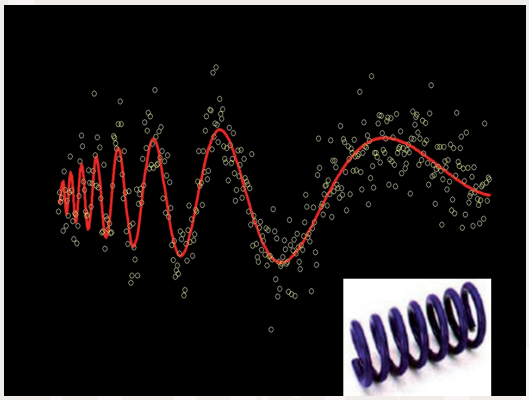
DOCTORAATSPROEFSCHRIFT

2008 | Faculteit Wetenschappen



Flexible Modelling Techniques and Use of Historical Controls in Animal Studies

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde, te verdedigen door:

John T. Maringwa

Promotor: Prof. dr. Helena Geys
Copromotor: Prof. dr. Christel Faes

universiteit hasselt
I-BioStat
Interuniversity Institute for Statistics and statistical Bioinformatics
KATHOLIEKE UNIVERSITEIT LEUVEN

DOCTORAATSPROEFSCHRIFT

2008 | Faculteit Wetenschappen

Flexible Modelling Techniques and Use of Historical Controls in Animal Studies

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde, te verdedigen door:

John T. Maringwa

Promotor: Prof. dr. Helena Geys
Copromotor: Prof. dr. Christel Faes

D/2008/2451/47

Acknowledgements

For the journey I started four years ago, it was clear that, although the destination would guarantee satisfaction, the journey was never a one-man show. At this juncture, I am deeply indebted to several people who have contributed immensely towards my arrival at the intended destination.

First and foremost I would like to thank my supervisors, Prof. Helena Geys and Prof. Christel Faes. Helena, it was a pleasure working with you throughout. You were always been there for me, patiently, I should say. Thanks for all the assistance and mentoring. Christel, I really appreciate the effort you put in helping and guiding me throughout. You always found time for me to iron out outstanding issues. Many thanks Christel for everything you did for me. I would also like to thank Prof. Ziv Shkedy for all the help, support, and encouragement he gave me during the past four years. Thanks for all the meetings, discussions, and the advice you always gave me Ziv!

I am also indebted to Prof. Marc Aerts and Prof. Geert Molenberghs for the very helpful discussions we had. Further, I wish to thank Prof. Carmen Cadarso-Suárez with whom we have had a fabulous collaboration. Thanks Carmen for the warm reception and hospitality during our visit in Spain. Many thanks also to my colleagues Abel Tilahun, Dr. Ariel Alonso, and Dr. Niel Hens for the joint work on various projects.

Special mention goes to a colleague Dr. Kristien Wouters and a former colleague Dr. Jan Serroyen. Thanks for the sacrifice you made before the 'minibus' to Beerse. And to the 'Janssening group', I enjoyed your company guys during our weekly trips to Beerse. I would also like to thank my room-mate Dr. Jośe Cortinas for the very interesting wide ranging discussions we had in the office, and for creating a conducive working environment. Thanks also to all the CENSTAT colleagues for creating such a friendly environment and for support in various forms. I would certainly not forget the secretarial staff in CENSTAT for all the assistance.

I gratefully acknowledge the financial support from Janssen Pharmaceutica without which this would be a non-event. In particular, I would like to thank Dr. Luc Bijnens for the productive collaboration.

I also wish to express my gratitude to one Dr. Ward Schrooten. After being my external supervisor for my Msc in Applied Statistics in 2003, you became a close friend. Thanks for helping me navigate my way in otherwise difficult circumstances. Many thanks to you and An. Lastly I would like to thank all my family members in Zimbabwe and my brothers abroad for being supportive in every respect. In particular, many thanks to my parents for investing in my education even under uncompromising conditions. Thank you! Special mention goes to my wife Emelda for all the love, encouragement, and support she has given me before and after joining me in Belgium. I honestly can't ask for more from you! Thanks Emelda!

JT

Diepenbeek, 26 September 2008

Contents

Publications

The material presented in the subsequent chapters is deeply rooted in the following scientific publications:

Maringwa, J.T., Faes, C., Geys, H., Aerts, M., Teuns, G., and Bijnens, L. (2007) On the use of historical control studies in pre-clinical safety studies. *Journal of Biopharmaceutical Statistics* **17**, 493-509.

Maringwa J.T., Faes, C., Geys, H., Hens, N and Cadarso-Suárez, C. (2008a) Bayesian adaptive penalized splines for non-normal data. *Submitted for publication.*

Maringwa, J.T., Faes, C., Geys, H., Molenberghs, G., Cadarso-Suárez, C., Pardo-Vázquez, J.L., Leborán, V., and Acuña, C. (2008b) Application of penalized smoothing splines in analyzing neuronal Data. *Accepted: Biometrical Journal.*

Maringwa, J.T., Faes, C., Geys, H., Shkedy, Z., Molenberghs, G., Aerts, M., and Bijnens, L. (2008c) Bayesian semiparametric modelling of univariate and bivariate longitudinal data. *Submitted for publication.*

Maringwa, J.T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Van Ammel, K., Teisman, A. and Bijnens, L. (2008d) Application of semiparametric mixed models and simultaneous confidence bands in a Cardiovascular safety experiment with longitudinal data. *Journal of Biopharmaceutical Statistics*, **18**, 000-000.

Maringwa, J.T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Van Ammel, K., Teisman, A. and Bijnens, L. (2006) Analysis of a cardiovascular safety experiment with longitudinal data using penalized splines. In: *Proceedings of the 21st International Workshop on Statistical Modelling.* Hinde, J., Einbeck, J. and Newell, J. (Eds.). Galway, Ireland. pp 346-353.

Maringwa, J.T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., G., Van Ammel, K., and Bijnens, L. (2008e) Analysis of cross-over designs with serial correlation within periods using semiparametric mixed models: Tutorial in Biostatistics. *Accepted: Statistics in Medicine.*

Tilahun, A., Maringwa, J.T., Geys, H., Alonso, A., Raeymaekers, L., Molenberghs, G., Kieboom, G.V., Drinkenburg, P., Bijnens, L. (2008) Investigating association between behavior, Corticosterone, Heart Rate, and Blood Pressure in rats using surrogate marker evaluation methodology. *Accepted: Journal of Biopharmaceutical Statistics.*

1

Introduction

Development of drugs in pharmaceutical companies is a rather long and complex process. One of the key aspects of the developmental process involves the initial stage of testing the drugs in animals. This stage is essentially important in order to determine the safety levels of the drugs of interest. Later stages would then involve tests carried out in humans.

Experiments carried out in animals, the main focus in this thesis, are normally referred to as pre-clinical or non-clinical experiments. Although pre-clinical studies can not replace clinical trials, they have some appealing features worthy mentioning, especially in comparison to clinical experiments. For example, pre-clinical experiments enable investigators to have a greater degree of control over the structure and size of the study as well as experimental subjects in comparison with clinical trials. There are some important features characteristic of experiments carried out in plants or animals which would not be associated with clinical trials. One issue is that, human response to medication tends to be more variable than in genetically identical animals or plants or from tightly controlled chemical or physical experiments. Further, investigators may not be able to control as many sources of variation compared to laboratories, therefore more subjects are needed to provide control over random error. Although still a controversial issue, ethics are rather more of an issue with clinical trials than with non-clinical trials.

In the following, a brief introduction to the different animal studies considered in

this thesis is given. The aims and objectives within each type of study will undoubtedly differ as will be seen shortly.

1.1 Cardiovascular Safety Experiments

A cardiovascular disease refers to the class of diseases that involve the heart or blood vessels, that is, arteries and veins. Technically, the term would refer to any disease that affects the cardiovascular system. As already pointed out, investigations of drugs aimed at such diseases start in non-clinical experiments, before finding their way to humans. Non-clinical studies involve different types of study designs. The type of outcome, i.e., whether continuous, discrete or categorical depends on the specifics of the study.

In this thesis, focus will be on both continuous and discrete outcomes. First, we give a brief overview on the continuous outcomes obtained from cardiovascular safety experiments carried out in dogs. The measurements from each subject are obtained repeatedly over a period of time, constituting longitudinal profiles. Two types of study design namely, the parallel and the cross-over design are considered. While the parallel design is somewhat standard, the cross-over design considered here differs from conventional cross-over studies due to the presence of longitudinal measurements within each treatment period. Intricacies including the possible presence of serial correlation within treatment periods, and carry-over effects further complicate the analysis. These are some of the issues that will be addressed in this thesis.

There are many reasons for collecting repeated measurements and naturally, these should influence the statistical analysis procedures. In such experiments, each subject produces a profile of repeated measurements, and the main goal of the analysis is usually to assess the effect of different treatment regimes on these profiles. Often, a simple and very useful approach is to summarize the data from each subject into a single summary statistic that is deemed relevant for the analysis. Not only does this lead to loss information, it also makes specific assumptions about the contribution of each subject to the summary statistic, which in certain cases, may be questionable.

Although the summary statistic approach may be convenient, the researchers' interest may be, for example, to access the effect of a compound over time. Indeed, interest therefore lies in the behavior of profiles, i.e., following the repeated measurements. In such situations, an analysis that takes into account the repeated measures structure of the data should be called into play. Tools like the linear mixed model (LMM, Verbeke and Molenberghs, 2000) for continuous outcomes and generalized

linear mixed models (GLMM, Molenberghs and Verbeke, 2005) for discrete data outcomes are obvious choices. For continuous outcomes, the normal distribution, with its desirable properties, ensures computations can be done with relative ease. The same is however not true when the outcome is discrete or categorical, essentially because the analogue of the multivariate normal distribution does not exist.

Unless the full factorial structure is applied, it is imperative to specify the function describing the evolution of profiles in time. After data exploration, one can decide on some parametric model, for example, a linear, quadratic or any polynomial of a specified degree. In most cases this works perfectly well. However, when data pose some irregular profiles or exhibit heterogeneous tendencies, appropriate parametric models may be difficult to obtain. This brings us to the issue of flexible modelling, wherein the data themselves play an integral role in determining the 'appropriate' function to describe the evolution. In this thesis, special emphasis is put on the use of semiparametric models in the form of penalized splines, primarily due to their desirable connection with mixed models. We construct hypothetical models by manipulating the structure of the construction of penalized splines model. Such models, would, in practice, help to define parsimonious mean structures. Special focus will be on adjusting existing methods of constructing simultaneous confidence bands for penalized splines (e.g., Ruppert *et al.*, 2003), with the aim of including within and between-subject variability, as well as variability arising from smoothing.

1.2 Electrophysiological Experiments

Neuronal data from an electrophysiological experiment carried out with a monkey will also be considered here. Neurons carry information by means of electrical signals (action potentials or spikes) which are transmitted across synapses. The spike is a pulse signal of about 1 ms duration and of the same amplitude. It constitutes the relevant signal for the interactions between neurons. In electrophysiological experiments, these spikes are recorded by microelectrodes inserted in the brain as they occur in the time course or time stamps. It is assumed that the number of spikes per time unit, the spike rate, produced by single neurons is a relevant parameter for the coding in the brain. Single-unit activity is irregular, both within and across trials; hence to obtain the regularity of the response, trials are repeated several times. Furthermore, to assess that the behavior of single neuron activity is present at population level, the statistical analysis should be extended to the population of neurons with the same properties (Kass *et al.*, 2005).

The focus in this thesis will be to appropriately describe the time evolution of profiles from such experiments. Comparison of different experimental conditions is also of interest. The primary analysis tool is the penalized spline model. Key issues include the transition from single neuron analysis to population level analysis, keeping in mind that the response considered now is of count form. Data from electrophysiological experiments often show profiles with temporal heterogeneity tendencies. This essentially means that a profile is more flat in some sections while tending to be more steep in others. Methods that acknowledge such type of behavior in the data are not only interesting, but sometimes necessary. Focusing on counts from the neuronal data, we propose a Bayesian model that will 'adaptively' fit such data. Simulations are used to compare the model with a closely related approach, emphasizing on non-normal data, for which literature is scarce.

1.3 Historical Control Data

Testing of chemicals for carcinogenic effects, for example, with mice and/or rats has produced large amounts of data. In these studies, animals in different experimental groups are followed for a fixed period of time, and often, each animal is then labeled as either 1/0 indicating presence/absence of a tumour or some form of defect of interest. Because such experiments are performed with fixed protocols, accumulated data from control groups in different experiments over time present an opportunity for use of such historical data. While seizing the self-presenting opportunity to use historical data in order to sharpen analysis of 'current' experiments, cautious use of such data is encouraged. Such is the focus of some of the work presented in this thesis. There is need to investigate plausible conditions for use of historical data. Issues like the number of historical studies one needs to use or when exactly one should use historical data require a closer look and are investigated herein.

In other settings, historical data accumulate over time, and using these data, the primary aim is to determine limits which will be used to validate or invalidate measurements from new studies. This in a sense, boils down to detecting which studies should be included or excluded in the historical data base. The lesser known, and often ignored tolerance limits (Hahn and Meeker, 1991) are one way of addressing this issue and will be studied in this thesis.

1.4 Organization of Subsequent Chapters

The work presented in this thesis can be considered as divided into two broad categories, the first, and larger part, dealing with flexible modelling techniques with special inclination towards penalized spline methodology. The second category involves use of historical information and the construction of tolerance intervals. The rest of this thesis is therefore structured in the following way. Chapter 2 gives an overview of all the motivating examples used in the thesis. In Chapter 3, general concepts pertaining to smoothing are briefly reviewed. Penalized spline smoothing of longitudinal data in the parallel design case occupies Chapter 4, while similar methodology, geared for the cross-over setting is encountered in Chapter 5. Still within the cross-over design setting, we combine aspects from validation of surrogate markers methodology with some flexible modelling techniques to quantify associations in Chapter 6. Application of the penalized spline methodology to non-normal data takes center stage in Chapter 7, where the neuronal data are analyzed. In Chapter 8 we take a step back to Chapter 4, and re-fit all the models encountered there, this time from a Bayesian perspective. Further, in the same chapter, joint modelling of two longitudinal responses is investigated, with particular attention falling on correlated smoothers. Continuing within the Bayesian framework, but shifting attention to non-normal data, adaptive penalized splines are discussed in Chapter 9. Use of historical control information and tolerance limits are the subjects of Chapter 10 and Chapter 11 respectively. We wrap up with some concluding remarks and possible directions for future research in Chapter 12.

2

Motivating Examples

This chapter presents a detailed explanation of the motivating examples used in this thesis. Specifically focusing on animal studies, and unless otherwise mentioned, all the experiments dealt with here emanate from studies carried out at Johnson and Johnson Pharmaceutical Research and Development in Beerse, Belgium.

2.1 Cardiovascular Safety Experiments Data

2.1.1 Parallel Design Case

One of the main goals of pre-clinical studies is to determine a drug's toxicity through animal testing. Often *in vivo* experiments are carried out with studies of toxicity, focusing on which organs of the body are targeted by the drug. The focus here is on a similar experiment, in the context of cardiovascular safety.

The data come from a two-group parallel design in a cardiovascular safety experiment conducted in dogs. Twenty-eight dogs were implanted with a device for telemetric studies and then orally dosed either with the vehicle (14 animals) or compound (14 animals). The primary objective of the study was to assess the effect of the compound on the QT, a measure of the complete electrical activity of the ventricle of the heart. A drug-induced prolongation of the ventricular repolarization and a concomitant QT prolongation is known to be associated with lethal arrhythmias. Several

other cardiovascular parameters of interest were recorded at 1 minute intervals for 4 hours, resulting in 240 time points per subject. For our purposes, first, attention is given to the heart rate profiles in the control and compound groups, measured as beats per second. In certain cases, administering such a compound may cause some kind of abnormal heartbeat that can be dangerous. One of the primary goals is to be able to detect specific sections, if any, where the compound group significantly differs from the control group. Hence, in line with detecting the possible cardiovascular effects of the compound, it is necessary to detect if the effect is in the earlier or later stages of the experiment. The top left and top right panels in Figure 2.1 show the heart

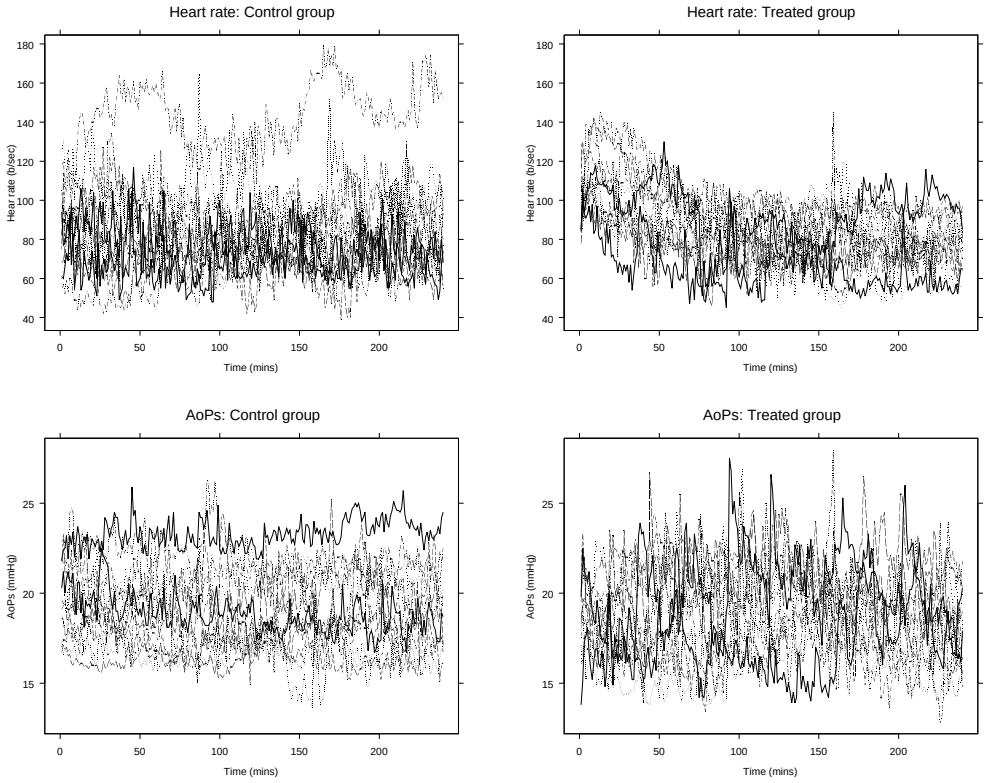


Figure 2.1: *Observed individual profiles for heart rate and AoPs in the control and compound groups.*

rate individual profiles from control and the compound groups respectively. The profiles show substantial between- as well as within-subject variability. These data are used in illustrating use of penalized spline methodology in the parallel design setting

in Chapter 4. In a later stage, another response, AoPs, which is the systolic blood pressure in mmHG (mm mercury) will also be of interest. The individual profiles for both groups are shown in the bottom panels of Figure 2.1. It is known that the heart rate and blood pressure sometimes influence each other, sometimes they compensate each other. In other circumstances they both go higher or lower. The relationship between the two is therefore not a fixed known relationship. These data are considered in Chapter 8, where a joint model between two longitudinally measured outcomes comes under the spotlight.

2.1.2 Cross-over Design Case

Similarly as in Section 2.1.1, the data we consider here were obtained from a cardiovascular safety experiment carried out in dogs. Although the primary objectives of the experiment remain essentially the same as in Section 2.1.1, it is the design of the experiment marking the major difference. The data emanate from a cross-over

Table 2.1: *Williams design for a cross-over study with four dose groups, control (C ≡1), low (L ≡ 2), medium (M ≡ 3), and high (H ≡ 4). The design is replicated twice resulting in a total of 8 animals being used.*

	Period			
Subject	1	2	3	4
1	H	M	C	L
2	M	L	H	C
3	L	C	M	H
4	C	H	L	M

study, where a balanced Latin square Williams design of four experimental groups and four periods is used (see Table 2.1). Eight female beagle dogs, with weights varying between 10.0 and 12.9 kg, were implanted with a device for telemetric studies. The animals were orally dosed with the vehicle or compound (low, medium and high doses) on four successive sessions, separated by a wash-out period of at least 3 days.

Cardiovascular parameters of interest were recorded at 5 minute intervals for 6 hours, hence 72 time points per subject per period. Several parameters were measured, and for our purposes, the response, Tau, is a measure of the relaxation capacity of the heart (in milliseconds) after a contraction. This is a measure of how good or bad a heart relaxes after a contraction. The question of interest here is twofold. First,

an overall measure of the difference between the compound groups and the control group is required. Second, there is the wish to detect specific sections, if any, where the compound groups significantly differ from the control group. Figure 2.2 shows the

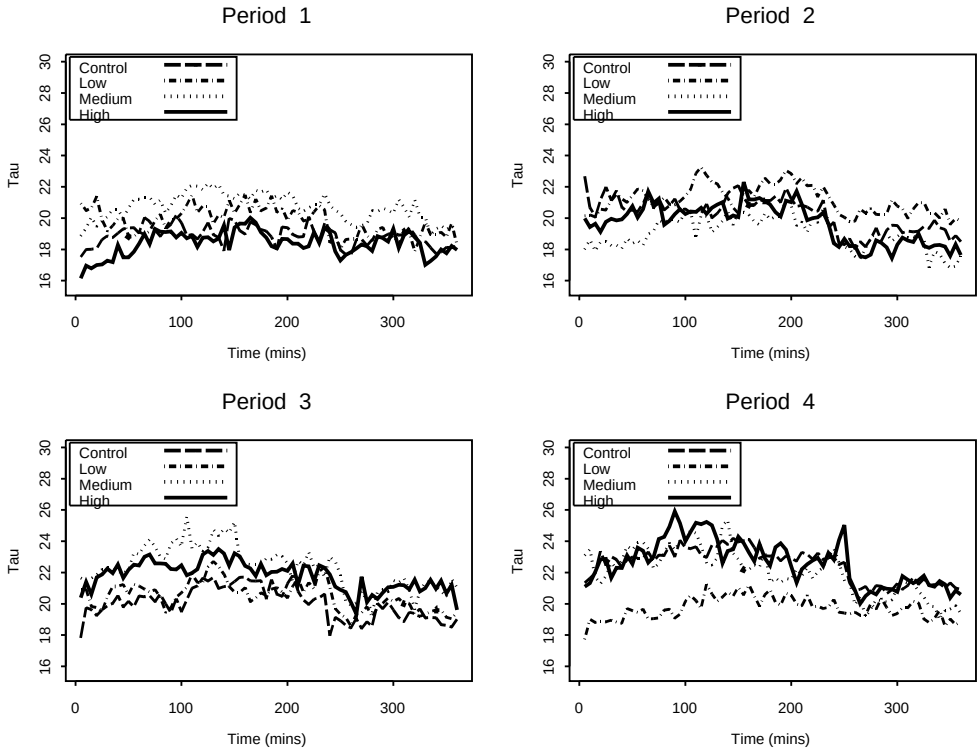


Figure 2.2: *Observed mean profiles for each period and experimental group at each experimental time point.*

observed mean profiles in the various experimental groups for each of the treatment periods. It appears the response tends to increase with period. There also appears to be a relatively large difference between the highest dose group and the low dose group in period 4. Note also that a parametric form for these mean profiles might not be easily determined, hence the need to use more flexible, semiparametric smoothing techniques as discussed in Chapter 5.

2.2 Swim-stress Study

These data come from a pre-clinical experiment with rats, investigating a compound under development for stress-related disorders. The objective of the experiment was to identify the effect of the compound on stress hormones and a series of physiological variables. In the experiment, stress is induced by forcing a rat to swim for 15 minutes in a bath of 20 cm high lukewarm water at a temperature of 25 degrees Celsius, according to a protocol as described by De Groote and Linthorst (2007). The experiment, set

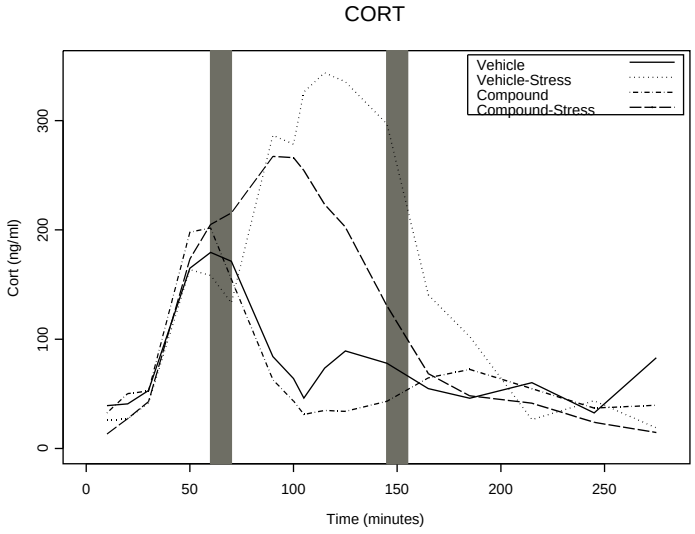


Figure 2.3: *Group-specific mean profiles for CORT values, averaged over different treatment periods. The shaded regions indicate the time windows in which activity was measured before and after the stress induction.*

up as a cross-over study, was designed according to a Latin square with 4 periods and 4 experimental groups (vehicle without stress, vehicle with stress, compound without stress, compound with stress). Forty-five minutes after randomization, the rats were injected with either a vehicle or the compound under consideration. Ten minutes later, half of the rats injected with the vehicle and half of the rats injected with the compound were subjected to the so-called 'swim stress', also depending on group membership. For all eight animals, measurements were taken in order to quantify their stress level. Telemetry measurements (such as heart rate and blood pressure) were recorded continuously and averaged every 5 minutes. Seventeen blood samples

were taken in a fully automated way, leaving the animals completely undisturbed and following a well-defined scheme to sample blood plasma from which corticosterone, henceforth abbreviated CORT, was later extracted and quantified. And finally, rats were also screened for their behavior in a 10 minutes interval by means of a video monitor. For each rat, the percentage of time it has been active was thus determined. The recording of behavior was done twice; a first time at 25 minutes after injection and a second time at 50 minutes after the end of the swim stress. These respective periods are indicated in Figure 2.3 as shaded bars. The graph shows mean profiles

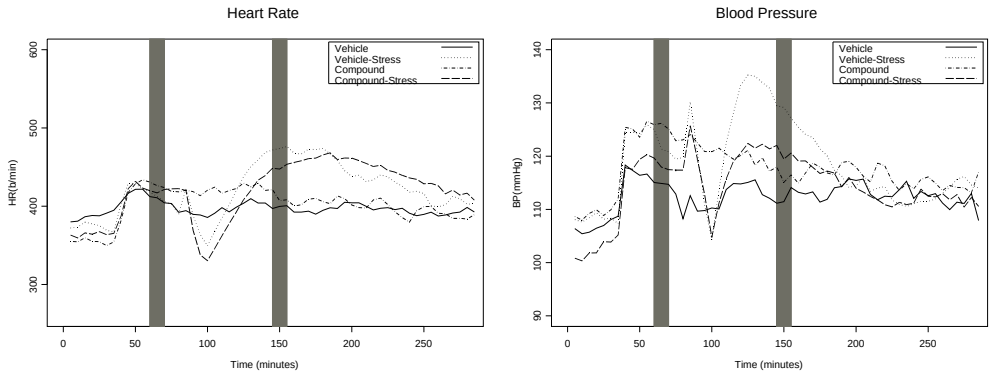


Figure 2.4: *Group-specific mean profiles for heart rate and blood pressure, averaged over different treatment periods. The shaded regions indicate the time windows in which activity was measured before and after the stress induction.*

for CORT in the four experimental groups, averaged over the four treatment periods. Group-specific mean profiles for heart rate and blood pressure are shown in Figure 2.4. The need for flexible modelling is apparent, since no clear parametric structure to model the mean easily comes into the picture. These data are revisited in Chapter 6.

2.3 Electrophysiological Experiment

The data we describe here come from an electrophysiological experiment carried out in the Departamento de Fisiologia, Universidad de Santiago de Compostela in Spain.

One monkey was trained to discriminate between different line orientations (stimuli) (Vazquez *et al.*, 2000). The stimuli (reference and test) consisted of stationary bright line segments presented on a monitor screen in front of the monkey. Reference stimuli were presented with three different orientations (85.5°, 90°, 94.5°). Eight test

stimuli per reference were presented rotated clockwise or counter-clockwise to the reference line in steps of 1.5°. Two bright circles were laterally displayed to the right and to the left of the center of the screen. A trial was initiated when the monkey fixated

Table 2.2: *The sequence of events over time in each trial, starting from -500 to 4500 ms. The region of interest here is 1000-2500 ms.*

Time period	Event
-500-0ms	Control period
0-500ms	Presentation of first (reference) stimulus
500-1500ms	Interstimulus interval (ISI) or delay period
1500-2000ms	Presentation of second (test) stimulus (comparison/decision period)
2000ms +	Subject makes a saccadic eye movement towards
	one of the 2 circles for reward

on a small line centered on the screen. Then, when the fixation line disappeared, the two stimuli, reference and test, each of 500 ms duration, appeared in sequence, separated by a fixed inter-stimulus interval (ISI, 1000 ms). At the end of the second stimulus, the subject had to make a saccadic eye movement towards one of the two circles to indicate whether the orientation of the second stimulus was clockwise (right) or counter-clockwise (left) to the reference stimulus. Monkeys were rewarded for correct discriminations (See Table 2.2). Once trained, extra-cellular single unit activity was recorded in the ventral premotor cortex (VPM).

Interest is in (1) determining the maximum peak activity when the monkey correctly decided that the test stimuli were to the right and to the left of the reference line, and (2) comparing the neural response between behaviorally relevant conditions. The data collected were summarized across the different trials in the form of spike counts per time unit. The period of analysis was between 1000–2500 ms. This time period occupies the last 500 ms of the ISI and the 500 ms of the comparison/decision period (2000–2500 ms), this being the relevant period for this analysis of the correct decisions to the left and to the right. Table 2.2 gives a summary of the events in each trial. The data described in this section is used in Chapter 7.

2.4 Incidences of Alopecia in Rats and Rabbits

Pre-clinical experiments are designed to investigate possible adverse effects of compounds of interest. Such experiments are carried out in animals before the compounds of interest are subjected to human beings. Similar such experiments, conducted under more or less similar conditions and accumulated over time are usually available, and constitute the so-called historical data. Historical data, in particular the control groups from different studies, may be incorporated in the analysis of a new experiment, which we term here the 'current' or 'examined' study. The particular current study considered here involves investigation of the occurrence of the parameter 'alopecia' in rabbits (Dom *et al.*, 2000). Alopecia is a hair loss condition which is characterized

Table 2.3: *Summary of alopecia incidence in the different dose groups of the examined study.*

		Adverse events	
Dose	Sample size	# of animals	%
0	25	1	4
10	20	1	5
40	20	2	10
80	25	5	25

by round patches of complete boldness. The doses selected for the compound of interest in the study were 0, 10, 40 and 80 mg/kg (of weight) with 25, 20, 20 and 25 animals randomly assigned to these dose groups, respectively. The data collected are summarized in Table 2.3. Data on 19 historical control studies are available. The incidence of alopecia in the historical control studies is summarized in Table 2.4. Of these studies, 12 experiments involve rabbits and the other 7 involve rats. Note that only incidences from dose level 0 are considered from the historical studies. In Table 2.4, the column labeled 'Frequency' gives the number of studies of a given sample size and a given number of animals having alopecia (e.g., there is only 1 study of size 10 with no adverse event). These data will be analyzed in Chapter 10.

2.5 In Vitro Ames Test

The data described in this section pertain to the Ames Test, carried out in different experiments over a period of time. The data therefore constitute an accumulation

Table 2.4: *Incidences of alopecia in the historical control studies. Frequency is the number of occurrence of a particular pair of sample size and number of animals having alopecia (the adverse event).*

	Adverse events			Adverse events	
Sample size	# of animals	Frequency	Sample size	# of animals	Frequency
10	0	1	24	1	2
12	3	1	24	2	2
20	1	5	24	3	1
20	2	1	25	1	1
22	2	1	40	4	1
24	0	2	40	9	1

of some form of historical data. The Ames Test is used to determine the mutagenic potential of a substance based on the mutation rate of bacteria that are exposed to the substance. In this particular study, five strains of Salmonella Typhimurium are used. Multiple strains are necessary because different strains mutate differently under different classes of compounds. The revertant bacteria can grow in the absence of an amino-acid (histidine). The bacteria are spread on an agar plate with a small amount of histidine to allow to grow for an initial time. When the histidine is depleted, only bacteria that have mutated to gain the ability to produce its own histidine will survive. The mutagenicity of a substance is proportional to the number of colonies observed. The data collected are the number of colonies on each of the three plates available in any particular experiment.

For our purposes, focus is put on one particular strain, TA100 from the spontaneous mutation group. Data from a total of 153 historical experiments have been collected. For each experiment, measurements from 3 plates, considered here as replicates, were taken. The main purpose of the current exercise is to construct reference ranges for historical control data. These ranges would then be used to validate or invalidate future observations taken from similar experiments. As such, these ranges are used for screening experimental data for atypical values (Amaratunga, 1997). These data are used in Chapter 11, where the issue of reference ranges is addressed by the less frequently used tolerance intervals.

3

General Concepts in Smoothing

Classical regression models are a common feature in the field of statistics. Usually, analysis involves relating a response variable as a function of at least one deterministic explanatory variable, via some parametric relationship. In such circumstances, the relationship is specifically predefined by the model, and the simplicity and interpretability, especially with simple linear regression models, make such models easy choices in practice. However, it is clear that more complex relationships between the response and the explanatory variables may exist, and may be difficult to pick using such parametric models. Use of higher order polynomials may, but not always, alleviate this problem. More flexible techniques to deal with such situations exist in literature. The list of such techniques includes but is not limited to nonparametric regression models, such as, kernel estimation (Azzalini and Bowman, 1993), local polynomial regression (Cleveland and Devlin, 1988; Fan and Gijbels, 1996) and spline smoothing (Freedman and Silverman, 1989). Spline smoothing related methodology is of particular interest in this thesis.

In general, spline smoothing may be divided into three broad categories. First, smoothing splines (see e.g., Green and Silverman, 1994) consider each of the observations as a knot point. When a fixed number of knot points is used, and the model

fitted using ordinary least squares, the second category known as regression splines surfaces. Next to regression splines, when some form of penalization of the knot coefficients accompanied by a roughness penalty is considered, the third class termed penalized splines is obtained (e.g., Eilers and Marx, 1996; Ruppert *et al.*, 2003). In this thesis, we focus on penalized spline smoothing related methodology, tailored in the mixed-model framework. This essentially capitalizes on the connection between linear mixed models and the penalized spline smoother, which has greatly necessitated fitting of smooth functions with relative ease and convenience using software primarily developed for mixed models.

Several issues are of interest regarding smoothing. Particular examples include the number and positioning of the knot points. A more subtle issue is the selection of the smoothing parameter. Different methods have been proposed in literature towards estimating the smoothing parameter as briefly reviewed in Section 3.1.5.

In the following section, we discuss the connection between spline smoothing and mixed models, a concept which will be used inexorably in this thesis. An illustration of some smoothing techniques, whose differences emanate from the way the smoothing parameter is estimated, is given in Section 3.2.

3.1 Spline Smoothing and Mixed-model Approach

Consider a pair of data points (t_j, y_j) of a continuous nature, as measurements of the explanatory and dependent variable, respectively. A nonparametric relationship between both variables may be defined as

$$y_j = f(t_j) + \varepsilon_j, \quad i = j, \ldots, T, \tag{3.1}$$

for some unknown function f, with the assumption that the residual errors ε_j follow a normal distribution with mean 0 and variance σ_ε^2. Determination of f may result from considering the solution of an optimization problem that aims to minimize the penalized residual sum of squares (Hastie and Tibshirani, 1990; Fan and Gijbels, 1996)

$$\sum_{j=1}^{T} (y_j - f(t_j))^2 + \lambda \int \{f''(t)\}^2 dt, \tag{3.2}$$

for a second derivative function $f''(\cdot)$ and the smoothing parameter $\lambda > 0$. The first part in this expression penalizes the lack of fit, and the second, puts a penalty on the roughness of the fit. Values of λ range from 0, corresponding to interpolation of the data, to infinity, implying an ordinary linear regression model. Values in between

provide different levels of smoothing on the function. Thus model complexity is effectively controlled by λ. It turns out that a solution to this minimization problem is the natural cubic spline (Hastie and Tibshirani, 1990).

Another possible way of achieving a smooth function is to allow discontinuities of the derivative function of the approximating function of polynomial functions at specific locations, resulting in polynomial splines. This issue is taken up further in the forthcoming sections.

3.1.1 Polynomial Basis

The piecewise polynomial smoother (Freedman and Silverman, 1989) can be defined as

$$f(t_j) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(t_j),$$

where t_j are design points, $\beta_0, \ldots, \beta_K$ are parameters to be estimated and, $\phi_k(\cdot)$, $k = 1, \ldots, K$ are known functions. A special case, the piecewise linear model, assumes that the basis function $\phi_k(t)$ takes the form

$$\phi_k(t) = (t - \kappa_k)_+ = \begin{cases} 0, & t \leq \kappa_k \\ t - \kappa_k, & t > \kappa_k. \end{cases}$$

The piecewise linear model consists of a set of K knot points in the range of t_j, such that, κ_k is the location of knot k. The piecewise linear smoother can then be expressed as

$$f(t_j) = \beta_0 + \beta_1 t_j + \sum_{k=1}^{K} b_k (t_j - \kappa_k)_+,$$

where, for notational convenience, b_k now denote the coefficients for the knot points. Extension to higher order polynomials is straightforward.

3.1.2 Radial Basis

Instead of the polynomial basis functions touched upon in the preceding section, one can also use the B-spline basis functions (Eilers and Marx, 1996), which apply a difference penalty on coefficients of adjacent B-splines. Yet another basis function, which we shall make use of in Chapter 7, is the so-called radial basis function. For the same set of knots defined as before, and for some univariate function r, and a degree $\wp$, the radial basis function takes the form

$$\phi_k(t) = |t - \kappa_k|^{\wp} = r(|t - \kappa_k|), \quad \text{where} \;\; r(u) = u^{\wp}. \tag{3.3}$$

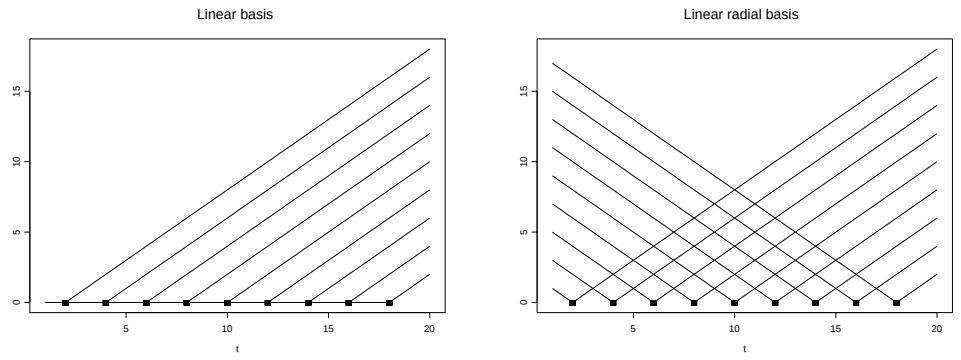


Figure 3.1: *Linear spline basis function (left) and linear radial basis function (right) where the black squares indicate 9 equally spaced knot points.*

The basis function then depends only on the distance $|t-\kappa_k|$ and r, a property which enables extension to higher dimensions with ease. Figure 3.1 illustrates, for 9 equally spaced knots, a linear spline basis and a linear radial basis ($p = 1$) function. Although in principle, different bases are not expected to give different fits, in practice, some bases tend to be numerically more stable in certain cases compared to others. It may therefore be worthwhile for one to experiment with a few different bases.

3.1.3 The Connection Between Penalized Splines and Mixed Models

The popularity of penalized splines derives partly from their connection with mixed models. Here we give a brief introduction to the synergy, a key component in the practical usage of penalized splines. For $1 \leq j \leq T$ and $1 \leq \kappa \leq K$, let us now adopt the following matrix notation, $\boldsymbol{X}_j = [1 \quad t_j]_{1\leq j\leq T}$, $\boldsymbol{Z}_j = [(t_j - \kappa_k)_+]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{'}$, $\boldsymbol{b} = (b_1, \ldots, b_K)^{'}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)$. Stacking these matrices, one below the other, we obtain the representation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}. \tag{3.4}$$

Note that, the generic penalized spline fitting criterion, liable for minimization, may be expressed as (Ruppert *et al.*, 2003)

$$\frac{1}{\sigma_\varepsilon^2} \parallel \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} \parallel^2 + \frac{\lambda^2}{\sigma_\varepsilon^2} \parallel \boldsymbol{b} \parallel^2, \quad \lambda > 0.$$

If one uses a large number of knot points, the model may overfit the data. Further, use of a large number of knots inherently increases the computational burden. A way

of circumventing this problem, is to treat the coefficients b_k as random, drawn from a normal distribution such that $b_k \sim \mathrm{N}(0, \sigma_b^2)$. In that case, (3.4) corresponds to the classical representation of the mixed model. It has been shown that the solution to penalized spline smoother just described corresponds to the BLUP of a mixed model (Eilers and Marx, 1996; Ruppert *et al.*, 2003; Ngo and Wand, 2004), a link enabling fitting penalized splines with mixed-model methodology. Although illustrated here with polynomial basis, a similar connection exists for other basis functions as well.

3.1.4 Penalized Splines from a Bayesian Perspective

To cast the penalized spline model in the Bayesian framework, let us re-visit the mixed-model representation in (3.4). A complete Bayesian model is obtained by assuming prior distributions for all parameters in the model. Specifically, and in many situations, each of the fixed-effects parameters in $\boldsymbol{\beta}$ can be assumed to follow a zero-mean normal distribution with a very large variance (e.g., Crainiceanu *et al.*, 2005*b*). In a similar fashion, the distribution for the variance components is assumed to be inverse gamma. Thus the precision parameter e.g., σ_b^{-2} is assumed to follow a gamma distribution. A choice of the parameters for the gamma distribution close to zero leads to a proper prior.

In classical Bayesian methodology, inference is based on the posterior distribution of parameters given the data. If we denote the data by $\boldsymbol{Y}$ and the vector of parameters by $\boldsymbol{\theta}$, the posterior density of the parameters given the data can be expressed as

$$[\boldsymbol{\theta}|\boldsymbol{Y}] = \frac{[\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{\int [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}] d\boldsymbol{\theta}}, \tag{3.5}$$

where $[\cdot]$ is used to denote a probability density. Therefore, reverting to our problem of expressing the penalized spline model in the Bayesian framework, the posterior distribution of the parameters given the data is such that

$$[\boldsymbol{\beta}, \boldsymbol{b}, \sigma_\varepsilon^2, \sigma_b^2|\boldsymbol{Y}] \propto [\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma_\varepsilon^2][\boldsymbol{b}|\sigma_b^2][\boldsymbol{\beta}][\sigma_\varepsilon^2],$$

which is reflective of the numerator of (3.5). The proportionality sign used comes from the fact that the denominator in (3.5) is a constant since it does not depend on $\boldsymbol{\theta}$. Computation of the denominator is, however, still required for inference. The idea is then to employ methods that can sample from a density that is known only up to a constant and literature abound (Metropolis *et al.*, 1953; Gelman *et al.*, 1995; Robert and Casella, 1999; Zhao *et al.*, 2006). The Bayesian approach will be encountered in Chapter 8 and 9.

3.1.5 Estimation of the Smoothing Parameter

Several aspects ought to be considered when implementing smoothing techniques. Some of these, and arguably the more important ones include, the number, and positioning of knot points, together with the choice of the smoothing parameter. It turns out that selecting the smoothing parameter is a more delicate issue, a subject that has already been extensively covered in literature. Different choices of the smoothing parameter λ will lead to different estimated models. In the context of smoothing splines, an automatic procedure, which leads to a data driven smoothing parameter, is the commonly used cross-validation method (e.g., Ruppert *et al.*, 2003), based on the concept of leaving out a single observation in turn. Alternatively, model selection criteria, e.g., Akaike's Information Criterion (AIC, Akaike (1973)) may be used.

When a linear mixed model is used as a scatterplot smoother, one does not need to use any additional procedure in order to select the smoothing parameter. The amount of smoothing is determined by the ratio of the (restricted) maximum likelihood estimates for both σ_ε^2 and σ_b^2. A similar approach follows from a Bayesian hierarchical model, where one can obtain the posterior mean for λ together with a density estimate for the posterior distribution of λ.

3.2 Illustration of Different Smoothing Techniques

In this section, we will give a brief illustration of the application of some of the smoothing techniques discussed in preceding sections on simulated data. We shall focus on the simple case of independent data, i.e., a single independent variable say t with a corresponding dependent variable y of a continuous nature. The main purpose of this exercise is to show the effect of some of the key factors associated with smoothing, namely, the smoothing parameter and the number of knot points. Following Hart (1997) we simulate data from the model

$$y_j = 3.0 + 0.3\,\sin(2\pi t_j) + \varepsilon_j, \quad j = 1, \ldots, 50, \tag{3.6}$$

where $t_j = (j - 0.5)/50$. The error terms, ε_j are assumed independently and identically distributed as $\mathrm{N}(0, 0.06^2)$. In the left panel of Figure 3.2, the effect of varying the smoothing parameter is illustrated using the cubic smoothing spline, based on the generalized cross-validation technique. For a different number of knots, the linear mixed-model approach is considered and the results are graphically depicted in the right panel of Figure 3.2. It is clear from Figure 3.2 that the smoothing parameter plays a more crucial role in determining the fit.

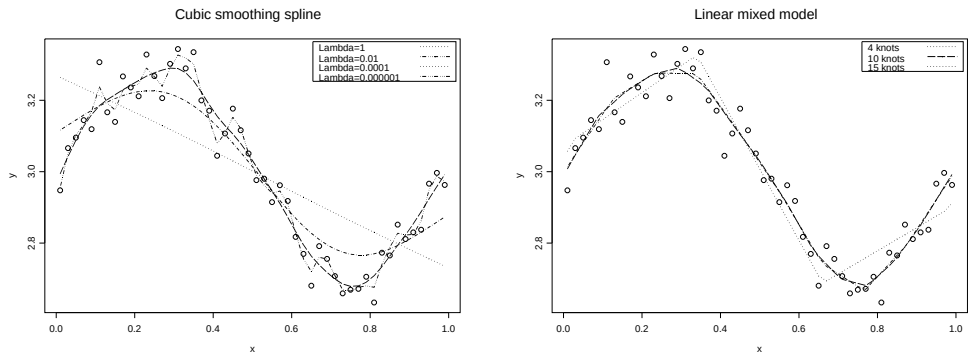


Figure 3.2: *Illustrating the effect of the smoothing parameter and the number of knots in the context of cubic smoothing spline (left) and the linear mixed-model approach (right).*

The number of knots appears to play a less crucial role. Indeed, only a certain minimum number of knots is required to satisfactorily describe the data, see for example, Ruppert (2002).

In all applications considered here, the number of knots will be considered fixed. The number will be generously chosen such that important features in the data are captured, without underestimating the computational complexity accompanying an excessive number of knots. Note however that computationally intensive methods, focusing on choosing the 'optimal' number and positioning of knot points, exist in literature, for example, DiMatteo *et al.* (2001).

Figure 3.3 illustrates fitted curves obtained from different smoothing techniques on a single simulated data set. The methods are; the cubic smoothing splines, the linear mixed-model approach (LMM) and the Bayesian hierarchical linear mixed model (BLMM). The LMM and the BLMM are based on the same 15 equally spaced knots, selected as quantiles of t, while the cubic spline uses all data points as knot points. Included also are the 95% credible intervals from the BLMM. The different smoothing approaches produce almost indistinguishable fitted curves. Of course, the choice of which method to use depends on the problem at hand, considering among many other factors, the type and complexity of the model. A trivial example is when one is confronted with longitudinal data, where smoothing may be warranted, the LMM may be the first port of call. The next chapter deals with smoothing of longitudinal data using penalized spline methodology.

