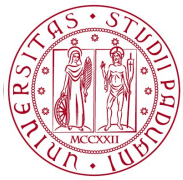

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche

SCUOLA DI DOTTORATO DI RICERCA IN: SCIENZE STATISTICHE

CICLO XXII

APPROXIMATE BAYES RANDOM EFFECTS MODELS FOR LARGE DATASETS

Direttore della Scuola: Ch.ma Prof. ssa ALESSANDRA SALVAN

Supervisore: Dott. BRUNO SCARPA

Co-supervisore: Prof. DAVID DUNSON

Dottorando: JAMES MBUGUA CIERA

Data: 31, January 2010

Abstract

Many medical studies collect functional data, such as trajectories in a biomarker over time. It is of interest to estimate the trajectories and identify or predict clinically-important features. Linear mixed effects (LME) models are commonly used in such cases, with non-linear effects easily incorporated through splines. However, for sufficient flexibility, it is often necessary to use adaptive splines in which the number and locations of knots is unknown and potentially varying across subjects. This can be accomplished with MCMC methodology, using reversible jump or stochastic search variable selection. However, such approaches are slow and infeasible to implement routinely, particularly for large data sets. Motivated by methods proposed in the machine learning literature for compressive sensing, we focus on relevant vector machine (RVM) methodology - a fast approximate Bayes functional data analysis approach that relies on sparseness-favouring hierarchical priors for basis coefficients. Recent literature on the use of RVM methodology is restricted to models that assume that the distribution of the basis coefficients is centered at zero with diagonal covariance. However, in many longitudinal and functional data analysis applications, centering at zero is an unrealistic assumption and does not allow shrinkage towards a population-averaged function. In this work, we develop a generalized multi-task relevant vector machine (MT-RVM) methodology that generates sparse functional linear mixed models to estimate both population-average and subject-specific curves. In particular, we first consider an LME model that assumes independent random effects and then extend the approach to a more generalized LME model with correlated random effects. Further, we extend the application of the generalized MT-RVM methodology into multi-level relevant vector machine (ML-RVM) methodology to generate a sparse multi-level functional mixed model. The analysis of basal body temperature curves over the menstrual cycle has been the motivating application for all the developed methods.

Riassunto

Molti studi medici raccolgono dati in forma funzionale come ad esempio le traiettorie in un bio-marcatore nel corso del tempo. Di questi dati di interesse stimare le traiettorie e individuare o predire caratteristiche clinicamente importanti. I modelli lineari ad effetti misti (LME) sono comunemente utilizzati in questi casi, anche utilizzando effetti non-lineari che si possono includere facilmente attraverso splines. Tuttavia, per ottenere una flessibilit adeguata, spesso necessario utilizzare splines adattive in cui il numero e la posizione dei nodi ignoto e potenzialmente variabile tra soggetti. In questo contesto si utilizzano strumenti di tipo MCMC (Markov Chain Monte Carlo), come ad esempio il reversible jump o la selezione di variabili attraverso ricerca stocastica. Questi approcci sono, tuttavia, lenti e difficilmente utilizzabili in contesti in cui si ripetono spesso le operazioni di stima, in particolare per grandi dati set. A partire dagli strumenti sviluppati nella letteratura del compressive sensing in ambito di machine learning, ci siamo concentrati sulle relevant vector machine (RVM) - un approccio di analisi di dati funzionali bayesiano che utilizza veloci approssimazioni che sfruttano distribuzioni a priori gerarchiche per i coefficienti delle basi che ne favoriscano la sparsit. La letteratura recente per l'uso della metodologia RVM limitata ai modelli che assumono che una distribuzione dei coefficienti base centrata sullo zero con matrice di varianze e covarianze diagonale. In molte applicazioni su dati longitudinali e funzionali, tuttavia, la centratura sullo zero risulta essere una ipotesi poco realistica non consentendo il restringimento ad una funzione centrata sulla media della popolazione. In questo lavoro, abbiamo sviluppato una "multi-task relevant vector machine" generalizzata (MT-RVM), che genera modelli funzionali lineari misti sparsi per stimare sia la curva della media della popolazione che la curva specifica per soggetto. In particolare, in primo luogo abbiamo considerato un modello LME che assume effetti casuali indipendenti e successivamente abbiamo esteso questo approccio ad un modello LME pi generalizzato con effetti casuali correlati. Inoltre, abbiamo esteso ls metodologia MT-RVM generalizzata alla situazione in cui sono disponibili diversi livelli di gerarchia, ottenendo una "multi-level relevant vector machine" (ML-RVM) che genera un modello multi-level funzionale sparso ad effetti misti. I metodi sviluppati sono stati motivati dal problema di analizzare le curve della temperatura basale durante il ciclo mestruale, e tale applicazione viene considerata come esemplificazione durante tutta la tesi.

Acknowledgements

My three year in Padova have been very challenging but fruitful. First, I would like to thank my advisors Dr. Bruno Scarpa, Prof. David Dunson, Prof. Bernardo Colombo and the director of the PhD school Prof. Alessandra Salvan. David, I treasure all the discussions we had, numerous emails that we exchanged and the courses you conducted. They really shaped my understanding of Bayesian Statistics. You were very patient with my numerous mistakes while writing my dissertation. I owe to you almost everything in this thesis. Bruno, you are one professor that every student would like to work with! Your teaching skills, encouragement and close companionship really made me feel at home while in Italy. Prof. Colombo, I'm grateful for allowing me to use the data from your database. Prof. Salvan, I cannot forget the statistical Inference course that shaped my understanding of statistical theory.

I'm greatly indebted to *Cassa di Risparmio di Padova e Rovigo* (CARIPARO) foundation who funded my PhD course in Italy. Many thanks to Prof. David Dunson for the efforts he provided to accommodate me for six months at NIEHS in North Carolina. It was a great honour to work with you at NIEHS and I appreciate all the support you provided during my internship period.

My classmates: Nadia, Vanna, Laura, Susanna, Nicola, Francesco and Daniele, I thank you for your warm friendship during my stay in Italy. Vanna, thanks for helping me negotiate for my Permisso di Soggiorno at the Questura. Many thanks to my colleagues in Morgangni, Goito and Galileo ESU residences. It was fun having you around.

Special thanks goes to my father Ciera, who was persistent in encouraging me to work diligently and remain focused upon the promises, my mother, Wairimu, for her consistent mothering spirit, my sisters, Wangari, Wanjiru, Ngina and Muthoni for their strong moral support. Above all, I lack words to express my gratitudes to my loving God. I believe You when you say, "Behold, I have set before you an open door, which no one is able to shut. I know that you have but little strength ...". Help me to remain strong in You till the end. I dedicate this work to You!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In many clinical studies, data are collected repeatedly from many subjects over a period of time. Using massive datasets, physicians require fast automated tools to estimate data trajectories and predict clinically important events for a current patient. For example, in reproductive studies, trajectories of hormonal level or daily basal body temperature (bbt) among women can help to identify or predict early pregnancy loss and occurrence of the ovulation day (Bigelow and Dunson, 2008). Hence, to estimate the shape and predict the location of such features in the function, there is a need for fast algorithms for estimating functional trajectories while borrowing information from other patients.

Our research is motivated by the bbt data from European fecundability study (Colombo and Masarotto, 2000). The study enrolled women aged between 18 and 40 years, were not taking hormonal medications or drugs affecting fertility, and had no known impairment of fecundity. The participants kept daily records of daily basal body temperature. We consider data from women that contributed temperature measurements from at least one menstrual cycle. The data are characterized with unequal cycle lengths and unequally-spaced measurements causing problems in estimating the bbt curves. Thus, rapid estimation of accurate and smooth curves is not a trivial problem especially when working in Bayesian framework.

In this thesis we use Functional data analysis (FDA) methods (Ramsay and Silverman, 2005) to estimate curves. In particular, we restrict ourselves to Bayesian framework where posterior sampling is usually based on slow Markov Chain Monte Carlo (MCMC) algorithms. This raises a practical motivation for fast approximate Bayes approaches that bypass MCMC while maintaining some of the benefits of a Bayesian analysis. For

fast parameter estimation, we use Relevant Vector Machine (RVM) (Tipping, 2001) which is one among fast Bayesian methods that promote sparseness in estimation of the basis coefficients, providing a more flexible alternative to Support Vector Machines (SVM) (Burges, 1998) and LASSO (Tibshirani, 1996), leading to a sparser solution that is more robust to outliers (Tipping, 2001; Ji, Dunson and Carin, 2009). RVM is based on Empirical Bayes methodology and penalizes the basis coefficients through a scale mixture of normal priors, which is carefully-chosen so that maximum a posteriori (MAP) estimates of many of the coefficients are zero. The work in this thesis is divided in seven chapters.

Chapter two and three describe the motivating application. In particular, chapter two highlights some background information about the biological processes that lead to a biphasic pattern in basal body temperature as well as other related fertility biomarkers while chapter three gives a brief review about the European fecundability study (Colombo and Masarotto, 2000). Chapter four reviews recent developments in functional data analysis in estimating non-linear curves.

In chapter five we give an application of the multi-task relevant vector machine (MT-RVM) methodology (Ji, et al., 2009) into functional random coefficients and linear mixed models. The models are based on basis functions generated using natural cubic smoothing spline. To allow implementation of the RVM algorithm in linear mixed models, we consider a simple mixed model that assumes independent random effects. The approach is used to rapidly estimate population and individual-specific functions based on basal body temperature data.

Chapter six presents a more flexible generalization of the MT-RVM in linear mixed models that allows shrinkage towards a non-zero mean and non-zero covariance in the random effects. Since the random effects are normally correlated, we use the approach of Chen and Dunson (2003) to generate uncorrelated latent variables that can easily be implemented in MT-RVM methodology. We also present a simulation study and an application to real data. The simulation study compares the performance of the MT-RVM method relative to Functional Principal Component Analysis (Crainiceanu, 2009), Functional mixed model of Durban et al. (2005) and Bayesian functional mixed model of Wand and Ormerod (2008).

Chapter seven contains an extension of the MT-RVM model into multi-level relevant vector machine (ML-RVM) methodology. This approach is used to model multi-level data where measurements are nested within cycles and cycles are nested within subjects. The ML-RVM method is used to generate a sparse hierarchical mixed model by selecting relevant fixed and random effects. Based on the bbt data, we estimate population-average, subject and cycle specific curves.

1.2 Main Contributions of the Thesis

The main contribution of this work is to extend Multi-Task Relevant Vector Machine (MT-RVM) approach introduced by Ji, et al.(2009) into functional mixed models. Multi-Task Relevant Vector Machine methodology is an extension of Relevant Vector Machine method (Tipping, 2001) that provides a natural and fast mechanism in selection of the basis functions. The Ji, et al. (2009) approach allows basis function selection within a restricted class of models that assumes that the distribution of the basis coefficients is centered at zero with diagonal covariance. However, centering at zero does not allow shrinkage towards a population-averaged curve and independence of the random effects which is an unrealistic assumption in many longitudinal and functional data analysis applications. The generalized MT-RVM methodology is used to select basis functions in functional mixed models to generate sparse models that are easy to fit and take shorter time relative to classical Bayesian MCMC based methods. In summary the contribution of this work can be described as follows:

- Since Multi-Task Relevant Vector Machine has been used to estimate and reconstruct multiple signals based on wavelet bases, our first goal is to implement the MT-RVM approach in reproductive studies to estimate the bbt curves. We use functional random coefficients model based on linear combination of cubic B-spline basis functions that are commonly used to estimate non-linear curves.

- The generalized MT-RVM methodology is then used to generate sparse functional linear mixed models that can estimate both population-average and subject-specific curves. In particular, we consider two cases: (i) A linear mixed models (LME) that assumes independent random effects. (ii) A general LME model with correlated random effects. To facilitate the extension, we rely on a modified Cholesky decomposition proposed by Chen and Dunson (2003).

- In the final chapter, we extend the application of the generalized MT-RVM methodology to fit hierarchical functional mixed models. In particular, we consider cases where cycles with unequal number of measurements are nested within subjects.

This work demonstrates the use of a fast Empirical Bayes method as an alternative to computer intensive methods that rely on MCMC. The method is fast and can be used to rapidly approximate curves for massive datasets. The advantage of our approach is not only on computational speed, but it also allows for better generalization performance which leads to sparse generalized linear mixed models. This aspect can provide inference for a wide variety of models at a moderate computational cost. The approach can be

extended to accommodate probit models where multiple binary categorical outcomes can be handled using data augmentation (Albert and Chib, 1993).

Chapter 2

A review on human fertility

2.1 Introduction

Human reproductive study is one among many scientific fields that have been studied and significantly benefited from recent technological innovations. The key factor that has stimulated the advancement is the exponential population growth throughout the world and in particular the desire for couples to regulate and limit their family size. This has lead to the development of advanced statistical models that have been implemented in devices used to predict the most fertile period among women and can be used at home by couples in all parts of the world. With these advancements, it is technically possible to predict the most fertile period within days although there are still limitations on accuracy on predicting the exact moment when fertility is at its peak.

In many text books, human reproductive studies are mostly focused on women and narrows down to human fertility or child-bearing capacity of a woman. However, both men and women have different underlying biological capabilities that determine the child-bearing processes. For instance among women, fertility is the ability to become pregnant while infertility is the inability to initiate or sustain a pregnancy after one year of attempt (Tuerlings, 2000). Women have complex biological processes which regulates their fertility cycles lengths and the associated physiological changes. Studies have shown that various factors affect the child-bearing process and some of these factors include: physical, psychological, biological factors, and the age of the woman (Holman, O'Connor and Wood, 2006). Fertility among women peaks at late teens and deteriorates after the age of thirty. For a detailed review on physiological and biological processes related to human fertility, see (Wood, 2001; Rodgers and Kohler, 2003).

The fertile period that has the highest likelihood of pregnancy resulting from sexual intercourse is believed to be between 5 days before and 2 days after the release of a

mature egg (Rodgers and Kohler, 2003). Accurate timing for this period is of great importance and can benefit both infertile and fertile couples to achieve and avoid pregnancy respectively. For example, in the management of infertile couples, right timing is essential for artificial insemination and retrieval of oocytes at appropriate maturity for *in vitro* fertilization (Edirisinghe, Murch, Junk and Yovich, 1997; Wood, 2001). For the fertile couples, right timing can help to determine the right moment for child bearing which is mostly geared toward spacing and limiting purposes.

Although knowledge on the development of follicle (nests of cells that contain primitive egg) in the ovary has increased considerably especially the developments that occur some days prior to the most fertility period, there is a great need for simpler and reliable methods on predicting and identifying the fertility peak. However, there is a lot of variations on cycles lengths and the levels of fertility hormones excreted in the body which are widely used to track fertility among women. These variations can cause problems in predicting the fertility peak and hence raising a practical motivation to develop better and easy to use methods that couples can use to identify the fertile and infertile days in the cycle. Most methods are based on observations made from one or more primary fertility indicators which include: basal body temperature, cervical mucus, and cervical position (Halberg et al., 2000; Halberg et al., 2001; Wood, 2001). These primary fertility indicators will be discussed in the next section. Other modern methods involves urine test kits that consist of tools used to detect the surge in fertility hormones that occurs prior to the peak of fertility (Holman and Wood, 2001; O'Connol et al., 2006). Advancement of these tools has lead to computerized devices that can interpret daily measurements from the basal body temperatures and fertility hormones.

In this study we shall focus on studying curves obtained from basal body temperature (bbt) a key fertility indicator that varies systematically and periodically and is widely used to monitor underlying fertility levels among women. In natural family planning methods which entirely rely on visible physiological characteristics to track fertility levels, the bbt is used to confirm the occurrence of the fertility period. Our study aims to develop a fast and efficient statistical method that can be used by physician to rapidly estimate the bbt curves resulting from daily collection of the bbt measurements. Before we get into developing the anticipated statistical methodologies, we shall give a brief overview on menstrual cycle, indicators used to monitor fertility level among women, the biological mechanism underlying the bbt curve and also highlight some findings from recent studies related to basal body temperature.

2.2 Biological processes in a menstrual cycle

Menstrual cycle is a recurring process of physiological changes that occur among reproductive aged females and on average it lasts for about 28 days. A normal menstrual cycle is divided into four phases; the bleeding phase (menstruation), pre-ovulation, ovulation and post-ovulation phases. Ovulation is the time when a mature egg is released for fertilization and is the optimal time of fertility. The length of the fertile phase is believed to remain constant for each woman, but the cycle length and the ovulation day can change depending on the underlying biological processes taking place in a woman. Irregular cycles are common among teens, women approaching menopause, those that are breastfeeding and those coming off the pills.

A hormone is a substance which provides means of communication in that it travels from a special tissue, where it is released into the bloodstream, to distant responsive cells where the hormone exerts its characteristic effects (Sperroff et al., 1994). Several hormones are known to regulate different activities in the menstruation cycles and in particular on the shape of the bbt curve. These hormones are: the Follicle Stimulating Hormone (FSH), the Luteinising Hormone (LH), pituitary, progesterone, estrogen and oestriodiol. The levels of FSH and LH hormones in the blood are regulated by hypothalamus which are highly specialized brain cells. A decease or increase of any of the hormones leads to changes in the menstruation cycle and periodical fluctuations take place throughout the menstrual cycle. Figure 2.1 shows the levels of different hormones during a menstrual cycle.

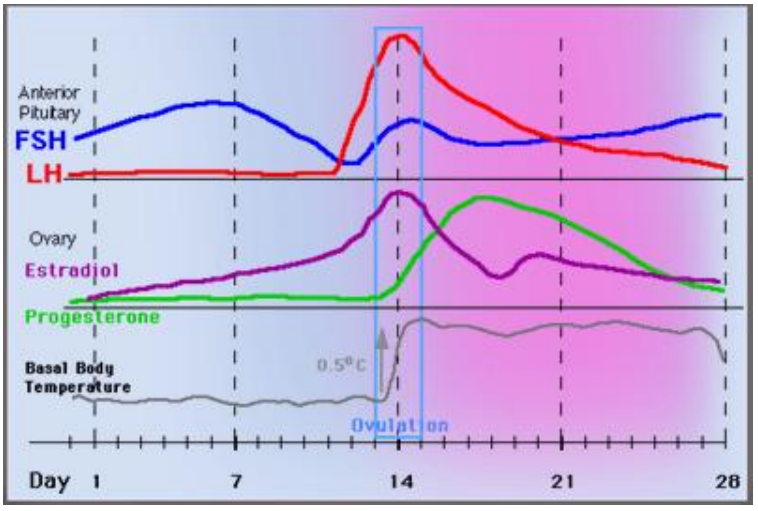

FIGURE 2.1: Levels of different hormones in an menstrual cycle.
Source: http://www.early-pregnancy-tests.com/progesterone.html

The menstruation phase is the first stage in the cycle and is characterized by shedding of blood from the vagina. The released blood is dominated by the shed the uterus lining that was developed in the previous cycle. The beginning of this phase marks the end of

the previous cycle and the beginning of a new one. On average it takes about four days, but it can also take a shorter or longer duration depending on the length of the cycle. However, bleeding can occur without the release of an egg and is commonly caused by inadequacy of sufficient hormones to causes the release of an egg. This is a common phenomenon among women at the extreme ends of the fertility period (teen-age and menopause).

Pre-ovulation phase is characterized with the rise of FHS hormone levels in the blood resulting into production of pituitary hormones in pituitary glands which initiates the development of follicle in the ovary. The developing follicle produces oestriodiol hormones that activate the cervix to produce mucus at the opening of the vagina. The release of mucus heralds the beginning of the fertility phase. While the egg in the follicle is maturing, the follicle moves toward the surface of the ovary in order to be release into the fallopian tube. At this moment, the levels of the released oestriodiol hormones from the follicle rises further leading to an increase in fertility characteristics on the mucus. At the same time the endometrial lining of the uterus begins to grow.

Ovulation takes place immediately after maturity of the egg in the follicle which is then released into the fallopian tube. This activity takes one day of the cycle and the most fertile type of mucus is evident at the lining of the cervix and the vagina. This mucus type facilitates the sperm with guidance, transportation, protective environment along the passage and supplies the required nutrition needed for fertilization. In absence of this special type of mucus, the sperms cannot survive in this passage and hence no fertilization can take place. At the same time the egg drift along the fallopian tube ready to meet the sperm. If the egg is not fertilized, the egg dies and breaks down into its constituent components and is re-absorbed by the body. But if the released egg comes into contact with the male's sperm, fertilization takes place and foetal growth is initiated. The union leads to the development of an embryo which latter continue moving towards the uterus. In preparation for a pregnancy, progesterone hormone is released by the corpus luteum (empty follicle left in the ovary). The release of progesterone leads to a sharp rise in the body temperature of a woman.

Post ovulation is the last stage of the menstrual cycle. This stage is commonly characterized with an elevated plateau of the bbt level. It is also characterized with either the development of the embryo if fertilization took place or the disposal of the released egg and the uterus lining. When fertilization takes place the embryo moves from the fallopian tube to the uterus and after about six days it is embedded on the uterus lining. Implantation of the embryo is complete in about 12 days after ovulation. The uterus lining thickens and is filled with nutrients suitable for the implanted embryo. When

fertilization does not result after ovulation, the levels of progesterone and estrogen hormones decline in the blood resulting to breaking away of the uterus lining from the uterus wall. The resulting refuse is commonly released as menstrual on the fourteenth day after ovulation.

2.3 Indicators used to monitor fertility level

Many methods have been developed to monitor the fertility level within a menstrual cycle. Many indicators are characterized with observable characteristics that women can identify and detrmine the occurrence of the ovulation day. Modern indicators consist of direct measurements of levels of estrogen and progesterone in urine or blood and the use of ultrasound scanning to monitor ovarian activities. Traditional indicators include: the rhythm calender, basal body temperature, cervical mucus and a combination of both the mucus and the temperature indicators. We shall give a brief overview of each one of these indicators.

The levels of LH, FHS, estrogen and progesterone hormones in the blood have been used to monitor the fertility level during the menstrual cycle. Their quantities can be measured in blood by radio-immunoassay or their metabolites which can be measured in urine. Blood assays can be used but require tedious daily sampling which is necessary to provide an accurate picture of ovarian activity around ovulation. In most cases analysis is done using urine assays. Women collect urine early in the morning daily and the assays are simplified to the stage that women themselves can do accurate testing at home using the Home Ovarian Monitor. A surge on secretion of these hormones is a prerequisite of ovulation. For example, the rising levels of oestradiol secreted about 5 to 7 days prior to ovulation and a surge on FSH/LH hormone heralds pre-ovulatory events of a normal cycle (Ross et al., 1970).

Ultrasound scanning is one of the modern indicators that is believed to predict ovulation accurately though expensive. Ultrasound scanning technology helps in visualizing the activity taking place on the follicles and corpus luteum. Mostly it is concerned with visualizing the actual rupture of the follicle, the extrusion of the ovum and follicular fluid, the development of a corpus luteum, the blood supply to these structures and the degree of stimulation of the uterine endometrium as a result of the hormones produced. The method has played an important role in providing basic information on all phases of ovarian activity, and its agreement with the findings based on the hormone patterns and mucus symptoms. For daily application, ultrasound scanning is expensive and other methods are used to assess ovarian activity but used latter as a final confirmation that ovulation is imminent.

The rhythm calendar indicator is based on calculating the ovulation pattern and specifying the interval when ovulation would occur. In most cases it is believed that ovulation occurs between 11 to 16 days before the beginning of menstrual bleeding. Using this pattern, a woman can calculate the expected range of days when ovulation occurs. However, biological processes in the body changes with the underlying environmental and psychological factors. Therefore, the menstrual cycles are not always constant and there are tendency for a woman to experience short and long menstrual cycles. Since the underlying assumption in calculation is based on the average cycle length, there are problems with identifying short and long cycles. Hence, the use of this indicator can lead to unwanted pregnancies or force couples to abstain when there is no risk of pregnancy.

Cervical mucus indicator is based on changes that occur on vaginal discharges during the menstrual cycle. Cervical mucus is a type of hydrogel fluid produced by the cervical glands and heralds the beginning of the fertility phase of a cycle (Odeblad, 1994). It is release is determined by the rise of the levels of estrogen hormone in the blood. This fluid prepares the birth channel to receive any sperms that are introduced into the channel. Its primary function is to facilitate fertilization by nourishing and protecting sperms as they travel through the reproductive system. The cervical mucus tracks the cyclic changes in oestrogen and progesterone released by the ovaries.

The mucus changes in quality and quantity before and during ovulation. Cervical mucus is not identical for all women but is believed to be characterized with four common states. These states are: state 1 that is characterized with dry, rough and itchy feeling or nothing is felt or no mucus (occurs during the infertile phases), state 2 that is characterized with damp feeling, nothing is seen or there is no mucus (early fertile days), state 3 has a damp feeling, the mucus is thick, creamy, whitish, yellowish, not stretchy but sticky. State 4 is the peak fertile level of the mucus where the mucus is characterized with wet, slippery, smooth feeling, is transparent, like raw egg white, stretchy/elastic, liquid, watery, reddish with some blood (Colombo and Masarotto, 2000). With proper personal instructions, women can recognize these changes and can correctly identify the pre-ovulation infertility period as well as post-ovulation phase of the menstrual cycle (Billings, Billings, Brown, and Burger, 1972; Billings and Billings, 1983).

Like other biomarkers, the bbt is an indicator that responds regularly to changes in progesterone hormone levels during the menstrual cycle. Basal body temperature is defined as the temperature a body has at the time you wake up each day. During the pre-ovulation period, the progesterone level in the blood is very low, and this is characterized by low bbt measurements. As ovulation approaches, the level of the progesterone hormone starts to rise in the blood and this leads to a sharp rise in the basal body temperature. After ovulation, the progesterone level is relatively high and this corresponds

to a high plateau on the bbt curve. As the end of the menstrual cycle approaches, both the bbt and progesterone levels decrease in preparation for the next cycle. The patterns of the progesterone trajectory can also facilitate in explaining the biphasic shape of the bbt curve (Marshall, 1968). Taking the bbt measurements daily immediately after waking up and noting the day that the temperature level changes, a woman can easily determine the phases of her menstrual cycle. The work in this thesis is based on the basal body temperature data. First, we will give a detailed review of characteristics associated with the bbt indicator in the next section.

The cervical mucus, basal body temperature, and rhythm calendar indicators are commonly used in Natural family planning (NFP) methods to predict and determine the occurrence of ovulation. Natural family planning methods are collection of methods that relies on natural observable signs and symptoms of the fertile and infertile phases of the menstrual cycle. The NFP methods give provision of avoiding or achieving pregnancy naturally through conforming to the women's reproductive cycle without the use of drugs or devices. Hence, they can comfortably be used in diverse populations with varied religious/ethical beliefs. Other advantages address the need to provide alternative for couples who want to use natural methods due to medical or personal reasons.

2.4 Researches on basal body temperature

We shall begin by giving a historical review on researches involving the use of basal body temperature in determining ovulation. Researches on establishing the relationship between the bbt and fertility levels in woman have received attention since the 19 century when Squire in 1868 reported that the bbt has a biphasic pattern during a menstrual cycle. In 1905 a Dutch gynecologist van de Velde, established that there is a relationship between the change in bbt curve and the timing of ovulation. Early extensive reviews on bbt patterns and menstrual cycles have been documented by Marshall (1965), Vollman (1977) and Zuspan & Zuspan (1979). Further, Moghissi et al. (1976), established that the rise in bbt level happens after an LH surge and that a significant rise in bbt coincide with rise in progesterone and urinary pregnanediol.

A standard shape of bbt measurements recorded daily in a menstrual cycle is characterized with a biphasic curve (see figure 2.2). In most cases, trajectories of bbt curves from a healthy woman have identical shapes. A typical curve is characterized with two main sections, pre and post ovulation regions. The pre-ovulation region is characterized with a low plateau while the post ovulation region is characterized with an elevated plateau. The transition region between the pre and post ovulation is relatively short and coincides with the ovulation period. There is a wide fluctuation of the bbt measurements

but the transition phase can be identified from the rest of the regions by a sharp drop (nadir) in temperature that is followed by a sharp rise (Dunlop et al., 2005; Scarpa and Dunson, 2009).

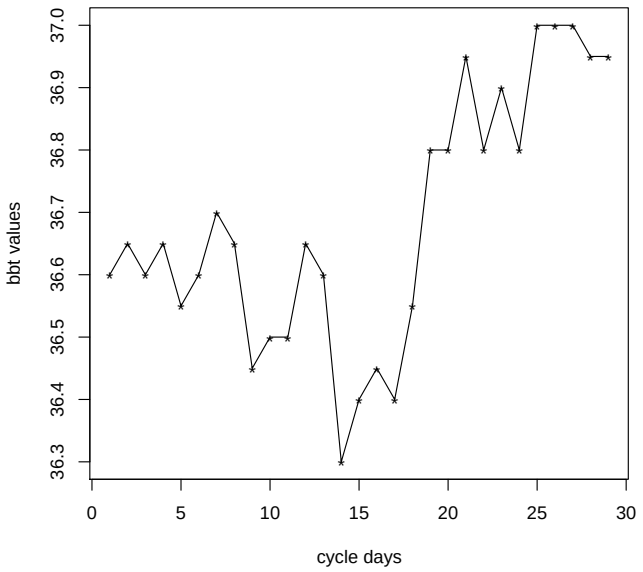

FIGURE 2.2: A typical bbt curve.

However, the existence of a nadir prior to ovulation has drawn criticism since the inception of the bbt method. Marshall (1968) conducted a detailed study focusing on thermal changes in menstrual cycles with biphasic bbt patterns and observed that an acute rise with an elevated level of at least $0.2C$ ($0.4F$) between two consecutive days occurred in 80% of the menstrual cycle but only 10% of the cycles had a dip preceding the temperature rise. Further, Hilgers and Bailey (1980), working on both bbt and hormonal measurements observed that a dip in bbt curves preceding ovulation is about 15% hormonally normal cycles. Thus concluding that a significant elevated shift in bbt occurring in 48 hours or less in which three consecutive daily temperatures are at least $0.2C$ ($0.4F$) than the last six daily measurements is a more acceptable measure that confirms ovulation (Marshalls, 1968; Jeffcoate, 1983).

The average body temperature is known to vary from person to person. In most women the body temperature during the follicular phase ranges between $36.37 \pm 0.12C$ ($97.48 \pm 0.25F$) and $36.72 \pm 0.12C$ ($98.09 \pm 0.22F$) in the luteal phase. A shift in the bbt following ovulation should occur within a period of 48 hours or less depending on the length of the cycle. In most ovulatory cycles, there are three consecutive daily bbt measurements that are between $0.2 - 0.5C$ ($0.4 - 1.0F$) higher than the previous six daily temperatures which leads to a three by six method which is commonly used to identify occurrence

of ovulation (Jeffcoate, 1983). By recording and charting the body temperature for each day, the change on the trajectory of the temperature curve can be used by couples to confirm the onset of ovulation and be able to predict with some probability when conception will occur in the future cycles (for more details refer to: Jeffcoate, 1983; Wood, 2001).

Predicting the ovulation day has been of interest in many reproductive studies. However, the temporal relationship between ovulation and a shift in bbt measurements and ovulation day has been extensively been investigated but the relationship appears to be a variable. To uncover the relationship, the biphasic pattern has been studied and provides a basis for algorithms in identifying the ovulation day in a menstrual cycle (Scarpa and Dunson, 2009). Based on these and other algorithms variety of statistical models have been developed to identify the shift in bbt curve which is commonly linked to ovulation day. For example, Royston and Adams developed a CUSUM algorithm (Royston and Adams, 1980), while Carter and Blight introduced a Bayesian rule to detect the bbt shift (Carter and Blight, 1981). Colombo and Masarotto 2000, identified the sharp rise of the bbt curve at the last day of hypothermia prior to the post-ovulatory rise, as a good marker of the day of ovulation.

Observing the bbt curve to determine fertility has its own limitations. Information from the bbt chart is retrospective and normally informs the user that ovulation has already occurred. Moreover, the bbt measurements can be influenced by psychological, physical or health disorders e.g. sickness, fever or stress, among other factors. Therefore, it is advisable for users to note special events in the chart to help make proper interpretation of the bbt curve in the future. Women in their teenage and menopause can not fully rely on the method since there is random fluctuation of both the cycle length and the bbt measurements due to change in biological functioning of their bodies.

If the bbt measurements for all cycles result to smooth curves identical to the one in figure 2.2 and none of the cited limitations interfere with trajectories of the bbt curves, then it can be easy to identify the ovulation day. However, in real life situation it is very rare to have such a smooth curve. Moreover, most bbt measurements are noisy and majority of curves from a single woman may have different patterns. Thus, the relationship between the ovulation day and the bbt shift becomes a variable that keeps on changing depending on the underlying biological factors taking place in the body. These complex circumstances raises a practical motivation leading to explore for better approaches to estimate smooth biphasic bbt curves.

Previous researches on bbt have been on estimating the ovulation day but little has been done on generating smooth bbt curves that can be used to solve other problems. In particular for women attempting to conceive, the pattern of the estimated curve can

be used to distinguish healthy ovulatory cycles from those characterized with menstrual disorders. Hence, it is of interest to estimate the trajectory in bbt over the menstrual cycle. The aim of this work is to provide a fast statistical methodology that can be used to generate curves while borrowing information flexibly from other cycles in the data base. Such automated tool can be used to generate smooth curves from noisy bbt measurements and physicians can easily approximate the ovulation day and also distinguish dysfunctional cycles from normal ones. These statistical methodologies can be extended to fit curves generated from hormonal measurements leading to better and more efficient methods to predict both ovulation day and menstrual disorders including dysfunctional menstrual cycles.

Chapter 3

The dataset

3.1 Introduction

In this chapter we will give a review of the study design, sampling process and related issues linked to the collection of data in a multi-center study called The European Study of Daily Fecundability. One of the primary goals in the study was to investigate and predict the daily probability of conception within the fertile window of a menstrual cycle among healthy women using body basal temperature data and cervical mucus characteristics. The dataset is a result of a multi-center study that has been discussed in Colombo and Masarotto (2000). The investigation was planned as a prospective cohort study that started in 1992 through 1996 on subjects interested in learning about the fertile phase of the woman and the use of a Natural Family Planning method to avoid or achieve pregnancies.

3.2 Study Population

The recruitment process enrolled 782 women from The European Study of Daily Fecundability in collaboration with seven European centers (Milan, Verona, Lugano, Dsseldorf, Paris, London and Brussels) that provided services on fertility awareness and natural family planning. The 782 women from the European centres were recruited between 1992 through 1996. The research protocol was reviewed and approved by the Institutional Review Boards of Fondazione Lanza (Padua, Italy) and Georgetown University (Washington D.C., U.S.A.). The study was co-ordinated by Professor Bernardo Colombo from the Department of Statistical Sciences of the University of Padua (Padua, Italy) (Colombo and Masarotto, 2000).

The selection/entry criteria for the subjects were: women experienced in use of a Natural Family Planning method; married or in a stable relationship; aged between 18 and 40 years at admission; having at least had one menses after cessation of breastfeeding or after delivery; not currently taking hormonal medication or drugs affecting fertility. Neither partner could be permanently infertile nor both had to be free from any illness that might cause sub-fertility, e.g., endocrine disorders. Another requirement was that couples were not supposed to have a habit of mixing incidences of unprotected and protected intercourse. Any woman was excluded if any one of the discussed criteria was not fulfilled (Colombo and Masarotto, 2000).

To increase the sample size, additional data for 99 women was included retrospectively from a prospective study that was conducted in Auckland, New Zealand between 1979 and 1985. The main focus of the New Zealand study was to investigate the relationship between the interval from intercourse to fertilization and the sex of the baby conceived (Colombo and Masarotto, 2000). For the New Zealand study, recruitment process involved couples that contemplated having more pregnancies and had proven fertility. The two studies had similarities in that the couples were instructed to recognize the fertile period using the daily changes in cervical mucus and patterns of the basal body temperature. However, the couples in the study were restricted to only one act of intercourse during the fertile phase of the cycle (France et al, 1984, France et al, 1992; Colombo and Masarotto, 2000). This condition was not properly observed in some cases and was believed to be one of the causes that lead to frequently dropping out of the study when couples failed to achieved a pregnancy after 3-4 cycles of attempting pregnancy.

3.3 Study Design and Data Collection

Subjects in the European Study of Daily Fecundability were selected from different centres. The subjects were screened and selected by natural family planning teachers for admission. These teachers were trained by the local principal investigator in each centre on the purpose and requirements of the study. When a subject satisfied the above entry criteria, she was requested to give a written informed consent for her to be enrolled into the study. Since the study collected sensitive personal nature which encompassed sexual behaviour subject's, anonymity and confidentiality was ensured by assigning a study number to each woman and only the NFP teacher maintained a personal relationship with the subject.

On data collection, the study collected data related to the menstrual cycle, pregnancy, basal body temperature, cervical mucus characteristics, and other demographic characteristics e.g. marital status, contraceptive use, etc. A menstrual cycle was defined as

the interval that begins on the first day of vaginal bleeding until the commencement of the next menstrual cycle. Day 1 of a menstrual cycle was considered as the first day of fresh red bleeding that excluded any preceding days with spotting. A conception was assumed in the presence of a pregnancy going on at 60 days from the onset of the last menses or when before that term a miscarriage was clinically detected.

For each woman, the following information was collected: the month and year of birth of the woman and of her partner; the number of previous pregnancies, if any; the date of her last delivery (or miscarriage) and of the end of breastfeeding period, if relevant; the date of last oral contraceptive pill taken, if any. Subsequently, after the collection of data had begun, it was decided to add the date of marriage for married couples and the sex of any baby conceived and born during the period of the study. This latter information is available for a large proportion of subjects.

On bbt, cervical mucus and other related data, the woman was asked to record on a chart the days of her period and of any disturbance such as illness, broken sleep. Subsequently, the woman was asked to collect the bbt measurements every morning before engaging in any activity. She was asked to record her basal body temperature on a chart (as shown on figure 3.1) for as many days as necessary to determine a clear post-ovulatory rise. These measurements were charts and sent periodically to the Department of Statistics at the University of Padua. The charts were then evaluated for all cases of the recorded bbt measurements to ensure data consistency.

Similarly, the data for the mucus typology and texture was recorded by the subjects. To collect the cervical mucus data, a woman was asked to observe and chart her cervical mucus symptoms daily during the cycle. The study required a record of every episode of coitus, with specification on whether the act was unprotected or protected (barrier methods, withdrawal, and others). The reliability of the information recorded of acts of intercourse was checked by the teacher in discussion with subjects at the end of each cycle. To promote data consistency, the investigators in the study excluded any cycle that was characterized with a single act of protected intercourse or any simple genital contact. Mucus characteristics coding were done in the local centres in accordance with agreed common rules. The charts for both bbt and cervical mucus were sent to the co-ordinating investigators in Padua for processing and entry into the data base (Colombo and Masarotto, 2000).

The bbt shift based on “three over six rule” (Marshall 1968) was used to identify the start of the infertile period following ovulation. The shift was defined as the first time in the cycle that three temperatures were recorded all of which were above the level of the immediately preceding six daily temperature recordings (Marshalls, 1968; Colombo and Masarotto, 2000). However, several exemptions were allowed in the study. i) if there

was one "spike" temperature among the six at the lower level (a spike temperature was defined as a temperature which was 0.2^0 C or more above both its immediate neighbouring temperatures); ii) or, a cycle in which the impact of illness or other disturbances could be discounted, if there were at least six lower temperatures recorded before the upward shift.

