BAYESIAN SEMIPARAMETRIC METHODS FOR FUNCTIONAL DATA

by
Jamie Lynn Bigelow

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2005

Approved by:

Dr. Amy Herring, Advisor
Dr. David Dunson, Advisor
Dr. Chirayath Suchindran, Committee Member
Dr. Haibo Zhou, Committee Member
Dr. John Thorp, Committee Member

ABSTRACT
JAMIE LYNN BIGELOW: BAYESIAN SEMIPARAMETRIC METHODS FOR FUNCTIONAL DATA.
(Under the direction of Dr. Amy Herring and Dr. David Dunson.)

Motivated by studies of reproductive hormone profiles in the menstrual cycle, we develop methods for hierarchical functional data analysis. The data come from the North Carolina Early Pregnancy Study, in which measurements of urinary progesterone metabolites are available from a cohort of women who were trying to become pregnant. Methods for menstrual hormone data are needed that avoid standardizing menstrual cycle lengths while also allowing for flexible relationships between the hormones and covariates. In addition, it is necessary to account for within-woman dependency in the hormone trajectories from multiple cycles. All of the methods are developed for and applied to menstrual hormone data, but they are general enough to be applied in many other settings.

The statistical approach is based on a hierarchical generalization of Bayesian multivariate adaptive regression splines. The generalization allows for an unknown set of basis functions characterizing both the overall trajectory means and woman-specific covariate effects and allows for the complex dependency structure of the data. To relax distributional assumptions, we use a Dirichlet process prior on the unknown distribution of the random basis coefficients in the spline model. This requires the development of methodology for the use of the Dirichlet process on the distribution of a parameter of varying dimension. While modeling the curves nonparametrically, the Dirichlet process also identifies clusters of similar curves. Finally, we combine our approach with Bayesian methods for generalized linear models, developing a procedure that clusters trajectories while jointly estimating the response distribution of each cluster.

In all of the models, a reversible jump Markov chain Monte Carlo algorithm is developed for posterior computation. Applying the methods to the progesterone data, we investigate differences in progesterone profiles between conception and non-conception cycles, identify clusters of pre-ovulatory progesterone, and demonstrate the ability of the joint model to distinguish early pregnancy losses from clinical pregnancies.

ACKNOWLEDGMENTS

Many thanks to David Dunson and Amy Herring for their ideas and guidance from start to finish. I would like to thank my committee for their time and support. I am grateful to Donna Baird, Clarice Weinberg, and Allen Wilcox for providing the NC-EPS data and for their helpful comments.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEW & INTRODUCTION

1.1 Motivating Example

1.1.1 Hormones and the menstrual cycle

Certain physiological characteristics, such as hormone levels, basal body temperature, follicle diameter and cervical mucus are known to vary in accordance with the menstrual cycle (Stanford et al., 2002). It is of interest to understand how these characteristics tend to vary within and across different menstrual cycles and in relationship to predictors, such as age or environmental exposures. Also of interest is the inter-relationship among these different physiological characteristics and their association with fertility and pregnancy outcomes. In this paper, we develop models appropriate for longitudinal menstrual data and apply the models to progesterone trajectories. The models can be generalized to other menstrual and non-menstrual applications based on the concepts discussed in Section 1.1.3.

A brief review of the menstrual cycle is warranted before discussion of methodology. The menstrual cycle has traditionally been defined to start at the onset of a menstrual

bleed and to end the day before the onset of the next menstrual bleed. The first phase of the cycle, occuring up to the day of ovulation, is termed the follicular phase. Ovulation marks the transition into the luteal phase, which continues until menstrual bleeding marks the start of the next cycle (Murphy et al., 1995).

Progesterone levels tend to be low at the cycle start, rise to a peak, and then decrease Baird et al. (1997). The location of the peak is related to the timing of ovulation within the cycle. Menstrual cycles tend to vary among and even within women with respect to cycle length and timing of ovulation. Thus modeling the progesterone trajectory is not a simple matter of the estimation of a function over a fixed amount of time, with a fixed ovulation point. Researchers have approached this problem using several different methods.