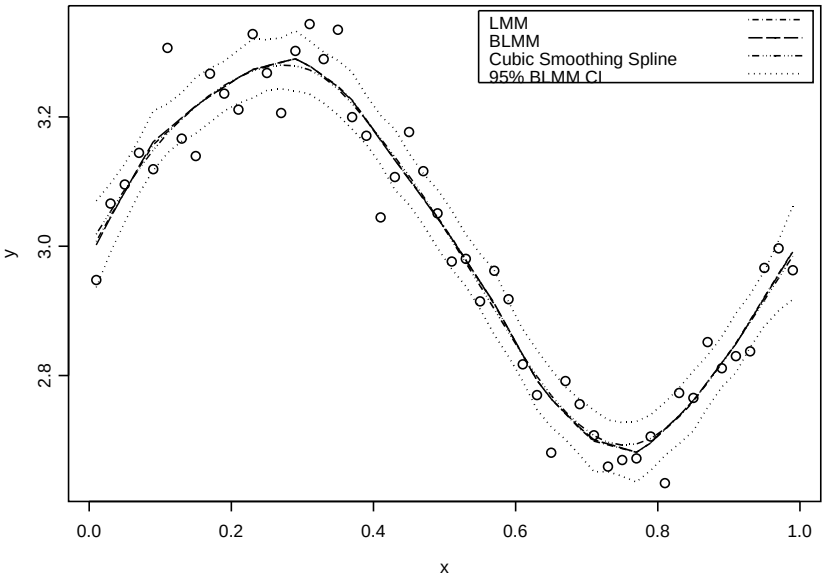


Figure 3.3: *An illustration of using cubic smoothing splines, the linear mixed-model approach and a fully Bayesian mixed model together with the 95% credible intervals from the BLMM.)*

4

Penalized Splines Smoothing of Longitudinal Data

Several pharmacological studies involve experiments aimed at testing for a difference between experimental groups wherein the data are longitudinal in nature, frequently with long sequences per subject. Oftentimes, treatment effect, if present, is not constant over time. This chapter focuses on the analysis of one such experiment, consisting of a long sequence of repeated measurements per subject over time, as shown in Figure 2.1. The present chapter will restrict attention to heart rate profiles in Figure 2.1, addressing the objectives outlined in Section 2.1.1.

For inferences involving the longitudinal nature of the data, it is necessary that one properly captures the form of the evolution of profiles over time. In certain situations, the mean profile can easily be estimated by some parametric function, for example, Dunsmore (1981), Grender and Johnson (1994), and Putt and Chinchilli (1999). These authors use parametric trends in time to model the mean evolution in crossover settings with repeated measurements. However, imposing a parametric function may not always yield satisfactory results. Moreover, finding a suitable parametric form may not be easy. A more flexible approach to modelling the mean, which is situated within the mixed-model framework, is through penalized splines (Eilers and Marx, 1996; Ruppert *et al.*, 2003). Applications of this nature can be found, for ex-

ample, in Zeger and Diggle (1994), Verbyla *et al.* (1999), Ruppert *et al.* (2003) and Durbán *et al.* (2005). The core of the material presented in this chapter is contained in Maringwa *et al.* (2007d).

While a test for a difference between the average profiles gives an overall impression about the equality or inequality of the two functions, it may be of interest to detect particular sections of the profiles that show significant differences between the two groups. This can be achieved, for example, by applying the Wilcoxon- Mann-Whitney test (Lehman and D'Abrera, 1975) at each time point or a parametric version of such a test. However, such an approach suffers from a multiple comparisons problem, especially with long sequences per subject. Importantly, also the correlation structure among the observations is completely ignored.

A comparison based on confidence bands is an attractive alternative. Our approach is to formulate a series of models exhibiting how the group-specific mean profiles could possibly differ. Once an appropriate model is chosen, interest lies in identifying specific time points where the groups differ. For this purpose, we propose the use of simultaneous confidence bands around the fitted models wherein the bands take into account within and between-subject variability, as well as variability arising from smoothing. Such confidence intervals follow as an adjustment to the confidence bands of Ruppert *et al.* (2003) to accommodate the longitudinal nature of the data at hand. Specifically, the bands include components of the variance of the subject-specific effects, used to capture the correlation structure amongst the observations.

Confidence intervals have been used in similar applications, for example, by Lin and Zhang (1999) who, in the context of generalized linear mixed models (Molenberghs and Verbeke, 2005), discuss both frequentist and Bayesian confidence intervals around fitted functions. Guo (2002) constructs Bayesian confidence intervals around the fitted functions in the different groups, while Wood (2004) constructs confidence intervals for generalized additive models fitted using penalized splines.

From Figure 2.1 (top left panel), one profile, in the control group, appears to be outlying. An analysis excluding this particular subject reveals no change in the overall conclusions, therefore the subsequent discussion is based on all data. However, focus on this potentially outlying observation may be an intriguing topic of further research.

Section 4.1 gives a brief review of linear mixed models and semiparametric mixed models. Section 4.2 focuses on the formulation of the semiparametric models considered herein, whilst in Section 4.3 aspects of inference are discussed. An application of the models under discussion is the subject of Section 4.4 with emphasis on a single longitudinally measured response, heart rate, in this case.

4.1 Semiparametric Mixed Models

The data considered in this chapter, described in Section 2.1.1, fall within the framework of continuous longitudinal data, and hence can be modeled by use of a linear mixed model. The general linear mixed effects model can be represented as (Verbeke and Molenberghs, 2000)

$$\begin{cases} \boldsymbol{Y}_i = X_i\boldsymbol{\beta}_i + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\varepsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i), \quad \boldsymbol{b}_1, \ldots, \boldsymbol{b}_n, \quad \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \text{ are independent,} \end{cases} \tag{4.1}$$

where $\boldsymbol{Y}_i$ is the m_i-dimensional response vector of measurements for dog $i (i = 1, \ldots, n)$, X_i and Z_i are $m_i \times p$- and $m_i \times q$-dimensional matrices of known covariates (e.g., time), respectively, $\boldsymbol{\beta}_i$ is a p-dimensional vector of fixed effects, $\boldsymbol{b}_i$ is q-dimensional dog specific vector of random effects and $\boldsymbol{\varepsilon}_i$ is an m_i-dimensional vector of residuals. The matrix $\boldsymbol{G}$ is a general $q \times q$ covariance matrix and $\boldsymbol{\Sigma}_i$ is an $m_i \times m_i$ covariance matrix. Often, $\boldsymbol{\Sigma}_i$ is assumed to be equal to $\sigma_\varepsilon^2 \boldsymbol{I}_{m_i}$, resulting in the so-called conditional independence model.

Given the mean profiles in the top panel of Figure 4.1, it appears a suitable parametric function to describe the mean evolution may not be easily deduced. An appealing alternative is to model the mean with a semiparametric smooth function, $f(t)$, which can be estimated, among others, with penalized splines. Here, we build on the foundation set in Chapter 3.

Let y_{ij} denote the response taken from dog i at time t_{ij} $(j = 1, \ldots, m_i)$. The model of interest can be expressed as $Y_{ij} = f(t_{ij}) + b_{0_i}$, for a smooth function $f(\cdot)$ and subject-specific random intercepts b_{0_i}, accounting for the clustered nature of the observations. The penalized spline representation, based on a truncated power basis, can be written as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \ldots, + \beta_p t_{ij}^{\wp} + \sum_{k=1}^{K} b_k (t_{ij} - \kappa_k)_+^{\wp} + b_{0_i} + \varepsilon_{ij}, \tag{4.2}$$

where $\kappa_1, \ldots, \kappa_K$ are a set of distinct knots in the range of t_{ij}, with $u_+ = \max(0, u)$. Making the assumption $b_k \sim \mathrm{N}(0, \sigma_b^2)$, and $b_{0_i} \sim \mathrm{N}(0, \sigma_{b_0}^2)$ gives rise to the so-called semiparametric mixed model. The truncated lines basis ($\wp = 1$) is simple in formulation, performs adequately in many circumstances (Ngo and Wand, 2004), and therefore is a sensible choice, provided a sufficiently large number of knots is used.

For ease of notation, we adopt the following matrix notation (see also Durbán *et al.*, 2005). Let $\boldsymbol{Y} = [y_{ij}]_{1 \le i \le n, 1 \le j \le m_i}$ be the vector of stacked subject-specific responses,

$\boldsymbol{X} = \left[\begin{array}{cc} 1 & t_{ij} \end{array} \right]_{1\le i\le n, 1\le j\le m_i}$ the corresponding design matrix, and $\boldsymbol{\beta} = \left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right]$ the vector of fixed effects to be estimated. Further, define

$$\boldsymbol{Z}_i = \left[\begin{array}{c} (t_{ij} - \kappa_k)_+ \end{array} \right]_{1\le j\le m_i,\ 1\le \kappa\le K}, \quad \boldsymbol{Z} = \left[\begin{array}{ccccc} \boldsymbol{Z}_1 & \mathbf{1}_1 & 0 & \ldots & 0 \\ \boldsymbol{Z}_2 & 0 & \mathbf{1}_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Z}_n & \vdots & \vdots & \ldots & \mathbf{1}_n \end{array} \right],$$

where $\mathbf{1}$ is an m_i-dimensional column of ones,

$$\boldsymbol{b} = \left[\begin{array}{c} b_1, \ldots, b_K, b_{0_1}, \ldots, b_{0_n} \end{array} \right]', \quad \text{and} \quad \boldsymbol{\varepsilon} = \left[\begin{array}{c} \varepsilon_{11}, \ldots, \varepsilon_{nm_n} \end{array} \right]'.$$

Using this notation, a stacked version of (4.1) becomes $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}$. The correspondence between the penalized spline smoother and the optimal predictor in a mixed-model framework enables conventional software tools for mixed models, e.g., S-Plus, R, or SAS, to be used for fitting the penalized spline model. In particular, we use the function `lme ()` in S-Plus and the MIXED procedure in SAS to fit the models as exemplified in the appendix. Parameters are estimated through maximum likelihood (ML). We make this choice because some of our comparisons include models with different fixed-effects structures, precluding the use of restricted maximum likelihood (REML, Verbeke and Molenberghs (2000)).

Fitting penalized splines by the mixed-model approach has some appealing advantages, such as the automatic determination of the smoothing parameter, a unified framework for inference, and the flexibility with which the models can be extended.

4.2 Semiparametric Models for Mean Evolution

We are interested in investigating whether there is a difference between the two experimental groups, that is, in comparing their average profiles (see top panel of Figure 4.1). Further, we intend to investigate which specific sections of the profiles exhibit significant differences. To test whether the means of the two groups are equal, without loss of generality, we formulate the hypotheses

$$\begin{aligned} &H_0 : f_A(t) = f_B(t), \text{ for all } t, \\ &H_1 : f_A(t) \neq f_B(t), \text{for at least one value of } t. \end{aligned} \tag{4.3}$$

In certain situations, the null hypothesis may be composite, consisting of both tests for fixed effects as well as variance components. The null hypothesis obviously implies a common mean for both groups.

The semiparametric model discussed in Section 4.1 implies that the mean response for each treatment group can be represented by an additive model of two components, a linear component and a smooth component. Figure 4.1 (bottom) illustrates, with hypothetical examples, several possible scenarios related to the evolution of the means over time. In all examples, the mean is a sum of a linear part and a smooth part. In panel A, the two groups have the same mean, implying that the null hypothesis in (4.3) is satisfied. Panel B reveals a pattern in which the means of the two groups differ only by a constant, while in panel C the groups are different in the linear part but the smooth component of the mean is identical. Finally, panel D reveals a pattern in which the means of the two groups have different evolutions over time and the groups are different in both the linear and smooth parts.

In what follows, we formulate linear mixed models following each of the scenarios illustrated in Figure 4.1 (bottom), and in Section 4.3, we discuss corresponding approaches to inference based on these models.

4.2.1 Model 1: Single Curve for Both Groups

Under the null hypothesis in (4.3), it is assumed that there is no difference between the treatment groups, requiring the fit of a single, common curve only (see panel A, Figure 4.1). The following model, based on the linear spline basis is considered:

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k (t_{ij} - \kappa_k)_+}_{f(t_{ij})} + b_{0_i} + \varepsilon_{ij},$$

where $f(t_{ij})$ is the semiparametric smooth function, b_{0_i} is a subject-specific random intercept and ε_{ij} are residuals. The covariance matrix for the random effects $(b_1, \ldots, b_K, b_{0_1}, \ldots, b_{0_n})$ is defined as

$$\boldsymbol{H} = \begin{bmatrix} \sigma_b^2 \boldsymbol{I}_K & 0 \\ 0 & \sigma_{b_0}^2 \boldsymbol{I}_n \end{bmatrix}, \tag{4.4}$$

where $\sigma_b^2 = \text{var}(b_k)$ and $\sigma_{b_0}^2 = \text{var}(b_{0_i})$.

4.2.2 Model 2: Separate Curves With No Time Interaction

In Model 2, we fit two separate curves to the two groups, which are assumed to be 'parallel'. The random effects are assumed to vary at the highest level of the model, hence there is no grouping structure in the corresponding matrix $\boldsymbol{Z}$, defined

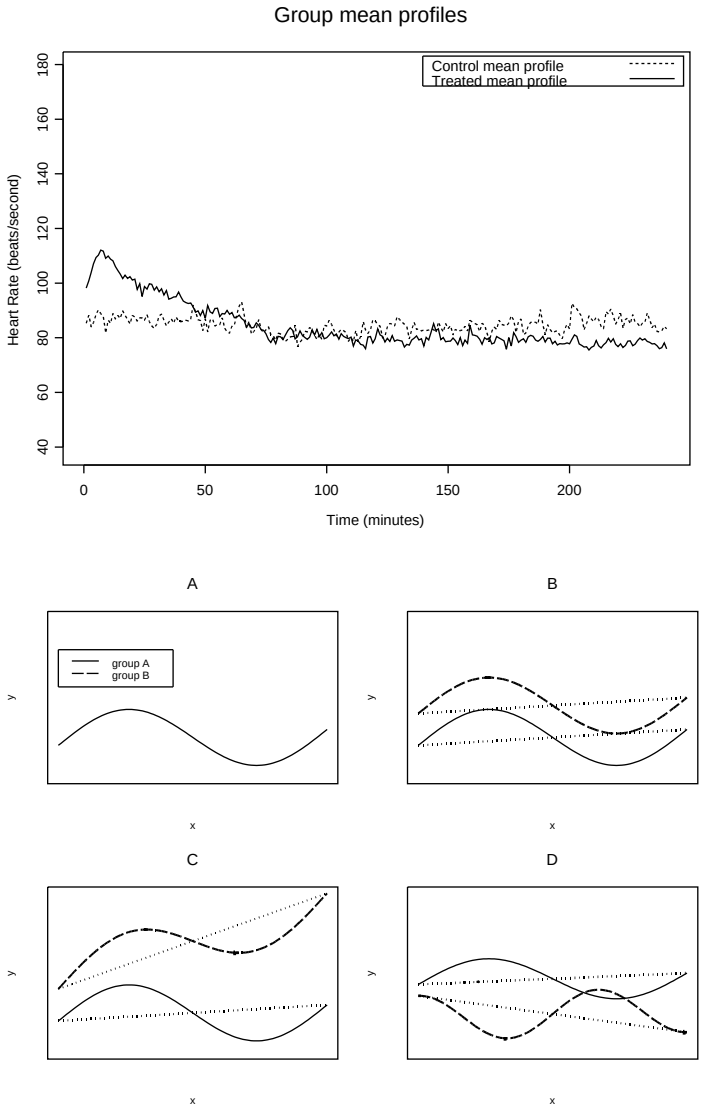


Figure 4.1: *Heart rate group-specific mean profiles (top) and hypothetical examples of the semiparametric models showing the linear and nonparametric parts of the model (bottom). The models in panels A-D illustrate how the group specific could possibly differ.*

in Section 4.1. The penalized spline formulation of this model is

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k (t_{ij} - \kappa_k)_+ + b_{0_i} + \varepsilon_{ij}, & \text{Group A,} \\ (\beta_0 + \beta_{01}) + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k (t_{ij} - \kappa_k)_+ + b_{0_i} + \varepsilon_{ij}, & \text{Group B.} \end{cases}$$

Here, β_{01} is the difference in the group-specific intercepts. Figure 4.1 (panel B) graphically illustrates such a situation. The covariance matrix structure for the random effects is equal to (4.4).

4.2.3 Model 3: Separate Curves With Different Linear Effects but Equal Nonparametric Part

Model 3 assumes that the fixed parts (the linear trends), are different across the two groups but the non-parametric component, responsible for the smoothing, is the same (Figure 4.1, panel C). Hence, the random effects responsible for smoothing vary at the highest level of the model. The models in the two groups can be expressed as

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k (t_{ij} - \kappa_k)_+ + b_{0_i} + \varepsilon_{ij}, & \text{Group A}, \\ (\beta_0 + \beta_{01}) + (\beta_1 + \beta_{11}) t_{ij} + \sum_{k=1}^{K} b_k (t_{ij} - \kappa_k)_+ + b_{0_i} + \varepsilon_{ij}, & \text{Group B}. \end{cases}$$

The covariance matrix for the random effects equals (4.4) as well.

4.2.4 Model 4: Separate Curves Smoothed Separately with the Same Smoothing Parameter

In this model, the smoothed functions are different in the two groups, although the level of smoothing is assumed common. It is further assumed that the random effects are independent. Unlike the models discussed sofar, the matrix $\boldsymbol{Z}$ now has a grouping structure. Suppose we have n_A and n_B subjects in the first and second groups, respectively, with $n = n_A + n_B$, and represent the design matrices as follows,

$$\boldsymbol{Z}_A = \begin{bmatrix} \boldsymbol{Z}_1 & \mathbf{1}_1 & 0 & \dots & 0 \\ \boldsymbol{Z}_2 & 0 & \mathbf{1}_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Z}_{n_A} & \vdots & \vdots & \dots & \mathbf{1}_{n_A} \end{bmatrix}, \boldsymbol{Z}_B = \begin{bmatrix} \boldsymbol{Z}_{n_A+1} & \mathbf{1}_{n_A+1} & 0 & \dots & 0 \\ \boldsymbol{Z}_{n_A+2} & 0 & \mathbf{1}_{n_A+2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Z}_n & \vdots & \vdots & \dots & \mathbf{1}_n \end{bmatrix},$$

which form a block-diagonal matrix

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_A & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_B \end{bmatrix}.$$

The random-effects vector now is

$$\boldsymbol{b} = \left[b_1^A, \dots, b_K^A, b_1^B, \dots, b_K^B, b_{0_1}, \dots, b_{0_n} \right]'.$$

The covariance matrix is a block-diagonal matrix with different entries for the variance components corresponding to the two groups and the random intercept, given by

$$\boldsymbol{H} = \begin{bmatrix} \sigma_b^2 \boldsymbol{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_b^2 \boldsymbol{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_{b_0}^2 \boldsymbol{I}_n \end{bmatrix}.$$

A convenient way to represent this model is by its penalized spline representation

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k^A (t_{ij} - \kappa_k)_+ + b_{0_i} + \varepsilon_{ij}, & \text{Group A,} \\ (\beta_0 + \beta_{01}) + (\beta_1 + \beta_{11}) t_{ij} + \sum_{k=1}^{K} b_k^B (t_{ij} - \kappa_k)_+ + b_{0_i} + \varepsilon_{ij} & \text{Group B,} \end{cases} \tag{4.5}$$

where $\text{var}(b_k^A) = \text{var}(b_k^B) = \sigma_b^2$.

4.2.5 Model 5: Separate Curves Smoothed Separately with Different Smoothing Parameter

The fixed part of Model 5 remains the same as in Model 4 but the smoothed functions are different in the two groups, and also, the level of smoothing is allowed to differ. Hence, the smoothing parameter differs by group. Interestingly, the matrix $\boldsymbol{Z}$ is similar to that in Model 4. However, the covariance matrix is expressed as

$$\boldsymbol{H} = \begin{bmatrix} \sigma_{b^A}^2 \boldsymbol{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_{b^B}^2 \boldsymbol{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_{b_0}^2 \boldsymbol{I}_n \end{bmatrix}.$$

The penalized spline representation of this model is similar to (4.5) and the reference panel is D in Figure 4.1, which also serves as the reference panel for Model 4.

4.3 Model Selection, Inference, and Confidence Intervals and Bands

We consider two inferential approaches. The first is based on formal hypotheses tests, the second on the use of confidence intervals and/or confidence bands. Issues of model selection and hypotheses testing are discussed in Section 4.3.1. Confidence intervals and bands follow in Section 4.3.2.

4.3.1 Model Selection and Hypotheses Testing

The null hypothesis specified in (4.3) implies a common mean for both treatment groups. The alternative models given in Section 4.2 describe a hierarchy of more complicated models. To test for a difference between the two groups, first select the best fitting model among the models described in Section 4.2, by employing a commonly used selection criterion, AIC (Akaike, 1973; Burnham and Anderson, 2002). The smaller the AIC value, the better the model. Next to this, if a formal test is required, the model declared 'best' may then be compared to Model 1, the null model, using e.g., a likelihood ratio test.

Models we consider possess differences in both fixed effects and the nonparametric part, as reflected by the complexity of the smoothing matrix. As mentioned by Eilers and Marx (1996), the idea behind the AIC is to correct the log-likelihood of a fitted model for the effective number of parameters. To this end, we consider the effective number of parameters in the model (Eilers and Marx, 1996; Ruppert *et al.*, 2003; Lee *et al.*, 2006). For convenience and to distinguish it from the marginal AIC (Wager *et al.*, 2005) reported by the mixed-model software, denote the resulting AIC as an adjusted AIC, abbreviated AIC_{adj}.

Let $\boldsymbol{C} = [\boldsymbol{X} \quad \boldsymbol{Z}]$ be the design matrix with appropriate fixed-effects components and the corresponding smoothing matrix, as defined in the different models. The effective number of parameters is then defined as

$$\text{E}_p = \text{trace}\left((\boldsymbol{C}^T\boldsymbol{C} + \hat{\lambda}\boldsymbol{D})^{-1}\boldsymbol{C}^T\boldsymbol{C}\right), \tag{4.6}$$

where $\hat{\lambda} = \widehat{\sigma_\varepsilon^2}/\widehat{\sigma_b^2}$ and $\boldsymbol{D} = \text{diag}(0, 0, 1, 1, 1, \ldots, 1)$, with dimension $K + 2$. For some log-likelihood LL, an adjusted AIC is then

$$\text{AIC}_{adj} = -2LL + 2E_p.$$

Note that while the marginal AIC penalizes only for the number of parameters in the model (fixed effects and variance components), the penalty term of AIC_{adj} in (4.6) takes smoothing into account by including the design matrix for smoothing, $\boldsymbol{Z}$ via the matrix $\boldsymbol{C}$. Note also that for other models, for example, Model 4, a suitable variation of $\boldsymbol{D}$ is required. Since the smoothing matrix is block-diagonal, we can define $\boldsymbol{D}^* = \boldsymbol{I}_2 \otimes \boldsymbol{D}$, where $\otimes$ defines a kronecker product with an identity matrix of dimension 2.

For tests involving fixed effects only, a conventional likelihood ratio test with an appropriate chi-square distribution is considered. Table 4.1 gives an overview of how one can test each of the alternative models against the null model. Appropriate null

hypotheses of interest, involving variance components in our case, focus on equality of variance components and not on zero variances. This is a non-boundary situation and hence conventional chi-squared null distributions apply. For example, for Model 5, involving different variance components in the two groups, a test involving both fixed effects and variance components is required. In such a case, the null hypothesis can be formulated as

$$H_0 : \beta_{01} = 0, \quad \beta_{11} = 0, \quad \sigma^2_{b^A} = \sigma^2_{b^B}. \tag{4.7}$$

For $\sigma^2_{b^A} = \sigma^2_{b^B} + \Delta$, the hypothesis $\sigma^2_{b^A} = \sigma^2_{b^B}$ is equivalent to testing $H_0 : \Delta = 0$. The distribution of the LRT statistic is then χ^2_3 since (4.7) is equivalent to

$$H_0 : \beta_{01} = \beta_{11} = \Delta = 0.$$

Table 4.1: *Inference about the mean structure and variance components. Illustrating the null hypotheses corresponding to testing each of the alternative models against Model 1.*

Model	Model description	Null hypothesis	Null distribution
5	Different smoothing parameters	$H_0 : \beta_{01} = \beta_{11} = 0,\ \sigma^2_{b^A} = \sigma^2_{b^B}$	χ^2_3
4	Same smoothing parameters	$H_0 : \beta_{01} = \beta_{11} = 0$	χ^2_2
3	Different linear components	$H_0 : \beta_{01} = \beta_{11} = 0$	χ^2_2
2	Parallel means	$H_0 : \beta_{01} = 0$	χ^2_1
1	Common model		

4.3.2 Pointwise Confidence Intervals and Simultaneous Confidence Bands

Consider the penalized spline model in the mixed-model form as in Section 4.1, re-expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_b\boldsymbol{b}_b + \overbrace{\boldsymbol{Z}_{b_0}\boldsymbol{b}_{b_0} + \boldsymbol{\varepsilon}}^{\boldsymbol{\varepsilon}^*}, \tag{4.8}$$

such that,

$$\text{Cov}(\boldsymbol{\varepsilon}^*_i) = \boldsymbol{M}_i + \boldsymbol{\Sigma}_i = \boldsymbol{R}^*_i, \tag{4.9}$$

where $\boldsymbol{Z}_b$ and $\boldsymbol{Z}_{b_0}$ are matrices corresponding to smoothing and random intercepts, respectively, $\boldsymbol{M}_i = \sigma^2_{b_0}\boldsymbol{J}$ with $\boldsymbol{J}$ an $m_i \times m_i$ matrix of ones. Construction of pointwise

confidence intervals as well as simultaneous confidence bands requires the covariance for the vector of contrasts between the estimated and true parameters for the fixed and random effects such that (Ruppert *et al.*, 2003)

$$\mathrm{Cov}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{b}}_b - \boldsymbol{b}_b \end{bmatrix}\right) \simeq \left(\boldsymbol{C}^T \hat{\boldsymbol{R}}^{-1} \boldsymbol{C} + \hat{\boldsymbol{B}}\right)^{-1}, \tag{4.10}$$

where $\boldsymbol{C}$ is a design matrix containing linear time effects and a truncated line basis, $\hat{\boldsymbol{R}}$ is the residual covariance and $\hat{\boldsymbol{B}}$ is a matrix constructed from variance components corresponding to smoothing. Assuming a conditional independence model with random intercept only, and an equal number ($m = m_i$) of measurements per subject, it can be shown that

$$\left(\boldsymbol{C}^T \hat{\boldsymbol{R}}^{-1} \boldsymbol{C} + \hat{\boldsymbol{B}}\right)^{-1} = \hat{\sigma}_\varepsilon^2 \left[\sum_{i=1}^{n} \left\{ \boldsymbol{C}_i^T \left(\boldsymbol{I}_{m\times m} - \frac{\hat{\sigma}_{b_0}^2}{\hat{\sigma}_\varepsilon^2 + m\hat{\sigma}_{b_0}^2} \boldsymbol{J}_{m\times m} \right) \boldsymbol{C}_i \right\} + \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_b^2} \boldsymbol{D} \right]^{-1}, \tag{4.11}$$

where

$$\boldsymbol{C}_i \equiv \begin{bmatrix} 1 & t_1 & (t_1 - \kappa_1)_+ & \dots & (t_1 - \kappa_K)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_m & (t_m - \kappa_1)_+ & \dots & (t_m - \kappa_K)_+ \end{bmatrix}, \quad \boldsymbol{C} \equiv [\boldsymbol{C}_i]_{1\le i\le n},$$

and $\boldsymbol{D} = \mathrm{diag}(0, 0, 1, \dots, 1)$. The simultaneous confidence bands are based on simulations assuming a multivariate normal distribution for the vector of contrasts between the estimated and true parameters for both fixed and random effects. Such bands allow joint statements, for example, that $f(t_1)$ is contained in some interval and simultaneously $f(t_2)$ is contained in another interval with some level of confidence (e.g., 95%).

Let $\boldsymbol{g} = (g_1, \dots, g_T)$ be a set of values for which a simultaneous confidence band for $\boldsymbol{f_g}$ is required. It is assumed that, approximately (Ruppert *et al.*, 2003)

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{b}}_b - \boldsymbol{b}_b \end{bmatrix} \sim \boldsymbol{N}\left\{\boldsymbol{0}, \left(\boldsymbol{C}^{\mathbf{T}} \hat{\boldsymbol{R}}^{-\mathbf{1}} \boldsymbol{C} + \hat{\boldsymbol{B}}\right)^{-\mathbf{1}}\right\}. \tag{4.12}$$

A 100(1-α)% simultaneous confidence band for $\boldsymbol{f_g}$ can be obtained as

$$\left[\hat{f}(g_l) \pm \tilde{h}_{(1-\alpha)} \widehat{\mathrm{stdev}}\{\hat{f}(g_l) - f(g_l)\}\right]_{1\le l\le T},$$

where $\tilde{h}_{(1-\alpha)}$ is the $1-\alpha$ quantile of (Ruppert *et al.*, 2003)

$$\sup_{t\in\chi}\left|\frac{\hat{f}(t)-f(t)}{\widehat{\text{stdev}}\{\hat{f}(t)-f(t)\}}\right| \approx \max_{1\le l\le T}\left|\frac{\left(\boldsymbol{C_g}\begin{bmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{b}}_b-\boldsymbol{b}_b\end{bmatrix}\right)_l}{\widehat{\text{stdev}}\{\hat{f}(g_l)-f(g_l)\}}\right|, \qquad (4.13)$$

with $\boldsymbol{C_g}$ constructed in a similar way to $\boldsymbol{C}_i$ above. Simulations from (4.12) and computation of (4.13) can be repeated for N times to obtain $\tilde{h}^1_{1-\alpha},\ldots,\tilde{h}^N_{1-\alpha}$. The value with rank $(1-\alpha)N$ becomes our $\tilde{h}_{(1-\alpha)}$.

Confidence Bands on the Difference Between Both Groups

For comparison of the group-specific curves, confidence intervals on the difference between both curves may be more informative. Let $\boldsymbol{f}_A$ and $\boldsymbol{f}_B$ be the respective group-specific profiles. Using appropriately defined matrices, we can write,

$$\begin{aligned}\begin{bmatrix}\boldsymbol{f}_A\\ \hline \boldsymbol{f}_B\end{bmatrix} &= \boldsymbol{X\beta}+\boldsymbol{Zb}\\ &= \boldsymbol{C}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{b}\end{bmatrix}, \quad \text{where} \quad \boldsymbol{C}=[\boldsymbol{X}\quad\boldsymbol{Z}].\end{aligned}$$

The difference $\boldsymbol{f}_d=\boldsymbol{f}_A-\boldsymbol{f}_B$ can be obtained by defining an appropriate contrast matrix $\boldsymbol{L}$. Let $\boldsymbol{I}_T$ be an $T\times T$ identity matrix. Define the contrast matrix $\boldsymbol{L}=[\boldsymbol{I}_T \quad -\boldsymbol{I}_T]$. It then follows that the difference is

$$\begin{aligned}\boldsymbol{f}_d=\boldsymbol{f}_A-\boldsymbol{f}_B &= [\boldsymbol{LX}\quad\boldsymbol{LZ}]\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{b}\end{bmatrix}\\ &= [\boldsymbol{X}^*\quad\boldsymbol{Z}^*]\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{b}\end{bmatrix},\end{aligned}$$

which takes a familiar look of the linear mixed model. The construction of pointwise confidence intervals and simultaneous confidence bands now follows exactly as defined for the group-specific profiles, for example, with $\boldsymbol{C}_i$ replaced with elements from $\boldsymbol{C}^*=[\boldsymbol{X}^*\quad\boldsymbol{Z}^*]$.

4.4 Application to the Cardiovascular Safety Experiment Parallel Design Case

The models as well as the inferential machinery discussed in this chapter are illustrated on the data example described in Section 2.1.1. Aspects of model selection, hypotheses testing and simultaneous confidence bands occupy the forthcoming sections of the present chapter. Although the essence of this thesis is to focus on data driven flexible modelling techniques, a comparison with some classical parametric models is not harmful.

All the models are fitted with the time scale in hours. For reasons of flexibility, smoothing is done with 40 equally spaced knots, selected as quantiles of the time variable (Ruppert, 2002).

4.4.1 Model Fitting and Selection

In this section, without losing focus on the semiparametric models, which are of more interest here, a brief comparison of these models with some possible parametric models one can fit to the data is given.

The various semiparametric models fitted are shown in Figure 4.2. The fitted models in the two groups for Model 2 practically overlap, similar in appearance to Model 1; we therefore omit the plot of Model 1. It can be observed that the fitted functions for Models 4 and 5 are very similar, suggesting that different levels of smoothing in the two groups may not be necessary.

A formal test for this can be conducted, a point to which we will return. The marginal and adjusted AIC values in Table 4.2 are used as exploratory tools for discriminating amongst candidate models. We first illustrate the difference between the AIC and AIC_{adj}. Using the marginal AIC, Model 4 and 5 have the smallest but very similar AIC values, however, we are inclined to select Model 4, the more parsimonious between them, as the 'best'. Recall, in Model 4 we smooth both groups separately, with a single smoothing parameter, while Model 5 allows for varying levels of smoothing in the two groups. Compared with Model 3, Model 4 has a smaller marginal AIC. However, this will always be the case since both models have the same number of parameters while Model 4 has a higher likelihood because of the separate smoothers for the groups. Hence, use of the marginal AIC may not be appropriate, instead, the adjusted AIC should be used in this case. When AIC_{adj} is used, Model 5 has a slightly lower AIC_{adj} value than Model 4 (see Table 4.2) and could therefore be preferred. A formal test between both models follows in Section 4.4.2.

Under the parametric models, a full factorial structure in time, i.e., unstructured

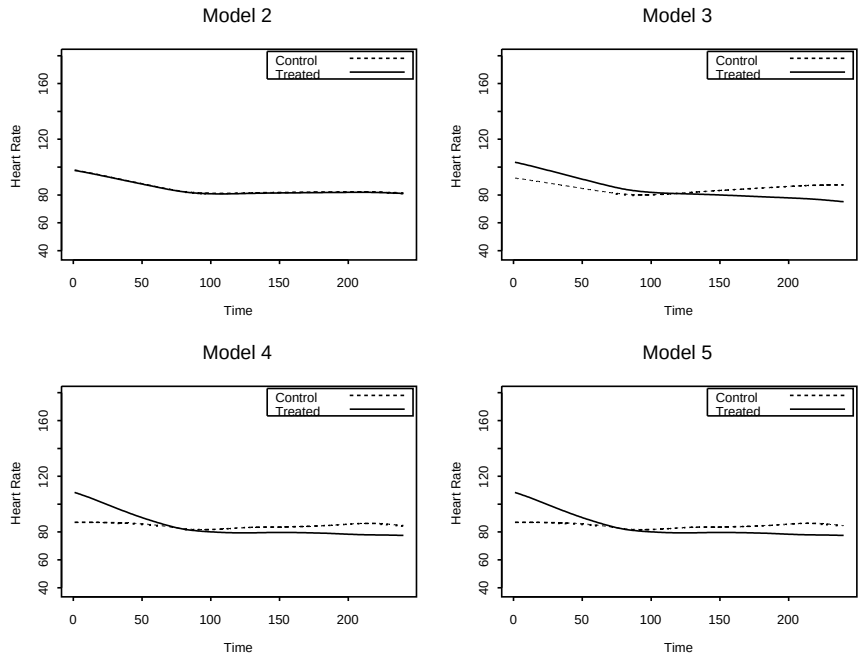


Figure 4.2: *Fitted models in the control and treated groups for the different models under the alternative hypothesis.*

mean, and classical polynomials up to order four are considered. The AIC and -2LL values for the parametric models are also given in Table 4.2. In all cases, a random intercept model is considered. While according to this example, at least a cubic polynomial appears to provide a better fit than Model 4, this may not be the case in other examples where non-linear relationships are more complex. Even in this case, it is important to note that Model 4 tells us more about the underlying structural relationship between the two groups (in terms of smoothing) compared to, for example, the knowledge that a cubic polynomial fits better in both groups. In general, we feel use of semiparametric models offers more flexibility, and the formulation considered here provides a better understanding of the underlying profiles. As such, we revert to the main focus of this chapter, the use of semiparametric mixed models and in the following, more focus is put on Model 4.

A random-intercept model only assumes a shift in subject-specific profiles, a rather restrictive assumption. More complex models, for example, including subject-specific random intercepts and slopes, can be considered. In addition to Model 4, let us

Table 4.2: *Marginal AIC values denoted by AIC and minus twice loglikelihood values for each of the different models. AIC_{adj} gives an adjusted AIC based on the effective number of parameters in the model.*

	Semiparametric models				
	Model 1	Model 2	Model 3	Model 4	Model 5
-2LL	52648.8	52648.8	52074.0	51947.6	51947.7
AIC	52658.8	52660.8	52088.0	51961.6	51963.7
AIC_{adj}	52664.9	52669.9	52094.7	51978.5	51977.8
	Parametric models				
	Unstruct. mean	Linear	Quadratic	Cubic	4^{th} order
-2LL	51649.6	52536.0	52031.7	51935.1	51920.8
AIC	52613.6	52546.0	52047.7	51955.1	51944.8

include subject-specific random slopes b_{1_i} and express the so-obtained Model 4*a* as

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k^A (t_{ij} - \kappa_k)_+ + b_{0_i} + b_{1_i} t_{ij} + \varepsilon_{ij}, & \text{A,} \\ (\beta_0 + \beta_{01}) + (\beta_1 + \beta_{11}) t_{ij} + \sum_{k=1}^{K} b_k^B (t_{ij} - \kappa_k)_+ + b_{0_i} + b_{1_i} t_{ij} + \varepsilon_{ij} & \text{B,} \end{cases}$$

with covariance matrix

$$\boldsymbol{H} = \begin{bmatrix} \sigma_b^2 \boldsymbol{I}_K & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_b^2 \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \text{Blockdiag}(\boldsymbol{G})_{1 \le i \le n} \end{bmatrix},$$

where

$$\boldsymbol{G} = \text{Cov}(b_{0_i}, b_{1_i})' \equiv \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1}^2 \\ \sigma_{b_0 b_1}^2 & \sigma_{b_1}^2 \end{bmatrix}.$$

This model will be revisited and discussed further in Section 4.4.2.

4.4.2 Hypotheses Testing and Confidence Intervals

Let us now focus on hypotheses testing and on the construction of confidence intervals/bands for the selected Model 4. The conclusion that we do not require separate smoothing parameters in both groups can be substantiated by a formal test. Following notation in Sections 4.2.4 and 4.2.5, the estimated variance components for

Model 4 and Model 5 are given by

$$\hat{\boldsymbol{H}} = \begin{bmatrix} 7.87\boldsymbol{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 7.87\boldsymbol{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 245.78\boldsymbol{I}_n \end{bmatrix}, \quad \hat{\boldsymbol{H}} = \begin{bmatrix} 7.36\boldsymbol{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 9.39\boldsymbol{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 246.13\boldsymbol{I}_n \end{bmatrix},$$

respectively. A likelihood ratio test for the null hypothesis of non-differential variance components asymptotically follows a χ^2_1. The associated p-value is 0.80, implying that there is no need for separate smoothing parameters for the two groups.

The fact that we have selected a model other than Model 1 already suggests a difference between the two groups. However, a formal test supporting this claim may be more appealing. Thus, we test the hypothesis that we have one common average curve against separate average curves in the two groups that exhibit the same amount of smoothing. The difference in the double loglikelihood values between the two models equals 701.20, which is highly significant compared to a χ^2_2 (see Table 4.1). Therefore, we need separate curves in the two groups, implying the two groups are not the same. Note that, apparently, the null hypothesis should also include the restriction, $b_K^A = b_K^B$. However, we argue that these coefficients are not treated as parameters in the model, only their variance is. Indeed, this is to be understood in the same spirit as for the random intercepts b_{0_i}, where the variance of b_{0_i} is the model parameter, not the b_{0_i} themselves.

Turning to Model 4a, the estimated random-effects covariance matrix is

$$\boldsymbol{H} = \begin{bmatrix} 9.28\boldsymbol{I}_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 9.28\boldsymbol{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \text{Blockdiag}(\hat{\boldsymbol{G}})_{1\leq i\leq n} \end{bmatrix}, \text{ where,}$$

$$\hat{\boldsymbol{G}} = \begin{bmatrix} 297.85 & -29.09 \\ -29.09 & 16.96 \end{bmatrix}.$$

A formal test to determine the need for the random slope can be conducted between Models 4 and 4a, using a likelihood ratio test based on a mixture of chi-squares, $\frac{1}{2}\chi^2_2 + \frac{1}{2}\chi^2_3$ (Verbeke and Molenberghs, 2003). Note that the null Model 4 contains two variance components. The test statistic is 1110.14, yielding a highly significant result, implying the need for random slopes. Further, Model 4a has a relatively lower AIC_{adj} (50870.9) compared to other models in Table 4.2. Figure 4.3 displays the observed mean profiles in the two groups as well as fitted functions from Model 4 and Model 4a. The fitted curves appear to describe the mean evolution rather well.

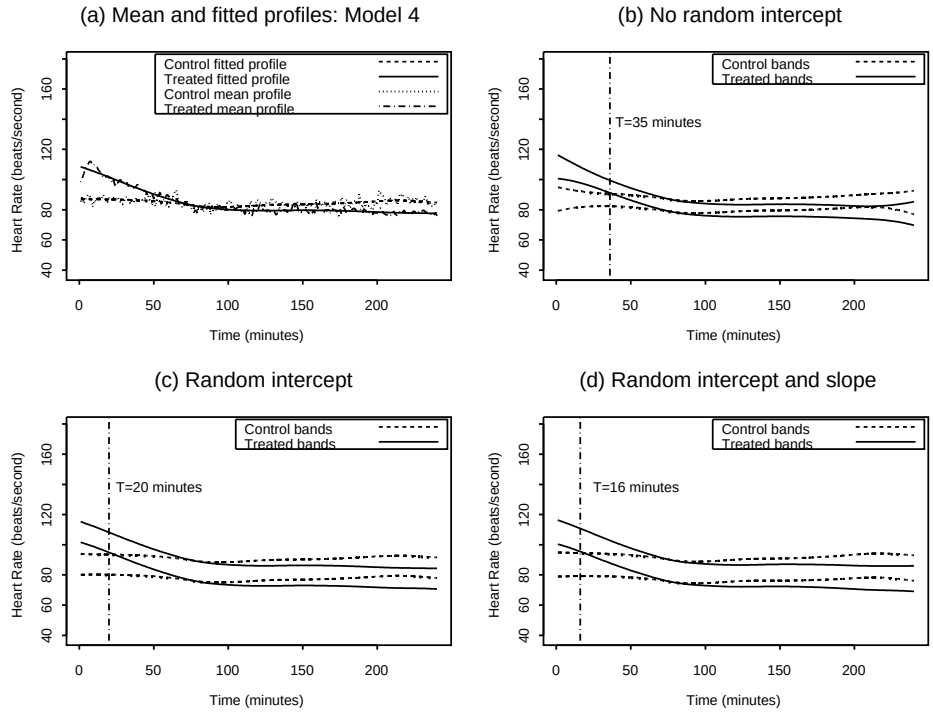


Figure 4.3: *Fitted profiles in the control and treated group and corresponding confidence bands constructed with and without subject-specific random effects.*

The next step involves construction of confidence bands to identify particular sections of the profiles where significant differences occur. By observing the mean profiles in two groups (Figure 4.1), one can expect that between 50 and 200 minutes, the two profiles would not be significantly different from each other, while earlier than 50 minutes, significant differences are not uncommon. However, to have a clearer picture as well as support for such conclusions, using results in Section 4.3, 95% confidence bands for the population profiles are constructed. Three independent simulations yield values of $\tilde{h}_{(1-\alpha)}$ taking values 2.2360, 2.2261 and 2.2256, therefore the simultaneous bands are estimated to be approximately $2.23/1.96 = 1.14$ times wider than the corresponding pointwise confidence intervals. Although simultaneous bands are slightly wider (as expected), there is no major difference in conclusions between the pointwise and the simultaneous bands in this case.

The construction of confidence bands for Model 4a requires an appropriate modi-

fication of (4.9). For any time points q and r $(1 \leq q, r \leq m)$, $q \leq r$, the (q, r) element of the matrix $\boldsymbol{M}_i$ is given by:

$$\sigma^2_{b_0} + \sigma^2_{b_0 b_1}(t_q + t_r) + \sigma^2_{b_1}(t_q t_r).$$

The resulting $\boldsymbol{R}^*$ does not yield a readily invertible form, unlike in the random-intercept model. However, calculations can still be done using the general form (4.10), with the rest following the steps in the random-intercept model.

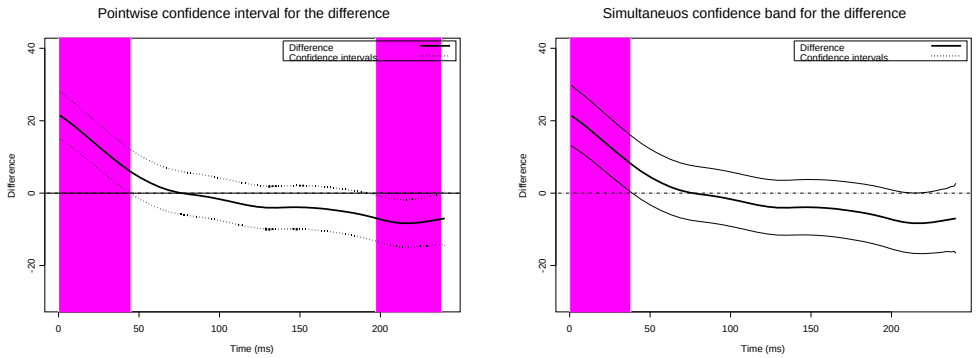


Figure 4.4: *Random intercept and slope model (Model 4a): 95% pointwise confidence intervals and simultaneous bands for the difference between the compound and control groups. The shaded areas indicate sections where significant differences between the groups are observed.*

Figure 4.3 (panel (b)) shows confidence bands, constructed without accounting for the random intercept. This only serves to show the effect of underestimating variability by not including the variance for the random intercept as indicated by the shift in the vertical lines, approximating the start of non-overlapping sections. It can be observed that, when the random intercept is included (panel (c)), more overlapping sections appear and the line is shifted more to the left. Similarly, inclusion of the random slopes results in wider bands with even more overlapping sections, as depicted in Figure 4.3, panel (d).

Further, confidence intervals and bands for the difference between both groups have also been calculated and displayed in Figure 4.4. The difference between the two groups is more visible from Figure 4.4, where significant differences appear approximately within the first 35 minutes. Following the above discussion, we conclude that the two groups appear to differ significantly only in the early stages of the experiment, i.e., approximately within the first 35 minutes. It appears the heart rate

for the treated animals is higher, but only very early in the experiment, while no significant differences are observed at any other time during the experiment. Such a joint conclusion may only be made with reference to simultaneous confidence bands and not pointwise intervals.

4.5 Discussion

The models fitted here make use of the correspondence between the linear mixed model and the penalized spline smoother. We have shown that one can formulate different possible situations, illustrating how groups can differ, under the alternative hypothesis. From these, a 'suitable' model was selected based on the AIC criterion and all inference based on that particular model.

Our aim was to compare group-averaged profiles. As such, although models involving subject-specific curves may be appealing, it is not our interest to predict subject-specific profiles. As a result, we have focused on the random-intercept model, extending the discussion to a random-intercept and slope model.

The problem of testing for a difference in the average profiles, by first fitting an overall common average curve under the null, may in certain cases involve testing for both fixed effects and variance components. This can also be done by way of simulations, as suggested by Ruppert *et al.* (2003) and Crainiceanu *et al.* (2005a), but we have considered likelihood ratio tests based on asymptotic chi-square distributions.

The models we have considered could also be tested for in a hierarchical way, where one would attempt to reduce the most complex Model 5 in a number of steps. Suppose a comparison between Model 4 and Model 3 is required. It is then interesting to note that the two models contain exactly the same number of parameters. From a parametric point of view, although the group-specific random effects for smoothing for Model 4 are independent, only a single variance component is estimated, as in Model 3. Hence, in a testing problem, the fact that $b_k^A \neq b_k^B$ in Model 4 is of no consequence since these coefficients are not treated as parameters. This situation presents a challenge in case one wants to make a formal test (e.g., a likelihood ratio test) to move from Model 4 to Model 3. The difference in the number of parameters is 0 and a formal parametric test is not straightforward. In such situations, differentiation between the models can be done based on information criteria adjusted for the effective number of parameters in the model.

