MASTER'S THESIS

Anna-Leena Orsama

Semiparametric models for the analysis of longitudinal weight measurement data

University of Tampere
School of Information Sciences
Statistics

May 2011

University of Tampere

School of Information Sciences

Orsama, Anna-Leena: Semiparametric models for the analysis of longitudinal weight measurement data

Master's thesis, 60 p., 3 appendix p.

Statistics

May 2011

Abstract

This thesis discusses semiparametric regression models that provide a flexible tool for modelling longitudinal data. Nonparametric models can be used for exploring data when parametric assumptions are too restricted to provide an adequate fit. In this thesis splines, piecewise defined polynomials, are discussed with emphasis on penalised splines. A penalised spline model can be connected to the widely known linear mixed-effects models that are a powerful tool for analysing clustered unbalanced data. Partitioning spline components into fixed effects and random effects, fusion between a parametric and a nonparametic model can be obtained. The advantage is that through the linear mixed-effect model presentation a nonparametric fit can be extended to account for the longitudinal nature of the data.

For model estimation and inference, likelihood based methods maximum likelihood and restricted maximum likelihood are discussed. Testing the significance of a spline model involves testing if a variance component of the corresponding LME model is zero. Testing is nonstandard since basic assumptions do not hold; in longitudinal data measurements are not independent or identically distributed and under the null hypothesis the test parameter is on the boundary of its parameter space.

The presented theory is applied to longitudinal weight measurements data to explore the weekly rhythm of weight. The rhythm is explored in the population level and in two subgroups; among subjects who have lost weight and among subjects who have gained weight. We find a rhythm that shows weight to be higher after weekends, on Sundays and on Mondays, and decrease during weekdays. Furthermore, in the whole population level and in the loss group, there seems to be a slight but significant increase at the end of the week, on Fridays and Saturdays. There is no difference in the shape of the estimated profile curves between the groups.

Keywords linear mixed-effects model, penalised spline, semiparametric regression, weekly rhythm

Contents

1 Introduction **7**
- 1.1 Background of the application . . . 8
- 1.2 Structure of the thesis . . . 9

2 Linear mixed-effects model **11**
- 2.1 Estimation . . . 13
  - 2.1.1 Restricted maximum likelihood . . . 15
- 2.2 Inference . . . 17
  - 2.2.1 Likelihood ratio test . . . 18
  - 2.2.2 Restricted likelihood ratio test . . . 18
  - 2.2.3 Use of simulations for the restricted likelihood ratio test 20

3 Scatterplot smoothing **21**
- 3.1 Regression spline . . . 21
- 3.2 Smoothing spline . . . 24
- 3.3 Penalised spline . . . 25

4 Penalised splines in the linear mixed-effects model framework **27**
- 4.1 Estimation of penalised splines . . . 28
- 4.2 Hypothesis testing . . . 31
- 4.3 Extension of penalised spline models . . . 32
  - 4.3.1 Interaction models . . . 32
  - 4.3.2 Subject-specific curves . . . 35

5 Application **37**
- 5.1 Materials and methods . . . 38
  - 5.1.1 Data description . . . 38
  - 5.1.2 Derivation of variables . . . 39
- 5.2 Analysis of the weekly rhythm of weight . . . 41
  - 5.2.1 Weekday effect . . . 42
  - 5.2.2 Parametric model versus nonparametric model . . . 42
  - 5.2.3 Individual effects . . . 44
  - 5.2.4 Group effect . . . 46

6 Results **47**
- 6.1 Weekday effect . . . 47
- 6.2 Parametric model versus nonparametric model . . . 47

6.3 Individual effects 49
6.4 Group effect 49

7 Discussion **51**
7.1 Finding the weekly rhythm 51
7.2 Longitudinal nature of the data 52
7.3 Group effect on the weekly rhythm of weight 54

8 Conclusions **55**

9 Acknowledgements **56**

Bibliography **57**

Appendix A: Model output **61**

Appendix B: R-code **63**

Abbreviations

BP	Best Predictor
BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Predictor
CSS	Cubic Smoothing Spline
DSS	Decision Support System
i.i.d.	independent and identically distributed
LME	Linear Mixed-Effects
LR	Likelihood Ratio
LRT	Likelihood Ratio Test
ML	Maximum Likelihood
PLS	Penalised Least Squares
REML	Restricted Maximum Likelihood
RLRT	Restricted Likelihood Ratio Test

1 Introduction

In longitudinal data a certain outcome variable of an individual is followed over a certain period of time. Measurements within subjects are repeated several times, producing data which consists of multiple time series. The grouped nature of the data violates the basic assumption of independence that leads to the requirement of special statistical methods to describe the data and to make valid inferences. In addition, the fact that data can be unbalanced and measurements unequally spaced sets certain requirements for the analysis method.

In longitudinal data analysis measurements of each subject are treated as a sequence $\mathbf{y_i} = (y_{i1},...,y_{im})$, where y_{im} denotes the mth measurement of the ith subject. The response variable $\mathbf{Y}$ comprises the complete set of all observations so that $\mathbf{Y} = (\mathbf{y_1}, .., \mathbf{y_n})$ where n is the number of subjects. Measurements within subjects are often correlated like in time series data but instead of studying the pattern of a single subject, the pattern of the overall study population is explored. Thus, the analysis combines elements of multivariate and time series methods.

Linear mixed-effects (LME) models introduced by Laird and Ware (1982) are widely known in the analysis of longitudinal data which inherits variability from different sources: within subjects and between subjects. LME models enable modelling of clustered data by adding random components to the model. The random components account for the homogeneity of measurements that are repeated within a subject. However, in repeated measurements data the response variable may depend on covariates in a more complicated way which is difficult to model parametrically. Nonparametric regression methods (Wahba 1990; Hastie & Tibshirani 1990) allow a more flexible approach where the relationship between the response variable $\mathbf{y}$ and the covariate $\mathbf{x}$ is modelled by an unspecified function $f(x)$ that is unrestricted to parametric assumptions. In this thesis splines, piecewise defined functions, are discussed as a nonparametric smoothing method with emphasis on a penalised spline model, (Eilers & Marx 1996).

Recent advances exploit the connection between penalised spline smoothing and the linear mixed-effects model. Through this fusion $f(x)$ can be estimated in the LME framework by representing it as $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zb}$ where fixed component $\mathbf{X}\boldsymbol{\beta}$ handles covariates that enter the model linearly while random effects component $\mathbf{Zb}$ and the corresponding covariance matrix handle the non-linear effects through spline bases functions. Thus, a penalised spline smoother can be seen as a best predictor (BP) (Speed 1991). This introduces semiparamet-

ric models (Ruppert, Wand & Carrol 2003) that include both parametric and nonparametric components making flexible modelling possible but at the same time preserving the effectiveness and interpretability of the parametric models. LME model framework has an advantage of allowing a straightforward extension to models with more a complicated random effects structure and with additional covariates.

In this thesis we discuss presentation of a penalised spline model as an LME model, its implemention with standard softwares and inference procedures for testing the significance of a spline. The inference for a penalised spline model is not straightforward. First, it involves testing to determine whether the variance component differs from zero. However, the test parameter under the null hypothesis is not an interior point of the parameter space that prevents the use of the asymptotic chi-square distribution. Secondly, in longitudinal data, measurements are not independent, which prevents the use of asymptotic chi-square mixture distribution presented by Self and Liang (1987). The behaviour of the likelihood based test statistic in nonparametric models is discussed by Crainiceanu, Ruppert, Claeskens & Wand (2001), Crainiceanu & Ruppert (2004) and Greven (2007), among others. The authors suggest the use of simulations to obtain the exact significance level for the observed test statistic.

1.1 Background of the application

Human body weight is one of the best-known indicators of health (Doll, Petersen & Stewart-Brown 2000, Mokdad et al. 2003). In several studies the impact of different treatments or interventions are investigated and weight change analysed over the study period. But little is known about weight variation inside of these periods, for example within a week. A week is a period of time that strongly affects our life (Zerubavel, 1985). This seven-day cycle has a strong impact on our activities. For example, it determines our sleeping patterns (Monk, Buysse, Rose, Hall & Kupfer 2010) and eating habits. Nutrition intake is reported to have a weekly rhythm (De Castro 1992) which shows that subjects have greater calorie intake on weekends. This thesis investigates whether the weekly rhythm can be found from longitudinal weight measurement data. The interest lies in profiling the rhythm on the whole population level as well as among subjects who lost weight and those who gained weight. We are also interested in the possible group differences in weight behaviour during this seven-day period.

Within-day-variation of several physiological variables has been investigated and diurnal rhythms are reported by several authors, e.g. Turjanmaa, Kalli, Majahalme, Saranummi $ Uusitalo (1987) and Kanabrocki et al. (1990). As far as is known, however, studies of weekly rhythm are rarely reported, specifically regarding weight variation. One of the few studies of the weekly rhythm of weight is reported by Tuomisto et al. (2006). The authors found differences in

weight between working days and weekends. Lappalainen, Pulkkinen, van Gils, Pärkkä & Korhonen (2005) previously made the same finding when analysing self-monitored weight data for a female subject over a period of ten years. Mattila, Lappalainen, Pärkkä, Salminen & Korhonen (2010) continued the work by analysing the weekly rhythm in subgroups; groups of subjects who lost weight versus those who did not lose weight were explored to see if the weekly rhythm is related to weight management. The authors found a significant correlation between the groups and weekdays. In the group of subjects who lost weight a similar rhythm was detected as in earlier studies, whereas the group of subjects who did not lose weight lacked a clear rhythm.

In this thesis, self-monitored weight measurements from 69 subjects are used to explore how weight varies within a week-length cycle. Using a weight change -based categorisation the relationship between the rhythm, weight loss and weight gain will be explored. Possible differences in the shape of curves between the groups will also be analysed. The analysis is conducted by fitting smooth curves through penalised splines. Modelling with contiguous accounts for the effect of previous days. We will analyse the data in the population level but the main interest lies in profiling the rhythm in subgroups and testing the difference between the groups.

The research for this thesis is funded by the Care4Me (Cooperative Advanced REsearch for Medical Efficiency) project as a part of the ITEA 2 programme of the VTT Technical Research Centre of Finland. The purpose of this thesis is to extract features from self-monitored weight measurement data for the development of a decicion support system (DSS). A DSS provides reminders that are based on patient's personal time series data. There are different kinds of contents from supportive feedback to notifications and alarms to raise one's knowledge and interest of the current health condition. We want to study the weekly rhythm of weight and use the results in the development of this future reminder system. Profiling patients according to their weekly rhythm of weight allows for more personalised feedback.

1.2 Structure of the thesis

In Chapter 2 we discuss linear mixed-effects (LME) models that are a central tool for analysing longitudinal data. Estimation and inference are represented and the challenge of testing the significance of random effects is discussed. Chapter 3 is concentrated on spline models to give readers an overview of smoothing methods. The emphasis of the chapter is on penalised splines (Section 3.3), which are later used in the application part. Chapter 4 discusses the central idea of the thesis; the presentation of a penalised spline model in a linear mixed-effects model framework. We look at the advantages of this fusion and the possibilities that LME framework provides for the inference of penalised splines and to the model extension. We discuss how to test the difference between two smooth curves. Section 5.2.3 shows how to build a flexible

random effects structure to model subject-specific curves. In Chapter 5, presented methods are applied to the longitudinal weight measurement data to analyse the weekly rhythm of weight. Results of the applications part are in Chapter 6 and discussed in Chapter 7. Finally, Chapter 8 concludes the methods and main results and gives a short overlook for future analysis of the weekly rhythm of weight.

2 Linear mixed-effects model

A linear mixed-effects (LME) model is an extension of a regression model that incorporates a random component in the model. LME models are a powerful tool for longitudinal data analysis, particularly through their ability to model the within-subject variation. In longitudinal data, measurements are repeated over a certain period and, thus, correlated within cluster. This violates the assumption of independence in classical regression models and can lead to misleading inference if the dependence is ignored (Fitzmaurice, Laird & Ware 2004). The underlying idea of LME model is that some of the regression parameters are fixed as in the original regression model while others are allowed to vary randomly from cluster to cluster to account for the heterogenity of measurements between the clusters. Random effects cover the variance-covariance structure of the response variable $\mathbf{y}$ whereas fixed effects model the mean of the data.

An introduction to the general design of linear mixed-effects models is provided by Robinsson (1991) and various models and methods for dealing with longitudinal data are given by Diggle, Liang & Zeger (1994). LME models have since been studied and applied by several authors. Pinheiro and Bates (2000) provide a clear theoretical overview to LME models but emphasise practical examples using S-Plus (included the S-code). Another good practical book is from Fitzmaurice et al. (2004). A more theoretical approach to the estimation and inference can be found in McCulloch, Searle & Neuhaus (2008).

Longitudinal data includes variability from following the sources that can be accounted for with the use of LME models (Fitzmaurice et al. 2004):

- between subject variation: difference between clusters. This is present in cross-sectional studies.
- within-subject variation: measurements within subject might be expected to be more similar than between subjects. Therefore, the presence of the dependence needs to be taken into account.
- residual error: unexplained variation in the response after modelling within subjects and between subjects variation. This is present in all statistical models.

LME models do not assume all subjects to be measured at simultaneous times nor that there is an equal number of observations available for subjects. This makes them perfectly suited for the analysis of longitudinal data where individuals may enter and withdraw from the study at any time (Fitzmaurice et

al. 2004). LME model techniques can also be incorporated into nonparametric regression models and they are closely related to penalised spline models shown in Chapter 4.

The formulation of a linear mixed-effects (LME) model, introduced by Harville (1976,1977) and Laird and Ware (1982), is

$$\begin{aligned} \mathbf{y_i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, i = 1, ..., n \\ \mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}), \epsilon_i \sim N(\mathbf{0}, \mathbf{R_i}), \end{aligned} \tag{2.1}$$

where $\mathbf{b_i}$ and $\boldsymbol{\epsilon_i}$, are assumed to be independent of one another

$$\begin{pmatrix} \mathbf{b_i} \\ \boldsymbol{\epsilon_i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D}, \mathbf{0} \\ \mathbf{0}, \mathbf{R_i} \end{pmatrix} \right). \tag{2.2}$$

Independence between different subjects y_i and y_j is also assumed, meaning that

$$\mathbf{Cov}(\mathbf{y_i}, \mathbf{y_j}) = 0. \tag{2.3}$$

In the model formulation (2.1) $\mathbf{y_i}$ denotes the response vector for ith subject, $\epsilon_\mathbf{i}$ denotes the corresponding measurement error for the subject. $\boldsymbol{\beta}$ is the p-dimensional parameter vector for fixed effects indicating the population parameters, and $\mathbf{b_i}$ is the q-dimensional vector for random effects (individual parameters). $\mathbf{X_i}$ and $\mathbf{Z_i}$ are associated model matrices for $\boldsymbol{\beta}$ and $\mathbf{b}$. The matrices $\mathbf{D}$ and $\mathbf{R_i}$ are the positive definite covariance matrices for the random effects and the error terms determining variance components. Through $\mathbf{D}$ the similarity of the measurements within subjects is handled. $\mathbf{R_i}$ is usually assumed to have a simple structure, identity matrix. But the assumption can be relaxed to model nonconstant group variances or serially correlated measurement errors (Pinheiro & Bates 2000).

The expected mean of $\mathbf{y_i}$, in the population level, is

$$E(\mathbf{y_i}) = \mathbf{X_i}\boldsymbol{\beta}, \tag{2.4}$$

where the effect of random effects $\mathbf{b_i}$ is not accounted for. Taking into account the individual effect, the expected value for $\mathbf{y_i}$ is the conditional mean of $\mathbf{y_i}$, given $\mathbf{b_i}$

$$E\left(\mathbf{y_i}|\mathbf{b_i}\right) = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i}. \tag{2.5}$$

Similarly the marginal and the conditional covariances can be distinguished. The population-level covariance of $\mathbf{y_i}$ is

$$\begin{aligned} Cov(\mathbf{y_i}) &= Cov(\mathbf{Z_i}\mathbf{b_i}) + Cov(\boldsymbol{\epsilon_i}) \\ &= \mathbf{Z_i}Cov(\mathbf{b_i})\mathbf{Z_i}' + Cov(\boldsymbol{\epsilon_i}) \\ &= \mathbf{Z_i}\mathbf{D}\mathbf{Z_i}' + \mathbf{R_i} \end{aligned} \tag{2.6}$$

and the conditional covariance of $\mathbf{y_i}$, given $\mathbf{b_i}$ is

$$Cov(\mathbf{y_i}|\mathbf{b_i}) = Cov(\boldsymbol{\epsilon_i}) = \mathbf{R_i}. \tag{2.7}$$

Equation (2.6) shows that LME models distinquish the variability from different sources; between-subject and within-subject variability. Let $\mathbf{\Sigma_i}$ denote a covariance matrix of the response variable $\mathbf{y_i}$,

$$\mathbf{\Sigma_i} = \mathbf{Z_i D Z_i}' + \mathbf{R_i}. \tag{2.8}$$

Based on the presented assumptions, $\mathbf{y}$ is distributed as

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma}).$$

The LME model for n subjects can be written in matrix notation as follows:

$$\begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_n} \end{pmatrix} = \begin{pmatrix} \mathbf{X_1} & \mathbf{Z_1} & \mathbf{0}\ldots & \mathbf{0} \\ \mathbf{X_2} & \mathbf{0} & \mathbf{Z_2}\ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ X_n & \mathbf{0} & \ldots & \mathbf{Z_n} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b_1} \\ \vdots \\ \mathbf{b_n} \end{pmatrix} \begin{pmatrix} \epsilon_\mathbf{1} \\ \vdots \\ \epsilon_\mathbf{n} \end{pmatrix}. \tag{2.9}$$

or more compactly

$$\begin{gathered} \mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \epsilon \\ \mathbf{b} \sim N(\mathbf{0}, \mathbf{D}), \epsilon \sim N(\mathbf{0}, \mathbf{R}). \end{gathered} \tag{2.10}$$

This is due to the independence of subjects $\mathbf{\Sigma} = diag[\mathbf{Cov}(\mathbf{y_1}), \mathbf{Cov}(\mathbf{y_2}), ..., \mathbf{Cov}(\mathbf{y_n})]$ where $\text{Cov}(\mathbf{y_i}) = \mathbf{Z_i D Z_i}' + \mathbf{R_i}$. Formulation of (2.9) shows that data may be unbalanced and subjects may have an unequal number of measurements but, nevertheless, can be modelled efficiently.

