HIERARCHICAL AND MULTIDIMENSIONAL SMOOTHING WITH APPLICATIONS TO LONGITUDINAL AND MORTALITY DATA

Viani Aimé Djeundje Biatat

Submitted for the degree of
Doctor of Philosophy
on completion of research in the
Department of Actuarial Mathematics & Statistics,
School of Mathematical and Computer Sciences,
Heriot-Watt University

November 2011

Abstract

This thesis is concerned with two themes: (a) smooth mixed models in hierarchical settings with applications to grouped longitudinal data and (b) multi-dimensional smoothing with reference to the modelling and forecasting of mortality data.

In part (a), we examine a popular method to smooth models for longitudinal data, which consists of expressing the model as a mixed model. This approach is particularly appealing when truncated polynomials are used as a basis for the smoothing, as the mixed model representation is almost immediate. We show that this approach can lead to a severely biased estimate of the group and subject effects, and to confidence intervals with undesirable properties. We use penalization to investigate an alternative approach with either B-spline or truncated polynomial bases and show that this new approach does not suffer from the same defects. Our models are defined in terms of B-splines or truncated polynomials with appropriate penalties, but we re-parametrize them as mixed models and this gives access to fitting with standard procedures.

In part (b), we first demonstrate the adverse impact of over-dispersion (and heterogeneity) in the modelling of mortality data, and describe the resolution of this problem through a two-stage smoothing of mean and dispersion effects via penalized quasi-likelihoods. Next, we propose a method for the joint modelling of several mortality tables (e.g. male and female mortality in Demography, mortality by lives and by amounts in Life Insurance, etc) and describe how this joint approach leads to the classification and simple comparison of these tables. Finally, we deal with the smooth modelling of mortality improvement factors, which are two-dimensional correlated data; here we first form a basic flexible model incorporating the correlation structure, and then extend this model to cope with cohort and period shock effects.

Acknowledgements

First and foremost, I would like to thank my advisor Iain Currie; his kindness, encouragement and commitment to my personal and scientific development have been invaluable to me.

My thanks go also to the Engineering and Physical Sciences Research Council, and the Continuous Mortality Investigation for their financial support. I extend my gratitude to colleagues who were funded by the Spanish Ministry of Science and Innovation (project MTM 2008-02901); especially Paul Eilers and Maria Durban for insightful comments and discussions. I acknowledge the helpful discussion with Angus Macdonald and Stephen Richard regarding the work in Section 5.3.

I am greatly indebted to my family (Janine, mum, Cyrille, Charles,...) for their unconditional love and support. I wish also to thank many student colleagues and surrogate family (Antoine, Issa, Kokouvi, Christine, Elizabeth,...) for making my time in Edinburgh so much fun. I do not forget my "Taps' Cops" friends spread all over the world.

Above all, glory and praise be to the Designer of all "things".

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In many areas where Applied Statistics plays a key role, data often have various and complex shapes causing parametric modelling attempts to fail. Nonparametric or smoothing techniques provide an attractive solution in such situations in that the shape of the functional relationships is not predetermined, but are driven by the data; ie, can adjust to capture unusual or unexpected features in the data. More interestingly, smoothing can be embedded in any application area of Statistics that uses regression-type analysis. Hence smoothing methods play a pivotal role in modern Statistics. This thesis is concerned with smooth models in two settings: (a) a hierarchical setting with applications to longitudinal data and (b) a multi-dimensional setting with reference to the modelling and forecasting of mortality data. In this introduction we will first outline the motivation of our interest in these two domains and then sketch the plan of the thesis.

1.1 Modelling longitudinal data

Repeated measurements on subjects over time are common in many areas and the analysis of such data is referred to as a longitudinal study. By virtue of the repeated observations at the subject level, longitudinal studies have more features than cross-sectional observational studies, since they allow us to distinguish short from long-term phenomena. Because of this benefit, they play a prominent role in Statistics and provide important information about individual change.

Very often, longitudinal data have a grouped structure and so, the central interest

of the study usually lies in the simultaneous estimation of the group and subject effects. In this context, mixed models represent a powerful tool for data analysis, and in certain cases, a parametric mixed model is sufficient to summarize the desired effects from the data (Searle et al., 2006; Pinheiro and Bates, 2000). In practice however, the patterns at both the group and subject levels are generally unknown, and one common way to tackle this complexity is to incorporate flexibility or smoothing into the modelling process; Coull et al. (2001b) and Durban et al. (2005) refer to this modelling framework as subject-specific curves or factor-by-curve interaction models. A well known method in this setting (as explored by these authors as well as Ruppert et al. (2003), among others) uses a smooth mixed model, with the curves described in terms of truncated lines and the randomness expressed in terms of normal distributions. This method is attractive, first in its simplicity, and second in that it offers the possibility to estimate the group and subject effects through standard mixed model packages. However, this approach has received very little attention in terms of its ability to appropriately identify these effects.

1.2 Modelling mortality data

During the last century, large increases in life expectancy have followed medical and scientific breakthroughs unimagined a hundred years ago and this increasing tendency shows no sign of slowing down in the near future (Willets, 1999). Such an improvement is welcomed from the individual point of view as survival is a very potent human instinct. However, this benefit has brought with it stress and real challenges to governments and to the Insurance Industry. Indeed, for governments, planning public services relies heavily on future life expectancy, and for the insurance industry, future mortality rates play a consequential role in the pricing process. As a result, modelling and forecasting mortality has gained particular attention and become an area of intensive investigation.

In recent decades, a wide range of mortality models has been proposed and discussed, the best known being the Lee-Carter model, introduced and developed by Lee and Carter (1992). This model finds its merits in its simplicity and its robustness. Intrinsically, it involves a (simple) bilinear function of age and time, and uses decompositions to extract a single time-varying mortality index with a time series model.

In other words, this model assumes that the mortality dynamism over time at all ages is driven by a single time varying index. Consequently, forecasting mortality follows immediately from the time series forecasting of the time index. Many variants of Lee-Carter's original idea have been investigated by several authors. While some of these authors (Booth et al., 2006) questioned the fitting period as well as the adjustment of the time index, others such as Brouhns et al. (2002) investigated the issue of the distribution of the number of deaths. Specifically, these latter authors assumed that the number of deaths is Poisson distributed, and this allowed them to fit the model by maximum likelihood. Many extensions of the Lee-Carter model are based on this Poisson assumption; see for example Cairns et al. (2009) and references therein.

One criticism of the Lee-Carter framework is the strong assumption made about the functional form of the mortality surface. Alternatively, fully two-dimensional approaches have been proposed (de Boor, 1978; Dierckx, 1996), but most of them have focussed on the modelling process and did not consider the extrapolation problem. Currie et al. (2004) presented an approach that allows for extrapolation in a straightforward manner. Specifically, these authors used regression splines and looked at the mortality as a smooth surface sited appropriately on top of a two dimensional Kronecker B-spline basis in age and time. In this context, they addressed the smoothness issue using the difference penalty, which was originally introduced by Eilers and Marx (1996). Beside the flexibility, one nice feature of this approach is that forecasting follows naturally from the smoothing process. Although this approach can be applied to a wide range of distributions, Currie et al. (2004) fitted the mortality surface under the Poisson assumptions as suggested by Brouhns et al. (2002).

However, mortality data are often classified by age at death and year of death. Such a classification results in a heterogeneous risk set and this makes the Poisson assumption awkward for these data. Also, some mortality tables, such as those of males and females, have some connections between them; modelling such tables independently can result in serious incongruity. In addition to these problems, the existence of correlations structures, as well as the simultaneous presence of cohort and period effects in some mortality data poses serious challenges for model building.

1.3 Guide to this thesis

The work presented in this thesis has led to

- V. A. D. Djeundje and I. D. Currie (2010). Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, 4, 1202-1224.

- V. A. D. Djeundje and I. D. Currie (2010). Smoothing dispersed counts with applications to mortality data. *Annals of Actuarial Science*, 5, 33-52.

- V. Biatat and I. D. Currie (2010). Joint models for classification and comparison of mortality in different countries. *Proceedings of 25rd International Workshop on Statistical Modelling*, Glasgow, 89-94.

- V. A. D. Djeundje and I. D. Currie (2011). Fitting subject-specific curves to grouped longitudinal data. *Proceedings of 58th World Statistics Congress, Dublin.*

- V. A. D. Djeundje and I. D. Currie (2011). Smooth mixed models for nested curves. *Proceedings of 26th International Workshop on Statistical Modelling, Valencia.*

and two working papers entitled:

- Penalized spline smoothing for hierarchical curves with applications to grouped longitudinal data.

- Smoothing correlated data: the mortality improvement factor.

The core of this dissertation consists of six Chapters divided into two parts. The first part, comprising Chapters 2 to 4, presents the background material and investigates smooth models in a hierarchical setting with applications to longitudinal data. The second part, consisting of Chapters 5 to 7, is concerned with multi-dimensional smoothing with reference to mortality data.

In Chapter 2, we lay the groundwork and notation for subsequent Chapters, motivate the need to move from the parametric framework to the smoothing setting, and review some common smoothing methods with particular attention to penalized splines based on truncated polynomials and B-splines.

In Chapter 3, we investigate the modelling of longitudinal data in which one is interested in estimating both the population and subjects effects in a smooth fashion. Here, we first present some problems arising from fitting a well known standard smooth mixed model based on truncated lines. These problems motivate us to develop an alternative approach via penalties on truncated polynomials or B-splines bases. Although this alternative approach is constructed via penalty arguments, we give its re-parametrisation and interpretation as a mixed model. This Chapter is based on Djeundje and Currie (2010a), and the results are presented here for balanced data in which the same number of observations are made on each subject at the same time points.

In Chapter 4, we extend Chapter 3 to multiple groups and unbalanced data. We start by demonstrating in more detail the problems arising from the standard approach with truncated lines, first on real data, and then through a simulation study. Next, we extend the penalty approach, describe its implementation via restricted likelihood with best linear unbiased predictor, discuss the computational demands, illustrate its consistency on real and simulated data, and finally outline its generalization to the multivariate setting. Most of the material in this Chapter has appeared in Djeundje and Currie (2011a,b).

In Chapter 5, we turn to the modelling of mortality data. First, we outline the formulation of penalized splines with B-spline bases for the modelling and extrapolation of two-dimensional mortality tables under the Poisson assumption. Second, we illustrate the negative impact of over-dispersion (and heterogeneity), and use quasi-likelihoods to describe a general class of smooth models for dispersed count data through a two-stage joint modelling of mean and dispersion effects. Some material in this Chapter has been described in Djeundje and Currie (2010b).

In Chapter 6, we propose a class of additive models for the economical joint modelling, comparison, and consistent forecasting of "similar" mortality tables. Here, the first component of our models describes a (common) two-dimensional smooth surface (viewed as the reference), and the remaining components depict the relative differences between these tables. Interestingly, this approach gives a straightforward and simple way of comparing and classifying populations (or mortality tables) into different categories. This Chapter is an extended version of Biatat and Currie (2010).

Finally in Chapter 7, we investigate the smoothing of two-dimensional correlated data, with reference to the estimation of mortality improvement factors. First, we derive a correlation structure in this setting, and set up a basic smooth model for the mortality improvement. Next, we discuss the limits of this basic model and its extension to an additive model with three components, where the first component portrays the underlying two-dimensional surface, the second component captures the period effects, and the last component takes care of the cohort effects.

We close with a summary and a future agenda in Chapter 8.

Chapter 2

Smoothing in one dimension

This chapter is designed to provide the necessary background material for this thesis. Here we motivate the need to move from the parametric framework to the smoothing platform, and give a brief review of popular smoothing methods with an emphasis on penalized splines based on two commonly used bases, namely truncated polynomials and B-splines.

We start in Section 2.1 by reviewing standard polynomial regression and then motivate the need for smoothing methods. In Section 2.2 we outline some well-known smoothing approaches, namely *local averaging*, *kernel smoothing* and *smoothing splines*, which are examples of the so-called *full rank* methods. In Section 2.3 we turn to *low rank* methods (as opposed to the full rank methods), especially *penalized splines* based on truncated polynomials and B-splines. In Sections 2.4 and 2.5, we discuss issues related to model selection and confidence bands. In Section 2.6 we look at smoothing in the exponential family setting. In Section 2.7 we outline the idea of Bayesian smoothing, and close with a brief discussion on adaptive smoothing in Section 2.8.

2.1 From parametric to smoothing models

The fundamental aim of Statistics is to summarize data and explore potential relationships between variables, and a useful tool to achieve this purpose is found in regression models (Dobson, 1983; McCullagh and Nelder, 1989). Let us assume that we have measurements of a response variable $\boldsymbol{y} = (y_1, ..., y_n)'$ at the covariate design

points $\boldsymbol{x} = (x_1, ..., x_n)'$; for convenience, we suppose that these measurements are sorted by the covariate as $x_1 \leq \cdots \leq x_n$. For illustrations, we shall consider the data depicted in Figure 2.1. These data, which can be obtained from the `MASS` package in

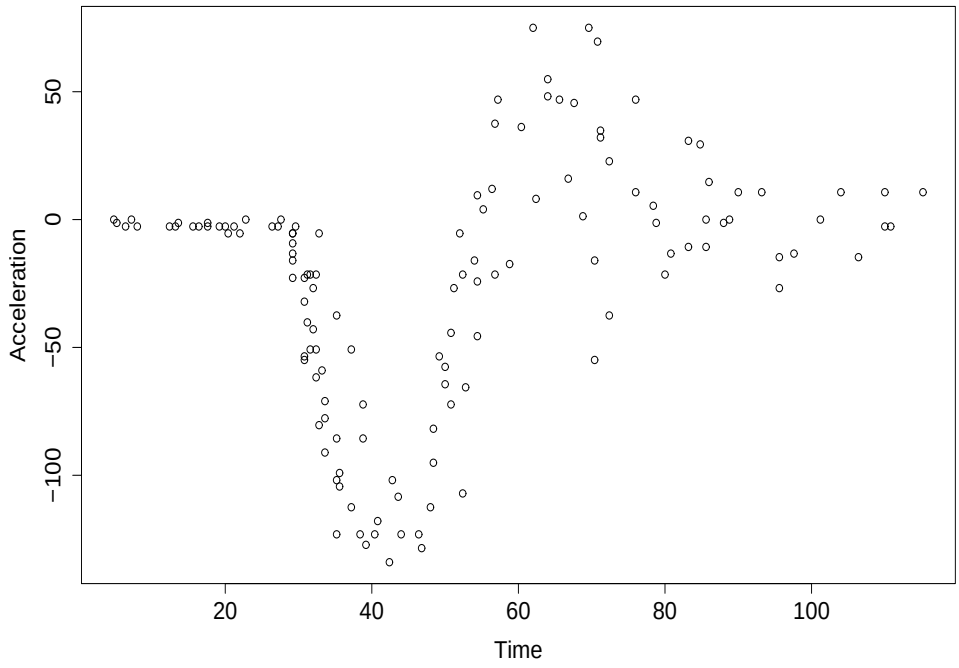

Figure 2.1: *Values of the acceleration (in* $9.81m/s^2$*) taken through time (in milliseconds) in an experiment on the efficacy of helmets for a motor-cycle crash.*

`R` (R Development Core Team, 2008), represent the values of the acceleration taken through time in an experiment related to the efficacy of helmets in a motor-cycle crash. A full description of these data is available in Silverman (1985) and Schmidt et al. (1981). We want to summarize the variation of $\boldsymbol{y}$ (acceleration) with respect to $\boldsymbol{x}$ (time). The simplest case of regression to accomplish this task assumes that the data have been generated according to the model

$$y_i = P(x_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0,\ \sigma^2), \quad i = 1, ..., n, \tag{2.1}$$

where

$$P(x) = \alpha_0 + \alpha_1 x, \quad \alpha_0, \alpha_1 \in \mathbb{R}, \tag{2.2}$$

is the linear predictor, ε_i is the noise, and the symbol $\mathcal{N}$ refers to the normal distribution. The black line in Figure 2.2 shows the fitted line and clearly, this line does not suit the data. A straightforward improvement can be obtained by changing the form of the linear predictor to some p-degree polynomial, ie,

$$P(x) = \sum_{r=0}^{p} \alpha_r x^r, \quad \alpha_0, ..., \alpha_p \in \mathbb{R}\cdot \tag{2.3}$$

The fitted curves corresponding to different values of p are shown in Figure 2.2. Although the polynomial approach can work well for some data, it suffers from various drawbacks due to the global dependence of polynomials on local properties of the data (de Boor, 1978). In other words, an individual observation can exert an unexpected influence on remote parts of the fitted predictor, and such a behaviour often yields an unstable predictor with poor interpolation properties, as illustrated in Figure 2.2.

An attractive way of circumventing this difficulty is found in nonparametric or smoothing methods. On the smoothing platform, the shape of the predictor is not predetermined by the model as in the parametric approach (2.2) or (2.3), but is driven by the data. This allows us to capture important local features in the data while still enforcing smoothness overall. Hence, in smoothing settings, model (2.1) is relaxed to

$$y_i = \mathcal{S}(x_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0,\ \sigma^2), \quad i = 1, ..., n, \tag{2.4}$$

for some unrestricted smooth function $\mathcal{S}(\cdot)$, which we will often refer to as the predictor or smoother.

Techniques for estimating $\mathcal{S}(\cdot)$ are available in many flavours including local averaging, kernel smoothing, smoothing splines, penalized splines, etc. These techniques are often classified according to the number of parameters that need to be estimated. From this perspective, a smoothing procedure is said to be of *full rank* if the number of parameters is at least equal to the number of data points n; otherwise it belongs to the *low rank* paradigm.

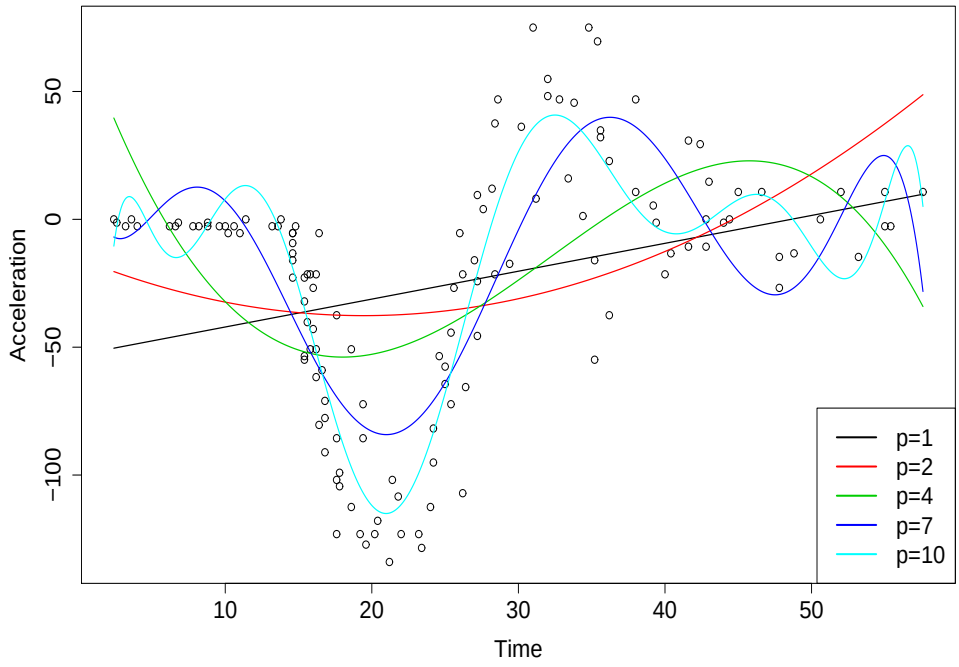

Figure 2.2: *Fitted polynomial curves of degree p to the motor-cycle data.*

2.2 Full rank smoothing

Although the main focus of this thesis is on low rank smoothing, we shall first give a brief overview of some common full rank methods.

2.2.1 Local averaging

One of the simplest full rank procedures is the running mean, also known as the moving average. Hastie and Tibshirani (1999, chap 2) and Keele (2008, sect 2.1) provide excellent descriptions of this approach. Here the central idea is to estimate the predictor $\mathcal{S}(\cdot)$ at each covariate point by the average of its nearest neighbours; ie, the estimate $\hat{\mathcal{S}}(x_i)$ of $\mathcal{S}(\cdot)$ at x_i is obtained by averaging the response values corresponding to the covariate values close to x_i. A important choice that the user has to make here is the structure of the *neighbourhood*, and one simple way consists of taking the target data point itself together with some k points on its left and on its right; clearly it would not be possible to have k points when approaching the endpoints, in which case we take as many points as possible. This leads to the general

expression

$$\hat{\mathcal{S}}(x_i) = \frac{\sum_{j=\max\{i-k,1\}}^{\min\{i+k,n\}} y_j}{\min\{i+k,n\} - \max\{i-k,1\} + 1}. \tag{2.5}$$

With this approach, the overall appearance of the smoother $\hat{\mathcal{S}}(\cdot)$ is governed by the

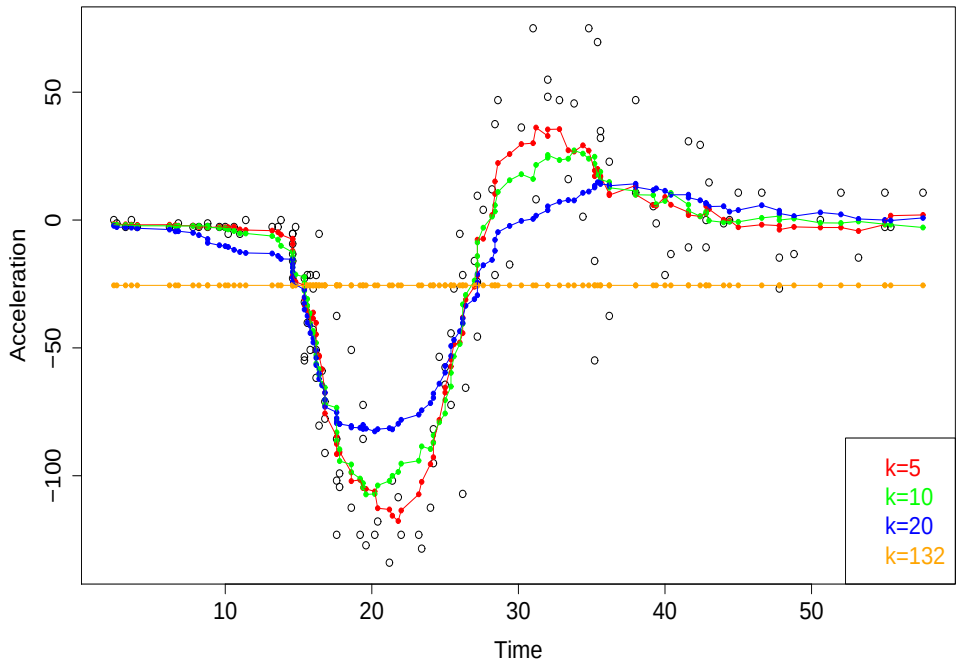

Figure 2.3: *Moving average corresponding to different values of neighbouring parameter k.*

relevant neighbourhood parameter k, in the sense that the larger the values of k, the more data participate, and so the smoother the estimate $\hat{S}(\cdot)$ will be. For instance, on the one extreme, if $k = 0$, then the neighbourhood of each point is reduced to that point alone, and so expression (2.5) yields $\hat{\mathcal{S}}(x_i) = y_i$; ie we have interpolated the data. On the other extreme, if $k \geq n - 1$, then all data are neighbours in which case expression (2.5) reduces to $\hat{\mathcal{S}}(x_i) = \sum_{j=1}^{n} y_j/n$, the standard data average. An illustration of this procedure is shown in Figure 2.3 for different values of k; on this graphic the individual local averages have been connected by a line.

The great advantage of the running mean is its simplicity; however it suffers from

serious drawbacks. For instance, the resulting smoother fluctuates across the data range (as shown in Figure 2.5) in such a way that it may hardly deserve the name smoother, and this approach performs poorly at the endpoints as discussed and illustrated in Hastie and Tibshirani (1999, chap 2). In addition to these problems, this method only provides a set of local averages at the observed values of the covariate; it is not able to produce direct estimates of $\mathcal{S}(\cdot)$ at unobserved data points, and therefore it is not suitable for prediction.

2.2.2 Kernel smoothing and local polynomials

Most of the problems with the running mean methodology as described above arise from the fact that all points in the relevant neighbourhood have the same weight and points outside are assigned zero weights. Kernel smoothing refines this approach and allows the evaluation of $\mathcal{S}(\cdot)$ at unobservable data points. Well-known references about kernel smoothing include Wand and Jones (1995) and Bowman and Azzalini (1997). In the kernel approach, all data have their "say" on the estimation of $\hat{\mathcal{S}}(\cdot)$ at any target point, but their contributions are modulated by suitably chosen weights "that decrease in a smooth fashion as one moves away from the target point" (Hastie and Tibshirani, 1999, pg 18). Here the weight $w_i(x)$ of the i th observation at a target point x usually takes the form

$$w_i(x) = K\left(\frac{x_i - x}{h}\right) \tag{2.6}$$

where $K(\cdot)$ is a predetermined function known as the *kernel function*, and h is a positive number known as the *bandwidth*. In practice, the kernel function is chosen to be symmetrical and in such a way that it attaches the greatest weights to observations that are closer to the target point and lesser weights to those that are further away; an example of such a kernel is the density of the standard normal distribution. The bandwidth quantifies the influence of the data points, and its choice turns out to be of primary importance as far as the smoothness appearance of $\hat{S}(\cdot)$ is concerned (Wand and Jones, 1995).

With these weights, one possibility is to update the simple local average given in

(2.5) to the weighted one as

$$\hat{\mathcal{S}}(x) = \frac{\sum_{i=1}^{n} w_i(x)\, y_i}{\sum_{i=1}^{n} w_i(x)}; \tag{2.7}$$

but in terms of bias reduction, one is usually interested in other estimators than the (weighted) mean. From this prospect, Cleveland (1979) proposed a local regression method, and this approach has been investigated intensively in the book by Bowman and Azzalini (1997). Essentially, this approach consists of estimating the smoother $\mathcal{S}(\cdot)$ by

$$\hat{\mathcal{S}}(x) = \hat{\alpha}_{x,0} + \hat{\alpha}_{x,1} x, \tag{2.8}$$

where the $(\hat{\alpha}_{x,0}, \hat{\alpha}_{x,1})$ minimise the weighted residual sum of squares given by

$$\sum_{i=1}^{n} w_i(x) \left[y_i - \alpha_{x,0} - \alpha_{x,1} x_i \right]^2 .$$

The green line in Figure 2.4 shows the fitted smoother from this approach with normal kernel and bandwidth set to 2, using the `sm.regression` function from the library `sm` in `R`. More generally, weighted local polynomial regressions can be considered as described by Cleveland (1979), although polynomials of degree higher that two do little in enhancing the estimated smoother (Keele, 2008, sect 2.2). For a detailed illustration of the benefits of local linear regressions over local averaging as well as their theoretical properties, we refer the reader to Fan and Gijbels (1992) and Fan (1993).

2.2.3 Smoothing splines

Before outlying the smoothing spline procedure, we give the formal definition of a spline.

Definition 2.1 (Dierckx, 1996)**.** A function $b(\cdot)$ defined on a finite interval $[a_1, a_2]$, is called a spline of degree p, having as knots the strictly increasing sequence $a_1 = \kappa_0 < \kappa_1 < \cdots < \kappa_q < \kappa_{q+1} = a_2$, if the following two conditions are satisfied:

(1) On each interval $[\kappa_l,\, \kappa_{l+1}]$, $b(\cdot)$ is given by a polynomial of degree p at most.

(2) The function $b(\cdot)$ and its derivatives up to order $(p-1)$ are all continuous at

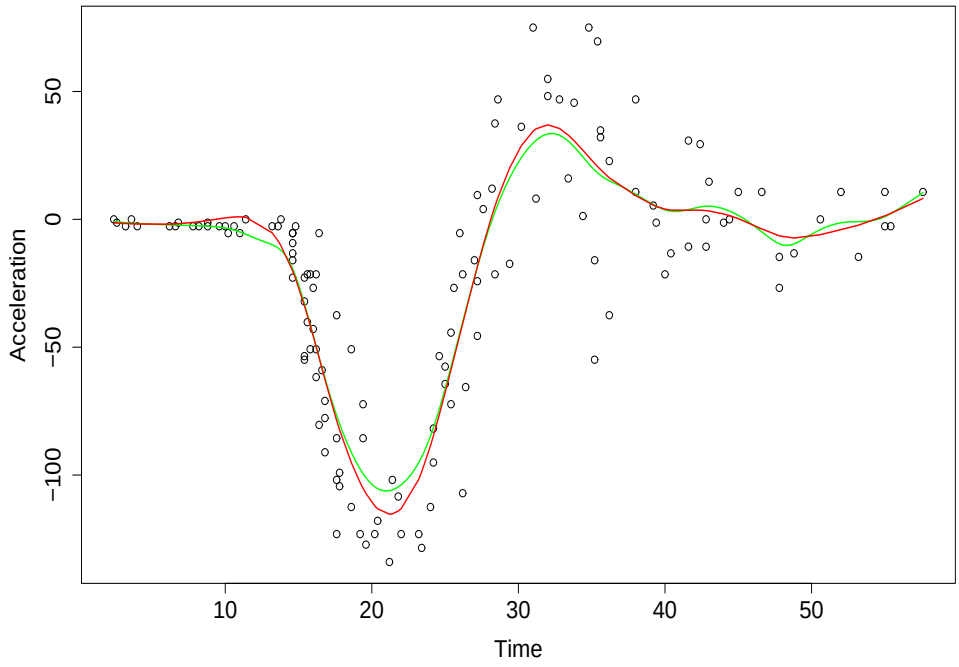

Figure 2.4: *Kernel smoother (green) with normal kernel and bandwidth* 2 *using the* `sm.regression` *function from the library* `sm` *in* `R`. *Smoothing splines (red) using the* `smooth.spline` *function in* `R` *with smoothing parameter selected by* GCV.

the knots.

Setting $p = 3$ and imposing the constraint $b''(a_1) = b''(a_2) = 0$ in this definition, we get a well known special case of splines called natural cubic splines. Technical references about spline functions include de Boor (1978) and Dierckx (1996).

Smoothing splines are constructed by introducing a *roughness* term to ensure that the estimate $\hat{\mathcal{S}}(\cdot)$ is not only determined by its goodness of fit to the data as quantified by the residual sum of squares, but also by the roughness term. A full description of this approach can be found in the book by Green and Silverman (1995), and there, a key point is the construction of the roughness term. Given that two functions that differ only by a linear term are usually seen as having the same level of roughness, and assuming that our smoother $\mathcal{S}(\cdot)$ is twice differentiable, we find that a natural measure of smoothness is defined by the integral $\int_{a_1}^{a_2} [\mathcal{S}''(x)]^2 dx$, which is often referred to as a *roughness penalty*. With this specification, the estimate of our smoother $\mathcal{S}(\cdot)$

can be chosen as the minimizer of

$$\sum_{i=1}^{n}(y_i - \mathcal{S}(x_i))^2 \quad + \quad \lambda \int_{a_1}^{a_2} [\mathcal{S}''(x)]^2 dx \tag{2.9}$$

among all twice differentiable functions on $[a_1, a_2]$, where λ is some positive real number. The amount of smoothing is controlled by λ which is therefore called the *smoothing parameter*.

For a given λ, it can be shown that the estimate of $\mathcal{S}(\cdot)$ that minimizes (2.9) is a natural cubic spline with knots at $x_1, ..., x_n$; Green and Silverman (1995). An illustration of the corresponding fitted smoother (produced using the `smooth.spline` function in `R`) is shown by the red colour in Figure 2.4; here the smoothing parameter has been chosen by minimizing the *Generalized Cross Validation* (GCV), which will be described in Section 2.4.2.

2.3 Penalized splines

Although the kernel approach and smoothing splines produced a satisfactory smoother for our data as shown in Figure 2.4, these full rank procedures can be computationally intensive since they require the estimation of (at least) as many parameters as the number of data points. The low rank approach ameliorates this potential difficulty. From now on, we shall concentrate on low rank methods, specifically penalized splines.

The idea of penalized splines is similar in spirit to that of smoothing splines in the sense that a penalty is designed to achieve smoothing. In penalized splines, the predictor $\mathcal{S}(\cdot)$ is expressed in terms of a linear combination of spline basis functions, ie,

$$\mathcal{S}(x_i) = \sum_{r=1}^{c} \alpha_r \, b_r(x_i) \tag{2.10}$$

where the $b_r(\cdot)$ are spline basis functions and the α_r are associated coefficients. Whatever basis one uses, the predictor (2.10) can be expressed in matrix form as

$$\mathcal{S}(\boldsymbol{x}) = \boldsymbol{\Omega}\boldsymbol{\alpha}, \tag{2.11}$$

where $\mathcal{S}(\boldsymbol{x})$ is the vector obtained by the element-wise action of $\mathcal{S}(\cdot)$ on the covariate

vector $\boldsymbol{x}$, $\boldsymbol{\Omega}$ is the $n \times c$ regression matrix of splines with $b_1(x_i)$ to $b_c(x_i)$ in row i, and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_c)'$ is the vector of spline coefficients.

With this representation, one can bring about the smoothness of the predictor $\mathcal{S}(\cdot)$ through parsimonious selection of the basis functions and careful knot arrangement (Friedman and Silverman, 1989). Alternatively, the penalized spline approach allows the user to relax concerning the exact number of knots or splines. The idea is to achieve smoothness through judicious constraints. Within this framework, standard constraints are obtained by requiring that a suitable combination $\mathcal{P}(\boldsymbol{\alpha})$ of the spline coefficients $\boldsymbol{\alpha}$ does not exceed some well chosen threshold; here $\mathcal{P}(\cdot)$ is some well chosen *penalty function*. Minimizing the residual sum of squares under this requirement is equivalent to the minimization of the so-called *penalized residual sum of squares* (PRSS) given by

$$\mathsf{PRSS} \;=\; \|\boldsymbol{y} - \boldsymbol{\Omega\alpha}\|^2 \;+\; \lambda \times \mathcal{P}(\boldsymbol{\alpha}). \tag{2.12}$$

In this expression, λ represents the smoothing parameter and quantifies the balance between the goodness of fit as measured by the residual sum of squares and the smoothness as measured by the penalty term $\mathcal{P}(\boldsymbol{\alpha})$. With the penalty functions used in this thesis, the penalty component $\lambda \times \mathcal{P}(\boldsymbol{\alpha})$ in (2.12) can be expressed in matrix form as

$$\lambda \times \mathcal{P}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha}, \tag{2.13}$$

in which case we shall refer to $\boldsymbol{P}$ as the *penalty matrix*. Hence, on minimizing this PRSS (2.12) with respect to $\boldsymbol{\alpha}$, one obtains the estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ as

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'\boldsymbol{y}. \tag{2.14}$$

That is, the smoother $\mathcal{S}(\boldsymbol{x})$ is estimated by

$$\hat{\mathcal{S}}(\boldsymbol{x}) = \boldsymbol{H}\boldsymbol{y}, \tag{2.15}$$

where $\boldsymbol{H}$ is given by

$$\boldsymbol{H} = \boldsymbol{\Omega}(\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'. \tag{2.16}$$

Since $\boldsymbol{H}$ maps the data vector $\boldsymbol{y}$ into the fitted smoother, it is usually refers to as the hat matrix or the smoother matrix.

In practice, the structure of the penalty function $\mathcal{P}(\cdot)$ (equivalently the form of the penalty matrix $\boldsymbol{P}$) is driven by the spline bases used. There is a large pool of spline basis functions that one can choose from, the most popular being truncated polynomials and B-splines. In the next sections, we shall outline how the above procedure applies to these two bases. But prior to that, it is important to mention that although the penalization diminishes the effect of the exact number of spline basis functions, a sufficiently "rich" basis is required, since a small basis can result in a predictor which is not flexible enough to capture the observed variability in the data. In view of this, a simulation study carried out by Ruppert (2002) in the context of penalized splines defined in terms of truncated lines bases suggests that working with approximately $\max\{4,\ \min\{n/4,\ 40\}\}$ basis functions provides satisfactory results.

Another essential point that needs clarification is the location of the knots. Although sophisticated algorithms have been proposed (see for example Yao and Lee, 2008), the two popular schemes used in practice are (i) equi-spaced knots as advised by Eilers and Marx (1996) in penalized splines via B-splines, and (ii) knot locations based on the quantiles of the covariate as advocated by Ruppert et al. (2003) in penalized splines with truncated polynomials. Recently however, Eilers and Marx (2010) suggested that in both smoothing approaches (ie, with truncated polynomials or B-splines bases), equi-spaced knots are to be preferred to quantile-based knots. Hence, we use equi-spaced knots throughout this thesis.

2.3.1 Penalized truncated polynomials

Smoothing with penalized truncated polynomials, as advocated by Ruppert et al. (2003), is a direct extension of polynomial regression. Here, instead of relying on high degree polynomials, important local features in the data are captured by adding truncated functions to a (low degree) polynomial basis. In this setting, the basis functions $b_r(\cdot)$ in (2.10) are expressly defined by

$$b_r(x) = \begin{cases} x^{r-1} & \text{for } r = 1, ..., p+1 \\ T_{r-p-1}(x) & \text{for } r = p+2, ..., c \end{cases} \tag{2.17}$$

where the $T_l(\cdot)$ are given by

$$T_l(x) = [(x - \kappa_l)_+]^p, \quad \text{with } u_+ = \max\{0,\, u\}\,. \tag{2.18}$$

In (2.18), $\kappa_1, ..., \kappa_{c-p-1}$ are equi-spaced knots, and we will denote by δ the distance between κ_l and κ_{l+1}, $l = 1, ..., c-p-2$. We refer to (2.18) as the *truncated polynomial basis* and to (2.17) as the *full truncated polynomial basis*. An illustration of these bases is shown in Figure 2.5 for $p = 1$ and $p = 3$. With this basis specification, the

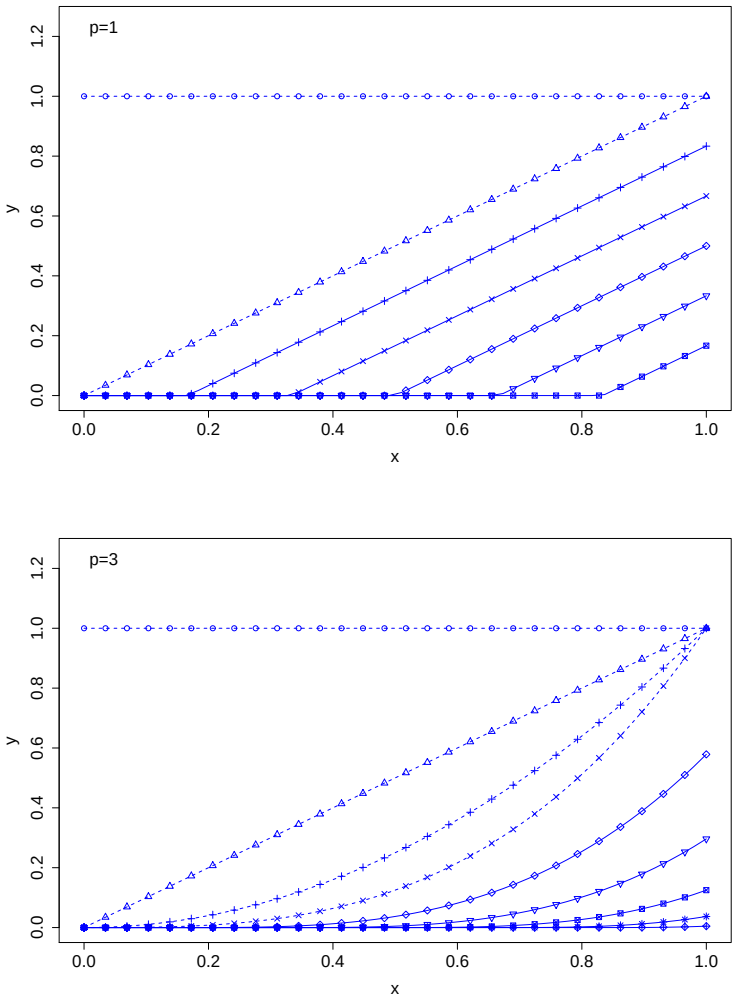

Figure 2.5: *Full truncated polynomial bases of degrees* $p = 1, 3$.

predictor becomes

$$\mathcal{S}(x_i) = \sum_{r=1}^{p+1} \alpha_r x_i^{r-1} + \sum_{r=p+2}^{c} \alpha_r T_{r-p-1}(x_i)\cdot \tag{2.19}$$

Setting $\boldsymbol{a} = vec(\alpha_1, ..., \alpha_{p+1})$ and $\boldsymbol{\xi} = vec(\alpha_{p+2}, ..., \alpha_c)$ yields the matrix representation

$$\mathcal{S}(\boldsymbol{x}) = \boldsymbol{X}_p \boldsymbol{a} + \boldsymbol{T}_p \boldsymbol{\xi}, \tag{2.20}$$

where the regression matrices $\boldsymbol{X}_p$ and $\boldsymbol{T}_p$ are defined as

$$\boldsymbol{X}_p = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p \end{bmatrix} \quad \text{and} \quad \boldsymbol{T}_p = \begin{bmatrix} T_1(x_1) & \cdots & T_{c-p-1}(x_1) \\ \vdots & \ddots & \vdots \\ T_1(x_n) & \cdots & T_{c-p-1}(x_n) \end{bmatrix}\cdot \tag{2.21}$$

The red line in Figure 2.6 shows the MLE of the predictor (2.20) using second degree full truncated polynomial basis with 40 equi-spaced knots. The more we increase the number of knots the wiggler the fitted predictor becomes. This roughness problem can be smoothed away by appropriate penalization/constraints. With a full truncated polynomial basis of degree p, the standard smoothness constraint is obtained by penalizing the jumps in the p-order derivative of the predictor at every knot. From (2.19) and (2.20), it is straightforward to see that this jump at the knot κ_r is equal to $p!\xi_r$; ie, the jumps are proportional to the corresponding coefficients, and so the standard penalty function used for penalized truncated polynomials is

$$\mathcal{P}(\boldsymbol{\alpha}) = \|\boldsymbol{\xi}\|^2 = \boldsymbol{\xi}'\boldsymbol{\xi}\cdot \tag{2.22}$$

Hence, fitting the model expressed in terms of truncated polynomial reduces to a special case of (2.14), where the regression matrix $\boldsymbol{\Omega}$, the coefficients $\boldsymbol{\alpha}$, and the penalty matrix $\boldsymbol{P}$ are given by

$$\boldsymbol{\Omega} = [\boldsymbol{X}_p : \boldsymbol{T}_p], \quad \boldsymbol{\alpha} = vec(\boldsymbol{a}, \boldsymbol{\xi}), \quad \boldsymbol{P} = \lambda \times blockdiag(\boldsymbol{0}_{(p+1)\times(p+1)}, \boldsymbol{I}_{c-p-1})\cdot \tag{2.23}$$

In these expressions, $\boldsymbol{0}_{r\times r}$ is the $s \times r$ matrix of zeros, $\boldsymbol{I}_s$ is the $s \times s$ identity matrix, and *blockdiag* is the block diagonal operator. The fitted curve from this approach is

shown by the blue line in Figure 2.7; here the smoothing parameter has been selected using the *Akaike Information Criterion* (AIC) which will be described in Section 2.4.2.

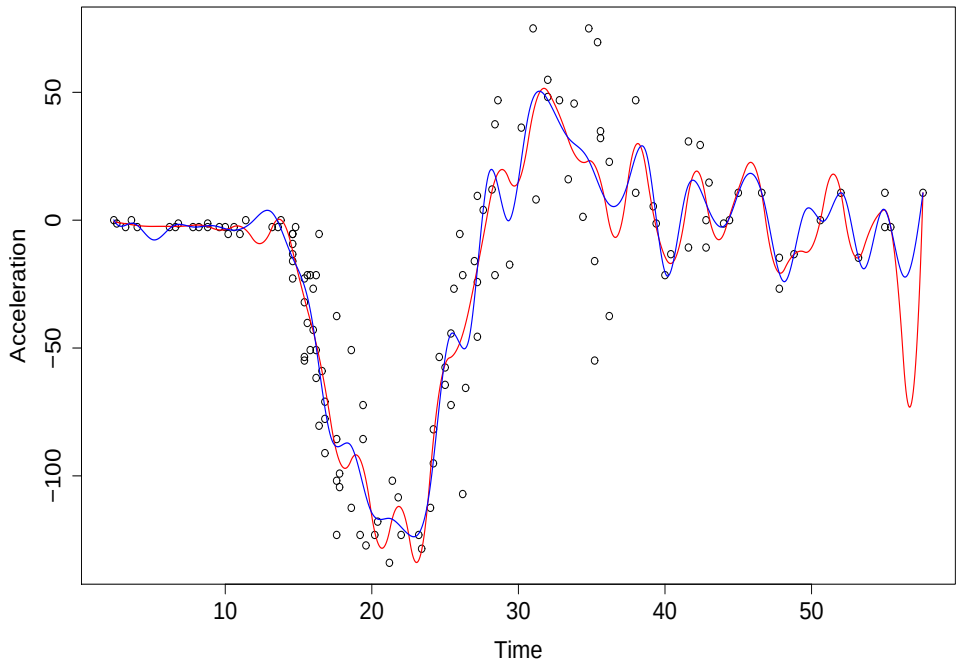

Figure 2.6: *Unpenalized predictor (using 40 equi-spaced knots) respectively with a quadratic full truncated polynomial basis (red), and a cubic B-spline basis (blue).*

2.3.2 Penalized B-splines

Smoothing by penalized B-splines (also known as P-splines) was developed by Eilers and Marx (1996). In this smoothing approach, the spline basis functions $b_r(\cdot)$ in (2.10) are replaced by B-splines, which we denote by $B_r(\cdot)$. Details on B-spline bases are available in de Boor (1978) and Dierckx (1996). Essentially, a B-spline basis of degree p can be defined either recursively using B-splines of lower degrees, or through differences of truncated polynomial functions of the same degree (see Eilers and Marx, 2010) as

$$B_r(x) = \frac{(-1)^{p+1}\ \Delta_{p+1} T_r(x)}{p!\ \delta^p}, \tag{2.24}$$

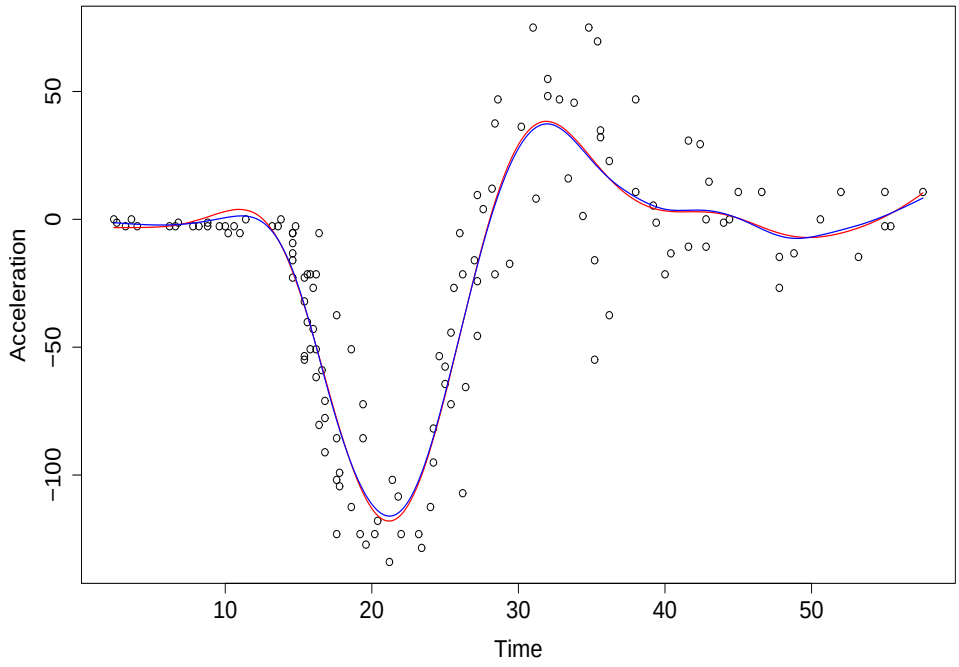

Figure 2.7: *Penalized quadratic truncated polynomials (red), and penalized cubic B-splines with second order difference penalty (blue), with smoothing parameter chosen by* AIC.

where the $T_r(\cdot)$ represent the truncated polynomial functions of degree p as described in Section 2.3.1, δ is the distance between two successive knots, and Δ is the difference operator defined recursively by

$$\Delta_1 T_r = T_r - T_{r-1} \text{ and } \quad \Delta_{d+1} T_r = \Delta_1(\Delta_d T_r)\cdot \tag{2.25}$$

An illustration of a basis of B-splines is shown in Figure 2.8. As we can see, this is a local basis in the sense that the support of each B-spline is compact; precisely, a B-spline of degree p is positive on a domain spanned by $(p+2)$ knots and it is zero everywhere else. With a B-spline basis, the predictor is expressed as

$$\mathcal{S}(x_i) = \sum_{r=1}^{c} \alpha_r \, B_r(x_i), \tag{2.26}$$

or compactly as

$$\mathcal{S}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\alpha}, \tag{2.27}$$

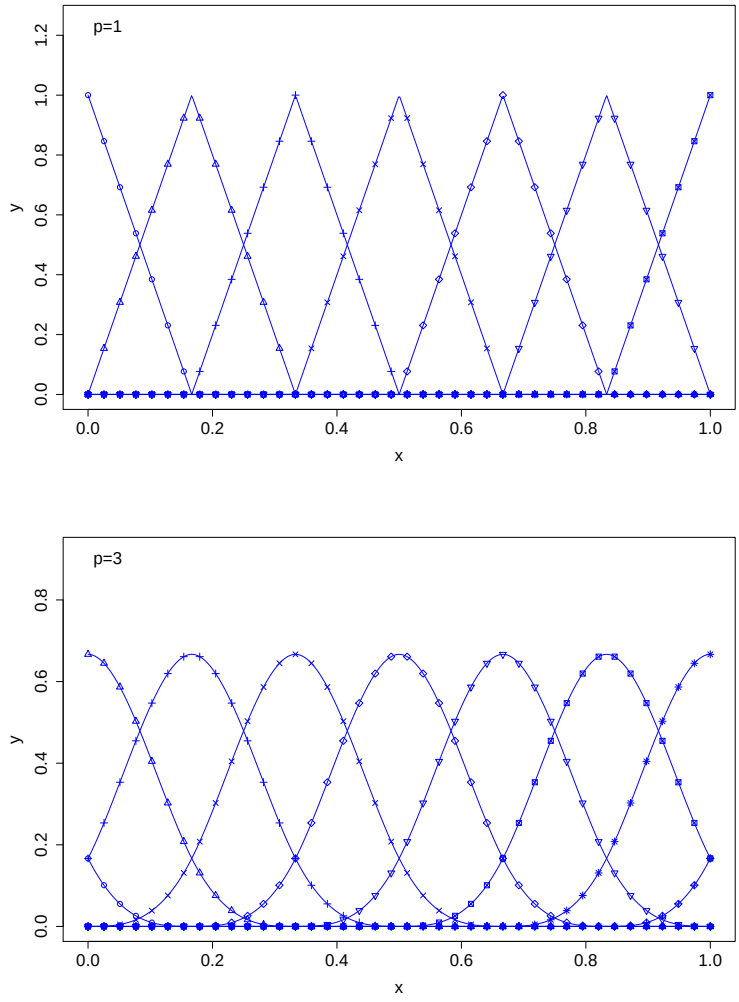

Figure 2.8: *B-spline bases of degree* $p = 1, 3$.

where $\boldsymbol{B}$ is the matrix of B-splines defined by

$$\boldsymbol{B} = \begin{bmatrix} B_1(x_1) & \cdots & B_c(x_1) \\ \vdots & \ddots & \vdots \\ B_1(x_n) & \cdots & B_c(x_n) \end{bmatrix} \tag{2.28}$$

and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_c)'$ now represents the vector of B-spline coefficients.