Since ovulation is an extremely important reference point in the cycle, marking the end of the fertile interval and known hormonal changes, many investigators have chosen simply to restrict consideration to a fixed interval surrounding ovulation, usually between seven and fourteen days in length (see Baird et al., 1997; Brumback and Rice, 1998; Massafra et al., 1999 for examples). This allows the incorporation of data from cycles of various lengths without complicated adjustment for cycle length. Mean progesterone profiles within the interval can easily be estimated by calculating the mean progesterone level for each day in the interval, and one can potentially incorporate covariates and allow for within-woman dependency by using a hierarchical model.

This method has several drawbacks. Truncation yields no information about progesterone levels on the days outside the chosen interval. In addition, if the interval is chosen to be too wide it may extend outside the current cycle, and thus reflect preceding or subsequent ovulations. In attempts to model progesterone over the whole cycle, some investigators have chosen to standardize all cycle lengths to 28 days, effectively shrinking the longer cycles and stretching shorter ones (see Zhang et al., 1998 for an ex-

ample). This maintains the relative length of the cycle phases, but masks any inherent relationship between progesterone and cycle length.

Harlow (1991) point out that, in studies of menstrual cycle diary data, long cycles tend to be underrepresented, and short cycles tend to be overrepresented. One reason for this is that long cycles are more likely to be truncated at the start or end of the study period. If all cycles are standardized to the same length in the analysis, then the results tend to be more representative of short cycles than of long cycles. Another source of this disproportionate representation of cycle lengths is that women who tend to have shorter cycles will contribute more over the study period. This is a problem of informative cluster size, with each cluster being the set of cycles contributed by a given woman. In scenarios like this, it is well known that failure to correct for cluster size can lead to biased inference (Dunson et al., 2003; Romero et al., 1992).

The justification for cycle standardization is entirely based on analytical convenience. That is, there is no evidence to suggest that cycles of different lengths will have similar characteristics. Studies have shown that menstrual cycle length varies among women according to BMI, exercise habits, and diet, (Kato et al., 1999) all of which also affect hormone levels (Unzer et al., 1995; Jasienska et al., 2000). Standardizing the cycle length obscures the relationship between the length of the menstrual cycle and the hormone levels.

The relationship between hormone levels and covariates has often been assessed by creating summary variables (i.e. mean progesterone level during the luteal phase, peak progesterone level, etc.) and comparing them across levels of the covariate of interest (see Baird et al., 1997, 1999). The simplicity of this method makes it an attractive way to model and draw inference about progesterone levels, and differences among women with respect to summary variables can be biologically informative. However, the creation of summary variables can mask the richness of the available data and result

in a loss of information.

A recent study by van Zonneveld et al. (2003) compared hormone trajectories and follicle development in young and older women of reproductive age. To deal with the differing cycle lengths and ovulation timings, they compared the two groups on each of three time-scales: days relative to the start of the cycle, days following the BBT-shift, and days leading up to the luteinizing hormone peak. They concluded that each time scale provided slightly different information about the hormone trajectory. this indicates that the understanding of hormones over the menstrual cycle is best understood through the consideration of multiple reference points.

1.1.2 Early pregnancy study

The data to we use to develop these methods are from the North Carolina Early Pregnancy Study (EPS). This prospective cohort study was conducted to determine the risk of early loss of pregnancy among healthy women. The study population, recruited in the early 1980s, consisted of 281 couples who were planning to become pregnant. The women collected daily first-morning urine samples from the time they stopped using birth control until six months had passed or until the eighth week of clinical pregnancy. The urine samples were assayed for pregnanediol-3-glucuronide (PdG), a progesterone metabolite, along with metabolites of many other hormones of interest. Study protocol and preliminary findings are detailed by Wilcox et al. (1985) and Wilcox et al. (1988), and the progesterone data are described in Baird et al. (1997).