The detection of particular sections of the profiles showing significant differences is achieved by constructing confidence intervals and bands. Pointwise confidence intervals suffer from the drawbacks associated with multiple comparisons, wherein the

overall significance level needs to be protected. To counterbalance this, simultaneous confidence bands have been discussed, specifically focusing on application to the random intercept and random intercept and slope models. Adjustments to the confidence bands of Ruppert *et al.* (2003) to include the random intercept or random intercept and slope in case of longitudinal data has been discussed. It is worth noting that more complex models, for example, models including serial correlation (Verbeke and Molenberghs, 2000), can also be considered. In such a case, parametric or semiparametric models for serial correlation would be interesting to investigate.

It is common in longitudinal studies to be confronted with missing data. A host of methods dealing with the problem exist in literature. Among them, likelihood-based approaches in longitudinal studies only require that the missing data mechanism can be considered as missing at random (MAR). We refer to Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005) and other missing data references cited there for an in-depth exposition.

5

Analysis of Cross-over Designs Using Semiparametric Mixed Models with Serial Correlation within Periods

Chapter 4 focuses on smoothing longitudinal data in a parallel design setting using penalized spline methodology, couched in the mixed-model framework. In this chapter, we demonstrate the versatility and extendability of such models to different study designs, in particular, to the often used cross-over design. Application is illustrated using the data described in Section 2.1.2. The current chapter hinges on material in Maringwa *et al.* (2008e).

In a cross-over trial, each unit or subject receives a sequence of experimental treatments, in randomized order. The main advantage of a cross-over trial is that treatments are compared within subject such that the difference between treatment measurements removes any subject effect from the comparison. Giving the treatments

in random order helps to minimize, remove and/or estimate effects due to time period or from carry-over treatment effect from earlier into later time periods. The theory is well established, whether for two treatments and two periods, or for higher-order designs (Jones and Kenward, 2003).

Here, a particular case of a cross-over design with a salient feature of a relatively long sequence of repeated measurements within treatment period is considered. Focus is put on modelling the mean evolution using semiparametric mixed models, accounting for correlation between observations through random effects. A considerable amount of literature with regards to repeated measures cross-over designs already exists, although mainly focusing on two treatment and two periods designs (see e.g., Wallenstein and Fisher, 1977; Patel and Hearne, 1980; Dunsmore, 1981; Grender and Johnson, 1994).

Dunsmore (1981) uses Bayesian growth curves of Fearn (1975), with a quadratic time effect in a two-period repeated measures cross-over design. Analyzing the same experiment as Dunsmore (1981), Grender and Johnson (1994) discuss a two stage approach wherein the repeated measures across time for each subject are modeled parametrically, also using a quadratic trend, and later analyzing the parameter estimates using multivariate methods. More recently, Putt and Chinchilli (1999) analyze a two treatment and four period design using a mixed effects model which eliminates the need for preliminary testing for nuisance factors e.g., carry-over. The response in their case is again modelled parametrically assuming a quadratic function in time.

As already mentioned, there are many practical situations where determining an appropriate parametric function for the mean may not be easy. As pointed out by Dunsmore (1981), checking for the assumption of the presumed time trend (quadratic, in that case) may be a difficult task. It is with such cases in mind that we propose modelling the mean evolution using flexible semiparametric models, riding of the need to specify any particular parametric form. Jones and Kenward (2003) for example, consider a cross-over with many periods and model the period effects using natural cubic splines.

After estimating the group mean profile using penalized splines, focus shifts to construction of confidence bands around the fitted functions. In this thesis, adaptations of the bands of Ruppert *et al.* (2003) are made to accommodate correlation between measurements through random effects, as well as more complex models for residual covariances, specifically, models including serial correlation and measurement error (Verbeke and Molenberghs, 2000). Indeed, experiments with long sequences of repeated measurements are bound to yield some form of residual dependencies, which, at least, should be modelled parametrically.

5.1 Analysis Using the AUC as Summary Statistic

In this section, we discuss application of the area under the curve (AUC) as one way of summarizing data from a repeated measures cross-over design. This however may be seen as loss of information. Indeed, if the aim is to compare evolution over time across the experimental groups, such an approach is not useful. However, for an overall profile comparison, the AUC may sometimes be a viable option. As pointed out by Jones and Kenward (2003), the approach makes few modelling assumptions about the joint behavior of the repeated measurements, making it robust. Also, given that in our situation the data are completely balanced, each subject provides approximately the same amount of information, a key assumption for the use of such a summary statistic (Jones and Kenward, 2003).

Let us now focus on the model considered for the AUC summary statistic. As is usually done, to uphold the ubiquitous assumption of normally distributed errors, the model is based on a log transformation of the AUC. Let Y_{ijv} denote the log of AUC for animal i in period j, receiving experimental group v, for $i = 1, \ldots, n$, $j = 1, \ldots, p$, and $v = 1, \ldots, g$. Taking the last period ($j = 4$) and the control group ($v = 1$) as reference categories, define P_j, G_v, and C_j as indicator variables for period, treatment group, and carry-over respectively, such that, for example, $P_j = 1$ if period $= j$, and 0 otherwise, for all $j \leq p-1$, with a similar definition for C_j and G_v and for $v = 2, 3, 4$. The model takes the form:

$$Y_{ijv} = \beta_0 + \alpha_j P_j + \tau_v G_v + \zeta_j C_j + b_{0_i} + \varepsilon_{ijv}, \tag{5.1}$$

where β_0 is an intercept, α_j is the effect associated with period j, τ_v is the effect associated with treatment group v, ζ_j is the carry-over effect in period j, b_{0_i} is the random intercept accounting for the correlation of observations from one subject, and ε_{ijv} is the random error term. Following Jones and Kenward (2003), we do not include an interaction between period and treatment group. Such an interaction may emanate from subjects being affected by some factors other than treatment, and/or when the effect of a treatment level might depend on the current state of the subjects Senn (1993). Needless to say that then the interpretation of results becomes difficult.

5.2 Analysis of the Cross-over Design Using Semi-parametric Mixed Models

In the discussion in Section 5.1, the repeated measurements for each dog were summarized using the log of AUC as a summary statistic. However, the data fall within the

realm of continuous longitudinal data and hence can be modelled by use of a linear mixed model (Verbeke and Molenberghs, 2000). A flexible route, situated within the framework of mixed models, utilizes penalized splines, as already seen in the previous chapters.

5.2.1 Formulation of the Models for the Cross-over Design

This section focuses on formulation of possible models which can be used to describe the data at hand. The model with a full factorial structure for treatment and time as in Jones and Kenward (2003) is a sensible starting point. We show how one can move from this very general model to more parsimonious models, based on describing the time evolution through penalized splines. The formulation of the models is similar in spirit to that in Chapter 4, also presented in Maringwa *et al.* (2008d).

Using appropriately constructed matrices, all the models given in this section can be represented using the matrix notation of Section 4.1. For the desired flexibility, 40 equally spaced knots, selected as quantiles of the time variable (Ruppert, 2002) are used. Models 1-6 in Section 5.2.2 are used to model the cross-over aspects of the experiment, i.e. $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I})$. In Section 5.2.3 the same set of models are considered focusing on decomposing the covariance matrix of $\boldsymbol{\varepsilon}$ into components of serial correlation and measurement error.

5.2.2 Modeling the Cross-over Aspect of the Design

Model 1: Full Factorial Structure for Treatment and Time

Let $Y_{ijv\ell}$ denote the measurement on subject i, in period j, corresponding to treatment group v at time point ℓ, for $i = 1, \ldots, n$; $j = 1, \ldots, p$; $v = 1, \ldots, g$; and $\ell = 1, \ldots, m$. Define t_ℓ as an indicator variable for time, such that $t_\ell = 1$ if time is ℓ and 0 otherwise, for $\ell \leq m - 1$. Consider a model with a full factorial structure (Jones and Kenward, 2003) for treatment group and time, expressed as

$$Y_{ijv\ell} = \beta_0 + \alpha_j P_j + \tau_v G_v + \lambda_\ell t_\ell + \gamma_{vl} G_v t_\ell + \psi_{jl} P_j t_\ell + \zeta_j C_j + b_{0_i} + \varepsilon_{ijv\ell}. \qquad (5.2)$$

The parameter λ_ℓ refers to the effect of time, $\gamma_{v\ell}$ denotes the interaction between treatment group and time, $\psi_{j\ell}$ is the interaction between period and time, and $\varepsilon_{ijv\ell}$ are random terms. Often, $\Sigma_i = \mathrm{Cov}(\boldsymbol{\varepsilon}_i)$ is assumed to be $\sigma_\varepsilon^2 \boldsymbol{I}$, resulting in a conditional independence model. A more general residual covariance structure, for example decomposing the vector $\boldsymbol{\varepsilon}_i$ into components of serial correlation, $\boldsymbol{\varepsilon}_{(1)i}$, and measurement error, $\boldsymbol{\varepsilon}_{(2)i}$, can be considered (Verbeke and Molenberghs, 2000) and will be

discussed in Section 5.2.3.

Given that for each subject, 72 measurements in each period are taken, Model 1 is bound to yield a large number of parameters, hence the need for refinement. The following models, adjusting for possible period effects, show different possible approaches to modeling the time evolution in the experimental groups.

Model 2: Single Curve for all Treatment Groups

In Model 2, it is assumed that the time evolution is the same in all treatment groups (see Figure 4.1, panel A). As such, smoothing of the time trend occurs at the highest level of the model, thereby ignoring the treatment groups. The model can be represented as:

$$Y_{ij\ell} = \alpha_j P_j + \beta_0 + \beta_1 t_{ij\ell} + \sum_{v=1}^{K} b_k (t_{ij\ell} - \kappa_k)_+ + b_{0_i} + \zeta_j C_j + \varepsilon_{ij\ell}, \tag{5.3}$$

where κ_k are knots, and the coefficients b_k are common to all treatment groups, such that $\text{Var}(b_k) = \sigma_b^2$. Model 2 can be expressed in matrix notation by adopting the following notation: $\boldsymbol{Y} = [Y_{ij\ell}]_{i,j,\ell}$ and $\boldsymbol{X} = [1, t_{ij\ell}, P_1, P_2, P_3, C_1, C_2, C_3]_{i,j,\ell}$. Further, define, for each subject, a smoothing matrix

$$\boldsymbol{Z_{bi}} = [t_{ij\ell} - \kappa_k)_+]_{1 \le \kappa \le K}, \quad \text{with stacked version} \quad \boldsymbol{Z_b} = \begin{bmatrix} \boldsymbol{Z}_{b1} \\ \boldsymbol{Z}_{b2} \\ \vdots \\ \boldsymbol{Z}_{bn} \end{bmatrix}. \tag{5.4}$$

Model 3: Groups Curves Differ Only by a Shift

Model 3 assumes that the underlying linear trends in the treatment groups differ by a shift only. However, the same non-parametric part is fitted to all treatment groups. This model assumes the difference amongst the treatment groups, if present, does not depend on time. A penalized spline representation of the model is

$$Y_{ijv\ell} = \alpha_j P_j + \tau_v G_v + \beta_0 + \beta_1 t_{ij\ell} + \sum_{k=1}^{K} b_k (t_{ij\ell} - \kappa_k)_+ + b_{0_i} + \zeta_j C_j + \varepsilon_{ijv\ell}. \tag{5.5}$$

This scenario corresponds to panel B of Figure 4.1. Compared with Model 2, the current model has additional fixed effects parameters, τ_v. Note that the covariance structure is the same as in Model 2.

Model 4: Different Linear Effects with Same Smooth Part

Here, it is assumed that the linear parts of the models differ, while the same smooth part is considered for all groups. This resembles a scenario where, relative to Model 3, profiles are tilted at some angle, such that treatment effect is no longer constant in time. A representation of such a model is

$$Y_{ijv\ell} = \alpha_j P_j + \tau_v G_v + \beta_0 + (\beta_1 + \beta_{1v} G_v) t_{iv\ell} + \sum_{k=1}^{K} b_k (t_{ij\ell} - \kappa_k)_+ + b_{0_i} + \zeta_j C_j + \varepsilon_{ijv\ell}, \quad (5.6)$$

with $\text{Var}(b_k) = \sigma_b^2$. Panel C of Figure 4.1 graphically illustrates such a scenario.

Model 5: Different Curves Smoothed Equally

All models considered sofar assume that the same smooth component is fitted to the different treatment groups. It is possible to go one step further and fit a model with different non-parametric parts of the model in the different treatment groups, although the same smoothing parameter is used. The linear parts of the models are assumed different and, although the random effects are assumed independent from group to group, a single parameter is used to smooth the groups. A representation of such a model is:

$$Y_{ijv\ell} = \alpha_j P_j + \tau_v G_v + \beta_0 + (\beta_1 + \beta_{1v} G_v) t_{iv\ell} + \sum_{k=1}^{K} b_{vk} (t_{iv\ell} - \kappa_k)_+ + b_{0_i} + \zeta_j C_j + \varepsilon_{ijv\ell}.$$

Note that part of the design matrix, $\boldsymbol{Z_b}$, corresponding to smoothing, is now block-diagonal with each diagonal entry corresponding to a particular treatment group and the coefficients for the truncated lines basis, b_{vk}, are now group-specific with $\text{Var}(b_{vk}) = \sigma_b^2$. The smoothing matrix is now given by

$$\boldsymbol{Z_b} = \begin{bmatrix} \boldsymbol{Z_b^1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z_b^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Z_b^3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{Z_b^4} \end{bmatrix},$$

where $\boldsymbol{Z_b^1}, \ldots, \boldsymbol{Z_b^4}$ are group-specific smoothing matrices each constructed by stacking the $\boldsymbol{Z}_{bi}$ as in (5.4). This situation is similar to the illustration in panel D of Figure 4.1.

Model 6: Different Curves with Varying Levels of Smoothing

A further step is to relax the assumption on the smoothing parameter and to assume that the groups can be smoothed separately but with different smoothing parameters. Hence, both the fixed effects part and the non-parametric part differ by group and four variance components corresponding to smoothing the different treatment groups are estimated. The penalized spline representation of this model and the $\boldsymbol{Z_b}$ matrix is the same as in Model 5, with $\text{Var}(b_{vk}) = \sigma^2_{vb}$. The covariance matrix pertaining to smoothing is given by

$$\begin{bmatrix} \sigma^2_{1b}\boldsymbol{I}_{K\times K} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{2b}\boldsymbol{I}_{K\times K} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma^2_{3b}\boldsymbol{I}_{K\times K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma^2_{4b}\boldsymbol{I}_{K\times K} \end{bmatrix}.$$

The reference panel for this model is panel D, as in the previous model, since the difference between these models cannot be seen graphically.

5.2.3 Modelling the Covariance Structure

Two types of covariance structures for measurements from a particular subject need to be accounted for in the analysis. First, the correlations amongst measurements across different treatment periods and then dependencies amongst measurements within one treatment period. Assuming that the covariances applying to one period are similar to those in other periods, the between and within-period covariance structures are separable (Jones and Kenward, 2003). As such, accommodation of between-period dependencies can be achieved by introducing the subject-specific random intercepts. Commonly used models for repeated measures covariance structures, for example, an AR(1) process can be used to model the remaining within-period dependencies. In particular, we consider the decomposition of the residual variance into components of serial correlation and measurement error, such that:

$$\text{Cov}(\boldsymbol{\varepsilon}_i) = \sigma^2_{\varepsilon}\boldsymbol{I} + \tau^2 H_i,$$

where elements of H_i, the serial correlation matrix are modeled by some function, particular cases of which are the exponential and Gaussian functions (Verbeke and Molenberghs, 2000).

Let the variance of the serial process be denoted by τ^2 and the rate of decay of correlations with distance $d_{\ell\ell'}$ between time point ℓ and $\ell^{'}$ by θ. Table 5.1 shows the

Table 5.1: *Different covariance structures for modelling residual covariance within periods.*

AR(1)	Gaussian Serial correlation	Exponential Serial correlation
$\tau^2\theta^{d_{\ell\ell'}}$	$\tau^2\exp\left(-d_{\ell\ell'}/\theta^2\right)$	$\tau^2\exp\left(-d_{\ell\ell'}/\theta\right)$

forms of the covariance structure we consider for within-period residual covariance. As mentioned before, owing to the length of the sequences of measurements per subject, one would expect the residuals to be serially correlated. To gain insight into this phenomenon, we fit an unstructured mean model that includes other fixed effects, like period and the necessary interactions, and assess the behavior of the residuals. Note that, at this stage, neither random effects are included nor covariance structure is modeled. Denote the residuals at time point ℓ by $r_{\ell\nu}$, $\nu = 1, \ldots, 32$. At each of the 72 distinct time points, there are 32 observations. Figure 5.1 shows a plot of $r_{1\nu}$ on the horizontal axis against $r_{\ell'\nu}$ on the vertical axis, with $\ell' = 2, 6, 10, \ldots, 62$, a selection of 16 time points. It is apparent from the residual plots that, after removing the mean structure, the residuals do not appear independent and, as expected, the dependencies tend to weaken with distance in time. As such, conditional independence models as in Section 5.2.2 may not be appropriate in this case, hence the modeling of serial correlation.

5.2.4 Constructing Confidence Intervals and Bands

This section focuses on the construction of confidence bands around the group-specific fitted functions, following closely the work in Chapter 4. Such bands can be used to compare the different treatment groups at specific time points, if necessary. Ruppert *et al.* (2003) give details for constructing such intervals or bands for smoothed functions. Our intention is to adapt their results to accommodate the correlation structure, as accounted for by the random intercept, and the residual covariances allowing for presence of serial correlation.

The model for the cross-over design under consideration may be expressed in the general formulation of a linear mixed model, whose random effects parts may well be partitioned into components corresponding to subject-specific effects and smoothing effects as in (4.8). Construction of pointwise confidence intervals as well as simultaneous confidence bands requires the covariance for the vector of contrasts between the estimated and true parameters for the fixed and random effects (Ruppert et al., 2003). Note, the generality of (4.10) makes it usable in different settings. For example, ex-

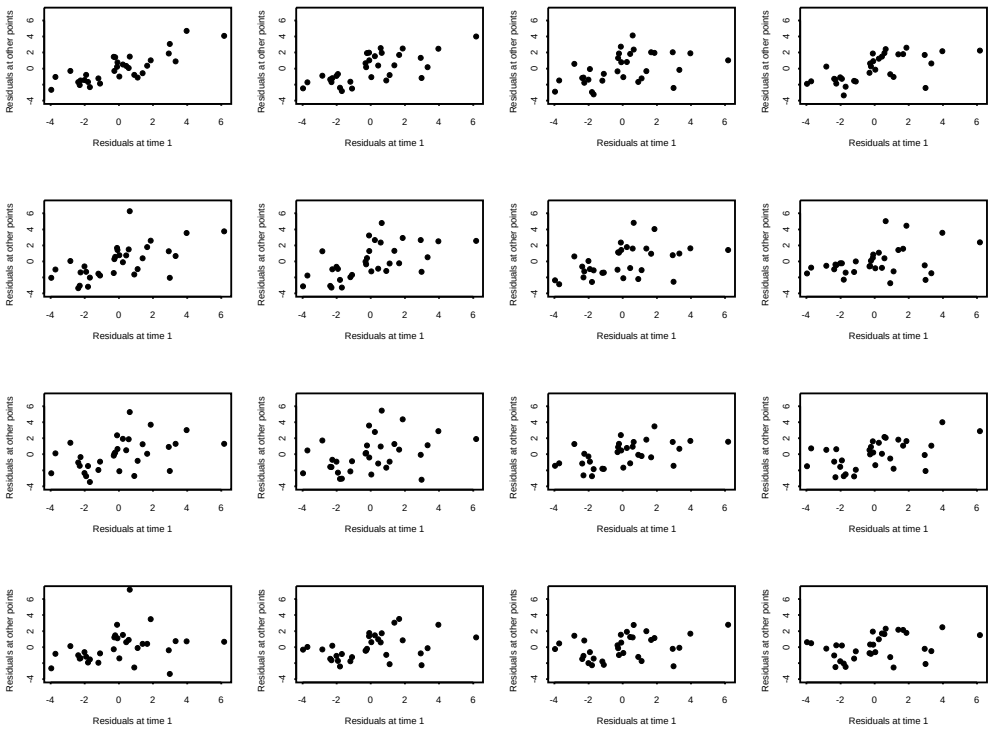


Figure 5.1: *Residuals at first time point plotted against residuals from other 16 selected time points (2, 6, 10,. . . , 62).*

pression (4.11) follows specifically from a random intercept model with independent errors, implying a compound symmetry structure for $\hat{\boldsymbol{R}}$, which is readily invertible. The following section focuses on how the preceding discussion can be extended to the particular case of a cross-over design.

Adaptation of the Confidence Bands to the Cross-over Setting

Consider a random intercept model as in Section 5.2.2. Further, assume the model also includes both components of serial correlation and measurement error in the residual covariance structure. The resulting $\widehat{\boldsymbol{R}}$ implies a simplified version of (4.10) may not be straightforwardly obtained. However, the matrix can still be used in its most general form. For illustrative purposes, consider a particular dog i. Assuming an exponential type of serial correlation, the part of $\text{Cov}(\boldsymbol{\varepsilon}_i^*)$ in the first period corresponding to the

first three observations is given by:

$$\mathbf{R}^*_{i[3]} = \begin{bmatrix} \sigma^2_{b_0} & \sigma^2_{b_0} & \sigma^2_{b_0} \\ \sigma^2_{b_0} & \sigma^2_{b_0} & \sigma^2_{b_0} \\ \sigma^2_{b_0} & \sigma^2_{b_0} & \sigma^2_{b_0} \end{bmatrix} + \tau^2 \begin{bmatrix} 1 & \exp(-d_{12}/\theta) & \exp(-d_{13}/\theta) \\ \exp(-d_{12}/\theta) & 1 & \exp(-d_{23}/\theta) \\ \exp(-d_{13}/\theta) & \exp(-d_{23}/\theta) & 1 \end{bmatrix}$$

$$+ \begin{bmatrix} \sigma^2_\varepsilon & 0 & 0 \\ 0 & \sigma^2_\varepsilon & 0 \\ 0 & 0 & \sigma^2_\varepsilon \end{bmatrix}, \tag{5.7}$$

where τ^2 is the variance for the serial correlation part, θ is the rate of decay of the correlations as a function of $d_{\ell\ell'}$, which is the Euclidean distance between coordinates of the time variable, and σ^2_ε is the variance of the measurement error. Therefore, $\boldsymbol{R}^*_i$ consists of matrices of the form (5.7) as diagonal elements, corresponding to each period, and matrices of the form $\sigma^2_{b_0}\boldsymbol{J}$ in the off diagonal entries. Now, replacing $\widehat{\boldsymbol{R}}$ with $\widehat{\boldsymbol{R}}^*$ in (4.10), where $\widehat{\boldsymbol{R}}^*$ is block-diagonal with diagonal elements $\widehat{\boldsymbol{R}}^*_i$, provides the expression which can used to construct either pointwise or simultaneous confidence bands.

5.3 Application to the Cardiovascular Safety Experiment Cross-over Case

This section focuses on the application of the methodology discussed in the previous sections to the data described in Section 2.1.2. Emphasis will be on the semiparametric mixed-model approach. Within the semiparametric models, a comparison of models assuming independent residual errors with models assuming some form of residual correlation structure will be undertaken. In addition, we briefly discuss the issue of carry-over where, loosely, the observed response in one period could be the result of the effect of the previously allocated treatment. Regardless of the wash-out period used in the study, carry-over effects are not unexpected. For models that do not require preliminary testing for carry-over, we refer to Putt and Chinchilli (1999). The approach we take is to select one of the semiparametric models with independent errors, based on AIC. The selected model will then be improved by modelling the covariance structure and all further inferences will be based on the model selected based on comparing covariance structures. The results from such a model are compared with the summary statistics analysis.

Table 5.2: *Minus twice loglikelihood values and AIC values for models in Section 5.2.2, assuming independent residual errors. For the null model (Model 2) as well as the model with the smallest AIC (Model 4), different covariance structures are considered. The boxed values indicate the best model under a particular covariance structure.*

		Model					
Within-period cov.	Crit.	1	2	3	4	5	6
Independent	-2 loglik	9025.3	7128.7	6922.9	6893.3	6941.8	6938.3
	AIC	10041.3	7150.7	6950.9	6927.3	6975.8	6978.3
AR(1)	-2 loglik		5590.7	5557.1	5552.5		
	AIC		5614.7	5587.1	5588.5		
Gaussian ser. corr.	-2 loglik		5650.4	5600.2	5593.9		
	AIC		5670.4	5632.2	5631.9		
Exp. ser. corr.	-2 loglik		5544.1	5521.6	5518.4		
	AIC		5570.1	5553.6	5556.4		

5.3.1 Model Fitting, Selection and Hypotheses Testing

Let us now focus on fitting and selection of the models discussed in Section 5.2.2. For each of the models, the independent errors structure, a commonly used approach in practice, is assumed and results are given in Table 5.2. A common issue with cross-over designs is carry-over (Senn, 1993; Jones and Kenward, 2003). Models including carry-over will be considered henceforth although a comparison with corresponding models excluding the effects may be interesting.

The exploratory comparison of the models with independent residual errors appears to indicate Model 4, with differing linear effects by group and the same non-parametric component is a plausible starting point. A formal likelihood ratio test (LRT) can be performed to see if indeed there is need to move from Model 3 to Model 4. Such a test, based on a χ^2_3, and a LRT statistic of $29.6 = 6922.9 - 6893.3$ yields a highly significant result ($p = 0.0001$). Hence, a model with different linear effects by group fits better. Note, the test between Model 3 and 4 is based on fixed effects only; no variance components are involved.

Different ways of modelling the residual covariance are applied and the results in Table 5.2 indicate a substantial improvement in the fit of Model 4 upon modelling

of the residual covariance structure. In particular, the model with exponential-type serial correlation appears to fit better than other models. However, the group-by-time interaction is now insignificant, implying Model 3. Note that, the group-by-time interaction is the characteristic separating Models 3 and 4. Thus a formal test between both models would be based on χ^2_3, with a LRT statistic of $3.2 = 5521.6 - 5518.4$ and $p = 0.3618$, corroborating that a constant difference in time suffices in this situation. Henceforth, Model 3, with exponential serial correlation, becomes our chosen model and any further inferences will be based upon this model.

The selection of Model 3 already suggests presence of treatment effect. However, a formal test may be required, and that translates to testing the chosen model against Model 2 (the null model). The hypothesis of interest then becomes, following (5.3) and (5.5): $H_0 : \tau_v = 0, v = 2, 3, 4$. The null model was fitted under the various covariance structures, such that appropriate comparisons of fixed effects against any of the (chosen) alternative models may be effectuated. In this case, a test between Models 3 and 2, both under the exponential serial correlation would be appropriate. Again, there are no variance components equated to zero in this test and therefore it is based on χ^2_3. The LRT statistic is 22.5, which is significant ($p < 0.0001$), hence groups differ.

We present the fixed-effects parameter estimates for Model 3, p-values from the associated t-tests, and variance components in Table 5.3. For the sake of comparison, we have included parameter estimates from the model with independent residual errors. Focusing on the model with serial correlation in Table 5.3, it can be observed that only the medium dose group differs from the control group ($p = 0.0014$). Let us focus on comparing this model with the model assuming independent errors. While parameter estimates do not change much, it is the standard errors that substantially change, rendering some previously significant effects insignificant, such as, for example, the difference between the low and high doses. This highlights the problem of underestimating variability, often ignored when models such as the conditional independence model are applied in practice.

Table 5.3 also gives results from the AUC analysis. Although the time dimension is lost in this analysis, an overall comparison amongst the doses can be salvaged. Note the parameter estimates for the AUC are not directly comparable to those from the semi-parametric mixed models since they are on the log scale. Similar to the conclusion made above, the results indicate a significant difference between the medium dose and the control group, albeit with weakened evidence ($p = 0.0129$). Hence, the animals receiving the medium dose group tend to have higher values of the measure of relaxation capacity of the heart than the control group, and the difference

Table 5.3: *Fixed-effects parameter estimates, standard errors (s.e) and p-values for associated t-tests corresponding to Model 3 with independent errors within periods, exponential serial correlation (ESC) and Model 3 with exponential serial correlation plus carry-over effects. Also included are estimates, standard errors and p-values for the AUC analysis. The last period and the control group are taken as reference categories.*

		Model 3 (Independent errors)		Model 3 (ESC)		Model 3 (ESC+carry-over)		AUC	
Effect	Par.	Est. (s.e.)	p	Est. (s.e.)	p	Est. (s.e.)	p	Est. (s.e.)	p
Fixed Effects:									
Intercept	β_0	20.679(0.567)	0.0001	20.767(0.587)	0.0001	20.654(0.596)	0.0001	8.930(0.027)	0.0001
Period 1	α_1	-2.122(0.079)	0.0001	-2.267(0.179)	0.0001	-2.082(0.216)	0.0001	-0.103(0.012)	0.0001
Period 2	α_2	-1.693(0.063)	0.0001	-1.646(0.179)	0.0001	-1.649(0.171)	0.0001	-0.080(0.010)	0.0001
Period 3	α_3	-0.569(0.063)	0.0001	-0.550(0.179)	0.0027	-0.551(0.171)	0.0017	-0.026(0.010)	0.0138
High	τ_4	0.232(0.066)	0.0012	0.164(0.179)	0.3594	0.197(0.179)	0.2746	0.010(0.010)	0.3539
Medium	τ_3	0.758(0.066)	0.0001	0.585(0.179)	0.0014	0.729(0.179)	0.0001	0.036(0.010)	0.0015
Low	τ_2	-0.136(0.066)	0.0383	-0.119(0.179)	0.5052	-0.10(0.179)	0.5555	-0.005(0.010)	0.6274
Time	β_1	1.206(1.184)	0.3212	1.353(1.064)	0.2236	1.350(1.069)	0.2260		
Variance components:									
$\text{Var}(b_k)$	σ_b^2	1.151(0.526)		0.888(0.546)		0.903(0.500)			
$\text{Var}(b_{0_i})$	σ_U^2	2.134(1.069)		2.088(1.060)		2.056(1.053)		0.005(0.003)	
Exponential serial correlation:									
$\text{Var}(\varepsilon_{(1)ijv\ell})$	τ^2			1.078(0.072)		1.144(0.073)			
Rate of decay	θ			0.386(0.043)		0.431(0.040)			
Measurement error variance:									
$\text{Var}(\varepsilon_{(2)ijv\ell})$	σ_ε^2	1.129(0.034)		0.163(0.023)		0.172(0.024)			

is constant over time.

Returning to the issue of carry-over, although the effect (parameter estimates not given) appears significant ($p = 0.0390$), no major changes are obtained in other parameter estimates, their standard errors, or in the conclusions that would have been reached should one have considered the model that excludes carry-over effects (see Table 5.3).

5.3.2 Confidence Intervals and/or Bands

The model selected as the 'best' explicitly implies that the treatment effect is constant in time, implying that for this situation, time point comparisons are redundant. We proceed to construct the confidence bands around the fitted profiles.

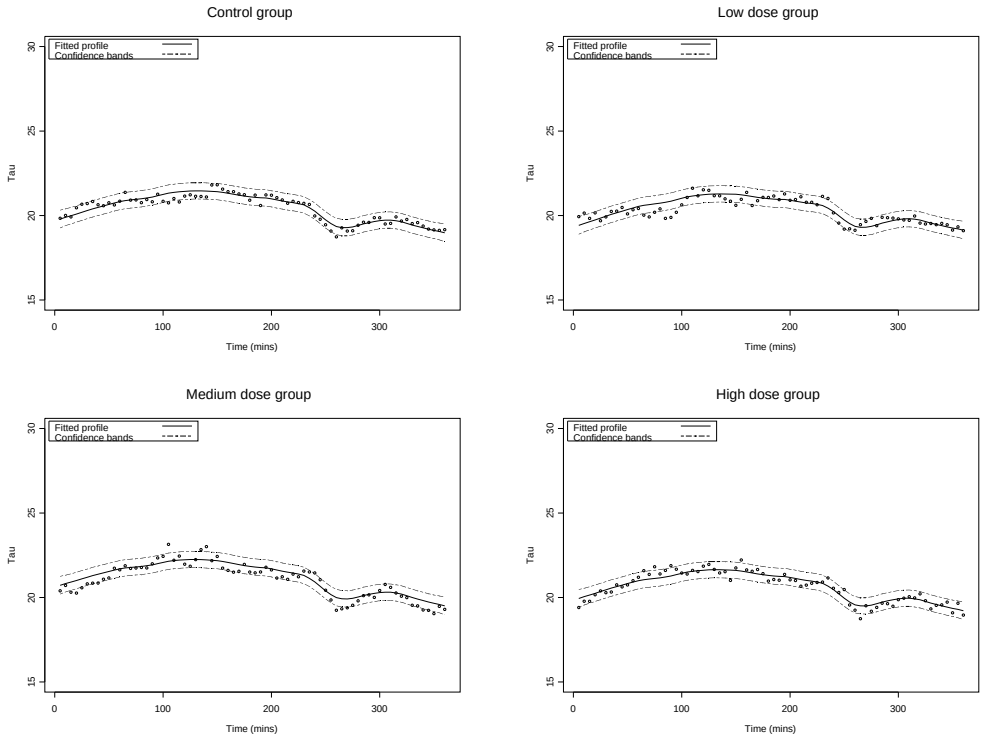


Figure 5.2: *Group-specific fitted profiles together with the corresponding 95% simultaneous confidence bands around them. The points are observed mean values at each time point.*

The construction of the confidence intervals and bands requires estimation of the

variance components pertaining to smoothing and random intercept, as well as the parameters associated with the exponential serial correlation (see Table 5.3). Note that smoothing is performed at the highest level of the model, and hence, one variance component is estimated for all four treatment groups.

Following the discussion in Sections 5.2.4, and using the estimates in Table 5.3, one can then construct the pointwise as well as simultaneous confidence bands. As in Chapter 4, constructing simultaneous confidence bands involves estimating a value $\tilde{h}_{(1-\alpha)}$, which would replace $Z_{(1-\alpha/2)}$ usually used in constructing confidence intervals under the normal distribution assumption (see Ruppert *et al.*, 2003).

Five independent simulations of 10 000 draws each, using results discussed in Section 5.2.4 are performed. We obtain $\tilde{h}_{0.95} \approx 2.4285, 2.4608, 2.4341, 2.4819$ and 2.4520. The minimum of these values, which is 2.5285, can be taken as the estimate of $\tilde{h}_{0.95}$, implying that the simultaneous confidence bands are about $2.4285/1.96 = 1.24$ times wider than the pointwise confidence bands. Figure 5.2 show the group-specific average profiles, fitted profiles and the 95% confidence bands. The model appears to fit well. The simultaneous confidence bands constructed enable one to make joint statements about the profiles' evolution in time. In case the model depicted treatment effect changing over time, such intervals could be used for example to compare each of the other treatment groups to the control and each time point.

5.4 Discussion

We have exemplified the flexibility of nonparametric smoothing techniques in terms of application to different types of study designs. In particular, penalized splines fitted within the linear mixed-model framework, in the context of cross-over designs, are used. We have illustrated that one can formulate different possible scenarios showing how the different treatment groups could possibly differ. Such an approach then enables one to select the model deemed 'best' according to some criterion, such as, for example, the AIC used here. Although we restricted attention to random intercepts, extension to more complicated models, including subject-specific spline models are possible.

Particular attention has also been given to models including serial correlation, wherein well-known functions for modelling it have been investigated for these data. Indeed, with relatively long sequences of repeated measurements, residual correlations are expected. As we have seen, ignoring such correlations can possibly lead to misleading results. It is worthy mentioning that, for future research, flexible models

considered here for the mean can as well be considered to model the serial correlation. This is a relatively new area of research worthwhile pursuing.

Often, researchers require comparisons of treatment groups at specific time points. This could possibly be done by fitting a full factorial structure in time and compare groups using appropriate contrasts. However, the large number of time points involved here makes such an approach prohibitive. An attractive alternative is the use of confidence intervals and/or confidence bands constructed around the fitted profiles.

Once a suitable model has been selected, confidence bands can then be constructed around the fitted functions. Focus has been on the adaptation of the confidence intervals and bands of Ruppert *et al.* (2003) for application in this specific situation of cross-over design. Using the confidence bands, one is able to identify specific sections where the bands do not overlap, indicating significant differences. Other than overcoming the disadvantages of the full factorial structure approach mentioned above, the problem of multiple comparison is also inherently solved here. Note that, as mentioned before, the model we have focussed on does not warrantee use of confidence bands for time point comparisons since treatment effect is constant in time.

6

Investigating Associations in Cross-over Designs Using Surrogate Marker Validation Methodology

The data considered in this chapter come from a cross-over design as in Chapter 5, and were described in Section 2.2. Although the objectives in both chapters differ substantially, the two chapters share common ground on study design, and the need for flexible modelling of the mean evolution in time. This chapter focuses on blending surrogate marker evaluation methodology with flexible modelling techniques in quantifying associations of interest.

The first step in the process of drug development is identifying promising compounds. Once a compound has been isolated for further scrutiny, it enters a rigorous testing and evaluation stage, the so-called pre-clinical phase. This stage is designed to assess the chemical properties of the new drug as well as to determine the steps for synthesis and purification. In this stage, the toxicological and pharmacological effects of the drug are evaluated through *in-vitro* and *in-vivo* animal testing. There

might be a variety of reasons hindering undertaking these tests directly on the clinically relevant outcome, even when the studies involve animals, necessitating the use of biomarkers.

Several challenges are encountered in the identification of biomarkers, including: understanding the role of a specific biomarker to a clinically relevant problem; developing either an indirect or a direct readout of physiologic state; determining the comparable pathways between animal models and humans; and finally embedding the biomarker into a robust assay and subsequent validation and approval of the assay in clinical applications (Pien *et al.*, 2005). Several attempts, from both a biological and a statistical angle, have been made to circumvent these challenges (Burzykowski *et al.*, 2005). Focusing on the statistical problem of identifying and validating a biomarker, statistical expertise, in particular paradigms designed to validate surrogate markers, might be handy tools to quantify the degree of association between the biomarker and the clinically relevant outcome.

In surrogate marker evaluation, two possible sources of evidence can be sought to validate a biomarker. The first is situated at the individual patient level and is concerned with the biological pathway from the surrogate to the true endpoint. The second possible source of evidence comes from the trial level, and quantifies the association between the treatment effects on the marker and clinical endpoint (Burzykowski *et al.*, 2005).

The focus of this chapter is to adapt existing surrogate marker validation methodology to quantify the degree of association between behavior, as measured by alertness, corticosterone levels, and telemetry measures such as heart rate and blood pressure of rats, with emphasis given to the prediction of one of the outcomes given the other in a single trial setting. In the process of adapting the surrogate marker methodology, we also adopt the relevant terminology. For example, while the term 'individual' is used to refer to a patient in the clinical trials setting, here it will be used to refer to a particular experimental unit, i.e., an animal. Similarly, a 'trial' is to be understood as referring to a particular experiment with animals. Note that, if there is an interest in the trial-level surrogacy, there is then need for repetition of the experiment, for example at different centers and/or by different investigators, or even through the conduct of a sequence of altogether different experiments. Here, attention will be primarily focused on individual level associations. The backbone of the contents of this chapter is the work presented in Tilahun *et al.* (2008).

6.1 Flexible Modelling of the Mean Using Fractional Polynomials

The primary goal here is to quantify the association between CORT, heart rate, and blood pressure via surrogate marker validation methods. However, a quick glance at the mean profiles in Figures 2.3 and 2.4 suggests that proper modeling of the mean evolution in time is necessary. One can get rid of the need to specify a parametric model through use of flexible modeling techniques. This has been the focus of Chapters 4 and 5, with emphasis on penalized splines methodology. While use of such methodology is also advocated for in this chapter, an alternative and frequently used approach is the so-called fractional polynomials of Royston and Altman (1994).

Fractional polynomials provide an extension to classical polynomials allowing for non-integer powers to the time covariate, thereby adding greater flexibility in capturing rather complex non-linear relationships. A brief description of fractional polynomials follows.

Let $\boldsymbol{t} = (t_{i1}, \ldots, t_{im})$ denote the set of time points pertaining to subject i. Royston and Altman (1994) define a fractional polynomial of degree q by

$$\phi_q(\boldsymbol{t}; \boldsymbol{\beta}, \boldsymbol{p}) = \sum_{r=0}^{q} \beta_r H_r(\boldsymbol{t}), \tag{6.1}$$

where q is a positive integer and $\boldsymbol{p} = (p_1, \ldots, p_q)$ is a real-valued set of powers such that $p_1 \leq \cdots \leq p_q$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)$ are real-valued coefficients. First, define the following transformation,

$$\boldsymbol{t}^{(p_r)} = \begin{cases} \boldsymbol{t}^{p_r} & \text{if } p_r \neq 0, \\ \ln(\boldsymbol{t}) & \text{if } p_r = 0. \end{cases}$$

For $r = 0$, $H_0(\boldsymbol{t}) = 1, p_0 = 0$, and for $r = 1, \ldots, q$

$$H_r(\boldsymbol{t}) = \begin{cases} \boldsymbol{t}^{(p_r)} & \text{if } p_r \neq p_{r-1}, \\ H_{r-1}(\boldsymbol{t}) \ln(\boldsymbol{t}) & \text{if } p_r = p_{r-1}. \end{cases}$$

As mentioned in Royston and Altman (1994), polynomials of a degree higher than 2 or 3 are rarely encountered in practice. The best power transformation is frequently found among the members of the list $\{-2, -1, -0.5, 0, 0.5, 1, \ldots, \max(3, q)\}$.

Note that the fractional polynomial model has been defined in its generic form and in analogy with penalized splines models, extension to include covariates other than time is possible. In such a situation, an extension of (6.1) may be obtained through

adding the fixed effects for treatment, period, and carry-over, together with relevant interactions. A more detailed treat on the analysis of cross-over designs with data of a longitudinal nature was given in Chapter 5. The current chapter essentially focuses on synthesizing such methodology with surrogate marker validation techniques in order to quantify associations of interest. The following section therefore gives a brief review of the relevant methodology from the world of surrogate marker validation techniques.

6.2 Validation Methods

In this section, we give a concise description of the various methods used in validating a surrogate endpoint, with emphasis on individual level surrogacy. Given the present situation of a single experiment, it suffices to consider surrogacy at the individual level.

6.2.1 Review of the Single Trial-based Validation Methods for Continuous Outcomes

Several methods have been suggested for the formal evaluation of surrogate markers. Some of these methods are based on a single trial while others, which are gaining momentum in the present day, are based on meta-analytic concepts. The first formal approach to evaluate markers is attributed to Prentice (1989), who has given a definition of surrogate endpoints, followed by a series of operational criteria to check whether the definition is fulfilled.

Freedman *et al.* (1992) have supplemented the hypothesis-testing-based criteria, which necessarily depend on the power of the test performed, with a quantity to be estimated. They suggested the use of the so-called *proportion of treatment effect explained* (PTE) by the surrogate as an alternative means of validation. The PTE faces serious drawbacks, against the background of which Buyse and Molenberghs (1998) have suggested the use of another quantity, the *relative effect* (RE), defined as the ratio of the treatment effect on the true endpoint to that on the surrogate endpoint. In turn, the RE is open to severe criticism as well. First, the RE's confidence intervals, like the ones for PTE, tend to be wide. While this could in principle be overcome, there is a second, more severe problem in the sense that the RE is useful for prediction of the true treatment effect from the surrogate treatment effect only when the relationship between both is multiplicative. This may be rightfully viewed as restrictive and, in any case, cannot be verified from a single trial.

Switching to the experimental animal level, the need might arise to quantify the association between the surrogate and the true endpoint after adjustment for the treatment effect. To this end, Buyse and Molenberghs (1998) suggested the use of the adjusted association.

Suppose we have a single experiment, let for subject i ($i = 1, \ldots, n$), S_i and T_i be the surrogate and true endpoint, respectively, and let G_i be a binary treatment group indicator. To compute the adjusted association, consider the following pair of models

$$\begin{aligned} T_i &= \mu_T + \alpha G_i + \varepsilon_{T_i} \\ S_i &= \mu_S + \beta G_i + \varepsilon_{S_i}, \end{aligned}$$

where (μ_T, μ_S, α, β) are intercepts and treatment effects on the true and surrogate endpoints, respectively, and the error terms have a joint zero-mean normal distribution with variance-covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{TS} & \sigma_{TT} \end{pmatrix}.$$

Then, the *adjusted association*, denoted R^2 can be computed as

$$R^2 = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}.$$

Note that the individual-level surrogacy is meant to measure the degree of correlation between the two endpoints after correcting for treatment and other possible effects.

6.2.2 Variance Reduction Factor

In this section, we review the variance reduction factor, suggested by Alonso *et al.* (2003) for the case of two repeatedly measured outcomes, where after we show how this method can be adapted to the situation where one of the two outcomes is cross-sectional. Let us assume that there are n subjects enrolled for a particular study and further suppose that t_{ij} is the time at which the j^{th} measurement of the i^{th} subject is taken. Let T_{ij} and S_{ij} be the true and the surrogate endpoints, respectively, and let G_i be a binary treatment indicator. Now, consider the following joint model for the true and surrogate endpoints

$$\begin{aligned} T_{ij} &= \mu_T + \alpha G_i + f(t_{ij}) + \varepsilon_{T_{ij}} \\ S_{ij} &= \mu_S + \beta G_i + f(t_{ij}) + \varepsilon_{S_{ij}}, \end{aligned} \tag{6.2}$$

where $f(t_{ij})$ is a flexible function in time, which can be modeled by fractional polynomials, penalized splines, or any flexible function in time. In principle, it is possible for the two endpoints to depend on time through different functions, in which case we will have $f_T(t_{ij})$ and $f_S(t_{ij})$ for the true and surrogate endpoint respectively. However, without loss of generality, let us assume that both depend on time through the same function. The error terms $(\varepsilon_{T_{ij}}, \varepsilon_{S_{ij}})$ are assumed to follow a zero-mean normal distribution with patterned variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}, \tag{6.3}$$

with obvious notation.

In this setting, Alonso *et al.* (2003) proposed to quantify the individual-level surrogacy using the so-called *variance reduction factor*, which is defined as

$$VRF = \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{T|S})}{\text{tr}(\Sigma_{TT})}, \tag{6.4}$$

where $\Sigma_{T|S}$ denotes the conditional variance-covariance matrix of T_{ij} given S_{ij}, i.e., $\Sigma_{T|S} = \Sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}$. Furthermore, these authors have shown that the VRF satisfies a set of properties that makes it practically applicable:

(i) VRF ranges between zero and one;

(ii) $VRF = 0$ if and only if the true and the surrogate endpoints are independent;

(iii) $VRF = 1$ if and only if there exists a deterministic relationship between the true and surrogate endpoint;

(iv) $VRF = R^2$ in the cross-sectional setting.

Note that, at the individual level, interest lies in the prediction of the true endpoint given the surrogate endpoint. In this regard, property (ii) shows that if the VRF equals zero, then no sensible prediction is possible, whereas a perfect prediction is attained if VRF equals one, as indicated by property (iii). Property (iv) establishes the link between this approach and the one suggested by Buyse *et al.* (2000) for univariate outcomes.

Let us now turn to the question as to how this approach can be used when one of the two endpoints is cross-sectional. Assume we have m measurements per subject for the longitudinal outcome.

Case 1: A Longitudinal Surrogate for a Cross-sectional True Endpoint

Let us assume that the surrogate endpoint is repeatedly measured over time with m repeated measures and that the true endpoint is cross-sectional. Model (6.2) now takes the form

$$\begin{aligned} T_i &= \mu_T^* + \alpha^* G_i + \varepsilon_{T_i}, \\ S_{ij} &= \mu_S^* + \beta^* G_i + f(t_{ij}) + \varepsilon_{S_{ij}}, \end{aligned} \tag{6.5}$$

and the error terms $(\varepsilon_{T_i}, \varepsilon_{S_{ij}})$ are assumed to follow a zero-mean normal with variance-covariance matrix Σ, which in this setting takes the form

$$\Sigma = \begin{pmatrix} \sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{pmatrix}. \tag{6.6}$$

Here, σ_{TT} denotes the variance of the true endpoint, Σ_{TS} is a $(1 \times m)$ vector containing the covariances between the true endpoint and the surrogate endpoint at different time points, and Σ_{SS} is a $(m \times m)$ variance-covariance matrix associated with the longitudinal surrogate endpoint. Then, the VRF_{indiv} for longitudinal surrogate and a cross-sectional true endpoint denoted by VRF_{ST}^{LC}, with a superscript 'L' ('C') reminiscent of 'longitudinal' ('cross-sectional'), can be computed as

$$VRF_{ST}^{LC} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{T|S})}{\text{tr}(\sigma_{TT})}, \tag{6.7}$$

where $\sigma_{T|S}$ denotes the conditional variance of T given S, that is,

$$\sigma_{T|S} = \sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}.$$

Using this expression, (6.7) can be re-written as

$$VRF_{ST}^{LC} = \frac{\text{tr}(\sigma_{TT}) - \text{tr}(\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\sigma_{TT})}. \tag{6.8}$$

Note that all matrices involved in the computation of VRF_{ST}^{LC} are of dimension (1×1) and hence the trace reduces to the corresponding scalar, offering the opportunity to simplify (6.8) to

$$VRF_{ST}^{LC} = \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}. \tag{6.9}$$

Notice that $VRF_{ST}^{LC} = 0$ if and only if $\Sigma_{ST} = 0$, i.e., when S and T are independent.