The blue line in Figure 2.6 shows the MLE estimate of the predictor (2.26) using a basis of cubic B-splines with 40 equi-spaced knots. In this case, the wiggliness problem is addressed by penalizing the differences in adjacent B-spline coefficients,

yielding the penalty function

$$\mathcal{P}(\boldsymbol{\alpha}) \;=\; \sum_r (\Delta_d \alpha_r)^2 = \boldsymbol{\alpha}'(\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d)\boldsymbol{\alpha}, \tag{2.29}$$

where $\boldsymbol{\Delta}_d$ is the $c \times (c-d)$ differencing matrix of order d. Hence, the model expressed in terms of B-splines reduces to a special case of (2.14), with the regression matrix $\boldsymbol{\Omega}$ and the penalty matrix $\boldsymbol{P}$ given by

$$\boldsymbol{\Omega} = \boldsymbol{B}, \;\; \boldsymbol{P} = \lambda \times \boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d. \tag{2.30}$$

In practice, the second order ($d = 2$) difference is frequently used in which case we have

$$\boldsymbol{\Delta}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 & -2 & 1 \end{bmatrix}. \tag{2.31}$$

An illustration of the fitted predictor from this approach is shown by the blue line in Figure 2.7.

2.3.3 Penalized truncated polynomials versus penalized B-splines: similarities and differences.

Following Eilers and Marx (2010), we shall use the notation PT-splines for penalized splines based on a full truncated polynomial basis, and PB-splines for penalized splines with a basis of B-splines. In both approaches, a rich basis together with an appropriate penalty are used to obtain the estimated smoother. As described in Section 2.3.1, the PT-spline approach is a direct extension of polynomial regression and for this reason, it is often seen as a simple introduction to smoothing. Also, the partition of the resulted predictor into an unpenalized term, $\boldsymbol{X}_p\boldsymbol{a}$, and a penalized component, $\boldsymbol{T}_p\boldsymbol{\xi}$, as displayed in equation (2.20) allows a direct implementation of the model with standard mixed model tools upon model specification; details on this implementation are given in the Appendix of the book by Ruppert et al. (2003).

However, as discussed by Eilers and Marx (1996, 2010), PB-splines offer several advantages. First, the B-spline coefficients are meaningful in the sense that the behaviour of the predictor $\boldsymbol{B\alpha}$ is very similar to that of the behaviour of the components of $\boldsymbol{\alpha}$; ie, the extent to which the B-spline coefficients α_j are smooth is in close agreement with the smoothness of the predictor $\boldsymbol{B\alpha}$. Second, the fact that B-splines are local functions provides PB-splines with excellent numerical properties. Third, the degree of the B-splines and the differencing order of the penalty can be chosen independently; this provides the modeller with additional flexibility. Fourth, PB-splines can be extended in a straightforward manner to cope with multidimensional problems, as we shall see in Chapter 5. For a full discussion and illustration of these properties, we refer the reader to Eilers and Marx (2010), Kirkby (2009), Currie et al. (2004). In addition to these benefits, PB-splines can also be partitioned into a penalized and an unpenalized component as

$$\boldsymbol{B\alpha} = \tilde{\boldsymbol{X}}_d\boldsymbol{\alpha}_1 + \tilde{\boldsymbol{T}}_d\boldsymbol{\alpha}_2, \quad \text{with } \tilde{\boldsymbol{X}}_d = \boldsymbol{X}_{d-1} \text{ and } \tilde{\boldsymbol{T}}_d = \boldsymbol{B}\boldsymbol{U}_d\boldsymbol{\Lambda}_d^{-1/2}. \tag{2.32}$$

In these expressions, $\boldsymbol{\Lambda}_d$ is the diagonal matrix formed by the positive eigenvalues of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$, $\boldsymbol{U}_d$ is the matrix whose columns are the eigenvectors corresponding to these positive eigenvalues, $\boldsymbol{\alpha}_1$ is a d-length vector containing the unpenalized coefficients, and $\boldsymbol{\alpha}_2$ is a coefficient vector subject to the transformed penalty matrix $\tilde{\boldsymbol{P}} = \lambda\boldsymbol{I}_{c-d}$. Additional details on the re-parametrisation of PB-splines can be found in Currie et al. (2006), and these re-parametrisations give access to the estimation of PB-splines with the standard mixed model tool as well.

Hence, PT-splines and PB-splines can be seen as mixed models, and in both cases, the smoother $\hat{\mathcal{S}}(\cdot)$ converges to the underlying unpenalized polynomial of degree $p = d-1$, as $\lambda \longrightarrow \infty$. However, such a re-formulation of smooth models as mixed models is not without controversy. Green (1999) commented on the Verbyla et al. (1999) paper as follows: "Formulating spline smoothing as a mixed model is simply a mathematical device; the suggested logical distinction between the fixed linear trend and the random smooth variation is artificial". According to Djeundje and Currie (2010a), Green's point is that "the randomness in the mixed model representation of smoothers is not assigned to units in a clear way as in the mixed models described in Searle et al. (2006, chap 1). Thus, smoothers usually have a mixed model representation but not

a mixed model interpretation in the original sense. Nowadays, one motivation behind the insertion of smooth models into the mixed model framework is the availability of standard computer packages for mixed models". We shall return to these issues in Chapter 3.

2.4 Model selection

At first sight, a good model needs to fit the data well. Obviously, by focusing only on this fitting requirement, the model will often be too flexible or complex; indeed in the limit, the model may simply reproduce the data. It is a good idea to incorporate into model choice: (a) parsimony and (b) capability of predicting a new observation that has not been used in the fitting of the model. This makes the choice tricky because it requires an optimal approach that should take into account both the fitting and the parsimony, as a gain in one involves a loss in the other, and vice-versa.

From the PRSS in (2.12), it is clear that the smoothing parameter λ quantifies the trade-off between the fidelity to the data as measured by the residual sun of squares, and the smoothness of the fitted predictor as measured by a penalty term. Hence, the smoothing parameter plays a central role in the model specification and its choice falls in the bias-variance trade-off paradigm. Various procedures have been proposed for the selection of λ, and these procedures are connected through the concept of the *effective dimension* of the model.

2.4.1 Effective dimension

If $\lambda = 0$, then we are back to the standard linear model setting and the dimension (or degrees of freedom) of the model is the number of linearly independent columns in the regression matrix; in our case, since $\boldsymbol{\Omega}$ has c columns then the dimension of the model is c. With penalization the flexibility of the model is reduced and so the dimension of the model is correspondingly reduced. In this case, the degrees of freedom (which we denote by ν), also known as the *effective dimension* of the model, is measured by

the trace of the hat matrix $\boldsymbol{H}$ (see Ye, 1998; Ruppert et al., 2003, sect 9.3). That is,

$$\begin{aligned} \nu &= tr(\boldsymbol{H}) \\ &= tr[(\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'\boldsymbol{\Omega}] \\ &= c - tr[(\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{P}]. \end{aligned} \tag{2.33}$$

This final form displays the reduction in the dimension of the model brought about by the penalization, with $\lambda = 0$ simplifying to $\nu = c$; hence (2.33) extends the concept of degrees of freedom from standard linear regression to the smoothing setting. For the sake of clarity in this Section, we will often write $\boldsymbol{H}(\lambda)$ and $\nu(\lambda)$ to enforce the dependence of $\boldsymbol{H}$ and ν on λ.

In practice, in order to obtain ν, it is common to first multiply the two matrices in (2.33), and then compute the trace of the result. This matrix multiplication can be avoided using the following: if $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are two matrices such that $\boldsymbol{A}_1'$ and $\boldsymbol{A}_2$ have the same dimension, then $tr(\boldsymbol{A}_2\boldsymbol{A}_1)$ is equal to the sum of the entries of $\boldsymbol{A}_2 * \boldsymbol{A}_1'$, where the symbol "$*$" refers to element-by-element multiplication.

2.4.2 Choosing the smoothing parameter

Several methods/indicators have been proposed for the automatic selection of the smoothing parameter, the best known being the Akaike information criterion (Akaike, 1974), the Bayesian information criterion (Schwarz, 1978), the generalized cross validation (Craven and Wahba, 1979), and the restricted likelihood (Patterson and Thompson, 1971).

Let $f(\cdot)$ denotes the true (unknown) density from which the data are generated, and $\tilde{f}(\cdot|\boldsymbol{\alpha})$ the density of our model defined through (2.4) and (2.11). The Kulbeck-Leibler discrepancy between $f(\cdot)$ and $\tilde{f}(\cdot|\boldsymbol{\alpha})$ is given by the following expectation

$$\begin{aligned} \mathcal{I}\left(f, \tilde{f}\right) &= \int \left[\log\left(f(y)\right) - \log\left(\tilde{f}(y|\boldsymbol{\alpha})\right)\right] \times f(y)dy \\ &= \int \log\left(f(y)\right) \times f(y)dy - \int \log\left(\tilde{f}(y|\boldsymbol{\alpha})\right) \times f(y)dy, \end{aligned} \tag{2.34}$$

and "provides a measure of how badly $\tilde{f}(\cdot|\boldsymbol{\alpha})$ matches the truth" (Wood, 2006,

pg 112). In (2.34), the first term is a constant, and the second one is known as the expected log-likelihood. The larger this latter term, the closer our model density $\tilde{f}(\cdot|\boldsymbol{\alpha})$ approaches the truth. However, this second term depends on the unknown true distribution $f(\cdot)$. By using the Monte Carlo method given the data $\boldsymbol{y} = (y_1, ..., y_n)'$, we can estimate this expected log-likelihood by

$$\frac{1}{n}\sum_{i=1}^{n} \log(\tilde{f}(y_i|\boldsymbol{\alpha})) = \frac{\ell(\boldsymbol{\alpha})}{n} \tag{2.35}$$

where $\ell(\boldsymbol{\alpha})$ is the log-likelihood of our model.

Thus, if $T_{\boldsymbol{\alpha}}(\boldsymbol{y})$ is the MLE estimator of $\boldsymbol{\alpha}$, then $\ell(T_{\boldsymbol{\alpha}}(\boldsymbol{y}))$ would be the resulting estimator of $n\int \log\left(\tilde{f}(y|T_{\boldsymbol{\alpha}}(\boldsymbol{y}))\right) \times f(y)dy$. As shown in Konish and Kitagawa (2008, chap 3), this estimator is biased with the bias approximately equal to the degrees of freedom of the model. In the smoothing setting, we already know that the degrees of freedom is approximated by the effective dimension and the estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is a function of the smoothing parameter, ie, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\lambda)$. Hence, the following indicator of the Kulbeck-Leibler discrepancy

$$\mathsf{AIC}(\lambda) = -2\ell(\hat{\boldsymbol{\alpha}}(\lambda)) \ + \ 2\nu(\lambda), \tag{2.36}$$

is derived, and is known as the *Akaike Information Criterion* (AIC).

The *Bayesian Information Criterion* (BIC), also called Schwarz's information, is similar in form to the AIC. It is derived using Bayesian arguments and is given by

$$\mathsf{BIC}(\lambda) = -2\ell(\hat{\boldsymbol{\alpha}}(\lambda)) \ + \ \log(n) \times \nu(\lambda). \tag{2.37}$$

Note that in both AIC and BIC, the effective dimension can be seen as a penalty term against the deviance term. For $n > 7$, the coefficient of this penalty term is larger in the BIC than in the corresponding AIC, and so the BIC will have a stronger preference for simpler/smoother models compared to those chosen by AIC.

We now turn to the generalized cross validation. For a given value of the smoothing parameter λ, let $\hat{\mathcal{S}}_{\lambda}^{(-k)}(\cdot)$ denote the estimated predictor based on the reduced data $\{(x_i, y_i),\ i \neq k\}$; then the overall efficacy of the model in terms of its ability to predict

an observation at random from the observed data can be measured by

$$\mathsf{CV}(\lambda) = \sum_{k=1}^{n} (y_k - \hat{\mathcal{S}}_{\lambda}^{(-k)}(x_k))^2/n, \tag{2.38}$$

known as the *cross validation* statistic. Clearly, a direct computation of the CV as displayed in (2.38) requires the estimation of n predictors $\hat{\mathcal{S}}_{\lambda}^{(-k)}(\cdot)$, $k = 1, ..., n$; this can make the optimization of the CV computationally heavy. Fortunately, it is shown in Green and Silverman (1995) that the CV reduces to

$$\mathsf{CV}(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mathcal{S}}_{\lambda}(x_i)}{1 - H_{ii}(\lambda)} \right)^2 /n, \tag{2.39}$$

where $\hat{\mathcal{S}}_{\lambda}(\cdot)$ is the estimated predictor based on the full data, and $H_{ii}(\lambda)$ is the i th diagonal element of the hat matrix $\boldsymbol{H}(\lambda)$. The *Generalized Cross Validation* (GCV) is obtained by replacing the diagonal elements $H_{ii}(\lambda)$ in (2.39) with their average; that is

$$\mathsf{GCV}(\lambda) = n \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mathcal{S}}_{\lambda}(x_i)}{n - \nu(\lambda)} \right)^2 . \tag{2.40}$$

For a full and coherent description of AIC, BIC and GCV we refer the reader to Konish and Kitagawa (2008). We postpone the description of the restricted likelihood to the more appropriate context in Chapter 4.

In practice, the choice of any of these criteria is problem-driven, and the appropriate value of the smoothing parameter is usually selected through an optimization (over positive values of the smoothing parameter) of the relevant criterion. However, such an optimization needs to be performed with caution since the starting values may turn out to be critical due to the multi-modal structure of the criterion, as investigated by Welham and Thompson (2009). From the same prospect, work carried out by Reiss and Ogden (2009) suggests that in some situations, optimization at the low values of the smoothing parameter may reflect random fluctuations that are unrelated to the fidelity-smoothness trade-off. In such cases, these last authors advise a graphical examination of the chosen criterion over plausible values of the smoothing parameters, rather than a blind/automatic choice of the optimal value. It may also be very helpful to compare the results obtained from several criteria.

2.5 Precision

An essential aid to the interpretation of the estimated predictor $\hat{\mathcal{S}}(\cdot)$ is its precision. In the spirit of Wahba (1983), by imposing a penalty on the coefficient, we are imposing some *prior* smooth belief about the form of the predictor. Consequently, to compute confidence intervals, it is common to step into the Bayesian framework and then rely on the posterior distribution of the spline coefficients $\boldsymbol{\alpha}$. From this perspective, it can be shown (Lin and Zhang, 1999; Wahba, 1983) that the effect of the penalty $\boldsymbol{P}$ in (2.12) is equivalent to that of the following prior distribution on $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} \propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha}\right). \tag{2.41}$$

Combining this prior distribution with (2.4) and (2.10) yields the posterior distribution

$$\boldsymbol{\alpha}|\boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\alpha}},\ \sigma^2(\boldsymbol{\Omega}'\boldsymbol{\Omega}+\boldsymbol{P})^{-1}\right), \tag{2.42}$$

from which the approximate covariance of the fitted predictor is usually computed as

$$cov[\hat{\mathcal{S}}(\boldsymbol{x})] = \sigma^2\boldsymbol{\Omega}(\boldsymbol{\Omega}'\boldsymbol{\Omega}+\boldsymbol{P})^{-1}\boldsymbol{\Omega}'. \tag{2.43}$$

Once the estimate of the smoother has been obtained for a given smoothing parameter, we use the diagonal elements of this covariance matrix to derive the confidence band around the estimated smoother, with an estimate of σ^2 plugged in. In line with the familiar unbiased estimate of the variance parameter in linear regression, σ^2 is usually estimated by

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y}-\boldsymbol{\Omega}\hat{\boldsymbol{\alpha}}\|^2}{n-\nu}, \tag{2.44}$$

although with penalization this estimate is not unbiased (Wood, 2006, chap 4).

2.6 Penalized Generalized Linear Models

We have illustrated how penalized splines apply to the basic linear model, which assumes that the data are normally distributed. In many situations, this Gaussian assumption does not hold; the response variable may even be categorical rather than continuous. In the mortality context for example, a standard assumption is that

the death counts are Poisson distributed. Generalized Linear Models (GLM) is an extension of the basic linear model, which allows for a wide family of distributions called the exponential family (McCullagh and Nelder, 1989). In this family, the density function $f(y_i;\theta)$ for the observation y_i given a certain parameter θ is expressed in the simple form as

$$f(y_i;\theta) = \exp\left(y_i\, h_1(\theta) + h_2(\theta) + h_3(y_i)\right), \tag{2.45}$$

where $h_1(\cdot)$, $h_2(\cdot)$ and $h_3(\cdot)$ are known functions; $h_1(\theta)$ is often refer to as the natural parameter. Additionally, this density may contain other parameters, in which case they are called nuisance parameters forming parts of the functions $h_1(\cdot)$, $h_2(\cdot)$ and $h_3(\cdot)$, (Dobson, 1983, chap 3). It is easy to check that this family encompasses many distributions including the normal, Poisson, binomial, etc.

In the GLM setting, the mean is related to the predictor through a monotone and differentiable function, $g(\cdot)$, called the *link function*, as

$$g(\mu_i) = \mathcal{S}(x_i), \quad \text{with } \mu_i = \mathbb{E}[y_i],$$

ie,

$$g(\boldsymbol{\mu}) = \mathcal{S}(\boldsymbol{x}) = \boldsymbol{\Omega\alpha}. \tag{2.46}$$

The penalized generalized linear model (PGLM) refers to the GLM with smoothness constraints on the predictor function $\mathcal{S}(\cdot)$, and these constraints simplify to the action of the penalty matrix $\boldsymbol{P}$ on the spline coefficient $\boldsymbol{\alpha}$. With analogy to the PRSS in the penalized linear model presented in Section 2.3, the spline coefficient $\boldsymbol{\alpha}$ is estimated in the PGLM setting by optimizing the *penalized log-likelihood*, $\ell_P(\cdot)$, defined as

$$\ell_P(\boldsymbol{\alpha}|\lambda) = \ell(\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{P\alpha}, \tag{2.47}$$

where $\ell(\boldsymbol{\alpha})$ is the ordinary log-likelihood based on (2.45) and (2.46). If we denote by $\boldsymbol{W}$ the diagonal weight matrix with

$$w_{ii} = \frac{1}{[g'(\mu_i)]^2\, var(y_i)}, \quad i = 1, ..., n$$

on the diagonal, it can be shown (Dobson, 1983; Eilers and Marx, 1996) that this optimization yields

$$\boldsymbol{\Omega}'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{P}\boldsymbol{\alpha}, \tag{2.48}$$

which is usually solved iteratively via the penalized version of the *scoring algorithm* given by

$$(\boldsymbol{\Omega}'\tilde{\boldsymbol{W}}\boldsymbol{\Omega} + \boldsymbol{P})\hat{\boldsymbol{\alpha}} = \boldsymbol{\Omega}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{2.49}$$

where

$$\tilde{\boldsymbol{z}} = \boldsymbol{\Omega}\tilde{\boldsymbol{\alpha}} + (\boldsymbol{y}-\tilde{\boldsymbol{\mu}}) * g'(\tilde{\boldsymbol{\mu}}) \tag{2.50}$$

is the *working vector*. In these expressions, tilde refers to the current estimate and hat refers to the update.

At convergence, the predictor is estimated by

$$\hat{\mathcal{S}}(\boldsymbol{x}) = \boldsymbol{\Omega}(\boldsymbol{\Omega}'\hat{\boldsymbol{W}}\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'\hat{\boldsymbol{W}}\hat{\boldsymbol{z}} \tag{2.51}$$

from which the hat matrix is obtained (Ye, 1998; Ruppert et al., 2003) as

$$\boldsymbol{H} \approx \frac{\partial\,\hat{\mathcal{S}}(\boldsymbol{x})}{\partial\hat{\boldsymbol{z}}} = \boldsymbol{\Omega}(\boldsymbol{\Omega}'\hat{\boldsymbol{W}}\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'\hat{\boldsymbol{W}}, \tag{2.52}$$

and $\nu = tr(\boldsymbol{H})$ provides a measure of the effective dimension. Also, by analogy to (2.43) and following Lin and Zhang (1999), we approximate the covariance of the fitted predictor $\hat{\mathcal{S}}(\boldsymbol{x})$ by

$$Cov[\hat{\mathcal{S}}(\boldsymbol{x})] \approx \boldsymbol{\Omega}(\boldsymbol{\Omega}'\hat{\boldsymbol{W}}\boldsymbol{\Omega} + \boldsymbol{P})^{-1}\boldsymbol{\Omega}'. \tag{2.53}$$

Since the GLM is a generalization of the basic linear model, the expressions in this Section simplify to those obtained in Sections 2.3 and 2.5 if we set the link function $g(\cdot)$ to identity and the weights w_{ii} to σ^{-2}.

2.7 Full Bayesian smoothing

In the classical estimation framework presented so far, the choice of the smoothing parameter via standard indicators like AIC, BIC or GCV is sometimes difficult, as

discussed in the final paragraph of Section 2.4.2. Even when this choice can be handled quickly, the selected smoothing parameter is treated as the true underlying smoothing parameter and then plugged into the appropriate formulae for the calculation of fitted values and variances. Such an approach operates as if the final model has been chosen in advance and ignores the inherent uncertainty introduced by the choice of the model (Hjort and Claeskens, 2003).

The smoothing parameter is a hyper-parameter that sits outside the likelihood, and such hyper-parameters are common in hierarchical Bayesian models. Hence these difficulties surrounding the choice of the smoothing parameter and the model uncertainty can be tackled by using a full Bayesian approach where all parameters in the model, including the smoothing parameter, are treated as random. This randomness is modelled by assigning prior distributions to characterize some knowledge about these parameters. For instance, in their Bayesian version of PB-splines, Lang and Brezger (2004) replaced the difference penalty with an equivalent Gaussian random walk prior (on the B-spline coefficients) and used a gamma-type prior for the smoothing parameters. More details on Bayesian smoothing within the spline context can be found in Crainiceanu et al. (2005), Baladandayuthapani et al. (2005), and references therein.

Whatever prior one uses, this full Bayesian smoothing approach yields a hierarchical Bayesian model and using Markov Chain Monte Carlo (MCMC) algorithms, one can then obtain simulations from the posterior distributions which take into account the uncertainty of all parameters. The MCMC credible intervals can also be computed using the quantiles of the sampled predictor. One advantage of this approach is that both the predictor and the underlying smoothing parameter are estimated simultaneously; however, it is computationally intensive especially for multidimensional problems.

2.8 Adaptive smoothing

So far, the extent to which the entire predictor has been smoothed is controlled by a single smoothing parameter. In many applications however, a simple plot of the data suggests that the predictor has varying smoothness, by which we mean that the predictor is changing rapidly in some regions while in other regions it is very

smooth. In these situations, the assumption of a global smoothing parameter may not be appropriate. One way to circumvent this complication is to choose the spline basis or knots adaptively to the requirement of the smoother (Luo and Wahba, 1995; Friedman and Silverman, 1989). Alternatively one can set a rich spline basis (as before) and then allow the smoothing parameter to vary locally. With this adaptive smoothing parameter approach, the model will now contain two smoothers: (i) the initial smoother/predictor $\mathcal{S}(\cdot)$ and (ii) the smoother that controls the smoothness of $\mathcal{S}(\cdot)$. There has been strong interest in this approach in recent past years, both from the classical point of view (Cardot, 2002; Krivobokova et al., 2008) and from the Bayesian perspective (Wood et al., 2002; Lang and Brezger, 2004).

Chapter 3

Appropriate covariance structure for smooth mixed models in longitudinal data analyses

Longitudinal studies in which we observe repeated measurements on subjects over time are common to many areas where Applied Statistics plays a pivotal role. In these studies, data can be divided into two categories: (i) *balanced data*, by which we mean that the same number of observations are made on each subject at the same time points, and (ii) *unbalanced data* which refer to data that are not balanced. In both cases, mixed-effect models (or simply mixed models) represent a powerful tool for data analysis. In its general form, a mixed model consists of expressing some linear predictor as a sum of two components: (a) the fixed effect, originally interpreted as the population/overall effect; and (b) the random effects, which result from the units drawn at random from the population. In some cases, a parametric approach is sufficient to summarize these effects from the data. Often however, parametric approaches do not seem appropriate, and then, smoothing is incorporated into the modelling process in order to extract the correct patterns from the data. Searle et al. (2006) investigate the basic concepts and theoretical aspects of mixed models, while Pinheiro and Bates (2000) mainly look at computational issues.

A key assumption for a mixed model is the structure of the covariance matrix of the random effects since its specification has important fitting and inferential consequences. In practice however, the real impact of the covariance structure on the

estimated effects has received very little attention as far as smooth mixed models for longitudinal data are concerned. In this Chapter, we shall consider the two bases described in Chapter 2, ie, truncated polynomial bases and the B-spline bases. First, we will demonstrate the unfortunate consequences that can occur when we use a well-known smooth mixed model with truncated lines for longitudinal data, and second, we will discuss the resolution of these problems with appropriate penalties, whether truncated polynomial or B-spline bases are used. Our aim is to present a smooth mixed model for longitudinal data with a natural, ie, non arbitrary, covariance structure and an immediately interpretable fixed effect; this covariance structure is derived from the penalties used to design the model.

Much of the material in this Chapter has appeared in Djeundje and Currie (2010a) with the important notational change that the data matrix treated here is the transpose of that in this paper; the reason for this change will become clear in Chapter 4.

This Chapter is structured as follows. Section 3.1 introduces the mixed model in its original setting. Section 3.2 presents a standard smooth mixed model for longitudinal data; we encounter some difficulties with this approach and use this to motivate a penalty approach which we examine in Section 3.3. Inference with the penalty approach and its mixed model interpretation are discussed in Sections 3.4 and 3.5, and we close this Chapter with a discussion in Section 3.6.

3.1 Mixed models

For the sake of simplicity, we start with balanced data and so assume that we have longitudinal data $\boldsymbol{Y} = (Y_{ij})$, $1 \leq i \leq n_1,\ 1 \leq j \leq n_2$, stored in the form of a matrix in such a way that rows are classified by subjects (i) and columns by time (t_j); that is, the row data $\boldsymbol{Y}_{i\bullet}$ are repeated measurements on the ith unit during time periods, $\boldsymbol{t} = (t_1, \ldots, t_{n_2})'$. A classic example is the well-known `pig.weights` data set available in the library `SemiPar` from `R`. This data set consists of the weight measurements on $n_1 = 48$ pigs (subjects) over a period of $n_2 = 9$ weeks (time); an overview is shown in the left panel in Figure 3.1. From this graphic we can see that the global effect looks linear even though the individual subject lines are quite variable, and so

it makes sense to consider models of the form

$$Y_{ij} = \{a_0 + a_1 t_j\} + \{\breve{a}_{i,0} + \breve{a}_{i,1} t_j\} + \varepsilon_{ij}, \tag{3.1}$$

where $a_0 + a_1 t$ describes the linear population/overall effect, $\breve{a}_{i,0} + \breve{a}_{i,1} t$ measures the deviations/departures of the i th subject/pig from the overall effect, and ε_{ij} represents the noise.

Clearly, we are not interested only in the specific 48 pigs involved in this study. The main motivation of mixed models is to enable our inference from (3.1) to apply to some population of pigs, and mixed models provide an attractive solution to this problem. We suppose that our sample of pigs is drawn at random from some population of pigs and that the impact of this randomness on model (3.1) is that the subject effects $\breve{\boldsymbol{a}}_i = (\breve{a}_{i,0},\, \breve{a}_{i,1})'$ are themselves random. A common specification of this randomness is that the $\breve{\boldsymbol{a}}_i$ are generated from a two-dimensional normal distribution with zero means. An important point is that this normal assumption solves the problem of non-identifiability of model (3.1); this same point will arise in Section 3.3, when we will see that penalties provide an alternative solution to the identifiability problem. Under the assumption of normality and homoskedasticity, model (3.1) can be written in matrix form as

$$\boldsymbol{Y}_{i\bullet} \,|\, \breve{\boldsymbol{a}}_i \sim \mathcal{N}\left(\boldsymbol{X}_1\boldsymbol{\beta} + \breve{\boldsymbol{X}}_1\breve{\boldsymbol{a}}_i,\; \sigma^2\boldsymbol{I}_{n_2}\right), \quad \breve{\boldsymbol{a}}_i \sim \mathcal{N}\left(\boldsymbol{0},\, \boldsymbol{\Psi}\right),$$

where $\boldsymbol{X}_1 = \breve{\boldsymbol{X}}_1 = [\boldsymbol{1}_{n_2} : \boldsymbol{t}]$, $\boldsymbol{\beta} = (a_0,\, a_1)'$, $\breve{\boldsymbol{a}}_i = (\breve{a}_{i,0},\, \breve{a}_{i,1})'$, $\boldsymbol{1}_n$ is the n-length vector of ones, and $\boldsymbol{\Psi}$ is a 2×2 symmetric, positive definite matrix. This leads to the standard mixed model representation

$$\boldsymbol{y} \,|\, \boldsymbol{u} \;\sim\; \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta} \;+\; \boldsymbol{Z}\boldsymbol{u},\; \sigma^2\boldsymbol{I}_{n_1 n_2}\right), \quad \boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0},\, \boldsymbol{\Phi}\right),$$

with

$$\boldsymbol{y} = vec(\boldsymbol{Y}'),\; \boldsymbol{X} = \boldsymbol{1}_{n_1} \otimes \boldsymbol{X}_1,\; \boldsymbol{Z} = \boldsymbol{I}_{n_1} \otimes \breve{\boldsymbol{X}}_1,\; \boldsymbol{u} = vec(\breve{\boldsymbol{a}}_1, \ldots, \breve{\boldsymbol{a}}_{n_1}),\; \boldsymbol{\Phi} = \boldsymbol{I}_{n_1} \otimes \boldsymbol{\Psi}.$$

Here, $\otimes$ represents the Kronecker product and $vec(\cdot)$ is the operator which stacks vectors or the columns of a matrix into a single vector. In the literature, $\boldsymbol{\beta}$ is known

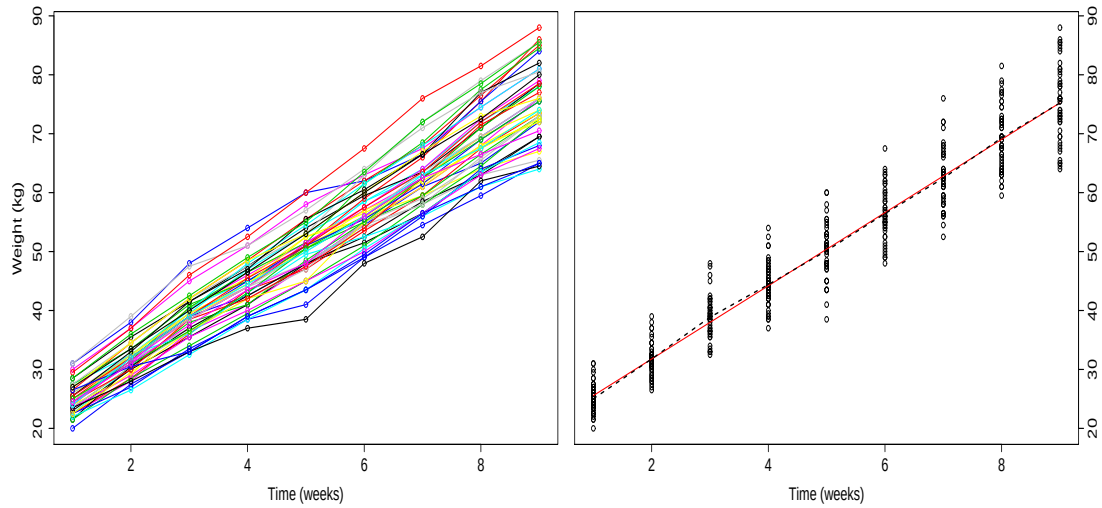

Figure 3.1: *Left: repeated measurements of the weight of 48 pigs over a period of 9 successive weeks (each continuous line refers to observations on the same pig). Right: fitted overall/population effect (red line) together with the observed point-wise average per week (black dashed line).*

as the *fixed effect* and $\boldsymbol{u}$ as the *random effect*.

The right panel in Figure 3.1 illustrates the estimated overall line fitted with the function `lme` from the `R` package `nlme`. Sub-models of (3.1) for the `pig.weights` data have been investigated by many authors; Ruppert et al. (2003, chap 4) among others implemented the case $\breve{a}_{i,1} = 0$, meaning that the subject departures from the overall effect are parallel. Such sub-models can be tested against model (3.1) to investigate the significance of this parallelism. However, as discussed in Self and Liang (1897) and Ruppert et al. (2003), this sort of test needs to be treated with care since the null hypothesis, $\mathrm{H}_0 : \boldsymbol{\Psi}_{1,2} = \boldsymbol{\Psi}_{2,2} = 0$, specifies that the non-negative $\boldsymbol{\Psi}_{2,2}$ is zero, and so sits on the boundary of the parameter space.

3.2 A standard smooth mixed model for longitudinal data

In the previous Section, we have assumed that both the overall and the subject effects can be captured linearly; this assumption suits the pigs data well. However, it is not tenable in general. Let us consider for example the left panel in Figure 3.2, which shows the daily average temperature (the averages are taken over the period 1960-1994) in 35 Canadian cities/subjects; these data are available from the the library

`fda` in `R`, and we shall refer to them as `CanadianWeather`. Clearly, for these data, the linear assumption fails (at least for the overall effect) and more flexibility is required to model the observed effects. In order to account for flexibility in such situations, both linear components in (3.1), ie, the population and the subject effects, are often extended using truncated lines as follows:

$$Y_{ij} = \left\{ a_0 + a_1 t_j + \sum_{r=1}^{q} \xi_r (t_j - \kappa_r)_+ \right\} + \left\{ \breve{a}_{i,0} + \breve{a}_{i,1} t_j + \sum_{l=1}^{\breve{q}} \breve{\xi}_{i,l} (t_j - \breve{\kappa}_l)_+ \right\} + \varepsilon_{ij}, \quad (3.2)$$

where $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_q\}$ and $\breve{\boldsymbol{\kappa}} = \{\breve{\kappa}_1, \ldots, \breve{\kappa}_{\breve{q}}\}$ are sets of equally spaced internal knots at the population and subject levels respectively. To be precise, let $\delta = (t_{n_2} - t_1)/(q+1)$ and $\breve{\delta} = (t_{n_2} - t_1)/(\breve{q}+1)$ be the distance between the knots at the population and subject levels, then the κ_r and the $\breve{\kappa}_l$ are defined by $\kappa_r = t_1 + r\delta$, $r = 1, \ldots, q$, and $\breve{\kappa}_l = t_1 + l\breve{\delta}$, $l = 1, \ldots, \breve{q}$.

Model (3.2) can be expressed compactly as

$$\boldsymbol{Y}_{i\bullet} = [\mathbf{1}_{n_2} : \boldsymbol{t}]\, \boldsymbol{a} + \boldsymbol{T}_1 \boldsymbol{\xi} \;+\; [\mathbf{1}_{n_2} : \boldsymbol{t}]\, \breve{\boldsymbol{a}}_i + \breve{\boldsymbol{T}}_1 \breve{\boldsymbol{\xi}}_i \;+\; \boldsymbol{\varepsilon}_{i\bullet} \quad (3.3)$$

where $\boldsymbol{T}_1$ and $\breve{\boldsymbol{T}}_1$ are the matrices of truncated lines at the population and subject levels. Within this setting, the smoothness of the estimates, as well as the identifiability of model (3.2), is frequently achieved by imposing the following normal constraints on the coefficients (Coull et al., 2001a; Ruppert et al., 2003, sect 9.3; Durban et al., 2005; etc):

$$\boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{0}, \sigma_P^2 \boldsymbol{I}_q\right), \quad \breve{\boldsymbol{a}}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \quad \breve{\boldsymbol{\xi}}_i \sim \mathcal{N}\left(\mathbf{0}, \breve{\sigma}^2 \boldsymbol{I}_{\breve{q}}\right), \quad (3.4)$$

where $\boldsymbol{\Sigma}$ is a 2×2 symmetric positive definite matrix. We refer to (3.4) as the *standard covariance* and to the model defined by (3.3) and (3.4) as the *standard model*.

One advantage of the standard model is that it can be fitted to data using standard functions like `lme`, as described in the Appendix of Durban et al. (2005). However, the investigation of this approach in terms of its ability to separate appropriately the two inter-connected effects (ie, the group and subject effects) has received little attention. We shall now illustrate some of the unfortunate consequences of the covariance assumption (3.4).

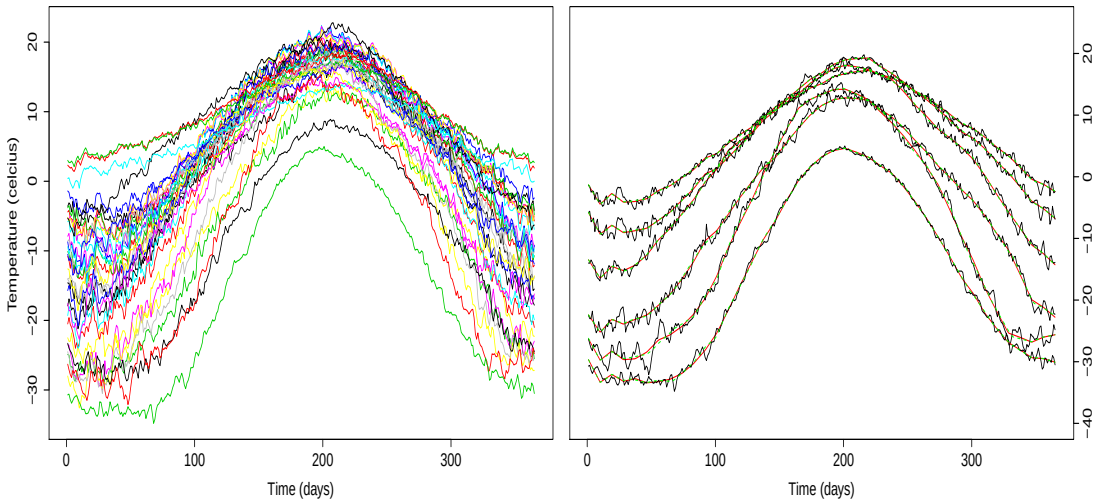

Figure 3.2: *Left: daily averages of temperature in 35 Canadian cities (each continuous line refers to observations on the same city). Right: the wiggly black lines are the observed values for selected cities; the red (smooth) lines correspond to model (3.4) fitted with* `lme` *under scenario 1; the green (dashed) lines correspond to model (3.4) fitted with* `lme` *under scenario 2 (largely hidden under the red lines).*

For the purpose of these illustrations, we consider the `CanadianWeather` data and use the `lme` function. The output of `lme` not only gives the estimates for the population and subjects effects, but also provides estimates for the variance parameters in (3.4), which we use to compute the bias corrected confidence intervals (Ruppert et al., 2003, sect 6.4) for the population and subjects effects. To illustrate our point, we first consider two knot-scenarios at the subject level. Guided by Ruppert (2002), we use $q = 39$ equi-spaced knots $\boldsymbol{\kappa}$ at the population level in both scenarios.

- Scenario 1: we use $\breve{q} = 19$ equi-spaced knots $\breve{\boldsymbol{\kappa}}$ at the subject level; in this case, $\breve{\boldsymbol{\kappa}} \subset \boldsymbol{\kappa}$.

- Scenario 2: we use $\breve{q} = 21$ equi-spaced knots $\breve{\boldsymbol{\kappa}}$ at the subject level.

The right panel in Figure 3.2 shows the fitted cities (obtained by adding the estimated population effect to the city effects) for both scenarios. As we can see from this graphic, the fits from both scenarios are almost identical and they look very satisfactory with regard to the data. One may be tempted to argue that this goodness of fit at the subject level induces a satisfactory fit at the population level. However, Figure 3.3 shows the fitted population effect for the two scenarios; we confirm the two observations of Heckman et al. (unpublished):

- the fitted population effect is very sensitive to the knot locations at the subject level, and
- the confidence bands exhibit a widening fan effect as we move from left to right.

Further, for a third scenario (not shown) with $\breve{q} = 20$, we observe upward bias, the opposite of that observed with $\breve{q} = 21$; for a fourth scenario, with $q = \breve{q} = 39$, we observe both severe bias and widening of the confidence interval. In all these scenarios, the behaviour (of the fitted population effect) is balanced by similar behaviour of the subject effects, in such a way that the global effect is recovered appropriately, as illustrated in the right panel of Figure 3.2. In other words, the action of the standard covariance (3.4) on the components of (3.2) turns out to be inconsistent. There are two main reasons for the choice in (3.4): first, the ridge penalty on a full truncated lines basis works well when we deal with smoothing at a single level (as in Chapter 2) and second, the simplicity of (3.4) is attractive; however, it does not appear capable of capturing appropriately the overall effect observed in the left panel in Figure 3.2.

One possibility is to use a full covariance matrix in (3.4) in place of $\breve{\sigma}^2 \boldsymbol{I}_{\breve{q}}$. This approach is not attractive since it has no obvious interpretation; it is also computationally very intensive. Thus, we are faced with one of the common challenges in mixed models: the appropriate specification of the covariance structure of the random effect. In the remainder of this Chapter, we do not rely on specifying the covariance structure directly; our plan is to work with appropriate penalties. The advantage of this approach is that we can discuss the modelling effects which we wish the penalties to have. Furthermore, we shall show how the penalty framework can be re-formulated as a mixed model, and then the covariance structure will follow naturally from the penalty structure and the bases.

3.3 Penalty approach

We consider the general structure

$$Y_{ij} = \mathcal{S}(t_j) + \breve{\mathcal{S}}_i(t_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,\, \sigma^2), \tag{3.5}$$

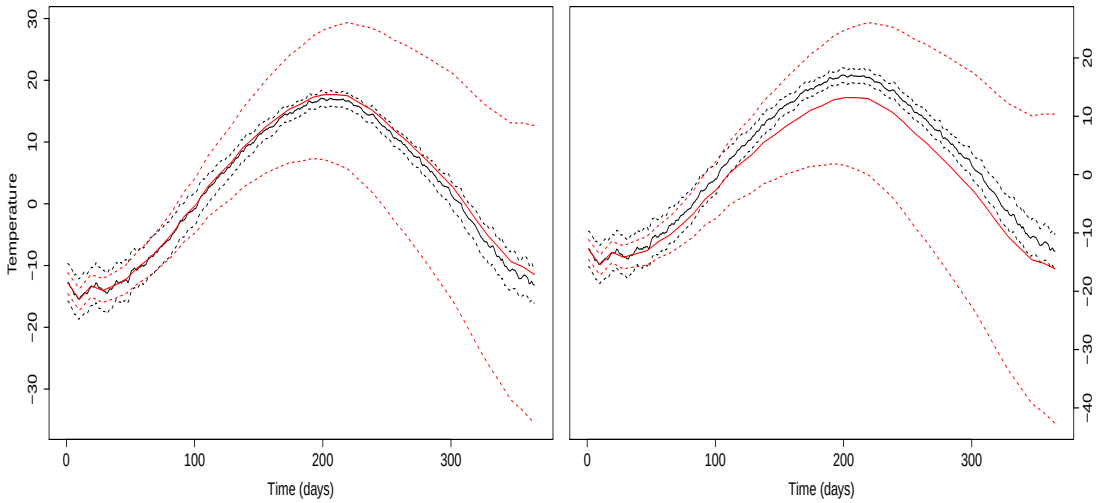

Figure 3.3: *Illustration of the sensitivity of the estimates of the population effect to the knot locations for the standard model (3.4). Left: scenario 1, ie, 39 and 19 inner knots at the population and subject levels respectively. Right: scenario 2, ie, 39 and 21 inner knots at the population and subject levels respectively. On both graphics, the black line is the observed (point-wise) mean effect with the associated empirical confidence band, while the red line is the fitted population effect.*

for some functions $\mathcal{S}(\cdot)$ and $\breve{\mathcal{S}}_i(\cdot)$ which quantify the population/overall effect and the deviations/departures of the i th unit from the population effect respectively. We view $\mathcal{S}(\cdot)$ as a smooth function (assigned to the population effect) and the $\breve{\mathcal{S}}_i(\cdot)$ as random smooth functions (assigned to the cities). At this stage, we do not make any distributional assumptions as in (3.4); these will come naturally out of our approach. In this section, we present different approaches for modelling $\mathcal{S}(\cdot)$ and $\breve{\mathcal{S}}_i(\cdot)$ and propose associated penalties for appropriate identification of these two components.

3.3.1 Penalties on B-spline bases

Here, we use B-spline bases to construct $\mathcal{S}(\cdot)$ and $\breve{\mathcal{S}}_i(\cdot)$; we start with B-spline bases because our approach with B-splines will motivate our solution with truncated polynomials. With B-splines, we will place separate penalties directly on the B-spline coefficients at the subject level, one to bring about smoothness (a difference penalty) and another to achieve identifiability by shrinkage (a ridge penalty). With truncated polynomials, it seems more difficult to achieve identifiability by direct shrinkage of the coefficients, as we have seen with the results of (3.2) in Figure 3.3. Indeed, with truncated polynomial bases we shall see in Section 3.3.2 that one way to achieve

both smoothness and identifiability is to introduce a second penalty, as we do with B-spline bases. Additional reasons for starting with B-splines are those discussed in Section 2.3.3. Hence, if we denote by $\mathcal{S}(\boldsymbol{t})$ the element-wise action of $\mathcal{S}(\cdot)$ on the time vector $\boldsymbol{t} = (t_1, \ldots, t_{n_2})'$, with a similar meaning for $\breve{\mathcal{S}}_i(\boldsymbol{t})$, then we write

$$\mathcal{S}(\boldsymbol{t}) = \boldsymbol{B}\boldsymbol{\theta} \quad \text{and} \quad \breve{\mathcal{S}}_i(\boldsymbol{t}) = \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}_i, \tag{3.6}$$

where $\boldsymbol{B}$ and $\breve{\boldsymbol{B}}$ are $n_2 \times c$ and $n_2 \times \breve{c}$ regression matrices of B-splines evaluated along $\boldsymbol{t}$, $\boldsymbol{\theta}$ is a vector of coefficients specifying the population effect, and the $\breve{\boldsymbol{\theta}}_i$ are random vectors of coefficients related to the subjects. We will refer to (3.5) and (3.6) as model $\texttt{M1} = \texttt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}})$.

Note that `M1` is not identifiable; indeed, if we add (for example) a constant to $\mathcal{S}(\cdot)$ and subtract the same constant from the $\breve{\mathcal{S}}_i(\cdot)$, then the predictor $\mathcal{S}(\cdot) + \breve{\mathcal{S}}_i(\cdot)$ remains unchanged. Thus, two issues need to be clarified in `M1`: *smoothness* and *identifiability*. In the context of nested curves, Brumback and Rice (1998) achieved smoothness via the smoothing spline approach (which can be very time consuming, specifically in the presence of a large data set as `CanadianWeather`); from the mixed model representation, they suggested using ANOVA-like identifiability constraints by requiring that the linear parts of the subject effects sum to zero at each level except the topmost level. Here we address smoothness and identifiability simultaneously via penalties as follows.

Let us first consider the overall effect $\mathcal{S}(\boldsymbol{t})$. For this component we take a sufficiently rich basis of B-splines and penalize the differences in adjacent components of $\boldsymbol{\theta}$ as described in Section 2.3.2. Thus the estimation of $\boldsymbol{\theta}$ will be subject to the penalty/constraint

$$\|\boldsymbol{\Delta}_d \boldsymbol{\theta}\|^2 \leq \rho,$$

where ρ quantifies the amount of smoothness applied to $\boldsymbol{\theta}$. A similar inequality constraint (but, without the difference operator) has been used on truncated line coefficients by Ruppert et al. (2003).

Given the above specification on the overall effect, we address the identifiability problem by centring the city effects via a shrinkage of the city coefficients $\breve{\boldsymbol{\theta}}_i$ as

$$\|\breve{\boldsymbol{\theta}}_i\|^2 \leq \breve{\rho}_2, \ i = 1, \ldots, n_1,$$

for some well chosen parameter $\breve{\rho}_2$.

The problem of smoothness of the city effects remains. Two possibilities are available:

(a) work with fewer B-splines (for the city effects) and only the ridge penalty, or

(b) take a rich set of B-splines as a basis (as for the overall effect) and apply a difference penalty (together with the ridge penalty) on the components of $\breve{\boldsymbol{\theta}}_i$; hence we further penalize the roughness of the $\breve{\boldsymbol{\theta}}_i$, ie,

$$\|\boldsymbol{\Delta}_2\breve{\boldsymbol{\theta}}_i\|^2 \leq \breve{\rho}_1,\ i = 1, \ldots, n_1.$$

Clearly, (b) is computationally more intensive than (a) since each city has its own (large) set of coefficients, while in comparison with (b), (a) is economical. However, (a) is open to the criticism that the selection of the number of B-splines is manual and artificial. Nonetheless, both approaches produce similar results (at least for `CanadianWeather` data), provided that a judicious choice of the number of B-splines is made at the subject level under (a). From now on, we will consider approach (b) only.

We remark that we have used a d-order penalty for smoothing at the population level since we may wish to have a specific fixed effect at this level; for instance, `CanadianWeather` data in Figure 3.2 suggest a quadratic fixed effect at the population level, in which case we take $d = 3$ (although a circular approach/penalty will be more appropriate as we shall discuss in Section 3.6). We have no particular form in mind for the city effects, and so we simply use a second order ($d = 2$) penalty to smooth these effects.

In summary for `M1`, (i) smoothing of the population effect is achieved by d-order penalization of the population coefficients, (ii) smoothing of the city effects is achieved by second order penalization of the city coefficients and (iii) identifiability is achieved by centring via a ridge penalty on the city coefficients. These three points are summarized as follows:

$$\texttt{C1} : \|\boldsymbol{\Delta}_d\boldsymbol{\theta}\|^2 \leq \rho, \quad \|\boldsymbol{\Delta}_2\breve{\boldsymbol{\theta}}_i\|^2 \leq \breve{\rho}_1, \quad \|\breve{\boldsymbol{\theta}}_i\|^2 \leq \breve{\rho}_2; \tag{3.7}$$

these constraints apply to the model `M1` and we refer to (3.7) as the constraints `C1`.

3.3.2 Penalties on truncated polynomial bases

Here we express $\mathcal{S}(\cdot)$ and $\breve{\mathcal{S}}_i(\cdot)$ in terms of a truncated polynomial and a truncated line basis respectively; ie, we set

$$\mathcal{S}(\boldsymbol{t}) = [\mathbf{1}_{n_2} : \boldsymbol{t} : \cdots : \boldsymbol{t}^p]\boldsymbol{a} + \boldsymbol{T}_p\boldsymbol{\xi} \;=\; \boldsymbol{X}_p\boldsymbol{a} + \boldsymbol{T}_p\boldsymbol{\xi} \;=\; \boldsymbol{L}_p\boldsymbol{\gamma}, \tag{3.8}$$

$$\breve{\mathcal{S}}_i(\boldsymbol{t}) = [\mathbf{1}_{n_2} : \boldsymbol{t}]\breve{\boldsymbol{a}}_i + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_i \;=\; \breve{\boldsymbol{X}}_1\breve{\boldsymbol{a}}_i + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_i \;=\; \breve{\boldsymbol{L}}_1\breve{\boldsymbol{\gamma}}_i, \tag{3.9}$$

say, where $\boldsymbol{T}_r$ and $\breve{\boldsymbol{T}}_r$ are matrices of truncated polynomials of degree r at the population and subject levels respectively. We will refer to (3.5), (3.8) and (3.9) as model $\texttt{M2} = \texttt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}})$.