2.1 Estimation

A widely known method for parameter estimation is the maximum likelihood (ML) method originally introduced by R.A. Fisher (1922). The fundamental idea behind the ML estimation is to find estimates for unknown parameters, say $\boldsymbol{\theta}$, so that they are most probable for the observed data. ML estimates of $\boldsymbol{\theta}$ maximise the joint probability density function $f(y_i)$ over all observations in $\mathbf{y}$. The likelihood function for obtaining the estimates is

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i). \tag{2.11}$$

In this section we will introduce maximum likelihood (ML) and its modification, restricted maximum likelihood (REML), both methods for parameter estimation. ML and REML are methods used for establishing estimates for fixed

and random effects in the LME model. Casella and Berger (1991, Chapter 7) give an introduction to ML estimation method.

To obtain an ML estimation, an assumption of the underlying probability distribution of the data is required. This is often chosen to be a normal distribution. Normality is appropriate for different kinds of data and, therefore, the assumption is unlikely to be seriously wrong (Searle, Casella & McCulloch 1992). For a normally distributed response variable $\mathbf{y}=(y_1,..,y_n])$ where $\mathbf{y_i} \sim N(\mathbf{X_i}\boldsymbol{\beta}, \mathbf{\Sigma_i})$ and $(\mathbf{y_i} \perp \mathbf{y_j})$, the likelihood function $L(\boldsymbol{\beta}, \mathbf{\Sigma})$ is defined as a product of the probability density functions of $\mathbf{y_i}$'s as

$$Ł(\boldsymbol{\beta}, \mathbf{\Sigma}|\mathbf{y}) = \prod_{i=1}^{n}\left[\frac{1}{2\pi^{1/2}\Sigma_i^{1/2}} exp\left\{-\frac{1}{2}(y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta)\right\}\right].$$

Fitting an LME model involves estimation of $\boldsymbol{\beta}$ and parameters involved with matrices $\mathbf{D}$ and $\mathbf{R}$, that often are $\sigma_{\mathbf{b}}$ and $\sigma_{\boldsymbol{\epsilon}}$. For simplicity, let denote $\boldsymbol{\theta}$, a vector of variance components that defines $\mathbf{D}$ and $\mathbf{R}$. This leads to the notation of $\mathbf{\Sigma}(\boldsymbol{\theta})$.

Under the normality assumption, the joint density function of $\mathbf{y}$ and $\mathbf{b}$ is

$$f(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{b}, \beta)f(\mathbf{b}|\theta), \tag{2.12}$$

where

$$f(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}) = \frac{1}{2\pi^{n/2}|\mathbf{R}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})\right\}$$

and

$$f(\mathbf{b}|\boldsymbol{\theta}) = \frac{1}{2\pi^{q/2}|\mathbf{D}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{b}'\mathbf{D^{-1}}\mathbf{b})\right\}.$$

q denotes the dimension of $\mathbf{b}$. For known $\mathbf{D}$ and $\mathbf{R}$, estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are obtained maximizing the log-likelihood, which minimizes the twice negative logarithm of the joint density function (2.12) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (i.e. $\mathbf{D}$ and $\mathbf{R}$). In practise, components in $\mathbf{D}$ and $\mathbf{R}$ are estimated from the data using the REML method, introduced later. The joint log-likelihood function is now:

$$\begin{aligned} l(\boldsymbol{\beta}, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y}) = log\ L(\boldsymbol{\beta}, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y}) \\ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) + \mathbf{b}'\mathbf{D}^{-1}\mathbf{b} + log|\mathbf{D}| + log|\mathbf{R}|. \end{aligned} \tag{2.13}$$

Since $\mathbf{b}$'s represents random effect parameter vectors, $l(\boldsymbol{\beta},\mathbf{b},\boldsymbol{\theta} \mid \mathbf{y})$ is not a conventional log-likelihood (Wu & Zhang, 2006). The first term of (2.13) is a weighted residual taking the within-subject variation into account, and $\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}$ is a penalty due to the random effects $\mathbf{b}$ accounting for the between-subject variation. To obtain estimators for $\boldsymbol{\beta}$ and $\boldsymbol{b}$, $\hat{\beta}$ and $\hat{b}$, ´the following requirements must be fulfilled:

$$\begin{cases} \frac{\partial l}{\partial \boldsymbol{\beta}} & = 0 \\ \frac{\partial l}{\partial \mathbf{b}} & = 0. \end{cases} \tag{2.14}$$

By differentiating (2.13) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ and using the usual rules for vector differentiation of scalar functions and equating the derivatives to zero, Hendersson's (1950) simultaneous equations, also known as mixed model equations (MME), are obtained (Robinsson 1991)

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{2.15}$$

MME equations give $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$, that are:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}^{-1}\mathbf{X}'\boldsymbol{\Sigma}-1\mathbf{y} \\ \hat{\mathbf{b}} &= \mathbf{D}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{2.16}$$

where $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$. These solutions are the best linear unbiased estimator (BLUE) and the best linear unbiased predictor (BLUP) for $\boldsymbol{\beta}$ and for $\mathbf{b}$. BLUP estimates are *linear* in the sense that they are linear functions of the data; *unbiased* meaning that the average value of the estimate equals the average value of the quantity being estimated; having a minimum mean squared error within the class of linear unbiased estimators makes it the *best* predictor; and the word predictor is used to distinquish it from the estimators of fixed effects. A fixed effect is considered to be a constant that is estimated but random effects are used as the basis for making inferences about the population from which they come. Data are used to make inferences about the variance of these random variables. The underlying distribution is predicted rather than estimated. (Robinsson 1991; McCulloch, Searle & Neuhaus 2008.)

To construct confidence intervals and test hypotheses about the estimated $\hat{\boldsymbol{\beta}}$, we need information about its variability

$$\hat{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1} \tag{2.17}$$

where $\hat{\boldsymbol{\Sigma}}$ is the REML estimate of $\boldsymbol{\Sigma}$.

2.1.1 Restricted maximum likelihood

The mixed model equations in (2.15) require estimates of $\hat{\mathbf{D}}$ and $\hat{\mathbf{R}}$ in order to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$. Using the normality assumption, the ML and REML methods can be used to estimate $\boldsymbol{\theta}$, the unknown components of $\mathbf{D}$ and $\mathbf{R}$. For the estimation we use the likelihood function defined in (2.13). By integrating out random components $\mathbf{b}$, the following likelihood function is obtained (Wu & Zhang 2006):

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}) &= \int \mathbf{L}(\boldsymbol{\beta},\mathbf{b},\mathbf{D},\mathbf{R}|\mathbf{y})\partial\mathbf{b} = \\ \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} & exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right). \end{aligned} \tag{2.18}$$

The ML estimation for variance components requires maximisation of the logarithm in (2.18) with respect to $\boldsymbol{\theta}$, given $\boldsymbol{\beta}$. However, joint maximation of variance components and fixed-effects parameters does not account for the loss of degrees of freedom. Thus, the estimates may be biased, specifically when the sample size is low but the number of fixed parameters increases.

An example by (Wood 2006) leads us to consider a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, I_\sigma)$ which is a simple form of LME model. The ML estimator for σ^2 can be obtained by differentiating the log-likelihood function and setting it to zero:

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \|\mathbf{y} - \mathbf{X}\beta\|^2 / \sigma^3 = 0 \rightarrow \hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2 / n, \tag{2.19}$$

whereas the unbiased estimator for σ^2 is

$$\hat{\sigma^2} = \frac{1}{n-p}\sigma^2,$$

where p denotes the dimension of $\boldsymbol{\beta}$ and n is the dimension of $\mathbf{y}$. This tendency to underestimate variance components, especially when p increases, is a general feature of ML estimators. Degrees of freedom lost by estimating fixed effects are not taken into account.

To avoid bias, the restricted maximum likelihood estimation (REML) method, introduced by Patterson and Thompson (1971) is favourable. The fundamental idea for estimating parameters in $\boldsymbol{\Sigma}$ is to eliminate $\boldsymbol{\beta}$ from the likelihood functions so that it is defined in terms of $\boldsymbol{\Sigma}$, therefore basing, the estimation only on the relevant part of the data (McCulloch et al. 2008). One way to implement the elimination is to convert the data to a set of linear combinations $\mathbf{K}'\mathbf{y}$ so that $\mathbf{K}'\mathbf{y}$ contains none of the fixed effects in β. McCulloch et al. (2008) describe a formulation of linear combinations in Section 6.9. Another way to eliminate $\boldsymbol{\beta}$ is described by Wu & Zhang (2006) and Wood (2006). The authors integrate $\boldsymbol{\beta}$ out from the likelihood function $L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y})$ to measure the fit of variance components over all possible values of $\boldsymbol{\beta}$. The REML criterion is

$$L_R(\boldsymbol{\Sigma}|y) = \int \mathbf{L}(\boldsymbol{\beta}, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y})\partial\beta$$

In section 6.2.5, Wood (2006) shows that

$$L_R(\boldsymbol{\Sigma}|\mathbf{y}) = \frac{e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})}}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \int e^{-(\beta-\hat{\beta})\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\beta-\hat{\beta})/2}\partial\beta \tag{2.20}$$

which can be recognised as a multivariate normal probability density function (the integral is 1) when divided by $(2\pi)^{p/2}|\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}|^{-1/2}$. This results in

$$L_R(\boldsymbol{\Sigma}|\mathbf{y}) = \frac{e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})}}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \sqrt{\frac{(2\pi)^p}{|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}|}} \tag{2.21}$$

The logarithm of the REML criterion is therefore

(2.22)
$$\log L_R(\boldsymbol{\Sigma}|\mathbf{y}) = \frac{p-n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta).$$

Equation (2.22) deviates from (2.18) by having an additional term

$$-\frac{1}{2}\log|\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}| = \frac{1}{2}\log|(\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X})^{-1}| = \log|\mathbf{Cov}(\hat{\boldsymbol{\beta}})|^{1/2} \tag{2.23}$$

which is the covariance of $\hat{\boldsymbol{\beta}}$ defined in (2.17). Consequently, the REML likelihood multiplies the usual ML likelihood by a factor that is the square root of the variance of $\boldsymbol{\beta}$. The result is the correction for biased values produced by ML estimation.

REML method can be applied to obtain estimators for random effects via maximisation (2.22) in a similar way to how ML function is maximised for $\boldsymbol{\beta}$s. it is interesting to note that REML cannot be used to obtain estimates for fixed effects since $\boldsymbol{\beta}$ is not involved in REML. The resulting estimators of variance components are invariant to the value of $\boldsymbol{\beta}$ (Searle, Casella & McCulloch, 1992). Once REML estimates for variance components in $\boldsymbol{\theta}$ ML estimates for fixed effects $\boldsymbol{\beta}$ are obtained, random effects $\mathbf{b}$ can be predicted using the conditional expectation.

2.2 Inference

Through maximum likelihood based estimation, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ have useful properties, such as establishment of confidence intervals and test hypotheses.

To make inferences of $\hat{\boldsymbol{\beta}}$, covariance matrix (2.17) is used to construct confidence intervals (variability measures) and test of hypothesis. The estimated standard deviation of the ith element of $\hat{\boldsymbol{\beta}}$, $\hat{\beta}_i$, is

$$st.dev(\hat{\beta}_i) = \sqrt{ith\ diagonal\ entry\ of(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}} \tag{2.24}$$

The classical test that is based on the probability of the observed $\hat{\beta}_i$ can be used here to evaluate some hypothesized value of β_i. We can test the null hypothesis H_0:$\beta_i = \beta_0$ versus an alternative H_a:$\beta_i \neq \beta_0$ for some specified value of β_0 by examining the probability of getting the observed $\hat{\beta}$ assuming H_0 to be true. This probability is known as a p-value of the test. It turns out that under H_0

$$Z = \frac{\hat{\beta} - \beta_0}{\sqrt{Var(\hat{\beta})}} \sim_{approx.} N(0,1) \tag{2.25}$$

Ruppert, Wand & Carroll (2003) have argued that the theoretical justification of (2.25) should be considered because of the dependence in $\mathbf{y}$ imposed by the random effects. Therefore, we continue with likelihood ratio test which can be used to compare models with different values of β_i and also to compare models with different random components.

2.2.1 Likelihood ratio test

The likelihood ratio test (LRT), introduced by Neyman and Pearson (1928) is carried out by comparing the maximised likelihood of two (nested) models. It is used to compare the fit of two models: the null (restricted) model against an alternative (unrestricted) model. The likelihood ratio of the models

$$LR(y) = L(\hat{\boldsymbol{\theta}}_{\mathbf{0}}; \mathbf{y})/L(\hat{\boldsymbol{\theta}}_{\boldsymbol{a}}; \mathbf{y}), \tag{2.26}$$

where $\mathbf{y}$ describes the data and $\boldsymbol{\theta}_{\mathbf{0}}$ and $\boldsymbol{\theta}_{\boldsymbol{a}}$ include the estimated parameters under the null model and under the unrestricted model. The likelihood ratio test (LRT) statistic can be constructed by taking twice the minus logarithm of the likelihoods, producing
LRT = -2 $log(LR)(\mathbf{y})$, which can be denoted as $l(\hat{\boldsymbol{\theta}}_{\mathbf{0}}, \mathbf{y}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{a}}, \mathbf{y})$. The larger the difference the stronger the evidence that the restricted model is inadequate.

A key feature for determining the significance of the observed value of the test statistic is that under H_0, LRT can be compared to the percentiles from a chi-squared distribution with q degrees of freedom (Ruppert et al. 2003).

$$LRT \sim_{approx} \chi^2_q, \tag{2.27}$$

where

$$\begin{aligned} q = &\text{ number of parameters in unrestricted model}- \\ &\text{number of parameters in null model.} \end{aligned} \tag{2.28}$$

An advantage of the χ^2-distribution is the connection with standardised normal distribution. χ_1 is the square of a standardised normal variable X,
X $\sim$ N(0,1). This makes the likelihood ratio test for testing the significance of fixed effect straightforward: for example for a hypothesis set-up

$$H_0 : \beta_i = 0 \qquad \text{versus} \qquad H_1 : \beta_i \neq 0$$

q is 1. Thus, p-value can be obtained as

$$\text{p-value} \approx 1 - \phi(\sqrt{-LRT}),$$

where ϕ is the cumulative density function of the normal distribution.

2.2.2 Restricted likelihood ratio test

For the test of significance of random effects, the REML function defined in (2.22) must be used instead of the ML function producing the restricted likelihood ratio test (RLRT) statistic:

$$RLRT(y) = REML(\hat{\boldsymbol{\theta}}_{\mathbf{0}}; \mathbf{y})/REML(\hat{\boldsymbol{\theta}}_{\boldsymbol{a}}; \mathbf{y}). \tag{2.29}$$

The use of RLRT is restricted in way that it can only be used to compare models which have identical fixed effects structures as stated in the previous chapter.

With the use of the LME model, several tests of interest involve testing the significance of variance components, i.e. testing wheter σ_b^2 differs significantly from zero. This test requires more careful consideration since null hypothesis specifies that the variance component is zero and, thus, on its boundary of parameter space (Verbyla, Cullis, Kenward & Welham 1997). As a result, we can not rely on the approximate chi-square distribution defined in (2.27).

Testing the significance of a random component b involves testing if the corresponding variance component is significantly different from zero.

$$H_0 : \sigma_b^2 = 0 \qquad \text{versus} \qquad H_a : \sigma_b^2 > 0.$$

The parameter space for σ_b^2 is]0,∞]. The testing problem is non-standard because the parameter vector is on the boundary (σ_b=0) under the null hypothesis. Thus, the null distribution is no longer chi-squared with degrees of freedom equal to the difference between the number of parameters of the alternative and the null model (Fitzmaurice et al. 2004). Self & Liang (1987) and Stram & Lee (1994) have discussed the asymptotic distribution of RLRT and shown that under the assumption that $\mathbf{y}$ is independent and identically (i.i.d.) distributed and the number of y_i increases to infinity, RLRT statistic distribution is approximately an equally weighted mixture of χ_0 and χ_1 distributions.

$$RLRT(\mathbf{y}) \approx \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 \tag{2.30}$$

meaning that the test statistic has an approximate density function that is a 50:50 mixture of the χ_1 and χ_0 densities. The corresponding p-value for the RLRT statistic is calculated as $1 - [0.5 + 0.5\mathbf{P}(\chi_1^2 \leq RLRT)]$. If the problem of null hypothesis being on the boundary of its parameter space were ignored, the corresponding p-value would be overestimated (pverly large p-values observed) which could lead to a selection for the covariance structure that is too parsimonious (Fitzmaurice et al. 2004).