3.4 Descriptive Statistics

In this section we present the descriptive statistics for the study population. Descriptive summaries are presented at centre level. The 881 subjects had a total of 7017 menstrual cycles. Figure 3.1 is an example of a typical menstrual cycle record that was used to record/chart different demographic and biomarker characteristics in a menstrual cycle. These menstrual cycle characteristics include: dates and month, the daily characteristics for the mucus and bbt measurements, sequence of the days from the estimated ovulation reference day and days with intercourse within the menstrual cycle. The cross on the date indicates the peak mucus day. The first panel shows the mucus characteristics that are coded between 0 and 4. Code 0 represents no information; code 1 represents dry, rough and itchy feeling or nothing felt or seen and sometimes no mucus; code 2 represents damp feeling with nothing seen or no mucus; code 3 represents damp feeling characterized with cervical mucus that is thick, creamy/whitish/yellowish, not stretchy/elastic, sticky; code 4 represents mucus that is wet, slippery, smooth feeling and transparent like raw egg white, stretchy/elastic, liquid, watery, reddish (with some blood). The second panel shows the trend of the daily bbt measurements that varied between 36.0 to 37.0 degrees centigrade. The lower panel shows the days in the menstrual cycle when an intercourse occurred.

Tables 3.1 presents basic demographic characteristics for couples at the beginning of the study. These demographic characteristics include: ages of both men and women, percentages of women with past pregnancies and use of hormonal contraceptives. From the demographic characteristics table, Milan had the highest number of women. The average age for women ranged between 28 and 29 years while the average age for men ranged between 30 and 34 years. Dusseldorf had the youngest age-group for both men and women. The average proportion of women that had at least one pregnancy was 44.6 for European centres and 97 for the New Zealand group. The average proportions of women with past use of hormonal contraception were 30.1 and 34.3 for the European centres and New Zealand respectively. Verona had the least proportions for women that had previous pregnancies and used hormonal contraceptives.

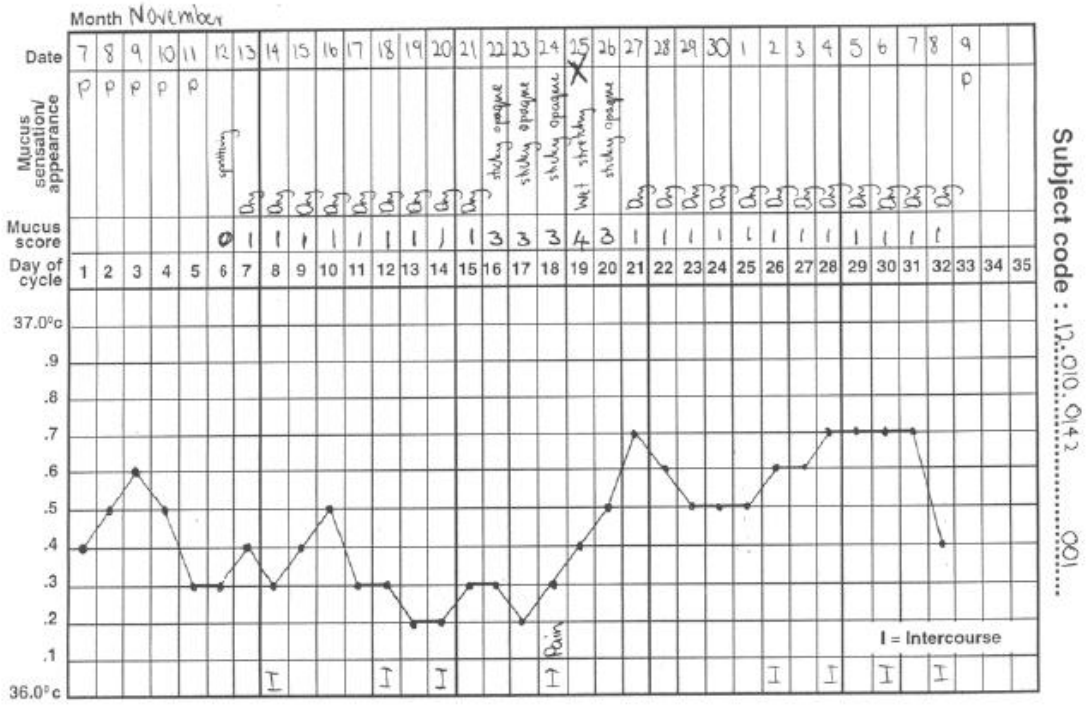

FIGURE 3.1: A typical menstrual cycle record chart.
Source: Colombo and Masarotto (2000)

Demographic characteristics of couples.					
Centres	No. of women	Age of women Mean(sd)	Age of men Mean(sd)	No. of women with at least past pregnancy (% of women)	No. of women with past use of hormonal contraception (% of women)
Verona	214	28.6 (3.54)	30.7 (4.16)	66 (30.8)	63 (29.4)
Milan	272	28.7 (3.56)	31.3 (4.73)	109 (40.1)	31 (11.4)
Lugano	13	29.3 (4.50)	32.1 (3.99)	5 (38.5)	4 (30.8)
Paris	104	29.3 (4.52)	31.4 (5.42)	76 (73.1)	38 (36.5)
Dusseldorf	105	28.2 (4.48)	30.4 (4.86)	44 (41.9)	59 (56.2)
London	45	31.6 (4.68)	34.0 (4.60)	29 (64.4)	24 (53.3)
Brussels	29	29.7 (4.52)	31.6 (3.78)	20 (69.0)	16 (55.2)
Total European	782	28.9 (4.00)	31.2 (4.70)	349 (44.6)	235 (30.1)
Auckland	99	29.9 (3.13)	32.3 (3.87)	96 (97.0)	34 (34.3)

TABLE 3.1: Characteristics of women and men participating in the study.
Source: Colombo and Masarotto (2000)

Table 3.2 shows the characteristics of menstrual cycles and their outcomes aggregated at centre levels. We include information on cycles with identified ovulation days based on the daily bbt measurements and the cervical mucus characteristics, cycles with at least one intercourse in the fertile window and the percentages of cycles with detected pregnancies and miscarriages. In particular, the ovulation identification was based on a pronounced bbt shift using the three over six rule (Marshalls, 1968). On average the percentage of menstrual cycles with a detectable bbt shift that identified the ovulation day were 96.4% and 94.8% for the European centres and New Zealand respectively.

Cycle characteristics						
Centres	No. of cycles	No. of cycles with identification of BBT ref. day (% of cycles)	Mucus reference (% of cycles)	No. of cycles with at least coition in the window	No. of detected pregnancy (% of cycles)	No. of miscarriages (% of pregnancies)
Verona	1279	1133 (97.9)	1246 (98.3)	827	171 (13.4)	11 (6.4)
Milan	3288	2840 (95.4)	3051 (95.8)	1351	151 (4.6)	20 (13.2)
Lugano	57	56 (98.2)	57 (100)	48	13 (22.8)	0 (0)
Paris	787	680 (95.8)	576 (74.0)	340	63 (8.0)	5 (7.9)
Dusseldorf	654	615 (97.8)	650 (99.4)	257	41 (6.3)	3 (7.3)
London	320	250 (95.8)	272 (96.1)	181	30 (9.4)	5 (16.7)
Brussels	339	286 (99.0)	314 (95.2)	171	18 (5.3)	3 (16.7)
Total European	6724	5860 (96.4)	6166 (94.1)	3175	487 (7.2)	47(9.7)
Auckland	293	238 (94.8)	285 (97.3)	215	88 (30.0)	2 (2.3)

TABLE 3.2: Characteristics of menstrual cycles and their outcomes.
Source: Colombo and Masarotto (2000)

Similarly, the percentage of cycles that had ovulation that could be determined using mucus characteristics were 94.1% and 97.3% for the two groups respectively. The average percentages from the European group are relatively lower than those from the New Zealand group due to low percentages from Paris subgroup. The number of cycles with at least one intercourse act in the fertile window are 3175 and 215; the number of detected pregnancy are 487 and 88 and the number of miscarriages of pregnancies are 47 and 2 for the European centres and New Zealand respectively. The 575 (i.e. 487+88) detected pregnancies in Table 3.2 include both those continuing at 60 days from the onset of the last menses and the 49 clinically recognized miscarriages of the same period (Colombo and Masarotto, 2000).

Table 3.3 presents characteristics of non-conception menstrual cycles based on the bbt ovulation reference day. These characteristics include the number of cycles in each centre, the total cycle lengths and the duration of both pre and post ovulation phases. The number of subjects and contributed cycles varied markedly between centres. In order to obtain meaningful fecundability patterns, the aggregate values in the tables for women from European centres were kept separate from those from New Zealand. Both groups (the European centres and New Zealand) had 5426 and 165 cycles respectively with a mean total length of 29 days per cycle. As expected, the lengths of the pre-ovulatory phase have relative higher variability than that of the post-ovulatory phase. On average the pre-ovulatory phase had a duration that varied between 16 and 18 days while the post- ovulatory phase had a duration that ranged between 12 and 13 days.

Table 3.4 gives the average number of intercourse acts per menstrual cycle in both conceptual and non-conceptual cycles based on the age-groups of the subjects (18-24, 25-29, 30-34, 35-39 and above 40 years). The averages are presented separately for both conceptual and non-conceptual cycles and grouped according to the age-groups of couples. The trend based on the age-groups was evaluated using the arithmetic average which is more preferred to the median for sake of better evidence. It is evident that the

Characteristics of non conception cycles.				
Centres	No. of cycles	Total length of cycles Mean(s.d.)	Duration of phases Pre-ovulatory Mean(s.d.)	Post-ovulatory Mean(s.d.)
Verona	982	29.0 (5.04)	16.4 (5.01)	12.6 (2.09)
Milan	2711	29.1 (3.89)	16.7 (3.93)	12.4 (2.09)
Lugano	44	27.2 (2.24)	14.7 (2.73)	12.5 (2.19)
Paris	620	29.3 (4.92)	17.1 (4.91)	12.2 (1.08)
Dusseldorf	574	28.3 (3.73)	16.3 (3.68)	12.0 (1.89)
London	224	29.8 (4.68)	17.2 (4.56)	12.5 (2.46)
Brussels	271	28.7 (3.63)	16.3 (3.74)	12.4 (1.94)
Total European	5426	29.0 (4.26)	16.6 (4.26)	12.4 (2.07)
Auckland	165	29.5 (4.37)	16.7 (4.64)	12.8 (2.36)

TABLE 3.3: Characteristics of non conception menstrual cycles with bbt reference day. Source: Colombo and Masarotto (2000)

Average number of intercourse acts.				
Age classes (years)	Intercourse of women in		Intercourse of men in	
	Conception cycles Mean (s.d.)	Non conception cycles Mean (s.d.)	Conception cycles Mean (s.d.)	Non conception cycles Mean (s.d.)
$18-24$	7.1 (3.19)	5.2 (3.10)	7.4 (3.86)	5.7 (3.47)
$25-29$	6.5 (3.08)	4.9 (2.82)	6.6 (3.17)	5.1 (3.08)
$30-34$	5.5 (3.03)	4.2 (2.73)	6.0 (3.00)	4.3 (2.54)
$35-39$	5.1 (2.30)	3.7 (1.96)	5.3 (2.65)	4.0 (2.52)
≥ 40			5.6 (2.62)	4.2 (2.19)
Total	6.2 (3.08)	4.5 (2.76)		

TABLE 3.4: Average number of acts of intercourse per menstrual cycle in the European centres. Source: Colombo and Masarotto (2000)

number of conceptual cycles is higher than the non conceptual cycles for both men and women. There is also a gentle decline in the frequency of intercourse with increase in the age for partners. The small variations between the male and the female findings, reflect differences on the number of subjects in various classes and on the whole study group. The higher coefficient of variation in non-conception cycles (61.3% vs. 49.7%), both support the reliability of the data collected (Colombo and Masarotto, 2000).

It was of interest to explore when the ovulation day occur within the menstrual cycle. Hence, we computed the percentages of cycles based on the day when ovulation was predicted to occur using the three over six rule (Marshalls, 1968). Figure 3.2 shows the distribution of the ovulation day (based on the three over six rule: refer to section 3.3) against the days of a menstrual cycle. It is evident that almost 14% of the menstrual cycles in the study have an ovulation day occurring on the 15 day of the cycle. The percentages lowers as we move away from the 15 day of the menstrual cycle.

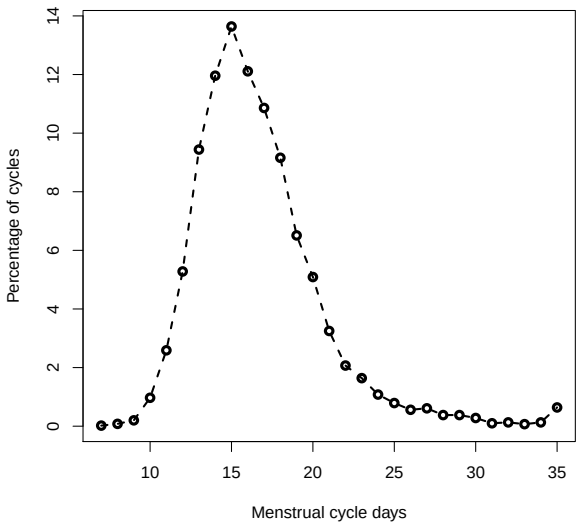

FIGURE 3.2: A graph for the percentages of the cycles against the occurrence of the ovulation day.

Chapter 4

A review on statistical methods

4.1 Introduction

This chapter gives a review of various methods that can be used to fit non-linear curves like the bbt trajectories. The chapter is divided into two main parts. The first part presents the existing literature on fitting non-linear curves. We start with a simple formulation of the bbt curve model and then give a brief review on various methods in the literature that can be adopted to estimate non-linear curves. In particular, our focus will be on non-parametric methods where we will highlight their use and briefly point out some advantages and limitations. The second part will discuss the existing literature on standard Bayesian approaches that can be used to estimate model parameters. We will give a review on how the standard Bayesian methods can be implemented to estimate the model parameters and highlight some of the problems that can be faced while using these standard approaches.

In the recent years, there has been an increased interest on researches involving correlated data that is characterized with curves or non-linear trajectories. In such cases data is collected sequentially from one or many subjects at small intervals over a period of time resulting to longitudinal correlated data. In the literature there exist many methods that can be used to model such data. For example, functional data analysis (FDA) is a common tool used to fit non-linear curves using high dimensional data models (Ramsay and Silverman, 2005). These approaches summarize the trend of the data into a curve, which is then used as a basic unit in data analysis. The features of the non-linear structure leads to the global curve having several segments joined together and weighted using carefully computed weights. In the next section we will consider a general functional data model to fit the basal body temperature data for one menstrual cycle.

4.2 Models for the bbt curve

The basal body temperature curves normally follow a trajectory that has a biphasic shape. During the follicular phase the bbt values tend to be low, with the nadir occurring close to the time of ovulation. Then after ovulation, the bbt values rises progressively before dropping prior to the next cycle. If there were no external factors interfering with the shape of the bbt curves, we can fit the curves using a simple parametric model. The model can be generated using a piecewise line composed of three parts: a first part defined as constant describing the low temperature plateau after menstruation, a second part linearly increasing, describing the sharp rise immediately following ovulation, and a third part constant, describing the high temperature plateau (Scarpa and Dunson, 2009).

However, the classic bbt pattern is difficult to replicate since there is a wide fluctuation in the bbt patterns from different cycles and subjects, therefore a parametric model cannot be adequate to fit the bbt curves. Moreover, if we assume a parametric model, it could be difficult to generate smooth curves and we would also be ignoring our uncertainty in specifying that model. A simple solution is to adopt a non-parametric approach that can accommodate uncertainty in model specification as well as generating smooth curves. Hence, non-parametric methods are commonly used since they offer more flexibility compared to parametric methods (Ruppert et al., 2003).

Let a menstrual cycle have T bbt measurements and be represented by a response vector $\boldsymbol{y} = (y_1, \cdots, y_T)'$. The covariate vector $\boldsymbol{z} = (z_1, \cdots, z_T)'$ contains the days of the menstrual cycle when the bbt measurements were collected while assuming that $z_1 < z_2 < \cdots < z_T$. A general functional model for the bbt curve can be written as,

$$y_j = f(z_j) + \epsilon_j, \qquad j = 1, \cdots, T, \tag{4.1}$$

where $f(.)$ is an unknown smoothing function at day z_j and $\{\epsilon_j\}$ are independent and normally distributed with error variance σ^2_ϵ, such that $\epsilon_j \sim N(0, \sigma^2_\epsilon)$.

The smoothing function $f(.)$ can be estimated by non-parametric approaches. Under non-parametric approach, the shape of the functional relationship between $f(.)$ and $\boldsymbol{y}$ is determined by the data while in parametric approach the shape is determined by a model. Non-parametric regression methods include: kernel methods (Wand and Jones, 1995), smoothing splines (Eubank 1988; Green and Silverman 1994), regression splines (Hastie and Tibshirani 1990; Friedman 1991) and penalized splines (Eilers and Marx, 1996; Ruppert, Wand, and Carroll, 2003). Kernel methods are mostly based on local likelihoods (Fan and Gijbels, 1996) while both smoothing splines and penalized splines

are based on penalized likelihoods. There is a strong connection between kernel and spline smoothing methods. Kernels methods and smoothing splines are asymptotically equivalent for independent data and splines are commonly viewed as higher-order kernels (Ramsay and Silverman, 2005).

4.2.1 Kernel smoothing methods

Kernel regression methods use local weighted averages to estimates a non-parametric regression function at a particular point ζ. Nadaraya-Watson estimator is one among many local weights used in kernel regression while local polynomials regression is the most commonly used method (Wand and Jones, 1994). Under the Kernel regression approach, the unknown smoothing function $f(z_j)$ in equation (4.1) is approximated locally around any arbitrary point ζ by a $d^{th} - order$ polynomial such that $f(z_j) \approx \alpha_0 + \cdots + \alpha_d(z_j - \zeta)^d = (\boldsymbol{z}_j - \zeta)^{'}\boldsymbol{\alpha}$ where $\boldsymbol{z}_j(\zeta) = \{1, \cdots, (z_j - \zeta)^d\}$ and the local weights $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_d)$. The weights $\boldsymbol{\alpha}$ are estimated by maximizing the local log-likelihood function

$$\frac{-1}{2\sigma^2}\sum_{j=1}^{T} K_h(\boldsymbol{z}_j - \zeta)\{(y_j - \zeta)^{'}\boldsymbol{\alpha}\}^2,$$

where $K_h(s) = h^{-1}K(s/h)$ such that h is a bandwidth and $K(.)$ is a kernel function. Most commonly chosen kernel functions have symmetric densities with mean zero e.g. Guassian, uniform and Epanechnikov kernels densities. The resulting kernel estimating equation is

$$\sum_{j=1}^{T} z_j(\zeta)K_h(z_j - \zeta)\{(y_j - \zeta)^{'}\boldsymbol{\alpha}\} = 0,$$

The orders of the kernel can be used to generate different estimators. For example, the local linear kernel estimator is of order $d = 1$ while the Nadaraya-Watson which is of order $d = 0$ results to

$$\hat{f}(\zeta) = \frac{\sum_{j=1}^{T} K_h(z_j - \zeta)y_j}{K_h(\boldsymbol{z}_j - \zeta)}.$$

Kernel smoothing methods mostly place more weight on observations when z_j values are in the neighbourhood of ζ, and less weight as observations move farther. Selection of the appropriate bandwidth h is important in kernel smoothing. Some of the most common selection methods are: cross-validation, plug-in estimators (Wand and Jones, 1994) and empirical bias bandwidth selection (Ruppert, 1997).

4.2.2 Smoothing splines

Smoothing splines is a curve fitting strategy that takes a spline with knots at every data point. It is used to estimates a non-parametric regression function $f(z)$ using a piecewise polynomial function with all observations used as knots (Green and Silverman, 1994). A spline function is a piece-wise polynomial with pieces defined by a sequence of $t_0, \cdots, t_k$ knots inside the range of the time points $1, \cdots, k$ and the pieces are joined smoothly at the knots. The most commonly used smoothing spline is the natural cubic smoothing spline, which assumes function $f(z)$ as a piecewise cubic function. A simple example of a cubic spline function $f(z)$ can be represented as a power series:

$$f(z) = p_3(z) + \sum_{l=1}^{k} \lambda_l (z - t_l)_+^3,$$

where $p_3(z)$ are cubic polynomial while λ_l are weights for $(z - t_l)_+^3$ such that

$$(z - t_l)_+^3 = \begin{cases} (z - t_l)^3 & if\ z > t_l, \\ 0 & otherwise. \end{cases}$$

Practically, a smooth line is estimated by minimizing the sum of squares errors plus a roughness penalty. A common approach on penalty is to integrate the squared second derivative, leading to minimizing the penalized least squares

$$\sum_{j=1}^{T} \{y_j - f(z_j)\}^2 + \tau \int_{g_1}^{g_2} \{f^{''}(z)\}^2 dz,$$

where g_1 and g_2 are the interval for which an estimate of $f(.)$ is sought, $f^{''}(z)$ is the second derivative of $f(z)$ and τ is the smoothing parameter that controls the mean square error and smoothness (Guo, 2004). As $\tau \to 0$ the effect of the penalty reduces to zero leading to a very close fit, but the curve follows every detail in the data leading to a very noisy curve. When parameter $\tau \to \infty$, the penalty dominates and the solution converges to the ordinary least square (OLS) line that has a very poor fit.

4.2.3 Regression splines and penalized splines

Smoothing splines become less practical when T is large leading to the use of $k = T$ knots. An alternative method is to use a regression splines, a curve fitting method where the number of knots are reduced using the least square methods. In particular, the number of knots is far much less than the number of data points ($k << T$). The

overall regression line is broken into $k+1$ line segments and each segment is linked to the subsequent segment at the knot to form a continuous line without discontinuities at the boundaries. Polynomial splines can easily be implemented under regression splines but the power series in their representation leads to computational problems due to high correlation between successive terms (for a more detailed reference see Ruppert, et al, 2003). To overcome computational problems, a more appealing representation of splines is a linear combinations of a set of basis splines called B-splines instead of polynomials. B-splines are spline functions that have minimal support with respect to a given degree, smoothness, and domain partition (deBoor, 2001).

Let $\boldsymbol{t} = (t_0, \cdots, t_{k-1})'$ be a knot vector such that $t_0 \leq t_1 \leq t_2, \cdots, t_{k-1} \leq t_k$. A B-spline function of degree n is a parametric curve composed of a linear combination of basis B-spline $B_{l,n}$ of degree n such that

$$f(z) = \sum_{l=0}^{k-n} P_l B_{l,n}(z), \qquad z \in [t_{n-1}, t_{k-n}]$$

where $P_0, P_1 \cdots, P_k$ are $k-n+1$ are the control points. The $k-n+1$ basis B-splines of degree n can be defined as

$$B_{l,0}(z) = \begin{cases} 1 & \textit{if } z \in [t_l, t_l + 1], \\ 0 & \textit{otherwise}. \end{cases}$$

$$B_{l,n}(z) = \frac{z - t_l}{t_{l+n} - t_l} B_{l,n-1}(z) + \frac{t_{l+n+1} - z}{t_{l+n+1} - t_{l+1}} B_{l+1,n-1}(z),$$

where $0 \leq k-n$ and $1 \leq n \leq k-1$. When the knots are located at the same distance, then the B-spline is called uniform, otherwise it is non-uniform. The use of spline demand additional computations, e.g. the computation of the number of knots k and determination of their location. The number of knots can be fixed or be placed at design points but a suitable number of knots should be select to allow flexibility in achieving a smooth curve (Botts and Daniels, 2008).

B-splines are easy to implement and available in many common statistical software. However, the functional form of B-splines is more complex compared to polynomials but they are the most commonly preferred splines due to their excellent numerical properties e.g. computational stability where each B-spline is non-zero over a limited range of knots (deBoor, 2001). One common problem with regression splines is determining where to position the knots. In many practical situations the knots are commonly placed at selected quintiles depending on the available number of knots. A more flexible strategy is to place more knots in regions where $f(.)$ is changing more rapidly to allow the smoothing function capture the trend of the data.

Penalized splines (P-splines) were introduced by Eilers and Marx (1996) for generalized linear smoothing. P-splines is a hybrid method that have important properties from both smoothing and regression splines. It combines the reduced number of knots ($k <<< T$) property in the regression splines and penalization of roughness that allows a smoothing curvature (Ruppert et al, 2003). The application of P-splines for smoothing data has been formulated using linear mixed model and extended into several areas including: generalized additive models (GAM) (Marx and Eilers, 1998), multivariate calibration and signal regression (PSR) (Marx and Eilers, 1999) and hybrid models that has building blocks chosen from GAM, PSR and varying-coefficient models (VCM) (Eilers and Marx, 2002). Ruppert et al (2003) has presented an excellent overview on applications and extensions of the penalized splines.

In many applications of Penalized splines, data smoothing is commonly done using B-splines with a discrete roughness penalty. Adequate number of basis function at equal spaced grid of knots are chosen to allow sufficient flexibility to avoid too much fluctuation (Eilers and Marx, 1996). Let $\boldsymbol{B}(z) = \{B_1(z), \cdots, B_k(z)\}$ be B-spline basis. A linear mixed model formulation of a penalized spline model has a mean function $f(z, \boldsymbol{\theta})$ represented as

$$f(z, \boldsymbol{\theta}) = \beta_0 + \beta_1 \boldsymbol{z}^{'} + \sum_{l=1}^{k} b_l B_l^{'}(\boldsymbol{z}) \tag{4.2}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{b}\}$ such that $\boldsymbol{\beta} = (\beta_0, \beta_l)^{'}$ are fixed effects while $\boldsymbol{b} = (b_1, b_2, \cdots, b_k)^{'}$ are random basis coefficients that have a distribution $b_l \sim N(0, \sigma_b^2)$. It is important to choose adequate number of basis functions k and in many practical applications, Ruppert (2000) recommended that a value between 20 and 40 is sufficient. The penalized least squares estimator leads to minimizing

$$\sum_{j=1}^{T} \{y_j - f(z, \boldsymbol{\theta})\}^2 + \tau \boldsymbol{\Omega},$$

where $\boldsymbol{\Omega}$ is symmetric, positive semi-definite $k \times k$ matrix such that $\boldsymbol{\Omega} = \boldsymbol{D}_m^T \boldsymbol{D}_m$ while $\boldsymbol{D}_m$ is the m^{th}-order differencing matrix and $\tau = \sigma_b^2/\sigma_\epsilon^2$ is the smoothing parameter. The differencing penalty is a discrete approximation to the integrated square of the m^{th} derivative of the B-spline smoother (Wand and Ormerod, 2008). The optimal solution for the penalized least squares estimator $\hat{\boldsymbol{b}}(\tau) = (\boldsymbol{B}^{'}\boldsymbol{B} + \tau\boldsymbol{\Omega})^{-1}\boldsymbol{B}^{'}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ where the penalty matrix gives additional continuous property that controls the smoothness of the estimated curve. The discreteness property of the penalty allows easy implementation in contrast to penalties that use the integral of squared higher order derivatives of the fitted function (O'Sullivan, 1986).

O'Sullivan penalized splines is a special type of P-splines that has been developed to

generalize smoothing splines that arise when maximal number of B-spline basis functions are included into the smoothing function (Wand and Ormerod, 2008). They have become one of the most widely used smoothing splines due to their attractive features e.g. natural boundary conditions (Green and Silverman, 1994), use of fewer basis functions, better numerical properties, availability in many popular statistical software, a straight forward representations in Bayesian models and their direct extension into more generalized models like generalized additive models (Wand and Ormerod, 2008).

In O'Sullivan penalized splines approach, we consider $\boldsymbol{B}(z)$ as cubic B-spline basis functions defined by the knots as described in Wand and Ormerod (2008). However, the penalty matrix changes from $\boldsymbol{\Omega}$ to $\boldsymbol{\Omega}^*$. The $(l, l^{'})$ element of the penalty matrix $\boldsymbol{\Omega}^*$ is defined by

$$\boldsymbol{\Omega}^*_{l,l'} = \int_{g_1}^{g_2} B_l^{''}(z) B_{l'}^{''}(z) dz,$$

where g_1 and g_2 are as described before. The computation of $\boldsymbol{\Omega}^*$ has been discussed by Wand and Ormerod (2008) and is readily available in many statistical software. The mean function in equation (4.2) can be generalized into linear mixed model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}, \tag{4.3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{b}_i$ are the coefficients for $\boldsymbol{X} = \{1, z_j\}_{j=1}^T$ and B-spline basis functions $\boldsymbol{Z} = \{B_1, \cdots, B_k\}$ respectively. The error terms in vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_T)^T$ have a multivariate normal distribution $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \boldsymbol{I})$. Computation of the coefficient estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ can be done either by using Bayesian or non-Bayesian methods. For example, in non-Bayesian framework the Best Linear Unbiased Predictor (BLUP) for the two sets of parameters result to

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{bmatrix} = \left(\boldsymbol{C}^T\boldsymbol{C} + \tau \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}^* \end{bmatrix} \right)^{-1} \boldsymbol{C}^T\boldsymbol{y}.$$

where $\boldsymbol{C} = \{\boldsymbol{X}|\boldsymbol{Z}\}$ and τ is the smoothing parameter. In Bayesian framework, parameters $\boldsymbol{\beta}$ and $\boldsymbol{b}$ in the models are treated as random variables. Prior distributions are assigned to parameter $\boldsymbol{\beta}$, $\boldsymbol{b}$ and τ in the model. The priors are then updated using the data likelihood to yield the anticipated posterior densities. The next section will discuss the implementation of the Bayesian methods to compute the posterior densities.

4.3 Bayesian methods

In this part we shall give a brief review on Bayesian methods. Tremendous improvement in computation mechanism within the last 20 years has made Bayesian models to gain

a lot of interest in data analysis work. Prior to that, Bayesian methodologies were out of reach due to complex models involved and computation burden. Typically, Bayesian statistic involves using probability distributions rather than point probabilities for all unknown quantities. Each parameter in the model is treated as a random variable and assumes some kind of distribution unlike the Frequentist approach that treats parameter as unknown constants. Several statistics text books have given a detailed review on both historical and data modelling methodologies e.g. Gelfand and Smith, (1990); Gill, (2002); Congdon, (2003); Gelman, et al., (2004); among others.

Bayesian inference is based upon Bayes' theorem to compute the posterior density which is a conditional density of unobserved quantity or parameter given the observed data. Let $\boldsymbol{y}$ be the observed data and $\boldsymbol{\theta}$ be the parameter of interest such that the parameters from the previous linear mixed model (4.3) can represented by $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{b}, \tau\}$. We assume that the probability model or conditional density for the data given parameter is $f(\boldsymbol{y}|\boldsymbol{\theta})$. The prior beliefs about the distribution of the parameter is expressed by a prior density $\pi(\boldsymbol{\theta})$. Using Bayes' theorem the posterior density for the parameters is expressed as a conditional density for $\boldsymbol{\theta}$ given $\boldsymbol{y}$ is $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and expressed as,

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{4.4}$$

Thus, we can summarize the computation procedure as $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ is the prior and $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood. The posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ can be computed analytically when the integration $\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is simple but in many practical applications, this is often a non-trivial problem requiring high-dimensional numerical integration. Recent research work has focused on this computation problem leading to development of computationally intensive numerical integration such as Markov chain Monte Carlo (MCMC) methods. Currently most attention is on Gibbs sampler and other related MCMC methods such as the Metropolis-Hastings algorithm (Hastings, 1970; Geman and Geman, 1984; Gelfand and Smith, 1990). Before discussing these recent computation methods, we shall briefly review the choice of prior π.

4.3.1 Choice for prior

Inference based on Bayes' theorem encounters controversies especially with the choice and interpretation of the prior π. The choice of prior density π depends on the prior information available about the data or the model used for data analysis. When reliable information is available, then *"informative"* priors are normally used as a building tool for the prior distribution. Otherwise when there is no reliable information about a

parameter, selection of a prior density π is a delicate step and in most cases it is common to choose *"uninformative"* prior densities.

In spite these controversies, the choice of prior π is very crucial if the goal is to arrive at an analytically feasible posterior density. The choice involves forming an important modelling assumption about the nature of the distribution of the parameters. In practice, *"Uninformative"* priors with large variance are commonly chosen to minimize the impact of the selected prior on the inference. Moreover, to make computation feasible, most common approaches exploit the mathematical relations among probability distributions leading to analytically feasible posterior densities that have known functional form. For example, many examples in most Bayesian textbooks work with exponential family of distributions (e.g. Gaussian, Gamma, Geometric, Poisson, Multinomial, etc.) since they have conjugate prior densities that result to analytically feasible posterior.

As described in Marin and Robert (2007), let an exponential family be described as,

$$f_\theta(\boldsymbol{y}) = h(\boldsymbol{y}) \exp\{\boldsymbol{\theta}.R(\boldsymbol{y}) - \psi(\boldsymbol{\theta})\}, \qquad \boldsymbol{\theta}, R(\boldsymbol{y}) \in \Re^p \tag{4.5}$$

where $\boldsymbol{\theta} \cdot R(\boldsymbol{y})$ in equation (4.5) denote a canonical scalar product in $\Re^p$. There exists a class of generic class of priors called conjugate priors,

$$\pi(\boldsymbol{\theta}|\xi, \eta) \exp\{\boldsymbol{\theta} \cdot \xi - \eta\psi(\boldsymbol{\theta})\},$$

which are parameterized by two quantities, $\eta > 0$ and ξ, that are of the same family as $R(\boldsymbol{y})$. These parameterized prior distributions on $\boldsymbol{\theta}$ are derived in a manner that the posterior distributions are of the same form

$$\pi(\boldsymbol{\theta}|\xi^{'}(\boldsymbol{y}), \eta^{'}(\boldsymbol{y})),$$

where $(\xi^{'}(\boldsymbol{y}), \eta^{'}(\boldsymbol{y}))$ is defined in terms of the observation $\boldsymbol{y}$. A prior density is said to be a natural conjugate with respect to a likelihood if it gives rise to a posterior density having the same parametric form as that of the prior. Hence, the conjugate priors allow both the prior and the posterior distributions to belong to the same parametric family of densities, though they have different parameters (Marin and Robert, 2007). This implies that the resulting estimates of the parameters in the posterior are just updates of parameters in the prior distribution by the information from the observations y. For example, a beta prior $\pi(\boldsymbol{\theta})$ and a binomial likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ yield a beta posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ while a normal prior and likelihood results to a normal posterior density. For more information about the choice of priors refer to Marin and Robert (2007) chapter 2.

4.3.2 MCMC methods

In many Bayesian problems, the computation of the posterior distribution often requires integration of high-dimensional functions which cannot be explicitly computed. MCMC methods provide ideal process to integrate high-dimensional functions. In brief, Markov chain Monte Carlo involves setting up a Markov chain in parameter space $\boldsymbol{\theta}$ with ergodic distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Then starting with some initial state $\boldsymbol{\theta}^0$, we simulate G transitions under this Markov chain and record the simulated states $\boldsymbol{\theta}^g$, where $g = 1, \cdots, G$ (Robert and Casella, 1999). MCMC methods use previous simulated values to randomly generate the next sample value. The transition probabilities between successive samples are functions of the most recent samples leading to generate a Markov chain. Several approaches of direct integration based on MCMC methods have been proposed. The most commonly used methods include: Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) and Metropolis-Hastings algorithms (Chib and Greenberg, 1995).

Historically, MCMC methods originated from Metropolis algorithm (Metropolis and Ulam 1949, Metropolis et al. 1953). This was an attempt by physicists to compute complex integrals by expressing them as expectations for some distribution and then estimate this expectation by drawing samples from that distribution (for a more detailed historical review see, Smith 1991, Evans and Swartz 1995, Tanner 1996). The Metropolis-Hastings algorithm was first introduced as Metropolis algorithm by Metropolis et al. (1953) and latter generalized by Hastings (1979). The algorithm has many applications and is commonly used for numerical integration and optimization. For a thorough review of the Metropolis-Hastings algorithms, the fundamental theory are well discussed in Chib and Greenberg (1995).

To sample from a target distribution $\pi(.) = \pi(\boldsymbol{\theta}|\boldsymbol{y})$, we start with sensible starting values $\boldsymbol{\theta}^0$ and run a Markov chain with transition matrix satisfying $\pi(i)P_{ij}(.) = \pi(j)P_{ji}(.)$ until the chain settles down to an equilibrium. For the g^{th} iteration of the $g = 1, ..., G$ simulations, we draw a proposal $\boldsymbol{\theta}^*$ from a known proposal distribution or instrumental function $P_g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{g-1})$. The proposal function can be any density function that is typically easy to simulate from than the target density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Moreover, the chosen proposal distribution should allow the chain explores the posterior distribution adequately. Common proposal densities are uniform distribution, multivariate normal or multivariate-t that is centred at the current location of the chain.

The realisations for the g^{th} iteration is $\boldsymbol{\theta}^g$ and can assume a proposal $\boldsymbol{\theta}^*$ with probability α^* or the previous realization $\boldsymbol{\theta}^{g-1}$ with probability $1-\alpha^*$. The probability α^* is defined as

$$\alpha^* = \min\{\frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})}{\pi(\boldsymbol{\theta}^{g-1}|\boldsymbol{y})}\frac{P_g(\boldsymbol{\theta}^{g-1}|\boldsymbol{\theta}^*)}{P_g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{g-1})}, 1\}.$$

The Metropolis-Hastings algorithm is relatively easy to implement since it requires the target density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ to be defined up to the normalising constant. The constant is usually dropped in the ratio $\pi(\boldsymbol{\theta}^*|\boldsymbol{y})/\pi(\boldsymbol{\theta}^{g-1}|\boldsymbol{y})$. The proposal $\boldsymbol{\theta}^*$ is automatically accepted when the ratio $\pi(\boldsymbol{\theta}^*|\boldsymbol{y})/P_g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{g-1})$ is increased relative to the previous $\pi(\boldsymbol{\theta}^{g-1}|\boldsymbol{y})/P_g(\boldsymbol{\theta}^{g-1}|\boldsymbol{\theta}^*)$. Hence, the performance of the acceptance rate depends on the choice of a particular proposal function. However, it is not necessary that high acceptance rates indicate that the algorithm is moving correctly. In some cases, such trends may indicate that the random walk is moving too slowly on the surface of the targeted distribution (Tanner 1996; Robert and Casella, 1999). For a simple one dimensional case, Gelman et al, (1996) suggested that an optimal jumping rule has an acceptance rate slightly under 0.5, but if the dimension of the parameter vector exceeds 5 the rate normally decrease to about 0.25. A heavier tailed proposal function that includes the support for the target distribution should be adopted to improve acceptance rates as well as reaching the targeted distribution fast. For an interested reader, Robert and Casella, (1999) has reviewed different approaches of choosing $P_g(.|.)$.

The Gibbs sampling algorithm (Gelman and Gelman, 1984) is another method similar to the Metropolis-Hastings algorithm. Gibbs sampling algorithm is simpler relative to Metropolis-Hastings algorithm and is the most implemented MCMC sampling method when the posterior distribution can be expressed in a fully conditional form. The use a Gibbs sampler leads to a sequence of draws from conditional distributions to characterize the joint target distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Suppose the targeted distribution has p parameters such that $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\theta_1, \theta_2, \cdots, \theta_p|\boldsymbol{y})$, the sampling process begins sampling θ_1 and choosing the starting values for the remaining $p-1$ parameter. In many practical situations, the starting values $\theta_2^0, \cdots, \theta_p^0$ are commonly chosen near the posterior mode or the maximum likelihood estimates. The whole process involves a repeated sequences of $g = 1, \cdots, G$ iterations.