With B-spline bases in the previous section, a polynomial fixed effect of degree $(d-1)$ at the population level was extracted by choosing a difference penalty of order d. An analogous thing is achieved in (3.8) by choosing the corresponding degree $p = d-1$ of the polynomial basis. Since the subject effects are likely (at least for `CanadianWeather`) to be quite different from one another, we simply use truncated lines at the subject level.

With a B-spline basis as in the previous section, the behaviour of $\breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}_i$ is very similar to that of $\breve{\boldsymbol{\theta}}_i$ in the sense that smoothness of $\breve{\boldsymbol{\theta}}_i$ implies the smoothness of $\breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}_i$, and shrinkage of $\breve{\boldsymbol{\theta}}_i$ implies shrinkage of $\breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}_i$. This is not entirely clear with truncated polynomial bases of degree p. For the latter, the coefficient vector $\breve{\boldsymbol{\xi}}_i$ reflects the jumps in the derivatives of order p at the corresponding knots (as discussed in Section 2.3.1) and so the smoothness of the population and subject effects is obtained by applying a ridge penalty on $\boldsymbol{\xi}$ and $\breve{\boldsymbol{\xi}}_i$, ie,

$$\|\boldsymbol{\xi}\|^2 \le \rho \quad \text{and} \quad \|\breve{\boldsymbol{\xi}}_i\| \le \breve{\rho}_1,\ i = 1, \ldots, n_1.$$

With B-splines, we have two penalties at the subject level, one for smoothness and one for identifiability. With this in mind, we address the identifiability issue by the introduction of a second penalty in `M2` at the subject level which shrinks the subject effects $\breve{\mathcal{S}}_i(\boldsymbol{t}) = \breve{\boldsymbol{L}}_1\breve{\boldsymbol{\gamma}}_i$ as

$$\|\breve{\boldsymbol{L}}_1\breve{\boldsymbol{\gamma}}_i\|^2 \le \breve{\rho}_2,\ i = 1, \ldots, n_1.$$

In summary for `M2`, smoothness at the population and subject level is obtained by

applying a ridge penalty on the truncated polynomial coefficients, and identifiability is achieved by shrinkage. We summarize these constraints in

$$\texttt{C2} : \|\boldsymbol{\xi}\|^2 \leq \rho, \quad \|\breve{\boldsymbol{\xi}}_i\| \leq \breve{\rho}_1, \quad \|\breve{\boldsymbol{L}}_1\breve{\boldsymbol{\gamma}}_i\|^2 \leq \breve{\rho}_2. \tag{3.10}$$

3.3.3 Penalties on a mixture of B-spline and truncated polynomial bases

Here we consider a mixture of B-splines and truncated polynomials. We start with

$$\mathcal{S}(\boldsymbol{t}) = \boldsymbol{B}\boldsymbol{\theta} \quad \text{and} \quad \breve{\mathcal{S}}_i(\boldsymbol{t}) = \breve{\boldsymbol{L}}_1\breve{\boldsymbol{\gamma}}_i \;=\; \breve{\boldsymbol{X}}_1\breve{\boldsymbol{a}}_i + \breve{\boldsymbol{T}}_1\breve{\boldsymbol{\xi}}_i, \tag{3.11}$$

say, where the components have been defined in Sections 3.3.1 and 3.3.2. We refer to (3.5) and (3.11) as model $\texttt{M3} = \texttt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}})$; for the same reasons detailed previously, smoothness and identifiability constraints on M3 are

$$\texttt{C3} : \|\boldsymbol{\Delta}_d\boldsymbol{\theta}\|^2 \leq \rho, \quad \|\breve{\boldsymbol{\xi}}_i\| \leq \breve{\rho}_1, \quad \|\breve{\boldsymbol{L}}_1\breve{\boldsymbol{\gamma}}_i\|^2 \leq \breve{\rho}_2. \tag{3.12}$$

Similarly, we consider the representation

$$\mathcal{S}(\boldsymbol{t}) = \boldsymbol{L}_p\boldsymbol{\gamma} \;=\; \boldsymbol{X}_p\boldsymbol{a} + \boldsymbol{T}_p\boldsymbol{\xi} \quad \text{and} \quad \breve{\mathcal{S}}_i(\boldsymbol{t}) = \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}_i; \tag{3.13}$$

we refer to (3.5) and (3.13) as $\texttt{M4} = \texttt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}})$; smoothness and identifiability constraints on M4 are as follows:

$$\texttt{C4} : \|\boldsymbol{\xi}\|^2 \leq \rho, \quad \|\boldsymbol{\Delta}_2\breve{\boldsymbol{\theta}}_i\|^2 \leq \breve{\rho}_1, \quad \|\breve{\boldsymbol{\theta}}_i\|^2 \leq \breve{\rho}_2. \tag{3.14}$$

3.3.4 Further possibilities

Models M1, M2, M3 and M4 with the associated constraints C1, C2, C3 and C4 are the main models that we investigate in the remainder of this Chapter. In all of these four cases, we achieve smoothness either by a difference penalty on the B-spline coefficients or a ridge penalty on the truncated polynomial coefficients. An alternative might be to smooth by applying a roughness (difference) penalty directly on the estimates, ie, on $\mathcal{S}(\boldsymbol{t})$ and $\breve{\mathcal{S}}_i(\boldsymbol{t})$, (whether a B-spline or a truncated polynomial basis is used). Note

also that solving the identifiability problem by shrinking the subject effects $\breve{\mathcal{S}}_i(\boldsymbol{t})$ is also applicable with a B-spline basis at the subject level. Furthermore, instead of applying shrinkage to $\breve{\mathcal{S}}_i(\boldsymbol{t})$, one may tackle the identifiability problem by applying the shrinkage at the knots only, ie, a ridge penalty to $\breve{\mathcal{S}}_i(\breve{\boldsymbol{\kappa}})$. These are topics for further research.

3.4 Inference and application

In Section 3.3, we presented four formulations of model (3.5) with penalized splines. Each of these formulations is expressed as

$$\boldsymbol{Y}_{i\bullet} = \boldsymbol{\Omega}_P\boldsymbol{\alpha}_P + \breve{\boldsymbol{\Omega}}\breve{\boldsymbol{\alpha}}_i + \boldsymbol{\varepsilon}_{i\bullet}, \quad \boldsymbol{\varepsilon}_{i\bullet} \sim \mathcal{N}(\boldsymbol{0},\, \sigma^2\boldsymbol{I}_{n_2}), \tag{3.15}$$

where $\boldsymbol{\Omega}_P$ and $\breve{\boldsymbol{\Omega}}$ are $n_2 \times c$ and $n_2 \times \breve{c}$ regression matrices at the population and subject levels, with the associated coefficients $\boldsymbol{\alpha}_P$ and $\breve{\boldsymbol{\alpha}}_i$. Specifically, we have:

$$(\boldsymbol{\Omega}_P, \boldsymbol{\alpha}_P) = \begin{cases} (\boldsymbol{B}, \boldsymbol{\theta}) & \text{under } \texttt{M1} \text{ or } \texttt{M3} \\ (\boldsymbol{L}_p, \boldsymbol{\gamma}) & \text{under } \texttt{M2} \text{ or } \texttt{M4} \end{cases} \tag{3.16}$$

$$(\breve{\boldsymbol{\Omega}}, \breve{\boldsymbol{\alpha}}_i) = \begin{cases} (\breve{\boldsymbol{B}}, \breve{\boldsymbol{\theta}}_i) & \text{under } \texttt{M1} \text{ or } \texttt{M4} \\ (\breve{\boldsymbol{L}}_1, \breve{\boldsymbol{\gamma}}_i) & \text{under } \texttt{M2} \text{ or } \texttt{M3}. \end{cases} \tag{3.17}$$

This general form (3.15) can be written compactly as

$$\boldsymbol{y} = \boldsymbol{\Omega}\boldsymbol{\alpha} + vec(\boldsymbol{\varepsilon}), \quad vec(\boldsymbol{\varepsilon}) \sim \mathcal{N}(\boldsymbol{0},\, \sigma^2\boldsymbol{I}_{n_1n_2}), \tag{3.18}$$

where

$$\begin{aligned} \boldsymbol{\Omega} &= \begin{bmatrix} \boldsymbol{\Omega}_P & \breve{\boldsymbol{\Omega}} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{\Omega}_P & \boldsymbol{0} & \breve{\boldsymbol{\Omega}} & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{\Omega}_P & \boldsymbol{0} & \boldsymbol{0} & \breve{\boldsymbol{\Omega}} & \ddots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{\Omega}_P & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \breve{\boldsymbol{\Omega}} \end{bmatrix} \\ &= [\boldsymbol{1}_{n_1} \otimes \boldsymbol{\Omega}_P : \boldsymbol{I}_{n_1} \otimes \breve{\boldsymbol{\Omega}}] \end{aligned} \tag{3.19}$$

is the full regression matrix, $\boldsymbol{\alpha} = vec(\boldsymbol{\alpha}_P, \breve{\boldsymbol{\alpha}}_1, \ldots, \breve{\boldsymbol{\alpha}}_{n_1})$ is the joint vector of population and subject coefficients, and $\boldsymbol{\varepsilon} = vec(\boldsymbol{\varepsilon}_{1\bullet}, ..., \boldsymbol{\varepsilon}_{n_1\bullet})$ is the noise vector. We note that model (3.18) is similar in form to that of (2.4) and (2.11) described in Chapter 2, except that the joint coefficient $\boldsymbol{\alpha}$ in (3.18) is controlled by multiple constraints.

It can be shown that the penalized residual sum of squares (PRSS) of (3.18), ie, the residual sum of squares (RSS) under constraints C1, C2, C3 or C4, is given by

$$\mathsf{PRSS} = \mathsf{RSS} + \boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha}, \quad \text{with} \quad \mathsf{RSS} = \|\boldsymbol{y} - \boldsymbol{\Omega}\boldsymbol{\alpha}\|^2, \tag{3.20}$$

where

$$\boldsymbol{P} = blockdiag(\boldsymbol{P}_P, \boldsymbol{I}_{n_1} \otimes \breve{\boldsymbol{P}}) \tag{3.21}$$

is the block diagonal penalty matrix, with

$$\boldsymbol{P}_P = \begin{cases} \lambda \boldsymbol{\Delta}_d' \boldsymbol{\Delta}_d & \text{for } \mathtt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}}) \text{ under } \mathtt{C1} \text{ or } \mathtt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}}) \text{ under } \mathtt{C3} \\ \lambda \boldsymbol{J}_d & \text{for } \mathtt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}}) \text{ under } \mathtt{C2} \text{ or } \mathtt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}}) \text{ under } \mathtt{C4} \end{cases} \tag{3.22}$$

$$\breve{\boldsymbol{P}} = \begin{cases} \breve{\lambda}_1 \boldsymbol{\Delta}_2' \boldsymbol{\Delta}_2 + \breve{\lambda}_2 \boldsymbol{I}_{\breve{c}} & \text{for } \mathtt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}}) \text{ under } \mathtt{C1} \text{ or } \mathtt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}}) \text{ under } \mathtt{C4} \\ \breve{\lambda}_1 \boldsymbol{J}_2 + \breve{\lambda}_2 \breve{\boldsymbol{L}}_1' \breve{\boldsymbol{L}}_1 & \text{for } \mathtt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}}) \text{ under } \mathtt{C2} \text{ or } \mathtt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}}) \text{ under } \mathtt{C3}. \end{cases} \tag{3.23}$$

Here, $\boldsymbol{J}_r$ is the identity matrix (of appropriate size) where the upper r diagonal elements have been set to zero, while λ and $\breve{\lambda}_1$ are the smoothing parameters at the population and subject level respectively, and $\breve{\lambda}_2$ is the shrinkage parameter of the subject effects; $(\lambda, \breve{\lambda}_1, \breve{\lambda}_2)$ plays (inversely) the equivalent role as $(\rho, \breve{\rho}_1, \breve{\rho}_2)$ used throughout Section 3.3. More precisely, increasing values of λ and $\breve{\lambda}_1$, (ie, decreasing the values of ρ and $\breve{\rho}_1$) induces more smoothness on the population and subject effects, while increasing the values of $\breve{\lambda}_2$, (ie, decreasing values of $\breve{\rho}_2$) corresponds to heavier shrinkage on the subject effects. At the limit, ie, $\breve{\lambda}_2 \to \infty$ (or equivalently $\breve{\rho}_2 \to 0$), we have $S_i(\cdot) \to 0$; this reduces the linear predictor of the model to the population effect. Given values of these parameters, we obtain

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{\Omega}'\boldsymbol{\Omega} + \boldsymbol{P})^{-1} \boldsymbol{\Omega}'\boldsymbol{y} \tag{3.24}$$

on minimizing the PRSS in (3.20). Note that the global regression matrix $\boldsymbol{\Omega}$ is very sparse as shown in (3.19), and this sparsity can be exploited to speed up (3.24).

Table 3.1: *Summary table for four models applied to* `CanadianWeather` *data.*

	$\texttt{M1}(\boldsymbol{B}, \breve{\boldsymbol{B}})$	$\texttt{M2}(\boldsymbol{T}, \breve{\boldsymbol{T}})$	$\texttt{M3}(\boldsymbol{B}, \breve{\boldsymbol{T}})$	$\texttt{M4}(\boldsymbol{T}, \breve{\boldsymbol{B}})$
$(\lambda, \breve{\lambda}_1, \breve{\lambda}_2)$	$(0.035, 20, 0.023)$	$(250, 1097, 7 \times 10^{-4})$	$(0.083, 1097, 5 \times 10^{-4})$	$(250, 20, 0.023)$
RSS	6904	6623	6635	6904
$tr(\boldsymbol{H})$	449	514	513	450
BIC	117179	117261	117267	117182
σ^2	0.56	0.54	0.54	0.56

For illustrations, we choose the smoothing/shrinkage parameters by minimizing the BIC. Nonetheless, since the BIC mostly addresses model choice at the global level, a more formal criterion for the selection of the shrinkage parameter may be derived by some partition of the BIC into a population and a subject component. Alternatively, one may prefer to choose a certain fixed amount of shrinkage, or to study (as a function of the shrinkage parameter) the departure of the fitted population effect from the observed mean, and then choose the amount of shrinkage that minimizes this departure.

Once the optimal smoothing/shrinkage parameters have been selected, the posterior covariance $cov(\boldsymbol{\alpha}|\boldsymbol{y})$ of the joint vector $\boldsymbol{\alpha}$ can be obtained as described in Section 2.5. Hence, the posterior covariance of the population coefficient $\boldsymbol{\alpha}_P$ (defined in (3.15)) follows immediately by taking the upper left $c \times c$ block of $cov(\boldsymbol{\alpha}|\boldsymbol{y})$. It is this posterior distribution that we use to compute the confidence band around the population effect.

We now apply this procedure to the `CanadianWeather` data. For each of our four models, we follow Ruppert (2002) and so use 39 equi-spaced internal knots at the population and subject levels respectively. The results are summarized in Table 3.1, and Figure 3.4 illustrates the fitted population effect with the associated confidence intervals. For our four models, `M1` and `M4` are the best ones (for `CanadianWeather` data) both in terms of BIC and parsimony. Note that the confidence bands for models `M1` and `M4`, the two models with $\breve{\boldsymbol{B}}$ as regression matrix at the subject level, are narrower than those of `M2` and `M3`. The right panel in Figure 3.5 displays the city effects as estimated under model `M1`; these city effects are essentially identical for all four models.

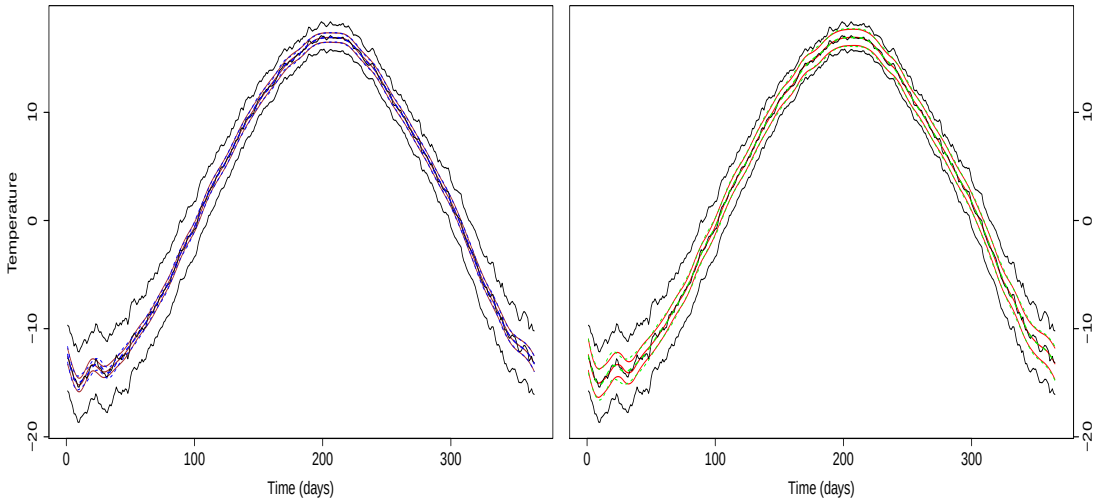

Figure 3.4: *Data and fitted population effect for our four models. The wiggly (black) line is the data (point-wise average) with the associated empirical confidence interval. Left:* `M1` *(brown) and* `M4` *(blue). Right:* `M2` *(red) and* `M3` *(green).*

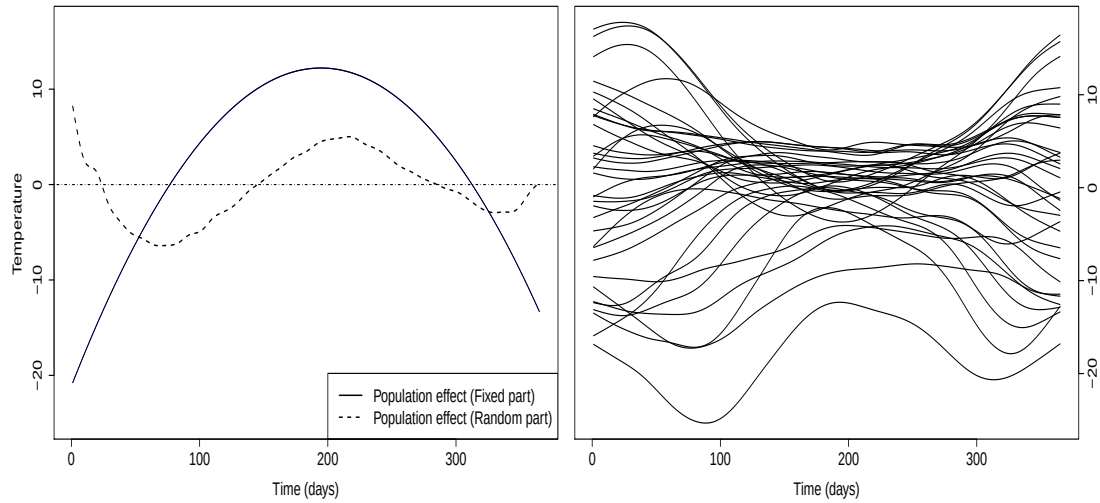

Figure 3.5: *The three components of model* `M1` *applied to the data* `CanadianWeather`. *Left: decomposition of the fitted population effect into the fixed (quadratic) component (continuous line) and the "random" component (dashed line). Right: fitted subject effects.*

In summary, our models all return essentially identical estimates of both the population and city effects; the width of the confidence intervals appears to depend on the basis used at the subject level. We will return to this point in Section 3.6.

We have also considered using different knot scenarios at the subject level from that at the population level. While this generally produces consistent estimates of the population and subject effects, this may adversely affect the confidence intervals at both levels for all four models, if the number of knots at the subject level is "too small" relative to that at the population level. Again, we will return to this point in Section 3.6.

3.5 Mixed model representation and interpretation

In the representation (3.15), it is natural to think of the coefficients $\breve{\boldsymbol{\alpha}}_i$ as random, since the subjects which they represent are randomly chosen from the population. The question is the following: from the smoothness and identifiability assumptions made so far, can we "naturally" derive the distributions which have generated these coefficients/subjects?

3.5.1 Mixed model representation for models M2 and M4

Recall that under M2 or M4, we have a truncated polynomial basis at the population level, ie, $\boldsymbol{\Omega}_P = \boldsymbol{L}_p = [\boldsymbol{X}_p : \boldsymbol{T}_p]$ and $\boldsymbol{\alpha}_P = \boldsymbol{\gamma} = vec(\boldsymbol{a}, \boldsymbol{\xi})$. Here, the mixed model representation is straightforward; this follows from the structure of the truncated polynomial basis at the population level. Indeed, the minimization of PRSS under M2 or M4 (with the associated constraints) is equivalent to the maximization of the log-likelihood which arises from the triplet $(\boldsymbol{y}, \boldsymbol{\xi}, \breve{\boldsymbol{\alpha}})$, where $\boldsymbol{\xi}$ and $\breve{\boldsymbol{\alpha}}$ are treated as a pair of independent random vectors under the distributional assumptions

$$\begin{aligned} \boldsymbol{Y}_{i\bullet} \,|\, \boldsymbol{\xi}, \breve{\boldsymbol{\alpha}}_i &\sim \mathcal{N}\left(\boldsymbol{X}_p\boldsymbol{a} + \boldsymbol{T}_p\boldsymbol{\xi} \;+\; \breve{\boldsymbol{\Omega}}\breve{\boldsymbol{\alpha}}_i,\;\; \sigma^2\, \boldsymbol{I}_{n_2}\right), \\ \boldsymbol{\xi} \sim \mathcal{N}\left(\boldsymbol{0},\; \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right),&\;\; \breve{\boldsymbol{\alpha}}_i \sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2\breve{\boldsymbol{P}}^{-1}\right), \end{aligned} \tag{3.25}$$

where $\breve{\boldsymbol{P}}$, defined in (3.23), depends on $\breve{\lambda}_1$ and $\breve{\lambda}_2$. We comment on this representation in Section 3.5.3.

3.5.2 Mixed model representation for models M1 and M3

Under both M1 and M3, we have a B-spline basis at the population level and so $\boldsymbol{\Omega}_P = \boldsymbol{B}$ and $\boldsymbol{\alpha}_P = \boldsymbol{\theta}$. In this case, the minimization of PRSS (with the associated constraints) is equivalent to maximizing the log-likelihood which arises from the triplet $(\boldsymbol{y}, \boldsymbol{\theta}, \breve{\boldsymbol{\alpha}})$, where $\boldsymbol{\theta}$ and $\breve{\boldsymbol{\alpha}}$ are treated as a pair of independent random vectors under the (improper for $\boldsymbol{\theta}$) distributional assumptions

$$\begin{aligned} \boldsymbol{Y}_{i\bullet} \,|\, \boldsymbol{\theta}, \breve{\boldsymbol{\alpha}}_i &\sim \mathcal{N}\left(\boldsymbol{B}\boldsymbol{\theta} \;+\; \breve{\boldsymbol{\Omega}}\breve{\boldsymbol{\alpha}}_i,\;\; \sigma^2\, \boldsymbol{I}_{n_2}\right), \\ \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2(\lambda \boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d)^-\right), &\;\; \breve{\boldsymbol{\alpha}}_i \sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2 \breve{\boldsymbol{P}}^{-1}\right). \end{aligned} \tag{3.26}$$

where $\breve{\boldsymbol{P}}$, defined in (3.23), depends on $\breve{\lambda}_1$ and $\breve{\lambda}_2$.

The roughness matrix $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$, which gives rise to the improper prior distribution for $\boldsymbol{\theta}$ in (3.26), is singular, symmetric and has rank $c-d$. Now the singular value decomposition of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$ is of the form $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ where $\boldsymbol{\Lambda} = diag(\rho_1, \ldots, \rho_{c-d}, 0, \ldots, 0)$ is the $c \times c$ diagonal matrix of eigenvalues arranged in non-increasing order; and $\boldsymbol{U}$ is the matrix with columns given by the eigenvectors of $\boldsymbol{\Delta}_d'\boldsymbol{\Delta}_d$. We will denote $diag(\rho_1, \ldots, \rho_{c-d})$ by $\boldsymbol{\Lambda}_d$. Hence, following Section 2.3.3, the smoother $\boldsymbol{B}\boldsymbol{\theta}$ can be re-parametrized as

$$\boldsymbol{B}\boldsymbol{\theta} = \boldsymbol{X}_p\boldsymbol{\theta}_1 + \boldsymbol{R}\boldsymbol{\theta}_2, \quad \text{with } \; \boldsymbol{R} = \boldsymbol{B}\boldsymbol{U}_d\boldsymbol{\Lambda}_d^{-1/2}, \tag{3.27}$$

where $\boldsymbol{U}_d$ is the sub-matrix of $\boldsymbol{U}$ corresponding to the diagonal elements of $\boldsymbol{\Lambda}_d$. With this decomposition, the (improper) normal assumption about $\boldsymbol{\theta}$ in (3.26) reduces to

$$\boldsymbol{\theta}_2 \sim \mathcal{N}\left(\boldsymbol{0},\; \frac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right).$$

Finally, minimizing the PRSS of M1 or M3 yields the mixed model representation

$$\begin{aligned} \boldsymbol{Y}_{i\bullet} \,|\, \boldsymbol{\theta}_2, \breve{\boldsymbol{\alpha}}_i &\sim \mathcal{N}\left(\boldsymbol{X}_p\boldsymbol{\theta}_1 + \boldsymbol{R}\boldsymbol{\theta}_2 + \breve{\boldsymbol{\Omega}}\breve{\boldsymbol{\alpha}}_i,\; \sigma^2\, \boldsymbol{I}_{n_2}\right), \\ \boldsymbol{\theta}_2 \sim \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right), &\quad \breve{\boldsymbol{\alpha}}_i \sim \mathcal{N}\left(\boldsymbol{0},\; \sigma^2 \breve{\boldsymbol{P}}^{-1}\right); \end{aligned} \tag{3.28}$$

we comment on this representation in the next Section.

3.5.3 Interpretation of the components

Clearly from (3.25) and (3.28), the mixed model representation of (3.15) for our four models `M1`, `M2`, `M3` and `M4` with the associated constraints has the form

$$\begin{aligned} \boldsymbol{Y}_{i\bullet} \,|\, \boldsymbol{b}, \breve{\boldsymbol{\alpha}}_i &\sim \mathcal{N}\left(\boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{Z}_p\boldsymbol{b} + \breve{\boldsymbol{\Omega}}\breve{\boldsymbol{\alpha}}_i,\ \sigma^2\,\boldsymbol{I}_{n_2}\right), \\ \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_{c-d}\right), &\quad \breve{\boldsymbol{\alpha}}_i \sim \mathcal{N}\left(\boldsymbol{0},\ \sigma^2\breve{\boldsymbol{P}}^{-1}\right), \end{aligned} \tag{3.29}$$

for appropriate $\boldsymbol{\beta}$, $\boldsymbol{Z}_p$, and $\boldsymbol{b}$. Hence, the model predictor is made up of three components:

- The first component, $\boldsymbol{X}_p\boldsymbol{\beta}$, represents the fixed overall effect. Motivated by the overview of the data in Figure 3.2, we require this component to be quadratic for `CanadianWeather`; this justifies the use of the third order difference penalty in `M1` and `M3`. An illustration of this first component under `M1` is shown by the continuous line in the left panel of Figure 3.5.

- The second component, $\boldsymbol{Z}_p\boldsymbol{b}$, which is shrunk towards $\boldsymbol{0}$, accounts for the flexibility of the population effect, and smoothly captures the deviation of the population effect from a simple quadratic curve. We do not view the normal constraint on this component as random behaviour, but just as a smoothing device. This component is illustrated (under `M1`) for `CanadianWeather` by the dashed line in the left panel in Figure 3.5.

- The third/random component, $\breve{\boldsymbol{\Omega}}\breve{\boldsymbol{\alpha}}_i$, measures the random departure of the subjects from the overall effect. The normal constraint on this component incorporates the random behaviour of the cities (controlled by $\breve{\lambda}_2$) as well as the smoothness of the city effects (as measured by $\breve{\lambda}_1$); these are shown for `CanadianWeather` (under `M1`) on the right panel in Figure 3.5.

In these illustrations, we have used the values of the smoothing and shrinkage parameters obtained from minimizing the BIC; one of the main reasons is that, due to the double constraints on the subject coefficients, current packages/softwares for fitting mixed models are not flexible enough to handle the model describe through (3.29).

We shall describe how to implement these models via restricted likelihood in the next Chapter.

3.6 Discussion

In this Chapter, we first illustrated some consequences of the mis-specification of the standard covariance structure (3.4) in a mixed model (for longitudinal data) defined using truncated line bases. One simple way of demonstrating the problem is to fit only the population effect. Here, truncated lines with a ridge penalty and B-splines with a difference penalty give almost identical answers, and both capture the population effect correctly with appropriate confidence intervals. However, when we add the city effects, the estimates of the population effect and its associated confidence intervals are distorted when truncated lines with the standard covariance structure (3.4) are used, as shown in Figure 3.3. No such distortion occurs with the penalty approach; the estimates of the population effect are identical whether city effects are included or not.

For the penalty approach, we first specify the bases, and then we design the components of the model (population, subject, etc, effects). With the components in place, we use penalties to bring about the model effects we wish to achieve. Even though the B-spline and truncated polynomial bases produced satisfactory results in the applications presented in this Chapter, we have a preference for B-splines bases, because of the direct connection between the regression coefficients and the penalty which is applied to these coefficients; for instance, with the B-spline basis, we can easily adjust the penalty to link the start and end of the year via a circular penalty, or to account for a periodic effect (for example if we are interested in modelling the temperatures collected over many years) by using a harmonic penalty. In the case of the `CanadianWeather` data we have used neither a circular nor a harmonic penalty since we used these data to illustrate some general points of fitting smooth subject-specific curves.

We return to two issues which we raised at the end of Section 3.4. First, our methods appear to be successful in recovering population and subject effects, and in solving the problem of the widening fan effect found with (3.4). However, the width of the associated confidence intervals arising from the use of BIC depends on whether

a B-spline or a truncated lines basis is used at the subject level. Nonetheless, it is possible to "play" with the values of the smoothing/shrinkage parameters when truncated lines are used and to produce the confidence intervals obtained with B-splines. A second difficulty arises when selecting the smoothing/shrinkage parameters by optimizing a deviance-type criterion like BIC. We found that, for some data such as the `CanadianWeather`, if the number of knots at the subject level is "too small" relative to the number at the population level, ie, $\breve{q} \ll q$, (for instance, $q = 39$ and $\breve{q} < 10$, for the `CanadianWeather` data), then the optimal values of the shrinkage/smoothing parameters as selected by BIC fall on the boundary of the parameters space; this can lead to unexpectedly wide confidence intervals at the population and subject levels. Hence, in practice, attention must be given to the choice of q and $\breve{q}$ (for example by following Ruppert (2002) both at the population and subject levels) as well as to the indicator used to select optimal values of these parameters.

Chapter 4

Penalized spline smoothing for hierarchical curves with applications to grouped longitudinal data

In the preceding Chapter we discussed the fitting of smooth models to balanced longitudinal data in which the same number of observations are made on each subject at the same time points. Throughout this discussion, we restricted our attention to the situation where all subjects were pooled in the same group. These restrictions are relaxed in this Chapter. Hence, let us suppose that data are collected on n subjects which are divided into m groups of sizes $r_1, ..., r_m$, and let $g(i)$ denote the group to which subject i belongs. The data on subject i are represented by $(g(i), t_{ij}, y_{ij})$, $j = 1, ..., \breve{n}_i$, where y_{ij} is the response measured at time t_{ij}. We will denote by $\boldsymbol{y}_i$ the response vector on subject i and by $\boldsymbol{t}_i$ the corresponding vector of time points. For notational convenience, we assume that the data are entered in group order, and within each group these data are stored in subject order.

Some of the material in this Chapter has been described is Djeundje and Currie (2011a,b).

Our first example, which is shown in Figure 4.1, corresponds to the `CanadianWeather` data used in Chapter 3, except that the cities are now classified into four regions (groups); these are balanced data, ie, $\boldsymbol{t}_1 = ... = \boldsymbol{t}_n$. Our second example, which is

displayed in Figure 4.2, are simulated heights of 197 children (subjects) suffering from acute lymphoblastic leukaemia, and receiving three different treatments (groups); a full description of these data can be found in Durban et al. (2005); these are unbalanced data and we refer to them as `ChildHeight`.

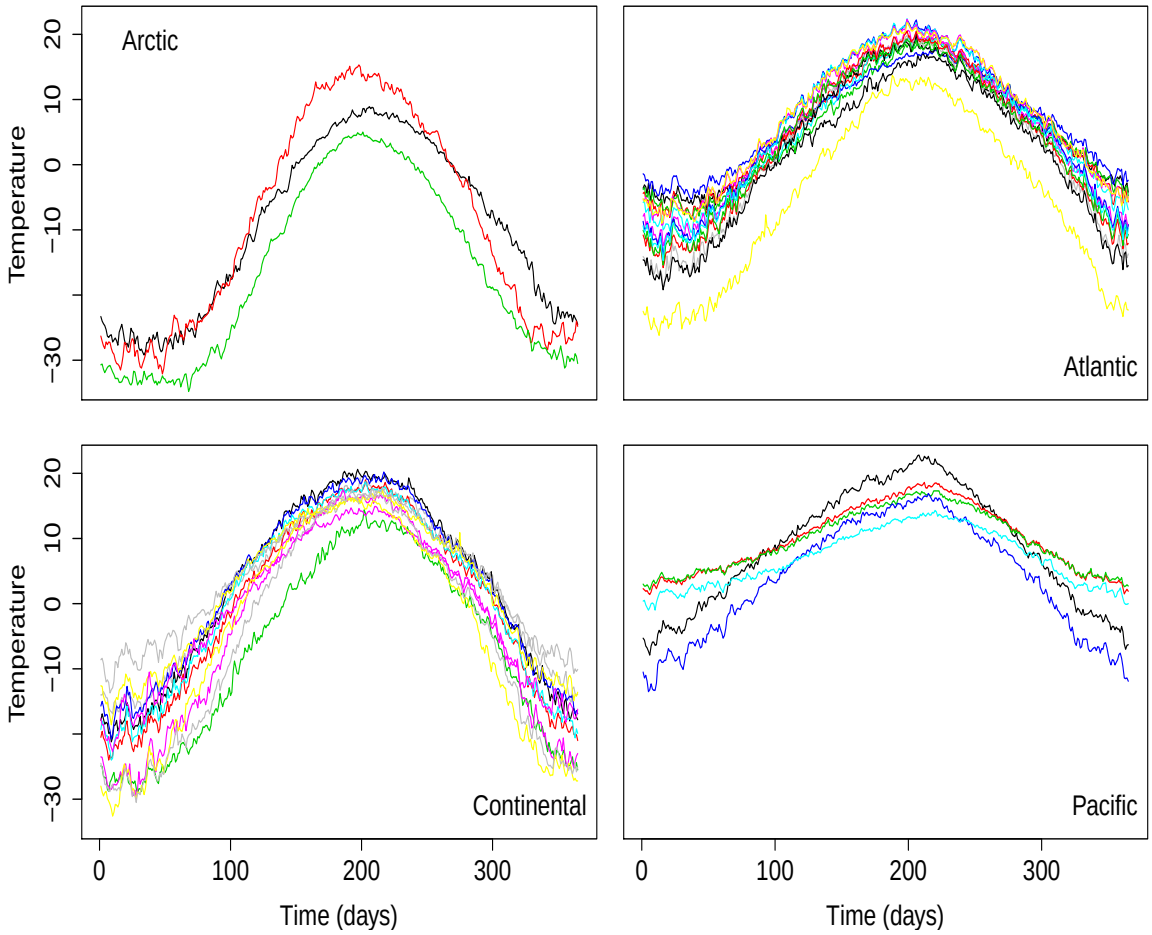

Figure 4.1: *Canadian weather data in 35 cities split into four regions.*

The usual interest lies in the group and subject effects for such data, and so we consider models of the form

$$\boldsymbol{y}_i = \mathcal{S}_{g(i)}(\boldsymbol{t}_i) + \breve{\mathcal{S}}_i(\boldsymbol{t}_i) + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{\mathcal{W}}_i^{-1}), \tag{4.1}$$

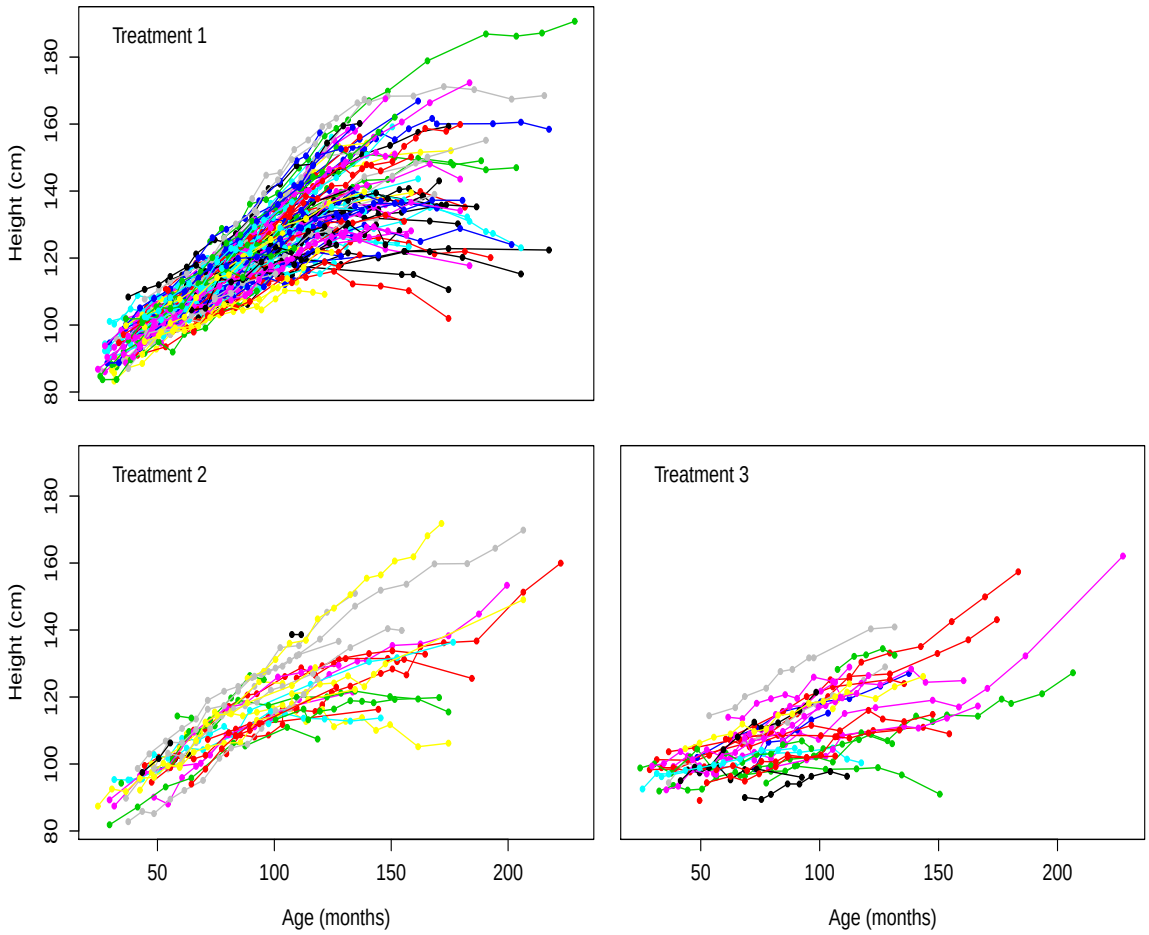

Figure 4.2: *Heights of 197 children suffering from acute lymphoblastic leukaemia, and receiving three different treatments.*

where $\mathcal{S}_{g(i)}(\cdot)$ measures the group effect to which subject i belongs, and $\breve{\mathcal{S}}_i(\cdot)$ captures the i th subject effect relative to its group effect. The precision matrices $\boldsymbol{\mathcal{W}}_i$ are set to the identity in many applications, although a more general covariance structure for the joint vector of observations is permitted (at least in principle).

In (4.1), the functions $\mathcal{S}_{g(i)}(\cdot)$ and $\breve{\mathcal{S}}_i(\cdot)$ are designed to capture the underlying patterns in the data, and if the structure of the data allows, a simple approach may be to treat them as straight lines or as some low degree polynomials. In general however we want $\mathcal{S}_{g(i)}(\cdot)$ and $\breve{\mathcal{S}}_i(\cdot)$ to be sufficiently flexible so they can reflect the true dynamism driving the data.

At first glance, model (4.1) looks equivalent to a data splitting approach where

data are split according to groups and then smooth models are fitted independently to each group. However, the covariance structure across groups in (4.1) offers several benefits: first, it can cope with both homoskedasticity and heteroskedasticity; second, it can handle both isotropic and anisotropic smoothing at both levels; third, it can be extended to cope with several covariates whose effects on the response can be group dependent, as we shall see in Section 4.6.

The structure of this Chapter is as follows. Section 4.1 investigates in detail the discrepancy and widening problems arising from the standard model described in Section 3.2. Section 4.2 extends the penalty approach of Section 3.3 to the setting of grouped and unbalanced data. Section 4.3 describes the implementation of the penalty approach via restricted maximum likelihood with best unbiased estimator/predictor. Section 4.4 deals with computational issues. Section 4.5 illustrates the consistency of the penalty approach, first on our two data sets, and next on simulated data. Section 4.6 outlines the extension of the model to the multivariate setting, and we close with a brief discussion in Section 4.7.

4.1 Standard model and more illustrations

In terms of truncated lines, the components of model (4.1) are obtained by extension of (3.3) as

$$\mathcal{S}_{g(i)}(\boldsymbol{t}_i) = [\mathbf{1}_{\breve{n}_i} : \boldsymbol{t}_i]\, \boldsymbol{a}_{g(i)} + \boldsymbol{T}_{g(i),i}\, \boldsymbol{\xi}_{g(i)} \quad \text{and} \quad \breve{\mathcal{S}}_i(\boldsymbol{t}_i) = [\mathbf{1}_{\breve{n}_i} : \boldsymbol{t}_i]\, \breve{\boldsymbol{a}}_i + \breve{\boldsymbol{T}}_i \breve{\boldsymbol{\xi}}_i, \tag{4.2}$$

where the $\boldsymbol{T}_{g(i),i}$ and $\breve{\boldsymbol{T}}_i$ are given by

$$\left.\begin{array}{rcl} \boldsymbol{T}_{g(i),i} & = & \begin{bmatrix} (t_{i,1}-\kappa_1)_+ & \cdots & (t_{i,1}-\kappa_q)_+ \\ \vdots & \ddots & \vdots \\ (t_{i,\breve{n}_i}-\kappa_1)_+ & \ldots & (t_{i,\breve{n}_i}-\kappa_q)_+ \end{bmatrix} \\ \breve{\boldsymbol{T}}_i & = & \begin{bmatrix} (t_{i,1}-\breve{\kappa}_1)_+ & \cdots & (t_{i,1}-\breve{\kappa}_{\breve{q}})_+ \\ \vdots & \ddots & \vdots \\ (t_{i,\breve{n}_i}-\breve{\kappa}_1)_+ & \ldots & (t_{i,\breve{n}_i}-\breve{\kappa}_{\breve{q}})_+ \end{bmatrix} \end{array}\right\}, \tag{4.3}$$

in which $\{\kappa_1, ..., \kappa_q\}$ and $\{\breve{\kappa}_1, ..., \breve{\kappa}_{\breve{q}}\}$ are sets of equi-spaced knots at the group and subject levels.

Clearly, expressions (4.3) use truncated lines which run from left-to-right with slope +1 as illustrated in the upper panel in Figure 2.5. Equally, we can set up a truncated lines basis which runs from right-to-left with slope −1, by replacing the $(t_{ij} - \kappa_r)_+$ and $(t_{ij} - \breve{\kappa}_r)_+$ with $(-t_{ij} + \kappa_r)_+$ and $(-t_{ij} + \breve{\kappa}_r)_+$ respectively. We refer to these bases as the *forward basis* (slope +1) and the *backward basis* (slope −1) respectively.

As in the case of a single group discussed in Chapter 3, a standard way to address smoothness and identifiability in model (4.1) is through the following normal constraints on the coefficients (Coull et al., 2001a; Ruppert et al., 2003, sect 9.3; Durban et al., 2005):

$$\boldsymbol{\xi}_{g(i)} \sim \mathcal{N}\left(\mathbf{0}, \sigma_P^2 \boldsymbol{I}_q\right), \quad \breve{\boldsymbol{a}}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \quad \breve{\boldsymbol{\xi}}_i \sim \mathcal{N}\left(\mathbf{0}, \breve{\sigma}^2 \boldsymbol{I}_{\breve{q}}\right). \tag{4.4}$$

Again, we will refer to (4.2) and (4.4) as the *standard model*.

In Chapter 3, we illustrated some problems encountered with this model in terms of its ability to appropriately extract the mean and subject effects in a single group. In the remainder of this Section, these problems are demonstrated in detail, first on our two data sets, and next through a simulation study.

4.1.1 Illustration 1: Canadian weather data

The `CanadianWeather` data displayed in Figure 4.1 contains 365 observations on each of 35 cities which are arranged in four groups. As in the case of one group in Section 3.2, we use the `lme` function to fit the standard model to these data. Figure 4.3 shows the data (wiggly black lines) for a subset of 12 cities, together with the corresponding global fit obtained under the forward (red dashed lines) and backward (green dashed lines) truncated line bases. Clearly, these global fits are nearly identical for both bases and the fitted curves are largely hidden under the data; ie, the overall fit to the data is very good. As in Section 3.2, it is easy to conclude that the goodness of fit at the global level implies that the underlying group and subject effects are correctly identified. However the two upper panels of Figure 4.4 show the fitted

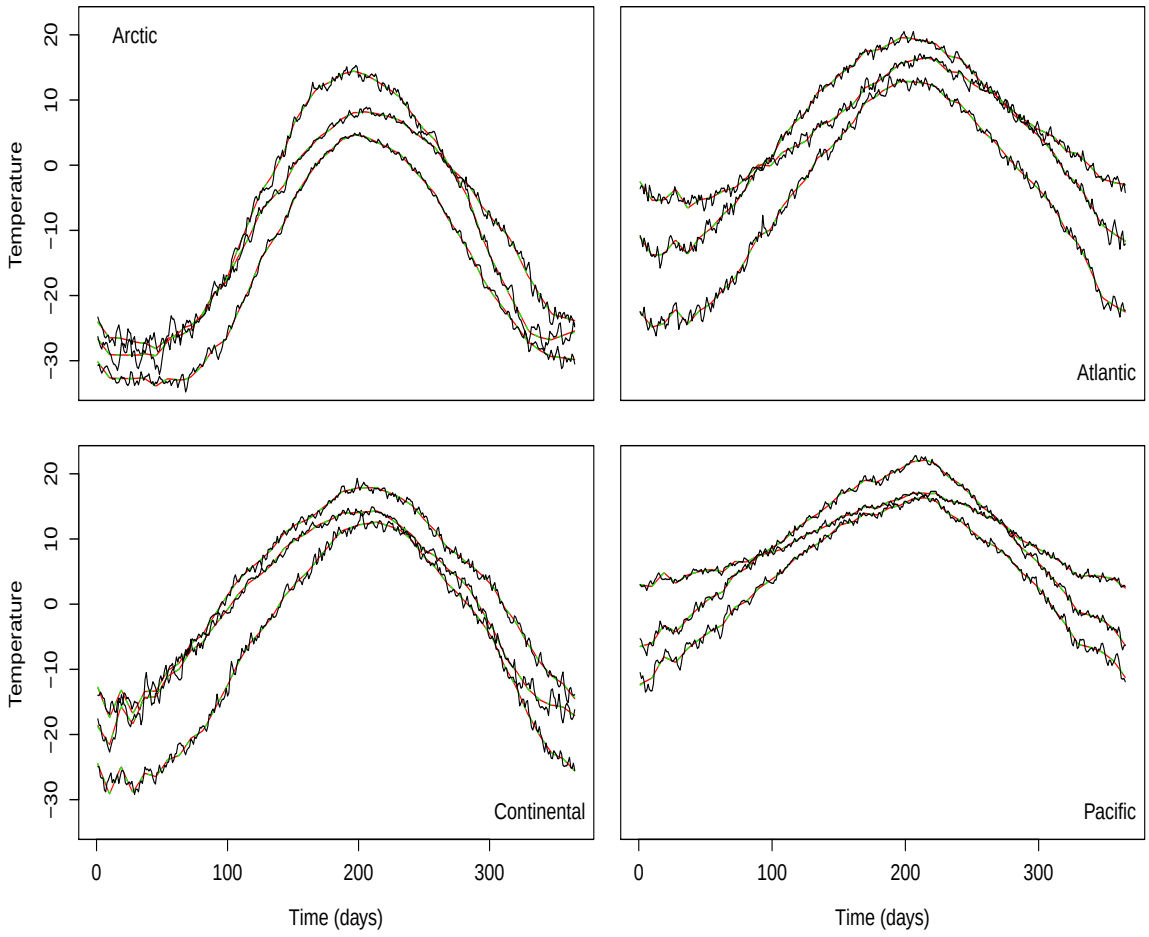

Figure 4.3: *Fitted cities using forward (red) and backward (green) bases, together with the observed data (black), for selected cities.*

group effects under the forward (red lines) and backward (green lines) bases, with the associated confidence bands (dashed lines); the group averages are also shown (black lines). Evidently, the fitted group effects are biased and basis-dependent, and the confidence bands exhibit a widening fan effect from left-to-right (under the forward basis) similar to that seen in Chapter 3, and from right-to-left (under the backward basis). The same behaviour is observed in the other two groups (not shown here). Further, we expect that within each group, the city effects should be roughly centred overall; yet, the lower panels of Figure 4.4 show that the fitted city effects (which are quite flexible in shape) are certainly not centred. In other words, the behaviour of the fitted group effects is balanced by a similar (opposite) behaviour at the city level,

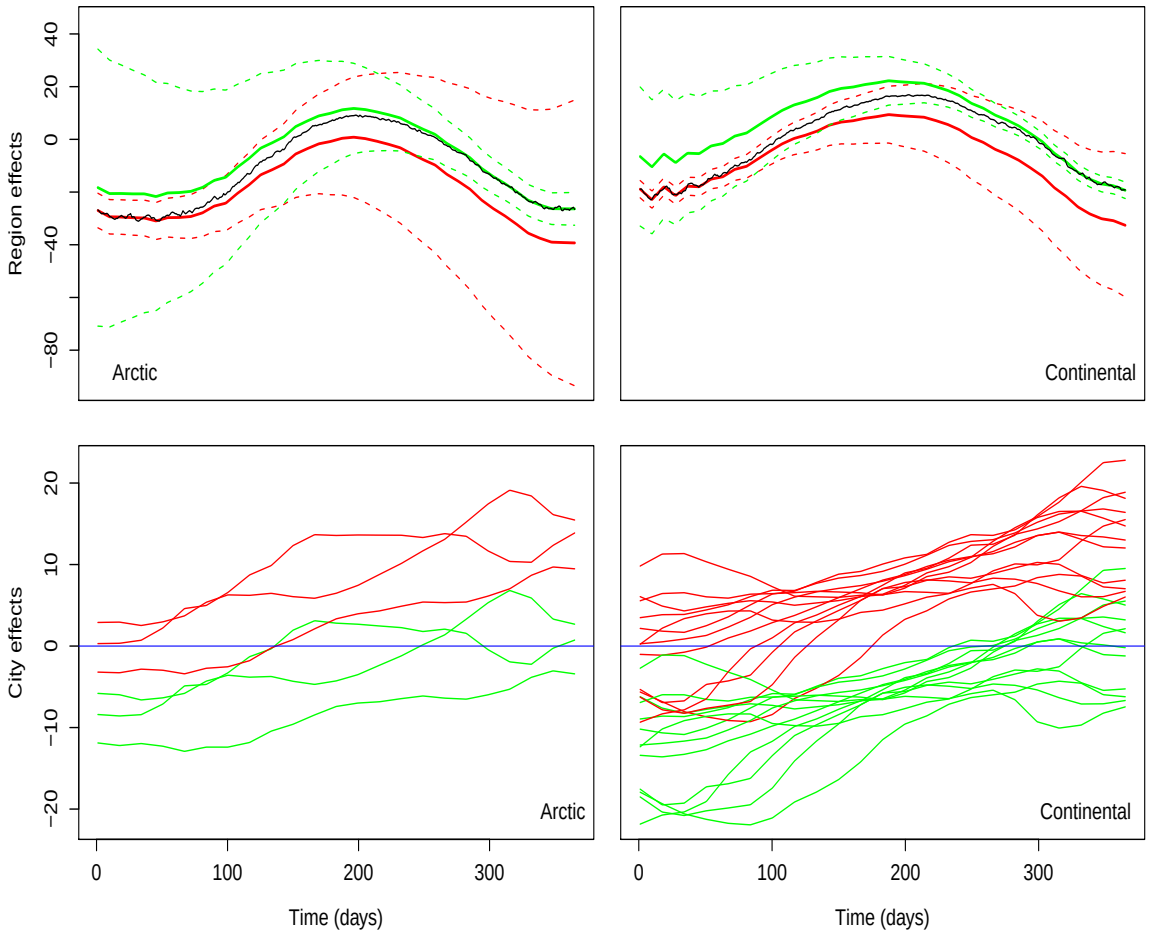

Figure 4.4: *The upper panels show the fitted region effects from the standard model under forward (red) and backward (green) bases, together with their confidence bands; the observed group average is also added (black). The lower panels show the fitted city effects.*

in such a way that the global fit is appropriately recovered as seen in Figure 4.3.

