sum_{i=1}^{n} w_i^* Z(s_i)$. The estimation $E[S(s_0) - \hat{S}(s_0)]^2$ is minimized subject to the constraint $\sum_{i=1}^{n} w_i^* f_l(s_i) = f_l(s_0)$, for $l = 0, 1, ..., L$. Let $m_l, l = 0, 1, ..., L$ be the Lagrange parameters, the Lagrange formalism is defined as

$$g(w_i^*, i = 1, ..., n; m_l, l = 0, ..., L) = E[S(s_0) - \hat{S}(s_0)]^2 + 2\sum_{l=0}^{L} m_l [\sum_{i=1}^{n} w_i^* f_l(s_i) - f_l(s_0)].$$

Define $C_S(s_i, s_j) = Cov[S(s_i), S(s_j)]$, also note $Cov[Z(s_i), S(s_j)] = C_S(s_i, s_j)$. By setting the partial derivatives of $g(w_i^*, i = 1, ..., n; m_l, l = 0, ..., L)$ with respect to w_i^* and m_l, then

$$\frac{\partial g}{\partial m_l} = 0 \Longrightarrow \sum_{i=1}^{n} w_i^* f_l(s_i) = f_l(s_0),\ l = 0, 1, ..., L.$$
$$\frac{\partial g}{\partial w_i^*} = 0 \Longrightarrow C_S(s_i, s_0) = \sum_{j=1}^{n} w_j^* C_Z(s_i, s_j) + \sum_{l=0}^{L} m_l f_l(s_i),\ i = 1, ..., n.$$

Hence, the variance of filtered universal kriging is

$$\sigma_{fk}^2 = E[S(s_0) - \hat{S}(s_0)]^2 = \sigma^2 - \sum_{i=1}^{n} w_i^* C_S(s_i, s_0) + \sum_{l=0}^{L} m_l f_l(s_0).$$

It also can be expressed as a form with variogram

$$\begin{cases} \sum_{i=1}^{n} w_i^* f_l(s_i) = f_l(s_0),\ l = 0, 1, ..., L. \\ \gamma_S(s_i, s_0) = \sum_{j=1}^{n} w_j^* \gamma_Z(s_i, s_j) - \sum_{l=0}^{L} m_l f_l(s_i),\ i = 1, ..., n. \end{cases}$$

The variance of filtered universal kriging is

$$\sigma_{fk}^2 = \sum_{i=1}^{n} w_i^* \gamma_S(s_i, s_0) + \sum_{l=0}^{L} m_l f_l(s_0) - \sigma_\epsilon^2.$$

Relations between regular kriging and filtered kriging

The relations between $C_Z(s_i, s_j)$ and $C_S(s_i, s_j)$ are as follows:

$$\begin{cases} C_Z(s_i, s_j) = C_S(s_i, s_j) & \text{if } s_i \neq s_j, \\ C_Z(s_i, s_j) = C_S(s_i, s_j) + \sigma_\epsilon^2 & \text{if } s_i = s_j. \end{cases} \tag{6.2}$$

It also can be expressed as a form with variogram as follows:

$$\begin{cases} \gamma_Z(s_i, s_j) = \gamma_S(s_i, s_j) & \text{if } s_i = s_j, \\ \gamma_Z(s_i, s_j) = \gamma_S(s_i, s_j) + \sigma_\epsilon^2 & \text{if } s_i \neq s_j. \end{cases} \tag{6.3}$$

Hence, at an unsampled site s_0, the predicted values in regular kriging and filtered kriging are the same, the only difference is they have different kriging variances, i.e., $\sigma_k^2 = \sigma_{fk}^2 + \sigma_\epsilon^2$. If the predicted site s_0 is one of the observed sites in $s_1, ..., s_n$, then the predicted value of regular kriging is the exact value at this site and $\sigma_k^2 = 0$; the predicted value of filtered kriging is different from the original value and smooths the value at this site, with larger values of σ_ϵ^2 resulting in more smoothing. Under the extreme case where all the variation is because of measurement error, the predicted value is just the average of all the known values (Cressie, 1993).

6.2.2 Dual kriging and spline smoothing

Consider a general spatial map $\{z(s) : s \in D\}$, $s = (x, y)$ denotes the coordinates of the sample site (x, y); D denotes the set of the region of interest, i.e. $s_1, ..., s_n$. There are two different but related smoothing approaches for this map (Watson, 1984). One is the *random function method*, i.e., kriging, which considers $\{z(s) : s \in D\}$ as a realization of a random function $\{Z(s) : s \in D\}$ with or without measurement error, and computes the best linear unbiased estimator to obtain a map as accurate as possible; the other is the *deterministic function method*, i.e., thin plate spline method, which uses a deterministic function to fit the map $\{z(s) : s \in D\}$ as "closely" as possible, and uses a penalized term to adjust the smoothness of the fitting.

Dual kriging

For a universal kriging model

$$Z(s_i) = \mu(s_i) + e(s_i) + \epsilon(s_i), \tag{6.4}$$

where $\mu(s_i) = \sum_{l=0}^{L} a_l f_l(s_i)$ is an unknown trend, $\epsilon(s_i)$ is a measurement error with $E[\epsilon(s_i)] = 0$ and $Var[\epsilon(s_i)] = \sigma_\epsilon^2$. The model can be expressed in matrix form as follows (Wackernagel, 2003). Define $\mathbf{Z} = [Z(s_i)]$ as a $n \times 1$ vector, $\mathbf{C} = [C(s_i, s_j)]$ as a $n \times n$ matrix, $\mathbf{c} = [C(s_i, s_0)]$ as a $n \times 1$ vector, $\mathbf{F} = [f_0(s_i), ..., f_L(s_i)]$ as a $n \times (L+1)$ matrix, $\mathbf{f} = [f_l(s_0)]$ as a $(L+1) \times 1$ vector. The noiseless universal kriging is $\hat{S}(s_0) = \mathbf{w}^T\mathbf{Z}$, where the parameters of weights $\mathbf{w}$ and Lagrange parameters $\mathbf{m}$ can be estimated by

$$\begin{bmatrix} \mathbf{C} + \sigma_\epsilon^2\mathbf{I} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{f} \end{bmatrix},$$

where $\mathbf{w} = [w_i]$ is a $n \times 1$ vector, $\mathbf{m} = [m_l]$ is a $(L+1) \times 1$ vector. Since the left hand side matrix does not depend on s_0, the formula of kriging can also be expressed in an alternative formulation, which is called *dual-kriging* (Journel, 1989; Cressie, 1993; Wackernagel, 2003).

Define an inverse matrix

$$\begin{bmatrix} \mathbf{S} & \mathbf{G} \\ \mathbf{G}^T & \mathbf{T} \end{bmatrix} = \begin{bmatrix} \mathbf{C} + \sigma_\epsilon^2 \mathbf{I} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix}^{-1} \Longrightarrow \begin{bmatrix} \mathbf{w} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{S} & \mathbf{G} \\ \mathbf{G}^T & \mathbf{T} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{f} \end{bmatrix}.$$

Note $\mathbf{w} = \mathbf{Sc} + \mathbf{Gf}$, if $\hat{S}(s_0) = \mathbf{w}^T\mathbf{Z} = \mathbf{Z}^T\mathbf{w}$, then $\hat{S}(s_0) = \mathbf{Z}^T\mathbf{Sc} + \mathbf{Z}^T\mathbf{Gf}$. Define $\mathbf{b}^T = \mathbf{Z}^T\mathbf{S}$ and $\mathbf{a}^T = \mathbf{Z}^T\mathbf{G}$, which both do not dependent on s_0, then

$$\begin{bmatrix} \mathbf{S} & \mathbf{G} \\ \mathbf{G}^T & \mathbf{T} \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} \Longrightarrow \begin{bmatrix} \mathbf{C} + \sigma_\epsilon^2 \mathbf{I} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix}.$$

Then the kriging can be expressed as the sum of the trend and a linear combination of the covariance structure.

$$\hat{S}(s_0) = \begin{bmatrix} \mathbf{c}^T & \mathbf{f}^T \end{bmatrix} \begin{bmatrix} \mathbf{C} + \sigma_\epsilon^2 \mathbf{I} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix}.$$

Thin-plate spline

Thin-plate spline refers to a physical analog involving the bending of a thin sheet of metal (Billings et al. 2002a), which is a two-dimensional analog of the cubic spline in one dimension. For $Z(s_i) = z(s_i) + \epsilon$, $i = 0, 1, ..., n$, the predicted value

$$\hat{Z}(s_0) = \sum_{i=1}^{n} b_i e(s_i, s_0) + \sum_{l=0}^{3} a_l f_l(s_0),$$

where $e(s_i, s_0) = ||s_i - s_0||^2 \log(||s_i - s_0||^2)$. Define $\mathbf{Z} = [Z(s_i)]$ as a $n \times 1$ vector, $\mathbf{E} = [e(s_i, s_j)]$ as a $n \times n$ matrix, $\mathbf{e} = [e(s_i, s_0)]$ as a $n \times 1$ vector, $\mathbf{b} = [b_i]$ as a $n \times 1$ vector, $\mathbf{F} = [f_0(s_i), ..., f_3(s_i)]$ as a $n \times 4$ matrix, $\mathbf{f} = [f_l(s_0)]$ as a 4×1 vector, $\mathbf{a} = [a_l]$ as a 4×1 vector. The penalized sum of squares criterion

$$(\mathbf{Z} - \mathbf{Eb} - \mathbf{Fa})^\mathbf{T}(\mathbf{Z} - \mathbf{Eb} - \mathbf{Fa}) + \lambda \mathbf{b}^\mathbf{T}\mathbf{Eb}$$

is minimized when

$$\begin{bmatrix} \mathbf{E} + \lambda \mathbf{I} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix},$$

where the nonnegative parameter λ controls the degree of smoothness, and measures the trade-off between the goodness of fit to the data and the roughness of the surface (Cressie, 1993), which is estimated by minimizing the generalized cross-validation. Then

$$\hat{Z}(s_0) = \begin{bmatrix} \mathbf{e}^T & \mathbf{f}^T \end{bmatrix} \begin{bmatrix} \mathbf{E} + \lambda \mathbf{I} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z} \\ \mathbf{0} \end{bmatrix}.$$

Relations between kriging and spline-smoothing

Roughly, the spline method is just a special case of kriging (Watson, 1984), which amounts to replacing the uncertain covariance function by a fixed smooth function. Hence, the spline method is equivalent to kriging with a given covariance (Dubrule, 1983). The penalized parameter λ can be considered as an alternative measurement error added to the variances at data locations, but not to the estimation locations (Wackernagel, 2003).

