On the Estimation of Smooth Autoregressive Parameter Fields with Applications in Ophthalmology

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund
vorgelegt von

Tillmann Krahnke

geboren in Soest, Westfalen

Dortmund 2001

Gutachter: Prof. Dr. Siegfried Schach
Prof. Dr. Ursula Gather

Tag der mündlichen Prüfung: 10. Dezember 2001

Para mi esposa - con todo el alma

Acknowledgements

I would like to thank Johannes Hüsing, University of Essen, for first drawing my attention to the analysis of AMD data, and Dr. Bernhard Jurklies and Malte Weismann, Eye Hospital of the University of Essen, for providing the AMD data sets analyzed. Helpful comments on the general structure of the manuscript were given by Dres. Roland Fried and Claudia Becker, University of Dortmund. Dr. Indranil Chakrabarti and Maritza Meléndez, M.Sc., M.Sc. patiently checked my use of the English language and took care of the readability of the result. Thanks also to Professor Dr. E.E. Sutter for kindly giving the permission to include Figure 2.2 into the manuscript. Matthias Schneider took a lot of effort to convert it into a printable format.

In also want to thank Professor Dr. Wolfgang Urfer for encouraging me to tackle spatiotemporal statistics in the first place, and Professor Dr. Siegfried Schach for fruitful discussions on the topic while preparing the thesis. My special thanks, however, go to Professor Dr. Ursula Gather, who gave me the opportunity to join her research group, and who guided me through the last stages of my work.

This doctorate thesis was written while the author was a member of the collaborative research centre 'Reduction of complexity in multivariate data structures', at the University of Dortmund, Germany. The funding was provided by the German Science Foundation (Deutsche Forschungsgemeinschaft - DFG) within the 'Sonderforschungsbereich 475'. I thankfully acknowledge the support of the DFG.

Contents

Chapter 1

Introduction

1.1 Motivation

The results presented in this document were motivated by a study carried out at the Department of Ophthalmology at the University Hospital of Essen, Germany. The goal was to investigate a type of retinal dysfunction called *Age-Dependent Macular Degeneration (AMD)*, which is a loosely defined set of visual deficiencies of the retina that occurs mostly in elderly people. AMD leads to a decline in bioelectrical response of retinal receptor cells to visual stimulation, and possibly even to complete loss of vision. Figure 1.1 shows an example of a retina in an early stage of AMD. In the central retinal region displayed, blood circulation is low (indicated by dark shades of grey), indicating a loss of retinal performance.

The bioelectrical functionality of a patient's eye can be evaluated by means of a so-called *Electroretinogram* (ERG), which records the electric potentials occurring on the eye ball when a well-defined set of visual stimuli is presented to the patient. The ERG is called *multifocal* if several distinct areas of the retina are examined simultaneously. The corresponding diagnostic technique is relatively new and was first introduced in the early nineties (Sutter and Tran 1992). While conventional ERG techniques allow for derivation of a single overall response of the complete retina only, the *multifocal* ERG enables the researcher to

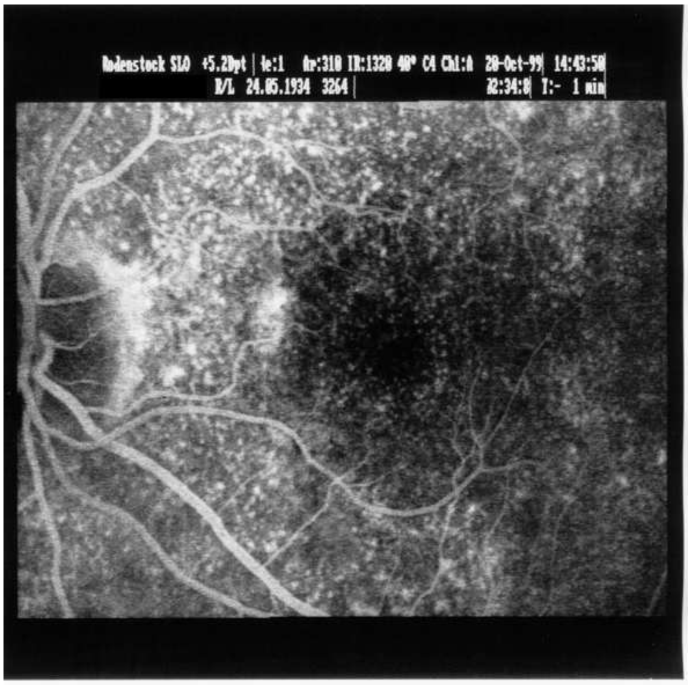

Figure 1.1: *Human retina with early age-dependent macular degeneration (AMD).*

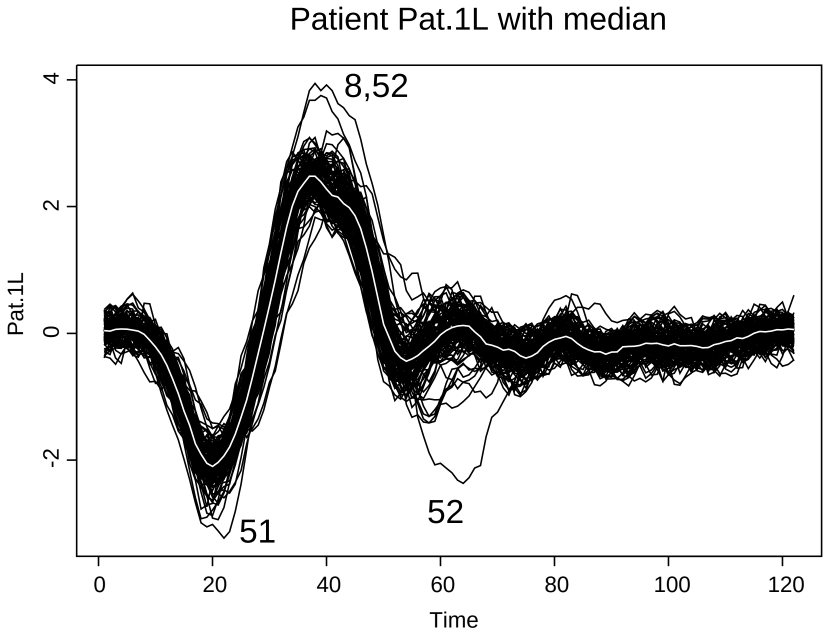

Figure 1.2: *Measurements obtained from the multifocal ERG for the retina depicted in Figure 1.1. The 103 time series represent non-overlapping areas. Data set is Pat.1L. The white line describes the pointwise median for each time point. The numbers 8, 51 and 52 identify curves with unusual or extreme behaviour.*

obtain measurements at a higher spatial resolution. Although several tools are available to analyze the temporal dynamics as well as the spatial features inherent in the data, the joint spatiotemporal characteristics of the observed values have rarely been addressed in the medical literature. The work described below is an attempt to improve on this situation using statistical techniques.

The data available were obtained as part of a larger study conducted at the University of Essen, Germany. Four data sets were collected using the VERIS® measuring system (EDI 1999) on the eyes of three different patients. The data represent the temporal evolvement over 122 msec of locally evoked potentials at each of 103 non-overlapping hexagonal areas of the retina. In a certain sense, each data set consists of a multivariate time series, since observations are available over a period of 122 milliseconds for each retinal area. Figure 1.2 shows the measurements obtained from the retina displayed in Figure 1.1.

Different views on the data are possible. In past clinical application and research, the analysis of multifocal ERG data often has been purely descriptive in nature and was mostly guided by the ophthalmologist's practical experience. Only few attempts are found in the medical literature to make use of more complex statistical methodology. The typical approach is to interpret the data obtained as a set of time series and to analyze some of their qualitative features. A certain amount of data reduction is sometimes gained by looking at amplitudes, or by averaging over adjacent groups of time series. It is known in ophthalmology that retinal features of healthy patients vary by distance from the center of the eye globe. Therefore, groups are often formed by selecting concentric rings around the retinal center. Provided that retinal dysfunction evolves concentrically as well, this is a sensible approach. However, in practice this is not necessarily the case. In consequence, important data features may be distorted by such kind of spatial averaging.

A first naive descriptive analysis of the temporal and spatial aspects of the empirical data sets given here indicates that dynamics both in time and space are involved. Therefore, application of suitable *spatiotemporal* statistical methods is called for. It is the major goal of this doctoral thesis to present a statistical analysis approach to multifocal ERG data which avoids the disadvantages of current analysis techniques. The resulting model should parsimoniously describe the biological process under study, while completely taking into account the spatial, temporal, and spatiotemporal information contained in the data. Techniques for doing this have been available in the statistical literature for several years, but have not been adapted to multifocal ERG data. One major reason appears to be that such techniques require the specification of the general dependency structure (i.e. covariance) in the data in advance, at least up to a small number of parameters. In order to circumvent this critical issue, a modified approach is proposed. It is built upon a combination of techniques from time series analysis, spatial statistics, and spline smoothing. The resulting method will be seen to provide good fit to the data, while yielding spatially smooth estimates of the parameters.

1.2 Overview

The focus of this doctoral thesis is on the analysis of data obtained from the multifocal ERG. An exploratory analysis of the data available is presented in Chapter 2. It is demonstrated that the data carry both spatial as well as temporal information. Therefore, a combination of techniques both from spatial statistics and time series analysis should be used to more adequately describe multifocal ERG data within a statistical framework.

Ordinary least squares estimates for parameters from autoregressive time series models are the starting point. They are introduced in Chapter 3. It is seen there that a purely temporal analysis only partly describes the dynamics in the multifocal ERG data sets at hand, and that the spatial layout of the data should be accounted for explicitly.

Chapter 4 therefore provides a spatial analysis of the data. Spatial smoothing is performed to remove noise inherent in the resulting estimators. The techniques of Kriging and Spline Smoothing are two candidates under study. They are introduced in Chapters 4 and 5, respectively. Applications to multifocal ERG data are added at the end of each of these chapters, and arguments are given why modifications are desirable.

A modified smoothing approach referred to as *smoothing of AR-parameter fields* is described in Chapter 6. It makes use of a fitting criterion that accounts both for spatial smoothness of the autoregressive parameter estimates as well as a satisfactory temporal fit to the observed data. It will be seen that the smoothed parameter estimates are well interpretable, while giving rise to only a small increase in the overall sum of squares for fit. Chapter 7 summarizes these results and gives some suggestions for future research. Empirical results obtained for the available data sets are combined in the appendix and complement the examples already described in the foregoing text.

Chapter 2

Characteristics of multifocal ERG Data

2.1 Study Background

The visual perception of the human eye is based on optic stimuli which fall onto the retina and are forwarded to the brain via the optic nerve. One of the first steps in this process is to transform the light projected onto the retina into electric signals. This is accomplished by photoreceptors situated in the inner eye ball, which activate or hamper corresponding neurons.

Degeneration of the retina leads to loss of visual ability. It is well known that, in particular, elderly people are affected by such a loss, the cause of which is not always easy to detect. This is somewhat reflected in the scientific name assigned to a large group of visual defects of the retina. They are combined under the name *Age-Dependent Macular Degeneration* (AMD).

At the Eye Hospital of the University of Essen a study is currently in progress to examine AMD closer. Electric potentials emitted at different regions of the retina of more than 150 patients have been recorded in a *Multifocal Electroretinogram* (MF-ERG) using the commercially available Visual Evoked Response Imaging System (VERIS®) version 4.0 (EDI 1999). One goal of this large study is to closely examine the possibilities offered by the relatively new diagnostic tools available as part of the VERIS® system, and to improve on diagnostic conclusions. In particular, it is hoped that better discrimination between different subclasses of AMD will be possible at some point in the future even in early stages of the disease. At present, qualitative characterization of the complete MF-ERG signal is still a major issue.

In a pilot study, four ERG data sets were made available to examine how statistical methods could be used for an efficient analysis of multifocal ERG data. The analysis of these data sets is described in this doctoral thesis. The pilot study was aimed at finding a parameterization for the patterns in ERG data which is parsimonious, accessible and informative to the ophthalmologist, all at the same time. Such parameterization could then possibly be used for characterization of data sets in later stages of the larger AMD study.

2.2 Electroretinography with the VERIS® System

The VERIS® System (EDI 1999) is an electrophysiological instrument used to evaluate visual perception by measuring electric potentials evoked on the eye ball by predefined visual stimuli. In this section an abbreviated description of the measuring process is given. Further details can be found in the software's manual.

2.2.1 Experimental Setup

Multifocal ERG data are obtained by presenting to the patient a sequence of flickering black and white hexagonal patterns on a regular monitor, similar to that shown in Figure 2.1. The person under study is seated in a darkened room and placed some 40 cm away from screen. The head is fixed to avoid movement. A so-called Burian-Allan contact lens electrode is placed on one eye to deduce the currents arising at the eye globe. The second eye is covered to prevent from blinking.

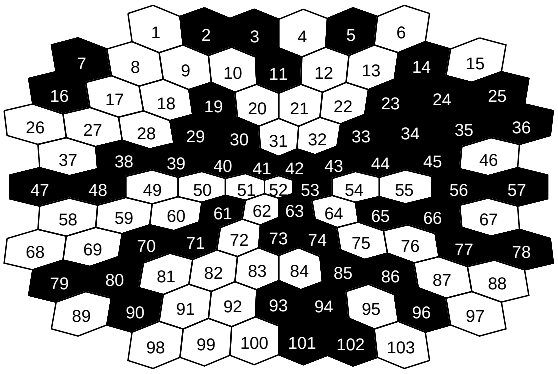

Figure 2.1: *Stimulus array of 103 hexagonals. Pattern is slightly distorted to account for shape of eye ball. Border lines between hexagonals are not visible during the experiment. Numbers are inserted here for easier reference, but are not visible in actual experiment.*

The potentials observed are transferred to a desktop computer and preprocessed for further analysis. This includes

- the removal of outliers
- the scaling of data according to the chosen hexagonal layout to obtain data on local luminance in a standardized unit (millivolt per unit area)
- the derivation of time series of electric potentials attributable to well specified hexagonal areas on the retina

Figure 2.2 gives a general impression of the measurement process. The processed data are made available to the analyst either via different visual displays on screen which are produced by the accompanying VERIS® software, or as a portable data set which may be saved in ASCII format and analyzed externally.

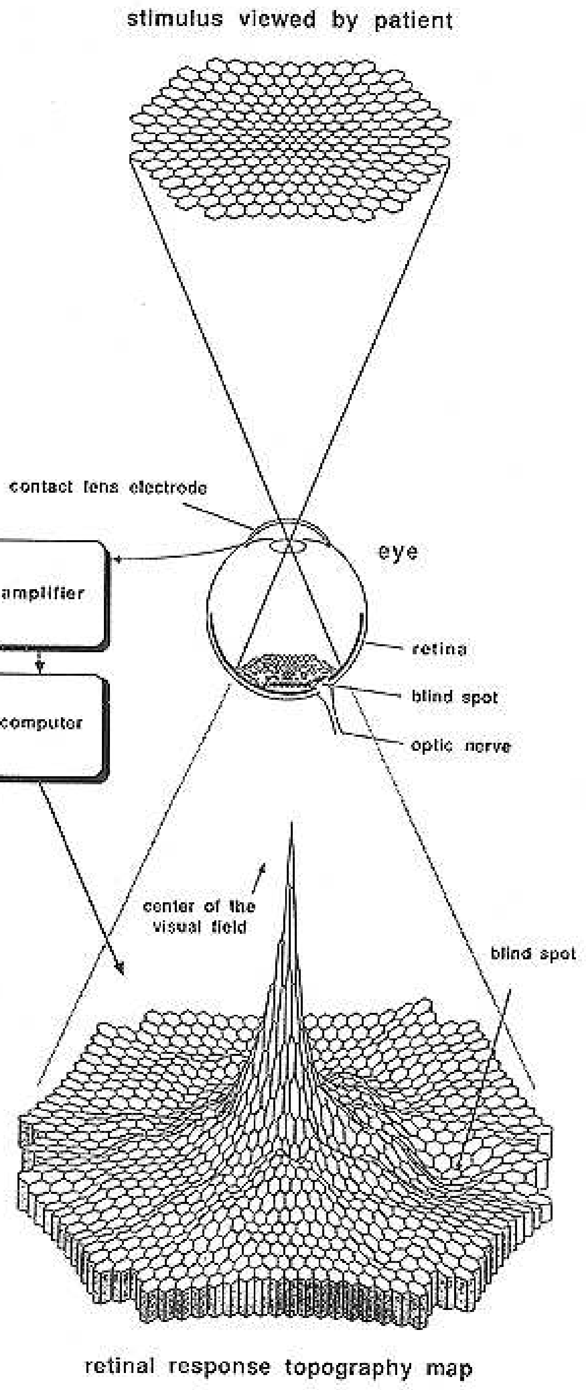

Figure 2.2: *Measurement process for multifocal ERG data. With kind permission of Dr. Erich E. Sutter (*`www.ski.org/EESutter_lab/ees1.html`*).*

2.2.2 ERG Analysis in Current Medical Research

It is common in medical literature to regard the overall *amplitudes* (i.e. range) of ERG measurements at a specific location as an indicator of the visual functionality of the retina. Other statistics used are *implicit times* or certain *inner product measures*. The former measure the time span between the onset of the time series and some characteristics of interest, e.g., the minimum or the maximum of the curve. The exact temporal occurrence of the relevant features is determined either automatically by the software, or it is specified manually by the ophthalmologist using some pointing device like a computer mouse.

Inner product measures give an indication of deviation from some standard response by calculating the distance between observed values and that response. Since they do not reflect the overall shape of the actual data curve, inner product measures are not analyzed here any further.

It appears that an overall model for a complete ERG data set has not been developed so far by ophthalmologists. A literature review did not provide any results towards that end. The following subsections briefly review the methods which are effectively applied when analyzing multifocal ERG data in practice. Most of them are readily available as part of the software accompanying the VERIS® measuring system.

Amplitude Measures

Multifocal ERG amplitudes are examined by Brown and Yap (1995), Mack et al. (1999) and Si et al. (1999), among others. An amplitude here is defined as the overall range of a sequence of data and is calculated separately for all hexagonal areas. Brown and Yap (1995) study closer the effect of environmental conditions on the response amplitudes, like target contrast and changes in local luminance. Amplitudes are plotted versus log-contrast (measured in percent) and simple linear regression is performed. The authors find linear dependencies between amplitude measure and log-contrast as well as log-luminance.

Mack et al. (1999) describe the major capabilities of the VERIS® software as it is used in clinical practice. Si et al. (1999) qualitatively describe changes in amplitudes measured before and after surgery. The analytical tools used are those already contained in the VERIS® software. This is typical for many articles on data obtained from the MF-ERG.

Curve Shape

The general curve shape of ERG time series is studied, for example, by Kondo et al. (1995) and Graham and Klistorner (1998). Kondo and group present a clinical evaluation of the multifocal ERG. They observe that different retinal malfunctions lead to qualitatively different waveforms of the resulting measurements. Their findings justify to some degree the use of the multifocal ERG to classify different types of AMD. The authors also study the reproducibility of measurements within subject by taking repeated measurements and calculating the local mean response and its standard deviation. They conclude that reproducibility of measurements within patients is satisfactory. However, Kondo et al. (1995) note some possible problems when using the multifocal ERG as a diagnostic aid. These are, among others, inter-subject variability, and difficulties in eye fixation for patients with severe visual defects.

It is stressed by Graham and Klistorner (1998) that the temporal aspect of multifocal-ERG measurements is a relatively recent additional information to ophthalmologists which was not available before. The authors qualitatively describe the ERG curve patterns obtained from different stimuli, and search for connections to certain physiological conditions or existing visual deficiencies. They are able to show that such connections indeed exist.

Special care was taken by the medical personal at Essen University when collecting the ERG data for their study to avoid the problems indicated by Kondo et al. (1995) and Graham and Klistorner (1998). All data sets were collected using the same type of visual stimulus, and eye fixation was constantly monitored during data recording.

Implicit Times

Implicit times, also referred to as latencies, describe the time between onset of the derived measurement, and occurrence of a certain curve feature. They are studied in several recent papers, including Aoyagi et al. (1998), Keating et al. (1998), Kretschmann et al. (1998), Parks et al. (1996), Palmowski et al. (1997), Seeliger et al. (1998a), Seeliger et al. (1998b), and Sutter and Bearse (1999).

In summary, it was discovered that the spatial distribution of implicit times in healthy patients follows a certain concentric pattern. Seeliger et al. (1998b) support this by means of power spectra of the response curves for each hexagonal area, as well as boxplots of locally (i.e., within each hexagon) observed implicit times obtained from a population of 30 patients. Seeliger et al. (1998a) found retinal asymmetry in that the nasal implicit times were longer than the temporal implicit times. Other authors confine themselves to describe the quantitative change in implicit times attributable to certain retinal disorders.

The analysis of latencies focusses on a single feature of MF-ERG data only. Only a small part of information available in the data is used. In addition, analysis results reported in the literature tend to be of a subjective nature, rather than yielding an objective and concise set of parameters describing the eyes' health status. This is an additional argument for a more involved analysis, possibly using spatiotemporal statistical methods.

The preceding brief literature review already indicates that multifocal ERG data contain both spatial and temporal features as important information. If only parts of the total information is needed, an overall model may still be helpful to estimate the parameters of interest more

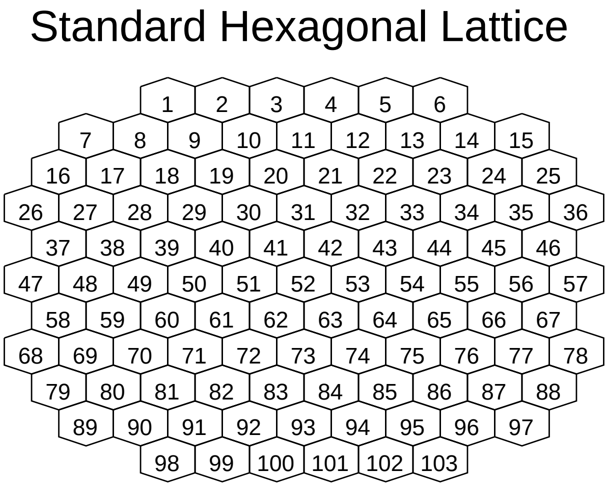

Figure 2.3: *Subdivision of the retina into 103 hexagonal areas as obtained by the multifocal ERG. Numbers are added for ease of reference.*

precisely. However, an analysis of the overall *spatiotemporal* dynamics involved was not found in the literature on multifocal ERG data.

2.3 A First Look at the Data

2.3.1 Data Layout

Four data sets were available for analysis in the pilot study, each consisting of 103 x 122 or 12,566 observations in total. The data were obtained using the VERIS® system (EDI 1999), exported into ASCII format and then analysed using the statistical software S-Plus® 2000 (Mathsoft 2000a). The data sets are referred to here as data set Pat.1R, Pat.1L, Pat.2 and Pat.3, respectively. As indicated by these names, observations in data sets Pat.1R and Pat.1L were obtained from the right and left eye of the same patient. However, in accordance to practical experience in ophthalmology, they may be regarded as two independent data sets. All data sets were automatically preprocessed by the VERIS® software for appropriate scaling and outlier removal.

In correspondence with the setup shown in Figure 2.3, the retina can be subdivided into a grid of 103 hexagonal areas onto which the stimulus is projected. The exact layout of the hexagonal grid is not known, except that (after automatic scaling) areas can be regarded as equally sized and regularly spaced. In particular, exact horizontal and vertical distances between hexagonal center points are not given by the measuring device. Therefore, dummy coordinates are used here. The innermost hexagon (number 52) was defined to have its center point at $(0, 0)$. Adjacent hexagonals in horizontal direction (referred to as X-coordinate) are one unit apart from their neighbors. Adjacent areas in vertical direction (referred to as Y-coordinate) have center points with coordinate one unit apart from their direct neighbor, but are shifted 0.5 units in horizontal direction. For example, area 52 has coordinates $(0, 0)$, while its left upper neighbor 41 is centered at $(-0.5, 1)$.

2.3.2 General Summary Statistics

Table 2.1 gives some overall descriptive statistics of the data sets under study. Generally, a series of 122 temporally ordered observations is given for each of the 103 hexagonal areas. The original data vary between -3280 mV/mm^2 (millivolt per millimeter squared) and 4260 mV/mm^2. The overall mean values of the four data sets vary between -0.0248 and 0.0148 mV/mm^2 and hence are negligible when compared to the data range.

The complete raw data sets were standardized by subtracting the overall mean and dividing by the overall empirical standard deviation. See Table 2.2 for results. Note that no trend components were removed at this point, so the estimated variances are difficult to interpret.

To gain some insight into the structures inherent in the data, an explorative graphical data analysis was performed. The following section shows how measurements vary over time.

Data Set	N	Min	Max	Range	Mean	Variance
Pat.1L	12566	-2540	3100	5640	-0.0248	618094.7
Pat.1R	12566	-2590	3010	5600	-0.0076	469097
Pat.2	12566	-3280	4260	7540	0.0148	1217845
Pat.3	12566	-1800	1670	3470	-0.0042	241372

Table 2.1: *Overall Summary Statistics for Raw Data*

Data Set	N	Min	Max	Range	Mean	Variance
Pat.1L	12566	-3.23	3.94	7.17	0	1
Pat.1R	12566	-3.78	4.39	8.18	0	1
Pat.2	12566	-2.97	3.86	6.83	0	1
Pat.3	12566	-3.66	3.40	7.06	0	1

Table 2.2: *Overall Summary Statistics for Standardized Data*

2.4 Exploring Temporal Aspects

2.4.1 A Temporal View at ERG Data

Multifocal ERG data can be regarded as a set of time series observed at nearby locations. A graphical display of the data as temporal sequences, or curves, gives some hints for suitable methods of analysis. Figures 2.4 and 2.5 show in their left column multiple line plots for each of the four data sets available.

Although a general damped sinusoidal appearance is common to all data sets, there are remarkable differences in other data features. The most prominent ones are

- variations in dispersion over time
- differences in overall curve amplitudes
- different ways in which the curves fade out

With regard to the four data sets available, there seems to be no such thing as a standard curve shape. It seems plausible to assume that this is due to the fact that patients with

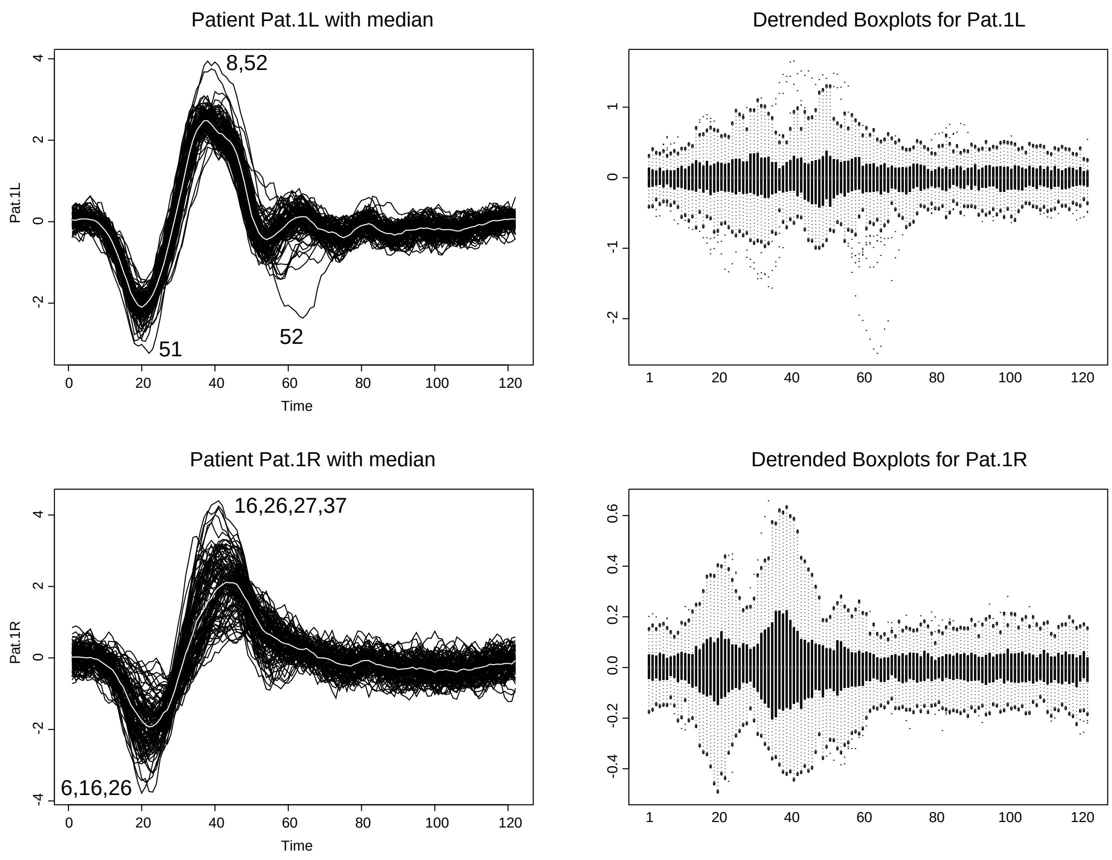

Figure 2.4: *Left Side: Multifocal ERG measurements, time series view of Pat.1L (top) and Pat.1R (bottom). Each line represents a time series measured at one of 103 locations. White line marks pointwise medians. Right Side: Boxplots with pointwise median removed.*

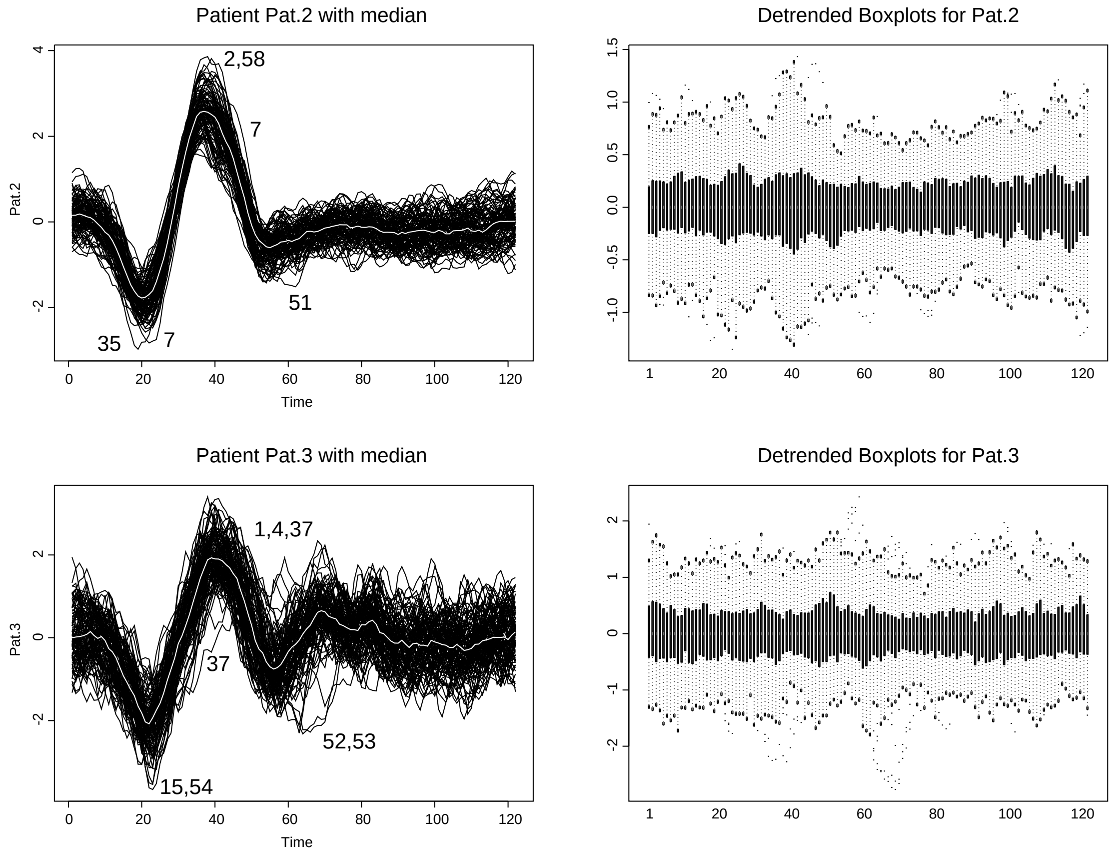

Figure 2.5: *Left Side: Multifocal ERG measurements, time series view of Pat.2 (top) and Pat.3 (bottom). Each line represents a time series measured at one of 103 locations. White line marks pointwise medians. Right Side: Boxplots with pointwise median removed.*

different defects were included in the study. However, within any single eye, most of the curves appear to show very similar features.

Side-by-side boxplots were created to study variability between time series within single time points. The pointwise median was taken as a robust measure of average trend within each time point (indicated as a white line in the graphics), and was subtracted for display. Obviously, boxplots do not take into account the temporal aspect of the curves. However, they highlight those time intervals where most of the differences between the 103 sequences occur.

When regarded as time series, individual curves show strong systematic changes, or temporal trend, but only relatively little random variation. Most of the pointwise spread visible in the above figures is due to the overlay of several curves, i.e., it is caused by variability between curves. Outliers within individual curves are not detectable. However, single curves occasionally behave differently than the vast majority. The number of some corresponding hexagonal areas is indicated in Figures 2.4 and 2.5. In several cases these are curves close to the retinal center. The following qualitative description of data sets goes more into details.

Patient 1 - Left Eye (Pat.1L)

Figure 2.4 shows in its upper left the temporal evolvement of data set Pat.1L. A damped sinusoidal curve shape is observable for the bulk of curves, with minimum around 20 msec and maximum around 40 msec after start of the sequence. This is emphasized by the white line in the graph representing the pointwise median of all curves at any given point of time. The curve with minimal value at 20 msec is curve 51. The maximum at 40 msec and the minimum at 60 msec is given by curve 52. These two curves are located just in the center of the hexagonal grid which is projected onto the retina (compare Figure 2.3). Most curves show a second local minimum after about 55 msec. This is followed by two smaller peaks at about 65 msec and slightly above 80 msec.

Pointwise boxplots for data set Pat.1L are also shown in Figure 2.4 (upper right). The pointwise median was removed before plotting. The variability between curves is highest between 25 and 35 msec, and about 45 to 55 msec. After about 60 msec the variability returns to values similar to those at the onset of the observation period. Curve 52 reappears as a sequence of potential outliers between 55 and 75 msec after onset. This stresses the fact that pointwise statistics should be used only with care, since by definition they do not take the overall temporal evolvement of single curves into account.

Patient 1 - Right Eye (Pat.1R)

Data set Pat.1R displayed in Figure 2.4 (lower left) shows less prominent oscillatory features than Pat.1L. The majority of univariate time series has a minimum at around 20 msec and a maximum at about 45 msec before slowly damping out. After 55 msec, a second local minimum can be observed. Two small bumps occur after 65 msec and 80 msec, but are hardly visible.

Figure 2.4 shows at its lower right side the boxplots for data set Pat.1R. The high variation around 20 and 40 msec indicate large differences between curves: Most of the time series are slowly fading out in the second half of the observed interval, while some others show similar oscillatory features to those seen in Pat.1L. In particular, a local minimum after 55 msec and two small bumps after about 75 and 90 msec occur. Later it will be examined how curves with similar features may be grouped together.

Patient 2 (Pat.2)

Data set Pat.2 shows yet another general shape of curves (Figure 2.5, upper left). Again, a first minimum after 20 msec followed by a maximum after 40 msec is observable. There exists also a local minimum after about 55 msec. However, after this second minimum the data values smoothly level out, roughly to the overall average. It is interesting to note that the somewhat extraordinary curve taking the minimal value at 60 msec is located at position 51, just to the left of the center of the retinal grid. Somewhat unusual behaviour can be seen for curve 7. It is located in the upper left corner of the hexagonal grid. Curve 7 is minimal around 25 msec after onset and has larger values than all other curves at around 50 msec.

The boxplots for Pat.2 (upper right) show a roughly constant pointwise variation, with exceptions at about 30 msec (decrease) and 40 msec (increase). Compared to the other data sets, the detrended Pat.2 data behave relatively homogeneous.

Patient 3 (Pat.3)

The data for the third patient are shown in Figure 2.5, second row. This data set has the smallest overall range before scaling. Again, prominent peaks after 20 msec (smallest values within areas 15 and 54) and 40 msec (maximum) can be seen in the multiple line plot. Two local maxima are given after about 70 msec, and possibly after 85 msec. At 40 msec, curve 37 takes unusual small values. After 70 msec, curves 52 and 53 are smallest. Thus, as in the three preceding data sets, centrally located curves show pronounced features which

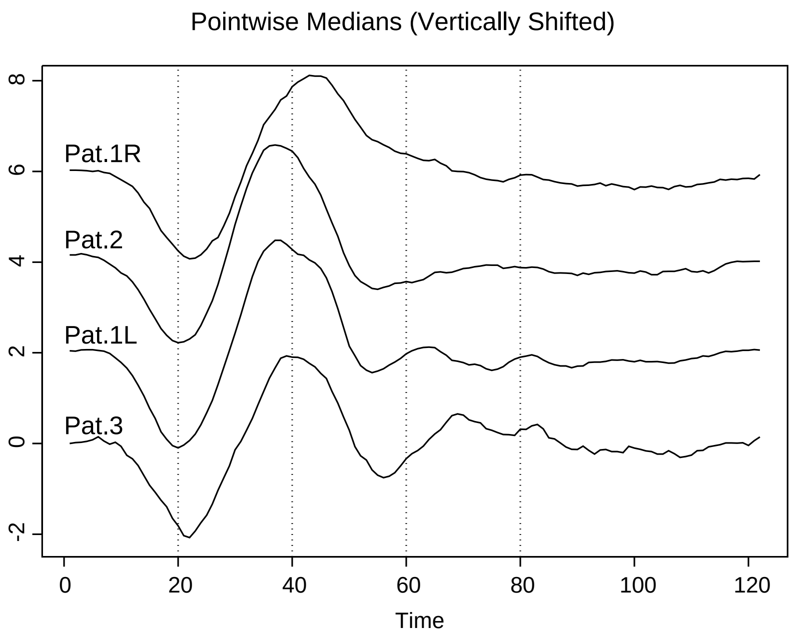

Figure 2.6: *Pointwise Medians for the four data sets under study. For clarity, curves were shifted vertically by 0, 2, 4, and 6 units.*

are different from other curves. This motivates to examine if spatial components should be incorporated into an overall data model for the ERG data.

A look at the pointwise boxplots for Patient 3 shows remarkable homogeneity in variation, but on a relatively high level. Extreme values around 70 msec are caused by the curves 37, 52 and 53, which were already identified as special. Note that this impression can only be tentative, since pointwise trend removal was done by simply subtracting the overall median, which is a rather crude approach.

2.4.2 Temporal Trend

The preceding subsection gave a general description of some major features in the four data sets analyzed. Although it became evident that certain curves do clearly not follow the overall general behaviour within the corresponding data set, a comparison between the overall

Group	Members (Hexagonal Areas)	N
1	41 42 51 52 53 62 63	7
2	30 31 32 40 43 50 54 61 64 72 73 74	12
3	19 20 21 22 29 33 39 44 49 55 60 65 71 75 82 83 84 85	18
4	9 10 11 12 13 18 23 28 34 38 45 48 56 59 66 70 76 81 86 91 92 93 94 95	24
5	1 2 3 4 5 6 7 8 14 15 16 17 24 25 26 27 35 36 37 46 47 57 58 67 68 69 77 78 79 80 87 88 89 90 96 97 98 99 100 101 102 103	42

Table 2.3: *Grouping of Hexagonal Rings.*

pointwise median lines still gives a rough impression of how retinal responses may differ between patients. Figure 2.6 shows all four pointwise median curves in the order Pat.3, Pat.1L, Pat.2, Pat.1R (top to bottom). For clear presentation, they were shifted vertically by 0, 2, 4 and 6 units, respectively. The order was chosen to emphasize the differences in dynamics in the second half of the observed time interval.

In terms of the overall temporal trend, it can be stated that within the first 50 msec, all four data sets roughly follow a sine curve. Between 50 and 90 msec the medians level out to the overall average. This happens in various ways: The values in data set Pat.1L fade out like a damped sine wave, whereas values in Pat.1R move towards the mean almost directly. A very flexible class of models has to be found which encompasses all these cases. The class of autoregressive moving average (ARMA) processes developed in time series analysis is a possible choice. It often provides a data description by relatively few parameters, particularly if sinusoidal components are involved. ARMA models rely on certain assumptions which will have to be checked before application.

Before this is done, the spatial features of the data should be examined. It was noted for all data sets under study that curves observed close to the retinal center show a somewhat unusual behavior when compared to their neighbors. For this reason, an exploratory analysis for *groups* of curves was performed and is described below.

2.5 Exploring Spatial Aspects

In clinical practice, grouping of retinal areas is often done by assigning concentric rings of hexagonals to groups. This is motivated by some findings on the physiological structure of the retina, which indicates highest receptor density close to the center. The grouping used in the current analysis is shown in Figure 2.7, lower right. The 103 hexagonals are sorted into five groups with group members given in Table 2.3. Note that group sizes vary considerably between 7 and 42.

2.5.1 Spatial Data Features

Figures 2.7, 2.9, 2.11 and 2.13 show the temporal evolvement within each group, with corresponding group median indicating how curves vary on the average by spatial position. The information included in these displays is described below for each data set.

Patient 1 - Left Eye

Multiple line plots of data set Pat.1L support the empirical findings that spatial location indeed has in influence on general curve shape (Figure 2.7). The median line for the innermost group (group 1) shows a higher amplitude than those of the other groups of this patient. The first group also differs in several other respects. For example, it includes curve 52, which has very low values around 60 msec. The other group members increase almost directly towards zero between 55 and 80 msec. In contrast, other groups show only a small local minimum around 55 msec, which is followed by a short oscillatory period until about 90 msec. It may also be noted that in group 4, two curves behave somewhat differently compared to the others. Curve 14 is lowest at around 40 msec, while curve 13 is highest at around 55 msec. These curves are located in the upper right region of the retina.

Temporal aspects of the data are ignored when looking at local *amplitudes* only. However,

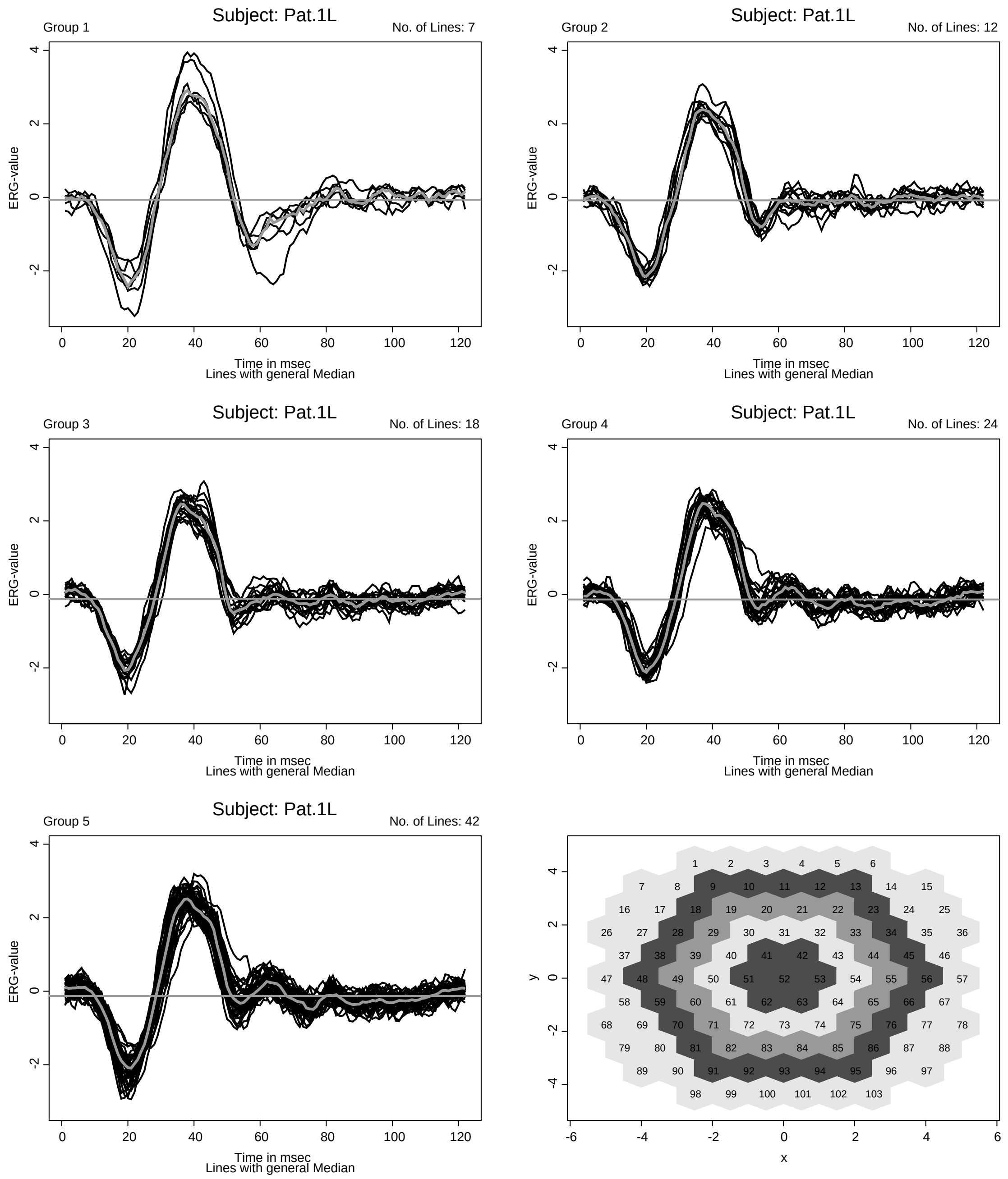

Figure 2.7: *Data set Pat.1L. Multiple time series plots grouped by concentric rings. Group-wise median line is added to highlight changes between groups.*

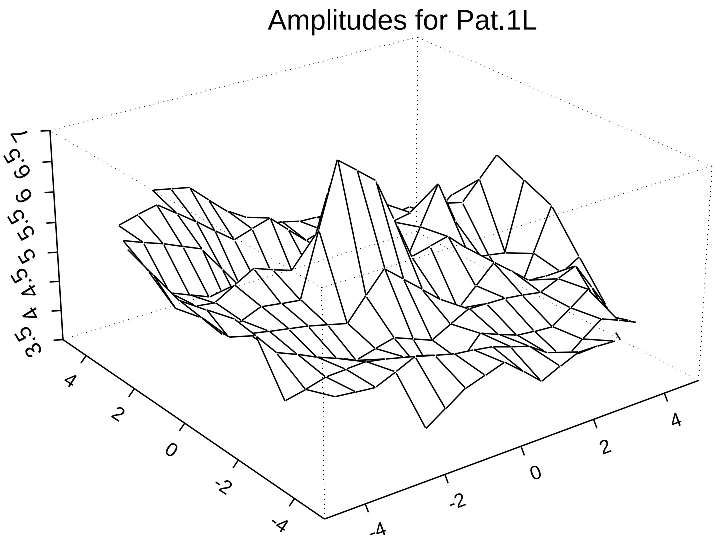

Figure 2.8: *Amplitudes for Pat.1L. Values are interpolated where necessary.*

this is often done in practice. Figure 2.8 shows a perspective plot of amplitudes for Pat.1L. Values are interpolated where necessary using the `interp()`-function which can be found in S-Plus® 2000. In the center of the retina, a peak is clearly visible. Also, some outer areas are somewhat elevated. Similar to the groupwise plots given above, the perspective plot also supports the impression that spatial features are indeed observable in this data set.