4.1.2 Illustration 2: Child height data

The `ChildHeight` data, displayed in Figure 4.2, are unbalanced with the number of observations per child varying from 1 to 21. The upper panels of Figure 4.5 show some discrepancy between the estimated group effects with the forward and backward bases but the difference is much less marked than for `CanadianWeather`. We note that with unbalanced data it is not easy to plot the treatment means directly from the data, as we could with `CanadianWeather`; thus, any bias in the estimated group effects is not

immediately evident. We also note that the fitted child effects in the lower panels look appropriately centered. In conclusion, unlike the analysis of `CanadianWeather` above, the application of the standard model to `ChildHeight` has been (arguably) successful. Why is this? A clue to the answer to this question can be seen by comparing the lower panels of Figure 4.4 and Figure 4.5. With `CanadianWeather` the subject effects are highly variable, but with `ChildHeight` they look close to linear. We will investigate this point through a simulation study.

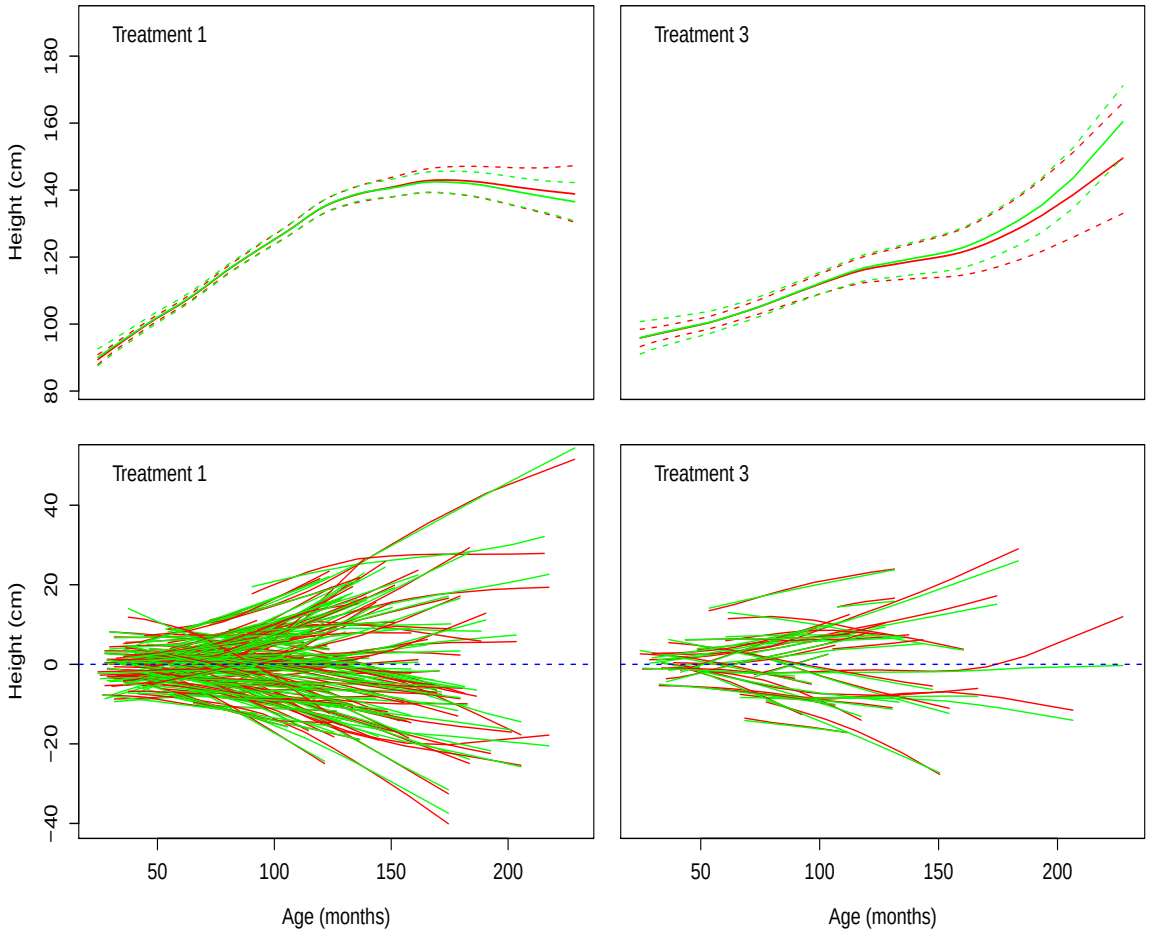

Figure 4.5: *The upper panels show the fitted treatment effects from the standard model under forward (red) and backward (green) bases. The lower panels display the fitted child effects.*

4.1.3 Illustration 3: Simulation study

For the purpose of these simulations, we suppose that subjects are divided into three groups ($m = 3$); we then define our true group effects by the following simple quadratic functions: $\mathcal{S}_k(t) = (t - \frac{k}{4})^2$, $t \in [0,\ 1]$, $k = 1, 2, 3$. We split our simulations into four scenarios.

- **Scenario 1:** We set the true subject curves to straight lines, defined by $\breve{\mathcal{S}}_i(t) = \breve{a}_{i,0} + \breve{a}_{i,1}t$, where $\breve{a}_{i,0} \sim \mathcal{N}(0,\ \sigma_0)$ and $\breve{a}_{i,1} \sim \mathcal{N}(0,\ \sigma_1)$, $i = 1, ..., n$. We then simulate (according to model (4.1)) balanced data for 3 groups of sizes 20, 25, and 15 subjects, with 25 observations on each subject at equi-spaced time points on $[0, 1]$. This is the balanced case with linear subject effects. The results presented in this Chapter use $(\sigma_0, \sigma_1, \sigma) = (0.25, 0.15,\ 0.1)$, and an illustration of such simulated data is shown in the upper left panel of Figure 4.6.

- **Scenario 2:** We set the true subject curves to straight lines as in scenario 1, and we then simulate (according to model (4.1)) unbalanced data for 3 groups in the same structure as `ChilHeight` data. Here we translate and scale the joint time vector so that it lies on $[0, 1]$. An illustration of such simulated data is shown in the lower left panel of Figure 4.6. This is the unbalanced case with linear subject effects.

- **Scenario 3:** This is similar to scenario 1, except that we define the true subject curves by $\breve{\mathcal{S}}_i(t) = \breve{A}_i \times \sin(\breve{\phi}_i t + \breve{\varphi}_i)$, where $\breve{A}_i \sim \mathcal{N}(0,\ \sigma_A)$ and $\breve{\phi}_i, \breve{\varphi}_i \sim \mathcal{U}(0, \varsigma)$, $i = 1, ..., n$. The results presented in this Chapter use $(\sigma_A, \varsigma, \sigma) = (0.25, 2\pi, 0.1)$, and such simulated data are shown in the upper right panel in Figure 4.6. This is the balanced case with flexible subject effects. Note that large values of ς would increase the flexibility in these curves.

- **Scenario 4:** This is similar to scenario 2 except that the true subject curves are define as in scenario 3; here we will use $(\sigma_A, \varsigma, \sigma) = (0.25, 4\pi, 0.1)$. This is the unbalanced case with flexible subject effects; see the lower right panel of Figure 4.6 for an illustration.

One important point about these functions is that $\mathbb{E}[\breve{\mathcal{S}}_i(\boldsymbol{t}_i)] = \mathbf{0}$ in all four scenarios; this condition is necessary so that on average, these subject curves truly quantify the subject departures from their group effects.

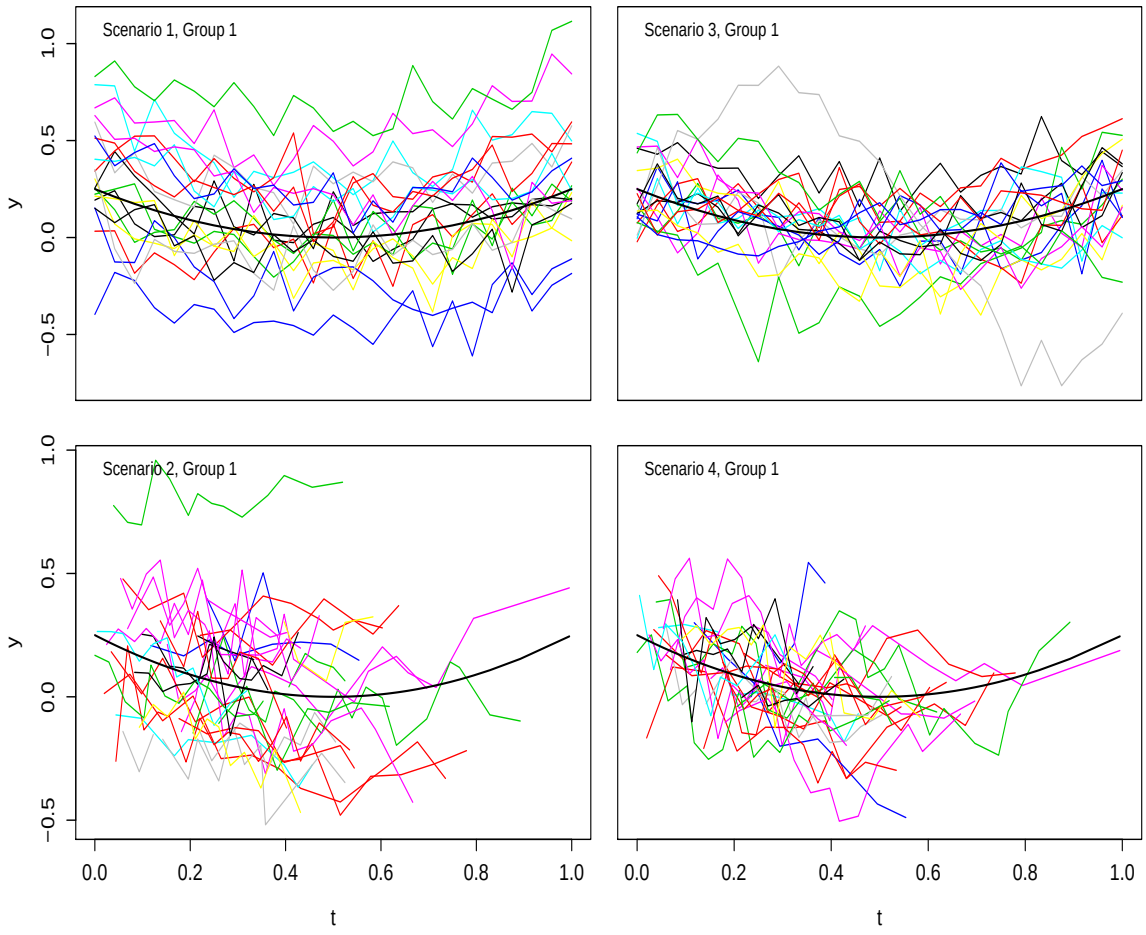

Figure 4.6: *An illustration of the simulated data under the four scenarios. The true group curves are also added (thick black line).*

We now want to evaluate the standard model in terms of its ability to recover the true underlying group curves as well as the behaviour of the related confidence bands. For this purpose, we define the vector of time points $\boldsymbol{x}_k = vec(\{\boldsymbol{t}_i : \text{ subject } i \in \text{group } k\})$ for each group $k \in \{1, 2, 3\}$; in other words, $\boldsymbol{x}_k$ is the joint vector of all time points for group k. Let n_k denotes the length of $\boldsymbol{x}_k$; ie n_k is the number of observations in group k. We proceed as follows:

- Under each scenario, we perform and store $N = 100$ sets of simulations
- For each simulated set r, $1 \leq r \leq N$,
  - Fit the standard model and obtain the fitted group effect $\hat{\mathcal{S}}_k^{(r)}(\boldsymbol{x}_k)$.

– Compute the standard deviation $SD_k^{(r)}(\boldsymbol{x}_k)$ about $\hat{\mathcal{S}}_k^{(r)}(\boldsymbol{x}_k)$ in each group k.

– Compute the mean square error

$$MSE_k^{(r)}(\boldsymbol{x}_k) = \|\hat{\mathcal{S}}_k^{(r)}(\boldsymbol{x}_k) - \mathcal{S}_k(\boldsymbol{x}_k)\|^2/n_k,$$

in each group.

– Obtain the overall mean square error as

$$MSE^{(r)} = \sum_{k=1}^{m} MSE_k^{(r)}(\boldsymbol{x}_k)/m\cdot$$

- For each group $k \in \{1, 2, 3\}$, compute the average standard deviation as

$$SD_k(\boldsymbol{x}_k) = \sum_{r=1}^{N^\star} SD_k^{(r)}(\boldsymbol{x}_k)/N\cdot$$

In the upper panels of Figure 4.7, the left and middle boxplots show the $MSE^{(r)}$ obtained under the forward and backward bases in scenarios 1 and 2, while the boxplot on the right refers to the penalty approach (which will be detailed below). Clearly, in these two scenarios respectively, the $MSE^{(r)}$ from the three approaches (forward, backward, penalty) look similar; the same closeness is observed about mean square error $MSE_k^{(r)}$ per group (not shown here). Furthermore, the upper panels of Figure 4.8 show the mean standard deviations $SD_k(\boldsymbol{x}_k)$ (on the scale of the group curves) in these two scenarios; here each line style refers to a specific group and the red lines are hidden by the green and black lines; again, these $SD_k(\boldsymbol{x}_k)$ from the three approaches are very similar (at least on this scale). In contrast, the equivalent graphics obtained in scenarios 3 and 4 are shown in the lower panels; these graphics show that the standard deviations are highly basis-dependent and they exhibit the same kind of widening effect encountered with `CanadianWeather` in Section 4.1.1. However, we see that the black lines, which correspond to the penalty approach, give a more reassuring result. Additionally, the $MSE^{(r)}$ from the penalty approach (boxplots on the right in the lower panels of Figure 4.7) are significantly smaller than those obtained from the standard model (with both the forward and backward bases). Investigation of different values of the parameters $(\varsigma, \sigma_0, \sigma_1, \sigma_A, \sigma)$ supports the conclusion that with

the standard model the discrepancy increases as the underlying subject curves become more flexible. Unfortunately, in the unbalanced case we will rarely have a clear view

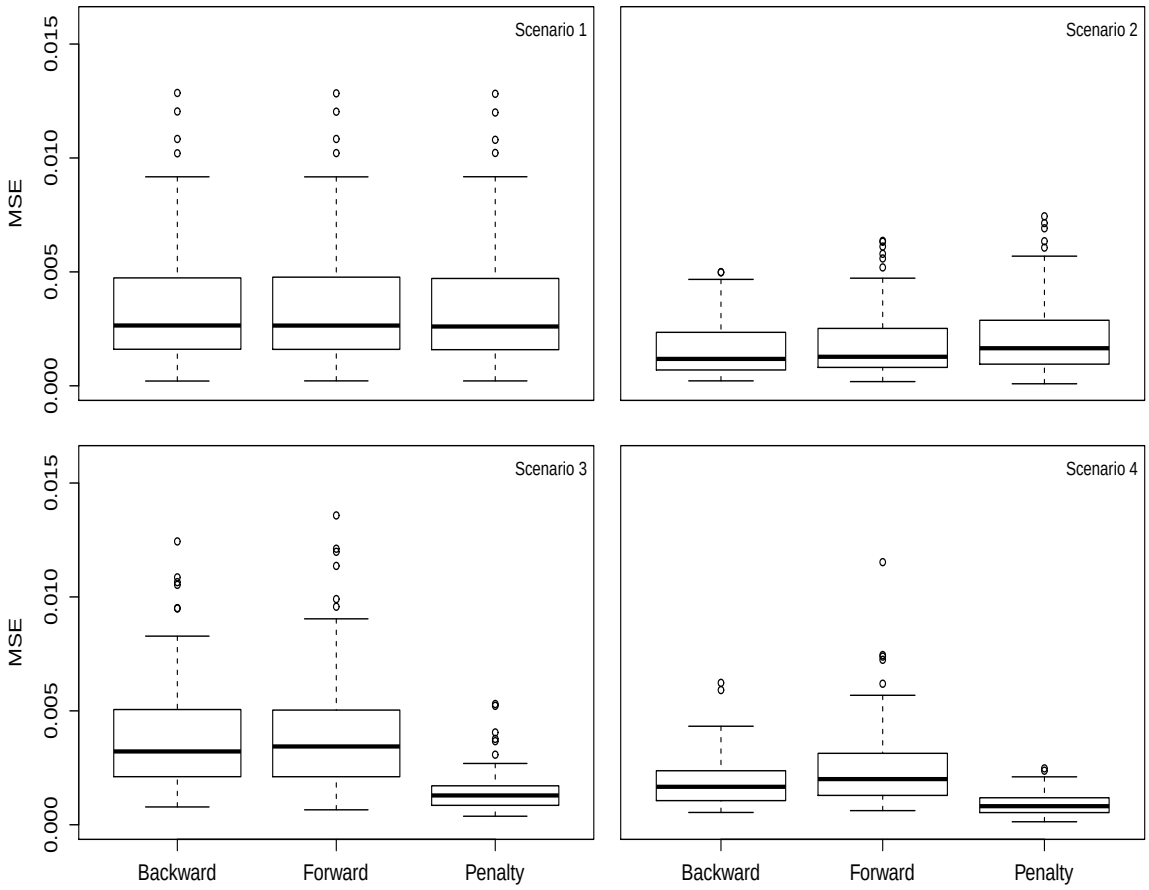

Figure 4.7: *Boxplots of the mean square errors, $MSE^{(r)}$, respectively from the standard model (with forward and backward bases), and the penalty approach.*

of the flexibility of these effects prior to fitting them. In the balanced case subtracting the group means from the observed values does give an idea of the flexibility of the subject effects; however, when these effects are found to be flexible the standard model is unable to capture them appropriately.

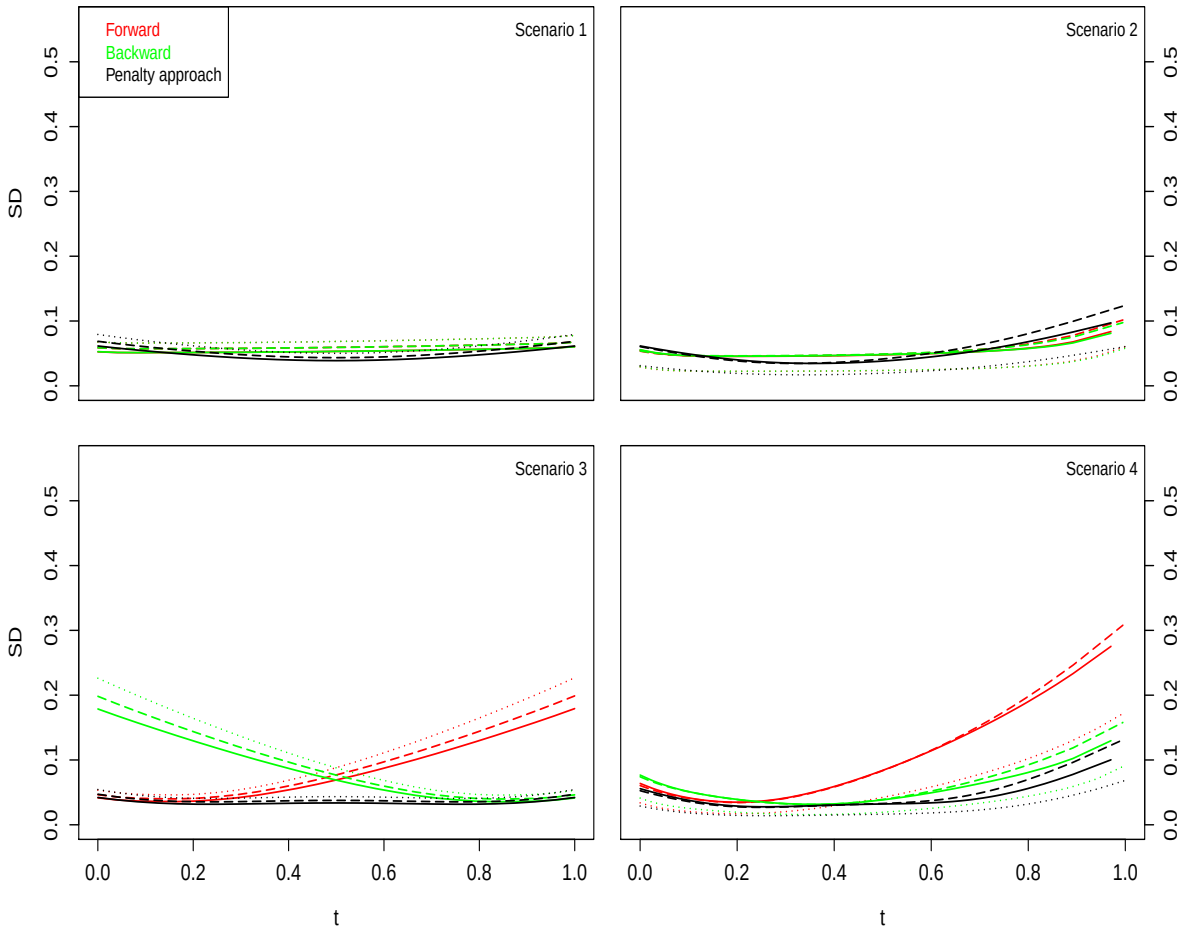

Figure 4.8: *Mean standard deviations $SD_k(\boldsymbol{x}_k)$ (on the scale of the group effects) respectively from the standard model (red, green), and the penalty approach (black). Here, each line style refers to a specific group, and the red lines are largely hidden by the green and black ones in the upper panel panels.*

In the remainder of this Chapter, we address these problems by extending the penalty approach presented in Chapter 3 to the unbalanced and grouped setting. We also deal with a computational issue: with unbalanced data it is inefficient to use the same set of knots irrespective of the number of observations on a particular subject; for example, with the `ChildHeight` data, some subjects have just one observation, while others have more than twenty. Finally, although our models are derived from penalty arguments, we will detail the implementation using the mixed model representation.

4.2 Adaptive knots and penalty approach

We recall that in the context of one level smoothing, as discussed in Chapter 2, a penalty is placed on a rich spline basis to achieve smoothness. With this argument, one is allowed to use more knots than the number of data points since the penalty allows us to smooth away the wiggliness of the predictor and to take care of any singularity problem that may occur (provided that the number of data points is higher than the degree of the underlying unpenalized polynomial). However, in the context of hierarchical curves, each group/subject brings an additional set of parameters into the model and consequently, a large number of knots/basis per subject makes the fitting computationally intensive. A common approach (Coull et al., 2001a; Ruppert et al., 2003, sect 9.3; Durban et al., 2005; etc) is to use the same sets of knots $\boldsymbol{\kappa} = \{\kappa_1, ..., \kappa_q\}$ for all groups, and $\breve{\boldsymbol{\kappa}} = \{\breve{\kappa}_1, ..., \breve{\kappa}_{\breve{q}}\}$ for all subjects, as described in Chapter 3 and in expression (4.3). Such a knot specification is reasonable in balanced situations; but it turns out to be inefficient for unbalanced data since the number of observations can differ considerably from one group/subject to the other.

In order to avoid unnecessarily large numbers of knots/basis, we can use the following simple procedure. For each subject i, we set $\max\{4, \min\{\text{round}(\breve{n}_i/4), 40\}\}$ equi-spaced knots on the range of $\boldsymbol{t}_i$. In this way, individual knots are placed only in the useful range of the covariate, and their number depends only on the number of observations on this subject. Moreover this knot allocation can be streamlined by simply fitting a point or a line to subjects with one or two observations respectively. Evidently, a similar procedure can be applied at the group level; ie for a given group k, we can set $\max\{4, \min\{\text{round}(\tilde{n}_k/4), 40\}\}$ equi-spaced knots on the range of $\boldsymbol{x}_k$, where $\boldsymbol{x}_k$ and $\tilde{n}_k$ represent the joint time vector and the number of unique time points in group k respectively. We refer to this scheme as *adaptive knots*.

Hence, extending the penalty approach presented in Section 3.3 and summarised by equation (3.15), we express the components of model (4.1) as

$$\mathcal{S}_{g(i)}(\boldsymbol{t}_i) = \boldsymbol{\Omega}_{g(i),i}\, \boldsymbol{\alpha}_{g(i)} \quad \text{and} \quad \breve{\mathcal{S}}_i(\boldsymbol{t}_i) = \breve{\boldsymbol{\Omega}}_i \breve{\boldsymbol{\alpha}}_i, \tag{4.5}$$

where $\boldsymbol{\Omega}_{g(i),i}$ and $\breve{\boldsymbol{\Omega}}_i$ are spline matrices (with adaptive knots) at the group and subject levels, built along the time vector $\boldsymbol{t}_i$, and $\boldsymbol{\alpha}_{g(i)}$ and $\breve{\boldsymbol{\alpha}}_i$ are the associated regression

coefficients. We will denote by c_k and $\breve{c}_i$ the lengths of $\boldsymbol{\alpha}_k$ and $\breve{\boldsymbol{\alpha}}_i$ respectively, $k = 1, ..., m$ and $i = 1, ..., n$. That is, the dimensions of $\boldsymbol{\Omega}_{g(i),i}$ and $\breve{\boldsymbol{\Omega}}_i$ are $\breve{n}_i \times c_{g(i)}$ and $\breve{n}_i \times \breve{c}_i$ respectively. For balanced data sets such as `CanadianWeather`, all the group regression matrices $\boldsymbol{\Omega}_{g(i),i}$ become identical; the same remark applies to the subject regression matrices $\breve{\boldsymbol{\Omega}}_i$.

With these components in place, we address smoothness and identifiability as in Section 3.3, depending on whether B-splines or truncated polynomials are used. With B-spline bases for instance, we obtain smoothness by a difference penalty on the $\boldsymbol{\alpha}_{g(i)}$ and $\breve{\boldsymbol{\alpha}}_i$, and for identifiability we additionally place a ridge penalty on the subject coefficients $\breve{\boldsymbol{\alpha}}_i$ as in (3.7). With full truncated polynomial bases, we deal with the smoothness issue through a ridge penalty on the components of $\boldsymbol{\alpha}_{g(i)}$ and $\breve{\boldsymbol{\alpha}}_i$ corresponding to the truncated basis, and we solve the identifiability problem by shrinking the fitted subject effects as in (3.10). In both cases, the resulting penalized weighted residual sum of squares takes the form

$$\mathsf{PRSS} = \sum_{i=1}^{n} \|\boldsymbol{\mathcal{W}}^{1/2}(\boldsymbol{y}_i - \boldsymbol{\Omega}_{g(i),i}\,\boldsymbol{\alpha}_{g(i)} - \breve{\boldsymbol{\Omega}}_i\breve{\boldsymbol{\alpha}}_i)\|^2 + \sum_{k=1}^{m} \boldsymbol{\alpha}_k'\boldsymbol{P}_k\boldsymbol{\alpha}_k + \sum_{i=1}^{n} \breve{\boldsymbol{\alpha}}_i'\breve{\boldsymbol{P}}_i\breve{\boldsymbol{\alpha}}_i, \quad (4.6)$$

which is an extension of the penalized residual sum of squares given by (3.20). In (4.6), $\boldsymbol{P}_k$ is the penalty matrix achieving the smoothness of the k th group effect in a similar fashion as $\boldsymbol{P}_P$ does in (3.22), except that the size of the $\boldsymbol{P}_k$ are group-dependent. If we assume isotropic smoothing for all groups, then these $\boldsymbol{P}_k$ can be expressed as $\boldsymbol{P}_k = \boldsymbol{P}_k(\lambda)$, where λ is the common smoothing parameters for the group curves. The $\breve{\boldsymbol{P}}_i$ control the smoothness of the subject effects and the identifiability of the model in the same way as $\breve{\boldsymbol{P}}$ in (3.23). Hence, these $\breve{\boldsymbol{P}}_i$ take the form $\breve{\boldsymbol{P}}_i = \breve{\boldsymbol{P}}_i(\breve{\lambda}_1, \breve{\lambda}_2)$, where $\breve{\lambda}_1$ and $\breve{\lambda}_2$ represent the smoothing and shrinkage parameters on the subject effects.

With these details, we can fit the model in the weighted least squares sense by minimizing (4.6) with respect to the regression parameters, and use criteria like AIC, BIC or GCV to estimate the smoothing/identifiability parameters as we did in Chapter 3. Nowadays however, smooth models are often expressed as mixed models, since estimates of the variance/smoothing parameters obtained from the mixed model representation and restricted likelihood tend to behave well (Reiss and Ogden, 2009; Wang, 1998). Further, our data sets and simulation study display a mixed model structure; ie, it is reasonable to assume that the subjects are a random sample from

some population of subjects.

4.3 Mixed model representation and inference

Whether B-spline bases or full truncated polynomial bases are used, the penalty matrices $\boldsymbol{P}_k$ are singular. Hence, following Sections 2.3.3 and 3.5, we partition the group effects $\boldsymbol{\Omega}_{g(i),i}\,\boldsymbol{\alpha}_{g(i)}$ into unpenalized and penalized components as

$$\boldsymbol{\Omega}_{g(i),i}\,\boldsymbol{\alpha}_{g(i)} = \boldsymbol{X}_{g(i),i}\,\boldsymbol{\beta}_{g(i)} + \boldsymbol{Z}_{g(i),i}\,\boldsymbol{b}_{g(i)}, \tag{4.7}$$

where the $\boldsymbol{\beta}_{g(i)}$ represent the vectors of unpenalized coefficients, and the coefficient vectors $\boldsymbol{b}_{g(i)}$ are subject to the transformed penalty $\tilde{\boldsymbol{P}}_{g(i)} = \lambda \boldsymbol{I}_{c_{g(i)}-p-1}$. Also, $\boldsymbol{X}_{g(i),i}$ and $\boldsymbol{Z}_{g(i),i}$ are appropriate regression matrices; precisely, we have $\boldsymbol{X}_{g(i),i} = [\boldsymbol{1}_{\breve{n}_i} : ... : \breve{\boldsymbol{t}}^p_{\breve{n}_i}]$, and the explicit form of $\boldsymbol{Z}_{g(i),i}$ is similar to that of $\boldsymbol{Z}_p$ in equation (3.29), depending on whether B-splines or truncated polynomial bases are used. With this re-parametrization, expression (4.6) becomes

$$\begin{aligned} \mathsf{PRSS} \;=\; & \sum_{i=1}^{n} \|\boldsymbol{\mathcal{W}}_i^{1/2}(\boldsymbol{y}_i - \boldsymbol{X}_{g(i),i}\,\boldsymbol{\beta}_{g(i)} - \boldsymbol{Z}_{g(i),i}\,\boldsymbol{b}_{g(i)} - \breve{\boldsymbol{\Omega}}_i\breve{\boldsymbol{\alpha}}_i)\|^2 \\ & +\; \lambda \sum_{k=1}^{m} \boldsymbol{b}_k'\boldsymbol{b}_k \;\;+\;\; \sum_{i=1}^{n} \breve{\boldsymbol{\alpha}}_i\breve{\boldsymbol{P}}_i\breve{\boldsymbol{\alpha}}_i. \end{aligned} \tag{4.8}$$

With $\boldsymbol{y} = vec(\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$, $\boldsymbol{\beta} = vec(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m)$ and $\boldsymbol{u} = vec(\boldsymbol{b}_1, ..., \boldsymbol{b}_m, \breve{\boldsymbol{\alpha}}_1, ..., \breve{\boldsymbol{\alpha}}_n)$, minimizing the PRSS (4.8) with respect to the regression coefficients corresponds to the maximization of the log-likelihood of $(\boldsymbol{y}, \boldsymbol{u})$ arising from the mixed model representation

$$\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\; \sigma^2\boldsymbol{\mathcal{W}}^{-1}), \;\; \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0},\; \sigma^2\boldsymbol{P}^{-1}), \tag{4.9}$$

in which $\boldsymbol{\beta}$ is the fixed effect, $\boldsymbol{u}$ is the random effect,

$$\boldsymbol{\mathcal{W}} = blockdiag(\boldsymbol{\mathcal{W}}_1, ..., \boldsymbol{\mathcal{W}}_n) \tag{4.10}$$

is the global precision matrix, and

$$\boldsymbol{P} = blockdiag(\lambda\boldsymbol{I}_{c_1-p-1}, ..., \lambda\boldsymbol{I}_{c_m-p-1},\; \breve{\boldsymbol{P}}_1, ..., \breve{\boldsymbol{P}}_n) \tag{4.11}$$

is the full penalty matrix, where the $\breve{\boldsymbol{P}}_i$ are similar in form to $\breve{\boldsymbol{P}}$ defined in (3.23), the only difference being that the sizes of the $\breve{\boldsymbol{P}}_i$ are subject-dependent.

In this mixed model representation, the regression matrix $\boldsymbol{X}$ for the fixed effect is defined by

$$\boldsymbol{X} = blockdiag(\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_m); \tag{4.12}$$

where $\boldsymbol{\mathcal{X}}_k$ is obtained by stacking the matrices $\boldsymbol{X}_{g(i),i}$, with $g(i) = k$, on top of each other. Also, the regression matrix $\boldsymbol{Z}$ for the random effects is partitioned into the regression matrix for the group random effects, $\boldsymbol{Z}_{\text{gp}}$, and the regression matrix for the subject random effects, $\boldsymbol{Z}_{\text{subj}}$, as

$$\boldsymbol{Z} = [\boldsymbol{Z}_{\text{gp}} : \boldsymbol{Z}_{\text{subj}}], \tag{4.13}$$

with $\boldsymbol{Z}_{\text{gp}}$ and $\boldsymbol{Z}_{\text{subj}}$ defined as

$$\boldsymbol{Z}_{\text{gp}} = blockdiag(\boldsymbol{\mathcal{Z}}_1, \ldots, \boldsymbol{\mathcal{Z}}_m) \tag{4.14}$$

$$\boldsymbol{Z}_{\text{subj}} = blockdiag(\breve{\boldsymbol{\Omega}}_1, \ldots, \breve{\boldsymbol{\Omega}}_n), \tag{4.15}$$

where $\boldsymbol{\mathcal{Z}}_k$ is obtained by stacking the matrices $\boldsymbol{Z}_{g(i),i}$, with $g(i) = k$, on top of each other.

We shall now describe the fitting of model (4.5) or equivalently model (4.9) via restricted maximum likelihood with best linear unbiased estimator/predictor. This description, which we adapt from Bates (2011), provides a convenient computational platform.

4.3.1 Best linear unbiased estimator/predictor

There are two random vectors in (4.9): $\boldsymbol{y}$ and $\boldsymbol{u}$; to simplify the notation, we do not distinguish between random variables and their observations. $\boldsymbol{y}$ has been observed, but $\boldsymbol{u}$ is not observed. Statistical inference in this context is based on the conditional distribution $f_{\boldsymbol{u}|\boldsymbol{y}}(\cdot)$ of $\boldsymbol{u}$ given $\boldsymbol{y}$; see Bates (2011). After re-arrangement, we find

from (4.9)

$$\begin{aligned} f_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{u}) &\propto f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}) \times f_{\boldsymbol{u}}(\boldsymbol{u}) \\ &= |\boldsymbol{\mathcal{W}}|^{\frac{1}{2}} \times |\boldsymbol{P}|^{\frac{1}{2}} \times \left(2\pi\sigma^2\right)^{\frac{-N-K_2}{2}} \times \exp\left(-\frac{\mathsf{PRSS}(\boldsymbol{\beta},\boldsymbol{u})}{2\sigma^2}\right) \end{aligned} \tag{4.16}$$

where N represents the total number of observations, K_2 is the number of columns in $\boldsymbol{Z}$, and

$$\mathsf{PRSS}(\boldsymbol{\beta},\boldsymbol{u}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})'\boldsymbol{\mathcal{W}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}. \tag{4.17}$$

The coefficients ($\boldsymbol{\beta}$, $\boldsymbol{u}$) can now be estimated by their best linear unbiased estimator/predictor (BLUE/BLUP), denoted by $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{u}})$, which are defined as the conditional mode of $\boldsymbol{u}$ based on (4.16); (Pinheiro and Bates, 2000, chap 2). That is, $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{u}})$ minimize the PRSS (4.17), and so they satisfy the following system of Henderson equations

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{Z}+\boldsymbol{P} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{y} \end{pmatrix} \tag{4.18}$$

conditional on the parameters $(\lambda, \breve{\lambda}_1, \breve{\lambda}_2)$ and any other parameter vector $\boldsymbol{\omega}$ that $\boldsymbol{\mathcal{W}}$ might depend on. We will estimate these parameters by maximizing the restricted likelihood. In this way, although λ and $\breve{\lambda}_1$ will be treated as variance parameters in the estimation process, they are purely smoothing parameters, in the sense that they act on the shape of the corresponding effects only. In contrast, $\breve{\lambda}_2$ is a variance parameter for the subjects in the original sense of a mixed model; ie, $\breve{\lambda}_2$ controls the overall size of the subject effects, in the same way that the departures of the response data from the mean are modulated by σ^2. We will denote by $\boldsymbol{\lambda} = (\lambda, \breve{\lambda}_1, \breve{\lambda}_2, \boldsymbol{\omega}')'$ the vector containing these parameters.

4.3.2 Restricted maximum likelihood estimate for the smoothing and identifiability parameters

For a fixed value of $\boldsymbol{u}$, $f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y})$ corresponds to the likelihood of the parameters σ, $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, given the data $\boldsymbol{y}$. Since $\boldsymbol{u}$ is a non-observable random variable and $\boldsymbol{y}$ is observed,

the marginal likelihood, $L(\cdot)$, is obtained by integrating $f_{\boldsymbol{y}|\boldsymbol{u}}(\cdot)$ with respect to the marginal density $f_{\boldsymbol{u}}(\cdot)$ of $\boldsymbol{u}$; this gives

$$L(\sigma, \boldsymbol{\lambda}, \boldsymbol{\beta}) = \int_{\mathbb{R}^{K_2}} f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}) f_{\boldsymbol{u}}(\boldsymbol{u})\, d\boldsymbol{u}\cdot \tag{4.19}$$

On maximizing (4.19), we get the MLE for $(\sigma, \boldsymbol{\lambda}, \boldsymbol{\beta})$, and the estimate of $\boldsymbol{\beta}$ obtained in this way coincides with its BLUE in (4.18). However, the MLE of the variance parameters discards the degrees of freedom involved in the estimation of $\boldsymbol{\beta}$ and as a result, the MLE of the variance parameters tend to be underestimated (Pinheiro and Bates, 2000, chap 2). A solution to this issue is provided by the so-called restricted likelihood, L_R; Patterson and Thompson (1971). From a computational perspective (Laird and Ware, 1982), a convenient derivation of $L_R(\cdot)$ consists of assuming a uniform prior distribution about $\boldsymbol{\beta}$ and then integrating it out of $L(\cdot)$, ie

$$L_R(\sigma, \boldsymbol{\lambda}) = \int_{\mathbb{R}^{K_1}} L(\sigma, \boldsymbol{\lambda}, \boldsymbol{\beta})\, d\boldsymbol{\beta}, \tag{4.20}$$

where K_1 denotes the length of $\boldsymbol{\beta}$.

The standard approach to evaluate this integral is to first evaluate (4.19) and then (4.20). We can skip this first step and focus directly on the restricted likelihood (4.20) as follows. Let us denote by $\boldsymbol{L}$ the left Choleski factor of the matrix in (4.18), ie

$$\boldsymbol{L}\boldsymbol{L}' = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{Z} + \boldsymbol{P} \end{pmatrix}, \tag{4.21}$$

then, using the orthogonal projection properties of the least squares solution, the PRSS can be re-written as

$$\mathsf{PRSS}(\boldsymbol{\beta},\, \boldsymbol{u}) \;=\; \mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}}) + \|\boldsymbol{L}'(\boldsymbol{v} - \tilde{\boldsymbol{v}})\|^2, \quad \text{with } \boldsymbol{v} = vec(\boldsymbol{\beta},\, \boldsymbol{u})\cdot \tag{4.22}$$

Hence integral (4.20) becomes

$$\begin{aligned} L_R(\sigma, \boldsymbol{\lambda}) \quad = \quad & |\boldsymbol{\mathcal{W}}|^{\frac{1}{2}} \times |\boldsymbol{P}|^{\frac{1}{2}} \times (2\pi\sigma^2)^{\frac{-N-K_2}{2}} \times \exp\left(-\frac{\mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}})}{2\sigma^2}\right) \\ & \times \int_{\mathbb{R}^{K_1+K_2}} \exp\left(-\frac{\|\boldsymbol{L}'\boldsymbol{v}\|^2}{2\sigma^2}\right) d\boldsymbol{v} \end{aligned}$$

$$
\begin{aligned}
&= |\boldsymbol{\mathcal{W}}|^{\frac{1}{2}} \times |\boldsymbol{P}|^{\frac{1}{2}} \times (2\pi\sigma^2)^{\frac{-N-K_2}{2}} \times \exp(-\frac{\mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}})}{2\sigma^2}) \times \frac{(2\pi\sigma^2)^{\frac{K_1+K_2}{2}}}{|\boldsymbol{L}|} \\
&= (2\pi\sigma^2)^{\frac{-N+K_1}{2}} \times |\boldsymbol{\mathcal{W}}|^{\frac{1}{2}} \times |\boldsymbol{P}|^{\frac{1}{2}} \times |\boldsymbol{L}|^{-1} \times \exp(-\frac{\mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}})}{2\sigma^2}).
\end{aligned}
$$

ie,

$$
\begin{aligned}
-2\log L_R(\sigma, \boldsymbol{\lambda}) = \;& (N-K_1)\log(2\pi\sigma^2) - \log|\boldsymbol{\mathcal{W}}| - |\log|\boldsymbol{P}| \\
&+ \; 2\log|\boldsymbol{L}| + \frac{\mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}})}{\sigma^2}
\end{aligned}
\tag{4.23}
$$

For a given $\boldsymbol{\lambda}$, we obtain $\sigma^2 = \sigma^2(\boldsymbol{\lambda}) = \mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}})/(N-K_1)$ on minimizing (4.23). On substitution back into (4.23), we obtained the profile deviance

$$
\begin{aligned}
\ell_R(\boldsymbol{\lambda}) &= -2\log L_R(\sigma(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \\
&= -\log|\boldsymbol{\mathcal{W}}| - \log|\boldsymbol{P}| + 2\log|\boldsymbol{L}| + (N-K_1)\log\Big(\mathsf{PRSS}(\tilde{\boldsymbol{\beta}},\, \tilde{\boldsymbol{u}})\Big) \;+\; \text{cst},
\end{aligned}
\tag{4.24}
$$

which is a function of $\boldsymbol{\lambda}$ alone. The REstricted Maximum Likelihood (REML) estimate for $\boldsymbol{\lambda}$ is then obtained as the minimizer of $\ell_R(\cdot)$.

We prefer (4.24) over the standard form

$$
\ell_R(\boldsymbol{\lambda}) = \log|\boldsymbol{V}| + \boldsymbol{y}'\boldsymbol{V}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})\boldsymbol{y} + \log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| + \text{cst} \tag{4.25}
$$

(with $\boldsymbol{V} = cov(\boldsymbol{y}) = \sigma^2\boldsymbol{Z}'\boldsymbol{P}^{-1}\boldsymbol{Z} + \sigma^2\boldsymbol{\mathcal{W}}^{-1}$) used by Ruppert et al. (2003) and Currie et al. (2006) because, besides the simplicity of (4.24), the efficient computation of its components is straightforward, as we shall see in Section 4.4.

4.3.3 Bias adjusted confidence bands

Using Bayesian arguments as in Section (2.5), we obtain the posterior covariance of the regression coefficients $vec(\boldsymbol{\beta}, \boldsymbol{u})$ as

$$
cov[vec(\boldsymbol{\beta}, \boldsymbol{u})|\boldsymbol{y}] = (\boldsymbol{\Omega}'\boldsymbol{\mathcal{W}}\boldsymbol{\Omega} + \sigma^2\boldsymbol{S})^{-1}, \tag{4.26}
$$

with

$$
\boldsymbol{\Omega} = [\boldsymbol{X} : \boldsymbol{Z}] \;\text{ and }\; \boldsymbol{S} = blockdiag(\boldsymbol{0}_{K_1},\, \boldsymbol{P}).
$$

The covariance matrix for any sub-vector from $vec(\boldsymbol{\beta}, \boldsymbol{u})$ is then obtained by taking the appropriate block diagonal components of $cov[vec(\boldsymbol{\beta}, \boldsymbol{u})|\boldsymbol{y}]$ in (4.26). For instance the bias adjusted covariance for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ corresponds respectively to the upper $K_1 \times K_1$ and lower $K_2 \times K_2$ block diagonal components in $cov[vec(\boldsymbol{\beta}, \boldsymbol{u})|\boldsymbol{y}]$. It is this procedure that we use to derive the covariance matrices of the group/subject coefficients, which are used in turn to compute confidence bands around the corresponding smoothers/effects. We remark that the confidence bands obtained from (4.26) are identical to the bias adjusted confidence bands described in Ruppert et al. (2003, chap 6), provided the same values of the variance parameters are used in the two perspectives.

4.4 Computational considerations

We divide the computations into three steps:

Step 1: Obtain the REML estimate of $\hat{\boldsymbol{\lambda}}$ using (4.24).

Step 2: Set $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\lambda}})$, $\hat{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\hat{\boldsymbol{\lambda}})$ and $\hat{\sigma} = \sigma(\hat{\boldsymbol{\lambda}})$

Step 3: Compute the effects with their confidence bands.

The first step is the most intensive one as it involves optimization of the REML criterion $\ell_R(\cdot)$, and this latter depends on many (potentially large) components. Our approach to this optimization problem depends on whether the precision component $\boldsymbol{\mathcal{W}}$ is known or not. In many applications, $\boldsymbol{\mathcal{W}}$ is the identity matrix, but in a more general framework, it may be unknown with some specific structure. Succinctly, we proceed as follows:

- If $\boldsymbol{\mathcal{W}}$ is known, then

  Compute and store the invariant components $\boldsymbol{X'\mathcal{W}X}$, $\boldsymbol{Z'\mathcal{W}Z}$, $\boldsymbol{Z'\mathcal{W}X}$, $\boldsymbol{X'\mathcal{W}y}$ and $\boldsymbol{Z'\mathcal{W}y}$.

  For each trial value of $\boldsymbol{\lambda}$,

  (1) Compute $\boldsymbol{P}$ in (4.11) and form the left Choleski factor $\boldsymbol{L}$ in (4.21)

  (2) Solve $\boldsymbol{L\theta} = \begin{pmatrix} \boldsymbol{X'\mathcal{W}y} \\ \boldsymbol{Z'\mathcal{W}y} \end{pmatrix}$ in $\boldsymbol{\theta}$, and then $\boldsymbol{L'} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \boldsymbol{\theta}$

(3) Compute $\log|\boldsymbol{P}|$, $\log|\boldsymbol{L}|$ and $\mathsf{PRSS}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{u}})$

(4) Evaluate the REML criterion $\ell_R(\boldsymbol{\lambda})$ using (4.24)

- Otherwise,

  For each trial value of $\boldsymbol{\lambda}$,

  (i) Compute $\boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{X}$, $\boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{Z}$, $\boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{\mathcal{W}}\boldsymbol{y}$ and $\boldsymbol{Z}'\boldsymbol{\mathcal{W}}\boldsymbol{y}$. In practice, a direct plug-in computation of these components for all trial values of $\boldsymbol{\lambda}$ can be time consuming. However, in many cases the efficiency can be improved depending on the specific structure of both the data and $\boldsymbol{\mathcal{W}}$.

  (ii) Perform (1), (2), (3) and (4) above.

In some applications, $\boldsymbol{P}$ can be a large matrix and the computer may return $\log|\boldsymbol{P}| = \infty$ (if computed directly), while $|\boldsymbol{P}| < \infty$. However, $\boldsymbol{P}$ has a block diagonal structure and each of its block components is generally of moderate size; hence $\log|\boldsymbol{P}|$ can be obtained efficiently as the sum of the logarithm of the determinants of these blocks. Also, $\boldsymbol{L}$ is a triangular positive definite matrix and so, $\log|\boldsymbol{L}|$ is simply obtained as the sum of the logarithm of its diagonal elements. Further, since the penalty does not act on the fixed coefficient $\boldsymbol{\beta}$, then, if K_1 (the length of $\boldsymbol{\beta}$) is large, the above algorithm can be refined via a partition of $\boldsymbol{L}$ with regard to the blocks of the right hand side matrix in (4.21); this will then allow one part of $\boldsymbol{L}$ to be computed and stored outside of the optimization routine, and then only the remaining part will be computed for each trial value of $\boldsymbol{\lambda}$.

4.5 Applications

In this Chapter, we are not specifically interested in the interpretation of the polynomial part of the group effects as in Chapter 3, and so, for our illustrations here, we shall simply use cubic B-splines with second order difference penalty at both levels; also, we will set $\boldsymbol{\mathcal{W}}$ to the identity matrix.

First, we consider `CanadianWeather` data. Two fitted region effects are shown by the red lines in the upper panels of Figure 4.9; the point-wise average of the data per region is also added (black line). In comparison with the results obtained from the standard model (in Figure 4.4), it is clear that the penalty approach gives a more

acceptable result. This is also confirmed by the fitted city effects in the lower panels of Figure 4.9. Indeed, on these lower panels, the green line is the horizontal line passing through zero and the red dashed one is the point-wise average of the fitted city effects; these lines show that the fitted city effects from the penalty approach are appropriately centred (as opposed to the skewed ones obtained from the standard model in the lower panels of Figure 4.4).