The smoothing patterns in the two methods are quite similar. In the kriging method, $E[Z - \hat{S}]^2 = E[S - \hat{S}]^2 + \sigma_\epsilon^2$, where $E[S - \hat{S}]^2/\sigma_\epsilon^2$ is controlled by the measurement error. In the spline method, $E[Z - \hat{Z}]^2 = Var[\hat{Z}] + bias[\hat{Z}]^2$, where λ controls the trade-off between variance and bias of the estimator.

6.2.3 Smoothing ratio (SR)

Regular signal to noise ratio (SNR)

The term *signal to noise ratio* (SNR) has been widely used in many areas, but it also has different definitions and meanings in different contexts, which may cause confusion. To classify my definition of smoothing ratio in a statistical way, I introduce the regular signal to noise ratio at first.

In imaging processing, the *signal to noise ratio* is usually estimated by the ratio of the signals' mean to the square root of variability in the data (Curran and Dungan, 1988; Huettel et al., 2004). In geostatistics, it is estimated by the signals' mean divided by the square root of the nugget effect, which is called "geostatistical method" in the literature (Van Der Meer and

De Jong, 2001; Curran and Dungan, 1988). This method estimate the intra-signal variability and the random sensor noise by using the variogram model fitting from a transect of signals, and considers the estimated nugget effect as the variability. In fMRI data analysis, the raw signal is not as important as the magnitude of the intensity difference in the signal between different conditions. Hence researchers usually use *functional signal to noise ratio* in fMRI. *Functional signal to noise ratio*, also called *dynamic contrast to noise ratio*, is defined as the difference between two states of the signal divided by the square root of the variability (Huettel et al, 2004). All the above *signal to noise ratios* have a common property, that is, noise varies little but signal will change by states and time (Curran and Dungan, 1988; Huettel et al., 2004). Hence, *signal to noise ratio* varies across different states and time.

In statistics, signal to noise ratio has totally different meanings and properties, and it is usually defined as a noise to signal ratio. To avoid confusion, I call it *smoothing ratio*(SR) in the following introduction.

Smoothing ratio (SR) in spline smoothing

In the spline smoothing method, the degree of smoothness can be measured by the *smoothing ratio* (SR) (Zhen and Basher, 1995), which provides a good tool to control the roughness of a fitted model. *Smoothing ratio* is defined as the ratio of the degrees of freedom for the noise to the degrees of freedom for the signal. It also can be considered as an alternative *noise to signal ratio* in statistics.

Following Wahba (1990), define

$$[\hat{Z}(s_1), ..., \hat{Z}(s_n)]^T = \mathbf{A}(\lambda)[Z(s_1), ..., Z(s_n)]^T,$$

where $\mathbf{A}(\lambda)$ is a smoothing matrix controlled by λ. Mean square residual (MSR) is

$$R(\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}[Z(s_i) - \hat{Z}(s_i)]^2.$$

Estimated error variance is

$$\hat{\sigma}^2(\hat{\lambda}) = nR(\hat{\lambda})/Tr[\mathbf{I} - \mathbf{A}(\hat{\lambda})].$$

Expectation of the predictive mean square error (PMSE) is

$$E[T(\hat{\lambda})] = E\left[\frac{1}{n}\sum_{i=1}^{n}[\hat{Z}(s_i) - E(\hat{Z}(s_i))]^2\right] = \hat{\sigma}^2(\hat{\lambda}) - R(\hat{\lambda}).$$

Then the *smoothing ratio*(SR) is

$$\begin{aligned} SR &= MSR/PMSE \\ &= R(\hat{\lambda})/[\hat{\sigma}^2(\hat{\lambda}) - R(\hat{\lambda}))] \\ &= Tr[\mathbf{I} - \mathbf{A}(\hat{\lambda})]/Tr[\mathbf{A}(\hat{\lambda})], \end{aligned} \tag{6.5}$$

where $Tr[\mathbf{A}(\hat{\lambda})]$ is the degrees of freedom for signal, $Tr[\mathbf{I} - \mathbf{A}(\hat{\lambda})]$ is the degrees of freedom for error. In the spline smoothing methods, the error variance is determined by the penalized parameter λ in the generalized cross validation. Actually, *smoothing ratio* is a monotonic function of the penalized parameter λ, but it provides a useful meaning to understand and control the nature of the surface fitting (Zhen and Basher, 1995).

Smoothing ratio (SR) in filtered kriging

Since filtered kriging and spline smoothing are consistent for a given covariance structure, I introduce the *smoothing ratio* from spline smoothing to filtered kriging. For the filtered kriging $\hat{S}(s_i) = \mathbf{w}_i^T\mathbf{Z}$, $i = 1, ...n$, write

$$[\hat{S}(s_1), ..., \hat{S}(s_n)]^T = \mathbf{W}[Z(s_1), ..., Z(s_n)]^T,$$

where the weight matrix $\mathbf{W} = (\mathbf{w}_1^T, ..., \mathbf{w}_n^T)^T$. Similarily, the degrees of freedom for signal is $Tr[\mathbf{W}]$, which is the sum of the weights at the original sample sites; the degrees of freedom for noise is $Tr[\mathbf{I} - \mathbf{W}]$, which is the sum of the weights at the neighbors of the original sites. Hence the *smoothing ratio*(SR) in the filtered kriging is defined as

$$SR = Tr[\mathbf{I} - \mathbf{W}]/Tr[\mathbf{W}], \tag{6.6}$$

which is a function controlled by the nugget effect σ_ϵ^2 in the variogram .

In the spline smoothing method, the smoothing matrix $\mathbf{A}(\lambda)$ does not have a physical meaning, it just provides considerable computational convenience for calculations (Wahba,

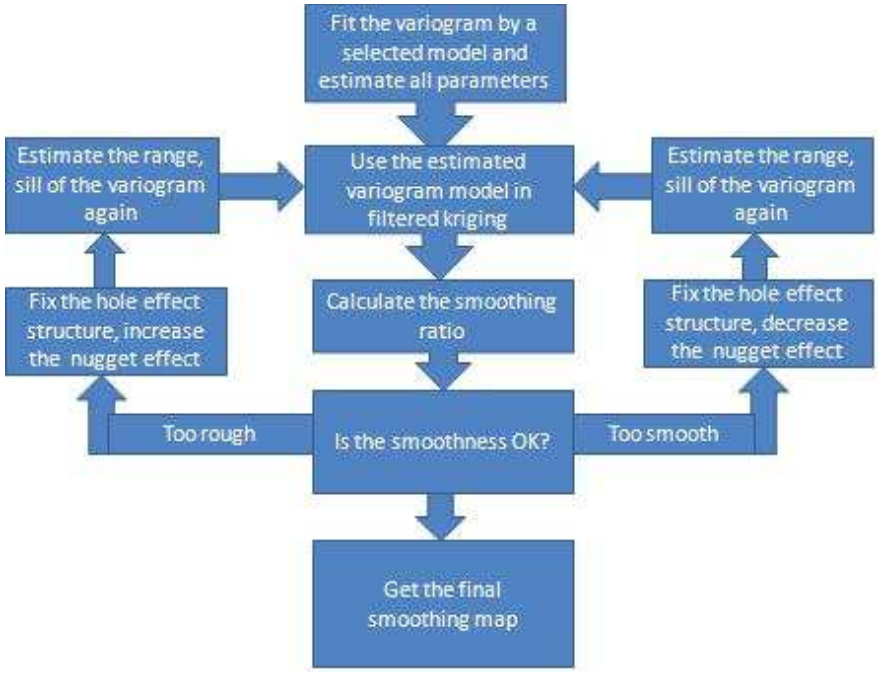

Figure 6.1: Steps of filtered kriging by controlling the smoothing ratio

1990; Zhen and Basher, 1995). But in filtered kriging, the study of the kriging weights is very informative. The weight matrix **W** has a clear statistical meaning and gives a better interpretation for the *smoothing ratio*. Here the *smoothing ratio* can be considered as the contribution of the neighbors in the prediction of the attribute at the original sampled sites. If a variogram only has nugget effect structure, then the filtered kriging gives the sample mean at the sampled sites (Schabenberger and Pierce, 2002); if a variogram has no nugget effect, then the filtered kriging interpolates the data at the predicted sites.