We restrict attention to the subset of menstrual cycles for which a hormonally-determined day of ovulation is available. The ovulation day was estimated by the ratio of estrogen and progesterone metabolites in urine, which decreases abrubtly in response to ovulation (Wilcox et al., 1998). This measure of ovulation is superior to BBT-based estimates, and nearly as accurate as estimates based on the urinary luteinizing hormone

surge (Ecochard et al., 2001).

In Chapter 2, we examine hormone profiles over the entire menstrual cycle, comparing conception and non-conception cycles. In Chapter 3, we consider progesterone pre-implantation, which we take to be the follicular phase until two days after ovulation. Finally, in Chapter 4, we consider post-ovulatory progesterone in an analysis of cycles that resulted in a clinical pregnancy and cycles that resulted in an early loss.

1.1.3 Multiple reference point data

Longitudinal data problems can be adapted to a spatial framework through the identification of multiple reference points in time. van Zonneveld et al. (2003) indicate that understanding hormones over the menstrual cycle requires the incorporation of multiple reference points. For simplicity, suppose we consider two reference points: the start of the cycle, and the day of ovulation. We expect that both of these points will be informative about progesterone. Each measurement can be given a set of coordinates (r, s), where r is the cycle day, and s is the day relative to ovulation. If we are correct in our assumptions about the importance of these reference points in predicting progesterone, then measurements on days with similar coordinates will tend to be similar, and our model should reflect this.

Multiple reference points are not unique to menstrual data. They are also important in epidemiologic studies of exposures. Consider a study with the goal of determining the relationship between exposure and disease. Subjects in the study provide their date of exposure and other time-independent covariates of interest. At follow-up visits, the age of the subject and the disease status are collected. Each measurement can then be given a set of coordinates (x, y), where x is the number of years since exposure, and y is the age of the patient at measurement. In this case, the date of exposure and date of birth are the reference points for a given subject. The investigators may have reason

to believe that time since first exposure, age at first exposure, and current age all play a role in disease status. All of this information is contained in the coordinate pair for each measurement and if the investigator is correct, we would expect disease status measurements with similar coordinates to be similar.

There is a need for innovative methods for multiple reference point data. These methods should allow the response to vary flexibly in relation to the reference points, accommodate the estimation and testing of covariate effects, and account for within-trajectory dependency. In terms of the progesterone data, this model will incorporate the timing relative to the start of the cycle and the day of ovulation. The model should be flexible enough to account for the various trajectory shapes that are seen in women, and to allow for dependency within women and within cycles.

If we choose to think of the model in this spatial framework, then the methods we develop will have applicability to any problem where the framework applies. This includes regression settings without reference points and not necessarily longitudinal as well as true spatial settings, such as image analysis or environmental modeling.

The remainder of this chapter summarizes statistical methods that may be adapted to the multiple reference point setting. Section 1.2 discusses methods for spatial modeling, and Section 1.3 discusses regression methods that may be adapted to the multiple reference point and spatial settings. Section 1.6 is a summary of preliminary work I have done to this end. Section 1.7 outlines the dissertation project, which includes the development of a regression method appropriate for multiple reference point data, a method for clustering based on the regression coefficients, and a joint model describing the relationship between trajectories and outcomes of interest.

1.2 Methods for spatial modeling

After demonstrating the spatial nature of these data, a natural first step was to explore methods for spatial analysis. The model we choose should model the data flexibly, but should also be easily interpretable in a non-spatial manner that is relevant to the context of the original data problem. In this section, I discuss several methods that may be adapted to achieve this goal.