Patient 1 - Right Eye

The groupwise median lines for data set Pat.1R are displayed in Figure 2.9. They show a steady increase in local amplitudes from the center towards the edge of the retinal area under study. Other features are very similar for all five groups. The choice of groups by concentric rings seems to be reasonable.

The perspective plot of amplitudes for Pat.1R in Figure 2.10 looks quite different from what has been observed before. A roughly quadratic spatial trend is present, with minimum in the central region of the retina. The central peak observed for Pat.1L is not present. The

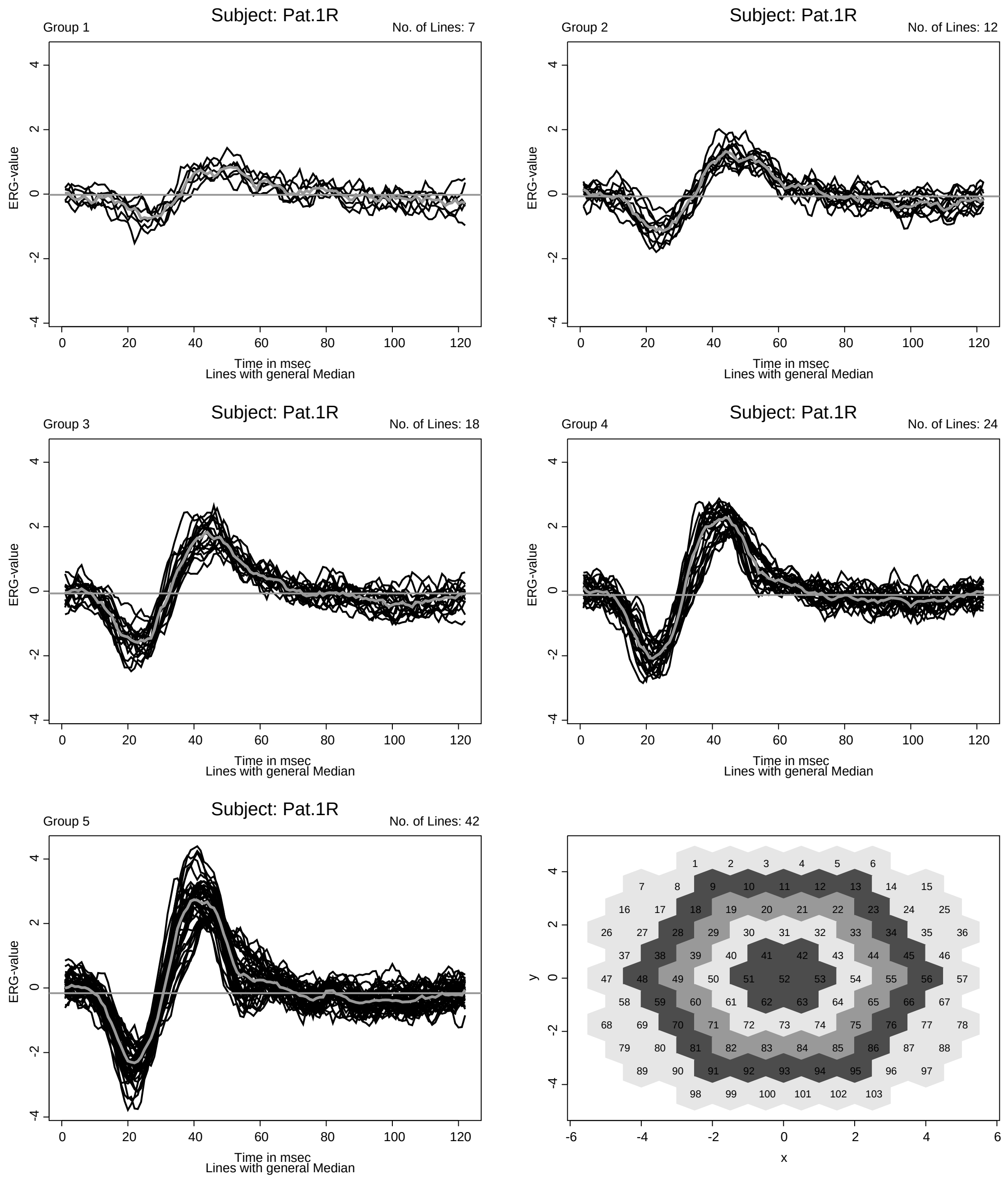

Figure 2.9: *Data set Pat.1R. Multiple time series plots grouped by concentric rings. Group-wise median line is added to highlight changes between groups.*

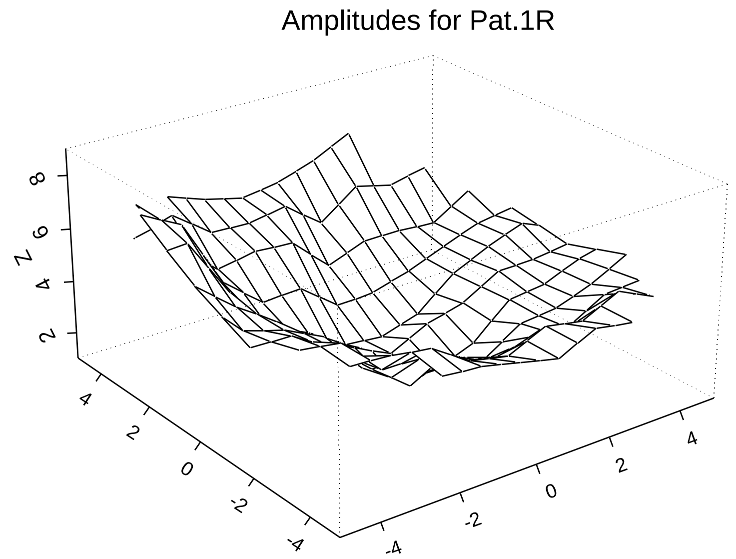

Figure 2.10: *Amplitudes for Pat.1R. Values are interpolated where necessary.*

amplitudes for data set Pat.1R also show a clear spatial component, although different from what was observed for the left eye of the same patient.

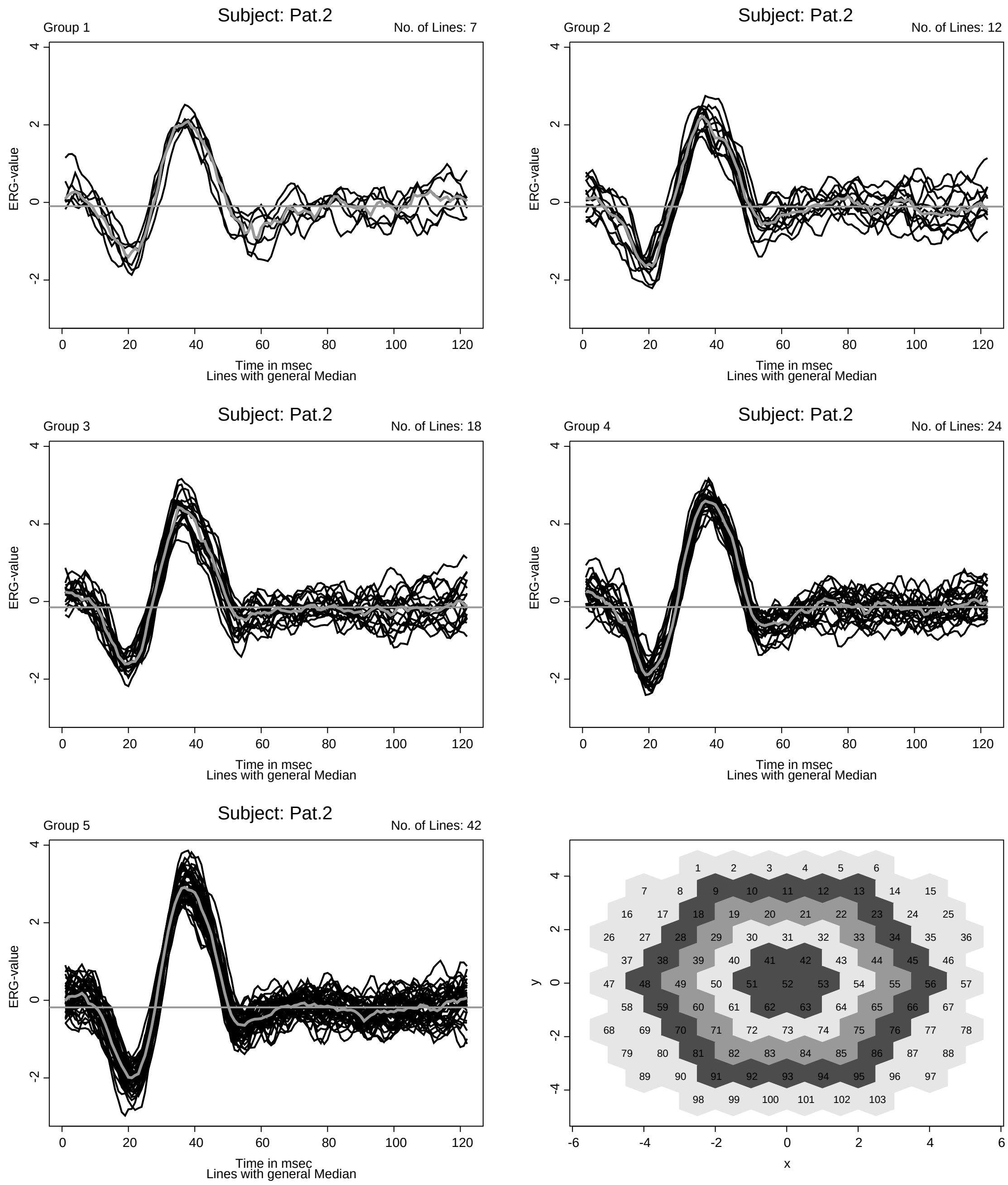

Figure 2.11: *Data set Pat.2. Multiple time series plots grouped by concentric rings. Group-wise median line is added to highlight changes between groups.*

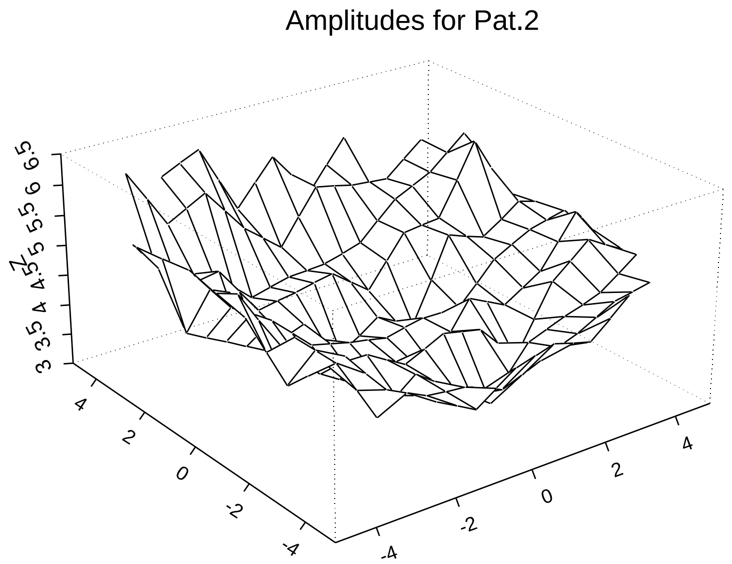

Figure 2.12: *Amplitudes for Pat.2. Values are interpolated where necessary.*

Patient 2

Data set Pat.2 (Figure 2.11) shows several features similar to data set Pat.1R. The median curves for the 5 groups look qualitatively similar when compared to each other. Amplitudes increase from the midpoint of the retina towards the edge. It is interesting to note that the median line for group 2 (upper right) shows a clear bump at 40 msec. A similar bump of much smaller size can also be observed in group 3 (central left) at about 45 msec. This indicates that concentric grouping for Patient 2 may not be fully appropriate, since heterogeneous time series are combined.

The perspective plot of amplitudes for Patient 2 in Figure 2.12 looks similar to that of Pat.1R, but the general quadratic spatial trend with minimum at the retinal center is not as clearly visible. Data set Pat.2 seems to be spatially less homogeneous than the preceding data sets. Nevertheless, the impression that position should be considered in a model as an important factor is confirmed again in this case.

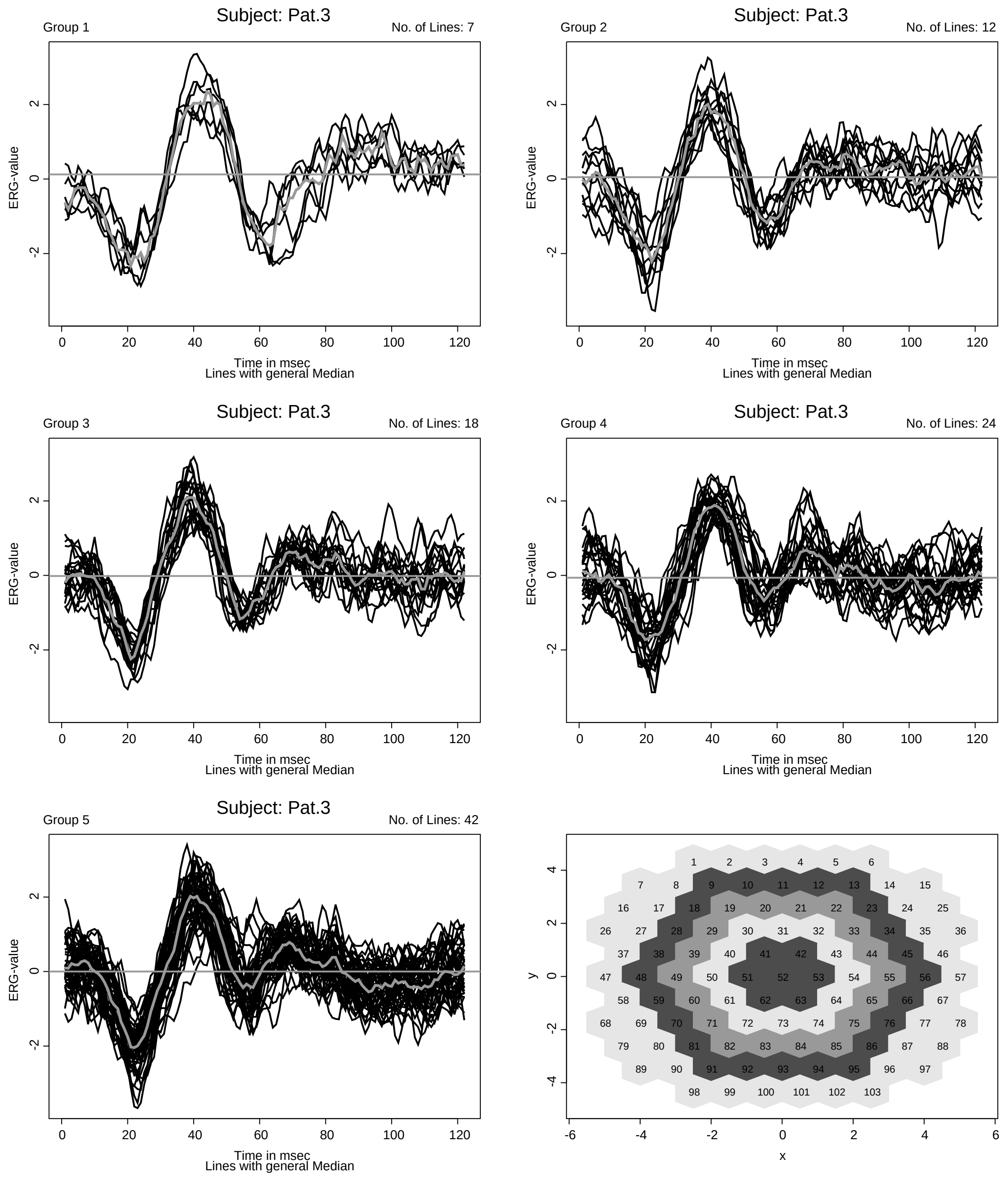

Figure 2.13: *Data set Pat.3. Multiple time series plots grouped by concentric rings. Group-wise median line is added to highlight changes between groups.*

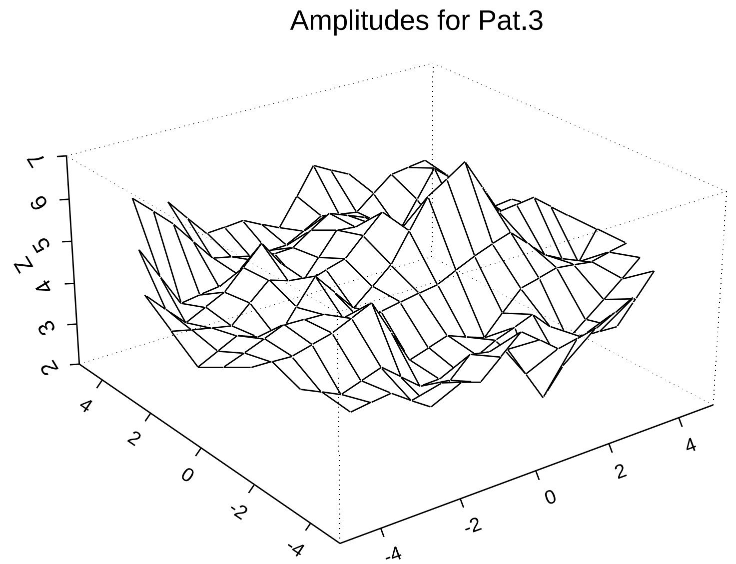

Figure 2.14: *Amplitudes for Pat.3. Values are interpolated where necessary.*

<u>Patient 3</u>

Finally, for Patient 3, the 5 groups of curves are displayed in Figure 2.13. They show rather strong differences among each other. Similar to Pat.1R, there is a roughly linear increase in individual curves between 60 and 90 msec in the first group, while other groups seem to slowly oscillate in this interval with local minimum around 80 msec. The second minimum in the pointwise median line for group 1 occurs around 65 msec, whereas the overall minimum for the complete data set occurs about 10 msec before. Note that the median is somewhat influenced by the outer groups, which contain the majority of group members. Peaks at about 70 and 85 msec are most clearly visible in group 4 (lines 11, 13 and 17 within this group take highest values at 70 msec), although curves with a similar peak in different groups may be hidden by other group members.

The spatial layout of curve amplitudes for Patient 3 differs markedly from the preceding two data sets Pat.1R and Pat.2. Although a spatial trend may be crudely approximated by a quadratic trend surface, it has its maximum at the retinal center, instead of its minimum (see

Figure 2.14). There is no such clear peak as in data set Pat.1L, though.

2.5.2 Spatial Trend in Amplitudes

Summarizing the observations just described, there is clearly some evidence for spatial location having an influence on the observed data values and their amplitudes. Concentric rings are commonly suggested as a possible choice to form groups. It was noted that such grouping may be improved on, since in some cases data curves did not go parallel with the bulk of curves in their respective group.

A first check on the appropriateness of grouping can be done by plotting data values against distance from origin. Ignoring the temporal components and looking only at amplitudes, a roughly quadratic trend was observed in several cases in the explorative analysis above. Such a spatial trend should also be reflected in plots of local amplitudes versus distance or versus angle, provided it is strong enough. Corresponding plots are given in Appendix A. It can be seen from the plots of amplitudes versus angle that for Pat.1R, concentric grouping is quite satisfactory (Figure A.2), while, e.g., for Pat.2, smooth lowess-curves (Cleveland 1979) for groups 2 and 3 intersect and therefore indicate that grouping may be done differently in this case (Figure A.6).

2.6 Spatiotemporal Aspects

2.6.1 Space-Time Data Visualization

It was seen that temporal features like curve shape may vary in space. To see that spatial features may also vary with time, wireframe plots of the observed values were produced for different time points. Figure 2.15 gives an example using data set Pat.1L.

Wireframe plots obtained from the other data sets can be found in Appendix B. Combining

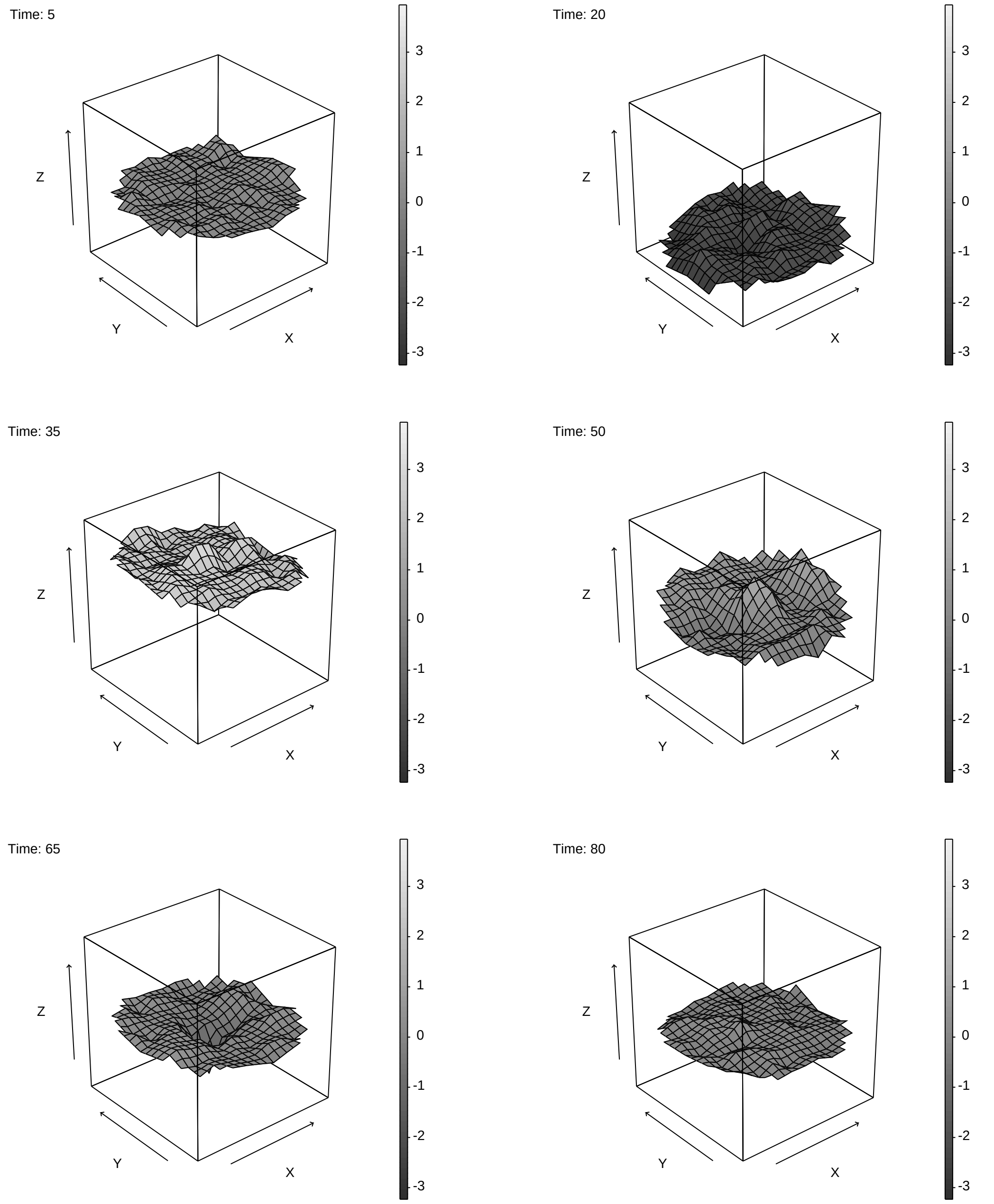

Figure 2.15: *Patient Pat.1L. Measurements at times 5, 20, 35, 50, 65 and 80 milliseconds. Values were linearly interpolated for graphical display.*

	μ	x	y	xy	x^2	y^2	xy^2	yx^2	x^3	y^3
Pat.1L	•	•		•	•	•	•		•	
Pat.1R	•	•		•	•	•	•	•	•	
Pat.2	•		•		•	•		•		•
Pat.3	•	•	•	•	•	•	•			

Table 2.4: *Coefficients exceeding pointwise confidence intervals for 5 or more times.*

wireframes of all 122 time points results in an animation which shows even clearer that the underlying processes are only suboptimally described in general by a purely spatial or purely temporal analysis, respectively. It seems to be more appropriate to fit a model which encompasses both of these two features and therefore allows for a spatiotemporal description of the data.

2.6.2 Spatiotemporal Data Features

Spatial trend components vary markedly in time for all four data sets. This behaviour was explored somewhat closer by fitting a polynomial spatial trend up to order 3 both in x and y direction. The estimated parameter values for Pat.1L are displayed together with pointwise intervals at 1.96 times their estimated standard error in Figures 2.16 and 2.17. In this naive exploratory analysis, the simplifying assumption of independently distributed random errors was made. Both the estimators and the associated intervals would have to be modified under an improved error model. Corresponding graphical displays for the other data sets are given in Appendix C. Table 2.4 shows for all four data sets which coefficients exceed the pointwise confidence regions for 5 or more times over a period of 122 msec.

For all data sets, each of the ten sequences of regression coefficients exceeds the pointwise confidence intervals at least once over the measuring period of 122 msec, the only exception being the cubic coefficient in x for data set Pat.3. In all four case, the mean as well as the quadratic terms both in x- and y-direction seem to be important for modeling trend over time.

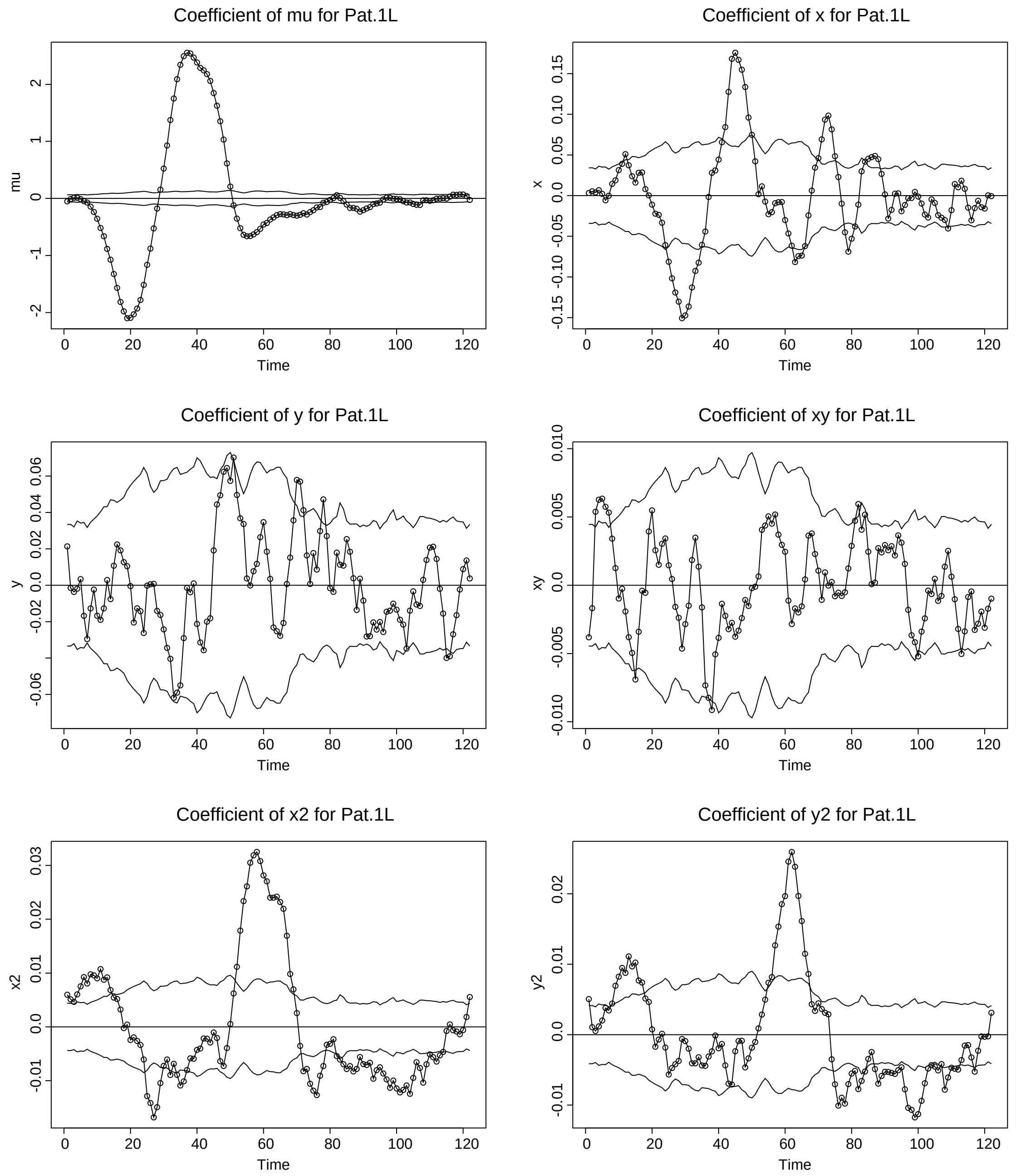

Figure 2.16: *Patient Pat.1L. Coefficients for intercept, x, y, $x * y$, x^2 and y^2. Pointwise 95 percent reference intervals under the assumption of independence.*

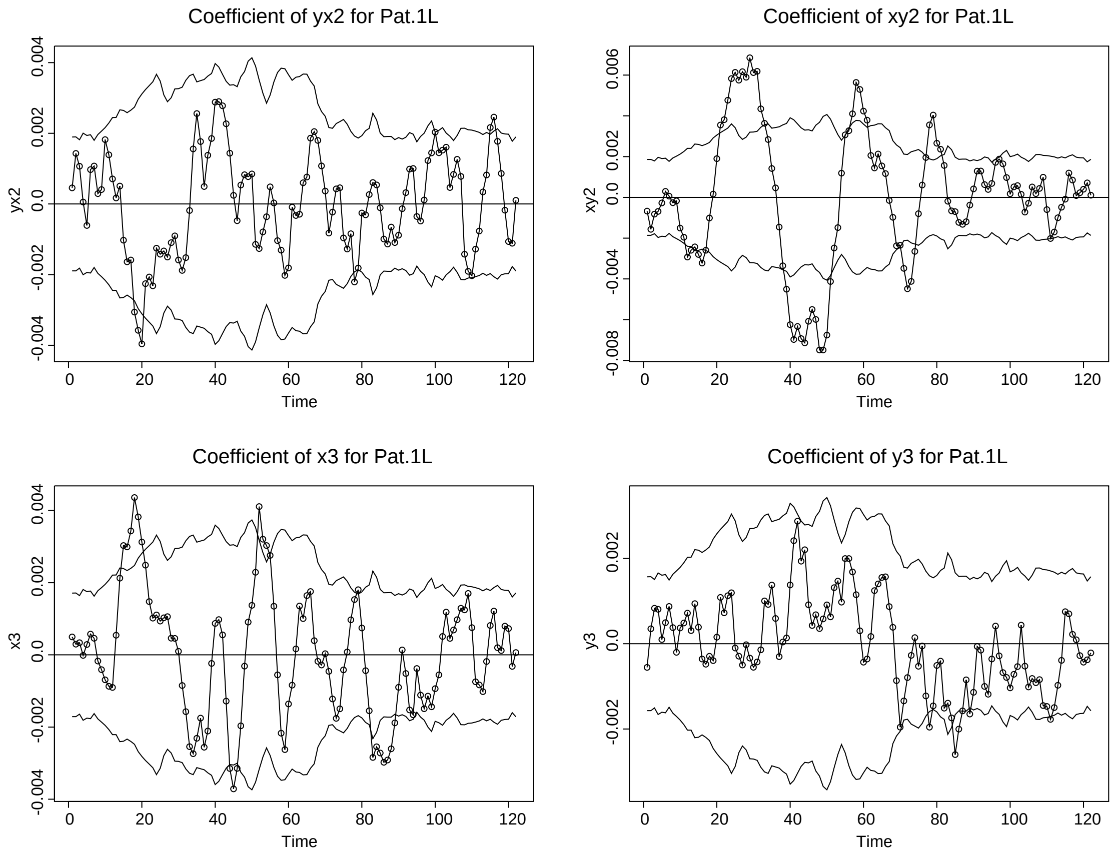

Figure 2.17: *Patient Pat.1L. Coefficients for intercept, yx^2, xy^2, x^3 and y^3. Pointwise 95 percent reference intervals under the assumption of independence.*

Other features common to two or more data sets were not detected, even though some interesting patterns within single data sets are observable. Examples are the damped sinusoidal behaviour of the coefficient estimate for the xy interaction for Pat.1R (Figure C.1, central right), or the striking downward trend in the same coefficient for Pat.2 (Figure C.3).

In summary, a crude fit of polynomial spatial trend confirms that there are variations over time which are very heterogeneous between the four patients under study. This result was obtained with a simple model of the error process involved, which possibly also inhibits both spatial and temporal dependencies. Note that a polynomial trend is not adapting well to localized trend features. It may be helpful in describing the overall structure in a data set, but at the same time may hide certain local features like, for example, a consistently low response at the hexagonal area which includes the blind spot. When prediction of the response is of major interest, this can be accounted for by appropriate joint modeling of trend and error process, as in Berke (1998). However, in the AMD study emphasis was put on characterization of the data sets. A first step towards this end is to parameterize the temporal evolvement inherent to the data by using techniques from time series analysis. This will be the starting point for a more involved spatiotemporal data description.

Chapter 3

Temporal Data Analysis

The analysis of spatiotemporal data is based on a combination of theoretical work from both spatial statistics and time series analysis. In this chapter, a short review is given of the concepts of univariate and multivariate time series analysis.

Focus is put on the estimation of so-called autoregressive parameters, which will serve as the building block of the spatiotemporal analysis of multifocal ERG data. Definitions given here are mainly based on the monographs by Box, Jenkins, and Reinsel (1994), and Reinsel (1997).

3.1 Stochastic Processes

A univariate time series may be regarded as a realization of a sequence of random variables that take values over time. This point of view is taken by the theory of stochastic processes, which builds the base of modern time series analysis.

Definition 3.1.1. A sequence of random variables $(Y_t)_{t\in\mathbb{T}}$ with time index t and index set $\mathbb{T} \subseteq \mathbb{Z}$ is called a *stochastic process*. A realization of a stochastic process is called a *time series* and is denoted by (y_t). ♦

In what follows, $\mathbb{T}$ represents a finite subset of the set of all integers $\mathbb{Z}$ with cardinal number

T. It is assumed that observations are taken at discrete and equally spaced points in time. The elements of the process $(Y_t)_{t\in\mathbb{T}}$ are assumed to take real values and follow a normal (or *Gaussian*) distribution with expected value $E(Y_t) = \mu_t$ and finite variance $E(Y_t - \mu_t)^2 = \sigma_t^2$, unless noted otherwise. The linear dependency between random variables Y_u and Y_v ($u, v \in \mathbb{T}$) is measured by their (temporal) covariance $\gamma(u, v)$ which will be defined below. Note that in general the mean, the variance and the covariance of a stochastic process may vary in time, unless further assumptions are made.

Temporal stochastic processes differ from other data structures in that they may exhibit dependencies over time. Otherwise, they could be simply modeled as a sequence of independent and possibly identically distributed random variables. Instead, often values measured within a short time interval are assumed to be more alike than those further apart. Models for dependencies of such type will be considered below. Before this can be done, some additional definitions are introduced.

3.1.1 Stationary Processes

If the distribution of every element Y_u of the stochastic process $(Y_t)_{t\in\mathbb{T}}$ was allowed to have parameters freely varying over time, parameter estimation would be virtually impossible. Some additional assumptions have to be made to reduce the number of parameters, or to restrict the way they may behave. This leads to the concept of stationarity.

Definition 3.1.2. The stochastic process $(Y_t)_{t\in\mathbb{T}}$ is said to be *stationary in mean* if for the corresponding expected values it holds true that $\mu_t = \mu$ for all $t \in \mathbb{T}$.
♦

Every process with *known* sequence of means can easily be transformed into a mean stationary process by simply subtracting the mean for each t. Thus, it is common to assume $\mu = 0$ without loss of generality. The second distributional parameter of interest is the variance.

Definition 3.1.3. The stochastic process $(Y_t)_{t\in\mathbb{T}}$ is called *stationary in variance* if the variance $\sigma_t^2 = \sigma^2$ for all $t \in \mathbb{T}$.
♦

If a process is stationary both in mean and variance, the situation still differs from the independent and identically distributed (i.i.d.)-case, since dependencies between elements in $(Y_t)_{t\in\mathbb{T}}$ are still possible. A considerable simplification is achieved by introducing stationarity in covariance as well. *Autocovariances* are the tool to do so. They describe the linear relation between temporally lagged components of $(Y_t)_{t\in\mathbb{T}}$.

Definition 3.1.4. For $u, v \in \mathbb{T}$, the *autocovariance* $\gamma(u, v)$ between the two elements Y_u and Y_v of a variance stationary stochastic process $(Y_t)_{t\in\mathbb{T}}$ with constant mean μ and variance σ^2 is given by

$$\gamma(u, v) = E\left[(Y_u - \mu)(Y_v - \mu)\right].$$

The corresponding *autocorrelation* is then given by

$$\rho(u, v) = \frac{\gamma(u, v)}{\sigma^2}$$

A subscript may be added to identify the process involved.
♦

Definition 3.1.5. The stochastic process $(Y_t)_{t\in\mathbb{T}}$ is said to be *stationary in covariance* (or covariance stationary) if for all pairs (Y_u, Y_v) with $u, v \in \mathbb{T}$ the sequence of covariances $\gamma(u, v)$ depends only on the difference, or *lag*, between u and v, i.e.,

$$\gamma(u, v) = \gamma(h)$$

with $h = |u - v|$. In this case, $\rho(u, v) = \rho(h)$.
♦

Clearly, covariance stationarity includes variance stationarity by setting $u = v$. Combining the above properties results in the definition of stationarity:

Definition 3.1.6. A stochastic process is called *(weakly* or *second order) stationary*, if it is stationary both in mean and covariance.
♦

When working with normally distributed data, weak stationarity is an assumption that simplifies the analysis considerably, since the normal distribution is completely specified by moments up to second order. A stationary process is thus one with equal distributional properties at all points in time.

In general, however, moments of higher order than two are necessary to describe the probability density of a stochastic process. Such cases are dealt with by the following definition.

Definition 3.1.7. A stochastic process $(Y_t)_{t\in\mathbb{T}}$ is called *strictly stationary*, if the joint probability distribution of any finite n-subset of random variables in this process is invariant to time shift h, i.e.,

$$P\left(Y_{t_1} \leq y_1, \cdots, Y_{t_n} \leq y_n\right) = P\left(Y_{t_1+h} \leq y_1, \cdots, Y_{t_n+h} \leq y_n\right)$$

for all $t_1, \ldots, t_n$ in $\mathbb{T}$ and $h \in \mathbb{Z}$.
♦

Under the assumption of a normal distribution, second order and strict stationarity coincide. In what follows, *stationarity* will always refer to second order stationarity, and normality is assumed.

3.1.2 Nonstationary Processes

The assumption of stationarity is deliberately restrictive. A few comments are in order about the theoretical tools available for processes that are *not* stationary. Nonstationary processes are primarily defined by the absence of stationarity, see for example Priestley (1988, Chap. 6). A common strategy to tackle them is to concentrate on subclasses of (possibly) nonstationary processes which obey certain conditions on their behaviour, which in turn make their parameters estimable. Examples for models for nonstationary processes are, among others, ARIMA-models (Box, Jenkins, and Reinsel 1994, Chapter 4) or Priestley's State Dependent Models (Priestley 1980; Priestley 1988, Chapter 5).

A somewhat different approach to the analysis of time series is to use wavelets for estimation. See Mallat (1989), Daubechies (1992), Chui (1992). Krahnke (1997) describes an application. Wavelets help to find a so-called *time-scale* decomposition of the data which differs from the commonly used time-frequency decomposition in that it is well localized both in time *and* space. However, Priestley (1996) notes that the concept of frequency, which is widely used in time series analysis, is not exchangeable with the idea of different scales. An additional drawback is that when applying wavelet methods to a time series of finite length, edge effects may arise which possibly yield misleading results.

Wavelets may generally be more useful for spatial smoothing rather than temporal decomposition, since they can be chosen in such a way that the resulting fit shows a certain degree of smoothness. The amount depends on the number of existing derivatives of the underlying mother wavelet which was used to construct the underlying wavelet basis. This is somewhat similar to splines (see, e.g., Green and Silverman 1994) in that some smoothness is built-in into the basis functions used. However, a considerable amount of sampling locations (or time points, respectively) is needed for a satisfactory fit of wavelets to the data, and edge effects may occur. In the multifocal ERG data sets at hand, only 103 locations in two-dimensional space are given. This is quite a small number, which is a major practical reason why wavelet techniques will not be pursued here any further.

3.2 Univariate Autoregressive and Moving Average Processes

3.2.1 Basic Model Formulation

A possible way to characterize the structure of a stochastic process is to allow for past values of Y_{t-k} $(k > 0)$ to enter linearly into the current value of Y_t. A random component ϵ_t may be added to allow for an additional random change.

Definition 3.2.1. A stochastic process $(Y_t)_{t\in\mathbb{T}}$ is called an *autoregressive process of order p*, or AR(p) process for short, if it follows the relation

$$Y_t = \sum_{k=1}^{p} \alpha_k Y_{t-k} + \epsilon_t \tag{3.2.1}$$

Here, the α_k describe the linear dependence on preceding realizations, $(\epsilon_t)_{t\in\mathbb{T}}$ is a *White Noise* process (WNP), i.e., it has expected value $E(\epsilon_t) = 0$ and variance $Var(\epsilon_t) = \sigma_\epsilon^2$. The components of the noise process are assumed to be pairwise stochastically independent, and so are the pairs (Y_t, ϵ_{t+k}) for all $t \in \mathbb{T}$ and $k > 0$.
♦

In what follows it is assumed that the random shocks ϵ_t are normally distributed, hence zero-correlation between them coincides with stochastic independence. In addition, only *causal* processes will be treated, where causality is given when the summation in (3.2.1) does not extend to future (unobserved) values. An alternative model for stochastic processes is the perturbation model. The observed random variable is taken to come from a series of random shocks which are adding up over time.

Definition 3.2.2. A stochastic process $(Y_t)_{t\in\mathbb{T}}$ is called a *moving average process of order q*, or MA(q)-process for short, if it obeys the relation

$$Y_t = \epsilon_t - \sum_{l=1}^{q} \beta_l \epsilon_{t-l} \tag{3.2.2}$$

where $(\epsilon_t)_{t\in\mathbb{T}}$ is a WNP.
♦

The MA-model and the AR-model can be combined, yielding an even more general class of stochastic processes.

Definition 3.2.3. A (univariate) *autoregressive moving average process of order p and q*, or ARMA(p,q)-process for short, follows the relation

$$Y_t = \sum_{k=1}^{p} \alpha_k Y_{t-k} + \epsilon_t - \sum_{l=1}^{q} \beta_l \epsilon_{t-l}$$

with coefficients $\beta_l \in \mathbb{R}$, and $(\epsilon_t)_{t\in\mathbb{T}}$ denoting a WNP.
♦

3.2.2 Covariance Structure of ARMA Processes

Linear dependencies between temporally lagged values of a univariate time series can be described by means of autocovariances and autocorrelations. The autocovariance function γ of a zero mean stationary ARMA(p,q) process $(Y_t)_{t\in\mathbb{T}}$ can be derived from the relation (with $\beta_0 = -1$)

$$\begin{aligned}
\gamma(k) &= E\left[Y_{t-k} Y_t\right] && (3.2.3)\\
&= E\left[Y_{t-k}\left(\sum_{i=1}^{p} \alpha_i Y_{t-i} - \sum_{j=0}^{q} \beta_j \epsilon_{t-j}\right)\right] && (3.2.4)\\
&= \sum_{i=1}^{p} \alpha_i \gamma(i-k) - \sum_{j=0}^{q} \beta_j \gamma_{Y,\epsilon}(j-k) && (3.2.5)
\end{aligned}$$

where $\gamma_{Y,\epsilon}(k) = E[Y_t \epsilon_{t+k}]$, i.e. the *cross-covariance* between the observed values and the current shock ϵ_t. For an AR(p) process, for example, one obtains

$$\gamma(k) = \sum_{i=1}^{p} \alpha_i \gamma(i-k) \tag{3.2.6}$$

or, equivalently, in terms of the autocorrelations,

$$\rho(k) = \sum_{i=1}^{p} \alpha_i \rho(i-k) \tag{3.2.7}$$

These equations ar known as Yule-Walker equations and can be useful when deriving preliminary estimators for the autoregressive coefficients.

Once estimates of the autocorrelation coefficients are available, partial autocorrelations can be approximated using a recursive formula due to Durbin (1960) to support model identification.

The univariate analysis of the ERG data observed for specific hexagonal areas provides a first impression of the complexity of the locally underlying physiological process. However, it does not make use of the fact that neighboring areas may behave similar. Cross-correlations between univariate time series should be inspected and possibly incorporated into the analysis. This can be done using vector time series models.

3.3 Vector ARMA-Processes

3.3.1 Multivariate Model Formulation

Multivariate time series can be conveniently handled in vector form. They are described in several books, for example Hannan (1970) and Reinsel (1997). Many aspects of univariate time series carry over to the multivariate case. However, special care has to be taken with respect to parameter estimation. For example, solutions to multivariate ARMA-equations do not generally uniquely identify the underlying process, as they do in the univariate case. To fix notation, different kinds of vector processes are defined, before estimation in the multivariate case is addressed.

Definition 3.3.1. A *d-dimensional stochastic vector process* $\left(\vec{Y}_t\right)_{t\in\mathbb{T}}$ consists of vector-valued components $\vec{Y}_t = (Y_t(1), \ldots, Y_t(d))' \in \mathbb{R}^d$, where each element $(Y_t(s))_{t\in\mathbb{T}}$, $1 \leq s \leq d$ itself represents a univariate stochastic process.
♦

Each element $Y_t(s), s \in \{1, \ldots, d\}$, has associated with it the expected value and variance

$$E\left[Y_t(s)\right] = \mu_t(s) \tag{3.3.8}$$

$$Var\left[Y_t(s)\right] = \sigma_t^2(s) \tag{3.3.9}$$

The expected value of the vector $\vec{Y_t}$ is given by $E[\vec{Y_t}] = \vec{\mu}_t = (\mu_t(1), \ldots, \mu_t(d))'$. In the sequel, mean stationarity will be assumed for all components of $\vec{Y_t}$ for all t, implying $\vec{\mu}_t = \vec{\mu}$. Therefore, $\vec{\mu} = 0$ can be assumed without loss of generality.

If $\left(\vec{Y_t}\right)_{t\in\mathbb{T}}$ is a purely nondeterministic stationary process, Wold's Theorem in its multivariate version shows that it can be represented as an infinite vector moving average process (cf. Reinsel 1997, Sec. 1.2.1).

Definition 3.3.2. A causal infinite *Vector Moving Average Process* of (possibly infinite) order q (VMA(q) process) is defined through the relation

$$\vec{Y_t} = \vec{\mu} + \sum_{j=0}^{q} \mathbf{B}_j \vec{\epsilon}_{t-j}$$

where the $\mathbf{B}_j, j \in \mathbb{Z} \cup \{\infty\}$, are coefficient matrices in $\mathbb{R}^{d\times d}$ with $\sum_{j=0}^{q} ||\mathbf{B}_j||^2 < \infty$, and $\mathbf{B}_0 = I_{d\times d}$ is defined to be the identity matrix. The error process $(\vec{\epsilon_t})_{t\in\mathbb{T}}$ is a vector valued White Noise Process, such that $E(\vec{\epsilon_t}) = \vec{0} \in \mathbb{R}^d$, $E[\vec{\epsilon}_t\vec{\epsilon'}_t] = \Sigma_\epsilon$ is positive definite, and $E[\vec{\epsilon}_t\vec{\epsilon'}_{t+h}] = \mathbf{0}$ for $h \neq 0$.
♦

The often more parsimonious finite-dimensional vector ARMA representation has the following form.