Steps of filtered kriging by controlling the smoothing ratio

In the variogram modeling, all the parameters, i.e., sill, range, nugget effect and number of sign changes are estimated at the same time by the least squares. When this estimated variogram is used in filtered kriging, different parameters have different effects in the process, i.e., number of sign changes determines the hole effect structure of the data; range determines

the effective neighbors of the target point; nugget effect determines the measurement error of the target point; sill controls the kriging variance. Actually, number of sign changes, range and nugget effect are more important in the filtered kriging, because they are directly related to the predicted (smoothed) value at the target point.

Figure 6.1 gives the steps of filtered kriging by controlling the smoothing ratio. Note when smoothing ratio is considered in the procedure, the case is a little bit different. If we change the nugget effect only, there must be literally an infinite number of possibilities of the other estimated parameter values, i.e., different combinations of the range and the number of sign changes, leading to the same smoothing ratio. Hence I fix the number of sign changes, and only estimate nugget effect, range and sill again, as shown in Figure 6.1. Because sill is not related to the predicted (smoothed) value, and the MSR is monotonically increasing as the nugget effect increases, we have a relationship between nugget effect and range for a fixed hole effect structure (number of sign changes). By changing the different measurement errors (nugget effects), we can reassign the different weights among the neighbors of the target point again and again, i.e., get different estimated ranges, thereafter find an ideal smoothing ratio in the filtered kriging.

Temporal consistency of the smoothing ratio

Keeping the *temporal consistency* of the smoothing ratio is an important consideration for smoothing when a sequence of time related data sets is being considered (Zhen and Basher, 1995), which is different from regular signal to noise ratio. For example, when we try to krige a sequence of time related data, the estimated nugget effect at each separate time point may vary. This will result in the degrees of the smoothness at these time points being different. Because we usually want all the kriged maps to have the same degree of smoothness for any further analyses, we may first estimate the variogram model and do filtered kriging in a sequence of temporal data respectively, then identify the value of the smoothing ratio at each time point which provides an acceptable result. After that, adjust the measurement

error σ_ϵ^2, and characterize the model fitting at each time point to obtain the desired value of smoothing ratio.

6.3 Data analysis

I choose time points 61, 64, 65, 70, 71, 76, 80, 83, 85, 89 at the fourth slice with variogram model BG-200 as examples for demonstration purposes.

6.3.1 Relations between nugget and smoothing ratio

To see the relations between nugget and smoothing ratio in filtered kriging, I arbitrarily choose time point 64 as a first example. Figure 6.2 shows BG-200 variogram model fits in the x and y directions with fixed $b = 1.9986$ and nugget effect $\sigma_\epsilon^2 = 0, 2, 4, 6, 12, 24, 36, 48$ respectively. The effective lag distance is chosen as 19 in both x and y directions in the model fitting. The estimated values of the other parameters are listed in Table 6.1. Next I use the estimated variogram to do the filtered kriging on the 10 time points; the kriged maps are shown in Figure 6.3. When the nugget effect $\sigma_\epsilon^2 = 0$, filtered kriging interpolates the map. When the nugget effect $\sigma_\epsilon^2 > 0$, filtered kriging smooths the map. The degree of smoothness increases as the value of the nugget effect increases. The different nugget effects of the BG-200 model and the corresponding smoothing ratios in filtered kriging at time point 64 are shown in Table 6.1. Note that the smoothing ratio is a monotonic function of the nugget effect. A larger nugget effect results in more weights on the neighbors of the target points and therefore results in a larger smoothing ratio. It is also noted that mean squared residual (MSR) is a monotonic function and estimated ranges of the variogram do not change too much. I will discuss this issue later.

In geostatistics, the nugget effect is just interpreted as a discontinuous nugget at the origin of the variogram model; it has no intrinsic meaning (Figure 6.2). In filtered kriging, the nugget effect is considered to be measurement error and it prevents the regular kriging procedure from degenerating into exact interpolation (Figure 6.3). But the smoothing ratio

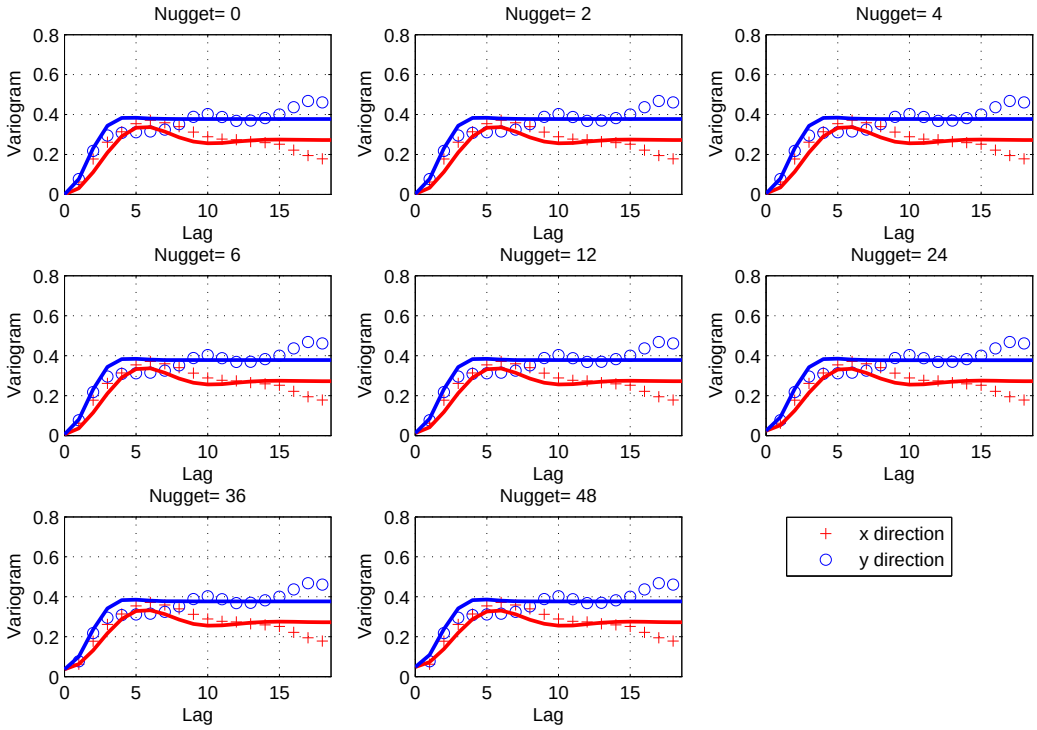

Figure 6.2: At time point 64, BG-200 variogram model fits in x and y directions with fixed $b = 1.9986$ and nugget effect $\sigma_\epsilon^2 = 0, 2, 4, 6, 12, 24, 36, 48$ respectively. The effective lag distance is chosen as 19 for both x and y directions in the model fitting. The estimated values of the other parameters are listed in Table 6.1. Note the sills and ranges in the two directions only have minor changes as the nugget effect changes.

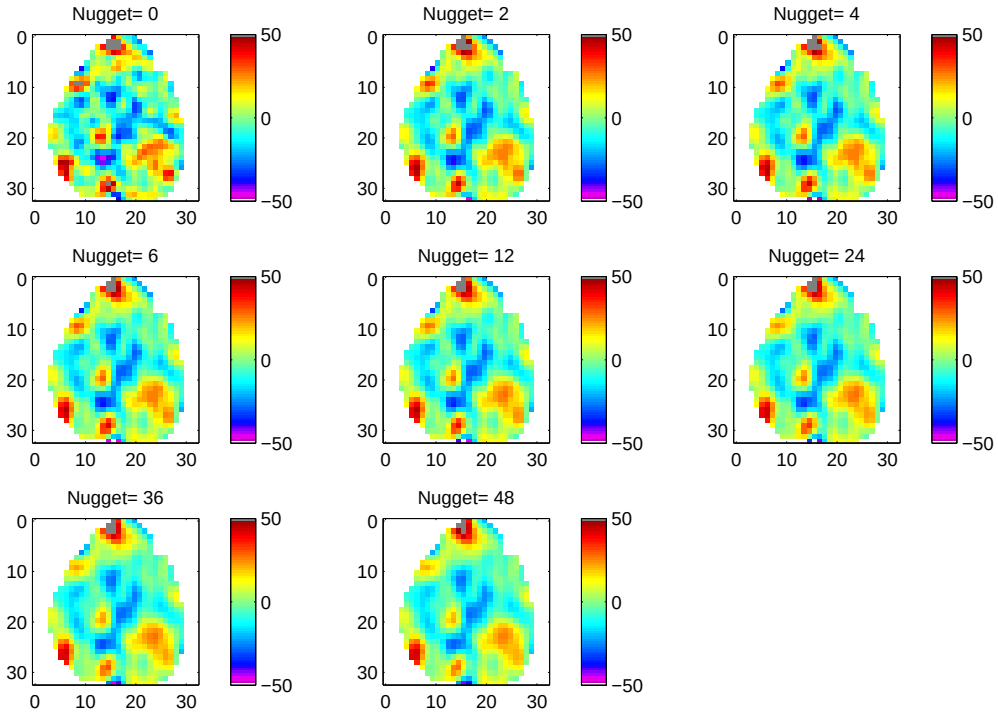

Figure 6.3: At time point 64, filtered kriging maps for different nugget effects 0, 2, 4, 6, 12, 24, 36, 48. When nugget effect $\sigma_\epsilon^2 = 0$, filtered kriging interpolates the map. When nugget effect $\sigma_\epsilon^2 > 0$, filtered kriging smooths the map. The degree of smoothness increases as the nugget effect increases.