Intuitively, (6.9) quantifies how much of the total variability of the true endpoint is explained by the surrogate endpoint, after adjusting for treatment effects and repeated measures of the surrogate endpoint.

Case 2: A Cross-sectional Surrogate for a Longitudinal True Endpoint

Next, let us consider a role reversal, such that the true endpoint is repeatedly measured over time with m repeated measures, whilst having the surrogate endpoint in cross-sectional form. Model (6.2) becomes

$$\begin{aligned} T_{ij} &= \mu_T^* + \alpha^* G_i + f(t_{ij}) + \varepsilon_{T_{ij}}, \\ S_i &= \mu_S^* + \beta^* G_i + \varepsilon_{S_i}. \end{aligned} \tag{6.10}$$

The error terms ($\varepsilon_{T_{ij}}$,ε_{S_i}) are zero-mean normally distributed with variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \sigma_{SS} \end{pmatrix}. \tag{6.11}$$

Now, the VRF_{indiv} for this case takes the form

$$\begin{aligned} VRF_{ST}^{CL} &= \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{T|S})}{\text{tr}(\Sigma_{TT})} \\ &= \frac{\text{tr}(\Sigma_{TT}) - \text{tr}(\Sigma_{TT}) + \text{tr}(\Sigma_{TS}\sigma_{SS}^{-1}\Sigma_{ST})}{\text{tr}(\Sigma_{TT})} \\ &= \frac{\text{tr}(\Sigma_{TS}\Sigma_{ST})}{\sigma_{SS}\text{tr}(\Sigma_{TT})}. \end{aligned} \tag{6.12}$$

From (6.9) and (6.12), it is clear that there is asymmetry in the VRF calculations. Results differ depending on which of the two endpoints is the cross-sectional one. This is in line with our expectations. In the case of a longitudinal true endpoint, the VRF measures the ability of the cross-sectional endpoint to predict the longitudinal outcome at each time point, whereas when the longitudinal sequence is treated as surrogate endpoint, the VRF measures the adequacy of the longitudinal sequence to predict the cross-sectional outcome. It is therefore imperative to determine in advance which of the two outcomes is treated as true when applying this procedure to quantify association. Either way, a VRF value close to one indicates that the surrogate is a 'good' predictor of the true endpoint at the individual level, while values close to zero indicate 'poor' prediction. In any case however, the values of the VRF have to be complemented with expert opinion before passing judgment on the adequacy of the surrogate to predict the true endpoint.

6.2.3 The Measure R_Λ^2

As can be seen from (6.4), the VRF summarizes the variability of the two endpoints using the trace of the corresponding variance-covariance matrices. In multivariate

analysis, there is no unique way of defining a generalized variance, the trace is one of the classical ways of doing so, while another common definition uses the determinant. Interestingly, using the trace or the determinant to summarize the variability of the endpoints has important ramifications for analysis and leads to two totally separate measures with different interpretations. To this end, Alonso *et al.* (2006) have suggested another measure, the so-called R^2_Λ, which uses this alternative definition of the generalized variance. Like the VRF, this measure can be derived based on Model (6.2), as follows:

$$R^2_\Lambda = 1 - \frac{|\Sigma|}{|\Sigma_{TT}||\Sigma_{SS}|}. \tag{6.13}$$

The authors have shown that this measure enjoys the following desirable properties;

(i) R^2_Λ is symmetric and invariant with respect to linear bijective transformations;

(ii) R^2_Λ ranges between zero and one;

(iii) $R^2_\Lambda = 0$ if and only if the error terms are independent;

(iv) $R^2_\Lambda = 1$ if and only if there exist a and b so that $a^T\varepsilon_{S_{ij}} = b^T\varepsilon_{T_{ij}}$ with probability one, and;

(v) $R^2_\Lambda = R^2$ in the cross-sectional setting.

All of these properties, except the fourth property are shared with the VRF. The fourth property, however, differs in important ways from the VRF. Indeed, whereas the VRF takes the value 1 when there is a deterministic relationship between both endpoints, R^2_Λ is 1 whenever there is a deterministic relationship between two linear combinations of both endpoints, allowing us to uncover strong association in cases where the VRF might fail to do so. This is not a disadvantage of one or the other proposal, but rather underscores their focusing on different aspects. The expression for R^2_Λ clearly shows that, unlike the VRF, this measure treats both endpoints symmetrically. To clarify this further, let us first consider the surrogate to be longitudinal and the true endpoint cross-sectional, and thereafter reverse the roles.

Case 1: A Longitudinal Surrogate for a Cross-sectional True Endpoint

Consider Model (6.5) and the corresponding variance-covariance matrix (6.6). The R^2_Λ for a longitudinal surrogate and a cross-sectional endpoint is given by

$$R^{2,LC}_{\Lambda,ST} = 1 - \frac{|\Sigma|}{|\sigma_{TT}||\Sigma_{SS}|}, \tag{6.14}$$

where σ_{TT}, Σ_{SS}, and Σ are as defined in (6.6). Note that

$$|\Sigma| = |\Sigma_{SS}||\Sigma_{T|S}| = |\Sigma_{SS}| \cdot |\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}|,$$

and substituting this in (6.14), we obtain

$$\begin{aligned} R^{2,LC}_{\Lambda_{ST}} &= 1 - \frac{\sigma_{TT} - \Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}} \\ &= \frac{\Sigma_{TS}\Sigma_{SS}^{-1}\Sigma_{ST}}{\sigma_{TT}}, \end{aligned} \tag{6.15}$$

since all matrices involved are of dimension one.

Case 2: A Cross-sectional Surrogate for a Longitudinal True Endpoint

Now, turning to the model in (6.10), the R^2_Λ for a longitudinal true and a cross-sectional surrogate endpoint is

$$\begin{aligned} R^{2,CL}_{\Lambda,_{ST}} &= 1 - \frac{|\Sigma\,|}{|\Sigma_{TT}\,||\sigma_{SS}\,|} \\ &= 1 - \frac{|\sigma_{SS} - \Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}\,|}{|\sigma_{SS}\,|} \\ &= \frac{\Sigma_{ST}\Sigma_{TT}^{-1}\Sigma_{TS}}{\sigma_{SS}}. \end{aligned} \tag{6.16}$$

Comparing (6.15) with (6.16) establishes that $R^{2,LC}_{\Lambda,_{ST}} = R^{2,CL}_{\Lambda,_{ST}}$. In the first case, we used σ_{TT} and Σ_{SS} as component variances, of scalar and matrix type, respectively. These roles are reversed in the current, second case. Nevertheless, we obtain the same final expression for R^2_Λ as is, of course, entirely in line with the original, symmetric definition (6.13) of the quantity.

Furthermore, note that R^2_Λ and VRF are equal when the surrogate is longitudinal and the true endpoint cross-sectional. This implies that, only the VRF with the surrogate cross-sectional and the true endpoint longitudinal will be different from all of the others. This again highlights the feature that, for a longitudinal true endpoint, the VRF studies prediction of the entire sequence, while the R^2_Λ assesses how well an optimal linear combination of the true endpoint profile can be predicted. Both may be useful, but definitely are different. Moreover, one would expect the VRF to be well below the R^2_Λ in many applications, since prediction of an entire longitudinal sequence from a cross-sectional quantity is a tall order, whereas it might well be feasible to predict a particular linear combination.

The choice between the two measures lies in the objective to be attained. If the objective is to measure the strength of the surrogate to predict the entire sequence of the true endpoint, then VRF will be an ideal choice. However, when this seems an attainable goal or when we are rather interested in predicting some linear combination of the true endpoint, then we can resort to R^2_Λ.

6.3 Application to the Swim-stress Study

As described in Section 2.2, the focus of analysis is in two periods namely pre- and post-stress periods. Before some animals are subjected to stress, two experimental groups, the treatment (or compound) and vehicle groups are present, while after stress induction, four experimental groups appear. The four groups would therefore be; (1) vehicle alone, (2) vehicle and stress, (3) compound alone, and (4) compound and stress.

Figure 2.3 shows the group-specific mean profiles of CORT measurements, averaged over the four treatment periods. The plot depicts the average CORT values per treatment group at each time point, essentially showing how, on average, CORT values evolve over time in each treatment group. The need for flexible modelling tools is apparent from Figure 2.3, hence, as mentioned before, we discuss results emanating from an application of surrogate marker validation methodology in conjunction with flexible modelling techniques (penalized splines and fractional polynomial based), meant to appropriately capture trends over time. The fractional polynomial and pe-

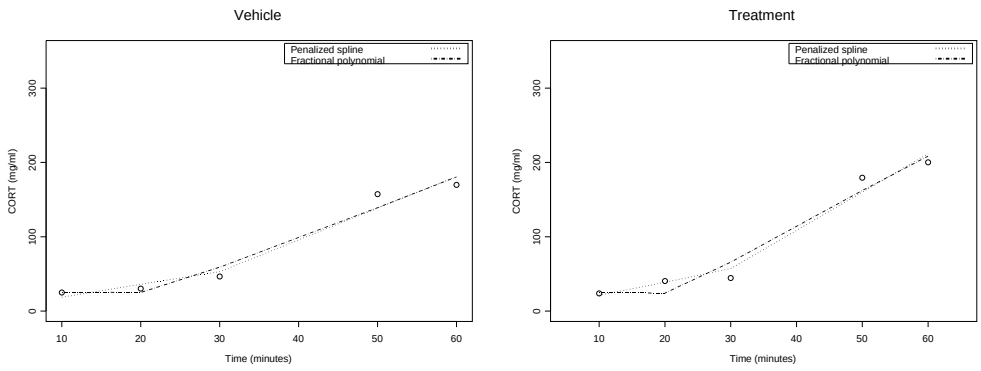


Figure 6.1: *Penalized spline and fractional polynomial fit to the data in the pre-stress period. The dots indicate the mean values at each time point averaged over the different periods.*

nalized splines approaches gave comparable fit to the data. For illustrative purposes, let us show some results for CORT in the pre-stress period. This period consists of 5 unique time points, and 4 knots, selected as quantiles of time were used. For the fractional polynomial, $p_1 = 0$ and $p_2 = 0$ were obtained, resulting in a quadratic effect in log(time) appearing in the model. The results from the two approaches are illustrated in Figure 6.1, and show a similar fit to the data.

For purposes of comparison, an unstructured mean model or a full factorial structure for time is also considered. However, this approach often yields excessively large numbers of parameters, thereby rendering it less desirable. The researchers wished to assess the association between the different responses before and after stress was induced. Thus, the results for pre- and post-stress correspond to the associations measured between the different responses before and after the stress with the treatment variable (G), having two possible values for pre-stress and having four different possible values after stress as explained above.

The VRF and R^2_Λ approaches have been applied to the dataset introduced in Section 2.2. The variance-covariance matrices, based upon which the VRF and R^2_Λ are computed, are estimated using maximum likelihood. The variance-covariance matrices can assume general structures unless the data suggests otherwise. In such cases, simple covariance structures, such as auto-regressive or compound symmetry, might be considered. For the purpose of our application, a number of models with different variance-covariance structures has been fitted. The best model, here being an unstructured variance-covariance structure, was chosen based on the AIC.

The results of the analysis for the association of telemetry and behavior as well as that of CORT and behavior are summarized in Table 6.1. We should like to point out that it is not a trivial task to derive a closed-form expression for the standard errors of VRF and R^2_Λ for the particular case we have considered here. However, fortunately, Alonso *et al* (2006) have shown that the VRF and R^2_Λ are special cases of the so-called *Likelihood Reduction Factor*, which is based on the information-theory approach. These authors have derived an asymptotic solution for standard errors for the LRF. Hence, by virtue of the relationship of these measures with the LRF, we have been able to provide asymptotic standard errors based on the information-theory approach. Standard errors for the estimates can also be obtained using bootstrap techniques.

There are no general guidelines as to how large a VRF and R^2_Λ should be in order to be considered sufficiently large. However, since the VRF and R^2_Λ are R-square type measures, it might be possible to make some general remarks concerning the degree of association based on their magnitude. Since such a degree of association

Table 6.1: *VRF and R^2_{indiv} values (asymptotic standard errors) under pre-stress and post-stress conditions, for a variety of true and surrogate endpoints, using unstructured, fractional polynomial, and penalized-splines models, and based on both VRF and R^2_Λ.*

endpoint		unstructured		fract. pol.		pen. splines	
true	surrogate	VRF	R^2_Λ	VRF	R^2_Λ	VRF	R^2_Λ
Pre-stress							
behavior	CORT	0.433(0.1178)	0.433(0.1178)	0.372(0.1174)	0.372(0.1174)	0.402(0.1179)	0.402(0.1179)
CORT	behavior	0.060(0.0632)	0.433(0.1178)	0.039(0.0533)	0.372(0.1174)	0.026(0.0463)	0.402(0.1179)
behavior	heart rate	0.807(0.0724)	0.807(0.0724)	0.816(0.0702)	0.816(0.0702)	0.798(0.0745)	0.798(0.0745)
heart rate	behavior	0.119(0.0850)	0.807(0.0724)	0.069(0.0669)	0.816(0.0702)	0.071(0.0677)	0.798(0.0745)
behavior	blood pressure	0.571(0.1105)	0.571(0.1105)	0.586(0.1091)	0.586(0.1091)	0.408(0.1179)	0.408(0.1179)
blood pressure	behavior	0.081(0.0717)	0.571(0.1105)	0.073(0.0685)	0.586(0.1091)	0.011(0.0823)	0.408(0.1179)
Post-stress							
behavior	CORT	0.386(0.1177)	0.386(0.1177)	0.499(0.1156)	0.499(0.1156)	0.359(0.1171)	0.359(0.1171)
CORT	behavior	0.038(0.0528)	0.386(0.1177)	0.045(0.0563)	0.499(0.1156)	0.032(0.00497)	0.359(0.1171)
behavior	heart rate	0.913(0.0415)	0.913(0.0415)	0.984(0.0108)	0.984(0.0108)	—	—
heart rate	behavior	0.227(0.1063)	0.913(0.0415)	0.126(0.0868)	0.984(0.0108)	—	—
behavior	blood pressure	0.343(0.1164)	0.343(0.1164)	0.513(0.1149)	0.513(0.1149)	—	—
blood pressure	behavior	0.079(0.0709)	0.343(0.1164)	0.160(0.0947)	0.513(0.1149)	—	—

arguably would vary from application to application, the final decision has to be made in consultation with the experts, regardless of the value. Having this in mind, from the results for the pre- and post-stress, we might infer that there is a rather weak relationship between behavior and CORT. However, strong and moderate relationships were observed between heart rate and behavior, and between blood pressure and behavior, respectively. Recall that behavior is measured cross-sectionally while CORT, heart rate, and blood pressure are longitudinal outcomes.

In this regard, when the cross-sectional outcome was used as a possible surrogate for the longitudinal outcomes, the VRF produced very low values, as anticipated in the previous section. Indeed, it is very difficult to predict the subtleties and richness of a longitudinal sequence from a single, cross-sectional measure. We consider this a desirable feature of the VRF. The R^2_Λ on the other hand, states that, although still small for some of the endpoints, there is better hope to predict a particular linear combination of the longitudinal outcomes from the cross-sectional outcome. As such, VRF and R^2_Λ both provide useful but totally *different* pieces of information. When there is role reversal, that is, when the longitudinal outcomes were treated as a possible surrogates for the cross-sectional outcome, the VRF values coincides with the R^2_Λ. This underscores that the VRF does not treat both endpoints symmetrically. The R^2_Λ, however, stayed the same even when there was role reversal, as expected from its construction.

The higher VRF and R^2_Λ values obtained when the longitudinally measured heart rate and blood pressure were used as surrogate endpoints for the cross-sectionally measured behavior, establish the possibility of predicting behavior using some linear combination of the longitudinal sequence.

Zooming in on the association between telemetry and CORT, both longitudinal in nature, we learn that there is a very weak association, with a maximum $\widehat{R}^2_\Lambda = 0.2314$ and maximum $\widehat{VRF} = 0.0513$, between the three modelling approaches. This is an indication that there is a very limited overlap in information between both outcomes, inhibiting comfortable prediction of one from the other.

In conclusion, the analysis has revealed that the longitudinally measured CORT level offers limited opportunity for prediction of activity, which is measured by the degree of alertness expressed in terms of the percentage of minutes the rats have been awake. We learn that heart rate and blood pressure are weakly related to CORT but have a strong predictive ability for behavior. The results advice against the use of activity to predict the longitudinal CORT level, heart rate, and blood pressure at each time point. These findings, however, have to be complemented with expert opinion before the results are to be practically used.

6.4 Discussion

In this chapter, we have adapted surrogate marker evaluation methods, originally designed to handle two repeated measures sequences, to the case of one cross-sectional and one longitudinal outcome, where either of these can be used as the surrogate. The methods have been applied to quantifying association between longitudinally measured CORT level, heart rate, and blood pressure, with cross-sectional behavior measured by the level of activity, expressed as the percentage of time experimental rats have been active after exposure to treatment followed by stress. The methods appear to work adequately for this particular mix of longitudinal and cross-sectional endpoints.

The various theoretical properties of the methods have manifested themselves in the results of the data analysis. In particular, it has been nicely confirmed that the VRF focuses on the prediction of a longitudinal sequence as a whole by a cross-sectional outcome, while R^2_Λ is concerned with the prediction of an optimal linear combination of the longitudinal outcome.

In the case of two longitudinal outcomes, the optimal linear combinations from the two outcomes are the first canonical variates. In the context of a longitudinal true and cross-sectional surrogate endpoint, the optimal linear combination could be the first principal component or any other summary measure of the longitudinal measurements, thereby maximally retaining information. Thus, optimality in this context refers to finding a linear combination that best summarizes the repeated measures.

The longitudinal outcomes were modeled using flexible modeling tools such as fractional polynomials, penalized splines, and a general unstructured mean where the time trend is not modeled but rather an analysis-of-variance type approach is followed. This offers the possibility of fitting different models and then selecting the best one according to some model selection tool such as, for example, AIC. It is, indeed, important to conduct proper modelling before moving into quantifying surrogacy, because the results may critically depend on the model's goodness-of-fit.

In all cases, VRF or R^2_Λ estimates close to one are indicative of 'good' surrogacy, with the reverse holding for values close to zero. Evidently, it is difficult to provide general advice as to how large is large enough. Arguably, the statistical evaluation of a surrogate can be an important component in the decision making process, but at least equally important is expert opinion coming in from pharmacological, biological, clinical, ethical, and health economy considerations.

7

Smoothing Neuronal Data with Penalized Splines

Penalized spline smoothing methodology and related applications have been discussed in the previous chapters. In all cases considered this far, the response has been assumed to be normally distributed, facilitating use of the linear mixed model. The current chapter shifts focus to non-normal data, specifically, Poisson counts, where the generalized linear mixed model (GLMM) paradigm becomes inevitable. The motivating example, as described in Section 2.3 comes from an electrophysiological experiment carried out with a monkey.

The data considered here involve the electrical activity in 20 different neurons recorded in the ventral premotor cortex (VPM) while a monkey performs a continuous discrimination task (CD task). In this task, the monkey reports a decision, based on the comparison of the orientation of two visual stimuli shown sequentially, separated by a delay. The main determinant of the neuron's discharge was whether the second stimulus (test) was to the left or right of the first (reference). Several trials are performed, classified according to orientation of the second stimulus with reference to the first (left or right) and the degree of difficulty (easy, difficult). For each trial, within the experimental period, i.e., 1000 ms to 2500 ms, time points where electrical activity was noted are recorded. To reduce the computational burden, the time scale

is subdivided into 20 ms bins and each bin represented by the median time point of that bin. Full details of the experiment are given in Section 2.3.

Usually, data from such electrophysiological experiments is summarized using a raster plot, displaying the complete set of spikes for each of the trials (Kass *et al.*, 2005). Also the peristimulus histogram (Gerstein and Kiang, 1960) can be used to summarize the overall activity and evolution in time by counting spikes in intervals of a certain width. Several ways to smooth instantaneous firing rates have been studied in the literature. An overview of the application of smoothing techniques in neuronal data can be found in Kass *et al.* (2003). Cadarso-Suárez *et al.* (2006) and Roca-Pardinas *et al.* (2006), for example, employ a flexible modelling technique based on the logistic Generalized Additive Model (GAM) with local linear kernel smoothers. Faes *et al.* (2007) apply a flexible method based on natural cubic splines to model synchrony in neuronal firing. Other recent techniques in this context include the Bayesian adaptive regression splines (DiMatteo *et al.*, 2001; Behseta and Kass, 2005; Behseta *et al.*, 2005). Flexible regression-based techniques come out favorable since they enjoy the flexibility of capturing the temporal evolution without the restriction of parametric modelling as well as the possibility to include covariate or factor information. We revisit this aspect in Section 7.1 where the models discussed happen to share similar properties.

The number of spikes accumulated over the different trials can be assumed to come from an inhomogeneous Poisson counting process (Ventura *et al.*, 2002; Cadarso-Suárez *et al.*, 2006), and our interest lies in estimating the instantaneous firing rate, denoted by λ_j. Estimating the instantaneous firing rate is necessary, especially in our situation where one of the main research goals is to determine the time trend and the time of maximal firing rate. Capturing the temporal structure with a parametric function may prove difficult or unsatisfactory. For example, Ventura *et al.* (2002) define a piecewise parametric function to describe the mean of the intensity function. One of the problems they face is that, for some neurons, the proposed model does not conform to the observed pattern. An attractive alternative is to model the time evolution by a flexible semiparametric function estimated through use of penalized splines. Molenberghs and Verbeke (2005) present an example together with SAS code for analyzing an ordinal outcome in a clinical-trial setting.

One of the main objectives of the study is to summarize certain characteristics of the time evolution of activity in the neurons from a population (all neurons) point of view. The important characteristics under consideration include time at which the maximum firing rate is observed. In addition, confidence intervals on the time of maximum firing rate are also required. We consider modelling the data from

two perspectives, namely, single-neuron analysis and a population-averaged approach. Modelling of the time evolution in both perspectives essentially follows a similar route as will be discussed shortly.

The contents of the present chapter are mainly based on the paper of Maringwa *et al.* (2008b).

7.1 Single Neuron Analysis

Let us introduce the methodology, first in the context of a single neuron, followed by the extension to the population of neurons. It should be noted that extension of the model to several neurons, in matrix notation, simply involves stacking together matrices corresponding to the different neurons.

7.1.1 Penalized Splines with Radial Basis

Let $y_j (j = 1, \ldots, T)$ represent the total count of activities recorded in bin j, aggregated over all the trials, t_j the median time point of bin j and $\kappa_1, \ldots, \kappa_K$ be a set of knots in the range of t_j. To flexibly model the response $\boldsymbol{y} = (y_1, \ldots, y_T)'$, consider the model

$$h(E[\boldsymbol{y}]) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_k\boldsymbol{b}\ , \tag{7.1}$$

where $h(\cdot)$ is an appropriate link function, such as the log link, and $\boldsymbol{X}$ and $\boldsymbol{Z}_k$ are design matrices of the form

$$\boldsymbol{X} = \left[1\ t_j\ \ldots\ t_j^{q-1}\right]_{1\le j\le T}, \quad \boldsymbol{Z}_k = \left[|t_j - \kappa_k|^{2q-1}\right]_{1\le j\le T,\quad 1\le k\le K}. \tag{7.2}$$

The parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are fixed and random effects, respectively. It is assumed that the random-effects vector $\boldsymbol{b}$ has a zero mean vector and covariance matrix (Ruppert *et al.*, 2003)

$$\mathrm{Cov}(\boldsymbol{b}) = \sigma_b^2(\boldsymbol{\Omega}_k)^{-1/2}(\boldsymbol{\Omega}_k^{-1/2})^T, \quad \text{with} \quad \boldsymbol{\Omega}_k = \left[|\kappa_k - \kappa_{k'}|^{2q-1}\right]_{1\le k,k'\le K}.$$

This defines the radial basis spline function of degree $2q-1$, for some positive integer q. Note that using a large number of unrestricted knots results in a wiggly fit. The constraint mentioned above therefore diminishes the effects, resulting in a smooth fit. To fit the model using standard mixed-model software, the transformation $\boldsymbol{Z} = \boldsymbol{Z}_k{\boldsymbol{\Omega}_k}^{-1/2}$ is applied, resulting in an equivalent model

$$h(E[\boldsymbol{y}]) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \quad \text{where} \quad \mathrm{Cov}(\boldsymbol{b}) = \boldsymbol{G} = \sigma_b^2\boldsymbol{I}. \tag{7.3}$$

Estimation of the parameters in the model above are obtained through the SAS procedure GLIMMIX, which uses pseudo-likelihood estimation techniques (Wolfinger and O'Connell, 1993). In particular, estimation is based on linearization of the outcome variable (see e.g., Molenberghs and Verbeke, 2005), resulting in the application of weighted least squares (McCullagh and Nelder, 1989). We return to the general case of this in Section 7.2.

Since interest is in a comparison of the time trend among different experimental conditions (e.g., left versus right, and easy versus difficult), a complex model is assumed which considers different smooth functions in each of the experimental conditions. Such a model can be said to have factor by factor by curve interactions (Ruppert *et al.*, 2003).

Let $\lambda_j = E(y_j)$ denote the mean at time point j. From (7.2), for $q = 2$, the penalized spline model defined for each experimental condition takes the form:

$$\log(\lambda_j) = \begin{cases} \beta_0^{LE} + \beta_1^{LE} t_j + \sum_{k=1}^{K} b_k^{LE} |t_j - \kappa_k|^3, & \text{if Left-Easy,} \\ \beta_0^{LD} + \beta_1^{LD} t_j + \sum_{k=1}^{K} b_k^{LD} |t_j - \kappa_k|^3, & \text{if Left-Difficult,} \\ \beta_0^{RE} + \beta_1^{RE} t_j + \sum_{k=1}^{K} b_k^{RE} |t_j - \kappa_k|^3, & \text{if Right-Easy,} \\ \beta_0^{RD} + \beta_1^{RD} t_j + \sum_{k=1}^{K} b_k^{RD} |t_j - \kappa_k|^3, & \text{if Right-Difficult,} \end{cases}$$

where, for example, $\beta_0^{LE}, \beta_1^{LE}$, and b_k^{LE} are the intercept, slope and knot coefficients for the left-easy combination of the experimental conditions. Note the model has been defined in an over-parameterized form and fitting such a model requires appropriate constraints on some of the parameters (see, e.g., Ruppert *et al.*, 2003).

It is assumed that the random-effects vectors $\boldsymbol{b}^{LE}$, $\boldsymbol{b}^{LD}$, $\boldsymbol{b}^{RE}$, and $\boldsymbol{b}^{RD}$ follow zero-mean normal distributions with equal variance-covariance matrix $\sigma_b^2 \boldsymbol{\Omega}_k^{-1}$. Although a common variance is assumed, the approach allows for different functions in each of the experimental conditions via independent sets of random effects. The most general model, which varies levels of smoothing by experimental condition, led to computational problems with some neurons, hence the approach above.

Using appropriate design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, the model can be written as (7.3), paving the way for implementation with mixed-model software. The model considered can easily be extended or reduced in several ways. For example, one can assume a constant shift in the curves, if that is considered a reasonable assumption.

For comparison of curves from different experimental conditions, we propose use of bias-adjusted simultaneous confidence bands around the fitted curves. Construction of such intervals requires the use of the variance-covariance matrix (see Ruppert *et*

al., 2003; SAS Institute Inc., 2004):

$$\boldsymbol{V} \;=\; \mathrm{Cov}\begin{bmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \hat{\boldsymbol{b}}-\boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{S}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{S}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{S}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{S}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix}^{-}.$$

Here, the matrix $\boldsymbol{S}$ is the conditional variance of the pseudo-data generated during the fitting process. A detailed discussion about this issue can be found in Molenberghs and Verbeke (2005).

Let $\boldsymbol{g} = (t_1, \ldots, t_T)$ be a set of values for which a simultaneous confidence band for $\boldsymbol{f_g} = \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_T) \end{bmatrix}$ is required. It can be assumed that, approximately,

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \hat{\boldsymbol{b}}-\boldsymbol{b} \end{bmatrix} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}\right). \tag{7.4}$$

For a true function value $f(t_j)$, denote the fitted value by $\hat{f}(t_j)$, and its bias-adjusted standard deviation by $\widehat{\mathrm{stdev}}\{\hat{f}(t_j) - f(t_j)\}$, which can easily be calculated using the corresponding entries in $[\boldsymbol{X} \quad \boldsymbol{Z}]$ and $\boldsymbol{V}$, as in (7.6). Simultaneous confidence bands for $\boldsymbol{f_g}$ can then be obtained as

$$\left[\hat{f}(t_j) \pm \tilde{h}_{(1-\alpha)}\widehat{\mathrm{stdev}}\{\hat{f}(t_j) - f(t_j)\}\right]_{1\le j\le T}, \tag{7.5}$$

where $\tilde{h}_{(1-\alpha)}$ is determined as in Chapter 4. The construction of confidence bands can be performed on the scale of the linear predictor and then transformed to intervals for the mean of the firing rate.

7.1.2 Derivation of the Time of Maximal Firing Rate and its Confidence Interval

One of the goals of this research is to detect the time at which the maximal firing rate occurs, together with the corresponding confidence interval. The derivative of λ_j, upon which construction of confidence intervals for the time of maximal firing is based, can be obtained explicitly based on the penalized spline representation as exemplified in Figure 7.3. First, we provide a description of how to obtain the maximal firing time, followed by a discussion about the derivative of the firing rate.

After estimating the instantaneous firing rate, optimization and specifically, the conjugate gradient method (Gill *et al.*, 1981; Fletcher, 1987) is applied to determine

the maximal firing time. The optimization of the firing rate function is implemented in the SAS procedure IML using the NLPCG subroutine. The maximum firing rate is an immediate by-product of the maximization procedure. As is commonly encountered with non-unimodal optimization problems, the optimization algorithm implemented converges towards local rather than global optima. The smallest local minimum of an objective function is called the global minimum, and the largest local maximum of an objective function is called the global maximum. It is therefore not unusual that the algorithm occasionally fails to obtain the global optimum. Therefore, several starting values within the time range of interest are used. The objective function is evaluated over all these possible candidates, and the time that gives the maximum function value is taken as the time resulting in the maximal firing rate.

This approach can be considered along the lines of a general concept for looking for features such as peaks, often referred to in literature as bump hunting (e.g., Heckman, 1992). Related approaches also include tests for monotonicity of regression functions (Gijbels *et al.*, 2000).

Let us now shift attention to the derivative of the objective function. Denote the derivative of $h(E[\boldsymbol{y}])$ with respect to time by $h^{'}(E[\boldsymbol{y}])$. Further, let $\boldsymbol{X}^d$ and $\boldsymbol{Z}^d$ be matrices containing derivative elements of $\boldsymbol{X}$ and $\boldsymbol{Z}$ as defined in (7.2). In general, $\boldsymbol{X}^d = \left[0 \quad 1 \quad 2t_j \ldots (q-1)t_j^{(q-1)}\right]_{1\leq j\leq T}$, and

$$\boldsymbol{Z}_k^* = \left[(2q-1)(t_j-\kappa_k)|t_j-\kappa_k|^{2q-3}\right]_{1\leq j\leq T,\ 1\leq k\leq K}.$$

Here we consider $q = 2$, and therefore,

$$\boldsymbol{X}^d = [0 \quad 1]_{1\leq j\leq T}, \quad \boldsymbol{Z}^d = \boldsymbol{Z}_k^*\boldsymbol{\Omega}_k^{-1/2}, \text{ where } \boldsymbol{Z}_k^* = [3(t_j-\kappa_k)|t_j-\kappa_k|]_{1\leq j\leq T,\ 1\leq k\leq K}.$$

It then follows from Section 7.1.1 that $h^{'}(E[\boldsymbol{y}]) = \boldsymbol{X}^d\boldsymbol{\beta}+\boldsymbol{Z}^d\boldsymbol{b}$. The derivative function should be zero at the time corresponding to the maximum firing rate. To construct a confidence interval for time of maximal firing, a confidence interval for the derivative function is constructed. Defining $g(\lambda_j) = h^{'}(E[y_j])$, the variance function at a particular time point j is (Ruppert *et al.*, 2003)

$$\text{var}\{(\hat{g}(\lambda_j) - g(\lambda_j))\} \simeq \boldsymbol{C}_j^T\boldsymbol{V}\boldsymbol{C}_j\ , \tag{7.6}$$

where $\boldsymbol{C}_j = \left[\boldsymbol{X}_j^d\ \boldsymbol{Z}_j^d\right]$. Construction of simultaneous confidence now follows the discussion in Section 7.1.1, with appropriate adjustments involving $\boldsymbol{X}^d$ and $\boldsymbol{Z}^d$. Confidence limits for the time of maximal firing rate are then taken as the points where the so-obtained confidence limits for the derivative function cross the zero line. Note that the first-order derivative function at the scale of the link function is readily obtained

by substituting the parameters by their estimates obtained from the model. Due to the monotonicity property of the link function, confidence intervals constructed for $h^{'}(\cdot)$ therefore suffice in this situation.

Similar approaches, making use of the first-order derivative function have been used in the literature. For example, Ganguli and Wand (2007) apply tests for feature significance using the significant zero crossings methodology (SiZer), owing to Chaudhuri and Marron (1999). Harezlak *et al.* (2007) employ bootstrap techniques to construct a test for bump hunting with penalized spline regression methodology. While these authors also make use of the first derivative of the objective function in the construction of their test, the main focus here is to determine the time corresponding to the maximal firing rate accompanied by a confidence interval.

7.2 Population-averaged Model: Combining Information from Different Neurons

Our discussion until this far has focused on data from a particular neuron. However, it is the goal to combine the information from different neurons, resulting in a so-called population based analysis. We propose the use of a marginal or population-averaged model (Molenberghs and Verbeke, 2005).
Formally, let $\boldsymbol{y}_i$ denote the vector of outcomes for neuron i ($i = 1, \ldots, 20$), $\boldsymbol{X}_i$ the matrix with covariate information, and $\boldsymbol{Z}_i$ the smoothing matrix. Stacking these matrices, a representation of the model for all neurons takes matrix form (7.3). For the current case of data assumed to follow a Poisson distribution, the distribution of the stacked data vector $\boldsymbol{y}$ is assumed to come from the 1-parameter exponential family distribution (McCullagh and Nelder, 1989) such that

$$f(\boldsymbol{y}|\boldsymbol{b}) = \exp\left\{[\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b})] - \mathbf{1}^T\eta(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}) + \mathbf{1}^T\vartheta(\boldsymbol{y})\right\}, \tag{7.7}$$

where, for the Poisson case, $\eta(x) = \exp(x)$, and $\vartheta(\cdot)$ is a function of the data. Penalizing the likelihood according to the distribution of the random effects $\boldsymbol{b}$ leads to estimates for $(\boldsymbol{\beta}, \boldsymbol{b})$ being the values maximizing

$$\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}) - \mathbf{1}^T\eta(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}\boldsymbol{G}^{-1}\boldsymbol{b}, \tag{7.8}$$

which is the penalized log-likelihood (Green, 1987). Parameter estimates are then obtained by using a linearized version of the response, $\boldsymbol{y}^* = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}^*$, where the pseudo-errors $\boldsymbol{\varepsilon}^*$ are assumed to be normally distributed with zero mean and a constant variance. Estimation proceeds by repeatedly fitting linear mixed models to

the pseudo-data (Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005) until convergence.

Note that the general case of the GLMM is of course a useful paradigm. Thus, one can consider a random-effects model, where neuron-specific random effects are used to account for the association. Such models are also handled by the procedure GLIMMIX. The marginal average evolution can then be obtained by averaging the conditional means over the random effects (Molenberghs and Verbeke, 2005), essentially integrating over them. Here, the random effects implied are neuron-specific effects, e.g., random intercepts, and not the random coefficients for smoothing. The former approach, which directly results in a population-averaged fit, is preferred in this situation. Within that approach, correlation between neurons can be directly specified in the modelling process.

Information from different experimental conditions from the different neurons is therefore combined to obtain condition-specific population-averaged profiles from which aspects of interest will be calculated. Since for each neuron, the number of trials per experimental condition varies, we fit the model with the number of trials as an offset variable.

7.3 Application to the Electrophysiological Experiment

Let us now turn to the application of the penalized splines methodology to the data described in Section 2.3. First, single neuron analysis is encountered in Section 7.3.1, followed by the population-averaged model in Section 7.3.2.

7.3.1 Single Neuron Analysis

Figure 7.1 displays data from a particular neuron, selected from the 20 neurons performing the CD task. The graph shows a raster plot from 176 trials as well as the fitted curves for each experimental condition, obtained using the penalized spline model. The model was fitted with 25 knots, obtained as equally spaced quantiles of the time (Ruppert, 2002; Ngo and Wand, 2004). For this particular neuron, one can observe the increased activity around 2000 ms into the experiment. The maximum firing rate occurs in this period. For inference, first, pointwise confidence intervals, followed by the simultaneous bands, are constructed. The pointwise confidence intervals are constructed using (7.5), with $\tilde{h}_{1-\alpha}$ replaced by $z_{1-\alpha/2}$. The more relevant

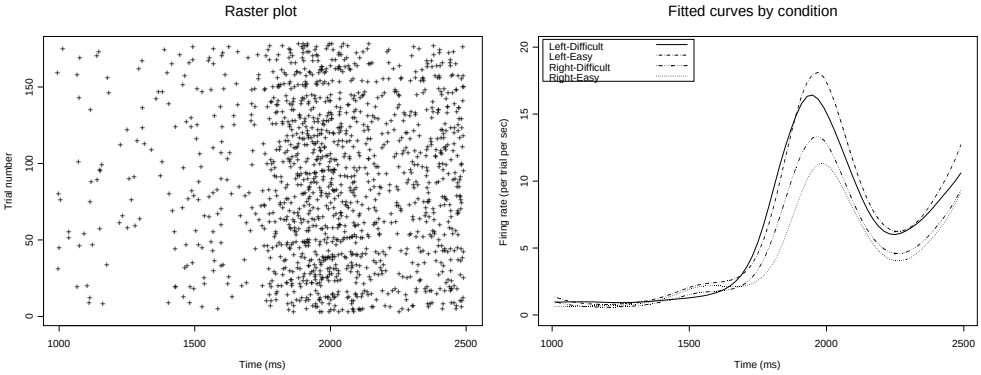


Figure 7.1: *Raster plot (left panel) and the corresponding fitted profiles by experimental condition (right) for a particular selected neuron.*

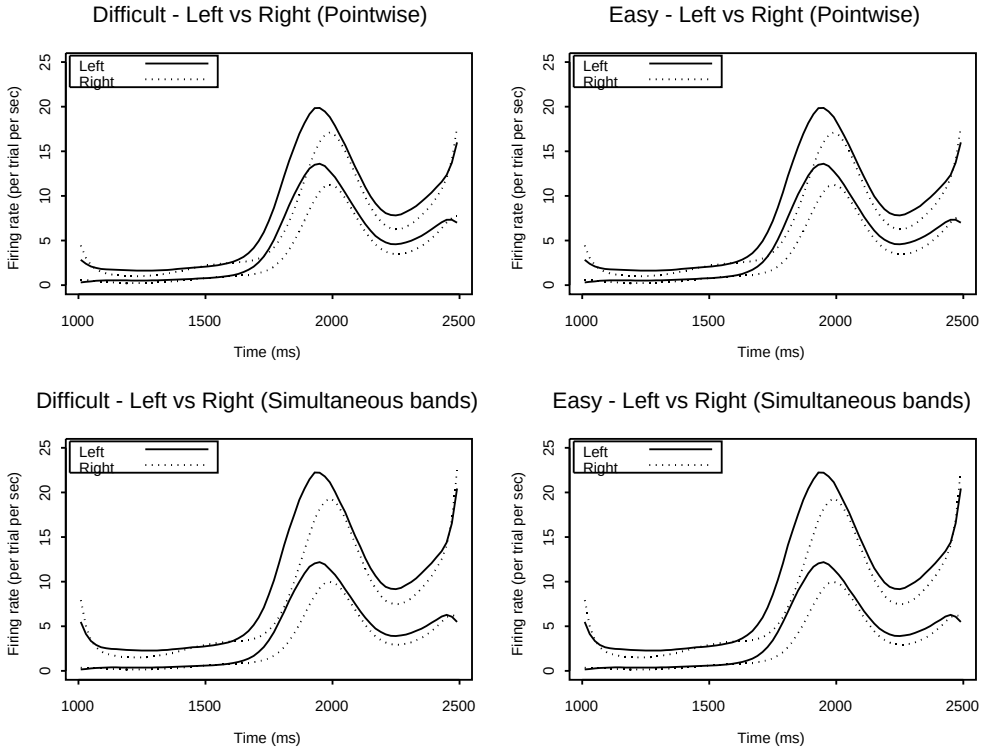


Figure 7.2: *Pointwise confidence intervals and simultaneous bands comparing the left and right orientation for a fixed level of difficulty.*

comparison of the left and right decisions, for a fixed level of difficulty, is based on pointwise confidence intervals and simultaneous confidence bands. Other than the small section between about 1750 and 2000 ms, which appears to show a difference, the pointwise confidence intervals appear to suggest no significant differences between the left and right decisions in this case. Using the simultaneous confidence bands, which are expected to be wider than the pointwise counterparts, the apparent difference mentioned above is absorbed and the hypothesis of no difference between left and right is upheld. These comparisons are graphically presented in Figure 7.2.

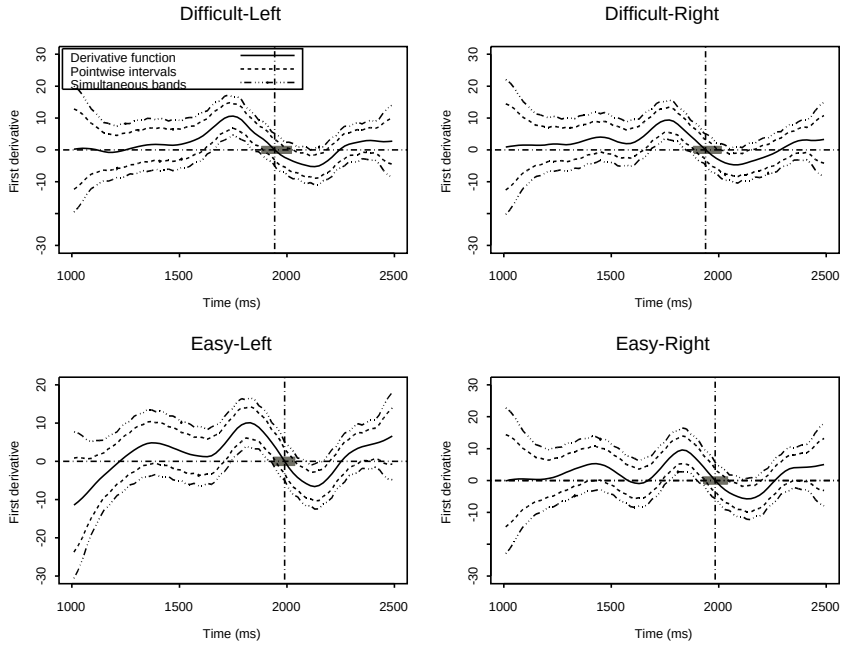


Figure 7.3: *First-order derivative functions (continuous line) for each condition together with corresponding 95% pointwise and simultaneous confidence intervals. The vertical dashed line indicates the time of maximal firing rate and the shaded regions indicate its corresponding confidence interval, approximated by the pointwise intervals.*

Figure 7.3 shows the first-order derivative function corresponding to the fitted profiles in Figure 7.1, together with their 95% pointwise confidence intervals and simultaneous bands. One can then approximate the confidence intervals for the time of maximal firing rate as suggested by the shaded areas in Figure 7.3. We return to the use of the derivative function and its confidence interval in the following section, where

a marginal model that combines information from different neurons is considered.

Table 7.1: *Maximal firing times and the corresponding approximate 95% derivate-based and bootstrap-based confidence limits. Limits from the Bayesian approach are obtained from the appropriate 95% credible intervals.*

		T_{max}(ms)	Derivative-based	Bayesian derivative-based	Bootstrap-based
Left	Difficult	1943	(1885; 2020)	(1890; 2030)	(1927; 1965)
	Easy	1939	(1885; 2010)	(1880; 2020)	(1920; 1965)
Right	Difficult	1990	(1945; 2035)	(1950; 2040)	(1970; 2010)
	Easy	1984	(1935; 2040)	(1935; 2040)	(1960; 2014)

To compare our confidence intervals with other approaches, a fully Bayesian hierarchical model (Gelman *et al.*, 1995; Ruppert *et al.*, 2003), as well as a nonparametric bootstrap approach (Efron and Tibshirani, 1993) have been applied. Attention is restricted to the first-order derivative function, focusing on limits of the maximal firing time. The bootstrap approach is based on resampling trials from each experimental condition and neuron. Aggregating activities from the obtained samples produces a bootstrap sample reminiscent of the original data, to which the penalized spline model is fitted. A total of 1000 bootstrap samples were used. The results from the pointwise intervals have been summarized in Table 7.1. The results show a close comparison between the limits on the time of maximal firing from the approach proposed here, and a fully Bayesian approach. The bootstrap technique appears to yield narrower limits compared to the other two approaches. It should be noted that the simplest case of percentile bootstrap intervals was applied here, and ways of improving such intervals are detailed in Efron and Tibshirani (1993).

To give an overview on individual neurons, a similar analysis has been performed on each of the other neurons. There appears to be relatively large variability between neurons (see also Figure 7.6). A comparison of left and right orientations based on confidence intervals for each neuron separately is performed and the resulting plots are given in Figures 7.4 and 7.5. The results indicate that although in most neurons, differences between left and right occur in the region 1500-2000 ms, for some neurons, differences occur elsewhere. Moreover, the maximal firing rates in the sections showing differences are highly variable, suggesting that different neurons have different peaks. As such, in a population analysis, it may be difficult to detect differences between left and right with differences in different places tending to cancel each other.

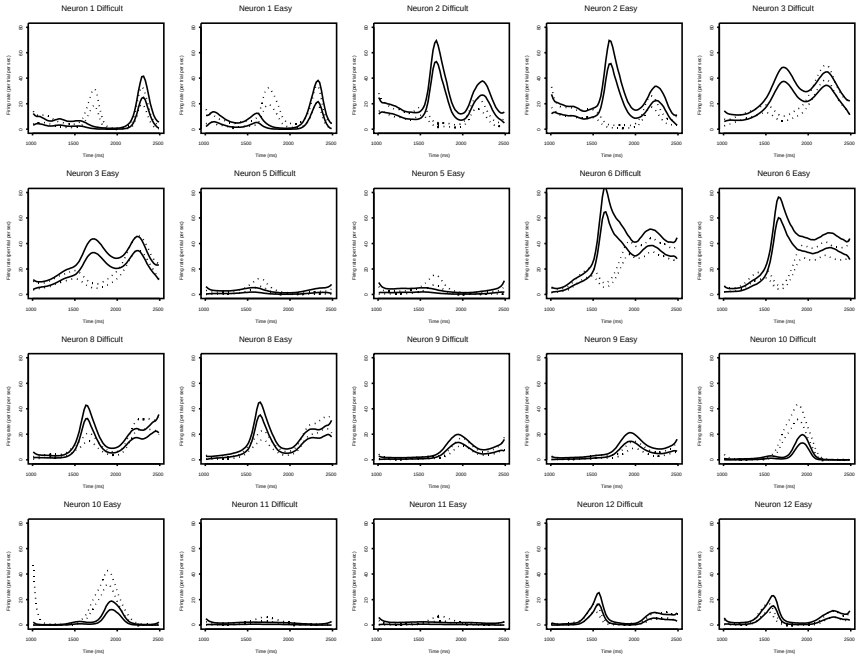


Figure 7.4: *Confidence intervals comparing left vs right decisions in individual neuron data: 1 (D represents "difficult" and E represents "Easy").*

7.3.2 Overall Average Profile

In this section, focus is put on the marginal or population-averaged model. Essentially, all data from the different neurons have been combined to produce condition-specific profiles. The top panel of Figure 7.6 shows individual neuron profiles for each of the experimental conditions wherein within-neuron variability appears substantial.