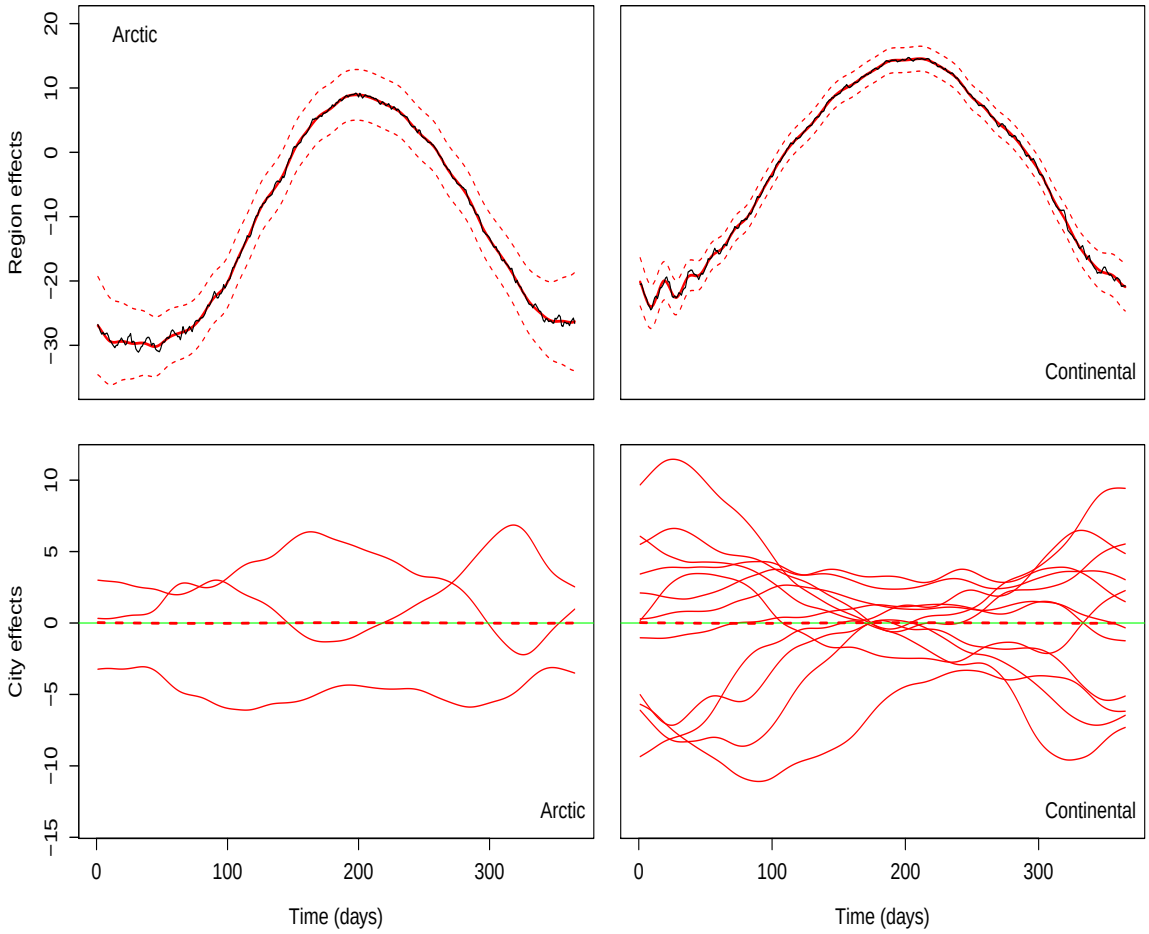

Figure 4.9: *Graphic related to* `CanadianWeather` *data. The red lines in the upper panels show two fitted region effects obtained from the penalty approach; the wiggly back lines represents the data averaged per region. The lower panels show the fitted city effects in these two regions; on these lower panels, the green line is the horizontal line passing through zero, and the red dashed one is the point-wise average of the fitted city effects.*

Second, the upper panels in Figure 4.10 show the data together with the global fit and the treatment effects from the penalty approach on `ChildHeight` data; the child effects are also shown in the lower panels. Although we cannot compute the observed mean per treatment in this case, it is clear that our fitted treatment effects look consistent with the data.

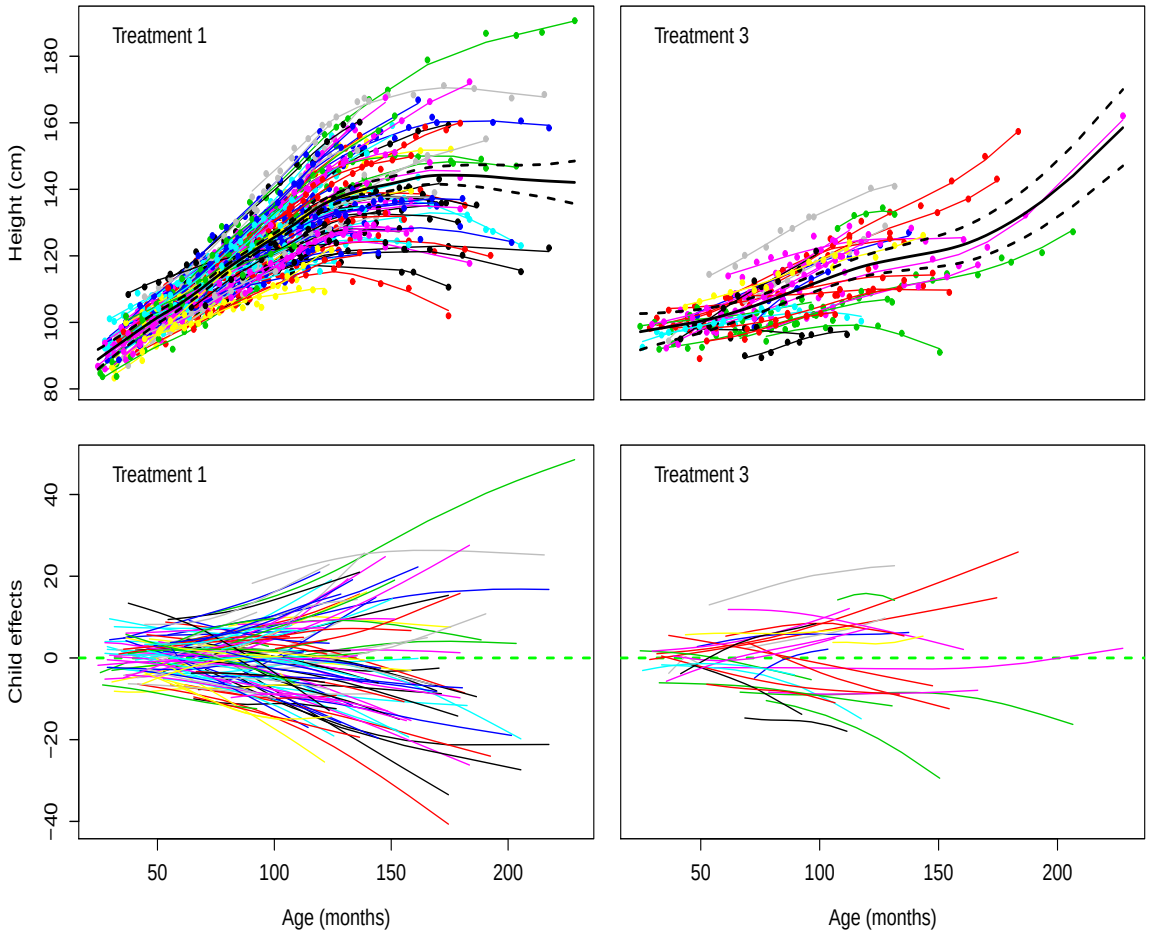

Figure 4.10: *Graphic related to* `ChildHeigth` *data. The upper panels show two fitted treatment effects obtained from the penalty approach; the data and global fit are also added. The lower panels show the fitted child effects.*

Finally, we turn to the simulated data described in Section 4.1.3. The black lines in Figure 4.8 illustrate the standard deviation $SD_k(\boldsymbol{x}_k)$ from the penalty approach,

compared to those obtained from the standard model (under forward and backward bases). Comparative graphics of the mean square error, $MSE^{(r)}$, and associated boxplots are also provided in Figures **??** and 4.7. All these graphics illustrate the good properties of the penalty approach (at least for the data sets and simulated data considered in this thesis).

The general conclusion drawn in the course of this simulation exercise is that the discrepancy of the fitted group/subject effects using the standard model with forward or backward bases increases as the subject effects become more flexible. In contrast, whatever this flexibility was, the results from the penalty approach were consistent. Additionally, it is worth pointing out that whenever the fitted effects from the standard approach under forward bases coincided with that under backward bases, then the fitted effects obtained from the penalty approach matched them (although some relatively small differences in the standard errors similar to those in the upper panels of Figure 4.8 remain).

4.6 Multivariate subject-specific curves

In Chapter 3 and so far in this Chapter, we have been looking at the response as a function of the time covariate in a hierarchical setting. More generally, the response data $\boldsymbol{y}_i$ on each subject i can vary non-parametrically as a function of several covariates $\boldsymbol{x}_i^{[1]} = (x_{i,1}^{[1]}, ..., x_{i,\check{n}_i}^{[1]})', ..., \boldsymbol{x}_i^{[G]} = (x_{i,1}^{[G]}, ..., x_{i,\check{n}_i}^{[G]})'$. In this case, we extend model (4.1) to the general structure

$$\boldsymbol{y}_i = \sum_{s=1}^{G} \mathcal{F}_i^{[s]}(\boldsymbol{x}_i^{[s]}) + \boldsymbol{\varepsilon}_i, \;\; i = 1, ..., n; \tag{4.27}$$

where the individual functions $\mathcal{F}_i^{[s]}(\cdot)$ are smooth functions acting on the s th covariate. If we assume that these n subjects are partitioned into m groups as before, then our interest may lie either in the group effects only, in the subject effects only, or in both effects. Without loss of generality, let us suppose that our concern is about the group effects alone for the first G_1 covariates, about the marginal subject effects alone for the next G_2 covariates, and about both the group and subject effects for the remaining $(G - G_1 - G_2)$ covariates. That is, by analogy to model (4.1), the individual functions

$\mathcal{F}_i^{[s]}$ are structured as

$$\mathcal{F}_i^{[s]}(\cdot) = \begin{cases} \mathcal{S}_{g(i)}^{[s]}(\cdot) & \text{for } s = 1, ..., G_1 \\ \breve{\mathcal{S}}_i^{[s]}(\cdot) & \text{for } s = G_1 + 1, ..., G_1 + G_2 \\ \mathcal{S}_{g(i)}^{[s]}(\cdot) + \breve{\mathcal{S}}_i^{[s]}(\cdot) & \text{for } s = G_1 + G_2 + 1, ..., G. \end{cases} \tag{4.28}$$

Hence, by extension of (4.5), we derive the following matrix representation

$$\mathcal{F}_i^{[s]}(\boldsymbol{x}_i^{[s]}) = \begin{cases} \boldsymbol{\Omega}_{g(i),i}^{[s]} \boldsymbol{\alpha}_{g(i)}^{[s]} & \text{for } s = 1, ..., G_1 \\ \breve{\boldsymbol{\Omega}}_i^{[s]} \breve{\boldsymbol{\alpha}}_i^{[s]} & \text{for } s = G_1 + 1, ..., G_1 + G_2 \\ \boldsymbol{\Omega}_{g(i),i}^{[s]} \boldsymbol{\alpha}_{g(i)}^{[s]} + \breve{\boldsymbol{\Omega}}_i^{[s]} \breve{\boldsymbol{\alpha}}_i^{[s]} & \text{for } s = G_1 + G_2 + 1, ..., G \end{cases} \tag{4.29}$$

where $\boldsymbol{\Omega}_{g(i),i}^{[s]}$ and $\breve{\boldsymbol{\Omega}}_i^{[s]}$ are spline regression matrices, with associated coefficients $\boldsymbol{\alpha}_{g(i)}^{[s]}$ and $\breve{\boldsymbol{\alpha}}_i^{[s]}$.

Under smoothness constraints for $\mathcal{F}_i^{[s]}(\cdot),\ s = 1, ..., G$, identifiability constraints on the components of $\mathcal{F}_i^{[s]}(\cdot),\ s = G_1 + G_2 + 1, ..., G$, and following the re-parametrization in (4.7), we can re-parametrize the components in (4.29) as

$$\mathcal{F}_i^{[s]}(\boldsymbol{x}_i^{[s]}) = \begin{cases} \boldsymbol{X}_{g(i),i}^{[s]} \boldsymbol{\beta}_{g(i)}^{[s]} + \boldsymbol{Z}_{g(i),i}^{[s]} \boldsymbol{b}_{g(i)}^{[s]} & \text{for } s = 1, ..., G_1 \\ \breve{\boldsymbol{X}}_i^{[s]} \breve{\boldsymbol{\beta}}_i^{[s]} + \breve{\boldsymbol{Z}}_i^{[s]} \breve{\boldsymbol{b}}_i^{[s]} & \text{for } s = G_1 + 1, ..., G_1 + G_2 \\ \boldsymbol{X}_{g(i),i}^{[s]} \boldsymbol{\beta}_{g(i)}^{[s]} + \boldsymbol{Z}_{g(i),i}^{[s]} \boldsymbol{b}_{g(i)}^{[s]} + \breve{\boldsymbol{\Omega}}_i^{[s]} \breve{\boldsymbol{\alpha}}_i^{[s]} & \text{for } s = G_1 + G_2 + 1, ..., G. \end{cases} \tag{4.30}$$

Note that, in this representation, we have partitioned the subject effects in the second line into penalized and unpenalized components because we do not need identifiability constraints for these effects; in other words these effects are subject only to the smoothness penalty matrices which are singular and therefore, such a partition is important for the mixed model representation. In contrast, on the third line, we need both smoothness and identifiability constraints, leading to full rank penalty matrices.

With the representation (4.30), we show that model (4.27)-(4.29) falls in the range of the mixed model (4.9), with appropriate components $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\boldsymbol{P}$, $\boldsymbol{X}$, and $\boldsymbol{Z}$. Conse-

quently, we can essentially adopt the computational scheme of Section 4.4 to fit the extended model (4.27).

4.7 Conclusion

In this Chapter, we have generalized Chapter 3 to grouped and unbalanced data. We started with some further illustrations of the discrepancy arising from fitting the standard model (4.2)-(4.4). Next, we extended the penalty approach and described its implementation on the mixed model platform. In summary, our approach consists of designing the group and subject components, and then addressing the smoothing and/or identifiability problems via appropriate penalization. Although this approach produced interesting results and outperformed the standard method, some points still need to be addressed. For example, in Section 4.6, we outlined the extension of these nested curves to the multivariate setting, but we gave no illustration. Additionally we have concentrated our investigation on normal responses; although this has been sufficient to demonstrate, first the problem arising form the standard model, and second, the attractive impact of the penalization, a more general class of distribution needs to be investigated. We shall return to these points in the final Chapter.

Chapter 5

Smoothing dispersed counts with applications to mortality data

Modelling and forecasting mortality is a problem of fundamental importance to actuaries, demographers and governments. Data for this purpose come largely from two sources: (a) population mortality data available from either the Human Mortality Database (2009) or government offices of statistics, and in this case, data are available in terms of death counts and exposure-to-risk of deaths; (b) insurance data collected and collated by some central agency; the Continuous Mortality Investigation (CMI) fulfils this role in the UK, in which case, data are available in terms of claims on policies and number of policy-years live. We will often use the terminology deaths and exposures to refer either to real deaths and exposure-to-risk of deaths in the case of population data, or to claims on policies and number of policy-years live in the insurance context.

In both cases, these data are generally available at the aggregate level, ie, deaths and exposures are often arranged in two-way tables of deaths and exposures, classified by age at death and year of death. We shall denote by $\boldsymbol{x} = (x_1, ..., x_{n_x})'$ and $\boldsymbol{t} = (t_1, ..., t_{n_t})'$ the vectors of age and year indices. Also, $\boldsymbol{D}$ will represent the $n_x \times n_t$ table of deaths and D_{ij} the entry of $\boldsymbol{D}$ corresponding to age x_i in calendar year t_j. Similarly, $\boldsymbol{E}$ and E_{ij} will represent corresponding quantities in the exposure data; ie, $\boldsymbol{E}$ will be the $n_x \times n_t$ matrix of exposure-to-risk of death and E_{ij} its entry corresponding to age x_i and calendar year t_j.

Figure 5.1 shows the observed mortality rates (ie, the number of deaths D_{ij} divided

by the corresponding exposure E_{ij}) on the log scale, for male assured life data from age 25 to 95 and calendar years from 1947 to 2006. These rates show some random variability across ages and years. As a result it is not appropriate to use them directly for planning purposes. A reasonable and simple view is to suppose that there is an underlying true surface driving these rates, such that these observed values represent a distorted image of this surface; the observed rates can then be seen as a twinning of that underlying surface and some noise.

Standard approaches for estimating this surface assume that the deaths D_{ij} are Poisson distributed. This provides satisfactory results for some data sets, but in general, mortality data show more variability than expected under the Poisson assumption. The main purpose of this chapter is first to show how ignoring such variability in the modelling process can have a dramatic impact on the fitted surface, and second to propose a solution to this problem.

This Chapter is structured as follows. Section 5.1 extends Section 2.3.2 and outlines the formulation of PB-splines in two dimensions, with reference to the modelling and forecasting of mortality data. Section 5.2 describes and illustrates the impact of high variability and heterogeneity on the fitted surface. Section 5.3 develops a solution to the problem of heterogeneity and over-dispersion via quasi-likelihoods. Section 5.4 concentrates on applications. Section 5.5 discusses the use of the negative binomial distribution with regard to over-dispersion, and we close with a brief discussion in Section 5.6.

5.1 Modelling and forecasting mortality data

The Poisson distribution is a popular choice when dealing with count data. Hence, many models for mortality data (Brouhns et al., 2002; Currie et al., 2004; Kirkby, 2009; Cairns et al., 2009) assume that the number of deaths are Poisson distributed as

$$D_{ij} \sim \mathcal{P}oiss(\tau_{ij} \times E_{ij}), \tag{5.1}$$

where τ_{ij} is the force of mortality at age x_i in year t_j.

Our broad interest lies in the estimation and projection of the force of mortality, τ_{ij}, in a smooth fashion. Using the Poisson canonical link, this estimation corresponds to

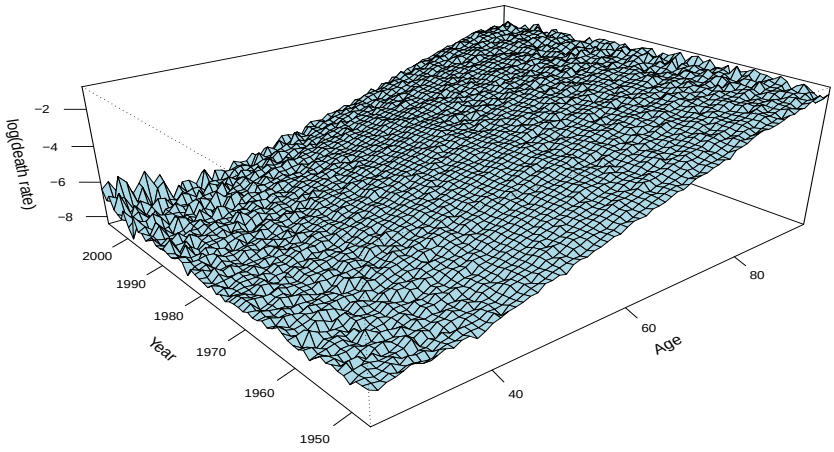

Figure 5.1: *Male assured lives data; age: 25 - 95, year: 1947 - 2006.*

the fitting of the model defined through

$$\mathbb{E}[D_{ij}] = E_{ij}\tau_{ij}, \quad \log(\tau_{ij}) = \mathcal{S}(x_i, t_j), \tag{5.2}$$

in which we want to estimate and extrapolate the two-dimensional function $\mathcal{S}(\cdot)$. We will ensure smoothness of $\mathcal{S}(\cdot)$ by penalization as in Section 2.6 and so model (5.2) will become a PGLM. This Section outlines the extension of PB-splines to the smoothing and forecasting of two-dimensional data in the grid format of mortality data, and sketches the array method that will be used in this and subsequent Chapters to speed up the computations.

5.1.1 Two-dimensional smoothing for grid data

In general, let us suppose that we have grid data (y_{ij}, x_i, t_j) where the y_{ij} are realizations of some random variables at the covariate points (x_i, t_j), with $\mathbb{E}[y_{ij}] = \mu_{ij}$. We assume that these random variables are independent and distributed according to a particular distribution from the exponential family, and we consider the regression problem with predictor defined by

$$\mathbb{E}[y_{ij}] = \mu_{ij}, \quad \text{with } \; g(\mu_{ij}) = \mathcal{S}(x_i, t_j) \tag{5.3}$$

for some smooth function $\mathcal{S}(\cdot)$ which we wish to estimate; here, $\mathcal{S}(\cdot)$ is a bivariate smooth function.

As in the one-dimensional case, there are several approaches in the literature for estimating $\mathcal{S}(\cdot)$, one of the well-known being the extension of PB-splines as proposed by Currie et al. (2004). This extension has two advantages: first, it benefits from the features of B-spline bases and difference penalties, as discussed in Section 2.3.3, and second, it allows forecasting to take place in a straightforward manner.

Let $\boldsymbol{B}_x$ and $\boldsymbol{B}_t$ represent our B-spline bases of sizes $n_x \times c_x$ and $n_t \times c_t$ respectively in age and time; we will often refer to them as *marginal bases*. In the two-dimensional PB-splines approach, the regression matrix is constructed as the Kronecker product of such marginal B-spline bases. Precisely, the two dimensional B-spline basis, $\boldsymbol{B}$, corresponding to (5.3) is obtained as

$$\boldsymbol{B} = \boldsymbol{B}_t \otimes \boldsymbol{B}_x. \tag{5.4}$$

To emphasise that this regression matrix is derived from B-splines, we write it as $\boldsymbol{B}$ rather that $\boldsymbol{\Omega}$, as in Chapter 2. Hence, by definition of the Kronecker product, $\boldsymbol{B}$ is a $n_x n_t \times c_x c_t$ regression matrix associated with a $c_x c_t$-length vector $\boldsymbol{\theta}$ of regression coefficients (to be estimated), and the entries of $\boldsymbol{B}$ are obtained by multiplying the entries of $\boldsymbol{B}_t$ with those of $\boldsymbol{B}_x$. In other words, we can think of $\boldsymbol{B}$ as a multiplicative table with $\boldsymbol{B}_x$ and $\boldsymbol{B}_t$ sited in the two directions, yielding the two-dimensional basis, $\boldsymbol{B}$, from which a subset is illustrated in Figure 5.2. With this representation, each entry of the regression vector $\boldsymbol{\theta}$ is associated with a summit of this basis, and so it becomes intuitively appealing to think of these coefficients as a $c_x \times c_t$ matrix $\boldsymbol{\Theta}$ such that $vec(\boldsymbol{\Theta}) = \boldsymbol{\theta}$.

Turning to the estimation of our two-dimensional smooth surface $\mathcal{S}(\cdot)$, one can now proceed by analogy with the one-dimensional case. That is, rich marginal bases are taken in age and time, a two dimensional rich basis is formed using (5.4), and the roughness is penalized in each row and each column of the coefficient matrix $\boldsymbol{\Theta}$. In order to remain consistent with the fact that the regression matrix $\boldsymbol{B}$ as defined in (5.4) is a $n_x n_t \times c_x c_t$ matrix, Currie et al. (2004) showed that this penalization of

rows and columns of $\boldsymbol{\Theta}$ is identical to the action of the penalty matrix

$$\boldsymbol{P} = \lambda_x(\boldsymbol{I}_{c_t} \otimes \boldsymbol{\Delta}_x'\boldsymbol{\Delta}_x) \; + \; \lambda_t(\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t \otimes \boldsymbol{I}_{c_x}) \tag{5.5}$$

on the coefficient vector $\boldsymbol{\theta}$. In this expression, $\boldsymbol{\Delta}_x$ and $\boldsymbol{\Delta}_t$ are the $(c_x - d) \times c_x$ and $(c_t - d) \times c_t$ difference matrices of order d respectively for age and time, and λ_x and λ_t are smoothing parameters in the age and time direction respectively.

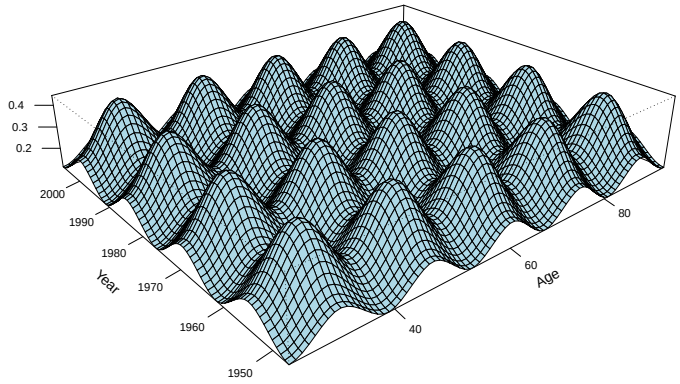

Figure 5.2: *A subset of a two-dimensional basis of B-splines,* $\boldsymbol{B} = \boldsymbol{B}_t \otimes \boldsymbol{B}_x$.

With these details, fitting model (5.3) reduces to the estimation of a one dimensional PGLM with response data $vec(\boldsymbol{Y})$, link function $g(\cdot)$, regression matrix $\boldsymbol{B}$ defined in (5.4), and penalty matrix $\boldsymbol{P}$ given by (5.5). Applying this procedure to the Poisson mortality model (5.1)-(5.2) via Section 2.6, we find a succinct summary of the estimation process as follows:

- Data:

$$vec(\boldsymbol{D}), \quad vec(\boldsymbol{E}), \tag{5.6}$$

- Model predictor:

$$\log[vec(\boldsymbol{\mu})] = \log[vec(\boldsymbol{E})] + \log[vec(\boldsymbol{\tau})] = \log[vec(\boldsymbol{E})] + \boldsymbol{B}\boldsymbol{\theta}, \tag{5.7}$$

where $\boldsymbol{\tau}$ and $\boldsymbol{\mu}$ are $n_x \times n_t$ matrices of τ_{ij} and μ_{ij}, $1 \leq i \leq n_x$, $1 \leq j \leq n_t$.

- Iterative equations:

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{\theta}} \;\approx\; \boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{5.8}$$

  with $\tilde{\boldsymbol{W}} = diag[vec(\tilde{\boldsymbol{\mu}})]$, and $\tilde{\boldsymbol{z}} = \boldsymbol{B}\tilde{\boldsymbol{\theta}} + vec\,[(\boldsymbol{D} - \tilde{\boldsymbol{\mu}})/\tilde{\boldsymbol{\mu}}]$.

- Variance of the fitted values at convergence:

$$var(\boldsymbol{B}\hat{\boldsymbol{\theta}}) = diag[\boldsymbol{B}\, cov(\hat{\boldsymbol{\theta}})\, \boldsymbol{B}'], \quad \text{with} \quad cov(\hat{\boldsymbol{\theta}}) \approx (\boldsymbol{B}'\hat{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})^{-1}. \tag{5.9}$$

An illustration of the fitted mortality surface is shown by the light-blue colours in Figure 5.3; some profile views from this surface are also depicted by the black lines in Figure 5.4.

Though the above procedure produces satisfactory results, we should bear in mind that the Kronecker product underpinning the B-spline regression matrix $\boldsymbol{B}$ is very memory hungry; ie, the size of $\boldsymbol{B} = \boldsymbol{B}_t \otimes \boldsymbol{B}_x$ increases rapidly with the size of its marginals. This can yield serious difficulties both in computational time and even in storage, since two smoothing parameters must be selected by solving a large system of iterative equations. This awkwardness becomes more worrying when we deal with the joint modelling of several mortality tables in the next Chapter.

Clearly, our data $\boldsymbol{D}$ and $\boldsymbol{E}$, as well as the regression coefficients $\boldsymbol{\Theta}$ are in the form of matrices, yet the above formulation has almost discarded the benefits of this matrix structure. It was Eilers et al. (2006) and Currie et al. (2006) who first showed how such matrix structures, together the Kronecker product in the regression matrix can be exploited efficiently in the smoothing context. More generally, these authors developed an array method that leads to low storage and high speed computations for multi-dimensional PGLM where data have an array structure and the model matrix can be written using the Kronecker product. They referred to this approach as Generalized Linear Array Models (GLAM).

5.1.2 Generalized Linear Array Models

The main motivation of GLAM stems from observing that the left hand side of the well known formula

$$vec(\boldsymbol{A}_2\boldsymbol{A}\boldsymbol{A}_1') = (\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)vec(\boldsymbol{A}) \tag{5.10}$$

has fewer multiplications and lower storage requirements that its right hand side, for any conformable matrices $\boldsymbol{A}$, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. With this formula, it is obvious that the $n_x n_t \times 1$ predictor vector $\boldsymbol{B\theta}$ contains the same elements as the $n_x \times n_t$ matrix $\boldsymbol{B}_x\boldsymbol{\Theta}\boldsymbol{B}_t'$; we write

$$(\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta},\ n_x n_t \times 1 \ \equiv\ \boldsymbol{B}_x\boldsymbol{\Theta}\boldsymbol{B}_t',\ n_x \times n_t, \tag{5.11}$$

where the symbol "$\equiv$" means that although the matrices on the left and right have different dimensions, they contain the same elements. Besides being more efficient than its left hand, the right hand in (5.11) is conceptually interesting since it returns the fitted values in the matrix structure of the original data. A similar representation can be used to compute the right hand side of the iterative equation (5.8).

The next component that needs computational care is the weighted inner product $\boldsymbol{B'WB}$. Prior to giving its GLAM representation, let us recall that if $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are $n \times p$ and $n \times q$ matrices, then the *row tensor product* (Eilers et al., 2006) of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, which we denote by $\boldsymbol{A}_1 \square \boldsymbol{A}_2$, is the matrix defined as

$$\boldsymbol{A}_1 \square \boldsymbol{A}_2 = (\boldsymbol{A}_1 \otimes \boldsymbol{1}_q') * (\boldsymbol{1}_p' \otimes \boldsymbol{A}_2); \tag{5.12}$$

ie, $\boldsymbol{A}_1 \square \boldsymbol{A}_2$ is the $n \times pq$ matrix containing the element-by-element multiplication of each column of $\boldsymbol{A}_1$ by each column of $\boldsymbol{A}_2$.

With this definition, Currie et al. (2006) showed that

$$(\boldsymbol{B}_t \otimes \boldsymbol{B}_x)'\boldsymbol{W}(\boldsymbol{B}_t \otimes \boldsymbol{B}_x),\ c_x c_t \times c_x c_t \ \equiv\ (\boldsymbol{B}_x \square \boldsymbol{B}_x)'\boldsymbol{\mathcal{W}}(\boldsymbol{B}_t \square \boldsymbol{B}_t),\ c_x^2 \times c_t^2, \tag{5.13}$$

where $\boldsymbol{\mathcal{W}}$ represents the $n_x \times n_t$ matrix containing the diagonal elements of the $n_x n_t \times n_x n_t$ diagonal weight matrix $\boldsymbol{W}$ such that $vec(\boldsymbol{\mathcal{W}}) = diag(\boldsymbol{W})$.

Finally, at convergence, the variance of the fitted values can be obtained efficiently

and in the matrix structure of the data as

$$diag[cov(\boldsymbol{B}\hat{\boldsymbol{\theta}})],\ n_x n_t \times 1 \ \equiv\ (\boldsymbol{B}_x \square \boldsymbol{B}_x)\hat{\boldsymbol{S}}_{\boldsymbol{\theta}}(\boldsymbol{B}_t \square \boldsymbol{B}_t)',\ n_x \times n_t, \tag{5.14}$$

where $\hat{\boldsymbol{S}}_{\boldsymbol{\theta}}$ is a $c_x^2 \times c_t^2$ matrix obtained by re-arranging the elements of the $c_x c_t \times c_x c_t$ covariance matrix $cov(\hat{\boldsymbol{\theta}})$.

A full description on how the rearrangements related to $\hat{\boldsymbol{S}}_{\boldsymbol{\theta}}$ and (5.13) are achieved, as well as proofs of these GLAM formulae, are given in Currie et al. (2006). An important point is that these rearrangements are very efficient. For detailed illustrations of the computational and storage benefits, as well as extensions of these formulae to other matrix and array structures (specially for non-diagonal weight matrices), we refer the reader to Kirkby (2009).

5.1.3 Forecasting

Actuaries need to project mortality rates into the far future for calculating present values of pension and annuity liabilities. With PB-splines, projection as introduced in Currie et al. (2004) is treated as a missing data problem and the penalty is used to fill in the missing data. This approach is best explained through an example. Let us assume that we have mortality data in the age range $\boldsymbol{x} = (x_1, ..., x_{n_x})'$ and years $\boldsymbol{t} = (t_1, ..., t_{n_t})'$. To forecast for n_t^+ years into the future, these n_t^+ future years are first appended to $\boldsymbol{t}$ and then the marginal B-spline basis in time is computed along the augmented time index $(t_1, ..., t_{n_t}, t_{n_t+1}, ..., t_{n_t+n_t^+})'$. That is, the initial $n_t \times c_t$ marginal B-spline matrix $\boldsymbol{B}_t$ becomes a $(n_t + n_t^+) \times (c_t + c_t^+)$ matrix of B-splines in time, which we denote by $\boldsymbol{B}_t^\star$. Consequently, the two-dimensional B-spline basis $\boldsymbol{B}$ is replaced by

$$\boldsymbol{B}^\star = \boldsymbol{B}_t^\star \otimes \boldsymbol{B}_x. \tag{5.15}$$

Second, the response is correspondingly augmented with dummy data, leading to the adjusted penalized scoring algorithm

$$((\boldsymbol{B}^\star)'\boldsymbol{V}\tilde{\boldsymbol{W}}\boldsymbol{B}^\star + \boldsymbol{P})\hat{\boldsymbol{\theta}} \ \approx\ (\boldsymbol{B}^\star)'\boldsymbol{V}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{5.16}$$

where $\boldsymbol{V} = diag[vec(\mathbf{1}_{n_x n_t}, \mathbf{0}_{n_x \times n_t^+})]$, and $\hat{\boldsymbol{\theta}}$ contains both the fitted and forecast coefficients. Thus, these iterative equations are used to fit and forecast simultaneously. A similar procedure can be used for backward projections as well as extrapolation in the age direction. Note that if we replace the last n_t^+ row of $\boldsymbol{B}_t^\star$ in (5.15) by zeros, then the adjusted penalized scoring algorithm (5.16) becomes identical to the standard penalized scoring algorithm (5.8), but with $\boldsymbol{B}$ replaced by $\boldsymbol{B}^\star$.

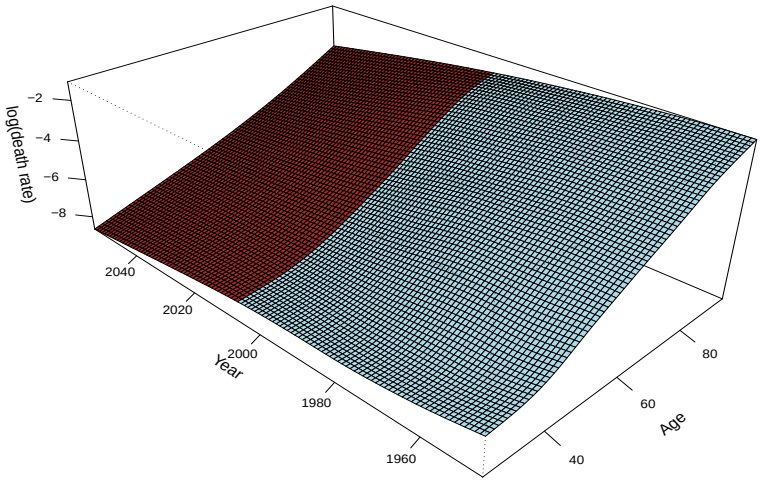

Figure 5.3: *Fitted (light-blue) and forecast (brown) mortality surface using cubic B-splines with second order difference penalty. Male assured lives data; age: 25 - 95, year: 1947 - 2006.*

Although the order of the penalty has no discernible effect on the smoothing of the observed data, it has a consequential impact on the extrapolated trends/surface. In the one dimensional case specifically, using a d order penalty would result in a $(d-1)$ degree polynomial forecast defined exactly by the last d coefficients in the fitting region. This exact property does not hold in two dimensions since the penalty function in this case tends to maintain the structure of the whole surface in the age and year directions. In practice the second order penalty is often preferred. There is no mathematical reason for this preference; to our knowledge, there is currently no mathematical method nor formal criterion for choosing the order of differencing. However the second order difference penalty is straightforward to manage, it produces sufficient

flexible models over the fitting range, and when using it for forecasting, the shape of the extrapolated coefficients ties reasonably well with the shape of the fitted coefficients, provided there is a sufficiently strong signal in the forecasting direction. An illustration of the extrapolated surface 50 years into the future is shown by the brown colour in Figure 5.3. Some profile views from this surface with associated confidence bands are also shown in Figure 5.4. These confidence intervals have been computed using (5.14) and based on the asymptotic normal distribution of the predictor $\boldsymbol{B}\hat{\boldsymbol{\theta}}$.

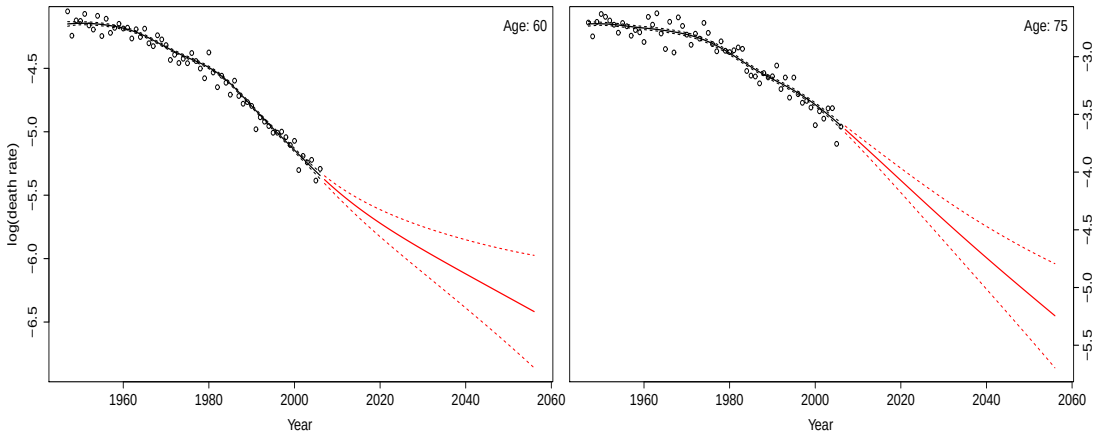

Figure 5.4: *Profile views from the fitted (black lines) and forecast (red) force of mortality using two-dimensional PB-spline model with Poisson errors. Male assured lives data; age: 25 - 95, year: 1947 - 2006.*

5.2 The impact of heterogeneity and over-dispersion

The fact that mortality data are generally available at the aggregate level (ie, deaths and exposures are classified by age at death and year of death) gives rise to two problems for model building. For population data, the risk set for each age and year of death is heterogeneous with respect to mortality since it contains smokers and non-smokers, different social classes, etc. For insurance data, the risk set is subject to an additional source of heterogeneity: 'deaths' are claims on policies, and 'exposure-to-risk' is the number of policy-years lived. Very often, some policyholders have more than one policy and so, for these policyholders, a single death gives rise to multiple claims; this is known as the *problem of duplicates*. Ideally the data would be de-

duplicated, ie, policies held by a single life would be consolidated into a single policy; Richards (2008) describes such a process. Unfortunately, such consolidation is not available for historical data such as collected by the CMI. Further, de-duplication does not address the problem of the heterogeneity of mortality across the risk set.

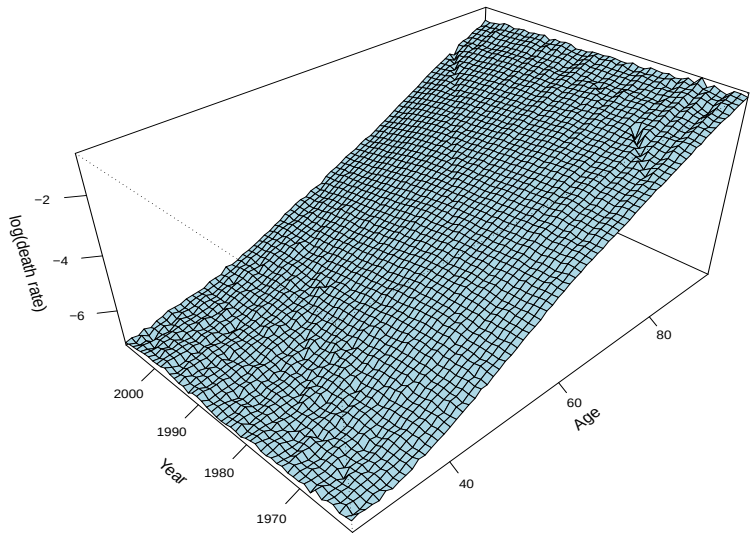

Figure 5.5: *ONS mortality data for males in England & Wales; age: 25 - 95, year: 1961 - 2007.*

Let us consider the ONS (Office for National Statistics in the UK) mortality data for males in England and Wales, from age 25 to 95 and year 1961 to 2007, depicted in Figure 5.5. The red lines in Figure 5.6 show the profile views of the fitted and forecast force of mortality for these data using the two-dimensional PB-splines model under the Poisson assumption (5.1)-(5.2). In the data region (ie, from year 1961 to 2007), the fit seems perfectly reasonable and we could hope to forecast the fitted trends successfully as we did with the assured lives data in Figure 5.4. But, there are two worrying features.

- First, the smoothing parameters (selected by minimizing the BIC) are 394 for age and 0.01 for year. Such a low value indicates a weak signal in the year direction; volatile and unreliable forecasts are a likely consequence; we return to this issue in section 5.4.2.2.

- Second, how plausible is our Poisson assumption? Does the fitted curve provide a satisfactory smoothing of the observed rates? The Poisson distribution has equal mean and variance, and if this property holds, the standardized residuals $r_{ij} = (D_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}}$ should have an approximate $\mathcal{N}(0,1)$ distribution. But for the ONS data set, many of the r_{ij}^2 are above 15, with an average value of 3.20 (in comparison to 1.68 for the assured lives data). In other words, the observed variation from the fitted surface seems to be way in excess of what is reasonable under the Poisson assumption; this is the phenomenon of over-dispersion.

These two features have a catastrophic impact on the projections, as shown by the red lines in Figure 5.6.

To cope with the problem of heterogeneity within the Lee-Carter family, Li et al. (2009) incorporated gamma distributed random variables within each cell from which it follows that the number of deaths/claims has a negative binomial distribution with variance larger than the mean, as required. Other approaches to the problem of over-dispersion can be found in Williams (1982), Breslow and Clayton (1993), Hinde and Demetrio (1998), for example. In the next section, we tackle this problem with a two-stage joint-modelling of mean and dispersion through the extended quasi-likelihood, similar to that described in the parametric setting by McCullagh and Nelder (1989, chap 2). Our reason for adopting the quasi-likelihood approach is that it allows us to remain within the exponential family, and, as a result, we can essentially adopt a generalized linear model approach to model building. Renshaw (1992) also used a similar approach (in the parametric setting) in his paper on the graduation of mortality data in the presence of duplicates. Our contribution is to extend this work to the smoothing of count data, especially the smoothing and forecasting of two-dimensional mortality tables.

5.3 Dispersed counts and quasi-likelihoods

The work in this Section is based on Djeundje and Currie (2010b).

We consider independent observations $\boldsymbol{y} = (y_1, ..., y_n)'$. In general the y_k can be counts but we will concentrate our description to the case when they are counts of deaths or claims. We supposed that these data can be partitioned into $\mathcal{K}$ classes

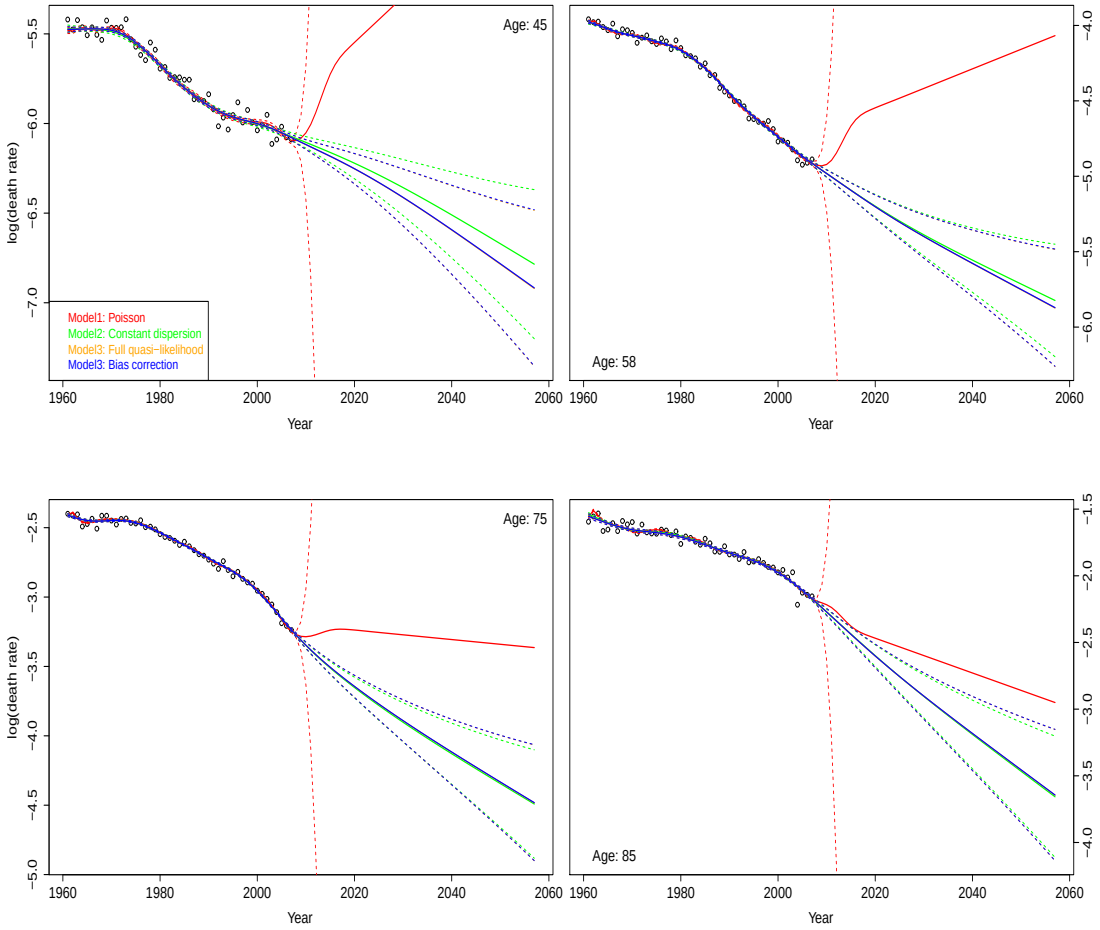

Figure 5.6: *Profile views from the fitted and forecast force of mortality using two-dimensional PB-spline model ignoring over-dispersion (red) and incorporating over-dispersion (green, blue, orange; the orange lines are hidden by the blue ones). ONS data for males in England & Wales, age: 25 - 95, year: 1961 - 2007.*

$\mathcal{C}_1, ..., \mathcal{C}_\mathcal{K}$, $\mathcal{K} \leq n$, in such a way that each class is homogeneous (by homogeneous we mean that the level of dispersion within each class can be assumed constant). For example, at one extreme if $\mathcal{K} = n$ then each count has its own dispersion parameter while at the other extreme if $\mathcal{K} = 1$ then a single dispersion parameter applies to all counts. We will be particularly interested in the intermediate case where the dispersion parameter is age-dependent in a mortality table; this is consistent with the approach of Forfar et al. (1988), Renshaw (1992) and Li et al. (2009) who use age-dependent dispersion parameters. We will denote by $\phi_{\mathcal{C}_k}$ the over-dispersion parameter in the class $\mathcal{C}_k$; we note in passing that $\phi_{\mathcal{C}_k}$ could in theory be less than one, so this approach can also deal with under-dispersion. Let $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_\mathcal{K}\}$ denote the set of classes and let $\varphi : \{1, ..., n\} \longrightarrow \mathcal{C}$ assign observations to classes.

We include dispersion in the model through the first and second moment assumptions

$$\mathbb{E}\left[y_k\right] = \mu_k, \quad var(y_k) = \phi_{\varphi(k)} \times v(\mu_k), \quad k = 1, ..., n, \quad g\left(\boldsymbol{\mu}\right) = \boldsymbol{B\theta}, \tag{5.17}$$

where $\boldsymbol{\mu} = vec(\mu_1, ..., \mu_n)$, $v(\cdot)$ is the variance function, $g(\cdot)$ is the link function, $\boldsymbol{\theta}$ is the unknown vector of coefficients, and $\boldsymbol{B}$ is the regression matrix. In the case of a mortality table for example, $\boldsymbol{B}$ becomes the Kronecker product of the marginal bases as defined in (5.4).

The Poisson assumption is given by $\phi_{\varphi(k)} = 1, \forall k$, and $v(\cdot)$ equal to the identity function, in which case the model can be estimated via maximum (penalized) likelihood; this requires the distribution that has generated the data. Unfortunately, such a distribution is not available for (5.17). An alternative is the quasi-likelihood framework of Wedderburn (1974), an extension of the familiar likelihood function that allows estimation to take place in more general settings such as (5.17). The quasi-likelihood framework shares several interesting properties of the ordinary likelihood; see McCullagh and Nelder (1989, pg 325). Under model (5.17), the *quasi-likelihood* (or more correctly the *log quasi-likelihood*) of a single observation y_k is defined as

$$\mathcal{Q}(\mu_k \,|\, y_k) \;=\; \frac{1}{\phi_{\varphi(k)}} \int_{y_k}^{\mu_k} \frac{y_k - t}{v(t)}\, dt \;=\; -\frac{1}{2\phi_{\varphi(k)}} d_k, \tag{5.18}$$

where

$$d_k = -2 \int_{y_k}^{\mu_k} \frac{y_k - t}{v(t)}\, dt \tag{5.19}$$

is the *deviance component*. The estimates of the dispersion parameters $\phi_{\varphi(k)}$ are based on the d_k. (The normal distribution is a well-known example here since when $v(t) = 1$ we have $d_k = (y_k - \mu_k)^2$, the k th component of the residual sum of squares.) The quasi-likelihood of the sample $\boldsymbol{y}$ is

$$\mathcal{Q}(\boldsymbol{\mu} \,|\, \boldsymbol{y}) \;=\; \sum_{k=1}^{n} \mathcal{Q}(\mu_k \,|\, y_k). \tag{5.20}$$

If the dispersion parameters are known, the fitting of model (5.17) is reduced to the optimization of the quasi-likelihood (5.20). However, since these parameters are generally unknown they also need to be estimated. Thus we also need the derivative

of $\mathcal{Q}(\cdot)$ to behave like a log-likelihood with respect to the ϕ_u derivatives; ie, if T_u represents our estimator of ϕ_u, we want

$$\mathbb{E}\left[\frac{\partial\,\mathcal{Q}}{\partial\phi_u}\right]_{\phi_u=T_u} = 0,\ \forall u \in \mathcal{C}.$$

For this to be achieved, the quasi-likelihood is usually adjusted (see Nelder and Pregibon, 1987) to the so-called *extended quasi-likelihood*, $\mathcal{Q}^+(\cdot)$, as follows:

$$\mathcal{Q}^+(\boldsymbol{\mu}, \boldsymbol{\phi}\,|\,\boldsymbol{y}) \;=\; \mathcal{Q}(\boldsymbol{\mu}\,|\,\boldsymbol{y}) \;+\; f\,(\boldsymbol{\phi})\,,\ \text{with}\ \boldsymbol{\phi} = (\phi_{\mathcal{C}_1}, \ldots, \phi_{\mathcal{C}_\mathcal{K}})', \tag{5.21}$$

for some well chosen $\mathcal{K}$-variate function $f(\cdot)$. A standard candidate is

$$f(\boldsymbol{\phi}) \;=\; -\frac{1}{2}\sum_{k=1}^{n} \log\left(2\pi\phi_{\varphi(k)}\,\psi(y_k)\right),$$

where $\psi(\cdot)$ is a positive function.