SR	MSR	Nugget	Sill	Range in x	Range in y	b
0	0	0	324.7866	2.6856	0.9344	1.9986
2.7602	27.7430	2	324.7699	2.6975	0.9385	1.9986
3.1727	33.6589	4	324.7532	2.7094	0.9426	1.9986
3.4355	37.0032	6	324.7358	2.7215	0.9466	1.9986
3.9345	43.0341	12	324.6809	2.7581	0.9557	1.9986
4.6083	50.9338	24	324.5935	2.8173	0.9837	1.9986
5.1394	57.7743	36	324.4915	2.8840	1.0065	1.9986
5.6252	64.2208	48	324.4292	2.9514	1.0297	1.9986

Table 6.1: At time point 64, the different parameters of the BG-200 model and the corresponding smoothing ratios and MSRs in filtered kriging. Note smoothing ratio (SR) is a monotonic function of the nugget effect. For each different nugget effect, b is fixed and the other parameters of the variogram model are estimated. The values of these other parameters do not change much. A larger nugget results in more weights on the neighbors of the target points and therefore results in a larger SR and a larger MSR.

has a more clear statistical meaning than the nugget effect; it is interpreted as the weight that the neighbors of a target point contribute compared with the weight of the target point (Figure 6.3). When the smoothing ratio is small, the neighbors contribute less weight to the target point, the map is less smoothed. The extreme case is when the smoothing ratio equals to 0, which means the map is just interpolated and the neighbors of each target point contribute nothing. When the smoothing ratio is larger, the neighbors contribute more weights to the target point, and hence the map is more smoothed. The extreme case is that all the points, including each target point and its neighbors, contribute the same weights in the estimated map. So the map is just averaged by all points.

6.3.2 Temporal consistency of the smoothing ratio

As I mentioned earlier, keeping the temporal consistency in filtered kriging is very important. Here I first give the estimated smoothing ratios and the corresponding parameters of the BG-

200 variogram model at the 10 different time points (Table 6.2). Note the smoothing ratios vary from 1.2525 to 9.4196 because of the different estimated nugget effects. The average of the 10 different smoothing ratios is 4.4861. The filtered kriging maps at the 10 different time points are shown in Figure 6.4. Since the ratios are different, the filtered kriging maps at the 10 different time points exhibit substantial variability of smoothness. Compared with other time points, the maps at time points 61, 71 are oversmoothed; the maps at time points 76, 83 are undersmoothed (Figure 6.4). Hence it is necessary to have temporal consistency in smoothing. Since the mean value of the smoothing ratio in Table 6.2 is around 4.5, I adjust all smoothing ratios to this value at each time point by specifying different nugget effects and other parameters in the BG-200 model (Table 6.3). The filtered kriging maps at the 10 different time points with an adjusted smoothing ratio 4.50 are shown in Figure 6.5. These maps have a consistent degree of smoothness, which should benefit further analysis.

Time point	SR	Nugget	Sill	Range in x	Range in y	b
61	9.4196	53.5500	438.2500	23.0823	3.6586	2.2061
64	4.5352	24.1000	324.8000	9.1949	2.8945	1.9986
65	5.3383	25.9000	307.8000	3.2488	10.8047	2.0803
70	3.0922	22.9500	282.8500	5.5177	2.6077	1.8173
71	7.4068	45.9000	296.5000	2.5222	16.9359	1.4084
76	1.2521	7.7000	352.9000	2.6443	3.2528	1.9264
80	4.2186	34.3500	408.7000	5.5902	3.4581	1.7075
83	1.5253	7.8000	546.1000	3.2014	3.5067	2.0169
85	3.9757	30.6000	294.3500	2.9028	6.4040	1.8466
89	4.0974	21.2500	264.3000	6.3621	3.4596	2.1461

Table 6.2: The smoothing ratio (SR) and the corresponding parameters of the BG-200 variogram model at the 10 different time points. Note the SRs are different because of the different estimated nugget effects. The average value of the ten SRs is 4.4861.

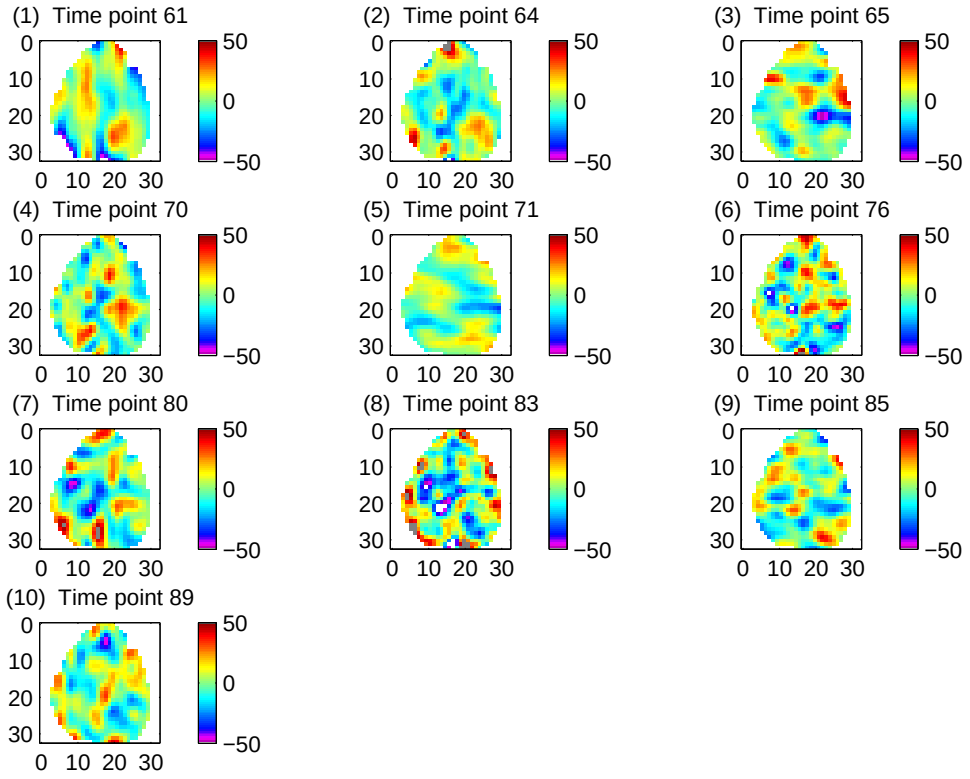

Figure 6.4: The filtered kriging maps at the 10 different time points. Since the smoothing ratios are different, the brain maps at the 10 different time points exhibit substantial variability of smoothness. Obviously, the maps at time points 61, 71 are oversmoothed; the maps at time points 76, 83 are undersmoothed, compared with other time points. Hence it is necessary to have temporal consistency in smoothing.

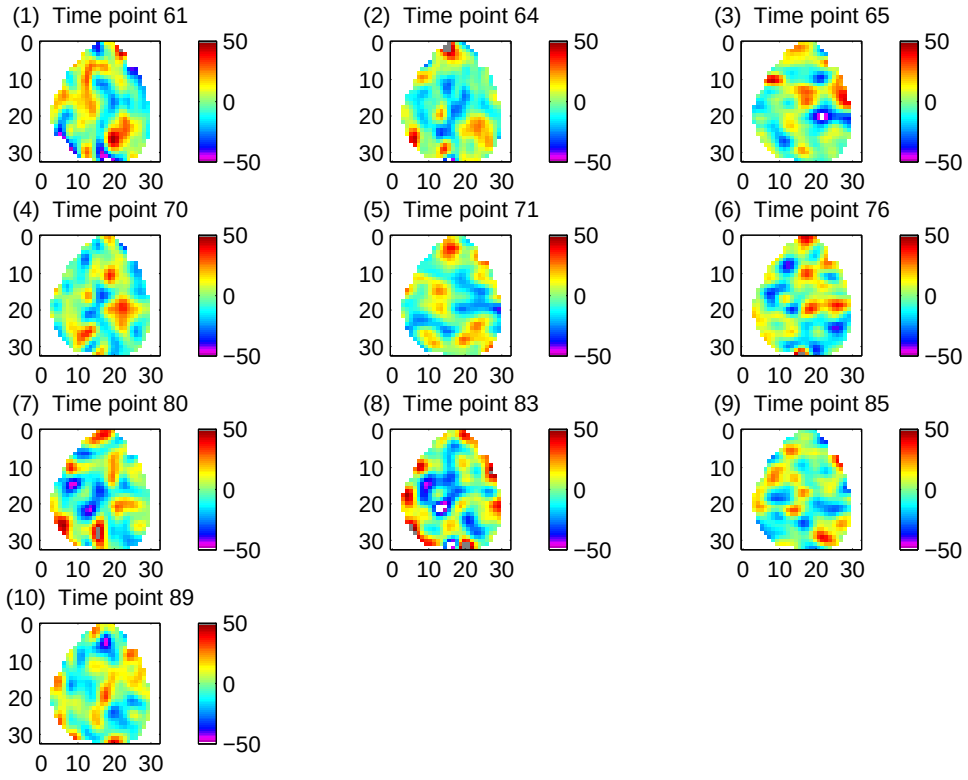

Figure 6.5: The filtered kriging maps at the 10 different time points with an adjusted smoothing ratio 4.50. These maps have a consistent degree of smoothness, which should benefit further analysis.