The weakness of the presentation of Self & Liang and Stram & Lee is the assumption of the independence of the $\mathbf{y}$ meaning that $\mathbf{y}$ can be partitioned into subvectors that are independent. This assumption does not hold in the LME model framework, specifically not for the alternative model (Crainiceanu & Ruppert 2004c), meaning that one does not get chi-square mixtures as the asymptotic null distribution.

Crainiceanu & Ruppert (2004c) suggest the elimination of the i.i.d. assumption that changes the null distribution theory. These authors as well as Crainiceanu, Ruppert & Vogelsang (2002) suggest the use of simulations to determine the null distribution of the likelihood ratio test statistic. Crainiceanu and Ruppert (2004b) derive an algorithm to simulate the finite sample distribution of the RLRT statistic for a model with one variance component (p. 4).

If there are more than one variance component of interest, which is the case when the random effects structure is extended, the asymptotic distribution theory becomes more complicated. Craniceanu & Ruppert (2004c) recommend the use of simulations that can be computed efficiently through their algorithm.

2.2.3 Use of simulations for the restricted likelihood ratio test

Computing the RLRT statistic is easy using available standard software but determining the asymptotic distribution of the observed test statistic is not. Several papers have been published related to this problem (see Ruppert Crainiceanu, Ruppert, Claeskens & Wand (2002); Crainiceanu & Ruppert (2004a,b,c) among others). As a conclusion the writers suggest the use of parametric bootstrap.

The underlying idea is to simulate observations under the null hypothesis, fit them according to the alternative and obtain the likelihood ratio test statistic between the models. A large number, say 10,000, of independent data sets are simulated with fixed parameters at their estimated values added by values of $\mathbf{ZDZ}' + \mathbf{R}$ that are generated according to their estimated variances $\hat{\sigma}_\epsilon$ and $\hat{\sigma}_b$. The likelihood test statistic is calculated for each data set and the exact p-value is obtained as a proportion of the simulated values of the test statistic that exceeded the value of observed RLRT (Faraway 2006). In statistical software R, there is RLRsim library (Scheipl 2011) that can be used for testing the significance of random components. Functions in the library are based on the algorithm from Craniceanu & Ruppert (2004c) which makes simulation efficient using matrix computation techniques.

3 Scatterplot smoothing

Longitudinal data often exhibit a relationship between the response and the explanatory variable that is best described by nonparametric models. Parametric models tend to be too limited whereas nonparametric methods are flexible and robust against parametric assumptions. The basic idea of the nonparametric approach is to let the data determine the most suitable form for the function but at the same time preserve the summarising characteristic of statistical modelling. These models are unspecified to a probabilistic model meaning that it is not assumed that $\mathbf{y}$ and $\mathbf{x}$ have a predefined form of dependence.

Let us define $\mathbf{y} = [y_1, ..., y_n]$ as a response variable the variation of which is explained by $\mathbf{x} = [x_1, .., x_n]$. A simple model could be

$$y_i = f(x_i) + \epsilon_i, \tag{3.1}$$

where f is a smooth function and $\boldsymbol{\epsilon}$ follows the normal distribution N$\sim$ $(0, \sigma^2)$. In the spirit of least squares, $\hat{f}$ would simply be a function that minimises the mean squared error $\frac{1}{n}\sum(y_i - f(x_i))^2$. Without any further restrictions this would lead to the solution $\hat{f}(x_i) = y_i$ which, clearly, does not summarise the variation of $\mathbf{y}$ but follows every point of the data precisely. Therefore, smooth functions need to be restricted to control the roughness of the fit.

In this chapter splines, piecewise defined functions, are discussed as a nonparametric mean to smooth data. Eubank (1999) provides a deeper insight into smoothing and offers more theoretical aspects but the presentation is kept simple throughout the work. Another book about nonparametric regression methods with emphasis on applications is provided by Keele (2008).

Splines form just one class of the large collection of scatterplot smoothers. As an alternative one can use local polynomial fitting (Wand & Jones 1995) or wavelet smoothing (Daubechies 1992; Härdle, Kerkyacharian, Picard & Tsybakov 1998), among others.

3.1 Regression spline

Let

$$\tau_0, \tau_1, \tau_2, ..., \tau_p$$

denote a set of locations in a closed interval [a,b] covering the range of the explanatory variable x. These locations are called knots. The knots divide the

interval into p contiguous subintervals when $\tau_1<...<\tau_p$. Neighbouring knots, τ_p and τ_{p+1}, are known as local neighbourhoods and intervals between neighbouring knots are modelled piecewise using k-th degree polynomial as a basis function. To ensure the contiguity of the curve, derivatives are required to be contiguous up to the k-1 -times derivative at each knot. For the basis to build these models, we introduce a truncated power basis.

Let us define the k-th degree truncated power basis with p knots as

$$1, x, ..., x^k, (x-\tau_0)^k, (x-\tau_1)^k, ..., (x-\tau_p)^k. \tag{3.2}$$

Any linear combination of these basis functions is called a spline where

$$(x-\tau_r)_+ = \begin{cases} 0 & x < \tau_r \\ x-\tau_r & x \geq \tau_r. \end{cases} \tag{3.3}$$

Using (3.2), a k-th degree regression spline with p knots can be expressed as

$$f(x) = \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{p} \beta_{k+r}(x-\tau_r)_+^k, \tag{3.4}$$

where first k +1 coefficients, [β_0,...,β_k], are for the polynomial, and β_{k+r}, ($1 \leq$ r $\leq$ p) are for the truncated line functions.

The truncated power basis is conceptually simple but numerical computing may cause problems. Specifically, large values of k can lead to rounding problems (Hastie & Tibshirani 2001). As an alternative there are a variety of other widely used basis functions such as B-splines, wavelet basis or Fourier series available in literature. Wood (2006) and Hastie, Tibshirani and Friedman (2001) provide a description of different basis functions with a good list of references.

The regression spline defined in (3.4) can be written as

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \sum_{r=1}^{p} \beta_{k+r}(x-\tau_r)_+^k. \tag{3.5}$$

As can be seen in (3.5) the shape of a spline depends on the degree of the polynomial basis function and on the location of the knots. The degree of regression splines is often taken as 1, 2 or 3 for computational convenience as discussed earlier. These splines are called linear, quadratic and cubic regression splines (Wu & Zhang 2006). Hastie et al. (2001) have stated that it is hardly ever necessary to use spline models of a higher degree than three.

Clearly, a spline can only change its shape after passing a knot which means that the positioning of knots is an essential factor for determining the shape of the regression spline. When the number of knots equals all time points in the range of x, the curve becomes rough, showing all of the smallest changes, whereas a large distance between knots produces an estimate with low variance

but potentially high bias. A general overview of different knot selection methods is provided by Wu and Zhang (2006) in Section 3.3.3. Ruppert (2002) and Wand (2003) discuss the selection and placements of knots in more detail. Wand has proposed a simple rule for knot specification that suits well for most problems.

$$\tau_p = \frac{p+1}{K+1}\text{th sample quantile of unique x} \qquad 1 < \text{p} < \text{K}, \tag{3.6}$$

with

$$K = \begin{cases} 5, & \frac{1}{4}N < 5 \\ \frac{1}{4}N, & 5 \leq \frac{1}{4}N \leq 35, \\ 35, & \frac{1}{4}N > 35 \end{cases}$$

where N is the number of unique values of x. The specification above was determined for penalised splines but similar specifications are generally used for other spline models as well. For example the *smooth.spline*-function (Ripley & Maechler) in *R* uses a rule similar to this specification.

To write (3.5) in linear form

$$y = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{3.7}$$

the model matrix X and vector $\boldsymbol{\beta}$ need to be defined. Let us denote the truncated line basis in (3.2) for p knots as

$$\boldsymbol{\Psi}_r(x) = [1, x, .., x^k, (x-\tau_0)^k, ..., (x-\tau_p)^k]' \tag{3.8}$$

where $r = p + k + 1$ involving the number of basis functions and

$$\boldsymbol{\beta} = [\beta_0, ..., \beta_k, \beta_{k+1}, ..., \beta_{k+p}]'$$

are associated coefficients. Now, the regression spline can be re-expressed as

$$f(x) = \boldsymbol{\Psi}_{\mathbf{k}}(\mathbf{x})'\boldsymbol{\beta}.$$

The matrix $\boldsymbol{\Psi}$ contains linearly independent basis functions and therefore we have a linear model matrix of full rank and thus invertible. These definitions allow us to write a regression spline in linear form $\mathbf{y} = \mathbf{X}\beta$ where

$$\begin{aligned} \mathbf{y} &= (y_1, .., y_n)' \\ \mathbf{X} &= \boldsymbol{\Psi}_{\mathbf{k}}(\mathbf{x})' \\ \boldsymbol{\beta} &= [\beta_0, ..., \beta_k, \beta_{k+1}, ..., \beta_{k+p}]' \\ \boldsymbol{\epsilon} &= (\epsilon_1, ..., \epsilon_n)'. \end{aligned}$$

The estimator for $\boldsymbol{\beta}$ can be obtained by using the ordinal least squares method resulting in

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Fitted values for $\mathbf{y}$ are

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.9}$$

and the influence matrix is

$$\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{3.10}$$

$\mathbf{A}$ is also called a smoother matrix. The trace of A = r, which is the degrees of freedom of the smoother, that measures the complexity of the fitted curve (Wu & Zhang 2006).

3.2 Smoothing spline

Regression splines require care in terms of knot locations and the fit of the model depends rather strongly on these locations. Another way of smoothing scatterplots is smoothing splines, where the amount of knots is high and smoothness is controlled by adding a penalty term for the roughness. Roughness of the fitted curve, f, can be defined as an integral of its squared k-times (k≥1) derivative

$$\int \left[f^k(x)\right]^2 dx. \tag{3.11}$$

Let us define the penalty term as

$$P = \lambda \int \left[f^k(x)\right]^2 dx, \tag{3.12}$$

where $\lambda > 0$ and $f(x)$ is continuous up to the k-th derivative. P sums the measure of curvature of f along its entire range and λ is a mutable parameter that controls the relative weight to be given to the P. When λ is set small, the curve follows data points carefully and is less biased but $\hat{f}$ has greater variance. The bigger the λ the smoother the curve will be but the bias increases along (Keele 2008). The fitting problem has extended from ordinary least squares to minimising the penalised least squares (PLS)

$$\sum_{i=1}^{n}(y_i - f(x_i)) + \lambda \int \left[f^k(x)\right]^2 dx \qquad k \geq 1. \tag{3.13}$$

To minimise the criterion in (3.13), integral in (3.11) needs to be computed which is a challenging computational issue. However, when k=2, the associated cubic smoothing spline (Green & Silverman 1994) is computationally faster which makes them popular in statistical applications. It is shown by Green & Silverman (1994) that the roughness term (3.11) of a cubic smoothing spline can be expressed as

$$\int_a^b \left[f''(x)\right]^2 dx = \mathbf{f}'\mathbf{G}\mathbf{f}, \tag{3.14}$$

where $\mathbf{G}$ is a positive semi-definite roughness matrix. Therefore, the PLS criterion for a cubic smoothing spline (CSS) is

$$\|\mathbf{y} - \mathbf{f}(x)\|^2 + \lambda \mathbf{f}'\mathbf{G}\mathbf{f} \tag{3.15}$$

and, respectively, the expression for smoother f_{css} is

$$f_{css} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{G})^{-1}\mathbf{X}'\mathbf{y}, \tag{3.16}$$

where $\mathbf{X}$ denotes the truncated power basis defined in (3.2). The associated smoothing matrix $\mathbf{A_{css}}$ is

$$\mathbf{A_{css}} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{G})^{-1}\mathbf{X}.$$

The PLS criterion can only estimate model coefficients, $\boldsymbol{\beta}$, given λ. However, the choice of the smoothing parameter λ plays an important role for smoothing splines. A good selector usually tries to select a favorable value for λ, for the trade-off between the goodness of fit and the complexity of the model. For smoothing splines, a good choice can be obtained by applying smoothing parameter selectors such as cross-validation (CV) or generalized cross-validation (GCV). The main idea is to fit the smooth function by leaving out each measurement at a time and calculating a difference between the missing value and its fitted value. Differences are summed together over all measurements. A more precise description of the optimisation methods for smoothing splines can be found in Wood (2006).

3.3 Penalised spline

Smoothing splines are performed with a large amount of knots, which leads to a great number of estimated parameters which are therefore expensive to compute. An alternative to smoothing splines are penalised splines proposed by Eilers and Marx (1996). It is a method for fitting a smoothing spline using knots, where the roughness of the fit is controlled via imposing constraints on part of the coefficient in $\boldsymbol{\beta}$. The selection of knots for penalised splines is discussed in Ruppert (2002), who finds that the number of knots has little effect on the smooth curve, providing that the number is at least the mimimum value of 5. On the other hand, one can use large number of knots even if it is not necesssary to do so. The amount of smoothing depends on the trace of the smoother matrix $\mathbf{G}$ which is defined below in (3.17). The advantage of penalised splines is that the number of parameters can be kept reasonably small, which makes rapid computation feasible and allows large datasets to be smoothed, while the accuracy is as good as with smoothing splines (Howell 2007).

We use earlier defined regression spline to construct a smooth function which is to be penalised to its coefficients to avoid overfitting;

$$f(x) = \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{p} \beta_{k+r}(x - \tau_r)^k.$$

Taking k-times derivative of $f(x)$, with predefined knots τ_r for r=1,...,p, we obtain

$$f^k(\tau_r-) = k! \sum_{s=0}^{r-1} \beta_{k+s} \qquad \text{and}$$
$$f^k(\tau_r+) = k! \sum_{s=0}^{r} \beta_{k+s}.$$

Therefore,

$$f^k(\tau_t+) - f^k(\tau_t-) = k!\beta_{k+r},$$

that is, $f^k(x)$ jumps at τ_r with amount of $k!\beta_{k+r}$ (Wu & Zhang 2006). This means that a regression spline of degree of k with defined knots $\tau_1,...,\tau_p$ has continuous derivatives up to k-1 times and has a discontinuous k-times derivative meaning that β_{k+r} measures how large the jump is. Therefore, when the number of knots is fixed, controlling the roughness of $f(x)$ can be imposed by controlling the size of $|\beta_{k+r}|$. Ruppert et al. (2003) suggest the following constraint on the coefficients:

$$\sum_{r=1}^{p} \beta_{k+r}^2 \leq C \tag{3.17}$$

This can be implemented by defining matrix $\mathbf{G}$ as follows:

$$\mathbf{G} = \begin{pmatrix} \mathbf{0}_{(k+1)\times(k+1)} & \mathbf{0}_{(k+1)\times K} \\ \mathbf{0}_{K\times(k+1)} & \mathbf{I}_K \end{pmatrix} \tag{3.18}$$

where $\mathbf{G}$ is positive semi-definite implying that penalty is always non-negative. Now the constraint in (3.17) can be expressed as $\boldsymbol{\beta}'\mathbf{G}\boldsymbol{\beta} < \mathrm{C}$ which, with proper choice of C, modifies the least squares problem as

$$min \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right\|^2 \qquad \text{subject to} \qquad \boldsymbol{\beta}'\boldsymbol{G}\boldsymbol{\beta}, \tag{3.19}$$

$\mathbf{X}$ denoting the truncated power basis in (3.2). Trace of $\mathbf{G}$ is the effective degrees of freedom that defines the smooth of the function.

The choice for $\mathbf{G}$ is not unambiguos but depends both on the form of the penalty (3.17) imposed to coefficients (see other penalties from Ruppert et al. 2003, p. 65) and on the basis being used to construct the smooth. The left corner of $\mathbf{G}$ is specified for the truncated power basis (Crainiceanu and Ruppert 2004c).

Using a Lagrange multiplier optimisation method, the minimisation problem in (3.19) is equivalent to choosing β to minimise following penalised least squares criterion:

$$\left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \boldsymbol{\beta}'\mathbf{G}\boldsymbol{\beta}. \tag{3.20}$$

The estimation problem of penalised splines can be conducted in the linear mixed-effects model framework which is a central approach in this thesis. In next chapter, we introduce the estimation procedure for spline coefficients and for λ and look at the possibilities that fusion with mixed-effects model environment provides to nonparametric modelling.

4 Penalised splines in the linear mixed-effects model framework

Nonparametric smoothing techniques can be connected to parametric environment through a fusion between a penalised spline model and a linear mixed-effects model. Penalised spline model coefficients are partitioned in a specific way and the shrinkage property of LME model utilised. The LME model approach provides an analytic framework for the construction of penalised spline where the amount of smoothing is directly related to variance components. This introduces semiparametric regression (Rupper et al. 2003), an extension of parametric regression that uses penalised spline functions to achieve greater flexibility. It enables simultaneous modelling of several variables where those features of the data that are suitable for parametric modeling are modeled that way and nonparametric components are used only where needed (Ruppert, Wand & Carroll 2009).