If we set

$$d_u = \frac{\sum\limits_{k\in\varphi^{-1}(u)} d_k}{n_u},\ \text{ where }\ n_u = |\varphi^{-1}(u)|,\ \forall u \in \mathcal{C}, \tag{5.22}$$

then,

$$\frac{\partial\,\mathcal{Q}}{\partial\phi_u}|_{\phi_u=d_u} = 0$$

and, at the true value of $\boldsymbol{\mu}$, (McCullagh and Nelder, 1989, chap 10),

$$\mathbb{E}\,[d_u] \simeq \phi_u,\ \forall u \in \mathcal{C}. \tag{5.23}$$

We make two comments on (5.22). First, there is a possible confusion of notation; we have adopted the convention that the suffix k, $k = 1, \ldots, n$, refers to observations, while the suffix u, $u \in \mathcal{C}$, refers to classes. Second, with the normal distribution, d_u reduces to the familiar maximum likelihood estimate of σ^2 in the class u.

Now, corresponding to the model (5.17) for the mean $\boldsymbol{\mu}$, we model the dispersion parameters (under the assumption that the structure of the classes allows us to do so) with

$$h(\boldsymbol{\phi}) = \breve{\boldsymbol{B}}\boldsymbol{\beta} \tag{5.24}$$

for some suitable link function $h(\cdot)$, and regression matrix $\breve{\boldsymbol{B}}$. (In the intermediate

case of age-dependant dispersions in a mortality table, we shall set $\breve{\boldsymbol{B}} = \boldsymbol{B}_x$.) Within this setting, fitting model (5.17) is reduced to the optimization of the extended quasi-likelihood (5.21) with respect to the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. This optimization yields the inter-dependent equations

$$\sum_{k=1}^{n} \frac{y_k - \mu_k}{\phi_{\varphi(k)}\, v(\mu_k)} \frac{\partial \mu_k}{\partial \theta_l} = 0, \quad l = 1, ..., c, \tag{5.25}$$

$$\sum_{u \in \mathcal{C}} \frac{n_u (d_u - \phi_u)}{\phi_u^2} \frac{\partial \phi_u}{\partial \beta_l} = 0, \quad l = 1, ..., \breve{c}, \tag{5.26}$$

where c and $\breve{c}$ are the lengths of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. Equations (5.26) correspond to the quasi-likelihood estimating equations based on independent responses d_u with $\mathbb{E}[d_u] = \phi_u$ and $var(d_u) = \phi_u^2/n_u$. In the GLM setting, equations (5.26) are identical to the estimating equations based on gamma responses d_u, with shape parameter n_u and scale parameter ϕ_u/n_u. The canonical link for the gamma distribution is the negative inverse function (see McCullagh and Nelder, 1989, chap 2), so we refine (5.24) as

$$d_u \sim Gamma\left(n_u, \frac{\phi_u}{n_u}\right), \quad u \in \mathcal{C}, \quad h(\boldsymbol{\phi}) = -1/\boldsymbol{\phi} = \breve{\boldsymbol{B}}\boldsymbol{\beta}, \tag{5.27}$$

although, for computational reasons, one might consider using other link functions such as log, instead of the canonical one.

In the parametric setting, we have generalized the two-stage joint modelling of mean and dispersion described in McCullagh and Nelder (1989, chap 10), and used by Renshaw (1992) for graduation in life insurance. However, mortality data often reveal complex patterns which suggest that a smoothing rather than a parametric approach is more appropriate. In the next section, we extend the above results to PB-splines.

5.3.1 Extended quasi-likelihood and PB-splines

Taking rich B-spline bases $\boldsymbol{B}$ and $\breve{\boldsymbol{B}}$, we apply roughness penalties to their coefficients $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ to achieve smoothness. Combining this penalization with the extended quasi-likelihood (5.21), we derive an optimization criterion, the *penalized extended*

quasi-likelihood,

$$\mathcal{Q}_P^+(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\phi}(\boldsymbol{\beta}) | \, \boldsymbol{y}) = \mathcal{Q}^+(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\phi}(\boldsymbol{\beta}) \, | \, \boldsymbol{y}) \; - \; \frac{1}{2} \left(\boldsymbol{\theta}' \boldsymbol{P}_{\boldsymbol{\lambda}_\theta} \boldsymbol{\theta} + \boldsymbol{\beta}' \boldsymbol{P}_{\boldsymbol{\lambda}_\beta} \boldsymbol{\beta} \right), \tag{5.28}$$

where $\boldsymbol{P}_{\boldsymbol{\lambda}_\theta}$ and $\boldsymbol{P}_{\boldsymbol{\lambda}_\beta}$ are penalty matrices acting on $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. Here $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_\beta$ represent the vectors of smoothing parameters; the length of $\boldsymbol{\lambda}_\theta$ depends of the structure of the data $\boldsymbol{y}$ and the model matrix $\boldsymbol{B}$ while that of $\boldsymbol{\lambda}_\beta$ is a function of the structure of the $\mathcal{C}_j$'s and the model matrix $\breve{\boldsymbol{B}}$. For instance, in the case of a mortality table with age-dependant dispersion parameters, $\boldsymbol{P}_{\boldsymbol{\lambda}_\theta}$ is given by (5.5), and $\boldsymbol{P}_{\boldsymbol{\lambda}_\beta} = \lambda_\beta \boldsymbol{\Delta}_x' \boldsymbol{\Delta}_x$ (under the assumption that $\breve{\boldsymbol{B}} = \boldsymbol{B}_x$).

Optimizing $\mathcal{Q}_P^+(\cdot)$ with respect to $\boldsymbol{\theta}$ yields the penalized iterative equation

$$\left(\boldsymbol{B}' \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}} \, \boldsymbol{B} \, + \boldsymbol{P}_{\boldsymbol{\lambda}_\theta} \right) \hat{\boldsymbol{\theta}} \; \approx \; \boldsymbol{B}' \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}} \, \tilde{\boldsymbol{z}}, \tag{5.29}$$

where $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$ represents the diagonal weight matrix in the quasi-likelihood model (5.17) based on the response $\boldsymbol{y}$ with the $(\tilde{w}_{\boldsymbol{\phi}})_{kk} = 1/[\phi_{\varphi(k)} v(\tilde{\mu}_k)(g'(\tilde{\mu}_k))^2]$ on the diagonal positions, and $\tilde{\boldsymbol{z}} = \boldsymbol{B}\tilde{\boldsymbol{\theta}} + (\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) * g'(\tilde{\boldsymbol{\mu}})$ is the working variable. This form is similar to the penalized scoring algorithm encountered in the PGLM setting, the difference being that the dispersion parameters are involved in the smoothing process through the weight matrix $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$.

Similarly, optimizing $\mathcal{Q}_P^+(\cdot)$ with respect to $\boldsymbol{\beta}$ yields the penalized iterative equation

$$\left(\breve{\boldsymbol{B}}' \tilde{\boldsymbol{W}}_{\boldsymbol{d}} \, \breve{\boldsymbol{B}} \, + \, \boldsymbol{P}_{\boldsymbol{\lambda}_\beta} \right) \hat{\boldsymbol{\beta}} \; \approx \; \breve{\boldsymbol{B}}' \tilde{\boldsymbol{W}}_{\boldsymbol{d}} \, \tilde{\boldsymbol{z}}_{\boldsymbol{d}}, \tag{5.30}$$

where $\tilde{\boldsymbol{W}}_{\boldsymbol{d}}$ is the diagonal weight matrix in the GLM based on a gamma response $\boldsymbol{d} = (d_{\mathcal{C}_1}, ..., d_{\mathcal{C}_\mathcal{K}})'$ from (5.27), and $\tilde{\boldsymbol{z}}_{\boldsymbol{d}}$ is the corresponding working variable. That is, $\tilde{\boldsymbol{z}}_{\boldsymbol{d}} = \breve{\boldsymbol{B}}\breve{\boldsymbol{\beta}} + (\boldsymbol{d} - \boldsymbol{\phi}) * h'(\boldsymbol{\phi})$, and the diagonal elements of $\tilde{\boldsymbol{W}}_{\boldsymbol{d}}$ are given by $(\tilde{w}_{\boldsymbol{d}})_{u,u} = n_u \phi_u^2$.

We note that equations (5.29) and (5.30) are the penalized versions of the scoring equations arising from (5.25) and (5.26) but written in matrix form. The precise form of the weight matrices $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{W}}_{\boldsymbol{d}}$ depends on the form of the link functions $g(\cdot)$ and $h(\cdot)$.

For given values of the smoothing parameters $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_\beta$, the estimation process consists of iterating between (5.29) (the $\boldsymbol{\theta}$-step) and (5.30) (the $\boldsymbol{\beta}$-step) until conver-

gence is achieved. For the estimation of $\boldsymbol{\lambda}_\theta$ and $\boldsymbol{\lambda}_\beta$, one can step outside the likelihood framework and use BIC (as we did in Section 5.1), which is given under the Poisson assumption by:

$$\mathsf{BIC} \;=\; \mathsf{D} \;+\; \log(n) \times \nu, \tag{5.31}$$

where $\mathsf{D} = \sum \hat{d}_k$ is the residual deviance, and ν is the effective dimension of the model. For Poisson count data, (5.31) is appropriate when the value of the dispersion is close to 1; however, if the data are over(under)-dispersed, the deviance will tend to be large(small), with the result that the deviance will also tend to be over(under)-weighted in (5.31). This implies that the effective dimension will also tend to be large(small); we end up by under(over)-smoothing our data, as shown by the red lines in Figure 5.6. We correct this inappropriate weighting by adjusting the deviance in each class; this gives the scaled BIC as

$$\mathsf{BICS} \;=\; \sum_{k=1}^{n} \frac{\hat{d}_k}{\phi_{\varphi(k)}} \;+\; \log(n) \times \nu, \tag{5.32}$$

a generalization of the scaled criterion used by Heuer (1997). Clearly, if there is no over(under)-dispersion in the data, then BIC and BICS are equivalent. The complete estimation algorithm can now be summarized as follows.

(1) Initialize $\phi_u = 1$, $\forall u \in \mathcal{C}$.

(2) Update $\boldsymbol{\mu}$ by solving (5.29) in $\boldsymbol{\theta}$ with $\boldsymbol{\lambda}_\theta$ selected by minimizing (5.32).

(3)
- If $|\mathcal{C}|$ is small, update the ϕ_u to their extended quasi-likelihood estimates given by (5.22), ie $\hat{\phi}_u = d_u$.
- Else, update $\boldsymbol{\phi}$ by solving (5.30) in $\boldsymbol{\beta}$, with $\boldsymbol{\lambda}_\beta$ chosen by minimizing the BIC related to (5.27).

(4) Repeat (2) and (3) until convergence is achieved.

In our applications, we will refer to this algorithm as the *full extended quasi-likelihood scheme*.

5.3.2 Bias adjustment

We have already remarked after (5.22) that d_u reduces to the maximum likelihood estimate of σ^2 in the normal distribution case. This estimate is biased downward and, in the same way, the maximum extended quasi-likelihood estimate of the dispersion parameters also tends to be biased downward (see Figure 5.9). This stems from the fact that (5.23) holds only at the true value of $\boldsymbol{\mu}$ while $\boldsymbol{\mu}$ is generally unknown. An alternative approach is to estimate $\boldsymbol{\mu}$ by maximizing criterion (5.28) as before, ie, by solving the iterative equation (5.29), but to look for a different estimate for $\boldsymbol{\phi}$. A potential candidate (analogous to the unbiased estimate of σ^2 in standard normal regression) is the *bias corrected mean Pearson statistic* in each class:

$$d_u^* = \frac{1}{n_u - \nu_u} \sum_{k \in \varphi^{-1}(u)} \frac{(y_k - \hat{\mu}_k)^2}{v(\hat{\mu}_k)}, \quad u \in \mathcal{C}, \tag{5.33}$$

where ν_u is the contribution of the class u to the total effective dimension ν. Intuitively from (2.52), we estimate these ν_u's by

$$\nu_u \; = \sum_{k \in \varphi^{-1}(u)} \frac{\partial\, g(\hat{\mu}_k)}{\partial \hat{z}_k} \; = \sum_{k \in \varphi^{-1}(u)} H_{kk}, \quad u \in \mathcal{C}, \tag{5.34}$$

where the H_{kk} are the diagonal elements of the hat matrix $\boldsymbol{H}$ related to (5.29).

If the number of classes, $\mathcal{K}$, is small then ϕ_u is estimated by d_u^*; otherwise, we proceed as follows. Instead of relying on the (penalized) extended quasi-likelihood of model (5.17) to estimate $\boldsymbol{\phi}$, we assume a full quasi-likelihood framework for the 'observations' $\boldsymbol{d}^* = (d_{\mathcal{C}_1}^*, ..., d_{\mathcal{C}_\mathcal{K}}^*)'$:

$$\mathbb{E}\,[d_u^*] = \phi_u, \quad var(d_u^*) = \tau^* \times v^*(\phi_u),\ \forall u \in \mathcal{C}, \quad h(\boldsymbol{\phi}) = -1/\boldsymbol{\phi} = \breve{\boldsymbol{B}}\boldsymbol{\beta}, \tag{5.35}$$

for some variance function $v^*(\cdot)$ to be specified, and the additional nuisance parameter τ^* to be estimated. In summary, the estimation algorithm is modified as follows:

(1) Initialize $\phi_u = 1$, $\forall u \in \mathcal{C}$.

(2) Update $\boldsymbol{\mu}$ by solving (5.29) in $\boldsymbol{\theta}$ with $\boldsymbol{\lambda}_\theta$ selected by minimizing (5.32).

(3)
- If $|\mathcal{C}|$ is small, update the ϕ_u to their Pearson estimates given by (5.33), ie $\hat{\phi}_u = d_u^*$.

- Else, update $\boldsymbol{\phi}$ by fitting the smooth model (5.35), with $\boldsymbol{\lambda}_\beta$ chosen by minimizing the related BICS.

(4) Repeat (2) and (3) until convergence is achieved.

In our applications, we will refer to this algorithm as the *bias corrected scheme*.

5.4 Applications

In this section, we shall apply the full quasi-likelihood and the bias corrected schemes to mortality data. First, through a simulation in sub-section 5.4.1, we illustrate how over-dispersion affects the smoothing process and how the use of our two schemes leads to improved estimates. Second, in sub-section 5.4.2, we use both schemes to fit two-dimensional mortality data and describe how they lead to consistent forecasts, as opposed to the inconsistent forecast with the standard Poisson model shown by the red lines in Figure 5.6.

5.4.1 A simulation exercise

We conduct a simulation exercise with two aims: first, to illustrate how dispersion affects the smoothing process and second, to show how the use of the bias corrected scheme (as well as the full extended quasi-likelihood scheme) gives rise to an improved estimate of the true mortality curve. This simulation exercise will be split into two parts: first, a portfolio without duplicates, and second, one with duplicates (ie, over-dispersed).

5.4.1.1 A simulation exercise without duplicates

Figure 5.7 shows log mortality rates for years 1950 to 2006 for males aged 75 from the CMI assured lives data set. For the purpose of these simulation exercises, we suppose that underlying log mortality follows the fitted quadratic curve shown, ie, $\log \tau_{75,j} = Q(t_j)$. We now suppose that we have central exposure $E_{75,j} = 1000$ at each year t_j and assume that the number of deaths (claims) come from the Poisson distribution: $D_{75,j} \sim \mathcal{P}oiss\left(E_{75,j} \exp[Q(t_j)]\right)$. We simulate from this model and estimate the underlying mortality curve using PB-splines with cubic B-splines, second-order

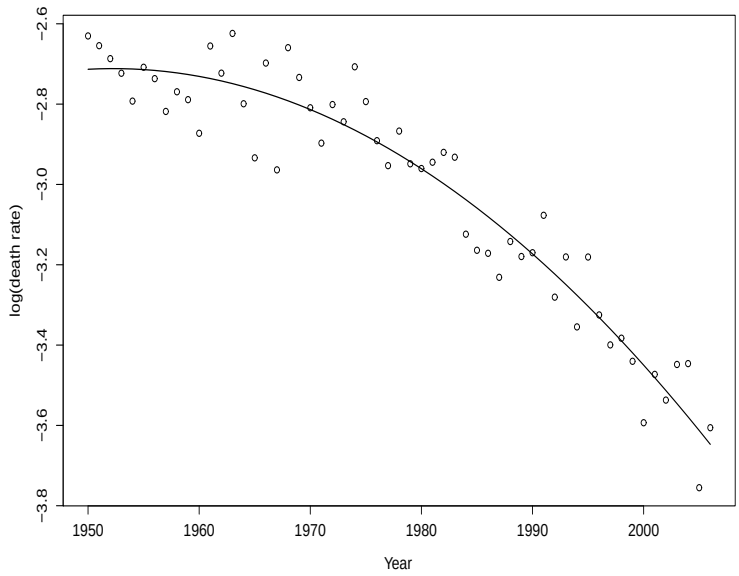

Figure 5.7: *Observed mortality rates for CMI assured lives, males age 75, together with the fitted quadratic curve.*

penalty and smoothing parameter chosen by minimizing BIC. This exercise is repeated 2000 times.

With Poisson errors we have $\phi = 1$. For each simulation s, $s = 1, ..., 2000$, we compute the mean square error

$$\mathsf{MSE} = \frac{1}{n_t}\sum_{j=1}^{n_t}(\log\,\hat{\tau}_{75,j} - \log\,\tau_{75,j})^2 = \frac{1}{n_t}\sum_{j=1}^{n_t}(\log\hat{\tau}_{75,j} - Q(t_j))^2, \tag{5.36}$$

an overall measure of the quality of the fit, and we also compute the bias corrected Pearson estimate $\hat{\phi}_s$ of ϕ using (5.33). The mean of the MSEs (over the 2000 simulations) was 0.001604.

We now perform a second round of smoothing for each of our 2000 simulations; for the s th simulation, we set $\phi = \hat{\phi}_s$ and re-estimate the force of mortality with the penalized iterative equation (5.29) and select the smoothing parameter with BICS defined in (5.32). The mean of the MSEs was very little changed at 0.001614.

In conclusion, since the quasi-likelihood extends the usual likelihood approach, the MSEs obtained with the two approaches are essentially equal when the Poisson assumption does hold. In the next section we discuss the situation when the presence of duplicates systematically introduces over-dispersion into the problem. Here we will

see a much stronger effect of over-dispersion, and both the bias corrected and the full extended quasi-likelihood schemes outperform the likelihood approach.

5.4.1.2 A simulation exercise with duplicates

In Section 5.4.1.1 we considered a portfolio of 1000 distinct policyholders in each year, where each policyholder was exposed to risk for one year. Now we consider males aged 75 again, but we suppose that we have a portfolio of 1000 policies in each year made up as follows: we have 200 policyholders with a single policy, 150 policyholders with two policies, 100 policyholders with three policies and 50 policyholders with four policies. Hence, we have 500 distinct policyholders (classified into four categories) with a total of 1000 policies, an average of two policies per policyholder; this is consistent with Richards and Currie (2009) where the average number of policies per person for the 5% of lives with the total largest pension is 1.84.

Denoting by $C_{75,j}$ the number of claims observed in year t_j, we have $\mathbb{E}[C_{75,j}] = 1000\tau_{75,j}$ and $var(C_{75,j}) = 2500\tau_{75,j}$. Thus, the (theoretical) dispersion parameter is $\phi = 2.5$ in each year, ie, the variance of claim numbers has been inflated by a factor of 2.5 relative to that of the real number of deaths. Finally we suppose that a policyholder in year t is subject to the same (quadratic) mortality as in the previous section and we repeat the previous simulation exercise for each category of policyholder, and derive the simulated number of claims.

The left boxplot in Figure 5.8 shows a summary of the MSEs arising from fitting the Poisson model to the simulated data. The mean of the MSEs was 0.00810 (compared to 0.001604 without duplicates); the presence of over-dispersion has had a negative impact on the smoothing process. We perform a second round of smoothing with the estimated values of the ϕ's incorporated into the estimation. The boxplot on the right in Figure 5.8 shows the MSEs after this second round of smoothing; the mean of the MSEs was 0.00396, a drop of more than 50%. The mean of the estimated $\hat{\phi}$'s (over the 2000 simulations) was 2.39; this is in agreement with the (theoretical) value $\phi = 2.5$.

The effective dimension of the fitted model gives another perspective on the effect of over-dispersion. Ignoring over-dispersion gave a mean effective dimension of 7.64, well in excess of 3, the true dimension of the model, while including over-dispersion reduced the mean effective dimension to 3.74. In general (by examining (5.29) and

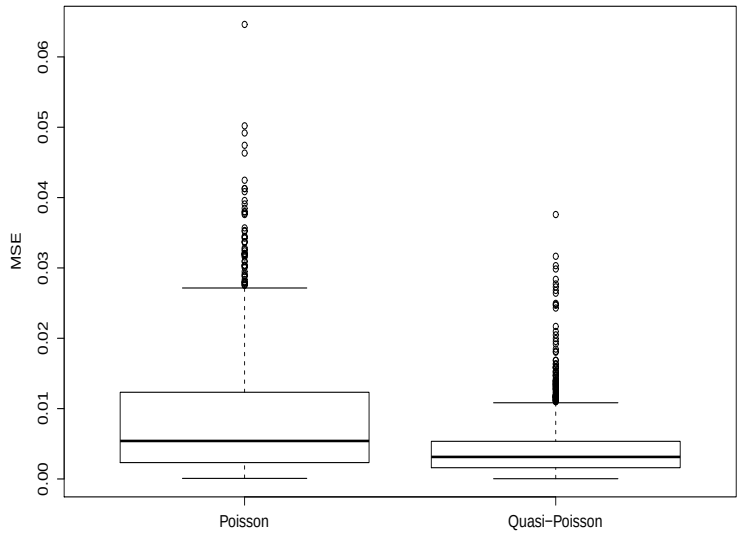

Figure 5.8: *Boxplot of* MSE *(mean square error) obtained from fitting the smooth Poisson model (left) and the smooth quasi-Poisson model (right) to the simulated/duplicated data described in Section 5.4.1.2.*

(5.32)), the flexibility of the fitted curve is reduced by the inclusion of over-dispersion parameters into the estimation process. This has important consequences for forecasting where less volatile curves lead to more stable forecasts.

5.4.2 Modelling and forecasting over-dispersed mortality tables

We shall now describe how our two schemes apply to mortality tables. For illustration, we return to the ONS mortality data for males in Figure 5.5 and consider three models specified as follows.

First, for comparison purposes, we re-consider the standard Poisson model described in (5.1)-(5.2), and we refer to it as Model1.

Second, we incorporate over-dispersion into the modelling process by replacing the Poisson assumption with a first and second moment assumption, as in (5.17). A starting model is to assume an over-dispersed Poisson model with a common over-dispersion parameter for all observations:

$$\text{Model2}: \ \mathbb{E}\left[D_{ij}\right] = E_{ij} \times \tau_{ij}, \ var(D_{ij}) = \phi \times \mathbb{E}\left[D_{ij}\right], \ \log(\boldsymbol{\tau}) = \boldsymbol{B}_x \boldsymbol{\Theta} \boldsymbol{B}_t'. \qquad (5.37)$$

Note that (5.37) is a special case of model (5.17), in which all the observations are assumed to belong to the same class; the variance function is the identity. The structure of the over-dispersion here is very simple, but it is useful for understanding the effect of the dispersion parameters on the smoothing and forecasting process. A refinement of Model2 is to allow the dispersion to be age dependent, that is

$$\text{Model3}: \ \mathbb{E}\left[D_{ij}\right] = E_{ij} \times \tau_{ij}, \ var(D_{ij}) = \phi_i \times \mathbb{E}\left[D_{ij}\right], \ \log(\boldsymbol{\tau}) = \boldsymbol{B}_x \boldsymbol{\Theta} \boldsymbol{B}_t'; \quad (5.38)$$

once again, (5.38) is a special case of model (5.17), where the classes in $\mathcal{C}$ comprise the observations of the same age, and the variance function is the identity.

We can extend Model3 by allowing the amount of dispersion to vary not only in age but also in time, leading to two surfaces to be estimated: the mortality surface and the dispersion surface. However as we shall see below, Model1, Model2 and Model3 are sufficient to illustrate the impact of the full quasi-likelihood scheme or the bias corrected and so, we shall omit this extension in this thesis.

5.4.2.1 Estimation

Model1 is fitted as described in Section 5.1. Model2 and Model3 are each fitted with both the full extended quasi-likelihood and the bias corrected schemes. In both schemes we applied the penalty matrix (5.5) to the coefficient $\boldsymbol{\theta} = vec(\boldsymbol{\Theta})$ to achieve smoothness, and the GLAM representation is used to speed up the computation. Model2 (with a single dispersion parameter) does not require second stage smoothing; we simply update (until convergence) the value of ϕ either to its extended quasi-likelihood estimate or to its Pearson estimate (as the case may be) given respectively by

$$2\sum_{i,j}\left(D_{ij}\log(\frac{D_{ij}}{E_{ij}\hat{\tau}_{ij}}) - D_{ij} + E_{ij}\hat{\tau}_{ij}\right)/n_x n_t \quad \text{and} \quad \sum_{i,j}\frac{(D_{ij} - E_{ij}\hat{\tau}_{ij})^2}{E_{ij}\hat{\tau}_{ij}}/(n_x n_t - \nu)\cdot$$

In contrast, for Model3 a second stage smoothing process is implemented to get smooth estimates of the dispersion parameters. This second stage modelling process is easier in the full extended quasi-likelihood scheme in comparison with the bias corrected scheme.

Indeed, in the latter approach, the "observed" dispersions given by (5.33) require

the estimation of the dimension contributions $\nu_u = \nu_i$. We compute them by adding up the appropriate entries on the diagonal of the hat matrix $\boldsymbol{H} = \boldsymbol{B}\, cov(\hat{\boldsymbol{\theta}})\, \boldsymbol{B}'\boldsymbol{W}_{\boldsymbol{\phi}}$, where $cov(\hat{\boldsymbol{\theta}})$ is given in (5.9). Using the GLAM idea and the fact that $\boldsymbol{W}_{\boldsymbol{\phi}}$ is diagonal, we write

$$\begin{aligned} diag(\boldsymbol{H}) &= diag[(\boldsymbol{B}\, cov(\hat{\boldsymbol{\theta}})\, \boldsymbol{B}')\boldsymbol{W}_{\boldsymbol{\phi}}] \\ &= diag[\boldsymbol{B}\, cov(\hat{\boldsymbol{\theta}})\, \boldsymbol{B}'] * diag(\boldsymbol{W}_{\boldsymbol{\phi}}),\ n_x n_t \times 1 \\ &\equiv [(\boldsymbol{B}_x \square \boldsymbol{B}_x)\hat{\boldsymbol{S}}_{\boldsymbol{\theta}}(\boldsymbol{B}_t \square \boldsymbol{B}_t)'] * \boldsymbol{\mathcal{W}}_{\boldsymbol{\phi}},\ n_x \times n_t, \end{aligned} \tag{5.39}$$

where the matrices $\hat{\boldsymbol{S}}_{\boldsymbol{\theta}}$ and $\boldsymbol{\mathcal{W}}_{\boldsymbol{\phi}}$ are derived as in (5.13) and (5.14). Thus, these contributions ν_i's can now be computed simply by adding up the entries of the corresponding rows in (5.39).

Additionally, the full quasi-likelihood uses the gamma distribution (5.27) with a known shape parameter, whereas the bias corrected sheme assumes an unknown dispersion parameter τ^* that needs to be estimated, especially if one want to compute the confidence band around the smoothed dispersions. In this case, τ^* can be estimated by optimizing the extended quasi-likelihood arising from model (5.35). In our applications we set the variance function, $v^*(\phi_u)$, in (5.35) to ϕ_u^2/n_u.

5.4.2.2 Results and comments

Figure 5.6 shows the profile views resulting from fitting Model1 (red), Model2 (green) and Model3 (blue and yellow; the yellow lines are hidden below the blue ones) to the ONS male data; some statistics are also provided in Table 5.1. As we can notice from these graphics, the estimated force of mortality obtained with the full extended quasi-likelihood scheme (yellow, hidden by the blue lines) can scarcely be distinguished by eye from those obtained with the bias corrected scheme (blue), since the difference between the estimated over-dispersion parameters from both schemes is not substantial as shown in Figure 5.9. However, this latter graphic illustrates how the full extended quasi-likelihood scheme tends to under-estimate the dispersion parameters compared to the bias corrected scheme (although the difference is very small here).

We make some comments on the results in Table 5.1. First, as measured by

Table 5.1: *Comparative statistics for* Model1, Model2 *and* Model3. *FEQS and BCS stand for the full extended quasi-likelihood and the bias corrected schemes respectively. Also,* $tr(\boldsymbol{h})$ *refers to the effective dimension of the smoothed dispersions.*

	Model1	**Model2**		**Model3**	
		FEQS	**BCS**	**FEQS**	**BCS**
$\boldsymbol{\lambda}_\theta = (\lambda_x, \lambda_t)$	(394, 0.01)	(433, 1471)	(394, 1485)	(546, 1233)	(546, 1232)
$\boldsymbol{\phi}$	1	3.58	3.66	see Figure 5.9	see Figure 5.9
$tr(\boldsymbol{H})$	163	71	70	75	74
Deviance	10755	12131	12147	12140	12154
BICS	12082	3958	3884	3956	3877
$tr(\boldsymbol{h})$	//	//	//	11.73	11.82

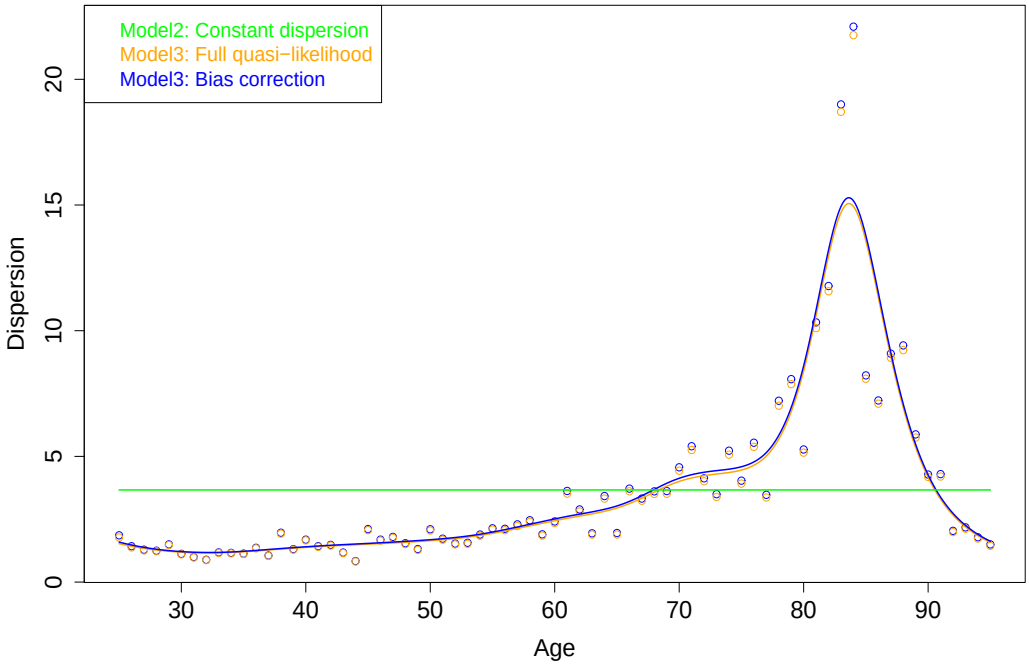

Figure 5.9: *"Raw" and smoothed estimates the* ϕ_x*'s from* Model3 *using respectively the full extended quasi-likelihood scheme (blue), and the bias corrected scheme (red). The green horizontal line corresponds to the estimated dispersion in* Model1 *(with constant dispersion, here estimated by its Pearson statistic at convergence).*

BICS, Model2 gives a much superior fit to the data compared to Model1, with Model3 a further improvement. Second, the less flexible the fitted model, the larger the deviance; however, the deviance in Model1 is computed under the assumption that $\phi = 1$, and the relative increase in deviance from 10755 to 12147 as we go from Model1 to Model2 is more than compensated for by the additional variance of Model2

(as measured by its estimated over-dispersion parameter $\hat{\phi} = 3.66$).

There are two important conclusions to be drawn from this example: the first concerns the central forecasts and the second the width of the confidence intervals. First, we consider the central forecasts. The effective dimension of the model under the Poisson assumption is 163. If we include the dispersion parameter in the estimation process, the effective dimension is reduced to 70 with Model2, and to 74 (or 75) with Model3; this corresponds to a more robust, ie, less volatile fit. In general, this seems to us to be a desirable property for a forecast. Second, we consider the effect on the confidence intervals. Taking account of the over-dispersion has led to narrower confidence intervals. At first sight, such narrowing of the width of the confidence interval arising from the inclusion of over-dispersion in the smoothing process may seem counter-intuitive from a stochastic point of view. We argue as follows: smoothing is a compromise between (a) increasing roughness, ie, improved fit to data and (b) increasing smoothness, ie, poorer fit to data. When we include ϕ we down-weight the fit to data (the deviance is increased from 10755 for Model 1 to 12147 for Model2 and 12154 for Model3) and so decrease the volatility of the fitted model. The width of the forecast confidence intervals reflects our faith in the selected model and we will have more faith in the future direction of a forecast in a less volatile model; we conclude that the width of the confidence interval will be decreased. Both of these effects can be seen in Figure 5.6.

5.5 Dispersion and the negative binomial

An alternative approach to account for over-dispersion is through the Negative binomial distribution ($\mathcal{NB}$), which arises as a mixture of Poisson and gamma distributions (Lawless, 1987; Thurston et al., 2000; Li et al., 2009). Let $\boldsymbol{y} = (y_1, ..., y_n)'$ be our count data as in Section 5.3. By introducing extra variables ς_k such that

$$y_k|\varsigma_k \sim \mathcal{P}oiss(\varsigma_k \times \mu_k), \quad \text{with} \quad \varsigma_k \sim Gamma(a, \frac{1}{a}), \tag{5.40}$$

it follows that the marginal distribution of y_k is the negative binomial:

$$y_k \sim \mathcal{NB}\,(\mu_k,\ a), \tag{5.41}$$

with associated probability mass function

$$f(y_k = y) = \frac{\Gamma(y+a)}{\Gamma(a)\Gamma(y+1)} \left(\frac{\mu_k}{\mu_k + a}\right)^y \left(\frac{a}{\mu_k + a}\right)^a. \tag{5.42}$$

Under this model, we have

$$\mathbb{E}[y_k] = \mu_k \quad \text{and} \quad var(y_k) = \mu_k(1 + \frac{\mu_k}{a}). \tag{5.43}$$

Thus, $\mathcal{NB}(\mu_k, a)$ has the same mean as $\mathcal{P}oiss(\mu_k)$, but with inflated variance. The level of over-dispersion in the data is then modulated by the parameter a in the sense that decreasing values of a correspond to increasing levels of over-dispersion; for this reason, a is usually refer to as the *dispersion index parameter* (Lawless, 1987). In the limit, ie, as $a \to \infty$, the $\mathcal{NB}$ tends to the underlying Poisson distribution. Thence, under the $\mathcal{NB}$ assumption, the adjustment for over-dispersion is incorporated in the distribution and so the resulting deviance given by

$$\mathsf{D} = 2\sum_k \left[y_k \log\left(\frac{y_k}{\hat{\mu}_k}\right) - (y_k + a)\log\left(\frac{y_k + a}{\hat{\mu}_k + a}\right)\right], \tag{5.44}$$

is correspondingly adjusted. For instance, we can show that the components $(y_k + a)\log[(y_k + a)/(\hat{\mu}_k + a)]$ in (5.44) are decreasing functions of a, and this property modulates the impact of over-dispersion on the resulted BIC.

For a fixed value of a, it is not difficult to see that the probability function (5.42) belongs to the exponential family. Hence, setting $\log(\boldsymbol{\mu}) = \boldsymbol{B\theta}$, with smoothness constraints on $\boldsymbol{\theta}$, and following Section 2.6 yields the penalized scoring algorithm

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P}_\theta)\hat{\boldsymbol{\theta}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{5.45}$$

where $\tilde{\boldsymbol{z}} = \boldsymbol{B}\tilde{\boldsymbol{\theta}} + (\boldsymbol{y} - \tilde{\boldsymbol{\mu}})/\tilde{\boldsymbol{\mu}}$, and $\tilde{\boldsymbol{W}} = diag\,[a\tilde{\boldsymbol{\mu}}/(a\mathbf{1}_n + \tilde{\boldsymbol{\mu}})]$. This penalized scoring algorithm is different from that obtained under the Poisson approach only through the weights.

In practice, the dispersion index parameter a is usually unknown and so it must be estimated too. For a fixed value of $\boldsymbol{\mu}$, this can be done by optimizing the profile

log-likelihood, which is given (up to an additive constant) by

$$n[a\log(a) - \log(\Gamma(a)] \; - \; (\boldsymbol{y} + a\mathbf{1}_n)' \log(\boldsymbol{\mu} + a\mathbf{1}_n) \; + \; \mathbf{1}_n' \, \Gamma(\boldsymbol{y} + a\mathbf{1}_n) \cdot \tag{5.46}$$

The complete estimation scheme (adapted from Thurston et al. (2000)) can now be summarised as follows:

(1) Initialize the estimate $\hat{a}$ of a to some large value (so that the corresponding $\mathcal{NB}$ approaches the underlying Poisson).

(2) Set $a = \hat{a}$, and get the estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ via the iterative equations (5.45).

(3) Set $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$, and update the estimate $\hat{a}$ of a, by optimizing the profile log-likelihood (5.46) with respect to a.

(4) Repeat (2) and (3) until convergence.

We have implemented this procedure on some mortality data (not shown here), and although it addresses the over-dispersion problem, it tends to produce fitted surfaces that are less robust than those obtained under the full quasi-likelihood or bias corrected scheme, especially as the amount of dispersion increases. Additionally, the level of heterogeneity can vary substantially across the data (e.g. over-dispersion in the ONS mortality data, illustrated in Figure 5.9), in which case a single dispersion index parameter a may be optimistic. One would ideally incorporate varying dispersion index parameters $a_{\varphi(k)}$ (eventually with some constraints on these $a_{\varphi(k)}$), where $\varphi(\cdot)$ assigns observations to classes as in Section 5.3. But in doing so, we should bear in mind that the conditional estimation of the $a_{\varphi(k)}$ (in step (3) above) becomes a multidimensional optimization problem. The advantage of the quasi-likelihood procedure in this context is that it reduces this optimization to a weighted least squares problem. Furthermore, the quasi-likelihood approach is able to detect and handle under-dispersion, which the negative binomial model cannot.

5.6 Conclusion

In this Chapter, we have illustrated some problems that arise from the smoothing and forecasting of mortality rates under the standard Poisson assumption. Motivated by

these problems, we have described a general class of models for count data through a two-stage joint modelling of mean and dispersion effects. Our approach is similar to that used by Renshaw (1992) for actuarial data. We have extended his work to models for two-dimensional data in age and time with general smooth functions for both the mean and dispersion parameters, and described how this leads to consistent smoothing and forecasting of mortality data. This joint smoothing is computational intensive if fitted with standard GLM procedures, since we must alternate between the smoothing of the mortality surface and the smoothing of dispersion parameters until convergence. Fortunately, the efficient array algorithms of Currie et al. (2006) enabled us to speed up the calculations; these algorithms impact on the computations but not on model formulation.

Chapter 6

Joint models for classification, comparison and forecasting of mortality tables

In Chapter 5, we dealt with the smoothing and forecasting of a mortality table. Suppose now that we are concerned with the modelling of several mortality tables. The simplest approach would be to address each table separately. However, observed mortality from different tables/populations can have some affinity/connection. For population data for instance, it is well known that male mortality is higher than that of female; additionally, mortality of these two groups often has some similarities in their dynamism. Analogously, the CMI pensioner data are of two types: data by lives and data by amounts. The first type consists of the number of claims (viewed as deaths by lives) and the number of policies at risk (viewed as exposure to risk by lives), similar to the assured lives data used in the previous Chapter. The second type consists of the total amount claimed (viewed as death by amounts) and the total amount at risk (viewed as exposure-to-risk by amounts). These two types of data lead to the concept of *mortality by lives* and *mortality by amounts*, and it is well known in the insurance world that the latter is lighter than the former.

In general, to what extent can the dynamism of mortality tables be similar? Can we build economical models for mortality tables which are similar (in some way)? In this chapter, we propose a class of additive models for the economical joint modelling, comparison, and consistent forecasting of "similar" mortality tables. The first com-

ponent of our models describes a (common) two-dimensional smooth surface (viewed as the reference), and the remaining components depict the relative differences (*gaps*) between these tables. As we shall see, this approach enables the classification of populations into different categories.

The work in this Chapter is an extended version of Biatat and Currie (2010). Section 6.1 introduces this joint modelling approach in the context of two populations. Section 6.2 presents the extension to multi-populations. Section 6.3 concentrates on the fitting and the computational demands. Section 6.4 presents some applications, and we close with a brief discussion in Section 6.5.

6.1 Joint modelling of two populations/tables

We consider mortality data for two populations, consisting of deaths and exposures, arranged in $n_x \times n_t$ matrices $\boldsymbol{D}^{[r]}$ and $\boldsymbol{E}^{[r]}$, $r = 1, 2$, such that the rows and columns of $\boldsymbol{D}^{[r]}$ and $\boldsymbol{E}^{[r]}$ are classified by ages and by years as in the previous Chapter. We suppose that the number of deaths $D_{ij}^{[r]}$ at age x_i in calendar year t_j in population r can be described approximately by the over-dispersed Poisson assumption as follows:

$$\mathbb{E}\left[D_{ij}^{[r]}\right] = \mu_{ij}^{[r]}, \quad var(D_{ij}^{[r]}) = \phi_{\varphi(r,i,j)} \times \mu_{ij}^{[r]}, \quad \text{with } \mu_{ij}^{[r]} = E_{ij}^{[r]} \times \tau_{ij}^{[r]}, \tag{6.1}$$

where $\tau_{ij}^{[r]}$ and $E_{ij}^{[r]}$ represent the force of mortality and the exposure-to-risk corresponding to $D_{ij}^{[r]}$, $\varphi(\cdot)$ is the function assigning observations to classes (similar to that in Section 5.3), and $\phi_{\varphi(r,i,j)}$ is the dispersion parameter in the cell corresponding to age x_i and year t_j in population r, $i = 1, ..., n_x$, $j = 1, ..., n_t$, $r = 1, 2$. We will denote by $\boldsymbol{\tau}^{[r]}$ and $\boldsymbol{\mu}^{[r]}$ the $n_x \times n_t$ matrices containing the $\tau_{ij}^{[r]}$ and $\mu_{ij}^{[r]}$ respectively. It is worth mentioning that the approach in this Chapter remains valid if one prefers the negative binomial model instead of (6.1).

The joint and economical aspects of the model are constructed through the predictor. The key idea is the following: if the dynamism of our two populations is similar, then the relative variation of their forces of mortality can be captured by a moderate number of parameters. In other words, if we set (conceptually) a two-dimensional smooth surface for the predictor of the force of mortality in population 1 (viewed as the reference), then the predictor of the force of mortality in population 2 can be

captured by adding a "simple" gap to this reference. Hence, we write

$$\begin{cases} \log(vec(\boldsymbol{\tau}^{[1]})) &= (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta}^{[1]} \\ \log(vec(\boldsymbol{\tau}^{[2]})) &= (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta}^{[1]} \;+\; \mathsf{Gap} \end{cases} \tag{6.2}$$

where the coefficient $\boldsymbol{\theta}^{[1]}$ is subject to the penalty matrix

$$\boldsymbol{P}^{[1]} = \lambda_{1,x}\,(\boldsymbol{I}_{c_t} \otimes \boldsymbol{\Delta}_x'\boldsymbol{\Delta}_x) \;+\; \lambda_{1,t}\,(\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t \otimes \boldsymbol{I}_{c_x})\cdot \tag{6.3}$$

We use this representation (6.2) to incorporate similarity or difference (between the two populations) in the modelling process, and this enables us to introduce concepts such as *strong similarity, similarity in age/time*, and *weak similarity.* As we shall see, this approach is well connected to the nested curves investigated in Chapter 3.

6.1.1 Strong similarity

We describe two populations as *strongly similar* if the relative variation of their mortality predictor is constant in age and time. That is, the corresponding gap can be expressed as

$$\mathsf{Gap} \;=\; (\mathbf{1}_{n_t} \otimes \mathbf{1}_{n_x})\theta^{[2]}. \tag{6.4}$$

The structure of the gap here is very simple and is governed by a single regression parameter $\theta^{[2]}$ and so, the mortality predictors of the two populations move in parallel in age and time.

6.1.2 Similarity in age/time

These populations would be viewed as *similar in time* if they are strongly similar in the time direction; ie their relative variation is constant in time, but flexibly smooth in age. In this case, the predictors of the two populations move in parallel in the time direction and we express the gap as

$$\mathsf{Gap} \;=\; (\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x)\boldsymbol{\theta}^{[2,x]}. \tag{6.5}$$

In order to obtain the smoothness of the gap component, we take a rich basis $\boldsymbol{B}_x$ and apply the difference penalty matrix, $\lambda_{2,x}\,\boldsymbol{\Delta}_x'\boldsymbol{\Delta}_x$, on the coefficient vector $\boldsymbol{\theta}^{[2,x]}$.

For simplicity, we have used the same B-spline basis $\boldsymbol{B}_x$ in both the two-dimensional reference surface (6.2) and the gap component (6.5); but in general we can allow them to be different.

Likewise, these populations would be considered as *similar in age* if the gap is constant in age and flexibly smooth in time. In this case, their mortality predictors move in parallel in the age direction, and the gap is structured as

$$\mathsf{Gap} \;=\; (\boldsymbol{B}_t \otimes \mathbf{1}_{n_x})\boldsymbol{\theta}^{[2,t]}; \tag{6.6}$$

smoothness in this case is achieved by applying the penalty matrix, $\lambda_{2,t}\,\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t$, on $\boldsymbol{\theta}^{[2,t]}$.

6.1.3 Weak similarity

We shall consider these populations as *weakly similar* if the relative variation of their mortality predictor is additively smooth in age and time, in which case the gap takes the form

$$\mathsf{Gap} \;=\; (\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x)\boldsymbol{\theta}^{[2,x]} \,+\, (\boldsymbol{B}_t \otimes \mathbf{1}_{n_x})\boldsymbol{\theta}^{[2,t]}. \tag{6.7}$$

A difficulty that arises from this representation is the identifiability of these two gap components. To cope with it, we let the first component $(\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x)\boldsymbol{\theta}^{[2,x]}$ capture both the constant and the smooth age-dependent components of the gap, while the second term $(\boldsymbol{B}_t \otimes \mathbf{1}_{n_x})\boldsymbol{\theta}^{[2,t]}$ models only the smooth year dependent component of this gap. Hence, we smooth $\boldsymbol{\theta}^{[2,x]}$ and $\boldsymbol{\theta}^{[2,t]}$, and for identifiability, we additionally shrink $\boldsymbol{\theta}^{[2,t]}$. This yields the penalty matrix $\lambda_{2,x}\,\boldsymbol{\Delta}_x'\boldsymbol{\Delta}_x$ on $\boldsymbol{\theta}^{[2,x]}$, and $\lambda_{2,t}\,\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t + \breve{\lambda}_{2,t}\,\boldsymbol{I}_{c_t}$ on $\boldsymbol{\theta}^{[2,t]}$.

Note that *strongly similar* populations are nested within *similar in time/age* populations, which are in turn nested within *weakly similar* populations. For populations which are not *weakly similar*, we might either model them independently, or consider a flexible gap consisting of some combination of an age-dependent component and a time-dependent component. A typical example here would be to set a gap as a bilinear function of age and time, similar to the Lee-Carter form and its extensions.

6.2 Unified representation and generalization

In each of the similarity scenarios discussed in Section 6.1, the joint predictor for the two mortality surfaces can be expressed in the form

$$\log(vec[\boldsymbol{\tau}^{[1]} : \boldsymbol{\tau}^{[2]}]) = \begin{bmatrix} \boldsymbol{B}_t \otimes \boldsymbol{B}_x & \mathbf{0} \\ \boldsymbol{B}_t \otimes \boldsymbol{B}_x & \boldsymbol{G}^{[2]} \end{bmatrix} \begin{pmatrix} \boldsymbol{\theta}^{[1]} \\ \boldsymbol{\theta}^{[2]} \end{pmatrix} \tag{6.8}$$

where $\boldsymbol{G}^{[2]}$, the regression matrix component corresponding to the gap, is given explicitly by

$$\boldsymbol{G}^{[2]} = \begin{cases} \mathbf{1}_{n_t} \otimes \mathbf{1}_{n_x} & \text{for strong similarity} \\ \mathbf{1}_{n_t} \otimes \boldsymbol{B}_x & \text{for similarity in time} \\ \boldsymbol{B}_t \otimes \mathbf{1}_{n_x} & \text{for similarity in age} \\ [\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x : \boldsymbol{B}_t \otimes \mathbf{1}_{n_x}] & \text{for weak similarity} \end{cases} \tag{6.9}$$

and $\boldsymbol{\theta}^{[2]}$, the corresponding regression vector, is subject to the penalty matrix $\boldsymbol{P}^{[2]}$ defined by

$$\boldsymbol{P}^{[2]} = \begin{cases} 0 & \text{for strong similarity} \\ \lambda_{2,x}\, \boldsymbol{\Delta}_x' \boldsymbol{\Delta}_x & \text{for similarity in time} \\ \lambda_{2,t}\, \boldsymbol{\Delta}_t' \boldsymbol{\Delta}_t & \text{for similarity in age} \\ blockdiag(\lambda_{2,x}\, \boldsymbol{\Delta}_x' \boldsymbol{\Delta}_x,\ \lambda_{2,t}\, \boldsymbol{\Delta}_t' \boldsymbol{\Delta}_t + \breve{\lambda}_{2,t}\, \boldsymbol{I}_{c_t}) & \text{for weak similarity} \end{cases} \tag{6.10}$$

Hence, these expressions (6.8)-(6.9)-(6.10) give a unified representation of the predictor for this joint modelling, under our four similarity regimes. Extension of this concept to more than two populations is straightforward. Let us consider the general case of p populations, $p > 1$. We set population 1 as the reference. Under an appropriate similarity regime, we model the mortality predictor in these p populations as

$$\begin{cases} \log(vec(\boldsymbol{\tau}^{[1]})) &= (\boldsymbol{B}_t \otimes \boldsymbol{B}_a)\, \boldsymbol{\theta}^{[1]} \\ \log(vec(\boldsymbol{\tau}^{[r]})) &= (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta}^{[1]} + \boldsymbol{G}^{[r]}\boldsymbol{\theta}^{[r]}, \quad r = 2, ..., p, \end{cases} \tag{6.11}$$

where $\boldsymbol{G}^{[r]}$, the regression matrix components for the gaps, are obtained as in (6.9), and the associated regression coefficients $\boldsymbol{\theta}^{[r]}$ are subject to appropriate smoothing and shrinkage constraints via the penalty matrix $\boldsymbol{P}^{[r]}$ derived as in (6.10).

In (6.11), we have a two-dimensional surface $(\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\,\boldsymbol{\theta}^{[1]}$ that describes an important part of the common dynamism in these populations. This surface corresponds to the predictor of the force of mortality in population 1 and it is viewed as the base reference for the others as well. The relative variation (with respect to this reference) of the mortality predictor in the r th population is then summarized by the corresponding gap $\boldsymbol{G}^{[r]}\boldsymbol{\theta}^{[r]}$.