Time point	SR	Nugget	Sill	Range in x	Range in y	b
61	4.5047	2.3000	436.9759	15.6791	3.6593	2.2061
64	4.5021	21.5000	324.6026	8.7420	3.0356	1.9986
65	4.4993	14.1000	307.7412	3.3626	8.9873	2.0803
70	4.5038	40.600	283.8281	5.4552	3.4617	1.8173
71	4.6176	18.9000	297.4787	2.6707	9.0827	1.4084
76	4.5020	90.6000	350.5176	3.3371	5.0223	1.9264
80	4.5019	51.8000	408.8418	4.9444	3.8217	1.7075
83	4.4993	120.2000	542.4088	3.9937	4.5525	2.0169
85	4.4997	43.6000	294.3770	3.2400	5.9598	1.8466
89	4.5022	33.0000	264.1362	5.6935	3.8702	2.1461

Table 6.3: Since the mean value of the smoothing ratio (SR) in Table 6.2 is 4.4861, I adjust all SRs to almost 4.5 at each time point by specifying different nugget effects and other parameters in the BG-200 model. Since there are no predefined methods in choosing the nugget effect near the origin of the variogram model, the ideal SR will depend on the characteristics of the map and how much smoothness is required.

6.4 Discussion and conclusion

6.4.1 Discussion

Choice of nugget effect From above, the choice of a behavior near the origin of the variogram is very important in data smoothing. In geostatistics, it has to be predefined by the researcher rather than entirely automatically because the variogram itself is not defined at lag distance zero (Chiles and Delfiner, 1999). Since it is difficult to determine whether the modeled variogram has been successfully estimated near the origin by the least squares method, some papers (for example, Billings et al., 2002a and 2002b) recommend to instead choose the nugget effect by cross-validation to achieve a better prediction performance through a bias-variance trade-off. But I find at least in our fMRI data set, this is not an effective way. This claim is also indicated by Goovaerts (1997). He pointed out one of the disadvantages

of cross-validation method in geostatistics is the relative nugget effect and the variogram behavior at the origin can not be cross validated.

The reason is that rescaling the variogram does not affect the weights in kriging, so does the smoothing ratio. In kriging, only the range is important for the relations between the target point and its neighbors. Hence changing the nugget effect or sill of the variogram without changing the range does not affect the kriging value, it only affect the kriging variance (Schabenberger and Pierce, 2002).

In our problem, the estimated ranges in the x and y directions only change slightly as the nugget effect changes (Table 6.1). This means that the number of effective neighbors and their weights for the target points are almost the same. By Wahba (1990), the generalized cross-validation (GCV) is

$$GCV(\lambda) = (n^2 \cdot MSR)/Tr[\mathbf{I} - \mathbf{A}(\hat{\lambda})]^2$$

in spline smoothing. Similarly, GCV in filtered kriging is defined as

$$GCV = (n^2 \cdot MSR)/Tr[\mathbf{I} - \hat{\mathbf{W}}]^2.$$

Mean square residual (MSR) is a monotonically increasing function of the nugget effect σ_ϵ^2. Note the weight matrix $\mathbf{W}$ is only related to the range; when the range is almost constant with the change of nugget effect, $Tr[\mathbf{I} - \hat{\mathbf{W}}]$ does not vary much. Hence the GCV function is just a monotonic function.

σ_ϵ^2	2	4	6	12	24	36	48
MSR	27.7430	33.6589	37.0032	43.0341	50.9338	57.7743	64.2208
DF	0.7341	0.7603	0.7745	0.7973	0.8217	0.8371	0.8491
GCV	51.4870	58.2207	61.6800	67.6890	75.4377	82.4445	89.0836

Table 6.4: At time point 64, MSR, DF, GCV values for different nugget effects σ_ϵ^2. Since MSR is a monotonic function and DF only has minor changes, GCV is a monotonic function too. Therefore cross-validation does not work for our data set.

Table 6.4 lists the MSR, DF, GCV values for different nugget effects σ_ϵ^2 at time point 64, where DF is defined as $\frac{1}{n^2}Tr[\mathbf{I}-\hat{\mathbf{W}}]^2$. Since MSR is a monotonic function and DF only fluctuates a little bit, GCV is a monotonic function too. Therefore cross-validation does not work for our data set.

An exception is that we can use cross-validation to do model selection for different variogram models, because different variogram models may have different ranges. This was done in the previous chapter.

Since there are no predefined methods in choosing the nugget effect near the origin of the variogram model, the ideal smoothing ratio will depend on the characteristics of the data map and how much smoothness the researcher wants. Currently I have to seek an acceptable result by the prior knowledge subjectively to maintain a similar smoothness of the data map. I will consider the choice of nugget effect in variogram model fitting in my future work.

6.4.2 Conclusion

In geostatistics, kriging interpolates the original data and the measurement error is usually ignored. Here I consider filtered kriging as a smoothing method to filter the original data, which can remove the measurement errors at the observed sites. This method incorporates spatial information of the data in the smoothing procedure to reduce noise in an intelligent way. I also compare the filtered kriging with the thin plate spline smoothing method, and creatively borrow the idea of the smoothing ratio from spline smoothing to control the degree of smoothness in filtered kriging. Use of smoothing ratio gives a clear statistical meaning and it is easier to interpret than the nugget effect in filtered kringing. In fMRI data analysis, it is common to smooth a sequence of time related data sets. Since the estimated nugget effects may vary at different time points, the degrees of smoothness are different as well. After adjusting the smoothing ratio to a fixed number in a sequence of time related fMRI data, my results show a consistent degree of smoothness.

Chapter 7

Conclusion and future work

7.1 Conclusion

This dissertation provides an exploration of the application of geostatistical methods more broadly in fMRI data analysis.

Clustering in fMRI time series is used to investigate and discover the salient features of the fMRI data by the temporal structure in brain activity. Current methods - either directly clustering the time series or through characteristic features such as the cross-correlation with the experimental protocol signal has drawbacks: clustering of the time series themselves may identify voxels with similar temporal behavior that is unrelated to the stimulus, whereas cross-correlation requires knowledge of the stimulus presentation protocol. In Chapter 3, I propose the use of autocorrelation structure instead - an idea borrowed from geostatistics; this approach does not suffer from the deficits associated with previous clustering methods. I first formalize the traditional classification methods as three steps: feature extraction, choice of classification metric, and choice of classification algorithm. The use of different characteristics to effect the clustering (cross-correlation, autocorrelation, and so forth) relates to the first of these three steps. I then demonstrate the efficacy of autocorrelation clustering on a simple visual task, and on resting data. A byproduct of my analysis is the finding that masking prior to clustering, as is commonly done, may degrade the quality of the discovered clusters, and I offer an explanation for this phenomenon.

The use of autocorrelation structural analysis in clustering provides an attractive framework in data-driven analysis, but it still has a weakness, which is the lack of dimension reduction for the ill-balanced data. When the idea of the subtraction paradigm is used in the

dimension reduction step (Sommer and Wichert, 2002) some prior knowledge about the data is needed. Sparse principal component analysis (SPCA) (Zou et al., 2006) is a relatively new technique which can combine dimension reduction, feature extraction and clustering together. SPCA uses the idea of LASSO in dimension reduction without any pre-requirements of the data. Also, SPCA produces comparable clustering results to the autocorrelation structural analysis. But since the informative content of the time courses is not taken into account in the SPCA process, the main disadvantage of SPCA is that the outcomes need to be judged and interpreted, which is quite similar to ICA in fMRI. Hence structural analysis in Chapter 3 is still needed after SPCA. In Chapter 4, SPCA and autocorrelation structural analysis are jointly used to cluster fMRI time series. This purely model-free approach not only changes the whole clustering process to be data-driven, but also offers a well-grounded framework for data clustering. Since both techniques consider the correlations among the time courses, they provide consistent results and the efficiency of the clustering procedure is greatly improved. Chapter 4 also shows that masking the brain prior to clustering is not necessarily effective for dimension reduction, consistent with the conclusion in the previous chapter.

Chapter 5 changes the point of view to the spatial analysis of the fMRI data. An axial image of the brain is chosen for demonstration purpose. The structural analysis of the empirical variograms during different time points is executed at first. This procedure is almost the same as that in Chapter 3, but aims at finding different spatial patterns instead of different temporal patterns, as was done in previous chapters. The results reflect the activations of the brain by the experimental task, which gives a good understanding of how the brain reacts to the experimental task.

In variogram modeling, the current choice in fMRI is either simply using a monotonic function or changing the lag distance from the physical location distance to the measured signal distance. Both approaches have drawbacks: use of a monotonic function in structural analysis simplifies the problem but it considers the physical neighbors only and ignores the functional connections among the distant voxels; changing the definition of the lag distance

may violate the assumption that there is no measurement error in the lag distance, because the measured signals are not as precise as measured physical locations. Hence auxiliary data is usually needed for estimating the lag distance. The variogram model is called "hole effect" when it shows a cyclical pattern with a "down-hole". I show that the use of the hole effect variogram model offers advantages in describing the structure of the fMRI data. Unlike the previously proposed, my approach considers both the nonlinear physical distance and functional distance inside the brain and does not need any auxiliary data.