The model fitted assumes independent sets of random effects for smoothing for each experimental condition, albeit with the same variance component. This effectively produces different curves for different experimental conditions. Although the random effects are different, a single smoothing parameter is used, implying similar amount of smoothing in all experimental conditions.

Different types of correlation structures, for example, compound-symmetry structure or AR(1), can be specified. We present results based on the compound-symmetry structure, under which the model could converge. Note that an unstructured variance-covariance matrix would yield a computationally prohibitive number of parameters, and therefore, may not be a good choice.

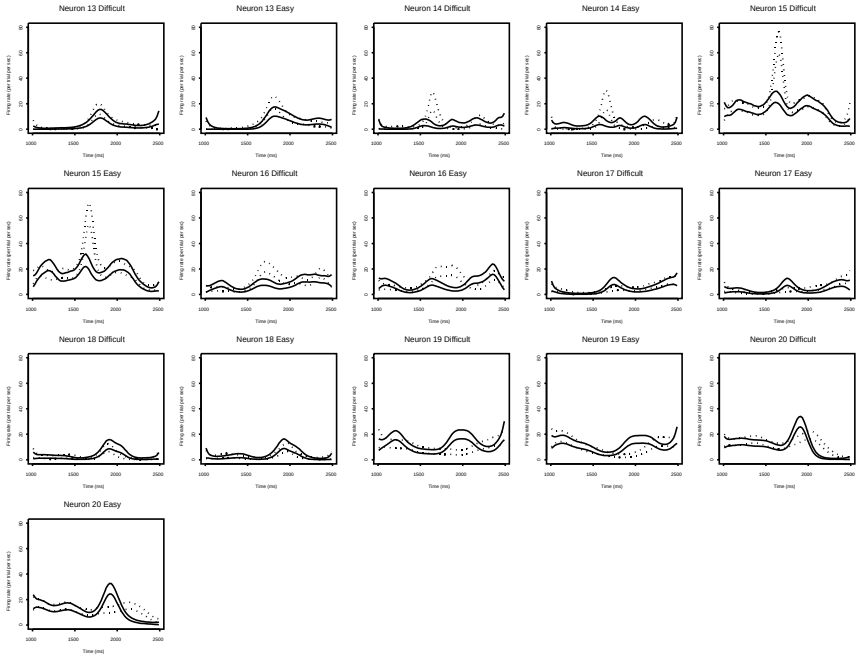


Figure 7.5: *Confidence intervals comparing left vs right decisions in individual neuron data: 2.*

The fitted profiles in each of the experimental conditions are given in Figure 7.6 (bottom panel). The plot shows that curves from the same decision (left or right) look rather similar, suggesting no differences between levels of difficulty for a fixed decision.

From Figure 7.6, one can observe increased firing activity for decisions to the left in the time period 1500 to 2000 ms. However, for decision to the right, no clear peak is evident, rather an overall increase is apparent between approximately 1500 and 2250 ms. The plot suggests that the time of maximal firing occurs earlier for decisions to the left compared to the right with the maximal firing rate being higher for decisions to the left.

Based on pointwise confidence intervals, apart from a small section between 1500 and 2000 ms, no differences between left and right are apparent in either of the levels of difficulty. Again, as one might expect, this apparent difference disappears as one considers simultaneous confidence bands. The simultaneous confidence bands used were found to be about 1.54 times wider than their pointwise counterparts. Note that

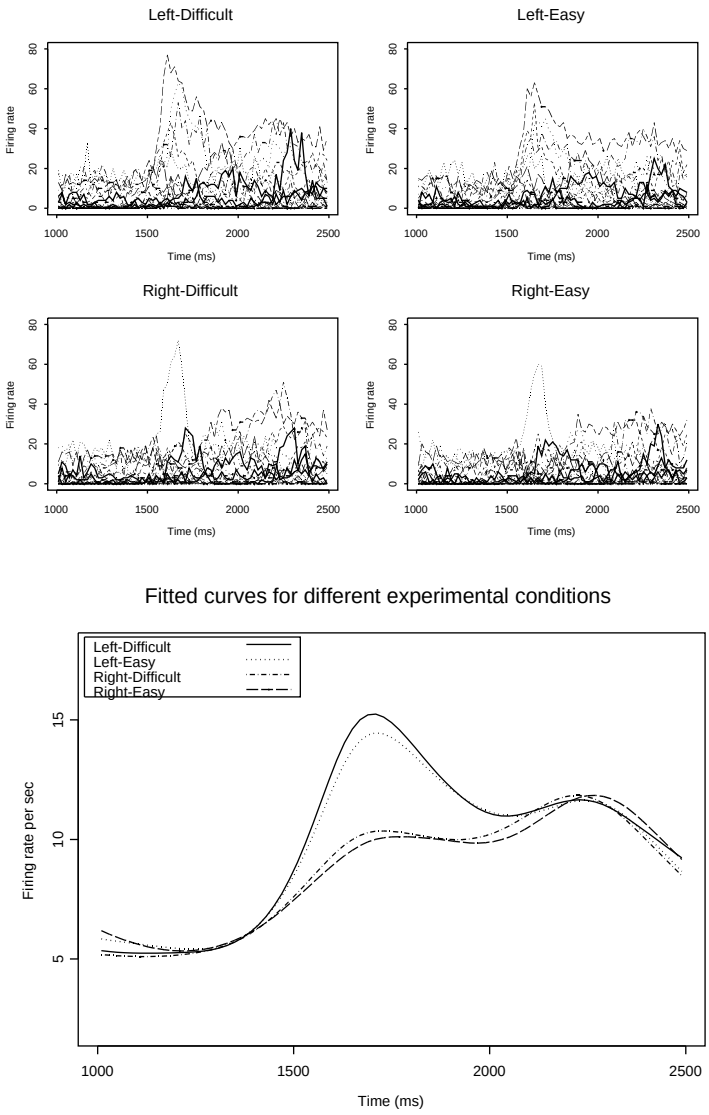


Figure 7.6: *Observed firing rates for all neurons in the different experimental conditions (top) and fitted curves in each condition obtained from the penalized spline model (bottom).*

the simultaneous confidence bands allow us to reach overall conclusions regarding differences or equality between the curves under comparison. Figure 7.8 shows the first-order derivative and its 90% pointwise and simultaneous confidence intervals,

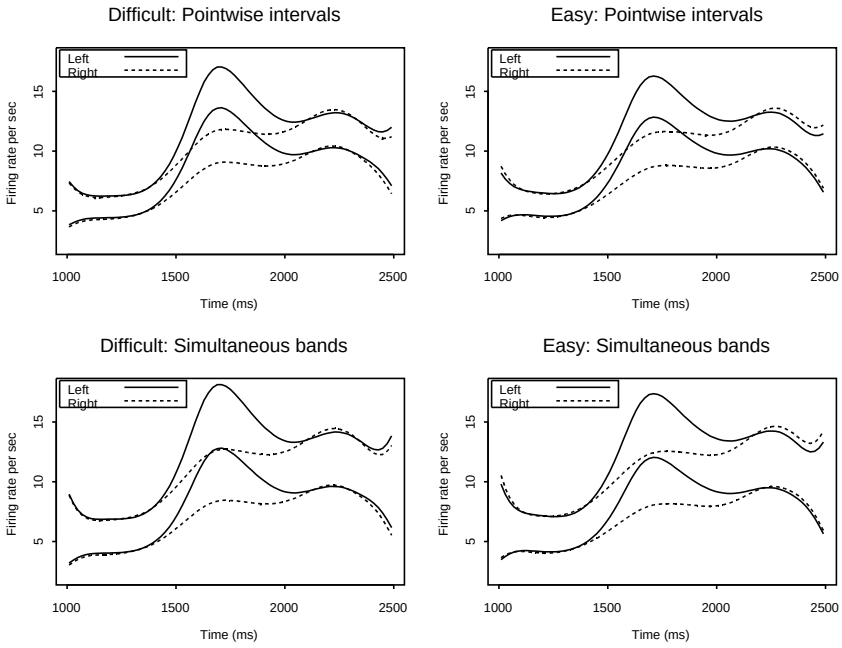


Figure 7.7: *Overall comparison of left and right orientations for a fixed level of difficulty using 95% pointwise and simultaneous confidence bands.*

together with an indication of the time of maximal firing. Since we already know the time of maximal firing, one would expect that its confidence interval may be deduced from the confidence interval of the derivative function, as exemplified by the shaded region in Figure 7.8. Note that in some experimental conditions, such as, for example, decisions to the right (see Figure 7.6), the time of maximal firing rate is not clear and may occur in a relatively wide region. In such instances, the upper and lower limits of the interval do not cross the zero line, leading to open-ended intervals. Here, we illustrate use of the proposed methodology using 90% confidence intervals. Such an interval can be interpreted as an interval such that in an indefinite repeat of similar experiments, 90% of the calculated confidence intervals for the maximal firing time will contain the true value of the time of maximal firing time. Table 7.2 displays the maximal firing times for each of the experimental conditions and the corresponding 90% pointwise and simultaneous confidence bands. The results in Table 7.2 suggest that for a fixed decision, there are no drastic differences between the levels of difficulty, neither in terms of the maximal firing rate nor the time of its

Table 7.2: *Maximal firing times and corresponding 90% pointwise and simultaneous confidence bands. Also given are the maximal firing times (in milli seconds) and maximal firing rates for the different experimental conditions.*

			Approximate 90% confidence intervals		
		T_{max}(ms)	Pointwise	Simult.	Firing rate
Left	Difficult	1703	(1668; 1765)	(1635; ∞)	15
	Easy	1712	(1668; 1795)	(1630; ∞)	14
Right	Difficult	2222	(1637; 2340)	(1550; ∞)	12
	Easy	2264	(1637; ∞)	(-∞; ∞)	12

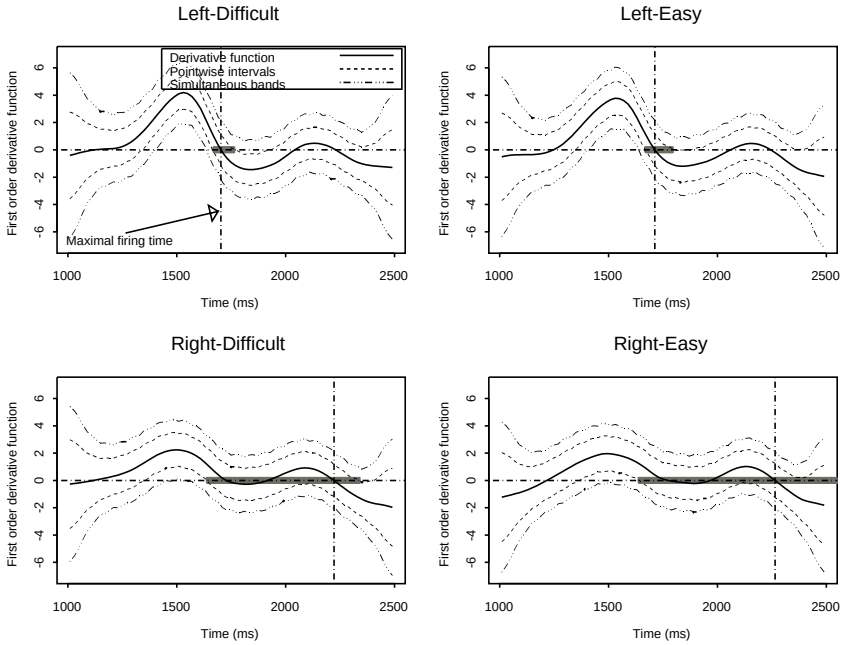


Figure 7.8: *First-order derivative functions (continuous line) for each condition together with corresponding 90% pointwise and simultaneous confidence bands. The vertical dashed line indicates the time of maximal firing rate and the shaded region indicates its corresponding pointwise confidence interval.*

occurrence. A similar conclusion may be drawn for the comparison between the left

and right-oriented decisions.

It is important to note the relatively wide confidence intervals/bands, especially for decision to the right. It is clear from the fitted curves in Figure 7.6 that for this condition, the maximum or peak may not be clearly determined, hence the wide confidence intervals on the time of maximal firing rate. The observed wide confidence intervals could be attributed to the large variability among the neurons, as mentioned before.

7.4 Discussion

We have considered an application of a flexible modelling technique, penalized splines, in smoothing neuronal data. This approach is convenient; it can be applied by means of widely available commercial software for mixed models. The models can also be fitted in the Bayesian framework, and consequently, WinBUGS, a publicly available free software package can be used. The models we have discussed are so general that they can be extended in several ways, exactly as required by the researcher. In particular, several possible scenarios depicting the evolution of curves in different experimental conditions, can be assumed. Differences or similarities can be assumed in the linear part of the models, the non-parametric part, or in both parts of the model. Extension of these models may also include variation of levels of smoothing in the different experimental conditions.

Focus has been on detecting the time of maximal firing rate and the maximal firing rate in a population of neurons, subjected to different experimental conditions. Moreover, we were also interested in comparing the temporal evolution across the experimental conditions, the comparison between the left- and right-oriented decisions being the main focus. The model we focused on is of a marginal or population-averaged type, wherein correlations of observations from independent subjects, neurons in this case, is specified. Different types of correlation structures can be used in this context. However, with the number of time points encountered in such electrophysiological experiments, some structures like the unstructured correlation are simply computationally infeasible. As a result, less computationally demanding structures, for example the compound-symmetry or the simple structure, can be used. To compare the curves in a moment by moment sense (Cardárso-Suarez *et al.*, 2006), simultaneous confidence around the fitted instantaneous firing rates are constructed. This effectively solves the problem of testing for a difference at multiple time points and therefore allows global conclusions in the time domain.

For the time of maximal firing one can fit the model and obtain the time corresponding to the maximum firing rate. However, the time of maximal firing rate does not necessarily have to be one of the design points, and therefore we implemented an optimization procedure based on the penalized spline model fitted, and confidence intervals on the time of maximal firing determined via the first order derivative function. It is clear that a number of properties of our proposed approach need to be investigated. For example, it is useful to assess the impact of the form of the underlying function, since it makes a difference whether it is constant, exhibits a single and small maximum, or features two local maxima. Also, the coverage probability of the confidence intervals need to be investigated. In addition, a comparison with alternative estimators at the population level is worth undertaking. Also, it is worth exploring further to what extent results depend on the choices made for the grid on the time variable. This is topic of further research.

The analysis performed here suggests no significant difference between the experimental conditions under consideration, both in terms of temporal evolution and the occurrence of the time of maximal firing rate. Three possible causes for the result in the population analysis may be anticipated: (1) there is temporal variation between neurons and this jittering provokes the lack of significance; (2) the different heights of discharge rate at single neuron level damped the differences at population level. If this were the case, perhaps normalizing the firing rates could solve it; (3) the firing rate maximum peaks for left and right are almost of the same height but occur at different times; as a consequence there is no statistical difference between the two peaks. In our situation the data may be considered as heterogeneous in some sense. This means that events occur at different times so peaks tend to cancel the possible differences. It would be interesting to compare the results with a population of neurons known to be homogeneous or in the same 'phase'.

It is important to mention that inference in general and hypothesis testing in particular is not straightforward due to the use of the pseudolikelihood. In general, conventional tools like the likelihood ratio test do not apply to semi-parametric models as discussed here (Ruppert *et al.*, 2003; Crainiceanu *et al.*, 2005a).

8

Bayesian Semiparametric Modelling of Univariate and Bivariate Longitudinal Data

In this chapter, we revisit the aspects discussed in Chapter 4 from another perspective, the Bayesian approach. Since penalized splines can be considered as BLUPS in the mixed-model framework, the models can be fitted using software developed for Bayesian analysis of mixed models. Some examples of work applying this methodology include Balandayuthapani *et al.* (2005), Crainiceanu *et al.* (2005b, 2007). The Bayesian approach becomes more attractive in this setting because one can directly monitor the difference between the groups, thereby rendering 'exact' inference. Moreover, the credible intervals derived thereof account for variability in all parameters in the model. In terms of modelling, similar settings, i.e., the same number and location of knots used in Chapter 4 will also be used here.

Within the context of smoothing longitudinal data, we also consider bivariate models for longitudinal processes. Several approaches may be considered for accounting for correlation between the responses (e.g., Thiebaut *et al.*, 2002; Molenberghs and Verbeke, 2005; Fieuws and Verbeke, 2006). In this chapter, we propose a bivariate model, wherein among other ways of accounting for correlation, correlation can be

imposed on the smoothers of the respective responses. This will be taken up further in Section 8.2. The work in the present chapter is also presented in Maringwa *et al.* (2008c).

8.1 Bayesian Approach to Semiparametric Mixed Models

8.1.1 Methodology

Unlike in frequentist and likelihood based statistics, where parameters are regarded as fixed but unknown quantities, a Bayesian approach treats the parameters as random. The classical mixed-model formulation (4.1), which already considers some parts as random, can be extended to a fully Bayesian model by taking all parameters in the model as random. Prior distributions are assumed on all parameters, thereby expressing some degree of knowledge about any particular parameter before data are available. The joint posterior distribution of the parameters given the data forms the basis for inference. When the dimension of the parameter vector grows, evaluating integrals appearing in the posterior density becomes non-trivial. Markov Chain Monte Carlo (MCMC) techniques can be applied to sample from the posterior distribution (Gelman *et al.*, 1995; Robert and Casella, 1999). The Bayesian inference for nonparametric models enjoys the flexibility of nonparametric models and the exact inferences provided by the Bayesian inferential machinery (Crainiceanu *et al.*, 2005b).

To provide a complete Bayesian specification of the model in (4.1), prior distributions on the parameters are required. Usually an improper uniform prior is considered for $\boldsymbol{\beta}$ and an inverse gamma distribution is assumed for the variance components. Reverting to the series of models defined in Chapter 4, a possible specification of the prior distributions for the parameters of Model 5, as an example, is as follows

$$\begin{cases} \beta_0 \sim \mathrm{N}(0, \sigma^2_{\beta_0}),\ \beta_1 \sim \mathrm{N}(0, \sigma^2_{\beta_1}),\ \beta_{01}, \sim \mathrm{N}(0, \sigma^2_{\beta_{01}}),\ \beta_{11} \sim \mathrm{N}(0, \sigma^2_{\beta_{11}}) \\ b^A_k \sim \mathrm{N}(0, \sigma^2_{b^A}),\ b^B_k \sim \mathrm{N}(0, \sigma^2_{b^B}),\ b_{0_i} \sim \mathrm{N}(0, \sigma^2_{b_0}) \\ \sigma^{-2}_{b^A}, \sigma^{-2}_{b^B}, \sigma^{-2}_{b_0}, \sigma^{-2}_{\varepsilon}, \sigma^{-2}_{\beta_0}, \sigma^{-2}_{\beta_1}, \sigma^{-2}_{\beta_{01}}, \sigma^{-2}_{\beta_{11}} \sim \mathrm{Gamma}(10^{-6}, 10^{-6}). \end{cases} \tag{8.1}$$

While we intend to fully apply the concepts of Bayesian inference, it is not our intention to discuss full details of the theory behind this wide subject here (see e.g., Gelman *et al.*, 1995; Robert and Casella, 1999). Extension to models discussed in Section 8.2 involves specification of the multivariate normal distribution on the effects, allowing correlated data structures. The models considered herein are all fitted in WinBUGS

(Lunn *et al.*, 2000), a freely available software package.

Out of the set of models discussed in Section 4.2, a single model is selected based on the Deviance Information Criterion (Spiegelhalter *et al.*, 2002), the smaller the DIC value, the better the model. Of particular interest in this study is determining differences, if present, between the group-specific profiles at each time point. As mentioned earlier, this can possibly be done in two ways, first based on credible intervals for the fitted functions in the two groups. The second approach is more direct and involves credible intervals for the population difference.

The Bayesian approach enables one to directly monitor the difference between the two functions and the resulting credible intervals can be used for inference. Such intervals emanate from the MCMC analysis. For example, the lower limit is the $\alpha/2$ sample quantile of the chain for the parameter of interest and the upper limit is the $1 - \alpha/2$ sample quantile. An appealing feature of this approach is that credible intervals allow for variability of each of the parameters, and do not use a so-called 'plug-in' approach (Crainiceanu *et al.*, 2005b).

Use of the Bayesian methodology depends on simulations from a presumed stationary distribution. As such, one needs to assess convergence properties. Here, we consider the diagnostic tool of Gelman and Rubin (1992), suitable for at least 2 chains. Basically, the method reports the ratio of the between-chain to within-chain variability. The comparison estimates the factor $\hat{R}$, by which the scale parameter of the marginal posterior distribution of each variable might be reduced if the chain were run to infinity (Best *et al.*, 1995). A factor of approximately 1 suggests that effective convergence may be assumed. Gelman *et al.* (1992) point out that a cut-off value of 1.2 works well for most cases, although, in other situations, a higher threshold may still be acceptable.

8.1.2 Application to the Cardiovascular Safety Experiment Parallel Design Case

For the single-variable longitudinal setting, we will focus on heart rate, applying the models discussed in Chapter 4. For clarity, the models are summarized in Table 8.1. Each of the five models is fitted and of particular importance is the DIC value for each model, used as an exploratory tool for discriminating amongst candidate models.

Based on 2 chains, each of 50 000 MCMC simulations, Table 8.2 gives the DIC values for the different models in the selection stage. The results indicate that Models 4 and 5 yield the lowest but very close DIC values, implying close similarity in the fit between the two models. For reasons of parsimony, one can focus mainly

Table 8.1: *Formulation of different semiparametric mixed models and the corresponding reference panel in Figure 4.1.*

Model	Description	Figure 4.1, panel
Model 1	Single curve for both groups	A
Model 2	Separate curves with no time interaction	B
Model 3	Separate curves with different linear effects but equal non-parametric part	C
Model 4	Separate curves smoothed separately with the same smoothing parameter	D
Model 5	Separate curves smoothed separately with different smoothing parameter	D

on Model 4, although it may also be interesting to compare results with those from Model 5. In particular, variation of the smoothing parameters between both models is investigated. For each of the models, it is safe to assume convergence since the $\hat{R}$ values obtained for all parameters in the model lie between 1.0 and 1.3. Figure 8.1 illustrates the fitted profiles from Model 4, which has lowest DIC value, as well as the 95% credible intervals for the fitted profiles. The model assumes separate curves for the two groups, which are smoothed separately with the same smoothing parameter. The credible intervals for Model 4 overlap in most sections of the experimental time period except the early stages of the experiment where a difference between the groups may not be unexpected. However, differences between the two profiles at each time point can be more easily noted in the profile of the difference, as illustrated in Figure 8.1 (right panel). The graph suggests significant differences between both groups in approximately the first 25 minutes of the experiment. The two models we have focused on involve smoothing the groups independently, first with the same smoothing parameter in the case of Model 4 and then varying the level of smoothing

Table 8.2: *DIC values for each of the five models.*

	Model 1	Model 2	Model 3	Model 4	Model 5
DIC	52489.7	52493.9	51917.6	51776.8	51779.0

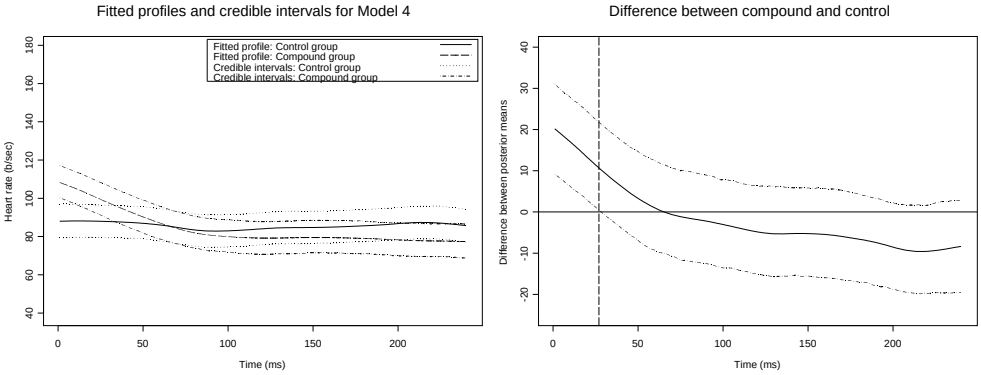


Figure 8.1: *Observed group-specific mean and fitted profiles for heart rate together with the difference between the groups.*

by group in Model 5. It may be enlightening to take a closer look at some parameters of interest from the two models. Of particular interest will be the residual variance as well as the variance of the random effects, responsible for smoothing, which can be used to determine the smoothing parameter using the relationship $\lambda = \sigma_\varepsilon^2/\sigma_b^2$. For purposes of comparison, we have also included parameter estimates obtained with the same data in the frequentist linear mixed-model approach (LMM). Both approaches lead to very similar results. The results in Table 8.3 show that when the groups are

Table 8.3: *Posterior mean and 95% credible intervals for parameters of interest in Model 4 and Model 5 from the Bayesian linear mixed models (BLMM) and parameter estimates and their standard errors for the same models obtained from the LMM. The LMM standard errors for λ, λ_A, and λ_B are obtained using the delta method.*

Model 4						Model 5					
	BLMM			LMM			BLMM			LMM	
Param.	Mean	2.5%	97.5%	Estim.	s.e	Param.	Mean	2.5%	97.5%	Estim.	s.e
$\sigma^2_{b_0}$	272	161.77	508.61	245.78	65.82	$\sigma^2_{b_0}$	270.25	166.39	522.76	245.34	65.59
σ^2_b	9.52	3.86	26.4	7.87	4.08	$\sigma^2_{b_A}$	8.84	0.54	79.94	5.32	5.39
						$\sigma^2_{b_B}$	9.21	3.64	26.97	8.99	6.03
σ^2_ε	129.10	124.7	133.80	129.06	2.23	σ^2_ε	129.00	124.8	133.60	129.09	2.24
λ	3.67	2.22	5.79	4.05	1.10	λ_A	3.81	1.28	15.54	4.93	7.79
						λ_B	3.74	2.19	6.00	3.79	1.30

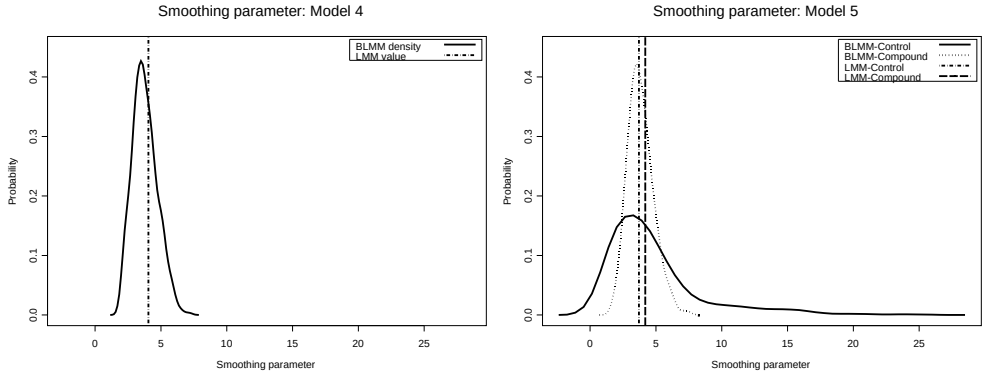


Figure 8.2: *The distributions of the smoothing parameters obtained from the Bayesian approach (BLMM) for Model 4 and 5 are shown, including values from the frequentist mixed-model approach (LMM).*

smoothed with varying levels, the corresponding variance components do not differ much. Indeed, such a conclusion was reported in Maringwa *et al.* (2008d), where a chi-square based test is used to test for the need to vary the amount of smoothing by group. However, the distribution of the variance component responsible for smoothing the control group appears more variable than for the compound group.

Figure 8.2 shows the distribution of the smoothing parameter(s) for models 4 and 5, complementing the similarity between the Bayesian and frequentist results. From Model 5, it is apparent that the smoothing parameter distribution for the control group is much more variable comparable to the compound group, resulting from the more widely varying distribution of $\sigma^2_{b^A}$, the variance component for smoothing the control group.

8.2 Joint Modelling of Bivariate Longitudinal Data

8.2.1 Methodology

Section 8.1.2 focuses on a single response measured in two independent experimental groups wherein smoothing in both groups may be assumed to be the same or varying. In this section we go a step further and consider bivariate smoothing. Specifically, we consider two longitudinal responses measured simultaneously on each subject in

the two experimental groups. Since the responses are recorded simultaneously, and to keep things relatively simple, we have decided to use the same number and location of knots for both responses in the two groups. The knots used are in fact the same as in Section 8.1.2. For each response separately, it is assumed the groups are smoothed independently, albeit the same smoothing parameter, as in Model 4 in Chapter 4. It is

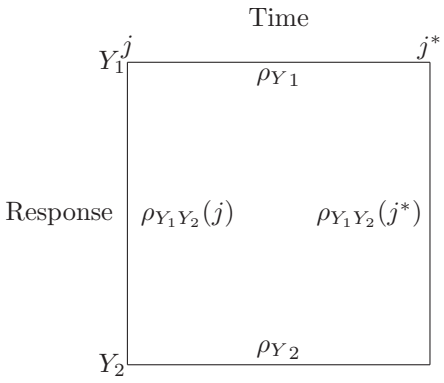


Figure 8.3: *Graphical illustration of the correlation structures of interest for two responses Y_1 and Y_2 at different time points j and j^*; ρ_Y denotes correlation of measurements within a particular response and $\rho_{Y_1Y_2}$ denotes correlation between both responses.*

expected that the two responses in a particular experimental group will be correlated and as such, the correlation ought to be accounted for. There are several ways of accounting for the correlation between the responses. Correlation can be imposed on residual errors across the responses, for example Molenberghs and Verbeke (2005) who jointly model a binary outcome and a continuous outcome in the context of surrogate markers for a clinical trials setting.

In the mixed-model framework, both responses may be jointly modelled by specifying a joint distribution for the random subject-specific effects. Examples of such an approach can be found in Thiebaut *et al.* (2002), Molenberghs and Verbeke (2005) and Fieuws and Verbeke (2006). In the same spirit, we propose another possible way of accounting for the correlation between both responses via the random effects respon-

sible for smoothing. It sounds logical that if responses evolve in a similar pattern, the knot coefficients for smoothing both groups will be correlated. This therefore yields the idea of imposing a correlation between the smoothers of both groups. Such a correlation may then be indicative of how strong the relationship between two responses' evolution over time is.

To formalize this discussion, let Y_{1ij} and Y_{2ij} $(i = 1, \ldots, n, j = 1, \ldots, m_i)$ be two different responses measured on the same subjects. For example, let Y_{1ij} denote the heart rate and Y_{2ij}, AoPs as discussed in Section 2.1.1. These two responses are measured simultaneously on each subject. Let us first consider Y_{1ij} and Y_{2ij} in one particular group, for example the control group. It is interesting to assess how the two responses jointly evolve in time, thereby capturing the time-varying relationship between the two. Earlier we have mentioned three possible ways of accounting for the correlation between the responses namely via correlated residual errors, correlated subject-specific random effects and correlated smoothers. We consider a number of models ranging from a model including at least one of these ways of accounting for correlation between responses to one having all three ways at the same time. For comparison purposes we also fit a model where no correlation is accounted for. A schematic representation of the possible correlations one can look at is given in Figure 8.3.

Suppose for each response, a random-intercept model is assumed. Further, response-specific residual errors are considered. Now consider a specific subject i with measurements for Y_1 and Y_2 at a particular time point j. The various penalized spline models at the fixed time point j can be expressed as:

$$\begin{cases} Y_{1ij} &= \beta_{01} + \beta_{11} t_{ij} + \sum_{k=1}^{K} b_{1k}(t_{ij} - \kappa_k)_+ + b_{0_{1i}} + \varepsilon_{1j}, \\ Y_{2ij} &= \beta_{02} + \beta_{12} t_{ij} + \sum_{k=1}^{K} b_{2k}(t_{ij} - \kappa_k)_+ + b_{0_{2i}} + \varepsilon_{2j}, \end{cases} \tag{8.2}$$

where $b_{0_{1i}}$, $b_{0_{2i}}$ are response-specific random intercepts, b_{1k} and b_{2k} are knot coefficients for smoothing the different responses. Let

$$\boldsymbol{Z}_{1j} = \boldsymbol{Z}_{2j} = \left[(t_{ij} - \kappa_1)_+ \ (t_{ij} - \kappa_2)_+ \ \ldots \ (t_{ij} - \kappa_K)_+ \right], \quad \boldsymbol{b}_1 = (b_{11}, \ldots, b_{1K})$$

and $\boldsymbol{b}_2 = (b_{21}, \ldots, b_{2K})$. Further, define matrices

$$\boldsymbol{Z}_j^b = \begin{bmatrix} \boldsymbol{Z}_{1j} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_{2j} \end{bmatrix}, \quad \boldsymbol{Z}^{b_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}, \quad \boldsymbol{b}_0 = \begin{bmatrix} b_{0_{1i}} \\ b_{0_{2i}} \end{bmatrix}.$$

Using matrix notation, (8.2) can easily be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$. From the

matrix definitions above, it then follows

$$\boldsymbol{D} = \text{Var}(\boldsymbol{b}) = \begin{bmatrix} \sigma_1^2 \boldsymbol{I}_K & \sigma_{12}^2 \boldsymbol{I}_K \\ \sigma_{12}^2 \boldsymbol{I}_K & \sigma_2^2 \boldsymbol{I}_K \end{bmatrix}, \quad \boldsymbol{G} = \text{Var}(\boldsymbol{b}_0) = \begin{bmatrix} \sigma_{b_{0_1}}^2 & \sigma_{b_{0_{12}}}^2 \\ \sigma_{b_{0_{12}}}^2 & \sigma_{b_{0_2}}^2 \end{bmatrix},$$

where $\boldsymbol{G}$ specifies the variance-covariance matrix of the subject-specific random effects and $\boldsymbol{D}$ represents the variance-covariance matrix corresponding to the smoothing terms. The variance-covariance matrix for the vector of measurements

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{1ij} \\ Y_{2ij} \end{pmatrix},$$

at a particular time point is

$$\begin{aligned} \text{Var}(\boldsymbol{Y}_i) &= \boldsymbol{Z}_j^b \boldsymbol{D} \boldsymbol{Z}_j^{b'} + \boldsymbol{Z}^{b_0} \boldsymbol{G} \boldsymbol{Z}^{b_0'} + \Sigma \\ &= \begin{bmatrix} \sigma_1^2 \boldsymbol{Z}_{1j} \boldsymbol{Z}_{1j}' & \sigma_{12}^2 \boldsymbol{Z}_{1j} \boldsymbol{Z}_{2j}' \\ \sigma_{12}^2 \boldsymbol{Z}_{2j} \boldsymbol{Z}_{1j}' & \sigma_2^2 \boldsymbol{Z}_{2j} \boldsymbol{Z}_{2j}' \end{bmatrix} + \begin{bmatrix} \sigma_{b_{0_1}}^2 & \sigma_{b_{0_{12}}}^2 \\ \sigma_{b_{0_{12}}}^2 & \sigma_{b_{0_2}}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_{12}}^2 \\ \sigma_{\varepsilon_{12}}^2 & \sigma_{\varepsilon_2}^2 \end{bmatrix}, \end{aligned}$$

where Σ is the variance-covariance matrix for residual errors. The correlation function is therefore given by

$$\rho_{Y_1 Y_2}^*(t) = \frac{\sigma_{12}^2 \sum_{k=1}^K (t_{ij} - \kappa_k)_+^2 + \sigma_{b_{0_{12}}}^2 + \sigma_{\varepsilon_{12}}^2}{\sqrt{\sigma_1^2 \sum_{i=1}^K (t_{ij} - \kappa_k)_+^2 + \sigma_{b_{0_1}}^2 + \sigma_{\varepsilon_1}^2} \sqrt{\sigma_2^2 \sum_{i=1}^m (t_{ij} - \kappa_k)_+^2 + \sigma_{b_{0_2}}^2 + \sigma_{\varepsilon_2}^2}}.$$

This expression is a general correlation function form for data from two responses, and therefore, represents a family of curves. The derived correlation function always depends on time via the spline formulation. In fact, since the spline coefficients are treated as random in this model, a particular correlation structure is implied.

Interpretation of this correlation is not straightforward. Note the dependence of the correlation function on time, a result of the 'sharing effect' at the level of the knot points. This correlation can be interpreted in two stages, in some hierarchical way. First, consider the correlation between the two splines, which reflects some sharing of knot penalties within and between both splines. This purely serves construction of both splines. Next, given the mean profiles, the correlation structure $\rho_{Y_1 Y_2}(t)$ takes the form

$$\rho_{Y_1 Y_2}(t) = \frac{\sigma_{b_{0_{12}}}^2}{\sqrt{\sigma_{b_{0_1}}^2 + \sigma_{\varepsilon_1}^2} \sqrt{\sigma_{b_{0_2}}^2 + \sigma_{\varepsilon_2}^2}},$$

which is similar to the correlation in the usual random-intercepts model. Note that the addition of higher order subject-specific effects, for example, random slope, induces a time-dependent correlation function.

8.2.2 Application to the Cardiovascular Safety Experiment Parallel Design Case

Let us now turn to our application. Of particular interest would be the correlation between the two responses of interest as well as the correlation between the smoothers. We consider correlated response-specific residual errors, correlated random intercepts and correlated smoothers. Our focus is to investigate how using one or a combination

Table 8.4: *DIC values from different models, where ρ_b, ρ_α and ρ_ε indicate presence (or absence) of correlation between smoothers, random intercepts and residual errors respectively. The estimated correlation between the two responses in the two experimental groups is denoted by $\hat{\rho}_{b_A}$ and $\hat{\rho}_{b_B}$ while $\hat{\rho}_{b_0}$ and $\hat{\rho}_\varepsilon$ estimate correlation between random intercepts and residual errors respectively.*

Model	ρ_b	ρ_{b_0}	ρ_ε	DIC	$\hat{\rho}_{b_A}$	$\hat{\rho}_{b_B}$	$\hat{\rho}_{b_0}$	$\hat{\rho}_\varepsilon$
I	0	0	0	75248.4				
II	0	1	0	75243.3			-0.34	
III	1	0	0	75250.3	0.54	0.71		
IV	1	1	0	75245.3	0.39	0.84	-0.34	

of these ways of accommodating correlated responses affects the implied or modelled correlation structure. Therefore a set of models ranging from a model with no correlation at all to a model including correlation between smoothers, random intercepts and residual errors is investigated. Table 8.4 shows all the different models under consideration together with their DIC values.

Based on the DIC values as shown in Table 8.4, we are inclined to consider Model II as the one describing the data best, based on DIC values. In this model, only the random intercepts are assumed to be correlated. Inclusion of a correlation between smoothers does not improve the fit. Table 8.5 gives estimates of some parameters of interest for Model II for the two responses considered. The estimates for heart rate do not differ much from those given in Table 8.3 for the univariate case.

For Model II, an indication of the correlation between heart rate and AoPs comes from the estimate of the correlation between the random intercepts, which is -0.23, the estimate of ρ (see also Table 8.5). The correlation is rather small.

Figure 8.4 shows plots of differences between experimental groups for each of the two responses obtained from the joint model. Differences between both groups can only be observed in heart rate and, moreover, only in the early stages of the profiles,

Table 8.5: *Posterior mean and 95% credible intervals for parameters for both responses from selected joint model. The correlation between the random intercepts for both responses is denoted ρ_{b_0}, while ρ denotes the marginal correlation of the observations.*

	Heart Rate			AoPs					
Parameter	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
$\sigma^2_{b_0}$	272.970	160.380	460.800	3.700	2.160	6.340			
ρ_{b_0}							-0.340	-0.620	-0.230
ρ							-0.230	-0.430	0.000
$\sigma^2_{b_A}$	13.180	1.380	56.030	0.150	0.002	1.010			
$\sigma^2_{b_B}$	8.910	3.300	20.660	5.720	2.260	11.560			
σ^2_{ε}	129.180	125.000	133.500	1.910	1.850	1.970			

in line with results in Section 8.1.2. A closer look at Table 8.4 reveals that not all the

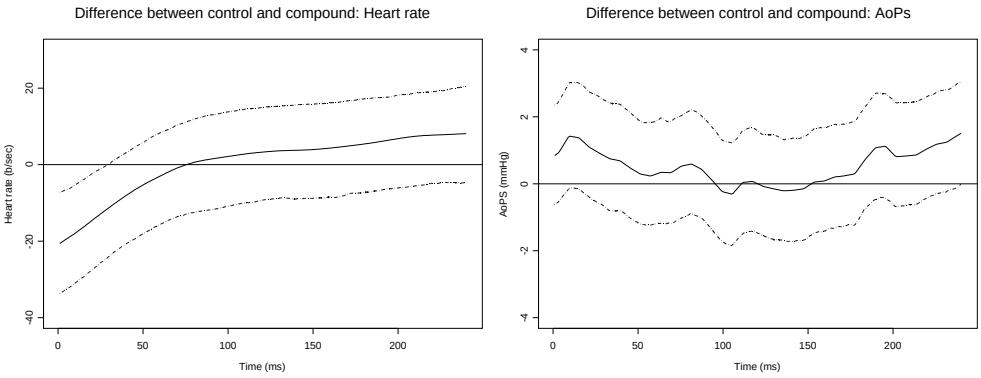


Figure 8.4: *Difference between group-specific profiles for heart rate and AoPs from the joint model, together with corresponding 95% credible intervals.*

anticipated eight models were fitted to the data. All the models including correlation between response-specific faced computational difficulties. This can be attributed to several things. The models considered are fairly complicated, and compounded by the relatively long sequences of measurements per-subject, computational difficulties may not be unexpected. To further investigate these models, a simulation study is set up as described in the following section.

8.2.3 Simulation Study

Simulation Settings

This section presents a simulation study aimed at assessing the plausibility of fitting the different models under consideration. In particular, it is investigated how well different parameter values are estimated, in light of the different forms of correlations specified. The setting of a longitudinal study is retained, although a relatively small number of time points is considered. Some of the models considered here involve com-

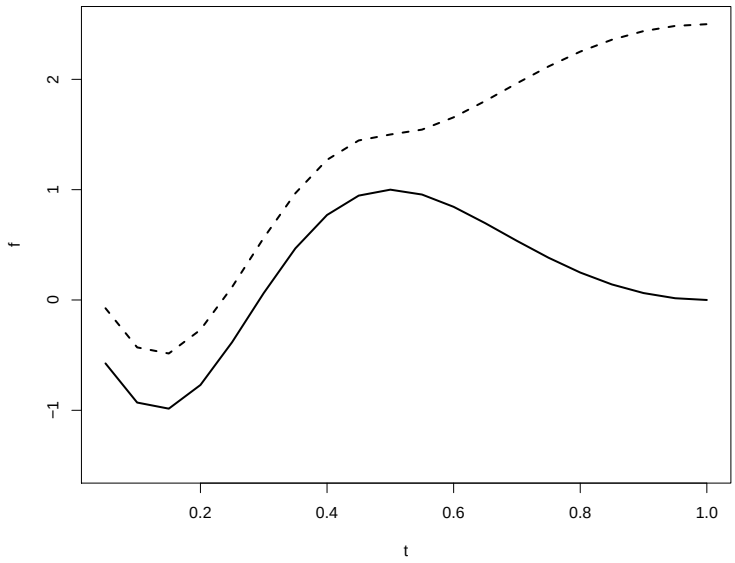


Figure 8.5: *The two functions used for generating data. The continuous line represents f_{1j} and the dashed line is for f_{2j}.*

plex combinations of correlations at different levels. The most complex model would include residual correlations between responses, correlations of the subject-specific random effects, as well correlations of the smoothers for the two responses. The main purpose of this exercise is to investigate how parameter estimates are affected by inclusion or exclusion of some of these types of correlations.

First, some convenient notation is introduced. The models considered here are denoted by M followed by three digits, each being either 1 or 0, indicating presence or absence of a certain correlation type in the model. The first digit represents

correlation between smoothers, the second, correlation between random intercepts of the two responses and the last denotes correlation between response-specific residuals. Thus, model M000 would represent the independence model, while M011 represents a model with correlated subject-specific effects as well as correlated residuals and so on. Note that all the different combinations result in eight models in total.

For purposes of this simulation study, data will be generated from model M011, and models M001, M011, and M111 are fitted to these data. Essentially, fitting models M001 and M111 is a way of investigating the effect of mis-specifying the model by; (1) exclusion of certain correlation types present in the data, i.e., in the case of M001 or (2) inclusion of a certain correlation, which is not there, in the case of M111.

Let us now provide some more detail on how data for two jointly measured longitudinal responses Y_{1ij} and Y_{2ij} $(i = 1, \ldots, n, j = 1, \ldots, m)$ is generated. It is assumed that $n = 20$ subjects are available, each with $m = 20$ measurements for each of the two responses. Let $t_{ij} \in [1/m, 1]$. Define two functions

$$\begin{aligned} f_1(t_{ij}) &= \sin(2\pi(1-t_{ij})^2) \\ f_2(t_{ij}) &= \begin{cases} 0.5 + \sin(2\pi(1-t_{ij})^2), & \text{if} \quad j \leq 10, \\ -(0.5 + f_{1j}) + 2f_1(10), & \text{if} \quad 11 \leq j \leq 20. \end{cases} \end{aligned}$$

The two functions are graphically depicted in Figure 8.5. Let variances be fixed such that $\sigma^2_{\varepsilon_1} = 0.10$, $\sigma^2_{\varepsilon_2} = 0.10$, $\sigma^2_{b_{0_1}} = 0.50$, and $\sigma^2_{b_{0_2}} = 0.50$. Further let the correlations take on values $\rho_\varepsilon = (0.50, 0.80)$ and $\rho_{b_0} = (0.50, 0.80)$, reflecting moderate and high levels of correlation. A combination of these settings results in four different cases to be investigated as shown in Table 8.6. Thus, as an example, for a particular subject i, ε_{1ij} and $b_{0_{1i}}$ are generated from multivariate normal distributions with zero-mean vectors, and respective covariance matrices

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{\varepsilon_1} & \rho_\varepsilon \sigma_{\varepsilon_1} \sigma_{\varepsilon_2} \\ \rho_\varepsilon \sigma_{\varepsilon_1} \sigma_{\varepsilon_2} & \sigma^2_{\varepsilon_2} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{G} = \begin{bmatrix} \sigma^2_{b_{0_1}} & \rho_{b_0} \sigma_{b_{0_1}} \sigma_{b_{0_2}} \\ \rho_{b_0} \sigma_{b_{0_1}} \sigma_{b_{0_2}} & \sigma^2_{b_{0_2}} \end{bmatrix}.$$

Thus data for the two responses is generated as $Y_{1ij} = f_1(t_{ij}) + b_{0_{1i}} + \varepsilon_{1_{ij}}$ and $Y_{2ij} = f_2(t_{ij}) + b_{0_{2i}} + \varepsilon_{2_{ij}}$. Note that, due to the way data is generated, imposing a correlation structure on the smoothers for both responses when generating data is not possible. However, when fitting the model, one can impose and estimate such a correlation structure, although of course there is no true value to compare with.

For a particular setting, 100 data sets are generated as described above and the model is fitted with 10000 iterations with a burn-in period of 1000 iterations. The results of interest include parameter estimates for specific components of the model.

Further, mean square error values split into squared bias and variance are also calculated.