If we now introduce the joint matrix of forces of mortality as $\boldsymbol{\tau} = [\boldsymbol{\tau}^{[1]} : ... : \boldsymbol{\tau}^{[p]}]$, then the joint predictor can be written compactly as

$$\log(vec(\boldsymbol{\tau})) = \boldsymbol{B}\boldsymbol{\theta}, \tag{6.12}$$

where the regression matrix $\boldsymbol{B}$ is defined as

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_t \otimes \boldsymbol{B}_x & \boldsymbol{0} \\ \boldsymbol{1}_{p-1} \otimes (\boldsymbol{B}_t \otimes \boldsymbol{B}_x) & \boldsymbol{G} \end{bmatrix}, \tag{6.13}$$

with $\boldsymbol{G} = blockdiag(\boldsymbol{G}^{[2]}, ..., \boldsymbol{G}^{[p]})$, and $\boldsymbol{\theta} = vec(\boldsymbol{\theta}^{[1]}, ..., \boldsymbol{\theta}^{[p]})$. These regression coefficients $\boldsymbol{\theta}$ are subject to the penalty matrix

$$\boldsymbol{P} = blockdiag(\boldsymbol{P}^{[1]}, ..., \boldsymbol{P}^{[p]}), \tag{6.14}$$

with $\boldsymbol{P}^{[1]}$ given by (6.3), and $\boldsymbol{P}^{[r]}$, $r = 2, ..., p$, defined as in (6.10).

We note that the approach in this Chapter can be seen as a generalization of the nested curves presented in Chapter 3, with the difference that here, subjects correspond to tables/populations, the common/overall effect is measured by a surface (confounded with the effect of the reference table/population) and the subject effects are measured by the gap components.

Also, we point out in passing that an alternative representation of the linear predictor (6.11) is

$$\begin{cases} \log(vec(\boldsymbol{\tau}^{[1]})) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_a)\,\boldsymbol{\theta}^{[1]} \\ \log(vec(\boldsymbol{\tau}^{[r]})) = \log(vec(\boldsymbol{\tau}^{[r-1]})) + \boldsymbol{G}^{[r]}\boldsymbol{\alpha}^{[r]},\ r = 2, ..., p. \end{cases} \tag{6.15}$$

These two representations are equivalent and they therefore lead to the same estimates. However, the interpretation of the gap components in these two representa-

tions is different. Indeed, in (6.11) each gap component $\boldsymbol{G}^{[r]}\boldsymbol{\theta}^{[r]}$ quantifies the gap between the corresponding population and the reference (population 1); whereas in (6.15) the gap components $\boldsymbol{G}^{[r]}\boldsymbol{\alpha}^{[r]}$ represent the gaps between consecutive surfaces. We prefer (6.11) to (6.15) because the regression matrix arising from (6.11) is more sparse than that from (6.15).

6.3 Estimation and computational considerations

The fitting can be based either on the full quasi-likelihood scheme or the bias corrected scheme. In both cases, the estimation of the joint coefficients $\boldsymbol{\theta}$ is concentrated around the optimization of the penalized quasi-likelihood as described in Section 5.3.

Let us denote by $\boldsymbol{\phi}^{[r]}$ the $n_x \times n_t$ matrix with entries $\phi_{ij}^{[r]} = \phi_{\varphi(i,j,r)}$. If we set $\boldsymbol{\phi} = [\boldsymbol{\phi}^{[1]} : ... : \boldsymbol{\phi}^{[p]}]$, $\boldsymbol{D} = [\boldsymbol{D}^{[1]} : ... : \boldsymbol{D}^{[p]}]$, $\boldsymbol{E} = [\boldsymbol{E}^{[1]} : ... : \boldsymbol{E}^{[p]}]$, and $\boldsymbol{\mu} = \boldsymbol{E} * \boldsymbol{\tau}$ with $\boldsymbol{\tau} = [\boldsymbol{\tau}^{[1]} : ... : \boldsymbol{\tau}^{[p]}]$, then this optimization results in the penalized iterative equations

$$\left(\boldsymbol{B}'\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\,\boldsymbol{B} + \boldsymbol{P}\right)\hat{\boldsymbol{\theta}} \;\approx\; \boldsymbol{B}'\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\,\tilde{\boldsymbol{z}}, \tag{6.16}$$

with $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}} = diag[vec(\tilde{\boldsymbol{\mu}}/\boldsymbol{\phi})]$, and $\tilde{\boldsymbol{z}} = \boldsymbol{B}\tilde{\boldsymbol{\theta}} + vec\,[(\boldsymbol{D} - \tilde{\boldsymbol{\mu}})/\tilde{\boldsymbol{\mu}}]$.

As for the estimation of the dispersion parameters, a second stage smoothing may be necessary depending on the structure of $\varphi(\cdot)$, especially in the intermediate case where the dispersion is age-dependent in these populations, ie $\varphi(r, i, j) = (r, i)$. Whatever structure one chooses, the computational burden from one population to this joint modelling of multi-populations increases dramatically. For instance, the joint model for two populations under the scenario of *weak similarity* evolves 4 smoothing parameters and in the general case of p populations, we have $2p$ smoothing parameters. The search for the optimal values of these parameters in this context can be very time consuming. This problem is further complicated by the fact that we must alternate the estimation of these p mortality surfaces with that of the dispersion parameters, until convergence. The use of the GLAM idea offers a huge benefit here.

For instance, let us consider the weighted inner product $\boldsymbol{B}'\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\,\boldsymbol{B}$ in (6.16). We split it as

$$\boldsymbol{B}'\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\,\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}' & \boldsymbol{\Omega}_{22} \end{bmatrix}, \tag{6.17}$$

with

$$\begin{cases} \boldsymbol{\Omega}_{11} = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)' \sum_{r=1}^{p} \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[r]} (\boldsymbol{B}_t \otimes \boldsymbol{B}_x) \\ \boldsymbol{\Omega}_{22} = blockdiag((\boldsymbol{G}^{[2]})' \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[2]} \boldsymbol{G}^{[2]}, ..., (\boldsymbol{G}^{[p]})' \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[p]} \boldsymbol{G}^{[p]}) \\ \boldsymbol{\Omega}_{12} = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)' [\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[2]} \boldsymbol{G}^{[2]} : ... : \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[p]} \boldsymbol{G}^{[p]}] \end{cases} \tag{6.18}$$

where $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[r]}$ is the diagonal weight (extracted from $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$) corresponding to the r th population. The three matrices $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{12}$ and $\boldsymbol{\Omega}_{22}$ in (6.18) are made up of components of the form $(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)' \boldsymbol{A} (\boldsymbol{A}_3 \otimes \boldsymbol{A}_4)$, for some $n_x n_t \times n_x n_t$ diagonal weight matrix $\boldsymbol{A}$ and conformable matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$ and $\boldsymbol{A}_4$. In order to compute such inner products efficiently, we extend the GLAM formula (5.13) to

$$(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)' \boldsymbol{A} (\boldsymbol{A}_3 \otimes \boldsymbol{A}_4), c_1 c_2 \times c_3 c_4 \equiv (\boldsymbol{A}_4 \square \boldsymbol{A}_2)' \boldsymbol{\mathcal{A}} (\boldsymbol{A}_3 \square \boldsymbol{A}_1), c_2 c_4 \times c_1 c_3, \tag{6.19}$$

where c_i represents the number of columns in $\boldsymbol{A}_i$, and $\boldsymbol{\mathcal{A}}$ is the $n_x \times n_t$ matrix obtained by re-arranging the diagonal elements of $\boldsymbol{A}$ such that $vec(\boldsymbol{\mathcal{A}}) = diag(\boldsymbol{A})$.

Additionally, the iterative equations (6.16) can be expressed as

$$\begin{bmatrix} \boldsymbol{\Omega}_{11} + \boldsymbol{P}^{[1]} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}' & \boldsymbol{\Omega}_{22} + \boldsymbol{P}_{\boldsymbol{g}} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}}^{[1]} \\ \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \end{pmatrix} \approx \begin{pmatrix} \boldsymbol{r}^{[1]} \\ \boldsymbol{r}_{\boldsymbol{g}} \end{pmatrix} \tag{6.20}$$

with

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} = vec(\hat{\boldsymbol{\theta}}^{[2]}, ..., \hat{\boldsymbol{\theta}}^{[p]}), \quad \boldsymbol{P}_{\boldsymbol{g}} = blockdiag(\boldsymbol{P}^{[2]}, ..., \boldsymbol{P}^{[p]}),$$

$$\boldsymbol{r}^{[1]} = (\boldsymbol{B}_t \otimes \boldsymbol{B}_a)' \sum_{r=1}^{p} \tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}^{[r]} \tilde{\boldsymbol{z}}^{[r]} \quad \text{and} \quad \boldsymbol{r}_{\boldsymbol{g}} = vec((\boldsymbol{G}^{[2]})' \boldsymbol{W}_{\boldsymbol{\phi}}^{[2]} \tilde{\boldsymbol{z}}^{[2]}, ..., (\boldsymbol{G}^{[p]})' \boldsymbol{W}_{\boldsymbol{\phi}}^{[p]} \tilde{\boldsymbol{z}}^{[p]}), \tag{6.21}$$

where $\tilde{\boldsymbol{z}}^{[r]}$ is the vector made up of the entries of $\tilde{\boldsymbol{z}}$ corresponding to the r th population. With this decomposition, if we set $\boldsymbol{\Omega}_{\boldsymbol{g}} = \boldsymbol{\Omega}_{22} + \boldsymbol{P}_{\boldsymbol{g}}$, then the iterative equations (6.16) can be partitioned as

$$\begin{cases} \left(\boldsymbol{\Omega}_{11} + \boldsymbol{P}^{[1]} - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{\boldsymbol{g}}^{-1} \boldsymbol{\Omega}_{12}'\right) \hat{\boldsymbol{\theta}}^{[1]} \approx \boldsymbol{r}_1 - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{\boldsymbol{g}}^{-1} \boldsymbol{r}_g \\ \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \approx \boldsymbol{\Omega}_{\boldsymbol{g}}^{-1} \left(\boldsymbol{r}_{\boldsymbol{g}} - \boldsymbol{\Omega}_{12}' \hat{\boldsymbol{\theta}}^{[1]}\right). \end{cases} \tag{6.22}$$

Note that each component in (6.21) can be computed efficiently in the GLAM framework using (5.11). Further, $\boldsymbol{\Omega}_{\boldsymbol{g}}$ is a block diagonal matrix and so, its inverse required in (6.22) is readily obtained by taking the inverse of these diagonal blocks.

6.4 Applications

We shall now illustrate this joint modelling approach on some population and actuarial data. The graphics presented here use a single dispersion for each population, ie $\varphi(r, i, j) = r$; hence, the $\phi_{\varphi(r,i,j)}$ reduce to $\phi_{\varphi(r,i,j)} = \phi_r$.

6.4.1 Population data

We first consider the male and female mortality in Japan from ages 30 to 90 and years 1960 to 2005. The residuals from our model corresponding to *weak similarity* applied to the mortality of these two groups show that the model fits "well" (some profile views are shown in Figure 6.1); we conclude that these two populations are *weakly similar*. As a result, the relative variation between their mortality predictor is simply summarized by the smooth age-dependent gap component (left panel in Figure 6.2) and the smooth year-dependent gap component (right panel in Figure 6.2), where the female mortality has been set as the reference. Similarly, the residuals and profile views indicate that the mortality dynamism for males of Italy, Denmark and US are *weakly similar*. This is illustrated by some profile views in Figure 6.3. Here we consider Italy as the reference; the smooth age-dependent gap component as well as the smooth year-dependent gap component are then shown in Figure 6.4.

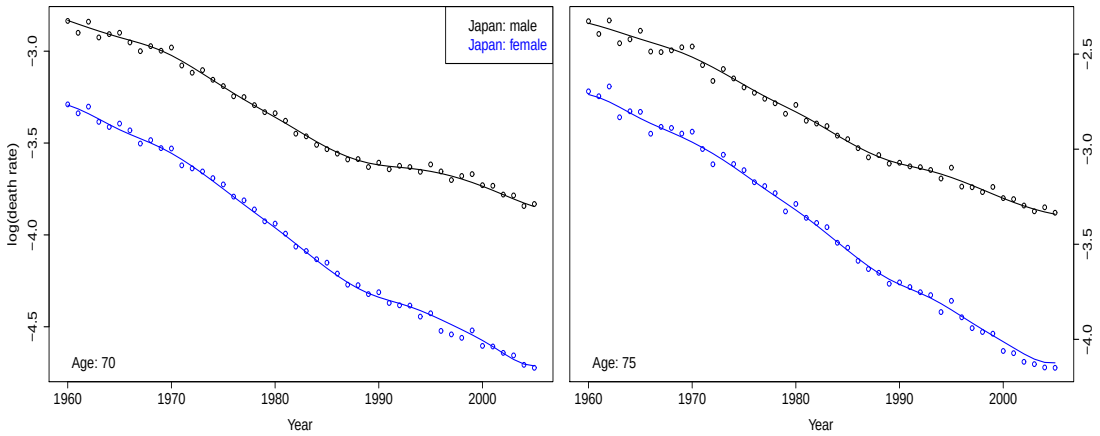

Figure 6.1: *Profile views from the joint modelling of male and female mortality in Japan, using the model corresponding to weak similarity.*

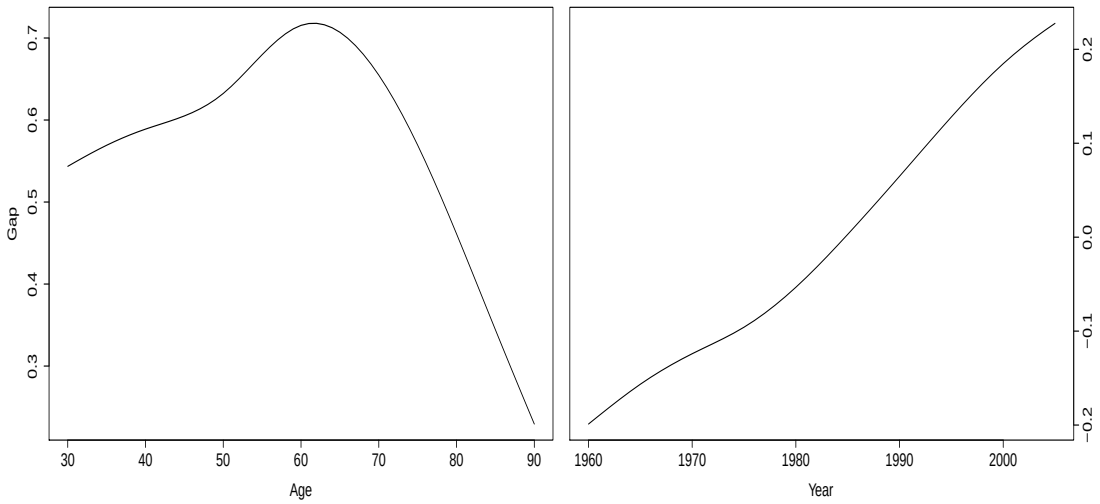

Figure 6.2: *The smooth age-dependent gap component (left) and the smooth year-dependent gap component (right) in the male and female populations in Japan; here the female mortality is set as the reference.*

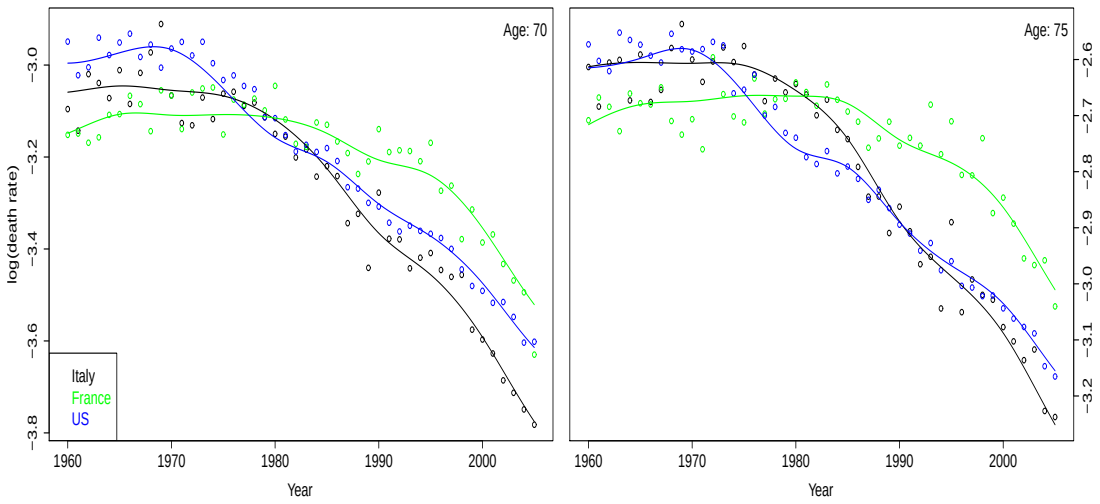

Figure 6.3: *Profile views from the joint modelling of male mortality in Italy, Denmark and US, using the model corresponding to weak similarity.*

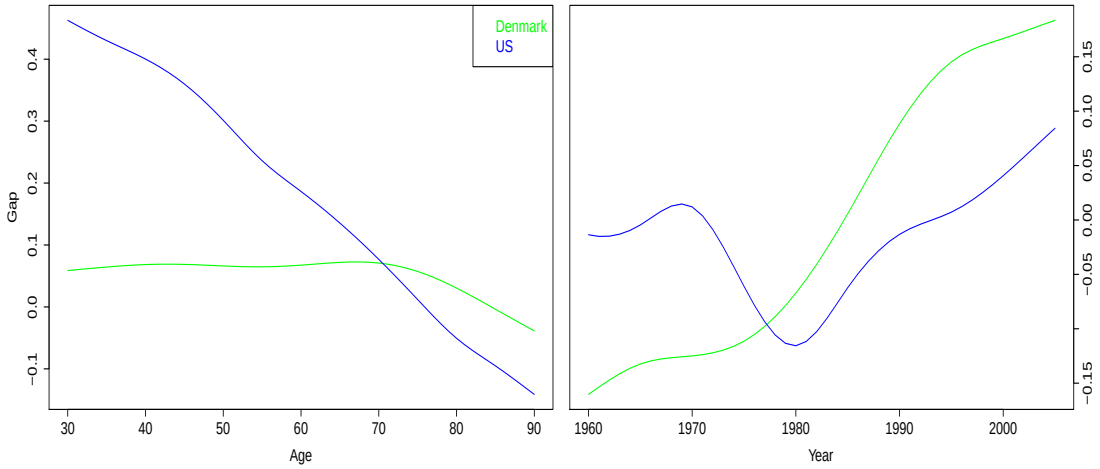

Figure 6.4: *The smooth age-dependent gap component (left) and the smooth year-dependent gap component (right) for male mortality of Italy, Denmark and US; here Italy is set as the reference.*

6.4.2 Actuarial data

We now turn to the CMI data. Most of the material in this Section was part of Djeundje and Currie (2010b). As we have already mentioned in the first paragraph of this Chapter, the CMI data are available both by lives and amounts, and we denote by $\boldsymbol{D}^{[L]}$ and $\boldsymbol{E}^{[L]}$ the matrices of deaths and exposures by lives; likewise $\boldsymbol{D}^{[A]}$ and $\boldsymbol{E}^{[A]}$ will represent the equivalent matrices by amounts. As we have discussed in Section 5.2, the appropriate modelling of the lives data suffers from the problem of duplicate policies and so, we consider the quasi-Poisson model

$$\mathbb{E}[\boldsymbol{D}_{ij}^{[L]}] = \boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[L]}; \quad var(\boldsymbol{D}_{ij}^{[L]}) = \phi^{[L]}\boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[L]}, \tag{6.23}$$

where the $\tau_{ij}^{[L]}$ represent the force of mortality by lives and $\phi^{[L]}$ quantifies the dispersion in the lives data; we will return to this model.

For data by amounts, the problem of duplicates arises in two ways:

(a) the original problem of duplicate policies in the portfolio, and

(b) the much larger problem that amounts data by its very nature contains duplication on a grand scale, since a single death, even of a pensioner with a single pension, generates not one but multiple claims, namely the amount of pension at risk; see also Forfar et al. (1988).

To model these amounts data, let $\boldsymbol{M} = \boldsymbol{E}^{[A]}/\boldsymbol{E}^{[L]}$ be the mean amount at risk per life; the matrix of raw mortality by amounts is

$$\boldsymbol{D}^{[A]}/\boldsymbol{E}^{[A]} = (\boldsymbol{D}^{[A]}/\boldsymbol{M})/(\boldsymbol{E}^{[A]}/\boldsymbol{M}) = \boldsymbol{D}^{*}/\boldsymbol{E}^{[L]}, \tag{6.24}$$

where $\boldsymbol{D}^{*} = \boldsymbol{D}^{[A]}/\boldsymbol{M}$. If all policies are of the same amount, then $\boldsymbol{D}^{*} = \boldsymbol{D}^{[L]}$ and so we may assume that

$$\boldsymbol{D}^{*}_{ij} \sim \mathcal{P}oiss(\boldsymbol{E}^{[L]}_{ij}\tau^{[A]}_{ij}), \tag{6.25}$$

where $\tau^{[A]}_{ij}$ is the force of mortality by amounts. This is known as the method of scaling. This approach attempts to solve problem (b), but it ignores (a). Hence we update (6.25) to

$$\mathbb{E}[\boldsymbol{D}^{*}_{ij}] = \boldsymbol{E}^{[L]}_{ij}\tau^{[A]}_{ij}, \quad var(\boldsymbol{D}^{*}_{ij}) = \phi^{[*A]}\boldsymbol{E}^{[L]}_{ij}\tau^{[A]}_{ij}, \tag{6.26}$$

where $\phi^{[*A]}$ is the dispersion in the scaled amounts data. That is, in (6.26), we address problem (b) by the method of scaling and we tackle problem (a) through quasi-likelihoods.

Alternatively to (6.26), we can follow Renshaw and Hatzopoulos (1996) and consider the model defined through

$$\mathbb{E}[\boldsymbol{D}^{[A]}_{ij}] = \boldsymbol{E}^{[A]}_{ij}\tau^{[A]}_{ij}, \quad var(\boldsymbol{D}^{[A]}_{ij}) = \phi^{[A]}\boldsymbol{E}^{[A]}_{ij}\tau^{[A]}_{ij}, \tag{6.27}$$

where $\phi^{[A]}$ is the dispersion parameter in the amounts data. That is, in (6.27), we address problems (a) and (b) directly through quasi-likelihoods.

With these details, we can estimate the force of mortality surface by lives independently of that by amounts as described in the previous Chapter. But, in the Insurance world, it is well known that mortality by lives is higher than that by amounts, and such an independent modelling approach turns out to be inappropriate. Combining (6.26) or (6.27) with (6.23) and using the modelling framework in Section 6.1, we

consider two joint models for the lives and amounts data defined as

$$\begin{cases} \mathbb{E}[\boldsymbol{D}_{ij}^{[L]}] = \boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[L]}; \quad var(\boldsymbol{D}_{ij}^{[L]}) = \phi^{[L]}\boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[L]} \\ \\ \mathbb{E}[\boldsymbol{D}_{ij}^{*}] = \boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[A]}; \quad var(\boldsymbol{D}_{ij}^{*}) = \phi^{[*A]}\boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[A]} \end{cases} \tag{6.28}$$

and

$$\begin{cases} \mathbb{E}[\boldsymbol{D}_{ij}^{[L]}] = \boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[L]}; \quad var(\boldsymbol{D}_{ij}^{[L]}) = \phi^{[L]}\boldsymbol{E}_{ij}^{[L]}\tau_{ij}^{[L]} \\ \\ \mathbb{E}[\boldsymbol{D}_{ij}^{[A]}] = \boldsymbol{E}_{ij}^{[A]}\tau_{ij}^{[A]}; \quad var(\boldsymbol{D}_{ij}^{[A]}) = \phi^{[A]}\boldsymbol{E}_{ij}^{[A]}\tau_{ij}^{[A]}. \end{cases} \tag{6.29}$$

The estimates from the fitting of these two joint models under the *similarity in time* scenario shows that the dynamisms in the mortality by lives and by amounts are *similar in time*; profile views from the fitting of (6.28) are shown in Figure 6.5; the graphics obtained from (6.29) are very similar to these ones and so we have omitted them here. Thus, the mortality predictor surface by lives sits on top of the surface by amounts in such a way that the cross-sections in time by age are parallel.

An important point about these two joint models is the values of $\phi^{[L]}$, $\phi^{[A]}$ and $\phi^{[*A]}$. With (6.28) we find $(\hat{\phi}^{[L]}, \hat{\phi}^{[*A]}) = (1.54, 7.9)$, whereas with (6.29) we get $(\hat{\phi}^{[L]}, \hat{\phi}^{[A]}) = (1.55, 9299)$. As we can see, $\hat{\phi}^{[L]}$ is essentially the same in the two models; however, $\hat{\phi}^{[A]}$ is much larger than $\hat{\phi}^{[*A]}$, the reason being that $\hat{\phi}^{[*A]}$ accounts only for problem (b) whereas $\hat{\phi}^{[A]}$ accounts for both problems (a) and (b).

This joint model corresponding to the *similarity in time* scenario has a particular importance for forecasting in life insurance, since it ensures that the extrapolated trends in time at different ages for mortality by lives and by amounts do not cross each other; this property is illustrated in Figure 6.6.

6.5 Conclusion

In this chapter we have proposed a class of joint models for classifying populations according to their mortality tables. When two (or more) populations turn out to be similar in some way, our joint models lead to simple and graphical comparisons of these tables. An attractive feature of these models is that, once the components are set up, the fitting is reduced to the penalized scoring algorithm with appropriate

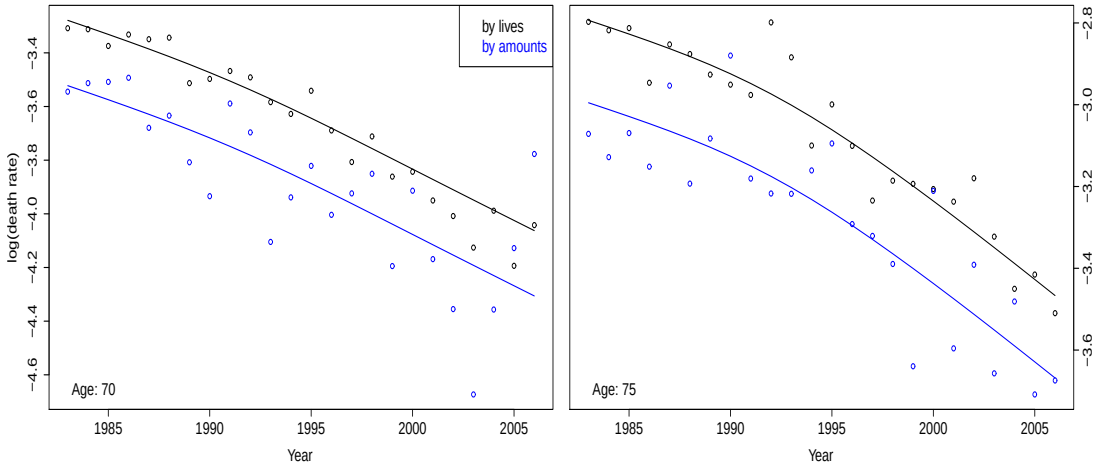

Figure 6.5: *Profile views of the CMI mortality by lives and by amounts, using the joint model corresponding to similarity in time.*

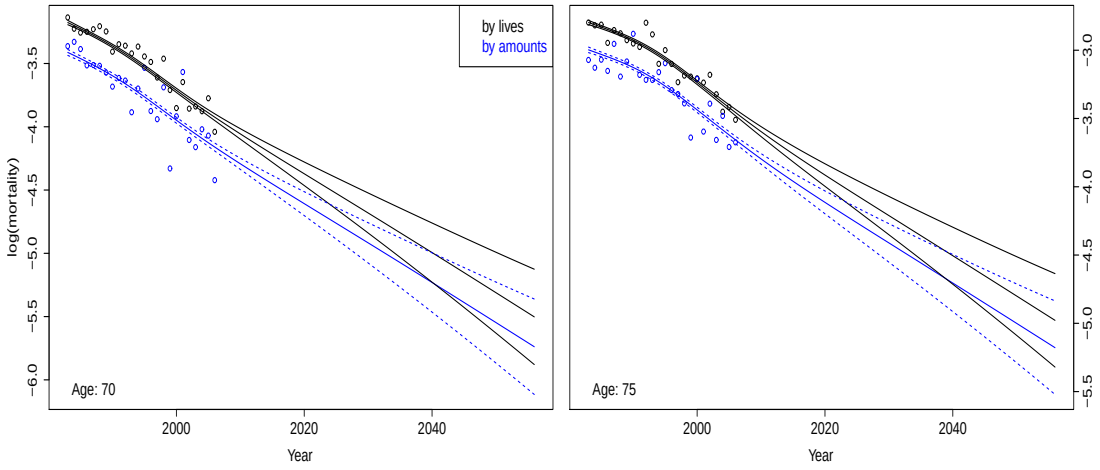

Figure 6.6: *These profile views illustrate how the joint model corresponding to similarity in time allows us to preserve the dynamism in the CMI mortality by lives and by amounts over the extrapolated range.*

components. In our formulation, all populations/tables have the same dimension $(n_x \times n_y)$, and they are equally important in the model in the sense that they are all involved (simultaneously) in the fitting of both the underlying two-dimensional reference surface and the gap components.

In some situations, we may have at our disposal some large and robust population together with a small and less robust one. A typical context is Life Insurance, where young companies have very little mortality data. In such cases, we do not treat these

populations equally. Instead, we can follow Currie (2009) and "piggy-back" the small population on the large one as follows: first, we fit a smooth surface to the robust population; second, we set this fitted surface as an offset in the modelling of the small population, and finally, the small population data are used to estimate the gap component(s), depending on the appropriate similarity regime.

Another area where this joint modelling technique can be exploited is the modelling of mortality by cause of death. Here, we would split the joint data (according to different causes of death) into many tables possibly with small counts, and then model these data/tables using the idea of reference table and gap components.

Chapter 7

Smoothing correlated data: the mortality improvement factor with period and cohort effects

In the two preceding Chapters, we investigated the dynamism of mortality through the smoothing and forecasting of the force of mortality. Another well known indicator of this dynamism is the *mortality improvement factor* (MIF); see Willets (1999) for example. As we shall describe in Section 7.1 below, the MIF is used to quantify the improvement or the decline in mortality as we move from one year to the next. Very often, the observed MIFs show high variability (see the upper panel in Figure 7.1) and so, it seems appropriate to smooth these observed values.

To our knowledge, previous work on smoothing MIFs has been developed (a) under isotropic smoothing, ie, the same amount of smoothing in both the age and year directions, and (b) with the assumption that the observed MIFs are independent (Andreev and Vaupel, 2005). Assumption (a) may suit some data sets, but in general, this assumption is questionable because it is not clear that the table of MIFs should be smoothed equally in the age and year directions. Further, assumption (b) is inappropriate because, from the definition of the MIF as we shall see in expressions (7.1) below, the observed MIFs across years are correlated for any given age; more precisely (as we shall prove in Section 7.1) each observed MIF is correlated with its predecessor and its successor in time. Ignoring this correlation structure may have a substantial fitting and inferential impact; see Wang (1998) for some illustrations.

In Section 7.1, we first derive an appropriate covariance structure for the MIF. Next, we formalize a smooth model for the MIF using two-dimensional PB-splines, and due to the correlation structure, we express the model in the mixed model framework and apply it to some data in Section 7.2. In Section 7.3, we capture both the year shocks and cohort effects observed in the MIF of some populations by extending our initial model to a *smooth-period-cohort* effects model. In Section 7.4, we use the bootstrap method to compute standard errors for this extended model, and we close with some concluding remarks in Section 7.5.

7.1 Modelling the mortality improvement factor

We consider mortality data stored in $n_x \times n_t$ matrices of deaths, $\boldsymbol{D}$, and exposures, $\boldsymbol{E}$, and we denote by $\boldsymbol{R}$ the corresponding matrix of death rates. That is, the entries of $\boldsymbol{R}$ are given by $R_{ij} = D_{ij}/E_{ij},\ i = 1, ..., n_x,\ j = 1, ..., n_t$. Following Willets (1999) and Andreev and Vaupel (2005), we define the mortality improvement factor at age x_i and calendar year t_j, $(i = 1, ..., n_x,\ j = 1, ..., m_t$, with $m_t = n_t - 1)$ by

$$\mathsf{MIF}(x_i, t_j) = \frac{R_{i,j+1} - R_{ij}}{R_{ij}} = \frac{R_{i,j+1}}{R_{ij}} - 1\cdot \tag{7.1}$$

Thus, a negative value of $\mathsf{MIF}(x_i, t_j)$ indicates a reduction in the mortality rate as we move from year t_j to t_{j+1} in age x_i, and the larger this negative value, the sharper this reduction will be.

An illustration of the MIFs for CMI pensioner males is shown in the upper panel in Figure 7.1. Our aim here is to smooth these observed values and one apparent way may be to smooth the rates R_{ij} as in Chapter 5, from which we compute the smooth MIFs using (7.1). Alternatively, we can follow Andreev and Vaupel (2005) and address directly the observed MIF values. In this way, we hope to cope with the inherent high volatility that may arise from the quotient in (7.1).

As in Chapters 5 and 6, we assume that the death counts D_{ij} are independent, and then it is clear from (7.1) that the quantity $1 + \mathsf{MIF}(x_i, t_j)$ is the ratio of two independent random variables. This suggests a log transformation approach. Thus,

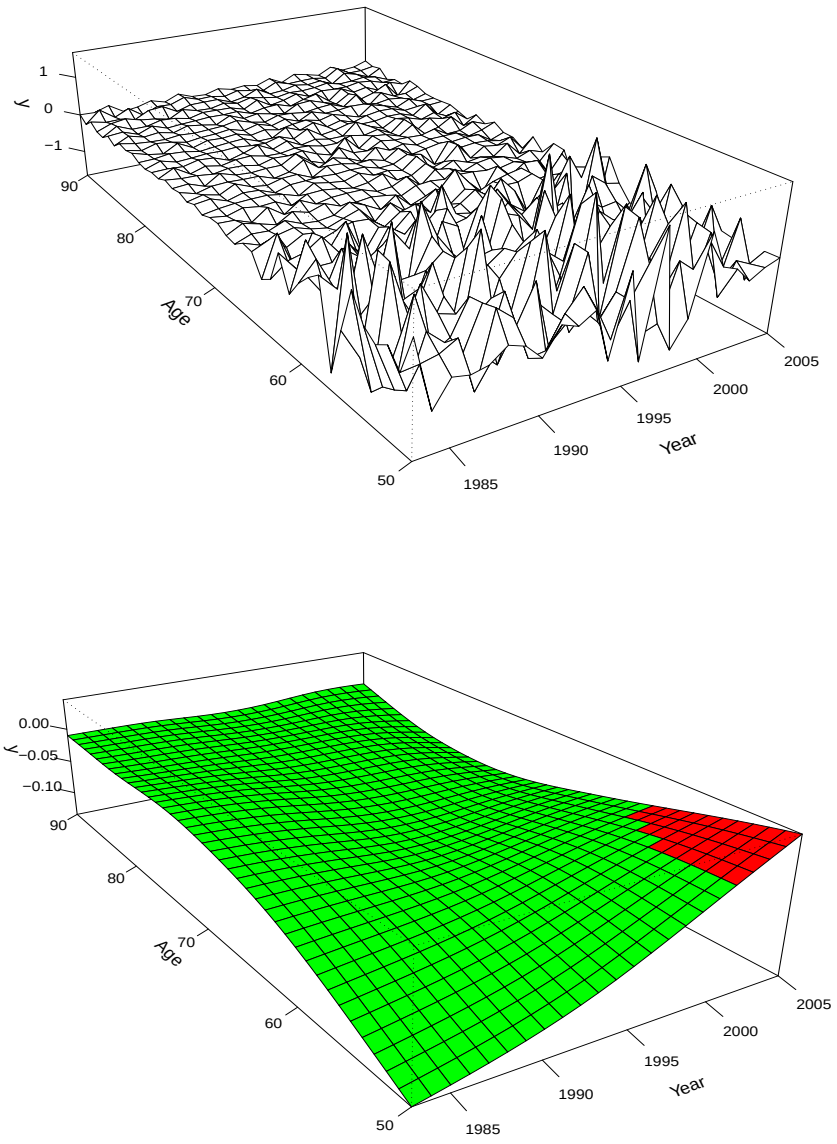

Figure 7.1: MIF *indicator,* $\boldsymbol{Y}$*, for CMI pensioner males. Upper: observed. Lower: fitted surface.*

we introduce dependent variables, denoted by Y_{ij}, as

$$Y_{ij} = \log(1 + \mathsf{MIF}(x_i, t_i)) = \log(R_{i,j+1}) - \log(R_{ij}), \tag{7.2}$$

and we refer to these Y_{ij} as the MIF *indicators*. We will denote by $\boldsymbol{Y}$ the $n_x \times m_t$ matrix of Y_{ij}, and by $\boldsymbol{Y}_{i\bullet}$ the row of $\boldsymbol{Y}$ corresponding to age x_i. Note that these MIF indicators have an equivalent interpretation as MIFs in the sense that

$$Y_{ij} > 0 \Longleftrightarrow \mathsf{MIF}(x_i, t_j) > 0\cdot$$

Prior to specifying a model for $\boldsymbol{Y}$, we first examine the correlation structure in these data.

7.1.1 Correlation structure

We consider two models for the D_{ij}: the quasi-Poisson assumption and the Negative Binomial one.

Under the quasi-Poisson model

$$\mathbb{E}[D_{ij}] = \mu_{ij}, \quad var(D_{ij}) = \phi_{\varphi(i,j)} \times \mu_{ij}, \text{ with } \mu_{ij} = E_{ij}\tau_{ij}, \tag{7.3}$$

we have,

$$\begin{aligned}
cov(Y_{i,j}, Y_{i',j'}) &= cov\left(\log(R_{i,j+1}) - \log(R_{ij}),\ \log(R_{i',j'+1}) - \log(R_{i',j'})\right) \\
&= cov\left(\log(R_{i,j+1}),\ \log(R_{i',j'+1})\right) - cov\left(\log(R_{i,j+1}),\ \log(R_{i',j'})\right) \\
&\quad - cov\left(\log(R_{i,j}),\ \log(R_{i',j'+1})\right) + cov\left(\log(R_{i,j}),\ \log(R_{i',j'})\right) \\
&= \begin{cases} var(\log(R_{i,j+1})) + var(\log(R_{i,j})) & \text{if } i = i' \text{ and } j = j' \\ -var(\log(R_{i,\max\{j,j'\}})) & \text{if } i = i' \text{ and } |j - j'| = 1 \\ 0 & \text{otherwise} \end{cases} \\
&\approx \begin{cases} \dfrac{\phi_{\varphi(i,j)}}{\mu_{ij}} + \dfrac{\phi_{\varphi(i,j+1)}}{\mu_{i,j+1}} & \text{if } i = i' \text{ and } j = j' \\ -\dfrac{\phi_{\varphi(i,\max\{j,j'\})}}{\mu_{i,\max\{j,j'\}}} & \text{if } i = i' \text{ and } |j - j'| = 1 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{7.4}$$

In a similar fashion, the negative binomial assumption

$$D_{ij} \sim \mathcal{NB}\left(\mu_{ij},\ a_{\varphi(i,j)}\right), \quad \text{with } \mathbb{E}[D_{ij}] = \mu_{ij} \text{ and } var(D_{ij}) = \mu_{ij} + \frac{\mu_{ij}^2}{a_{\varphi(i,j)}}, \tag{7.5}$$

leads to

$$cov(Y_{i,j}, Y_{i',j'}) \approx \begin{cases} \dfrac{1}{a_{\varphi(i,j)}} + \dfrac{1}{a_{\varphi(i,j+1)}} + \dfrac{1}{\mu_{ij}} + \dfrac{1}{\mu_{i,j+1}} & \text{if } i = i' \text{ and } j = j' \\ -\dfrac{1}{a_{\varphi(i,\max(j,j'))}} - \dfrac{1}{\mu_{i,\max(j,j')}} & \text{if } i = i' \text{ and } |j - j'| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{7.6}$$

Hence, in both cases, expressions (7.4) and (7.6) show that the MIFs are independent from age to age, and correlated across years. More precisely, for any given age x_i, Y_{ij} is correlated to its predecessor and to its successor in time. This results in a covariance matrix $\mathbf{\Sigma}_i := cov(\boldsymbol{Y}_{i\bullet})$ with a tri-diagonal structure for age x_i.

We introduce the joint vector of MIF indicators $\boldsymbol{y}$, defined by

$$\boldsymbol{y} = vec(\boldsymbol{Y}'); \tag{7.7}$$

ie, $\boldsymbol{y}$ is obtained by stacking the rows $\boldsymbol{Y}_{i\bullet}$ of matrix $\boldsymbol{Y}$ on top of each other into a vector. We choose to stack rows (instead of columns as usual) because this leads to the global covariance matrix $\mathbf{\Sigma} = cov(\boldsymbol{y})$ with a block diagonal structure, as

$$\mathbf{\Sigma} = cov(\boldsymbol{y}) = blockdiag(\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{n_x}), \tag{7.8}$$

indicating the independence of the MIF from age to age. Such a block diagonal structure is very convenient for handling and speeding up the estimation process.

7.1.2 Basic smooth model for the mortality improvement factor

We now suppose that the data vector $\boldsymbol{y}$ follows approximately a multivariate normal distribution

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mathcal{B}\theta},\ \sigma^2\mathbf{\Sigma}), \tag{7.9}$$

where $\boldsymbol{\theta}$ is the vector of unknown regression coefficients, and $\boldsymbol{\mathcal{B}} = \boldsymbol{B}_x \otimes \boldsymbol{B}_t$ is the two-dimensional B-spline basis built on the marginal bases. We remark that the order of the two marginal bases in this Kronecker product basis is reversed, ie $\boldsymbol{\mathcal{B}} = \boldsymbol{B}_x \otimes \boldsymbol{B}_t$ (as opposed to $\boldsymbol{B} = \boldsymbol{B}_t \otimes \boldsymbol{B}_x$ used in Chapters 5 and 6); this is a direct consequence of our definition of $\boldsymbol{y}$ in (7.7).

Andreev and Vaupel (2005) estimated the smooth MIF surface by minimizing the penalized residual sum of squares. This is implicitly equivalent to the fitting of model (7.9), but with the covariance matrix $\boldsymbol{\Sigma}$ replaced by the identity matrix, ie, assuming that the observed MIFs are independent. In contrast, we incorporate the correlation structure through the covariance matrix $\boldsymbol{\Sigma}$.

We now turn to the estimation of model (7.9), and we break it down into two stages. The first stage concerns the estimation of the covariance matrix $\boldsymbol{\Sigma}$ and we proceed as follows: we estimate smooth values $\hat{\tau}_{ij}$ of the force of mortality under the quasi-Poisson assumption (7.3) or the negative binomial assumption (7.5) as described in Chapter 5. We then compute the estimates of the μ_{ij} as $\hat{\mu}_{ij} = E_{ij}\hat{\tau}_{ij}$, plug them into formulas (7.4) or (7.6), as the case may be, and this provides us with an estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$.

The second stage concentrates on the estimation of $\boldsymbol{\theta}$. We use PB-splines, and as usual, we take rich marginal bases in age and time, and penalize the roughness of $\boldsymbol{\theta}$ with the penalty matrix

$$\boldsymbol{P}_{\boldsymbol{\theta}} = \lambda_t \boldsymbol{I}_{c_x} \otimes \boldsymbol{\Delta}_t' \boldsymbol{\Delta}_t + \lambda_x \boldsymbol{\Delta}_x' \boldsymbol{\Delta}_x \otimes \boldsymbol{I}_{c_t}.$$

Due to the presence of correlation, it is appropriate to express the model in the mixed model framework, a convenient setting for smoothing in the presence of correlated data (Wang, 1998; Krivobokova and Kauermann, 2007). Hence, following Currie et al. (2006) and Lee and Durban (2011), we use the singular value decomposition of the components of $\boldsymbol{P}_{\boldsymbol{\theta}}$ and re-parametrize the predictor $\boldsymbol{\mathcal{B}\theta}$ as

$$\boldsymbol{\mathcal{B}\theta} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u}, \tag{7.10}$$

where the coefficient vector $\boldsymbol{\alpha}$ is unpenalized and $\boldsymbol{u}$ is subject to the transformed

penalty matrix $\boldsymbol{P_u}$ defined as

$$\boldsymbol{P_u} = blockdiag\left(\lambda_t(\boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_t),\ \lambda_x(\boldsymbol{\Lambda}_x \otimes \boldsymbol{I}_2),\ \lambda_x(\boldsymbol{\Lambda}_x \otimes \boldsymbol{I}_{c_t-2}) + \lambda_t(\boldsymbol{I}_{c_x-2} \otimes \boldsymbol{\Lambda}_x)\right), \tag{7.11}$$

in which $\boldsymbol{\Lambda}_t$ and $\boldsymbol{\Lambda}_x$ represent the diagonal matrices of positive eigenvalues in the singular value decomposition of $\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t$ and $\boldsymbol{\Delta}_x'\boldsymbol{\Delta}_x$ respectively. Full details on such a re-parametrization as well as the explicit form of the regression matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are computed as in Currie et al. (2006) and Lee and Durban (2011), except that the age and time dimensions must be permuted since our data matrix $\boldsymbol{Y}$ has been transposed as shown in (7.7). We find

$$\begin{cases} \boldsymbol{X} = [\boldsymbol{1}_{n_x} : \boldsymbol{x}] \otimes [\boldsymbol{1}_{m_t} : \boldsymbol{t}] \\ \boldsymbol{Z} = [[\boldsymbol{1}_{n_x} : \boldsymbol{x}] \otimes \boldsymbol{B}_t\boldsymbol{U}_t \ :\ \boldsymbol{B}_x\boldsymbol{U}_x \otimes [\boldsymbol{1}_{m_t} : \boldsymbol{t}] \ :\ \boldsymbol{B}_x\boldsymbol{U}_x \otimes \boldsymbol{B}_t\boldsymbol{U}_t] \end{cases} \tag{7.12}$$

where $\boldsymbol{x} = (x_1, ..., x_{n_x})'$, $\boldsymbol{t} = (t_1, ..., t_{m_t})'$, and $\boldsymbol{U}_x$ and $\boldsymbol{U}_t$ are the $n_x \times (c_x - 2)$ and $m_t \times (c_t - 2)$ matrices whose columns are the eigenvectors corresponding to the positive eigenvalues of $\boldsymbol{\Delta}_t'\boldsymbol{\Delta}_t$ and $\boldsymbol{\Delta}_x'\boldsymbol{\Delta}_x$.

With this re-parametrization, fitting model (7.9) becomes equivalent to the estimation of the mixed model

$$\boldsymbol{y}|\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u},\ \sigma^2\boldsymbol{\Sigma}), \quad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0},\ \sigma^2\boldsymbol{P}_{\boldsymbol{u}}^{-1}). \tag{7.13}$$

As a result, our model can now be fitted using the restricted maximum likelihood with best linear unbiased estimator/predictor as described in Sections 4.3.1 and 4.3.2. In summary, the complete estimation machine can be written as follows:

(1) Obtain the smooth estimates $\hat{\tau}_{ij}$ of the force of mortality using either the full extended quasi-likelihood scheme, the bias corrected scheme, or the negative binomial model as described in Chapter 5.

(2) Set $\hat{\mu}_{ij} = E_{ij}\hat{\tau}_{ij}$, plug these $\hat{\mu}_{ij}$ into formula (7.4) or (7.6), and obtain an estimate $\hat{\boldsymbol{\Sigma}}$ of the covariance matrix $\boldsymbol{\Sigma}$.

(3) Plug $\hat{\boldsymbol{\Sigma}}$ into model (7.13), and fit this model using restricted maximum likelihood with best linear unbiased estimator/predictor.

Clearly, our data have a matrix structure and the regression matrix (7.12) is expressed in terms of Kronecker product; hence the GLAM representation can be used in the estimation process. In this case however, the GLAM algorithm as presented by Eilers et al. (2006) and Currie et al. (2006) needs adjusted so that it can handle the non-diagonal covariance matrix $\boldsymbol{\Sigma}$. Details of this adjustment were given by Kirkby (2009).

7.2 Applications and the need for an extension

Our first application uses the CMI pensioner data for males, from age 50 to 90 and calendar years 1983 to 2006. The corresponding MIF indicator is shown in the upper panel in Figure 7.1. We can see from this graphic that the observed MIFs at "younger" ages (less than 65) are more unstable, partly due to the relatively small number of deaths at these ages. The smoothed MIF indicator $\boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}}$ is shown on the lower panel in Figure 7.1; some profile views with associated confidence intervals are also presented in Figure 7.2. An exploration of such profile views across ages and years shows that the model fits the CMI pensioner male data well. Also, the resulting smoothed MIFs are negative (except for few cells illustrated by the red piece in the lower panel in Figure 7.6), implying that the mortality rate decreases (on average) from year to year. We now consider the England & Wales data for males. Some profile views from its smoothed MIF indicator are illustrated in Figure 7.3 and compared to those of CMI pensioners males. For the central trend, there seems to be no general conclusion that we can draw from the relative variation of the MIF in these two groups. However, the confidence intervals for the CMI are wider than those of the population data. One major justification of this difference in the standard errors is the large exposure in the population data relative to that of the CMI. Indeed from expression (7.4) or (7.6), it is clear that a large exposure leads to a small variance.

For the population data however, the fit does not look satisfactory across ages for some years, as illustrated by the profile views for years 1986 and 1992 in Figure 7.4. We have observed similar lack of fit in other population data (e.g. France, Japan, etc). This problem seems to be caused by two features in the data: (i) the presence of MIF shocks in some years, and (ii) the presence of cohort effects. These points can be seen in the observed MIF indicator in Figure 7.5, where the main cohort effect corresponds to men born in 1920 and 1921. In such a situation, a smooth surface

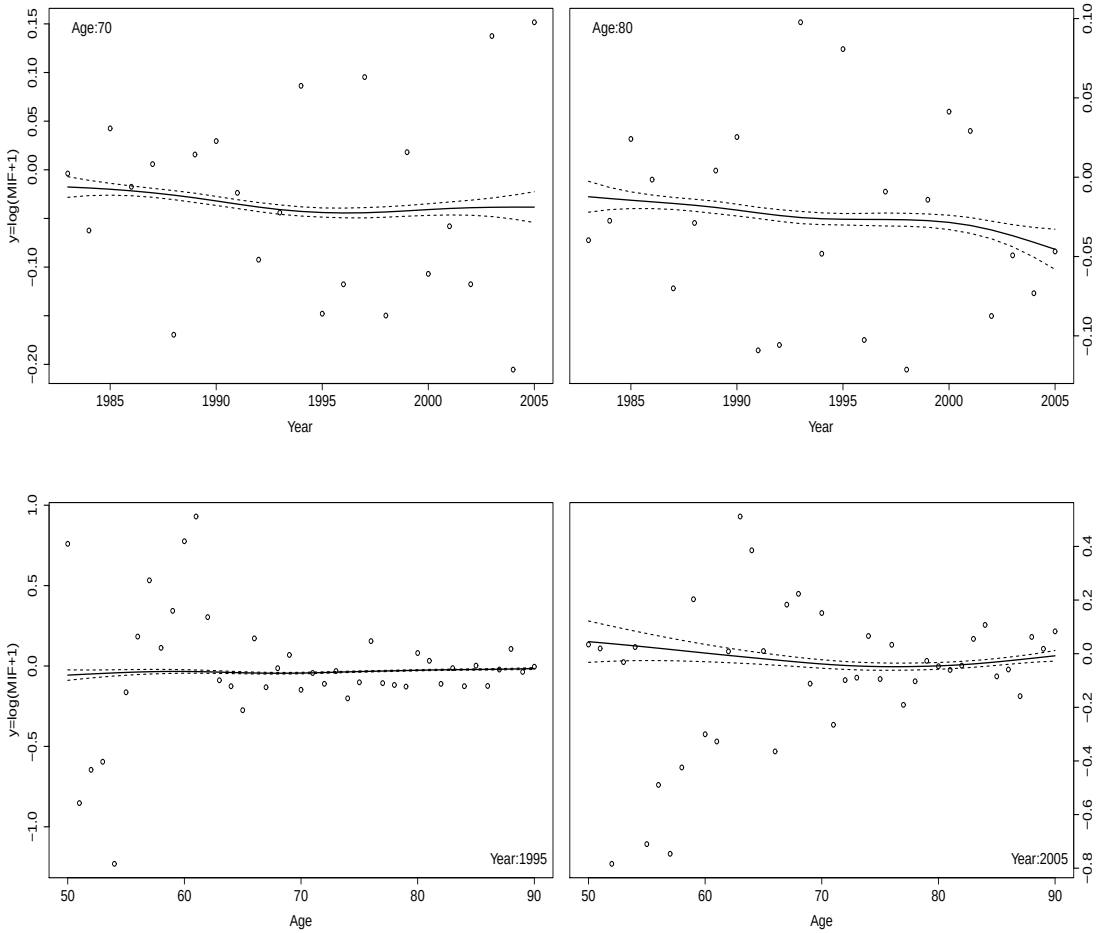

Figure 7.2: *Profile views of the* MIF *indicator,* $\boldsymbol{y}$*, for the CMI pensioner males smoothed under model (7.9).*

alone is not able to capture appropriately the observed effects. In the next Section, we incorporate both period/year effects and cohort effects into the model.