Chapter 6 considers *filtered kriging* as a smoothing method to filter the original data, which can remove the measurement errors at the observed sites. This method incorporates combining spatial information of the data in the smoothing procedure and can reduce the noise of the data in an intelligent way. I also compare filtered kriging with the thin-plate spline smoothing method, and creatively borrow the idea of the smoothing ratio from spline smoothing to control the degree of smoothness in filtered kriging. Use of the smoothing ratio gives a clear statistical meaning and it is easier to interpret than the use of nugget effect in filtered kringing. In fMRI data analysis, it is common to smooth a sequence of time related data sets. Since the estimated nugget effects may vary at different time points, the degrees of smoothness are different also. In filtered kriging, I adjust the smoothing ratio to a fixed number in a sequence of time related fMRI data, which takes into account the temporal information in the spatial analysis. My results show a consistent degree of smoothness among a sequence of time related data in the data processing procedure.

In summary, I mainly contribute two different geostatistical approaches to fMRI data analysis: one is concentrated on temporal analysis, i.e., clustering of fMRI time series; the other aims at spatio-temporal analysis of fMRI data, i.e., structural analysis and filtered kriging in a sequence of time related data sets. As a complement, two charts representing the methods I use in the dissertation are graphically presented in Figures 7.1 and 7.2.

Figure 7.1 includes the procedures covered in Chapter 3 and Chapter 4. By jointly using SPCA method and geostatistical method in the clustering procedure, my new method is

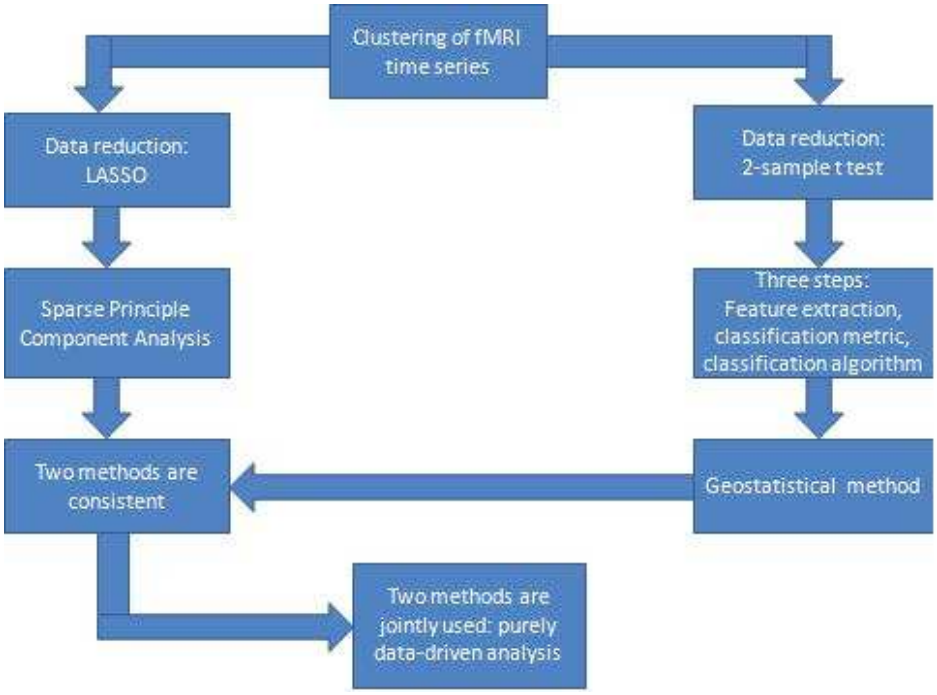

Figure 7.1: Steps in clustering of fMRI time series

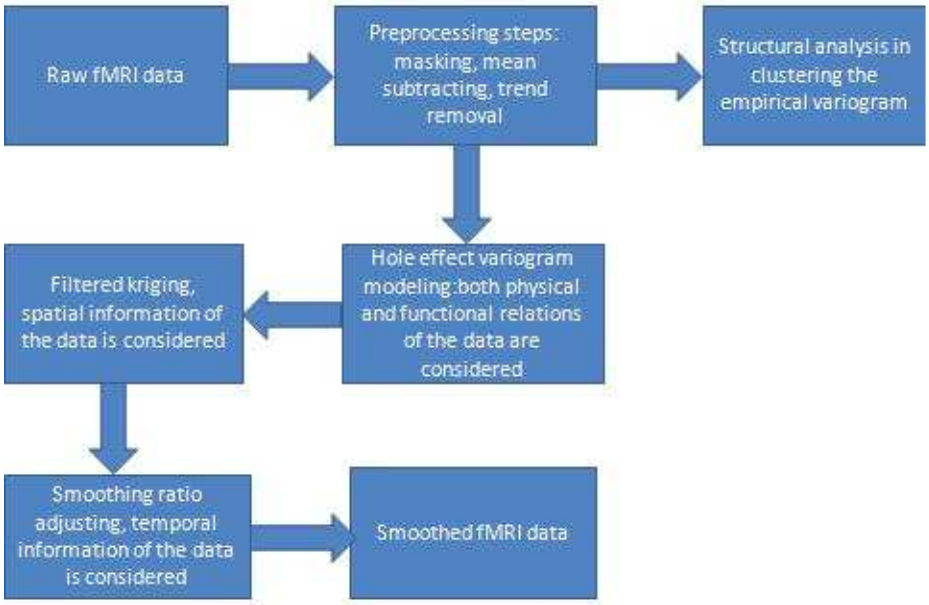

Figure 7.2: Steps in spatio-temporal analysis of fMRI data

purely data-driven and is superior to other current clustering methods in fMRI. Figure 7.2 gives a big picture of the smoothing procedure of a sequence of time related data sets. By considering the hole effect structure and controlling the smoothing ratio, both spatial and temporal information of the data are taken account in the smoothing procedure. Especially, the functional relations inside the brain are considered without any auxiliary data, which is new in fMRI. I also give a new explanation of the smoothing ratio based on the geostatistical view, which has not been seen in literature.

7.2 Future work

In the near future, I will consider the following post-dissertation works:

In the clustering of fMRI time series, I use the empirical autocorrelation as the main feature, which is easy to use and very popular in geostatistics. But voxel values at adjacent time points are likely correlated in the presence of noise. Modeling the autocorrelation may remove the noise and improve the precision of clustering results. No such parametric models have been proposed and proven to be valid for fMRI data. I will consider such models in future. Also, in the clustering of fMRI time series, the silhouette values may be negative in some clusters (e.g, Figure 4.12, graphs (5) and (6)), which means some voxels are not well-classified. I will investigate the relationships between the clustering method and number of negative silhouette values by simulations in the future.

In describing the physical and functional relations inside the brain, the choice of hole effect variogram is very important. But hole effect structural analysis is a relatively unexplored area and is usually ignored in geostatistics. How to well define a suitable variogram model and keep the balance between the physical and functional relations in the model fitting will be considered in future work.

Cokriging is the logical extension of kriging to situations where two or more variables are spatially interdependent. Hence cokriging can be considered as the multivariate case of kriging. Cokriging is the extension of kriging to the situation where auxiliary variables can

be used to improve the kriging estimate. In my opinion, cokriging can be applied in the following two different cases for fMRI: one is in Chapter 5, where only one slice of data is considered for kriging. It is possible to consider the whole volume or several relevant slices together by cokriging, where different slices are considered as different variables. The other is in Chapter 6. We may treat the spatio-temporal random function $Z(s,t)$ as a collection of a finite number T of temporally correlated spatial random functions $Z(s)$, i.e., the variables of cokriging are the different time points. Overall, we can address the kriging in multi-slice or multi-time studies. These will also be my future works.

Bibliography

Akritas M. G. and Arnold, S. F. (1994). Fully nonparametric hypotheses fro factorial designs I: multivariate repeated measures designs, *Journal of the American Statistical Association*, **89**, 336-343.

Atkinson, P. M. and Lewis, P. (2000). Geostatistical classification for remote sensing: an introduction, *Computers and Geostatistics*, **26**, 361-371.

Backfriender, W., Baumgartner, R., Stamal, M., Moser, E. and Bergmann, H. (1996). Quantification of intensity variations in functional MR images using rotated principal components, *Physics in Medicine and Biology*, **41**, 1425-1438.

Bandettini, P. A., Jesmanowicz, A., Wong, E. C. and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, **30**, 161-173.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation, *IEEE Transaction on Information Theory*, **37**, 1034-1054.

Baudelet, C. and Gallez, B. (2003). Cluster analysis of BOLD fMRI time series in tumors to study the heterogeneity of hemodynamic response to treatment. *Magnetic Resonance in Medicine*, **49**, 985-990.

Baumgartner R., Somorjai, R., Summers, R., Richter, W. and Ryner, L. (2000). Correlator beware: correlation has limited selectivity for fMRI data analysis, *NeuroImage*, **12**, 240-243.

Billings, S. D., Beatson, R. K. and Newsam, G. N. (2002a). Interpolation of geophysical data using continuous global surfaces, *Geophysics*, **67**, 1810-1822.

Billings,S. D., Newsam,G. N. and Beatson,R. K. (2002b). Smoothing fitting of geophysical data using continuous global surfaces, *Geophysics*, **67**, 1823-1834.

Bourgault, G., Marcotte, B. and Legendre, P. (1992). The multivariate (co)variogram as a spatial weighting function in classification methods, *Mathematical Geology*, **24**, 463-478.