Simulation Results

For each of the settings described in Section 8.2.3, it is investigated how parameter estimates would be affected under different model assumptions. Particular interest is on variance-covariance parameters as well as some correlations. As already mentioned, data are generated from model M011 and models M000, M011 and M111 are fitted to the same data. Table 8.6 shows the parameter estimates obtained for the different models. The results indicate that parameter estimates do not differ extensively from

Table 8.6: *Parameter estimates for models M001, M011 and M111 fitted from data generated using model M011. Gaps indicate parameters which need not be estimated or which do not appear in the particular model being fitted.*

$\sigma^2_{b_{0_1}}$	$\widehat{\sigma^2_{b_{0_1}}}$	$\sigma^2_{b_{0_2}}$	$\widehat{\sigma^2_{b_{0_2}}}$	$\rho_{b_{0_{12}}}$	$\widehat{\rho_{b_{0_{12}}}}$	$\sigma^2_{\varepsilon_1}$	$\widehat{\sigma^2_{\varepsilon_1}}$	$\sigma^2_{\varepsilon_2}$	$\widehat{\sigma^2_{\varepsilon_2}}$	$\rho_{\varepsilon_{12}}$	$\widehat{\rho_{\varepsilon_{12}}}$	$\widehat{\rho_b}$
Fit model M001												
0.50	0.5446	0.50	0.5529			0.10	0.1044	0.10	0.1043	0.50	0.4973	
0.50	0.5686	0.50	0.5458			0.10	0.1073	0.10	0.1075	0.80	0.7865	
0.50	0.5421	0.50	0.5483			0.10	0.1051	0.10	0.1046	0.50	0.5017	
0.50	0.5573	0.50	0.5291			0.10	0.1074	0.10	0.1079	0.80	0.7895	
Fit model M011												
0.50	0.5563	0.50	0.5863	0.50	0.4221	0.10	0.1032	0.10	0.1028	0.50	0.4864	
0.50	0.5845	0.50	0.5768	0.50	0.4245	0.10	0.1059	0.10	0.1061	0.80	0.7859	
0.50	0.5721	0.50	0.5625	0.80	0.6788	0.10	0.1045	0.10	0.1050	0.50	0.4959	
0.50	0.6012	0.50	0.6242	0.80	0.7098	0.10	0.1038	0.10	0.1049	0.80	0.7803	
Fit model M111												
0.50	0.5865	0.50	0.5952	0.50	0.4474	0.10	0.1027	0.10	0.1027	0.50	0.4547	0.8808
0.50	0.5999	0.50	0.5598	0.50	0.4688	0.10	0.1037	0.10	0.1040	0.80	0.8086	0.9034
0.50	0.6191	0.50	0.6152	0.80	0.7179	0.10	0.1026	0.10	0.1040	0.50	0.4717	0.8815
0.50	0.6068	0.50	0.6099	0.80	0.7233	0.10	0.1038	0.10	0.1049	0.80	0.7676	0.9076

the true values for all models. It appears there is no drastic influence on parameters estimated in the mis-specified models. However, variances of different components tend to be overestimated while correlations between effects tend to be underestimated.

For an in-depth assessment of how parameter are affected under different models, mean square error values split into its components of squared bias and variance are calculated and presented in Table 8.6. The results suggest that it is feasible to fit a model with all the three different forms of correlation discussed above. There appears to be no indication of large differences in mean square error values in comparison with other models. Hence, although in this particular case, M111 is a mis-specified model, parameter estimates tend to be estimated appropriately. Thus, one can actually fit such a complicated model without seriously affecting the different parameters in the model. The level of correlation either between random intercepts or between response-specific residuals also does not appear to have any bearing on bias, variance or the mean square error.

8.3 Discussion

This chapter has illustrated an application of the Bayesian inference methodology in the context of smoothing longitudinal data. The approach enables one to account for uncertainty associated with estimating all parameters in the model. Different models hypothesizing how the profiles in different groups can possibly evolve were considered. To narrow down the scope of the models, selection of the best model describing the data is done based on DIC values, readily obtained as part of the MCMC results. Focus turns to the selected model for inference. Note that selection of any other model other than the null already suggests difference between the groups, the form of which is determined by the model selected. Since our main aim was to detect specific time points exhibiting differences, attention is put on credible intervals around both the fitted functions and the difference between the group-specific profiles.

In the same context of smoothing longitudinal data using penalized spline smoothing, formulated as fully Bayesian hierarchical models, the possibility of jointly modelling two longitudinal profiles via imposition of correlated smoothers was investigated. Ideally, if profiles in a bivariate model tend to evolve in the same way, this should be captured by including a correlation between smoothers of both responses. Our simulations have shown that models with such type of correlation, on top of other types of correlation in the same model, are feasible, especially with relatively short sequences of repeated measurements. Although the approach sounds logically motivated, it is not short of pitfalls. Including such type of correlation obviously complicates models. While addressing the issue of correlation, the approach lands itself in problems of interpretation. By default, due to the construction of the penalized

spline model, a time dependent correlation always comes out. One possible route to circumvent this problem, as we have already seen, would to be to base conclusions on the conditional correlation. The possibility of including correlation on smoothers in tandem with higher order subject-specific random effects is also interesting to investigate. We believe further research along these lines is worthy indulging in.

Table 8.7: *Mean square errors, squared bias and variances for estimated parameters under the different model assumptions.*

	$\rho_{b_0} = 0.50, \rho_\varepsilon = 0.50$			$\rho_{b_0} = 0.50, \rho_\varepsilon = 0.80$			$\rho_{b_0} = 0.80, \rho_\varepsilon = 0.50$			$\rho_{b_0} = 0.80, \rho_\varepsilon = 0.80$		
Param.	var	bias^2	mse	var	bias^2	mse	var	bias^2	mse	var	bias^2	mse
					Fit model M001							
$\sigma^2_{b_{0_1}}$	0.0356	0.0020	0.0376	0.0359	0.0047	0.0406	0.0310	0.0018	0.0327	0.0431	0.0033	0.0464
$\sigma^2_{b_{0_2}}$	0.0371	0.0028	0.0399	0.0322	0.0021	0.0343	0.0307	0.0023	0.0330	0.0334	0.0008	0.0343
$\sigma^2_{\varepsilon_1}$	7.75e-05	1.89e-05	9.64e-05	6.01e-05	5.29e-05	1.13e-04	6.22e-05	2.55e-05	8.77e-05	5.38e-05	5.46e-05	1.08e-04
$\sigma^2_{\varepsilon_2}$	4.77e-05	1.89e-05	6.66e-05	6.78e-05	5.60e-05	1.23e-04	6.10e-05	2.12e-05	8.22e-05	5.69e-05	6.21e-05	1.19e-04
$\rho_{\varepsilon_{12}}$	1.24e-03	7.27e-06	1.25e-03	3.32e-04	1.82e-04	5.14e-04	1.40e-03	2.94e-06	1.41e-03	2.84e-04	1.09e-04	3.94e-04
					Fit model M011							
$\sigma^2_{b_{0_1}}$	0.0230	0.0032	0.0261	0.0374	0.0071	0.0445	0.0258	0.0052	0.0310	0.0259	0.0103	0.0362
$\sigma^2_{b_{0_2}}$	0.0318	0.0074	0.0393	0.0416	0.0059	0.0475	0.0219	0.0039	0.0258	0.0324	0.0154	0.0478
$\sigma^2_{\varepsilon_1}$	6.11e-05	1.01e-05	7.12e-05	6.26e-05	1.39e-05	7.65e-05	6.76e-05	3.58e-05	7.81e-05	6.73e-05	1.47e-05	8.20e-05
$\sigma^2_{\varepsilon_2}$	6.33e-05	7.84e-06	7.11e-05	4.81e-05	3.81e-05	8.63e-05	4.41e-05	2.55e-05	6.96e-05	5.68e-05	2.44e-05	8.12e-05
$\rho_{b_{0_{12}}}$	0.0340	0.0060	0.0400	0.0205	0.0056	0.0262	0.0116	0.0146	0.0263	0.0085	0.0081	0.0166
$\rho_{\varepsilon_{12}}$	0.0011	0.0002	0.0013	0.0003	0.0002	0.0005	1.44e-03	1.64e-05	1.46e-03	0.0005	0.0003	0.0008
					Fit model M111							
$\sigma^2_{b_{0_1}}$	0.0257	0.0075	0.0332	0.0262	0.0099	0.0361	0.0336	0.0141	0.0478	0.0338	0.0113	0.0451
$\sigma^2_{b_{0_2}}$	0.0289	0.0091	0.0380	0.0240	0.0036	0.0276	0.0302	0.0133	0.0435	0.0284	0.0121	0.0404
$\sigma^2_{\varepsilon_1}$	4.58e-05	7.44e-06	5.33e-05	6.26e-05	1.39e-05	7.65e-04	4.031e-05	6.92e-06	4.72e-05	6.57e-05	1.44e-05	8.01e-05
$\sigma^2_{\varepsilon_2}$	5.42e-05	7.05e-06	6.12e-05	4.57e-05	1.61e-05	6.18-05	4.36e-05	1.59e-05	5.95e-05	4.77e-05	2.43e-05	7.20e-05
$\rho_{b_{0_{12}}}$	0.0235	0.0027	0.0263	0.0208	0.0009	0.0218	0.0086	0.0067	0.0153	0.0078	0.0058	0.0137
$\rho_{\varepsilon_{12}}$	0.0017	0.0004	0.0021	0.0004	0.0004	0.0008	0.0012	0.0002	0.0014	0.0004	0.0003	0.0007

9

Bayesian Adaptive Penalized Splines for Non-normal Data

As already pointed out, flexible modelling techniques have become a common feature in statistical analysis. In particular, the use of penalized spline methodology has received wide attention in different applications requiring nonparametric smoothing.

Some of the key issues associated with smoothing literature include the positioning and the number of knot points to be used as well as the selection of the smoothing parameter. It turns out that choosing the smoothing parameter is a more subtle aspect (Ruppert, 2002). Regarding the knots, it is generally believed that with a certain minimum number of knots, an acceptable fit can always be obtained. The basic idea behind the penalized splines methodology is to shrink the coefficients of the knot points towards zero using some common variance, an aspect which can be referred to as global smoothing. Essentially one assumes all coefficients are drawn from a common distribution. This is exactly what we have been doing with penalized splines in all the previous chapters. However, with data exhibiting heterogenous tendencies, assuming a global smoothing parameter may be restrictive. Indeed, global smoothing tends to ignore the spatial variability in the data.

A host of methods have appeared in literature addressing this issue. Some of the work relates to juggling around with the number of knots and their positioning, for ex-

ample, the computationally intensive Bayesian Adaptive Regression Splines (BARS) of DiMatteo *et al.* (2001). A broad section of methods has mainly focused on relaxing the common variance assumption on the knot coefficients, for example, Lang and Brezger (2004). Some of the proposed methods focus on modelling the resultant knot-specific variances (penalty parameters) as a function of the independent variable. In this context, Ruppert and Carroll (2000) developed spatially adaptive penalty parameters in a frequentist setting with normal data. Baladandayuthapani *et al.* (2005) address the same problem but from a Bayesian perspective. Also from a Bayesian perspective, Crainiceanu *et al.* (2007) develop spatially adaptive parameters, extending the methods by further modelling of the error terms, also using penalized splines. Krivobokova *et al.* (2008) consider spatially adaptive parameters with an approximation to the marginal likelihood based on the Laplace transformation for non-normal data. This chapter proposes using the Bayesian inferential tools aiming at filling the gap of Bayesian spatially adaptive penalized splines for non-normal data. To the best of our knowledge, only the work of Krivobokova *et al.* (2008) addresses a similar problem with the Laplace approximation approach. The case for normal data from a frequentist approach has been studied for example by Ruppert and Carroll (2000). For normally distributed data in the Bayesian framework, we refer to Baladandayuthapani *et al.* (2005).

The work presented here follows closely that of Baladandayuthapani *et al.* (2005) and Krivobokova *et al.* (2008). The latter provide a fast way of fitting adaptive penalized splines for non-normal data by the use of a pseudo-quasi likelihood. However, it is well known that crude approximations to the marginal likelihood, especially for binary data, often give inaccurate results (Zhao *et al.*, 2006). The use of Markov Chain Monte-Carlo (MCMC) methodology (see e.g., Gelman *et al.*, 1995; Robert and Casella, 1999), which provides a source for 'exact' inference, is an appealing alternative. One can simultaneously estimate the function of interest, the penalty curve as well as their uncertainty bounds. Such methodology also provides tools for model comparison, e.g., the Deviance Information Criteria (DIC, Spiegelhalter *et al.*, 2002), which is not available in the penalized quasi-likelihood methodology.

The methodology is illustrated via an application to real data coming from the electrophysiological field. Further, some simulations to compare our proposed model with models already existing in literature are performed. For ease of exposition, attention will be restricted to cross-sectional data. It goes without saying that the methodology can seamlessly be used with longitudinal data as in the previous chapters. The contents of this chapter can also be found in Maringwa *et al.* (2008a).

9.1 Formulation of the Adaptive Penalized Spline Model

The main purpose here is to focus on adaptive smoothing in the case of non-normal responses. As such, consider a pair of data points $(y_j, t_j), j = 1, \ldots, T$, where y_j denotes the response variable and t_j is the independent variable. The model of interest can be expressed as $h(E(y_j)) = f(t_j)$, for a smooth function $f(\cdot)$, where $h(\cdot)$ is an appropriate link function, for example, the log link in the case of poisson counts. The penalized spline representation for $f(\cdot)$, based on a truncated linear basis, can be written as

$$f(t_j) = \beta_0 + \beta_1 t_j + \sum_{k=1}^{K} b_k (t_j - \kappa_k)_+,$$

where $\kappa_1, \ldots, \kappa_K$ are K distinct knots in the range of t_j $(j = 1, \ldots, T)$, with $u_+ = \max(0, u)$, and the knot coefficients, b_k, are assumed normally distributed with mean 0 and common variance σ_b^2, i.e., $b_k \sim \mathrm{N}(0, \sigma_b^2)$. The normality assumption on the parameters b_k assures that the data is not overfitted, but rather a smooth function is obtained.

The knots used here are selected using the quantile spacing approach (Ruppert, 2002) and a certain minimum number of knots is required to obtain an acceptable fit. The truncated lines basis, which is simple in formulation is used for explaining the methodology.

As before, we adopt the following matrix notation. Let

$$\boldsymbol{Y} = \left[\begin{array}{c} y_j \end{array} \right]_{1 \le j \le T}, \quad \boldsymbol{X} = \left[\begin{array}{cc} 1 & t_j \end{array} \right]_{1 \le j \le T}, \text{ and } \boldsymbol{\beta} = \left[\begin{array}{cc} \beta_0 & \beta_1 \end{array} \right]'.$$

Further, define

$$\boldsymbol{Z} = \left[\begin{array}{c} (t_j - \kappa_k)_+ \end{array} \right]_{1 \le j \le T,\ 1 \le \kappa \le K}, \quad \boldsymbol{b} = \left[\begin{array}{c} b_1, \ldots, b_K \end{array} \right]'.$$

The model may then be expressed in matrix notation as

$$h(E(\boldsymbol{Y})) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b},$$

with $\boldsymbol{b} \sim \mathrm{N}(0, \sigma_b^2)$. Note that, a common variance σ_b^2 is assumed for the different knot coefficients. This essentially shrinks the knot coefficients to zero, using a global variance component. This however, in certain cases, turns out to be restrictive. Indeed, often, profiles tend to have sharper turns or curves in some parts compared to other sections. A global smoothing parameter, which tends to combine information

borrowed from different sections of the profile obviously ignores such spatial variability. A way out is to relax this assumption and assume variable variance components for each knot point. The variance components, numbering up to the number of knot points K, are then modelled, again with a penalized spline model. The basis function for the spline model modelling the variance components can either be the same or different from the one for modelling the mean.

To fix ideas, let us explicitly formulate the model under consideration. The starting point to spatially adaptive smoothing is to assume a non-constant variance for the knot coefficients, such that, $b_k \sim \mathrm{N}(0, \sigma_b^2[\kappa_k])$, clearly emphasizing the dependence of the smoothing parameter on knot location. The K knots now define a new set of design points, denoted $t_1^c, \ldots, t_K^c$. In a similar way as above, another set of (sub) knots, numbering K^c and denoted $\kappa_s^c,\ s = 1, \ldots, K^c$, is selected. The variances responsible for smoothing the mean are then related to the independent variable using another penalized spline model,

$$\begin{aligned} \log(\sigma_b^2[\kappa_k]) &= \log(\sigma_b^2[t_k^c]) \\ &= \beta_0^c + \beta_1^c t_k^c + \sum_{s=1}^{K^c} c_s (t_k^c - \kappa_s^c)_+. \end{aligned}$$

Similarly as above, define the following matrices

$$\boldsymbol{Y}^c = \left[\ \log(\sigma_b^2[\kappa_k])\ \right]_{1\le k\le K}, \quad \boldsymbol{X}^c = \left[\ 1 \quad t_k^c\ \right]_{1\le k\le K}, \text{ and } \boldsymbol{\beta}^c = \left[\ \beta_0^c \quad \beta_1^c\ \right]^{'}.$$

Further, also define

$$\boldsymbol{Z}^c = \left[\ (t_k^c - \kappa_s^c)_+\ \right]_{1\le k\le K,\ 1\le s\le K^c}, \quad \boldsymbol{c} = \left[\ c_1, \ldots, c_{K^c}\ \right]^{'},$$

resulting in a similar mixed-model representation, $\boldsymbol{Y}^c = \boldsymbol{X}^c\boldsymbol{\beta}^c + \boldsymbol{Z}^c\boldsymbol{c}$. The coefficients c_s for the subknots κ^c are assumed to be normally distributed with mean zero and a constant variance σ_c^2, i.e., $c_s \sim \mathrm{N}(0, \sigma_c^2)$. The complete list of all parameters to be estimated to implement the spatially adaptive penalized spline model is $\boldsymbol{\beta}, \boldsymbol{\beta}^c, \boldsymbol{b}, \boldsymbol{c}, \sigma_c^2$.

The simplicity of the linear basis makes them conceptually easier to understand and implementation is rather straightforward. However, the models considered here are fitted using the radial basis (Ruppert *et al.*, 2003), which tends to be more numerically stable (Crainiceanu *et al.*, 2005b). Needless to mention that the basic idea remains exactly the same. The construction of radial basis functions was given in Chapter 7.

Consider a distribution from the exponential family with the canonical link function, for example, the Poisson distribution. In general, the distribution of $\boldsymbol{Y}$ is

$$[\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{b}] = \exp\{\boldsymbol{Y}^{'}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}) - \mathbf{1}^{'}\eta(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}) + \mathbf{1}^{'}\vartheta(\boldsymbol{Y})\},$$

where $\mathbf{1}$ is the unit vector and the distribution of the random effects $\boldsymbol{b}$ is assumed to be distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{G})$. The expression $\eta(\cdot)$ depends on the type of response where for example, in the Poisson case $\eta(x) = e^x$ and $\eta(x) = \log(1 + e^x)$ for the Bernoulli distribution. In a traditional penalized spline model with a global smoothing parameter, $\boldsymbol{G}$ will be diagonal with equal entries. However for the adaptive case, the issue is further complicated by the fact that diagonal entries of $\boldsymbol{G}$ are not only different, but also dependent on the independent variable $\boldsymbol{X}$.

9.2 Implementation as a Bayesian Model

The model proposed here is applicable in both the normal and non-normal cases. The main focus here is to specifically address the issue of Bayesian adaptive penalized splines (BAPS) for non-normal data, an area in which literature is not in abundance. To fully describe the model discussed Section 9.1 in the Bayesian framework, prior distributions for all parameters i.e., the fixed effects vectors $\boldsymbol{\beta}, \boldsymbol{\beta}^c$ as well as the variance component σ_c^2. The fixed effects are assumed to be independently normally distributed with a large variance, suggesting noninformative priors. For the variance components, an inverse gamma (IG) is considered for each variance component.

The model discussed here is a particular case of the general design Bayesian generalized linear mixed models of Zhao *et al.* (2006). Let us combine some parameters of interest in a single vector $\boldsymbol{\theta} = [\boldsymbol{\beta} \;\; \boldsymbol{b}]^{'}$. Further, let $\boldsymbol{C} = [\boldsymbol{X} \;\; \boldsymbol{Z}]$, and define $\boldsymbol{V}$ as a block diagonal matrix containing variances from fixed effects and variance components. The posterior distribution of the parameters given the data can then be written as (Zhao *et al.*; 2006)

$$[\boldsymbol{\theta}|\boldsymbol{Y}] = \frac{\int \exp\left(\boldsymbol{Y}^{'}\boldsymbol{C}\boldsymbol{\theta} - \mathbf{1}^{'}\eta(\boldsymbol{C}\boldsymbol{\theta}) - \frac{1}{2}(\log|\boldsymbol{G}| + \boldsymbol{\theta}^{'}\boldsymbol{V}^{-1}\boldsymbol{\theta})\right)[\boldsymbol{G}]d\boldsymbol{G}}{\int\int \exp\left(\boldsymbol{Y}^{'}\boldsymbol{C}\boldsymbol{\theta} - \mathbf{1}^{'}\eta(\boldsymbol{C}\boldsymbol{\theta}) - \frac{1}{2}(\log|\boldsymbol{G}| + \boldsymbol{\theta}^{'}\boldsymbol{V}^{-1}\boldsymbol{\theta})\right)[\boldsymbol{G}]d\boldsymbol{G}d\boldsymbol{\theta}}.$$

Evaluating such integrals analytically is an insurmountable task. Instead, alternative ways have been developed, making use of the MCMC methodology. In particular, one would employ algorithms that obtain samples from the required distribution. An example of such procedures is the so-called Gibbs sampler, based on conditional distributions of certain parameters given the other parameters in the model. For example, the conditional distribution of $\boldsymbol{\theta}$ is

$$[\boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{Y}] \propto \exp(\boldsymbol{Y}^{'}\boldsymbol{C}\boldsymbol{\theta}) - \mathbf{1}^{'}\eta(\boldsymbol{C}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^{'}\boldsymbol{V}^{-1}\boldsymbol{\theta}).$$

This conditional distribution does not resemble any standard distribution and therefore sampling from it becomes difficult (Ruppert *et al.*, 2003; Zhao *et al.*, 2006). The

Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) is one technique of obtaining samples from a distribution only known up to a certain constant. In brief, the technique generates samples from some distribution convenient to sample from and in comparison with the target distribution, the new sampled values are either accepted or rejected. The Bayesian models considered here are all fitted in WinBUGS (Lunn *et al.*, 2000). WinBUGS uses the Gibbs sampling technique, a technique that hinges on availability of full conditionals. Therefore Gibbs sampling becomes impossible when some or all of the required conditionals are difficult to sample from, which is the case for non-normal responses. In such cases, the Metropolis-Hastings algorithm may be called into play to provide approximations. Note that the Gibbs sampling technique is a special case of the Metropolis-Hastings algorithm. From a practical point of view, fitting Bayesian models in WinBUGS only requires one to correctly define and specify the model, estimation of parameters will be done automatically.

9.3 Application to the Electrophysiological Experiment Data Example

From the data described in Section 2.3, a single neuron with which a total of 185 trials were conducted, is selected and is the focus of interest in this chapter. The

Table 9.1: *Comparison of the Bayesian P-splines, Bayesian Adaptive P-Splines and KCK methods. The three methods are evaluated based on MSE and the Bayesian models are compared based on DIC.*

	Bayesian P-splines	Bayesian Adaptive P-splines	KCK
MSE	19.6444	19.4183	19.8178
DIC	1598.74	1591.19	

raw data for the particular neuron, in the form of raster plot, are shown in the left panel of Figure 9.1. The raster plot displays the complete set of spikes for each of the trials (Kass *et al.*, 2005). The right panel of Figure 9.1 shows the peristimulus time histogram (PSTH), which displays the number of spikes per second occurring in every 20 ms bins, averaged over all trials. Both plots show increased activity just after 1500 ms and also between 2000 and 2500 ms, possibly suggesting the need for differential smoothing in the time domain. Indeed, the profile appears flat in the beginning, sharply rising to a peak around 1500 ms. The intention is to illustrate the

use of Bayesian adaptive penalized spline methodology for smoothing complex data structures like the present situation. Data from the selected neuron, summarized

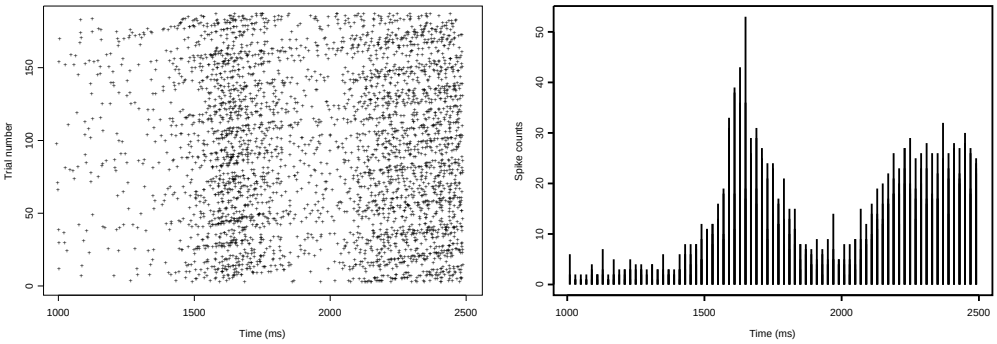


Figure 9.1: *Left: Raster plot showing the observed times at which activity was recorded. The plot shows for each single trial, the time at which an activity was recorded. Right: Peristimulus time histogram for the same data showing activity in 20 ms bins, averaged over all the different trials.*

over the different trials is shown in Figure 9.2 where the number of spike counts with time is shown. The adaptive Bayesian penalized spline methodology is illustrated on these data. Although the Bayesian adaptive spline model described in Section 9.1 is formulated based on the polynomial linear basis, the models fitted here are based on the radial basis. The basic concepts however remain the same regardless of the basis used. Following experimentation with a few different bases, it turns out that, for practical implementation, the radial basis is preferred due to numerical stability.

To achieve the desired flexibility, we use $K = 25$ and $K^c = 5$ where the knots are selected as quantiles of the time variable. The model is fitted with 4 MCMC chains, each with 50 000 iterations and a burn-in period of 10 000. The top left panel in Figure 9.2 shows the fit obtained from the Bayesian adaptive penalized spline model, together with the corresponding 95% credible intervals in the top right panel. Example WinBUGS code for fitting the model is given in the appendix.

For comparison, results of the method of Krivobokova *et al.* (2008), which for convenience, shall sometimes be referred to as the KCK method, are also presented. The KCK method was implemented using the R package `AdaptFit` written by the authors of that article and the results are presented in the bottom left and right panels of Figure 9.2. The top right and bottom right panels of Figure 9.2 show plots of the

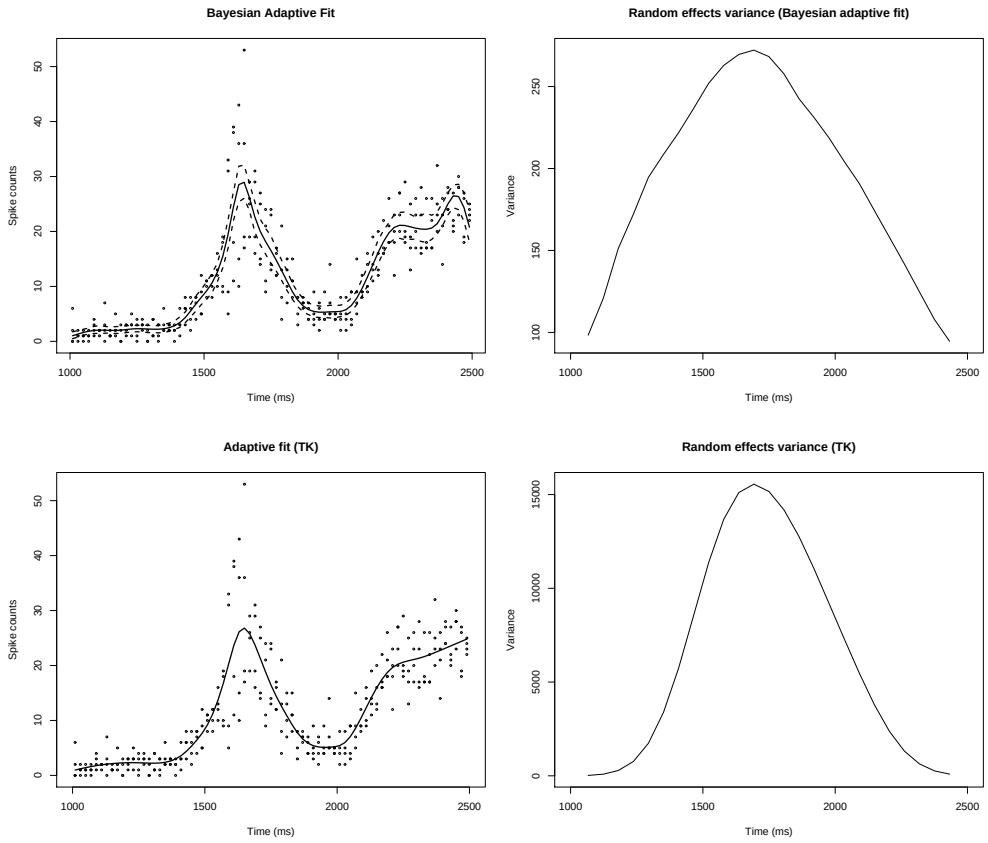


Figure 9.2: *Top: Fitted Bayesian Adaptive P-splines model together with 95% credible intervals as well as the variance function of the random effects for smoothing. Bottom: Fitted curve based on the KCK method and the corresponding variance of random effects function.*

random effects variance for the two approaches. The peak between 1500 ms and 2000 ms is associated with high random effects variances relative to other sections suggesting differential levels of smoothing. The differences in the range of values for random effects variances between both approaches is attributed to the difference in formulation of the basis functions. To compare these approaches on the current data, the mean squared error (MSE) is computed for each method. This is done by simply squaring the difference between the observed and the model predicted values averaged over the number of time points. Also considered for comparison purposes is the traditional Bayesian penalized spline model without modeling of the penalty

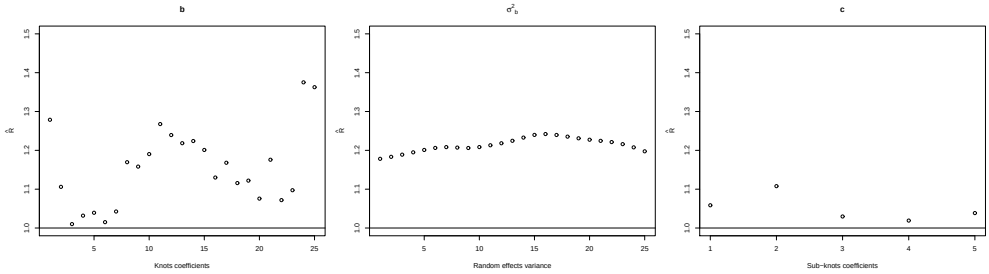


Figure 9.3: *Assessing convergence of knot coeffients* $\boldsymbol{b}$*, their variances* $\sigma_b^2[\kappa_k]$ *and sub-knots coefficients* $\boldsymbol{c}$ *using the* $\hat{R}$ *measure, which should be close to 1 if convergence is to be assumed.*

parameters (e.g., Crainiceanu *et al.*, 2005b). The results obtained from applying the different models are given in Table 9.1. The results indicate the adaptive methods give comparable results although the Bayesian approach appears to have a slight edge in terms of MSE values. The superiority of the Bayesian adaptive modelling in comparison with the traditional Bayesian penalized splines, which does not model penalty parameters, is evident from the MSE and DIC values. However, for a more informed comparison and or evaluation, a simulation study is required.

The DIC provides a way of formally checking for the need of an adaptive as compared to the simple Bayesian penalized spline. This is not possible in the KCK method, since the penalized quasi-likelihood cannot be used for comparison with other models. Note that one can also compare profiles based on confidence intervals or credible intervals in the Bayesian realm. Although this can be done, confidence intervals are not readily available from the method of KCK. On the other hand, credible intervals are inherently obtained from the Bayesian analysis. Figure 9.3

Table 9.2: *Assessing convergence status of the fixed effects parameters and the variance component for the sub-knots using the* $\hat{R}$*, which should be close to 1 for convergence to be safely assumed.*

	β_0	β_1	β_0^c	β_1^c	σ_c^2
$\hat{R}$	1.22	1.25	1.05	1.03	1.02

and Table 9.2 show an assessment of convergence using $\hat{R}$ for all model parameters for the Bayesian adaptive penalized spline model. Note that the vector of random effects $\boldsymbol{b}$ is of length 25 (left panel of Figure 9.3), resulting in the same number of

variance components (center panel of Figure 9.3). For the sub-knots vector $\boldsymbol{c}$ (right panel of Figure 9.3), the length is 5 but the corresponding variance components is a single value whose $\hat{R}$ value, together with the corresponding values for fixed effects parameters, are given in Table 9.2. Note that each and every parameter in the model has a value for $\hat{R}$. The results in Figure 9.3 and Table 9.2 suggest that convergence can reasonably be assumed since all $\hat{R}$ values do not deviate substantially from the reference value of 1. Gelman *et al.* (1992) suggest an umbrella cut-off value of 1.2, although they mention that, in complicated models, a higher threshold may still be acceptable.

The next section focuses on a simulation study focusing on comparing the proposed Bayesian adaptive method with similar methods currently existing in literature.

9.4 Simulation Study

To evaluate and compare the BAPS method with other existing methods, a simulation study is performed. For the sake of completeness, the case for normally distributed data is considered first, and in a second step, a similar exercise for non-normal data follows. For evaluation of the performance of the models, the squared bias, variance and the mean squared error for the estimated function are calculated. At each point t_j, one can define the local versions of the three quantities mentioned above. The mean of the estimated values at a particular point is $\bar{\hat{f}}(t_j) = \sum_{i=1}^{N} \hat{f}_i(t_j)/N$, where N is the number of simulation runs. The local squared bias, variance and mean squared error can then be obtained in a straightforward way as

$$\text{bias}_j^2 = (f(t_j) - \bar{\hat{f}}(t_j))^2, \quad \text{var}_j = \sum_{i=1}^{N} (\hat{f}_i(t_j) - \bar{\hat{f}}(t_j))^2/N, \quad \text{and} \quad \text{mse}_j = \text{bias}_j^2 + \text{var}_j,$$

where $f(\cdot)$ and $\hat{f}(\cdot)$ denote the true and model predicted values respectively. The global versions of the squared bias, variance and mean squared error are obtained by averaging the local values over the number of time points.

9.4.1 Simulation Settings

Let us give a description of the simulation settings considered here. The most commonly encountered data distributional assumptions namely the normal, bernoulli and the poisson are investigated.

The Normal Case

The function we consider for generating the data has already been used by several authors including Ruppert and Carroll (2000), Crainiceanu *et al.* (2007) as well as Krivoboka *et al.* (2008). Assume the model

$$y_j = f(t_j) + \varepsilon_j, \quad \text{for} \quad 1 \leq j \leq T,$$

where $\varepsilon_j \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and

$$f(t) = \sqrt{t(1-t)} \ \sin\left(\frac{2\pi(1+2^{(9-4\jmath)/5}}{t+2^{(9-4\jmath)/5}}\right), \tag{9.1}$$

with $\jmath$ controlling the level of spatial heterogeneity. For 400 values of t_j, assumed to be fixed and equally spaced on $[0, 1]$, it is assumed that $\sigma_\varepsilon^2 = 0.04$ and $\jmath$ is taken to be 3, allowing moderate spatial variability.

For the ordinary Bayesian penalized spline model, the number of knots is fixed at $K = 40$, selected as equally spaced quantiles of t. The same number of knots is also used in the adaptive version of the model, with the number of knot points for smoothing the variance components fixed at $K^c = 4$. Whilst we make an attempt to obtain a desired fit, it is also apparent that K^c should be much smaller than K for obvious reasons of computational complexity. For each simulated data, the Bayesian models are fitted with 10 000 iterations with a burn-in period of 1000. For the same data, the method proposed by Krivobokova *et al.* (2008) is also considered, using the same values of K and K^c. In all cases, 100 data sets are simulated, keeping the sample size at 400. Results of the simulation exercise are given in Section 9.4.2.

The Binary Data Case

Let us now shift attention to the case of non-normal data, and in particular, for this section, consider binary data. We seek to evaluate the proposed method using a simulation study, in comparison to the other existing methods already mentioned. Generation of the data follows the following simple procedure. Let p denote some proportion and, making use of (9.1), one can define $p(t) = \exp(f(t))/(1 + \exp(f(t)))$, which is considered a true known function of the proportion. The next step then involves generating data, i.e., either 1 or 0 at each value of t, from a bernoulli distribution with probability $p(t)$, thereby generating a sequence of binary outcomes.

The proposed method is first illustrated, in comparison to the method of Krivobokova *et al.* (2008). We generate a single data set with a binary response as explained above. The true function generating the data is shown in Figure 9.4. For two different sample

sizes ($T = 400$ and $T = 3000$), the model proposed in this chapter as well as that of Krivobokova *et al.* (2008) are fitted and graphically presented in Figure 9.4. The number of knots are kept the same as in the previous section.

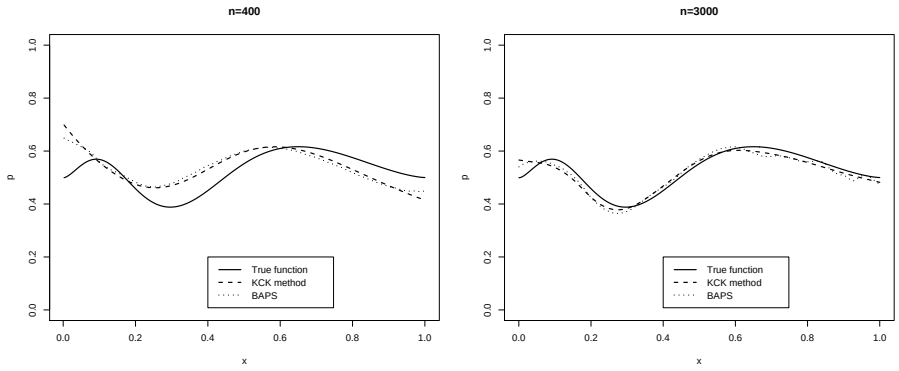


Figure 9.4: *Illustrating adaptive fitting methods on a single simulated data set. The methods are compared on two different sample sizes of* $T = 400$ *(left panel) and* $T = 3000$ *(right panel).*

With regards to smoothing of binary data, it appears the gain or the use of the adaptive methods largely depends on sample size. While Krivobokova *et al.* (2008) mention that their method does not necessarily require large samples, our application of the method shows poor performance in relatively small samples. This is evident in the left panel of Figure 9.4, which is based on a sample size of 400. Although the model gives a similar fit to the Bayesian model proposed here, both models appear to perform poorly in estimating the true function. The situation however improves when the sample size is taken up to $T = 3000$, albeit the same setting for the knots. As can be observed in the right panel of Figure 9.4, a more pleasing fit from both models is obtained. Of course, a proper simulation exercise can enlighten this situation. It is important to note that simulations involving such Bayesian models, especially with large sample sizes, can be computationally hard and time consuming.

Let us now discuss a proper simulation study to compare the different models. Here we assume $T = 1000$, which relatively, reduces the computational burden whilst still large enough to guarantee the possibility of an acceptable estimate of the true function. Indeed, we have seen that sample sizes as low as 400 can result in poor estimation of the true function, while relatively large sample sizes render a simulation study infeasible.

Again as above, consider p as the true function on the logit scale. We generate 100

data sets, each with a binary response generated at each time point using probabilities $p(t)$. For each data set, the ordinary Bayesian and the Bayesian adaptive models, together with the approach of Krivobokova *et al.* (2008) are fitted and the results are summarized in Table 9.3.

The Poisson Case

As a follow-up to the binary data case, this section extends the discussion to the case of Poisson distributed data. Once again, we make use of the function in (9.1). At each point t_j, a Poisson variate with mean $\exp(f(t_j))$ is generated. A total of 100 data sets are simulated. Similarly as in the normal and binary cases, we compare the Bayesian adaptive splines proposed here with the traditional Bayesian penalized splines without spatial adaptation, as well as the method of Krivobokova *et al.* (2008). The knot settings are kept the same as in the two preceding sections. Figure 9.5 illustrates the adaptive methods considered here on a single simulated data set with Poisson data for $T = 1000$. Also shown are the variance functions for the random effects for smoothing. The random effects variances suggest more curvature at the beginning, reducing with the increasing values of t. Note the difference in the range of the random effects variances between both approaches is due to differences in formulation of the basis functions.

9.4.2 Simulation Results

In this section, results obtained from fitting the model proposed in this chapter, in comparison to related methods in the literature, are summarized. In particular, the mean square error values, split into squared bias and variance, for the different models under the different response distributional assumptions are calculated. Results are summarized in Table 9.3 for the normal, binary and the Poisson cases. A graphical presentation of all the fitted values from the simulation runs is shown in Figure 9.6, where, for each scenario, the results are summarized in the form of a boxplot. One can observe the similarity of the results between the BAPS and the KCK method especially for the normal and binomial cases.

First, a comparison between the BAPS model with the method of Krivobokova *et al.* (2008) shows very comparable results in all three cases. The BAPS however tends to have smaller mean square values, although for the normal and binary cases, the approach is associated with more bias while the approach of Krivobokova *et al.* (2008) shows more variability of the estimates. This phenomenon is however not evident in the Poisson case and as such, a more extensive simulation exercise may be necessary

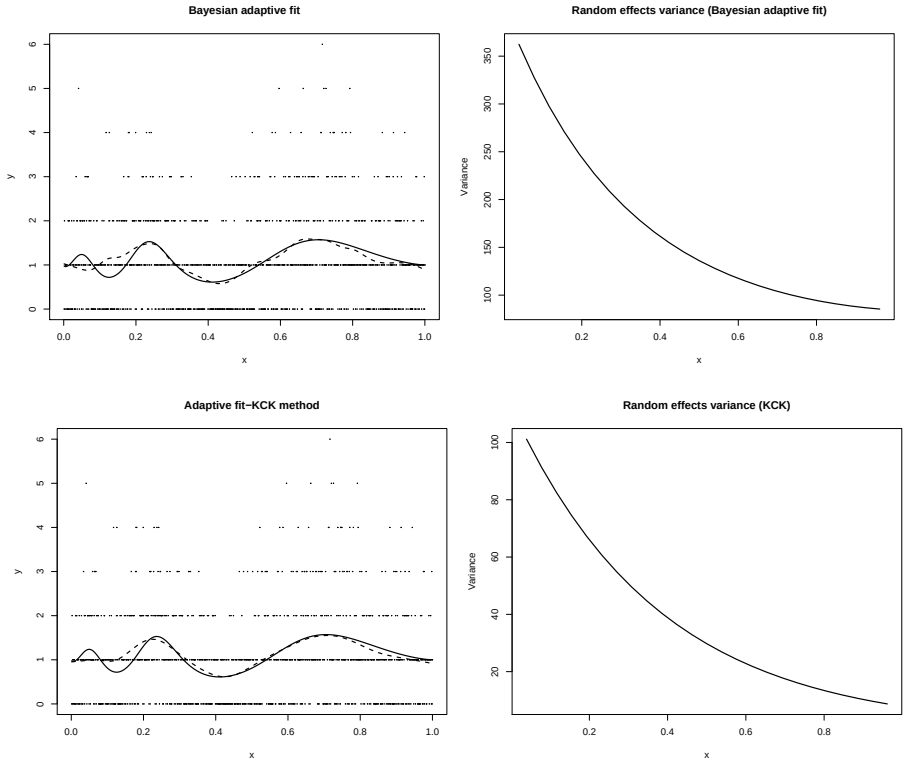


Figure 9.5: *Left: Illustrating adaptive fitting methods on Poisson count data on a single simulated data set. The continuous line shows the true generating function and the dotted line is the fitted model. The dots show the simulated data points. Right: Variances of random effects responsible for smoothing as a function of t for the two adaptive smoothing approaches.*

to ascertain or clarify the issue. The superiority of both adaptive methods compared to the non-adaptive version is clear from the mean square values in the normal and binary cases. A comparison between the BAPS and the non-adaptive version of the models can be obtained using the DIC. The results point to a better performance of the adaptive methods.

The results in Table 9.3 appear to suggest a better fit for the non-adaptive Bayesian method compared to the adaptive counter parts in the case of the Poisson distribution. In general, for non-normal data, a relatively large sample size is required to realize the full benefit of the adaptive methods. Indeed, as already seen in Section 9.4.1, small sample sizes can lead to poor estimation of the underlying function. Thus, in this case,

Table 9.3: *Comparison of the Bayesian P-splines, Bayesian Adaptive P-Splines and KCK methods based in simulated data. The three methods are evaluated based on MSE and the Bayesian models are compared based on the DIC.*

	Bayesian P-splines	Bayesian Adaptive P-splines	KCK
		Normal case	
bias^2	0.00051	0.00017	0.00006
var	0.00184	0.00077	0.00088
MSE	0.00235	0.00094	0.00095
DIC	-126.55	-146.43	
		Binary case	
bias^2	0.00140	0.00144	0.00140
var	0.00066	0.00032	0.00046
MSE	0.00206	0.00176	0.00186
DIC	1366.98	1359.19	
		Poisson case	
bias^2	0.00696	0.00887	0.01362
var	0.02567	0.02854	0.02385
MSE	0.03263	0.03741	0.03747
DIC	1085.36	1086.89	

it is difficult to clearly see the advantage of the adaptive methods. Krivobokova *et al.* (2008) also mention the need for large sample sizes to achieve optimal results with adaptive methods in the case of non-normal data. It is worthy mentioning that for purposes of simulations, very large sample sizes become computationally prohibitive with Bayesian methods involving complicated models.

9.5 Discussion

Adaptive penalized splines in a Bayesian framework were considered in this chapter. Such models are useful in nonparametric regression settings when data tend to show heterogeneous tendencies, i.e., in some regions, the mean changes rapidly yet remains rather smooth in other regions. Particular interest was on Bayesian adaptive penalized splines for non-normal data; the case for normal data having been studied in detail

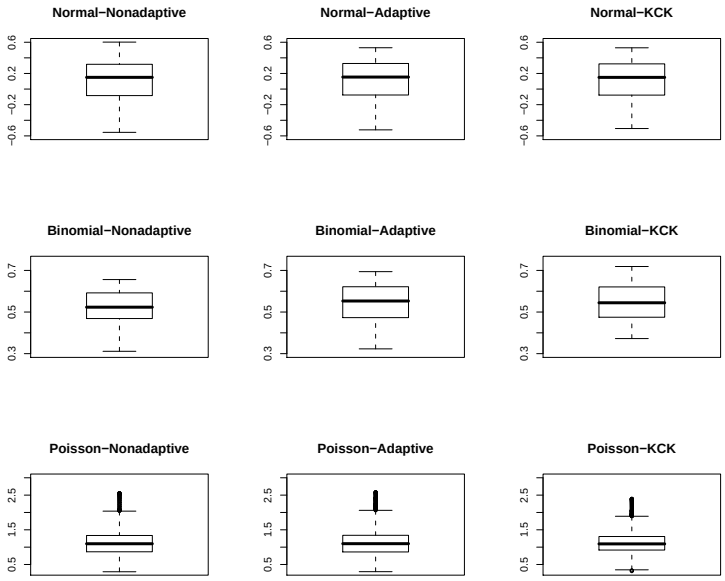


Figure 9.6: *For each of the methods and each distributional assumption, the fitted values across all simulation runs are summarized by boxplots.*

in literature. This chapter illustrates use of such models for binary and Poisson counts, in comparison to a related method based on a Laplace approximation of the likelihood. The two approaches give comparable results. It is important to mention that, while the Bayesian models considered here come in with some advantages, such as, accounting for variability in all parameters estimated in the model, the models can be computationally demanding. Obviously, the computational burden grows with sample size and complexity of the models. A major advantage with the formulation of the Bayesian models considered here is the extendability to more complex situations. For example, in an experiment with three experimental groups, smoothing can be done adaptively at different levels of smoothing for each of the experimental groups. It is envisaged that extendability of the model can also be taken into the direction of additive models as in Baladandayuthapani *et al.* (2005) or spatial smoothing as in Krivobokova *et al.* (2008). Further, the Bayesian models considered offer a readily available platform for model comparison in the form of the DIC. Other than using the mean square error, it may be difficult to compare two models using the approach of, for example, Krivobokova *et al.* (2008). In any case, a comparison based on mean

square error does not take into account model complexity. Useful tools for inference, easily obtainable from Bayesian methodology machinery, are the credible intervals. Such intervals can be used to compare or assess moment-by-moment differences in experimental groups. Note that, although one can still calculate such intervals, they are not readily available in the software package developed by Krivobokova *et al.* (2008).

A possible topic of further research would be to consider model selection at the level of smoothed variance components. As we have seen, one can employ the same or different basis function to model the mean and the penalty parameters. A closer look at how different basis functions for the penalty parameters would influence the results may be interesting.

Several aspects are of interest with electrophysiological experiments. For example, interest may lie in determining the maximum peak activity when the monkey correctly decided that the test stimuli were to the right and to the left of the reference line. Further, one may also be interested in comparing neuronal response between behaviorally relevant condition. Following the discussion of Chapter 7, such objectives can be tackled using the methodology of the current chapter.