7.3 Period and cohort effects

The period and cohort effects encountered in the previous Section about the MIF can be tackled by adding a year/period dependent component and a cohort component to the predictor surface.

7.3.1 *Smooth-period* model

We first ignore the cohort effects and add the period effects alone to the model. That is, we substitute the smooth surface $\boldsymbol{\mathcal{B}\theta}$ with $\boldsymbol{\mathcal{B}\theta}+\boldsymbol{\mathcal{C}_1}$, where $\boldsymbol{\mathcal{C}_1}$ represents the period

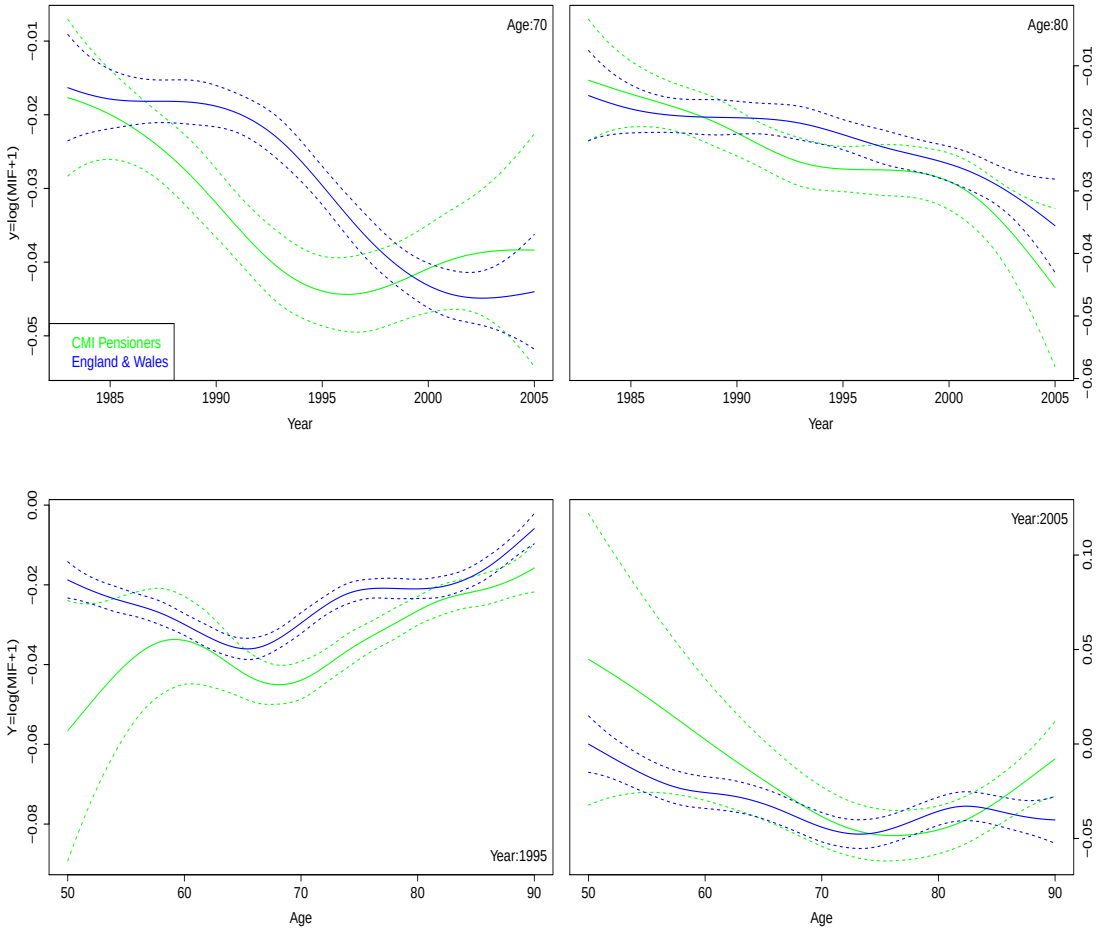

Figure 7.3: *Comparison of* MIF *indicators for CMI pensioner males and England & Wales males, under model (7.9).*

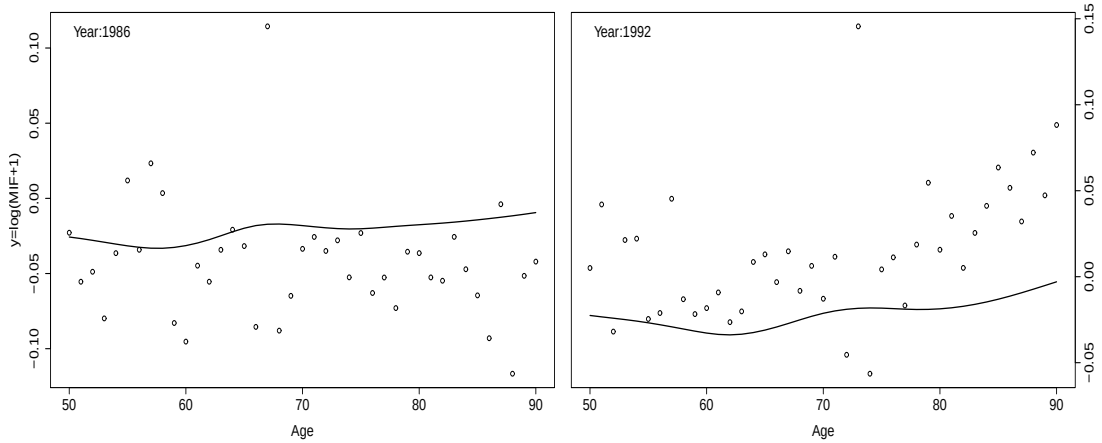

Figure 7.4: *Profile views from the estimated* MIF *indicator provided by model (7.9); England & Wales males.*

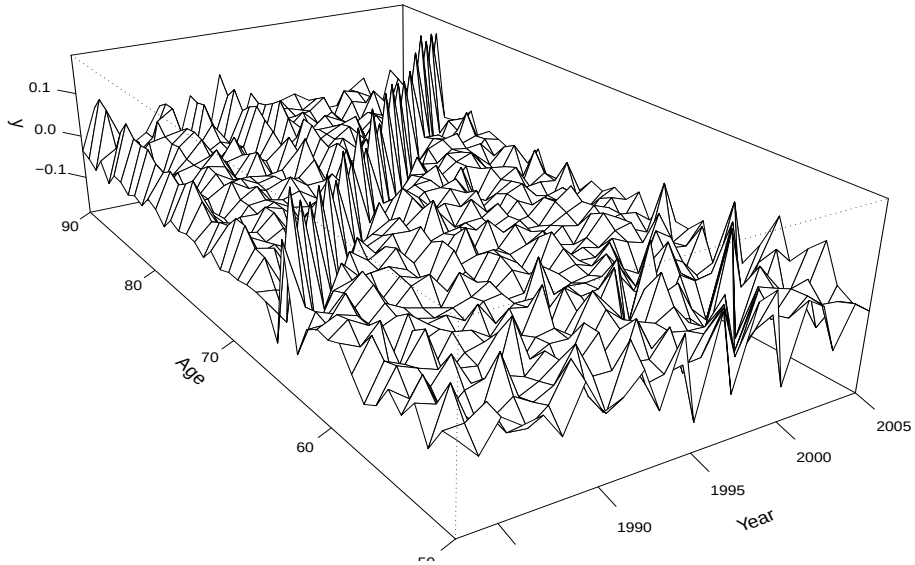

Figure 7.5: *Observed* MIF *indicator,* $\boldsymbol{y}$*, for males in England & Wales.*

dependent component. Hence, the initial model (7.9) becomes

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mathcal{B}\theta} + \boldsymbol{\mathcal{C}_1},\ \sigma^2\boldsymbol{\Sigma})\cdot \tag{7.14}$$

How should we model the effects $\boldsymbol{\mathcal{C}_1}$? A starting/simplistic point is to assume some parametric form for these effects in each year; this may produce satisfactory results for some particular data; but it will not be appropriate in a general setting because, for some data, the period effects can be very flexible across ages and very different as one moves from year to year. Kirkby and Currie (2010) referred to such effects as *period shocks* and they suggested modelling them in each year using B-splines. We follow their idea and so express $\boldsymbol{\mathcal{C}_1}$ as

$$\boldsymbol{\mathcal{C}_1} = \breve{\boldsymbol{\mathcal{B}}}\breve{\boldsymbol{\theta}}, \quad \text{with} \quad \breve{\boldsymbol{\mathcal{B}}} = \breve{\boldsymbol{B}}_x \otimes \boldsymbol{I}_{m_t} \ \text{ and } \ \breve{\boldsymbol{\theta}} = vec(\breve{\boldsymbol{\theta}}_1, ..., \breve{\boldsymbol{\theta}}_{m_t}); \tag{7.15}$$

where $\breve{\boldsymbol{B}}_x$ is a $n_x \times \breve{c}$ B-spline matrix in age, and $\breve{\boldsymbol{\theta}}_j$ is a $\breve{c}$-length vector representing the shock coefficients in year t_j. It is not difficult to see that the computational burden from the initial model (7.9) to the period shock model (7.15) increases substantially. Indeed, in addition to the $c_t \times c_x$ parameters involved in $\boldsymbol{\mathcal{B}}$, each year brings an additional set of $\breve{c}$ parameters into the model, which results in $m_t \times \breve{c}$ parameters brought

by the period components. It seems reasonable to require these shock components to be smooth in age separately for each year, and this can be obtained by taking a moderate number of splines for these components. For the identifiability of the model, we shrink the shock coefficients as in Kirkby and Currie (2010). Finally, we use the

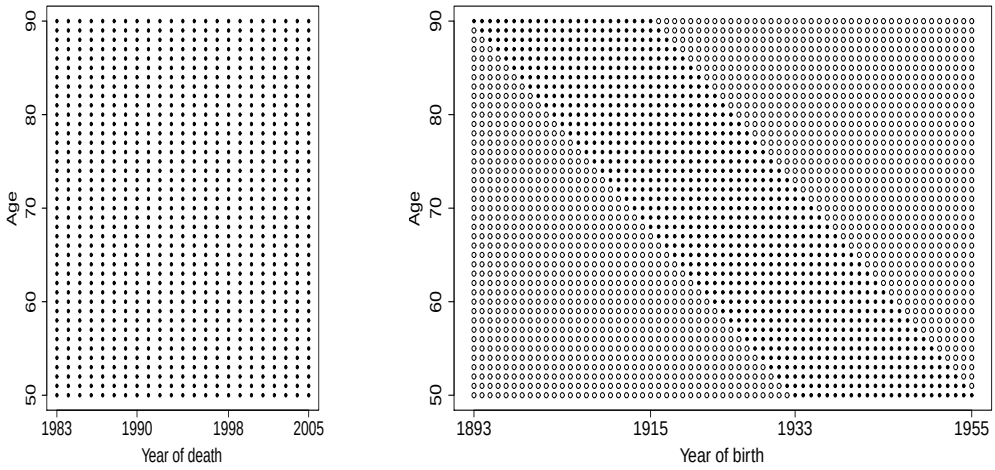

Figure 7.6: *Left: data (• • •) in the original structure, ie by age and year of death. Right: data (• • •) arranged by age at death and year of birth, together with dummy data (○ ○ ○).*

decomposition (7.13), and re-write model (7.14) as

$$\boldsymbol{y}|\tilde{\boldsymbol{u}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\alpha} + \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{u}},\ \sigma^2\boldsymbol{\Sigma}), \quad \tilde{\boldsymbol{u}} \sim \mathcal{N}(\boldsymbol{0},\ \boldsymbol{P}_{\tilde{u}}^{-1}), \tag{7.16}$$

with $\tilde{\boldsymbol{Z}} = [\boldsymbol{Z} : \breve{\boldsymbol{\mathcal{B}}}]$, $\tilde{\boldsymbol{u}} = vec(\boldsymbol{u}, \breve{\boldsymbol{\theta}})$, and $\boldsymbol{P}_{\tilde{\boldsymbol{u}}} = blockdiag(\boldsymbol{P}_{\boldsymbol{u}}, \breve{\lambda}\boldsymbol{I}_{\breve{c}\times m_t})$, where $\breve{\lambda}$ is the shrinkage parameter. This corresponds to a mixed model with fixed effect $\boldsymbol{\alpha}$ and "random" effect $\tilde{\boldsymbol{u}}$. Note that the period shock effects here are naturally interpreted as random effects, in the sense that they are part of the random component $\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{u}}$.

7.3.2 *Smooth-period-cohort* model

Let $\boldsymbol{\mathcal{C}_2}$ represents the cohort components. If we ignore the period component, then from the initial model (7.9), the adjusted model incorporating the cohort effects can

be expressed as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mathcal{B}\theta} + \boldsymbol{\mathcal{C}_2},\ \sigma^2\boldsymbol{\Sigma}), \tag{7.17}$$

with smoothness constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\mathcal{C}_2}$. One problem with this model in the present context is that the cohort effects run along diagonals and not down columns as the period effects. This makes the GLAM representation (which depends on the row and column structure of both data and model) of model (7.17) laborious; extra work needs to be done to put the data in the appropriate structure, ie the matrix structure that enables the GLAM representation. This re-structuring (as discussed by Kirkby 2009) together with the fitting can be summarized as follows. First, we represent the data by age at death and year of birth (instead of age at death and year of death, as in the original data); as a result, the data move from the rectangular structure to the parallelogram structure (see Figure 7.6). Second, we fill the two empty triangles in the two opposite corners of this parallelogram with dummy data; as a result, the data becomes a matrix again, but larger than the initial matrix (see Figure 7.6). With this representation, the cohort effects are then tackled identically to the period shock effects described in Section 7.3.1, with the additional aspect that the dummy data must be weighted out throughout the estimation process so that they do not affect the estimates. Finally, we return the data and the estimates to the original matrix structure (ie, to the the age at death and year of death structure).

We now extend the initial model (7.9) to incorporate both the period and cohort effects. The extended model can be represented as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mathcal{B}\theta} + \boldsymbol{\mathcal{C}_1} + \boldsymbol{\mathcal{C}_2},\ \sigma^2\boldsymbol{\Sigma}), \tag{7.18}$$

with smoothness constraints on $\boldsymbol{\theta}$, $\boldsymbol{\mathcal{C}_1}$ and $\boldsymbol{\mathcal{C}_2}$. We refer to (7.18) as the *smooth-period-cohort* model. We would ideally use the GLAM representation of this extended model. Unfortunately, it is not possible to write simultaneously the period and the cohort sub-regression matrices of this model using the Kronecker product. Furthermore, the structure of the joint penalty matrix on the period and the cohort effects is tricky. All these points make the simultaneous estimation of $\boldsymbol{\theta}$, $\boldsymbol{\mathcal{C}_1}$ and $\boldsymbol{\mathcal{C}_2}$ problematic. We get around these difficulties by adopting a profile-like methodology, as follows:

(1) Fit the period shock model (7.14) to the data $\boldsymbol{y}$; this gives the estimates $\hat{\boldsymbol{\theta}}$

and $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1}}$.

(2) Fit a cohort effects model to the data $\boldsymbol{y}$ with $\boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}}+\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1}}$ set as an offset; this gives the estimate $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2}}$.

(3) Fit the period shock model to the data $\boldsymbol{y}$ with $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2}}$ set as an offset; this gives the updated estimates for $\boldsymbol{\mathcal{B}\theta}$ and $\boldsymbol{\mathcal{C}}_{\mathbf{1}}$.

(4) Alternate between (2) and (3) until convergence.

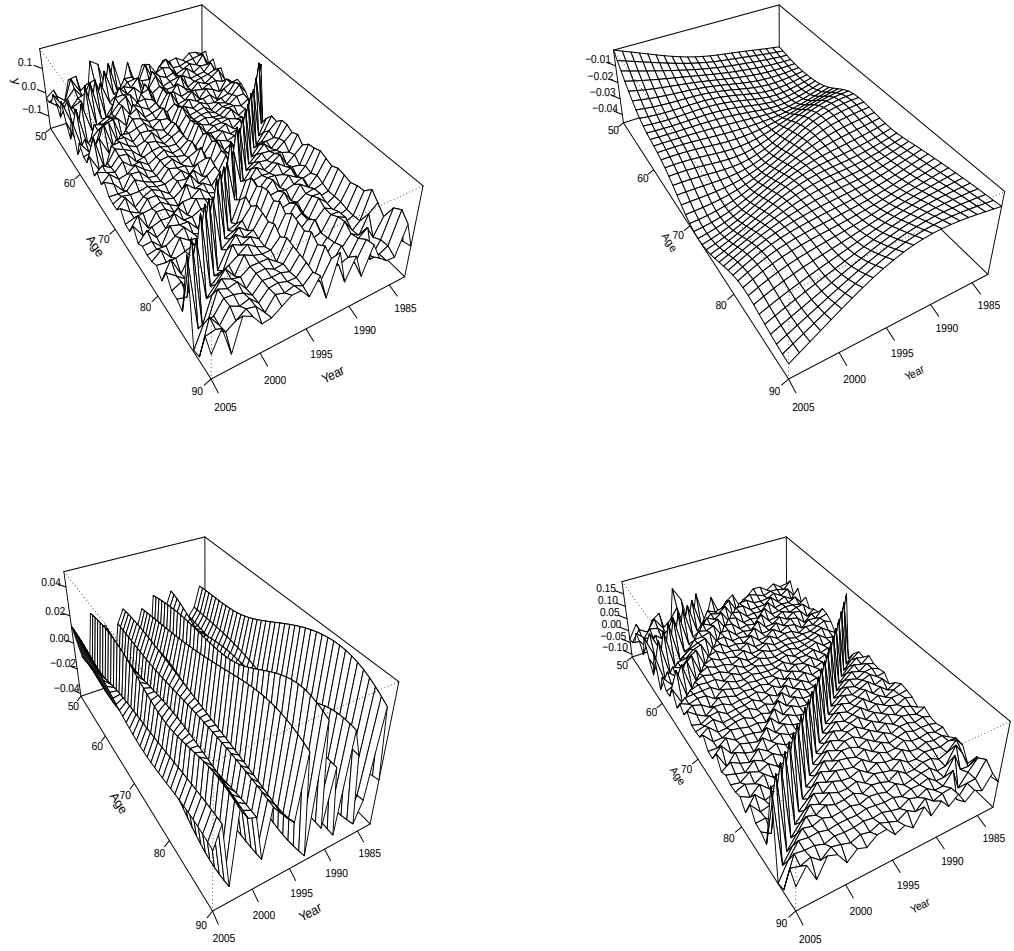

Figure 7.7: *Fitted* MIF *indicator and its components for the smooth-period-cohort model (7.18) applied to England & Wales males. Upper left: fitted* MIF *indicator,* $\hat{\boldsymbol{y}} = \boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\mathcal{C}}}_1 + \hat{\boldsymbol{\mathcal{C}}}_2$. *Upper right: smooth surface,* $\boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}}$. *Lower left: period component,* $\hat{\boldsymbol{\mathcal{C}}}_1$. *Lower right: cohort component,* $\hat{\boldsymbol{\mathcal{C}}}_2$.

The upper left panel in Figure 7.7 displays the predictor of the MIF indicator, which is the sum of its three components: the underlying two-dimensional smooth surface $\boldsymbol{\mathcal{B}\theta}$ in the upper right panel, the period effects $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1}}$ in the lower left panel, and the cohort effects $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2}}$ in the lower right panel. These three components are also illustrated by some profile views in Figure 7.8. From these graphics, we see that the underlying smooth surface alone (ie model (7.9)) is not able to provide a good summary for these data (blue lines). By adding the period component (ie model (7.14)), we obtain a substantial improvement (green lines), but this cannot capture the cohort effect. The *smooth-period-cohort* model (7.18) captures both the period and cohort effects as shown by the red lines on these graphics. However, with this *smooth-period-cohort* model fitted as described above, the standard error of the predictor estimate $\hat{\boldsymbol{y}} = \boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1}} + \hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2}}$ becomes tricky since our alternating method does not provide an explicit formula for $\hat{\boldsymbol{y}}$. One possibility is to use the bootstrap method to calculate standard errors.

7.4 Bootstrap standard errors for the *smooth-period-cohort* model

With the advance of computers, the bootstrap approach has become a very useful tool in statistics, specifically for its simplicity. We distinguish the *non-parametric bootstrap* in which we generate bootstrap samples with replacement from the initial data, and the *parametric bootstrap* which is based on generating bootstrap samples from some parametric distribution, with estimates of the parameters from the data plugged in (Efron and Tibshirani, 1993, chap 2,3). In the setting of mixed models, the bootstrap approach is not without controversy. On the one hand, the parametric bootstrap relies heavily on the normal assumptions like (7.13); hence the bootstrap estimates may turn out to be inconsistent if these Gaussian assumptions are suspect. On the other hand, the non-parametric bootstrap relies on samples drawn from the initial data; this usually assumes that the data are independent, which is not true in general, specifically in the mixed model framework where correlations and grouping structures are present (Morris, 2002). Since it is standard to assume that the deaths D_{ij} are independent, we proceed as follows. Firstly, from the original death counts $\boldsymbol{D}$ and

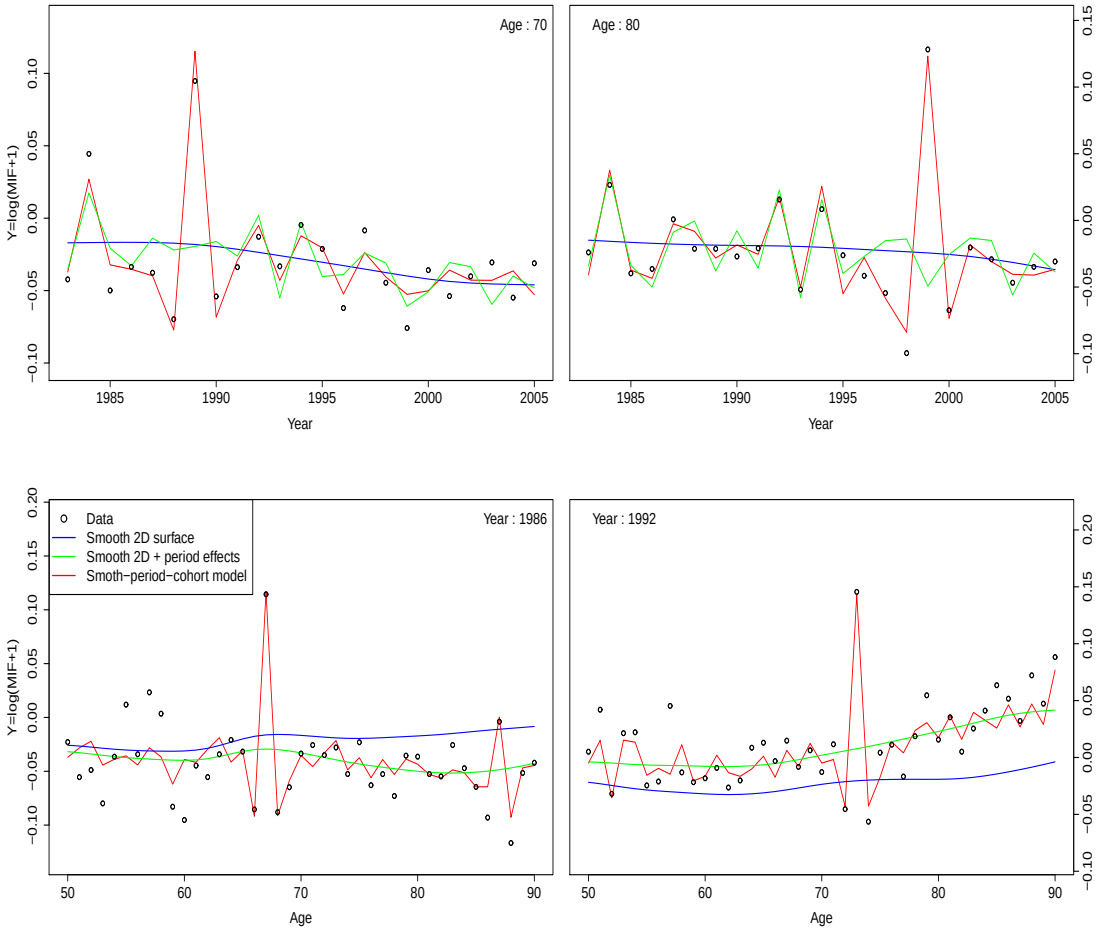

Figure 7.8: *Profile views from fitting the smooth-period-cohort model to England & Wales males. The blue lines represent the underlying two-dimensional smooth surface. The green lines illustrate the underlying two-dimensional surface + period effects, and as we can see from these green lines on the lower panels, the period shocks are clearly captured. The red lines correspond to the global predictor from the smooth-period-cohort model, and we can see from these lines how the cohort effects are identified by this extended model.*

exposures $\boldsymbol{E}$, we estimate the smooth forces of mortality $\hat{\tau}_{ij}$ and the dispersion index parameters $\hat{a}_{\varphi(i,j)}$ as described in Section 5.5, and compute the estimates $\hat{\mu}_{ij} = E_{ij}\hat{\tau}_{ij}$. Secondly, we simulate r parametric bootstrap samples, $\boldsymbol{D}_1^*, ..., \boldsymbol{D}_r^*$, of death counts based on the negative binomial model (7.5) with the estimates $\hat{\mu}_{ij}$ and $\hat{a}_{\varphi(i,j)}$ plugged in. Thirdly, we derive the bootstrap samples of the MIF indicator, $\boldsymbol{y}_1^*, ..., \boldsymbol{y}_r^*$, using (7.1). Finally, we fit the *smooth-period-cohort* effects model (7.18) to each of the bootstrap samples $\boldsymbol{y}_s^*$, which gives: - the smooth two-dimensional coefficient estimate $\hat{\boldsymbol{\theta}}_s^*$ - the period shock estimate $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1},s}^*$ - the cohort estimate $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2},s}^*$ - the predictor estimates $\hat{\boldsymbol{y}}_s^*$, where $\hat{\boldsymbol{y}}_s^* = \boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}}_s^* + \hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1},s}^* + \hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2},s}^*$, $s = 1, ..., r$. The bootstrap standard error on the

predictor $\hat{\boldsymbol{y}}$ is then obtained from the sample variance of $\{\hat{\boldsymbol{y}}_s^*,\ s = 1, ..., r\}$. Similarly, the bootstrap confidence intervals for the three components $\boldsymbol{\mathcal{B}}\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\mathcal{C}}}_{\mathbf{2}}$ can be computed. The result for England & Wales males is illustrated by the profile views in Figure 7.9.

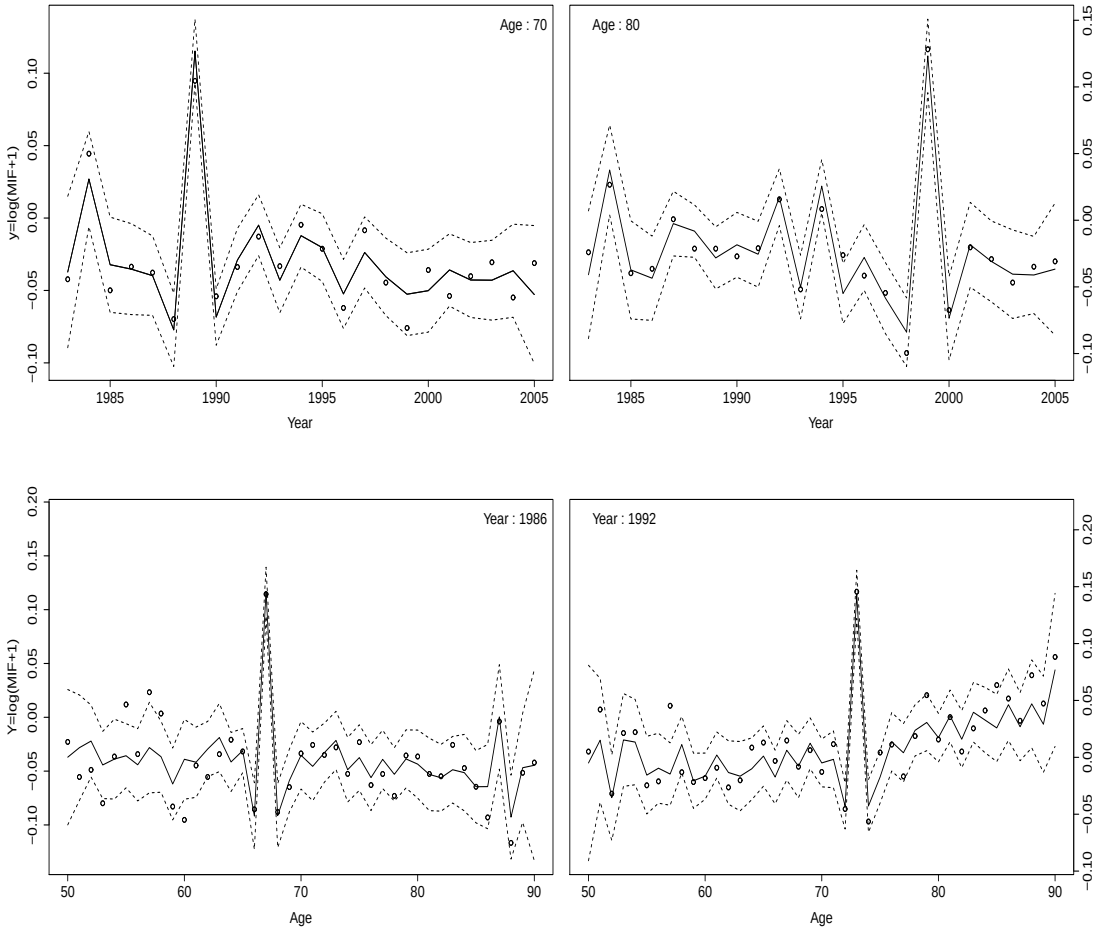

Figure 7.9: *Profile views of the fitted* MIF *indicator together with the bootstrap confidence intervals provided by the smooth-period-cohort model. England & Wales males.*

One advantage of using the parametric bootstrap approach here is that it leads to bootstrap samples in the same matrix format, which allows the GLAM representation to be used. Alternatively, the non-parametric bootstrap can also be implemented since the death counts D_{ij} are assumed to be independent. However, this alternative will lead to scattered data and so the matrix structure of the generated data will not be

preserved. In this case, we replace the Kronecker product basis with the row tensor product basis described in Eilers et al. (2006).

7.5 Conclusion

The central point of this Chapter was the modelling of the mortality improvement factor (MIF). We first derived the covariance structure of the MIF data and then formalized a basic model for the smoothing of these data. This basic model was sufficient to summarize the MIF for CMI pensioner males. For the population data however, the period and cohort effects needed to be taken into account; this has resulted in the *smooth-period-cohort* model, an extension of the period shock model in Kirkby and Currie (2010).

Chapter 8

Summary and future agenda

In this final Chapter, we summarize the main points of the thesis and discuss some topics needing further investigation.

8.1 Summary

Nonparametric or smoothing methods are increasingly becoming a standard tool in modern Statistics. In this dissertation, we investigated smooth models in hierarchical and multidimensional settings, with application to longitudinal and mortality data respectively.

We started in Chapter 2 by reviewing some well-known full rank smoothing methods before focusing on penalized splines based on two popular bases, namely truncated polynomials (PT-splines) as suggested by Ruppert et al. (2003), and B-splines (PB-splines) in the sense of Eilers and Marx (1996). In terms of results, there is little to choose between these two approaches, but technically, the PB-splines method offers several benefits as discussed in Section 2.3.3.

The main feature of the PT-splines approach resides in the simplicity of its relation to polynomial regression. Because of this simplicity, truncated lines are widely used to model the population/group and subjects effects in longitudinal data (Coull et al., 2001a; Ruppert et al., 2003, sect 9.3; Durban et al., 2005; etc). Here, two issues must to be addressed: (a) the smoothness of the population and subject effects and (b) their identifiability. A standard way of circumventing these two points is through the set of normal distributions given by (3.4). Using some balanced `CanadianWeather` data, we

have shown that this approach leads to inconsistent fitted effects, and to confidence intervals with undesirable properties, as illustrated in Figure 3.3. The reason for these problems is the misspecification on the standard covariance structure (3.4).

In Section 3.3 we proposed a solution to these problems via penalty arguments on truncated polynomials and/or B-splines bases. Essentially, we addressed the smoothness issue through ridge penalties on the truncated line coefficients or difference penalties on the B-spline coefficients, and we tackled the identifiability problem via an appropriate amount of shrinkage either on the subject B-spline coefficients, or on the fitted subject effects. This shrinkage allowed us to give a random interpretation to the subjects effects in the original sense of a mixed model. The results from this method were shown in Figure 3.4, and we referred to this new approach as the penalty approach. Although this penalty approach is constructed using penalty arguments, we showed in Section 3.5 how it can be re-formulated and interpreted as a mixed model.

We started Chapter 4 by investigating the standard approach in more detail, first on grouped balanced data, then unbalanced data, and finally through a simulation study. From this investigation, we concluded that the inconsistencies in this approach increase with the flexibility of the subject effects. We then extended the penalty approach to this grouped unbalanced setting in Section 4.2, described its implementation from the mixed model perspective in Section 4.3, illustrated its consistency in Section 4.5, and generalized it to the multivariate setting in Section 4.6. This closed the first part of the thesis.

In the second part, we moved to two-dimensional smoothing with reference to mortality data. In the first Sections of Chapter 5, we outlined the formulation of two-dimensional PB-splines, discussed the computational demands via the generalized linear array representation of Currie et al. (2006), looked at the extrapolation issue, and illustrated the method on mortality data under the Poisson assumption. In Section 5.2, we investigated the adverse impact of over-dispersion (and heterogeneity) as shown in Figure 5.6, and we then developed a class of models for dispersed count data in Section 5.3, through a joint smooth modelling of the mean and dispersion effects, using quasi-likelihoods. The effectiveness of this approach has been illustrated on real data and through simulations in Section 5.4.

We devoted Chapter 6 to the construction of a class of joint models for "similar"

mortality tables. Essentially, we proposed to model these tables in terms of a common surface and then portray their relative differences by appropriate simple gaps. We introduced this approach for two populations in Section 6.1, extended it to the general setting in Section 6.2, discussed the implementation in Section 6.3, and applied it on several data sets in Section 6.4. The key attractive feature of this joint approach is that it facilitates the classification of these tables and allows a simple comparison through the gap components.

The goal of Chapter 7 was the smoothing of two-dimensional correlated data, especially mortality improvement factors. First, we derived an appropriate covariance structure in this setting, and set up a basic model. Due to the presence of correlation, we re-parametrised the basic model as a mixed model and described its implementation. However, for certain mortality data, the presence of period and cohort effects made this basic model questionable. We then extended it in Section 7.3 to the *smooth-period-cohort* model, an additive model with three components which enabled us to capture appropriately the underlying smooth surface, the period effects and the cohort effects.

8.2 Future agenda

Some ideas in this dissertation can be improved.

First, the penalty approach in Chapters 3 and 4 has been presented for normal response data. This approach can be extended to the exponential family. In this case, deriving a closed form of the restricted likelihood criterion becomes tricky since the integral in (4.19) turns out to be intractable. However, a good approximation of this integral can be obtained via the Laplace approximation; details on this approximation can be found in Breslow and Clayton (1993) and Bates (2011), in which case, the computational scheme presented in Section 4.4 needs adjusted.

Second, we have assumed isotropic smoothing for all subjects. On the one hand, the individual subject effects can display different amounts of flexibility and so, working under isotropic smoothing may sound optimistic. On the other hand, optimizing deviance based criteria such as AIC, BIC, or restricted likelihood under individual smoothing parameters is computationally demanding. One possibility is to embark on the full Bayesian framework as outlined in Section 2.7. We have implemented

Bayesian fitting on some small data sets using `R2WinBUGS`, an `R` package that calls `WinBUGS` from `R` (Sturtz et al., 2005), but the fitting process was very slow. More work needs to be done here, especially for large data sets such as `CanadianWeather` and `ChildHeight`.

Third, in Section 4.6, we sketched the generalisation of subject-specific curves to the multivariate setting. However, we gave no practical illustration. Although the implementation can borrow substantially from the material in Sections 4.3 and 4.4, a considerable amount of work needs to be done on the efficient implementation; some elements in the work by Coull et al. (2001a) would be helpful here.

For the sake of clarity, we have restricted the attention to the situation where subjects are nested within population or groups. The idea can be relaxed to more than two levels. A straightforward example is: subject within groups, and groups within population. In this case, we will need an additional constraint in order to separate the group effects from the population effect, and this constraint will depend on whether the groups are viewed as fixed or random. In the former view, we can require that the vectors of spline coefficients at the group level (or the fitted group effects) sum up to zero, and in the latter, we will place an appropriate shrinkage at the group level.

In Chapter 6, we approached the analysis of multiple mortality tables by fitting nested models. This allowed us to compare such models by residual and graphical methods. Hypothesis testing is a more rigorous approach to such comparisons and our models give a platform for the development of these testing procedures. One problem that will need to be addressed is the very large power that our extensive datasets would give to any such test.

Finally, in the modelling of mortality improvement factors in Chapter 7, where we were interested in extracting the underlying smooth surface, together with the period and cohort effects, we have been confronted by the difficulty of a GLAM representation as well as the structure of the global penalty matrix on these components (and the complexity of these problems increase if we try to incorporate age effects). Although we proposed a profile-type approach that worked reasonably well on all our data sets, we think that the work in this Chapter deserves additional attention.

Appendix A

Notation and abbreviation

A.1 Notation

$/$	element-wise division
$*$	element-wise multiplication
$\otimes$	Kronecker product
$\diamond$	row tensor product
$\sim$	distributed as
i.i.d.	independent and identically distributed as
$\approx$	approximately equal to
tr	trace
$\mathbb{R}$	set of all real numbers
λ	smoothing parameter
$\breve{\lambda}_1$	smoothing parameter at the subject level
$\breve{\lambda}_2$	shrinkage or identifiability parameter
$\boldsymbol{\lambda}$	vector containing all smoothing, identifiability and variance parameters
ν	effective dimension
$\mathcal{S}(\cdot)$	smooth function
$\mathcal{S}_k(\cdot)$	smooth function characterising the k th group mean
$\breve{\mathcal{S}}_i(\cdot)$	smooth function characterising the i th subject effect
$\mathbf{0}_{s\times s}$	$s \times s$ matrix of zeros
$\boldsymbol{I}_s$	$s \times s$ identity matrix
$\mathbf{1}_n$	n-length vector of ones

$\boldsymbol{H}$	hat matrix
$\boldsymbol{P}$	penalty matrix
$\boldsymbol{B}$	B-spline matrix
$\boldsymbol{T}_p$	truncated polynomial matrix of degree p
$\boldsymbol{\Omega}$	regression matrix of B-splines or truncated polynomials
$\breve{\boldsymbol{\Omega}}_i$	regression matrix of B-splines or truncated polynomials for subject i
$\mathcal{P}(\cdot)$	penalty function
$\breve{\boldsymbol{P}}_i$	penalty matrix for the i th subject effect
$\boldsymbol{\Delta}$	difference matrix
$\boldsymbol{Y}_{i\bullet}$	i th row of $\boldsymbol{Y}$
$\boldsymbol{Y}_{\bullet j}$	j th column of $\boldsymbol{Y}$
$\boldsymbol{A}'$	transpose of $\boldsymbol{A}$
$diag(\boldsymbol{A})$	diagonal elements of $\boldsymbol{A}$
$vec(\boldsymbol{A})$	vector obtained by stacking the columns of $\boldsymbol{A}$ on top of each other
$vec(\boldsymbol{A}_1, ..., \boldsymbol{A}_n)$	$= (vec(\boldsymbol{A}_1)', ..., vec(\boldsymbol{A}_n)')'$
$blockdiag(\boldsymbol{A}_1, .., \boldsymbol{A}_n)$	block diagonal matrix with the $\boldsymbol{A}_i$ in the block diagonal positions
$\boldsymbol{A}_1 \equiv \boldsymbol{A}_2$	$\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ have different dimensions but they contain the same elements
$\mathbb{E}[\boldsymbol{\alpha}]$	expect value of $\boldsymbol{\alpha}$
$cov(\boldsymbol{\alpha})$	covariance matrix of $\boldsymbol{\alpha}$
$\mathcal{U}(a_1, a_2)$	uniform distribution on $[a_1, a_2]$
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathcal{P}oiss(\mu)$	Poisson distribution with mean μ
$Gamma(a_1, a_2)$	gamma distribution with mean $a_1 a_2$ and variance $a_1 a_2^2$
$\mathcal{NB}(a_1, a_2)$	negative binomial distribution with mean a_1 and variance $a_1 + a_1^2/a_2$
PT-splines	penalized splines via truncated polynomials
PB-splines	penalized splines via B-splines

A.2 Abbreviation

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BICS	Scaled Bayesian Information Criterion
CV	Cross Validation
GCV	Generalized Cross Validation
BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Predictor
REML	REstricted Maximum Likelihood
GLM	Generalized Linear Model
PGLM	Penalized Generalized Linear Model
GLAM	Generalized Linear Array Model
MCMC	Markov Chain Monte Carlo
MIF	Mortality Improvement Factor
MSE	Mean Square Error
MLE	Maximum Likelihood Estimation
RSS	Residual Sum of Squares
PRSS	Penalized Residual Sum of Squares
SD	Standard Deviation
ANOVA	ANalysis Of VAriance
CMI	Continuous Mortality Investigation
ONS	Office of National Statistics in the UK

References

A. Agresti. *Categorical data analysis*. Wiley, 1990.

H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

K. Andreev and J. Vaupel. Patterns of mortality improvement over age and time in developed countries: Estimation, presentation and implications for mortality forecasting. Technical report, Department of Community Health and Epidemiology, Queen's University, Kingston, Canada; and Max Planck Institute for Demographic Research, Rostock, Germany, 2005.

V. Baladandayuthapani, B. K. Mallick, and R. J. Carroll. Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computaional and Graphical Statistics*, 14:378–394, 2005.

D. Bates. Computational methods for mixed models. 2011. URL `http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf`.

V. Biatat and I.D. Currie. Joint models for classification and comparison of mortality in different countries. *Proceedings of 25rd International Workshop on Statistical Modelling, Glasgow*, pages 89–94, 2010.

H. Booth, R. J. Hyndman, L. Tickle, and P. de Jong. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, 15: 289–310, 2006.

A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-plus illustrations*. Oxford University Press, 1997.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

N. Brouhns, M. Denuit, and J. K. Vermunt. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31:373–393, 2002.

B. A. Brumback and J. A. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93: 961–976, 1998.

A. J. G. Cairns, D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, A. Ong, and I. Balevich. A quantitative comparison of stochastic mortality models using data from England & Wales and the United States. *North American Actuarial Journal*, 13:1–35, 2009.

H. Cardot. Spatially adaptive splines for statistical linear inverse problems. *Journal of Multivariate Analysis*, 81:100–119, 2002.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

B. A. Coull, D. Ruppert, and M. P. Wand. Simple incorporation of interactions into additive models. *Biometrics*, 57:539–545, 2001a.

B. A. Coull, J. Schwartz, and M. P. Wand. Respiratory health and air polution: additive mixed model analyses. *Biostatistics*, 2:337–349, 2001b.

C. M. Crainiceanu, D. Ruppert, and M. P. Wand. Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14:1–24, 2005.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

I. D. Currie. Adjusting for bias in mortality forecasts. 2009. URL `http://www.ma.hw.ac.uk/~iain/research/talks/Currie_Longevity_4up.pdf`.

I. D. Currie, M. Durban, and P. H. C. Eilers. Smoothing and forecasting mortality rates. *Statistical Modelling*, 4:279–298, 2004.

I. D. Currie, M. Durban, and P. H. C. Eilers. Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society (Series B)*, 68:259–280, 2006.

C. de Boor. *A practical guide to splines*. Springer, 1978.

P. Dierckx. *Curve and surface fitting with splines*. Oxford Science Publications, 1996.

V.A.B. Djeundje and I. D. Currie. Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, 4:1202–1224, 2010a.

V.A.B. Djeundje and I.D. Currie. Smoothing dispersed counts with applications to mortality data. *Annals of Actuarial Science*, 5:33–52, 2010b.

V.A.B. Djeundje and I.D. Currie. Smooth mixed models for nested curves. *Proceedings of 26th International Workshop on Statistical Modelling, Valencia.*, 2011a.

V.A.B. Djeundje and I.D. Currie. Fitting subject-specific curves to grouped longitudinal data. *Proceedings of 58th World Statistics Congress, Dublin.*, 2011b.

A. J. Dobson. *An introduction to statistical modelling*. Chapman and Hall, 1983.

M. Durban, J. Harezlack, M. P. Wand, and R. J. Carroll. Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24:1153–1167, 2005.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, 1993.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Sciences*, 11:89–121, 1996.

P. H. C. Eilers and B. D. Marx. Splines, knots, and penalties. *Computational Statistics*, 2:637–653, 2010.

P. H. C. Eilers, I. D. Currie, and M. Durban. Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, 50:61–76, 2006.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21:196–216, 1993.

J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20:2008–2036, 1992.

D. O. Forfar, J. J. McCutcheon, and A. D. Wilkie. On graduation by mathematical formula. *Journal of the Institute of Actuaries*, 115:1–149, 1988.

J. H. Friedman and B. W. Silverman. Flexible parsimonious smoothing and additive modelling. *Technometrics*, 31:3–21, 1989.

P. J. Green. Discussion on 'The analysis of designed experiments and longitudinal data by using smoothing splines'(by A. P. Verbyla, B. R. Cullis, M. G. Kenward and S. J. Whelam). 48:304–305, 1999.

P.J. Green and B.W. Silverman. *Nonparametric regression and generalized linear models*. Chapman and Hall, 1995.

S. Greven. *Non-standard problems in inference for additive and linear mixed models*. PhD thesis, Fakultät für Mathematik, Informatik and Statistik der Ludwig-Maximilians-Universität, Mücheun, 2007.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman and Hall, 1999.

N. Heckman, R. Lockhart, and J. D. Nielsen. Penalized regression, mixed effects models and appropriate modelling. *Unpublished report*.

C. Heuer. Modelling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, 53:161–177, 1997.

J. Hinde and C. G. B. Demetrio. Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, 27:151–170, 1998.

N. L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899, 2003.

L. Keele. *Semiparametric regression for the Social Sciences*. John Wiley and Sons, 2008.

J. Kirkby. *Array methods is statistics with applications to the modelling and forecasting of mortality*. PhD thesis, Heriot-Watt University, Edinburgh, 2009.

J. G. Kirkby and I. D. Currie. Smooth models of mortality with period shocks. *Statistical Modelling*, 10:177–196, 2010.

S. Konish and G. Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

T. Krivobokova and G. Kauermann. A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102:1328–1337, 2007.

T. Krivobokova, C. M. Crainiceanu, and G. Kauermann. Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, 17:1–20, 2008.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

S. Lang and A. Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13:183–212, 2004.

J. F. Lawless. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15:209–225, 1987.

D. J. Lee and M. Durban. P-spline ANOVA-type interaction models for spacio-temporal smoothing. *Statistical Modelling*, 11:49–69, 2011.

R. D. Lee and L. R. Carter. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87:659–671, 1992.

J. S. H. Li, M. R. Hardy, and K. S. Tan. Uncertainty in mortality forecasting: an extension to the classic Lee-Carter approach. *Astin Bulletin*, 39:137–164, 2009.

X. Lin and D. Zhang. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society (Series B)*, 61:381–400, 1999.

Z. Luo and G. Wahba. Hybrid adaptive splines. *Journal of the American Statistical Association*, 92:107–116, 1995.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman and Hall, 1989.

J. S. Morris. The BLUEs are not best when it comes to bootstrapping. *Statistics and Probability Letters*, 56:425–430, 2002.

J. A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74:221–232, 1987.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 56:545–554, 1971.

Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford Science Publications, 2004.

J. C. Pinheiro and D. M. Bates. *Mixed-effects models in S and S-PLUS*. Springer, 2000.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`.

P. T. Reiss and R. T. Ogden. Smoothing parameter selection for a class of semi-parametric linear models. *Journal of the Royal Statistical Society (Series B)*, 71: 505–523, 2009.

A. E. Renshaw. Joint modelling for actuarial graduation and duplicate policies. *Journal of the Institute of Actuaries*, 119:69–85, 1992.

A. E. Renshaw and P. Hatzopoulos. On the graduation of amounts. *British Actuarial Journal*, 2:185–205, 1996.

Richards and I. D. Currie. Longevity risk and annuity pricing with the lee-carter model (to appear). *British Actuarial Journal*, 2009.

S. J. Richards. Applying survival models to pensioner mortality data. *British Actuarial Journal*, 14:257–326, 2008.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757, 2002.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Cambridge University Press, 2003.

G. Schmidt, R. Mattern, and F. Schuler. Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under effects of impact. EEC Research Program on Biomechanics of Impacts, Final report, Phase III, Project G5, Institut für Rechtsmedizin, Universität Heidelberg, Heidelberg, Germany. 1981.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. Wiley, 2006.

S. G. Self and K. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610, 1897.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society (Series B)*, 47:1–52, 1985.

S. Sturtz, F. Statistik, U. Ligges, and A. Gelman. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, pages 1–16, 2005.

S. W. Thurston, M. P. Wand, and J. K. Wiencke. Negative binomial additive models. *Biometrics*, 56:139–144, 2000.

A. P. Verbyla, B. R. Cullis, M. G. Kenward, and S. J. Welham. The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society (Series C)*, 48:269–311, 1999.

G. Wahba. Bayesian confidence intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society (Series B)*, 45:133–150, 1983.

M. Wand and M. Jones. *Kernel smoothing*. Chapman and Hall, 1995.

Y. Wang. Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93:341–348, 1998.

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–447, 1974.

S. J. Welham and R. Thompson. A note on bimodality in the log-likelihood function for penalized spline mixed models. *Computational Statistics and Data Analysis*, 53: 920–931, 2009.

R. C. Willets. Mortality in the next millennium. *Staple Inn Actuarial Society*, 1999.

D. A. Williams. Extra binomial variation in logistic linear models. *Journal of the Royal Statistical Society (Series C)*, 31:144–148, 1982.

S. Wood, W. Jiang, and M. Tanner. Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, 89:513–528, 2002.

S. N. Wood. *Generalized additive models*. Chapman and Hall, 2006.

F. Yao and T. C. M. Lee. On knot placement for penalized spline regression. *Journal of the Korean Statistical Society*, 37:259–267, 2008.

J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998.