1.2.1 Random fields

A random field is a region in space over which random variables may be observed. It may be finite or infinite, discrete or continuous, and of any dimension. Every observation from a random field will be associated with a location in space. The interpretation of random field data depends on these known locations and the spatial correlation structure of the field. Random fields are most directly seen in true spatial or space-time data, and are readily applicable to the study of meteorological phenomena (Handcock and Wallis, 1994), analysis of agricultural field experiments (Allcroft and Glasbey, 2003), image reconstruction (Besag, 1986) and modeling of disease incidence (Waller et al., 1997; Knorr-Held et al., 2002; Knorr-Held and Richardson, 2003).

Spatial data often has the property that observations that are near each other tend to be similar. Modeling and inference through the use of a random field is a powerful way to use this property. The simplest case of this is data smoothing. For example, a meteorologist may collect air temperature information over a region, then employ an algorithm to predict the temperature where there were no observations. This is often accomplished by a technique known as kriging, in which the analyst uses available data to model the correlation structure as a function of distance between points. The correlation structure is then used in an iterative procedure to predict the value of

the random field at a large number of unobserved locations given the available data. (Handcock and Stein, 1993)

A stationary random field has the property that covariance among points is constant across the field. That is, two points with the same relative position will have the same correlation regardless of their exact location in the field. The following discussion assumes that random fields are stationary, although models and computational methods could be adjusted to relax this assumption.

1.2.2 Markov random fields

Besag (1974) describes a special case of a random field consisting of a finite number of sites with a univariate random variable observed at each of the sites. For simplicity, suppose this random field is 1-dimensional with an infinite number of regularly spaced sites indexed by the integers. Each site r has an associated random variable y_r. This field is called a Markov random field (MRF) if the distribution of y_r given $\mathbf{y_{-r}}$, the observed values at all other sites, depends only upon the values at a finite set of sites that are 'neighbors' of r. A simple case is that when the distribution of y_r given $\mathbf{y_{-r}}$ depends only upon y_{r+1} and y_{r-1}. I use ∂r to denote the set of neighbors of site r, and $\mathbf{y_{(r)}}$ to denote the set of observed values at those sites. Neighbor definition is often arbitrary, and little is known about the impact the chosen neighborhood structure can have on inference (Assunção et al., 2002).

In this one-dimensional field, distances between locations are fixed. The correlation between points is only meaningful at specified distances between points, so there is no need to model correlation as a continuous function of distance. As described in Besag et al. (1991), the pairwise difference prior can be used to induce this type of correlation structure. In the context of the one-dimensional random field described above, each site is defined by a unique integer i. If we let $w_{ij} = w_{ji}$ be the weight that describes

the relative degree of association between site i and j, the pairwise difference prior is:

$$p(\mathbf{x}) \propto exp\left\{ -\sum_{i<j} w_{ij}\phi(y_i - y_j) \right\} \tag{1.1}$$

where ϕ is a function that increases with the absolute value of its argument. We wish to adapt the pairwise difference prior to induce a Markov random field. If we let w_{ij}=0 if $j \notin \partial i$, the pairwise difference prior reduces to:

$$p(\mathbf{x}) \propto exp\left\{ -\sum_{i\sim j} w_{ij}\phi(y_i - y_j) \right\} \tag{1.2}$$

where $i \sim j$ indicates all pairs of i and j such that $i < j$ and i and j are neighbors. This results in a conditional density where the value of y_r depends only on the values at neighboring sites, and thus we have a Markov random field:

$$p(x_r|x_{-r}) \propto exp\left\{ -\sum_{j\in\partial i} w_{ij}\phi(y_i - y_j) \right\} \tag{1.3}$$

When this conditional distribution is normal, we are working in a Gaussian MRF. Choosing $\phi(u) = \tau * u^2/2$ in the pairwise difference prior will induce a Gaussian MRF with the following conditional distribution:

$$x_i|x_{-i} \sim N\left(\frac{\sum_{j\in\partial i} w_{ij}y_j}{\sum_{j\in\partial i} w_{ij}}, \tau^{-1}\sum_{j\in\partial i} w_{ij} \right) \tag{1.4}$$

In words, a value at a point given values at all other points depends only on its neighbors, and is normally distributed with the mean corresponding to a weighted average of the neighbors. This normality is desirable, as many methods are available for sampling and inference in the Gaussian distribution.