Definition 3.3.3. A *Vector Autoregressive Moving Average process* of order p and q ($p, q \in \mathbb{N}_0$), or VARMA(p,q) process for short, is defined by

$$(\vec{Y_t} - \vec{\mu}) - \sum_{k=1}^{p} \mathbf{A}_k (\vec{Y}_{t-k} - \vec{\mu}) = \vec{\epsilon}_t - \sum_{j=1}^{q} \mathbf{B}_j \vec{\epsilon}_{t-j} \tag{3.3.10}$$

where $E\left[\vec{Y_t}\right] = \vec{\mu}$ for all t, and $\mathbf{A}_k$ and $\mathbf{B}_j$ are coefficient matrices in $\mathbb{R}^{d\times d}$, and $(\vec{\epsilon_t})$ is a vector valued WNP. Without loss of generality, $\vec{\mu}$ may be taken to equal zero. If all $\mathbf{B}_j = \mathbf{0}$,

one obtains a *Vector Autoregressive Process* of order p, or VAR(p) process.
♦

The motivation to introduce vector processes was to allow for dependencies between neighboring univariate time series. These are specified by means of covariances. The next section gives the multivariate formulation of the covariance structure of vector processes in general terms.

3.3.2 General VARMA Covariance Structure

The associated autocovariance matrix of two elements of a zero-mean vector process $\left(\vec{Y}_t\right)_{t\in\mathbb{T}}$ at times $(t-h)$ and t, say, is given by

$$\Gamma(t-h,t) = E[\vec{Y}_{t-h}\vec{Y}_t'] \tag{3.3.11}$$

The concept of stationarity carries over from the univariate to the multivariate case. The multivariate stochastic process $(\vec{Y}_t)$ is said to be stationary if the probability distribution of $\vec{Y}_t$ is the same as the distribution of $\vec{Y}_{t+h}$ for all $h \in \mathbb{Z}$, where h is the time lag. For second order stationary processes, this means that the cross-covariance between components $\vec{Y}_t(i)$ and $\vec{Y}_{t+h}(j)$ only depends on time lag h, not on time t. Hence, it can be written as $\Gamma(t-h,t) = \Gamma(h)$.

As in the univariate case, only stationary processes will be considered here. Since several univariate processes are involved, a description of their association has to be given to completely characterize the process.

Definition 3.3.4. The *cross-covariance* between two zero-mean univariate jointly covariance stationary time series $Y_t(i)$ and $Y_{t+h}(j)$ at lag $h \in \mathbb{Z}$ is given by

$$\gamma_{i,j}(h) = Cov(Y_t(i), Y_{t+h}(j)) = E\left[(Y_t(i))(Y_{t+h}(j))\right]$$

These components define the *cross-covariance matrix* at lag $h \in \mathbb{Z}$,

$$\Gamma(h) = E[\vec{Y}_t \vec{Y}'_{t+h}] = \begin{pmatrix} \gamma_{1,1}(h) & \gamma_{1,2}(h) & \dots & \gamma_{1,d}(h) \\ \gamma_{2,1}(h) & \gamma_{2,2}(h) & \dots & \gamma_{2,d}(h) \\ \vdots & \vdots & \dots & \vdots \\ \gamma_{d,1}(h) & \gamma_{d,2}(h) & \dots & \gamma_{d,d}(h) \end{pmatrix}$$

and the *cross-correlation matrix*

$$\mathbf{R}(h) = \mathbf{D}^{-1/2}\Gamma(h)\mathbf{D}^{-1/2} = \begin{pmatrix} \rho_{1,1}(h) & \rho_{1,2}(h) & \dots & \rho_{1,d}(h) \\ \rho_{2,1}(h) & \rho_{2,2}(h) & \dots & \rho_{2,d}(h) \\ \vdots & \vdots & \dots & \vdots \\ \rho_{d,1}(h) & \rho_{d,2}(h) & \dots & \rho_{d,d}(h) \end{pmatrix}$$

where $\rho_{i,j}(h) = \frac{\gamma_{i,j}(h)}{\sqrt{\gamma_{i,i}(0)\gamma_{j,j}(0)}}$ and $\mathbf{D}^{-1/2} = Diag\left(\sqrt{\gamma_{1,1}(0)}, \dots, \sqrt{\gamma_{d,d}(0)}\right)^{-1}$.
♦

Note that $\Gamma(h)$ is not symmetric in general, but rather

$$\Gamma'(h) = \Gamma(-h) \text{ and } \mathbf{R}'(h) = \mathbf{R}(-h).$$

3.4 Parameter Estimation for VAR-Models

The Wold decomposition of a VARMA process can be obtained by defining the errors $\vec{\epsilon}_t$ in the infinite representation of a purely nondeterministic stationary process as the residuals of the best linear one-step ahead prediction $\hat{\vec{Y}}_{t-1;1}$ of $\vec{Y}_t$. This predictor can be expressed as

$$\hat{\vec{Y}}_{t-1;1} = \sum_{j=1}^{\infty} \mathbf{B}_j(\hat{\vec{Y}}_{t-j} - \hat{\vec{Y}}_{t-j-1;1}) = \sum_{j=1}^{\infty} \mathbf{B}_j \vec{\epsilon}_{t-j} \tag{3.4.12}$$

The matrices $\mathbf{B}_j$ may therefore be interpreted as the regression matrices of $\vec{Y}_{t-1}$ on the errors $\vec{\epsilon}_{t-j}$. Along the lines of standard regression theory, they thus take the form

$$\mathbf{B}_j = Cov(\vec{Y}_t, \vec{\epsilon}_{t-j})\Sigma_\epsilon^{-1}. \tag{3.4.13}$$

The coefficient matrices $\mathbf{B}_j$ can be estimated in several different ways, for example using iterative procedures as described in Reinsel (1997, Sec. 5.1). The resulting estimator of the complete set of VARMA-parameters has no closed form. For this reason, only autoregressive processes will be considered below for parameter estimation. The estimation approaches reviewed are the Yule-Walker equations, the method of conditional ordinary least squares (OLS), and the maximum likelihood (ML) technique.

3.4.1 The Yule-Walker Equations

Given the coefficient matrices $\mathbf{A}_k$ and $\mathbf{B}_j$ of a VARMA(p,q)-process $(\mathbf{Y}_t)_{t\in\mathbb{T}}$ it is possible to obtain the autocovariance matrices $\Gamma(h)$ by solving simultaneously a set of linear equations (e.g. Reinsel 1997, Appendix A.2.3). For $h \in \{0, 1, \dots, p\}$, the $\Gamma(h)$ have the representation

$$\Gamma(h) = \sum_{k=1}^{p} \Gamma(h-k)\mathbf{A}'_k - \sum_{j=h}^{q} \mathbf{B}_{h-j}\Sigma_\epsilon \mathbf{B}'_j \tag{3.4.14}$$

$$= \sum_{k=1}^{p} \Gamma(h-k)\mathbf{A}'_k + \mathbf{G}_h \tag{3.4.15}$$

where $\mathbf{B}_0 = -I$ and $\mathbf{G}_h = -\sum_{j=h}^{q} \mathbf{B}_{j-h}\Sigma_\epsilon \mathbf{B}'_j$ for $h \leq q$ and $\mathbf{G}_h = 0$ if $h >$ q. In the VAR(p)-case all $\mathbf{B}_j = 0$ for $j > 0$, and only the AR-coefficient matrices and Σ_ϵ are necessary to determine the autocovariances. Equation (3.4.14) then simplifies to the *Yule-Walker equations* (Durbin 1960, Walker 1931, Yule 1927)

$$\sum_{k=1}^{p} \Gamma(h-k)\mathbf{A}'_k = \Gamma(h) \tag{3.4.16}$$

for $h = 1, 2, \dots$. The error covariance Σ_ϵ enters into the recursion only through $\Gamma(0) = \sum_{k=1}^{p} \Gamma(-k)\mathbf{A}'_k + \Sigma_\epsilon$.

If instead the autocovariance matrices $\Gamma(k)$ are known for $k \in \{0, 1, \dots, p\}$, the AR-parameter matrices $\mathbf{A}_k$ and the error covariance Σ_ϵ can be obtained from the same set of equations. Using the notation

$$\mathbf{A} = \left(\mathbf{A}_1', \ldots, \mathbf{A}_p'\right)' \in \mathbb{R}^{dp\times d} \tag{3.4.17}$$

$$\Gamma_p = (\Gamma(1)', \ldots, \Gamma(p)')' \in \mathbb{R}^{dp\times d} \tag{3.4.18}$$

$$\Gamma = \begin{pmatrix} \Gamma(0) & \Gamma(1)' & \Gamma(2)' & \ldots & \Gamma(p-1)' \\ \Gamma(1) & \Gamma(0) & \Gamma(1)' & \ldots & \Gamma(p-2)' \\ \Gamma(2) & \Gamma(1) & \Gamma(0) & \ldots & \Gamma(p-3)' \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Gamma(p-1) & \Gamma(p-2) & \Gamma(p-3) & \ldots & \Gamma(0) \end{pmatrix} \in \mathbb{R}^{dp\times dp} \tag{3.4.19}$$

the set of equations is written in matrix form as

$$\Gamma\mathbf{A} = \Gamma_p \tag{3.4.20}$$

with solution

$$\mathbf{A} = \Gamma^{-1}\Gamma_p \tag{3.4.21}$$

The estimate for Σ_ϵ is finally obtained from

$$\Sigma_\epsilon = \Gamma(0) - \sum_{k=1}^{p} \Gamma(-k)\mathbf{A}_k' \tag{3.4.22}$$

$$= \Gamma(0) - \sum_{k=1}^{p} \Gamma(k)'\mathbf{A}_k' \tag{3.4.23}$$

$$= \Gamma(0) - \Gamma_p'\mathbf{A} \tag{3.4.24}$$

$$= \Gamma(0) - \Gamma_p'\Gamma^{-1}\Gamma_p \tag{3.4.25}$$

The autocovariance at lag h can be estimated consistently from a finite sample of T observations $\left(\vec{Y}_1, \ldots, \vec{Y}_T\right)'$ by their sample covariance matrix

$$\hat{\Gamma}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h}(\vec{Y}_t - \bar{\vec{Y}})(\vec{Y}_{t+h} - \bar{\vec{Y}})' \tag{3.4.26}$$

for $h \in \mathbb{N}_0$, where $\bar{\vec{Y}}$ is the arithmetic mean over all vector observations. The elements of $\hat{\Gamma}(h)$ are

$$\hat{\gamma}_{i,j}(h) = \frac{1}{T-h} \sum_{t=1}^{T-h} (Y_t(i) - \bar{Y}(i))(Y_{t+h}(j) - \bar{Y}(j)) \tag{3.4.27}$$

with $\bar{Y}(i)$ denoting the sample mean over the i-th component. The properties of $\hat{\gamma}_{i,j}(h)$'s are examined in Jenkins and Watts (1968). Inserting them into the Yule-Walker equations (3.4.16) allows for efficient estimation of the process parameters. However, the resulting estimates are very sensitive to rounding errors if the process is close to nonstationarity (Box et al. 1994, p.88; Reinsel 1997, p.91). Conditional ordinary least squares estimates are reported to be favorable in this case.

3.4.2 Conditional Ordinary Least Squares

An alternative approach to the Yule-Walker equations (3.4.16) are regression-like estimates calculated conditional on the p observations preceding time t. Given a stationary VAR(p) process with zero mean, set up the equations

$$\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{E} \tag{3.4.28}$$

with

$$\begin{aligned}
\mathbf{X}_{(T-p)\times dp} &= \begin{pmatrix} \vec{Y}_p' & \dots & \vec{Y}_1' \\ \vdots & \ddots & \vdots \\ \vec{Y}_{T-1}' & \dots & \vec{Y}_{T-p}' \end{pmatrix} && (3.4.29) \\
\mathbf{A}_{dp\times p} &= (\mathbf{A}'_1, \dots, \mathbf{A}'_p)' \quad \text{with blocks } \mathbf{A}_k \in \mathbb{R}^{d\times d} && (3.4.30) \\
\mathbf{Y}_{(T-p)\times p} &= \left(\vec{Y}_{p+1}, \dots, \vec{Y}_T\right)' && (3.4.31) \\
\mathbf{E}_{(T-p)\times p} &= (\vec{\epsilon}_{p+1}, \dots, \vec{\epsilon}_T)' && (3.4.32)
\end{aligned}$$

Along the lines of multivariate linear model theory, the least squares estimator of $\mathbf{A}$ is given by

$$\hat{\mathbf{A}}_{dp\times p} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \tag{3.4.33}$$

provided the inverse exists, where

$$\tilde{\mathbf{Y}} = \left(\vec{Y}_{p+1} - \bar{Y}_{(0)}, \ldots, \vec{Y}_T - \bar{Y}_{(0)}\right)' \tag{3.4.34}$$

and $\tilde{\mathbf{X}}$ has rows $\left((\vec{Y}_{t-1} - \bar{Y}_{(1)})', \ldots, (\vec{Y}_{t-p} - \bar{Y}_{(p)})'\right)$. Here, $\bar{Y}_{(i)} = \frac{1}{T-p}\sum_{t=p+1}^{T}\vec{Y}_{t-i}$. Note that the number of parameters in $\mathbf{A}$ is d^2p. For a VAR(1) process with $d = 103$ components and no restrictions on the covariance structure, this leads to $d^2p = 10609$ parameters for $\mathbf{A}$, and $d(d+1)/2 = 5356$ parameters for the error covariance matrix Σ_ϵ. Such an unrestricted model is not estimable with the given MF-ERG data sets which consist of only 12,566 data points. This is a common problem in spatial data analysis. A possible solution is to make certain structural assumptions on the covariances. These will be treated in some more detail below, when aspects of spatial statistics are considered.

Variance of least squares estimates

Hannan (1970, Chap. 6) shows that $\hat{\vec{\alpha}} = vec(\hat{\mathbf{A}})$ converges to a normally distributed random variable if suitably scaled. Here, the $vec(.)$-operator stacks the columns of a matrix beneath each other to form a vector. More precisely,

$$\frac{1}{\sqrt{T-p}}(\hat{\vec{\alpha}} - \vec{\alpha}) \to N\left(\vec{\mathbf{0}}, \Sigma_\epsilon \otimes \Gamma^{-\mathbf{1}}\right) \tag{3.4.35}$$

with $T \to \infty$ and Γ as defined in (3.4.19), and $\otimes$ denoting the *Kronecker matrix product*, i.e., for two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}_{m\times n} \otimes \quad \mathbf{B}_{p\times q} = \begin{pmatrix} a_{11}\mathbf{B} & a_{1n}\mathbf{B} & \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{mn}\mathbf{B} & \end{pmatrix}_{(mp)\times(nq)}$$

Approximate covariance estimates for the parameters are given by

$$\hat{\Sigma}_\epsilon \otimes \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tag{3.4.36}$$

using the estimate of the error covariance matrix

$$\hat{\Sigma}_\epsilon = \frac{1}{T-dp}\sum_{t=p+1}^{T} \tilde{e'}_t\tilde{e}_t \tag{3.4.37}$$

with residuals $\tilde{e}_t = (\vec{Y}_t - \bar{Y}_{(0)}) - \hat{\mathbf{A}}' \left[(\vec{Y}_{t-1} - \bar{Y}_{(1)})', \dots, (\vec{Y}_{t-p} - \bar{Y}_{(p)})'\right]'$ for $t = p+1, \dots, T$.

To avoid unnecessary cluttering of notation, centering of the observations by their empirical mean is implicitly assumed to have been performed before the analysis, and the tilde symbols indicating demeaning are dropped hereafter.

3.4.3 Maximum Likelihood

There are two different approaches to maximum likelihood (ML) estimation commonly used in time series analysis. The *exact* ML estimation treats all observed values similarly and includes them into the estimation procedure, while the *conditional* approach considers the first few observations as fixed, and performs estimation conditional on their value.

Conditional ML Estimation

Given normally distributed observations $\vec{Y}_t, t \in \{1,\dots,T\}$ from a VAR(p) process with mean $\vec{\mu}$ and nonsingular error covariance matrix Σ_ϵ, the log-likelihood function conditional on p observations $\vec{Y}_{1-p},\dots,\vec{Y}_0$ preceding the actual observed values can be written as

$$l = -\frac{T}{2}log|\Sigma_\epsilon| - \frac{1}{2}\sum_{t=1}^{T}\vec{\epsilon}'_t\Sigma_\epsilon^{-1}\vec{\epsilon}_t \tag{3.4.38}$$

For fixed autoregressive parameters, the log-likelihood is minimized by $\hat{\Sigma}_\epsilon = \frac{1}{T}\sum_{t=1}^{T}\hat{\vec{e}}_t\hat{\vec{e}}'_t$. Taking partial derivatives with respect to $\vec{\alpha} = vec\,[(\mathbf{A}'_1,\dots,\mathbf{A}'_p)']$ with $\mathbf{A}$ from (3.4.30) eventually leads to the likelihood equation

$$\frac{\partial l}{\partial \vec{\alpha}} = \left(\mathbf{I}_{dp}\otimes\Sigma_\epsilon^{-1}\right)\Xi'\left(\vec{y}-\Xi\vec{\alpha}\right) = \mathbf{0} \tag{3.4.39}$$

With L denoting the *lag-operator*, i.e., $L^k\vec{Y}_t = \vec{Y}_{t-k}$, the components of the likelihood equation are given by

$$\mathbf{I}_{dp} = \text{Identity matrix in } \mathbb{R}^{dp\times dp} \tag{3.4.40}$$

$$\vec{y} = (\vec{Y}'_1,\cdots,\vec{Y}'_T)' \in \mathbb{R}^{Td\times 1} \tag{3.4.41}$$

$$\mathbf{Y} = \left(\vec{Y}_1,\cdots,\vec{Y}_T\right)' \in \mathbb{R}^{T\times d} \tag{3.4.42}$$

$$\mathbf{X} = (L\mathbf{Y},\dots,L^p\mathbf{Y}) \in \mathbb{R}^{T\times dp} \tag{3.4.43}$$

$$\Xi = \mathbf{X}\otimes\mathbf{I}_d \in \mathbb{R}^{Td\times d^2} \tag{3.4.44}$$

$$= [(L\mathbf{Y}\otimes\mathbf{I}_d),\dots,(L^p\mathbf{Y}\otimes\mathbf{I}_d)] \tag{3.4.45}$$

$$\vec{\alpha} = vec\,[(\mathbf{A}'_1,\cdots,\mathbf{A}'_p)'] \in \mathbb{R}^{1\times p^3} \tag{3.4.46}$$

The solution is found to be

$$\hat{\vec{\alpha}} = \left(\Xi'\Xi\right)^{-1}\Xi'\vec{y} \tag{3.4.47}$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\otimes\mathbf{I}_d\right)\vec{y} \tag{3.4.48}$$

or

$$
\begin{aligned}
\hat{\mathbf{A}} &= \left(\hat{\mathbf{A}}_1', \cdots, \hat{\mathbf{A}}_p'\right)' && (3.4.49)\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} && (3.4.50)
\end{aligned}
$$

which is the same result as that obtained in (3.4.33) in the realm of least squares estimation.

Exact ML Estimation

As described for example in Reinsel (1997, Sec. 5.3), the exact likelihood of a VAR(p) process can be written under normality as

$$
\begin{aligned}
L &= |\Gamma(0)|^{-1/2}|\Sigma_\epsilon|^{-(T-p)/2}\\
&\times \exp\left[-\frac{1}{2}\left(\vec{y'}_p\Gamma(0)\vec{y}_p + \sum_{t=p+1}^{T}\vec{\epsilon'}\Sigma_\epsilon^{-1}\vec{\epsilon}\right)\right] && (3.4.51)
\end{aligned}
$$

where $\vec{y}_p = \left(\vec{Y}_1', \cdots, \vec{Y}_p'\right)'$ and $\Gamma(0) = Cov(\vec{y}_p, \vec{y}_p)$. The likelihood may be maximized by nonlinear maximization algorithms.

In later stages of the analysis, smoothing techniques will be applied to estimated autoregressive parameters. The necessary calculations can be expressed as linear operations. For this reason, least squares estimators are preferred below, since their closed form allows for a simple representation of the results. As was just shown, there is a close connection between these estimators and the conditional maximum likelihood solution, provided the underlying random process can be assumed to be normal.

3.5 Diagnostic Checking

Once the VARMA(p,q) orders have been chosen and the model is fit, it is important to check its adequacy by means of some diagnostic checks. These are often based on the residual autocorrelations, which may exhibit some form of systematic behaviour if the fit is insufficient. A first visual check is performed by looking at plots of the univariate autocorrelations which should resemble those of a White Noise process.

A formal check on the correct univariate model is provided by the Ljung-Box-Pierce test (Ljung and Box 1978). It tests if the residuals can be modeled as White Noise. Special multivariate tests are also available. A version due to Hosking (1980) and Li and McLeod (1981) is described below.

3.5.1 Assessing the Residual Autocorrelation Function

The standard univariate *autocorrelation function* is a plot of the autocorrelations against their lag h. In case of a White Noise process, it can be shown (Quenouille 1949, Ali 1989, Box and Pierce 1970) that for $h > 0$, the estimated autocorrelations $\hat{\rho}(h)$ have expected value zero and upper bound $1/\sqrt{T}$ for their standard deviation at lower lags, where T denotes the sample size, i.e.,

$$Var(\hat{\rho}(h)) \leq \frac{1}{T} \tag{3.5.52}$$

A rule of thumb is that univariate sample autocorrelations should be smaller than about $2/\sqrt{T}$. The same holds true for the *partial autocorrelations* of the residuals of the fit, which should also resemble that of White Noise for a model to be acceptable, i.e.,

$$Var(\hat{\rho}_{\text{partial}}(h)) \leq \frac{1}{T} \tag{3.5.53}$$

For the cross-correlation function of two series $Y_t(i)$ and $Y_t(j)$ with $i, j \in \{1, \ldots, d\}$ one would also expect small values, again in the range

$$Var(\hat{\rho}_{i,j}(h)) \leq \frac{1}{T} \tag{3.5.54}$$

It has to be kept in mind that the cross-correlation function is not symmetric for fixed lag h in general, i.e., $\hat{\rho}_{i,j}(h) \neq \hat{\rho}_{i,j}(-h)$, so both of these estimates need to be checked.

3.5.2 Testing the Goodness of Fit

Instead of looking at single sample correlations one by one, one may consider taking into account several estimators at once. This was proposed by Box and Pierce (1970, p. 314). The test statistic to check if the univariate series of residuals $(e_t)_{t\in\mathbb{T}}$ can be modeled as White Noise is given by

$$Q = T\sum_{k=1}^{M}\hat{\rho}_e^2(k) \tag{3.5.55}$$

which is distributed as $\chi^2(M-p-q)$. Here, M is taken to be sufficiently large, and $\hat{\rho}_e(k)$ denotes the estimated lag k autocorrelation of the residuals. A slightly modified version is known as the Ljung-Box-Pierce statistic (Ljung and Box 1978) and is given by

$$\tilde{Q} = T(T+2)\sum_{h=1}^{M}(T-h)^{-1}\hat{\rho}_e^2(h) \tag{3.5.56}$$

The distribution of $\tilde{Q}$ is approximated more precisely by a χ^2-distribution with $(M-p-q)$ degrees of freedom than the distribution of Q.

A multivariate test statistic for

$$H_0 \;:\; \left(\vec{Y}_t\right)_{t\in T} \textit{ is VARMA(p,q)}$$

versus

$$H_1 \;:\; \left(\vec{Y}_t\right)_{t\in T} \textit{ is not VARMA(p,q)}$$

is also based on residuals and was developed by Hosking (1980), Hosking (1981), and Li

Data Set	p=2	p=3	p=4	p=5	p=6
Pat.1R	10	7	9	6	8
Pat.1L	28	22	16	10	6
Pat.2	19	19	14	14	6
Pat.3	14	6	9	7	5

Table 3.1: *Number of rejected univariate Ljung-Box-Pierce tests out of 103. Critical value was the 95 percent quantile of* $\chi^2(10)$*, i.e. it was chosen as* $M = p + 10$*.*

and McLeod (1981). It is defined as

$$Q_M = T^2 \sum_{h=1}^{M} \frac{1}{T-h} tr\left(\hat{\mathbf{R}}_e(h)\hat{\mathbf{R}}_e(0)^{-1}\hat{\mathbf{R}}_e(-h)\hat{\mathbf{R}}_e(0)^{-1}\right) \tag{3.5.57}$$

which is approximately $\chi^2(d^2(M-p-q))$-distributed under the null hypothesis, and provided M is large enough. Here $\hat{\mathbf{R}}_e(h)$ denotes the estimated residual cross-correlation matrix at lag h.

3.6 Application: Estimation of AR-Coefficients

Any ERG data set can be regarded in a first approximation as a multivariate time series with fixed cross-correlations, possibly induced by spatial proximity. However, in practice the covariance structure is not known in advance. In fact, spatial heterogeneity in covariance would not be completely surprising, since the retinal area under study is expected to consist of both functioning and AMD-affected regions at unknown locations, and is thus spatially inhomogeneous. For this reason, the approach taken in a first step is to fit time series models only locally for each of the 103 series. In a second step, the resulting estimates will be smoothed spatially.

An autoregressive model of order 3 was fit locally to the data sets at hand. This choice was based on multiple Ljung-Box-Pierce tests at 95 percent significance level applied to residuals of models of different order. Table 3.1 shows for how many univariate time series the test rejected when an AR-model of order $p = 2, \dots, 6$ was fitted. For data sets Pat.1R and Pat.3, the choice $p = 3$ lead to rejection in only 7 and 6 cases, respectively, which is little more than 5 percent of the total number of 103 series, justifying this choice of p. However, for Pat.1L and Pat.2, the test was rejected in 22 and 19 cases. Rejection rates around 5 percent are found at an order as high as $p = 6$. However, $p = 3$ was chosen here as in the other cases, both to reduce numerical cost in subsequent crossvalidation steps, and since univariate models of higher order did not qualitatively improve the fit. Residual plots were compared for autoregressive orders up to $p = 12$, but no significant improvement of the overall fit was visible.

As an example for the obtained estimates, Figure 3.1 displays the estimated AR-parameters for data set Pat.1R in the appropriate spatial layout. Note that results for the other data sets are displayed in Appendix E. The spatially ordered set of first AR-parameters is referred to as first *AR-parameter field*, the set of second AR-parameters as second AR-parameter field, and so on. Values are linearly interpolated for display only. Any calculations were done using the original grid. A considerable amount of spatial variation makes a spatial structure somewhat difficult to detect.

The characteristic roots associated with the estimates for Pat.1R are all greater than one in absolute value, and thus correspond to stationary univariate processes. However, the lower limit of the absolute roots is 1.04, which is quite close to the critical boundary. Similar results hold true for the other three data sets.

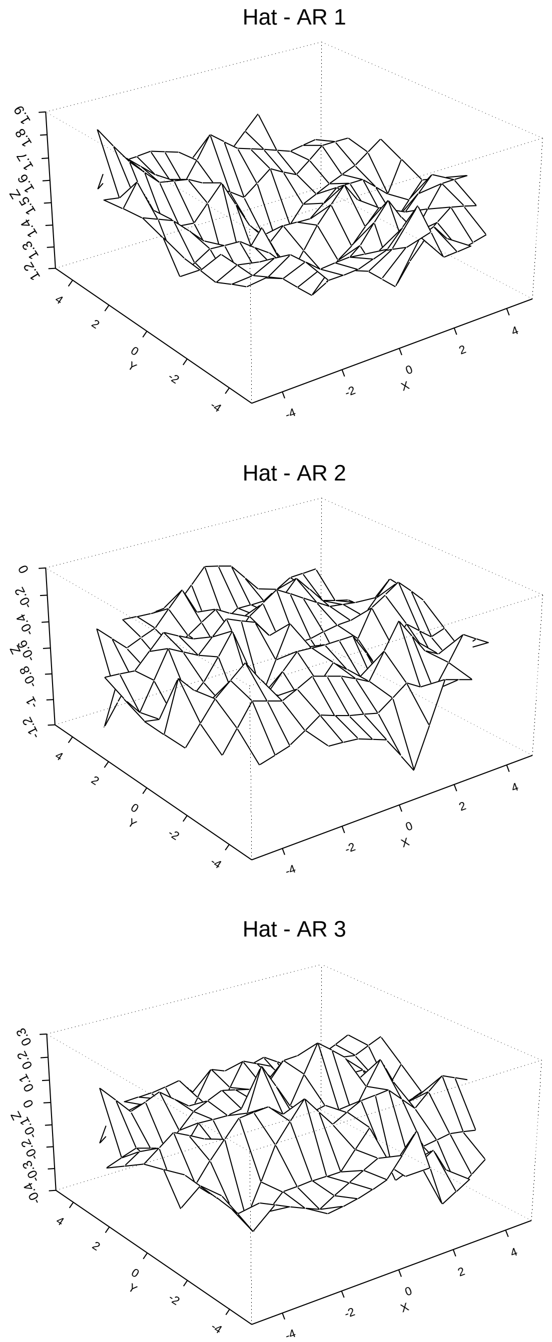

Figure 3.1: *OLS estimates for three AR parameter fields estimated from data set Pat.1R.*

Chapter 4

Spatial Data Analysis

In this chapter models and techniques are reviewed for the analysis of purely spatial data, i.e. data without any temporal component. The major task is to extract trend and variance estimates from data obtained at a finite set of spatial locations, and to allow for optimal estimation and prediction in space. MF-ERG amplitudes serve as an example.

4.1 Basic considerations

4.1.1 Typology of Spatial Statistics

Methods in spatial statistics were specifically developed to adequately take into account the location where data are observed. Spatial position can enter into the analysis in different forms, leading to three major categories of spatial statistics (Cressie 1993, Sec. 1.2):

- **Geostatistics** is widely used in mining and geology. Measurement locations are allowed to vary continuously in space. In principle, they can be chosen by the person conducting the experiment. This distinguishes a geostatistical analysis from other categories described below, although the difference is not always clear-cut. In terms of statistical modeling, the pronounced feature of geostatistical data is that there is variability both on a large scale (referred to as trend) and a small scale. The latter is incorporated into the variance-covariance structure of the process and modeled by a random field approach. The inclusion of small scale variation distinguishes geostatistical models from other trend-surface models which frequently assume that errors are independently distributed.

- **Analysis of lattice data** applies to data that can be assigned to a grid of spatial points in $\mathbb{R}^m$, often representing different administrative regions. There is a neighboring structure between regions which may be defined by geographical proximity. Typical examples of lattice data are found in epidemiology. Disease rates from epidemiological studies are collected within administrative regions. The neighboring structure is induced here by shared boarders, or by distance of regional centers. Other lattice structures are possible as well, like chessboard-type regular lattices as in agricultural field trials. The hexagonal grid used in the ERG experiments is also an admissible regular lattice structure.

- **Point pattern analysis** focuses on the locations where events occur. It addresses questions like spatial randomness or searches for other patterns in the distribution of locations. Point pattern analysis is not dealt with in this doctoral thesis and mentioned here only for the sake of completeness.

4.1.2 Modeling Aspects for Multifocal ERG Data

Data from the multifocal ERG can be directly linked to spatial statistics. On one hand, the experimental setup allows to identify measurements with center points of hexagonals. Figure 2.3 on page 14 shows the layout of hexagons chosen in the experiments. A univariate time series can be ascribed to each of the 103 areas. A set of statistics deduced from these time series (like amplitudes) may then be analyzed as a spatial data set. The hexagonal design allows to define a simple neighborhood structure very easily by determining areas to the left, to the right, to the upper left etc. from any given hexagonal region, or from its center point, respectively. In this respect, the spatial layout coincides with the *analysis of lattice data*-situation.

On the other hand, a *geostatistical approach* appears to be meaningful as well. The data obtained reflect the sum of a large number of small potentials evoked at the receptor cells of the retina. In a healthy eye, these cells are relatively evenly distributed over the whole retina, the major exception being the *blind spot* where the optic nerve emerges from the eye globe and visual stimulation is physically impossible. Since the hexagonal grid shown in Figure 2.3 is introduced only artificially by the experimental setup, the resulting measurements may be interpreted as a snapshot of a spatially continuous process.

There is some interest in predicting electric potentials at *every* location on the retina, not just at the hexagonal center points. Good predictions should help to identify regions with suboptimal bioelectrical functionality, an important aspect in medical diagnostics. A widely used method for prediction in spatial statistics is Kriging. Besides minimizing the mean squared prediction error, Kriging has the advantage of yielding estimates of variability for the resulting spatial predictions. Such estimates are usually not obtained by other approaches, such as smoothing splines which will be discussed in later chapters. However, under certain circumstances spline estimates and Kriging estimates coincide.

4.2 Deterministic Trend and Random Variation

The main goal of spatial data analysis is to obtain an estimate of the underlying structure of the observed process, like spatial trend and covariance. The following model can serve as a starting point.

Definition 4.2.1. Denote a *spatial stochastic process*, or *random field*, by $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$, where $\mathbb{D} \subset \mathbb{R}^m$ is a *spatial domain* of dimension m, and $\vec{s}$ is the *location*. A general model for the components $Z(\vec{s}) \in \mathbb{R}$ of this process is of the form

$$Z(\vec{s}) = \mu(\vec{s}) + \eta(\vec{s}) \qquad \vec{s} \in \mathbb{D}$$

Here, $\mu(\vec{s}) \in \mathbb{R}$ denotes the deterministic large scale variation or *trend* at location $\vec{s}$, and $\eta(\vec{s}) \in \mathbb{R}$ encompasses the stochastic small scale variation, or *random noise*.
♦

Since characterization of the major features within a multifocal ERG data set is of primary interest, particular focus will be on the trend component. A first step in isolating it may be to fit a low-order polynomial over $\mathbb{D}$ using standard techniques from linear regression. Several difficulties arise: This approach does not take into account any stochastic spatial dependency between locations unless this is modeled explicitly through the covariances involved. In addition, a polynomial fit is susceptible for extreme values and outliers. Finally, it is unclear in general which part of the data should be attributed to trend, and which part is noise, unless the covariance structure of the data is known in advance. The latter is usually not the case.

The approach taken by Cressie (1993) is to explicitly *define* what is trend and what should be attributed to random variation. This approach indeed leads to sensible predictions of $Z(\vec{s})$. The definition of trend really depends on the application at hand, and on the assumptions the scientist is willing to make. A polynomial trend may be meaningful in some situations. A robust alternative is given by *median polishing* (Tukey 1977; Emmerson and Hoaglin 1983). Once the trend model has been chosen and the trend is removed, the spatial covariance structure can be assessed by means of what is called *variogram estimation*. Appropriate combination of the resulting estimates then leads to optimal linear spatial prediction, referred to as *Kriging*.

4.2.1 A Decomposition of Variation

The major goal in the analysis of spatial data is to find a suitable decomposition of the observed values into a deterministic trend, or *large-scale variation*, and the stochastic *small-scale variation*. The standard linear model approach to spatial data analysis coincides with the assumption of a *White Noise* process for small scale variation. However, experience has shown that observations taken at sites nearby tend to exhibit higher correlation than those taken further apart, calling for the special techniques of spatial statistics to be applied.

It is common with spatial data that only single observations are available at each location. This makes direct estimation of measurement error difficult, if not impossible. In effect, the small-scale variation due to spatial proximity and the measurement error often are modeled jointly. This can be expressed as

$$Z(\vec{s}) = \mu(\vec{s}) + \underbrace{\xi(\vec{s}) + \epsilon(\vec{s})}_{\eta(\vec{s})} \tag{4.2.1}$$

with components

- $Z(\vec{s})$ denoting the measurement at spatial location $\vec{s} \in \mathbb{D} \subseteq \mathbb{R}^m$
- $\mu(\vec{s})$ representing deterministic trend
- $\eta(\vec{s})$ representing small scale variation
- $\xi(\vec{s})$ being a zero-mean intrinsically stationary process with variogram range larger than $min(\|\vec{s}_i - \vec{s}_j\|)$, $\vec{s}_i, \vec{s}_j \in \mathbb{D}$. (For a definition of intrinsic stationarity and the variogram see below.)
- $\epsilon(\vec{s})$ denoting a White Noise measurement error with mean zero and variance σ^2_{ME}, say.

If the underlying n basis functions of the trend are known up to some coefficient vector $\vec{\beta} = (\beta_1, \ldots, \beta_n)'$, the optimal prediction method is called *Universal Kriging*. If in addition the small-scale variation is White Noise, the parameter $\vec{\beta}$ for the mean structure can be estimated by ordinary least squares techniques.

However, an a priori assumption of a White Noise error process $\eta(\vec{s})$ would be a very restrictive one, because it excludes stochastic spatial dependence. Such an assumption may be sensible if it is the deterministic structure of the underlying process which is of major interest. Most of the variability in the data would be incorporated into the trend in this case. Obviously, the trend model then should be of considerable flexibility to closely fit the data. Otherwise, spatial dependencies will remain part of the residuals, giving a misleading impression of the trend. In summary, there seems to be no optimal model choice in this setting. Instead, Cressie (1993, p. 115) concludes:

> The criterion for choosing one model over another is at present a mixture of scientific context, familiarity, and intuition.

4.2.2 Median Polishing

Median Polishing (Tukey 1977; Emmerson and Hoaglin 1983) is favored by Cressie (1986) as a way to robustly remove spatial trend from data in a first analysis step, allowing for estimation of the spatial covariance structure in step two. In fact, Cressie preferably *defines* large scale variation by the result of median polishing (Cressie 1993, p.48). Median Polishing requires the data to be available on a regular lattice. It consists of a sequence of filtering operations, sweeping out row and column medians from the data. This results in a robust estimate of the underlying spatial trend. The remaining residuals (i.e., data minus estimated trend) are then used to model the covariance structure. Trend estimates are readily available at grid locations. Linear interpolation between grid points is done at locations where no observations were made, allowing for predictions over the whole spatial domain $\mathbb{D}$.

Definition 4.2.2 (Median Polishing). *Median Polishing (MP)* is performed by applying the following algorithm in a two-dimensional spatial domain.

0. Transform the data onto a regular coordinate grid, if necessary.
1. Define a two-dimensional array M of zeros representing the regular grid structure obtained in step 0. Use this to keep track of the trend removed in subsequent steps.
2. Take each row of the data and subtract the row median to obtain new data. Add the median to all elements of the corresponding row of M.
3. Take each column of the new data and subtract the column median. This again results in new data. Add the medians to the elements of the corresponding columns of M.
4. Repeat steps 2 and 3 until convergence of the medians to zero.

The difference between the original data values and the final values in M are the *median polishing residuals*. ⧫

The MP residuals are used in subsequent steps to estimate the spatial covariance structure. Practical experience shows that the algorithm usually converges after about 3 iterations (cf. Cressie 1993).

As a small example, the result of median polishing of the 103 amplitudes of the multifocal ERG data set Pat.1R is displayed in Figure 4.1. Amplitudes were chosen here because they are most commonly examined in the medical literature. Before polishing could be done, the original data were linearly interpolated onto a regular grid. Note that Median Polishing adapts relatively well to the quadratic polynomial trend in this particular data set, although it is a simple linear additive decomposition. However, data sets with more involved trend structure are not as well modeled by MP. See, for example, the results for Pat.3 in Figure 4.2. A disadvantage of MP is that a considerable number of estimated row and column median parameters is needed to achieve a certain amount of flexibility.

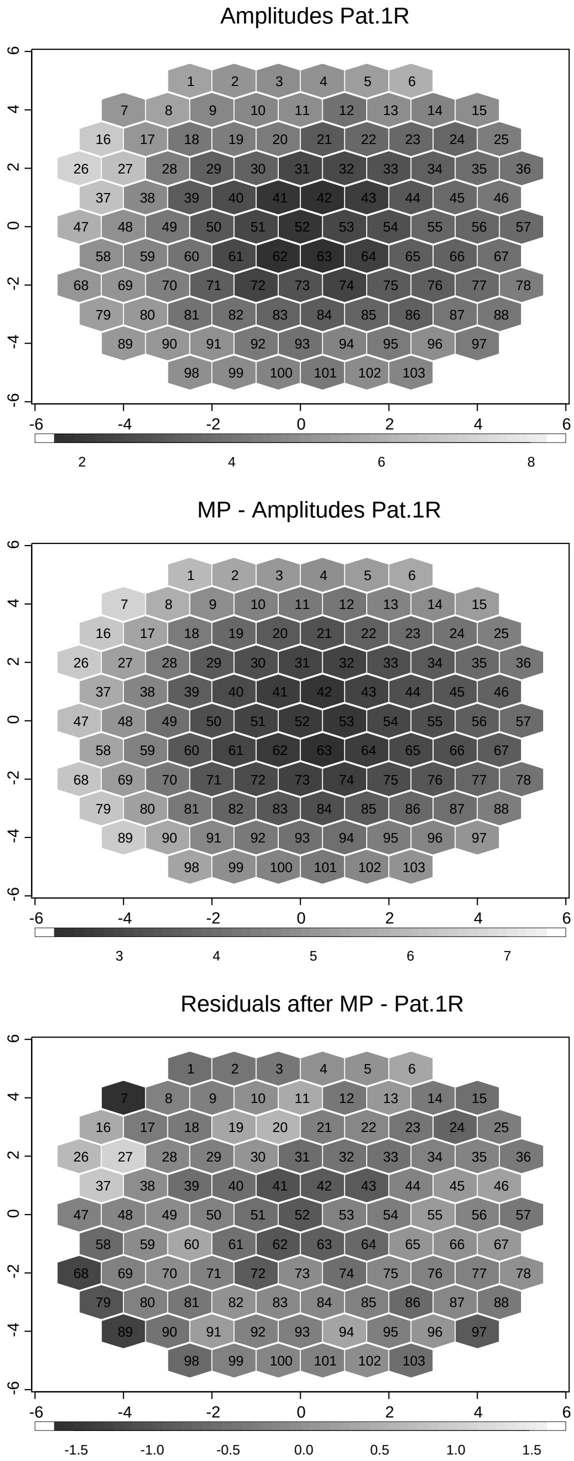

Figure 4.1: *Original amplitudes (top), estimated signal (middle), and noise (bottom) after Median Polishing. Patient is Pat.1R. Note the different scales.*

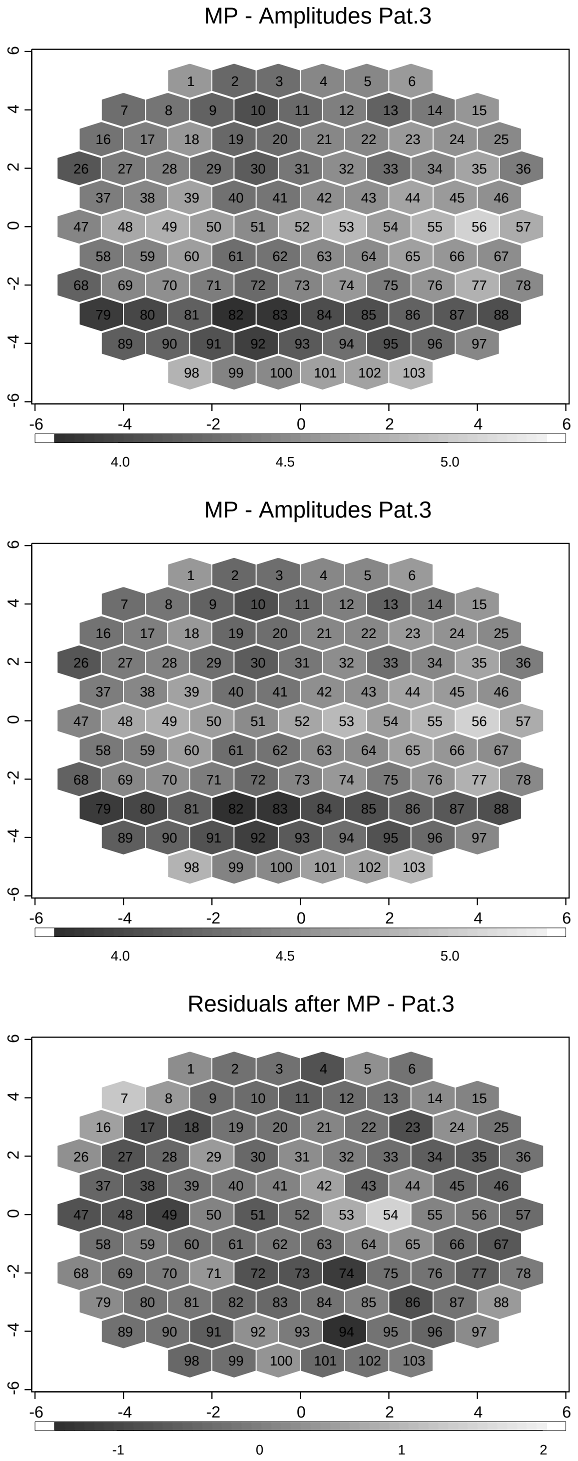

Figure 4.2: *Original amplitudes (top), estimated signal (middle), and noise (bottom) after Median Polishing. Patient is Pat.3. Note the different scales.*

4.3 Modeling Spatial Dependency

The model for dependencies between errors plays an important part in the modeling of spatial data, since it has direct impact on parameter estimation and resulting predictions. The main modeling approaches for spatial dependencies are either driven by intuition, prior knowledge, or suggested by the data.

4.3.1 The Variogram

It is common in geostatistics to define the small scale variation of a spatial process in terms of *differences* between values at locations which lie a certain distance apart. This is more general than using the autocovariances of $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ directly. A function that characterizes the small scale variation is the *variogram*.

Definition 4.3.1. Given locations $\vec{s}_i, \vec{s}_j \in \mathbb{D}$, the *variogram* $2\gamma(.,.)$ of a spatial process $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ is defined as

$$2\gamma(\vec{s}_i, \vec{s}_j) = Var(Z(\vec{s}_i) - Z(\vec{s}_j))$$

provided this variance exists. The quantity $\gamma(.,.)$ is called *semivariogram*. If the variogram does only depend on the vector $\vec{h} = \vec{s}_i - \vec{s}_j$ one may write

$$2\gamma(\vec{s}_i, \vec{s}_j) = 2\gamma(\vec{h})$$

and 2γ is said to be *stationary*. If furthermore, the direction of $\vec{h}$ does not provide any additional information, the variogram is called *isotropic*, and

$$2\gamma(\vec{h}) = 2\gamma(||\vec{h}||) =: 2\gamma(h)$$

where $||.||$ denotes the euclidean norm on $\mathbb{D}$, and $h = ||\vec{h}||$.
♦

The variogram plays a central role in spatial statistics. To be valid, it must be conditionally negative-definite, i.e., it must fulfill the condition

$$\sum_{i=1}^{d}\sum_{j=1}^{d} a_i a_j \gamma(\vec{s}_i, \vec{s}_j) \leq 0 \qquad (4.3.2)$$

for any finite number of locations $\vec{s}_i, \vec{s}_j \in \mathbb{D}$ with $i, j \in \{1, \ldots, d\}$, and real coefficients a_i with $\sum_{i=1}^{d} a_i = 0$ (Matheron 1963). Only valid variograms are considered here.