In general, we simulate θ_i from $\pi_g(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{y})$, which is called the full conditional distributions. The distribution is a function of θ_i and is obtained by ignoring all components in $\boldsymbol{\theta}_{-i}$ that do not depend on θ_i. Some of the most commonly used conditional distributions are multivariate normal, truncated normal, Gamma, etc. For a simple case that involves conjugate prior distributions, the full conditionals reduce to closed form distributions that is easy to simulate from. However, many practical situations involve complex models with full conditionals that cannot simplify to any analytically tractable expressions (Gilks et al, 1996). For a posterior density that has p parameters and can be expressed as a full conditional distributions, the Gibbs sampler at the g^{th} iteration works as follows,

$$
\begin{aligned}
&\theta_1^g \sim \pi_g(\theta_1|\theta_2^{g-1}, \cdots, \theta_p^{g-1}, \boldsymbol{y}),\\
&\theta_2^g \sim \pi_g(\theta_2|\theta_1^{g}, \theta_3^{g-1}, \cdots, \theta_p^{g-1}, \boldsymbol{y}),\\
&\theta_3^g \sim \pi_g(\theta_3|\theta_1^{g}, \theta_2^{g}, \theta_4^{g-1}, \cdots, \theta_p^{g-1}, \boldsymbol{y}),\\
&\cdots\\
&\cdots\\
&\cdots\\
&\theta_p^g \sim \pi_g(\theta_p|\theta_1^{g}, \theta_2^{g}, \cdots, \theta_{p-1}^{g}, \boldsymbol{y}).
\end{aligned}
\tag{4.6}
$$

The current sampled value $\boldsymbol{\theta}^g$ is always conditioning on past draws which leads to a sequence of samples in a Markov chain. To compute the posterior means, the Monte Carlo process require a large number of G iterations (e.g. $10,000$). It is a common practice to discard realizations for the first set (e.g $1,000$) of "burn-in" iterations. This helps to ensure that the posterior means are not influenced by the initial values that were assigned to the parameters. For more explanations on the choice of starting values and the number of iterations to be discarded, detailed reviews have been documented in Gelman and Rubin (1992); Casella and George (1999); Gelfand (2000) among others.

4.3.3 Diagnostic tools for MCMC methods

Inferences based on posterior density summaries that are commonly are generated using MCMC methods, cannot be trusted unless the Markov chain has reached steady state. In Bayesian analysis, we monitor the MCMC outputs to assess convergence to a distribution. This calls for specialized diagnostic tools to evaluate if the sampling process has generated a representative sample from the anticipated posterior distribution. However, since diagnostic tests do not provide proof of convergence, it is prudent to employ more than one method when assessing the quality of samples from an MCMC algorithm (Smith, 2007). There exist a lot of convergence assessment tools in the literature. Most convergence diagnostic tools evaluate the marginal posterior distributions. For example, some methods use different tests to assess the traceplots of outputs from the MCMC sampler and test if they are stationary. Others methods compare multiple runs from different starting values and using different random number seeds to determine whether the chains have converged (Gelman and Rubin, 1992). For an interested reader, Cowles & Carlin (1995) and Gelman, et al. (2004) have given classical reviews on various convergence diagnostic methods.

One of the commonly used convergence diagnostic tool is the Gelman & Rubin (1992) method. The method was first introduced to assess the convergence of individual model parameters based on the computation of two statistics: the potential scale reduction

factor (PSRF) and the corrected scale reduction factor (CSRF). The PSRF is used to measure convergence by comparing between-chain and within-chain variances while CSRF is used to account for the sampling variability in the estimates for the parameter of interest (Smith, 2007). The PSRF statistics is computed from m independent chains using the last n samples leading to

$$PSRF = \sqrt{\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}},$$

where B/n is the between-chain variance while W is the within-chain variance. The between-chain variance (B/n) should be smaller than the within-chain variance (W) leading to a PSRF that approaches 1. Any value larger than 1 suggests that convergence has not been attained. The CSRF accounts for the sampling variability and is computed as

$$CSRF = \sqrt{\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}}\sqrt{\frac{df+3}{df+1}},$$

where df represents the degrees of freedom based on a t approximation in the posterior inference. For more details, refer to Gelman & Rubin (1992) and Cowles & Carlin (1995) and Smith (2007).

The Raftery and Lewis (1992) method uses univariate analysis of a single parameter and chain. The method is used to determine the number of iterations that should be run. This is computed using the standard sample size formulas based on binomial variance (Cowles & Carlin, 1995). Other diagnostic methods of interest include Geweke (1992), Ritter & Tanner (1992), Roberts (1994), Yu & Mykland (1994) and Mykland, Tiermey,& Yu (1995).

4.3.4 Issues of concern with MCMC methods

The MCMC based algorithms are commonly used in many Bayesian data analysis. Over the last decade a number of software packages like WinBUGS (Thomas, Best, and Spiegelhalter, 2000), OpenBUGS (Thomas, O'Hara, Ligges, and Sturtz, 2006) and some libraries (e.g. R2WinBugs) in R (R Development Core Team 2006) have been developed to allow easy implementation of Bayesian methods. However, Bayesian analysis methods are considered by many to be too slow and far more complicated than classical non-Bayesian statistical methods. This might be one of the reasons why many commercial software for statistical analysis such as SAS, SPSS and Stata are mostly focused on non-Bayesian statistical methods.

One of the major setbacks of Bayesian methods is their dependence upon MCMC based algorithms that spend a lot of time when computing posterior estimates. Moreover, most

of the commonly used MCMC algorithms require fast and strong machines that have large storage capacity. For example, when implementing model selection procedures such as reversible jump (Green, 1995) or stochastic search variable selection (Smith and Kohn, 1996), these algorithms are computationally expensive, requiring large memory, fast processor and involving hours or days to implement especially when working with data sets that involve thousands of subjects. This is a major concern in data analysis and has lead to the development of fast approximate methods that bypasses MCMC algorithms. For example, in model selection procedures, most of the methods used are based on well known approximations methods like expectation maximization (EM) algorithm (Tipping, 2001) and the variational approximation (Faul and Tipping, 2002). Both expectation maximization and variational approximation approaches can be accelerated and they normally provide similar results.

Recently a fast and computationally efficient Bayesian procedure called Relevant Vector Machine (RVM) (Tipping, 2001) method has been introduced as an alternative. RVM approach is an Empirical Bayes method that promotes sparseness in estimation of the basis coefficients by computing the priors from the data and use classical parameter estimation methods to compute other model parameters while maintaining some of the benefits of Bayesian analysis methods. Relevance vector machine approach has similar functional form as the popular Support Vector Machine (SVM) in that it forms a linear combination of data-centred nonlinear basis functions, but has an advantage in providing a more flexible alternative in reducing the computational complexity of the Support Vector Machines (SVM) (Burges, 1998). The RVM method is also thought to have advantages over other sparse model generating methods like LASSO (Tibshirani, 1996) in leading to sparser models that are more robust to outliers (Tipping, 2001; Ji, et al, 2009). The implementation and extension of the Relevant Vector Machine (RVM) procedure into linear, linear mixed and multi-level models will be discussed in the next chapters.

Chapter 5

Fast bayesian functional data analysis of basal body temperature

5.1 Introduction

In this chapter, we present a general review of the link between the recent methods used to model the bbt curves and functional data analysis. In particular, the focus is on functional random coefficients and linear mixed models. This review include information on both Bayesian and non-Bayesian methods. We start with an introduction of a simple functional random coefficients model, followed by a motivation and a theoretical background of the Multi-task Relevant Vector machine (MT-RVM) method of Ji, et al (2009) and then implement the method to estimate the bbt curves. To demonstrate the usability of the approach, we give an extension of the MT-RVM method in functional linear mixed models that can be used to estimate both the population mean curve as well as subject-specific deviations.

In reproductive studies, tracking measurement patterns of hormonal level or daily basal body temperature (bbt) among women can help to identify or predict early pregnancy loss and occurrence of the ovulation day. In recent years there has been an increased interest on researches on the distribution of random curves describing such patterns (Collins, 1996; Dunson, *et al.*, 1999; Bigelow and Dunson, 2008). For example, it is known that a standard bbt curve from a healthy ovulating female has a biphasic pattern. This is characterized by a low plateau during the follicular phase, a temperature dip that occur prior to ovulation, and a sharp rise immediately after ovulation which is subsequently followed by a luteal phase plateau (e.g., Vincent, 1964; Marshall, 1979).

This characteristic pattern has been used to identify the ovulation day, since several studies have suggested that ovulation day corresponds to the low point prior to the rise in bbt (Marshall, 1979; Colombo and Masarotto, 2000). This result is often used as a basis for the identification of the fertile period during a cycle; in fact the probability of conception is lower when intercourse occurs outside of the six-day fertile interval ending on the ovulation day (Dunson et al., 1999). Therefore the estimate of a smooth trajectory for bbt over the menstrual cycle is of great interest for natural family planning, clinical and epidemiological applications.

In particular, it is convenient to rapidly predict features of interest based on data extracted from large database that involve many subjects. Unfortunately, due to heterogeneity among subjects, data collection problems, data entry and storage errors, it is common to have subjects that have sparse and unequally spaced measurements, and simple statistical tools seem not to work well. Royston and Abrams (1980) and Carter and Blight (1981) proposed approaches for estimating the shift in bbt as a marker of ovulation and Scarpa and Dunson (2009) uses a mixture of parametric and non parametric models to fit the curves, by clustering different shapes of functions.

Functional data analysis (FDA) is another ideal tool to estimate a smooth trajectories resulting from the bbt data characterized with sparseness and unequal cycle lengths. The main aim of FDA is to explore and highlight important features of a curve. A trajectory may consist of one or several segments weighted using functional coefficients (Ramsay and Silverman, 2005). Unfortunately, FDA relies on relatively large number of basis functions and estimation of functional coefficients becomes time consuming activity using standard software. A common approximation procedure in FDA is to consider only a subset of carefully chosen basis functions that can be used for approximation purpose. However, it can be difficult to choose the basis functions in advance, motivating the use of adaptive methods that allow uncertainty in basis function selection (Bigelow and Dunson, 2007; Johnson and Rosen, 2008).

In functional random coefficients models, various model reduction procedures have been introduced. In Bayesian framework, basis functions selection commonly rely on reversible jump algorithms (Green, 1995) or stochastic search variable selection methods (Smith and Kohn, 1996) that are computationally intensive. For example, a Bayesian versions of Multivariate adaptive regression spline (MARS) has been proposed by Denison *et al*, (2002) to automatically select the basis functions. The method has good performance in small to moderate dimensional random coefficients models but the posterior sampling is based on the slow reversible jump Markov Chain Monte Carlo (RJMCMC). The use of RJMCMC involves MCMC that requires hours to implement in data sets

involving thousands of subjects. Hence, raising a practical motivation for fast approximate Bayes approaches that bypass MCMC while maintaining some of the benefits of a Bayesian analysis. Recent MARS extensions (Sakamoto, 2007), also bypasses the MCMC algorithms and involves the use of an empirical Bayes approach for selecting basis functions and knots.

In functional linear mixed model, the selection of the random effects has complication since the null hypothesis lies on the boundary of the parameter space and the classical likelihood ratio test statistic no longer valid. To solve this problem Pauler et al. (1999) and Jiang, et al. (2008) introduced alternative approaches to reduce the dimension of the random effects model. Recent methods use functional principal component analysis (James, Hastie and Sugar, 2001; Yao, Muller, and Wang, 2005; Crainiceanu, 2009). The approaches have good performance in modest dimensional models with moderate number of subjects, but rapidly becomes computationally infeasible as the number of the subjects and candidate predictors increases.

In this chapter, we proposes to use a fast Bayesian methodology by approximating the bbt projectiles using Multi-Task Relevant Vector machine (MT-RVM) method an extension of Relevant Vector machine (RVM) method (Tipping, 2001; Ji, *et al*, 2009). RVM is a fast Bayesian method based on Empirical Bayes methodology and has featured mostly in Machine Learning especially in signal reconstruction and compressive sensing. Relevant Vector machine is one among several methods that promote sparseness in estimation of functional coefficients. Other similar methods include; Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM) (Tibshirani, 1996; Burges, 1998). Sparseness is a property where the fitted model retains the least number of basis functions (by having non-zero coefficients), while all the other basis functions are pruned by setting their corresponding coefficients to zero (Tzikas, *et al*, 2005). This property provides a natural mechanism in variable selection leading to a sparse model that is fast to compute.

We apply the proposed method to the basal body temperature (bbt) data for the European fecundability study (Colombo and Masarotto, 2000). The data are characterized by missing temperature measurements in some days and there is variability in cycle lengths among women such that, majority have cycles that ranges between 20 to 40 days. Besides measurement errors which is a common problem with many longitudinal data, unequal cycle lengths and data sparsity pose a great hindrance to analyze the data using most available standard software.

5.2 RVM in random coefficients model

Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})'$ and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iT_i})'$ be vectors for the response and covariates for the i^{th} woman. A functional model can be represented as

$$y_{it} = f_i(z_{it}) + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \qquad t = 1, \ldots, T_i, \qquad i = 1, \ldots, N. \tag{5.1}$$

where $f_i(.)$ is a smooth function at z_{it} for subject i and ϵ_{it} is a measurement error. The functional model in (1) can be represented as random coefficients or linear mixed model depending on whether the interest is on either subject-specific curves or both the population and subject-specific curves.

When the interest is to model the subject-specific curves, the functional model in (5.1) can be represented as a random coefficients model such that the smoothing function is described as a linear combination of M basis functions

$$f_i(z_{it}) = \sum_{j=1}^{M} \beta_{ij}\varphi_j(z_{it}) = \mathbf{x}_{it}{}'\boldsymbol{\beta}_i, \tag{5.2}$$

where $\mathbf{x}_{it} = (x_{it1}, \cdots, x_{itM})'$ are the values of the basis functions at z_{it}, parameter β_{ij} is the coefficient for the j^{th} basis function $\varphi_j(.)$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \cdots, \beta_{iM})'$. The basis functions φ_j can be generated using numerous methods that have been discussed in the literature (e.g.Hastie, et al. 2001; Ruppert, et al. 2003). Conditionally on the basis $\boldsymbol{\varphi}$, expression (5.1) can be expressed in the form of a random coefficients model.

The priors are $\beta_{ij} \sim N(0, \alpha_j^{-1})$, $\sigma_\epsilon^{-2} \sim Gamma(a, b)$ and $\alpha_j \sim Gamma(c, d)$. The parameters $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_M)'$ and σ_ϵ^{-2} are computed from the data as *maximum a posteriori* (MAP) estimates. Since these MAP estimates are estimated and shared among all the subjects, this leads to borrowing of strength across subjects in estimating subject-specific functions. To promote sparseness over the model coefficients $\boldsymbol{\beta}_i$, the hyper-parameters c and d are set close to zero leading to a distribution with a large spike concentrated at zero and a heavy right tail. The basis functions for which α_j is in the right tail have coefficients that are strongly shrunk toward zero.

5.2.1 Posterior Estimates

The commonly used approach to compute the joint posterior density $p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ cannot be implemented since the computation of the posterior density require a normalization that cannot be expressed analytically. An alternative approach is to compute

the posterior density based on the conditional distribution

$$p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y}) = p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y}) \prod_{i=1}^{N} p(\boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{\alpha}, \sigma_\epsilon^{-2}), \tag{5.3}$$

where $\boldsymbol{Y} = (\boldsymbol{y}, \cdots, \boldsymbol{y}_N)^{'}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots \boldsymbol{\beta}_N)$. The density function $p(\boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ is the posterior distribution for the random coefficients $\boldsymbol{\beta}_i$, while $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ is the posterior density for the variance components $\boldsymbol{\alpha}$ and σ_ϵ^{-2}.

The posterior density for the random coefficients $\boldsymbol{\beta}_i$ is a multivariate normal distribution

$$p(\boldsymbol{\beta}_i|\boldsymbol{Y}, \boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = \boldsymbol{N}(\boldsymbol{\beta}_i; \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i), \tag{5.4}$$

where $\hat{\boldsymbol{\mu}}_i = \sigma_\epsilon^{-2}\hat{\boldsymbol{\Sigma}}_i \boldsymbol{X}_i^{'}\boldsymbol{y}_i$ is the mean vector and $\hat{\boldsymbol{\Sigma}}_i = (\boldsymbol{A}+\sigma_\epsilon^{-2}\boldsymbol{X}_i^{'}\boldsymbol{X}_i)^{-1}$ is the covariance matrix such that $\boldsymbol{A} = diag\{\alpha_1, \cdots, \alpha_M\}$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{iM})^{'}$.

Since it is impossible to express the posterior density for the variance components $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ analytically, we use an Empirical Bayes approach to compute the posterior estimates for $\boldsymbol{\alpha}$ and σ_ϵ^{-2}. These estimates are computed as MAP estimates as will be discussed in the next section. The density function $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma_\epsilon^{-2})p(\boldsymbol{\alpha})p(\sigma_\epsilon^{-2})$ and both $p(\boldsymbol{\alpha})$ and $p(\sigma_\epsilon^{-2})$ are assumed to be Gamma density. To compute the estimates for $\boldsymbol{\alpha}$ and σ_ϵ^{-2}, we assume that the modes for $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ and $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma_\epsilon^{-2})p(\boldsymbol{\alpha})$ are equivalent and hence the MAP estimates for $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ are equivalent to the MLE estimates from $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ (Ji et al, 2009).

5.2.2 Empirical Bayes method

Expressing the posterior density for the variance components $p(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ is difficult analytically and the MAP estimates for $\boldsymbol{\alpha}$ and σ_ϵ^{-2} are computed from the marginal likelihood $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$, obtained after integrating out $\boldsymbol{\beta}_i$ from $p(\boldsymbol{Y}|\boldsymbol{\beta}_i, \sigma_\epsilon^{-2})$ such that

$$p(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = \int \prod_{i=1}^{N} p(\boldsymbol{Y}|\boldsymbol{\beta}_i, \sigma_\epsilon^{-2})p(\boldsymbol{\beta}_i|\boldsymbol{\alpha})d\boldsymbol{\beta}_i$$

This results to a normal density function $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = N(\boldsymbol{y}_i; \boldsymbol{0}, \boldsymbol{C}_i)$ where the covariance matrix $\boldsymbol{C}_i = \sigma_\epsilon^2 \boldsymbol{I}_{T_i} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{'}$. The expressions for the estimates of $\boldsymbol{\alpha}$ and σ_ϵ^{-2} are obtained from the log-likelihood function $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = \sum_{i=1}^{N} \log N(\boldsymbol{y}_i; \boldsymbol{0}, \boldsymbol{C}_i)$. We

differentiate the log-likelihood $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ with respect to parameters $\boldsymbol{\alpha}$ and σ_ϵ^{-2} respectively and equating the resulting score equations to zero. This results to

$$\hat{\alpha}_j = \frac{N}{\sum_{i=1}^{N} \mu_{ij}^2 + \boldsymbol{\Sigma}_{i,jj}}, \qquad j = 1, \cdots, M \tag{5.5}$$

$$\hat{\sigma}_\epsilon^{-2} = \frac{\sum_{i=1}^{N} ||\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\mu}_i||^2}{\sum_{i=1}^{N} (T_i - M - \boldsymbol{\alpha}^{-1} \boldsymbol{\Sigma}_{i,jj})}. \tag{5.6}$$

The estimates for $\boldsymbol{\alpha}$ and σ_ϵ^{-2} are inserted into $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ in equation (5.4) leading to an interactive procedure alternating between estimation of the parameters in equation (5.4) and (5.5-5.6) respectively.

However, two related problems arise while implementing the above empirical Bayes approach when the number basis functions is large. These problems are: estimability problems leading to lack of convergence and the computation process require large amount of time especially when dealing with large data sets. Such computation difficulties are commonly encountered when the dimension of the basis functions is large relative to the sample size, which is a common practice in many functional analysis work. For example, when M is large (e.g. $M > 10$), the inversion of $M \times M$ covariance matrix $\boldsymbol{\Sigma}_i$ becomes impossible leading to estimability problems resulting to lack of convergence of the procedure. Moreover, when the dataset consists of thousands of subjects, the computation process may take days.

Potentially these problems can be solved by using a MAP estimation approach that includes a proper prior to induce a penalty in the procedure that leads to shrinkage towards the prior and regularization. However, such an approach will be sensitive to hyper-parameter choice. An alternative approach is to adapt a fast algorithm that leads to a reduced model with dimension $m \times m$ for $\hat{\boldsymbol{\Sigma}}_i$ where $m << M$. The RVM iterative algorithm can generate such a sparse model and will be discussed in the next section.

5.2.3 A Fast Empirical Bayes method

Conditioning on the MLE estimates for σ_ϵ^{-2}, a fast approach to compute the elements of $\boldsymbol{\alpha}$ can be done sequentially. The algorithm is based on the dependence of the k^{th} component of $\boldsymbol{\alpha}$ upon the log-likelihood function

$$\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = -1/2 \sum_{i=1}^{N} \{T_i \log(2\pi) + \log |\boldsymbol{C}_i| + \boldsymbol{y}_i \boldsymbol{C}_i^{-1} \boldsymbol{y}_i\}. \tag{5.7}$$

However, the presence of matrix $\boldsymbol{C}_i$ in the log-likelihood function $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$ makes it impossible to express the log-likelihood function into two parts -one containing the k^{th}

component and the other one that does not. Hence, to allow such decomposition, we first decompose the variance matrix $\boldsymbol{C}_i$ into two part -with and without the contribution of the k^{th} basis function. This leads to $\boldsymbol{C}_i = \boldsymbol{C}_{i,-k} + \alpha_k^{-1}\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^{'}$ where $\boldsymbol{C}_{i,-k}$ is the part that does not have the contribution of the k^{th} basis function. The resulting decomposed log-likelihood function is

$$\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2}) = \ell(\boldsymbol{\alpha}_{-k}, \sigma_\epsilon^{-2}) + \frac{1}{2}\sum_{i=1}^{N}\left(\log\alpha_k - \log|\alpha_k + s_{ik}| + \frac{q_{ik}^2}{\alpha_k + s_{ik}}\right)$$

where $\ell(\boldsymbol{\alpha}_{-k}, \sigma_\epsilon^{-2})$ is the part without the contribution of the k^{th} basis function, $s_{ik} = \boldsymbol{x}_{ik}^{'}\boldsymbol{C}_{i,-k}^{-1}\boldsymbol{x}_{ik}$ and $q_{ik} = \boldsymbol{x}_{ik}^{'}\boldsymbol{C}_{i,-k}^{-1}\boldsymbol{y}_i$. Differentiating $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^2)$ with respect to α_k and setting the result to zero yield the score equations,

$$\frac{\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2})}{\partial\alpha_k} = \sum_{i=1}^{N}\frac{s_{ik}^2/\alpha_k + s_{ik} - q_{ik}^2}{2(\alpha_k + s_{ik})^2} = 0.$$

The solutions for the score equations are infeasible to express analytically except for a trivial case when $\alpha_k = \infty$. The exact solutions require finding the zeros of a polynomial of degree $2N - 1$ which is computationally expensive. An alternative method to avoid such computation complexities is to assume that $\alpha_k << s_{ik}$ where $s_{ik} = \boldsymbol{x}_{ik}^{'}\boldsymbol{C}_{i,-k}^{-1}\boldsymbol{x}_{ik}$, leading to the approximate estimate

$$\hat{\alpha}_k \cong \begin{cases} \frac{N}{\sum_{i=1}^{N}(q_{ik}^2 - s_{ik})/s_{ik}^2} & if\ \sum_{i=1}^{N}\frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} > 0, \\ \infty & otherwise. \end{cases} \tag{5.8}$$

where $q_{ik} = \boldsymbol{x}_{ik}^{'}\boldsymbol{C}_{i,-k}^{-1}\boldsymbol{y}_i$ and $\boldsymbol{C}_{i,-k}$ is the component of $\boldsymbol{C}_i$ without the k^{th} basis function $\boldsymbol{x}_{ik}$. When $\hat{\alpha}_k = \infty$ the posterior mean $\hat{\mu}_{ik}$ becomes zero and the corresponding basis function $\boldsymbol{x}_{ik}$ is not in the model for all i. But when $\hat{\alpha}_k < \infty$ then $\hat{\mu}_{ik} \neq 0$ and the basis function $\boldsymbol{x}_{ik}$ is included in the model for all i.

The selection of the basis functions is done iteratively. We first start with an empty model and select the basis function that has the largest impact on the log-likelihood $\ell(\boldsymbol{\alpha}, \sigma_\epsilon^{-2})$. The subsequent steps on selection of the remaining basis functions involves three operations on $\boldsymbol{x}_{ik}$. Basically, the selection process involve; addition, deletion and updating $\hat{\alpha}_k$. Addition of the basis function $\boldsymbol{x}_{ik}$ occurs when $\sum_{i=1}^{N}\frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} > 0$ and $\boldsymbol{x}_{ik}$ is not in the model. An update for $\hat{\alpha}_k$ occurs when $\boldsymbol{x}_{ik}$ is already in the model and $\sum_{i=1}^{N}\frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} > 0$. We delete the basis function $\boldsymbol{x}_{ik}$ from the model when $\sum_{i=1}^{N}\frac{(q_{ik}^2 - s_{ik})}{s_{ik}^2} < 0$. The estimation process involves computation of $\hat{\sigma}_\epsilon^{-2}$ and $\hat{\alpha}_k$ in equations (5.5-5.6) that are used to update the mean vector $\hat{\mu}_i$ and covariance matrix $\hat{\Sigma}_i$ in

equation (5.4). For a concrete justification of this type of approximation refer to Ji et al 2009.

5.3 Extension to linear mixed model

The discussion in the previous section only allows estimation of subject specific curves but not the population average. To generate the population average curve we extend model (5.2) into linear mixed model that can captures both the subject specific and population average components. The functional model in equation (5.1) can be generalized into functional mixed model. The smoothing function is expressed as

$$f_i(z_{it}) = \sum_{j=1}^{M} \beta_j \varphi_j(z_{it}) + \sum_{j=1}^{M^*} b_{ij}\phi_j(z_{it}) = \boldsymbol{x}_{it}{}'\boldsymbol{\beta} + \boldsymbol{w}_{it}'\boldsymbol{b}_i,$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)'$ and $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iM^*})'$ are fixed and random effects respectively. Functions $\boldsymbol{\varphi} = \{\varphi_j\}_{j=1}^{M}$ and $\boldsymbol{\phi} = \{\phi_j\}_{j=1}^{M^*}$ are basis functions generated using methods discussed in the literature (Ramsay and Silverman, 2005). This results into the classical linear mixed model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_{T_i}), \quad i = 1, \ldots, N, \tag{5.9}$$

where $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \cdots, \boldsymbol{y}_{iM})'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT_i})'$, $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iT_i})'$ and $\boldsymbol{\epsilon}_i$ is a $T_i \times 1$ vector of error terms. Implementation of the RVM procedure require independence among all the random components $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iM^*})'$ and $\boldsymbol{\epsilon}_i$ leading to a diagonal covariance matrix $\boldsymbol{\Omega} = diag\{\omega_1, \cdots, \omega_{M^*}\}$. The priors are $\beta_j|\alpha_j \sim N(0, \alpha_j^{-1})$, $\alpha_j|c_1, d_1 \sim Gamma(c_1, d_1)$, $\omega_j|c_2, d_2 \sim Gamma(c_2, d_2)$ and $\sigma_\epsilon^{-2}|a, b \sim Gamma(a, b)$.

5.3.1 Parameter estimates

Inference in Bayesian data analysis is based on the posterior distribution of the parameters. The joint posterior distribution for the model parameters is,

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{b}, \sigma_\epsilon^{-2}) p(\boldsymbol{\beta}|\boldsymbol{\alpha}) p(\boldsymbol{b}|\boldsymbol{\omega}) p(\boldsymbol{\alpha}) p(\boldsymbol{\omega}) p(\sigma_\epsilon^{-2}),$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}\}$, $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)'$ and $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_{M^*})'$ are the diagonal elements of the covariance matrix $\boldsymbol{\Omega}$. The posterior $p(\boldsymbol{\Theta}|\boldsymbol{Y})$ is analytically intractable since the normalizing constant does not have a closed form solution. To approximate

$p(\boldsymbol{\Theta}|\boldsymbol{Y})$, we use the decomposition,

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) = p(\boldsymbol{b}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}|\boldsymbol{Y}), \tag{5.10}$$

with the first two terms obtained exactly as written in equation (5.4). The posterior distribution for $\boldsymbol{\beta}$ is,

$$p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) = \boldsymbol{N}(\boldsymbol{\beta}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}), \tag{5.11}$$

where $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\Sigma}}(\sum_{i=1}^N \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{y}_i)$ and $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{A} + \sum_{i=1}^N \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)^{-1}$. Matrix $\boldsymbol{V}_i = \sigma_\epsilon^2 \boldsymbol{I}_{T_i} + \boldsymbol{W}_i \boldsymbol{\Omega}^{-1} \boldsymbol{W}_i'$, where $\boldsymbol{A}$ is a diagonal matrix with elements $\boldsymbol{\alpha}$. The posterior distribution for the random effects is,

$$p(\boldsymbol{b}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) = \prod_{i=1}^N N(\boldsymbol{b}_i; \hat{\boldsymbol{\upsilon}}_i, \hat{\boldsymbol{\Omega}}_i), \tag{5.12}$$

where $\hat{\boldsymbol{\upsilon}}_i = \sigma_\epsilon^{-2} \hat{\boldsymbol{\Omega}}_i \boldsymbol{W}_i' (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})$ and $\hat{\boldsymbol{\Omega}}_i = (\boldsymbol{\Omega} + \sigma_\epsilon^{-2} \boldsymbol{W}_i' \boldsymbol{W}_i)^{-1}$.

Unfortunately, the posterior $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ for the variance components lacks a simple form. Therefore, we propose an empirical Bayes procedure for sparse MAP estimation in the next subsection.

5.3.2 Empirical Bayes for variance components

The posterior $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ is analytically intractable, and our goal is to rapidly conduct approximate Bayes inferences, while favouring sparsity. To accomplish this, we propose an empirical Bayes approach in which we obtain plug-in estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$ and σ_ϵ^2 that favor a sparse shrinkage structure, with many elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ set very close to zero. As empirical Bayes estimates of $\boldsymbol{\omega}, \sigma_\epsilon^{-2}$ and $\boldsymbol{\alpha}$, we use the modes of the generalized log-likelihood functions $l(\boldsymbol{\omega}, \sigma_\epsilon^{-2})$ and $l(\boldsymbol{\alpha})$ to be defined below.

We consider a joint density function for the variance parameters as $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}|\boldsymbol{Y}) \propto p(\boldsymbol{\alpha}) p(\boldsymbol{\omega}) p(\sigma_\epsilon^{-2}) p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$ where the density functions $p(\boldsymbol{\alpha})$, $p(\boldsymbol{\omega})$ and $p(\sigma_\epsilon^{-2})$ are the Gamma distributions that were defined in the previous section. The likelihood function $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^2)$ is obtained after integrating out $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iM^*})'$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)'$ such that,

$$p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^2) = \int \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i, \sigma_\epsilon^2) p(\boldsymbol{\beta}|\boldsymbol{\alpha}) p(\boldsymbol{b}_i|\boldsymbol{\omega}) d\boldsymbol{b}_i d\boldsymbol{\beta}. \tag{5.13}$$

We assume that $p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = N(\boldsymbol{\beta}; \boldsymbol{0}, \boldsymbol{A}^{-1})$ where $\boldsymbol{A} = diag\{\alpha_1, \ldots, \alpha_M\}$, while the distribution for the j^{th} random effect is $b_{ij} \sim N(0, \omega_j^{-1})$ for $j = 1, \cdots, M^*$ and $i = 1, \cdots, N$.

Let $p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\omega},\sigma_\epsilon^2) = N(\boldsymbol{Y};\boldsymbol{0},\boldsymbol{C})$, where $\boldsymbol{C} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{'}$, $\boldsymbol{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N)^{'}$ is the design matrix and $\boldsymbol{V} = diag(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N)$. We follow an empirical Bayes approach and choose non-informative priors for $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$ and σ_ϵ^2 by setting all the gamma hyper-parameters equal to zero. The mode of the posterior $p(\boldsymbol{\alpha},\boldsymbol{\omega},\sigma_\epsilon^{-2}|\boldsymbol{Y})$ is then equivalent to the maximum of $\log p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\omega},\sigma_\epsilon^{-2})$. This can be obtained by alternating conditional maximization iterating between calculating the conditional MLE of $\boldsymbol{\alpha}$ holding $\boldsymbol{\omega}$ and σ_ϵ^{-2} fixed and the conditional MLE of $\boldsymbol{\omega},\sigma_\epsilon^{-2}$ holding $\boldsymbol{\alpha}$ fixed.

We consider to maximize $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\sigma_\epsilon^{-2}) = \frac{-1}{2}\{N\log(2\pi) + \log|\boldsymbol{C}| + \boldsymbol{Y}^{'}\boldsymbol{C}^{-1}\boldsymbol{Y}\}$ to obtain the estimates for $\boldsymbol{\alpha}$. Unfortunately the estimates for $\boldsymbol{\alpha}$ cannot be computed analytically due to the presence of the square matrix $\boldsymbol{C}$ in the log-likelihood function $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\sigma_\epsilon^{-2})$. Following a parallel approach as in Ji et al. (2009) while taking fixed values of $\boldsymbol{\omega}$ and σ_ϵ^2, we re-write the log-likelihood function in a decomposed representation leading to,

$$\begin{aligned}\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\sigma_\epsilon^{-2}) = \frac{-1}{2}\Big\{&N\log(2\pi) + \log|\hat{\boldsymbol{\Sigma}}^{-1}| + \log|\boldsymbol{A}| \\ &+\log|\boldsymbol{V}^{-1}| + \hat{\boldsymbol{\mu}}\boldsymbol{A}\hat{\boldsymbol{\mu}} + (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\mu}})^{'}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\mu}})\,\}. \end{aligned} \tag{5.14}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the posterior mean and covariance for $\boldsymbol{\beta}$ as in equation (5.5). The estimate for the j^{th} element of $\boldsymbol{\alpha}$ is,

$$\hat{\alpha}_j = \frac{1}{\hat{\mu}_j^2 + \hat{\Sigma}_{jj}} \qquad j = 1,\ldots,M, \tag{5.15}$$

where $\hat{\mu}_j$ and $\hat{\Sigma}_{jj}$ are the j^{th} elements of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ respectively. Therefore, the estimates for $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1,\ldots,\hat{\alpha}_M)^{'}$ are functions of both the mean and covariance of $\boldsymbol{\beta}$. The computation of $\hat{\alpha}_j$ involves an iterative procedure that estimates the variance hyper-parameter α_j in equation (5.15) and updates the mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ as in equation (5.11).

Let $\ell(\boldsymbol{\omega},\sigma_\epsilon^{-2};\boldsymbol{\alpha}) = -1/2\sum_{i=1}^{N} T_i\log(2\pi) + \log|\boldsymbol{V}_i^{-1}| + (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})^{'}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})$ be a conditional log-likelihood for $\boldsymbol{\omega}$ and σ_ϵ^{-2} given $\boldsymbol{\alpha}$. We maximize $\ell(\boldsymbol{\omega},\sigma_\epsilon^{-2};\boldsymbol{\alpha})$ to obtain the estimates for $\boldsymbol{\omega}$ and σ_ϵ^{-2}. Unfortunately, the presence of matrix $\boldsymbol{V}_i$ in the log-likelihood function causes problem in the computation of $\boldsymbol{\omega}$ and σ_ϵ^2. But upon decomposing the log-likelihood $\ell(\boldsymbol{\omega},\sigma_\epsilon^{-2};\boldsymbol{\alpha})$ we obtain,

$$\begin{aligned}\ell(\boldsymbol{\omega},\sigma_\epsilon^{-2};\boldsymbol{\alpha}) = \frac{-1}{2}\sum_{i=1}^{N}&\log|\hat{\boldsymbol{\Omega}}_i^{-1}| + log|\boldsymbol{\Omega}^{-1}| + \log|\sigma_\epsilon^2\boldsymbol{I}| \\ &-\sigma_\epsilon^2||\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}} - \boldsymbol{W}_i\hat{\boldsymbol{v}}_i||^2 + \hat{\boldsymbol{v}}_i^{'}\boldsymbol{\Omega}\hat{\boldsymbol{v}}_i,\end{aligned}$$

where $\hat{\boldsymbol{\upsilon}}_i$ and $\hat{\boldsymbol{\Omega}}_i^{-1}$ are the posterior mean and variance for $\boldsymbol{b}_i$. We then obtain,

$$\hat{\omega}_j = \frac{N}{\sum_{i=1}^{N}(\hat{\upsilon}_{ij}^2 + \hat{\Omega}_{ijj})}, \tag{5.16}$$

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{N} ||\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}} - \boldsymbol{W}_i\hat{\boldsymbol{\upsilon}}_i||^2}{\sum_{i=1}^{N}(T_i - M + \sum_{j=1}^{n}\omega_j\hat{\Omega}_{ijj})}, \quad i = 1, \ldots N, j = 1, \ldots M, \tag{5.17}$$

where $\hat{\upsilon}_{ij}$ is the j^{th} component of $\hat{\boldsymbol{\upsilon}}_i$ and $\hat{\Omega}_{ijj}$ is the j^{th} diagonal element of $\hat{\boldsymbol{\Omega}}_i$. Both $\hat{\omega}_j$ and $\hat{\sigma}_\epsilon^2$ are functions of $\hat{\boldsymbol{\upsilon}}_i$ and $\hat{\boldsymbol{\Omega}}_i$, which leads to an iterative algorithm that alternates between updating $\hat{\omega}_j$ and $\hat{\sigma}_\epsilon^2$ via equations (5.16)-(5.17) and updating the mean and covariance of the random effects as in equation (5.12).

Two problems arise when implementing the above empirical Bayes approach when the number of fixed and random effects is large. First, the computational time increases dramatically for large M and/or M^* due to the need to invert $M \times M$ and $M^* \times M^*$ matrices at each step of the iterative procedure. In addition, when M and/or M^* are large relative to the sample size, estimability problems can arise that lead to lack of convergence of the procedure. Potentially this can be solved by using a MAP estimation approach that includes a proper prior to induce a penalty in the procedure that leads to shrinkage towards the prior and regularization. However, such an approach will be sensitive to hyper-parameter choice. An alternative is to adapt a fast algorithm that will bypass the inversion step leading to a reduced model with dimension $m \times m$ and $m^* \times m^*$ for both $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Omega}}_i$ respectively where $m << M$ and $m^* \leq m$. The RVM iterative algorithm can generate such a sparse model and will be discussed in the next section.

5.3.3 A fast MT-RVM method

A fast MT-RVM approach can be used to hasten the estimation process for $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ while overcoming computational problems that were highlighted before. The RVM approach reduces the dimensions of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ by discarding $M - m$ columns of $\boldsymbol{X}_i$ and $M^* - m^*$ columns of $\boldsymbol{W}_i$. These columns correspond to fixed and random effects that can be excluded, since their posteriors are concentrated at zero. The posterior for β_j is concentrated at zero when $\hat{\alpha}_j^{-1} = 0$, while the posterior for b_{ij} is concentrated at zero for all i when $\hat{\omega}_j^{-1} = 0$.