Bowman, F. D., Patel, R. and Lu, C. (2004). Methods for detecting functional classifications in neuroimaging data, *Human Brain Mapping*, **23**, 109-119.

Bowman, F. D. (2005). Spatio-temporal modeling of localized brain activity, *Biostatistics*, **6**, 558-575.

Bowman, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data, *Journal of the American Statistical Association*, **102**, 442-453.

Brunner, E., Domhof, S. and Langer, F. (2002). *Nonparametric Analysis of Longitudinal data in factorial experiments*, John Wiley and Sons, Inc., New York, NY.

Calhoun, V. D., Adali, T., Pearlson, G. D. and Pekar, J. J. (2001). Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms, *Human Brain Mapping*, **12**, 43-53.

Calhoun, V. D., Adali, T., Pekar, J. J. and Pearlson, G. D. (2003). Latency (in) sensitive ICA group independent component analysis of fMRI data in the temporal frequency domain, *NeuroImage*, 30, 1661-1669.

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics*, **3**, 1-27.

Carew, I. D., Wahba, G., Xie, X., Nordheim, E. V. and Meyrand, M. E. (2003). The optimal spline smoothing of fMRI time series by generalized cross-validation, *NeuroImage*, **18**, 950-961.

Carr, J. M. (1996). Spectral and textural classification of single and multiple band digital images, *Computer and Geosciences*, **22**, 849-865.

Cattell, E. M. (1966). The scree test for the number of factors, *Multivariate Behavioral Research*, **1**, 245-276.

Chen, H. and Yao, D. (2004). Discussion on the choice of separated components in fMRI data analysis by spatial independent component analysis, *Magnetic Resonce Imaging*, **22**, 827-833.

Chen, X.(2005). Statistical and geostatistical features of streambed hydraulic conductivities in the platte river, Nebraska, *Environmental Geology*, **48**, 693-701.

C.-Olmo, M. and A.-Hernandez, F. (2000). Computing geostatistical image texture for remotely sensed data classification, *Computers and Geosciences* , **26**, 373-383.

Cherry, S., Banfield, J. and Quimby, W. F (1996). An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes, *Journal of Applied Statistics*, **23**, 435-449.

Chiles, J. P. and Delfiner, P.(1999). *Geostatistics: Modeling Spatial Uncertainty*, John Wiley and Sons, Inc., New York, NY.

Christidis, P. and Reynolds, R. (2004). AFNI HowTO03: background general information fMRI experimental design.

(http://afni.nimh.nih.gov/pub/dist/HOWTO//howto/)

Cliff, A. D. and Ord, J. K. (1973). *Spatial Autocorrelation*, Pion Limited, London.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, revised edition, John Wiley and Sons, Inc., New York, NY.

Curran, P. J. and Dungan, J., L. (1988). Estimating the signal to noise ratio of AVIRS data, *NASA Technical Memorandum 101035*.

(http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19890006072-1989006072.pdf)

Davis, B. M. (1987). Use and abuse of cross-validation in geostatistics, *Mathematical Geology*, **19**, 241-248.

Descombes, X., Kruggel, F. and Von Cramon, D. Y. (1998). FMRI signal restoration using a spatio-temporal markov random field preserving transitions, *NeuroImage*, **8**, 340-349.

Deutsch, C. V. and Journel, A. G.(1998). *GSLIB Geostatistical Software Library and User's Guide*, second edition, Oxford University Press, Inc., New York, NY.

Dubrule, O. (1983). Two methods with different objectives: splines and kriging, *Mathematical Geology*, **15**, 245-257.

Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process, *Journel of Agricultural, Biological, and Environmental Statistics*, **4**, 347-369.

Eddy, W. F., Fitzgerald, M., Genovese, C., Lazar, N., Mockus, A. and Welling, J. (1999). The Challenge of functional Magnetic Resonce Imaging, *Journel of Computational and Grapgical Statistics*, **8**, 545-558.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of statistics*, **32**, 407-451.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker Inc., New York, NY.

Fadili, M. J., Ruan, S., Bloyet, D. and Mazoyer, B. (2000). A multistep unsupervised fuzzy clustering analysis of fMRI time series, *Human Brain Mapping*, **10**, 160-178.

Fischer, B. and Everling, S. (1998) The antisaccade, a review of basic research and clinical studies.

(http://www.optomotorik.de/blicken/anti-rev.htm)

Friman, O., Borga, M., Lundberg, P. and Knutsson, H.(2002a). Detection of neural activity in fMRI using maximum correlation modeling, *NeuroImage* **15**, 386-395.

Friman, O., Borga, M., Lundberg, P. and Knutsson, H.(2002b). Exploratory fMRI analysis by autocorrelation maximization, *NeuroImage*, **16**, 454-464.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. -P., Frith, C. D. and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach, *Human Brain Mapping*, **2**, 189-210.

Genovese, C. R. (2000). A bayesian time-course model for functional magnetic resonance imaging data, *Journal of the American Statistical Association*, **95**, 691-719.

Gibbons, R. D., Lazar, N. A, Bhaumik, D. K., Sclove, S. L., Chen, H. Y., Thulborn, K. R., Sweeney, J. A., Hur, K. and Patterson, D. (2004). Estimation and classification of fMRI hemodynamic response patterns, *NeuroImage*, **22**, 804-814.

Glover, D.M., Jenkins , W.J. and Doney, S.C.(2006). Modeling, data analysis and numerical techniques for geochemistry.

(http://w3eos.whoi.edu/12.747/mfiles.html)

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, Oxford University Press, Inc., New York, NY.

Gordon, A. D. (1999). *Classification*, 2nd edition, Chapman and Hall/CRC.

Gorsich, D. J. and Genton, M. G. (2000). Variogram model selection via nonparametric derivative estimation, *Mathematical Geology*, **3**, 249-270.

Goutte. C., Hansen, L. K., Liptrot, M. G. and Rostrup, E. (2001). Feature space clustering for fMRI meta analysis, *Human Brain Mapping*, **13**, 165-183.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. A. and Hansen, L. K. (1999). On clustering fMRI time series, *NeuroImage*, **9**, 298-310.

Gringarten, E. and Deutsch, C. V. (2001). Teacher's aide variogram interpolation and modeling, *Mathematical Geology*, **33**,507-534.

Guitton, D., Buchtel, H. A. and Douglas, R. M. (1985). Frontal lobe lesions in man cause difficulties in suppressing reflexive glances and generating goal-directed saccades, *Experimental Brain Research*, **58**, 455-472.

Haacke, E. M., Brown, R. W., Thompson , M. L. and Venkatesan, R. (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*, John Wiley and Sons, Inc., New York, NY.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions, *Vision Research*, **18**, 1279-1296.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning, Data mining, Inference, and Prediction*, Springer Science, New york, NY.

Hartvig, N. V. and Jensen, J. L. (2000). Spatial mixture modeling of fMRI data, *Human Brain Mapping*, **11**, 233-248.

Huettel, S.C., Song, A. W. and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging*, Sinauer Associates Inc., Sunderland, MA.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*, John Wiley and Sons, Inc., New York, NY.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, Inc., New York, NY.

Jackson, J. E. (1991). *A User's Guid to Principal Components*, John Wiley and Sons, Inc., New York, NY.

Jezzard, P., Matthews, P. M. and Smith, S. M. eds. (2001). *Functional MRI: An Introduction to Methods*, Oxford University Press Inc., New York, NY.

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

Journel, A. and Huijbregts, Ch. J. (1978). *Mining Geostatistics*, Academic Press, Inc., New York, NY.

Journel, A. G. (1989). *Fundamentals of geostatistics in five lessons*, American Geophysical Union, Washington, D. C.

Katanoda, K., Matsuda, Y. and Sugishita, M. (2002). A spatio-temporal regression model for the analysis of functional MRI data, *NeuroImage*, **17**, 1415-1428.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons, Inc., Hoboken, NJ,

Keogh, B. P. and Cordes, D. (2007). Quantitative approaches to functional application in epliesy, *Eliesy*, **48**, 27-36.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information, *Problems of Information and Transmission*, **1(1)**, 1-7.

Kosslyn, S. M. (1999). If neuroimaging is the answer, what is the question, *Philosophical Transactions of the Royal Society of London B Biological Sciences*, **354**, 1283-1294.

Kruggel, F., Von Cramon, D. Y. and Descombes (1999). Comparison of filtering methods for fMRI datasets, *NeuroImage*, **10**, 530-543.

Kyriakidis, P. C. and Journel, A. G.(1999). Geostatistical space-time models: a review, *Mathematical Geology*, **31**, 651-684.

Lai, S. H. and Fang, M. (1999). A novel local PCA-based method for detecting activation signals in fMRI, *Magnetic Resonance Imaging*, **17**, 827-836.

Lange, N. and Zeger, S. L. (1997). Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging, *Applied Statistics*, **46**, 1-29.

Lazar, N. A. (2008). *The Statistical Analysis of Functional MRI Data*, Springer Science, New york, NY.

Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993). Canonical correlation analysis when data are curves, *Journal of the Royal Statistical Society*, **55**, 725-740.

Liu, Y., Gao, J. H., Liu, H. L. and Fox, P. T. (2000). The temporal response of the brain after eating revealed by functional MRI, *Nature*, **405**, 1058-1062.

Luss, R. and d'Aspremont, A. (2006). DSPCA: a toolbox for sparse principal component analysis.