There is a connection between the cross-covariance used in time series analysis, and the *covariogram* in spatial statistics:

Definition 4.3.2 (Covariogram, Correlogram and Isotropy). The *covariogram* $C(.,.)$ of a spatial process $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ is defined as

$$C(\vec{s}_i, \vec{s}_j) = Cov(Z(\vec{s}_i), Z(\vec{s}_j))$$

for $\vec{s}_i, \vec{s}_j \in \mathbb{D}$. If $C(\vec{s}_i, \vec{s}_j)$ depends only on the difference of $\vec{s}_i$ and $\vec{s}_j$, one may write $C(\vec{s}_i, \vec{s}_j) = C(\vec{s}_i - \vec{s}_j)$. If in addition $C(\vec{s}_i - \vec{s}_j) = C(h)$ where $h = ||\vec{s}_i - \vec{s}_j||$, the covariogram is called *isotropic*. Scaling the covariogram of an isotropic process by $C(0) > 0$ results in the *correlogram*

$$\rho(h) = \frac{C(h)}{C(0)}$$

♦

The assumption that $C(\vec{s}_i, \vec{s}_j) = C(\vec{s}_i - \vec{s}_j)$ is often made in spatial statistics to allow for estimation of the covariogram and will be adopted here as well.

Stationarity of a spatial process can now be described in terms of the covariogram.

Definition 4.3.3 (Spatial Stationarity). If for the random process $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ with $E(Z(\vec{s})) = \mu$ the condition

$$Cov(Z(\vec{s}_i), Z(\vec{s}_j)) \quad = \quad C(\vec{s}_i - \vec{s}_j) \quad \text{for all } \vec{s}_i, \vec{s}_j \in \mathbb{D}$$

holds true, then the process is called *second order* (or weak) *spatially stationary*.
♦

If the process $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ is second order stationary and isotropic, the relation

$$2\gamma(\vec{h}) = 2C(\vec{0}) - 2C(\vec{h})$$

can be derived. This immediately follows from

$$var(Z(\vec{s}_i) - Z(\vec{s}_j)) = var(Z(\vec{s}_i)) + var(Z(\vec{s}_j)) - 2Cov(Z(\vec{s}_i), Z(\vec{s}_j))$$

However, the variogram exists for an even wider class of processes (possibly with non-existing variance), which has the property of what is called *Intrinsic Stationarity*. This property simplifies estimation considerably, since it implies a certain homogeneity of the process under study. It is defined as follows.

Definition 4.3.4 (Intrinsic Stationarity). A spatial random process $\{Z(\vec{s}) : \vec{s} \in \mathbb{D}\}$ is said to be *intrinsically stationary*, if it fulfills the conditions

$$\begin{aligned} E\left(Z(\vec{s}+\vec{h}) - Z(\vec{s})\right) &= 0 \\ Var\left(Z(\vec{s}+\vec{h}) - Z(\vec{s})\right) &= 2\gamma(\vec{h}) \end{aligned}$$

for all $\vec{s} \in \mathbb{D}$ and $\vec{h} \in \mathbb{R}^m$.
♦

To describe the properties of the variogram, it is helpful to define a few additional parameters, which characterize its general shape.

Definition 4.3.5 (Nugget effect, sill and range). If for an isotropic spatial process

$$\gamma(h) \longrightarrow c_0 \neq 0 \text{ for } h \longrightarrow 0$$

then c_0 is called the *nugget effect*. The *sill* σ^2 is defined as

$$\sigma^2 = lim_{h\to\infty}\gamma(h)$$

if the limit exists. The smallest distance r at which the semivariogram reaches the sill is called the *range* r_0, i.e.,

$$r_0 = min\{r \in \mathbb{R} : \gamma(r(1+\epsilon)) = 2C(0) \text{ for all } \epsilon > 0\}$$

♦

The classical estimator for the variogram was proposed by Matheron (1962). It is of the form

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(\vec{s}_i) - Z(\vec{s}_j))^2 \tag{4.3.3}$$

where $N(h) = \{(\vec{s}_i, \vec{s}_j) : \|\vec{s}_i - \vec{s}_j\| = h\}$ describes the set of locations within a certain distance from each other, and $|N(h)|$ denotes its cardinal number. A more robust version was proposed by Cressie and Hawkins (1980) as

$$2\bar{\gamma}(h) = \frac{\{\frac{1}{|N(h)|} \sum_{N(h)} \|Z(\vec{s}_i) - Z(\vec{s}_j)\|^{\frac{1}{2}}\}^4}{0.457 + \frac{0.494}{|N(h)|}} \tag{4.3.4}$$

The properties of this estimator are discussed in Hawkins and Cressie (1984). The idea behind it is to relate the scale estimation problem in (4.3.3) to that of a location estimate of a suitably scaled χ^2-distribution.

In practice, the two variogram estimators just considered can only be estimated at a finite set of distances. However, for modeling of continuous spatial processes, one needs continuous variogram models. Care has to be taken to guarantee that an estimated variogram is valid in the sense that it results in a nonnegative definite covariance matrix. The following models are admissible candidates.

Isotropic Variogram Models

In the independent error case, errors at nearby locations are uncorrelated and the covariance matrix of a set of observations from $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ has the form $\Sigma_Z = \sigma_Z^2 I$, if it exists. In terms of the variogram, stochastic independency is formulated as

$$\gamma(\vec{s}_i, \vec{s}_j; \theta) = \begin{cases} \theta & \text{, if } \vec{s}_i = \vec{s}_j \\ 0 & \text{, if } \vec{s}_i \neq \vec{s}_j \end{cases} \tag{4.3.5}$$

where $\theta = c_0 \geq 0$ is the variance of the measurement process.

If nearby observations are not independent, various other models are valid. See Journel and Huijbregts (1978, pp.161-195) or Cressie (1993, Sec. 2.5) for details. Basic models for isotropy are the linear, spherical and exponential variogram. These models mainly differ in the way $C(\|\vec{h}\|)$ changes with $\|\vec{h}\| \to \infty$. With $h = \|\vec{h}\|$ as before, the *linear variogram* is given by

$$\gamma(h;\theta) = \begin{cases} 0 & , h = 0 \\ c + bh & , h \neq 0 \end{cases} \tag{4.3.6}$$

with parameter vector $\theta = (c, b)'$ and $c > 0, b > 0$. It takes its minimal value c at $h = 0$ and increases linearly to infinity for $h > 0$. The linear variogram is valid for all $m \geq 1$. The *spherical variogram* is defined as

$$\gamma(h;\theta) = \begin{cases} 0 & h = 0 \\ c + c_{sph}\left(\frac{3}{2}\frac{h}{a_{sph}} - \frac{1}{2}\left(\frac{h}{a_{sph}}\right)^3\right) & 0 < h \leq a_{sph} \\ c + c_{sph} & h > a_{sph} \end{cases} \tag{4.3.7}$$

This variogram is valid in $\mathbb{R}^m$ for $m = 1, 2, 3$. The parameter θ contains the elements $(c, c_{sph}, a_{sph})'$, all being greater or equal to 0. The *exponential variogram* increases exponentially, and is given by

$$\gamma(h;\theta) = \begin{cases} 0 & h = 0 \\ c + c_{exp}\left(1 - exp\{1 - \frac{h}{a_{exp}}\}\right) & h \neq 0 \end{cases} \tag{4.3.8}$$

For the ERG data, the spherical variogram proved to be the most appropriate type of variogram.

Anisotropic Variogram Models

The class of variograms can be extended by allowing for *geometric anisotropy*, defined by the relation

$$2\gamma(h) = 2\gamma_0(\|\mathbf{G}\vec{h}\|) \tag{4.3.9}$$

where $\vec{h} \in \mathbb{R}^m$, and $\mathbf{G} \in \mathbb{R}^{m \times m}$ describes a linear matrix transformation. Geometrical anisotropy is checked in practice by estimation of empirical variograms in different directions. If $\mathbf{G}$ is invertible, anisotropy can be corrected for by appropriate scaling of the data locations by $\mathbf{G}^{-1}$. The transformed process may then be analyzed using an isotropic variogram as defined above.

4.3.2 Nearest Neighbors

Practical experience shows that measurements at nearby locations generally tend to be similar.

A possible estimation approach for the variogram is to assume that the covariance between random variables observed at two locations depends solely on their distance, being essentially zero beyond a certain point. The spherical variogram model, for example, reflects this in case of a continuous spatial domain. The critical distance beyond which the covariance is zero is the range of the variogram.

If grid locations are fixed and regularly spaced, determining the range of a variogram is essentially equivalent to define a set of *nearest neighbors* which have an impact on the observation at location of interest $\vec{s}_0$, say. A formal theoretic treatment of nearest neighbors has been done in the context of the analysis of lattice data using Markov random fields (Besag 1974). Cressie (1993, Chapter 6) gives a detailed account of this approach. Although applicable in principle, the nearest neighbor approach seems not to be sensible for the MF-ERG data. Residual plots indicate that spatial dependencies are too heterogeneous over the domain of interest to allow for nearest neighbor methods to be sensibly applied.

4.4 Optimal Spatial Prediction

Kriging yields the optimal linear spatial predictor while taking spatial variation into account. The name of this method was proposed by Matheron (1963) after the South African mining engineer D.G. Krige. The origins of Kriging are described in Cressie (1990a). Different modifications of this technique are available. *Simple Kriging* assumes a zero-mean spatial process, while *Universal Kriging* allows for a spatial trend of prespecified structure, for example a polynomial of known degree. A combination of trend estimation and variogram estimation finally results in the predictor.

4.4.1 Simple Kriging

Recall the decomposition of the spatial process $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ into components

$$Z(\vec{s}) = \mu(\vec{s}) + \epsilon(\vec{s}) \qquad , \vec{s} \in \mathbb{D} \tag{4.4.10}$$

assuming that $\mu(\vec{s})$ is a noiseless trend component and $\epsilon(\vec{s})$ describes White Noise. The goal is to linearly predict $\mu(.)$ in an optimal way from the available data $\vec{Z} = (Z(\vec{s}_1), \cdots, Z(\vec{s}_d))'$. The criterion for goodness of fit chosen is the mean squared prediction error.

Definition 4.4.1. The *Mean Squared Prediction Error* (MSPE) of the predictor $p(\vec{Z}, \vec{s}_0)$ at location $\vec{s}_0 \in \mathbb{D}$ is defined as

$$MSPE = E[(Z(\vec{s}_0) - p(\vec{Z}; \vec{s}_0))^2 | \vec{Z}]$$

♦

The best linear predictor $p(\vec{Z}; \vec{s}_0)$ at location $\vec{s}_0$ is known to be the conditional expectation,

$$p(\vec{Z}; \vec{s}_0) = E(Z(\vec{s}_0)|\vec{Z}) \tag{4.4.11}$$

which can be expressed as a linear combination of the $Z(\vec{s}_i)$, $i = 1, \ldots, d$, plus some constant k,

$$p(\vec{Z}; \vec{s}_0) = \sum_{i=1}^{d} a_i Z(\vec{s}_i) + k \tag{4.4.12}$$

Assuming the trend $\mu(\vec{s}_i) = E(Z(\vec{s}_i))$ is *known*, the optimal weights are given by

$$\begin{aligned} k &= \mu(\vec{s}_0) - \sum_{i=1}^{d} a_i \mu(\vec{s}_i) & (4.4.13) \\ \vec{a} &= (a_1, \ldots, a_d)' \\ &= \Sigma_Z^{-1} \vec{c} & (4.4.14) \end{aligned}$$

where $\vec{c} = Cov(Z(\vec{s}_0), \vec{Z})$. This results in the optimal linear *simple Kriging predictor*

$$p_{SK}(\vec{Z}; \vec{s}_0) = \mu(\vec{s}_0) + \vec{c}' \Sigma_Z^{-1} (\vec{Z} - \vec{\mu}) \tag{4.4.15}$$

with $\vec{\mu} = (\mu(\vec{s}_1), \ldots, \mu(\vec{s}_d))'$ and MSPE

$$\sigma_{SK}^2(\vec{s}_0) \equiv Var(Z(\vec{s}_0)) - \vec{c}' \Sigma_{\vec{Z}}^{-1} \vec{c} \tag{4.4.16}$$

This type of prediction was called *Simple Kriging* (SK) by Matheron (1962), since the mean structure is taken to be known. It is extended by introducing trend estimates into the prediction process as follows.

4.4.2 Ordinary Kriging

Ordinary Kriging (OK) is a slightly generalized version of Kriging and has two assumptions at its base (Matheron 1971;Journel and Huijbregts 1978, pp.304). The model is given by

$$Z(\vec{s}) = \mu + \eta(\vec{s}) \tag{4.4.17}$$

where $\vec{s} \in \mathbb{D}$. The first assumption is that $\mu \in \mathbb{R}$ is *unknown*, but constant. Secondly, the predictor is restricted to be linear and unbiased for all μ, i.e. given the observations $\vec{Z} = (Z(\vec{s}_1), \ldots, Z(\vec{s}_d))'$, one requires

$$p(\vec{Z}; \vec{s}_0) = \sum_{i=1}^{d} \lambda_i Z(\vec{s}_i) \tag{4.4.18}$$

with

$$\sum_{i=1}^{d} \lambda_i = 1 \tag{4.4.19}$$

where the latter condition ensures unbiasedness. If the spatial covariance structure is determined by some valid variogram $2\gamma(\vec{h})$, the optimal weights for Ordinary Kriging under squared error loss can be found using the following result (Cressie 1993).

Theorem 4.4.1. *The* Ordinary Kriging Equations *provide the optimal Kriging weights for (4.4.18) and are given by*

$$\vec{\lambda}_{OK} = \Gamma_{OK}^{-1} \vec{\gamma}_{OK} \tag{4.4.20}$$

with components

$$\vec{\lambda}_{OK} = (\lambda_1, \ldots, \lambda_d, u)' \tag{4.4.21}$$

$$u = \text{a Lagrange multiplier ensuring unbiasedness} \tag{4.4.22}$$

$$\vec{\gamma}_{OK} = (\gamma(\vec{s}_0 - \vec{s}_1), \ldots, \gamma(\vec{s}_0 - \vec{s}_d), 1)' \tag{4.4.23}$$

$$\Gamma_{OK} = \begin{cases} \gamma(\vec{s}_i - \vec{s}_j) & for \quad i = 1, \ldots, d; j = 1, \ldots, d \\ 1 & for \quad i = d+1; j = 1, \ldots, d \\ 0 & for \quad i = d+1; j = d+1 \end{cases} \tag{4.4.24}$$

♦

For a derivation of this result see Cressie (1993, Sec. 3.2). The (minimized) MSPE resulting from these equations can be specified directly:

Lemma 4.4.2. *The(minimized) MSPE obtained from (4.4.20) is called the* Kriging variance *and is given by*

$$\begin{aligned} \sigma_{OK}^2(\vec{s}_0) &= \vec{\lambda}'_{OK} \vec{\gamma}_{OK} \\ &= \sum_{i=1}^{d} \lambda_i \gamma(\vec{s}_0 - \vec{s}_i) + u \end{aligned} \tag{4.4.25}$$

♦

From this, pointwise prediction intervals for the process $(Z(\vec{s}))_{\vec{s}\in\mathbb{D}}$ can be constructed. The variogram does not have to be stationary, although this simplifies estimation.

If measurement error is to be allowed for, and $Z(\vec{s}_0)$ is to be predicted, the equations to be solved need to be modified to become

$$\vec{\nu}_{OK} = \Gamma_{OK}^{-1}\gamma_{OK} \tag{4.4.26}$$

where now

$$\begin{aligned}
\vec{\nu}_{OK} &= (\nu_1, \ldots, \nu_d, u)' && (4.4.27)\\
u &= \text{the Lagrange multiplier ensuring unbiasedness} && (4.4.28)\\
\gamma_{OK}(\vec{s}_0 - \vec{s}_i) &= \begin{cases} \gamma(\vec{s}_0 - \vec{s}_i) & for \quad \vec{s}_0 \neq \vec{s}_i, i \in \{1, \ldots, d\} \\ \sigma^2_{ME} & for \quad \vec{s}_0 \in \{\vec{s}_1, \ldots, \vec{s}_d\} \end{cases} && (4.4.29)
\end{aligned}$$

where σ^2_{ME} is the variance for measurement error, and Γ_{OK} as in (4.4.24). The minimized MSPE is then given by

$$\tau^2_{OK}(\vec{s}_0) = \vec{\nu}'_{OK}\vec{\gamma}_{OK}(\vec{s}_0 - \vec{s}_i) + u - \sigma^2_{ME} \tag{4.4.30}$$

Note that by varying σ^2_{ME}, the ordinary Kriging predictor changes from an exact interpolator ($\sigma^2_{ME} = 0$) to a non-exact interpolator ($\sigma^2_{ME} \neq 0$), or *smoother*. This will be of importance when comparing Kriging estimates to smoothing splines.

4.4.3 Universal Kriging

The Kriging approach is generalized to *Universal Kriging* (UK) by allowing for a generally *non*-constant mean process $\mu(\vec{s})$, which is an unknown linear combination of n known basis functions f_j (Huijbregts and Matheron 1971). This can be formulated as

$$Z(\vec{s}) = \sum_{j=0}^{n} f_j(\vec{s})\theta_j + \eta(\vec{s}) \quad , \vec{s} \in \mathbb{D} \tag{4.4.31}$$

where the parameter vector $\vec{\theta} = (\theta_0, \ldots, \theta_n)' \in \mathbb{R}^{n+1}$ has to be estimated. The error process $(\eta(\vec{s}))_{s\in\mathbb{D}}$ has zero mean and variogram $2\gamma(.)$. This can be written in matrix notation as

$$\vec{Z} = \mathbf{X}\vec{\beta} + \vec{\eta} \tag{4.4.32}$$

where $\mathbf{X} \in \mathbb{R}^{d\times(n+1)}$ has elements $x_{i,j} = f_{j-1}(\vec{s}_i)$. In this case, a uniformly unbiased linear predictor has to fulfill the condition

$$p(\vec{Z}, \vec{s}_0) = \sum_{i=1}^{d} \lambda_i Z(\vec{s}_i) \quad \text{with} \quad \mathbf{X}'\vec{\lambda} = \vec{x} \tag{4.4.33}$$

where $\vec{x} = (f_0(\vec{s}_0), \ldots, f_n(\vec{s}_0))'$. See e.g. Searle (1971, p. 88) for details about unbiasedness conditions. Ordinary Kriging is a special case with $\mathbf{X} = 1_d$, the vector of d ones.

Theorem 4.4.3. *The* Universal-Kriging equations *for solving (4.4.31) under condition (4.4.33) take the form*

$$\vec{\lambda}_{UK} = \Gamma_{UK}^{-1}\vec{\gamma}_{UK} \tag{4.4.34}$$

where $\vec{\lambda}_{UK}$ and $\vec{u}$ are as in Theorem 4.4.1, and

$$\vec{\gamma}_{UK} = (\gamma(\vec{s}_0 - \vec{s}_1), \ldots, \gamma(\vec{s}_0 - \vec{s}_d), 1, f_1(\vec{s}_0), \ldots, f_n(\vec{s}_0))' \tag{4.4.35}$$

$$\Gamma_{UK} = \begin{cases} \gamma(\vec{s}_i - \vec{s}_j) & i = 1, \ldots, d; j = 1, \ldots, d \\ f_{j-1-d}(\vec{s}_i) & i = 1, \ldots, d; j = d+1, \ldots, d+n+1 \\ 0 & i = d+1, \ldots, d+n+1; \\ & j = d+1, \ldots, d+n+1 \end{cases} \tag{4.4.36}$$

and $f_0(\vec{s}) = 1$ for all $\vec{s}$.
♦

This result is derived in Cressie (1993, Sec. 3.4). The above form is referred to as the *variogram formulation*, since the equations are given in terms of $\gamma(.)$. For estimation of a noiseless version $S(\vec{s}_0)$ of $Z(\vec{s}_0)$, σ^2_{ME} is to be inserted into the diagonal of Γ_{UK}. The resulting *universal Kriging variance* is given by

$$\begin{aligned}\sigma^2_{UK}(\vec{s}_0) &= \vec{\lambda}'_{UK}\vec{\gamma}_{UK} \\ &= 2\sum_{i=1}^{d}\lambda_i\gamma(\vec{s}_0-\vec{s}_i) - \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j\gamma(\vec{s}_i-\vec{s}_j)\end{aligned} \tag{4.4.37}$$

If the process $(Z(\vec{s}))_{s\in\mathbb{D}}$ is second order stationary, the covariance does exist, and the process can be predicted by generalized least squares techniques given $Var(\vec{Z}) = \Sigma_Z$ is *known*. The estimator of the trend parameters then is

$$\hat{\theta}_{gls} = \left(\mathbf{X}'\Sigma_Z^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma_Z^{-1}\vec{Z} \tag{4.4.38}$$

The existence of the covariances also allows for a *covariance formulation* of the Kriging equations, which results in the solution

$$\vec{\lambda}_{UK} = \Sigma_{UK}^{-1}\vec{c}_{UK} \tag{4.4.39}$$

where now

$$\vec{\lambda}_{UK} = (\lambda_1,\ldots,\lambda_d,u_1,\ldots,u_{n+1})' \tag{4.4.40}$$

$$\vec{u} = \text{vector of (n+1) Lagrange multipliers for unbiasedness} \tag{4.4.41}$$

$$\vec{c}_{UK} = (C(\vec{s}_0,\vec{s}_1),\ldots,C(\vec{s}_0,\vec{s}_d),1,f_1(\vec{s}_0),\ldots,f_n(\vec{s}_0))' \tag{4.4.42}$$

$$\Sigma_{UK} = \begin{cases} C(\vec{s}_i,\vec{s}_j) & i=1,\ldots,d; j=1,\ldots,d \\ f_{j-1-d}(\vec{s}_i) & i=1,\ldots,d; j=d+1,\ldots,d+n+1 \\ 0 & i=d+1,\ldots,d+n+1; \\ & j=d+1,\ldots,d+n+1 \end{cases} \tag{4.4.43}$$

A different representation of the Universal Kriging equations is also possible. It will be of particular interest later when comparing Kriging to spline smoothing. It is obtained by noting that the prediction $\hat{Z}(\vec{s}_0)$ at location $\vec{s}_0 \in \mathbb{R}^m$ can be written as linear combination

$$\begin{aligned} \hat{Z}(\vec{s}_0) &= (Z(\vec{s}_1), \ldots, Z(\vec{s}_d), 0, \ldots, 0)\, \Sigma_{UK}^{-1} \vec{c}_{UK} && (4.4.44) \\ &= \vec{v'}_1 \vec{c} + \vec{v'}_2 \vec{x} && (4.4.45) \end{aligned}$$

The vector $\vec{x}$ in (4.4.45) is given by $\vec{x} = (f_0(\vec{s}_0), \ldots, f_n(\vec{s}_0))'$, and $\vec{v} = (\vec{v'}_1, \vec{v'}_2)'$ is the solution of the dual system of equations

$$\begin{pmatrix} \Sigma_{UK} & \mathbf{X}' \\ \mathbf{X} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vec{v}_1 \\ \vec{v}_2 \end{pmatrix} = \begin{pmatrix} \vec{Z} \\ \vec{0} \end{pmatrix} \tag{4.4.46}$$

The equations (4.4.45) and (4.4.46) are known as the *Dual Kriging equations*. They also have a variogram formulation, which is formally obtained by replacing covariances with the corresponding values of the variogram.

The most flexible Kriging method introduced here obviously is Universal Kriging. Note that its adaptivity in trend estimation crucially depends on the choice of basis functions. This can be regarded both as an advantage and disadvantage in comparison to Median Polishing, depending on how much is known about the trend structure a priori.

4.5 Application: Kriging of ERG-Amplitudes

When it comes to parameterization of an underlying trend, Universal Kriging differs from the generalized least squares approach essentially only in the way the covariance matrix of the process is determined. As noted by Cressie (1993), this can be regarded as an advantage, since an additional modeling step is involved.

To clarify the process of Kriging, and for comparison with Median Polishing and the spline smoothing approach introduced in later chapters, this section describes an application of Universal Kriging to the multifocal ERG data available for Pat.1R (Figure 2.9). Focus again is on amplitudes only, since they ignore temporal features and constitute a purely spatial data set. Results for the other patients are presented in Appendix D.

UK with a polynomial trend of order two was done using the software S+SpatialStats® (Mathsoft 2000b), an add-on module to S-Plus® (Mathsoft 2000a). For comparison with other data sets, the original amplitudes are shown in a perspective plot in Figure 4.3. A spherical variogram model was chosen for all four data sets with parameters estimated by an approximate weighted least squares approach (Cressie 1985). Figure 4.4 shows that a simpler linear variogram may have been possible in some cases as well, but was not used for consistency with the other data sets. The estimated UK trend surface with polynomial trend with degree up to order two is presented in Figure 4.5 in a perspective plot. Coefficients are presented in Table 4.1. There was no indication of anisotropy. Due to the polynomial fit, the estimated trend is very smooth. The residuals (data - trend) are displayed in Figure 4.6. The Universal Kriging predictions are shown in Figure 4.7. They fit the data very well. The local Kriging variance can be assessed by (4.4.37), allowing to evaluate the prediction quality. However, the deterministic trend is of major interest in the ERG case. It can be seen that the parametric fit of the trend by itself gives a somewhat simplistic view of the data. This is certainly due to the limited choice of basis functions, which in addition have to be

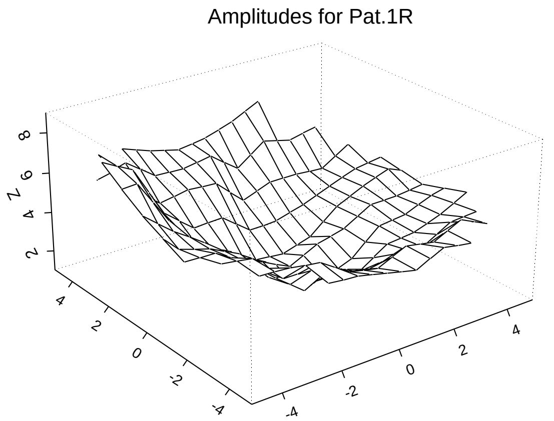

Figure 4.3: *Amplitudes for data set Pat.1R. Values are linearly interpolated for better graphical presentation.*

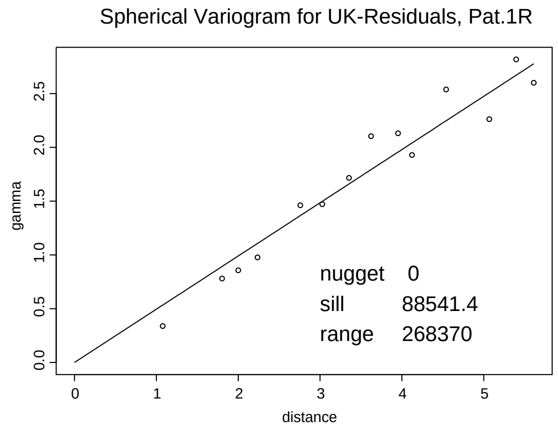

Figure 4.4: *Estimated spherical variogram from UK residuals to Pat.1R.*

chosen in advance. Since standard polynomials are locally not very adaptive, they may not be appropriate to closely describe the trend: Appendix D displays results on amplitudes of Pat.1L and Pat.3 which are not fitted well in the retinal center. An additional problem is that for application in clinical practice, variogram fitting needs to take place in an automated way. This allows for possible over-parameterization as in the case of Pat.1R above, where a spherical variogram was used when a linear variogram would suffice.

The conclusion is that Universal Kriging is only of limited use for trend estimation in MF-

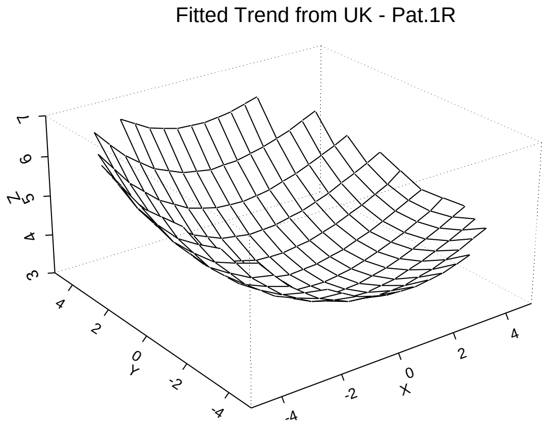

Figure 4.5: *Estimated trend from UK for amplitudes in data set Pat.1R.*

Figure 4.6: *Residuals after UK for amplitudes in data set Pat.1R.*

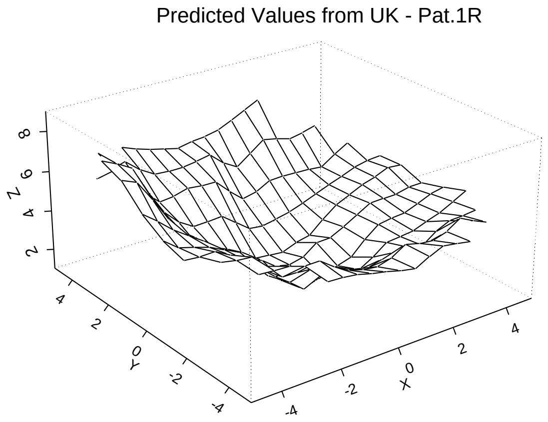

Figure 4.7: *Predictions obtained from UK for amplitudes in data set Pat.1R.*

	μ	x	y	x^2	y^2	x^3	y^3
Pat.1R	3.350	-0.195	0.331	1.953	1.795	-0.643	0.311
Pat.1L	4.771	-0.307	0.338	0.361	-0.183	-0.104	-0.214
Pat.2	3.791	0.552	0.306	1.336	1.232	-1.022	0.034
Pat.3	4.373	0.330	0.493	0.333	0.385	-0.345	-0.513

Table 4.1: *Scaled coefficient estimates from Universal Kriging.*

ERG data, since too little a priori information is available about the shape of the trend surface. In consequence, the chosen set of basis functions may easily be too restricted to encompass the full large scale variation. Therefore, a more flexible and locally adaptive procedure for trend estimation is sought for.

Chapter 5

Concepts of Spline Smoothing

In the analysis of multifocal ERG data, emphasis is on the determination of a parsimonious set of parameters which adequately reflect the functionality of the retina and allow for sensible interpolation. The true parameter values are likely to vary locally due to different degrees of functionality. However, there is no *a priori* knowledge available on how parameters vary locally within a particular eye, since the exact local retinal state is unknown before examination. Therefore, a locally adaptive fit seems to be a sensible approach.

Penalized smoothing splines (Green and Silverman 1994) are proposed here as a way to circumvent restrictive *a priori* structural assumptions on the shape of the parameter trend surface. The use of splines in statistics is by no means new, but rather dates back at least to Whittaker (1923). The basic model starts out with an underlying fixed function g, say, which is observed with White Noise measurement error ϵ, resulting in observations

$$Y(\vec{s}) = g(\vec{s}) + \epsilon(\vec{s}) \tag{5.0.1}$$

at $\vec{s} \in \mathbb{D}$. It is argued here that in the ERG problem splines indeed are an appropriate tool for trend estimation. It is not suggested that they are the only sensible tool. In fact, it can be shown that under certain conditions trend estimates obtained from penalized splines are even identical to those obtained from Kriging with nugget effect.

A classic mathematical reference on the construction of cubic splines is De Boor (1978). Several aspects from a statistical point of view can be found in Silverman (1985), Green and Silverman (1994) and Wahba (1990b), among others. The basic idea when using penalized splines is to relax the assumptions of classical linear regression towards a more adaptive, data driven method. While splines are certainly not the only possible alternative to classical linear regression, they have certain desirable properties which will be described in more detail below. For sake of clearer presentation, basic concepts of smoothing splines are first introduced in the univariate setting. Extensions to the bivariate case relevant to ERG data are then presented in Section 5.2 and later.

5.1 Some Fundamentals on Splines

5.1.1 Cubic Spline Basis Functions

Cubic splines are widely used in univariate nonparametric regression. The following definition is adapted here.

Definition 5.1.1. Let $[a, b]$ describe some interval on the real line with interior points $a < s_1 < \cdots < s_d < b$. The points $s_i \in \mathbb{R}$ are called *knots*. A function $g \in \mathbb{R}$ is called a *natural cubic spline*, if it satisfies the following conditions:

- g is piecewise polynomial of order 3 in each of the intervals $[a, s_1], [s_1, s_2], \ldots, [s_d, b]$
- g and its first and second derivatives g' and g'' are continuous at each point s_i
- $g''' = g'' = 0$ at a and b

The last condition is called the *natural boundary condition*. It ensures linearity beyond the interval $[a, b]$.
♦

Cubic splines therefore consist of local cubic polynomials which are connected in a smooth way in the sense that their first and second derivatives are continuous. It is sometimes convenient to present a cubic spline in the *value-second derivative representation* (Green and

Silverman 1994, p.12). If g is a cubic spline with knots $s_1 < \cdots < s_d$, define

$$\begin{aligned} g_i &= g(s_i) \quad \text{for } i = 1, \ldots, d \\ \vec{g} &= (g_1, \ldots, g_d)' \\ \vec{g''} &= (g''(s_2), \ldots, g''(s_d))' \end{aligned}$$

By definition, $g''(s_1) = g''(s_d) = 0$. Then $\vec{g}$ and $\vec{g''}$ specify g completely. For $\vec{g}$ and $\vec{g''}$ to define a cubic spline, certain necessary and sufficient conditions have to be fulfilled. Define

$$\begin{array}{lll} h_i & = s_{i+1} - s_i & \text{for } i = 1, \ldots, d-1 \\ q_{j-1,j} & = h_{j-1}^{-1} - h_j^{-1} & \text{for } j = 2, \ldots, d-1 \\ q_{j,j} & = -h_{j-1}^{-1} - h_j^{-1} & \text{for } j = 2, \ldots, d-1 \\ q_{j+1,j} & = h_j^{-1} & \text{for } j = 2, \ldots, d-1 \\ q_{ij} & = 0 & \text{for } |j-i| \geq 2 \end{array}$$

and take $\mathbf{Q} \in \mathbb{R}^{d \times (d-2)}$ to be the matrix with entries q_{ij}. In addition, define the symmetric, positive definite matrix $R \in \mathbb{R}^{(d-2)\times(d-2)}$ with elements

$$\begin{array}{rcll} r_{ii} & = & \frac{1}{3}(h_{i-1} + h_i) & \text{for } i = 2, \ldots d-1, \\ r_{i,i+1} & = & \frac{1}{6}h_i & \text{for } i = 2, \ldots d-2, \\ r_{i+1,i} & = & r_{i,i+1} & \text{for } i = 2, \ldots d-2, \\ r_{ij} & = & 0 & \text{for} |i-j| \geq 0. \end{array}$$

The condition for $\vec{g}$ and $\vec{g''}$ to define a natural cubic spline can then be formulated as (Green and Silverman 1994, Theorem 2.1)

$$\mathbf{Q}'\vec{g} = \mathbf{R}\vec{g''} \tag{5.1.2}$$

It is important to note that when defining

$$\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}' \in \mathbb{R}^{d \times d} \tag{5.1.3}$$

the integral over the second derivative of g can be represented in matrix form as

$$\int_a^b g''(t)^2 dt = (\vec{g''})'\mathbf{R}\vec{g''} = \vec{g}'\mathbf{K}\vec{g} \tag{5.1.4}$$

This integral can be interpreted as a *roughness* measure of the curve g, since smooth functions g will result in small values of $\vec{g}'\mathbf{K}\vec{g}$.

5.1.2 Optimal properties of Cubic Splines

It turns out that within the class of all differentiable functions with absolutely continuous first derivative on the interval $[a, b]$, the unique function interpolating data points $(s_i, Y_i) \quad (i = 1, \dots d)$ while simultaneously minimizing (5.1.4) is a cubic spline. See Green and Silverman (1994, Section 2.2) for a proof. In this sense, natural cubic splines are interpolators of functions in $\mathbb{R}$ with optimal smoothness.

If *smoothing* is desired instead of direct interpolation, a weighted sum of the roughness penalty (5.1.4) and a least squares penalty term may be considered. This is done by introducing a parameter λ controlling the influence of the penalty on the overall fit. A motivation for smoothing from the statistical point of view is that measurements are usually distorted by some random error and often do not represent the true value. Hence, the data should not necessarily be interpolated exactly. A penalized sum of squares allows to balance the two goals of close fit to the data versus smoothness of the estimated function.

Definition 5.1.2. Let $[a, b]$ be an interval on the real line with knots $s_i, i = 1, \dots, d$, where $a < s_1 < \cdots < s_d < b$. Denote the d observations at the knots by $Y(s_1), \dots, Y(s_d)$. Let $S_2[a, b]$ denote the space of all functions that are differentiable on $[a, b]$ with absolutely continuous first derivative. In addition, define the *Penalized Sum of Squares*

$$PSS_\lambda(g) = \sum_{i=1}^{d} \{g_i - Y(s_i)\}^2 + \lambda \int_a^b \{g''(x)\}^2 dx \tag{5.1.5}$$

with $\lambda > 0$. The curve estimate $\hat{g}$ is then defined as the minimizer of $PSS_\lambda(g)$ within $S_2[a, b]$. The parameter λ is called *smoothing parameter*.
♦

Extensive use of $PSS_\lambda(g)$ is made in the *Roughness Penalty Approach* described by Green and Silverman (1994). The authors show (p. 18) that the solution $\hat{g}$ to (5.1.5) must be a natural cubic spline for any fixed λ. The argument is as follows.

Assume g is any curve, not being a natural cubic spline, observed at knots s_i. Take $\bar{g}$ to be the natural cubic spline interpolant of g at the knots. Then $g_i = \bar{g}_i$ for all i, and hence their (unpenalized) sums of squares are equal. However,

$$\int_a^b \bar{g}''(x)dx < \int_a^b g''(x)dx \tag{5.1.6}$$

implying that $PSS(\bar{g}) < PSS(g)$ for any λ. In other words, for any function g which is not a cubic spline itself, there is always a natural cubic spline $\bar{g}$ with a truly smaller value of PSS. If two functions g and $\bar{g}$ take the same values at the knots, the penalized sum (5.1.5) will always chose the natural cubic spline as its minimizer. As a result, the estimate must be a cubic spline.

Following the above remarks, the search for a solution $\hat{g}$ to $PSS_\lambda(g)$ in the univariate case can be confined to the class of natural cubic splines. This makes the search much more feasible. To show that a solution for $PSS_\lambda(g)$ indeed exists and is unique, consider the following arguments. Let $\vec{g} = (g(s_1), \cdots, g(s_d))'$ denote the vector of evaluations of g at the knots, and $\vec{y} = (y_1, \cdots, y_d)'$ the vector of observed data. Then, from (5.1.4), the penalized sum can be written as

$$\begin{aligned} PSS_\lambda(g) &= (\vec{y} - \vec{g})' (\vec{y} - \vec{g}) + \lambda \vec{g}' \mathbf{K} \vec{g} \qquad (5.1.7) \\ &= \vec{g}' (I + \lambda \mathbf{K}) \vec{g} - 2\vec{y}' \vec{g} + \vec{y}' \vec{y} \qquad (5.1.8) \end{aligned}$$

with positive definite matrix $(I + \lambda \mathbf{K})$ (since $\lambda \mathbf{K}$ is nonnegative definite). The unique solution to the minimization problem is therefore given by

$$\hat{\vec{g}} = (I + \lambda \mathbf{K})^{-1} \vec{y} \tag{5.1.9}$$

Note that with this solution, $PSS_\lambda(g)$ may be rewritten as

$$PSS_\lambda(g) = (g - \hat{g})' \, (I + \lambda \mathbf{K}) \, (g - \hat{g}) + constant(\vec{y}) \tag{5.1.10}$$

where the constant depends on $\vec{y}$ only. An efficient algorithm to solve for $\vec{g}$ was originally proposed by Reinsch (1967). Green and Silverman (1994, Section 2.3.3) are a more recent reference. The main idea is to make use of certain band structures of the matrices involved.

5.1.3 Choice of the Smoothing Parameter λ

So far, the smoothing parameter λ was implicitly assumed to be fixed and known. In practice, it usually has to be estimated. It may be argued that subjectively fixing λ at a finite set of values and then minimizing the penalized sum (5.1.5) will allow for exploration of the data on different scales. However, for practical application an automatic selection procedure is called for. Craven and Wahba (1979) apply a method called *Generalized Cross-Validation* (GCV) which solves this problem in common spline smoothing. It is introduced below, starting with a slightly simpler version, the *Ordinary Cross Validation*.

Lemma 5.1.1 (Ordinary Cross-Validation). *Define the matrix*

$$S(\lambda) = (I + \lambda K)^{-1} \tag{5.1.11}$$

with diagonal elements $\mathbf{S}_{ii}(\lambda)$. *Let* $\vec{s} = (s_1, \ldots, s_d)'$ *be the vector of all knots, while* $\hat{g}^{(-i)}(\vec{s}; \lambda)$ *denotes the estimated fit to the data if observation* y_i *at knot* s_i *is omitted. The* Ordinary Cross-Validation Score $OCV(\lambda)$ *is given by*

$$\begin{aligned} OCV(\lambda) &= \frac{1}{d} \sum_{i=1}^{d} (Y(s_i) - \hat{g}^{(-i)}(\vec{s}; \lambda))^2 \\ &= \frac{1}{d} \sum_{i=1}^{d} \left(\frac{Y(s_i) - \hat{g}(s_i; \lambda)}{1 - \mathbf{S}_{ii}(\lambda)} \right)^2 \end{aligned}$$

Ordinary Crossvalidation *(OCV) proceeds by minimizing* $OCV(\lambda)$ *numerically as a function of* λ.

♦

A derivation of the above representation of $OCV(\lambda)$ is given in Wahba (1990b). Keeping $\mathbf{S}_{ii}(\lambda)$ fixed, the crossvalidation score depends on the residuals of the fit involving *all* data points. This reduces computations considerably. The diagonal elements $\mathbf{S}_{ii}(\lambda)$ can be calculated efficiently by an algorithm proposed by Hutchinson and de Hoog (1995). The $\mathbf{S}_{ii}(\lambda)$'s correspond to the leverage values known in standard regression (cf. Cook and Weisberg 1982). An additional simplifying assumption leads to the following generalized version of $OCV(\lambda)$ proposed by Craven and Wahba (1979).

Definition 5.1.3. The *Generalized Cross-Validation Score* is given by

$$GCV(\lambda) \quad = \quad \frac{1}{d}\sum_{i=1}^{d}\left(\frac{Y(s_i)-\hat{g}(s_i;\lambda)}{1-d^{-1}tr(S(\lambda))}\right)^2 \tag{5.1.12}$$

GCV consists in minimizing $GCV(\lambda)$ as a function of λ.
♦

In (5.1.12), the diagonal elements of $S(\lambda)$ are replaced by their average value. This downweights deleted residuals with large leverage values. Craven and Wahba (1979) show that the GCV-approach should asymptotically give the best value of λ, i.e. the smoothing parameter which minimizes the average squared error at the points $s_1, \ldots, s_d$.

From a practical point of view, and in the realm of generalized additive models (GAM), Hastie and Tibshirani (1990, pp. 50) note that empirical studies suggest that GCV tends to undersmooth, particularly in small data sets. However, there is no real alternative to the GCV score available, so these authors state they tend to rely somewhat on graphical methods to choose λ, while taking the so-called *equivalent degrees of freedom* into account when judging the goodness of fit. Paralleling definitions in linear regression, Hastie and Tibshirani (1990, Sec. 3.5) suggest to use the following definitions:

Definition 5.1.4. The *equivalent degrees of freedom for a linear smoother* $S(\lambda)$ are defined as

$$df(\lambda) = tr(S(\lambda))$$

The *equivalent degrees of freedom for error* are defined as

$$df^{err}(\lambda) = n - tr(2S(\lambda) - S(\lambda)S(\lambda)')$$

The *equivalent degrees of freedom for the variance* are defined as

$$df^{var}(\lambda) = tr(S(\lambda)S(\lambda)')$$

♦

All three functions are decreasing in λ, and $tr(S(\lambda)S(\lambda)') \leq tr(S(\lambda)) \leq tr(2S(\lambda) - S(\lambda)S(\lambda)')$. See Buja et al. (1989) for a discussion of equivalent degrees of freedom.

5.1.4 Bias and Variance

Spline smoothing attempts to find a compromise between goodness of fit to the data, and some degree of smoothness. Wahba (1990b) shows that for some differential operator L_1 the (univariate) minimizer $\hat{g}$ of

$$PSS_\lambda(g) = \frac{1}{d}\sum_{i=1}^{d}(Y_i - g(s_i))^2 + \lambda \int (L_1 g)^2(s)ds \tag{5.1.13}$$

has bias

$$\begin{aligned} Bias^2(\hat{g}(s)) &= E[\hat{g}(s) - g(s)]^2 & (5.1.14) \\ &\leq \int L_1 g(s)ds & (5.1.15) \end{aligned}$$

See Definition 5.2.1 below for a precise description of L_1. The interpretation is that for functions g which approximately annihilate the penalty, the bias is small. The upper bound allows a certain degree of control over the bias and at least partly justifies the use of spline smoothing for function estimation.

From the preceding result it follows that the *Integrated Mean Squared Error* (IMSE) approximates the error variance σ_ϵ^2 well, if the (multivariate) linear differential operator L_τ

approximately annihilates the function g. It is given by

$$IMSE(\hat{g}) = \int E\left[\hat{g}(s) - g(s)\right]^2 ds \tag{5.1.16}$$

The IMSE can be kept small by an appropriate choice of the penalty term involved in the smoothing process. If some prior knowledge is given about the shape of g, it should be incorporated into the penalized sum of squares term via some operator L_τ to obtain a better estimate of g. See Ramsay and Silverman (1997) for examples on this.

Meiring et al. (1998, p.204) use splines to fit what they call a deformation plane, or *D-plane*. After projection onto the D-plane, the transformed data can be modeled by an isotropic random field, and then transformed back to the original coordinates. This approach was first suggested in Sampson and Guttorp (1992) and Guttorp and Sampson (1994). It is similar in spirit to the concept of geometric anisotropy, but is more general, since it includes a wider class of transformations than just linear ones. The motivation behind it is that variogram estimation can be performed on the deformed plane in a straightforward manner. The actual process covariance structure is then obtained from a back-transform to the original scale. Spatial deformations are also used by Schmidt and O'Hagan (2000) in a Bayesian setup to estimate nonstationary spatial covariance structure.

5.2 Thin Plate Splines

5.2.1 Definition

Thin Plate Splines are a natural generalization to splines to two (or more) dimensions. They are treated in some detail in Wahba (1990b, Sec. 2.4). The underlying model again is

$$Y(\vec{s}) = g(\vec{s}) + \epsilon(\vec{s}) \tag{5.2.17}$$

where g is the function of interest, but $\vec{s} = (s_{(1)}, \ldots, s_{(m)})$ now is a vector in $\mathbb{R}^m$ (with $m = 2$ in the MF-ERG case), and $\epsilon(\vec{s})$ denotes the error term. Changing to higher dimensions requires a generalized definition of smoothness as compared to the roughness measure (5.1.4):

Definition 5.2.1. The *Multivariate Linear Differential Operator* (multivariate LDO) applied to a function $g : \mathbb{R}^m \mapsto \mathbb{R}$ is defined as

$$L_\tau g(\vec{s}) = \sum_{\nu_1+\cdots+\nu_m=\tau} \frac{\tau!}{\nu_1! \cdots \nu_m!} \left(\frac{\partial^\tau g(\vec{s})}{\partial s_{(1)}^{\nu_1} \cdots \partial s_{(m)}^{\nu_m}} \right)^2$$

with $\vec{s} \in \mathbb{D}$. The *General Roughness Penalty* is given as

$$J_\tau(g(\vec{s})) = \int_{\mathbb{R}^m} L_\tau g(\vec{s}) \quad ds_{(1)} \ldots ds_{(m)}$$

with integers $\nu_i \geq 0$ for all $i \in \{1, \ldots, m\}$ to be chosen.
♦

Only choices of τ with $2\tau - m > 0$ guarantee a continuous solution.