To estimate $\boldsymbol{\alpha}$, we consider a conditional log-likelihood $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \sigma_\epsilon^{-2})$ and partition it into two parts; one with and one without the k^{th} element of $\boldsymbol{\alpha}$. To achieve this decomposition,

the covariance matrix has to be partitioned into $\boldsymbol{C} = \boldsymbol{C}_k + \boldsymbol{C}_{-k}$, such that $\boldsymbol{C}_k = \alpha_k^{-1}\boldsymbol{X}_{.k}\boldsymbol{X}_{.k}^{'}$ and $\boldsymbol{C}_{-k} = \boldsymbol{V} + \sum_{j\neq k}^{M}\alpha_j^{-1}\boldsymbol{X}_{.j}\boldsymbol{X}_{.j}^{'}$ where $\boldsymbol{X}_{.j}$ is the j^{th} column of matrix $\boldsymbol{X}$. The partitioned log-likelihood function becomes,

$$\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\sigma_\epsilon^{-2}) = \ell(\boldsymbol{\alpha}_{-k};\boldsymbol{\omega},\sigma_\epsilon^{-2}) + \frac{1}{2}\left(\log\alpha_k - \log|\alpha_k + s_k| + \frac{q_k^2}{\alpha_k + s_k}\right),$$

where $s_k = \boldsymbol{X}_{.k}^{'}\boldsymbol{C}_{-k}^{-1}\boldsymbol{X}_{.k}$ and $q_k = \boldsymbol{X}_{k}^{'}\boldsymbol{C}_{-k}^{-1}\boldsymbol{Y}$. The estimate for α_k is

$$\hat{\alpha}_k = \begin{cases} \frac{s_k^2}{q_k^2 - s_k} & \textit{if } q_k^2 > s_k, \\ \infty & \textit{otherwise.} \end{cases} \tag{5.18}$$

This estimate determines the value of the fixed effect $\hat{\beta}_k$. Depending on the values of q_k and s_k, three operations can take place on $\boldsymbol{X}_{.k}$, including addition, deletion or update of the coefficient. At the beginning of the computation process we fix all $\hat{\alpha}_k = \infty$ which corresponds to an empty model with $\boldsymbol{\beta} = \boldsymbol{0}$. Subsequent iterations involve selection of a candidate $\boldsymbol{X}_{.k}$ that has the largest contribution to the log-likelihood $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\sigma_\epsilon^{-2})$. After the selection, we compute the values of q_k and s_k. When $q_k^2 > s_k$ we add $\boldsymbol{X}_{.k}$ into the model or update $\hat{\alpha}_k$ if $\boldsymbol{X}_{.k}$ was already in the model. Deletion occurs when $q_k^2 < s_k$ and $\boldsymbol{X}_{.k}$ is currently in the model.

The components of $\boldsymbol{\omega}$ can be computed based on the conditional log-likelihood function $\ell(\boldsymbol{\omega};\boldsymbol{\alpha},\sigma_\epsilon^2) = -1/2\sum_{i=1}^{N} T_i\log(2\pi) + \log|\boldsymbol{V}_i^{-1}| + (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})^{'}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})$. Similarly, the log-likelihood is partitioned into two parts; one with and one without the k^{th} component of $\boldsymbol{\omega}$. To allow such decomposition, we first partition matrix $\boldsymbol{V}_i$ into two parts such that $\boldsymbol{V}_i = \boldsymbol{V}_{i,-k} + \boldsymbol{V}_{ik}$, where $\boldsymbol{V}_{i,-k} = \sigma_\epsilon^2\boldsymbol{I} + \sum_{j\neq k}^{M^*}\alpha_j^{-1}\boldsymbol{w}_{ij}\boldsymbol{w}_{ij}^{'}$ and $\boldsymbol{V}_{ik} = \alpha_k^{-1}\boldsymbol{w}_{ik}\boldsymbol{w}_{ik}^{'}$. The decomposed log-likelihood becomes,

$$\ell(\boldsymbol{\omega};\boldsymbol{\alpha},\sigma_\epsilon^2) = \ell(\boldsymbol{\omega}_{-k};\boldsymbol{\alpha},\sigma_\epsilon^2) + \frac{-1}{2}\sum_{i=1}^{N}\left(\log\omega_k - \log|\omega_k + s_{ik}^*| + \frac{q_{ik}^{*2}}{\omega_k + s_{ik}^*}\right)$$

where $s_{ik}^* = \boldsymbol{w}_{ik}^{'}\boldsymbol{V}_{i,-k}^{-1}\boldsymbol{w}_{ik}$ and $q_{ik}^* = \boldsymbol{w}_{ik}^{'}\boldsymbol{V}_{i,-k}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})$. Differentiating $\ell(\boldsymbol{\omega};\boldsymbol{\alpha},\sigma_\epsilon^2)$ with respect to ω_k and setting the result to zero yields the score equations,

$$\frac{\ell(\boldsymbol{\omega};\boldsymbol{\alpha},\sigma_\epsilon^2)}{\partial\omega_k} = \sum_{i=1}^{N}\frac{s_{ik}^{*2}/\omega_k + s_{ik}^* - q_{ik}^{*2}}{2(\omega_k + s_{ik}^*)^2} = 0.$$

The solutions from the above equation are infeasible to express analytically except for a trivial case where $\omega_k = \infty$. The exact solutions require finding the zeros of a polynomial of degree $2N-1$ which is computationally expensive. A method to avoid such complexities is to assume that $\omega_k << s^*_{ik}$, leading to the approximate estimate

$$\hat{\omega}_k \cong \begin{cases} \frac{N}{\sum_{i=1}^{N}(q^{*2}_{ik}-s^*_{ik})/s^{*2}_{ik}} & if\ \sum_{i=1}^{N}\frac{(q^{*2}_{ik}-s^*_{ik})}{s^{*2}_{ik}} > 0, \\ \infty & otherwise. \end{cases} \tag{5.19}$$

For a justification of this type of approximation, refer to Ji et al. (2009).

The computation of the random effects is conditioned on the fixed effects that are already in the model and it involves three operations: add, update and delete. Addition occurs when $\sum_{i=1}^{N}\frac{(q^{*2}_{ik}-s^*_{ik})}{s^{*2}_{ik}} > 0$ and $\boldsymbol{w}_{ik}$ is not in the model, while an update occurs when $\boldsymbol{w}_{ik}$ is in the model and $\sum_{i=1}^{N}\frac{(q^{*2}_{ik}-s^*_{ik})}{s^{*2}_{ik}} > 0$. Deletion occurs when $\sum_{i=1}^{N}\frac{(q^{*2}_{ik}-s^*_{ik})}{s^{*2}_{ik}} < 0$ and $\boldsymbol{w}_{ik}$ is currently in the model. Updating vectors $\boldsymbol{s}^*$ and $\boldsymbol{q}^*$ is based on the values from the previous iteration. The final model tends to have most of the $\omega_j = \infty$, which corresponds to $b_{ij} = 0$ for all i (exclusion of the random effects).

At each iteration, the algorithm computes or updates one component of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ at a time and this helps to overcome the problem encountered while trying to invert matrices $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Omega}}_i$. However, the computation process in this sequential algorithm requires a lot of iterations to reach convergence. To accelerate convergence, we choose and update the fixed and random effects that lead to the largest increase in the log-likelihood functions $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\sigma_\epsilon^{-2})$ and $\ell(\boldsymbol{\omega};\boldsymbol{\alpha},\sigma_\epsilon^2)$, respectively. The algorithm is computationally demanding when working with large datasets but it is an improvement over the previous approach that required inversion of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Omega}}_i$. The resulting model has m instead of the original M fixed effects and m^* instead of original M^* random effects. For the selections of the basis functions for the fixed and the random effects, we follow the algorithm discussed in Tipping (2001) and Ji et. al (2009) using the simplified expression for different quantities discussed in Appendix A.

5.4 Results

The research is motivated by the basal body temperature data for the European fecundability study (Colombo and Masarotto, 2000). There were 880 women in the study aged between 18 and 40 years, who were not taking hormonal medications or drugs affecting fertility, and had no known impairment of fecundity. The subjects kept daily records of cervical mucus or basal body temperature measurements from at least one menstrual

cycle, and they recorded the days during which intercourse and menstrual bleeding occurred. For more details about the study, refer to Colombo and Masarotto (2000). In this chapter we considered the bbt measurements from 520 menstrual cycles where each subject contributed one menstrual cycle.

Typically a standard bbt curve has biphasic shape that is characterized with three phases representing the pre-ovulation, ovulation and post-ovulation periods. Identification of the ovulation day was based on three over six rule or a dip that is followed by a sharp rise in bbt (Colombo and Masarotto, 2000). Practically it is common to observe many menstrual cycles with wide fluctuations in bbt measurements resulting to false nadirs and peaks. Hence, it is difficult to replicate a standard bbt pattern from data collected from many cycles. Hormonal fluctuation is one among many causes that can interfere with the pattern of a bbt curve. Other causes include: less amount of sleep, sleep disturbances, ambient bedroom temperature, food ingestion and fluctuating emotional state (Colombo and Masarotto, 2000).

Other functional data analysis methods have been proposed to model the bbt data. For example, Scarpa and Dunson (2008) proposed a Bayesian semi-parametric model based on nonparametric contamination of a linear mixed effects model. The implementation of this approach relies on a highly computational intensive MCMC algorithm and our goal is to obtain a fast approximate Bayes approach that can be implemented much more rapidly, while obtaining smooth bbt trajectories. Hence, to compare the performance of the RVM approach with an MCMC based approach, we used the Wand and Ormerod (2008) subject specific approach instead of the Scarpa and Dunson (2008) method since it cannot generate smooth curves.

5.4.1 Subject-specific profiles

To implement the RVM procedure into a random coefficients model, we used the cubic B-splines (Ramsay and Silverman, 2005) to generate the basis functions $\boldsymbol{\varphi}$. Following the Wand and Ormerod (2008) approach, the basis functions were generated based on the standardized values of time covariate ($\boldsymbol{z}_i$). The number of the generated basis functions was 27 cubic B-splines with 23 interior knots. In addition, we added two columns containing $1's$ and $\boldsymbol{z}_i$ (i.e. $\{1, z_{it}\}_{t=1}^{T_i}$). Hence, the dimension for the design matrix was $\boldsymbol{X}_i$ is $T_i \times 29$ where $M = 29$.

Table 5.1 presents the computed posterior mean estimates $\hat{\boldsymbol{\mu}}_i$ for two menstrual cycles generated from the random coefficients model using both RVM and MCMC procedures. The MCMC based curves are estimated using 29 non-zero basis coefficients while the RVM method uses only three non-zero basis coefficients. We implemented the two

	Parameter estimates			
Basis no.	RVM_1	RVM_2	$MCMC_1$	$MCMC_2$
1	0.000	0.000	-0.036 (-0.219, 0.159)	0.025 (-0.102, 0.152)
2	0.684 (0.645, 0.722)	0.661 (0.623, 0.699)	0.647 (0.437, 0.849)	0.631 (0.426, 0.835)
3	0.000	0.000	-0.082 (-13.254, 11.721)	0.032 (-5.108, 5.172)
4	0.000	0.000	-0.175 (-13.875, 13.596)	-0.312 (-24.737, 24.113)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
25	0.000	0.000	4.104 (-2.507, 14.383)	0.451 (-0.275, 1.177)
26	0.000	0.000	-0.627 (-5.773, 4.770)	1.429 (-10.299, 13.157)
27	9.215 (8.492, 9.938)	6.188 (5.702, 6.674)	7.860 (2.597, 11.446)	3.708 (1.225, 6.190)
28	2.295 (1.922, 2.667)	0.000	2.150 (0.207, 3.961)	1.970 (0.189, 3.750)
29	0.000	1.356 (1.135, 1.576)	-0.292 (-1.018, 0.399)	1.258 (-1.869, 4.385)
Time spent	0.59 sec	0.57 sec	19.30 sec	16.34 sec

TABLE 5.1: A table for the parameter estimates for two bbt cycles.

methodologies using R software (version 2.8.1) on Pentium IV, 2.4GHz, 512MB, Windows XP computer platform. On time factor, the MCMC based method takes 19.30 and 16.34 seconds while the RVM method takes 0.59 and 0.57 seconds to estimate the two bbt curves respectively. Figure 5.1 presents estimated curves for the two bbt cycles using the two procedures. The continuous and dashed lines represent the estimated curves generated by the MCMC based and RVM methods respectively. The two sets of thin dashed lines and the grey region in the plots represent the 95% credible band for the RVM and the MCMC based methods respectively. Hence, the RVM method takes a shorter time relative to the MCMC based method given that the quality of the fitted curves is almost the same.

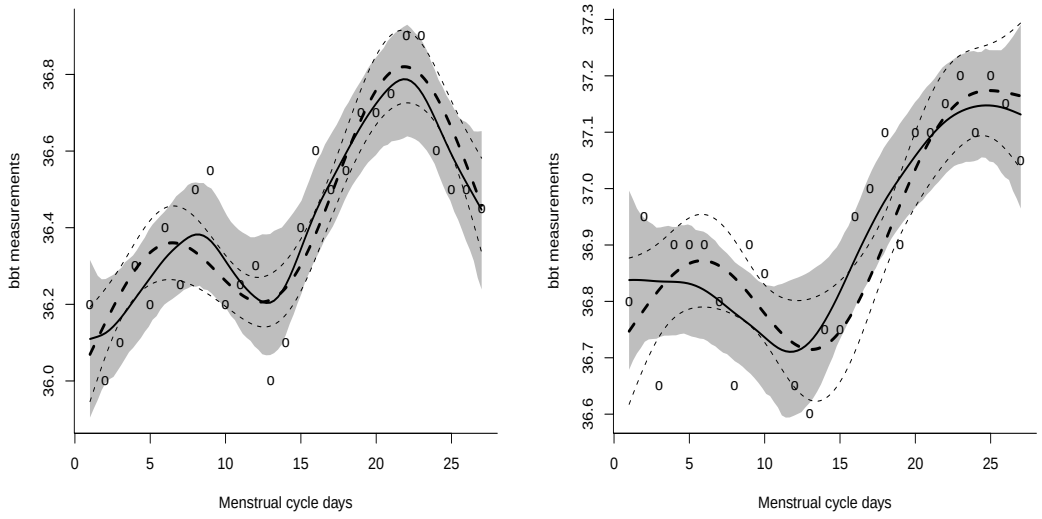

FIGURE 5.1: Plots of bbt curves

5.4.2 Effects of adding more observations

It might be of interest to evaluate the performance of the RVM method with the increase of the number of observations. Potentially, we expect that the gap between the curves from the two methods to narrow down as the number of observations increases. Similarly, the estimated non-zero random coefficients from the two methods are also expected to be identical at after a certain number of observations. However, we do not know the threshold number of observations when the two curves or non-zero parameter estimates look identical.

To assess the performance of the MT-RVM with the increase in the number of observations per cycle, we generate 30 biphasic curves that mimic the shape of the bbt curve. The curves are generated using a sine function,

$$y_{it} = v_i + \rho_i z_{it} \sin(10 z_{it} - r_i) + \epsilon_{it}, \quad t = 1, 2, \ldots 27, \quad i = 1, 2, \ldots 30,$$

where covariate $z_{it} \sim unif(0, 1)$, while $v_i \sim unif(-1, 1)$ and $r_i \sim unif(-1, 1)$ are the vertical and horizontal shift parameters and $\rho_i \sim unif(0.5, 1.5)$ controls the amplitude of the curves. Each curve had 27 observations and we generated the basis functions using the method used in the previous section. Figure 5.2 shows one of the curves generated using the sine function. The plot shows three curves namely: the true curve, the estimated curve using MCMC and MT-RVM methods.

The computation of the basis coefficients is based on the two methods -an MCMC based and the MT-RVM methods. To compare the estimation of the basis coefficients from the two methods as the number of observations increases, we computed Reconstructive Error defined as

$$RE_l = \frac{1}{N} \sum_{i=1}^{N} \frac{||\boldsymbol{\beta}_i^{RVM} - \boldsymbol{\beta}_i^{MCMC}||}{||\boldsymbol{\beta}_i^{RVM}||}, \qquad l = 27, \ldots 227.$$

Reconstructive error is a measure that captures the differences between the two sets of parameters estimates from the two methods. After the computation of the coefficient estimates and the reconstructive error for the initial model, we subsequently simulated additional 200 observations for each cycle. After each increment, we re-computed the basis functions $\boldsymbol{X}_i$ for the new data, computed the basis coefficients using the two methods and then re-computed the reconstructive error value. We plot the RE_l against the number of observations (l) as shown in figure 5.3.

From the plots, it is evident that an increase in the number of observations leads to a gradual decrease in the reconstructive error but the decreasing trend reaches to a constant value after about 150 observations. However, the reconstructive error curve

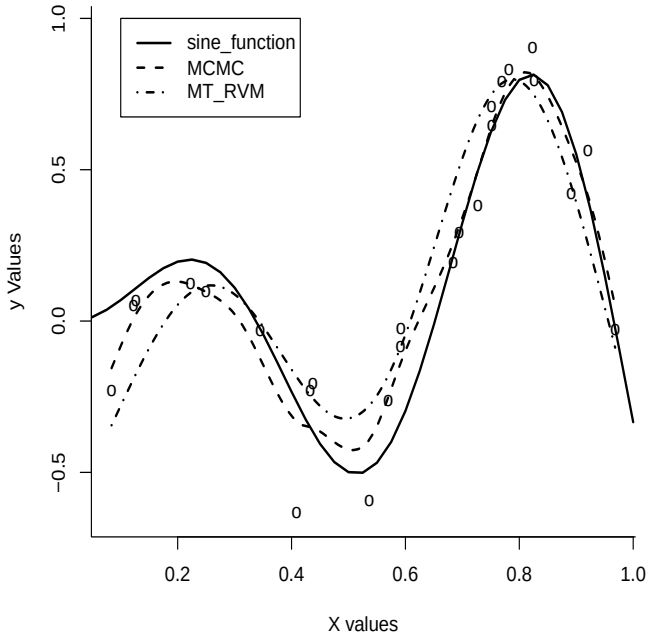

FIGURE 5.2: Estimated sine curves.

remains constant at a non-zero RE value since most of the non-relevant basis coefficients from the MCMC based method are non-zero while their corresponding basis coefficients from the MT-RVM method are zero.

5.4.3 Subject-specific and population average profiles

Similarly to implement the RVM procedure for a linear mixed model, the cubic B-splines (Ramsay and Silverman, 2005) were used to generate the basis functions $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$. Following Wand and Ormerod (2008) approach, the two sets of the basis functions were generated based on the standardized values of time covariate ($\boldsymbol{z}_i$). Both design matrices $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$, have a total of $M = M^* = 29$ columns. The first two columns in both matrices contain $1's$ and $\boldsymbol{z}_i$ (i.e. $\{1, z_{it}\}_{t=1}^{T_i}$) while the remaining columns are generated from 27 cubic B-splines with 23 interior knots. Hence, both $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ matrices have dimensions $T_i \times 29$ such that $M = M^* = 29$.

The RVM and the MCMC based procedures for the linear mixed model were implemented on data for the 520 bbt cycles. The final RVM model has $m = 3$ fixed effects (basis functions: 1, 2 and 28) and $m^* = 10$ random effects (basis functions: 1, 2, 28,

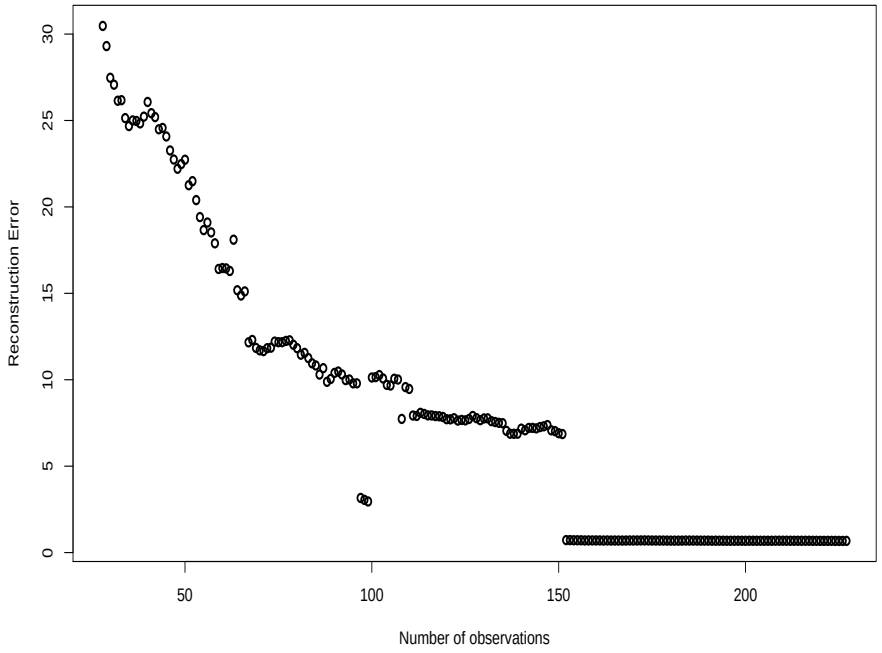

FIGURE 5.3: A plot for RE against the number of observations.

29, 9, 14, 27, 12, 11 and 16). Figure 5.4 shows some estimated bbt curves from six randomly selected subjects based on the RVM method. The continuous line represent the estimated bbt curve while the gray region represents the 95% confidence band. Figure 5.5 shows a plot for the population and subject specific curves based on the estimates from the RVM procedure. The thick black curve represents the population average bbt curve while the thin gray curves represent the estimated subject specific bbt curves. The interval that is characterized with a gentle rise in bbt curve is beleived to the most probable period in the menstrual cycle when majority of women experience ovulation.

5.4.4 Prediction

To evaluate the predictive ability of the proposed RVM procedure, we conducted an out-of-sample prediction. Typically, the process involves randomly dropping a certain percentage of observations (e.g. 20%, 30%, 50%) in each cycle, generate the basis functions, fit a model using an appropriate method and then estimate the curves. Based on the parameter estimates generated using the reduced data, we predict the dropped observations. The computation involves estimation of both the mean and the credibility interval for the predicted values. In this application, we dropped 20% of the total observations chosen at random from each women and then the RVM procedure was used

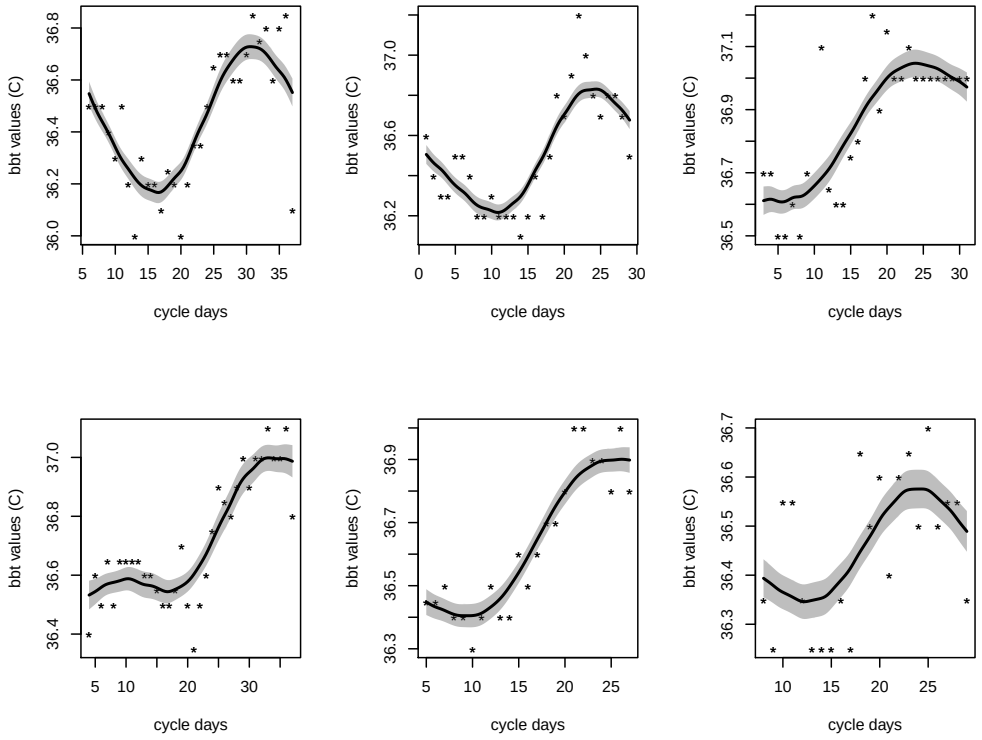

FIGURE 5.4: Estimated bbt curves and the 95% confidence band from the RVM procedure.

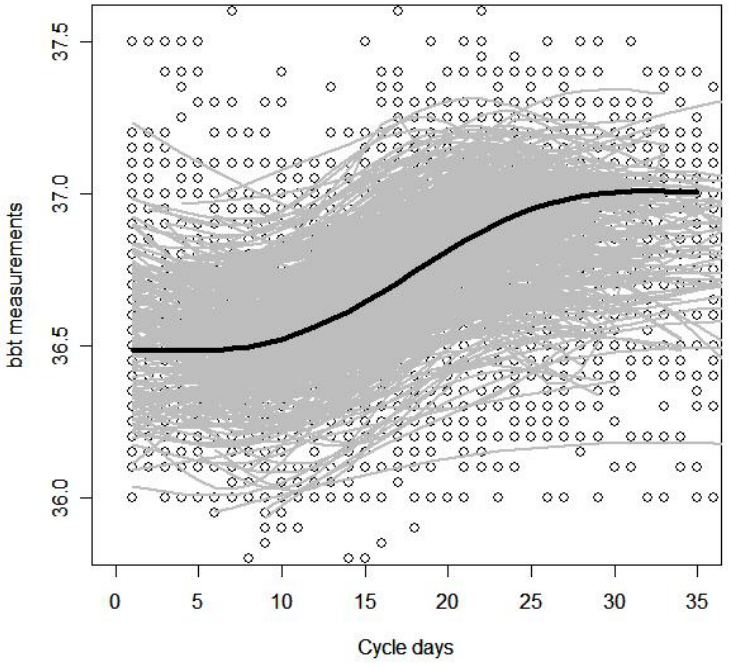

FIGURE 5.5: Estimated population and subjects specific bbt curves from the RVM procedure.

to estimate the model parameters. Based on the parameter estimates from the RVM procedure, we predicted the 20% dropped bbt values. To assess the predictive ability of the RVM procedure, we computed the correlation value between the predicted 20% that were dropped and their corresponding fitted values when all observations are present. The correlation value was 0.80.

Figure 5.6 shows the predicted bbt values from four randomly selected subjects based on the RVM procedure. The thick line and the gray region in each plot represents the estimated curve and the 95% credibility band based on all observations. The "+" sign along the estimated curve shows the mean estimate of the bbt value that would latter be dropped and then predicted. The star ($*$) at the middle of a vertical line represent the predicted bbt value when 20% of the observations are excluded and the small vertical lines represent the 95% predictive credibility intervals. However, we note that for those predicted bbt values that have narrow credibility intervals, the star and the "+" sign are more visible than the horizontal lines (credibility intervals) which are concealed within the star ($*$).

A high correlation value between the fitted and the predicted bbt values has substantial clinical implications. This is because women may be able to collect fewer bbt observations without greatly reducing the accuracy of the estimated bbt curve over the cycle. Since most of the data collected in reproductive studies are commonly sparse, the results from the out-of-sample prediction for the proposed procedure suggest that moderate data sparseness does not have a strong effect on the accuracy of the estimated curves.

5.5 Discussion

In the literature, many data smoothing procedures have been proposed to estimate non linear curves. For example, Brumback and Rice, (1998), used penalized smoothing spline mixed-model to generate smooth curves for multi-level data. To reduce the dimension of the data Crainiceanu (2009) used Functional Principal Component Analysis (FPCA) on linear mixed models to generate sparse models that can approximate non-linear curves. However, these MCMC based approaches do not allow fast computation since the computation relies on slow and computational intensive MCMC algorithms. The use of the RVM method can helps hasten the computation process leading to a faster model building process that can simultaneously address the two variable selection problems at a less computational cost.

Multi-task Relevant Vector Machine (MT-RVM) approach has been used in machine learning especially in signal reconstruction and compressive sensing applications Ji et

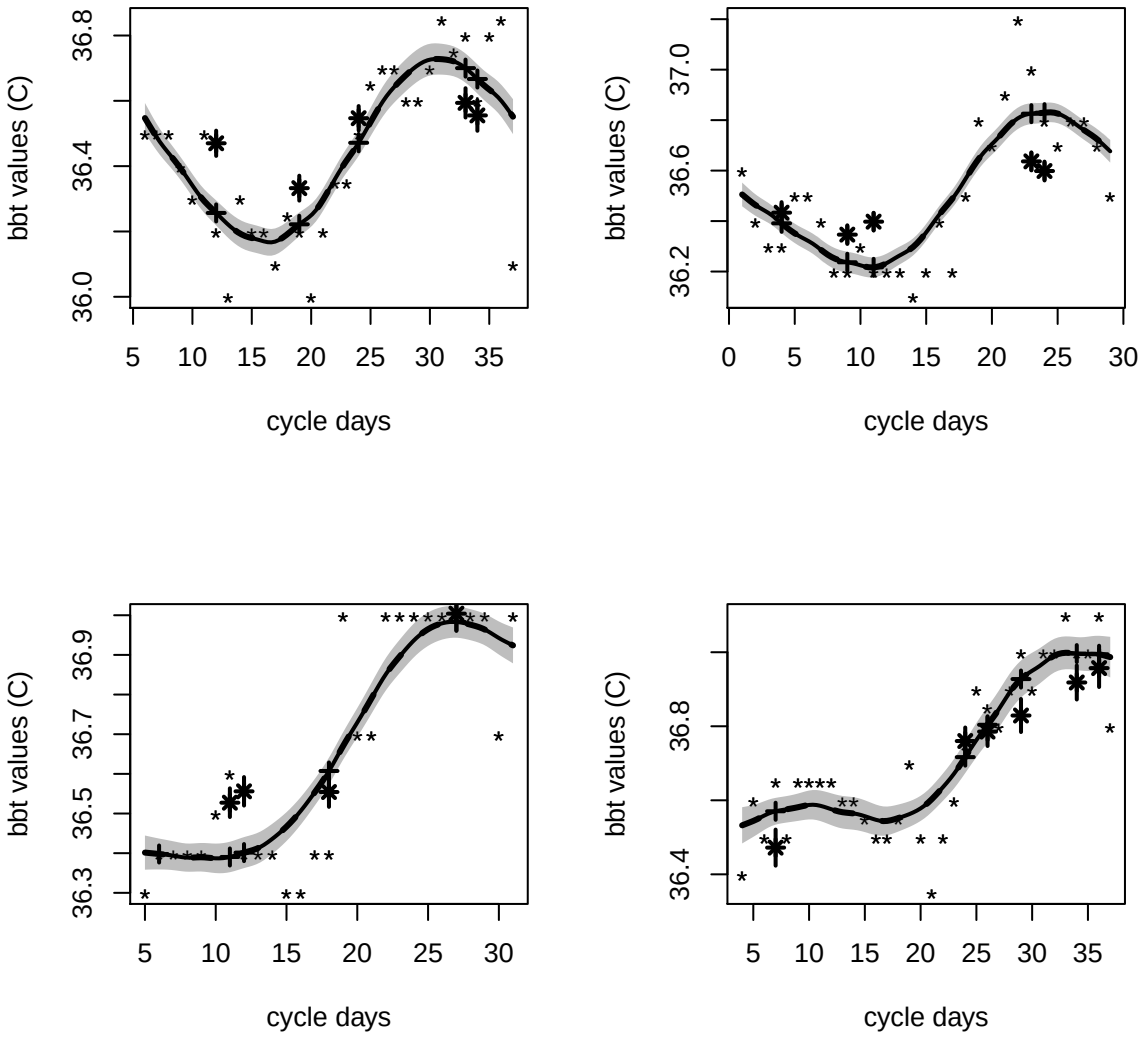

FIGURE 5.6: Estimated bbt curves based on the RVM method and the predicted 20% out of sample bbt values.

al, (2009). However, the application of this approach has not featured in applications that involve smoothing non-linear curves using penalized spline basis functions. In this chapter we demonstrate the use of RVM as an alternative approach to computer intensive methods that rely on MCMC. The method is fast and can be used in large dimensional functional models to generate sparse random coefficients and linear mixed models that can be used to rapidly estimate non-linear curves from massive datasets. In this application the RVM procedure was used to smoothen non-linear bbt curves that feature commonly in many reproductive studies.

In linear model there exist numerous variable selection procedures where variable selection procedure can be based on likelihood ratio tests, goodness-of-fit criteria and other

methods that are commonly applied in linear regression models. However, in linear mixed model framework, the variable selection procedures for the random effects normally fail due to complications that arise since the null hypothesis lies on the boundary of the parameter space. Hence, the likelihood ratio (LR) test statistic is no longer valid. This work proposes a fast approximate Bayes approach for simultaneous selection of both fixed and random effects to fit a linear mixed effects model. In adapting RVM method to perform variable selection, we considered a linear mixed model that assumes independent random effects. This is not always possible in many practical situations; hence it is of interest to extended the use of RVM methodology into linear mixed model with correlated random effects.

The advantage of our approach is not only on computational speed but it also allows for better generalization performance which leads to sparse generalized linear mixed models. This aspect can provide inference for a wide variety of models at a moderate computational cost. For example, this approach can easily be extended to accommodate multiple predictors, linear mixed effect models and probit models where multiple binary categorical outcomes can be handled using data augmentation (Albert and Chib, 1993). In the next chapter, we generalize the implementation of the MT-RVM method into classical linear mixed model with correlated random effects to generate sparse linear mixed model models. Such generalization can easily be extended into LME models that can handle hierarchical data where a woman can have data from multiple cycles.

Chapter 6

Fast approximate bayesian functional mixed effects model

6.1 Introduction

In this chapter we present an extension of the MT-RVM procedure. We develop a more flexible generalization of the MT-RVM in linear mixed models that allows shrinkage towards a non-zero covariance in the random effects. In particular, we implement the RVM procedure to a general LME model that has correlated random effects. We use the approach of Chen and Dunson (2003) to generate uncorrelated variables that can easily be implemented in MT-RVM methodology. We also present a simulation study and an application to real data. In the simulation study, we compares the performance of the MT-RVM method relative to other related competitive methods.

In clinical studies, it is routine to collect repeated measurements of a biomarker or other variable. Although the measurements for an individual are often sparse and unequally-spaced, with routine entry of patient information into computer data bases, it is increasingly common to have data available for massive numbers of individuals. Ideally, physicians would have automated tools available for utilizing information in the data base to rapidly estimate and predict the trajectory for a current patient.

Our motivation is drawn from reproductive applications collecting basal body temperature (bbt) or reproductive hormone measurements over the menstrual cycle. For women attempting conception, it is important to identify the day of ovulation, as intercourse has a near zero probability of resulting in conception if it occurs outside of the six-day fertile interval ending on the ovulation day (Dunson, Weinberg, Perreault and Chapin, 1999). In addition, healthy menstrual cycles exhibit characteristic trajectories in bbt

and hormones, so that the trajectories can be used to distinguish healthy ovulatory cycles from cycles with possible dysfunction. Hence, it is of interest to estimate the trajectories over the menstrual cycle based on the data available for that cycle, while borrowing information flexibly from other cycles in the data base.

As for other types of functional data (Ramsay and Silverman, 2005), it is common to have missing data in bbt or hormone data, and there are different numbers and spacings of measurements for different individuals. In addition, there is substantial heterogeneity in the curve shapes, with parametric normal random effects models providing a poor characterization of the data for unhealthy individuals (Scarpa and Dunson, 2008). For sparse functional data, it is common to rely on functional mixed effect models, which characterize a baseline curve and additive covariate effects using splines (Lin and Zhang, 1999; Guo, 2004). Additional discussions on connection between spline smoothing methods and mixed models can be found in Brumback and Rice (1998); Rice and Wu (2001); Durban, Harzelak, Wand and Carrol (2005). However, it can be difficult to choose the basis functions in advance, motivating the use of adaptive methods that allow uncertainty in basis function selection (Bigelow and Dunson, 2007; Thompson and Rosen, 2008). Other dimension reduction strategies include the use of functional principal component analysis (James, Hastie and Sugar, 2001; Yao, Muller and Wang, 2005; Crainiceanu, 2009).

Bayesian methods are useful for accommodating uncertainty in basis selection, with posterior computation relying on Markov chain Monte Carlo (MCMC) algorithms, such as reversible jump (Green, 1995) or stochastic search variable selection (Smith and Kohn, 1996). However, implementation of the algorithms is computationally expensive. Hence, there is a clear practical motivation for fast approximate Bayes approaches that bypass MCMC while maintaining some of the benefits of a Bayesian analysis. In the setting of estimation of a single function, such as a non-linear regression curve or multivariate regression surface, a rich variety of approaches have been proposed. For example, Krivobokova, Crainiceanu and Kauermann (2008) recently proposed a method for locally adaptive smoothing using P-splines implemented with the Laplace approximation to the marginal likelihood. Sakamoto (2007) proposed an empirical Bayes approach for selecting basis functions and knots using an alternative approximation to the marginal likelihood.

In the machine learning literature, the relevance vector machine (RVM) (Tipping, 2001) is widely used for function estimation. RVM penalizes the basis coefficients through a scale mixture of normals prior, which is carefully-chosen so that maximum a posteriori (MAP) estimates of many of the coefficients are zero. RVM is one among several methods that promote sparseness in estimation of basis coefficients, providing a more flexible

alternative to Support Vector Machines (SVM) (Burges, 1998). RVM is also thought to have advantages over LASSO (Tibshirani, 1996) in leading to a sparser solution that is more robust to outliers (Tipping, 2001; Ji, Dunson and Carin, 2009). Sparseness is a property where the fitted model retains the least number of basis functions (non-zero basis coefficients), while all the other basis functions are pruned by setting their corresponding coefficients to zero.

Motivated by bbt data from the European fecundability study (Colombo and Masarotto, 2000), this chapter introduces a more flexible generalization of the RVM applied to automated selection of fixed and random effects in a functional mixed model with correlated random effects. A related problem was considered by Ji et al. (2009) who proposed a multi-task relevance vector machine (MT-RVM) procedure based on wavelet bases to reconstruct multiple signals from compressive sensing measurements. Their approach allows basis function selection within a restricted class of models that assumes that the distribution of the basis coefficients is centered at zero with diagonal covariance. Centering at zero does not allow shrinkage towards a population-averaged curve and independence of the random effects is an unrealistic assumption in many longitudinal and functional data analysis applications. The generalization of the MT-RVM methodology to allow separate fixed and random effects and correlated random effects is not at all straightforward and represents the primary methodological advance in this chapter. To facilitate the extension, we rely on a modified Cholesky decomposition proposed by Chen and Dunson (2003) and further justified by Pourahmadi (2007).

6.2 Functional Data Analysis Model

6.2.1 Background and motivation

Functional data analysis models have been discussed extensively in the literature (Ramsay and Silverman, 2005; Ruppert and Carroll, 2000; Ruppert, et al. 2003). We consider observations from the i^{th} subject with response vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})^{'}$ and covariate vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iT_i})^{'}$. Using the t^{th} observation for the i^{th} subject, a simple functional model is represented as,

$$y_{it} = f_i(z_{it}) + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \quad t = 1, \ldots, T_i, \quad i = 1, \ldots, N. \tag{6.1}$$

where $f_i(.)$ is a smooth function for subject i and ϵ_{it} is a measurement error.