Early contributions to the LME model representation to nonparametric curve fitting include Speed (1991) and Verbyla (1994). Connection to the penalised splines was introduced by Speed. Commenting on Robinssons's paper, he showed that it was possible to fit penalised splines as LME models and, thus, splines can be considered as BLUPs. Later, the fusion has been discussed by many researchers e.g Eilers and Marx (1996), among others. Ruppert et al. (2003) discuss the method in their book *Semiparametric Regression* (2003) with a great variety of extensions. A practical point of view using S-Plus is provided by Ngo and Wand (2004).

Representation of the penalised splines as BLUPs is useful because it allows practical implementation to be conducted using standard software such as R. This facilitates the use of LME model's diagnostics, inference and model selection as well as modelling the hierarcial structure of longitudinal data with correlated measurements. Estimated model coefficients do not have direct meaning but standard errors can be calculated and used for the inference procedure.

Later in this chapter, we discuss methods of extending the presented semiparametric regression technique. The group-level penalised spline model is extended to account for the individual effects and it is shown how to test factor influence in nonparametric models.

4.1 Estimation of penalised splines

As discussed in the previous chapter, penalised splines offer an efficient and relatively simple mean to construct a nonparametric model when parametric models do not provide a suitable fit. A penalised spline model

$$f(x) = \sum_{s=0}^{k} \beta_s x^s + \sum_{r=1}^{p} \beta_{k+r}(x - \tau_r)^k \tag{4.1}$$

is fitted through least squares where constraints are added to the truncated line coefficients to control the roughness of the curve. As shown in the previous chapter, the penalised least squares (PLS) criterion is

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}' \boldsymbol{G} \boldsymbol{\beta}$$

where $\mathbf{G}$ is the constraint matrix and λ denotes the weight given to penalisation of the roughness.

Wand (2003) has shown that modelling spline coefficients both through fixed and random effects make the fit smoother compared to the situation where the whole spline is fitted using fixed effects. A shrinkage property of a normally distributed variable with zero mean and a known covariance matrix, say $\mathbf{R}$, makes variable estimates shrink towards zero which is favorable for prediction. Wand takes advantage of this property and divides a regression spline into the fixed and random components of an LME model. His approach is presented here through a descriptive example which is also discussed in his article (p. 7-8).

Let us consider a linear penalised spline, which is obtained from (4.1) setting k=1. To demonstrate the difference between the two fitting approaches, the truncated line basis functions $(x - \tau_k)_+$ are fitted through fixed components and random components. A fit that is obtained by modelling truncated lines as fixed effects is shown in the left panel of Figure 4.1. The fit is rather rough due to the large number of truncated line basis functions which are shown in the bar at the base of the panel.

The roughness of the fit can be remedied imposing a restriction

$$b_r \sim N(0, \sigma^2_{\mathbf{b}}). \tag{4.2}$$

This makes $\mathbf{b}$ a random component. The corresponding variance component $\sigma^2_{\mathbf{b}}$ has a finite parameter space ($\sigma^2_{\boldsymbol{b}} < \infty$) imposed from (4.2). b_k's they are no longer allowed to vary freely between -∞ and ∞ but they must obey the normal distribution with zero mean. This makes b_k's to shrink toward zero and leads to a smoother fit. The fit is shown in the right panel in Figure 4.1. This approach makes penalised spline modelling straightforward since it relies on the well-known LME theory and, thus, model can be estimated with standard LME softwares such as lme-function in nlme library (Pinheiro et al. 2011) in *R*.

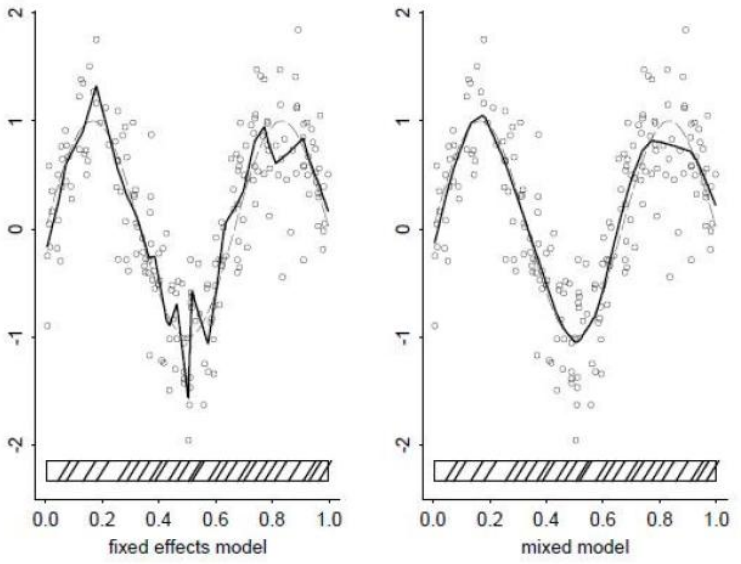

Figure 4.1. A penalised spline model fitted in two ways: through fixed effects (left panel) and through a combination of fixed and random effects (right panel).

To present a penalised spline as an LME model, we need to reparameterize vectors $\boldsymbol{\beta}$ and $\mathbf{b}$ and model matrices $\mathbf{X}$ and $\mathbf{Z}$ in a way that they can be connected to LME model. The coefficient vector $\boldsymbol{\beta}$ is split into two parts in a following way:

$$\boldsymbol{\beta} = [\beta_0, \beta_1, .., \beta_k]' \qquad \mathbf{b} \ = [b_1, b_2, .., b_r]',$$

where $\boldsymbol{\beta}$ consists of the coefficients of (k+1) polynomial basis functions and $\mathbf{b}$ corresponds the truncated line coefficients β_{k+r} in (4.1). The model matrices $\mathbf{X}$ and $\mathbf{Z}$ are defined in a way that $\mathbf{X}$ consists of the first (k+1) columns from the truncated power basis functions in (3.2) and $\mathbf{Z}$ consists of the remaining columns.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & \dots & x_1^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^k \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} (x_1 - \tau_1)^k & \dots & (x_1 - \tau_p)^k \\ \vdots & \ddots & \vdots \\ (x_n - \tau_1)^k & \dots & (x_n - \tau_p)^k \end{pmatrix}.$$

Accounting for the special structure of $\mathbf{G}$ defined in (3.17), estimators for spline coefficients $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{b}}$ can be obtained from

$$\begin{aligned} & min \, \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\| \\ & = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) + \lambda \mathbf{b}'\mathbf{b}. \end{aligned} \tag{4.3}$$

This equals to the LME likelihood function in (2.13) when $\mathbf{y}|\mathbf{b} \sim (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R})$ where $\mathbf{b} \sim (\mathbf{0}, \mathbf{G})$. Assuming $\mathbf{b}$ and $\boldsymbol{\epsilon}$ to be independent and normally

distributed, a penalised spline model can be presented as a LME model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \qquad \text{where} \qquad Cov\begin{pmatrix}\mathbf{b}\\ \boldsymbol{\epsilon}\end{pmatrix} = \begin{pmatrix}\sigma_{\mathbf{b}}^{\mathbf{2}}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\sigma}_{\boldsymbol{\epsilon}}^{\mathbf{2}}\boldsymbol{I}\end{pmatrix}. \tag{4.4}$$

This implies that the penalised spline is the BLUP for $\mathbf{y}$. Identity matrix $\mathbf{I}$ for error terms, $\boldsymbol{\epsilon}$, is assumed for simplicity. λ can be obtained as the ratio of the two variance components

$$\lambda = \sigma^2/\gamma^2.$$

However, λ has no interpretation as a smoothing parameter since the spline is penalised to its coefficients but λ is a by-product.

The solution for (4.3) is

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{b}}\end{pmatrix} = [\mathbf{C}'\mathbf{C} + \lambda\mathbf{G}]^{-1}\mathbf{C}\mathbf{y} \tag{4.5}$$

where $\mathbf{C}$=[$\mathbf{X}$ $\mathbf{Z}$], Wand (2003). Correspondingly, the fitted values $\hat{\mathbf{f}}$ can be written as

$$\hat{\mathbf{f}} = \mathbf{C}(\mathbf{C}'\mathbf{C} + \lambda\mathbf{C})^{-1}\mathbf{C}'\mathbf{y}. \tag{4.6}$$

With LME representation we can obtain variability estimates for the fitted curve. Randomness of $\mathbf{b}$ is a device used to model the curvature of $f(x)$, whereas $\boldsymbol{\epsilon}$ accounts for the variability of the curve. Therefore, variance of $\hat{\mathbf{y}}$ should be calculated with respect to the conditional distribution of $\mathbf{y}$ with given $\mathbf{b}$.

$$\text{Var}(f(x)|\mathbf{b}) = \begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix} Cov\left[\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{b}}\end{pmatrix}|\mathbf{b}\right]\begin{pmatrix}\mathbf{X} & \mathbf{Z}\end{pmatrix}' = \mathbf{C}Cov\left[\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{b}}\end{pmatrix}|\mathbf{b}\right]\mathbf{C}' \tag{4.7}$$

where $\mathbf{C}$=[$\mathbf{X}$ $\mathbf{Z}$] (Ruppert el al. 2003). Following Ruppert et al. (2003, Section 4.7), the conditional covariance matrix of $\boldsymbol{\beta}$ and $\mathbf{b}$ is

$$Cov\left[\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{b}}\end{pmatrix}|\mathbf{b}\right] = \hat{\sigma}_\epsilon^2(\mathbf{C}'\mathbf{C} + \frac{\hat{\sigma}_\epsilon}{\hat{\sigma}_b^2}\mathbf{R})^{-1}\mathbf{C}'\mathbf{C}(\mathbf{C}'\mathbf{C} + \frac{\hat{\sigma}_\epsilon}{\hat{\sigma}_b^2}\mathbf{R})^{-1}.$$

Assuming that $\boldsymbol{\epsilon} \sim (\mathbf{0}, \mathbf{I}\sigma_{\boldsymbol{\epsilon}}^2)$,

$$\hat{f}(x)|\mathbf{b} \sim N(E(\hat{f}(x)|\mathbf{b}), Var(\hat{f}(x)|\mathbf{b}))$$

and 100(1-α)% confidence interval for $\hat{f}(x)$ is

$$\hat{f}(x) \pm \boldsymbol{\phi}(1 - \frac{\alpha}{2})st.dev(\hat{f}(x)|\mathbf{b})$$

where

$$st.dev.(\hat{f}(x)|\mathbf{b}) = \sqrt{Var(\hat{f}(x)|\mathbf{b})} = \hat{\sigma}_\epsilon^2\sqrt{\mathbf{C}(\mathbf{C}'\mathbf{C} + \frac{\hat{\sigma}_\epsilon}{\hat{\sigma}_b^2}\mathbf{R})^{-1}\mathbf{C}'\mathbf{C}(\mathbf{C}'\mathbf{C} + \frac{\hat{\sigma}_\epsilon}{\hat{\sigma}_b^2}\mathbf{R})^{-1}\mathbf{C}'}.$$

4.2 Hypothesis testing

Determining test of the significance of a penalised spline model involves testing whether the fitted spline deviates significantly from the corresponding polynomial basis function. As will be shown, this test is equivalent to testing if the variance component determining the coefficients for truncated lines differs from zero (Crainiceanu and Ruppert 2004a). Therefore, considerations regarding tests of random components need to be accounted for in the RLRT test for penalised spline models. Let us consider the following model

$$y_i = f(x_i) + \epsilon_i$$

where $f(x)$ is an undefined function under interest. We are interested in testing the hypothesis whether $f(x)$ is a k-th degree polynomial

$$H_0 : f(x) = \beta_0 + \beta_1 x + ... + \beta_k x^k$$

versus an alternative, that $f(x)$ is a more flexible model that is obtained by a penalised spline

$$H_1 : f(x) = \beta_0 + \beta_1 x + ... + \beta_k x^k + \sum_{r=1}^{p} b_r (x - \tau_r)_+^k,$$

where b_r's account for departures from the polynomial. This is equivalent to testing

$$H_0 : \sigma^2_{\mathbf{b}} = 0 \qquad \text{versus} \qquad H_1 : \sigma^2_{\mathbf{b}} > 0. \tag{4.8}$$

Since $\mathbf{b}$ has zero mean and covariance matrix $\mathbf{I}\sigma^2_{\mathbf{b}}$ (defined in the previous section), the condition that $\sigma^2_b = 0$ under H_0, is equivalent to the condition that all truncated polynomial coefficients b_r's are identically zero. If $\sigma^2_{\mathbf{b}} > 0$, then any open set of truncated line coefficients deviates from zero meaning that there is a need for a more complex structure than can be obtained with the polynomial.

For testing the hypothesis set-up in (4.8), we consider the likelihood ratio test. Since the fixed-effect structure remains same under the null and the alternative hypothesis, the restricted likelihood test is appropriate. The RLRT statistic is

$$RLRT(\mathbf{y}) = -2[l_{H0}(0, \hat{\sigma^2_{\boldsymbol{\epsilon}}}; \mathbf{y}) - \ l_{H1}(\hat{\sigma^2_{\mathbf{b}}}, \hat{\sigma^2_{\boldsymbol{\epsilon}}}; \mathbf{y})].$$

Computing RLRT is straightforward; with most softwares REML or ML can be derived directly from the model output. The challenge is to find the null distribution of the test statistic. As stated in Chapter 3, the chi-square distribution can not be used since under H_0, $\sigma_{\mathbf{b}}$ is on its boundary of the parameter space. Moreover, Crainiceanu, Ruppert & Vodelgsang (2002) have shown that the chi-square mixture $0.5\chi_q + 0.5\chi_{q+1}$ by Self & Liang (1987) as an asymptotic distribution approximation can be poor because of the lack of independence in the response variable.

RLRTs in nonparametric longitudinal models are discussed in several papers: Crainiceanu, Ruppert, Claeskens & Wand (2002), Crainiceanu & Ruppert (2004a,b,c). Crainiceanu et al. (2002) discuss the null finite sample for RLRT for testing polynomial against a general alternative. Crainiceanu & Ruppert (2004b) created an algorithm that took an advantage of spectral composition of LRT which makes the simulations from distributions remarkably faster. Crainiceanu & Ruppert (2004c) discuss the problem in with several variance components. As summed up in Section 2.2.2, the recommendation is to use parametric bootstrap to obtain the finite sample distribution of RLRT. Using the simulated quantiles, exact p-values can be extracted.

4.3 Extension of penalised spline models

The fusion between penalised splines and LME models allows easy extension to more complex modelling problems. In this section we show how to model an interaction between a categorical variable and a smoothing curve. We also show how to account for subject-specific trajectories through penalised splines.

4.3.1 Interaction models

When analysing longitudinal data, we can easily imagine a need for a model that incorporates parametric and nonparametric covariates to describe how a smooth curve varies across groups. Following Coull, Ruppert & Wand (2001) and Ruppert et al. (2003), we will describe models that incorporate factor-by-curve interaction into smoothing methods. LME model representation of penalised splines makes this extension to (4.1), which included only a single covariate, straightforward. The interaction between two continuous covariates, known as the varying coefficient model (Hastie and Tibshirani 1993), is excluded from this section. In varying coefficient models, the effect of one variable is linear but intercept and slope depend nonparametrically on the second variable. A general overview can be found in Ruppert et al. (2003) in Section 12.5. For a detailed description see Hastie & Tibshirani.

A binary-by-continuous interaction model can be used to model smooth curves for two different groups and to test the difference between them. A model which incorporates an interaction between a binary factor and a nonparametric function of a continuous variable is

$$y_i = f_{z_i}(x_i) + \epsilon_i, \tag{4.9}$$

where f_{z_i} is a smooth curve depending on the value of z_i, and z_i is the categorical variable (Coull et al. 2001). To describe how to build a structure for (4.9), we start with its simplification $y_i = \gamma_0 z_i + f(x_i) + \epsilon_i$, where $\gamma_0 z_i$ shows the vertical shift between two curves. The corresponding penalised spline model is

$$y_i = \beta_0 + \sum_{s=1}^{k} \beta_s x_i^s + \sum_{r=1}^{p} b_r (x_i - \tau_r)_+^k + \gamma_0 z_i + \epsilon_i. \tag{4.10}$$

Defining the indicator of the lth location, l$\in$ {1,2}, as

$$z_{il} = \begin{cases} 1, & \text{if } z_i = l \\ 0, & \text{otherwise} \end{cases} \tag{4.11}$$

the interaction model in (4.9) can be written in a following way:

$$y_i = \beta_0 + \sum_{s=1}^{k} \beta_s x_i^s + \sum_{r=1}^{p} b_r (x_i - \tau_r)_+^k + \sum_{l=1}^{2} z_{il} \left(\gamma_{0l} + \sum_{s=1}^{k} \gamma_{sl} x_i^s + \sum_{r=1}^{p} \upsilon_{rl} (x_i - \tau_r)_+^k \right) + \epsilon_i \tag{4.12}$$

that is subject to the constraints

$$\sum_{r=1}^{p} b_r^2 < C \qquad \text{and} \qquad \sum_{r=1}^{p} \upsilon_{rl}^2 < C_l \tag{4.13}$$

for some constants C and C_l, l=1,2. For simplicity, the polynomial part in (4.12) is often taken to be linear (Coull et al. 2001) which produces a linear penalised spline model:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{r=1}^{p} b_r (x_i - \tau_r)_+ + \sum_{l=2}^{2} (z_{il}(\gamma_{0l} + \gamma_{1l} x_i) + \sum_{l=1}^{2} z_{il} \left(\sum_{r=1}^{p} \upsilon_{rl} (x_i - \tau_r)_+ \right) + \epsilon_i. \tag{4.14}$$

In (4.14), $(\gamma_{0l} + \gamma_{1l} x_i)$ models the linear deviation between f_1 and f_2, whereas $\sum_{r=1}^{p} \upsilon_{rl} (x_i - \tau_k)_+$ represents deviations from the overall smooth term $\sum_{r=1}^{p} b_r (x_i - \tau_r)_+$. The constraints in (4.13) induce the penalty to control the smoothness of f_l.