(http://www.princeton.edu/ aspremon/DSPCAUserGuide.pdf)

Ma, Y. Z. and Ione, T. A. (2001). Teacher's aide modeling hole-effect variogram of lithology-indicator variables, *Mathematical geology*, **33**, 631-648.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

Marchini, J. L. and Ripley, B. D. (2000). A new statistical approacj to detecting significant activation in functional MRI, *NeuroImage*, **12**, 366-380.

Marcotte, D. (1991). Cokriging with Matlab, *Computers and Geosciences*, **17**, 1265-1280.

Marcotte, D. (1996). Fast variogram compution with FFT, *Computers and Geosciences*, **22**, 1175-1186.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters, *Journal of the Society of Industrial and Applied Mathematics*, **11**, 431-441.

Matheron, G. (1962). Traite de geostatistique appliquee, Tome I, *Memoires du Bureau de Recherches Geologiques et Minieres*, **24**, Editions Technip, Paris.

McBratney, A. B. and Webster, R. (1983). Optimal interpolation and isarithmic mapping of soil properties, V. Co-regionalization and multiple sampling strategy, *Journal of Soil Science*, **34**, 137-162.

McKeown, M. J., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A. and Sejnowski, T. (1998). Analysis of fMRI data by blind separation into independent spatial components, *Human Brain Mapping*, **6**, 160-188.

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

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrics*, **44**, 23-34.

Milot, S. (1998). Dot user's guide revision 1.8.0.

(http://www.bic.mni.mcgill.ca/users/sylvain/doc/html/dot/node79.html)

Morris, M. D. (1991). On counting the number of data pairs for semivariogram estimation, *Mathematical Geology*, **23**, 929-943.

Noll, D. C. (2001). A primer on MRI and functional MRI.

(http://www.eecs.umich.edu/ dnoll/primer2.pdf)

Oliver, M. A. and Webster, R. (1989). A geostatistical basis for spatial weighting in multivariate classification, *Mathematical Geology*, **21**, 15-35.

Otten, L. J., Henson, R. N. A. and Rugg, M. D. (2002). State-related and item-related neural correlations of successful memory encoding, *Nature Neuroscience*, **5**, 1339-1344.

Petersen, K., Hansen, L., Koleda, T., Rostrup, E. and Strother, S. (2000). On the independent components of functional neuroimages, *Procedings of the Third International Conference on Independent Component Analysis and Blind Source Separation (ICA2000)*, 615-610.

Preisendorfer, R. W. (1988). *Principal component analysis in meteorology and oceanography*, Elsevier Science Publishing Company Inc., New York, NY.

Pyrcz, M. J. and Deutsch, C. V. (2007). The whole story on the hole effect. (http://gaa.org.au/pdf/gaa-pyrcz-deutsch.pdf)

Qian, G., Gabor, G., Gupta, R. P.(1994). Principal components selection by the criterion of the minimum mean difference of complexity, *Journal of Multivariate Analysis*, **49**, 55-75.

Radeloff, V.C., Miller, T.F., He, H.S. and Mladenoff, D.J. (2000). Periodicity in spatial data and geostatistical models: autocorrelation between patches, *Ecography* **23**, 81-91.

Ramsay, J.O. (2005). *MATLAB, R and S-PLUS functions for functional data analysis*, McGill University.

Ramsay, J.O. and Silverman, B.W.(2005). *Functional Data Analysis*, 2nd edition, Springer Science Business Media, Inc., New York, NY.

Rencher, A. C. (1998). *Multivariate Statistical Inference and Applications*, John Wiley and Sons, Inc., New York, NY.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, 2nd edition, John Wiley and Sons, Inc., New York, NY.

Ripley, B. D. (1981). *Spatial Statistics*, John Wiley and Sons, Inc, New York, NY.

Rissanen, J. (1986). Stochastic complexity and modeling, *The Annuals of Statistics*, **14**, 1080-1100.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*, World Scientific Publishing Co. Pte. Ltd., Singapore.

Rosen, B. R., Buckner, R. L. and Dale, A. M. (1998). Event-related functional MRI: past, present, and future. *Proceedings of the National Academy of Sciences*, **95**, 773-780.

Rossi, R. E., Mulla, D. J., Journel, A. G. and Franz, E. H. (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence, *Ecological Monographs*, **62**, 277-314.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20**, 53-65.

Saad, Z. S., Ropella, K. M., Cox, R. W. and Deyoe, E. A. (2001). Analysis and use of fMRI response delays, *Human Brain mapping*, **13**, 74-93.

Schabenberger, O. and Gotway, C. A. (2005). *Spatial Methods for Spatial Data Analysis*, Chapman and Hall/ CRC Press, Boca Raton, FL.

Schabenberger, O. and Pierce, F. J. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences*, CRC Press, Boca Raton, FL.

Shapiro, A. and Botha, J. D. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions, *Computational Statistics and Data Analysis*, **11**, 87-96.

Shibli, S. A. R. (2003) Geostatistics FAQ – frequently asked questions. (http://www.ai-geostats.org/fileadmin/Documents/FAQ/FAQ-Geostatistics-01.pdf)

Sjstrand, K., Stegmann, M. B. and Larsen, R. (2006). Sparse principal component analysis in medical shape modeling, *Medical Imaging 2006: Image Processing, Proceedings of the SPIE*, **6144**, 1579-1590.

Sommer, F. T. and Wichert A., eds. (2002). *Exploratory Analysis and Data Modeling in Functional Neuroimaging*, The MIT press, Cambridge, Massachusetts, London, England.

Souza, L. E. D., Costa, J. F. C. L. and Koppe, J. C. (2004). Uncertainty estimation in resources assessment: a geostatistical contribution, *Natural Resources Research*, **13**, 1-15.

Spence, J. S., Carmack, P. S., Gunst, R. F., Schucany, W. R., Woodward, W. A. and Haley, R. W. (2007). The accounting for spatial dependence in the analysis of SPECT brain imaging data. *Journal of the American Statistical Association*, **102**, 464-473.

Stanberry, L., Nandy, R. and Cordes, D. (2003). Cluster analysis of fMRI data using dendrogram sharpening , *Human Brain Mapping* **20**, 201-219.

Stein, M. L. (1999). *Interpolation of Spatial Data, Some Theory for Kriging*, Springer-Verlag New York, Inc., New York, NY

Stone, J. V., Porril, J., Porter, N. R. and Wilkinson, I. D. (2002). Spatio-temporal independent component analysis of event-related fMRI data using skewed probility density functions, *NeruoImage*, **15**, 407-422.

Strauss, J. S., Bartko, J. J. and Carpenter Jr., W. T. (1973). The use of clustering techniques for the classification of psychiatric patients, *The British Journal of Psychiatry*, **122**, 531-540.

Strupp, J. P. (1996). Stimulate: a GUI based fMRI analysis software package, *NeuroImage* **3**, S607.

Tatsuoka, M. M. and Lohnes, P. R. (1988). *Multivariate analysis*, 2nd edition, Macmillan Publishing Company, New York, NY.

Thiron, B. (2003). fMRI data analysis: statistics, information and dynamics.

Tibshirani, R. (1996). Regression shrinkage and selection via LASSO, *Journal of Royal Statistical Society: Series B Methodological*, **58**, 267-288.

Tibshirani, R., Walther G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society*, **63**, 411-423.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, **46**, 234-240.

Van Der Meer, F. D. and De Jong, S. M. eds. (2001). *Imaging Spectrometry: Basic principales and Prospective Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Viviani, R., Gron, G. and Spitzer, M. (2005). Functional principal component analysis of fMRI data, *Human Brain Mapping* **24**, 109-129.

Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis, *Geoderma*, **62**, 83-92.

Wackernagel, H. (2003). Multivariate Geostatistics, An Introduction with Applications, 3rd edition, completely revised Edition, Springer-Verlag Berlin Heidelberg New York.

Wahba, G. (1990). *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics.

Wang, Y. M., Schultz, R. T., Constable R. T. and Staib, L. H. (2003). Nonlinear estimation and modeling of fMRI data using spatio-temporal support vector regression, *Lecture Notes in Computer Science*, **2732**, 647-659.

Wang, D., Shi, L., Yeung, D. S., Tsang, E. C. C. and Heng, P. A. (2007). Ellipsoidal support vector clustering for functional MRI analysis, *Pattern Recognition*, **40**, 2685-2695.

Watson, G. S. (1984). Smoothing and interpolation by kriging and with splines, *Mathematical Geology*, **16**, 601-615.

Webster, R. and Oliver, M. A. (2001). *Geostatistics for Environmental Scientists*, John Wiley and Sons, Ltd., Baffins Lane, Chichester, West Sussex, England.

Wynn, G. (2000). Image for idiots, a medical imaging learning resource.

(http://cal.man.ac.uk/student-projects/2000/mmmr7gjw/menu.htm)

Yee, S. H., and Gao, J. H. (2002). Improved detection of time windows of brain responses in fMRI using modefied temporal clustering analysis, *Magnetic resonance Imaging*, **20**, 17-26.

Yeo, B. T. and Ou, W. (2004). Clustering fMRI time series.

(http://people.csail.mit.edu/ythomas/6867fMRI.pdf)

Zhen, X. and Basher, R. (1995). Thin-plate smoothing spline modeling of spatial climate data and its application to mapping south pacific rainfalls, *Monthly Weather Review*, **123**, 3086-3102.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, **15**, 265-286.