All polynomial components of g of order less than τ do not contribute to the penalty term, since their τ-th derivative is zero. In what follows, the penalty

$$J_2(g(\vec{s})) = \int\int \left(\frac{\partial^2 g(\vec{s})}{\partial s_{(1)}^2} \right)^2 + 2 \left(\frac{\partial^2 g(\vec{s})}{\partial s_{(1)} \partial s_{(2)}} \right)^2 + \left(\frac{\partial^2 g(\vec{s})}{\partial s_{(2)}^2} \right)^2 ds_{(1)} ds_{(2)} \tag{5.2.18}$$

will be used. It has several advantages:

- The penalty measures departure from local linearity.
- It is invariant to rotation in $\mathbb{R}^2$.
- It is always non-negative, and zero if and only if g is linear.

The fitting criterion in $\mathbb{R}^2$,

$$PSS^2_\lambda(g) = \sum_{i=1}^{d} ||g_i - Y_i||^2 + \lambda J_2(g) \quad \text{for } \lambda > 0 \tag{5.2.19}$$

involves the bivariate penalty $J_2(g)$ from (5.2.18) and results in a so-called *Thin Plate Spline*. For a definition of thin plate splines a couple of additional function definitions are needed. Define

$$\begin{aligned}
\eta(r) &= \frac{1}{16\pi} r^2 log(r^2) \quad \text{for } r > 0 \\
\eta(0) &= 0 \\
\vec{s}_i &= (s_{(1)i}, s_{(2)i})' \quad \text{for } \vec{s}_i \in \mathbb{R}^2 \text{and } i = 1, \dots, d \\
\phi_1(\vec{s}) &= 1 \\
\phi_2(\vec{s}) &= s_{(1),j} \\
\phi_3(\vec{s}) &= s_{(2),j} \\
\mathbf{H} &= (\phi_i(\vec{s}_j))_{i=1,2,3;j=1,\dots,d} \\
&= \begin{pmatrix} 1 & 1 & \dots & 1 \\ \vec{s}_1 & \vec{s}_2 & \dots & \vec{s}_d \end{pmatrix}
\end{aligned} \tag{5.2.20}$$

The function $\eta(r)$ is a *radial basis function*. In addition, denote the Euclidean norm of $\vec{z}$ by $||\vec{z}|| = (\vec{z}'\vec{z})^{1/2}$. Then, thin plate splines are characterized as follows.

Definition 5.2.2. A function g is a *Thin Plate Spline* over the points $\vec{s}_1, \dots, \vec{s}_d \in \mathbb{R}^2$ if and only if it is of the form

$$g(\vec{s}) = \sum_{i=1}^{d} \delta_i \eta(||\vec{s} - \vec{s}_i||) + \sum_{j=1}^{3} \vartheta_j \phi_j(\vec{s}) \tag{5.2.21}$$

for $\vec{s} \in \mathbb{R}^2$ and some constants δ_i and ϑ_j in $\mathbb{R}$. It is a *natural* thin plate spline, if in addition $\mathbf{H}\vec{\delta} = 0$ with $\vec{\delta} = (\delta_1, \dots, \delta_d)'$ and $\mathbf{H}$ from (5.2.20).

♦

Among thin plate splines, only natural thin plate splines yield a finite value of the penalty $J_2(g)$ (cf. Green and Silverman 1994, Theorem 7.1). In this case, the penalty can be expressed as

$$\vec{\delta}'\mathbf{K}\vec{\delta} \tag{5.2.22}$$

where $\mathbf{K} \in \mathbb{R}^{d\times d}$ with $K_{ij} = \eta(\|\vec{s_i} - \vec{s_j}\|)$, and $\vec{\delta}$ from (5.2.21).

The criterion $PSS_\lambda^2(g)$ from (5.2.19) can now be represented in matrix notation as follows. Let $\vec{Y}$ denote the vector of observations, $\vec{\delta}$ the vector of weights for the radial basis function η, and $\vec{\vartheta}$ the vector of coefficients for the polynomial basis functions ϕ_j. Define the matrices $\mathbf{H} \in \mathbb{R}^{\tau\times d}$ with $d = \binom{m+\tau-1}{\tau}$ and $\mathbf{K} \in \mathbb{R}^{d\times d}$ with elements

$$\begin{aligned} H_{ij} &= \phi_i(\vec{s}_j) \\ K_{ij} &= \eta(\|\vec{s}_i - \vec{s}_j\|) \end{aligned}$$

Then, with $\vec{Y} = (Y(\vec{s_1}), \ldots, Y(\vec{s_d}))'$, the criterion $PSS_\lambda^2(g)$ can be written as

$$PSS_\lambda^2(g) = (\vec{Y} - \mathbf{H}\vec{\vartheta} - \mathbf{K}\vec{\delta})'(\vec{Y} - \mathbf{H}\vec{\vartheta} - \mathbf{K}\vec{\delta}) + \lambda\vec{\delta}'\mathbf{K}\vec{\delta} \tag{5.2.23}$$

which allows for efficient calculation if λ is fixed.

5.2.2 Splines for Interpolation and Smoothing

The penalized sum of squares for smoothing $PSS_\lambda^2(g)$, with fixed smoothing parameter $\lambda > 0$, is minimized by a natural thin plate spline (cf. Green and Silverman 1994, pp.147-148). It turns out that the solution $\hat{g}$ is uniquely determined by the system of equations

$$\begin{pmatrix} \mathbf{K} + \lambda\mathbf{I} & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\vec{\delta}} \\ \hat{\vec{\vartheta}} \end{pmatrix} = \begin{pmatrix} \vec{Y} \\ \vec{0} \end{pmatrix} \tag{5.2.24}$$

with matrix $\mathbf{K} = \{\frac{1}{12}|s_i - s_j|^3\}$. An exact interpolator is obtained by setting $\lambda = 0$.

A problem which did not occur in the one-dimensional case is the choice of the region over which the penalty is calculated. Edge effects are much more influential in higher dimensions than in $\mathbb{R}^1$. In fact it can be shown that the fit in the univariate case is independent of the interval $[a, b]$ which includes the knots for the spline. This is treated in some more detail in Green and Silverman (1994, Section 7.7).

On the other hand, the solution to (5.2.19) in practice does depend on the boundary of the region over which the penalty is evaluated. However, Green and Silverman argue that the effect of calculating the roughness only over a finite window is 'not enormous' (p. 153), and refer the reader to the literature for empirical support of their statement. In the applications to ERG data, edge effects therefore will be ignored.

5.2.3 Assessment of Spline-Residuals

When assessing the importance of individual observations on the overall fit, the so-called *hat-matrix* is a helpful tool in common regression diagnostics. This is the matrix $\mathbf{S}$ from (5.1.11), which maps the observation vector into the parameter space. Wahba (1978) deduces an estimate of the error variance paralleling similar results in common regression, suggesting

$$\hat{\sigma}_\epsilon^2 = \frac{\sum [z(s_i) - \hat{g}(s_i)]^2}{n - tr(\mathbf{S}(\lambda))} \tag{5.2.25}$$

Here, $tr(.)$ denotes the trace of a matrix. Simulation studies have shown that this gives indeed a good estimate of $Var(\epsilon_i) = \sigma^2$ (Wahba 1983). This estimator is the basic building block of Generalized Crossvalidation, which was introduced in Section 5.1.3.

Silverman (1985, Sec. 5.2) gives a brief treatment of regression diagnostics for splines and highlights some of the differences to the usual regression case, as described for example by

Cook and Weisberg (1982, Chapter 2). However, plots of residuals against knots, residuals against predicted values, and residuals against observed values can be interpreted in the same manner. To account for estimation bias, Silverman suggests to studentize the residuals, for example using the variance estimate from (5.2.25). He proposes

$$r(s_i) = \frac{[z(s_i) - \hat{g}(s_i)]}{\hat{\sigma}_\epsilon \left[1 - n^{-1} tr(\mathbf{S}(\lambda))\right]^{1/2}} \tag{5.2.26}$$

as studentized residuals at locations s_i.

5.3 Kriging and Splines

Several versions of Kriging were introduced in Chapter 4. Under certain circumstances, spline smoothing and Kriging yield the same estimates. This section gives some details on this connection.

5.3.1 Formal equivalence

The link between Kriging and splines is as follows. Consider the model

$$Z(\vec{s}_j) = g(\vec{s}_j) + \epsilon(\vec{s}_j) \tag{5.3.27}$$

at locations $\vec{s}_j \in \mathbb{R}^m$ for $j = 1, \ldots, d$, with trend function $g : \mathbb{R}^m \longmapsto \mathbb{R}$ with expected value $E[g(\vec{s})] = \sum \theta_k f_k(\vec{s})$. The errors $\epsilon(\vec{s}_j)$ are assumed to resemble *uncorrelated* noise with $E[\epsilon(\vec{s}_j)] = 0$ and $Var(\epsilon(\vec{s}_j)) = \sigma_\epsilon^2$ for all j. Denote the covariance between values of $Z(\vec{s}_j)$ at two knots by

$$Cov(Z(\vec{s}_i), Z(\vec{s}_j)) = \sigma_Z(\vec{s}_i, \vec{s}_j) \qquad (i, j = 1, \ldots, d) \tag{5.3.28}$$

and the matrix with elements $\sigma_Z(\vec{s}_i, \vec{s}_j)$ by Σ.

The Universal Kriging solution is obtained by solving the dual Kriging equations (4.4.46), which can be expressed as

$$\begin{pmatrix} \Sigma & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\vec{v}}_1 \\ \hat{\vec{v}}_2 \end{pmatrix} = \begin{pmatrix} \vec{Z} \\ \vec{0} \end{pmatrix} \tag{5.3.29}$$

Identifying $\mathbf{K} + \lambda I$ in (5.2.24) with $\Sigma = \tilde{\Sigma} + c_0 I$ with measurement error c_0 shows that Kriging with covariance matrix Σ coincides with spline smoothing with smoothing parameter $\lambda = c_0$. Hence, the estimates obtained from Kriging and spline smoothing coincide everywhere in this particular case.

The question might be posed if Kriging is to be preferred to spline smoothing, or vice versa. The basic view in geostatistics has been expressed by Matheron (1967). He discusses the use of Kriging versus polynomial fits in $\mathbb{R}$ with particular reference to geology, arguing that in geology, the former is more appropriate due to lack of the necessary theoretical background which would justify the assumption of a deterministic polynomial trend plus White Noise. Since trend as well as error structure both are of stochastic nature in geology, Kriging seems to be more appropriate than polynomial fitting. While both give weighted averages of the observed values, Kriging gives the solution with minimal prediction variance. This is not only of mathematical interest, but is also a convincing economical argument in geological mining applications. Additional arguments to decide on the appropriate choice between Kriging and splines can be found in (Cressie 1990b) and (Wahba 1990a).

5.4 Some Alternatives to Splines

The choice of splines for smoothing is somewhat arbitrary, although it has several advantages in terms of local adaptivity, interpretability and computational feasibility. In this section, some alternative methods are briefly addressed. They are mostly based on the notion that the trend to be estimated can be described by means of several different function bases. Thin plate splines are just one possible choice.

5.4.1 Local Polynomials

Instead of splines, local polynomials of low order could be used to interpolate the data. However, a local polynomial basis would not incorporate restrictions on derivatives at the knots which guarantee a certain degree of smoothness of the result. In addition, polynomials have some other disadvantages which make them unattractive in many situations:

- Polynomials are known to be highly influenced by extreme values.
- The degree of polynomials must be integer, so there is no smooth transition between different choices of polynomials.

In summary, piecewise polynomials are not as smooth and not as flexible as penalized splines, and therefore are less favourable.

5.4.2 Wavelets

Wavelets (Mallat 1989) have been of some interest in recent years due to their localizing property both in time (or space, for that matter) and frequency. Daubechies (1992) and Chui (1992) are two introductory references. Several univariate wavelet basis functions with compact support are available (Daubechies 1988), allowing for efficient and precise wavelet decomposition of functions under study. For higher dimensions, Rioul and Vetterli (1991) propose a way to construct appropriate basis functions. The result of a wavelet analysis is a so-called multiresolution decomposition, which decomposes the underlying function into several scales. A scale may very roughly be compared to a frequency band in Fourier analysis, although Priestley (1996) shows that the correspondence is not exact. A multiresolution analysis can be formulated as a linear filtering operation, which allows for efficient computation.

A disadvantage of wavelet methods is that a relatively large number of observations and a regular grid of locations is required. Edge effects may occur, although they can be accounted for if certain assumptions are made, like for example periodicity of the underlying function. Since the multifocal ERG provides only relatively few spatial locations, which in addition are given on a hexagonal grid, wavelet analysis does not seem to be appropriate for the data available.

5.4.3 Splines and Kernel Methods

There is a close connection between splines and kernel methods. This is most easily seen by expressing spline smoothing as a linear operation as in (5.1.9). One obtains the *equivalent kernel* (cf. Hastie and Tibshirani 1990) as the weights $\vec{\delta}$ in (5.2.21) of the linear spline operator at the knots $\{s_1, \ldots, s_d\}$. See Silverman (1984) for further details. The choice of the bandwidth in kernel smoothing then corresponds to the choice of the smoothing parameter λ in spline regression. However, the explicit incorporation of derivative properties into the penalty term makes splines to become a more attractive choice for the analysis of MF-ERG data sets.

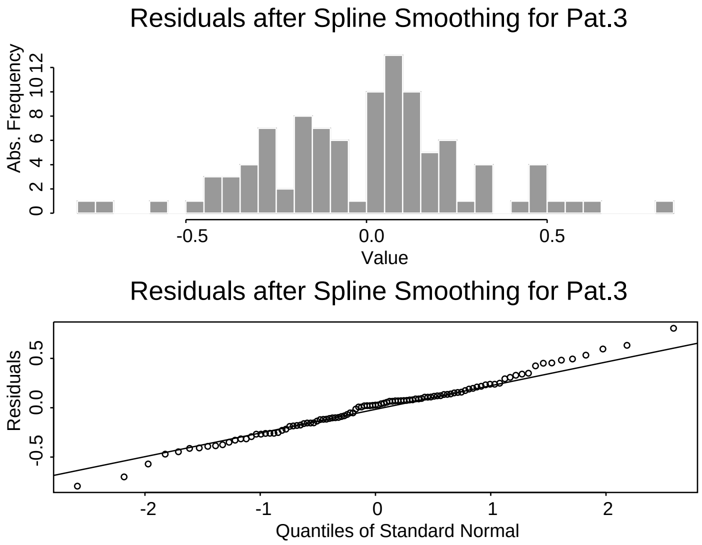

Figure 5.1: Histogram and QQ-plot for the residuals of the fit.

5.5 Application: Smoothing of Amplitudes

The spline smoothing approach is demonstrated on amplitudes derived from data set Pat.3, because they show a certain amount of irregular local variation in trend. Figures 5.2 and 5.3 show the original amplitudes and their smoothed version. The roughness parameter λ was chosen by generalized crossvalidation as 0.50679 using the S-Plus® software library FUNFITS provided by and described in Nychka et al. (2000). The residuals from the fit are displayed in a perspective plot in Figure 5.4. Estimated values are satisfactory when taking the residual distribution as a measure of fit (Figure 5.1), although the hollow in the histogram for values close to zero is remarkable and calls for further investigation. The graphical display of fitted values (Figure 5.3) has a smooth appearance, yet conserving certain local features. The effective degrees of freedom for fit are 45, compared to 5 for polynomial trend plus 3 for variogram fit in the Universal Kriging approach. This indicates the price that is paid for the local adaptivity of splines.

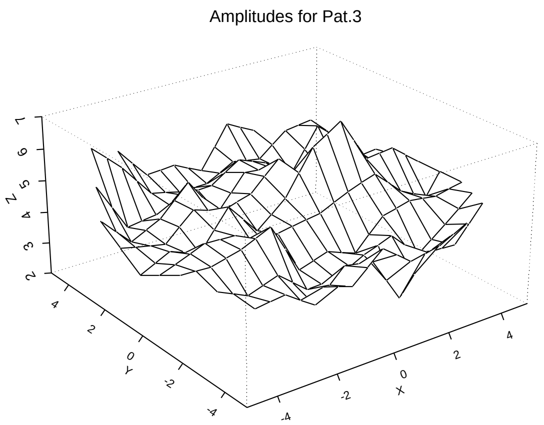

Figure 5.2: Original amplitudes for data set Pat.3.

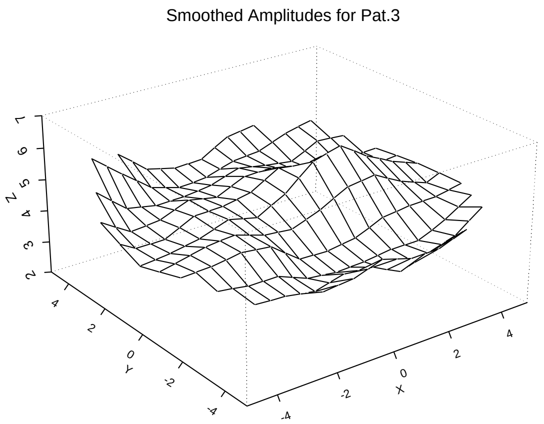

Figure 5.3: Smoothed amplitudes for data set Pat.3 with parameter $\lambda = 0.50679$.

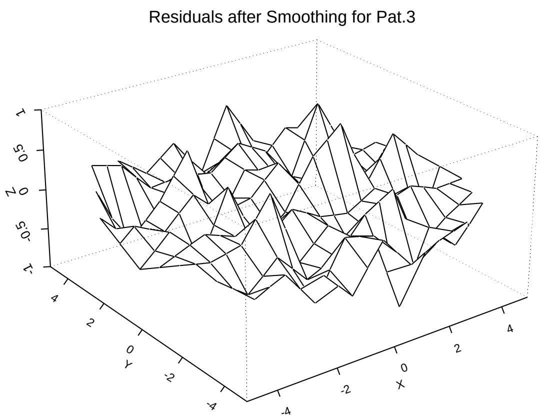

Figure 5.4: Residuals after spline smoothing of amplitudes for data set Pat.3.

Chapter 6

Smoothing of AR-Parameter Fields

This chapter tries to combine and modify some of the statistical methods presented above to find a set of AR-parameter estimates which reflect the spatiotemporal dynamics in the data, while being interpretable to the investigator. The approach suggested is to use a combination of locally applied time series techniques within each hexagon of the underlying grid, and spatial spline smoothing over the resulting parameter estimates.

In particular, autoregressive models of fixed order p will be fit to each of the 103 time series available. The result is interpreted as a set of estimates of p underlying parameter surfaces, or *AR-parameter fields*, which will be smoothed in a second step. This approach differs from other methods for spatiotemporal data analysis in that it avoids explicit a priori modeling of spatial dependencies.

The goal of many commonly used techniques is to optimally predict values of the observed process at future times or at unobserved locations. Main interest is *not* in a spatially smooth set of parameters. For example, consider the work of Meiring et al. (1998) and Sampson et al. (1994) which is based on Sampson and Guttorp (1992). The authors estimate AR-parameters from time series analysis and interpolate them deterministically to obtain a trend estimate. Error estimates obtained by Kriging are added for prediction. Although this is perfectly reasonable for prediction, it does not give in general a good description of the deterministic trend in the AR parameters. It is not apparent which part of the predicted AR coefficient is due to random fluctuation, and which part can be ascribed to deterministic trend.

In a different approach, Huang and Cressie (1996) assume constant AR-parameters and incorporate spatiotemporal dynamics via temporally varying spatial covariances. Estimation is done using Kalman-filter recursions (Kalman 1960). Smoothing as a means to provide trend estimates for the parameters themselves is not done. In regard to the ERG application at hand, it seems not unlikely that the driving forces behind the physiological process exhibit some degree of smoothness in space. Hence, spatial smoothing of these parameters is meaningful. An appropriate estimator is proposed and applied below.

6.1 Model Formulation

6.1.1 Basic Notation

A purely temporal univariate autoregressive process $(Z_t)_{t\in\mathbb{T}}$ can be described by the equation

$$Z_t = \mu + \sum_{k=1}^{p} \alpha_k Z_{t-k} + \epsilon_t$$

where ϵ_t is taken from a White Noise process. To introduce a spatial component, an argument $\vec{s}$ is added as spatial index, resulting in the representation

$$Z_t(\vec{s}) = \mu(\vec{s}) + \sum_{k=1}^{p} \alpha_k(\vec{s}) Z_{t-k}(\vec{s}) + \epsilon_t(\vec{s}) \tag{6.1.1}$$

Without loss of generality, $\mu(\vec{s}) = 0$ is assumed for all $\vec{s} \in \mathbb{D}$. This can always be achieved in practice by removing the overall mean before performing the actual analysis. In the sum on the right hand side, other locations than just $\vec{s}$ could be considered as influential and MA-type dependencies could be included, leading to STARMA-type models (Pfeiffer and Deutsch 1980a, Pfeiffer and Deutsch 1980b). However, these models require a priori structural assumptions on spatial dependencies, and the resulting model quickly becomes considerably more cluttered and difficult to interpret. In fact, determination of an influential neighborhood is not always straight forward, and the search for a correct model of neighborhood structure may be a major problem. Spatial dependencies like in the STARMA case will therefore not be considered here.

The class of linear spatiotemporal models under study can be formulated in matrix notation as follows. Take $\vec{Z}_t$ to be the vector of all observations at time t over the finite set of locations $\{\vec{s}_1 \ldots, \vec{s}_d\} \in \mathbb{D}$, with corresponding error term $\vec{\epsilon}_t$,

$$\vec{Z}_t = (Z_t(\vec{s}_1), \ldots, Z_t(\vec{s}_d))' \in \mathbb{R}^{d \times 1} \tag{6.1.2}$$

$$\vec{\epsilon}_t = (\epsilon_t(\vec{s}_1), \ldots, \epsilon_t(\vec{s}_d))' \in \mathbb{R}^{d \times 1} \tag{6.1.3}$$

Further, let

$$\mathbf{Z} = \begin{pmatrix} \vec{Z'}_{p+1} \\ \vdots \\ \vec{Z'}_T \end{pmatrix}_{(T-p)\times d} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \vec{Z'}_p & \ldots & \vec{Z'}_1 \\ \vdots & \ldots & \vdots \\ \vec{Z'}_{T-1} & \ldots & \vec{Z'}_{T-p} \end{pmatrix}_{(T-p)\times dp} \tag{6.1.4}$$

Define the parameter matrix $\mathbf{A}$ with submatrices $\mathbf{A}_k$ in $\mathbb{R}^{(d\times d)}$ for all $k = 1, \ldots, p$, and the error matrix $\mathbf{E}$ with rows $\vec{\epsilon'}_t$ in $\mathbb{R}^{(1\times d)}$ for all $t = p+1, \ldots, T$ by

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_p \end{pmatrix}_{dp\times d} \quad \text{and} \quad \mathbf{E} = \begin{pmatrix} \vec{\epsilon'}_{p+1} \\ \vdots \\ \vec{\epsilon'}_T \end{pmatrix}_{(T-p)\times d} \tag{6.1.5}$$

Following the notation of the VAR(p)-model from Chapter 3, the spatiotemporal model can now be written as

$$\mathbf{Z} = \mathbf{X}\mathbf{A} + \mathbf{E} \tag{6.1.6}$$

where $\mathbf{X}$ contains lagged versions of $\mathbf{Z}$, and the parameters in $\mathbf{A}$ can be identified with their corresponding spatial locations. If only temporal autoregressive dependencies are assumed without influence by neighboring areas, the $\mathbf{A}_k$ are diagonal matrices, i.e.,

$$\mathbf{A}_k = Diag(\vec{\alpha}_k) = \begin{pmatrix} \alpha_k(s_1) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \alpha_k(s_d) \end{pmatrix} \tag{6.1.7}$$

for $k = 1, \dots, p$. This is a crucial assumption in the estimation steps to follow. It reflects the fact that the 103 time series initially are modeled separately. However, the above matrix form will simplify notation in later steps.

6.1.2 Penalizing the Sum of Squares

Although in a first step temporal dependencies will be assumed to exist only locally within each of the 103 time series, spatial dependencies will be introduced indirectly through smoothness conditions imposed on the resulting estimated AR-parameter field. Since $\mathbb{D}$ is countable, the diagonal elements of each $\mathbf{A}_k$ may be combined into a vector $\vec{\alpha}_k$. The AR coefficients of p fields may in turn be combined into a vector $\vec{\alpha} = \left(\vec{\alpha}'_1, \dots, \vec{\alpha}'_p\right)'$. Smoothing may then be done within each field via the roughness penalty approach described in Chapter 5. The physiological argument for smoothing of ERG data is given by the fact that the response density on the retina of a healthy eye varies continuously according to

$$\text{response density} \quad = \quad ae^{bx} + c \tag{6.1.8}$$

where a, b and c are constants, x is retinal eccentricity in degrees, and e denotes Euler's constant (Verdon and Haegerstrom-Portnoy 1998).

As was shown above, the penalty (5.2.18) combined with a least squares fitting criterion results in a solution within the class of thin plate splines. It provides a means to control the spatial smoothness of a given parameter field estimator via its derivatives. Note that in principle, other penalty terms are equally applicable which incorporate additional information on the overall shape of the field.

When the goal is to smooth the autoregressive parameters separately for each $k = 1, \ldots, p$, a first naive attempt may be to look for p vector-valued functions

$$\breve{\alpha}_k = (\breve{\alpha}_k(\vec{s}_1), \ldots, \breve{\alpha}_k(\vec{s}_d))' \tag{6.1.9}$$

$(k = 1, \ldots, p)$ which minimize

$$\begin{aligned} PENSS(\breve{\mathbf{A}}) &= \sum_{k=1}^{p} (\breve{\alpha}_k - \vec{\alpha}_k)'(\breve{\alpha}_k - \vec{\alpha}_k) + \sum_{k=1}^{p} \lambda_k \int_{\mathbb{R}^2} L_2 \left(\breve{\alpha}_k(\vec{s})\right) d\vec{s} && (6.1.10) \\ &= \sum_{k=1}^{p} PSS(\breve{\mathbf{A}}_k) && (6.1.11) \end{aligned}$$

where the penalty functional is taken from (5.1.5). In practice however, direct observation of the $\vec{\alpha}_k$ involved in $PENSS(\breve{\mathbf{A}})$ is not possible. In addition, the above target function does not take into account at all the quality of the *temporal* fit. Therefore, (6.1.10) is not what is sought for.

An alternative solution may again be based on the commonly used penalized sum of squares from multivariate spline smoothing (Hastie and Tibshirani 1990; Ramsay and Silverman 1997), but now including an additional penalty for the temporal fit. This leads to the following modification.

Definition 6.1.1 (Penalized Sum of Squares for $\mathbf{A}^*$). Let $\otimes$ denote the Kronecker matrix product, and adapt the following notation:

$$\begin{aligned}
\alpha_k^* &= (\alpha_k^*(\vec{s}_1), \ldots, \alpha_k^*(\vec{s}_d))' \in \mathbb{R}^d \\
\alpha^* &= \left(\alpha_1^{*'}, \ldots, \alpha_p^{*'}\right)' \in \mathbb{R}^{pd} \\
\mathbf{A}_k^* &= Diag(\alpha_k^*) \in \mathbb{R}^{d\times d} \\
\mathbf{A}^* &= \left(\mathbf{A}_1^*, \ldots, \mathbf{A}_p^*\right)' \in \mathbb{R}^{pd\times d}
\end{aligned}$$

$$\begin{aligned}
\mathbf{Z}^* &= \mathbf{X}\mathbf{A}^* \quad \text{with } \mathbf{X} \text{ from (6.1.4)} \\
\vec{\lambda} &= (\lambda_1, \ldots, \lambda_p)' \\
\Lambda &= Diag(\vec{\lambda}) \otimes \mathbf{K} \in \mathbb{R}^{pd\times pd} \\
\mathbf{U} &= \frac{1}{d}\left(\mathbf{I}_p \otimes \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix}\right)_{dp\times dp} \\
\bar{\alpha}^* &= U\mathbf{A}^* 1_{dp}
\end{aligned}$$

Then the *Penalized Sum of Squares for α^** is defined as

$$\begin{aligned}
PENSS^*(\alpha^*) &= tr\left\{(\mathbf{Z}-\mathbf{Z}^*)'(\mathbf{Z}-\mathbf{Z}^*)\right\} \\
&\quad + \sum_{k=1}^{p} \lambda_k \int_{\mathbb{R}^2} L_2\left[\alpha_k^*(\vec{s})\right] d\vec{s} \qquad (6.1.12) \\
&= tr\left\{(\mathbf{Z}-\mathbf{Z}^*)'(\mathbf{Z}-\mathbf{Z}^*)\right\} \\
&\quad + (\alpha^* - \bar{\alpha}^*)' \,\Lambda\, (\alpha^* - \bar{\alpha}^*) \qquad (6.1.13)
\end{aligned}$$

♦

This sum indeed involves components both on the temporal level (via $\mathbf{Z}$) and the parameter level (via α^*). It differs from the penalized sum usually chosen in the multivariate spline smoothing framework. To see this, recall that the goal in multivariate spline smoothing is to find functions $g_k(.)$ in $\mathbb{R}^m$ within the model framework

$$\vec{y} = \sum_{k=1}^{p} g_k(\vec{s}) + \text{White Noise} \qquad (6.1.14)$$

with dependent variable $\vec{y}$, and regressors g_k which are *independent* of $\vec{y}$ (Hastie and Tibshirani 1990). In the standard approach, g_k is deterministic, and a roughness penalty is imposed on the estimates of the g_k. At first sight, one could be tempted to identify $\vec{y}$ with $\vec{Z}_{t_0}$ for any t_0 fixed , and $g_k(\vec{s})$ with $\mathbf{A}_k^*\vec{Z}_{t_0-k}$, relating the autoregressive model (6.1.1) with the common spline smoothing model. However, the goal followed in the ERG case is to smooth the parameters in $\mathbf{A}_k^*$ directly, rather than smoothing the product $\mathbf{A}_k^*\vec{Z}_{t_0-k}$.

Note that the first summand in $PENSS^*$ in (6.1.12), which is just the sum of squared residuals, is clearly minimized by the ordinary least squares estimate $\hat{\mathbf{A}}_k$, but for $\lambda_k \neq 0$ for all k, some deviation from $\hat{\mathbf{A}}_k$ is to be expected because of the smoothness condition incorporated in the penalty term.

The OLS estimate $\hat{\mathbf{A}}$ by definition is the optimal solution (in the least squares sense) when only the *temporal* fit to the data is of interest. It is therefore proposed here to consider a penalized sum of squares which treats the least squares solution as observed values, or *pseudo-observations*. These constitute the starting point for smoothing of the respective AR-parameter field.

The fit at the $\mathbf{Z}$-level (i.e. in time) is then taken into account by incorporating a term of squared residuals $tr\left\{(\mathbf{Z}-\tilde{\mathbf{Z}})'(\mathbf{Z}-\tilde{\mathbf{Z}})\right\}$ as a summand, where $\tilde{\mathbf{Z}}$ represents the predicted values of the observations obtained under an adapted penalized sum. The process is more clearly described by the following definition.

Definition 6.1.2. [Space-Time Penalized Sum of Squares] Define

$$\tilde{Z}_t = \sum_{k=1}^{p} \tilde{\mathbf{A}}_k \vec{Z}_{t-k} \tag{6.1.15}$$

to be the predicted value of $\vec{Z}_t$ from (6.1.2) for some estimator $\tilde{\mathbf{A}}_k$ of $\mathbf{A}_k$. In particular, take $\tilde{\mathbf{A}} = \left(\tilde{\mathbf{A}}_1', \ldots, \tilde{\mathbf{A}}_p'\right)'$ with $\tilde{\mathbf{A}}_k$ being diagonal matrices for all $k = 1, \ldots, p$ with elements

$\tilde{\alpha}_k = (\tilde{\alpha}_k(\vec{s}_1), \ldots, \tilde{\alpha}_k(\vec{s}_d))'$ on its diagonal, and zero otherwise. In addition, take $\vec{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ and $\vec{\nu} = (\nu_1, \ldots, \nu_p)'$ to be vectors of smoothing parameters which take only nonnegative values. Then the *Space-Time Penalized Sum of Squares* (STPSS) corresponding to $\tilde{\alpha} = (\tilde{\alpha}_1', \ldots, \tilde{\alpha}_p')'$ and with fixed smoothing parameters $\vec{\lambda}$ and $\vec{\nu}$ both in $\mathbb{R}^p$ is defined as

$$\begin{aligned} STPSS_{\vec{\lambda},\vec{\nu}}(\tilde{\alpha}) &= tr\left\{(\mathbf{Z} - \tilde{\mathbf{Z}})'(\mathbf{Z} - \tilde{\mathbf{Z}})\right\} \\ &+ \sum_{k=1}^{p} \lambda_k \int_{\mathbb{R}^d} L_2(\tilde{\alpha}_k(\vec{s}))d\vec{s} \qquad (6.1.16) \\ &+ \sum_{k=1}^{p} \nu_k (\hat{\alpha}_k - \tilde{\alpha}_k)'(\hat{\alpha}_k - \tilde{\alpha}_k) \end{aligned}$$

Here $\hat{\alpha} = (\hat{\alpha}_1', \ldots, \hat{\alpha}_p')'$ represents the OLS parameter estimate for (6.1.15).
♦

In matrix notation, with $\Lambda = Diag(\vec{\lambda}) \otimes \mathbf{K}$ with $\mathbf{K}$ from (5.1.3), and

$$\begin{aligned} \mathbf{V} &= Diag(\vec{\nu}) \otimes I_d \\ \mathbf{U} &= \frac{1}{d}\left(\mathbf{I}_p \otimes \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix}_{d \times d}\right)_{dp \times dp} \\ \tilde{\tilde{\alpha}} &= U\tilde{\alpha} \end{aligned}$$

one obtains the alternative representation

$$\begin{aligned} STPSS_{\vec{\lambda},\vec{\nu}}(\tilde{\alpha}) &= tr\left((\mathbf{Z} - \tilde{\mathbf{Z}})'(\mathbf{Z} - \tilde{\mathbf{Z}})\right) \qquad (6.1.17) \\ &+ (\tilde{\alpha} - \tilde{\tilde{\alpha}})'\Lambda(\tilde{\alpha} - \tilde{\tilde{\alpha}}) \qquad (6.1.18) \\ &+ (\tilde{\alpha} - \hat{\alpha})'\mathbf{V}(\tilde{\alpha} - \hat{\alpha}) \qquad (6.1.19) \end{aligned}$$

The newly introduced weights ν_k allow to balance the close fit on the $\mathbf{Z}$-level versus a smooth spatial fit to the pseudo-observations in $\hat{\alpha}$. To see more clearly what STPSS is actually doing, write $\lambda_k = \nu_k \tilde{\lambda}_k$ with $\tilde{\lambda}_k = \frac{\lambda_k}{\nu_k} > 0$ where $\nu_k > 0$. It is then immediate that

$$\begin{aligned} STPSS_{\tilde{\lambda},\vec{\nu}}(\tilde{\alpha}) &= tr\left\{(\mathbf{Z}-\tilde{\mathbf{Z}})'(\mathbf{Z}-\tilde{\mathbf{Z}})\right\} \\ &+ \sum_{k=1}^{p} \nu_k \Big\{\tilde{\lambda}_k \int_{\mathbb{R}^2} L_2 \tilde{\alpha}_k(\vec{s}) d\vec{s} \qquad (6.1.20) \\ &\qquad + (\hat{\alpha}_k - \tilde{\alpha}_k)'(\hat{\alpha}_k - \tilde{\alpha}_k)\Big\} \\ &= SS(\tilde{\mathbf{Z}}) + \sum_{k=1}^{p} \nu_k PSS_{\tilde{\lambda}_k}(\tilde{\alpha}_k) \qquad (6.1.21) \end{aligned}$$

with PSS as in (6.1.11), but using pseudo-observations $\hat{\alpha}$. Hence, STPSS is a weighted average between the sum of squares for fit to the data, and the weighted regular penalized sums of squares for fit to the parameter fields. Note that $\tilde{\mathbf{Z}}$ directly depends on $\tilde{\alpha}$, and both summands in STPSS have to be taken into account jointly when searching for values of $\vec{\nu}$ and $\vec{\lambda}$ yielding an optimal fit.

6.2 Derivation of Estimators

6.2.1 Solution for known Smoothness Parameters

The penalized sum STPSS($\tilde{\alpha}$) from (6.1.16) includes the OLS case if $\vec{\lambda} = \vec{\nu} = \vec{0}$, and the penalized sum PENSS(α^*) from (6.1.12) if $\vec{\nu} = \vec{0}$. Starting out with

$$\begin{aligned} STPSS_{\tilde{\lambda},\vec{\nu}}(\tilde{\alpha}) &= tr\left((\mathbf{Z}-\tilde{\mathbf{Z}})'(\mathbf{Z}-\tilde{\mathbf{Z}})\right) \\ &+ (\tilde{\alpha}-\tilde{\tilde{\alpha}})'\Lambda(\tilde{\alpha}-\tilde{\tilde{\alpha}}) \qquad (6.2.22) \\ &+ (\tilde{\alpha}-\hat{\alpha})'\mathbf{V}(\tilde{\alpha}-\hat{\alpha}) \end{aligned}$$

with $\Lambda = Diag(\vec{\lambda}) \otimes \mathbf{K}$, and letting $\odot$ denote the Hadamard (i.e. elementwise) matrix product, the derivatives with respect to $\tilde{\alpha}$ are given by

$$\begin{aligned}
\frac{\partial}{\partial \tilde{\alpha}} STPSS(\tilde{\alpha}) \quad = \quad & - \quad 2 \underbrace{\left[\mathbf{X}'\mathbf{Z} \odot \begin{pmatrix} \mathbf{I}_d \\ \vdots \\ \mathbf{I}_d \end{pmatrix}_{dp \otimes d} \right]}_{=:\widetilde{\mathbf{X}'\mathbf{Z}}} 1_d \\
& + \quad 2 \underbrace{\left[\mathbf{X}'\mathbf{X} \odot \begin{pmatrix} \mathbf{I}_d & \dots & \mathbf{I}_d \\ \vdots & \ddots & \vdots \\ \mathbf{I}_d & \dots & \mathbf{I}_d \end{pmatrix}_{dp \otimes dp} \right]}_{=:\widetilde{\mathbf{X}'\mathbf{X}}} \tilde{\alpha} \\
& + \quad 2\Lambda\tilde{\alpha} - 4\Lambda\mathbf{U}\tilde{\alpha} + 2\mathbf{U}\Lambda\mathbf{U}\tilde{\alpha} \\
& + \quad 2\mathbf{V}\tilde{\alpha} - 4\mathbf{V}\hat{\alpha}
\end{aligned}$$

For ease of notation, the copies of $\mathbf{X}'\mathbf{Z}$ and $\mathbf{X}'\mathbf{X}$ which only contain the corresponding block diagonals (and zeros in the off-diagonals) are denoted by $\widetilde{\mathbf{X}'\mathbf{X}}$ and $\widetilde{\mathbf{X}'\mathbf{Z}}$, respectively. This corresponds to the assumption that all pairs of time series are uncorrelated. Equating to zero yields the solution

$$\tilde{\alpha}_{min} \quad = \quad \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{V} + \Lambda + (\mathbf{U} - 2\mathbf{I})\Lambda\mathbf{U} \right]^{-1} \left[\widetilde{\mathbf{X}'\mathbf{Z}} 1_d + \mathbf{V}\hat{\alpha} \right] \tag{6.2.23}$$

where $\hat{\alpha}$ denotes the ordinary least squares solution from a VAR(p)-model with d independent series. See Appendix F for details on how this solution is derived.

To simplify notation, the matrix

$$\mathbf{C} = \Lambda + (\mathbf{U} - 2\mathbf{I})\Lambda\mathbf{U} \tag{6.2.24}$$

is introduced which can be shown to be nonnegative definite. Setting $\vec{\lambda}$ and $\vec{\nu}$ to zero where appropriate, using known results from matrix algebra (e.g. Harville 1997, pp.419), and assuming the inverse matrices involved do exist, the following representations are possible for the three estimators given in equations (3.4.33), (6.1.12) and (6.2.23):

$$\begin{aligned}
\hat{\alpha} &= \left[\widetilde{\mathbf{X}'\mathbf{X}}\right]^{-1} \widetilde{\mathbf{X}'\mathbf{Z}} 1_d && (6.2.25)\\
\alpha^*_{min} &= \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right]^{-1} \widetilde{\mathbf{X}'\mathbf{Z}} 1_d \\
&= \left[\mathbf{I} - \left(\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right)^{-1} \mathbf{C}\right] \hat{\alpha} && (6.2.26)\\
\tilde{\alpha}_{min} &= \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C} + \mathbf{V}\right]^{-1} \left[\widetilde{\mathbf{X}'\mathbf{Z}} 1_d + \mathbf{V}\hat{\alpha}\right] \\
&= \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C} + \mathbf{V}\right]^{-1} \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{V}\right] \hat{\alpha} \\
&= \left[\mathbf{I} + \left(\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{V}\right)^{-1} \mathbf{C}\right]^{-1} \hat{\alpha} \\
&= \left[\mathbf{I} - \left(\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C} + \mathbf{V}\right)^{-1} \mathbf{C}\right] \hat{\alpha} && (6.2.27)
\end{aligned}$$

These representations show that the last two estimators which result in smoothed AR-parameters are directly expressible in terms of the OLS estimate $\hat{\alpha}$. This considerably facilitates comparison. Before this is done, the practical problem of choosing the smoothness parameters will be treated.

6.2.2 Determination of Smoothness via Crossvalidation

Smoothing parameters have to be chosen in a satisfactory, objective and computationally feasible way. Originally in the setup of univariate spline smoothing, Craven and Wahba (1979) proposed a version of crossvalidation for this choice. Their approach is slightly adapted here to the spatiotemporal setting and to the estimator $\tilde{\alpha}$ described above.

The general idea of crossvalidation is to evaluate the quality of an estimator by its ability to predict new values. In practice, this is done by leaving out a single observation, calculating the estimator of interest, and predicting the value which was set aside. This results in a residual. Repeating the same procedure for all other observations and summing up the squared residuals obtained gives an indication of how well an estimator performs. If the estimator depends on some parameter vector $\vec{\lambda}$, say, then one may attempt to minimize the resulting sum of squares after crossvalidation as a function of that vector to find an optimal parameter estimate.

The appropriateness of the choice of a given (univariate) smoothing parameter λ_k can be judged according to the value of a *generalized crossvalidation function* evaluated at λ_k. Let $\hat{\alpha}_k$ represent the k-th AR-parameter field, and take $\lambda_k \in \mathbb{R}_+$ to be a smoothing parameter. An optimal smoother g_{λ_k} for the parameter field is then given by the minimizer of

$$V_{GCV}(\lambda_k) = \frac{1}{d}\sum_{i=1}^{d}\left[\hat{\alpha}_k(\vec{s}_i) - g_{\lambda_k}(\vec{s}_i)\right]^2 / \left[1 - \frac{1}{d}tr(\mathbf{S}(\lambda_k))\right]^2 \tag{6.2.28}$$

This was already shown in Chapter 5 on spline smoothing. Recall now the representation (6.1.21) of $STPSS_{\vec{\lambda},\vec{\nu}}(\tilde{\alpha})$ in terms of the sum of $SS(\tilde{\mathbf{Z}})$ and a weighted sum of the $PSS_{\tilde{\lambda}_k}(\tilde{\alpha}_k)$ with $\tilde{\lambda}_k = \frac{\lambda_k}{\nu_k}$. It incorporates two aspects of fitting, which are the fit of the original data $\vec{Z}_t$, and the smoothness and fit within the p AR-parameter fields. The problem now is to determine parameter vectors $\vec{\lambda}$ and $\vec{\nu}$ which lead to an optimal fit with respect to $STPSS_{\vec{\lambda},\vec{\nu}}(\tilde{\alpha})$. As a solution, the *Space-Time Crossvalidation Function* is introduced.

Definition 6.2.1. Let $V_{GCV}(\tilde{\lambda}_k)$ denote the ordinary crossvalidation score from (5.1.12) evaluated at $\tilde{\lambda}_k = \frac{\lambda_k}{\nu_k}$ for the k-th AR-parameter field. Let $SS_{\vec{\lambda},\vec{\nu}}(\tilde{Z}^{[j]})$ denote the sum of squares for the observations $\vec{Z}_t$ obtained from solution $\tilde{\alpha}_{min}$ to (6.1.21) given data at all locations except $\vec{s}_j$, and inserting the predictions of the $\alpha_k(\vec{s}_j)$'s for $k = 1, \ldots, p$. Then, the *Space-Time Crossvalidation Function* (STCVF) $V_{ST}(\vec{\lambda}, \vec{\nu})$ is defined as

$$V_{ST}(\vec{\lambda}, \vec{\nu}) = \frac{1}{d}\sum_{j=1}^{d} SS_{\vec{\lambda},\vec{\nu}}(\tilde{Z}^{[j]}) + \sum_{k=1}^{p}\nu_k V_{GCV}(\tilde{\lambda}_k)$$

where d is the number of locations where observations were made. The vector $\vec{\zeta}_{min} = (\vec{\lambda}'_{min}, \vec{\nu}'_{min})'$ minimizing $V_{ST}(\vec{\lambda}, \vec{\nu})$ is called the *Space-Time Crossvalidation Minimizer*. The process of finding $\vec{\zeta}_{min}$ is called *Space-Time Crossvalidation* (STCV).
♦

Note that $V_{ST}(\vec{\lambda}, \vec{\nu})$ requires $\nu_k > 0$ for all k to be sensibly defined. If $v = 0$, it will be implicitly assumed that the penalty $PENSS^*$ from (6.1.12) is applied. The computational effort to calculate $V_{ST}(\vec{\lambda}, \vec{\nu})$ are quite high, since a closed form as for the Generalized Cross-validation Score (5.1.12) is not known.

From an interpretational point of view, the resulting Space-Time Crossvalidation Minimizer balances out the quality of fit to the data with the smoothness of the estimated AR-parameter fields in such a way that the data are optimally fit.

6.3 Comparison of Estimators

6.3.1 Direct Differences

The differences between the estimators $\hat{\alpha}$, α^*_{min} and $\tilde{\alpha}_{min}$ can be expressed as follows.

$\underline{\alpha^*_{min} - \hat{\alpha}}$

With α^*_{min} as derived in (6.2.26) one obtains

$$\alpha^*_{min} - \hat{\alpha} \quad = \quad - \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right]^{-1} \mathbf{C}\hat{\alpha} \tag{6.3.29}$$

showing that $\Lambda \to \mathbf{0}$ results in convergence of α^*_{min} to $\hat{\alpha}$.