The extension of the binary-by-continuous model to the discrete-by-continuous interaction model is straightforward. The level of groups increases, which makes slight modifications to (4.12) and to the linear penalised spline model in (4.14). Let x_i represent a continuous predictor, and $z_i \in \{1,...,\text{L}\}$ denote a factor variable, where L $>$ 2. The formulation of the model is similar to (4.9),

$$y_i = f_{z_i}(x_i) + \epsilon_i, \tag{4.15}$$

but now there are L different smooth functions, f_1,...,f_L, depending on the value of z_i. The corresponding linear penalised model is

$$y_i = \beta_0 + \beta_1 x_i + \sum_{r=1}^{p} b_r (x_i - \tau_r)_+ + \sum_{l=1}^{L} z_{il} (\gamma_{0l} + \gamma_{1l} x_i) + \sum_{l=1}^{L} z_{il} \left(\sum_{r=1}^{p} \upsilon_{rl} (x_i - \tau_r)_+ \right) + \epsilon_i. \tag{4.16}$$

Following Coull et al. (2001), b_r and υ_{rl} are taken to be normally distributed,

$$b_r \sim N(0, \sigma_b^2) \qquad \text{and} \qquad \upsilon_{rl} \sim N(0, \sigma_\upsilon^2). \tag{4.17}$$

Hence, we can write (4.16) as LME model notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \gamma_{02}, ..., \gamma_{0L}, \gamma_{12}, ..., \gamma_{1L}]',$$
$$\mathbf{b} = [b_1, ..., b_p, \upsilon_{11}, ..., \upsilon_{p1}, \upsilon_{12}, ..., \upsilon_{p2}]',$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & z_{12} & \dots & z_{1L} & z_{12}x_1 & \dots & z_{1L}x_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_n & z_{n2} & \dots & z_{nL} & z_{n2}x_n & \dots & z_{nL}x_n \end{pmatrix},$$

$$\mathbf{Z} = \begin{pmatrix} (x_1-\tau_1)_+ & \dots & (x_1-\tau_r)_+ & z_{11}(x_1-\tau_1)_+ & \dots & z_{11}(x_1-\tau_r)_+ & \dots & z_{1L}(x_1-\tau_r)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (x_n-\tau_1)_+ & \dots & (x_n-\tau_r)_+ & z_{n1}(x_n-\tau_1)_+ & \dots & z_{n1}(x_n-\tau_r)_+ & \dots & z_{nL}(x_n-\tau_r)_+ \end{pmatrix}$$

and

$$\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I} \end{pmatrix}\right)$$

with $\mathbf{G} = \text{diag}(\sigma^2_b\mathbf{I_r}, \sigma^2_{v1}\mathbf{I}_r, ..., \sigma^2\mathbf{I}_r)$.

In order for the fixed-effects parameters to be identified, constraints on γ_{sl} need to be imposed. In presented notations we have assumed γ_{01}=γ_{11}=0, which makes z_1 a reference group to which other groups are compared. Therefore,

$$\beta_0 + \beta_1 x_i + \sum_{r=1}^{p} b_r(x_i - \tau_r)_+$$

is the fitted curve for $l = 1$ and

$$\sum_{l=2}^{L}(z_{il}(\gamma_{0l} + \gamma_{1l}x_i) + \sum_{l=1}^{L} z_{il}\left(\sum_{r=0}^{p} \upsilon_r^l(x_i - \tau_r)_+\right)$$

is the difference betwwen the reference group and the group where $l = L$. Constraints could have been chosen in another way but it does not have an effect on the shape of curves but on the interpretation of parameters (Rupper et al. 2003).

For testing the significance of the interaction model, one can fit a model with one common average curve and compare it to the model with the factor-by-curve interaction. For the linear penalised spline model defined in (4.16), the null and the alternative hypotheses are

$$\begin{aligned} &H_0 : \gamma_{jl} = 0 \qquad j = 0, 1 \qquad l = 1, ..., L \qquad \text{and} \qquad \sigma^2_v = 0 \\ &H_1 : \gamma_{jl} \neq 0 \qquad \text{or} \qquad \sigma^2_v \neq 0. \end{aligned}$$

For comparison of the models, a difference between restricted log-likelihoods [-2log(RLRT)] is calculated. To assess the significance of the observed test statistic, Coull, Schwartz & Wand (2001) recommend a parametric bootstrap.

4.3.2 Subject-specific curves

Penalised spline model in (4.1) fits the population mean function but the dependence between measurements derived from same subject is ignored. In a longitudinal context it is often sensible to assume measurements to be more similar within subjects. This is implemented through random-effect structure. The correspondence between a penalised spline smoother and the LME model allows us to easily extend the individual effects to smoothing methods. The origin of modelling subject-specific deviations as smooth curves goes back to Brumback and Rice (1998) who extended the group-level smoothing spline model for individual progesterone curves in menstrual cycle data.

Let us consider a model

$$y_{ij} = b_i + f(x_{ij}) + \epsilon_{ij} \qquad i = 1, ..., n \qquad j = 1, ..., m \tag{4.18}$$

that accounts for the individual variation with the added random intercept. Individuals have their own functions but the functions differ from one another only by their intercept. The individual functions can be made more flexible by extending the random effect structure. In most flexible models, subject-specific deviations are nonparametric functions.

Subject-specific smooth functions can be added to a spline model as penalised splines producing the the following model:

$$y_{ij} = f(x_{ij}) + g_i(x_{ij}) + \epsilon_{ij}, \qquad \text{subject to} \sum_{i=1}^{n} g_i = 0 \tag{4.19}$$

where $f(x)$ is the population mean function and $g_i(x)$ is the subject-specific deviation function. Since trajectories vary from one subject to another, it is natural to assume g_i's to be random functions. As a general feature of random components, g_i's are assumed to sum up to 0 (i.e. to have a zero mean) to reflect that $f(x)$ is the population mean function.

Similarly to the smooth functions in previous sections, k-th degree truncated power basis $\boldsymbol{\Psi_r(x)}$ in (3.8) is employed, with q knots $\delta_1, .., \delta_q$, to construct individual smooth functions g_i's. In general $q < K$, that is the number of knots used for the population mean function since individual curves are estimated with less data than the population curve (Wu & Zhang 2006).

Individual trajectories can be written as penalised splines:

$$g_i(x) = \boldsymbol{\Psi(x)'\mathbf{u_i}} = \sum_{s=1}^{k+1} u_{is} x^{s-1} + \sum_{c=1}^{q} v_{ic}(x - \delta_r)_+^k \tag{4.20}$$

where constraints are imposed on v_{ic}'s. The population-level spline model had two components: a fixed component to construct the polynomial part and a random component to model the deviations from this polynomial (see (4.1)). Equation (4.20) is also constructed from linear and nonlinear components but remarkable is that the both components are now random (Ruppert et al. 2003).

Coefficients have following distributions (Durban, Harezlak, Wand & Carrol 2005):

$$(u_{i1}, .., u_{ik}) \sim (0, \mathbf{\Sigma_u}), \qquad v_{ic} \sim (0, \sigma_v). \tag{4.21}$$

a detailed description of the matrix formulations and the estimation procedure can be found in the book of Wu & Zhang (2006). The authors describe subject-specific curve fitting through penalised splines but also through other smooth functions.

To describe model construction in practise, we use a linear penalised spline for both the population curve and the individual trajectories. The linear penalised spline model with subject-specific smooth functions is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{r=1}^{p} b_r (x_{ij} - \tau_r)_+ + u_{i1} + u_{i2} x_{ij} + \sum_{c=1}^{q} v_{ic} (x_{ij} - \delta_c)_+ + \epsilon_{ij} \tag{4.22}$$

By defining the following **Z**-matrix, the coefficient vector **b** and the corresponding covariance matrix **G**, (Durban et al. 2005), (4.22) can be implemented in R:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z_1} & \mathbf{X_1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{Zs_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{Z_2} & \mathbf{0} & \mathbf{X_2} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{Zs_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z_m} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z_m} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Zs_m} \end{pmatrix}$$

$$\mathbf{b} = [b_1, ..., b_p, u_{11}, ..., u_{m1}, u_{12}, ..., u_{m2}, v_{11}, ..., v_{mq}]'$$

$$\mathbf{G} = Cov(\mathbf{b}) = \begin{pmatrix} \sigma_b^2 \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & blockdiag(\mathbf{\Sigma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_v^2 \mathbf{I} \end{pmatrix}.$$

Test of the significance of subject-specific smooth curves is performed with the restricted likelihood ratio test. The model in (4.22) is compared to a model with a more simple random effects structure, that is a model with random intercepts and random slopes, u_{i1} and u_{i2}. Following models are compared:

$$\begin{aligned} M_0 &: y_{ij} = f(x_{ij}) + u_{i0} + u_{i0}(x) + \epsilon_{ij} \\ M_1 &: y_{ij} = f(x_{ij}) + g_i(x_{ij}) + \epsilon_{ij}, \qquad \text{where} \\ & g_i(x_{ij}) = u_{i0} + u_{i1}(x) + \sum_{c=1}^{q} \upsilon_{ic} (x_{ij} - \delta_c)_+ + \epsilon_{ij}. \end{aligned}$$

The corresponding hypothesis set-up is:

$$\begin{aligned} H_0 &: \sigma^2_{\upsilon_{ic}} = 0 \\ H_1 &: \sigma^2_{\upsilon_{ic}} > 0. \end{aligned}$$

Difference in models' likelihoods is calculated and simulations used to obtain the significance for the observed difference.

5 Application

In this chapter, the represented theory is applied to describe the weekly rhythm of weight measurements. We assume an underlying, regular weekly rhythm to be found in weight variation. Regarding earlier studies, such as Mattila et al. (2010), weight seems to increase at the weekend and decrease during weekdays. In this thesis, the interest lies in characterising the rhythm in specific subgroups; among subjects who have lost weight and among subjects who have gained weight, and testing the difference in the group's average curves. We will also explore the rhythm in the population level in our data set. There are four main hypotheses that are explored using the weight data.

1. Weekday effect

The first hypothesis involves testing whether weekdays have a linear effect on weight or whether the expected weight is constant during the week.

$H_0 : E[\text{weight}|\text{weekday}] = \beta_0$

$H_a : E[\text{weight}|\text{weekday}] = \beta_0 + \beta_1 \times weekday, \qquad \beta_1 \neq 0$

2. Nonparametric model versus parametric model

With the second hypothesis the form of the dependence between weight and weekdays is determined. A linear model is tested against its extension, a linear penalised spline model that corrects departures from the linearity.

$H_0 : E[\text{weight}|\text{weekday}] = \beta_0 + \beta_1 \times weekday$

$H_a : E[\text{weight}|\text{weekday}] = \beta_0 + \beta_1 \times weekday + \sum_{r=1}^{p} b_i(weekday - \tau_r)_+$

3. Individual effects

The group-average model of loss and gain groups is extended to account for the subject-specific effects. Three models nested to their random effects structure are fitted for both groups. The third hypothesis involves testing if the added variance component ($\sigma_{u_{ik}}$) differs significantly from zero.

$H_0 : \sigma_{u_{ik}} = 0$

$H_a : \sigma_{u_{ik}} > 0$

4. Group effect

The fourth hypothesis tests the significance of the group to the shape of the curves.

$H_0 : f(weekday_{loss}) = f(weekday_{gain}) = f(weekday_{maintain})$

$H_a : f(weekday_{loss}) \neq f(weekday_{gain})$ or $f(weekday_{loss}) \neq f(weekday_{maintain})$ or $f(weekday_{gain}) \neq f(weekday_{maintain})$

5.1 Materials and methods

Linear regression models and linear penalised spline models are used to find an appropriate curve estimates. Models are implemented and estimated using statistical software R version 2.11.1 (http://www.r-project.org/). Significance tests are conducted with the risk level α=0.05. Spline functions are formed using the truncated power basis and estimated in the linear mixed-effects framework with nlme library (Pinheiro et al. 2011). Through the LME environment likelihood based model testing can be used and construction of variability bands for the curve estimate is straightforward. We will explore the rhythm in the population level but due to the longitudinal nature of the data, we also expect great subject-to-subject variation. Subject-specific variation is modelled through random components of the LME model; we construct subject-specific penalised splines that allow flexible modelling of the individual trajectories as described in 4.3. The difference between groups' profile curves is analysed using an entirely parametric model, a cubic polynomial.

To my knowledge there was no R library available to fit penalised splines in the mixed model framework so that models could be extended to subject-specific curves and two smooth curves could be compared. I ended up to write the R-code myself. According to the theory, I exploited the nlme library for mixed-effects models to estime the smooth functions and for the inference procedure. Crucial aid for the code implementation was provided by Howel (2007), Ngo & Wand (2004) and Durban et al. (2005) who published R-code with their papers. Creating the code was challenging but contributed greatly to my understanding of the theoretical part of this thesis.

5.1.1 Data description

Data were collected from five different studies conducted by the VTT Technical Research Centre of Finland in co-operation with other institutes in 1996-2009. All studies involved health management by self-monitoring health-related variables, weight being one of them. For detailed descriptions of the studies see the following articles: Tuomisto et al. (2006), Mattila et al. (2008), Ahtinen et al. (2009), Kaipainen (2009). Table 5.1 presents the size and the duration of the studies. In all, the studies resulted in 7409 weight measurements.

In four of the studies, the subjects used a mobile phone application, Wellness

Table 5.1. The number of study subjects and the duration of the studies from which longitudinal weight measurements were collected.

	Study I	Study II	Study III	Study IV	Study V	Total
number of subjects	14	27	29	119	23	212
duration of the study	50-79 days	12 weeks	3 months	332-445 days	4 months	

Diary (Nokia, Helsinki, Finland), to record self-observations related to their wellbeing including exercise, weight and eating, and they received automatic graphical feedback based on these observations. In the remaining study subjects were given a wellness monitoring system for monitoring several health- related physiological and psychological variables simultaneously. Weight monitoring was done by weight scale which automatically transferred the information to a computer.

In each study, the subjects were advised to monitor their weight regularly among other health variables with weight itself being the primary outcome only in one of the studies. As a result of the different backgrounds of the studies, the study groups are heterogeneous and the amount of observed weight measurements varies widely between studies and between study subjects. The length of the monitoring time varied from a few weeks to almost a year. The most active users monitored their weight almost every day within a year whereas some of the subjects had only a few measurements.

5.1.2 Derivation of variables

Only measurements taken at least on seven sequential days were included in the analysis. The minimum monitoring time was required to last 14 days to describe any rhythms from at least two weeks. These requirements resulted in observations from 69 subjects, the length of subjects' time series varying from 15 to 330 days. Available background variables were not consistent within the studies. To describe the overall features of the data we use gender, age, and body mass index (BMI) which were available for most cases. Table in 5.2 represents summary statistics of relevant variables in the whole study group and in the subgroups of weight gainers and weight losers.

The response variable was derived by subtracting trend from time series of each subject to eliminate the effects of linear weight gain or linear weight loss. For example, a rapidly decreasing trend may cause weight to appear to be higher in the beginning of each week-length period and, therefore, the underlying "real" rhythm can be confounded with the trend-effected rhythm. The trend component, m_x, was estimated by applying a two-sided moving average filter (Brockwell & Davis 1996).

$$\hat{m}_x = (2q+1)^{-1} \sum_{j=-q}^{q} X_{x-j}, \qquad q+1 \leq t \leq n-q$$

where period d = 2q+1 = 7. Measurements were also normalised with respect

Table 5.2. Baseline characteristics of the study subjects.

		n	mean	median	sd	minimum	maximum
Gender							
	female	31					
	male	39					
Age	**		45,80	49,00	9,33	25,00	62,00
Weight in the beginning			79,91	80,00	14,41	53,00	113,40
BMI	***		27,69	28,04	3,45	20,02	33,46
	BMI < 25	11					
	BMI 25 -30	27					
	BMI > 30	15					
Number of monitored days		total = 4452	84	70	60,47726846	19	330
Number of monitored weeks		total = 736					
Relative weight change (%)		69,0000	-0,0145	-0,0090	0,0269	-0,1335	0,0187
Weight change group							
	weight loss	12					
	weight maintain	51					
	weight gain	6					

**missing from 14 subjects ***missing from 16 subjects

	Weight gain					Weight loss			
		n	mean	median	sd	n	mean	median	sd
Gender									
	female	0				6			
	male	6				6			
Age	*		43,50	45,50	11,09		45,27	49,00	10,50
Weight in the beginning			95,27	92,05	14,05		81,37	82,20	19,13
Relative weight change (%)			0,0123	0,0117	0,0042		-0,0585	-0,0454	0,0326
BMI	*		31,49	31,52	1,65	**	27,28	28,10	3,59
	BMI < 25	0				2			
	BMI 25 -30	1				6			
	BMI > 30	3				3			
Number of monitored days		total=498	110,17	68,50	108,85	total=1051	113,92	88,00	78,23
Number of monitored weeks		total = 82				total = 171			

*missing from 2 subjects **missing from 1 subject

to each subject's time series by centering and dividing by the corresponding standard deviation.