This prior specification illustrates the two ways in which a MRF structure can be

induced. Besag and Kooperberg (1995) call these intrinsic and conditional autoregressions. (The terminology is inconsistent throughout the literature, I will use Besag and Kooperberg's distinction). Under intrinsic autoregressive model specification, the joint prior is specified so that a MRF structure results. This is what we do when we choose the pairwise difference prior and Markov neighbor structure. Another approach would have been to specify the prior entirely in terms of a dependent set of conditional distributions that described a MRF. This method, known as conditional autoregression or auto-modeling, often leads to an easier computational algorithm. This is not an issue here, as this special case of the pairwise difference prior (an intrinsic autoregression) leads to a simple conditional structure. The conditional structure induced by the special case of the pairwise difference prior above has also been called a Gaussian conditional autoregression (Besag, 1974), or an auto-Gaussian model (Cressie and Chan, 1989). Kaiser and Cressie (2000) discuss the specification of models through conditional distributions and conditions when and how these conditional distributions correspond to a joint density. In the specifying of a joint prior, we eliminate the need for this consideration.

Besag et al. (1991) point out that the pairwise difference prior is improper because it doesn't address actual values of random variable across the field, only differences among the values. However, in the presence of informative data, the resulting posterior distributions will be proper. If necessary, another way to avoid impropriety would be to restrict any one of the field values to a plausible finite interval.

This model is easily extended to higher dimensions. On the two-dimensional regular lattice, each site is indexed by an integer pair (r, s) and has an associated random variable y_{rs}. This field is a MRF if the distribution of y_{rs} given the observed values at all other sites depends only upon the values at a finite set of 'neighboring' sites, ∂_{rs}. Neighboring sites can be defined in various ways, the simplest being the "first order"

setting, where the the distribution of y_{rs} given the observed values at all other sites depends only on the values of $y_{r+1,s}$, $y_{r-1,s}$, $y_{r,s+1}$, and $y_{r,s-1}$. When the field is finite, this can be simplified to the one-dimensional case by implementing a 1-1 transformation from the coordinate pairs into the integers, and defining neighbors appropriately. This can be extended to describe a MRF in n dimensions, although neighbor definition becomes more complex as the number of dimensions increases.

The adaptability of MRFs to longitudinal data has been demonstrated. Besag et al. (1995) performed Bayesian logistic regression on longitudinal data, with the incorporation of unobserved covariates to account for extra-binomial variation. The data are from a cohort study of prostate cancer deaths, and a spatial model is used to incorporate both information about age at observation and birth cohort. In other words, the researchers expected the death rate in 50-year-old men in 1940 to be different from that of 50-year-old men in 1980, so age alone was not sufficient information. They found a frequentist logistic regression model with age and cohort to be inadequate (i.e. there was extra-binomial variation), and chose to use Bayesian methods to account for this. They treated the data as arising from a binomial MRF, where the coordinates of the three-dimensional field were age group (i), observation year (j), and cohort number (k). Note that cohort number is uniquely determined by year and age group. Where z_{ij} is the random unobserved covariate for coordinate (i,j,k) and p_{ij} is the probability of prostate cancer at coordinate (i,j,k), the model was:

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \theta_i + \phi_j + \psi_k + z_{ij} \tag{1.5}$$

In words, the effect of any one of the three coordinates was the same regardless of the values of the other two. Thus the age effect was the same for all years and cohorts. A pairwise difference prior was put on θ, ϕ, and ψ independently, so that each of these effect was thought to come from a distinct one-dimensional MRF. An alternative

approach would have been to model the log-odds directly, setting up a two-dimensional neighbor structure with a pairwise-difference prior, under the model:

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu_{ij} + b_k \tag{1.6}$$

In this model, b_k is a cohort random effect with prior mean 0 to allow for the dependence among observations from the same cohort. This modeling procedure does not separate the age and year parameters and could allow for a non-linear relationship between the log-odds, age, and year.