The functional data model in equation (6.1) can be generalized into a functional mixed model. Let $\boldsymbol{\varphi} = \{\varphi_j\}_{j=1}^{M}$ and $\boldsymbol{\phi} = \{\phi_j\}_{j=1}^{M^*}$ be a collection of basis functions for the fixed

and random effects components. The basis functions can be generated using numerous methods that have been discussed in the literature (e.g. Hastie, et al. 2001; Ruppert, et al. 2003). Function $f_i(.)$ can be described as a linear combination of basis functions

$$f_i(z_{it}) = \sum_{j=1}^{M} \beta_j \varphi_j(z_{it}) + \sum_{j=1}^{M^*} b_{ij} \phi_j(z_{it}), \qquad i = 1, \ldots, N,$$

where $\sum_{j=1}^{M} \beta_j \varphi_j(z_{it}) = \mathbf{x}_{it}{}'\boldsymbol{\beta}$ and $\sum_{j=1}^{M^*} b_{ij}\phi_j(z_{it}) = \boldsymbol{w}_{it}'\boldsymbol{b}_i$ such that $\boldsymbol{x}_{it} = (x_{it1}, \cdots, x_{itM})'$ and $\boldsymbol{w}_{it} = (w_{it1}, \cdots, w_{itM^*})'$ are the values of the basis functions at z_{it}. Parameter $\boldsymbol{\beta}$ is a vector of unknown population-specific parameters controlling the average curve while $\boldsymbol{b}_i$ are random effects capturing the systematic departure of the i^{th} subject. To express the linear mixed effects model, we let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT_i})'$ and $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iT_i})'$, then

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_{T_i}), \quad i = 1, \ldots, N, \tag{6.2}$$

where $\boldsymbol{\epsilon}_i$ is a $T_i \times 1$ vector of independent error terms. When the basis functions are pre-specified and few, standard Bayes and frequentist methodologies can be used for estimation and inference. However, for sufficient flexibility it is typically necessary to include many basis functions leading to a high-dimensional model. To reduce the dimension of the model, Smith and Kohn (1996) proposed the use of stochastic search variable selection (George and McCulloch, 1993), with Morris (2006) proposing a related approach.

Motivated by its conjugacy property, the inverse-Wishart distribution is commonly used as a prior for covariance matrix $\boldsymbol{\Omega}$. However, the inverse-Wishart distribution is too inflexible as a shrinkage prior for a high- dimensional covariance matrix in that all the diagonal entries have a common degree of freedom, and adaptive shrinkage of certain elements is not possible. To address this problem, a number of authors have proposed shrinkage priors for covariance matrices with MCMC used for posterior computation (Daniels and Zhao, 2003; Daniels and Kass, 1999; Morris and Carroll 2006).

To avoid MCMC, we focus on generalizing the MT-RVM methodology to allow shrinkage estimation of $\boldsymbol{\Omega}$ following common practice that rely on decomposition (Smith and Kohn, 2002; Chen and Dunson, 2003; Fruhwirth-Schnatter and Tuchler, 2004; Kinney and Dunson, 2006). We place carefully chosen shrinkage priors on the parameters in the decomposition.

6.2.2 Re-parameterization of $\boldsymbol{\Omega}$

The covariance matrix $\boldsymbol{\Omega}$ is re-parameterized based on a modified Cholesky decomposition that was proposed by Chen and Dunson (2003) and further justified by Pourahmadi (2007). The parameterization used in both papers can lead to problems if implemented using the MT-RVM algorithm. A modified version leads to latent variables with free variances making it possible to estimates the variances within the RVM procedure.

Let $\boldsymbol{y}_i$ be a vector of response variables while $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ are defined as in the previous section. We adopt a model,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{\Gamma}^{'}\boldsymbol{h}_i + \boldsymbol{\epsilon}_i, \tag{6.3}$$

where $\boldsymbol{\Gamma}$ is a lower triangular matrix that contains off-diagonal elements $\boldsymbol{\gamma} = (\gamma_{jl} : j = 2, \ldots, M^*; l = j+1, \ldots, M^*-1)^T$ and diagonal entries $\gamma_{jj} = 1$. Vector $\boldsymbol{h}_i = (h_{i1}, \ldots, h_{iM^*})^{'}$ consist of independent latent variables such that the random effects for the i^{th} subject are $\boldsymbol{b}_i = \boldsymbol{\Gamma}^{'}\boldsymbol{h}_i$. The prior for the latent variables is $\boldsymbol{h}_i \sim N(\boldsymbol{0}, \boldsymbol{H}_0^{-1})$ where $\boldsymbol{H}_0 = diag(\omega_1, \cdots, \omega_{M^*})$ and the j^{th} element $\omega_j \geq 0$.

The variance for the random effect b_{ij} is

$$var(b_{ij}) = \omega_j^2 \left(1 + \sum_{r=1}^{j-1} \gamma_{jr}^2\right) \qquad for \;\; j = 1, \cdots, M^*.$$

Matrix $\boldsymbol{\Gamma}$ is related to the degree of within-subject dependence in the random effects, with the correlation between the j^{th} and the l^{th} random effects

$$corr(b_{ij}, b_{il}) = \frac{\gamma_{jl} + \sum_{r=1}^{l-1} \gamma_{lr}\gamma_{jr}}{\sqrt{(1 + \sum_{r=1}^{l-1} \gamma_{lr}^2)(1 + \sum_{r=1}^{j-1} \gamma_{jr}^2)}}.$$

Moreover, the estimates for the j^{th} row and column in $\boldsymbol{\Gamma}$ depend upon the estimates of ω_j such that if $\omega_j^{-1} = 0$ then $\gamma_{lj} = \gamma_{jl} = 0$. Equation (6.3) implies that,

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \sum_{j=1}^{M^*} w_{itj} \left(h_{ij} + \sum_{l=j+1}^{M^*} \gamma_{jl} h_{il}\right) + \epsilon_{it}, \qquad t = 1, \cdots, T_i \tag{6.4}$$

where w_{itj} is the j^{th} element in the t^{th} row of matrix $\boldsymbol{W}_i$ and γ_{lj} are the off-diagonal elements in matrix $\boldsymbol{\Gamma}$. When all off-diagonal elements in $\boldsymbol{\Gamma}$ are zeros ($\gamma_{lj} = 0$ for $l \neq j$), equation (6.4) reduces to a simpler LME model where the random effects $\boldsymbol{b}_i$ are assumed to be independent.

6.2.3 Sparse functional mixed model estimation

Inference in Bayesian data analysis is based on the posterior distribution of the parameters. We consider the functional mixed model in equation (6.3), the joint posterior distribution for the model parameters can be expressed as

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{h},\boldsymbol{\gamma},\sigma_\epsilon^{-2})p(\boldsymbol{\beta}|\boldsymbol{\alpha})p(\boldsymbol{h}|\boldsymbol{\omega})p(\boldsymbol{\alpha})p(\boldsymbol{\omega})p(\boldsymbol{\gamma})p(\sigma_\epsilon^{-2}),$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\beta},\boldsymbol{h},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2}\}$, $\boldsymbol{Y} = (\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N)'$ and $\boldsymbol{h} = (\boldsymbol{h}_1,\cdots,\boldsymbol{h}_N)'$. The prior for the fixed effects is $\boldsymbol{\beta}|\boldsymbol{\alpha} \sim N(\boldsymbol{0},\boldsymbol{A}_0^{-1})$ such that $\boldsymbol{A}_0 = diag(\alpha_1,\ldots,\alpha_M)$. Gamma priors are chosen for the diagonal elements of $\boldsymbol{A}_0$ and $\boldsymbol{H}_0$, the elements of $\boldsymbol{\gamma}$ and σ_ϵ^{-2}, with $\sigma_\epsilon^{-2}|a,b \sim Gamma(a,b)$, $\alpha_j|c_1,d_1 \sim Gamma(c_1,d_1)$ for $j = 1,\ldots M$, $\omega_j|c_2,d_2 \sim Gamma(c_2,d_2)$ and $\gamma_{jl}|c_3,d_3 \sim Gamma(c_3,d_3)$ for $j = 1,\ldots,M^*$,$l = j+1,\ldots,M^*$. The posterior $p(\boldsymbol{\Theta}|\boldsymbol{Y})$ is analytically intractable since the normalizing constant does not have a closed form solution.

We approximate the joint posterior density by decomposing $p(\boldsymbol{\Theta}|\boldsymbol{Y})$ into conditional distributions,

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) = p(\boldsymbol{\beta}|.)p(\boldsymbol{h}|.)p(\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2}|\boldsymbol{Y}), \tag{6.5}$$

where $p(\boldsymbol{\beta}|.) = p(\boldsymbol{\beta}|\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2},\boldsymbol{Y})$ and $p(\boldsymbol{h}|.) = \prod_{i=1}^N p(\boldsymbol{h}_i|\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2},\boldsymbol{Y})$ are posterior distributions for the fixed effects and latent variables, while $p(\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2}|\boldsymbol{Y})$ is the joint density function for the variance components. The posterior distribution for the latent variables $\boldsymbol{h}_i$ is

$$p(\boldsymbol{h}_i|\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2},\boldsymbol{Y}) = N(\boldsymbol{h}_i;\hat{\boldsymbol{h}}_i,\hat{\boldsymbol{H}}_i), \tag{6.6}$$

where $\hat{\boldsymbol{h}}_i = \sigma_\epsilon^{-2}\hat{\boldsymbol{H}}_i\boldsymbol{U}_i'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$, $\hat{\boldsymbol{H}}_i = (\boldsymbol{H}_0 + \sigma_\epsilon^{-2}\boldsymbol{U}_i'\boldsymbol{U}_i)^{-1}$ such that $\boldsymbol{U}_i = (\boldsymbol{u}_{i1},\ldots,\boldsymbol{u}_{iT_i})'$ and $\boldsymbol{u}_{it} = (w_{itj} + \sum_{l=1}^{j-1} w_{itl}\gamma_{lj} : j = 1,\ldots,M^*)'$. The posterior distributions for the fixed effects $\boldsymbol{\beta}$ is,

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2},\boldsymbol{Y}) = N(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{A}}), \tag{6.7}$$

where $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{A}}\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{V}_i^{*-1}\boldsymbol{y}_i$, $\hat{\boldsymbol{A}} = (\boldsymbol{A}_0 + \sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{V}_i^{*-1}\boldsymbol{X}_i)^{-1}$, $\boldsymbol{V}_i^* = \sigma_\epsilon^2\boldsymbol{I}_{T_i} + \boldsymbol{U}_i\boldsymbol{H}_0^{-1}\boldsymbol{U}_i'$ and $\boldsymbol{V}_i^*$ is the covariance matrix for the random effects for subject i.

The joint posterior density for the variance components $p(\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\gamma},\sigma_\epsilon^{-2}|\boldsymbol{Y})$ is analytically intractable and we use an empirical Bayes procedure to generate MAP estimates for the variance components. This leads to the computation of plug-in estimates for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and σ_ϵ^2 that favor a sparse shrinkage structure, with many elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ set very close to zero.

6.2.4 Empirical Bayes estimates

The joint density $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}|\boldsymbol{Y}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\omega})p(\boldsymbol{\gamma})p(\sigma_\epsilon^{-2})p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$ where $p(\boldsymbol{\alpha})$, $p(\boldsymbol{\omega})$, $p(\boldsymbol{\gamma})$ and $p(\sigma_\epsilon^{-2})$ are Gamma distributions as described before. Following an empirical Bayes approach, we take non-informative priors for $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$, σ_ϵ^{-2} and the mode for $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ is equivalent to the maximum of the likelihood $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$. The likelihood function $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$ is obtained after integrating out $\boldsymbol{\beta}$ and $\boldsymbol{h}$ from equation (6.3), such that

$$p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) = \int \prod_{i=1}^{N} p(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{h}_i, \boldsymbol{\gamma}, \sigma_\epsilon^2)p(\boldsymbol{\beta}|\boldsymbol{\alpha})p(\boldsymbol{h}_i|\boldsymbol{\omega})d\boldsymbol{h}_i d\boldsymbol{\beta}.$$

We consider $\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2} \sim N(\boldsymbol{0}, \boldsymbol{C}^{*-1})$ where the covariance $\boldsymbol{C}^* = \boldsymbol{X}\boldsymbol{A}_0^{-1}\boldsymbol{X}' + \boldsymbol{V}^*$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)'$ and $\boldsymbol{V}^* = diag(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_N^*)$. Practically, it is difficult to simultaneously estimate the variance parameters, so we use alternating conditional maximization based on $\log p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$. The procedure iterates between computing the conditional MLE of one parameter, while holding the remaining three parameters constant. The process continues until we attain convergence.

Let $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) = \log N(\boldsymbol{0}, \boldsymbol{C}^{*-1})$ be a conditional log-likelihood function of $\boldsymbol{\alpha}$ that depends upon some fixed values of $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and σ_ϵ^{-2}. Maximizing $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$ leads to the estimation of $\boldsymbol{\alpha}$ contained in $\boldsymbol{A}_0$ while holding the parameters in $\boldsymbol{V}^*$ constant. Unfortunately the estimates for $\boldsymbol{\alpha}$ cannot be computed analytically due to the presence of the square matrix $\boldsymbol{C}^*$ in the log-likelihood function $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$. Following a parallel approach as in Ji et al. (2009) while taking fixed values of $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and σ_ϵ^2, we re-write the log-likelihood function in a decomposed representation. The computation involves partitioning the covariance matrix $\boldsymbol{C}^*$ into $\boldsymbol{X}\boldsymbol{A}_0^{-1}\boldsymbol{X}'$ and $\boldsymbol{V}^*$ which results into the log-likelihood

$$\begin{aligned} \ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) \quad = \quad & \frac{-1}{2}\Big\{N\log(2\pi) + \log|\hat{\boldsymbol{A}}^{-1}| + \log|\boldsymbol{A}_0| \\ & + \log|\boldsymbol{V}^{*-1}| \hat{\boldsymbol{\beta}}\boldsymbol{A}_0\hat{\boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{V}^{*-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\Big\}, \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{A}}$ are the conditional posterior mean and covariance matrix for $\boldsymbol{\beta}$. Differentiating the log-likelihood function with respect to $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_M)'$ and setting the solution to zero leads to

$$\hat{\alpha}_j = \frac{1}{\hat{\beta}_j^2 + \hat{A}_{jj}}, \qquad j = 1, \ldots, M, \tag{6.8}$$

where $\hat{\beta}_j$ is the j^{th} element of $\hat{\boldsymbol{\beta}}$ and $\hat{A}_{jj}$ is the j^{th} diagonal element of $\hat{\boldsymbol{A}}$ in equation (6.7). The estimates for $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_M)^{'}$ are functions of both the mean and covariance of $\boldsymbol{\beta}$. The computation of $\hat{\alpha}_j$ involves an iterative procedure that estimates the variance hyper-parameter α_j in equation (6.8) and updates the mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{A}}$ as in equation (6.7).

Similarly, let $\ell(\boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}) = -1/2 \sum_{i=1}^{N} T_i \log(2\pi) + \log|\boldsymbol{V}_i^{*-1}| + (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})^{'}\boldsymbol{V}_i^{*-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})$ be a conditional log-likelihood function of $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and σ_ϵ^{-2} at a fixed value of $\boldsymbol{\alpha}$. To maximize $\ell(\boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha})$ we seek to estimate $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and σ_ϵ^{-2} in $\boldsymbol{V}^*$ to maximize the log-likelihood function while holding $\boldsymbol{A}_0$ constant. The computation leads to partitioning $\boldsymbol{V}_i^*$ such that $\boldsymbol{V}_i^* = \boldsymbol{U}_i\boldsymbol{H}_0\boldsymbol{U}_i^{'} +, \sigma_\epsilon^{-2}\boldsymbol{I}_{M^*}$ where $\boldsymbol{U}_i = (\boldsymbol{u}_{it}, \ldots, \boldsymbol{u}_{iT_i})^{'}$ and we let,

$$\ell(\boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}) = \frac{-1}{2}\sum_{i=1}^{N} \log|\hat{\boldsymbol{H}}_i^{-1}| + \log|\boldsymbol{H}_0^{-1}| + \log|\sigma_\epsilon^2\boldsymbol{I}| - \sigma_\epsilon^2||\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{U}_i\hat{\boldsymbol{h}}_i||^2 + \hat{\boldsymbol{h}}_i^{'}\boldsymbol{H}_0\hat{\boldsymbol{h}}_i,$$

where $\hat{\boldsymbol{h}}_i$ and $\hat{\boldsymbol{H}}_i^{-1}$ are the conditional posterior mean and variance for the latent variable $\boldsymbol{h}_i$. We differentiate the log-likelihood with respect to $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and σ_ϵ^{-2} while equating the solutions to zero. This leads to

$$\hat{\omega}_j = \frac{N}{\sum_{i=1}^{N}\hat{h}_{ij}^2 + \hat{H}_{ijj}}, \qquad j = 1, \ldots, M^*, \tag{6.9}$$

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{N}||\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{U}_i\hat{\boldsymbol{h}}_i||^2}{\sum_{i=1}^{N}(T_i - M^* + \sum_{j=1}^{M^*}\hat{\omega}_j\hat{H}_{ijj})}, \qquad i = 1, \ldots, N, \tag{6.10}$$

$$\hat{\gamma}_{jl} = \frac{\sum_{i=1}^{N}\left\{(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{U}_i\hat{\boldsymbol{h}}_i)^{'}\boldsymbol{w}_{ij}\hat{h}_{il} - \hat{H}_{i,ll}\boldsymbol{w}_{ij}^{'}(\boldsymbol{w}_{il} + \sum_{r=1}^{M^*}\boldsymbol{w}_{ir}\gamma_{rl})\right\}}{\sum_{i=1}^{N}\hat{H}_{i,ll}\boldsymbol{w}_{ij}^{'}\boldsymbol{w}_{ij}}, \quad r \neq j, l = j+1, \ldots, M^*, \tag{6.11}$$

where $\hat{A}_{jj}$ and $\hat{H}_{ijj}$ are the j^{th} diagonal elements of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{H}}_i$, respectively. The posterior estimates $\hat{\beta}_j$ and $\hat{h}_{ij}$ are the j^{th} components of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{h}}_i$, respectively.

Based on this estimation procedure, the computation proceeds iteratively applying equations (6.9)-(6.11), with the conditional posterior mean and covariance of $\boldsymbol{\beta}$ and $\boldsymbol{h}_i$ as in equations (6.6) and (6.7), respectively. Because the large dimensions of matrices $\boldsymbol{X}_i$ and $\boldsymbol{U}_i$ pose computation problems during the inversion of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{H}}_i$ in equations (6.6) and (6.7), we seek to find a fast method to improve both the speed and computation efficiency, while producing sparse estimates.

Two problems arise when implementing the above empirical Bayes approach when the number of fixed and random effects is large. First, the computational time increases

dramatically for large M and/or M^* due to the need to invert $M \times M$ and $M^* \times M^*$ matrices at each step of the iterative procedure. In addition, when M and/or M^* are large relative to the sample size, estimability problems can arise that lead to lack of convergence of the procedure. Potentially this can be solved by using a MAP estimation approach that includes a proper prior to induce a penalty in the procedure that leads to shrinkage towards the prior and regularization. However, such an approach will be sensitive to hyper-parameter choice. An alternative is to adapt a fast algorithm that will bypass the inversion step leading to a reduced model with dimension $m \times m$ and $m^* \times m^*$ for both $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{H}}_i$ respectively where $m << M$ and $m^* << M^*$. The RVM iterative algorithm can generate such a sparse model and will be discussed in the next section.

6.2.5 A fast MT-RVM method

A fast MT-RVM approach can be used to hasten the estimation process for $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ while overcoming the convergence, estimability and matrix inversion problems. The RVM approach reduces the dimensions of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ by discarding $M - m$ columns of $\boldsymbol{X}_i$ and $M^* - m^*$ columns of $\boldsymbol{W}_i$. These columns correspond to fixed and random effects that can be excluded, since their posteriors are concentrated at zero. The posterior for β_j is concentrated at zero when $\hat{\alpha}_j^{-1} = 0$, while the posterior for h_{ij} is concentrated at zero for all i when $\hat{\omega}_j^{-1} = 0$

The computation of α_k is based on the conditional log-likelihood function,

$$\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) = \frac{-1}{2}\left(N \log(2\pi) + \log|\boldsymbol{C}^*| + \boldsymbol{Y}'\boldsymbol{C}^{*-1}\boldsymbol{Y}\right).$$

This log-likelihood is partitioned into parts with and without the k^{th} component of $\boldsymbol{\alpha}$. We first let $\boldsymbol{C}^* = \boldsymbol{C}^*_{-k} + \alpha_k^{-1}\boldsymbol{X}_{.k}\boldsymbol{X}'_{.k}$ such that $\boldsymbol{C}^*_{-k} = \boldsymbol{V}^* + \sum_{l \neq k} \alpha_l^{-1}\boldsymbol{X}_{.l}\boldsymbol{X}'_{.l}$ where $\boldsymbol{X}'_{.l}$ is the l^{th} column of matrix $\boldsymbol{X}$. The resulting decomposed log-likelihood function is

$$\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) = \ell(\boldsymbol{\alpha}_{-k}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) + \frac{1}{2}\left(\log \alpha_k - \log|\alpha_k + s_k| + \frac{q_k^2}{\alpha_k + s_k}\right),$$

where $s_k = \boldsymbol{X}'_{.k}\boldsymbol{C}^{*-1}_{-k}\boldsymbol{X}_{.k}$ and $q_k = \boldsymbol{X}'_{.k}\boldsymbol{C}^{*-1}_{-k}\boldsymbol{Y}$. The estimate for α_k is

$$\hat{\alpha}_k = \begin{cases} \frac{s_k^2}{q_k^2 - s_k} & \text{if } q_k^2 > s_k, \\ \infty & \text{otherwise.} \end{cases} \tag{6.12}$$

The selection of a candidate $\boldsymbol{X}_{.k}$ is based on its contribution to the log-likelihood $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$, with the fixed effect having the largest contribution selected first. This

is followed by the computation of q_k and s_k, and based on these values we can either add, update or delete the selected fixed effect. We add $\boldsymbol{X}_{.k}$ into the model when $q_k^2 > s_k$. Alternatively, when $\boldsymbol{X}_{.k}$ is already in the model and $q_k^2 > s_k$, we update $\hat{\alpha}_k$ but if $q_k^2 < s_k$ we delete $\boldsymbol{X}_{.k}$ from the model. The process continues until convergence.

Similarly, we can reduce the dimension of the matrix $\boldsymbol{U}_i$ from M^* to m^* based on the estimates of $\boldsymbol{\omega}$ computed using the fast MT-RVM method. Let $\ell(\boldsymbol{\omega}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) = -1/2 \sum_{i=1}^{N} T_i \log(2\pi) + \log|\boldsymbol{V}_i^{*-1}| + (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})' \boldsymbol{V}_i^{*-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})$ be a conditional log-likelihood function of $\boldsymbol{\omega}$ at fixed values of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and σ_ϵ^{-2}. We first partition $\boldsymbol{V}_i^*$ into two parts such that $\boldsymbol{V}_i^* = \boldsymbol{V}_{i,-k}^* + \boldsymbol{V}_{ik}^*$, where $\boldsymbol{V}_{i,-k}^* = \sigma_\epsilon^2 \boldsymbol{I}_{M^*} + \sum_{j \neq k}^{M^*} \omega_j^{-1} \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}'$ and $\boldsymbol{V}_{ik}^* = \omega_k^{-1} \boldsymbol{u}_{ik} \boldsymbol{u}_{ik}'$. This results to

$$\ell(\boldsymbol{\omega}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) = \ell(\boldsymbol{\omega}_{-k}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2}) + \frac{-1}{2} \sum_{i=1}^{N} \left(\log \omega_k - \log|\omega_k + s_{ik}^*| + \frac{q_{ik}^{*2}}{\omega_k + s_{ik}} \right),$$

where $s_{ik}^* = \boldsymbol{u}_{ik}' \boldsymbol{V}_{i,-k}^{-1} \boldsymbol{u}_{ik}$ and $q_{ik}^* = \boldsymbol{u}_{ik}' \boldsymbol{V}_{i,-k}^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}})$. We differentiate $\ell(\boldsymbol{\omega}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$ and set the result to zero. An approximate solution can be obtained by assuming that $\omega_k << s_{ik}^*$ which leads to

$$\hat{\omega}_k \cong \begin{cases} \frac{N}{\sum_{i=1}^{N} (q_{ik}^{*2} - s_{ik}^*)/s_{ik}^{*2}} & \text{if } \sum_{i=1}^{N} \frac{(q_{ik}^{*2} - s_{ik}^*)}{s_{ik}^{*2}} > 0, \\ \infty & \text{otherwise.} \end{cases} \tag{6.13}$$

For a justification of this type of approximation, refer to Ji, et al. (2009).

The selection of the candidate $\boldsymbol{w}_{ik}$ is based on all 29 basis functions. Three operations can take place on $\boldsymbol{w}_{ik}$: add, update and delete. Column vector $\boldsymbol{w}_{ik}$ is added into the model when $\sum_{i=1}^{N} \frac{(q_{ik}^{*2} - s_{ik}^*)}{s_{ik}^{*2}} > 0$. An update occurs when $\sum_{i=1}^{N} \frac{(q_{ik}^{*2} - s_{ik}^*)}{s_{ik}^{*2}} > 0$ and $\boldsymbol{w}_{ik}$ is already in the model. Deletion occurs when $\sum_{i=1}^{N} \frac{(q_{ik}^{*2} - s_{ik}^*)}{s_{ik}^{*2}} < 0$ and $\boldsymbol{w}_{ik}$ is already in the model. The final model has few $\boldsymbol{w}_{ik}$ with $\omega_j < \infty$ while majority tend to have $\omega_j = \infty$ which correspond to $b_{ij} = 0$ for all i. Convergence is accelerated by choosing the fixed and random effects that lead to the largest increase in $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$ and $\ell(\boldsymbol{\omega}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \sigma_\epsilon^{-2})$, respectively. To compute (i.e. add, update, delete) the significant components of the prior vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$, we follow the steps of the algorithm discussed by Tipping (2001) and Ji et. al (2009) to select the basis functions for both the fixed and random effects. Appendix A contains simplified expressions for different quantities that can be used in the modified MT-RVM procedure discussed in this chapter.

6.3 Simulation Study

We simulate data designed to mimic the trajectories in bbt over the cycle. The i^{th} curve is generated from the model

$$y_{it} = v_i + \rho_i z_{it} \sin(10 z_{it} - r_i) + \epsilon_{it}, \qquad t = 1, 2, \ldots T_i, \quad i = 1, 2, \ldots N,$$

where covariate $z_{it} \sim unif(0, 1)$, parameter $\rho_i \sim unif(0.5, 1.5)$ controls the amplitude of the curve, $r_i \sim unif(-1, 1)$ and $v_i \sim unif(-1, 1)$ are horizontal and vertical shift parameters for the i^{th} curve and $\epsilon_{it} \sim N(0, \sigma^2_\epsilon = 0.10)$. The scale, horizontal and vertical shift parameters vary between subjects but they remain constant among observations within the same curve.

The smoothing basis functions $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$ were generated from the standardized values of $\boldsymbol{z}_i$ based on 27 cubic B-splines with 23 interior knots. For this particular case, the design matrix $\boldsymbol{X}_i$ is $T_i \times 29$ such that $M = 29$. We have 27 columns of basis functions and two additional columns containing 1's and $\boldsymbol{z}_i$ (i.e. $\{1, z_{it}\}_{t=1}^{T_i}$). Similarly, matrix $\boldsymbol{W}_i$ is of order $T_i \times M^*$ where $M^* = 29$.

The computation of the relevant fixed and random effects is based on the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$. We start with an empty model and select the first relevant fixed effect according to the discussion in section 6.2. This is followed by computation of the corresponding random effect. The subsequent iterations involve selection of relevant fixed effects followed by selection of the random effects.

The MT-RVM procedure discussed in section 2 was implemented to generate sparse LME models. In addition, we attempted to implement a variety of other reduced rank methods including functional principal components (FPCA) (James et al., 2001; Crainiceanu, 2009) and the adaptive fence method (Jiang et al., 2008). We do not show the results for the adaptive fence method since the approach required fitting of linear mixed models of varying dimension in implementing basis selection. We encountered convergence problems in fitting LME models containing most or all of the potential basis functions. In addition to the reduced rank methods, we considered recent approaches proposed by Durban et al. (2005) and Wand and Ormerod (2005), with the latter approach implemented using MCMC. Another alternative is the method of Scarpa and Dunson (2009), which allows a nonparametric contamination of a parametric hierarchical model and is quite computationally intensive. As the goal of the MT-RVM method is instead to obtain a fast approach for fitting of functional data in the absence of a known parametric model, we do not consider their method further. A final possibility we considered is a two stage approach in which the individual curves are smoothed, and then a model

is fitted on a fixed grid that is common across the subjects (Ramsay and Silverman, 1997). However, this type of approach is known to only perform well when there is a high signal-to-noise ratio and there are many observations per subject that are regularly spaced without regions of missing data. For the bbt data, there tends to be substantial noise and different observation times for the different cycles.

To investigate the performance of the RVM method we estimated the fitted curves for varying choices of the number of observations per subject (T_i) and the number of the subjects (N). The true curve is $f_i^{true}(z_{it}) = v_i + \rho_i z_{it} \sin(10 z_{it} - r_i)$. Estimates for f_i are obtained for the four procedures -RVM, FPCA (Crainiceanu, 2009, frequentist (Durban et al., 2005) and MCMC based (Wand & Ormerod, 2008) methods. We considered four cases for the number of subjects, $N = 25, 50, 100$ and 500. In each case, we vary the number of observations per subject such that $T_i = 10, 20, 30$. We generate 100 replications and each replication has a specified N and T_i, e.g. for a case with $N = 50$ and $T_i = 10$, the j^{th} replication has 50 subjects and each subject has 10 observations. In each replication we plot both the original and the fitted curve for each subject. We extract the values of f_i^{true} and $\hat{f}_i$ based on a grid $\mathbf{z}^* = (z_1^*, \cdots, z_T^*)$ where $\mathbf{z}^*$ is a vector containing values of $T = 100$ finely division points of a grid that covers the curves such that $T >> T_i$. To compare the estimated curves from the four procedures, we compute Mean Integrated Square Error (MISE) and Bias. The MISE and bias for the case with N subjects and T_i observations is

$$MISE = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \hat{f}_i(z_t^*) - f_i^{true}(z_t^*) \right\}^2,$$

$$Bias = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left| \hat{f}_i(z_t^*) - f_i^{true}(z_t^*) \right|.$$

On average, the RVM procedure selected $m = 3$ fixed and $m^* = 10$ random effects. Tables 6.1 and 6.2 present the results for the MISE and Bias respectively. Columns 1, 2, 3, and 4 in both tables present results for frequentist, MCMC based, RVM and FPCA procedures respectively. Both tables show that there is a gradual decrease in both MISE and Bias values as the number of observations increases and this hold for all procedures. Both smoothing approaches based on the full model perform well relative to the reduced model. Although both tables do not demonstrate a general improvement in performance for the proposed RVM approach over the FPCA method, it is important to consider two factors. First, as the number of observations increase, MISE and bias for both methods are almost identical. Second, when the number of observations and subjects increase (e.g. 500 subjects with 30 observations each), the FPCA approach fails due to convergence problems. The frequentist approach seems to be the best since

TABLE 6.1: A table for the Mean Integrated Squared Error for the four curve fitting procedures. CI's are the empirical 95% intervals for the MISE estimates for the simulation replicates.

		MISE			
N	T_i	*Frequentist* (CI)	$MCMC$ (CI)	RVM (CI)	$FPCA$ (CI)
25	10	0.021 (0.001,0.042)	0.029 (0.013,0.046)	0.110 (0.027,0.194)	0.045 (0.031,0.060)
	20	0.006 (0.004,0.008)	0.004 (0.003,0.006)	0.025 (0.005,0.041)	0.032 (0.021,0.042)
	30	0.004 (0.003,0.006)	0.004 (0.003,0.005)	0.019 (0.008,0.029)	0.019 (0.011,0.028)
50	10	0.023 (0.015,0.030)	0.034 (0.025,0.043)	0.095 (0.083,0.106)	0.031 (0.023,0.038)
	20	0.006 (0.005,0.007)	0.007 (0.004,0.011)	0.037 (0.025,0.050)	0.023 (0.019,0.028)
	30	0.004 (0.003,0.005)	0.005 (0.003,0.006)	0.023 (0.005,0.040)	0.019 (0.016,0.023)
100	10	0.016 (0.012,0.120)	0.021 (0.017,0.027)	0.085 (0.044,0.137)	0.033 (0.028,0.037)
	20	0.006 (0.005,0.007)	0.008 (0.006,0.010)	0.024 (0.019,0.029)	0.025 (0.021,0.029)
	30	0.004 (0.003,0.005)	0.004 (0.003,0.005)	0.020 (0.007,0.031)	0.021 (0.017,0.025)
500	10	0.017 (0.015,0.021)	0.018 (0.016,0.020)	0.040 (0.022,0.049)	0.033 (0.029,0.036)
	20	0.006 (0.005,0.006)	0.006 (0.004,0.007)	0.024 (0.023,0.026)	0.026 (0.025,0.027)
	30	0.004 (0.003,0.004)	0.004 (0.003,0.004)	0.005 (0.004,0.006)	–

TABLE 6.2: A table for the Bias for the four curve fitting procedures. CI's are the empirical 95% intervals for the simulation replicates.

		Bias			
N	T_i	*Frequentist* (CI)	$MCMC$ (CI)	RVM (CI)	$FPCA$ (CI)
25	10	0.094 (0.070,0.119)	0.117 (0.091,0.142)	0.244 (0.147,0.341)	0.160 (0.135,0.188)
	20	0.061 (0.053,0.069)	0.052 (0.045,0.059)	0.109 (0.068,0.150)	0.135 (0.112,0.160)
	30	0.051 (0.043,0.059)	0.051 (0.048,0.055)	0.104 (0.072,0.136)	0.104 (0.110,0.142)
50	10	0.099 (0.091,0.108)	0.108 (0.099,0.117)	0.218 (0.201,0.235)	0.130 (0.114,0.147)
	20	0.059 (0.055,0.063)	0.064 (0.056,0.072)	0.147 (0.124,0.170)	0.112 (0.101,0.124)
	30	0.049 (0.046,0.053)	0.053 (0.048,0.057)	0.114 (0.073,0.154)	0.102 (0.094,0.108)
100	10	0.088 (0.071,0.101)	0.102 (0.086,0.125)	0.200 (0.109,0.131)	0.133 (0.122,0.143)
	20	0.059 (0.057,0.062)	0.066 (0.062,0.071)	0.117 (0.108,0.123)	0.115 (0.106,0.124)
	30	0.050 (0.047,0.052)	0.051 (0.048,0.054)	0.107 (0.074,0.140)	0.105 (0.097,0.112)
500	10	0.091 (0.086,0.096)	0.091 (0.088,0.095)	0.116 (0.111,0.120)	0.117 (0.114,0.119)
	20	0.058 (0.057,0.060)	0.059 (0.057,0.061)	0.115 (0.113,0.119)	0.115 (0.110,0.120)
	30	0.050 (0.048,0.052)	0.050 (0.049,0.051)	0.052 (0.051,0.054)	–

it has the least MISE and bias. However, we note that as the number of subjects and observations increases we encountered convergence problems. Figure 6.1 presents a plot for one simulation case with 30 observations. The figure has five curves representing the true sine curve and the estimated curves based on the four curve fitting methods.

Table 6.3 presents the average time taken to fit a model for each simulated case based on the four procedures. The results were generated using an R software (version 2.8.1) on a Pentium IV, 2.4GHz, 512MB, Windows XP computer platform. Results show an increase on the average time spent to fit the models as the number of observations

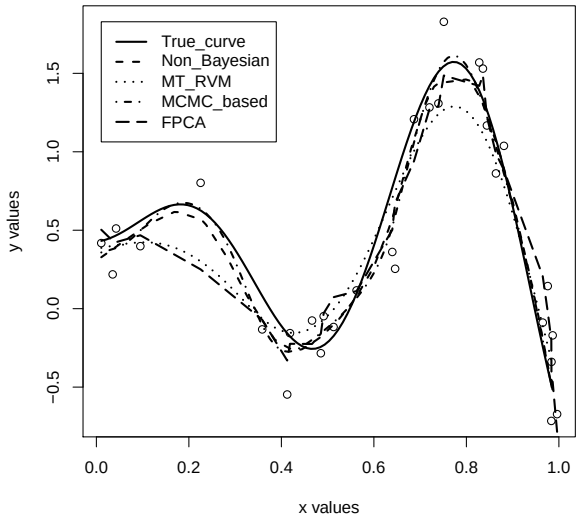

FIGURE 6.1: Estimated sine curve using the four procedures.

TABLE 6.3: A table for the average time taken by the four procedures to fit models in each simulation case.

		Computation time (in seconds)			
N	T_i	*Frequentist*	*MCMC*	*RVM*	*FPCA*
25	10	2.551	13.683	3.419	0.837
	20	2.817	21.013	2.922	1.419
	30	3.304	24.669	2.778	2.255
50	10	2.634	17.151	2.678	0.787
	20	3.616	22.784	2.448	1.385
	30	4.037	24.851	2.425	2.122
100	10	2.656	18.643	2.166	0.723
	20	3.696	22.661	1.901	1.426
	30	4.450	27.651	2.144	2.245
500	10	2.151	21.014	2.024	1.154
	20	3.819	23.121	2.157	2.358
	30	4.601	28.241	2.201	–

per cycle increased. Frequentist approach takes the least time in all cases while the MCMC based approach performs the worst. FPCA approach performs better that the RVM procedure but we note that as the number of subjects increases (e.g. over 100 subjects with at least 30 observations) the RVM approach performs better. Hence, the main advantages of the proposed RVM approach include the ability to obtain automated variable selection even in moderately high dimensional random effects models without facing convergence problems.

6.4 Application to the bbt measurements

Our research is motivated by the basal body temperature data from the European fecundability study (Colombo and Masarotto, 2000). The study enrolled 880 women, aged between 18 and 40 years, who were not taking hormonal medications or drugs affecting fertility, and had no known impairment of fecundity. The participants kept daily records of cervical mucus or basal body temperature measurements from at least one menstrual cycle, and they recorded the days during which intercourse and menstrual bleeding occurred. For more details about the study protocol, refer to Colombo and Masarotto (2000). In this study we considered bbt measurements from 520 menstrual cycles.

A standard bbt curve has biphasic shape and is characterized with three phases representing the pre-ovulation, ovulation and post-ovulation periods. The ovulation day is commonly identified using the three over six rule (Colombo and Masarotto, 2000) or by identifying the day that correspond to a dip that is followed by a sharp rise in the bbt curve. In reality, the classic bbt pattern is difficult to replicate and wide fluctuations in bbt, with many false nadirs and peaks, are commonly observed. Fluctuations result from a host of factors, other than hormonal fluctuations, that affect a womans bbt: amount of sleep, sleep disturbances, ambient bedroom temperature, convection currents, food ingestion and emotional state (Colombo and Masarotto, 2000).