To study group differences, subjects were divided into three subgroups according to their relative weight change derived as a difference between the first and last two measurements. Weight change was required to be *linear* to characterise weight patterns during loss and gain periods. Possible yoyo-type subjects where weight increased and decreased in cycles were excluded from both of the gain and loss categories since against our hypothesis they may have both of the rhythms and, therefore, confound the results. The categories of the corresponding group variable are following:

$loss$ = subjects who lost weight by more than three percent (3%)
$gain$ = subjects who gained weight by more than one percent (1%)
$maintain$ = subjects whose weight change varied
from minus three pro cent to one percent (-3% - 1%)

The weight limit for the gain group was set lower than the loss group for the reason that the gain group tended to remain reasonably small. Observing weight gain lowers the motivation for self-monitoring and those subjects are likely to stop monitoring. Thus, they had fewer data available.

For the weekday factor, levels were labelled as follows: 1=Sunday, 2=Monday, 3=Tuesday, 4=Wednesday, 5=Thursday, 6=Friday, 7=Saturday.

5.2 Analysis of the weekly rhythm of weight

The starting point for the analysis is presented in Figure 5.1. It describes weight profiles for three groups; weight losers, weight gainers and weight maintainers derived as an average weight of each weekday. According to Figure 5.1, there seems to be dependence between weight and weekdays and deviation from the linearity can be expected. Profile curves seem to have a similar shape but in vertical direction weight loss group has greater variation. This indicates a stronger rhythm that will be tested against other subgroups.

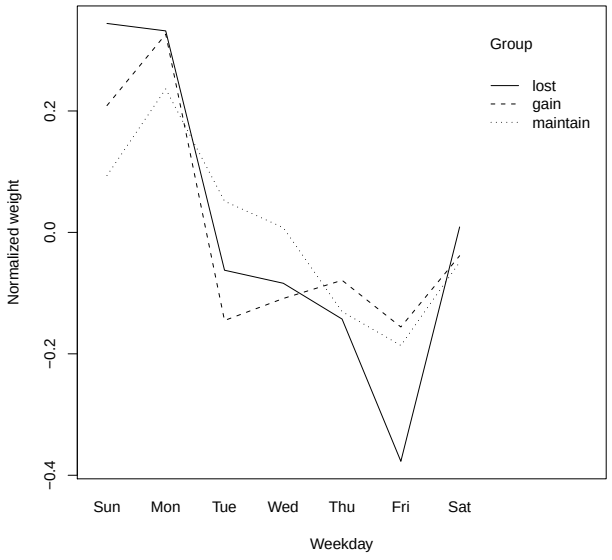

Figure 5.1. Weight profiles for the three groups derived as an average weight on each weekday

Table 5.3 shows frequencies for each weekday for the week's minimum and maximum weight. For the weight loss group, the week's maximum weight is most frequently on Sunday and in 62% of the cases week's maximum weight can be found either on Sunday or Monday. Correspondingly, minimum weight is reached mostly on Friday or Saturday, in 58% of the cases. This supports findings from the Figure 5.1 and is consistent with the results from earlier studies. In the weight gain group there is more variability and the phenomenon is not as obvious. Frequencies are distributed on several days with equal numbers. Minimum weight is most frequently on Sunday and maximum weight on the following day on Monday. But the difference from frequencies of other days is not apparent. For maximum weight there are high frequencies also in the beginning of the week and on weekends but on the other hand the same days gather frequencies for the week's minimum weight.

Table 5.3. Frequencies on weekdays for the week's minimum and maximum weight.

Weight losers					Weight gainers				
Weekday	minimum	maximum	min%	max%	Weekday	minimum	maximum	min%	max%
Sun	5	40	0,0505	0,4040	Sun	17	9	0,3148	0,1667
Mon	4	22	0,0404	0,2222	Mon	8	12	0,1481	0,2222
Tue	7	8	0,0707	0,0808	Tue	7	9	0,1296	0,1667
Wed	13	7	0,1313	0,0707	Wed	8	5	0,1481	0,0926
Thu	12	7	0,1212	0,0707	Thu	4	7	0,0741	0,1296
Fri	36	6	0,3636	0,0606	Fri	5	4	0,0926	0,0741
Sat	22	9	0,2222	0,0909	Sat	5	8	0,0926	0,1481

5.2.1 Weekday effect

The analysis is started by exploring the first hypothesis which involves testing whether the weekday has effect on weight's variation. The test is conducted for the whole data set but also separately for both groups under interest; subjects who have lost weight and subjects who have gained weight. The model formulations under H_0 and H_a are:

$$M_0 : weight_{ij} = \beta_0$$
$$M_1 : weight_{ij} = \beta_0 + \beta_1 weekday_{ij},$$

where i denotes the index of a subject and j is the measurement of the ith subject. The alternative hypothesis denotes a linear regression model to be fitted. To assess the significance of the covariate, a residual sum of squares based F-test (Zuur, 2009) is used. The three models under H_a are denoted as $M1_{all}$, $M1_{loss}$ and $M1_{gain}$ and the corresponding models under H_0 are $M0_{all}$, $M0_{loss}$ and $M0_{gain}$.

5.2.2 Parametric model versus nonparametric model

The analysis is continued by exploring the second hypothesis to assess the shape of the weekly rhythm. The linear effect of a weekday is modelled as a null model. Its general alternative, a nonparametric model ($M2$) under the H_a is

$$weight_{ij} = f(weekday_{ij}) + \epsilon_{ij}, \qquad 1 \leq j \leq n_i, \qquad 1 \leq i \leq 69 \tag{5.1}$$

where $weight_{ij}$ and $weekday_{ij}$ are jth measurements of normalised, trend deleted weight and weekday for subject i. f is the smooth function and ϵ_{ij} denotes measurement error. Again, the population-level curve as well as the group-level curves are fitted as separate models $M2-all$,$M2_{loss}$ and $M2_{gain}$. To construct the smooth function $f(weekday)$, a penalised spline is used, where the polynomial part is taken to be linear and truncated line functions $(weekday_{ij} - \tau_K)_+$ handle departures from the linear basis. Coefficients $\mathbf{b} = [b_1, .., b_6]$ are subject to the penalisation to avoid overfitting.

$$f(weekday) = \beta_0 + \beta_1 weekday_{ij} + \sum_{k=1}^{6} b_k (weekday_{ij} - \tau)_+. \tag{5.2}$$

The number and the locations of τ's are defined following the presentation in 3.1. With small adjustments, following cut points are used: $[1.5, 2.5, 3.5, 4.5, 5.5, 6.5]$. To form the spline model to correspond to the standard mixed-effects model structure, we define the model matrices as follows:

(5.3)

$$\mathbf{X} = \begin{pmatrix} 1 & weekday_{11} \\ \vdots & \vdots \\ 1 & weekday_{1i} \\ \vdots & \vdots \\ 1 & weekday_{j1} \\ \vdots & \vdots \\ 1 & weekday_{ji} \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} (weekday_{11} - 1.5)_+ & \dots & (weekday_{11} - 6.5)_+ \\ \vdots & \ddots & \vdots \\ (weekday_{1n_1} - 1.5)_+ & \dots & (weekday_{1n_1} - 6.5)_+ \\ \vdots & \ddots & \vdots \\ (weekday_{i1} - 1.5)_+ & \dots & (weekday_{i1} - 6.5)_+ \\ \vdots & \ddots & \vdots \\ (weekday_{im} - 1.5)_+ & \dots & (weekday_{im} - 6.5)_+ \end{pmatrix}$$

and parameter vectors $\boldsymbol{\beta} = [\beta_0, \beta_1]'$ and $\mathbf{b} = [b_1, b_2, b_3, b_4, b_5, b_6]'$. Thus,the model in (5.1) is convenient for implemention with lme function.

To test the significance of the penalised spline model, $M2$ is compared against the parametric model $M1$ which is equivalent to testing whether the curve is a linear ($M1$) or whether there is a need for a smoother fit ($M2$). The corresponding hypothesis is:

(5.4) $$H_0 : \sigma_b^2 = 0 \qquad \text{versus} \qquad H_1 : \sigma_b^2 > 0.$$

Testing problem is non-standard because the parameter vector is on the boundary of the parameter space ($\sigma_b = 0$) under H_0. The exactRLRT -function with 10000 simulations is used to obtain a p-value for the observed difference in the model's likelihoods.

As an example, R-output for the linear penalised spline model $M2_{loss}$ is printed in the following. Other model outputs and R-code are in Appendix A.

```
Linear mixed-effects model fit by REML
 Data: gdata
   logLik
  -1467.944
Random effects:
 Formula: ~-1 + Z | group
 Structure: Multiple of an Identity
               Z1        Z2        Z3        Z4        Z5        Z6  Residual
StdDev: 0.3310911 0.3310911 0.3310911 0.3310911 0.3310911 0.3310911 0.9697364
Fixed effects: y ~ x
                 Value Std.Error   DF    t-value p-value
(Intercept)  0.3756135 0.2428374 1049  1.5467696  0.1222
x           -0.0042283 0.1973397 1049 -0.0214266  0.9829
Number of Observations: 1051
Number of Groups: 1
```

REML is used for the estimation method since models include random components. With specific notations for the lme function parameters, the covariance

matrix for $\mathbf{Z}$ is determined to appear as defined in (5.3); $\mathbf{G}$ should be a multiple of an identity since data is clustered to subjects. The intercept is removed from the random structure to make the random component to correspond to the definition in (5.2). Model coefficients ($\boldsymbol{\beta}$) and ($\mathbf{b}$) do not have any particular meaning by themselves since the actual fit is determined by a combination of the fixed effects and the random effects. We use the coefficients to construct a plot that describes the fitted values with their variability bands. Fitted values are extracted by summing up the ML estimates of $\beta's$ and and REML estimates of b's.

5.2.3 Individual effects

Due to the longitudinal nature of the data, fitted models are extended to account for the within-subject dependence of measurements. The analysis is done only to the subgroups.

Figures 5.2 and 5.3 show individual average curves in the weight loss group and in the weight gain group. Clearly, the shape of the curves deviate from one another. This points to the extension of models that account for the individual effects.

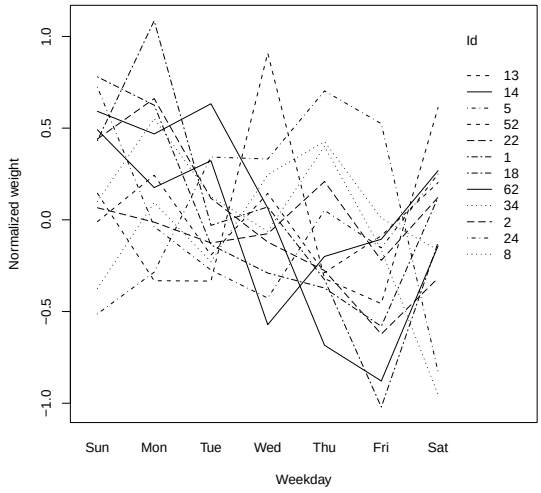

Figure 5.2. Individual average curves for loss group's subjects.

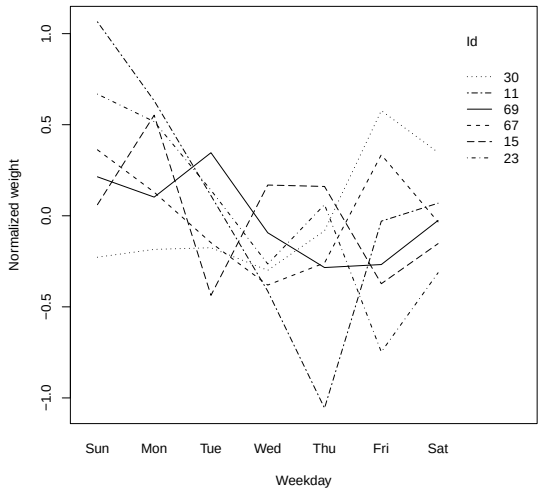

Figure 5.3. Individual average curves for gain group's subjects.

Three models with nested random-effects structure are fitted:

$$M3: \quad weight_{ij} = f(weekday_{ij}) + u_{i1} + \epsilon_{ij}$$

$$= \beta_0 + \beta_1 \times weekday_{ij} + \sum_{k=1}^{6} b_k(weekday_{ij} - \tau_k)_+ + u_{i1} + \epsilon_{ij}$$

$$M4: \quad weight_{ij} = f(weekday_{ij}) + u_{i1} + u_{i2} \times weekday_{ij} + \epsilon_{ij}$$

$$= \beta_0 + \beta_1 \times weekday_{ij} + \sum_{k=1}^{6} b_k(weekday_{ij} - \tau_k)_+ + u_{i1} + u_{i2} \times weekday_{ij} + \epsilon_{ij}$$

$$M5: \quad weight_{ij} = f(weekday_{ij}) + g_i(weekday_{ij}) + \epsilon_{ij}$$

$$= \beta_0 + \beta_1 \times weekday_{ij} + \sum_{k=1}^{6} b_k(weekday_{ij} - \tau_k)_+ +$$

$$u_{i1} + u_{i2} \times weekday_{ij} + \sum_{c=1}^{4} v_{ic}(weekday_{ij} - \tau_k)_+ + \epsilon_{ij}$$

where $f(weekday)$ is the most parsimonious population-average curve obtained from the first and second hypothesis set-ups. Individual effects are modelled by expanding the random-effects structure. The components u_{i1} and u_{i2} are random intercepts and slopes for the ith subject, and g_i is the subject-specific deviation curve of the ith subject. To test the significance of the individual effects involves testing following hypotheses including relevant variance parameters for the models.

$$H_0 : u_{i1} = u_{i2} = v_{ik} = 0$$
$$H_a : u_{i1} > 0 \text{ or } u_{i2} > 0 \text{ or } v_{ik} > 0$$

Because models are nested, likelihood based tests can be used to find an appropriate random structure. To find out the distribution of the test statistic, a bootstrap with 10,000 repetitions is used. The corresponding p_{boot}-value determines the significance of the statistic.

5.2.4 Group effect

According to the fourth hypothesis, the shape of the curves is compared. The factor-by-curve interaction model, presented in Section 4.4.4, was constructed to fit smooth curves for both groups simultaneously. However, the estimation in the lme environment did not converge and estimates for the model coefficients were not obtained.

We took another approach and carried out the testing with polynomials. The weekday variable has only seven unique values indicating that a polynomial can capture the curvature of the weekly rhythm. The advantage of splines over polynomials is through their feature of stable extrapolation but because we explore the profile curves, prediction is not the case.

To test the group effect, following hypothesis set-up is tested.

$$H_0 : weight_{ij} = f(weekday_{ij})$$
$$H_1 : weight_{ij} = f(weekday_{ij}) \times group_{iz}$$

where the group variable has three levels: *loss*, *gain* and *maintain*. Two models, $M6$ and $M7$, are fitted with different structures for $f(weekday)$. For $M6$ the stucture for $f(weekday)$ is determined as a simple linear regression model. $M7$ fits the most parsimonious structure that captures the curvature in weight variation adequately. Polynomials of second, third, and fourth degree and the corresponding penalised spline models were fitted and compared. This resulted in an adequate structure to be obtained with a cubic polynomial. For $M7$ $f(weekday)$=$\beta_0 + \beta_1 \times weekday + \beta_2 \times weekday^2 + \beta_3 \times weekday^3$.

Group effect is assessed using F-test. The null model for $M6$ is actually $M1_{all}$. For $M7$, the null model, is
$weight_{ij} = \beta_0 + \beta_1 \times weekday_{ij} + \beta_2 \times weekday_{ij}^2 + \beta_3 \times weekday_{ij}^3 + \epsilon_{ij}$. The comparison between the levels of the grouping variable (i.e. *loss*, *gain*, *maintain*) is derived using the gain group as a reference level to which the other levels are compared. The significance of the corresponding group coefficients is tested using the t-test.

6 Results

6.1 Weekday effect

The F-statistic for the comparison of population-level models M_{all} is 68.352 and the corresponding p-value ≈ 0. Respectively for M_{loss} and M_{gain} p-values are p=3.756e-09 and p=0.0195. In all cases, H_0 is rejected. Derived model coefficients, β_1's are following: -0.061647 for M_{all}, -0.09105 for M_{loss} and -0.05817 for M_{gain}.