10

On the Use of Historical Control Data in Pre-clinical Safety Studies

Typical pre-clinical safety experiments involve a study of a control group of untreated animals, and groups of animals exposed to increasing doses. The ultimate aim is to test for a dose related trend in the response of interest (e.g., a tumor of a certain type). Usually one would focus on one particular experiment. However, since such experiments are conducted in genetically homogeneous animal strains, historical control data from previous similar experiments can be helpful in interpreting results of a current study (Ibrahim and Ryan, 1996). For example, if a defect has never occurred in control animals, its occurrence even in only a few exposed animals may be a cause of concern although statistical tests may not yield a significant treatment effect in that study (Ryan, 1993). The issue of when one may use historical control data is still not clear. It is argued that one can gain efficiency from historical controls when the dose effect is not clear-cut (Dempster *et al.*, 1983; Ibrahim *et al.*, 1988) and when the control rates are low (Ibrahim *et al.*, 1988). Historical control data are also used to demonstrate that some tumors are species-specific and thus not compound related. Should one be in a position to use historical control data, it is also of interest to know

whether one should use all the available studies or possibly select only a subset of studies.
Our intention is to investigate these aspects and possibly provide general recommendations regarding the use of historical control data. We will focus on the logistic-normal model (Dempster *et al.*, 1983; Parise *et al.*, 2001) while monitoring the precision, bias and power (based on a likelihood ratio test) associated with estimation of treatment effect. Computer simulations are used for this purpose.

In Section 10.1, a brief description of the logistic-normal model is given and an application of the model is illustrated on a data example in Section 10.2. Section 10.3 focuses on the simulation study and in Section 10.4 we investigate the idea of using a selected subset of historical control studies. The contents of this chapter are based on the paper of Maringwa *et al.* (2007).

10.1 Incorporating Historical Controls Using a Logistic-Normal Model

Consider an experiment with dose groups $d_0 < d_1 < \cdots < d_k$ where $d_0 = 0$ is the control group and k is the number of other dose levels. Suppose we are interested in determining presence or absence of a certain feature in the different dose groups. Hence we aim to investigate whether the proportion of animals having the abnormality of interest increases with increasing dose. A common approach to model the proportion of animals developing the abnormality is by use of a logistic regression model. Let p_i denote the response probability of an animal in group i. A simple logistic model to describe the relationship between p_i and d_i can be written as

$$\text{logit}(p_i) = \beta_0 + \beta_1 d_i,$$

where β_0 and β_1 are parameters to be estimated, representing the background effect and the dose effect, respectively.

Suppose we intend to include historical control information in the analysis. If it can be assumed that the current control group as well as the historical controls are a random sample from the same population, one can pool the control rates together. However, to account properly for study to study variability, one can incorporate the historical control studies as random effects in the model. This provides a compromise between completely ignoring the historical controls and pooling together all the historical controls. Let n_{ij} denote the number of animals in group i of study j and y_{ij} denote the number of animals developing the abnormality in group i of study j for

$j = 0, 1, 2, \ldots, n_s$. The current experiment will be indexed by $j = 0$ and the historical controls are indexed $j = 1, 2, \ldots, n_s$ where n_s is the total number of historical studies included. Note that for historical studies, only the control group ($d_0 = 0$) is considered. Let b_{0_j} denote the random study component which is assumed to be normally distributed with mean 0 and variance $\sigma^2_{b_0}$. Conditional on the random intercept b_{0_j} we have:

$$y_{ij}|b_{0_j} \sim \text{Binomial}(n_{ij}, p_{ij}) \quad \text{and} \quad \text{logit}(p_{ij}) = \beta_0 + \beta_1 d_i + b_{0_j}, \tag{10.1}$$

for $i = 0, \ldots, k$ and $j = 0, \ldots, n_s$.

Parameters in this model can be obtained through maximum likelihood estimation using for example PROC NLMIXED in SAS. To obtain starting values for this model, the model (for the current study only) is fitted using PROC LOGISTIC and estimates obtained thereof are used as starting values in PROC NLMIXED. For the variance of the random effects, several different starting values including default settings are used to come to stable estimates.

To assess the adequacy of the model under consideration, it is advised to check the plausibility of the model assumptions. As such, we need to check whether the normal assumption assigned on the random effects is appropriate. Since interpretation of histograms and scatterplots of unstandardized Empirical Bayes (EB) estimates of the study-specific random intercepts is questionable (see Verbeke and Molenberghs, 2000), we follow Degruttola, Lange and Dafni (1991) in first standardizing the EB estimates and then constructing normal quantile plots to assess the normality assumption for the random effects. For the general fit of the model, we plot the fitted probabilities against the observed proportions at the different dose levels.

Our main interest is to determine the gains if any from the incorporation of historical controls. To that effect, we propose to assess the precision, bias as well as the power associated with the estimation of β_1, the main parameter of interest. For example, if inclusion of historical controls substantially improves the precision, with relatively low bias in the estimate as compared to the situation where no historical control information is used then it may be worthwhile including such information.

10.2 Application to Incidences of Alopecia Data Example

As an example, we consider data involving investigation of the occurrence of the parameter 'alopecia' in rabbits described in Section 2.4. These data involve historical

data from two different species. The occurrence of alopecia in these two species of animals is not expected to differ much and can therefore both be used as historical controls for the current or examined study. This is confirmed by a Wilcoxon Mann Whitney test applied to the proportions in two groups (p-value=0.2990). We will however include a variable in the model to account for possible species effects hence we have

$$y_{ij}|b_{0_j} \sim \text{Binomial}(n_{ij}, p_{ij}) \quad \text{and} \quad \text{logit}(p_{ij}) = \beta_0 + \beta_1 d_i + \beta_S S_j + b_{0_j}$$

where S_j identifies species and β_S is the corresponding coefficient.

Two logistic regression models, one ignoring the historical control studies and the other including them as random effects are fitted. The likelihood ratio test is used to test for species and treatment effects. The results are displayed in Table 10.1.

While the current study indicates a borderline treatment effect (p-value=0.0418), inclusion of historical control studies under the logistic-normal model appears to further weaken the evidence for treatment effect (p-value=0.0886). It can also be observed that the effect of species is non-significant (p-value=0.2207).

Note that parameter estimates and their standard errors have changed upon inclusion of historical controls. While the precision for estimation of treatment effect appears to have improved, the issue of bias needs to be investigated by use of simulations.

Figure 10.1 shows some diagnostic plots for the logistic-normal model considered. The top figures show the observed and fitted probabilities based on the current study only (left) and based on the logistic-normal model including historical control data (right). The model appears to fit well in the current study (Figure 10.1, top left). The effect of the random intercept for study is apparent in Figure 10.1 (top right) where the fitted values appear slightly shifted up, with the fitted proportion in dose 0 appearing to have been 'pulled' towards the relatively higher proportions in the historical studies.

The normal quantile plot for the standardized EB seems to suggest deviation from normality. However, this can possibly be attributed to one particular study identified as outlying (see Section 10.4.2). Indeed, removal of this study suggests no deviation from normality (see Figure 10.1, bottom right). The Shapiro-Wilk test (Shapiro and Wilk, 1965) fails to reject normality (p-value=0.1524) when the study is removed while it rejects the normality assumption (p-value=0.0045) when all historical studies are used. Results obtained from fitting models with and without this particular study are discussed in Section 10.4.2.

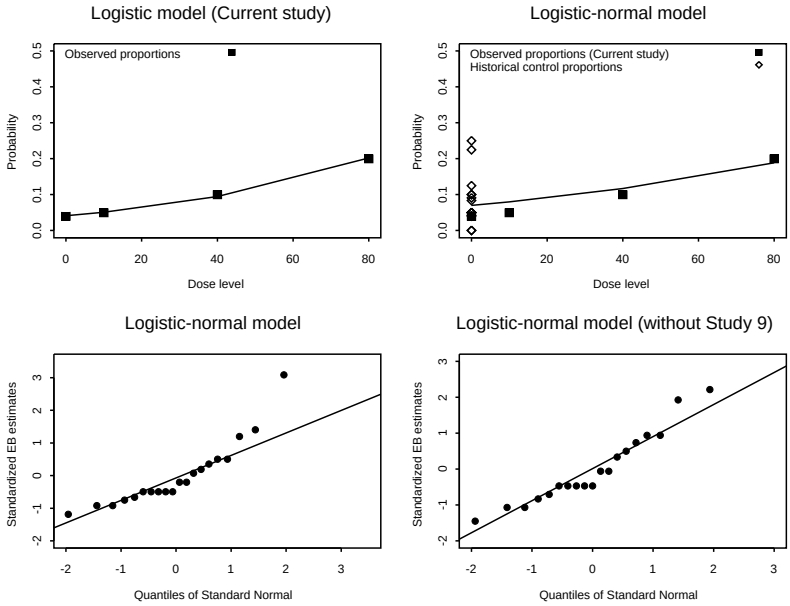


Figure 10.1: *Diagnostic plots: Top left panel shows the observed and fitted proportions (continuous line) in the current study while the top right panel shows the fitted proportions in the logistic normal model (i,e including historical studies). Proportions of Alopecia in historical studies are indicated by the diamonds and the bottom panel shows the normal quantile plot for standardized EB estimates for the logistic normal model.*

10.3 Simulation Study

To investigate the aspects discussed in Section 10.3.1, the study on occurrence of alopecia described in Section 10.2 is used to obtain realistic parameter settings for the simulations.

10.3.1 Monitoring Precision, Bias and Power

We are mainly interested in assessing how the precision associated with estimating treatment effect, power and bias are affected by use of historical control studies relative to considering only the current study. To that effect, some terminology is introduced. Using simulated data, we obtain the ratio of the standard error for the parameter of interest obtained by fitting the logistic-normal model to the combined data (current

study plus historical control studies) to the standard error obtained using only the examined study, which we define as relative efficiency (RE). In a similar way we calculate relative squared bias (RB) and relative power (RP). So, in summary, the lower the RE and RB (smaller than 1), and the higher the RP (larger than 1), the more beneficial the use of historical controls.

10.3.2 Simulation Settings

Using parameter estimates given in Table 10.1 as guidelines, we consider incidences of alopecia in the control group of approximately 2% ($\beta_0 = -4.0$), 12% ($\beta_0 = -2.0$) and 20% ($\beta_0 = -1.4$) representing low, medium and high background rates. We

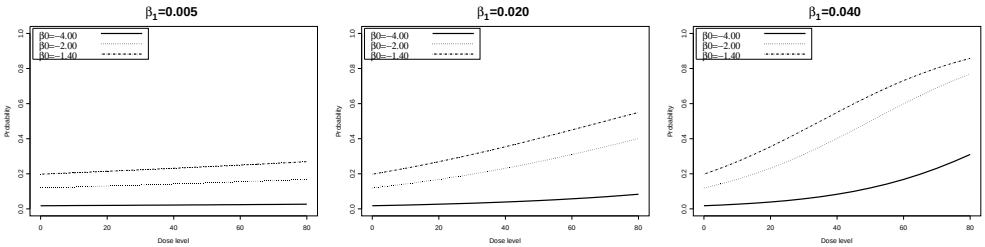


Figure 10.2: *Hypothesized models for simulations: probability of occurrence of alopecia as a function of dose for specified β_0 and β_1.*

assign values for β_1 equal to 0.005, 0.020 and 0.040 reflecting a low, borderline and high treatment effect respectively. Figure 10.2 graphically illustrates the hypothetical situations under consideration. To simulate historical control data, we use the logistic-normal model (10.1) with $d_i = 0$ and the control rates fixed as above. For variance of the random effects, we allow $\sigma^2_{b_0}$ taking values of 0.10, 0.30 and 0.90. The normal variate simulated for each study accounts for the study to study variation in the assumed background rate of alopecia. A fixed sample size of 30 animals in each group, both in the 'examined study' as well as historical control studies is assumed. For any fixed setting (see Figure 10.2), 200 data sets are randomly generated and the logistic regression model is fitted to the 'examined' study while the logistic-normal model is fitted to the combined data. The parameter values used to generate the data are used as starting values in fitting the models.

10.3.3 Simulation Results

For each specified level of treatment effect we investigate how the relative efficiency, bias and power change with increasing number of historical control studies, different background rates and variability amongst the historical studies (denoted by V).

Low Treatment Effect

Let us first assume a low treatment effect, $\beta_1 = 0.005$ (see Figure 10.2). From the left column of Figure 10.3, it can be observed that inclusion of historical control studies tends to improve precision since all RE values are below 1.0. The effect of variability amongst the studies on precision is evident from Figure 10.3 and is as expected. The more homogeneous the studies are, the more precise the estimation for treatment effect. However, for fixed variability amongst historical control studies the precision tends to decrease with increasing control rates.

It can also be observed that really large numbers of historical control studies do not necessarily lead to increased precision since the RE tends to level with increasing number of historical studies. Inclusion of historical studies also leads to bias reduction when the control rate is low (Figure 10.3, right column). For higher control rates however there is a price to pay for higher precision in terms of some bias.

From Figure 10.6 we further observe, particularly for a fairly homogeneous set of studies and a low background rate, an increase in power by about 50% as more historical control studies are incorporated. When 10 or more studies are taken into account, the power fluctuates around that same value and does not increase any further. For increasing background rates, this power advantage is less pronounced.

Borderline Treatment Effect

Let us now consider a borderline treatment effect where $\beta_1 = 0.02$. As in the previous section, variability amongst historical control studies plays an important role in the precision of estimating treatment effect (see Figure 10.4). It can however be observed that for a fixed β_0 and V, almost the same precision is gained as for the case of low treatment effect. Notice however the trade off between precision and bias in this case. As the precision improves with increasing number of historical studies, bias increases especially when the control rate is low. This increase in bias is however smaller for larger control rates. It can also be observed from Figure 10.6 that the incorporation of historical control studies in the model leads to a power advantage. Again we notice the leveling of the relative power which implies that continuously

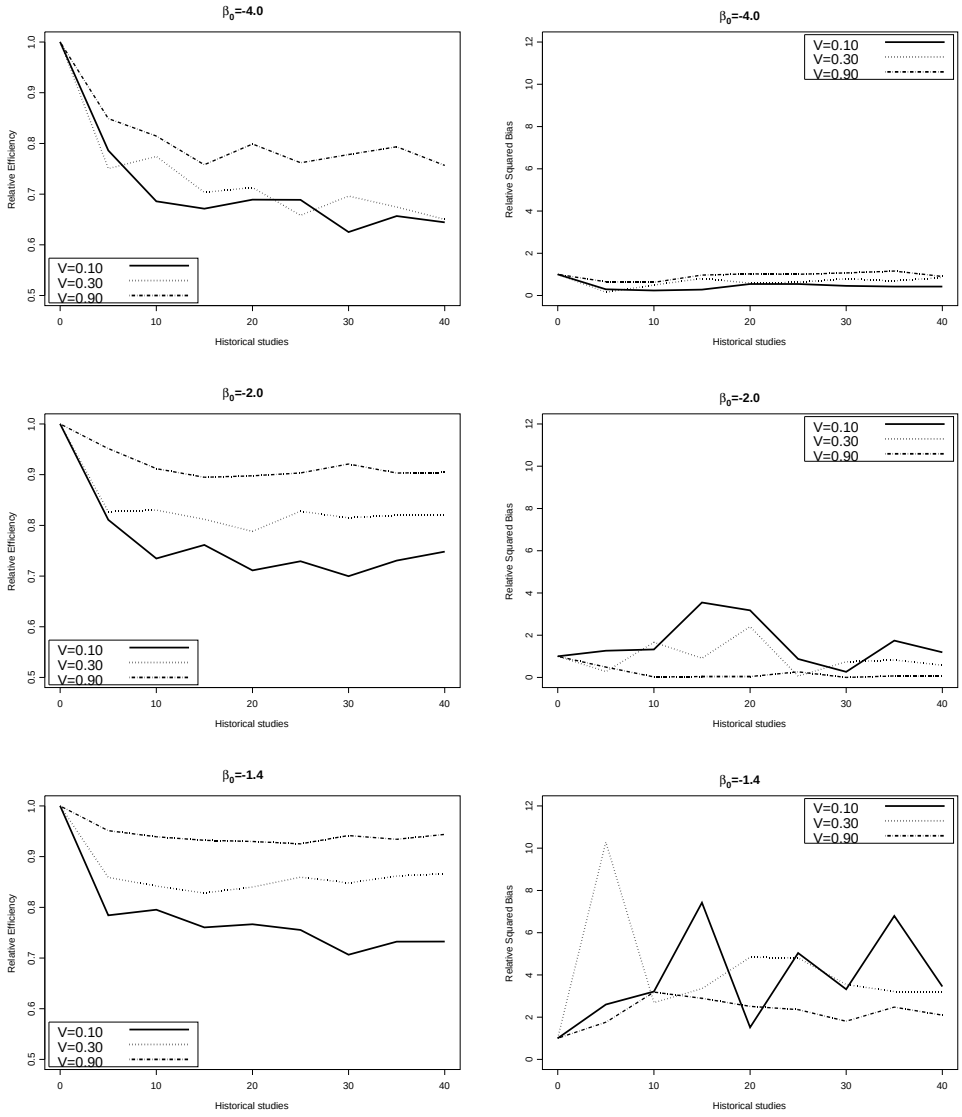


Figure 10.3: *Low treatment effect: Relative efficiency (left column) and Relative squared bias (right column).*

increasing the number of historical studies does little to further improve the power. We also notice that the increase in power decreases as the control rate increases and as the heterogeneity amongst the historical control studies increases.

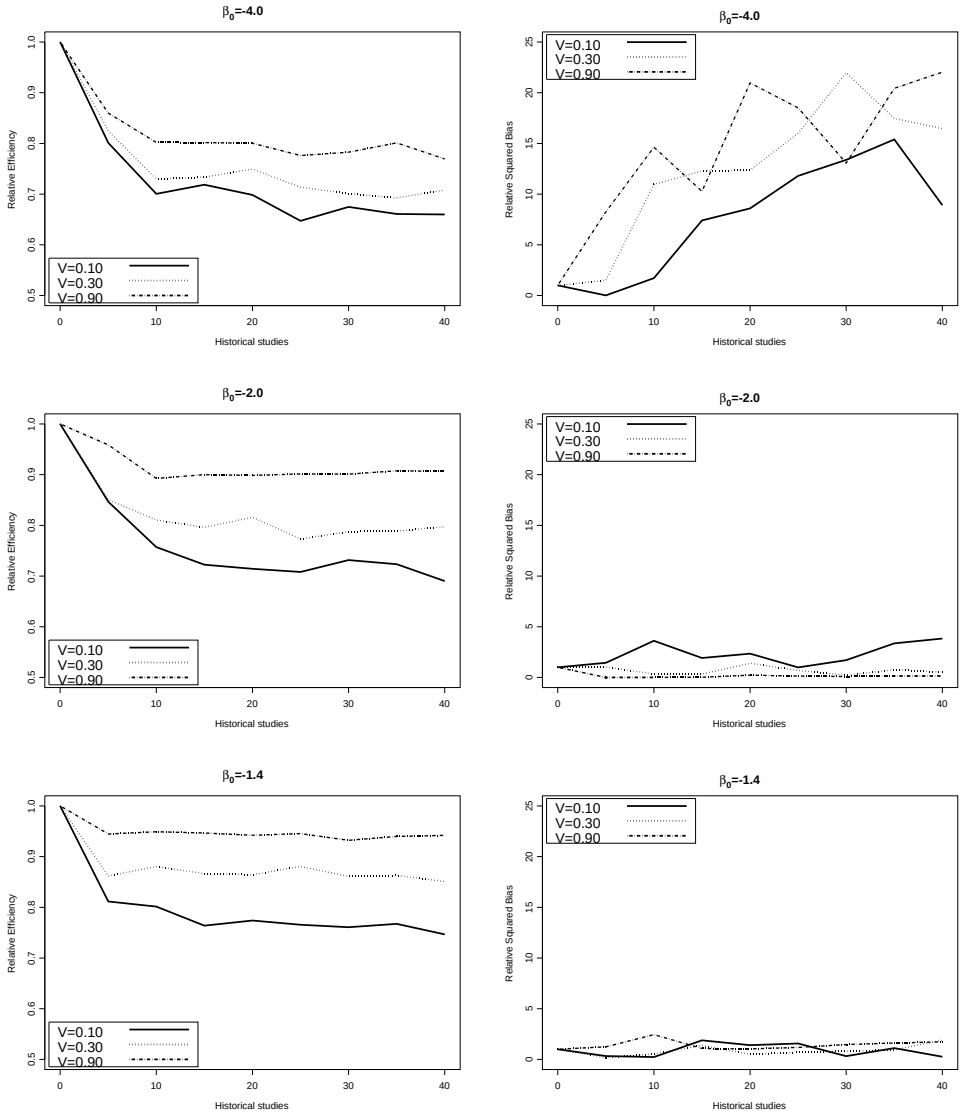


Figure 10.4: *Borderline treatment effect: Relative efficiency (left column) and Relative squared bias (right column).*

Strong Treatment Effect

Let us finally assume a strong treatment effect represented by $\beta_1 = 0.040$. The background rates are kept as before. As for the other two cases above, we vary

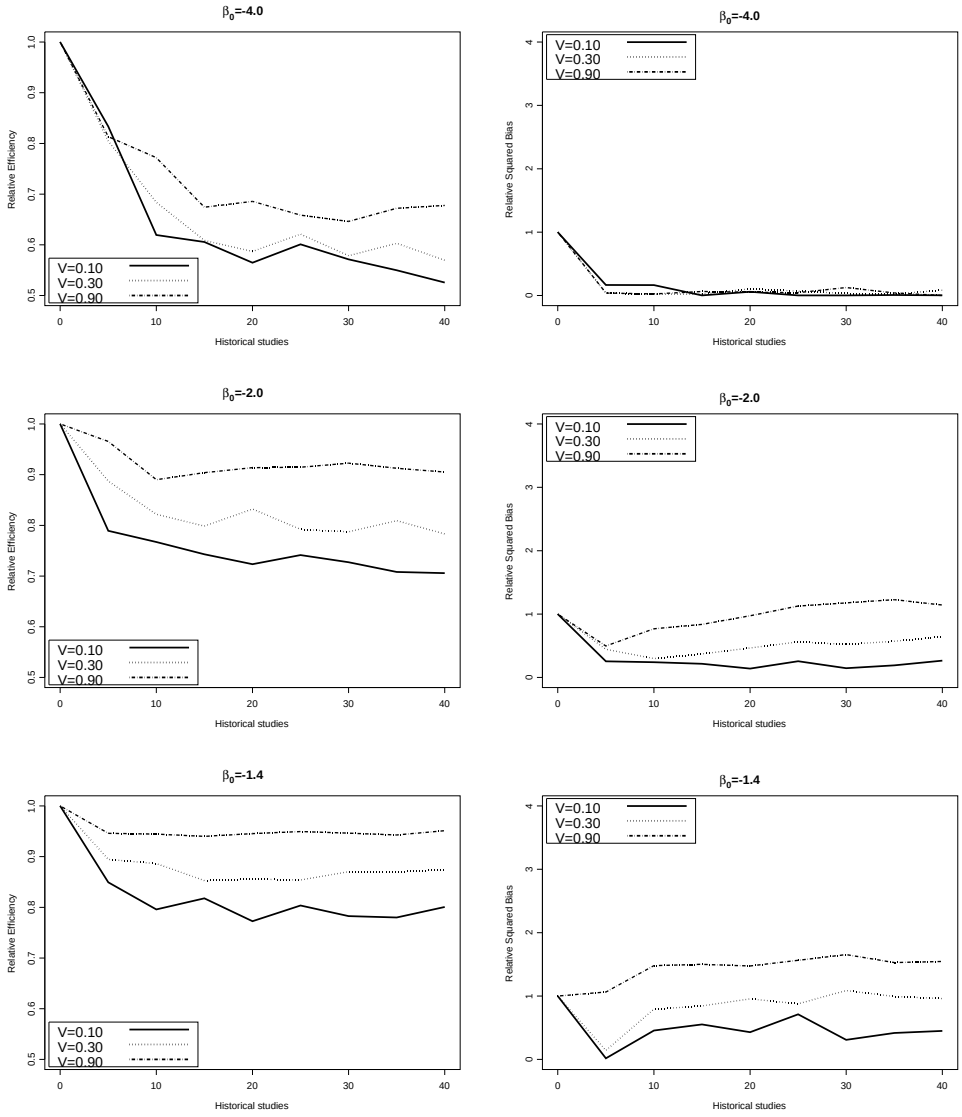


Figure 10.5: *Strong treatment effect: Relative efficiency (left column) and Relative squared bias (right column).*

variability amongst the historical studies and increase the number of studies. The results for precision and bias are summarized in Figure 10.5. The leveling of the precision as the number of historical control studies increases can also be observed in this case. An increase in precision is again associated with a decrease in bias. Figure

10.6 shows that there is no gain in power when treatment effects are strong. This is not surprising since when treatment effect is strong, it will be difficult not to detect it even without additional (historical) data since the examined study already has high power. However, for studies with low or borderline effect, the gain can be large.

10.3.4 Summarizing Gains from Incorporation of Historical Control Studies

Low variability amongst the historical controls is a key factor if we intend to maximize the gains from historical controls. Focusing on the lowest value of the variability

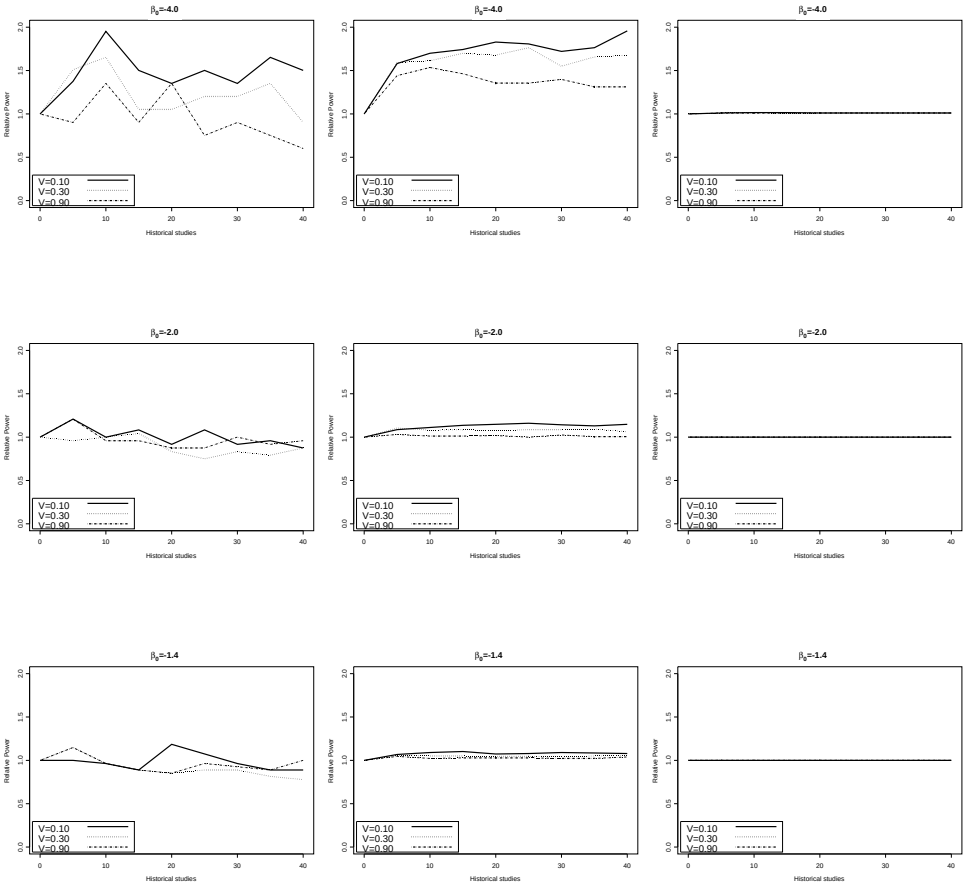


Figure 10.6: *Relative power: Low treatment effect (left column), Borderline effect (center column) and Strong treatment effect (right column).*

amongst studies considered, we can observe that including historical studies in general may lead to reduced standard errors for estimating treatment effect.

When treatment effect is low and control rate is low, other than gaining precision, there appears to be a substantial reduction in bias as well as gain in power. However, as the control rate increases, power decreases and bias increases also.

While we observe both an increase in precision and reduction in bias, we also observe that including historical data when treatment effect is high barely improves power (see Figure 10.5 and Figure 10.6).

In summary, one tends to gain more from historical data when control rates are low and when treatment effect is borderline although in the latter our simulations show that moderate control rates would be associated with less bias compared to low control rates. It is clear that continuously increasing the number of historical control studies will not offer any further improvement in terms of precision, bias or power.

One possible conclusion from the foregoing discussion is that one may use a relatively low number of historical control studies (e.g., 15 studies) provided that they are considered a homogeneous set of studies. This brings us to the issue of making a selection from the set of the historical control studies available. This will be discussed further in the next section.

10.4 Selection of a Subset of Historical Control Studies

While it may be ideal to have a large number of historical control studies, a leveling of profiles for precision, bias and power with increasing number of historical control studies could be observed. This implies we might possibly work with relatively fewer historical control studies. Therefore, the effect of using a selected subset of historical control studies is investigated .

10.4.1 Criterion for Selection

Selecting a subset of historical control studies may be performed in a number of ways. We propose use of the estimates for the study-specific random effects thereby remaining within the logistic-normal model framework. Therefore we fit the logistic-normal model in (10.1) to the historical control studies only, i.e, $d_i = 0$. Estimates for the study-specific random effects, which are termed Empirical Bayes (EB) estimates (Verbeke and Molenberghs, 2000) can be obtained from the model.

It is assumed that the study-specific random effects are in fact samples from a normal distribution with mean 0 and some variance, $\sigma^2_{b_0}$. As such, one can therefore use the estimated variance for the random effects, $\hat{\sigma}^2_{b_0}$, to construct selection bounds. One can make a selection of studies such that their Empirical Bayes estimates range between $-\delta\hat{\sigma}_{b_0}$ and $+\delta\hat{\sigma}_{b_0}$ for some $\delta > 0$.

10.4.2 Application to Incidences of Alopecia Data Example

The left panel of Figure 10.7 shows the distribution of the proportions of animals with alopecia in the 19 historical control studies. It can be observed that the distribution is skewed to the right with a possibility of outlying studies. The distribution of the estimates for the study specific random effects is shown in the right panel of Figure 10.7. Since the estimates are expected to be centered around 0, values greater than

Table 10.1: *Fitting the logistic regression model to the examined study and the logistic-normal model to combined data (i.e., examined study and all historical control studies). The parameter β_S represents the species effect.*

				$H_0 : \beta = 0$
Model for	Parameter	Estimate	$\widehat{se}$	p-value
Examined study	β_0	-3.1428	0.6903	
	β_1	0.0221	0.0113	0.0418
Examined study and all	β_0	-2.9270	0.3787	
historical studies	β_1	0.0141	0.0086	0.0886
	β_S	0.5202	0.4354	0.2207
	σ_{b_0}	0.3594	0.2638	
Examined study and selected historical	β_0	-2.8717	0.3426	
studies with EB estimate	β_1	0.0146	0.0069	0.0483
within $(-\hat{\sigma}_{b_0}, \hat{\sigma}_{b_0})$	β_S	0.2584	0.4171	0.5271
	σ_{b_0}	2.203E-7	0.1415	

0.4 (see right panel of Figure 10.7) could possibly represent outlying studies. In Table 10.1 a set of models fitted under different scenarios is presented. First the logistic regression model is fitted to the examined study (Table 2.3) and then the logistic-normal model fitted to the examined study combined with historical control studies data (Table 2.4). Finally the logistic-normal model is fitted to the examined study combined with only a subset of historical control studies whose random effects

estimates fall within the specified limits ($\delta = 1$). The estimate for the standard

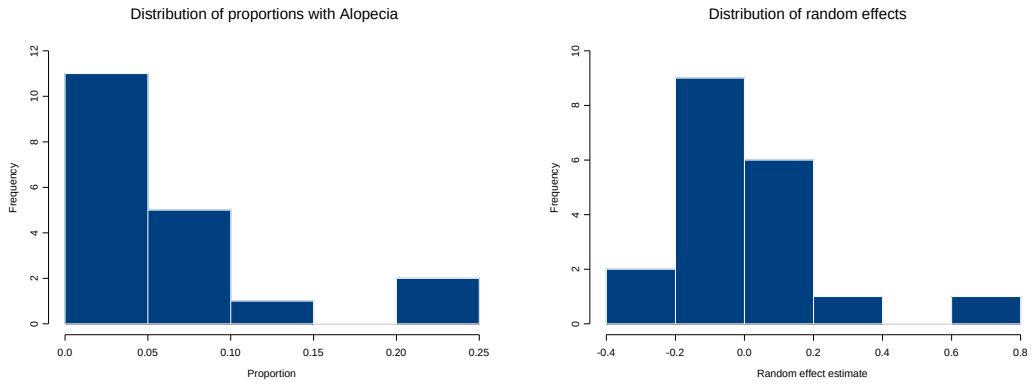


Figure 10.7: *Distribution of proportions (left) and distribution of study-specific random effects (right)*

deviation of the random effects is found to be $\hat{\sigma}_{b_0} = 0.3594$. Only one study with a proportion of adverse events (in the control group) of 9/40 was found to have its EB estimate absolute value larger than $\hat{\sigma}_{b_0}$. The results corresponding to the selection of studies in Table 10.1 are therefore obtained with 18 historical control studies.

The effect of including or excluding this particular 'outlying' study is outlined in Table 10.1. Analysis of the examined study only suggests presence of treatment effect although borderline (p-value=0.0418). Inclusion of selected historical control studies provides a similar conclusion (p-value=0.0483) with more precision in the estimation of treatment effect. However, including all the available historical studies tends to dilute the treatment effect (p-value=0.0886). Inclusion of the 'outlying' study potentially elevates the overall control mean, deviating further from the control mean in the current study (see also Figure 10.4 top right panel). The effect of omitting the outlying study can also be seen from the change of the estimate for the random effects. Omitting one particular study leads to the variance for random effects to be very small. This model is still to be preferred since even after removal of the outlying study, the heterogeneity amongst the remaining studies is still accounted for.

10.4.3 Simulation Study

We have seen from the previous section that excluding certain studies deemed out of range can influence conclusions. A simulation study is set up to further investigate this phenomenon. Data for the 'examined' study as well as for historical control

studies are generated using the estimates obtained from the examined study (Table 10.1) as the actual parameters. Values for the variance amongst the historical control studies were kept as in Section 10.3.

Assuming that the number of animals in the historical control studies follows a poisson distribution with mean λ, this model is fitted to the data in Table 2.4 and the maximum likelihood estimate for the mean ($\hat{\lambda} = 23$) is used to generate data. Keeping as close as possible to the data example, 20 historical control studies are generated. As in Section 10.3 we assume a fixed sample size of 30 animals in each dose group in the 'examined' study. For each of the 200 simulation runs performed, the models corresponding to the three situations as described in Table 10.1 are considered.

We assess how bias, precision and the mean squared error (MSE) can be affected by the selection procedure. For each simulation setting we also present the median p-value since the distribution of p-values is highly skewed. The results are summarized in Table 10.2. Note that when considering the selection of a subset of historical studies in the simulations ($\delta = 1$), some or none of the studies will be dropped and therefore we indicate the average number (n_s^{mean}) and the minimum (n_s^{min}) of studies that is used. As expected, the more heterogeneous studies are, the more the number

Table 10.2: *Selecting historical studies: Model 1-Examined study only ($\beta_1 = 0.0221$), Model 2-Examined study plus all historical studies and Model 3-Examined study plus a selected subset of studies with EB estimates within ($-\hat{\sigma}_{b_0}, \hat{\sigma}_{b_0}$).*

					Median		
V	Model	Bias	s.e(β_1)	MSE	p-value	n_s^{min}	n_s^{mean}
	1	0.00089	0.01231	0.00015	0.0494	0	0
0.1	2	-0.00066	0.00809	0.00007	0.0111	20	20
	3	0.00005	0.00798	0.00006	0.0087	16	19.560
0.3	2	-0.00121	0.00848	0.00007	0.0204	20	20
	3	-0.00023	0.00802	0.00006	0.0131	16	19.075
0.9	2	-0.00052	0.00953	0.00009	0.0339	20	20
	3	0.00011	0.00866	0.00007	0.0116	15	18.105

of studies left out as 'outlying'. The results also show that for any fixed variability, making a selection of studies performs better than using all studies available in terms of precision and bias. Overall, the mean square error tends to decrease with the selection of studies. Moreover making a selection of studies tends to provide more evidence for the apparently borderline treatment effect in the examined study (median

p-value=0.0494). A similar analysis based on historical control studies with EB within two standard deviations of the mean has also been carried out. The average numbers of studies used are 19.995, 19.985 and 19.905 for $V = 0.10,\ 0.30$ and 0.90, respectively. Essentially almost all the available studies are used and the result is close to using all the historical control studies.

The treatment effect in the examined study could already be considered as borderline significant and the effect of historical studies may not be that pronounced. Let us now show an example in which the use of historical control studies can really alter the conclusion. A relatively lower treatment effect is obtained by fixing the value of β_1 to 0.0175 and keeping β_0 as in Table 10.1. The same models as defined in Table 10.2 are fitted and the results obtained are summarized in Table 10.3. The results show a non-

Table 10.3: *Selecting studies: Illustrating how use of historical control studies can alter conclusion. Model 1-Examined study only ($\beta_1 = 0.0175$), Model 2-Examined study plus all historical studies and Model 3- Examined study plus a selected subset of studies with EB estimates within $(-\hat{\sigma}_{b_0}, \hat{\sigma}_{b_0})$.*

					Median		
V	Model	Bias	s.e(β_1)	MSE	p-value	n_s^{min}	n_s^{mean}
	1	-0.00246	0.01459	0.00022	0.08645	0	0
0.1	2	-0.00364	0.00961	0.00011	0.03100	20	20
	3	-0.00300	0.00963	0.00010	0.02541	17	19.480
0.3	2	-0.00347	0.01087	0.00013	0.04594	20	20
	3	-0.00236	0.01099	0.00013	0.02891	16	19.145
0.9	2	-0.00393	0.01222	0.00017	0.07302	20	20
	3	-0.00236	0.01178	0.00014	0.02787	15	18.195

significant treatment effect in the examined study (median p-value=0.0864) which is subsequently turned to a significant effect by use of historical studies except when using all historical studies with large variability amongst them ($V = 0.90$). For a fixed variability amongst historical control studies, more evidence for treatment effect arises from using a selection of historical control studies. We observe in both examples (i.e $\beta_1 = 0.0221$ and $\beta_1 = 0.0175$) that the mean squared error associated with a selected subset of studies is less than when considering all studies. Therefore, although we reduce the number of studies, it is evident that no serious bias is anticipated. In terms of decision making, selection of historical studies becomes more influential on the result for relatively low treatment effect and as the variability amongst historical

control studies increases. For example, in Table 10.3, the difference in the median p-values associated with a selection of studies and all available studies increases as variability amongst studies increases. For heterogeneous set of studies, two different conclusions can arise from the two approaches (see Table 10.3).

In a similar way as above, we also perform some simulations under the null hypothesis of no treatment effect ($\beta_1 = 0$) to assess how making a selection of historical studies may affect the Type 1 error. We therefore perform a simulation exercise for the three scenarios specified in Table 10.1, i.e using only the current study (Model 1), using all available historical studies (Model 2) and finally using a selected subsample of historical studies (Model 3). The results obtained using 200 simulated data sets for each scenario are summarized in Table 10.4. As in the previous two scenarios

Table 10.4: *Assessing the how use of a selected subset of historical control studies may influence Type 1 error. Model 1-Examined study only ($\beta_1 = 0$), Model 2-Examined study plus all historical studies and Model 3- Examined study plus a selected subset of studies with EB estimates within $(-\hat{\sigma}_{b_0}, \hat{\sigma}_{b_0})$.*

					Median			
V	Model	Bias	s.e($\hat{\beta}_1$)	MSE	p-value	n_s^{min}	n_s^{mean}	Type 1 error
	1	-0.00696	0.02614	0.00073	0.49810	0	0	0.07143
0.1	2	-0.00781	0.01894	0.00042	0.52986	20	20	0.06316
	3	-0.00674	0.01833	0.00038	0.52178	16	19.515	0.05789
0.3	2	-0.00857	0.02117	0.00052	0.48588	20	20	0.07895
	3	-0.00739	0.02006	0.00046	0.47937	16	19.195	0.06842
0.9	2	-0.00875	0.02303	0.00060	0.50815	20	20	0.07368
	3	-0.00700	0.02097	0.00049	0.51610	15	18.175	0.07895

discussed above, no serious bias is induced by reduction in the number of historical studies. The mean square error also tends to decrease when using a selection of more homogenous set of studies.

The results in Table 10.4 suggest that when variability is low, the Type 1 error does not deviate much from its nominal value of 0.05. When sampling studies with larger variability, an inflation in Type 1 error is noted. Our simulations indicate that the difference in Type 1 error when using all historical control studies compared to a selected subset of studies is relatively small.

10.5 Discussion

We have considered homogeneity amongst and number of historical control studies, differing background effects as well as varying treatment effect levels on the estimation and testing for treatment effect. As would be expected, variability amongst historical studies plays a pivotal role. Our results seem to concur with the notion that historical studies are more useful when treatment effect is borderline and when background rates are low. In general, precision for estimating treatment effect can be improved by using historical studies. The issue of how many and which studies to include in the analysis is a subject of debate. The number of studies to be used often depends on availability. Our simulations show that only a modest number of studies (e.g., 15) is sufficient since not much gain is realized by continuously increasing the number of historical studies. Note that we did not focus of the effect of sample size within historical control studies or within the 'examined' study. We assumed a sufficiently large number of 30 animals per group. However, when historical control studies are small one would probably need to select more than the suggested number of 15 studies.

Part of the judgement of which studies to include should obviously be decided by subject matter experts. The other part can be played by the statistician. Indeed, as shown from the use of study-specific random effects, studies considered to be outlying can be eliminated. As the data example shows, the conclusion one is bound to make can be altered. Our simulation results appear to corroborate this idea. Although a selection of studies results in reduction in number, we have seen through simulations that the use of large numbers of studies is not likely to be of any substantial benefit compared to a modest number of well selected homogeneous set of studies.

In conclusion we would recommend the use of historical studies when treatment effect is low or borderline and when the control rate is low. When using the logistic-normal model, it may be preferred to make a selection of a subset of historical studies based on the Empirical Bayes estimates. Our conclusions are based on the normality assumption for the random effects. In principal, other distributions can also be assumed and further investigation will be required. Other approaches for example the Empirical Bayes method of Tarone (1982) based on the Beta-Binomial distribution assumption for the historical controls can also be used.

While this paper focuses on several simulation settings to study the change in bias and precision when using the historical control data, it would be interesting to also analytically derive the differences among the two approaches. This is a challenging research question, since we are not in the (simple) continuous normally distributed setting. To this end, we should analytically derive the bias and precision of the dose-

parameter in case of a logistic regression (Cordeiro and McCullagh, 1991) and in case of a random-effects logistic regression (Breslow and Lin, 1995). This is a topic of further research.

One of the assumptions made is the absence of a trend in the occurrence of the parameter of interest in the control animals over the years. Should there be some trend, appropriate techniques incorporating such information would then be required.

11

Tolerance Intervals and Their Use in Pre-clinical Studies

Many practitioners have been exposed to some extent to confidence and prediction intervals for regression models. These, however, are only a few of the statistical intervals required in practice. Unfortunately, many important intervals, such as intervals for population percentiles and/or tolerance limits, are ignored in standard texts (Hahn and Meeker, 1991).

In pre-study method validation, tolerance limits are of utmost importance. Before an analytical procedure is used routinely on unknown samples, it is normal practice to perform a more or less extensive set of experiments to evaluate whether it will be able to meet the desired criteria. Those experiments are usually called 'pre-study validation' experiments. Since the bias and the precision of the intrinsic performance parameters are unknown, experiments are required so that the user can obtain estimates of these quantities before the method is used routinely. The objective of the pre-study validation phase is to evaluate whether, given the estimates of bias and standard deviation obtained, the proportion of measures of new unknown samples that will fall within the acceptance limits is greater than a predefined acceptance level.

11.1 Background

The intention in this chapter is to dwell upon the less known but often useful type of intervals, the tolerance intervals. Use of tolerance intervals dates back to the 1940s. Some of the earliest works emerge from Wilks (1941), Wald (1942, 1943) and Wald and Wolfowitz (1946), focusing mainly on simple random samples. More recent work in the same context include Hahn and Merker (1991), giving a detailed treat of parametric and distribution-free intervals. Amaratunga (1997) shows how to improve traditional ways of constructing tolerance intervals on simple random samples based on the normal distribution to use of distribution-free intervals.

Extensions to more complex designs have found their way into the literature. Mee and Owen (1983) derived one-sided (β, γ) content tolerance limits for balanced one-way ANOVA models, considering the case where the ratio of the between to within variance is known and estimated from the sample. Also in the context of balanced one-way ANOVA models, Mee (1984) discusses procedures for one- and two-sided β-expectation tolerance limits and extends the procedure of Mee and Owen (1983) to two-sided (β, γ) content tolerance intervals.

Some extensions to unbalanced data have also been considered, for example, Bhaumik and Kulkarni (1996) and Bagui *et al.* (1996). Beckman and Tietjen (1989) extend towards multi-way ANOVA models and construct two-sided approximate β-content tolerance limits for multiway balanced random-effects models. Liao and Iyer (2004) proposed a generalized two-sided tolerance interval for the normal distribution with several variance components.

Wolfinger (1998) addresses the same issue from a Bayesian standpoint, focusing on tolerance intervals for the so-called variance component models. Hoffmann and Kringle (2005) consider a procedure to construct two-sided (β, γ) tolerance intervals for general random effects models, in both balanced and unbalanced data scenarios.

In what follows, we intend to apply the approach of Wolfinger (1998) and that of Hoffman and Kringle (2005) on a data example. The two methods address the same issue from two very different perspectives. Next to that, a nonparametric method due to Hahn and Meeker (1991) is discussed. The nonparametric method adds a further dimension to the problem, and therefore provides an interesting comparison platform with the other two approaches. Thus the main focus here is to apply, in a comparative context, different methods of constructing tolerance intervals.

11.2 Approaches to Constructing Tolerance Intervals

In this section, a review of some particular methods of constructing tolerance limits is given. First, a brief introduction to the framework in which the methods are considered is given. The focus here is on the random ANOVA model, which can be written as

$$Y_{ij} = \beta_0 + b_{0_i} + \varepsilon_{ij}, \tag{11.1}$$

where Y_{ij} is the response of subject i $(i = 1, \ldots, n)$ for repeated measurement j $(j = 1, \ldots, m)$; β_0 is an overall mean; b_{0_i} and ε_{ij} represent subject-specific deviations from the overall mean and residual errors respectively. It is assumed that $b_{0_i} \sim N(0, \sigma^2_{b_0})$ and $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2_{\varepsilon})$. This particular model is used in the analysis of the data example in Section 11.3. A brief review of the methods of interest follows.

11.2.1 Wolfinger Approach

Wolfinger (1998) distinguishes between (β, γ) tolerance intervals, β-expectation tolerance intervals, and fixed-in-advance tolerance intervals. An interval (l, u) is termed a two-sided β-content, γ-confidence tolerance interval if, for some cumulative distribution function F,

$$P[F(u) - F(l) \geq \beta] = \gamma.$$

It can therefore be claimed that at least a proportion β of the population will lie within the interval (l, u) with a confidence coefficient of γ (Hoffman and Kringle, 2005). This type of interval is typically used in cases requiring long-run prediction about numerous observations from a process assumed to be in a state of statistical control (Wolfinger, 1998). The β-expectation type of intervals focus on prediction of one or a few observations from the process. The fixed-in-advance tolerance intervals start from predetermined limits, and will not be discussed further here. More details can be obtained in Wolfinger (1998).

Wolfinger (1998) describes ways for constructing all of the above mentioned types of tolerance intervals within the Bayesian framework, using a procedure termed Bayesian simulation. Here, we are interested mainly in the first two types of intervals, primarily because they relate more to the objectives of our data example.