Numerous methods have been proposed to estimate the shape of the bbt curves. For example, Scarpa and Dunson (2008) proposed a Bayesian semi-parametric model based on nonparametric contamination of a linear mixed effects model. However, implementation of the approach relies on a highly computationally intensive MCMC algorithm and it fails to produce smooth curves. To implement the RVM procedure, we used the cubic B-splines to generate the basis functions $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$ based on the standardized values of time covariate ($\boldsymbol{z}_i$). In total we generated $M = M^* = 29$ columns for matrices $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ respectively, where the first and the second columns contain values of 1's and $\boldsymbol{z}_i$. The RVM procedure was implemented on data for the 520 cycles and the final model has $m = 3$ fixed effects (basis functions: 1, 2 and 28) and $m^* = 10$ random effects (basis functions: 1, 2, 28, 29, 9, 14, 27, 12, 11 and 16). The expressions for the credible intervals for the RVM parameter estimates are presented in appendix B.

Figure 6.2 shows estimated curves based on the RVM method from four randomly selected cycles. Each plot shows the estimated bbt curve and the gray region represents the 95% confidence band. Figure 6.3 shows a plot for the population and subject specific curves based on the estimates from the RVM procedure. The thick black curve represents the population average bbt curve while the thin gray curves represent the estimated subject specific bbt curves. The population average curve shows a biphasic

shape that has a gentle rise starting from day 10 and reaches the climax around day 24. The interval characterized with a gentle rise in bbt curve is the most probable period in the menstrual cycle when majority of women experience ovulation.

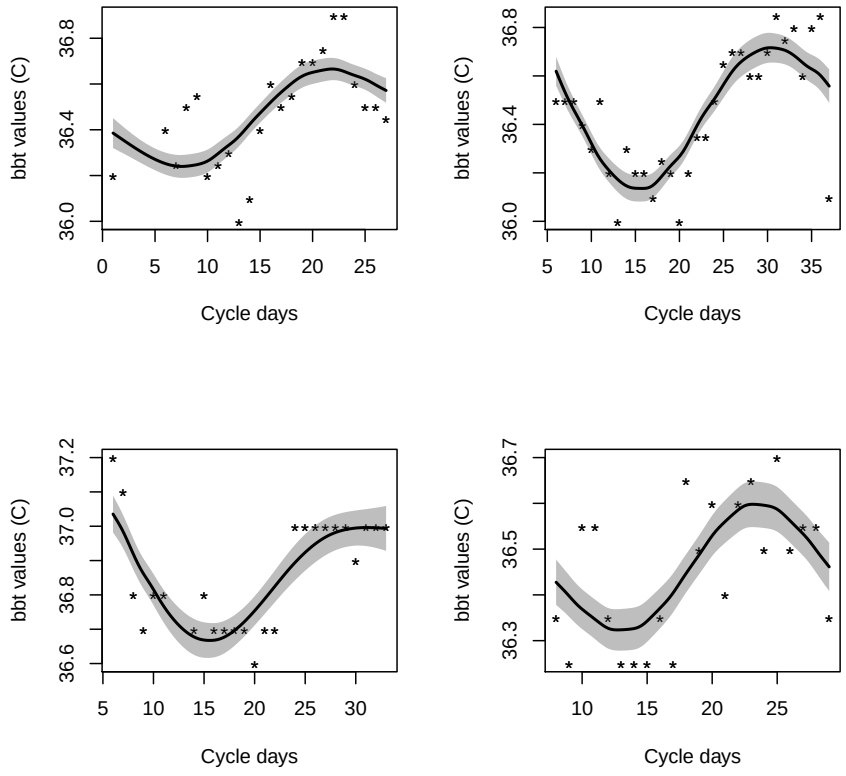

FIGURE 6.2: Estimated bbt curves and the 95% confidence band from the RVM procedure.

To evaluate the predictive ability of the proposed RVM procedure, we conducted an out-of-sample prediction process where we first dropped 10% of observations (chosen at random from among the different women) and then predict the bbt value at the times for these values. Figure 6.4 presents a plot for the ordered true observations against the predicted bbt values. The plot shows moderately high correlation between the predicted and observed bbt values with a correlation coefficient value of 0.82. We extended the evaluation by increasing the out-of-sample proportions to 25% and 50% and computed their correlation values. The correlation values for 25% and 50% out-of-sample proportions are 0.78 and 0.75 respectively, suggesting that accurate predictions can be obtained even with 50% of the data discarded. Figure 6.5 shows the estimated curves based on the RVM procedure from four randomly selected subjects. The thick line represent the estimated curve based on all observations in a cycle while the thin line represents the estimated curve when 20% of the observations are excluded. We predicted the 20% excluded observations and included them on the plots. The thick and

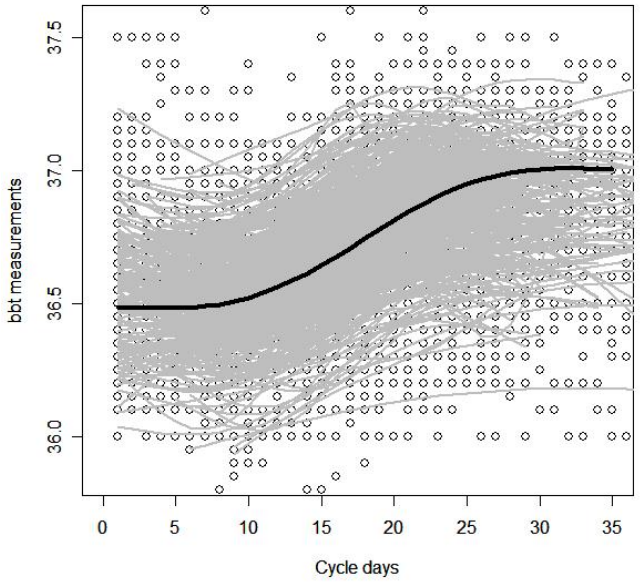

FIGURE 6.3: A plot for the population and subjects specific bbt curves from the model fitted using the RVM procedure.

small vertical lines represent the 95% confidence intervals while the thick star (∗) at the middle of the vertical lines represent the predicted estimates. This result has substantial clinical implications, as women may be able to collect fewer bbt observations without greatly reducing the accuracy of the estimated bbt curve over the cycle.

Basal body temperature curves provide a useful non-invasive marker of reproductive functioning, which can be quite informative to clinicians monitoring women attempting pregnancy for the purposes of providing guidance on highly fertile days of the woman's cycle, as well as inferring possible causes of a delay in conception if it occurs. Before recommending women having trouble conceiving to assistant reproductive technology (AFT), it has been increasingly recommended to follow the woman prospectively for at least several cycles using natural biomarkers, such as basal body temperature. The woman may provide unevenly spaced and sparse measurements on bbt across her cycle, which can be difficult for the clinician to interpret. By using our methodology, smoothed basal body temperature curves across the cycle can be estimated based on the available data, while borrowing information from the rich historical bbt data base to aid in filling in the gaps in the data. As shown for our out-of-sample prediction results, the proposed methods appear to do a good job in interpolating across regions of missing data in estimating the bbt curves. In performing online estimation of the bbt curves for new women, it is not necessary to re-estimate the population parameters, including the fixed effects and random effects covariance. Instead we can store the values of these parameters, estimated based on our data base and updated periodically. Then, a curve for

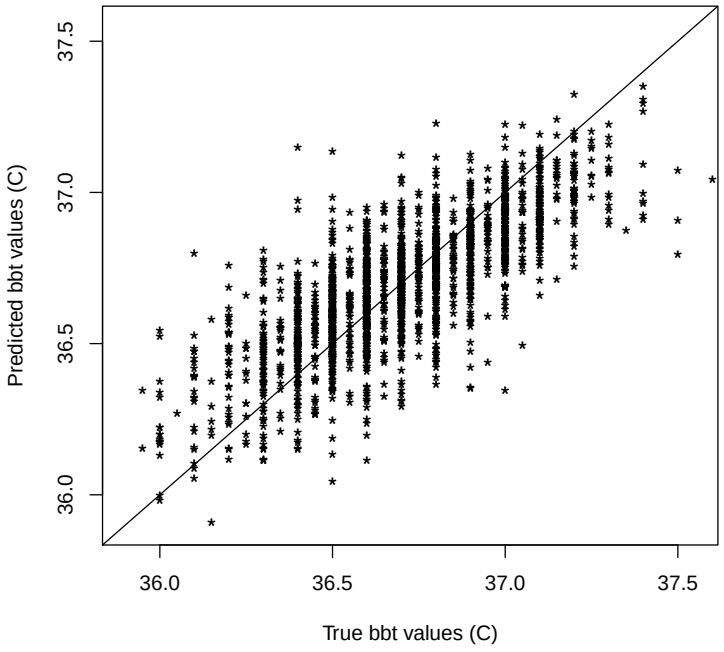

FIGURE 6.4: A plot for the predicted against true bbt values.

an incoming woman can be estimated extremely rapidly by a physician using a simple algorithm than could even be implemented in Excel.

6.5 Discussion

Our proposed MT-RVM method adds to the literature on variable selection procedures for the random effects component of the LME model (6.2). For fixed effects, variable selection can proceed using likelihood ratio tests, goodness-of-fit criteria and other methods applied routinely in linear regression models. In selecting the random effects, a complication arise since the null hypothesis lies on the boundary of the parameter space. Hence, the likelihood ratio (LR) test statistic no longer has an asymptotic chi-square distribution and the justification for the BIC and other criteria for model selection breaks down. A variety of approximations have been proposed for the distribution of the LR test statistic under the null hypothesis. The most widely used approach relies on the method of Stram and Lee (1994), which may not have good performance (Crainiceanu and Ruppert 2004).

In conducting random effects selection in linear mixed models, Jiang, et al. (2008) proposed the fence method that automatically selects a subset of predictors from the vast number of possible subsets under consideration. The method is easy to implement

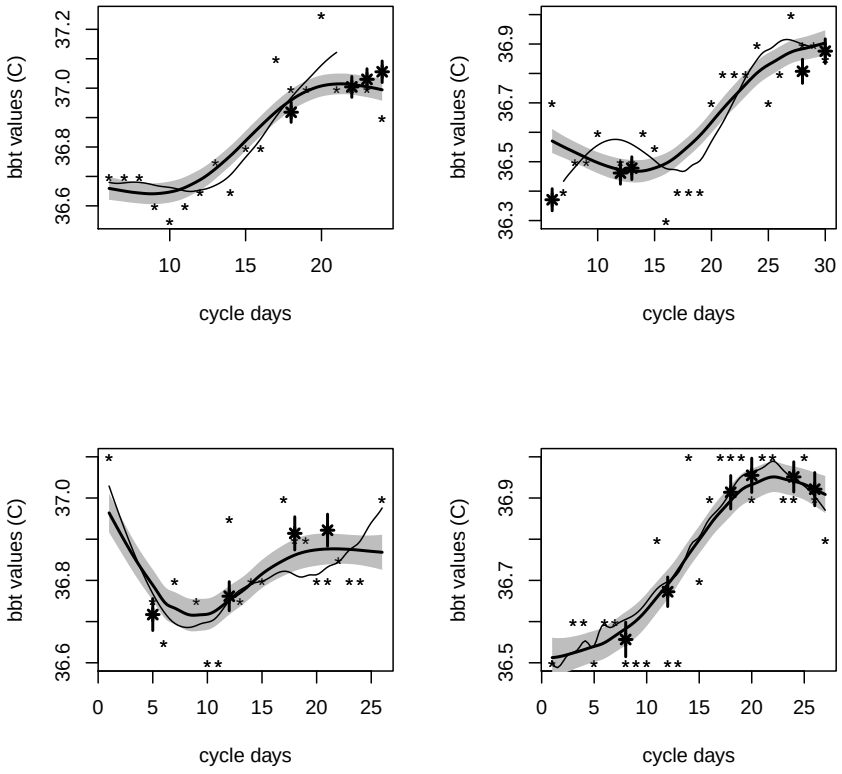

FIGURE 6.5: Estimated bbt curves using RVM (continuous curve), predicted plots (dotted curve) without 20% observations and the predicted out of sample observations.

and has good performance in modest dimensional model, but rapidly becomes computationally infeasible as the number of candidate predictors increases. Functional principal components analysis (FPCA) is another appealing approach that can be used to reduce the dimensionality of a functional linear mixed model. The approach improves numerical stability of parameter estimates through reduced rank models as motivated in James, Hastie and Sugar (2001) and others. However, fitting a FPCA model often relies on algorithms used for fitting linear mixed effects models, and we have encountered similar convergence problems. Moreover, there is the issue of how to select the number of principal components, with standard approaches being based on cross-validation, which can be time-consuming to implement. Often one tries increasing the number of principal components until the improvement in fit is negligible, though this is somewhat ad hoc. The simulation results for the FPCA procedure revealed that when the number of subjects and observations increases drastically the method fails.

In this chapter we have developed a fast approximate Bayes approach for simultaneous variable selection and fitting of linear mixed effects models, motivated by functional data

analysis applications in which candidate predictors correspond to different basis functions. The proposed approach is motivated by the MT-RVM Ji, et al. (2009) methodology, which was developed to borrow information across related signals in performing reconstruction based on compressive sensing measurements. The RVM approach has the positive features of allowing automatic basis selection without encountering convergence problems. In adapting RVM to perform variable selection in linear mixed models, it was necessary to incorporate two non-trivial modifications, with the first allowing non-zero values for the fixed effects and the second allowing non-diagonal random effects covariance. The resulting methodology can be implemented very rapidly and can accommodate high dimensions, leading to advantages over existing methods for variable selection in random effects models (Pauler et al., 1999).

Chapter 7

Multi-level relevance vector machine with applications to hierarchical functional data analysis

7.1 Introduction

In this chapter we extend the implementation of the MT-RVM procedure to multi-level data. In particular, we develop a multi-level relevance vector machine (ML-RVM) that can handle nested data. We will consider a model similar to Brumback and Rice (1998), where the measurements are nested withing cycles and cycles are nested within subjects. To avoid computation complexities resulting from additional data hierarchy, we consider a simple multi-level functional mixed model that assumes independent random effects at both cycle and subject specific levels. The ML-RVM approach is implemented on the bbt data and generate a flexible and sparse functional mixed model that can estimate population-average, subject and cycle specific curves.

Many medical and epidemiological studies collect massive multi-level functional data. The data results from repeated collection of measurements at different time points from series of clusters and subjects. For example, functional data that are collected for patients nested within studies as in many multi-center studies or functional data measured at repeated times for each patient e.g., medical images taken at each visit to the clinic and hormone curves for each cycle. In most cases, it is common to have measurements from different clusters and subjects often sparse and unequally-spaced due to routine

entry of patient information into computer data bases. Hence, difficult to accurately estimate trajectories from multiple clusters within a subject or center. For physicians working with many patients, automated tools are required to rapidly estimate and predict trajectories for a current patient while utilizing information in the data base.

Our motivation is drawn from multi-level data generated from reproductive studies. In particular, we consider the basal body temperature data (Colombo and Masarotto, 2000) collected to study the characteristic of the bbt trajectories from multiple cycles that can be used to predict the ovulation day or identify cycles with possible dysfunction. Hence, before recommending assistant reproductive technology (AFT) to women having trouble conceiving, it is advisable to first consider the available information from the bbt trajectories. The pattern of the bbt measurements over the menstrual cycle provides a natural informative marker of reproductive functioning that can provide guidance on highly fertile days of the woman's cycle as well as inferring possible causes of a delay in conception.

Functional data analysis (FDA) (Ramsay and Silverman, 2005) can be used to model sparse functional data. Numerous related approaches can be found in Rice and Wu (2001); Wu and Zhang (2002); Liang, Wu, and Carroll (2003); Wu and Liang (2004). The development of their approaches fall under the connection between mixed-effects models (Laird and Ware, 1982; Diggle et al., 1994) and smoothing spline functions (Wand, 2003) but their scope is limited to a single level of hierarchy. The extension of the FDA approaches into functional multi-level data can be found in Brumback and Rice (1998); Guo (2002); Morris et al. (2003); Durban, et al. (2005) among others. To model functional data occurring within nested hierarchy, Brumback and Rice (1998) introduced flexible smoothing spline method and considered individual specific trajectories as fixed instead of random effects. Guo (2002) introduced spline-based functional mixed model under a broad range of fixed and random effect structures while Morris et al. (2003) developed a functional mixed model based on wavelet-based methodology (Baladandayuthapani et al., 2008). Recent discussions on the use of spline smoothing methods based on mixed models can be found in Morris and Carroll (2006); Wand and Ormerod (2008); Crainiceanu (2009).

The hierarchical structure of the data, heterogeneity among the subjects' trajectories and the large dimension of the resulting functional mixed models make the approaches difficult to implement in Bayesian framework. A common solution is to reduce the dimension of the functional model. However, it is difficult to choose the basis functions in advance, motivating the use of adaptive methods that allow uncertainty in basis function selection using the time consuming reversible-jump markov chain monte carlo (Green, 1995) or stochastic search variable selection (Smith and Kohn, 1996) procedures (Bigelow

and Dunson, 2007; Thompson and Rosen, 2008). Functional Principal Component Analysis (FPCA) methodology has also been used extensively to reduce the dimension of the random effects (James, et al., 2001; Yao, et al., 2005; Crainiceanu, 2009). However, the posterior computation of model parameters in both classes of procedures are based on computationally expensive Markov Chain Monte Carlo (MCMC) algorithms. Hence, there is a clear practical motivation for fast approximate Bayes approaches that bypass MCMC while maintaining some of the benefits of a Bayesian analysis.

This paper proposes to use a large number of B-spline basis function for the multi-level functional mixed model and estimate a sparse model using relevance vector machine (RVM) methodology (Tipping, 2001). RVM is a fast approximate Bayes functional data analysis that relies on sparseness-favouring hierarchical priors for basis coefficients. The approach is widely used in machine learning and is one among fast Bayesian methods that promote sparseness in estimation of the basis coefficients, providing a more flexible alternative to Support Vector Machines (SVM) (Burges, 1998) and LASSO (Tibshirani, 1996), leading to a sparser solution that is more robust to outliers (Tipping, 2001; Ji, et al., 2009). The Ji et al. (2009) multi-task relevance vector machine (MT-RVM) approach implicitly assumed that the random effects distribution was centered at zero while Ciera and Dunson (2010) extended the approach to a non-zero mean that allow separate fixed and random effects and accommodate correlation among the random effects. We aim to extend the MT-RVM approach into multi-level relevant vector machine (ML-RVM) that can accommodate multi-level functional data. To avoid introducing additional computation complexities caused by the increase of hierarchy levels, our approach considers independence covariance structure for the random effects at both cycle and subject-specific levels. This is a slight deviation from Ciera and Dunson (2010) approach that allowed a general covariance structure for the random effects. Moreover, at the subjects level, we only consider two subject specific random effects (the intercept and the slope) since we assume that the average subjects' curves have a common biphasic pattern with varying intercept and slope. To allow more flexibility, this assumption can be relaxed and we consider subject specific random effects from all basis functions.

7.2 Functional Data Analysis Model

7.2.1 Motivating problem

In practice, the shapes of the average bbt trajectories at subject specific level from many women is mostly the same. However, it is common to encounter variations among cycle trajectories due to the day to day physical and psychological disorders that are

commonly experienced by women. For example, when working with bbt data, factors like hormonal fluctuations or sleep disturbances may cause abrupt temperature change leading to abnormal trend on the bbt curve in certain cycles from different women (Colombo and Masarotto, 2000).

To accommodate these cycle to cycle variations, we allow variations to occur among all random effects at cycle specific level. But to maintain the biphasic pattern assumption on the mean trajectories at subject specific level, we consider only two random effects (intercept and slope) at subject specific level. Hence, we consider a case where curves within a subject differ in shape but the average curve from different women are assumed to have a common biphasic structure that differ in height and slope within the x-y plane. Allowing such structural variability among subject and cycle specific curves, we can maintain the existing heterogeneity among subjects and variations among cycle trajectories that are caused by random factors that interferes with cycle trajectories within a subject.

7.2.2 Functional mixed effects model

We consider a general multi-level mixed model for data with three hierarchical levels such that observations (level 1) are nested within cycles (level 2) and cycles are nested within subjects (level 3). For example, in the bbt data application, the daily bbt measurements are nested within menstrual cycles and the cycles are nested within women. In such a framework, we have $i = 1, \cdots, N$ subjects, where the i^{th} subject has $j = 1, \cdots, n_i$ cycles and the j^{th} cycle has $t = 1, \cdots, T_{ij}$ bbt observations. Further, we denote that $T_i = \sum_{j=1}^{n_i} T_{ij}$ is the total number of observations within a subject.

A standard bbt curve can be estimated using functional data analysis (FDA) models described in the literature (Ramsay and Silverman, 1997; Ruppert, et al. 2003). Let $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijT_{ij}})'$ and $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijT_{ij}})'$ be the response and covariate vectors for the j^{th} cycle and the i^{the} subject such that $z_{ij1} < z_{ij2} < \ldots < z_{ijT_{ij}}$. A functional model for the t^{th} observation in the j^{th} cycle and the i^{th} subject can be represented as

$$y_{ijt} = f_{ij}(z_{ijt}) + \epsilon_{ijt}, \quad \epsilon_{ijt} \sim N(0, \sigma_\epsilon^2), \quad t = 1, \ldots, T_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, N. \tag{7.1}$$

where $f_{ij}(.)$ is a smooth function for cycle j and subject i while ϵ_{ijt} is a measurement error.

The functional data model in equation (7.1) can be generalized into functional mixed effects model for a multi-level hierarchical data. For a general case, we let the smoothing

function $f_{ij}(z_{ijt})$ be represented as a linear combination of two sets of basis functions

$$f_{ij}(z_{ijt}) = \sum_{l=1}^{M} \beta_l \varphi_l(z_{ijt}) + \sum_{l=1}^{M^*} (\gamma_{il} + \upsilon_{ijl}) \phi_l(z_{ijt}),$$

where $\sum_{l=1}^{M} \beta_l \varphi_l(z_{ijt}) = \boldsymbol{x}_{ijt}{}'\boldsymbol{\beta}$ and $\sum_{l=1}^{M^*} (\gamma_{il} + \upsilon_{ijl}) \phi_l(z_{ijt}) = \boldsymbol{w}_{ijt}{}'(\boldsymbol{\gamma}_i + \boldsymbol{\upsilon}_{ij})$ such that $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)'$ are fixed effects and $\boldsymbol{x}_{ijt} = (x_{ijt1}, \cdots, x_{ijtM})'$ are the values of the basis functions at z_{ijt}. Similarly, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \cdots, \gamma_{iM^*})'$ and $\boldsymbol{\upsilon}_{ij} = (\upsilon_{ij1}, \cdots, \upsilon_{ijM^*})'$ are the random effects at subject and cycle levels while $\boldsymbol{w}_{ijt} = (w_{ijt1}, \cdots, w_{ijtM^*})'$ are the values of the basis functions at z_{ijt}. The basis functions $\boldsymbol{\varphi} = \{\varphi_l\}_{l=1}^{M}$ and $\boldsymbol{\phi} = \{\phi_l\}_{l=1}^{M^*}$ are collection of basis functions for the fixed and random effects components and can be generated using numerous methods that have been discussed in the literature (e.g. Hastie, et al. 2001; Ruppert, et al. 2003).

To represent the smoothing function in a mixed model formulation where the random effects at subject level consist of the intercept and a slope, the design matrices for the fixed and random effects of the j^{th} cycle in the i^{th} subject can be represented by $\boldsymbol{X}_{ij} = (\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijT_{ij}})'$ and $\boldsymbol{W}_{ij} = (\boldsymbol{w}_{ij1}, \ldots, \boldsymbol{w}_{ijT_{ij}})'$ respectively. The design matrix for the random effects at subject level is represented by $\widetilde{\boldsymbol{W}}_{ij}$ that consist of $1's$ and $\boldsymbol{z}_{ij}$ columns for the i^{th} subject and j^{th} cycle. The functional mixed effects model for the j^{th} cycle of the i^{th} subject can be represented as

$$\begin{aligned} \boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{W}}_{ij}\boldsymbol{\gamma}_i + \boldsymbol{W}_{ij}\boldsymbol{\upsilon}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad \boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, \boldsymbol{E}^{-1}) \quad \boldsymbol{\upsilon}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}), \\ \boldsymbol{\epsilon}_{ij} \sim N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_{T_{ij}}) \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, N, \end{aligned} \tag{7.2}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)'$ are M population mean parameters, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2})'$ are the random effects for the i^{th} subject that capture deviations from the population mean while $\boldsymbol{\upsilon}_{ij} = (\upsilon_{ij1}, \cdots, \upsilon_{ijM^*})'$ are the random effects for the i^{th} subject and j^{th} cycle that capture deviations from the subject mean. The covariance matrices $\boldsymbol{E}$, and $\boldsymbol{\Omega}$ for the random effects are diagonal matrix with elements $(\alpha_{\gamma 1}, \alpha_{\gamma 2})$, and $(\omega_1, \cdots, \omega_{M^*})$ respectively. The random vectors $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \cdots, \epsilon_{ijT_{ij}})'$ are measurement errors that are assumed to be independent. We can stacking together $\boldsymbol{X}_{ij}$, $\widetilde{\boldsymbol{W}}_{ij}$ and $\boldsymbol{W}_{ij}$ for different cycles within a subject. The matrix representation for the subject specific functional

mixed model is

$$\begin{pmatrix} \boldsymbol{y}_{i1} \\ \boldsymbol{y}_{i2} \\ \boldsymbol{y}_{i3} \\ \vdots \\ \boldsymbol{y}_{in_i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{i1} \\ \boldsymbol{X}_{i2} \\ \boldsymbol{X}_{i3} \\ \vdots \\ \boldsymbol{X}_{in_i} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_M \end{pmatrix}' + \begin{pmatrix} \widetilde{\boldsymbol{W}}_{i1} & \boldsymbol{W}_{i1} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \widetilde{\boldsymbol{W}}_{i2} & \boldsymbol{0} & \boldsymbol{W}_{i2} & \dots & \boldsymbol{0} \\ \widetilde{\boldsymbol{W}}_{i3} & \boldsymbol{0} & \dots & \dots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widetilde{\boldsymbol{W}}_{in_i} & \boldsymbol{0} & \dots & \dots & \boldsymbol{W}_{in_i} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_i \\ \boldsymbol{\upsilon}_{i1} \\ \boldsymbol{\upsilon}_{i2} \\ \boldsymbol{\upsilon}_{i3} \\ \vdots \\ \boldsymbol{\upsilon}_{in_i} \end{pmatrix}' + \begin{pmatrix} \boldsymbol{\epsilon}_{i1} \\ \boldsymbol{\epsilon}_{i2} \\ \boldsymbol{\epsilon}_{i3} \\ \vdots \\ \boldsymbol{\epsilon}_{in_i} \end{pmatrix}$$

Following Hedeker and Gibbons (2006) approach, the above matrix representation can be expressed using the standard linear mixed effects model,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i^*\boldsymbol{\upsilon}_i^* + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\upsilon}_i^* \sim N(\boldsymbol{0}, \boldsymbol{D}^{*-1}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_{T_i}), \quad i = 1, \dots, N, \tag{7.3}$$

where $\boldsymbol{W}_i^* = \{\widetilde{\boldsymbol{W}}_i, diag(\boldsymbol{W}_{i1}, \cdots, \boldsymbol{W}_{in_i})\}$ such that $\widetilde{\boldsymbol{W}}_i = (\widetilde{\boldsymbol{W}}_{i1}, \cdots, \widetilde{\boldsymbol{W}}_{in_i})'$. The coefficients vector $\boldsymbol{\upsilon}_i^* = (\boldsymbol{\gamma}_i, \boldsymbol{\upsilon}_{i1}, \cdots, \boldsymbol{\upsilon}_{in_i})'$ contains all random effects for the i^{th} subject, covariance matrix $\boldsymbol{D}^* = diag(\boldsymbol{\alpha}_\gamma, \boldsymbol{\Omega}, \cdots, \boldsymbol{\Omega})$ and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}, \cdots, \boldsymbol{\epsilon}_{in_i})'$ are the error terms. We assume that the random components in $\boldsymbol{\upsilon}_i^*$ and $\boldsymbol{\epsilon}_i$ are independent.

Both frequentist and Bayesian methodologies can be used to estimate and make inference on parameters for a hierarchical functional mixed effects model. For example, in a frequentist framework, we can use the approach described in Hedeker and Gibbons (2006) to estimate parameters for a mixed-effects models for multi-level data. Similarly, in Bayesian environment we can use the MCMC based methods to compute the posterior estimates for the model parameters (see Gelman and Hill, 2007).

To compute the posterior estimates for parameters in the linear mixed effects model (3), we can specify the priors distributions $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{A}^{-1})$, $\boldsymbol{\upsilon}_{ij} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, $\boldsymbol{\gamma}_i \sim N(0, \boldsymbol{E}^{-1})$ and $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_{T_i})$ where $\boldsymbol{A} = diag(\alpha_1, \cdots, \alpha_M)$ while $\boldsymbol{E}$ and $\boldsymbol{\Omega}$ are as described before. The covariance matrices $\boldsymbol{E}$ and $\boldsymbol{\Omega}$ for the random effects can be assigned inverse-Wishart priors. Unfortunately, the inverse-Wishart priors cannot allow the shrinkage of the covariance matrix components (Chen and Dunson, 2003). To allow the shrinkage and accommodate the implementation of the RVM procedure, we assume that the random effects $\boldsymbol{\upsilon}_{ij}$ are independent resulting to a diagonal covariances matrix $\boldsymbol{\Omega} = diag(\omega_1, \cdots, \omega_{M^*})$ and $\boldsymbol{E} = diag(\alpha_{\gamma 1}, \alpha_{\gamma 2})$. To complete the specification of the priors, the variance hyper parameters can be assigned to non-informative gamma priors that are widely-used in hierarchical models but it is well known in the literature that gamma priors are inappropriate choice for variance components in random effects

models. Refer to Gelman (2006) for a discussion of this problem and description of alternatives.

7.3 Parameter estimates for functional mixed effects model

7.3.1 Posterior estimates

Inference in Bayesian data analysis is based on the posterior distribution of the parameters. We consider the mixed effects model in equation (7.2) or (7.3) and take $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\upsilon}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}\}$, where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_N)$ and $\boldsymbol{\upsilon} = (\boldsymbol{\upsilon}_i, \cdots, \boldsymbol{\upsilon}_N)^{'}$ such that $\boldsymbol{\upsilon}_i = (\upsilon_{i1}, \cdots, \upsilon_{iM^*})^{'}$ for all i subjects. Vectors $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_M)^{'}$ and $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_{M^*})^{'}$ and are the diagonal elements for the covariance matrix $\boldsymbol{A}$ and $\boldsymbol{\Omega}$ respectively. The priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\upsilon}$ are as described before. The priors for parameters $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$, $\boldsymbol{\alpha}_\gamma$ and σ_ϵ^{-2} are $\alpha_l|c_1, d_1 \sim Gamma(c_1, d_1)$, $\omega_l|c_2, d_2 \sim Gamma(c_2, d_2)$, $\alpha_{\gamma l}|c_3, d_3 \sim Gamma(c_3, d_3)$ and $\sigma_\epsilon^{-2}|a, b \sim Gamma(a, b)$ respectively.

Taking a response vector $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N\}$, the joint posterior distribution for the parameters in the mixed effects model is

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\upsilon}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}) p(\boldsymbol{\beta}|\boldsymbol{\alpha}) p(\boldsymbol{\upsilon}|\boldsymbol{\omega}) p(\boldsymbol{\gamma}|\boldsymbol{\alpha}_\gamma) p(\boldsymbol{\alpha}) p(\boldsymbol{\omega}) p(\boldsymbol{\alpha}_\gamma) p(\sigma_\epsilon^{-2})}{p(\boldsymbol{Y}|.)}$$

where $p(\boldsymbol{Y}|.)$ is the normalizing constant. However, the posterior density $p(\boldsymbol{\Theta}|\boldsymbol{Y})$ is analytically intractable since the normalizing constant does not have a closed form solution. To approximate $p(\boldsymbol{\Theta}|\boldsymbol{Y})$, we use the decomposition,

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) = p(\boldsymbol{\beta}|.) p(\boldsymbol{\gamma}|.) p(\boldsymbol{\upsilon}|.) p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}|\boldsymbol{Y}). \tag{7.4}$$

where the density functions $p(\boldsymbol{\beta}|.) = p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2})$, $p(\boldsymbol{\gamma}|.) = p(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2})$ and $p(\boldsymbol{\upsilon}|.) = p(\boldsymbol{\upsilon}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$. These are the posterior densities for the fixed and random effects at subject and cycle level respectively.

The posterior distributions for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\upsilon}$ are Gaussian densities. The posterior density for $\boldsymbol{\beta}$ is,

$$p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}) = \mathbf{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{7.5}$$

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}(\sum_{i=1}^N \boldsymbol{X}_i^{'} \boldsymbol{V}_i^{*-1} \boldsymbol{y}_i)$ and $\boldsymbol{\Sigma} = (\boldsymbol{A} + \sum_{i=1}^N \boldsymbol{X}_i^{'} \boldsymbol{V}_i^{*-1} \boldsymbol{X}_i)^{-1}$ such that matrix $\boldsymbol{V}_i^* = \sigma_\epsilon^2 \boldsymbol{I}_{T_i} + \boldsymbol{W}_i^* \boldsymbol{D}^{*-1} \boldsymbol{W}_i^{*'}$ and $\boldsymbol{I}_{T_i}$ is a $T_i \times T_i$ identity matrix. Matix $\boldsymbol{A}$ is a diagonal matrix with elements in vector $\boldsymbol{\alpha}$. The posterior distribution for the subject specific

random effects $\boldsymbol{\gamma}$ is,

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}) = \prod_{i=1}^{N} N(\boldsymbol{\gamma}_i; \hat{\boldsymbol{\gamma}}_i, \hat{\boldsymbol{\Psi}}_i), \tag{7.6}$$

where $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\Psi}}_i \widetilde{\boldsymbol{W}}_i^{'} \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ and $\hat{\boldsymbol{\Psi}}_i = (\boldsymbol{E} + \widetilde{\boldsymbol{W}}_i^{'} \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{W}}_i)^{-1}$ such that $\boldsymbol{V}_i = \sigma_\epsilon^2 \boldsymbol{I}_{T_i} + \boldsymbol{W}_i \boldsymbol{D}^{-1} \boldsymbol{W}_i^{'}$ and $\boldsymbol{E} = diag(\alpha_{\gamma 1}, \alpha_{\gamma 2})$. The dimensions of the diagonal matrix $\boldsymbol{D} = diag(\boldsymbol{\Omega}, \cdots, \boldsymbol{\Omega})$ may differ from subject to subject depending on the number of cycles in a subject and matrix $\boldsymbol{W}_i = diag(\boldsymbol{W}_{i1}, \cdots, \boldsymbol{W}_{in_i})$. The posterior distribution for the cycle specific random effects $\boldsymbol{\upsilon}_{ij}$ is

$$p(\boldsymbol{\upsilon}|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} N(\boldsymbol{\upsilon}_{ij}; \hat{\boldsymbol{\upsilon}}_{ij}, \hat{\boldsymbol{\Omega}}_{ij}), \tag{7.7}$$

where the posterior mean $\hat{\boldsymbol{\upsilon}}_{ij} = \sigma_\epsilon^{-2} \hat{\boldsymbol{\Omega}}_{ij} \boldsymbol{W}_{ij}^{'}(\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \widetilde{\boldsymbol{W}}_{ij}\boldsymbol{\gamma}_i)$ and posterior covariance $\hat{\boldsymbol{\Omega}}_{ij} = (\boldsymbol{\Omega} + \sigma_\epsilon^{-2} \boldsymbol{W}_{ij}^{'} \boldsymbol{W}_{ij})^{-1}$.

The posterior $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ for the variance components do not have a simple form. Therefore, to estimate the variance components we propose an empirical Bayes procedure for sparse MAP estimation that will be discussed in the next subsection.

7.3.2 Empirical Bayes for variance components

The posterior density $p(\boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ is analytically intractable, and we propose to use an empirical Bayes approach and compute plug-in estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_\gamma$, $\boldsymbol{\omega}$ and σ_ϵ^2. The estimates are carefully computed to favor a sparse shrinkage structure where many elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ are very close to zero. The empirical Bayes estimates are based on the modes of the generalized log-likelihood functions $l(\boldsymbol{\alpha})$, $l(\boldsymbol{\alpha}_\gamma)$ and $l(\boldsymbol{\omega}, \sigma_\epsilon^{-2})$ that will be defined latter in this section.

Let the joint density $p(\boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\omega}, \sigma_\epsilon^{-2}|\boldsymbol{Y}) \propto p(\boldsymbol{\alpha})p(\boldsymbol{\alpha}_\gamma)p(\boldsymbol{\omega})p(\sigma_\epsilon^{-2})p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$ where the density functions $p(\boldsymbol{\alpha})$, $p(\boldsymbol{\alpha}_\gamma)$, $p(\boldsymbol{\omega})$ and $p(\sigma_\epsilon^{-2})$ are the Gamma distributions that were defined in section 7.2.1. Following an empirical Bayes approach and choosing non-informative priors for $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$, $\boldsymbol{\alpha}_\gamma$ and σ_ϵ^2 , we set all the gamma hyper-parameters equal to zero, leading to the assumption that the modes for $p(\boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\omega}, \sigma_\epsilon^{-2}|\boldsymbol{Y})$ and $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$ are equivalent. Hence, the plug-in estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_\gamma$, $\boldsymbol{\omega}$ and σ_ϵ^2 can be computed as the maximum likelihood estimates (MLE) of the likelihood function $p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$.

The likelihood function is obtained after integrating out the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{\gamma}_i$ and $\boldsymbol{\upsilon}_{ij}$ such that,

$$p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2}) = \int \prod_{i=1}^{N}\prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij}|\boldsymbol{\beta},\boldsymbol{\upsilon}_{ij},\boldsymbol{\gamma}_i,\sigma_\epsilon^{-2}) p(\boldsymbol{\beta}|\boldsymbol{\alpha}) p(\boldsymbol{\gamma}_i|\boldsymbol{\alpha}_\gamma) p(\boldsymbol{\upsilon}_{ij}|\boldsymbol{\omega}) d\boldsymbol{\upsilon}_{ij} d\boldsymbol{\gamma}_i d\boldsymbol{\beta},$$

where the densities $p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \prod_{l=1}^{M} N(\beta_l; 0, \alpha_l^{-1})$ and $p(\boldsymbol{\upsilon}_{ij}|\boldsymbol{\omega}) = \prod_{l=1}^{M^*} N(\upsilon_{ijl}; 0, \omega_l^{-1})$, while $p(\boldsymbol{\gamma}_i|\boldsymbol{\alpha}_\gamma)$ is as defined in the previous section. The likelihood $p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2}) = N(\boldsymbol{Y}; \boldsymbol{0}, \boldsymbol{C}^{*-1})$, where the covariance matrix $\boldsymbol{C}^* = \boldsymbol{V}^* + \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}'$ such that $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)'$ is the design matrix and $\boldsymbol{V}^* = diag(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_N^*)$.