6.2 Parametric model versus nonparametric model

For the comparison of $M2$ and $M1$, the derived likelihood ratio in the population level is 23.2032 and its simulated p-value ≈ 0. For $M2_{loss}$ RLRT is 11.5081 and for $M2_{gain}$ 1.0347. The corresponding simulated p-values are p=3e-04 and p=0.0826. H_0 is rejected in population level and in loss group. For the gain group, H_0 is accepted but there for H_0 is not strong.

Fitted curves with variability bands are plotted in Figure 6.1. On the left the fit is located with jittered values of the observed weight on each weekday. On the right the same curves are plotted but with smaller range of y axis to describe the shape of the curves.

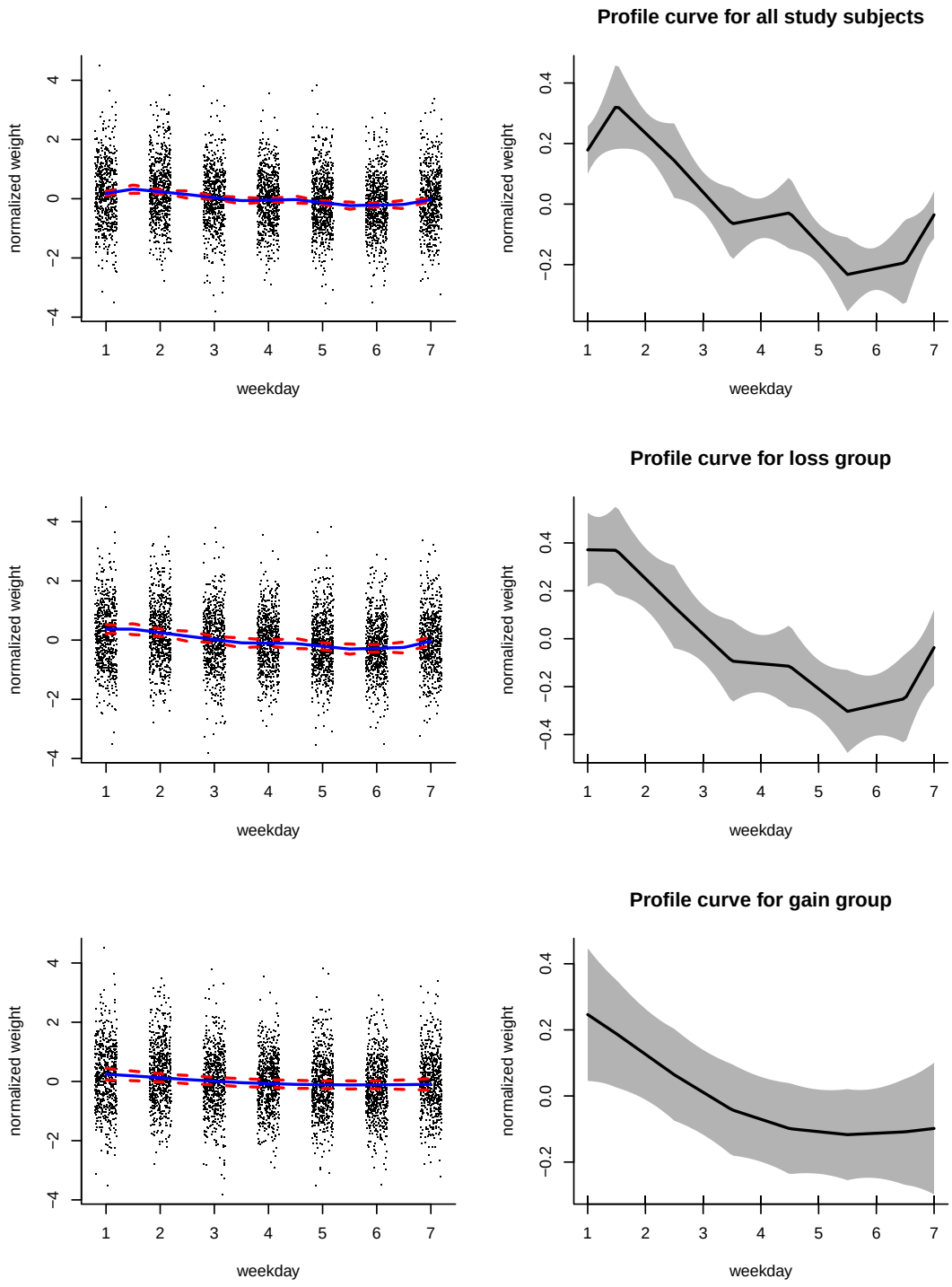

Figure 6.1. Fitted spline models for normalised weight measurements of the three models $M2_{all}$, $M2_{loss}$ and $M2_{gain}$

6.3 Individual effects

Table 6.1 gathers restricted log-likelihoods of models $M0$,$M2-M5$ for loss and gain groups. It is notable that log-likelihoods are the same under $M2$ and $M3$ and under $M4$ and $M5$ for both groups. Thus, the interest lies in comparing models M2 and M4. The hypothesis set-up is as follows:

$$H_0: \quad weight_{ij} = \beta_0 + \beta_1 \times weekday_{ij} + \sum_{k=1}^{5} b_k(weekday_{ij} - \tau_k)_+ + \epsilon_{ij}$$

$$H_a: \quad weight_{ij} = \beta_0 + \beta_1 \times weekday_{ij} + \sum_{k=1}^{5} b_k(weekday_{ij} - \tau_k)_+ + u_{i1} + u_{i2} \times weekday_{ij} + \epsilon_{ij}$$

Table 6.1. Log-likelihoods and estimated variance components of the fitted models, M0, M2-M5.

	Log-likelihood	Degrees of freedom	Variance components			
Lost -group						
M0_loss	-1485.27	2	σ_ε=0.99474			
M2_loss	-1467.944	4	σ_ε=0.94039	σ_b=0.109621		
M3_loss	-1467.944	5	σ_ε=0.94039	σ_b=0.109649	σ_u=7.8611e-10	
M4_loss	-1459.765	7	σ_ε=0.91394	σ_b=0.125797	σ_u=[0.11858,0.00731965]	
M5_loss	-1459.765	8	σ_ε=0.91394	σ_b=0.125791	σ_u=[0.11857,0.00731988]	σ_v=3.13157e-10
Gain -group						
M0_gain	-713.7517	2	σ_ε=1.03107			
M2_gain	-715.0887	4	σ_ε=1.01532	σ_b=0.0059012		
M3_gain	-715.0887	5	σ_ε=1.01532	σ_b=0.0059012	σ_u=8.7658e-10	
M4_gain	-711.0635	7	σ_ε=0.98533	σ_b=0.0060132	σ_u=[0.172494,0.0105764]	
M5_gain	-711.0635	8	σ_ε=0.98533	σ_b=0.0060132	σ_u=[0.172489,0.0105773]	σ_v=1.18838e-10

Table 6.2 shows the simulated p-values for the model comparisons. H_0 is rejected in both groups.

Table 6.2. P-values for RLRT using parametric bootstrap.

	logLik under H0	logLik under H1	-2 log(RLRT)	$\text{p-value}_{\text{boot}}$
Gain -group	-1467,99	-1459,765	8,050327	0,007
Loss -group	-715,0887	-711,0635	16,35714	0

6.4 Group effect

For the $M7$ F-statistic is 1.2192 and the p-value is 0.3003. For $M6$ the F-value is 1.6402 and corresponding p=value 0.1081. In both models null hypothesis is accepted.

In $M7$, the model coefficients for loss and maintain groups are not significantly different from the reference level. All p-values for the interaction terms are over the risk level. No difference between groups is detected.