Let us give a brief review of the approach by Wolfinger (1998) and refer to the article for a detailed account on the subject. Consider the model in (11.1), expressed

in the usual formulation of a linear mixed model already seen in previous chapters,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{11.2}$$

with the vectors retaining their obvious meanings. The Bayesian simulation methodology hinges on sampling from the posterior density of the mixed model parameters. Let $\boldsymbol{\theta}$ be a vector containing all the variance components in (11.2), which, based on (11.1) contains the elements $\sigma^2_{b_0}$ and σ^2_{ε}. The posterior density may then be factorized as

$$[\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta}|\boldsymbol{Y}] = [\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{Y}][\boldsymbol{\theta}|\boldsymbol{Y}], \tag{11.3}$$

which is a product of the conditional posterior density of $\boldsymbol{\beta}$ and $\boldsymbol{b}$, given the variance components and the marginal posterior density of the variance components. According to Wolfinger (1998), based on the factorization, one can obtain samples from the posterior using a rejection sampling algorithm. Based on the sampled values, one can then construct (β, γ) tolerance intervals as well as β-expectation tolerance intervals. A closer look at the construction of both types of intervals is taken up in the ensuing discussion.

(β, γ) Tolerance Intervals

Let us expound a bit on the construction of the (β, γ) tolerance intervals as discussed in Wolfinger (1998). As mentioned earlier, one can generate a sample from the posterior density in (11.3) and denote each of the sampled outcomes by $(\boldsymbol{\beta}^*, \boldsymbol{b}^*, \boldsymbol{\theta}^*)$, representing estimates for fixed effects, random effects parameters and variance components respectively. Note that following from (11.2), the idea is to make inference based on some normal distribution with mean $\boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{b}$ and variance $\boldsymbol{t}^T\boldsymbol{\theta}$ for certain fixed quantities $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{t}$. For the tolerance intervals under discussion, two quantiles for each of the sample values can be defined by

$$\begin{aligned} q^*_l &= \boldsymbol{x}^T\boldsymbol{\beta}^* + \boldsymbol{z}^T\boldsymbol{b}^* - q_{[(1+\beta)/2]}\sqrt{\boldsymbol{t}^T\boldsymbol{\theta}^*} \\ q^*_u &= \boldsymbol{x}^T\boldsymbol{\beta}^* + \boldsymbol{z}^T\boldsymbol{b}^* + q_{[(1+\beta)/2]}\sqrt{\boldsymbol{t}^T\boldsymbol{\theta}^*}, \end{aligned}$$

where q_π denotes the probit of a probability π. The next step involves forming a scatterplot of q^*_l against q^*_u, followed by construction of the reference line

$$q^*_l = -q^*_u + 2(\boldsymbol{x}^T\bar{\boldsymbol{\beta}^*} + \boldsymbol{z}^T\bar{\boldsymbol{b}^*}),$$

where $\bar{\boldsymbol{\beta}^*}$ and $\bar{\boldsymbol{b}^*}$ are the respective averages of the $\boldsymbol{\beta}^*$ and $\boldsymbol{b}^*$ values. Next, two lines, parallel to the axes and intersecting at the reference line are drawn. One then slides

the intersection point along the reference line until $100(1-\gamma)\%$ of the observations remain in the lower right portion of the graph. The coordinates of the obtained intersection point form the two-sided required (β, γ) tolerance limits. An illustration of this method follows in the application in Section 11.3.

β-expectation Tolerance Intervals

Construction of this type of intervals follows in a more straightforward manner by using the samples generated as discussed in the previous section. Specifically, simulations from the predictive distribution of some future observation y_F is required. Having already the posterior distribution $[\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta}|\boldsymbol{Y}]$, simulations from the predictive distribution $[y_F|\boldsymbol{Y}]$ of some future observation y_F, may be done by generating observations from $[y_F|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{Y}]$. From (11.1) it follows that the distribution for a future observation $[y_F|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{Y}]$ is $\mathrm{N}(\beta_0, \sigma^2_{b_0} + \sigma^2_\varepsilon)$. Observations from this distribution can therefore be generated from $\mathrm{N}(\beta_0^*, \sigma^{*2}_{b_0} + \sigma^{*2}_\varepsilon)$. Note that the values $(\beta_0^*, \sigma^{*2}_{b_0}, \sigma^{*2}_\varepsilon)$ would have already been generated as in Section 11.2.1, with the variance components contained in $\boldsymbol{\theta}^*$. The $(1-\beta)/2$ and $(1+\beta)/2$ quantiles of the generated values form the two-sided β-expectation limits. For more details we refer to Wolfinger (1998).

11.2.2 Hoffman and Kringle (HK) Approach

This approach is designed for general random effects models including balanced and unbalanced designs and addresses the issue of (β, γ) tolerance intervals. The HK method is based on the concept of effective sample size applied in conjunction with the Graybill and Wang (1980) method for constructing confidence intervals for variance components. Key aspects, necessary for the application of the method are briefly mentioned here, focusing on the balanced design case.

Let q denote the number of variance components in the model being considered and S_j^2 the sum of squares of the j^{th} component from a general ANOVA table for a random effects model. Note that for the model under consideration, $q = 2$, the length of vector $\boldsymbol{\theta}$. Further, denote the total variance of an arbitrary observation Y by σ^2_Y and variance for the mean overall $\bar{Y}$, by $\sigma^2_{\bar{Y}}$. For appropriate values of c_j and h_j, let $\sigma^2_Y = \sum_{j=1}^q c_j S_j^2$ and $\sigma^2_{\bar{Y}} = \sum_{j=1}^q h_j S_j^2$. Further, define the effective number of observations as $N_e = \frac{\sum_{j=1}^q c_j S_j^2}{\sum_{j=1}^q h_j S_j^2}$. According to Hoffmann and Kringle (2005), an approximate (β, γ) tolerance interval may be given by

$$\bar{Y} \pm z_{(1+\beta)/2}\sqrt{(1+N_e^{-1})}\sqrt{\hat{\sigma}^2_{\bar{Y}} + \left(\sum_{j=1}^q H_j^2 c_j^2 S_j^4\right)^{1/2}},$$

where $H_j = \frac{1}{F_{1-\gamma,n_j,\infty}}$ and $F_{1-\gamma,n_j,\infty}$ is a value from the F distribution with cumulative probability $1-\gamma$ and degrees of freedom n_j and ∞.

11.2.3 Distribution-Free Tolerance Intervals

The methods discussed in Section 11.2.1 and 11.2.2 depend of some distributional assumptions. When such assumptions can not be met, applicability of such methodology becomes questionable. This often happens for example in cases where sample sizes are relatively small. With the advent of distribution-free or the so-called nonparametric methods, one is spared the risk of mis-specifying a particular distribution. In this section, focus is put on one such method due to Hahn and Meeker (1991), applicable in the case of independent observations.

Let $t_1, \ldots, t_n$ be a random sample from any continuous distribution and $t_{(1)}, \ldots, t_{(n)}$ be the ordered sample. Hahn and Meeker (1991) define a two-sided tolerance interval designed to contain at least a proportion β of the sample with confidence coefficient γ. This therefore resonates with the idea of the (β, γ) tolerance intervals discussed in Section 11.2.1 and 11.2.2. The interval may be defined as $[t_{(l)}, t_{(u)}]$, for particular values l and u. Let the probability that the interval from a sample of size n, defined by the order statistics, will cover at least $100\beta\%$ of the population be defined by $P(n, l, u, \beta)$. The values l and u are chosen such that (Hahn and Meeker, 1991)

$$P(n, l, u, \beta) = B(u - l - 1; n, \beta) \geq \gamma, \tag{11.4}$$

for $0 \leq l < u \leq n + 1$ and $0 \leq p \leq 1$, with $B(u - l - 1; n, \beta)$ denoting the binomial based probability of observing at least $u - l - 1$ elements from a sample of size n with probability β. Thus, determination of the nonparametric interval, hence obtaining values for l and u essentially involves evaluating (11.4).

11.3 Application to the Ames Test Data Example

In this section, focus is put on illustrating some of the methods of constructing tolerance intervals. Particular attention will be given to the method of Wolfinger (1998), the approach of Hoffman and Kringle (2005) and the nonparametric method of Hahn and Meeker (1991). The data set described in Section 2.5 will be used for this purpose.

The example we consider involves 153 experiments and in each of the experiments, replicated measurements from three plates are available. One can therefore consider this example as an experiment involving repeated measures Y_{ij}, for subject $i(i =$

$1, \ldots, n$) and measurement $j (j = 1, \ldots, m)$. Here subject refers to the experiment and the repeated measurements refer to values from the different plates.

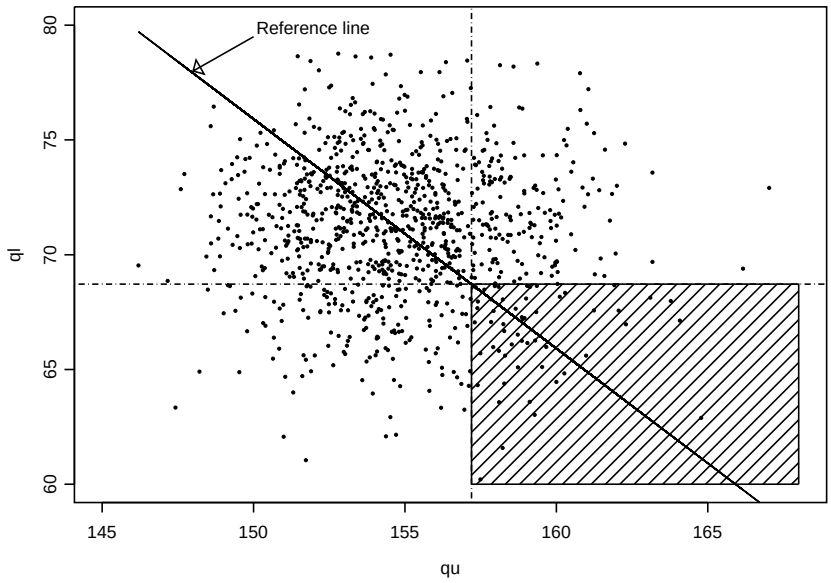


Figure 11.1: *Illustrating construction of the* $(\beta, \gamma) = (0.90, 0.95)$ *content tolerance interval of Wolfinger (1998) based on 1000 samples. The horizontal and vertical lines mark the required lower and upper limits respectively.*

Note that while the other two methods are suitable for repeated measures analysis, the nonparametric method deals with independent observations in a single sample. Therefore, an application of the nonparametric method in this case calls for summarizing the information from a particular subject into a single measurement, for example, using the mean.

The main question of interest was to determine, using the available historical data, limits which could be used to validate or invalidate measurements from future experiments. Using this data example, we will illustrate the different methods of constructing tolerance intervals as discussed earlier.

First, an illustration of the Bayesian simulation approach of Wolfinger (1998) is given. The process of constructing (β, γ) content tolerance intervals described in Section 11.2.1 is graphically shown in Figure 11.1. The shaded area in the right hand

bottom corner of the graph contains $100(1-\gamma)\%$ of the observations as described in Section 11.2.1. The dashed horizontal and vertical lines shown demarcate the two-sided end limits, that is, the respective lower and upper limit of the (β, γ) content tolerance interval, actual values of which are given in Table 11.1.

Limits from the other two methods under consideration are also given in Table 11.1. The choice for the values of β and γ follows traditionally used values. Other choices can as well be made. Although the different types of intervals are constructed

Table 11.1: *A comparison of the different intervals from the different approaches. For the parametric methods, a logarithmic transformation of the response is also used.*

Method	Transformation	β	γ	Lower limit	Upper limit
(β, γ) interval (Wolfinger)		0.90	0.95	68.57	151.21
(β, γ) interval (Wolfinger)	log	0.90	0.95	73.88	156.12
(β, γ) interval (HK)		0.90	0.95	68.17	151.61
(β, γ) interval (HK)	log	0.90	0.95	73.64	156.63
(β, γ) interval (Nonparametric)		0.90	0.95	73.00	157.00
β-expectation (Wolfinger)		0.90		62.18	157.11
β-expectation (Wolfinger)	log	0.90		70.24	164.56

from different perspectives, and in the case of (β, γ) and β-expectation, have different interpretations, for this example, the results in Table 11.1 do not indicate dramatic differences between the intervals. Thus, for each particular method, one would exclude observations that do not fall within the corresponding limits as per Table 11.1. Note that for the parametric methods, we have included interval calculation based on the logarithmic transformation of the response, aimed at upholding the normality assumption. It is also interesting to note the similarity between the intervals obtained under the log transformation and the nonparametric interval. However, note that, the non-parametric method is based on a summary of the data, hence does not take into account the repeated measures structure inherent in the data. Even though, an interesting observation is how closely the non-parametric approach compares with the other methods that take into account the structure in the data, especially for the transformed data.

11.4 Simulation Study

This section presents a simulation study to investigate how the different methods considered react to changes of certain factors. Of particular interest is the effect of correlation and the sample size on β and γ values in the (β, γ) tolerance intervals. We focus on the two methods that account for the clustering nature of the observations, i.e., the method due to Wolfinger (1998) and that of Hoffman and Kringle (2005).

To set up the simulation study, staying as close as possible to the data example, a linear mixed model is fitted to the data and resultant parameter estimates are used to generate data. In particular, data is generated from the model in (11.1), with the required inputs being the estimates from the data. Based on results from fitting the LMM to the data, it is assumed that $\beta_0 = 110$ and $\sigma_\varepsilon^2 = 180$. Values of $\sigma_{b_0}^2$ are chosen such that the intra-class correlation $\rho = \sigma_{b_0}^2/(\sigma_{b_0}^2 + \sigma_\varepsilon^2)$ takes on values 0.10, 0.50, and 0.80 reflecting low, medium, and high correlation. Also varied is the number of subjects, which is assumed to take on values of 20, 50, and 150. The number of repeated measurements per subject is kept fixed at 3, as in the data set. Of course this can be factored in as variable and its effect investigated in conjunction with the other factors mentioned above.

For a particular setting, a (β, γ) tolerance interval (l, u) is constructed. Following Hoffmann and Kringle (2005), the content of each interval is obtained by evaluating $\Phi(u) - \Phi(l)$, where $\Phi(\cdot)$ defines the cumulative distribution function of a standard normal distribution. The process is repeated for 10 000 times and the proportion of times the calculated content is at least β is obtained and represents the confidence level. The mean of the different content levels gives an estimate of the content level.

Simulation Results

In this section results of the simulation exercise described in the preceding section are summarized. A comparison between the method of Hoffman and Kringle (2005) and that of Wolfinger (1998) is conducted, focusing on (β, γ) tolerance intervals. The respective nominal values are taken to be $\beta = 0.90$ and $\gamma = 0.95$, values often used in the literature. For different values of ρ and sample sizes n, estimated values for β and γ are calculated and tabulated in Table 11.2.

First, from the method of Hoffman and Kringle (2005) it is evident that the estimated values for β and γ do not deviate much from their corresponding nominal values of 0.90 and 0.95 respectively. In general, the intervals tend to maintain their nominal content and confidence levels. It however appears that for relatively small

Table 11.2: *Investigating the effect of ρ and sample sizes on content (β) and confidence (γ) levels in the (β, γ) tolerance intervals. The indicated values for β and γ are the given nominal values.*

Hoffmann and Kringle method						
	$\beta = 0.90$			$\gamma = 0.95$		
	ρ			ρ		
n	0.10	0.50	0.80	0.10	0.50	0.80
20	0.9483	0.9483	0.9483	0.9703	0.9688	0.9689
50	0.9311	0.9312	0.9312	0.9592	0.9624	0.9632
150	0.9183	0.9183	0.9184	0.9569	0.9574	0.9582
Wolfinger method						
	$\beta = 0.90$			$\gamma = 0.95$		
	ρ			ρ		
n	0.10	0.50	0.80	0.10	0.50	0.80
20	0.9120	0.9060	0.9140	0.9353	0.9351	0.9360
50	0.8980	0.9100	0.9190	0.9227	0.9257	0.9309
150	0.9710	0.8670	0.8900	0.9211	0.9124	0.9138

sizes, $n = 20$ in the simulation study, the intervals tend to be conservative in the sense that the content and confidence levels exceed the corresponding nominal values. The effect of ρ is not clear from the results in Table 11.2. However, increasing the sample size appears to draw estimates closer to their nominal values. For the method of Wolfinger (1998), our simulations indicate that estimates for the content (β) are closer to their nominal level compared to the estimates for the confidence (γ) value. Further, content estimates tend to lie on both sides of β, implying no systematic under or overestimation of the nominal value. There however appears to be a clear underestimation of the confidence level. Our simulations therefore suggest that, in terms of the confidence level, the method of Wolfinger (1998) can be considered liberal, i.e., it tends to underestimate the nominal confidence level. Again, as in the Hoffman and Kringle (2005) method mentioned above, the effect of the correlation ρ is not evident here. A more elaborate simulation study may shade more light on this issue.

11.5 Discussion

This chapter has focused on application of tolerance limits in the context of historical data. In particular, limits which would in practice be used to validate or invalidate measurements were constructed. This was done under the notion that one requires certain limits which would contain a specified proportion of the population with a specified confidence. Three different methods addressing the same problem from different angles were considered. From the results obtained on the data example, the methods produced very comparable results. One of the methods considered is a nonparametric approach method, which normally, would be applied on independent data. In our case, the method was applied on a summary statistic of the repeated measures. The results from the nonparametric approach indicated comparable conclusions with the other methods which take into account the repeated measures nature of the data.

A simulation study to investigate the performance of the two methods that take into account the correlation structure in the data was set up. In general, the two methods tend to maintain their content nominal levels. The Bayesian simulation approach of Wolfinger (1998) was however found to be rather liberal in terms of confidence level. Of course the simulation settings considered here are not exhaustive, other effects as the number of replicated measurements could as well be investigated. The choice of the values for β and γ may be important, depending on the setting. Varying these values may also be interest. In general, a more detailed simulation study may be worthwhile pursuing to address some of the pertinent issues arising in this setting.

12

Concluding Remarks and Future Research

This thesis has touched upon two broad sections, the main part being flexible modelling techniques, with an in-depth focus on penalized spline methodology and the use of historical data in animal studies. We have demonstrated the versatility of the penalized spline based methodology. In particular, the use of the same basic model in different study designs and different intricacies is quite appealing. This follows the application of the methodology with continuous data to the parallel design case in Chapter 4, and also for the cross-over design in Chapter 5 and Chapter 6. The intermingling of the penalized spline methodology with surrogate marker validation techniques in Chapter 6 demonstrates its widespread applicability. Use of similar methodology with non-normal data was illustrated in Chapter 7 and Chapter 9.

The different models we proposed, coming from manipulating how penalized splines are constructed are applicable in many different scenarios. While the main inferential tool for data of a longitudinal nature was simultaneous confidence bands, accounting for various sources of variability, it will be interesting to compare the results obtained with other methods, for example, the approach of Behseta and Kass (2005). These authors propose a method for testing equality between functions, which they term a Gaussian test process. The approach is based on Hotelling's T^2 statistic, and, in-

terestingly, the method can be applied in conjunction with any smoothing technique, including penalized splines.

Particular attention has also been given to models including serial correlation, wherein time honored functions, such as exponential and Gaussian, have been considered. It is worth mentioning that flexible models considered here for the mean can be considered to model the serial correlation as well. This is something that has received limited attention thus far in the literature and could constitute a possible line of research.

In Chapter 8 we extended the discussion of smoothing of longitudinal data to the case of bivariate longitudinal outcomes in the Bayesian framework. Among other forms of correlation, imposition of correlation on response-specific smoothers was proposed. Further research in this area may be pursued in the direction of more than two longitudinally measured outcomes. This will inevitably increase the computational burden. Although not straightforward how it would fare with Bayesian models, the pairwise modelling approach of Fieuws and Verbeke (2006) could be a very useful route or starting point for this cause.

We have also proposed a Bayesian model for adaptive smoothing in Chapter 9, mainly geared for non-normal data. This is an extension to the conventional penalized spline model with global smoothing and has been studied quite in detail, especially for normally distributed data. However, at the moment, the only method known to us that deals with the exact problem we considered in Chapter 9 is based on an approximation to the likelihood. Our approach advocates for a fully Bayesian model. It will be interesting to perform extensive simulations expected to show the gain from the use of the Bayesian approach in view of the well documented problems associated with approximations to the likelihood. A couple of issues can be studied in this context. An investigation into the basis for the mean structure as well as for the penalty parameters can be considered. Specifically, one can investigate the performance of a combination of different bases as applied at the level of the mean and the penalty-parameter level. Although we demonstrated the approach on cross-sectional data, it is anticipated that extension to, for example, longitudinal data settings is straightforward. Extension of the approach to accommodate additive models as well spatial smoothing can also be considered.

The use of historical data has also been considered in this thesis. While in general, historical control data are expected to sharpen the estimation in current experiments, care should be taken in using them. Our simulations have shown that one make an informed selection of the historical control studies and use only those selected, instead of using all available historical studies. Scenarios where it would be encouraged to

make us of historical control information have also been suggested. Having focused on one particular model here, it is encouraged to consider other different types of models and investigate how different aspects play out. Extension in the direction of more complex designs, for example, clustered data can be considered. One can think of toxicity studies with animals where investigations are carried out on fetuses carried in wombs of their mothers. Here, various hierarchies exist and use of historical control data with such data maybe interesting to investigate. A different direction of research in this context would be to focus on more mathematical derivation of, for example, bias in the estimation of treatment effect. With the type of data normally encountered with such studies, the problem would reside outside the realm of the computationally friendly normal distribution, presenting a challenge.

A less frequently used type of intervals, known as tolerance intervals was also considered here. These are often used to detect if samples of interest lie within acceptable limits. Special focus was on reviewing available methods for implementing tolerance intervals as well illustrating some of the available methods. In line with this, we believe some extensive simulations focusing on some key properties of the intervals can enlighten on which particular methods are suitable under which conditions.

Appendix: Software Programs

Fitting Penalized Splines in SAS and S-Plus

Let us illustrate how the models discussed in Chapter 4 can be fitted using the MIXED and GLIMMIX procedures in SAS, as well as in S-Plus. In general, the concepts discussed here apply to other related models in the other chapters.

Selection of knots and knots location in SAS can be done using a SAS macro provided by Ruppert *et al.* (2003). In our SAS programs, we assume the matrix $\boldsymbol{Z}$ has been properly added to the data set, `thedata`, containing other variables of interest.

Model Fitting in SAS

Let us focus on Model 4*a*, a random-slope model where the two groups smoothed separately with the same smoothing parameter. The following SAS codes may be used to fit the model.

```
proc mixed data=thedata method=ml;
class dog group;
model hr=group time group*time/ solution;
random z1-z40/type=toep(1) subject=group s ;
random intercept time/type=un subject=dog;
run;
```

The linear part or the fixed effects part of the model is specified under the MODEL statement. The first RANDOM statement fits the non-parametric part of the model, with the option `subject=group` ensuring independent random effects by group but

with the same variance. Changing the option to `group=group` fits Model 5, with different smoothing parameters by group. The other models are more straightforward to fit. Should one opt to use the GLIMMIX procedure, the option `type=rsmooth`, implementing a radial basis, may be used. If the knots are already in a data set say `knotsdata`, the following option may be used: `type=rsmooth knotmethod=data();`. The complete code would then appear like:

```
proc glimmix data=thedata method=mmpl;
class dog group;
model hr=group time group*time/ solution;
random time/type=rsmooth knotmethod=data(knotsdata)
subject=group solution;
random intercept time/type=un subject=dog;
run;
```

To specify 40 equally spaced knots `knotmethod=equal(40);` can be used. One can also use the so-called kd-tree method of selecting knot points wherein one specifies a 'bucket size', for example `knotmethod=kdtree(bucket=50 knotinfo)`. The second RANDOM statement specifies the subject-specific random intercept and slope. All output of interest, especially for constructing confidence intervals and bands can be kept using the ODS OUTPUT statement.

Model Fitting in S-Plus

Here, we illustrate how the five models could be fitted in S-plus.

```
K<-40
knots<-quantile(unique(time),probs=seq(from=0.01,to=0.99,length=K))
Z<-outer(time,knots,"-")
Z<-Z*(Z>0)
indic<-factor(rep(1,length(hr)))
#No grouping structure in the observations,
#used for Models 1-3.

model.1<-lme(hr~time,random=list(indic=pdIdent(~Z-1),
dog=pdSymm(~time)),method="ML")
model.2<-lme(hr~group+time,random=list(indic=pdIdent(~Z-1),
dog=pdSymm(~time)),
method="ML")
```

```
model.3<-lme(hr~group*time,random=list(indic=pdIdent(~Z-1),
dog=pdSymm(~time)),
method="ML")
model.4<-lme(hr~group*time,random=list(group=pdIdent(~Z-1),
dog=pdSymm(~time)),
method="ML")

# In Model 4, 'group=pdIdent(~Z-1)' specifies independent random
effects by group, with the same smoothing parameter.
```

Model 5 requires a blocked structure for the matrix $\boldsymbol{Z}$. We can achieve this through the matrix zmat, obtained from,

```
k1<-k2<-length(knots)
timegrp<-time[grp==0]
z11<-z22<-outer(timegrp,knots,"-") z11<-z22<-z11*(z11>0)
d0<-matrix(0,nrow(z11),ncol(z11))
z1<-rbind(z11,d0)
z2<-rbind(d0,z11)
zmat<-cbind(z1,z2)
re.block.inds<-list(1:k1,(k1+1):(k1+k2)) z.bloc<-list() for(i in
1:length(re.block.inds))
z.bloc[[i]]<-as.formula(paste("~zmat[,c(",paste(re.block.inds[[i]],
collapse=","),")]-1"))

model.5<-lme(hr~group*time,random=list(group=pdBlocked(z.bloc,
pdClass="pdIdent"),dog=pdIdent(~1)),method="ML")
```

Fitting Adaptive Penalized Splines in WinBUGS

The following WinBUGS code illustrates how one may fit the adaptive penalized spline model discussed in Chapter 9. Fitting the non-adaptive version of the penalized spline model is also evident from these codes.

```
model{
for (i  in 1:T){Y[i] ~dpois(mu[i])
           mu[i] <-exp(linp[i])
               linp[i]<-fix1[i]+s1[i]+s2[i]+s3[i]
```

```
                fix1[i]<-beta[1]*X[i,1]+beta[2]*X[i,2]
s1[i]<-b[1]*Z[i,1]+b[2]*Z[i,2]+b[3]*Z[i,3]+b[4]*Z[i,4]+b[5]*Z[i,5]
+b[6]*Z[i,6]+b[7]*Z[i,7]+b[8]*Z[i,8]+b[9]*Z[i,9]+b[10]*Z[i,10]
s2[i]<-b[11]*Z[i,11]+b[12]*Z[i,12]+b[13]*Z[i,13]+b[14]*Z[i,14]
+b[15]*Z[i,15]+b[16]*Z[i,16]+b[17]*Z[i,17]+b[18]*Z[i,18]+b[19]*Z[i,19]
+b[20]*Z[i,20]
s3[i]<-b[21]*Z[i,21]+b[22]*Z[i,22]+b[23]*Z[i,23]+b[24]*Z[i,24]
+b[25]*Z[i,25]

}
for (j in 1:2){
beta[j]~dnorm(0,1.0E-6)
alpha[j]~dnorm(0,1.0E-6)}

for (j in 1:nknots)
{b[j]~dnorm(0,taub[j])
sigma2b[j]<-1/taub[j]
taub[j]<-exp(-fx[j])
fx[j]<-fix2[j]+s21[j]
fix2[j]<-alpha[1]*X1[j,1]+alpha[2]*X1[j,2]
s21[j]<-d[1]*Z1[j,1]+d[2]*Z1[j,2]+d[3]*Z1[j,3]+d[4]*Z1[j,4]
+d[5]*Z1[j,5]
}
for (k in 1:nknots1)
{d[k]~dnorm(0,taud)}

taud~dgamma(1.0E-6,1.0E-6)
sigma2d<-1/taud
}
```

SAS Macro to Calculate β-expectation and (β, γ) tolerance limits

In this section we present SAS codes illustrating how, in practice, one can implement the β-expectation and (β, γ) tolerance limits of Wolfinger (1998) seen in Chapter 11. The programs are run in a macro called by the call

```
FitWolf(data=,nsamp=,seed=,update=,alpha=);
```

Key inputs here are the number of samples to be generated `nsamp` and `alpha`$=1-\gamma$, assuming the data in `dat` follow the usual longitudinal format of stacking measurements from the different subjects.

```
%macro FitWolf(data=,nsamp=,seed=,update=,alpha=);
proc mixed data=dat;
class subject;
model resp=/s;
random subject/s;
prior /nsample=&nsamp update=&update seed=&seed out=myout;
run;

/*****************************/
/*   Beta-expectation limits   */
/*****************************/
data tol(keep=beta1 ql qu y);
   set myout;
   q90 = probit(.9);
   q95 = probit(.95);
   sd = sqrt(covp1+covp2);
   q = beta1 - q90*sd;
   ql = beta1 - q95*sd;
   qu = beta1 + q95*sd;
   y = beta1 + sd*rannor(1284703);
   output;
run;

proc univariate data=tol;
var y;
output  pctlpts=2.5 97.5 pctlpre=pp out=outlimit;
run;

data outlim;
set outlimit;
lower=exp(pp2_5);
upper=exp(pp97_5);
run;
```

```
/*************************/
/* Beta-Content limits     */
/*************************/
proc means data=tol;
var beta1;
output out=mout mean=m;
run;
data _null_;
set mout;
call symput('m',m);
run;

proc iml;
use tol;
read all var{ql} into ql;
read all var{qu} into qu;
qtemp=qu[1];
do until (f<=&alpha);
qtemp=qtemp+0.0001;
qlow=-qtemp+2*&m;
qup=qtemp;
f=abs((sum(ql<qlow & qu>qup)/&nsamp));
end;
lower=qlow;
upper=qup;
print lower upper;
quit;

%mend;

%FitWolf(data=dat,nsamp=10000,seed=1975,update=1e3,alpha=0.05);
```

References

Akaike, H. (1973) Maximum likelihood identification to Gaussian autoregressive moving average models. *Biometrika*, **60**, 255–265.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M.G. (2003) Validation of surrogate markers in multiple randomized clinical trials with repeated measures. *Biometrical Journal*, **45**, 931–945.

Alonso, A., Molenberghs, G., Geys, H., and Buyse, M. (2006) A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine*, **25**, 205–211.

Amaratunga, D. (1997) Reference ranges for screening preclinical drug safety data. *Journal of Biopharmaceutical Statistics*, **7**, 417–422.

Azzalin, A., and Bowman, A. (1993) On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society Series B*, **55**, 549–559.

Bagui, S.C., Bhaumik, D.K., and Parnes, M. (1996) One-sided tolerance limits for unbalanced m-way random effects ANOVA models. *Journal of Applied Statistical Science*, **3**, 135–148.

Balandayuthapani, V., Mallick, B.K., and Carroll, R.J. (2005) Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics,* **14**, 378-394.

Best, N.G., Cowles, M.K., and Vines, S.K. (1995) *CODA Manual Version 0.30. MRC Biostatitistics Unit.* Cambridge, UK.

Beckman, T.J., and Tietjen, G.L. (1989) Two sided tolerance limits for balanced random-effects ANOVA models. *Technometrics*, **31**(2), 185–197.

Behseta, S., and Kass, R.E. (2005) Testing equality of two functions using BARS. *Statistics in Medicine*, **24**, 3523–3534.

Behseta, S., Wallstrom, G.L., and Kass, R.E. (2005) Hierachical models for assessing variability among functions. *Biometrika*, **92**, 419–434.

Bhaumik, D.K., and Kulkarni, P.M. (1996) A simple and exact method of constructing tolerance intervals for the one-way ANOVA with random effects. *American Statistician*, **50**(4), 319–323.

Breslow, N. E., and Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81-91.

Burnham, K.P., and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005) *The Evaluation of Surrogate Endpoints.* New York: Springer.

Buyse, M., and Molenberghs, G. (1998) The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000) The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics*, **1**, 49–67.

Cadarso-Suárez, C., Roca-Pardinas, J., Molenberghs, G., Faes, C., Nacher, V., Ojeda, S., and Acuna, C. (2006) Flexible modeling of neuronal firing rates across different experimental conditions: an application to neural activity in the prefrontal cortex during a discrimination task. *Journal of the Royal Statistical Society Series C*, **55**, 431–447.

Chaudhuri P., and Marron, J.S. (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.

Cleveland, W.S., and Devlin, S.J. (1988) Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cleveland, W.S., and Grosse, E. (1991) Computational methods for local regression. *Statistics and Computing*, **1**, 47–62.

Cordeiro, G. M., and McCullagh, P. (1991) Bias correction in generalized linear models. *Journal of the Royal Statistical Society Series B*, **53**, 629-643.

Crainiceanu, C.M, Ruppert, D., Carroll, R.J., Joshi, A., and Goodner, B. (2007) Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, **16**, 265–288.

Crainiceanu, C.M., Ruppert, D., Claeskens, G., and Wand, M.P. (2005a) Exact likelihood ratio tests for penalized splines. *Biometrika*, **92**, 91–103.

Crainiceanu, C.M., Ruppert, D., and Wand, M.P. (2005b) Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software*, **14**. Available online at `http://www.jstatsoft.org/v14/i14/`

De Groote, L. and Linthorst, A.C. (2007) Exposure to novelty and forced swimming evoke stressor-dependent changes in extracellular GABA in the rat hippocampus. *Neuroscience*, **148**, 794–805.

De Gruttola, V., Lange, N., and Dafni, U. (1991) Modeling the progression of HIV infection. *Journal of the American Statistical Association,* **86**, 569-577.

Dempster, A.P., Selwyn, M.R., and Weeks, B.J. (1983) Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association*, **78**, 221-227.

DiMatteo, I., Genovese, C.R., and Kass, R.E. (2001) Bayesian curve-fitting with free-knot splines. *Biometrika*, **88**, 1055–1071.

Dom, P., Janssen, T. and Coussement, W. (2000) Oral developmental toxicity study in the Rabbit GLP-study. *Janssen Pharmaceutica N.V Non-Clinical Laboratory Study: Technical report.*

Dunsmore, I.R. (1981) Growth curves in two-period change over models. *Applied Statistics*, **30**, 223-229.

Durbán, M., Harezlak, J., Wand, M.P., and Carroll R.J. (2005) Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153–1168.

Eilers, P.H.C. and Marx B.D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Faes, C., Geys, H., Molenberghs, G., Aerts, M., Cadarso-Suárez, C., and Acuna, C., and Cano, M. (2007) A flexible method to measure synchrony in neuronal firing. *Journal of the American Statistical Association, 00,000-000*

Fan, J., and Gibjbels, I. (1996) *Local polynomial modelling and its applications.* Chapman & Hall.

Fearn T. (1975) A Bayesian approach to growth curves. *Biometrika*, **62**, 89–100.

Fieuws, S., and Verbeke, G. (2006) Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics*, **62**: 424-431.

Friedman, J., and Silverman, R. (1989) Flexible parsimonious smoothing and additive modelling. *Technometrics*, **31**, 3–39.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992) Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

Ganguli, B., and Wand, M.P. (2007) Feature significance in generalized additive models. *Statistical Computing*, **17**, 179-192.

Gelman, A., Carlin, J.B., and Stern, H.S. (1995) *Bayesian Data Analysis.* Chapman & Hall.

Gelman, J., and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

Gerstein G.L., and Kiang N.Y.S (1960) An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophysical Journal*, **1**, 15-28.

Gijbels, I., Hall, P., Jones, M.C., and Koch, I. (2000) Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, **87**(3), 663–673.

Green, P.J., and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models: a roughness penalty approach.* Chapman & Hall.

Graybill, F.A., and Wang, C.M. (1980) Confidence intervals on nonnegative linear combinations of variances. *Journal of American Statistical Association*, **75**, 869–873.

Grender, J.M., and Johnson, W.D. (1994) Fitting multivariate polynomial growth curves in two period crossover designs. *Statistics in Medicine*, **13**, 359-365.

Guo, W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121-128.

Guttman, I. (1970) *Statistical Tolerance Regions: Classical and Bayesian.* London: Charles W. Griffin and Co.

Hahn, G., and Meeker, W. (1991) *Statistical Intervals, A Guide for Practioners.* John Wiley and Sons.

Hart, J.D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer-Verlag: New York, Inc.

Harezlak, J., Naumova, E., and Laird, N.M. (2007) Long*CriSP*: A test for bump hunting in longitudinal data. *Statistics in Medicine*, **26**, 1383–1397.

Hastie, T.J., Tibshirani, R.J. (1990) *Generalized Additive Models.* Chapman & Hall.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Heckman, N.E. (1992) Bump hunting in regression analysis. *Statistics and Probability Letters*, **14**, 141–152.

Hoel, D. G. (1983). Conditional two sample tests with historical controls, In Sen, P. K.(Ed): *Contributions to Statistics: Essays in honour of Normal L. Johnson. Amsterdam: North Holland Publishing Company,* 229-236.

Hoffman, D., Kringle, R. (2005) Two-sided tolerance intervals for balanced and unbalanced random effects models. *Journal of Biopharmaceutical Statistics*, **15**, 283–293.

Ibrahim, J.G., Ryan, L.M. (1996) Use of historical controls in time-adjusted trend tests for carcinogenicity. *Biometrics,* **52**, 1478-1485.

Ibrahim, J.G., Ryan, L.M., Chen, M.H. (1998) Use of historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association,* **93**, 1282-1293.

Jones B., Kenward M.G. (2003) *Design and Analysis of Cross-Over Trials.* London: Chapman & Hall/CRC.

Kass, R.E., Ventura, V., Brown, E.N. (2005) Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology*, **94**, 8–25.

Kass, R.E., Ventura, V., Cai, C. (2003) Statistical smoothing of neuronal data. *Network: Computation in Neural Systems*, **14**, 5–15.

Kikuchi, Y., Yanagawa, T. (1991) Incorporating historical controls using a random effects models with a normal prior. *Communications in Statistics - Theory and Methods,* **20** (4), 1273-1291.

Krivobokova, T., Crainiceanu, C.M., Kauermann, G. (2008) Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17**(1), 1–20.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Lee, Y., Nelder, J.A., Pawitan, Y. (2006) *Generalized Linear Models with Random Effects: Unified analysis via H-likelihood.* Chapman & Hall/CRC.

Lehmann, E.L., D'Abrera, H.J.M. (1975) *Nonparametrics. Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Liao, C.T., Iyer, H.K. (2004) A tolerance interval for the normal distribution with several variance components. *Statistical Sinica,* **14**, 217–229.

Lin, X., Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B*, **61**: 384-400.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**: 325–337.

McCullagh, P., Nelder, J.A. (1989) *Generalized Linear Models.* London: Chapman & Hall.

Mee, R.W. (1984) β-Expectation and β-Content tolerance limits for balanced one-way ANOVA random model. *Technometrics*, **26**, 251–254.

Mee, R. W. (1989) Normal distribution tolerance limits for stratified random samples. *Technometrics*, **31**(1), 99–105.

Mee, R.W., Owen, D.B. (1983) Improved factors for one-sided tolerance limits for balanced one-way ANOVA random model. *Journal of the American Statistical Association*, **78**, 901–905.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Molenberghs, G., Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Ngo, L., Wand, M.P. (2004) Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1–56.

Parise, H., Wand, M. P., Ruppert, D., Ryan, L. (2001) Incorporation of historical controls using semi-parametric mixed models. *Applied Statistics,* **50** (1), 31-42.

Patel H.I, Hearne E.M. (1980) An application of multivariate analysis to the two-period repeated measures crossover design with application to clinical trials. *Communications in Statistics Theory and Methods*, **9**, 1919–1929.

Pien, H.H., Fischman, A.J., Thrall, J.H., Sorensen, A.G. (2005) Using imaging biomarkers to accelerate drug development and clinical trials. *Drug Disc Today*, **10**, 259–266.

Prentice, R.L. (1989) Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Putt, M., Chinchilli, V.M. (1999) A mixed effects model for the analysis of repeated measures cross-over studies. *Statistics in Medicine*, **18**: 3037-3059.

Robert, C.P., Casella, G. (1999) *Monte Carlo Statistical Methods.* New York: Springer-Verlag.

Roca-Pardinas, J., Cadarso-Suárez, C., Nacher, V., Acuña, C. (2006) Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. *Statistics in Medicine*, **25**, 2483-2501.

Royston, P., Altman, D.G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, **43**, 429–467.

Ruppert, D. (2002) Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

Ruppert, D., Carroll, R.J. (2000) Spatially adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **42**, 205–223.

Ruppert D., Wand M.P., Carroll R.J. (2003) *Semiparametric Regression.* Cambridge University Press.

Ryan, L. (1993) Using historical controls in the analysis of developmental toxocity data. *Biometrics*, **49**, 1126-1135.

SAS Institute Inc. (2004) *The GLIMMIX Procedure (Experimental).* Cary, NC: SAS Institute Inc.

Senn, S. (1993) *Cross-over Trials in Medical Research.* Chichester: John Wiley.

Shapiro, S.S., Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591-611.

Self, S.G., Liang, K. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*, **82**, 605-610.

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society Series B*, **47**, 1–52.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society Series B*, **64**: 583-640.

Tarone, R.E. (1982) The use of historical control information in testing for trend in proportions. *Biometrics,* **38**, 215-220.

Thiebaut, R., Jacqmin-Gadda, H., Chene, G., Leport, C., Commenges, D. (2002) Bivariate linear mixed models using SAS proc MIXED. *Computer methods and programs in Biomedicine*, **69**, 249-256.

Vazquez, P., Cano, M., Acuna, C. (2000) Discrimination of the line orientation in humans and monkeys. *Journal of Neurophysiology*, **83**, 2639–2648.

Ventura, V., Carta, R., Kass, R.E, Gettner, S.N., Olson, C.R. (2002) Statistical analysis of temporal evolution in single-neuron firing rates. *Biostatistics*, **1**, 1–20.

Verbeke, G., Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal data.* New York: Springer.

Verbeke, G., Molenberghs, G. (2003) The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262.

Verbyla, A.P, Cullis, B.R., Kenward, M.G., Welman, S.J. (1999) The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, **48**, 269–311.

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society Series Series B,* **40**, 364–372.

Wald, A. (1942) Setting of Tolerance Limits when the Sample is Large. *The Annals of Mathematical Statistics,* **13**, 389–399.

Wald, A. (1943) An Extension of Wilks Method for Setting Tolerance Limits, *The Annals of Mathematical Statistics,* **17**, 208–215.

Wald, A., Wolfowitz, J. (1946) Tolerance limits for a normal distribution. *Annals of Mathematical Statistics*, **17**, 2008–215.

Wager, C., Vaida, F., and Kauermann, V.G. (2005). Model section for P-spline smoothing using Akaike Information Criteria. *Technical Report*, Bielefeld University, http://www.wiwi.uni-bielefeld.de/ kauermann/research/WagerVaidaKauermann.pdf

Wallenstein S., Fisher A.C. (1977) The analysis of the two-period repeated measurements cross-over design with application to clinical trials. *Biometrics*, **33**, 251–269.

Wilks, S. (1941) Determination of Sample Sizes for Setting Tolerance Limits. *The Annals of Mathematical Statistics,* **12**, 91–96.

Wolfinger, R.D. (1998) Tolerance intervals for variance component models using Bayesian simulation. *Journal of Quality Technology*, **30**, 18–31.

Wolfinger, R.D., O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **4**, 233-243.

Wood, S. (2004) On confidence intervals for GAMs based on penalized regression splines. *Tech. Report*, Dept. Statistics, Univeristy of Glasgow.

Zhao, Y., Staudenmayer, J., Coull, B.A., Wand, M.P. (2006) General design Bayesian generalized linear mixed model. *Statistical Science*, **21**, 35–51.

Zeger, S.L., Diggle, P.J. (1994) Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, **50**: 689-699.

Samenvatting

In deze thesis staan twee onderwerpen centraal. Enerzijds worden flexibele modelleringstechnieken behandeld, met een diepgaande focus op de penalized spline methodologie; anderzijds wordt het gebruik van historische data in dierenstudies besproken. In deze thesis hebben we de veelzijdigheid van de penalized splines gebaseerde methodologie aangetoond. In het bijzonder toonden we dat hetzelfde basismodel erg aantrekkelijk is in studies met toch verschillende designs en/of verschillende complexiteiten. De penalized spline methodologie voor continue gegevens werd onderzocht in zowel de parallel design setting in hoofdstuk 4, als in de cross-over design in hoofdstukken 5 en 6. Het vermengen van de penalized spline methodologie met surrogaat respons validatie technieken in hoofdstuk 6 toont eveneens de brede toepasbaarheid van de methode. Het gebruik van gelijkaardige methodes voor niet-normale gegevens kwam aan bod in hoofdstukken 7 en 9.

De modellen die in deze thesis vooropgesteld worden vinden hun oorsprong in de manier waarop penalized splines geconstrueerd worden, en zijn toepasbaar in vele scenario's. In de thesis werd het gebruik van gezamenlijke betrouwbaarheidsbanden voor inferentie bij longitudinale gegevens onderzocht. In toekomstig onderzoek is het interessant om de verkregen resultaten te vergelijken met andere methodes, waaronder de methode van Behseta en Kass (2005). Deze auteurs stellen een methode voor om gelijkheid van twee functies te testen met behulp van een Gaussische test procedure, welke gebaseerd is op de Hotelling's T^2 statistiek. Deze methode kan eveneens worden toegepast bij het gebruik van smoothing methodes, waaronder penalized splines.

In de thesis werd ook bijzondere aandacht gegeven aan modellen met seriële correlatie, waarin de traditionele functies zoals de exponentiele en Gaussische functies werden onderzocht. Naast flexibele modellen voor de gemiddelde structuur kunnen ook flexibele modellen voor de seriële correlaties onderzocht worden, wat tot zover slechts weinig aandacht gekregen heeft in de literatuur.

In hoofdstuk 8 werd de bespreking van smoothing methodes voor longitudinale

gegevens uitgebreid naar het bivariaat modelleren van longitudinale uitkomsten in een Bayesiaanse setting. De correlatie tussen de verschillende uitkomsten werd opgelegd door onder meer een correlatie tussen de uitkomst-specifieke smoothing termen, en werd onderzocht. Verdere uitbreiding naar meerdere responsen is mogelijk maar vraagt verder onderzoek omwille van de toenemende computationele vereisten. Mogelijks is een oplossing de paarsgewijze modelleringsmethode voorgesteld door Fieuws en Verbeke (2008).

In deze thesis werd ook een Bayesiaan model voor adaptive smoothing voorgesteld (hoofdstuk 9), hoofdzakelijk in de context van niet-normale gegevens. De voorgestelde methode is een uitbreiding van de conventionele penalized spline modellen met globale smoothing. Een gelijkaardige methode werd voorgesteld in de literatuur, gebaseerd op een benadering van de likelihood. Uitgebreide simulaties welke de voordelen van de voorgestelde Bayesiaanse methode tonen in vergelijking met de bestaande methode is een belangrijk onderzoeksonderwerp, en moet nog worden uitgevoerd. We verwachten een voordeel van de Bayesiaanse methode ten opzichte van de methode gebruik makend van een benadering van de likelihood, omwille van de gekende problemen van deze benadering. Verscheidene onderwerpen kunnen verder onderzocht worden in deze context. Het zoeken naar de best geschikte basis functie gebruikt voor de hoofdstructuur en voor de penalizatie-parameters; het onderzoeken van de werking van een combinatie van basis functies op het niveau van het gemiddelde en het niveau van de penalizatie-parameter; het tonen van de toepasbaarheid van de methode in, bijvoorbeeld, de context van longitudinale data; het uitbreiden van de methode om zowel additieve modellen als spatiale smoothing toe te laten.

Tot slot werd het gebruik van historische gegevens onderzocht in deze thesis. Terwijl historische controle-gegevens in het algemeen de schatting in het huidige experiment verscherpen, moet men toch omzichtig omgaan met het gebruik van historische data. Het aantal beschikbare historische controle studies en de variabiliteit tussen deze studies hebben een belangrijk effect op de schatting. Het effect van het aantal historische studies, van de gelijkaardigheid van de historische studies, van de sterkte van het dosis effect en van het achtergrond-effect op de precisie van de schatting van een dosis effect werd onderzocht, en is beschreven in hoofdstuk 10. Uitbreiding van de methode in meer complexe settings, zoals bijvoorbeeld, geclusterde gegevens, kan worden beschouwd. Een voorbeeld is een ontwikkelings-toxicologische studie waar de foetussen van moederdieren worden onderzocht. Verscheidene hierarchieën zijn hierbij van belang, en het gebruik van historische data met dergelijk type data is niet vanzelfsprekend en moet verder worden onderzocht. Een ander mogelijk onderzoeksonderwerp is de mathematische afleiding van, bijvoorbeeld, de fout in de schatting

van het dosis-effect in een dergelijke studie.

www.uhasselt.be

Universiteit Hasselt | Campus Diepenbeek
Agoralaan | Gebouw D | B-3590 Diepenbeek | België
Tel.: +32(0)11 26 81 11