The MLE estimates for the four variance components cannot be computed simultaneously using the likelihood function $p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2})$. Therefore, we use an alternating conditional maximization process that iterates between calculating the conditional MLE of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_\gamma$ and $\boldsymbol{\omega}, \sigma_\epsilon^{-2}$ seperately. To compute $\boldsymbol{\alpha}$, we hold parameters $\boldsymbol{\omega}$, $\boldsymbol{\alpha}_\gamma$ and σ_ϵ^{-2} fixed and use the likelihood function $L(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2})$. Similarly, to compute the conditional MLE of $\boldsymbol{\alpha}_\gamma$ we fix the remaining variance parameters and use the conditional likelihood $L(\boldsymbol{\alpha}_\gamma; \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$. Computation of the conditional MLE for both $\boldsymbol{\omega}$ and σ_ϵ^{-2} is based on the likelihood $L(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$ while keeping parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_\gamma$ fixed. The three conditional likelihood functions are defined as follows.

The computation of the MLE estimates for $\boldsymbol{\alpha}$ involves maximizing the log-likelihood function $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}) = \frac{-1}{2}\{N\log(2\pi) + \log|\boldsymbol{C}^*| + \boldsymbol{Y}'\boldsymbol{C}^{*-1}\boldsymbol{Y}\}$. However, the presence of the covariance matrix $\boldsymbol{C}^*$ in the log-likelihood function makes it impossible to compute the estimates for $\boldsymbol{\alpha}$. Hence, following a similar approach as in Ji et al. (2009) and taking fixed values of $\boldsymbol{\omega}$, $\boldsymbol{\alpha}_\gamma$ and σ_ϵ^{-2}, the log-likelihood function $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2})$ is re-written in a decomposed representation

$$\begin{aligned}\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2}) = \frac{-1}{2}\Big\{ & N\log(2\pi) + \log|\hat{\boldsymbol{\Sigma}}^{-1}| + \log|\boldsymbol{A}| + \hat{\boldsymbol{\mu}}'\boldsymbol{A}\hat{\boldsymbol{\mu}} \\ & + \log|\boldsymbol{V}^{*-1}| + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\mu}})'\boldsymbol{V}^{*-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\mu}})\,\}. \end{aligned} \tag{7.8}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the posterior mean and covariance matrix defined in equation (7.5). The estimate for the l^{th} element of $\boldsymbol{\alpha}$ is,

$$\hat{\alpha}_l = \frac{1}{\hat{\mu}_l^2 + \hat{\Sigma}_{ll}} \qquad l = 1, \ldots, M, \tag{7.9}$$

where $\hat{\mu}_l$ is the l^{th} elements of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}_{ll}$ is the l^{th} diagonal element of $\hat{\boldsymbol{\Sigma}}$. The estimates for $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_M)'$ are functions of both the mean and covariance of $\boldsymbol{\beta}$. Hence the computation of $\hat{\alpha}_l$ involves an iterative procedure that estimates variance hyper-parameter α_l in equation (7.8) and updating the mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ as in equation

(7.5).

We maximize the log-likelihood function $\ell(\boldsymbol{\alpha}_\gamma; \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2})$ to obtain the estimates for $\boldsymbol{\alpha}_\gamma$. Let $\ell(\boldsymbol{\alpha}_\gamma; \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) = -1/2 \sum_{i=1}^{N} T_i \log(2\pi) + \log|\boldsymbol{C}_i^{-1}| + (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})' \boldsymbol{C}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}})$ be a log-likelihood function for $\boldsymbol{\alpha}_\gamma$ given $\boldsymbol{\omega}$ and σ_ϵ^{-2} and $\boldsymbol{\alpha}$ where $\boldsymbol{C}_i = \boldsymbol{V}_i + \alpha_\gamma^{-1}\widetilde{\boldsymbol{W}}_i\widetilde{\boldsymbol{W}}_i'$ where $\widetilde{\boldsymbol{W}}_i$ and $\boldsymbol{V}_i$ are as were defined previously. Like in the previous case, the presence of matrix $\boldsymbol{V}_i$ in the log-likelihood function causes problem in the computation of the MLE for $\boldsymbol{\alpha}_\gamma$, thus we follow a similar decomposition process that leads to

$$\ell(\boldsymbol{\alpha}_\gamma; \boldsymbol{\alpha}, \boldsymbol{\omega}, \sigma_\epsilon^{-2}) = \frac{-1}{2}\sum_{i=1}^{N} \log|\hat{\boldsymbol{\Psi}}_i^{-1}| + \log|\boldsymbol{\alpha}_\gamma| + \hat{\gamma}_i' \boldsymbol{E}\hat{\gamma}_i + \log|\boldsymbol{V}_i^{-1}| \\ +(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_i\hat{\gamma}_i)' \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_i\hat{\gamma}_i), \tag{7.10}$$

where $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\boldsymbol{\Psi}}_i$ are the posterior mean and variance for $\boldsymbol{\gamma}_i$ as in equation (7.6). The estimate for $\boldsymbol{\alpha}_\gamma$ is

$$\hat{\alpha}_{\gamma l} = \frac{N}{\sum_{i=1}^{N}(\hat{\gamma}_{il}^2 + \hat{\Psi}_{ill})} \qquad l = 1, 2. \tag{7.11}$$

The estimates for $\boldsymbol{\alpha}_\gamma$ are functions of both the mean $\hat{\boldsymbol{\gamma}}_i$ and variance $\hat{\boldsymbol{\Psi}}_i$ for the subject specific random effects. Therefore, the computation of $\hat{\boldsymbol{\alpha}}_\gamma$ involves an iterative procedure that estimates $\boldsymbol{\alpha}_\gamma$ in equation (7.10) and updating the estimates $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\boldsymbol{\Psi}}_i$ as in equation (7.6).

To obtain the estimates for $\boldsymbol{\omega}$ and σ_ϵ^{-2}, we maximize the log-likelihood $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma) = -1/2 \sum_{i=1}^{N}\{\sum_{j=1}^{n_i} T_{ij} \log(2\pi) + \log|\boldsymbol{V}_{ij}^{-1}| + (\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\boldsymbol{\gamma}}_i)' \boldsymbol{V}_{ij}^{-1}(\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\boldsymbol{\gamma}}_i)\}$ where $\boldsymbol{V}_{ij} = \sigma_\epsilon^2 \boldsymbol{I}_{T_{ij}} + \boldsymbol{W}_{ij}\boldsymbol{\Omega}^{-1}\boldsymbol{W}_{ij}'$. This is a log-likelihood of $\boldsymbol{\omega}$ and σ_ϵ^{-2} given some fixed values of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_\gamma$. Like in the previous two cases, the presence of matrix $\boldsymbol{V}_{ij}$ in the log-likelihood function causes problem in computing the estimates for $\boldsymbol{\omega}$ and σ_ϵ^{-2}. By decomposing the log-likelihood $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$ we obtain,

$$\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma) = \frac{-1}{2}\sum_{i=1}^{N}\{\sum_{j=1}^{n_i} \log|\hat{\boldsymbol{\Omega}}_{ij}^{-1}| + log|\boldsymbol{\Omega}| + \log|\sigma_\epsilon^2 \boldsymbol{I}_{T_{ij}}| + \hat{\boldsymbol{\upsilon}}_{ij}'\boldsymbol{\Omega}\hat{\boldsymbol{\upsilon}}_{ij} \\ -\sigma_\epsilon^2||\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\boldsymbol{\gamma}}_i - \boldsymbol{W}_{ij}\hat{\boldsymbol{\upsilon}}_{ij}||^2\}, \tag{7.12}$$

where $\hat{\boldsymbol{\upsilon}}_{ij}$ and $\hat{\boldsymbol{\Omega}}_{ij}$ are the posterior mean and covariance for the cycle specific random effects $\boldsymbol{\upsilon}_{ij}$. The MLE estimates are,

$$\hat{\omega}_l = \frac{\sum_{i=1}^{N} n_i}{\sum_{i=1}^{N}(\sum_{j=1}^{n_i}(\hat{\upsilon}_{ijl}^2 + \hat{\Omega}_{ijll}))}, \qquad l = 1, \ldots M^* \tag{7.13}$$

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{n_i} ||\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\gamma}_i - \boldsymbol{W}_{ij}\hat{\boldsymbol{v}}_{ij}||^2)}{\sum_{i=1}^{N}\{\sum_{j=1}^{n_i}(T_{ij} - M^* + \sum_{l=1}^{M^*}\omega_l\hat{\Omega}_{ijll})\}}, \tag{7.14}$$

where $\hat{v}_{ijl}$ is the l^{th} component of $\hat{\boldsymbol{v}}_{ij}$ and $\hat{\Omega}_{ijll}$ is the l^{th} diagonal element of $\hat{\boldsymbol{\Omega}}_{ij}$. Both estimates for $\boldsymbol{\omega}$ and σ_ϵ^2 are functions of $\hat{\boldsymbol{v}}_{ij}$ and $\hat{\boldsymbol{\Omega}}_{ij}$, which leads to an iterative algorithm that alternates between updating $\hat{\omega}_{ij}$ and $\hat{\sigma}_\epsilon^2$ as in equations (7.13)-(7.14) and updating the mean and covariance of the random effects as in equation (7.7).

While implementing the above empirical Bayes approach, two problems can arise when the number of the fixed and random effects is large. The first problem is related to the computational time. When $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Omega}}_{ij}$ have large dimensions ($M \times M$ and $M^* \times M^*$ respectively), the computational time increases dramatically while inverting $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Omega}}_{ij}$ at each step of the iterative procedure. Next, when dimensions M and/or M^* are large relative to the sample size, the computational efficiency worsens while estimating the elements of the covariance matrices. Such a problem can lead to lack of convergence of the procedure.

Potentially the computational efficiency problem can be solved by adapting a fast basis selection RVM algorithm that bypasses the inversion step leading to a reduced model with dimension $m \times m$ and $m^* \times m^*$ for both $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Omega}}_{ij}$ respectively where $m << M$ and $m^* << M^*$. However, the basis selection step typically adds to the computational burden, but this is still important to obtain a sparse yet flexible characterization of the data. Moreover, the RVM approach has another disadvantage in that typical Bayesian basis selection approach also allows both inferences on which variables are important and allows a better characterization of uncertainty. Unfortunately, the RVM approach does not have this latter advantage in ignoring uncertainty in the variable selection process. Hence, this is the price we pay for much greater computational efficiency relative to the MCMC approaches.

7.3.3 A fast Empirical Bayes approach

A fast Empirical Bayes procedure can be use to improve the computation efficiency while computing the prior parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$. Empirically, the local marginal maximization of the marginal log-likelihoods (7.10) and (7.12) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ respectively, can lead to a sparse model with reduced dimensions of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ by discarding $M-m$ columns of $\boldsymbol{X}_i$ and $M^* - m^*$ columns of $\boldsymbol{W}_i$. Conditional maximization of the two log-likelihood functions $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2})$ and $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$ can lead to highly sparse distributions with values of many hyper-parameters approach infinity. This allows the distribution for the posterior estimates to be infinitely peaked at zero for many fixed

and random effects with corresponding consequences of having few non-zero elements of posterior estimates $\boldsymbol{\mu}$ and $\hat{\boldsymbol{v}}_{ij}$ for all i and j.

To select the non zero elements of the population mean vector $\boldsymbol{\mu}$, we first maximize the log-likelihood function $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2}) = \frac{-1}{2}\left(N\log(2\pi)+\log|\boldsymbol{C}^*|+\boldsymbol{Y}^{'}\boldsymbol{C}^{*-1}\boldsymbol{Y}\right)$ with respect to $\boldsymbol{\alpha}$. To achieve optimal shrinkage, the maximization process is based on the k^{th} element of the log-likelihood function. For a better representation, the log-likelihood function is decomposed into two parts -with and without the k^{th} component of $\boldsymbol{\alpha}$. This is achieved by first partitioning the covariance matrix $\boldsymbol{C}^*$ such that

$$\boldsymbol{C}^* = \boldsymbol{V}^* + \sum_{l=1}^{M}\alpha_l^{-1}X_{.l}X_{.l}^{'} = \boldsymbol{V}^* + \sum_{l\neq k}^{M}\alpha_l^{-1}X_{.l}X_{.l}^{'} + \alpha_k^{-1}X_{.k}X_{.k}^{'},$$

where $X_{.l}$ is the l^{th} columns of the design matrix $\boldsymbol{X}$. Hence, the covariance matrix $\boldsymbol{C}^* = \boldsymbol{C}^*_{-k} + \alpha_k^{-1}X_{.k}X_{.k}^{'}$ such that $\boldsymbol{C}^*_{-k}$ is matrix $\boldsymbol{C}^*$ without the contribution of the k^{th} component of $\boldsymbol{X}$ and $\boldsymbol{\alpha}$. This partitioning process allows the decomposition of the log likelihood function leading to

$$\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2}) = \ell(\boldsymbol{\alpha}_{-k};\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2}) + \frac{1}{2}\left(\log\alpha_k - \log|\alpha_k+s_k| + \frac{q_k^2}{\alpha_k+s_k}\right),$$

where $s_k = X_{.k}^{'}\boldsymbol{C}_{-k}^{*-1}X_{.k}$ and $q_k = X_{.k}^{'}\boldsymbol{C}_{-k}^{*-1}\boldsymbol{Y}$. Differentiating the log-likelihood function $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2})$ with respect to α_k and equating the resulting normal equations to zero leads to the MLE estimate for $\alpha_k = \frac{s_k^2}{q_k^2-s_k}$. This results to a recursive RVM algorithm for the basis function selection that has

$$\hat{\alpha}_k = \begin{cases} \frac{s_k^2}{q_k^2-s_k} & \textit{if } q_k^2 > s_k, \\ \infty & \textit{otherwise.} \end{cases} \tag{7.15}$$

The basis selection process for a candidate $X_{.k}$ is based on the contribution of α_k to the log-likelihood $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2})$. To initiate the selection process, we first select the fixed effect that has the largest contribution to $\ell(\boldsymbol{\alpha};\boldsymbol{\omega},\boldsymbol{\alpha}_\gamma,\sigma_\epsilon^{-2})$. After the initial selection, we evaluate the contribution of all α_k to the log-likelihood function and select the one with the next largest contribution. Using this candidate basis function we compute the values of q_k and s_k which determine whether the candidate basis function is to be added, removed or update. Addition occurs when $q_k^2 > s_k$ and $X_{.k}$ is not in the model. But when the basis function $X_{.k}$ is already in the model, we can either update or delete the basis function. We update $\hat{\alpha}_k$ when $q_k^2 > s_k$. But if the quantity $q_k^2 < s_k$, we delete/remove $X_{.k}$ from the model. The process continues until the contribution of the candidate α_k to the log-likelihood function is negligible.

In a similar manner, a modified RVM procedure can hasten the process of dimension reduction of the random effects at cycle level from M^* to m^*. This process is based on a basis selection procedure that estimates the significant components of $\boldsymbol{\omega}$ by maximizing the log-likelihood function $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$. In particular, we fix the remaining parameter constant and evaluate the contribution of the k^{th} component of $\boldsymbol{\omega}$ on the log-likelihood function. We decompose the log-likelihood function $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma) = -1/2 \sum_{i=1}^{N} \{\sum_{j=1}^{n_i} T_{ij} \log(2\pi) + \log|\boldsymbol{V}_{ij}^{-1}| + (\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\boldsymbol{\gamma}}_i)' \boldsymbol{V}_{ij}^{-1} (\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\boldsymbol{\gamma}}_i)\}$ into two components - with and without the contribution of the k^{th} components of hyper-parameter $\boldsymbol{\omega}$. To allow such kind of decomposition, we first partitioning the covariance matrix $\boldsymbol{V}_{ij}$ into two parts - with and without the contribution of the k^{th} components. This leads to

$$\boldsymbol{V}_{ij} = \sigma_\epsilon^2 \boldsymbol{I}_{T_{ij}} + \sum_{l=1}^{M^*} \omega_l^{-1} W_{ijl} W_{ijl}^{'} = \sigma_\epsilon^2 \boldsymbol{I}_{T_{ij}} + \sum_{l \neq k}^{M^*} \omega_l^{-1} W_{ijl} W_{ijl}^{'} + \omega_k^{-1} W_{ijk} W_{ijk}^{'}.$$

where $\boldsymbol{V}_{ij} = \boldsymbol{V}_{ij-k} + \omega_k^{-1} W_{ijk} W_{ijk}^{'}$ such that $\boldsymbol{V}_{ij-k} = \sigma_\epsilon^2 \boldsymbol{I}_{T_{ij}} + \sum_{l \neq k}^{M^*} \omega_l^{-1} W_{ijl} W_{ijl}^{'}$. Decomposing the log likelihood function results to,

$$\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma) = \ell(\boldsymbol{\omega}_{-k}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma) + \frac{-1}{2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\log \omega_k - \log|\omega_k + s_{ijk}^*| + \frac{q_{ijk}^{*2}}{\omega_k + s_{ijk}^*} \right),$$

where $s_{ijk}^* = W_{ijk}^{'} \boldsymbol{V}_{ij-k}^{-1} W_{ijk}$, $q_{ijk}^* = W_{ijk}^{'} \boldsymbol{V}_{ij-k}^{-1} (\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\hat{\boldsymbol{\mu}} - \widetilde{\boldsymbol{W}}_{ij}\hat{\boldsymbol{\gamma}}_i)$ and $\ell(\boldsymbol{\omega}_{-k}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$ is the log-likelihood without the contribution of ω_k. We differentiate the log-likelihood function with respect to ω_k and set the result to zero. Since the resulting expression is complex to express, we approximate solution by assuming that $\omega_k << s_{ijk}^*$ resulting to $\hat{\omega}_k = \frac{N}{\sum_{l=1}^{N} \sum_{j=1}^{n_i} (q_{ijk}^{*2} - s_{ijk}^*)/s_{ijk}^{*2}}$. More details to justify the use of this approach can be found in Ji et al. (2009). This leads to a recursive algorithm that has,

$$\hat{\omega}_k \cong \begin{cases} \frac{N}{\sum_{i=1}^{N} \sum_{j=1}^{n_i} (q_{ijk}^{*2} - s_{ijk}^*)/s_{ijk}^{*2}} & \text{if } \sum_{l=1}^{N} \sum_{j=1}^{n_i} \frac{(q_{ijk}^{*2} - s_{ijk}^*)}{s_{ijk}^{*2}} > 0, \\ \infty & \text{otherwise.} \end{cases} \tag{7.16}$$

Like in the previous basis functions selection, the random effects at the cycle level are selected in a parallel manner similar to the fixed effects. We first select the k^{th} element of $\boldsymbol{\omega}$ that has the largest contribution to the log-likelihood function $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$. This is followed by the computation of q_{ijk}^* and s_{ijk}^* that are used to determine the appropriate operation to be done on the candidate basis function. We then evaluate the contribution of each ω_k towards the log-likelihood function and select the one with the next largest contribution. This becomes the candidate basis function W_{ijk}. Three operations can take place on W_{ijk}: add, update and delete. The basis function W_{ijk} is added into the

model when it is absent and $\sum_{l=1}^{N}\sum_{j=1}^{n_i}\frac{(q_{ijk}^{*2}-s_{ijk}^{*})}{s_{ijk}^{*2}} > 0$. An update operation occurs when $\sum_{l=1}^{N}\sum_{j=1}^{n_i}\frac{(q_{ijk}^{*2}-s_{ijk}^{*})}{s_{ijk}^{*2}} > 0$ and W_{ik} is already in the model. Deletion occurs when $\sum_{l=1}^{N}\sum_{j=1}^{n_i}\frac{(q_{ijk}^{*2}-s_{ijk}^{*})}{s_{ijk}^{*2}} < 0$ and the basis function W_{ijk} is already in the model. The basis functions selection process continues until the contribution of the candidate ω_k to the log-likelihood function is negligible. The final model has a few $\omega_l < \infty$ while majority tend to approach ∞. This corresponds to majority of the random effects at cycle level equal to zero and while just a few are non-zero for all subjects i and cycles j.

Appendix C contains simpler expressions for the quantities of interest that are used to compute: (1) changes in log likelihood functions in the three basis function operations (addition, deletion and update); (2) parameter estimates for the priors ($\boldsymbol{\alpha}$, $\boldsymbol{\omega}$), mean vectors ($\boldsymbol{\mu}$, $\boldsymbol{\upsilon}$) and covariance matrices ($\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$) as in equations (7.5, 7.7, 7.15, 7.16). These expressions help to minimize the computation burden involved in the estimation of the fixed and random effects for the sparse functional mixed model discussed in section 7.3.

7.4 Application to the bbt measurements

In this application, we consider the basal body temperature data from the European fecundability study (Colombo and Masarotto, 2000) that was used in the previous applications in this thesis. In previous chapters we considered measurements for one menstrual cycle from each woman, but in this case we generalize the application and consider a case where subjects contribute measurements from more than one menstrual cycle. However, since most of the 520 subjects used in the previous chapters have bbt measurements from a few menstrual cycles, we consider $N = 100$ subjects. Further, to make the programming work easier, we consider a fixed number of cycles n_i from each woman such that each contributes bbt measurements from $n_i = 5$ menstrual cycles.

To implement the proposed multi-level RVM procedure, we use the cubic B-splines (Ramsay, 2005) to generate the basis functions $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ for the fixed and random effects respectively. Following a similar approach like the one used by Wand and Ormerod (2008), the basis functions are generated using the standardized values of time covariate $\boldsymbol{z}_{ij} = (z_{ijt}, \cdots, z_{ijT_{ij}})'$. We generate 27 cubic B-splines with 23 interior equally spaced knots. This results to 27 column in the design matrices $\boldsymbol{X}_{ij}$ and $\boldsymbol{W}_{ij}$ for the fixed effects and cycle specific random effects respectively. We also include two additional columns generated from $1's$ and $\boldsymbol{z}_{ij}$ (i.e. $\{1, z_{ijt}\}_{t=1}^{T_{ij}}$). Hence, the two design matrices $\boldsymbol{X}_{ij}$ and $\boldsymbol{W}_{ij}$ have dimensions $T_{ij} \times M$ and $T_{ij} \times M^*$ respectively where $M = M^* = 29$ are the number of columns for $\boldsymbol{X}_{ij}$ and $\boldsymbol{W}_{ij}$ and T_{ij} is number of rows which vary from cycle to

cycle depending upon the number of observations in each menstrual cycle. The design matrix $\widetilde{\boldsymbol{W}}_{ij}$ for the subject specific random effects consists of 2 columns generated from $1's$ and $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \cdots, \boldsymbol{z}_{in_i})'$ leading to dimensions $T_i \times 2$, where T_i is the number of rows that correspond to the total number of the bbt measurements for all menstrual cycles n_i from the i^{th} subject.

The coding for the ML-RVM procedure was done using R software (version 2.8.1) on Pentium IV, 2.4GHz, 512MB, Windows XP computer. We implemented the MT-RVM procedure to the bbt data for $N = 100$ women. The final sparse Multi-level functional model has $m = 9$ basis functions for the fixed effects ($\# = 28, 26, 2, 4, 22, 29, 17, 27, 24$), two basis functions for the subject specific random effects and $m^* = 19$ basis functions for the cycle specific random effects ($\# = 17, 6, 7, 11, 23, 9, 15, 16, 19, 25, 21, 10, 4, 3, 2, 1, 8, 28, 29$). The indexes for the basis functions for the fixed and random effects in the final multi-level model, represent the order in which the basis functions were entered into the model at each level. The average time taken to estimate one bbt curve from the j^{th} cycle and i^{th} subject was 0.0742 seconds. As in other multi-level modeling cases, curves from the same subject are expected to cluster together. Hence, to reveal the clustering effects within the subjects' curves we plot the n_i curves from the same subject on the same plot. Figure 7.1 shows the estimated bbt cycles from 6 randomly selected women where each woman has five bbt curves. The thick line in all the six plots represent the estimated population average curves, while the five dotted curves in each plot represent the estimated cycle specific curves.

7.5 Discussion

In this chapter we have a developed a fast Bayesian method that generates a sparse multilevel functional mixed model. The methodology uses nested data to fit a multilevel functional mixed model similar to Brumback and Rice (1998). To allow for fast computation, we developed a dimensional reduction strategy for the multilevel model similar to James, et al., 2001, Yao, et al., 2005; Crainiceanu, 2009 but avoided their Functional Principal Component Analysis (FPCA) methodology and used the RVM (Tipping, 2001) procedure. Our approach establishes a close connection between the relevance machine methodology that is widely used in machine learning and the multi-level functional mixed models that are commonly used in biostatistics data analysis. It adds to the literature on dimensional reduction/variables selection procedures for the multi-level linear mixed model as well as providing an appealing alternative to the commonly used computer intensive reversible jump (Green, 1995) and stochastic search variable selection (Smith and Kohn, 1996) methods. The proposed ML-RVM approach provides a fast Bayesian

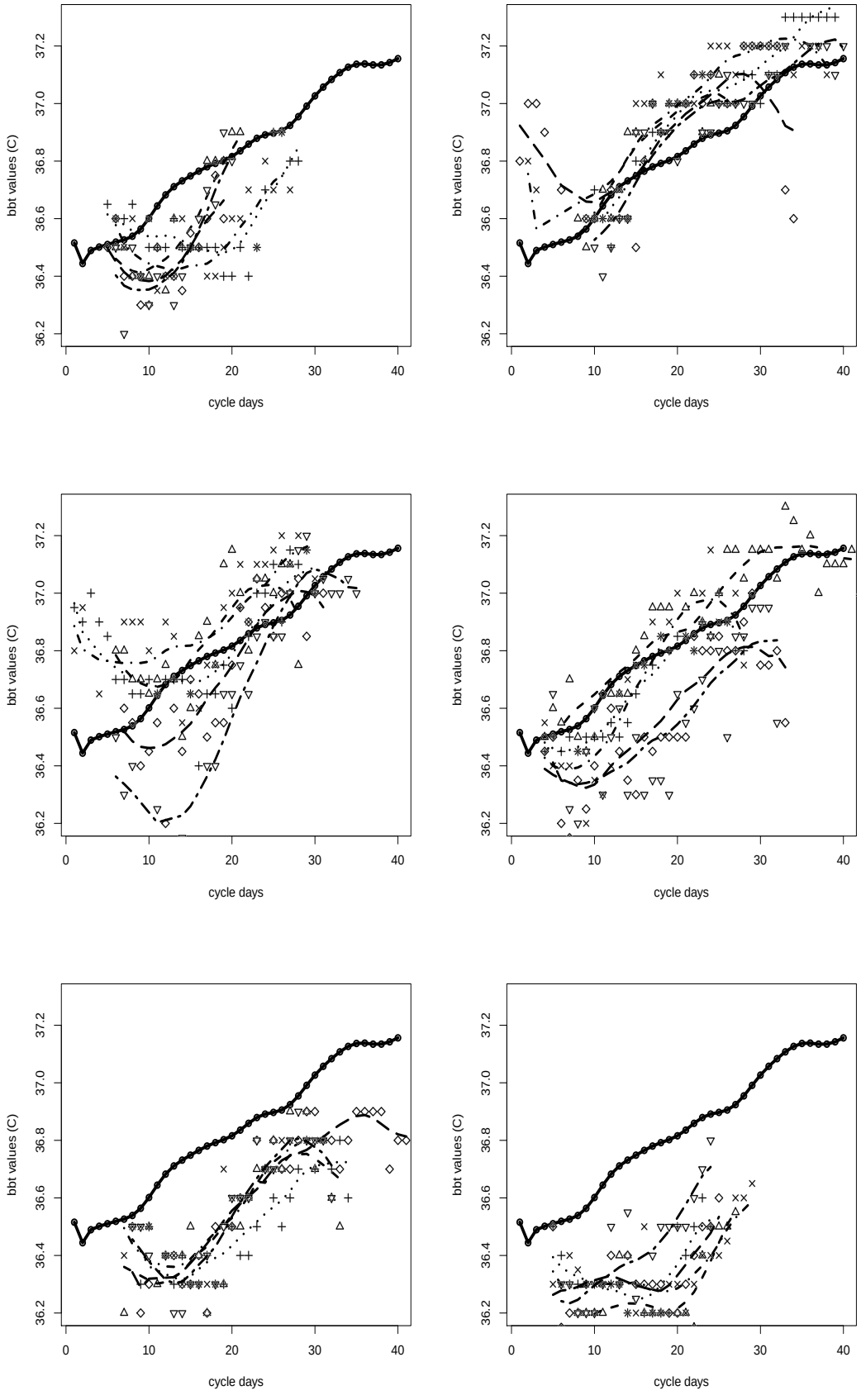

FIGURE 7.1: Plots for population average (thick curve) and 5 cycles specific curves (dotted) for each subject

method that combines the basis functions selection at both population and cycle specific levels and the parameter estimation process within a single step. The selection of the variables (fixed and random) is done at two levels (population average and cycle-specific level), but the number of the random effects at subject specific is fixed in that we only considered the intercept and the slope.

The proposed approach is appealing in developing a more flexible multi-level Bayesian methodology that bypasses the use of the MCMC algorithms and use the MAP plug-in estimates for the priors in both the fixed and random effects parameters. However, the approach has several limitations. First, to allow the implementation of the RVM procedure in a multi-level functional linear mixed model, we assumed (1) the daily bbt measurements are independent and (2) the random effects are also independent. The independence assumption among the bbt measurements and the random effects is an unrealistic assumption in many longitudinal and functional data analysis applications. Hence, a flexible approach should be adopted where the correlations among the measurements and the random effects should be incorporated into the model. To handle a more general correlation structure for the random effects, we can adopt and extend the approach used in Ciera nd Dunson (2010) that is based on a modified Cholesky decomposition that was proposed by Chen and Dunson (2003) and allows the decomposition of the covariance matrix for the random effects. Other candidate covariance matrix decomposition methodologies include Smith and Kohn (2002), Fruhwirth-Schnatter and Tuchler (2004) and Kinney and Dunson (2006). Second, we can considered a more generalized basis selection procedure that can be applied to the basis functions at all levels in the hierarchy (i.e. at the fixed effects and random effects at subject and cycle specific levels).

Despite of the cited limitations, the proposed approach offers a fast and flexible method to estimate the average population and subject specific curves in addition to the cycle specific curves. Moreover, the method is appealing in laying the foundation for developing fast multi-level Bayesian methodologies that bypasses the use of the MCMC algorithms and use the MAP plug-in estimates for the priors in both the fixed and random effects parameters. Since most theoretical properties of the application of the RVM procedure in functional mixed models and multilevel functional models are not yet known, there is need to investigate: (1) the sensitivity of the MAP estimates toward the initial values supplied to initiate the ML-RVM algorithm and (2) the performance of the approach while maximizing the two conditional log likelihood functions $\ell(\boldsymbol{\alpha}; \boldsymbol{\omega}, \boldsymbol{\alpha}_\gamma, \sigma_\epsilon^{-2})$ and $\ell(\boldsymbol{\omega}, \sigma_\epsilon^{-2}; \boldsymbol{\alpha}, \boldsymbol{\alpha}_\gamma)$ since there is a potential risk of arriving at local modes. Moreover, before adopting the proposed approach as a universal dimensional reduction and multilevel curve fitting method, further work is required to compare the performance

of the ML-RVM method relative to other related approaches in the literature. For example, on dimension reduction aspect, simulations studies need to be done to evaluate the performance of the ML-RVM procedure and compare it with other methods like the FPCA method (James, et al., 2001) and the fence method (Jiang et al, 2008) that are commonly used for model reduction in linear mixed model. Similarly, on curve fitting aspect, crucial properties like computational time, bias and Mean Integrated Squred Error (MISE) of the ML-RVM method need to be compared with other related methods in the literature e.g. Brumback and Rice (1998); Guo (2002); Morris et al. (2003); Durban, et al. (2005); Baladandayuthapani et al. (2008); Crainiceanu (2009) among others. We also note that, it would be great to find ways to simplify the RVM algorithm which many readers complain to be relatively complicated compared to other existing basis selection procedures.

7.6 Further work

Initial results on the application of the ML-RVM method in section 7.3 show that the approach has great potential in providing a fast method of fitting multilevel curves in Bayesian framework. Hence, to justify the application of the proposed method in other related problems, we aim to:

- Extend the basis selection procedure to all hierarchical levels of the random effects (i.e. implement the selection procedure to both subject specific and cycle specific levels).
- Evaluate the sensitivity of the MAP estimates by varying the initial values. In this case, we aim to vary the magnitudes of the initial values used to initiate the estimation process for parameters $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_\gamma$ and evaluate their effects on the computation time and the resulting MAP estimates.
- Compare the performance of the ML-RVM and other related approach with respect to the computation time, bias, and MISE. To achieve this, we can simulate data, implement the ML-RVM procedure and varying the number of observations per cycle, cycles per subject and the number of subjects in a study.
- Evaluate the robustness of the ML-RVM procedure against the sparseness of the data by conducting the out-of-sample prediction procedure that was used in chapter 5 and 6. We then evaluate the correlation coefficients between the fitted and the predicted values obtained after dropping different percentages of observations per cycles.

The proposed ML-RVM methodology can be extended to accommodate correlated random effects following the approach of Ciera and Dunson (2010) such that the covariance matrix for the random effects is non-diagonal. Further, the methodology can be modified to handle functional models that have demographical variables like age, number of previous births, marital status etc. Predicting non-dysfunctional menstrual cycles is an important task in many reproductive studies. Hence, the methodology can be extended to include a cycle clustering step that can be used to predict early early pregnancy loss or non-conceptional menstrual cycles. The clustering process can be based on the cycle specific parameters using parametric and non-parametric Bayesian clustering methods. The approach can also be generalize to cover many generalized linear mixed models. For example, when working with categorical outcomes, we can consider the Probit models approach (Albert and Chib 1993) to model a latent variable that can easily be used in the RVM methodologies.

Appendix A

Simplified computation methods for different quantities

In this appendix, we follow the approach of Tipping (2001) to provide simpler expressions for computing the changes in the log-likelihood functions ($\Delta\ell$ and $\Delta\ell_i$) and other quantities ($q_j,\ s_j,\ ,\ q_{ij}^*,\ s_{ij}^*$). The quantities are used while selecting the basis functions for the fixed and random effects in the modified MT-RVM procedure discussed in chapter 5. The expressions can easily be modified and be used for the functional mixed effects model discussed in chapter 6 that has correlated random effects. The appendix is divided into two parts, the first part containers expressions for quantities that are used in the selection of the fixed effects. The second part contains the expressions for the quantities used in basis functions selection for the random effects. Each part containers expressions for quantities used in the three operations (add, update, and delete) that are contained in the MT-RVM algorithm.

SELECTION OF THE BASIS FUNCTIONS FOR THE FIXED EFFECTS

To allow easy computation in the selection of the basis functions for the fixed effects, we avoid the use of expressions that have $\boldsymbol{C}_{-j}$ and instead we use the ones with $\boldsymbol{C}$. Hence, instead of using s_j and q_j that have $\boldsymbol{C}_{-j}$, we use $S_j = X_{.j}^{'}\boldsymbol{C}^{-1}X_{.j}$ and $Q_j = X_{.j}^{'}\boldsymbol{C}^{-1}\boldsymbol{Y}$. Both sets of variables are related such that $s_j = \frac{\alpha_j S_j}{\alpha_j - S_j}$ and $q_j = \frac{\alpha_j Q_j}{\alpha_j - S_j}$, when $\alpha_j = \infty$, $s_j = S_j$ and $q_j = Q_j$.

Adding a new basis function

$2\Delta\ell = \frac{Q_j^2 - S_j}{S_j} + \log\frac{S_j}{Q_j^2}$,

$$\widetilde{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{jj}\boldsymbol{\Sigma}\boldsymbol{X}^{'}X_{.j}\boldsymbol{V}X_{.j}^{'}\boldsymbol{V}\boldsymbol{X}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}_{jj}\boldsymbol{\Sigma}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j} \\ \boldsymbol{\Sigma}_{jj}\boldsymbol{\Sigma}(\boldsymbol{X}^{'}\boldsymbol{V}X_{.j})^{'} & \boldsymbol{\Sigma}_{jj} \end{bmatrix}$$

$$\widetilde{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu} \text{ - } \mu_j\boldsymbol{\Sigma}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j} \\ \mu_j \end{bmatrix}$$

$$\widetilde{S}_j = S_j - \boldsymbol{\Sigma}_{jj}(X_{.j}^{'}\boldsymbol{V}\boldsymbol{e}_j)^2$$

$$\widetilde{Q}_j = Q_j - \mu_j X_{.j}^{'}\boldsymbol{V}\boldsymbol{e}_j$$

Updating a basis function

$2\Delta\ell = \frac{Q_j^2(\widetilde{\alpha}_j-\alpha_j)}{S_j(\widetilde{\alpha}_j-\alpha_j)+\widetilde{\alpha}_j\alpha_j} - \log\{1+\frac{S_j(\widetilde{\alpha}_j-\alpha_j)}{\widetilde{\alpha}_j\alpha_j}\}$

$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - k_j\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_j^{'},$

$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - k_j\mu_j\boldsymbol{\Sigma}_j$

$\widetilde{S}_j = S_j + k_j(\boldsymbol{\Sigma}_j^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j})^2,$

$\widetilde{Q}_j = Q_j + k_j(\boldsymbol{\Sigma}_j^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j}),$

Deleting a basis function

$2\Delta\ell = \frac{Q_j^2}{S_j-\alpha_j} + \log(1-\frac{S_j}{\alpha_j}),$

$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \frac{1}{\boldsymbol{\Sigma}_{jj}}\boldsymbol{\Sigma}_j^{'},$

$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \frac{\mu_j}{\boldsymbol{\Sigma}_{jj}}\boldsymbol{\Sigma}_j,$

$\widetilde{S}_j = S_j + \frac{1}{\boldsymbol{\Sigma}_{jj}}(\boldsymbol{\Sigma}_j^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j})^2,$

$\widetilde{Q}_j = Q_j + \frac{\mu_j}{\boldsymbol{\Sigma}_{jj}}(\boldsymbol{\Sigma}_j^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j}),$

where $\boldsymbol{\Sigma}_{jj} = \frac{1}{\alpha_j+S_j}$, $\mu_j = \boldsymbol{\Sigma}_{jj}Q_j$, $\boldsymbol{e}_j \approx \boldsymbol{X}_{.j} - \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^{'}\boldsymbol{V}X_{.j}$, $k_j = \frac{\widetilde{\alpha}_j-\alpha_j}{\boldsymbol{\Sigma}_{jj}(\widetilde{\alpha}_j-\alpha_j)+1}$ and $\boldsymbol{\Sigma}_j$ is the j^{th} column of covariance matrix $\boldsymbol{\Sigma}$.

SELECTION OF THE BASIS FUNCTIONS FOR THE RANDOM EFFECTS

To allow easy computation, we avoid the expressions that contain $\boldsymbol{V}_{i,-j}$. Hence, instead of using s_{ij}^* and q_{ij}^* that have $\boldsymbol{V}_{i,-j}$, we use $S_{ij}^* = \boldsymbol{w}_{ij}^{'}\boldsymbol{V}_i^{-1}\boldsymbol{w}_{ij}$ and $Q_{ij}^* = \boldsymbol{w}_{ij}^{'}\boldsymbol{V}_i^{-1}\boldsymbol{y}_i^*$ where $\boldsymbol{y}_i^* = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\mu}$. Both sets of variables are related such that $s_{ij}^* = \frac{\omega_j S_{ij}^*}{\omega_j - S_{ij}^*}$ and $q_{ij}^* = \frac{\omega_j Q_{ij}^*}{\omega_j - S_{ij}^*}$, when $\omega_j = \infty$, $s_{ij}^* = S_{ij}^*$ and $q_{ij}^* = Q_{ij}^*$.