$\underline{\tilde{\alpha}_{min} - \hat{\alpha}}$

With $\mathbf{C}$ as in (6.2.24) and $\tilde{\alpha}_{min}$ as derived in (6.2.27) one obtains

$$\tilde{\alpha}_{min} - \hat{\alpha} \quad = \quad - \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C} + \mathbf{V}\right]^{-1} \mathbf{C}\hat{\alpha} \tag{6.3.30}$$

See Harville (1997, p. 419) for the inversion formula used. From this representation it is clear that the difference between the two estimators converges to zero if Λ and thus $\mathbf{C}$ go to zero, i.e.,

$$\tilde{\alpha}_{min} - \hat{\alpha} \longrightarrow \vec{0} \quad (\Lambda \to 0) \tag{6.3.31}$$

On the other hand, for Λ and $\mathbf{C}$ fixed, elementwise convergence of $\mathbf{V}$ to infinity has the same effect.

$$\tilde{\alpha}_{min} - \hat{\alpha} \longrightarrow \vec{0} \quad (\mathbf{V} \to \infty) \tag{6.3.32}$$

$\underline{\alpha^*_{min} - \tilde{\alpha}_{min}}$

Combining the two preceding results, and assuming the *same* matrix Λ is used in both cases, the two estimators α^*_{min} and $\tilde{\alpha}_{min}$ differ in the following way:

$$\begin{aligned} \alpha^*_{min} - \tilde{\alpha}_{min} &= \alpha^*_{min} - \hat{\alpha} - (\tilde{\alpha}_{min} - \hat{\alpha}) && (6.3.33)\\ &= \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right]^{-1} \mathbf{C}\hat{\alpha} - \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C} + \mathbf{V}\right]^{-1} \mathbf{C}\hat{\alpha} && (6.3.34)\\ &= \left\{\left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right]^{-1} - \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C} + \mathbf{V}\right]^{-1}\right\} \mathbf{C}\hat{\alpha} && (6.3.35)\\ &= \left\{\left(\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right)\left[\mathbf{V}^{-1}\left(\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{C}\right) + \mathbf{I}\right]\right\}^{-1} \mathbf{C}\hat{\alpha} && (6.3.36) \end{aligned}$$

The last equality follows from Harville (1997, p. 420). Again, letting Λ (and thus $\mathbf{C}$) converge to $\mathbf{0}$, the difference converges to zero as well. However, for $\mathbf{C}$ there remains a difference between the two estimators for every $\mathbf{V}$ with $v_{kk} > 0$ for all k, even in the limit (i.e., if $\vec{\nu} \to \infty$ in all its components).

In summary, one has

$$\hat{\alpha} \in \left\{\alpha^*(\vec{\lambda}) \mid \vec{\lambda} \in \mathbb{R}^p_{\geq 0}\right\} \subset \left\{\tilde{\alpha}(\vec{\lambda}, \vec{\nu}) \mid \vec{\lambda} \in \mathbb{R}^p_{\geq 0}; \vec{\nu} \in \mathbb{R}^p_{\geq 0}\right\}$$

All estimators can be expressed by appropriate weighting of $\hat{\alpha}$, where the weights involve the original observations via $\widetilde{\mathbf{X}'\mathbf{X}}$. The difference between any pair of α^*_{min}, $\tilde{\alpha}_{min}$ and the OLS-solution $\hat{\alpha}$ is readily available for given smoothing parameters $\vec{\lambda}$ and $\vec{\nu}$ and the matrix of inner products $\widetilde{\mathbf{X}'\mathbf{X}}$.

6.3.2 Estimators and Associated Sum of Squares

Since the class of estimators $\tilde{\alpha}$ obtained from the Space-Time Penalty includes the OLS estimator as well as α^*, it is the starting point for the derivation of the sum of squares between observations $\vec{Z}_t(s)$ and predicted values $\tilde{Z}_t(s)$ for all s and t. The following representation can be found in terms of the matrix $\tilde{\mathbf{A}} = (Diag(\tilde{\alpha}_1), \ldots, Diag(\tilde{\alpha}_p))'$.

$$\begin{aligned} SS(\tilde{\mathbf{A}}) &= tr\left((\mathbf{Z}-\tilde{\mathbf{Z}})'(\mathbf{Z}-\tilde{\mathbf{Z}})\right) && (6.3.37)\\ &= \underbrace{tr\left(\mathbf{Z}'\mathbf{Z} - 2\mathbf{Z}'\mathbf{X}\hat{\mathbf{A}} + \hat{\mathbf{A}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{A}}\right)}_{SS(\hat{\mathbf{A}})} \\ &+ 2tr\left((\tilde{\mathbf{A}}-\hat{\mathbf{A}})'(\mathbf{X}'\mathbf{X}\hat{\mathbf{A}} - \mathbf{X}'\mathbf{Z}) \right. && (6.3.38)\\ &+ \left. \frac{1}{2}(\tilde{\mathbf{A}}-\hat{\mathbf{A}})'\mathbf{X}'\mathbf{X}(\tilde{\mathbf{A}}-\hat{\mathbf{A}})\right) \\ &= SS(\hat{\mathbf{A}}) \\ &+ tr\left(\tilde{\mathbf{A}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{A}} - 2\tilde{\mathbf{A}}'\mathbf{X}'\mathbf{Z} + 2\hat{\mathbf{A}}'\mathbf{X}'\mathbf{Z} - \hat{\mathbf{A}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{A}}\right) && (6.3.39) \end{aligned}$$

From this, it follows that the difference in the sum of squares due to the three estimators $\hat{\mathbf{A}}$, $\mathbf{A}^*$ and $\tilde{\mathbf{A}}$ can be expressed as follows.

$$\begin{aligned} SS(\tilde{\mathbf{A}}) - SS(\hat{\mathbf{A}}) &= tr\left(2(\tilde{\mathbf{A}}-\hat{\mathbf{A}})'(\mathbf{X}'\mathbf{X}\hat{\mathbf{A}} - \mathbf{X}'\mathbf{Z}) \right. \\ &\quad + \left. (\tilde{\mathbf{A}}-\hat{\mathbf{A}})'\mathbf{X}'\mathbf{X}(\tilde{\mathbf{A}}-\hat{\mathbf{A}})\right) && (6.3.40) \end{aligned}$$

$$\begin{aligned} SS(\mathbf{A}^*) - SS(\hat{\mathbf{A}}) &= tr\left(2(\mathbf{A}^*-\hat{\mathbf{A}})'(\mathbf{X}'\mathbf{X}\hat{\mathbf{A}} - \mathbf{X}'\mathbf{Z}) \right. \\ &\quad + \left. (\mathbf{A}^*-\hat{\mathbf{A}})'\mathbf{X}'\mathbf{X}(\mathbf{A}^*-\hat{\mathbf{A}})\right) && (6.3.41) \end{aligned}$$

One has to keep in mind here that $\hat{\mathbf{A}}$ is not the usual least squares estimate, but has the form described by (6.1.5) and (6.1.7). Under the assumption that the weighting matrix Λ is fixed, one obtains the following expression for the difference between $SS(\tilde{\mathbf{A}})$ and $SS(\mathbf{A}^*)$.

$$\begin{aligned}
SS(\tilde{\mathbf{A}}) - SS(\mathbf{A}^*) &= tr\Big\{\tilde{\mathbf{A}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{A}} - \mathbf{A}^{*'}\mathbf{X}'\mathbf{X}\mathbf{A}^* \\
&\quad - 2(\tilde{\mathbf{A}} - \mathbf{A}^*)'\mathbf{X}'\mathbf{Z}\Big\} && (6.3.42) \\
&= tr\Big\{(\tilde{\mathbf{A}} - \mathbf{A}^*)'\mathbf{X}'\mathbf{X}(\tilde{\mathbf{A}} + \mathbf{A}^*) \\
&\quad - 2(\tilde{\mathbf{A}} - \mathbf{A}^*)'\mathbf{X}'\mathbf{Z}\Big\} && (6.3.43) \\
&= tr\Big\{(\tilde{\mathbf{A}} - \mathbf{A}^*)'(\mathbf{X}'\mathbf{X}\tilde{\mathbf{A}} - \mathbf{X}'\mathbf{Z}) \\
&\quad + (\tilde{\mathbf{A}} - \mathbf{A}^*)'(\mathbf{X}'\mathbf{X}\mathbf{A}^* - \mathbf{X}'\mathbf{Z})\Big\} && (6.3.44)
\end{aligned}$$

6.4 Application: AR-Fields

6.4.1 Fixed Smoothing Parameters

The data set Pat.1R is used here as an example to demonstrate the effect of spatial smoothing of autoregressive parameters. The values of the smoothing parameters chosen were $\vec{\lambda}' = (1, 1, 1)'$ and $\vec{\nu}' = (10, 10, 10)'$. Figures 6.1 to 6.3 show the results, and may be compared to Figure 3.1 on page 62. In general, the parameter fields appear smoother, and local features become more clearly visible. More detailed knowledge of the physiological state of the eye under study would allow to judge if these features correspond to certain retinal properties. Unfortunately, such information was not available.

6.4.2 Crossvalidation

The choice of roughness parameters $\vec{\lambda}$ and $\vec{\nu}$ was made by trial and error in the preceding section. When applying generalized crossvalidation on the AR-parameter fields directly, the results differ. Keeping $\nu = 10$ fixed, only $\vec{\lambda}$ was allowed to vary in crossvalidation. Table 6.1 on page 127 gives the estimates obtained, and Figures 6.4 to 6.6 display the results. Since $\vec{\nu}$ is not involved in the GCV score, fixing this was no restriction in this case. For spatiotemporal crossvalidation, recall from Equation (6.1.20) that the optimal choice of $\vec{\nu}$ and $\vec{\lambda}$ only depends on their (elementwise) ratio.

The values for $\vec{\lambda}$ for Pat.1R and Pat.2 are reasonable, although for Pat.2, the smoothing algorithm stops searching for $\lambda_{GCV}(2)$ at the boundary (indicated by the value 8040.751). The same happens for all three parameters for Pat.1L. This problem was not present when calculating $\vec{\lambda}_{opt}$, the parameter obtained from spacetime crossvalidation. The smoothing parameter becomes extremely large for Pat.1L and Pat.3, resulting in a very flat estimate of the smoothed parameter surfaces (see Appendix E). Apparently, the crossvalidation method is not performing well for these data sets, since the original OLS estimates show a high

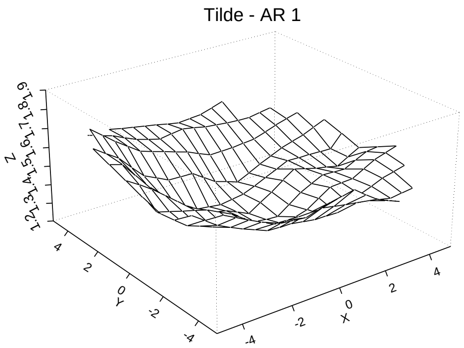

Figure 6.1: *GCV estimates for first AR parameter with* $\lambda = 1$ *and* $\nu = 10$. *Data set Pat.1R.*

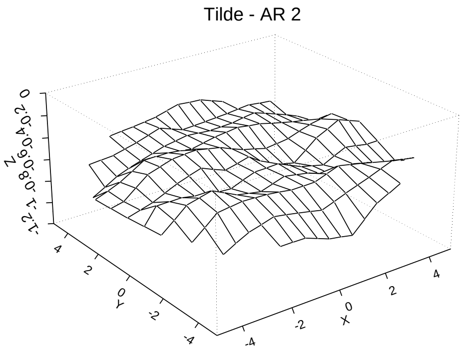

Figure 6.2: *GCV estimates for second AR parameter with* $\lambda = 1$ *and* $\nu = 10$. *Data set Pat.1R.*

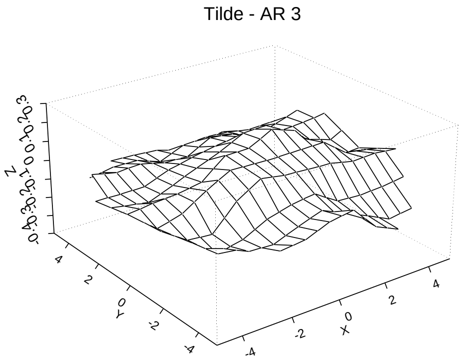

Figure 6.3: *GCV estimates for third AR parameter with* $\lambda = 1$ *and* $\nu = 10$. *Data set Pat.1R.*

Data Set	$\lambda_{GCV}(1)$	$\lambda_{GCV}(2)$	$\lambda_{GCV}(3)$	$\lambda_{opt}(1)$	$\lambda_{opt}(2)$	$\lambda_{opt}(3)$
Pat.1L	8040.751	8040.751	8040.751	100026	286585	116402
Pat.1R	7.011	266.634	7.011	1.565	1.538	1.539
Pat.2	95.963	8040.751	57.570	16.017	16.271	15.202
Pat.3	169.527	238.330	189.661	87906.98	362820.69	103717.46

Table 6.1: *Smoothing parameters resulting from GCV (λ_{GCV}) and STCV (λ_{opt})*

Data Set	Pat.1L	Pat.1R	Pat.2	Pat.3
$SS(OLS)$	210.5569	17.8976	14.8533	649.9345
$SS(GCV)$	212.9712	19.5333	15.1450	660.3875
$SS(OPT)$	213.1098	18.1769	14.9807	661.3359
$SS(GCV)/SS(OLS)$	1.0115	1.0914	1.0196	1.0161
$SS(OPT)/SS(OLS)$	1.0121	1.0156	1.0086	1.0175

Table 6.2: *Relative change in sum of squares for fit*

amount of variability. The residuals from OLS-fit to the data standardized on expected value zero and variance one are ranging between -0.531 and 0.638 for Pat.1L, and between -0.835 and 0.898 for Pat.3. The residuals for Pat.1R (-0.146 to 0.157) and Pat.2 (-0.136 to 0.140) are much smaller, indicating a better overall fit for the latter two by an autoregressive model. Room for improvement in the model is also indicated by the sum of squares after OLS fit which are given in Table 6.2. This confirms results already visible in Table 3.1 (see p.59) which support higher order autoregressive models.

Both after the application of GCV to the AR-parameter fields, and after spatiotemporal cross-validation, the overall fit to the data did not worsen much in the MSE sense. The sum of squares after standard GCV increased between 1 and 2 percent in most cases when compared to the OLS-fit, with the exception of Pat.1R with an increase of 9.1 percent. Spatiotemporal crossvalidation lead to an increase between 0.86 and 1.75 percent only when compared to the OLS fit (Table 6.2). However, the spatiotemporal approach results in the smoothest estimates for the parameter fields. While there is a close fit to the data, spatiotemporal cross-validation apparently involves some loss in local adaptivity. To overcome this critical issue, a modified way of choosing the smoothing parameters $\vec{\lambda}$ and $\vec{\nu}$ may be appropriate. Possible approaches involve a direct choice of the parameters as suggested by empirical experience. Unfortunately, a sufficiently large number of data sets was not available.

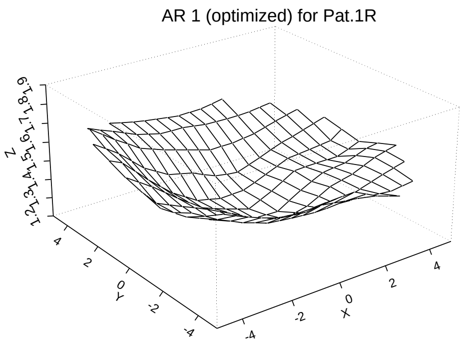

Figure 6.4: *Optimized GCV estimates for first AR parameter. Data set Pat.1R.*

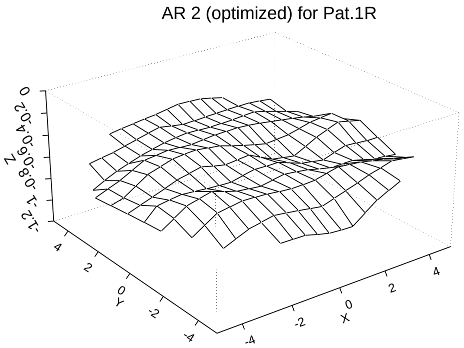

Figure 6.5: *Optimized GCV estimates for second AR parameter. Data set Pat.1R.*

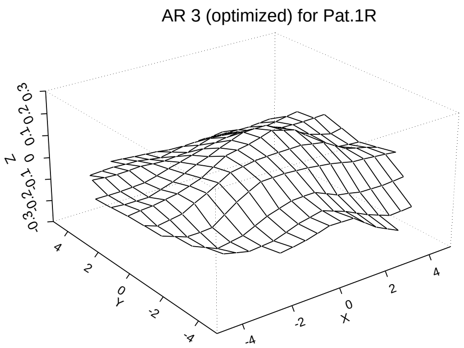

Figure 6.6: *Optimized GCV estimates for third AR parameter. Data set Pat.1R.*

Chapter 7

Summary and Outlook

The analysis of data obtained from the multifocal electroretinogram (MF-ERG) is a relatively new task in ophthalmology. The MF-ERG technique is applied at the University of Essen to patients with age-dependent macula degeneration in order to characterize the dynamics within the patients retina. In a pilot study, four data sets were made available to gain insight into the physiological process under study, and to develop appropriate methods of analysis for this type of data.

Multifocal ERG data sets are spatiotemporal data sets. Most analyses of such data found in medical literature are preceded by some form of data reduction, involving some loss of either spatial or temporal information. For example, often a set of derived amplitudes is examined, conveying only purely spatial information without any temporal component.

In an attempt to improve on this situation, penalized spline smoothing of AR-parameter fields is suggested. This approach combines univariate time series concepts with spline smoothing methodology. Explicit modeling of the covariance structure of the estimated AR-coefficients is avoided. However, smoothing at the parameter level may introduce covariances between residuals of the data and their predicted values.

There is a connection between Kriging methods used in spatial statistics, and spline smoothing. If information is available about the covariance structure of observations, and prediction is of primary interest, Kriging may be preferred to spline smoothing. Since this was not the case in the study at hand, main focus was put on the latter.

As part of the spline smoothing process, certain parameters have to be chosen to determine the degree of smoothness (or roughness) of the estimated parameter field. Smoothness may be measured by means of a linear differential operator. This can be added to a least squares fitting criterion to define an overall roughness penalty. The penalty indirectly determines the class of functions containing the possible solutions. In the ERG case, a penalty based on derivatives up to second order was chosen, resulting in a fit within the class of so-called thin plate splines. However, other penalties are equally valid. They may be used if there is more detailed a priori information available on the shape of the underlying function.

When performing smoothing, a parameter vector $\vec{\lambda}$ has to be chosen which controls the degree of smoothness of the fit. In the ERG application the components of $\vec{\lambda}$ were selected by a modified version of Generalized Crossvalidation. Since the ordinary least squares estimates of the underlying autoregressive parameters can be regarded as optimal, they were used as a starting point when defining a spatiotemporal penalized sum of squares. This penalty is a sum of the roughness measure for the parameters, plus the corresponding sum of squares for fit to the observations. Weighting can be done between these components via an additional parameter vector $\vec{\nu}$. For fixed $\vec{\lambda}$ and $\vec{\nu}$, a closed form estimator was derived. The smoothing procedure leads to an optimal fit to the data while yielding a smooth surface of autoregressive parameter estimates. An expression for the residuals between the smooth fit and the common OLS estimator was derived, and so was a formula for the change in sum of squares.

The empirical results found are twofold. Only four data sets were available. Two of these (Pat.1L and Pat.3) were not modeled well enough by an autoregressive process to allow for efficient spatiotemporal smoothing of the corresponding AR-parameter fields. The estimated parameter fields appeared to be over-smoothed in these two cases. However, it has to be noted that after smoothing, the sum of squared residuals from fit increased only slightly (about 1-2 percent) when compared to the OLS fit.

An increase of the same small magnitude was also found for the other two data sets, Pat.1R and Pat.2. In particular, data set Pat.1R was fit very well after spatiotemporal crossvalidation: Local features not clearly visible in the original data were found both after Generalized Crossvalidation of the AR-parameter fields, and spatiotemporal crossvalidation. Unfortunately, due to a lack of detailed knowledge on the physiological state of the eyes under study it was not possible to confirm if the detected local features actually correspond to physiological structures.

The sum of squares for fit increased by more than 9 percent after generalized crossvalidation, while spatiotemporal crossvalidation resulted in an increase of only 1.5 percent. A similar effect was also observed for Pat.2, where the sum of squares increased by 1.9 percent with GCV, but only 0.9 percent with spatiotemporal crossvalidation. In this regard, the proposed new method is preferable to the GCV approach.

These empirical results suggest that smoothing of AR-parameter fields is a promising approach to quantify spatiotemporal dynamics in multifocal ERG data. It can be extended to other applications by defining various kinds of derivatives for the roughness criterion to favor certain functional classes. Spline smoothing can also be seen from a Bayesian point of view (Wahba 1978), allowing for input of external information into the estimation process. This is of particular interest once a larger number of data sets of similar structure becomes available.

The choice of smoothing parameters is an issue not yet adequately solved. Spatiotemporal crossvalidation proofed to be very time-consuming, and the results were not very satisfactory for some of the data sets under study. A possible reason is the high variability in the least squares autoregressive parameter estimates used as pseudo-observations, leading to some degree of over-smoothing. This phenomenon should be studied further before smoothing of autoregressive parameter fields is applied to a large number of MF-ERG data sets. If high variability in the parameters is common, spatiotemporal crossvalidation certainly is not a good approach. Instead, determining the smoothness of the AR-parameter fields by allowing for a certain increase in the sum of squares for fit to the data may give reasonable results. This was the case for data set Pat.1R which gave a well localized smooth fit in the parameter domain, but resulted in the highest increase in sum of squares for temporal fit. From an empirical point of view, an algorithm seems to be desirable which allows to directly specify the admissible level of increase in sum of squares, and to derive the corresponding choices for the smoothing parameters $\vec{\nu}$ and $\vec{\lambda}$.

In terms of further theoretical developments, the properties of smoothed AR-parameter fields need to be studied further. Expected values and bias need to be evaluated to judge the results of the estimation process. A major problem is that the underlying AR-parameters themselves are autoregressive estimates. This makes evaluation of their expected values difficult.

Appendix A

Spatial Variation in Amplitudes

It is quite common that amplitudes of multifocal ERG data are analyzed, instead of complete data sets. Amplitudes can be deduced for each of the 103 hexagonal areas of the retina. Since they are spatially ordered, they may be analyzed using methods from spatial statistics. Empirical findings suggest that grouping of areas should be done be eccentricity from the retinal center. Plots of amplitudes versus distance from center, and versus angle towards the horizon were used to check if this is the case for the empirical data sets under study.

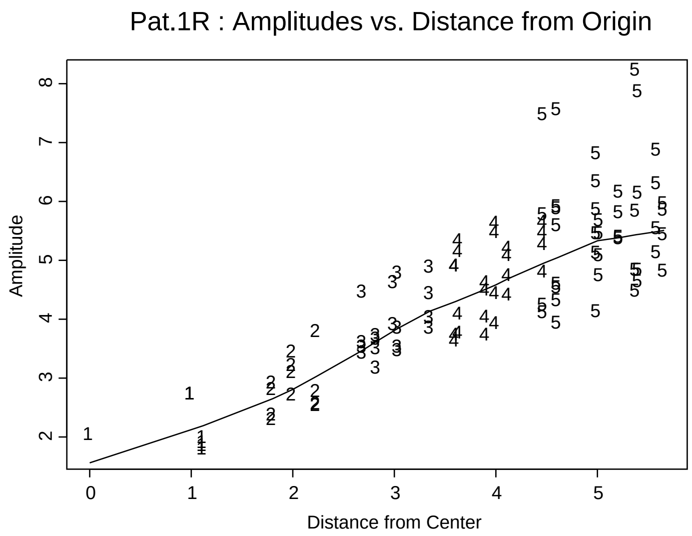

Figure A.1: Plot of amplitudes versus distance for data set Pat.1R. A lowess-curve indicates how the ERG varies with eccentricity. Plotting symbol reflects group number.

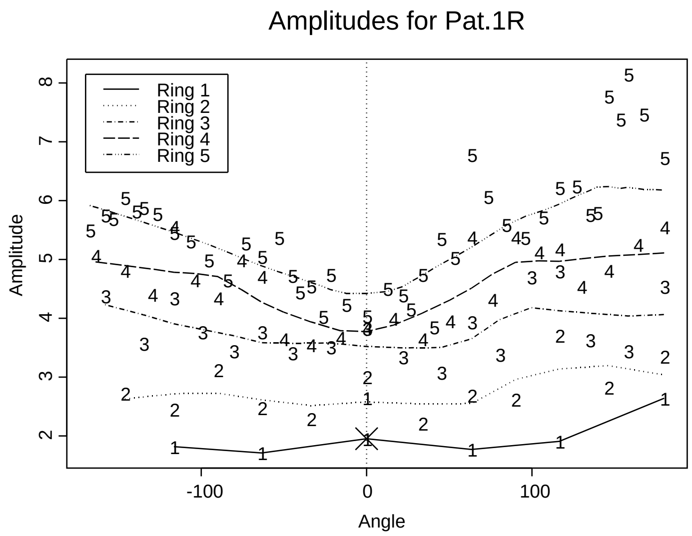

Figure A.2: Plot of amplitudes versus angle. Lowess-curves for each ring indicate how the ERG varies with eccentricity.

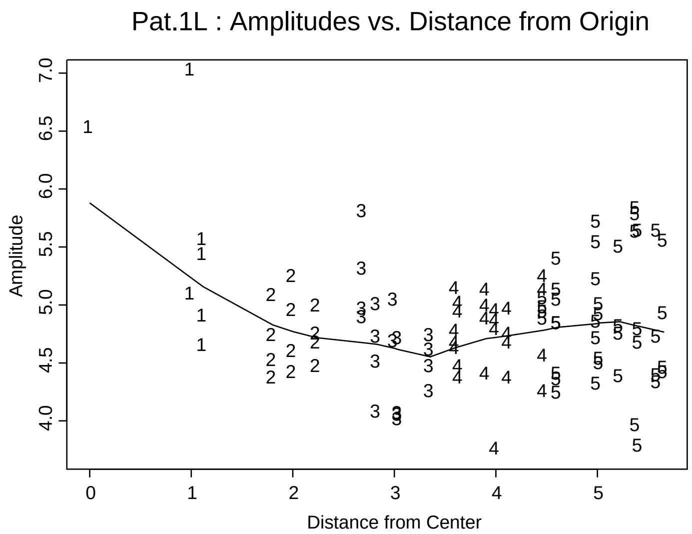

Figure A.3: Plot of amplitudes versus distance for data set Pat.1L. A lowess-curve indicates how the ERG varies with eccentricity. Plotting symbol reflects group number.

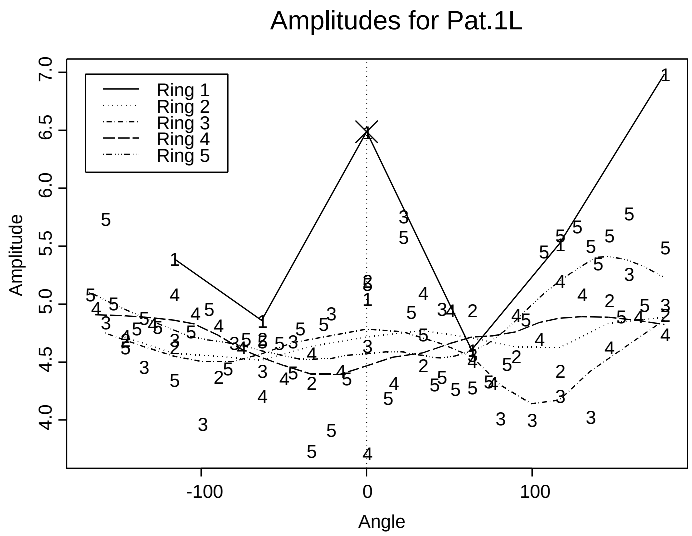

Figure A.4: Plot of amplitudes versus angle. Lowess-curves for each ring indicate how the ERG varies with eccentricity.

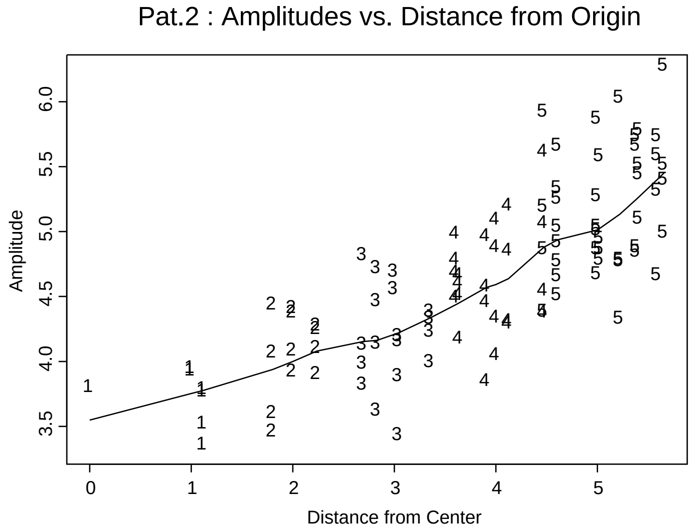

Figure A.5: Plot of amplitudes versus distance for data set Pat.2. A lowess-curve indicates how the ERG varies with eccentricity. Plotting symbol reflects group number.

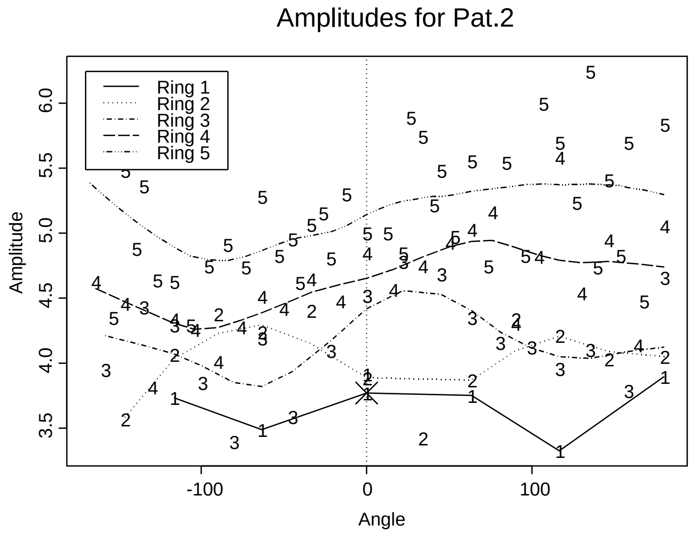

Figure A.6: Plot of amplitudes versus angle. Lowess-curves for each ring indicate how the ERG varies with eccentricity.

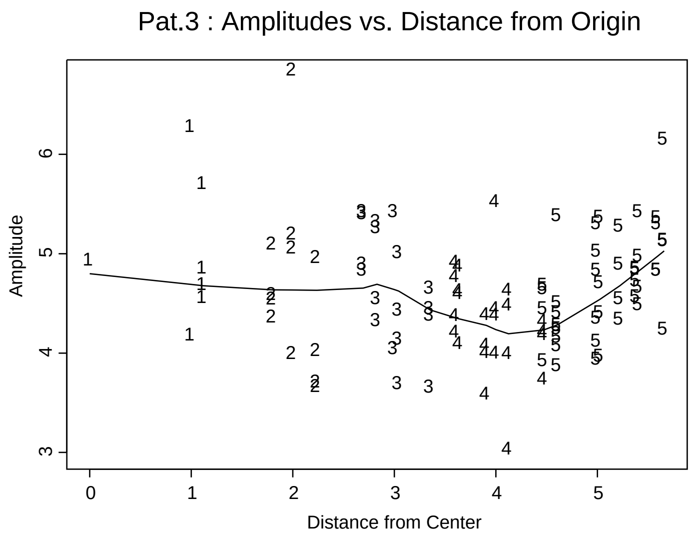

Figure A.7: Plot of amplitudes versus distance for data set Pat.3. A lowess-curve indicates how the ERG varies with eccentricity. Plotting symbol reflects group number.

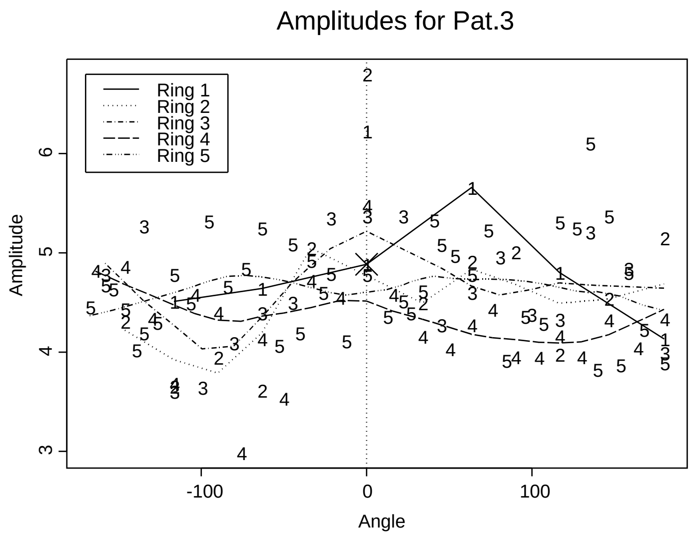

Figure A.8: Plot of amplitudes versus angle. Lowess-curves for each ring indicate how the ERG varies with eccentricity.

Appendix B

Spatiotemporal Wireframes

Figures B.1 to B.3 each show six spatial data sets obtained from two different eyes at times 5, 20, 35, 50, 65 and 80 milliseconds. They correspond to Figure 2.15 presented earlier in this document. It can be seen from the graphics that the responses vary both over time and space, suggesting a spatiotemporal approach to data analysis.

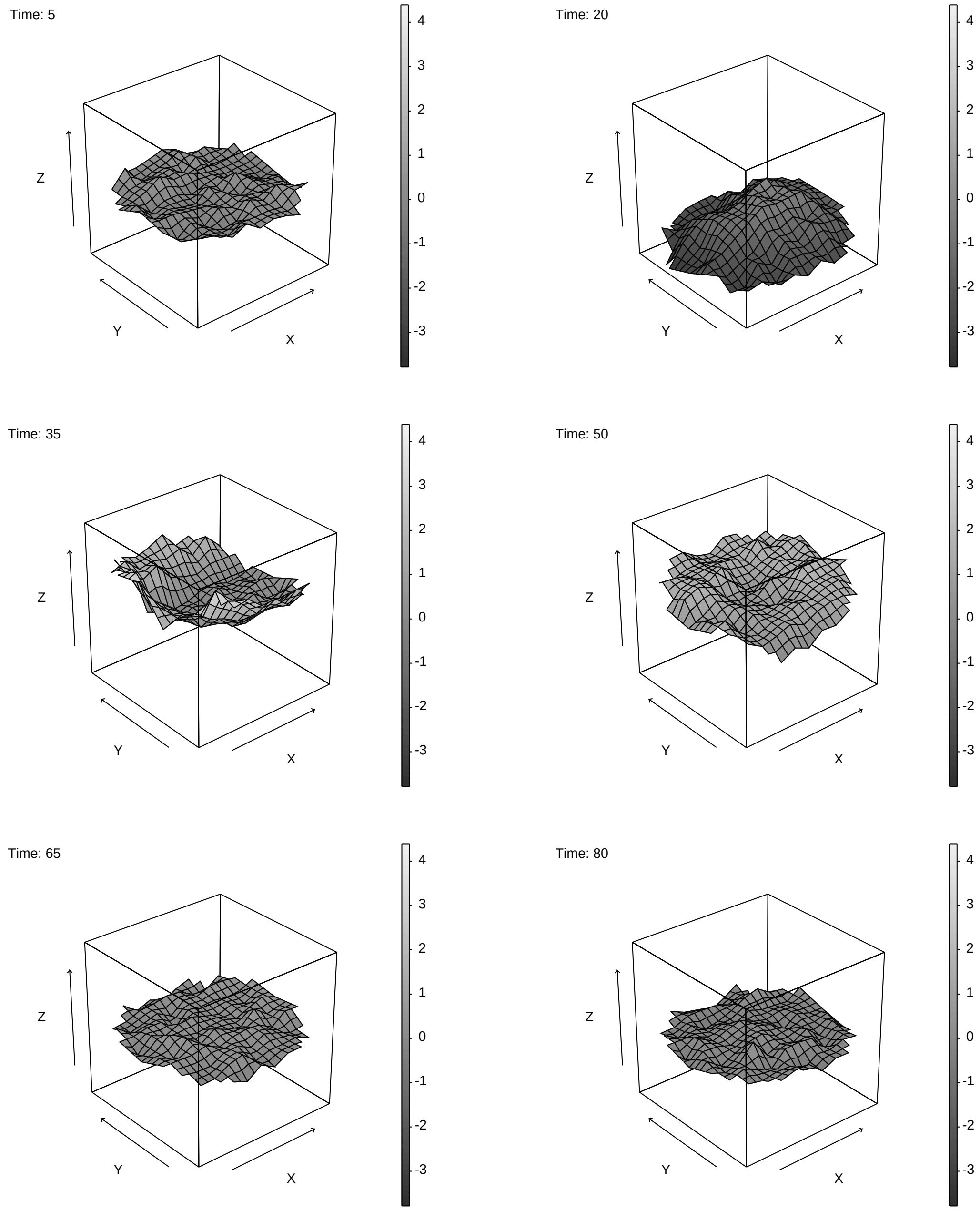

Figure B.1: Patient Pat.1R. Measurements at times 5, 20, 35, 50, 65 and 80 milliseconds. Values were linearly interpolated for graphical display.

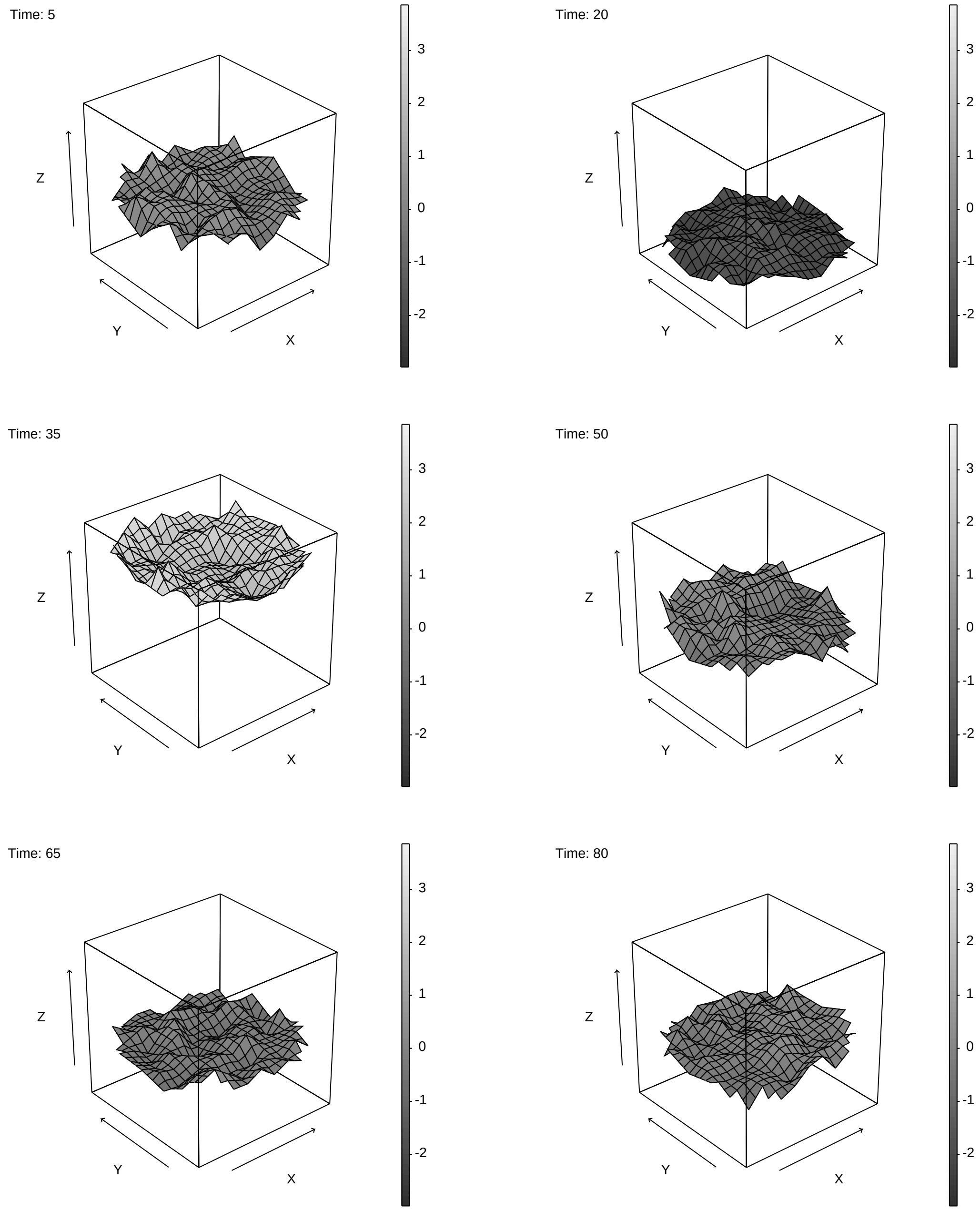

Figure B.2: Patient Pat.2. Measurements at times 5, 20, 35, 50, 65 and 80 milliseconds. Values were linearly interpolated for graphical display.

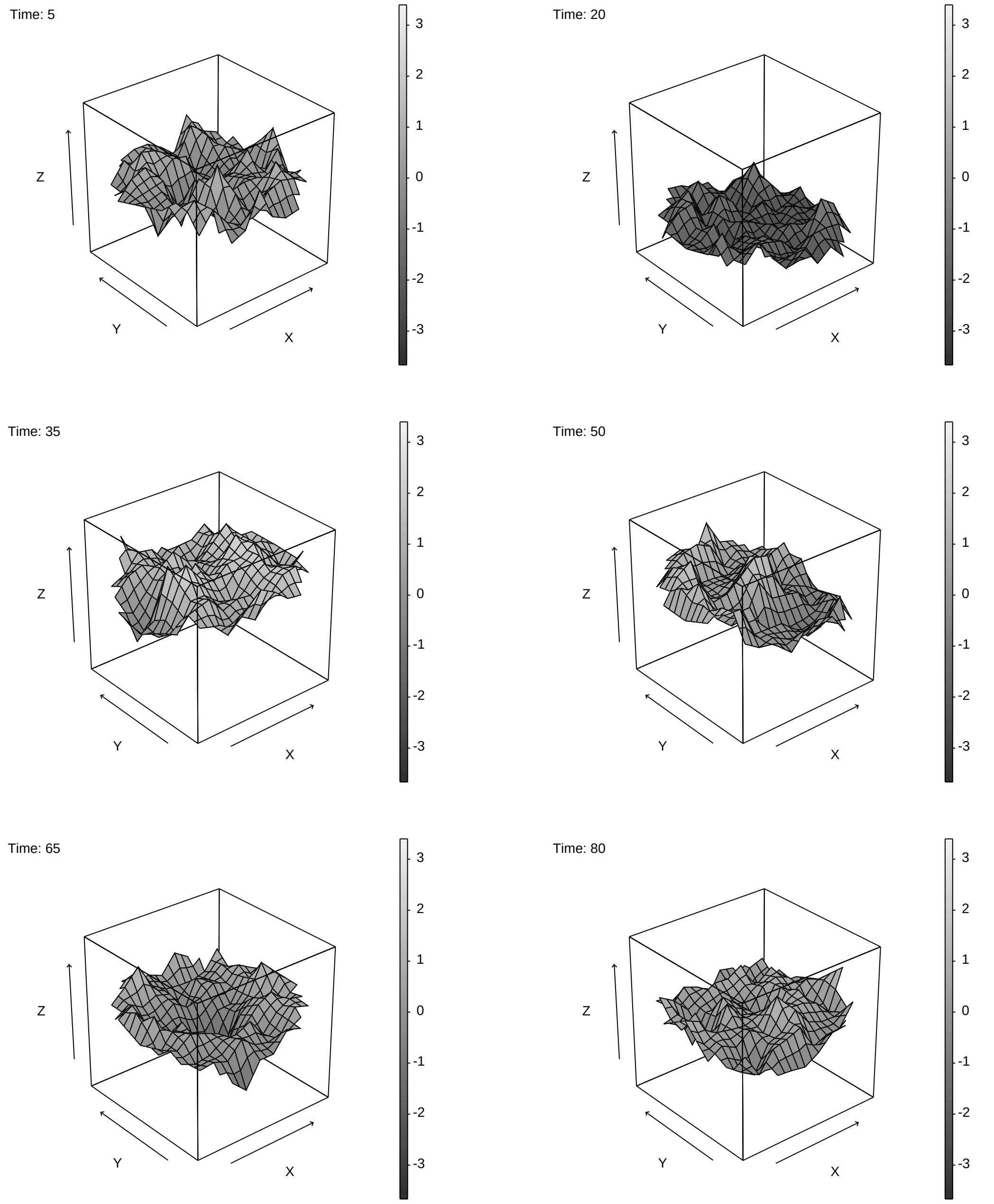

Figure B.3: Patient Pat.3. Measurements at times 5, 20, 35, 50, 65 and 80 milliseconds. Values were linearly interpolated for graphical display.

Appendix C

Exploratory Polynomial Fit

In an exploratory analysis of the spatial trend in the ERG data available, polynomial trend surfaces with components up to order 3 where fit to the data using ordinary least squares techniques. This was done for each point in time The estimated coefficients are displayed below as time series, with one plot for each coefficient.

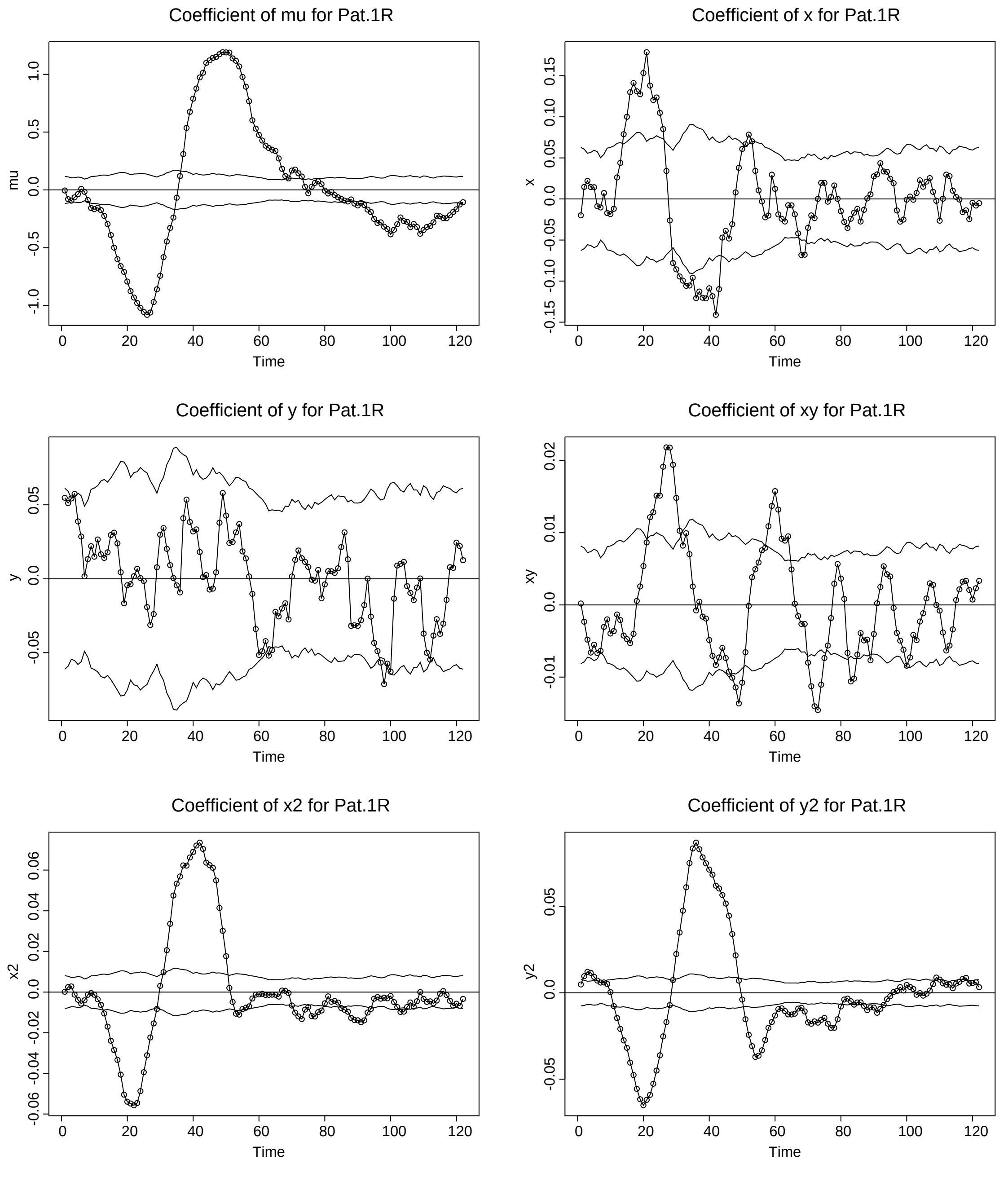

Figure C.1: Patient Pat.1R. Coefficients for intercept, x, y, $x * y$, x^2 and y^2. Pointwise 95 percent reference intervals under assumption of independence.