For the $M6$ the R-output for model coefficients is following

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.210372   0.041367   5.086 3.82e-07
maintain           0.019048   0.108647   0.175   0.8608
loss               0.152078   0.079884   1.904   0.0570
weekday           -0.051699   0.009215  -5.611 2.14e-08
maintain:weekday  -0.006471   0.024265  -0.267   0.7897
loss:weekday      -0.039347   0.017932  -2.194   0.0283
```

The slope and the interaction term of the loss group deviate significantly from zero.

7 Discussion

We discussed a relatively new method where representation of linear mixed-effects model is exploited in nonparametric curve fitting with penalised splines. This semiparametric regression technique has gained popularity among scientist in different fields. After releasing the first book-length publication *Semiparametric Regression* in 2003, Ruppert, Wand and Carrol (2009) discussed the development of semiparametric regression; the authors reviewed over 300 articles in 2003-2007 that had benefited from this approach. The presented theory was applied to explore weekly rhythm of weight from self-monitored weight measurements data. Modelling the rhythm with contiguous curves differed methodologically from the approach of previous studies by Tuomisto et al. (2006) and Mattila et al. (2010). By describing the rhythm through splines we were able to account for the effect of previous days in the shape of a curve rather than conduct multiple tests for different weekdays.

The choice of penalised splines was not obvious but it has advantages over other methods such as growth curve modelling (Pan & Zhang 2002), that requires balanced data, and it is beneficial in several application areas through its simplicity. Spline models allow data to be unbalanced which is a huge advantage specifically in longitudinal follow-up studies where subjects enter and withdraw in studies at different time points. On the other hand we wanted to present an alternative approach from polynomials. After all, splines are build on a polynomial basis that makes the shrinkage and comparison to the polynomial models straightforward as was seen in 5.2.4.

7.1 Finding the weekly rhythm

According to the previous studies, body weight was expected to vary within the week; weight was expected to increase during weekends and decrease during weekdays. The first two hypotheses involved finding the appropriate form of the dependence between weight and weekdays. Consistent with previous studies, we found that weekday has a significant effect on weight. Small p-values were derived for first models $M1_{all}$, $M1_{loss}$ and $M1_{gain}$ meaning a significant improvement from the null model where weight was expected to be constant during the week. With this data set, we obtained strong evidence that weight is not constant but varies within the week. The estimates for β_1s were negative in all models meaning that weight is higher with small values of the weekday variable (i.e. on Sunday and on Monday) and decreases during the week. The

smallest β_1 coefficient was detected in the loss group meaning a faster decrease of weight during the week.

Developing the models further, we found that for the population level and for the loss group the linear model was not adequate but there was a need for a more complicated structure to capture weight variation within a week. Corrections to the linearity through the linear penalised spline models were necessary. Figure 6.1 shows the fitted profile curves. For both groups weight is at its highest in the beginning of the week. It decreases during weekdays but shows a slight, but statistically significant, increase in the end of the week, on Fridays and on Saturdays. The observed weight gain on weekends is sensible and is supported by the findings on the changes in eating habits (De Castro 1992) as well as by lowered physical activity reported by Buchowski et al. (2004) and Treuth et al. (2007). In addition, as detected in Figure 6.1, weight is significantly higher on Sundays and Mondays than on Thursdays and on Fridays. To reach the highest level, it makes sense that weight starts to increase as early as Friday rather than increasing rapidly in just one day.

For the weight gain group, the p-value for the linear spline model, was slightly over 5% risk level. The linear model was adequate in this case but there is tendency that more curvature for the fit could be needed. For this reason we continue the model development in 5.2.3 with the linear penalised spline model also for the loss group. However, a similar pattern to the other groups was detected and this can be seen in Figure 6.1. There is a peak on Sunday after which weight starts to decrease until Thursday and after that plateaus. But what is remarkable is that the fitted values are gathered in the smaller range of the y-axis, (approx. -0.05 to 0.025) whereas the variability bands are wider. Wide bands indicate that measurements are distributed to a wider range on each weekday. Obviously, weight varies more since subjects belonging to the gain group are heavier (see Table 5.1) than subjects in the loss group. On the other hand, this may indicate that subjects are lacking a clear rhythm but weight varies without any regularity. The lack of the rhythm is also reported by Mattila et al. (2010) who analysed the rhythm of subjects unsuccessful in weight loss. Another reason that affects the fit, is the small group size. In the loss group, for example, there is twice as many measurement as in the gain group. This shrinks the variance.

7.2 Longitudinal nature of the data

Due to the longitudinal nature of the data, group-level fits in the loss and gain groups were extended to take into account the individual effect. This considered the third hypothesis.

The inherited results were interesting, deviating from the initial expectations. For both subgroups, the addition of a subject-specific slope did not actually affect the goodness of the fit but the log-likelihood remained same as it was in $M2_{loss}$ and $M2_{gain}$. The stability of the REML statistic can be explained

by the estimates of the additional variance component. Components, u_1's, for both models were practically zero.

$M4_{loss}$ and $M4_{gain}$ were extensions of $M3_{loss}$ and $M3_{gain}$ that allowed linear deviation for individuals from the group-average curves. This modification lowered the REML from -1485.27 to -1459.765 in the weight loss group and from -713.7517 to -711.0635 in the weight gain group. The changes were statistically significant. This means that allowing individual linear departures improved models significantly. The most flexible model $M5_{loss}$ and $M5_{gain}$, where individual trajectories were allowed to be nonparametric function did not improve the fits from $M4$ in either group. In fact, log-likelihoods remained the same.

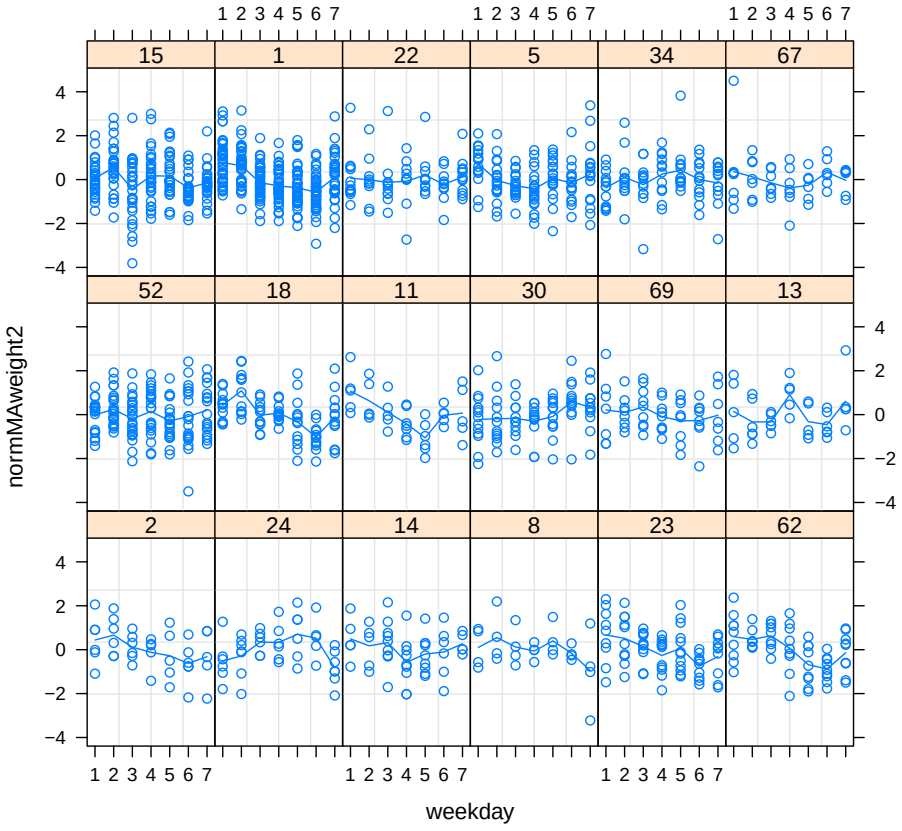

Figure 7.1. Subject-specific measurements and their average curves as a function of a weekday

The stability in log-likelihoods can partly be explained by the nested structure of the data. The data is nested in two levels since it consists of multiple monitoring weeks for each individual. Figure 7.1 illustrates the situation. Allowing subject-specific intercepts and slopes does not account for all the variability in specific time points (i.e. weekdays) but they actually model subject-specific individual average curves. In classical shape of the data, subjects have one mea-

surement at one time point. Therefore, the addition of individual effects have a direct effect on log-likelihoods. Ignoring this can confound results (Fitzmaurice et al. 2004). However, in our case the REML statistic remained the same under $M2$ and $M3$ and under $M4$ and $M5$ models due to the nested structure of the data.

The most parsimonius fit for both groups was obtained by allowing individual linear deviations. Despite this, the unexplained variance component ϵ still remained large. To improve models further, we could have added week-specific curves making random effects to be nested in two levels. But, according to our hypothesis we assumed that an underlying trend that repeats itself in the long run would ultimately be found. Thus, modelling week-specific curves would not have enhanced information to suit our requirements. Primarily we were interested in exploring the population-level profile which could be generalized and utilised in DSS development.

7.3 Group effect on the weekly rhythm of weight

From Figure 5.1 it can be seen that subgroups have similar weight patterns but in the loss group the variation seems to be greater in the vertical direction. According to the fourth hypothesis, the weekly rhythm was tested to determine whether it differed between the groups. We found that the addition of the group variable did not provide any additional information but the weekly rhythm can be modelled through one average curve for the whole data.

However, interesting results were obtained from model coefficients of the linear regression model, $M6$. Even though the overall effect of the group was not significant, slight indicators of the difference between loss and gain groups can be seen. The intercept and the slope of loss group significantly differed from the corresponding estimates of the reference group, which was the weight gain group. Subjects who lost weight had higher weights in the beginning of the week (on Sundays) than subjects who gained weight. The slope was negative and significantly smaller than in the gain group indicating faster decline of weight during the week.

The indicator from the possible group effect was inherited from a simple linear regression model whereas the most parsimonious showed no difference. Further analysis should be conducted with bigger group sizes to establish the possible difference in weight's weekly rhythm between subjects who lose and gain weight. For future research, categorisation of subjects need to be revised to raise the group sizes. The inference of the groups' differences was based on the information derived from 18 subjects, six subjects belonging to the weight gain group and 12 subjects belonging to the weight loss group.

8 Conclusions

In this thesis, we discussed a relatively new method where nonparametric smoothing is incorporated to a widely-known parametric environment, linear mixed-effects models. The fusion is obtained by partitioning a penalised spline into specific fixed and random components. The method is generally known as semiparametric regression method and it enjoys a great variety of extensions.

The presented theory was applied to the longitudinal weight measurement data to explore the weekly rhythm of weight. A rhythm that shows weight to be higher after weekends and decrease during weekdays was found. Furthermore, in the whole population level and in the loss group, there seems to be slight increase at the end of the week, during Fridays and Saturdays. However, no statistical difference was detected in the shape of profile curves of the three groups: loss, gain and maintain.

By analysing weight gain and weight loss groups separately, 82 and 171 weeks were explored, from six and twelwe study subjects respectively. Small group sizes were a weakness of this study. The derived results are statistically significant but generalisation for applications should be considered. The amount of weight measurements used for the separate analyses of subgroups were 498 and 1051. It is remarkable is that they were derived only from 18 subjects. Originally we had data from 7409 measurements from 69 subjects but requirements for linear weight development excluded most subjects from the subgroup analyses. Another challenge we faced was to obtain data from subjects who gained weight. Self-monitored data that is based on volunteer monitoring is difficult to obtain, specifically from subjects who have gained weight. Even without limitation of linear weight growth we had only 11 subjects whose weights increased in the end of the study.

For future research, the categorisation of subjects needs to be revised. In this thesis we expected weight to behave in a certain way according to the group to which a subject belonged to. Alternatively, time series could be segmented to the periods of weight loss and weight gain and be categorised to the corresponding groups. The segmentation would require a sophisticated algorithm that extracts time periods and categorizses those in relevant groups. The major challenge would, however, apply the methodological side; obviously the segmented periods are not independent and on the other hand one subject can belong to both groups. Construction of the covariance matrix requires a deeper familiarisation with the covariance models.

9 Acknowledgements

In conclusion, I want to say a few words to people who have contributed to this thesis.

I want to thank my thesis advisor Professor Erkki Liski, who introduced me to this interesting but challenging world of semiparametric models. I am also sincerely grateful to Arto Luoma for helping me to overcome barriers with the R-core implementation. Without his help these analyses would not look like they do now.

From the VTT Technical Research Centre of Finland, I want to express my gratitude to my advisors, Adjunct Professor Mark van Gils and Miikka Ermes for their invaluable guidance throughout the process. I wish to thank Elina Mattila for sharing her expertise in analysing weight behaviour and for answering my endless stream of questions. I also want to thank Jaakko Lähteenmäki for providing me with this possibility to work in the Care4me project throughout the year.

Finally, I want to thank my Toni for the support you gave me during the past year. Your contribution to the completion of this thesis was more important than you think.

Bibliography

Ahtinen, A., Mattila, E., Väätänen, A., Hynninen, L., Salminen, J., Koskinen, E. & Laine, K. (2009), "User experience of mobile wellness applications in health promotions", *3rd International Conference on Pervasive Computing Technologies for Healthcare*, London U.K.

Brockwell, P. J. & Davis, R. A. (1996) *Introduction to Time Series and Forecasting*, New York: Springer-Verlag.

Brumback, B. A. & Rice, J. A. (1998), "Smoothing splines models for the analysis of nested and cross samples of curves", *Journal of the American Statistical Association*, 93, 961–976.

Buchowski, M. S. & Acra, S., Majchrzar, K. M., Sun, M. & Chen K. Y. (2004), "Patterns of physical activity in free-living adults in the Southern United States", *European Journal of Clinical Nutriton*, 58,828–837 961–976.

Casella, G. & Berger, R. L. (1990), *Statistical Inference*, Wadsworth and Brooks, Pacific Grove, CA.

Coull, B. A., Ruppert, D. & Wand, M. P. (2001), "Simple incorporation of interactions into additive models", *Biometrics*, 57, 539–45.

Coull, B. A., Schwartz, J. & Wand, M. P. (2001), "Respiratory health and air pollution: Additive mixed model analyses", *Biostatistics*, 2, 337–49.

Crainiceanu, C. M., Ruppert, D., Cleaskens, G., Wand, M. P (2002), "Likelihood Ratio Tests of Polynomial Regression Against a General Nonparametric Alternative", *Biometrica*, 91, 35-42.

Crainiceanu, C. M. & Ruppert, D. (2004a), "Likelihood ratio tests for goodness-of-fit of a nonlinear regression model", *Journal of Multivariate Analysis*, 91, 35-42.

Crainiceanu, C. M. & Ruppert, D. (2004b), "Likelihood ratio tests in linear mixed models with one variance component", *Journal of the Royal Statistical Society*, B66, 165-185.

Crainiceanu, C. M. & Ruppert, D. (2004c), "Restricted likelihood ratio tests in nonparametric longitudinal models", *Statistica Sinica*, 14, 713–729.

Crainiceanu, C., Ruppert, D., Claeskens, G., & Wand, M. P. (2005), "Exact Likelihood Ratio Tests for Penalized splines", *Biometrika*, 92, 91–103.

Daubechies, I. (1992), "Ten Lectures in Wavelets", *Society for Industrial and Applied Mathematics*, Philadelphia:PA.

Diggle, P., Liang, K., & Zeger S. (1994), *Longitudinal Data Analysis*, Oxford University Press, Oxford.

Doll, H. A., Petersen, S. E. & Stewart-Brown, S. L. (2000), "Obesity and physical and emotional well-being: associations between body mass index, chronic illness,

and the physical and mental components of the SF-36 questionnaire", *Obesity Research*, 8 160–170.

de Castro, J. M. (1991), "Weekly rhythms of spontaneous nutrient intake and meal pattern of humans", *Physiological Behaviour*, 50, 729–38.

Durban, M., Harezlak, J., Wand M. P & Carroll, R. J. (2004), "Simple fitting of subject-specific curves for longitudinal data", *Statistics in Medicine*, 24, 1153–1167.

Eilers, P. H. & Marx, B. P. (1996), "Flexible smoothing with B-splines and penalties", *Statistical Science*, 11, 89–121.

Eubank, R. (1999), *Nonparametric regression and spline smoothing*, 2nd Edition, Texas: Marcel Dekker.

Faraway, J. J.(2006), *Extending the Linear Model with R*, Chapman & Hall/CRC.

Green, P. J. & Silverman B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, Boca Raton: Chapman and Hall.

Greven, S. (2008), *Non-Standard Problems in Inference for Additive and Linear Mixed Models*, PhD Thesis, Cuvillier: Verlag.

Fitzmaurice, G. M., Laird, N. N. & Ware, J. H. (2004), *Applied Longitudinal Analysis*, New Jersey: Wiley.

Hastie, T. & Tibshirani, R. (1993), "Varying-coefficients models", *Journal of Royal Statistical Society, Series B*, 55, 757-796.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hastie, T., Tibshirani, R. & Friedman (2001) *The Element of Statistical Learning*, New York: Springer-Verlag.

Howell, J. R. (2007) "Analysis of smoothing splines as implemented in LME() in R", *A project submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Master of Science*, Brigman Young University, Utah.

Härdle, W., Kerkyacharian, G., Picard, D. & Tsybakov, A. (1998) *Wavelets, Approximation and Statistical Applications*, New York: Springer-Verlag.

Kaipainen, K. (2009) *Design and implementation of web-based cognitive behavioural therapy intervention methods for management of mental wellbeing*, master's thesis, University of Tampere, Department of Computer Science. Available from Internet: http://tutkielmat.uta.fi

Keele, L. (2008) *Semiparametric Regression*, West Sussex: Wiley

Greven, S. (2007), *Applied Longitudinal Analysis*, New Jersey: Wiley.

Laird, N. M., & Ware, J. H. (1982), "Random effects models for longitudinal data" *Biometrics*, 38, 963–974.

Lappalainen, R., Pulkkinen, P., van Gils, M., Pärkkä, J. & Korhonen, I. (2005), "Long-term Self-monitoring of Weight: A Case Study", *Cognitive Behaviour Therapy*, 34, 108–114.

Mattila, E., Lappalainen, R., Pärkkä, J., Salminen J. & Korhonen, I. (2010), "Use of mobile phone diary for observing weight management and related behaviours", *Journal of Telemedicine and Telecare*, 00, 1-5.

Mattila, E., Pärkkä, J., Hermersdorf Marion., Kaasinen J., Vainio J., Samposalo K., Merilahti J., Kolari J., Kulju M., Lappalainen, R., & Korhonen, I. (2008), "Mobile Diary for Wellness Management - Results on Usage and Usability in Two User Studies", *IEEE Transaction on Information Technology in Biomedicine*, 12(4), 501-512.

McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008), "2nd ed.", *Generalized, linear, and Mixed Models*, New Jersey: Wiley.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. a., Marks, J. S.,(2003), "Prevalence of Obesity, Diabetes and Obesity-Related Health Risk Factors, 2001", *The Journal of the American Medical Association*, 289(1), 76-79.

Monk, R. H., Buysse, D. J., Rose, L. R., Hall, J. A. & Kupfer, D. J. (2000), "The sleep of healty people - a diary study", *Chronobiology Internationa*, 17(1), 49–60.

Ngo, L., & Wand, M. P. (2004), "Smoothing and Mixed Models Software", *Journal of Statistical Software*, 9, 1-54. Available from Internet: http://www.uow.edu.au/ mwand/papers.html

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. & The R core team (2011),"nlme 3.1-100" *R package* http://cran.r-project.org.

Pinheiro, J.C. & Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag

R Development Core Team (2008), "R: A Language and Environment for Statistical Computing", *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Robinson, G. K. (1991), "The BLUP Is a Good Thing: The Estimation of Random Effects", *Statistical Science*, 6, 15-32.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2009),"Semiparametric Regression during 2003-2007" *Electronic Journal of Statistics*, 3 ,1193-1256. Available in Internet: http://www.uow.edu.au/ mwand/publicns/Ruppert09.pdf

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, UK: Cambridge University Press.

Ruppert, D (2002),"Selecting the number of knots for penalized splines" *Journal of Computational and Graphical Statistics*,11, 735-757

Scheipl, F. (2011), "RLRsim 2.0-6" *R package* http://cran.r-project.org.

Self, S. G. & Liang, K. Y. (1987), "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions", *Journal of the American Statistical Association*, 82, 605–610.

Searle, S R., Casella, G., & McCulloch, C. E. (1992), *Variance Components*, Wiley, New York.

Shumway, R. H., & Stoffer, D. S. (2006), *Time series analysis and its application with R examples, 2nd edition*, Springer.

Stram, D. O & Lee, J. W. (1994), "Variance components testing in the longitudinal mixed effects model", *Biometrics*, 50, 1171–1177.

Treuth, M. S., Cetellier, D. SJ., Scimitz, K. H., Pate, R. R, Elder, J. P., McMurray, R. G., Blew, R. M., Yang, S. & Webber, L. (2007), "Weekend and Weekday

Patterns of Physical Activity in Overweight and Normal-weight Adolescent Girls", *Physiology & Behaviour*, 87, 650–658.

Tuomisto, M. T., Terho, I., Korhonen, I., Lappalainen, R., Tuomisto, T., Laippala, P., & Turjanmaa, V. (2006), "Diurnal and weekly rhythms of heatlh-related variables in home recordings for two months", *Physiology & Behaviour*, 87, 650–658.

Turjanmaa, V., Kalli, S., Majahalme. S., Saranummi, N., & Uusitalo, A. (1987), "Diurnal blood pressure profiles and variability in normotensive ambulant particpants", *Clinical Physiology*, 7(5), 389–401.

Verbyla, A. P., Cullis, B. R., Kenward M. G. & Welham, S. J. (1997), "The analysis of designed experiments and longitudinal data using smoothing splines", *Research Report*, 97.

Verbyla, A. P. (1995), "A mixed model formulation of smoothing splines and testing linearity in generalized linear models.", *Department of Statistics, The University of Adelaide*, Research Report 95/5.

Wahba, G. (1990) *Spline Models for Observational Data*, Philadelphia: SIAM

Wand, M. P., Coull, B.A., French, J.L., Ganguli, B., Kammann, E.E., Staudenmayer, J. and Zanobetti, A. (2005), "SemiPar 1.0. R package", *http://cran.r-project.org*.

Wand, M. P., & Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Wand, M. P. (2003), "Smoothing and Mixed Models", *Computational Statistics*, 18, 223–249. Available from Internet: http://www.uow.edu.au/ mwand/papers.html

Wood, S. J. (2006), *Generalized Additive Models*, Chapman & Hall/CTC

Wu, H. & Zhang, J-T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis*, New Jersey: Wiley.

Zerubavel, E. (1989), *The seven day circle: the history and the meaning of the week*, University of Chicago Press: Chigaco and London.

Appendix A: Model output

```
A linear penalized spline model for the whole population
Linear mixed-effects model fit by REML
 Data: gdata
       AIC     BIC    logLik
  12479.50 12505.1 -6235.749

Random effects:
 Formula: ~-1 + Z | group
 Structure: Multiple of an Identity
               Z1        Z2        Z3        Z4        Z5        Z6  Residual
StdDev: 0.3535024 0.3535024 0.3535024 0.3535024 0.3535024 0.3535024 0.9793325

Fixed effects: y ~ x
                 Value Std.Error   DF    t-value p-value
(Intercept) -0.1158690 0.1714851 4449 -0.6756796  0.4993
x            0.2943329 0.1529318 4449  1.9246021  0.0543
 Correlation:
  (Intr)
x -0.977

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.930178219 -0.635932724 -0.009896967  0.627097079  4.412858905

Number of Observations: 4451
Number of Groups: 1
********************************
A Linear penalized spline model for gain group

Linear mixed-effects model fit by REML
 Data: gdata
       AIC      BIC    logLik
  1438.177 1455.004 -715.0887

Random effects:
 Formula: ~-1 + Z | group
 Structure: Multiple of an Identity
                Z1         Z2         Z3         Z4         Z5         Z6
StdDev: 0.07681964 0.07681964 0.07681964 0.07681964 0.07681964 0.07681964
```

```
        Residual
StdDev: 1.007633

Fixed effects: y ~ x
                Value Std.Error  DF   t-value p-value
(Intercept)  0.3628822 0.1702841 496  2.131041  0.0336
x           -0.1165517 0.0893694 496 -1.304156  0.1928
 Correlation:
  (Intr)
x -0.884

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.79193720 -0.61883938 -0.01406034  0.60483785  4.22156687

Number of Observations: 498
Number of Groups: 1
```

Appendix B: R-code

```
#The number and placement of knots
K<-max(5,min(floor(length(unique(data1$weekday))/4),35))
knots<-c(quantile(unique(data1$weekday), seq(0,1,length=K+2))[-c(1,K+2)],7) - 0.5
#Model matrices
X<-cbind(rep(1,length(data1$weekday)),data1$weekday)
Z<-outer(data1$weekday,knots,"-")
Z<-Z*(Z>0)

#A linear penalized spline model, data in grouped form
fit1 <- lme(y ~ x, random=pdIdent(~-1+Z), method="REML", data=groupeddata)
#Test for variance components
exactRLRT(fit1)

#Extension of random effect structure
subject-specific slope
fits1 <- lme(y~x,random=list(n=pdIdent(~Z-1),subject=pdIdent(~1)))
subject-specific linear deviation
fits2 <- lme(y~x,random=list(n=pdIdent(~Z-1),subject=pdSymm(~x)))
subject-specific nonparametric curve
K.s<-max(3,min(floor(length(unique(data1$weekday))/4),35))
knots.s<-quantile(unique(data1$weekday), seq(0,1,length=K.s+2))[-c(1,K.s+2)]
Z.s<-outer(data1$weekday,knots.s,"-")
Z.s<-Z.s*(Z.s>0)
fits3 <- lme(y ~ x, random=list(n=pdIdent(~Z-1),subject=pdSymm(~x),subject=pdIdent(~Z.s

Varcorr(fits1)

#bootstrap
simulate.lme(fit1,nsim=10000,m2=fits3)
g.boot <- -2*(nl-al)
sum(g.boot > g.star)/10000


#quadratic and cubic splines in section 5.2.4
Z2<-outer(data1$weekday,knots,"-")
Z2<-Z2*(Z2>0)
Z <- Z2^2
Z3 <- Z^3
Z4 <- Z^4
```

```
fit2 <- lme(y ~ x+I(x^2), random=pdIdent(~-1+Z), method="REML",data=gdata)
fit3 <- lme(y ~ x+I(x^2) + I(x^3), random=pdIdent(~-1+Z3), method="REML",data=gdata)
fit4 <- lme(y ~ x+I(x^2) + I(x^3)+ I(x^4), random=pdIdent(~-1+Z4), method="REML",data=g
exactRLRt(fit2)
```