Adding a new basis function

$2\Delta\ell_i = \frac{Q_{ij}^{*2}-S_{ij}^*}{S_{ij}^*} + \log\frac{S_{ij}^*}{Q_{ij}^{*2}},$

$$\widetilde{\Omega}_i = \begin{bmatrix} \boldsymbol{\Omega}_i + \boldsymbol{\Omega}_{i,jj}\boldsymbol{\Omega}_i\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij}\boldsymbol{w}_{ij}^{'}\boldsymbol{W}_i\boldsymbol{\Omega}_i\sigma_\epsilon^{-2} & \text{-}\boldsymbol{\Omega}_{i,jj}\boldsymbol{\Omega}_i\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij}\sigma_\epsilon^{-2} \\ -\boldsymbol{\Omega}_{i,jj}\boldsymbol{\Omega}_i(\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij})^{'}\sigma_\epsilon^{-2} & \boldsymbol{\Omega}_{i,jj} \end{bmatrix}$$

$$\widetilde{\boldsymbol{v}}_i = \begin{bmatrix} \boldsymbol{v}_i - \upsilon_{ij}\boldsymbol{\Omega}_i\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij}\sigma_\epsilon^{-2} \\ \upsilon_{ij} \end{bmatrix}$$
$$\widetilde{S}_{ij}^* = S_{ij}^* - \boldsymbol{\Omega}_{i,jj}(\boldsymbol{w}_{ij}^{'}\boldsymbol{e}_{ij}\sigma_\epsilon^{-2})^2$$
$$\widetilde{Q}_{ij}^* = Q_{ij}^* - \upsilon_j\boldsymbol{w}_{ij}^{'}\boldsymbol{e}_{ij}\sigma_\epsilon^{-2}$$

Updating a basis function

$$2\Delta\ell_i = \frac{Q_{ij}^{*2}(\widetilde{\omega}_j-\omega_j)}{S_{ij}^*(\widetilde{\omega}_j-\omega_j)+\widetilde{\omega}_j\omega_j} - \log\{1+\frac{S_{ij}^*(\widetilde{\omega}_j-\omega_j)}{\widetilde{\omega}_j\omega_j}\}$$
$$\widetilde{\boldsymbol{\Omega}}_i = \boldsymbol{\Omega}_i - \boldsymbol{k}_{ij}\boldsymbol{\Omega}_{ij}\boldsymbol{\Omega}_{ij}^{'}$$
$$\widetilde{\boldsymbol{v}}_i = \boldsymbol{v}_i - \boldsymbol{k}_{ij}\upsilon_{ij}\boldsymbol{\Omega}_{ij}$$
$$\widetilde{S}_{ij}^* = S_{ij}^* + \boldsymbol{k}_{ij}(\boldsymbol{\Omega}_{ij}^{'}\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij})^2\sigma_\epsilon^{-2}$$
$$\widetilde{Q}_{ij}^* = Q_{ij}^* + \boldsymbol{k}_{ij}(\boldsymbol{\Omega}_{ij}^{'}\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij})\sigma_\epsilon^{-2}$$

Deleting a basis function

$$2\Delta\ell_i = \frac{Q_{ij}^{*2}}{S_{ij}^*-\omega_j} + \log(1-\frac{S_{ij}^*}{\omega_j}),$$
$$\widetilde{\boldsymbol{\Omega}}_i = \boldsymbol{\Omega}_i - \frac{1}{\boldsymbol{\Omega}_{i,jj}}\boldsymbol{\Omega}_{ij}^{'},$$
$$\widetilde{\boldsymbol{v}}_i = \boldsymbol{v}_i - \frac{\upsilon_{ij}}{\boldsymbol{\Omega}_{i,jj}}\boldsymbol{\Omega}_{ij},$$
$$\widetilde{S}_{ij}^* = S_{ij}^* + \frac{1}{\boldsymbol{\Omega}_{i,jj}}(\boldsymbol{\Omega}_{ij}^{'}\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij})^2\sigma_\epsilon^{-2},$$
$$\widetilde{Q}_{ij}^* = Q_{ij}^* + \frac{\upsilon_{ij}}{\boldsymbol{\Omega}_{i,jj}}(\boldsymbol{\Omega}_{ij}^{'}\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij})\sigma_\epsilon^{-2}$$

where $\boldsymbol{\Omega}_{i,jj} = \frac{1}{\omega_j+S_{ij}^*}$, $\upsilon_{ij} = \boldsymbol{\Omega}_{i,jj}Q_{ij}^*$, $\boldsymbol{e}_{ij} \approx \boldsymbol{w}_{ij} - \sigma_\epsilon^{-1}\boldsymbol{W}_i\boldsymbol{\Omega}_i\boldsymbol{W}_i^{'}\boldsymbol{w}_{ij}$, $\boldsymbol{k}_{ij} = \frac{\widetilde{\omega}_j-\omega_j}{\boldsymbol{\Omega}_{i,jj}(\widetilde{\omega}_j-\omega_j)+1}$ and $\boldsymbol{\Omega}_{ij}$ is the j^{th} column of covariance matrix $\boldsymbol{\Omega}_i$. A tilde on a parameter or variable indicates an updated quantity.

The described expressions are for the modified MT-RVM procedure used in the functional mixed effects model discussed in chapter 5. We note that the procedures for the modified MT-RVM discussed in chapters 5 and 6 are similar with slight different notations for the basis functions, the posterior mean and covariance matrices for the fixed and random effects. Since the quantities $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ in both chapters are are the same, then the expressions for the parameters and variables described in Appendix A can be used for the modified MT-RVM procedure discussed in chapter 6.

Appendix B

Approximate 95% credible intervals

The approximate 95% credible intervals for the parameters from the RVM procedure are:

- For parameter $\hat{\boldsymbol{\beta}}$ the approximate 95% credible intervals are,

$$\hat{\beta}_j \pm 1.96 se_{\beta_j} \qquad j = 1, \ldots, m$$

  where the standard error for β_j is $se_{\beta_j} = \sqrt{\hat{\boldsymbol{A}}_{jj}}$ such that $\hat{\boldsymbol{A}}_{jj}$ is the j^{th} diagonal elements of the estimated covariance matrix $\hat{\boldsymbol{A}}$ in equation (16).

- The approximate 95% credible intervals for $\hat{\omega}_j$ where $j = 1, \ldots, m^*$ are,

$$c_1\hat{\omega}_j\sqrt{\frac{\sum_{i=1}^{N} T_i - d^* - 1}{\chi^2_{\alpha/2,\sum_{i=1}^{N} T_i - d* - 1}}}, c_1\hat{\omega}_j\sqrt{\frac{\sum_{i=1}^{N} T_i - d^* - 1}{\chi^2_{1-\alpha/2,\sum_{i=1}^{N} T_i - d^* - 1}}}$$

  where $c_1 = \frac{2N\hat{\sigma}_\epsilon \sum_{i=1}^{N} \frac{\hat{\boldsymbol{y}}_i^{'}\hat{\boldsymbol{y}}_i}{s_i^{*2}}}{\sum_{i=1}^{N}(\frac{\hat{\boldsymbol{y}}_i^{'}\hat{\boldsymbol{y}}_i}{s_i^{*2}} - s_i^*)^2}$, $\hat{\boldsymbol{y}}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ and $d^* = m + m^* + \frac{m^*(m^*-1)}{2}$.

- The approximate 95% credible intervals for $\hat{\gamma}_{jl}$ are,

$$\hat{\gamma}_{jl} \pm 1.96 se_{\gamma_{jl}} \qquad j = 1, \ldots, m^* - 1 \quad \text{and} \quad l = j+1, \ldots, m^*$$

  where $se_{\gamma_{jl}}$ has a complex expression but can be approximated as

$$se_{\gamma_{jl}} \approx \frac{\hat{\sigma}_\epsilon \sum_{i=1}^{N} \boldsymbol{w}_{ij}\hat{h}_{il}}{\sum_{i=1}^{N} \hat{H}_{i,ll}\boldsymbol{w}_{ij}^{'}\boldsymbol{w}_{ij}}.$$

- The approximate 95% credible intervals for $\hat{\sigma}_\epsilon^2$ are,

$$c_2\hat{\sigma}_\epsilon\sqrt{\frac{\sum_{i=1}^{N} T_i - d^* - 1}{\chi^2_{\alpha/2,\sum_{i=1}^{N} T_i - d^* - 1}}}, c_2\hat{\sigma}_\epsilon\sqrt{\frac{\sum_{i=1}^{N} T_i - d^* - 1}{\chi^2_{1-\alpha/2,\sum_{i=1}^{N} T_i - d^* - 1}}}$$

where $c_2 = \frac{2\hat{\sigma}_\epsilon\sqrt{\sum_{i=1}^{N} \hat{\boldsymbol{y}}_i'\hat{\boldsymbol{y}}_i}}{\sum_{i=1}^{N}(T_i - M^* + \sum_{j=1}^{M^*}\hat{\omega}_j\hat{H}_{ijj})}$, $\hat{\boldsymbol{y}}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ and $d^* = m + m^* + \frac{m^*(m^*-1)}{2}$.

Appendix C

Simplified expressions for quantities used in the ML-RVM algorithm

Using the approach of Tipping (2001), we provide simpler expressions for computing the changes in the log-likelihood ($\Delta\ell$ and $\Delta\ell_{ij}$) and other quantities (q_l, s_l, , q^*_{ijl}, s^*_{ijl}) used in the ML-RVM algorithm discussed in section 7.3. The Appendix has two parts, the first part containers expressions for different quantities that are used in the selection of the fixed effects. The second part contains the expression for the quantities used in selection of the basis functions for the random effects at cycle specific level.

SELECTION OF THE BASIS FUNCTIONS FOR THE FIXED EFFECTS

The expressions for $s_k = X'_{.k}\boldsymbol{C}^{*-1}_{-k}X_{.k}$ and $q_k = X'_{.k}\boldsymbol{C}^{*-1}_{-k}\boldsymbol{Y}$ have matrix $\boldsymbol{C}^*_{-k}$ which is difficult to compute. To allow easy computations, we compute $S_k = X'_{.k}\boldsymbol{C}^{*-1}X_{.k}$ and $Q_k = X'_{.k}\boldsymbol{C}^{*-1}\boldsymbol{Y}$. Quantities s_k and q_k can be re-computed from S_k and Q_k using the expressions $s_k = \frac{\alpha_k S_k}{\alpha_k - S_k}$ and $q_k = \frac{\alpha_k Q_k}{\alpha_k - S_k}$. When $\alpha_k = \infty$, $s_k = S_k$ and $q_k = Q_k$.

For $l = 1, \cdots, M$, we can compute the changes of the log likelihood $\Delta\ell$ discussed in section 7.3.3 that is associated to the three operations: Add, Update and Delete. For each operations, we can be able to compute the changes for the log likelihood $\Delta\ell$, compute/update the mean vector ($\boldsymbol{\mu}$), covariance matrix ($\boldsymbol{\Sigma}$) and the estimates for the priors ($\boldsymbol{\alpha}$) as in equations (7.5, 7.15).

Adding a new basis function

$2\Delta\ell = \frac{Q_l^2 - S_l}{S_l} + \log\frac{S_l}{Q_l^2}$,

$\widetilde{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{ll}\boldsymbol{\Sigma}\boldsymbol{X}^{'}X_{.l}\boldsymbol{V}X_{.l}^{'}\boldsymbol{V}\boldsymbol{X}\boldsymbol{\Sigma} & \boldsymbol{\Sigma}_{ll}\boldsymbol{\Sigma}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l} \\ \boldsymbol{\Sigma}_{ll}\boldsymbol{\Sigma}(\boldsymbol{X}^{'}\boldsymbol{V}X_{.l})^{'} & \boldsymbol{\Sigma}_{ll} \end{bmatrix}$

$\widetilde{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu} - \mu_l\boldsymbol{\Sigma}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l} \\ \mu_l \end{bmatrix}$

$\widetilde{S}_l = S_l - \boldsymbol{\Sigma}_{ll}(X_{.l}^{'}\boldsymbol{V}\boldsymbol{e}_l)^2$

$\widetilde{Q}_l = Q_l - \mu_l X_{.l}^{'}\boldsymbol{V}\boldsymbol{e}_l$

Updating a basis function

$2\Delta\ell = \frac{Q_l^2(\widetilde{\alpha}_l - \alpha_l)}{S_l(\widetilde{\alpha}_l - \alpha_l) + \widetilde{\alpha}_l\alpha_l} - \log\{1 + \frac{S_l(\widetilde{\alpha}_l - \alpha_l)}{\widetilde{\alpha}_l\alpha_l}\}$

$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - k_l\boldsymbol{\Sigma}_l\boldsymbol{\Sigma}_l^{'}$,

$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - k_l\mu_l\boldsymbol{\Sigma}_l$

$\widetilde{S}_l = S_l + k_l(\boldsymbol{\Sigma}_l^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l})^2$,

$\widetilde{Q}_l = Q_l + k_l(\boldsymbol{\Sigma}_l^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l})$,

Deleting a basis function

$2\Delta\ell = \frac{Q_l^2}{S_l - \alpha_l} + \log(1 - \frac{S_l}{\alpha_l})$,

$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \frac{1}{\boldsymbol{\Sigma}_{ll}}\boldsymbol{\Sigma}_l^{'}$,

$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \frac{\mu_l}{\boldsymbol{\Sigma}_{ll}}\boldsymbol{\Sigma}_l$,

$\widetilde{S}_l = S_l + \frac{1}{\boldsymbol{\Sigma}_{ll}}(\boldsymbol{\Sigma}_l^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l})^2$,

$\widetilde{Q}_l = Q_l + \frac{\mu_l}{\boldsymbol{\Sigma}_{ll}}(\boldsymbol{\Sigma}_l^{'}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l})$,

where $\boldsymbol{\Sigma}_{ll} = \frac{1}{\alpha_l + S_l}$, $\mu_l = \boldsymbol{\Sigma}_{ll}Q_l$, $\boldsymbol{e}_l \approx X_{.l} - \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^{'}\boldsymbol{V}X_{.l}$, $k_l = \frac{\widetilde{\alpha}_l - \alpha_l}{\boldsymbol{\Sigma}_{ll}(\widetilde{\alpha}_l - \alpha_l) + 1}$ and $\boldsymbol{\Sigma}_l$ is the l^{th} column of covariance matrix $\boldsymbol{\Sigma}$.

SELECTION OF THE BASIS FUNCTIONS FOR THE RANDOM EFFECTS

We can avoid a direct computation of s_{ijl}^* and q_{ijl}^* since both quantities contain matrix $\boldsymbol{V}_{ij,-l}$ which has difficulties in computation. Instead, we compute $S_{ijl}^* = W_{ijl}^{'}\boldsymbol{V}_{ij}^{-1}W_{ijl}$ and $Q_{ijl}^* = W_{ijl}^{'}\boldsymbol{V}_{ij}^{-1}\boldsymbol{y}_i^*$ where $\boldsymbol{y}_{ij}^* = \boldsymbol{y}_{ij} - \boldsymbol{X}_i\boldsymbol{\mu} - \widetilde{\boldsymbol{W}}_i\boldsymbol{\gamma}_i$. We note that based on S_{ijl}^* and Q_{ijl}^*, we can compute both s_{ijl}^* and q_{ijl}^* using the expressions $s_{ijl}^* = \frac{\omega_l S_{ijl}^*}{\omega_l - S_{ijl}^*}$ and $q_{ijl}^* = \frac{\omega_l Q_{ijl}^*}{\omega_l - S_{ijl}^*}$ and when $\omega_l = \infty$, $s_l^* = S_{ijl}^*$ and $q_{ijl}^* = Q_{ijl}^*$.

Adding a new basis function

$2\Delta\ell_{ij} = \frac{Q_{ijl}^{*2} - S_{ijl}^*}{S_{ijl}^*} + \log\frac{S_{ijl}^*}{Q_{ijl}^{*2}}$,

$$\widetilde{\Omega}_{ij} = \begin{bmatrix} \boldsymbol{\Omega}_{ij} + \boldsymbol{\Omega}_{ij,ll}\boldsymbol{\Omega}_{ij}\boldsymbol{W}_{ij}^{'}W_{ijl}W_{ijl}^{'}\boldsymbol{W}_{ij}\boldsymbol{\Omega}_{ij}\sigma_{\epsilon}^{-2} & \text{-}\boldsymbol{\Omega}_{ij,ll}\boldsymbol{\Omega}_{ij}\boldsymbol{W}_{ij}^{'}W_{ijl}\sigma_{\epsilon}^{-2} \\ -\boldsymbol{\Omega}_{ij,ll}\boldsymbol{\Omega}_{ij}(\boldsymbol{W}_{ij}^{'}W_{ijl})^{'}\sigma_{\epsilon}^{-2} & \boldsymbol{\Omega}_{ij,ll} \end{bmatrix}$$

$$\widetilde{\boldsymbol{\upsilon}}_{ij} = \begin{bmatrix} \boldsymbol{\upsilon}_{ij} - \upsilon_{ijl}\boldsymbol{\Omega}_{ij}\boldsymbol{W}_{ij}^{'}W_{ijl}\sigma_{\epsilon}^{-2} \\ \upsilon_{ijl} \end{bmatrix}$$

$$\widetilde{S}_{ijl}^{*} = S_{ijl}^{*} - \boldsymbol{\Omega}_{ij,ll}(W_{ijl}^{'}\boldsymbol{e}_{ijl}\sigma_{\epsilon}^{-2})^{2}$$

$$\widetilde{Q}_{ijl}^{*} = Q_{ijl}^{*} - \upsilon_{jl}W_{ijl}^{'}\boldsymbol{e}_{ijl}\sigma_{\epsilon}^{-2}$$

Updating a basis function

$$2\Delta\ell_{ij} = \frac{Q_{ijl}^{*2}(\widetilde{\omega}_{l}-\omega_{l})}{S_{ijl}^{*}(\widetilde{\omega}_{l}-\omega_{l})+\widetilde{\omega}_{l}\omega_{l}} - \log\{1 + \frac{S_{ijl}^{*}(\widetilde{\omega}_{l}-\omega_{l})}{\widetilde{\omega}_{l}\omega_{l}}\}$$

$$\widetilde{\boldsymbol{\Omega}}_{ij} = \boldsymbol{\Omega}_{ij} - k_{ijl}\boldsymbol{\Omega}_{ijl}\boldsymbol{\Omega}_{ijl}^{'}$$

$$\widetilde{\boldsymbol{\upsilon}}_{ij} = \boldsymbol{\upsilon}_{ij} - k_{ijl}\upsilon_{ijl}\boldsymbol{\Omega}_{ijl}$$

$$\widetilde{S}_{ijl}^{*} = S_{ijl}^{*} + k_{ijl}(\boldsymbol{\Omega}_{ijl}^{'}\boldsymbol{W}_{ij}^{'}W_{ijl})^{2}\sigma_{\epsilon}^{-2}$$

$$\widetilde{Q}_{ijl}^{*} = Q_{ijl}^{*} + k_{ijl}(\boldsymbol{\Omega}_{ijl}^{'}\boldsymbol{W}_{ij}^{'}W_{ijl})\sigma_{\epsilon}^{-2}$$

Deleting a basis function

$$2\Delta\ell_{ij} = \frac{Q_{ijl}^{*2}}{S_{ijl}^{*}-\omega_{l}} + \log(1 - \frac{S_{ijl}^{*}}{\omega_{l}}),$$

$$\widetilde{\boldsymbol{\Omega}}_{ij} = \boldsymbol{\Omega}_{ij} - \frac{1}{\boldsymbol{\Omega}_{ij,ll}}\boldsymbol{\Omega}_{ijl}^{'},$$

$$\widetilde{\boldsymbol{\upsilon}}_{ij} = \boldsymbol{\upsilon}_{ij} - \frac{\upsilon_{ijl}}{\boldsymbol{\Omega}_{ij,ll}}\boldsymbol{\Omega}_{ijl},$$

$$\widetilde{S}_{ijl}^{*} = S_{ijl}^{*} + \frac{1}{\boldsymbol{\Omega}_{ij,ll}}(\boldsymbol{\Omega}_{ijl}^{'}\boldsymbol{W}_{ij}^{'}W_{ijl})^{2}\sigma_{\epsilon}^{-2},$$

$$\widetilde{Q}_{ijl}^{*} = Q_{ijl}^{*} + \frac{\upsilon_{ijl}}{\boldsymbol{\Omega}_{ij,ll}}(\boldsymbol{\Omega}_{ijl}^{'}\boldsymbol{W}_{ij}^{'}W_{ijl})\sigma_{\epsilon}^{-2}$$

where $\boldsymbol{\Omega}_{ij,ll} = \frac{1}{\omega_{l}+S_{ijl}^{*}}$, $\upsilon_{ijl} = \boldsymbol{\Omega}_{ij,ll}Q_{ijl}^{*}$, $\boldsymbol{e}_{ijl} \approx W_{ijl} - \sigma_{\epsilon}^{-1}\boldsymbol{W}_{ij}\boldsymbol{\Omega}_{ij}\boldsymbol{W}_{ij}^{'}W_{ijl}$, $k_{ijl} = \frac{\widetilde{\omega}_{l}-\omega_{l}}{\boldsymbol{\Omega}_{ij,ll}(\widetilde{\omega}_{l}-\omega_{l})+1}$ and $\boldsymbol{\Omega}_{ij,l}$ is the l^{th} column of covariance matrix $\boldsymbol{\Omega}_{ij}$. A tilde symbol on parameter or variable indicates an update.

Bibliography

Albert, J., and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669–679.

Baladandayuthapani, V., Mallick, B.K. and Carroll, R.J. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics*, **14**, 378–394.

Bigelow, J.L., and Dunson, D.B. (2007). Bayesian adaptive regression splines for hierarchical data. *Biometrics*, **63**, 724–732.

Bigelow, J.L., and Dunson, D.B. (2008). Posterior simulation across nonparametric models for functional clustering. *Sankhya*, revision submitted.

Billings, E.L. Billings, J.J., Brown, J.B., and Burger, H. (1972). Symptoms and hormonal changes accompanying ovulation. *Lancet*, 282–284.

Billings, J.J. (1983). *The Ovulation Method. Seventh Edition*, Advocate Press, Melbourne.

Botts, C.H. and Daniels, M.J. (2008). A flexible approach to Bayesian multiple curve fitting *Computational Statistics and Data Analysis*, **52**,(12) 5100–5120

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J*ournal of the American Statistical Association, **93**, 961–94.

Burges, C.J.C. (1998). *A Tutorial on Support Vector Machines for Pattern Recognition.* Kluwer Academic Publishers: Boston.

Carter, R.L. and Blight, B.J. (1981). A Bayesian change-point problem with an application to the prediction and detection of ovulation in women. *Biometrics*, **37**,(4), 743–751.

Chen, Z. and Dunson, B.D. (2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769.

Chib, S. and Greenberg, E., (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–335.

Ciera, J.M. and Dunson, D.B. (2010). Fast approximate bayesian functional mixed effects model, *Biometrics*, to appear.

Colombo, B. and Masarotto, G. (2000). Daily fecundability: First results from a new data base. *Demographic Research*, **3**,(5).

Congdon, P. (2003). *Applied Bayesian Modelling Chichester*. Wiley.

Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative study. *Journal of the American Statistical Association*. **91**, 883–904.

Crainiceanu, C.M. (2009). Bayesian Functional Data Analysis using WinBUGS. *Journal of Statistical Software*, to appear.

Crainiceanu, C.M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, B* **66**,(1) 165–185.

Crainiceanu, C., Ruppert, D. and Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, **14**, (14).

Daniels, M.J. and Kass, R.E. (1999). Nonconjugate Bayesian Estimation of Covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, **94**, 1254–1263.

Daniels, M. and Zhao, Y. (2003). Modelling the random effects covariance matrix in longitudinal data. *Statistics in Medicine*, **22**, 1631–1647.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Dunlop, A.L., Schultz, R. and Frank, E. (2005). Interpretation of the BBT Chart Using the "Gap" Technique Compared to the Coverline Technique. *Contraception*, **71**, 188–192.

Dunson, D.B., Weinberg, C.R., Perreault, S.D., and Chapin, R.E. (1999). Summarizing the motion of self-propelled cells: Applications to sperm motility. *Biometrics*, **55**, 537–543.

Durban, M., Harzelak, J., Wand, M. and Carroll, R. (2005). Simple fitting of subject-specifc curves for longitudial data. *Statistics in Medicine*, **24**, 1153–1167.

de Boor, C. (2001). *A Practical Guide to Splines. Revised edition. Applied Mathematical Sciences* **27**, Springer-Verlag, New York.

Edirisinghe, W.R., Murch, A., Junk, S. and Yovich, J.I. (1997). Cytogenetic abnormalities of unfertilized oocytes generated from in-vitro fertilization and intracytoplasmic sperm injection: a double-blind study. *Humuman Reproduction*, **12**, 2784–2791.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science*, **11**,(2) 89–121.

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

Evans, M., and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **10**, 254–272.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

Faul, A.C., and Tipping, M.E. (2002). Analysis of sparse Bayesian learning. *Advances in Neural Information Processing Systems*, **14**, 383–389.

France, J., Graham, F.M., Gosling, L. and Hair, P. (1984). A prospective study of the preselection of sex of offspring by timing intercourse relative to ovulation. *Fertility and Sterility*, **41**,(6) 894–900.

France, J., Graham, F.M., Gosling, L., Hair, P. and Knox, B.S. (1992). Characteristics of natural conceptual cycles occurring in a prospective study of sex preselection: Fertility awareness symptoms, normal levels, sperm survival, and pregnancy outcome. *Intern. Journ. of Fert.*, **37**,(4) 244–255.

Friedman, J.H. (1991). Multivariate Adaptive Regression Spline. *The Annals of Statistics*, **19**, 1.

Fruhwirth-Schnatter, S. and Tuchler, R. (2004). Bayesian Parsimonious Covariance Estimation for Hierarchical Linear Mixed Models. *Research Report Series*, **11**, WU: Vienna.

Gelfand, A.E. (2000). Gibbs Sampling, *Journal of the American Statistical Association*, **96**, 1300–1304.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*. **1**, 514–534.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. Second Edition, Chapman and Hall/CRC.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Geman, S. and Geman, A. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**,(6) 721–740.

Gelman, A., Meng, X.L. and Stern, H.S. (1996). Posterior predictive assessment of modelfitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733–807.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–473.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Geweke, J. (1992). *Bayesian Statistics*, Oxford University Press, New York.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*, Chapman and Hall, London.

Gill, J. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Second Edition, Chapman and Hall/CRC Press, New-York.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Guo, W.S. (2002). Functional mixed effects models. *Biometrics*, **58**, 121–128.

Guo, W.S. (2004). Functional data analysis in longitudinal settings using smoothing splines. *Statistical Methods in Medical Research*, **13**, 49–62.

Halberg, F., Cornelissen, G., Otsuka, K., et al. (2000). Cross-spectrally coherent 10.5 and 21-year biological and physical cycles, magnetic storms and myocardial infarctions. *Neuroendocrinol Letter*, **21**, 233–258.

Halberg, F., Cornelissen, G., Watanabe, Y., et al. (2001). Near 10-year and longer periods modulate circadians: intersecting anti-aging and chronoastrobiological research. J. Gerontol A *Biol. Sci. Med. Sci.* **56**, 304–324.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning; Data mining, Inference and Prediction*. Springer Verlag: New-York.

Hastings, W.K. (1979). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97–109.

Hedeker, D., Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Wiley: Hoboken, New Jersey.

Hilgers, T.W., Bailey, A.J. (1980). Natural family planning. II. Basal body temperature and estimated time of ovulation. *Obstetrics and Gynecology*, **55**,(3) 333–339.

Holman, D.J., Wood, J.W. (2001). Pregnancy loss and fecundability in women. In Ellison PT (ed.) *Reproductive Ecology and Human Evolution*, Aldine de Gruyter. Hawthorne, New-York, 15–38.

Holman, D.J., O'Connor, K.A. and Wood, J.W. (2006). *Age and Female Reproductive Function: Identifying the Most Important Determinants. In Sauvain-Dudgeril C, Leridon H, Mascie-Taylor N, (eds): Human Clocks: The Bio-Cultural Meanings of Age*, Oxford University Press, Oxford.

James, G., Hastie, T., Sugar, C. (2001). Principal component models for sparse functional data. *Biometrika*, *87*, 587–602.

Jeffcoate, S.L. (1983). *Use of rapid hormone assays in the prediction of ovulation. In Jeffcoate SL (ed): Ovulation: Methods for Its Prediction and Detection*, John Wiley and Sons,New-York.

Ji, S., Dunson, D.B. and Carin, L. (2009). Multi-task compressive sensing. *IEEE Transactions on Signal Processing*, **57**,(1) 92–106.

Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, **36**, 1669–1692.

Kinney, S. and Dunson, D.B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690–698.

Krivobokova, T., Crainiceanu, C.M. and Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17**, 1–20.

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

Liang, H., Wu, H. and Carroll, R.J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics*, **4**, 297–312.

Lin, X.H. and Zhang, D.W. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society B*, **61**, 381–400.

Marin, J.M. and Robert, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer, New-York.

Marshall, J. (1965). *Analyse statistique du moment de la concpton en rlation avec l'lvation de la temprature sur 5013 cycles, Actes du Congrs Mondial la Population, Belgrade, 30 Aot-10 Septembre. 1965. Vol. II: Fcondit, Planification de la famille, Mortalit 1967*, Nations Unies: New-York, 305–307.

Marshall, J. (1968). A field trial of the basal body-temperature method of regulating births. *The Lancet*, **2**, 810.

Marx, B.D. and Eilers, P.H.C. (1998). Direct generalized additive modelling with penalized likelihood. *Computational Statistics and Data Analalysis*, **28**, 193–209.

Marx, B.D. and Eilers, P.H.C. (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics*, **41**, 1–13.

Metropolis, N. and Ulam, S. (1949). The Monte-Carlo method. *Journal of the American Statistical Association*, **44**, 335–341.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1091.

Moghissi, K.S. (1976). Accuracy of basal body temperature for ovulation detection. *Fertility Sterilization*, **27**, 1415.

Morris, J.S., Vannucci, M., Brown, P.J. and Carroll, R.J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association*, **98**, 573–583.

Morris, J.S. and Carroll, R.J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series, B*, **68**,(2) 179–199.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov Chain Samplers. *Journal of the American Statistical Association*, **90**, 233–241.

O'Connor K.A., Brindle. E., Miller, R.C., Shofer, J.B., Ferrell, R.J., Klein, N.A., Soules, M.R., Holman, D.J., Mansfield, P.K., Wood, J.W., and Weinstein. (2006). Ovulation detection methods for urinary hormones: Precision, daily and intermittent sampling, and a combined hierarchical method, *Human Reproduction*, **21**,(6):1442–1452.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistic Sciences*, **1**, 505–527.

Odeblad, E. (1994). The discovery of different types of cervical mucus and the Billings Ovulation Method. *Bulletin of the Natural Family Planning Council of Victoria* **21**, 3.

Pauler, D.K., Wakefield, J.C. and Kass, R.E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association*, **94**, 1242–1253.

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

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika*, **94**,(4) 1006–1011.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raftery A.L. and Lewis S. (1992). *Bayesian Statistics*, Oxford University Press, New York.

Ramsay, J.O., and Silverman, B.W. (2005). *Functional Data Analysis*, Springer-Verlag, New-York.

Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy Gibbs sampler. *Journal of the American Statistical Association*, **87**, 861–868.

Roberts, G.O. (1994), Methods for Estimating L2 Convergence of Markov Chain Monte Carlo. *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner, eds. D. Berry, I. Chaloner, and J. Geweke*, Amsterdam: North Holland.

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New-York.

Rodgers, J.L. and Kohler, H.P. (2003). *The biodemography of human reproduction and fertility*, Kluwer academics publishers.

Ross, G.T., Cargille, C.M., Lipsett, M.B., Rayford, P.L., Marshall, J.R., Strott, C.A. and Rodbard, D. (1970). Pituitary and gonadal hormones in women during spontaneous and induced ovulatory cycles. *Recent Prog. Horm. Res.*, **26**, 1–62.

Royston, J.P. and Abrams, R.M. (1980). An Objective Method for Detecting the Shift in Basal Body Temperature in Women, *Biometrics*, **36**,(2), 217–224.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* **92**, 1049–1062.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. *Unpublished manuscript.*

Ruppert, D. and Carroll, R.J. (2000). Spatially adaptive penalties for spline fitting. *Australia and New Zealand Journal of Statistics*, **42**, 205–223.

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

Sakamoto, W. (2007). Selecting basis functions and knots with an empirical Bayes method. *Computational Statistics*, **22**,(1) 583–597.

Scarpa, B., and Dunson, D.B. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, **65**,(3) 772–780.

Smith, B.J. (2007). boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *Journal of Statistical Software*, **21**,(11).

Smith, A.F.M. (1991). Bayesian computational methods. *Phil. Trans. Royal Society*, London.

Smith, M. and Kohn, R. (1996). Nonparametric regression via Bayesian variable selection. *Journal of Econometrics*, **75**, 317–344.

Smith, M. and Kohn, R. (2002). Parsimonious Covariance Matrix Estimation for Longitudinal Data. *Journal of the American Statistical Association*, **97**, 1141–1153.

Sperroff, L., Glass, R.H. and Kase, N., (1994). *Regulation of the menstrual cycle. Clinical gynecologic endocrinology and fertility*, 5th edition, Baltimore: Williams and Wilkins, 809.

Stram, D. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**,(3) 1171–1177.

Tanner, M., (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Third edition, Springer-Verlag, New-York.

Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing*, **10**,(4), 325–337.

Thomas, A., O'Hara, B., Ligges, U., Sturtz, S. (2006). Making BUGS Open. *R News*, **6**,(1), 12–17.

Thompson, W.K. and Rosen, O. (2008). A Bayesian model for sparse functional data. *Biometrics*, **64**, 54–63.

Tibshirani, R.J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B* **58**, 267–288.

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, **1**, 211–244.

Tuerlings, J.M. (2000). *Genetic Aspects of Male Factor Infertility*, Durk.

Vollman, R.F. (1977). Assessment of the fertile and sterile phases of the menstrual cycle, *Intern. Rev. of Nat. Fam. Plann.*, **1**,(1) 40–47.

Wand, M.P. (2003) Smoothing and mixed models. *Computational Statistics*, **18**, 223–249.

Wand, M.P. and Jones, M.C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, **9**, 97–116.

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

Wand, M.P. and Ormerod, J.T. (2008). On O'Sullivan penalised splines and semiparametric regression. *Australian and New Zealand Journal of Statistics*, **50**, 179–198.

Wu, H. and Liang, H. (2004). Backfitting Random Varying-Coefficient Models with Time-Dependent Smoothing Covariates. *Scand. J. of Statist.*, **31**, 3-19.

Wu, H.L. and Zhang, J.T. (2002). Local Polynomial Mixed-Effects Models for Longitudinal Data, *Journal of the American Statistical Association*, **97**, 883-897.

Yao, F., Muller, H.G. and Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**,(470) 577–591.

Yu, B. and Mykland, P. (1994), Looking at Markov Samplers Through Cusum Path Plots: A Simple Diagnostic Idea, *Technical Report No. 413*, Department of Statistics, University of California at Berkeley.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93**, 710–719.

Zuspan, K.J. and Zuspan, F.P. (1979). *Basal Body Temperature. In Human Ovulation: Mechanisms, Predictions, Detection and Induction*, Ed. ESE Hafez. Elsevier, North Holland.

James Mbugua Ciera - Curriculum Vitae

Personal information

Address:	Department of Statistical Sciences, University of Padua. Via Cesare Battisti 241-243, 35121 Padova - ITALY.
Email:	ciera@stat.unipd.it or jmciera@yahoo.com
Phone:	+39 - 346 854 8441

Education

Jan, 2007 - Dec, 2009:	Department of Statistical Sciences, University of Padua (Italy), PhD. Statistical Sciences, Approximate Bayes Random Effects models for large datasets. Advisors: Dr. Scarpa Bruno and Prof. David Dunson.
Oct, 2001 - Sept, 2003:	Limburgs Universitair Centrum (Belgium), Msc. Biostatistics, Msc. Applied Statistics.
Mar, 1996 - Dec, 1999:	Jomo Kenyatta University of Agriculture and Technology (Kenya), Bsc. Science (Statistics and Computer Sceince).

Work Experience

Sept, 2005 - Dec, 2006:	African Population and Health Research Center (Kenya), Data Analyst.
Oct, 2003 - Aug, 2005:	Kenya Methodist University (Kenya), Tutorial Fellow.
May - Aug, 2000:	Micro-LAN Kenya Limited (Kenya), Assistant computer programmer.

Research Periods Abroad

12 - 18 July 2009: Summer School,
at Institute on Reproductive and Perinatal Epidemiology,
Montreal, Canada.

April-Sept. 2008: Internship program with David Dunson (*Duke University, USA*),
at National Institute of Environmental Health Sciences,
Research Triangle, Durham, NC, USA.

Conference and Seminar Presentations

14-17 Sept. 2009: Milan, Italy. (*S. Co. conference*)
"Fast Bayesian Functional Data Analysis: Application to basal body temperature data". (*Oral presentation*).

20-25 July 2009: New York, USA. (*IWSM workshop*)
"Fast Bayesian Functional Data Analysis: Application to basal body temperature data".(*Poster presentation*).

12-18 June 2009: Turin, Italy: (*7th Workshop on Bayesian Non-parametrics*)
"Fast Approximate Bayesian Functional Mixed Effects Model". (*Poster presentation*).

09 Jan. 2009: Padova, Italy. (*PhD Seminars,-Seminari intermedi dei dottorandi del XXII ciclo*)
"A Fast Approximate Bayesian Functional Model". (*Oral presentation*).

24-26 Oct. 2007 Rome, Italy: (*5th International Conference on Behavioural Risk factor Surveillance*)
Presentations I (oral): "The Nairobi Urbani Health and Demographic Surveillance System (NUHDSS): platform for monitoring health outcomes."
Presentations II (poster): "The utility of demographic surveillance systems (DSS) for chronic disease risk factor surveillance in developing countries."

Publications

Ciera, J.M. and Dunson, B.D. (2010). Fast Approximate Bayesian Functional Mixed Effects Model. *Biometrics*, Submitted.

Ciera, J.M. (2010). Fast Bayesian Functional Data Analysis of Basal Body Temperature. *Complex data modeling and computationally intensive statistical methods*, Submitted.

Ciera, J.M., Scarpa, B. and Dunson, D.B. (2009). Fast Bayesian Functional Data Analysis: Application to basal body temperature data. *Proceedings of the 24th International Workshop on Statistical Modelling, Ithaca, NY, USA; July 20-24, 2009.*

Ciera, J.M., Scarpa, B. and Dunson, B.D. (2009). Fast Bayesian Functional Data Analysis: Application to basal body temperature data. *Department of Statistical Sciences, University of Padova: Working Paper Series; 18, November, 2009.*

Fotso, J.C., Ezeh, A., Madise, J.N. and **Ciera, J.** (2007). Progress towards the child mortality millennium development goal in urban sub-Saharan Africa: the dynamics of population growth, immunization, and access to clean water. *B.M.C. Public Health* **7**: 218.

Madise,J.N., Zulu, E. and **Ciera J.** (2007). Is Poverty a Driver for Risky Sexual Behaviour? Evidence from National Surveys of Adolescents in four African Countries. *African Journal of Reproductive Health; African Journal of Reproductive Health*, ISSN: 1118–4841, **11**:3, 83–98.