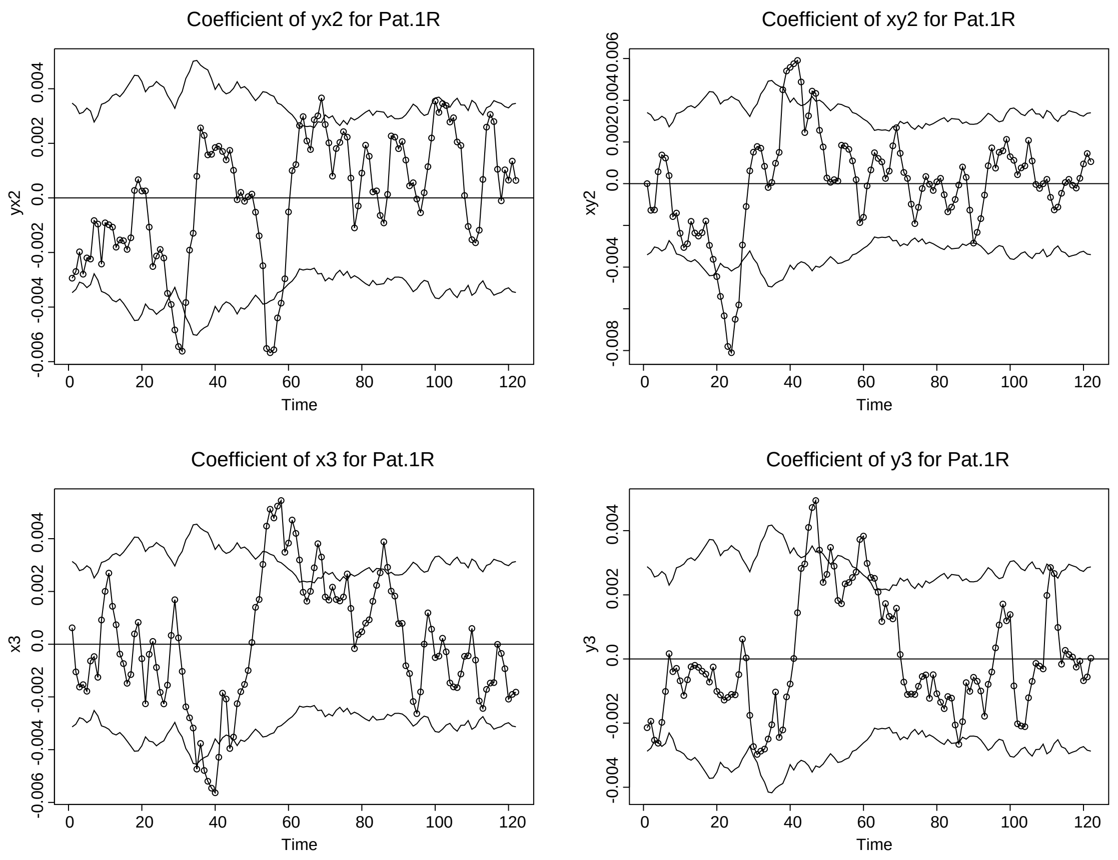

Figure C.2: Patient Pat.1R. Coefficients for intercept, yx^2, xy^2, x^3 and y^3. Pointwise 95 percent reference intervals under assumption of independence.

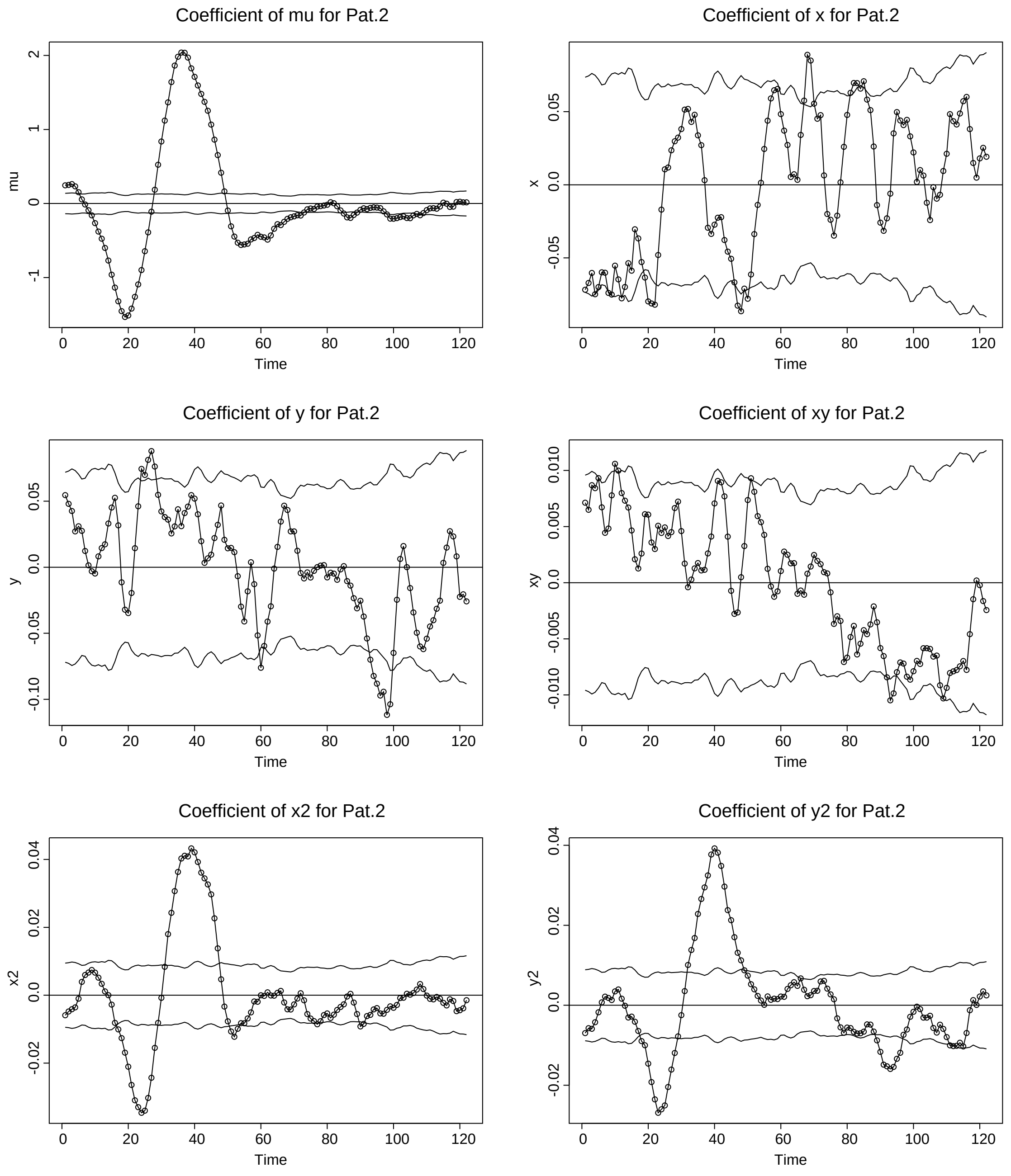

Figure C.3: Patient Pat.2. Coefficients for intercept, x, y, $x * y$, x^2 and y^2. Pointwise 95 percent reference intervals under assumption of independence.

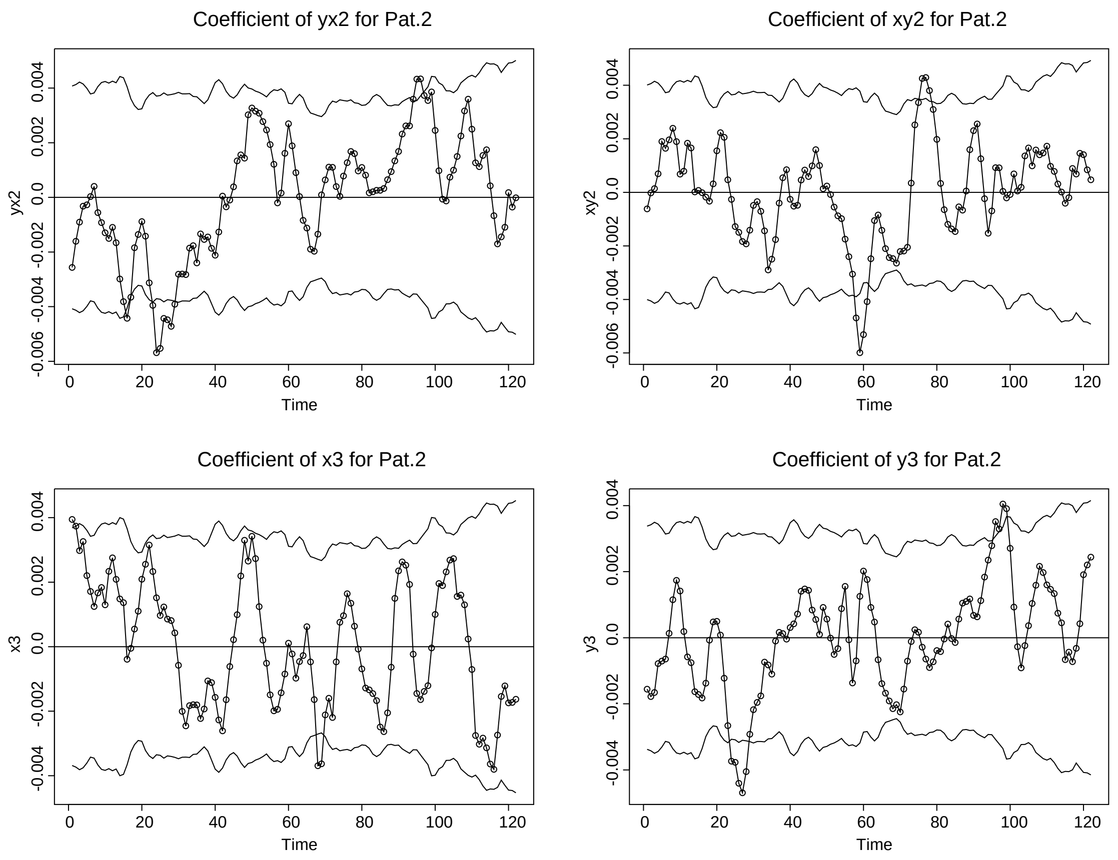

Figure C.4: Patient Pat.2. Coefficients for intercept, yx^2, xy^2, x^3 and y^3. Pointwise 95 percent reference intervals under assumption of independence.

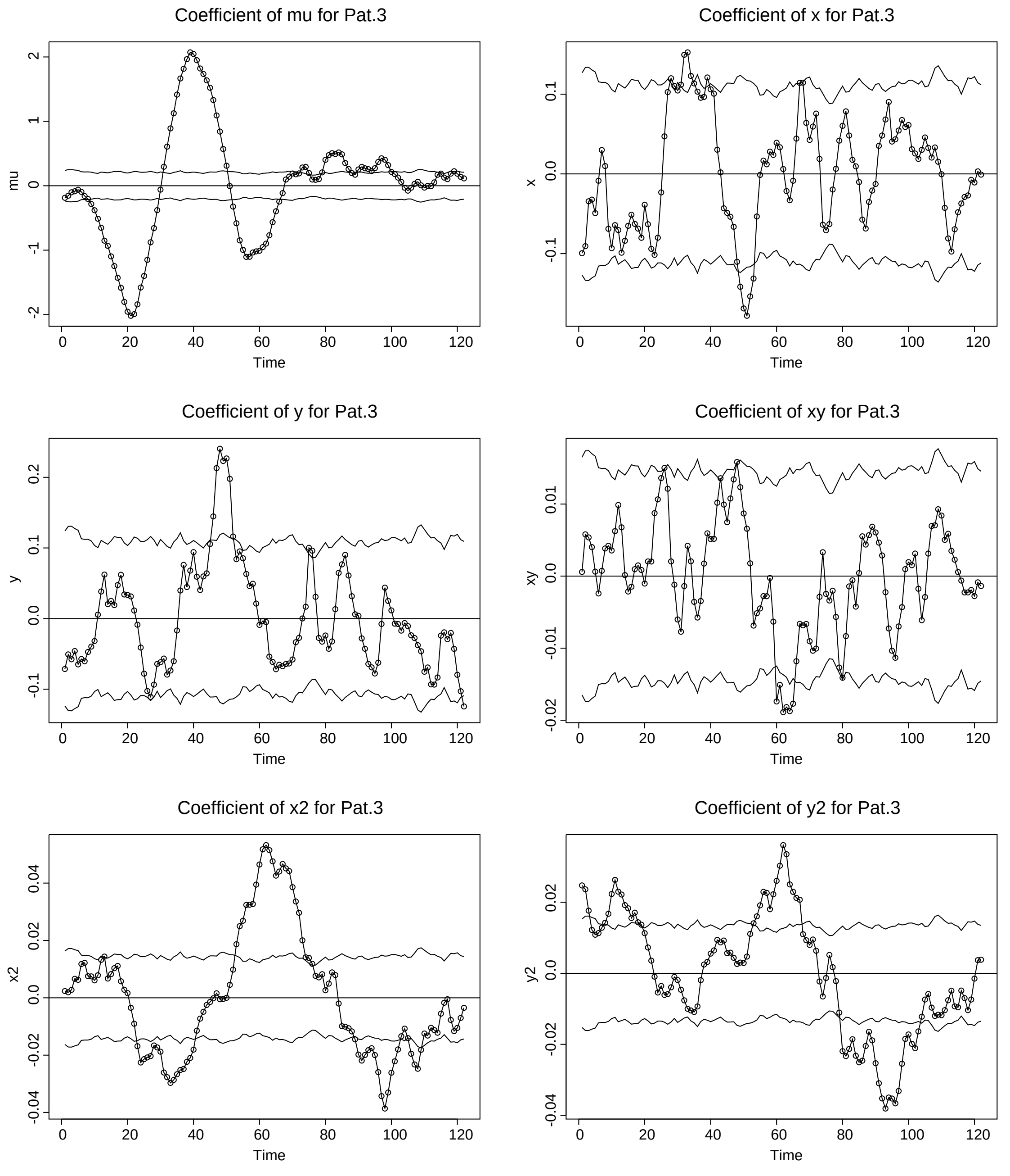

Figure C.5: Patient Pat.3. Coefficients for intercept, x, y, $x * y$, x^2 and y^2. Pointwise 95 percent reference intervals under assumption of independence.

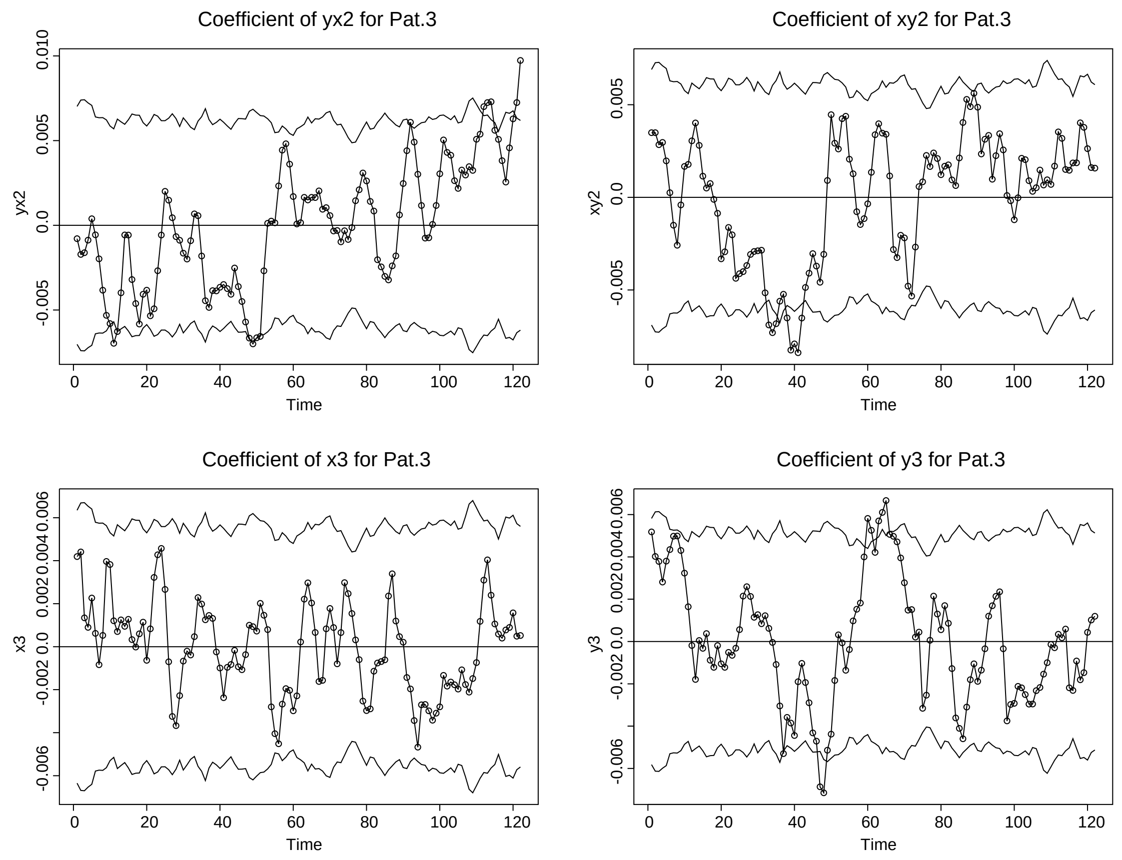

Figure C.6: Patient Pat.3. Coefficients for intercept, yx^2, xy^2, x^3 and y^3. Pointwise 95 percent reference intervals under assumption of independence.

Appendix D

Universal Kriging Results

Universal Kriging of ERG-amplitudes was used as an example to point to strengths and weaknesses of the Kriging approach. Results for patients Pat.1L, Pat.2 and Pat.3 are given here. In particular, the estimated variogram is given together with perspective plots of the estimated trend, residuals, and predictions from Universal Kriging.

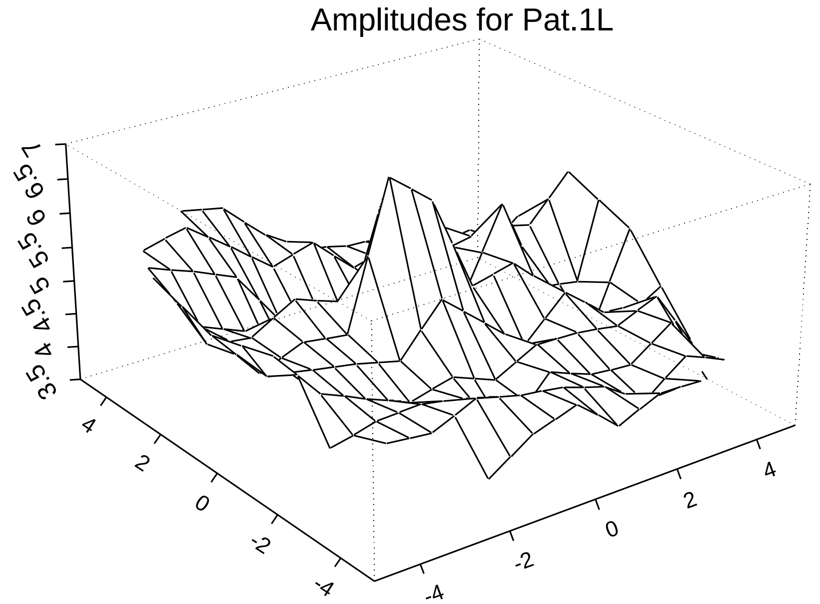

Figure D.1: Amplitudes for data set Pat.1L. Values are linearly interpolated for better graphical presentation.

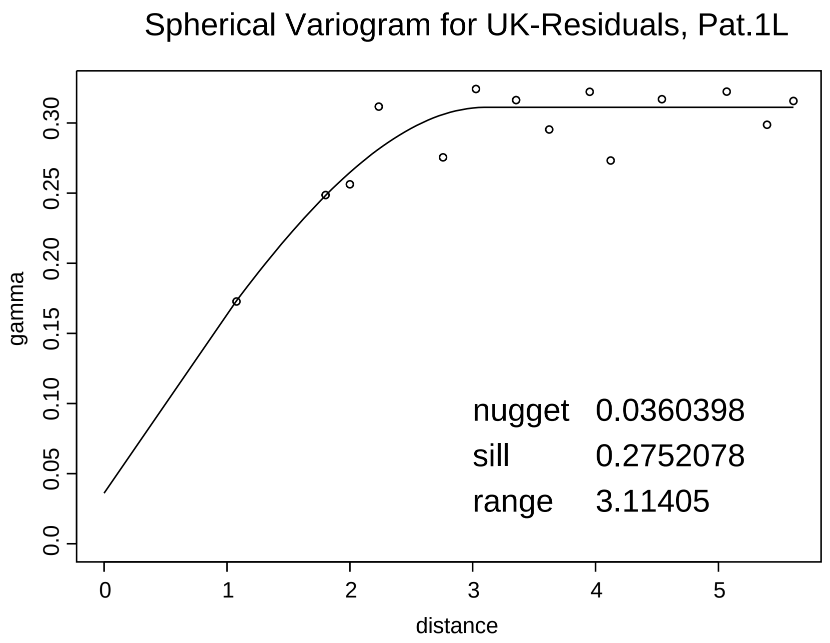

Figure D.2: Estimated spherical variogram from UK residuals to Pat.1L.

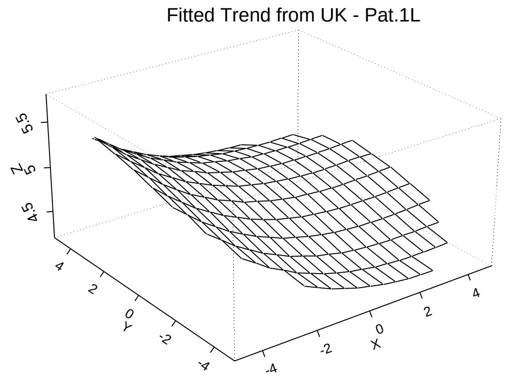

Figure D.3: Estimated trend from UK for amplitudes in data set Pat.1L.

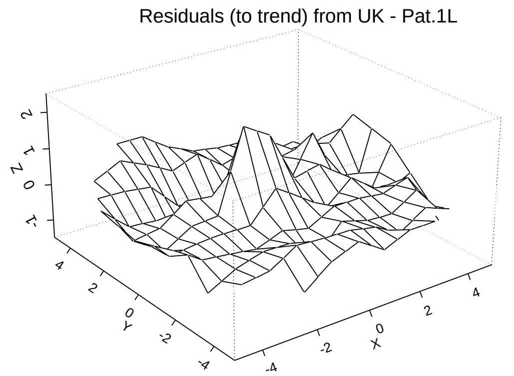

Figure D.4: Residuals after UK for amplitudes in data set Pat.1L.

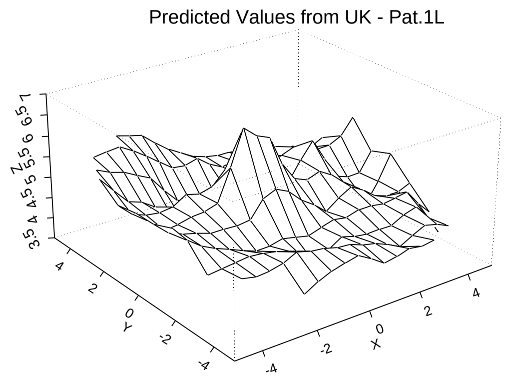

Figure D.5: Predictions obtained from UK for amplitudes in data set Pat.1L.

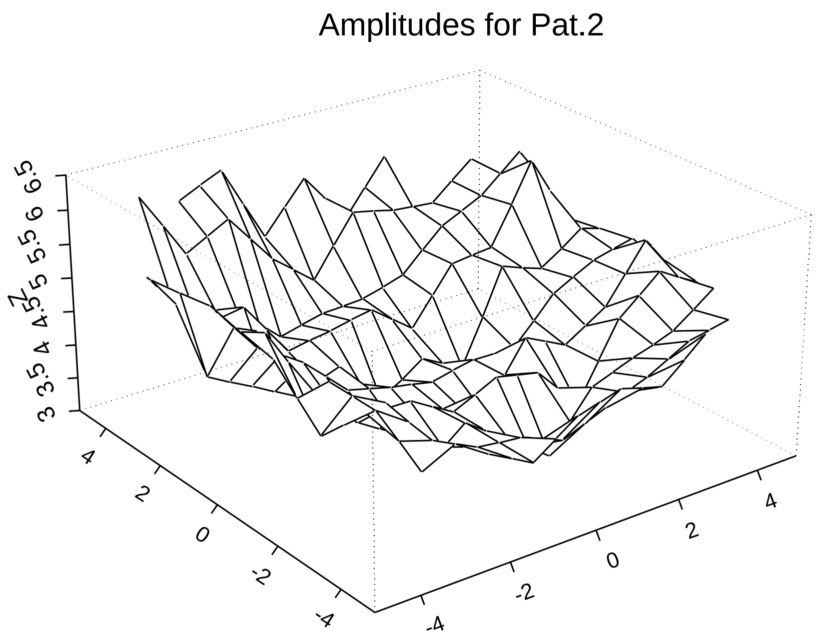

Figure D.6: Amplitudes for data set Pat.2. Values are linearly interpolated for better graphical presentation.

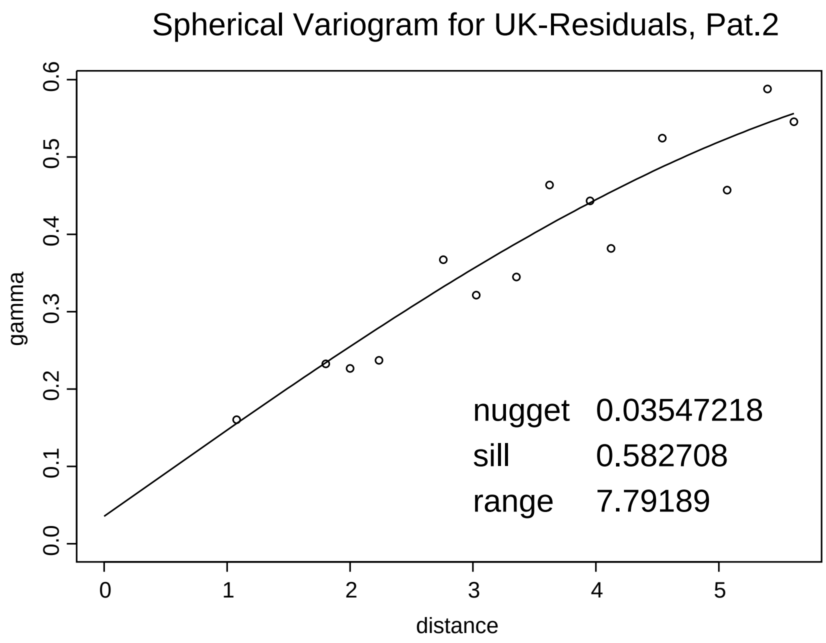

Figure D.7: Estimated spherical variogram from UK residuals to Pat.2.

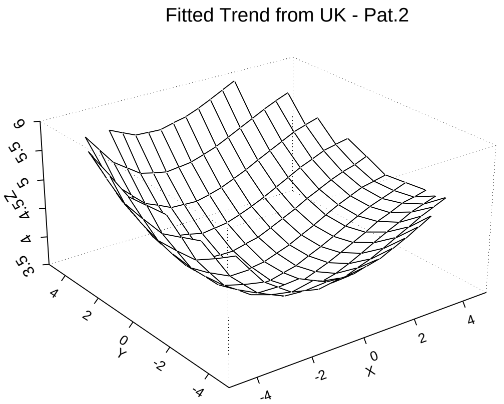

Figure D.8: Estimated trend from UK for amplitudes in data set Pat.2.

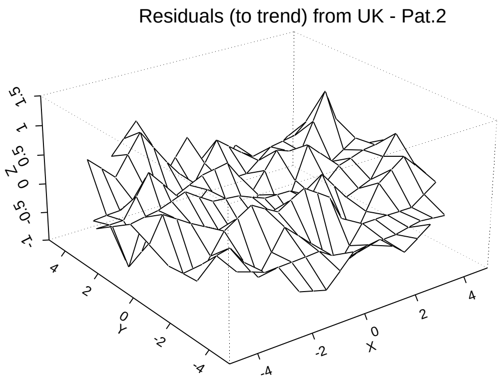

Figure D.9: Residuals after UK for amplitudes in data set Pat.2.

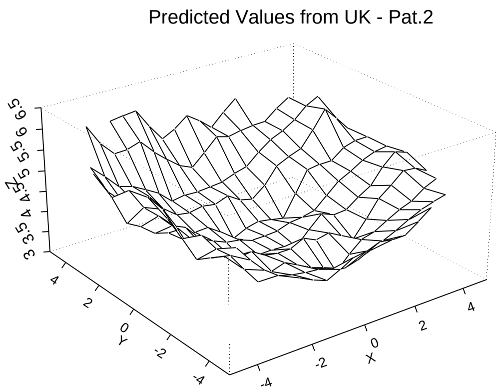

Figure D.10: Predictions obtained from UK for amplitudes in data set Pat.2.

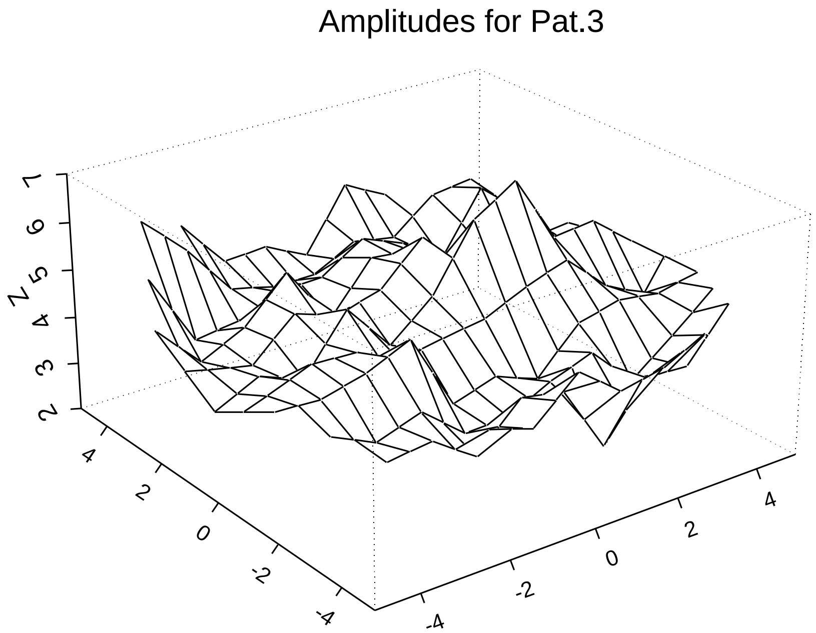

Figure D.11: Amplitudes for data set Pat.3. Values are linearly interpolated for better graphical presentation.

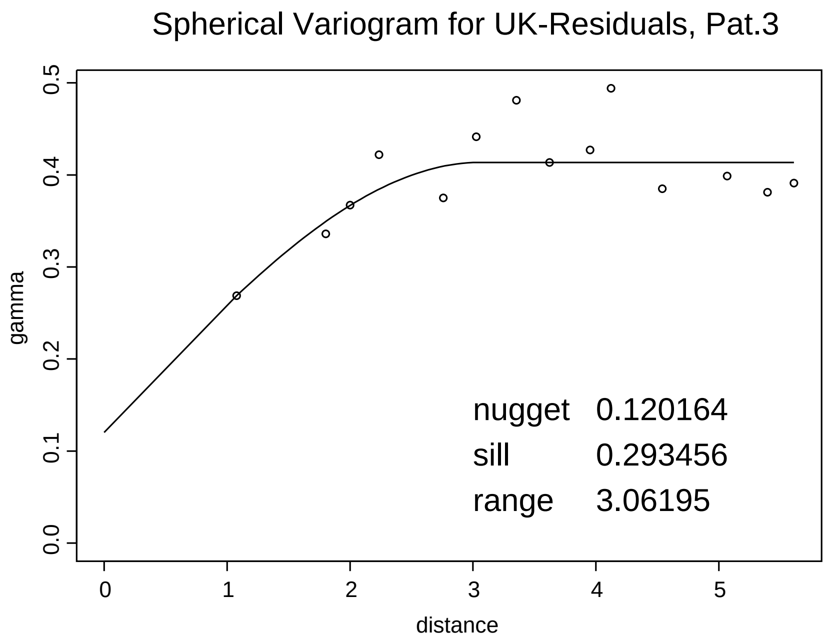

Figure D.12: Estimated spherical variogram from UK residuals to Pat.3.

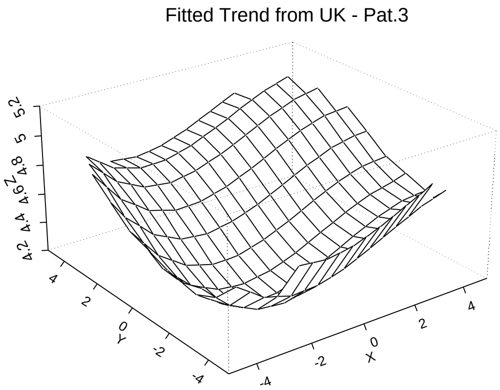

Figure D.13: Estimated trend from UK for amplitudes in data set Pat.3.

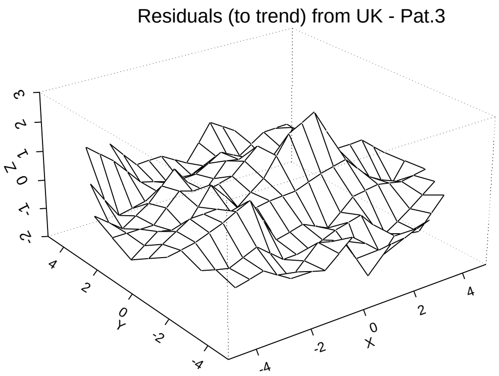

Figure D.14: Residuals after UK for amplitudes in data set Pat.3.

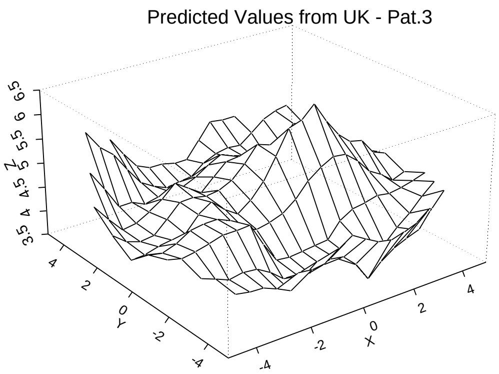

Figure D.15: Predictions obtained from UK for amplitudes in data set Pat.3.

Appendix E

Smoothed AR-Parameter Fields

In this part of the appendix, the OLS estimates of the AR-parameter fields are displayed together with their smoothed versions after Generalized Crossvalidation and Spatiotemporal Crossvalidation. The smoothing parameters $\vec{\lambda}$ used are given in the footnotes, but can also be found in Table 6.1 on page 127. The elements of $\vec{\nu}$ were set to 10 were needed.

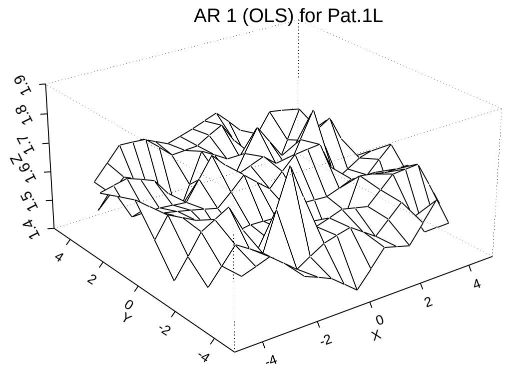

Figure E.1: Least squares estimates for first AR-parameter for data set Pat.1L.

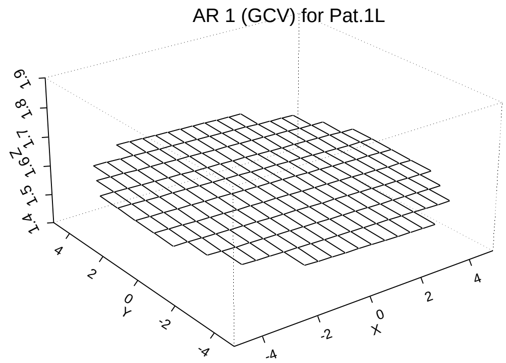

Figure E.2: Smoothed first AR-field for data set Pat.1L. Choice of $\lambda_{GCV}(1) = 8040.751$ by GCV.

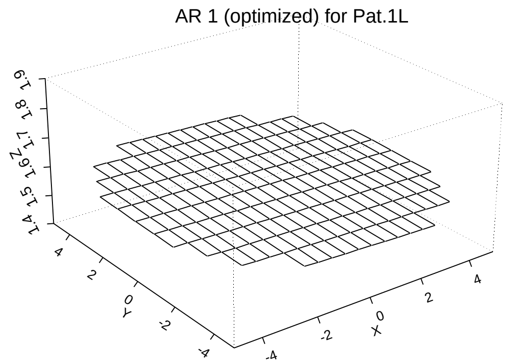

Figure E.3: Smoothed first AR-field after spatiotemporal crossvalidation for data set Pat.1L with $\lambda_{opt}(1) = 100026$.

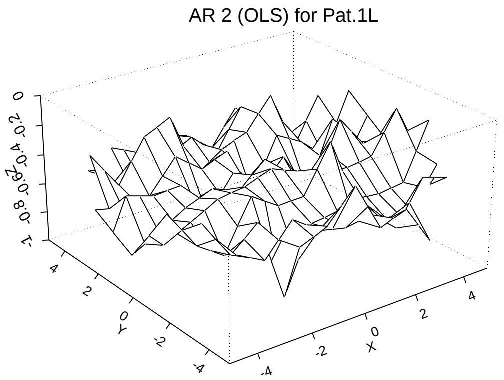

Figure E.4: Least squares estimates for second AR-parameter for data set Pat.1L.

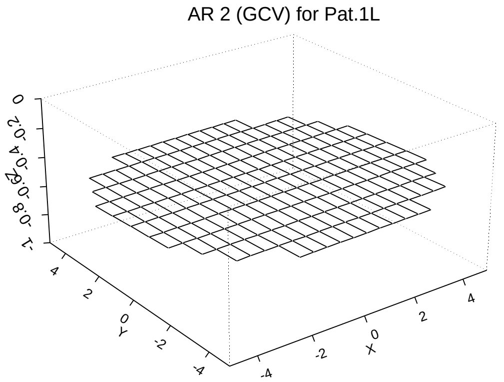

Figure E.5: Smoothed second AR-field for data set Pat.1L. Choice of $\lambda_{GCV}(2) = 8040.751$ by GCV.

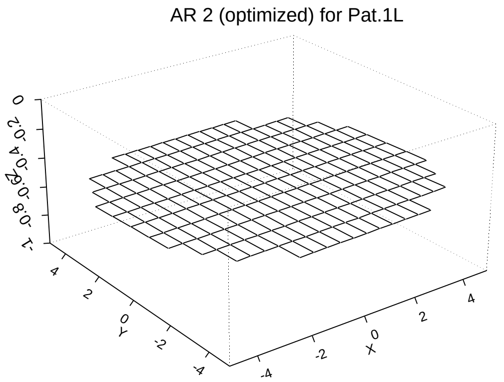

Figure E.6: Smoothed second AR-field after spatiotemporal crossvalidation for data set Pat.1L with $\lambda_{opt}(2) = 286585$.

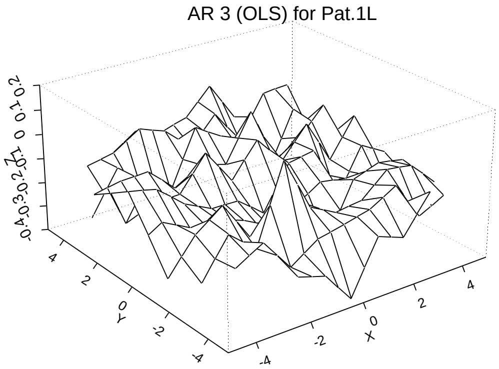

Figure E.7: Least squares estimates for third AR-parameter for data set Pat.1L.

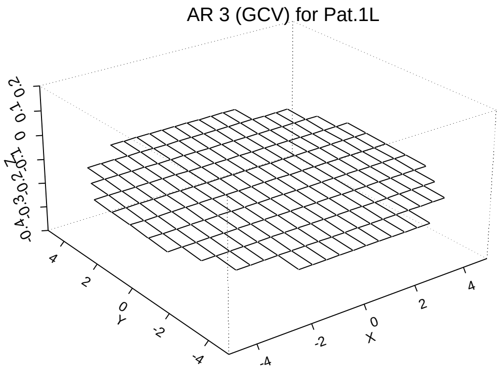

Figure E.8: Smoothed third AR-field for data set Pat.1L. Choice of $\lambda_{GCV}(3)$ = 8040.751 by GCV.

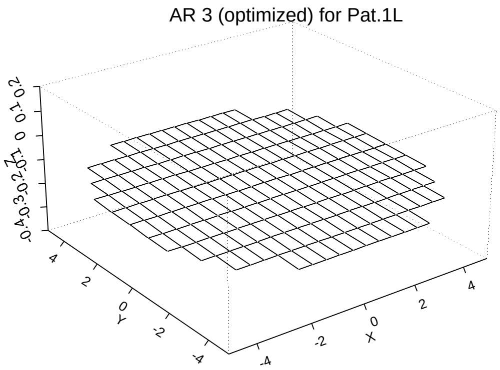

Figure E.9: Smoothed third AR-field after spatiotemporal crossvalidation for data set Pat.1L with $\lambda_{opt}(3)$ = 116402.

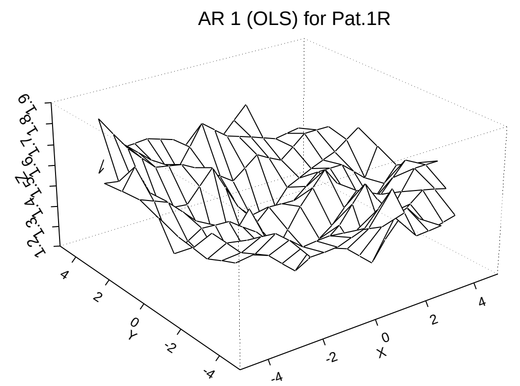

Figure E.10: Least squares estimates for first AR-parameter for data set Pat.1R.

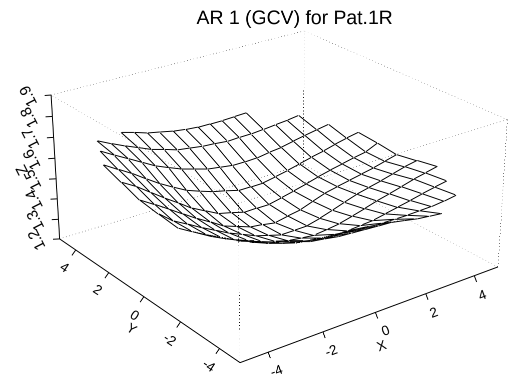

Figure E.11: Smoothed first AR-field for data set Pat.1R. Choice of $\lambda_{GCV}(1) = 7.011$ by GCV.

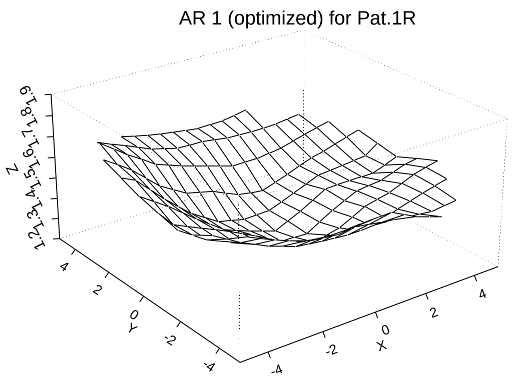

Figure E.12: Smoothed first AR-field after spatiotemporal crossvalidation for data set Pat.1R with $\lambda_{opt}(1) = 1.565$.

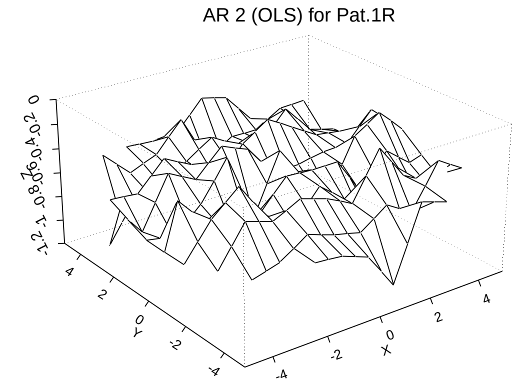

Figure E.13: Least squares estimates for second AR-parameter for data set Pat.1R.

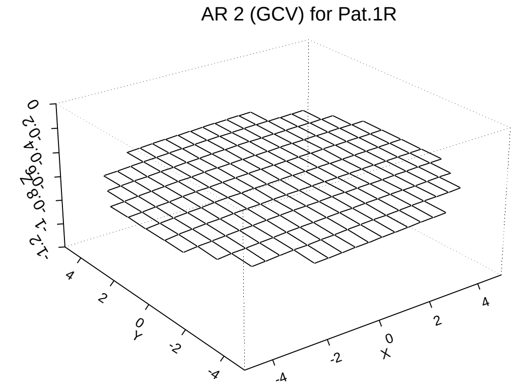

Figure E.14: Smoothed second AR-field for data set Pat.1R. Choice of $\lambda_{GCV}(2) = 266.634$ by GCV.

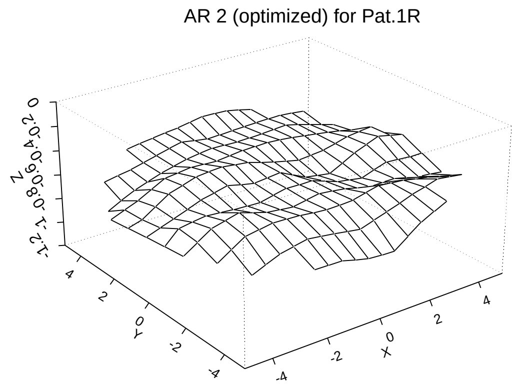

Figure E.15: Smoothed second AR-field after spatiotemporal crossvalidation for data set Pat.1R with $\lambda_{opt}(2) = 1.538$.

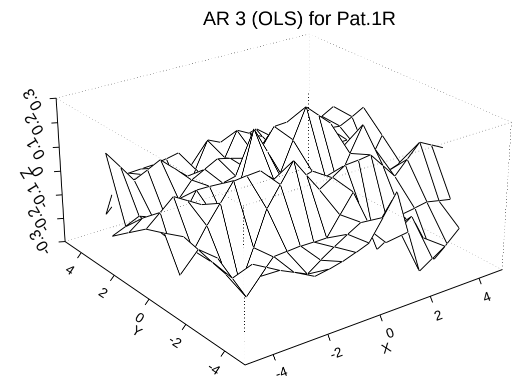

Figure E.16: Least squares estimates for third AR-parameter for data set Pat.1R.

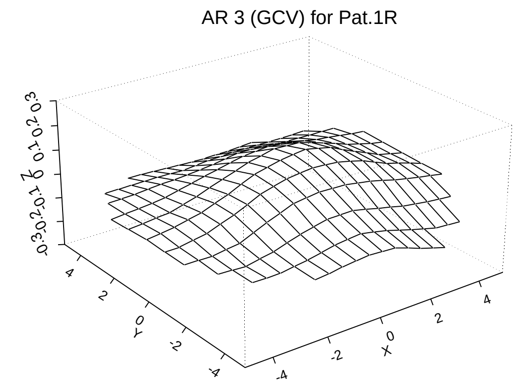

Figure E.17: Smoothed third AR-field for data set Pat.1R. Choice of $\lambda_{GCV}(3) = 7.011$ by GCV.

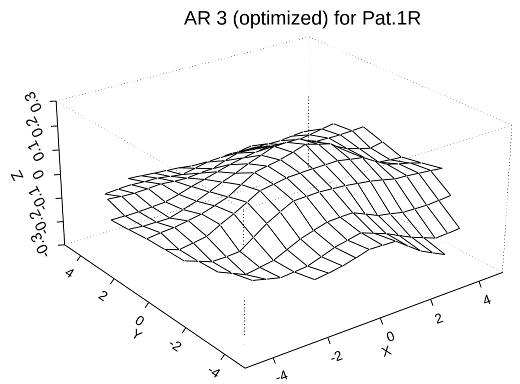

Figure E.18: Smoothed third AR-field after spatiotemporal crossvalidation for data set Pat.1R with $\lambda_{opt}(3) = 1.539$.

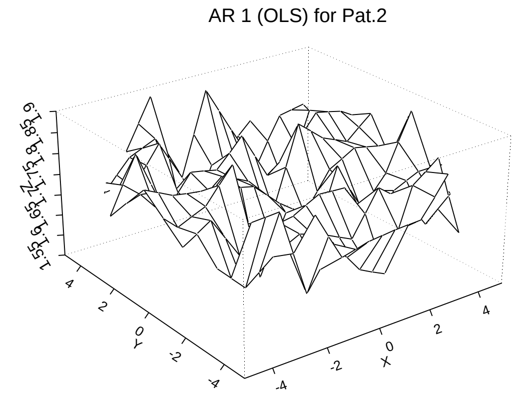

Figure E.19: Least squares estimates for first AR-parameter for data set Pat.2.

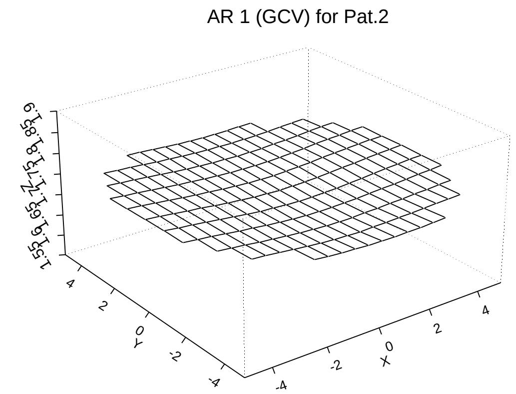

Figure E.20: Smoothed first AR-field for data set Pat.2. Choice of $\lambda_{GCV}(1) = 95.963$ by GCV.

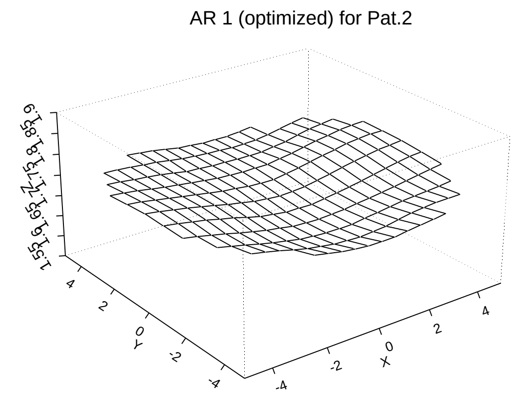

Figure E.21: Smoothed first AR-field after spatiotemporal crossvalidation for data set Pat.2 with $\lambda_{opt}(1) = 16.017$.

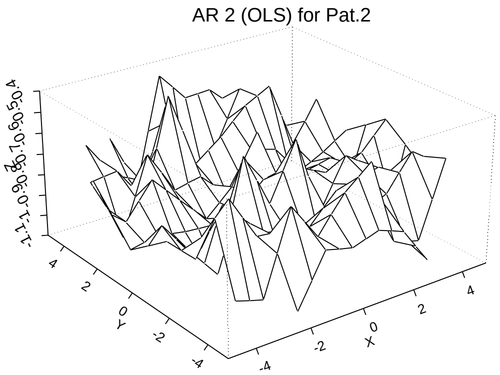

Figure E.22: Least squares estimates for second AR-parameter for data set Pat.1L.

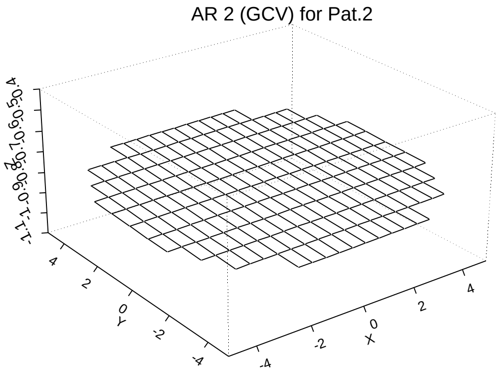

Figure E.23: Smoothed second AR-field for data set Pat.2. Choice of $\lambda_{GCV}(2) = 8040.751$ by GCV.

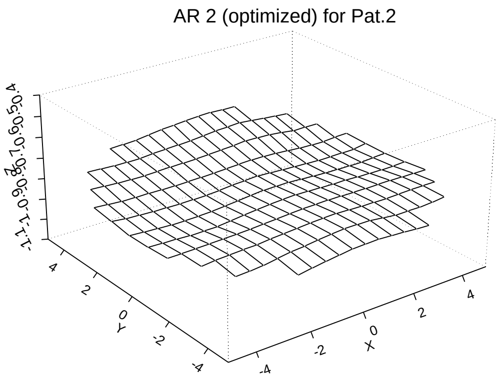

Figure E.24: Smoothed second AR-field after spatiotemporal crossvalidation for data set Pat.2 with $\lambda_{opt}(2) = 16.271$.

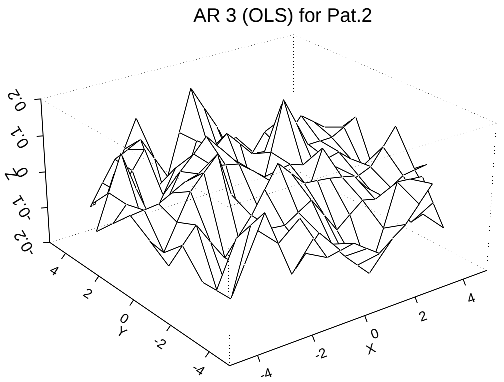

Figure E.25: Least squares estimates for third AR-parameter for data set Pat.2.

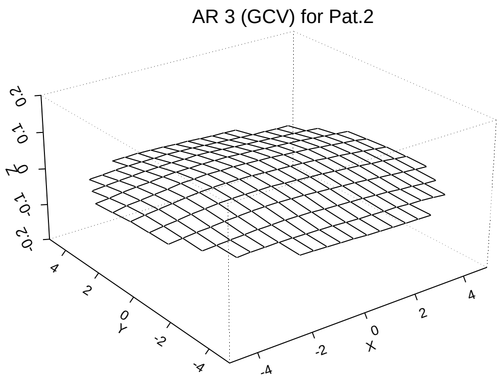

Figure E.26: Smoothed third AR-field for data set Pat.2. Choice of $\lambda_{GCV}(3) = 57.570$ by GCV.

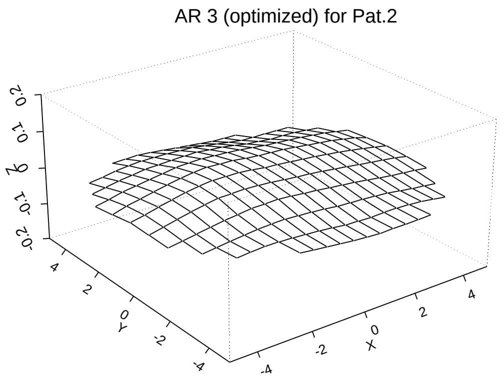

Figure E.27: Smoothed third AR-field after spatiotemporal crossvalidation for data set Pat.2 with $\lambda_{opt}(3) = 15.202$.

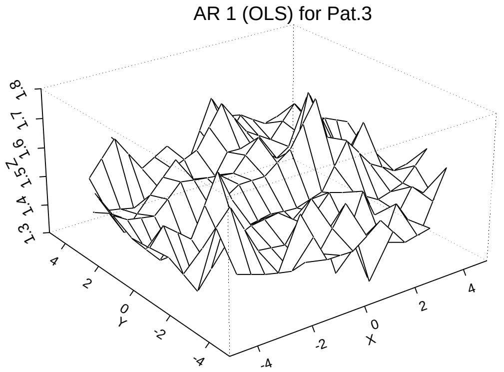

Figure E.28: Least squares estimates for first AR-parameter for data set Pat.3.

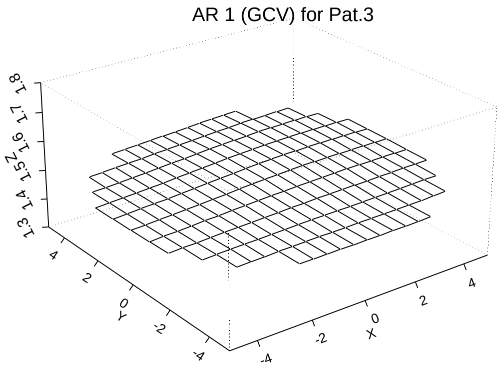

Figure E.29: Smoothed first AR-field for data set Pat.3. Choice of $\lambda_{GCV}(1) = 169.527$ by GCV.

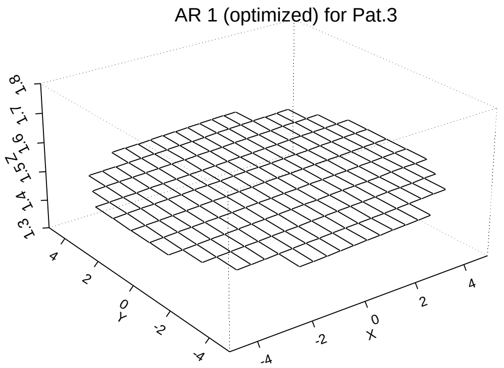

Figure E.30: Smoothed first AR-field after spatiotemporal crossvalidation for data set Pat.3 with $\lambda_{opt}(1) = 87906.98$.

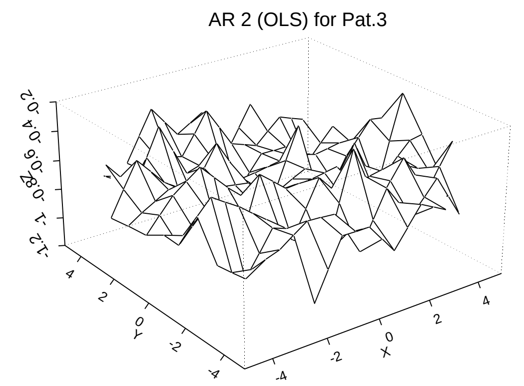

Figure E.31: Least squares estimates for second AR-parameter for data set Pat.3.

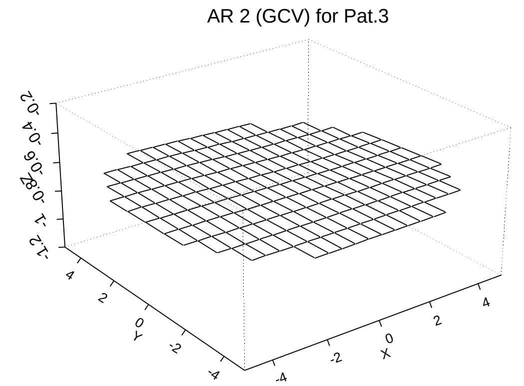

Figure E.32: Smoothed second AR-field for data set Pat.3. Choice of $\lambda_{GCV}(2) = 238.330$ by GCV.

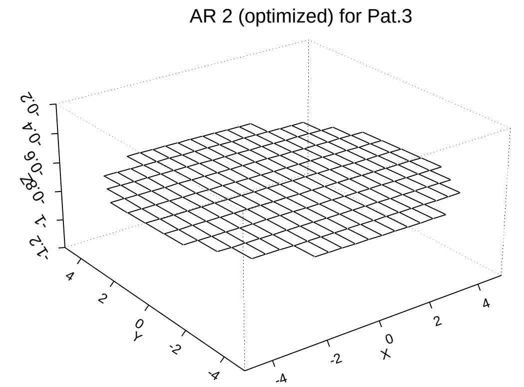

Figure E.33: Smoothed second AR-field after spatiotemporal crossvalidation for data set Pat.3 with $\lambda_{opt}(2) = 362820.69$.

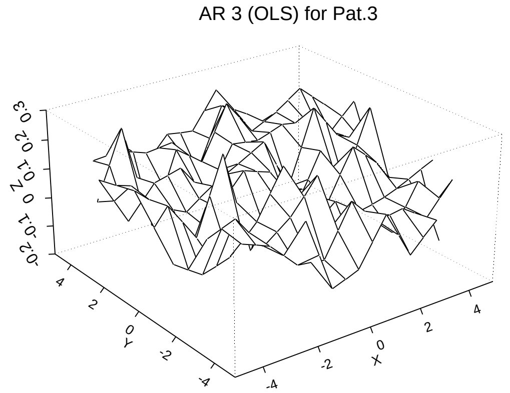

Figure E.34: Least squares estimates for third AR-parameter for data set Pat.3.

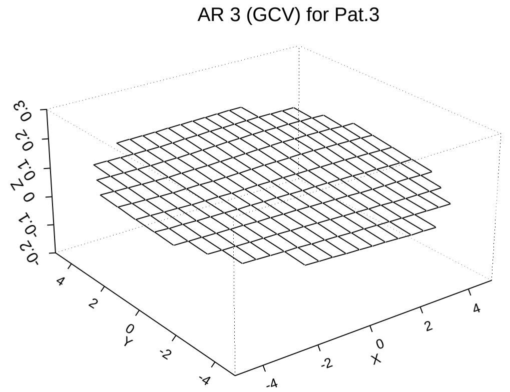

Figure E.35: Smoothed third AR-field for data set Pat.3. Choice of $\lambda_{GCV}(3) = 189.661$ by GCV.

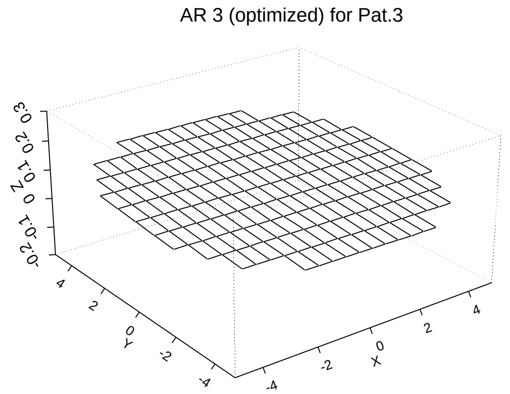

Figure E.36: Smoothed third AR-field after spatiotemporal crossvalidation for data set Pat.3 with $\lambda_{opt}(3) = 103717.46$.

Appendix F

Solution for Space-Time Sum of Squares

Below a detailed derivation can be found of the least squares solution $\tilde{\alpha}$ which minimizes the space-time penalized sum of squares given in Equation (6.1.17), i.e.,

$$\begin{aligned} STPSS_{\vec{\lambda},\vec{\nu}}(\tilde{\alpha}) &= tr\left((\mathbf{Z}-\tilde{\mathbf{Z}})'(\mathbf{Z}-\tilde{\mathbf{Z}})\right) \\ &+ (\tilde{\alpha}-\bar{\alpha})'\Lambda(\tilde{\alpha}-\bar{\alpha}) \\ &+ (\tilde{\alpha}-\hat{\alpha})'\mathbf{V}(\tilde{\alpha}-\hat{\alpha}) \end{aligned}$$

Following the standard procedure, the optimal solution $\tilde{\alpha}$ can be found by finding the first derivative and equating to zero. With

$$\begin{aligned} \mathbf{U} &= \frac{1}{d}\left(\mathbf{I}_p \otimes \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}\right)_{dp\times dp} \\ \bar{\alpha} &= U\vec{\alpha} \end{aligned}$$

and using $\frac{\partial \vec{x}'\mathbf{A}\vec{x}}{\partial \vec{x}} = (\mathbf{A}+\mathbf{A}')\vec{x}$ one obtains for the second and third summand of $STPSS_{\vec{\lambda},\vec{\nu}}(\alpha)$

$$
\begin{aligned}
\frac{\partial}{\partial \vec{\alpha}}(\vec{\alpha}-\bar{\alpha})'\Lambda(\vec{\alpha}-\bar{\alpha}) &= \frac{\partial}{\partial \vec{\alpha}}\left[\vec{\alpha}'\Lambda\vec{\alpha} - 2\vec{\alpha}'\Lambda\mathbf{U}\vec{\alpha} + \vec{\alpha}'\mathbf{U}'\Lambda\mathbf{U}\vec{\alpha}\right] \\
&= \left[\Lambda + \Lambda'\right]\vec{\alpha} - 2\left[\Lambda\mathbf{U} + \mathbf{U}'\Lambda'\right]\vec{\alpha} \\
&\quad + \left[\mathbf{U}'\Lambda\mathbf{U} + \mathbf{U}'\Lambda\mathbf{U}\right]\vec{\alpha} \\
&= 2\Lambda\vec{\alpha} - 4\Lambda\mathbf{U}\vec{\alpha} + 2\left[\mathbf{U}'\Lambda\mathbf{U}\right]\vec{\alpha}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial \vec{\alpha}}(\vec{\alpha}-\hat{\alpha})'\mathbf{V}(\vec{\alpha}-\hat{\alpha}) &= \frac{\partial}{\partial \vec{\alpha}}\left(\vec{\alpha}'\mathbf{V}\vec{\alpha} - 2\vec{\alpha}'\mathbf{V}\hat{\alpha} + \vec{\alpha}'\mathbf{V}\hat{\alpha}\right) \\
&= (\mathbf{V}+\mathbf{V}')\vec{\alpha} - 2\mathbf{V}\hat{\alpha}
\end{aligned}
$$

In a next step, the first summand is reexpressed as a sum before forming the derivative, starting with

$$
\begin{aligned}
\mathbf{Z} - \mathbf{X}\mathbf{A} &= \begin{pmatrix} \vec{Z}'_{p+1} \\ \vdots \\ \vec{Z}'_{T} \end{pmatrix} - \begin{pmatrix} \vec{Z}'_{p} & \dots & \vec{Z}'_{1} \\ \vdots & \dots & \vdots \\ \vec{Z}'_{T-1} & \dots & \vec{Z}'_{T-p} \end{pmatrix} \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_p \end{pmatrix} \\
&= \mathbf{Z} - \begin{pmatrix} \sum_{k=1}^{p} \vec{Z}_{p-k}(\vec{s}_1)\alpha(s_1) & \dots & \sum_{k=1}^{p} \vec{Z}_{p-k}(\vec{s}_d)\alpha(s_d) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} \vec{Z}_{T-k}(\vec{s}_1)\alpha(s_1) & \dots & \sum_{k=1}^{p} \vec{Z}_{T-k}(\vec{s}_d)\alpha(s_d) \end{pmatrix}
\end{aligned}
$$

The jj-th diagonal element of $(\mathbf{Z}-\mathbf{X}\mathbf{A})'(\mathbf{Z}-\mathbf{X}\mathbf{A})$ is

$$
\sum_{t=p+1}^{T}\left[\vec{Z}_t(\vec{s}_j) - \sum_{k=1}^{p}\vec{Z}_{t-k}(\vec{s}_j)\alpha(s_j)\right]^2
$$

resulting in the trace

$$
tr\left[(\mathbf{Z}-\mathbf{X}\mathbf{A})'(\mathbf{Z}-\mathbf{X}\mathbf{A})\right] = \sum_{j=1}^{d}\sum_{t=p+1}^{T}\left[\vec{Z}_t(\vec{s}_j) - \sum_{k=1}^{p}\vec{Z}_{t-k}(\vec{s}_j)\alpha(s_j)\right]^2
$$

The derivative of this trace with respect to some $\alpha_{k_0}(s_{j_0})$ is

$$\frac{d}{d\alpha_{k_0}(s_{j_0})} tr\left[(\mathbf{Z}-\mathbf{X}\mathbf{A})'(\mathbf{Z}-\mathbf{X}\mathbf{A})\right]$$

$$= -2\sum_{t=p+1}^{T} \vec{Z}_{t-k_0}(\vec{s}_{j_0})\left[\vec{Z}_t(\vec{s}_{j_0}) - \sum_{k=1}^{p} \vec{Z}_{t-k}(\vec{s}_{j_0})\vec{\alpha}_k(\vec{s}_{j_0})\right]$$

The trace now can be expressed as

$$\begin{aligned} tr\left\{(\mathbf{Z}-\mathbf{X}\mathbf{A})'(\mathbf{Z}-\mathbf{X}\mathbf{A})\right\} &= -2\underbrace{\left[(\mathbf{X}'\mathbf{Z}) \odot \begin{pmatrix} \mathbf{I}_d \\ \vdots \\ \mathbf{I}_d \end{pmatrix}\right]}_{\widetilde{\mathbf{X}'\mathbf{Z}}} \mathbf{1}\vec{1}_d \\ &\quad +2\underbrace{\left\{(\mathbf{X}'\mathbf{X}) \odot \left[\begin{pmatrix} \mathbf{I}_d \\ \vdots \\ \mathbf{I}_d \end{pmatrix}\begin{pmatrix} \mathbf{I}_d & \dots & \mathbf{I}_d \end{pmatrix}\right]\right\}}_{\widetilde{\mathbf{X}'\mathbf{X}}} \vec{\alpha} \end{aligned}$$

For ease of notation, results of the above Hadamard products are symbolized by corresponding tilde symbols in the sequel. The derivatives of the three summands are combined and equated to zero. This yields the following system of equations.

$$\widetilde{\mathbf{X}'\mathbf{Z}}\vec{1}_d + \mathbf{V}\hat{\alpha}$$

$$= \left\{\widetilde{\mathbf{X}'\mathbf{X}} + \Lambda - 2\Lambda\mathbf{U} + \mathbf{U}'\Lambda\mathbf{U} + \mathbf{V}\right\}\vec{\alpha}$$

which is solved by

$$\tilde{\alpha} = \left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{V} + \Lambda + (\mathbf{U}-2\mathbf{I})\Lambda\mathbf{U}\right]^{-1}\left\{\widetilde{\mathbf{X}'\mathbf{Z}}1_d + \mathbf{V}\hat{\alpha}\right\}$$

provided the inverse exists. To see that this is indeed the case, take a look at the components involved:

- If appropriately scaled, $\widetilde{\mathbf{X}'\mathbf{X}}$ converges to a matrix of covariances which is invertible
- $\mathbf{V}$ is an invertible diagonal matrix, provided all weights in the vector $\vec{v}$ are larger than zero (as was assumed throughout)
- The matrix $\mathbf{C} = \Lambda + (\mathbf{U} - 2\mathbf{I})\Lambda\mathbf{U}$ is block diagonal with jj-th block matrix

$$\mathbf{C}_{jj} \quad = \quad \lambda_j \left[\mathbf{K} + \frac{1}{d}\mathbf{1}_{d\times d}\mathbf{K}\mathbf{1}_{d\times d}\frac{1}{d} - \frac{2}{d}\mathbf{K}\mathbf{1}_{d\times d} \right]$$

Given distinct spatial locations where data are observed, the matrix $\mathbf{K}$ is known to be invertible with $\vec{x'}\mathbf{K}\vec{x} \geq 0$ for all vectors $\vec{x}$ in $\mathbb{R}^d$, where equality holds only for $\vec{x} = 0$. In particular, this condition is fulfilled if $\vec{x}$ is a vector with all elements being equal. Such vectors have the property $\vec{x} = \frac{1}{d}\mathbf{1}_{d\times d}\vec{x}$. From this it follows that

$$\vec{x'}\frac{1}{d}\mathbf{1}_{d\times d}\mathbf{K}\mathbf{1}_{d\times d}\frac{1}{d}\vec{x} \geq 0$$

where equality holds if the sum of elements of $\vec{x}$ is zero. Therefore, this summand must be nonnegative definite. The same argument can be applied to the summand $\frac{1}{d}\mathbf{K}\mathbf{1}_{d\times d}$. Finally, since the sum of an invertible matrix and a nonnegative definite matrix remains invertible, $\mathbf{C}_{jj}$ is invertible, and so are $\mathbf{C}$ and $\left[\widetilde{\mathbf{X}'\mathbf{X}} + \mathbf{V} + \Lambda + (\mathbf{U} - 2\mathbf{I})\Lambda\mathbf{U}\right]$.

List of Figures

List of Tables

List of Symbols

This section provides a list of symbols used in the preceding text. Scalars and indices are denoted by lower case roman letters. Vectors are symbolized by an arrow as in $\vec{x}$. Matrices are denoted by upper case letters and printed in bold face like $\mathbf{X}$. Greek symbols denote parameters.

	<u>Sets</u>	
$\mathbb{D}$	Spatial domain.	p. 66
$\mathbb{N}_0$	Set of positive integer numbers, including zero.	
$\mathbb{R}$	Set of real numbers.	
$\mathbb{T}$	Countable index set for discrete time points. $\mathbb{T} \subset \mathbb{Z}$.	p. 39
$\mathbb{Z}$	Set of all integer numbers.	p. 39

	<u>Scalars and Indices</u>	
T	Cardinal number of $\mathbb{T}$. Total number of time points.	p. 40
d	Dimensionality of a multivariate vector process.	p. 46
	Cardinal number of $\mathbb{D}$.	
i, j, k, l, r	Indices in $\mathbb{Z}$.	
m	Spatial Dimensionality, e.g. $m = 2$ for a plane.	p. 66
n	Number of basis functions in Universal Kriging.	p. 82
p	Order of (vector) AR-process.	p. 44
q	Order of (vector) MA-process.	p. 44
t	Time index, $t \in \mathbb{T}$.	p. 39

	Spatial Statistics	
c_0	Nugget effect.	p. 74
$\vec{h}$	Spatial or temporal lag.	p. 72
r_0	Range.	p. 74
s_i	Interior knot for spline function.	p. 90
$\vec{\delta}$	Spline coefficients of radial basis function.	p. 99
σ^2	Sill.	p. 74
$\vec{\theta}$	Coefficients of basis functions in Universal Kriging.	p. 82
$\vec{\vartheta}$	Spline coefficients of basis functions.	p. 99
	Splines	
$GCV(\lambda)$	Crossvalidation score.	p. 95
$OCV(\lambda)$	Crossvalidation score.	p. 94
$PENSS^*$	Temporally summed roughness penalty.	p. 114
$PSS_\lambda(g)$	Penalized sum of squares.	p. 92
$STPSS_{\vec{\lambda},\vec{\nu}}(\tilde{\alpha})$	Space-Time penalty.	p. 116
$J_\tau(g(\vec{s}))$	General roughness penalty.	p. 98
M	Number of summands in Ljung-Box test statistics.	p. 58
$S_2[a,b]$	Space of differentiable functions on [a,b] with	p. 92
$\mathbf{V}$	Smoothing Matrix.	p. 116
$[a,b]$	Interval on the real line.	p. 90
$df(\lambda)$	Equivalent degrees of freedom for λ.	p. 95
$df^{err}(\lambda)$	Equivalent degrees of freedom for error.	p. 96
$df^{var}(\lambda)$	Equivalent degrees of freedom for λ.	p. 96
$f_j(\vec{s})$	Basis function.	p. 82
g	Function.	p. 90
$\hat{\vec{g}}$	solution of spline smoothing problem.	p. 93
$\vec{g}$	Vectorized cubic spline.	p. 90
$\vec{g''}$	Vectorized second derivatives of cubic spline g.	p. 91
s_i	Location. Interior knot for spline function.	p. 90
$s_{(j)i}$	Component j of location vector $\vec{s}_i$.	p. 99
$\lambda, \vec{\lambda}$	Smoothing parameter, multivariate.	p. 92
$\nu, \vec{\nu}$	Smoothing parameter, multivariate.	p. 116
Λ	Smoothing Matrix.	p. 116

	Univariate Stochastic Processes	
$\|\|.\|\|$	Euclidian norm.	p. 99
$\otimes$	Kronecker-Operator.	p. 54
$\odot$	Hadamard-Operator.	p. 117
$vec(.)$	Operator to stack columns of a matrix.	p. 53
$C_Z(\vec{s}_i, \vec{s}_j)$	Covariance of $Z(\vec{s}_i)$ and $Z(\vec{s}_j)$.	p. 102
$C(\vec{s}_i, \vec{s}_j)$	Covariogram.	p. 73
$C(h)$	Stationary covariogram.	p. 73
$V_{GCV}(\tilde{\lambda}_k)$	Generalized crossvalidation function.	p. 120
Y_t	Random variable Y observed at time $t \in \mathbb{T}$.	p. 40
$(Y_t)_{t\in\mathbb{T}}$	Stochastic process.	p. 39
l	Log-likelihood.	p. 55
(y_t)	Time series.	p. 39
α_k	k-th Autoregressive coefficient.	p. 44
$\alpha_k(\vec{s})$	k-th Autoregressive coefficient at location $s \in \mathbb{D}$.	p. 110
$\hat{\alpha}$	OLS-Solution in diagonal VAR-model.	p. 119
α^*_{min}	Solution in diagonal VAR-model under $PENSS^*$.	p. 119
β_l	l-th Moving average coefficient.	p. 44
$(\epsilon_t)_{t\in\mathbb{T}}$	White Noise with mean 0 and constant variance σ^2_ϵ.	p. 44
$\eta(.)$	Radial basis function.	p. 99
$\gamma(u, v)$	Autocovariance of Y_u and Y_v.	p. 41
$\gamma(h)$	Autocovariance at lag $h = \|u - v\|$.	p. 41
$2\gamma(\vec{h})$	Stationary variogram.	p. 72
$\gamma_{Y,\epsilon}(k)$	Cross-covariance between Y_{t-k} and ϵ_t, i.e. $E[Y_{t-k}\epsilon_t]$ for zero mean processes (Y_t) and (ϵ_t).	p. 45
$\gamma_{i,j}(h)$	Cross-covariance at lag h between $Y_t(i)$ and $Y_{t+h}(j)$.	p. 48
$\gamma_{i,j}(h)$	Cross-covariance estimate.	p. 52
$2\gamma(\vec{s}_i, \vec{s}_j)$	Variogram.	p. 72
$2\hat{\gamma}(h)$	Classical variogram estimator.	p. 75
$2\bar{\gamma}(h)$	Robust variogram estimator.	p. 75
$\vec{\gamma}_{OK}$	Solution vector for semivariogram, ordinary Kriging.	p. 80
$\vec{\gamma}_{UK}$	Solution vector for semivariogram, universal Kriging.	p. 82

μ, μ_t	Expected value in $\mathbb{R}$ (at time t).	p. 40
$\mu(\vec{s})$, $\mu_t(\vec{s})$	Expected value at location $\vec{s} \in \mathbb{D}$ (at time t).	p. 110
$\vec{\mu}$	Expected value in $\mathbb{R}^d$.	p. 47
$\vec{\mu}_t$	Expected value in $\mathbb{R}^d$ at time t.	p. 47
$\rho(u, v)$	Autocorrelation between Y_u and Y_v.	p. 41
$\rho_{i,j}(h)$	Cross-correlation Y_i and Y_j measured at distance h.	p. 49
$\rho(h)$	Autocorrelation at lag $h = \lvert u - v \rvert$.	p. 41
$\hat{\rho}_{\text{partial}}(h)$	Estimated partial correlation at lag h.	p. 57
σ^2, σ_t^2	Variance (at time t).	p. 40
$\sigma^2_{ME}(\vec{s})$	Variance of measurement error.	p. 67
$\sigma^2_{SK}(\vec{s})$	Variance for Simple Kriging.	p. 79
$\sigma^2_{OK}(\vec{s})$	Variance for Ordinary Kriging.	p. 80
$\sigma^2_{UK}(\vec{s})$	Variance for Universal Kriging.	p. 83
$\tau^2_{OK}(\vec{s})$	Variance for Ordinary Kriging with nugget effect.	p. 81
$\xi(\vec{s})$	Error process.	p. 67

<u>Vector-Valued Stochastic Processes</u>

$\mathbf{A}$	Matrix of all coefficients of AR-process.	p. 111
$\hat{\mathbf{A}}$	Matrix of AR-coefficients (OLS-solution).	p. 53
$\mathbf{A}_k$	k-th coefficient matrix from VAR-representation.	p. 47
$\hat{\mathbf{A}}$	Matrix of (local) OLS estimates.	p. 53
$\mathbf{B}_j$	j-th coefficient matrix from VMA-representation.	p. 47
$\mathbf{C}$	Matrix of Covariances.	p. 102
$\mathbf{D}$	Scaling Matrix.	p. 49
$\mathbf{E}$	Matrix of Errors.	p. 52
$\mathbf{H}$	Matrix of basis functions.	p. 99
$\mathbf{K}$	Weight matrix for roughness penalty.	p. 91
$\mathbf{Q}$	Matrix.	p. 91
$\mathbf{R}$	Matrix.	p. 91

$\mathbf{R}(h)$	Cross-correlation matrix at lag h.	p. 49
$\mathbf{S}(\lambda)$	Smoothing Matrix.	p. 94
$\mathbf{U}$	Centering matrix.	p. 114
$\mathbf{X}$	Matrix of lagged observations in VARMA-model.	p. 52
$\vec{Y}_t$	Vector-valued random variable observed at time t.	p. 46
$\mathbf{Z}$	Matrix of vector observations.	p. 111
$\vec{Z}_t(\vec{s})$	Vector-valued random variable at time t and location $\vec{s}$.	p. 110
$\tilde{\mathbf{Z}}$	Predicted values under $STPSS$.	p. 115
$Z(\vec{s})_{\vec{s}\in\mathbb{D}}$	Spatial stochastic process at location $\vec{s}\in\mathbb{D}$	p. 66
$\vec{c}_{UK}$	Solution vector for covariances, universal Kriging.	p. 83
$\tilde{e}_t, \vec{e}_t$	Residuals.	p. 54
$\vec{h}$	Multivariate spatial distance.	p. 72
$s_{(i)}$	i-th element of spatial location vector $\vec{s}$.	p. 90
$\vec{s}$	Location index in $\mathbb{D}\subset\mathbb{R}^m$.	p. 66
$\vec{\alpha}_k$	Vector of elements of k-th AR-parameter field.	p. 112
$\vec{\alpha}$	Elements of all p AR-parameter fields.	p. 112
$\tilde{\alpha}_{min}$	Matrix of coefficients minimizing $STPSS$.	p. 118
α^*_{min}	Matrix of coefficients minimizing $PENSS^*$.	p. 119
$\tilde{\alpha}_{min}$	Solution in diagonal VAR-model under $STPSS^*$.	p. 119
$\eta(\vec{s})$	Spatial small scale variation, or *noise*, at location $\vec{s}\in D$.	p. 66
$\vec{\lambda}_{OK}$	Solution of Ordinary Kriging Equation.	p. 80
$\vec{\lambda}_{UK}$	Solution of Universal Kriging Equation.	p. 82
$\mu(\vec{s})$	Spatial large scale variation, or *trend*, at location $\vec{s}\in D$	p. 66
$\Gamma(h)$	Cross-covariance matrix at lag h.	p. 49
Γ	Matrix (in $\mathbb{R}^{dp\times dp}$) of autocovariances up to order $(p-1)$.	p. 51
Γ_p	Matrix of dimensions $dp\times d$ with stacked autocovariance matrices $\Gamma(1),\ldots,\Gamma(p)$.	p. 51
Γ_{OK}	Ordinary Kriging matrix.	p. 80
Γ_{UK}	Universal Kriging matrix.	p. 82
Λ	Weight matrix for smoothing.	p. 114
Σ_ϵ	Covariance matrix of vector White Noise Process $(\vec{\epsilon})$	p. 47
$\hat{\Sigma}_\epsilon$	Covariance matrix estimate.	p. 54
Σ_{UK}	Universal Kriging matrix.	p. 83

Bibliography

Ali, M. (1989). Tests for autocorrelation and randomness in multiple time series. *JASA 84*, 533–540.

Aoyagi, K., Y. Kimura, H. Isono, and H. Akiyama (1998). Multifocal electroretinogram in central serous chorioretinopathy. *I.O.V.S. 39*, 185.

Berke, O. (1998). *Über die statistische on-line-Prädiktion von Umweltmonitoringdaten im Rahmen von dynamischen linearen Raum-Zeit Modellen*. Ph. D. thesis, University of Dortmund, Dortmund, Germany.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *JRSS-B 36*, 192–225.

Box, G., G. Jenkins, and G. Reinsel (1994). *Time Series Analysis: Forecasting and Control* (third ed.). Prentice-Hall Inc.

Box, G. and D. Pierce (1970). Distribution of residual autocorrelations in autoregressive–integrated moving average time series models. *JASA 65*, 1509–1526.

Brown, B. and M. Yap (1995). Contrast and luminance as parameters defining the output of the VERIS® topographical ERG. *Ophthalmic and Physiological Optics 16*, 42–48.

Buja, N. A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *Ann.Statist. 17*, 453–555.

Chui, C. (1992). *An Introduction to Wavelets* (Second ed.). Academic Press.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *JASA 74*, 829–836.

Cook, R. and S. Weisberg (1982). *Residuals and Influence in Regression*. Chapman and Hall.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numer.Math. 31*, 377–390.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology 17*(7), 563–586.

Cressie, N. (1986). Kriging nonstationary data. *JASA 81*, 625–634.

Cressie, N. (1990a). The origins of kriging. *Mathematical Geology 22*, 239–252.

Cressie, N. (1990b). Reply to Wahba's letter. *American Statistician 44*, 256–258.

Cressie, N. (1993). *Statistics for Spatial Data* (Second ed.). Wiley.

Cressie, N. and D. Hawkins (1980). A robust / resistant spatial analysis of soil-water infiltration. *Water Resources Research 23*, 911–1017.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics XLI*, 909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF regional conference series in applied mathematics.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Durbin, J. (1960). The fitting of time-series models. *Rev. Internat. Statisti. Inst. 28*, 233–244.

EDI (1999). *VERIS® Science 4.0 Manual*. Electro-Diagnostic Imaging, Inc., San Mateo, CA.

Emmerson, J. and D. Hoaglin (1983). Analysis of two-way tables by medians. In D. Hoaglin, F. Mosteller, and J. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis*, pp. 166–210. Wiley.

Graham, S. and A. Klistorner (1998). Electrophysiology: a review of signal origins and applications to investigating glaucoma. *Australian and New Zealand Journal of Ophthalmology 26*, 71–85.

Green, P. and B. Silverman (1994). *Nonparametric Regression and Generalized Linear Models - A Roughness Penalty Approach*. Chapman and Hall.

Guttorp, P. and P. Sampson (1994). Methods for estimating heterogeneous spatial covariance functions with environmental applications. In G. Patil and C. Rao (Eds.), *Handbook of Statistics XII: Environmental Statistics*, pp. 663–90. Elsevier/North Holland.

Hannan, E. (1970). *Multiple Time Series*. Wiley.

Harville, D. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall.

Hawkins, D. and N. Cressie (1984). Robust kriging – a proposal. *Journal of the International Association for Matheatical Geology 16*, 3–18.

Hosking, J. (1980). The multivariate portmanteau statistic. *JASA 75*, 602–608.

Hosking, J. (1981). Equivalent forms of the multivariate portmanteau statistics. *JRSS-B 43*, 261–262.

Huang, H. and N. Cressie (1996). Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics and Data Analysis 22*, 159–175.

Huijbregts, C. and G. Matheron (1971). Universal kriging (an optimal method for estimation and contouring in trend surface analysis). In *Proceedings of Ninth International*

Symposium on Techniques for Decision-Making in the Mineral Industry, Volume 12, pp. 159–169. The Canadian Institute of Mining and Metallurgy.

Hutchinson, M. and F. de Hoog (1995). Smoothing noisy data with spline functions. *Numer. Math. 47*, 99–106.

Jenkins, G. and D. Watts (1968). *Spectral Analysis and its Applications*. Holden-Day.

Journel, A. and C. Huijbregts (1978). *Mining Geostatistics*. Academic Press.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering. 82*, 35–45.

Keating, D., S. Parks, J. McDonagh, A. Evans, A. Elliot, and J. Jay (1998). Appropriate interpretation of the multifocal electroretinogram in a wide range of disease states. *I.O.V.S. 39*, 185.

Kondo, M., Y. Miyake, M. Horiguchi, S. Suzuki, and A. Tanikawa (1995). Clinical evaluation of multifocal electroretinogram. *Investigative Ophthalmology and Visual Science 36*, 2146–2150.

Krahnke, T. (1997). Rekursive Spektralschätzung und Wavelet-Analyse am Beispiel psychophysiologischer Daten. Master's thesis, Universität Dortmund, Dortmund, Germany.

Kretschmann, U., M. Seeliger, K. Ruether, T. Usui, E. Apfelstedt-Sylla, and E. Zrenner (1998). Multifocal electroretinography in patients with stargardt's macular dystrophy. *British Journal of Ophthalmology 82*, 267–275.

Li, W. and A. McLeod (1981). Distribution of the residual autocorrelations in multivariate arma time series models. *JRSS-B 43*, 231–239.

Ljung, G. and G. Box (1978). On a measure of lack of fit in time series models. *Biometrika 66*, 297–303.

Mack, G., D. H., F. J., M.-S. S., and S. J. (1999). A new mode of recording retinal activity: Multifocal ERG. *Journal of French Ophthalmology 22*, 221–225.

Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence 11*, 674–693.

Matheron, G. (1962). *Traite de Geostatistique Appliquee, Tome I.* Number 14 in Memoires du Bureau de Recherches Geologiques et Minieres. Editions Technip, Paris.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology 58*, 1246–1266.

Matheron, G. (1967). Kriging, or polynomial interpolation procedures? A contribution to polemics in mathematical geology. *Transactions, Ecole National Superieure des Mines, Paris LXX*, 240–244.

Matheron, G. (1971). *The Theory of Regionalized Variables and its Applications*. Fontainebleau, France.

Mathsoft (2000a). *S-Plus® 2000 Professional for Micrososft Windows*. Mathsoft Inc., Seattle.

Mathsoft (2000b). *S+SPATIALSTATS Version 1.5*. MathSoft, Seattle, WA.

Meiring, W., P. Guttorp, and P. Sampson (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics 5*, 197–222.

Nychka, D., B. Bailey, S. Ellner, P. Haaland, and N. O'Connell (2000). *FUNFITS: Data analysis and Statistical Tools for Estimating Functions*. http://www.cgd.ucar.edu/stats/Funfits/index.shtml.

Palmowski, A., E. Sutter, M. Bearse, and W. Fung (1997). Mapping of retinal function in diabetic retinopathy using the multifocal electroretinogram. *Investigative Ophthalmology and Visual Science 38*, 2586–2596.

Parks, S., D. Keating, T. Williamson, A. Evans, A. Elliott, and J. Jay (1996). Functional imaging of the retina using the multifocal electroretinograph: a control study. *British Journal of Ophthalmology 80*, 831–834.

Pfeiffer, P. and S. Deutsch (1980a). Identification and interpretation of first order space-time arma models. *Technometrics 22*(3), 397–408.

Pfeiffer, P. and S. Deutsch (1980b). A three-stage procedure for space-time modeling. *Technometrics 22*, 35–47.

Priestley, M. (1980). State-dependent models: A general approach to non-linear time series analysis. *Journal of Time Series Analysis 1*, 47–71.

Priestley, M. (1988). *Non-Linear and Non-Stationary Time Series Analysis*. Academic Press.

Priestley, M. (1996). Wavelets and time-dependent spectral analysis. *Journal of Time Series Analysis 17*(1), 85–103.

Quenouille, M. (1949). Approximate tests of correlations in time-series. *JRSS - B 11*, 68–84.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis*. Springer.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math 10*, 177–183.

Reinsel, G. (1997). *Elements of Multivariate Time Series Analysis* (Second ed.). Springer.

Rioul, O. and M. Vetterli (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine Oct 1991*, 14–37.

Sampson, P. and P. Guttorp (1992). Nonparametric estimation of nonstationary spatial covariance structure. *JASA 87*, 108–119.

Sampson, P., P. Guttorp, and W. Meiring (1994). Spatio-temporal analysis of regional ozone data for operational evaluation of an air quality model. In *ASA 1994 Proceedings of the Section on Statistics and the Environment*, pp. 46–55. American Statistical Association.

Schmidt, A. and A. O'Hagan (2000). Bayesian inference for nonstationary spatial covariance structure via spatial deformations. Technical report, School of Mathematics and Statistics, University of Sheffield. Research Report 498 / 00.

Searle, S. (1971). *Linear Models*. Wiley.

Seeliger, M., U. Kretschmann, E. Apfelstedt-Sylla, K. Ruther, and E. Zrenner (1998a). Implicit time topography of multifocal electroretinograms. *Investigative Ophthalmology and Visual Science 39*, 718–723.

Seeliger, M., U. Kretschmann, E. Apfelstedt-Sylla, K. Ruther, and E. Zrenner (1998b). Multifocal electroretinography in retinitis pigmentosa. *American Journal of Ophthalmology 125*, 2844–2851.

Si, Y., S. Kishi, and K. Aoyagi (1999). Assessment of macular function by multifocal electroretinogram before and after macular hole surgery. *British Journal of Ophthalmology 83*, 420–424.

Silverman, B. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist. 12*, 898–916.

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *JRRS-B 47*, 1–52.

Sutter, E. and M. Bearse (1999). The optic nerve head component of the human ERG. *Vision Research 39*, 419–436.

Sutter, E. and D. Tran (1992). The field topography of ERG components in man - i. the photopic luminance response. *Vision Research 32*, 433–446.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesly.

Verdon, W. and G. Haegerstrom-Portnoy (1998). Topography of the multifocal electroretinogram. *Documenta Ophthalmologica 95*, 73–90.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *JRRS-B 40*, 364–372.

Wahba, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *JRRS-B 45*, 133–150.

Wahba, G. (1990a). Letter to the editor. *The American Statistician 44*, 255–256.

Wahba, G. (1990b). *Spline Models for Observational Data*. SIAM, Philadelphia.

Walker, G. (1931). On periodicity in time series of related terms. *Proc. Roy. Soc. A 131*, 518–532.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematics Society 41*, 63–75.

Yule, G. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philos. Trans. Roy. Soc. A 226*, 267–298.