As in the two models above, non-spatial data with multiple reference points in time (or space) can be adapted to the spatial paradigm for ease of modeling or to counteract modeling problems. In the current problem, I examine daily measurements of progesterone levels over the menstrual cycle. Progesterone levels tend to be low at the start and end of the cycle, with a peak near ovulation. The location of a given measurement can be classified relative to three reference points: cycle start day, day of ovulation, and cycle end day.

An important goal of the project is to incorporate covariates that do not vary systematically with time (i.e. covariates other than the reference points). Assunção et al. (2002) examine fertility rates across a region of Brazil, fitting a Poisson regression model for the number of births in each small region. They put a pairwise difference prior on the model coefficients, allowing the coeffiicients to be more similar in neighboring regions, but to vary across the entire region. This allows the incorporation of covariate effects in the usual GLM manner, but gives flexibility in that the covariate effects can vary across the region. In the multiple reference point problem, these types of covariate effects would need to be combined with the effects of location relative to the reference point.

A hidden Markov model (HMM) is a model in which the observed data are not direct

realizations of the Markov random field. Rather, the field is an underlying process, and the observed data provide information about this process. In the current problem, we say that the the mean progesterone value at each site is a realization of a MRF. We observe several observations at each set of coordinates, which will give us information about the mean (the MRF), but we have no direct observations of the MRF. More complex HMMs have applications in speech recognition and image analysis (Kunsch et al., 1995). Given the hierarchical structure of HMMs, Bayesian hierarchical modeling through MCMC algorithms is well-suited for computation.

The flexibility of the Bayesian hierarchical model is also useful in dealing with other non-standard features of the progesterone data. Some sites in the region will have no observations at all, so it is necessary to estimate the mean at these sites through their correlation with the means at neighboring sites. Additionally, data are not independent across the field. Each menstrual cycle contributes multiple data points, and it is unreasonable to assume that observations within a cycle have the same correlation structure as observations from different cycles.

The current problem would require a random field over a regular lattice. When coordinates are not equally spaced, a generalization of the pairwise difference prior can be employed. Berthelson and Moller (2003) discuss MCMC inference based on this prior, which takes into account the exact distance between two points in calculating their degree of association. The MRF neighborhood structure can still apply here, but careful problem-specific definition is required.

1.3 Regression methods

Rather than thinking of our covariates and times relative to reference points as coordinates in space, and alternative is to think of them as covariates in a regression model.

This works conceptually, as we would expect observations with similar times relative to reference points to be similar in the same way as a regression model tends to assign similar response values to observations with similar covariates. This section contains a review of some flexible regression methods that may be adapted to this setting.

1.3.1 Multivariate linear splines

Holmes and Mallick (2001) developed a semiparametric regression methodology for modeling a response as a function of a design matrix. The response is modeled using piecewise linear splines, so that the regression surface consists of hyperplanes across the covariate space. The number and location of splines are treated as random and are updated using the reversible jump sampler. The resulting samples are non-smooth surfaces, but Bayesian model averaging can be applied to the samples to produce a smooth regression surface.

Bayesian model averaging techniques were developed as a method for dealing with the uncertainty about model correctness. Often, there is uncertainty about which model to choose, but the standard response is to pick the 'best' one and use it for all inference. Bayesian model averaging allows for the incorporation of model uncertainty into inference (Raftery et al., 1997). Holmes and Mallick (2001) capitalized on the properties of Bayesian model averaging to create a smooth regression surface from a sample of implausible but informative non-smooth surfaces.

Holmes and Mallick (2003) generalized this method to the setting where the outcome is non-normal and multivariate. It remains, however, to implement this method when the independent response data are vectors of varying lengths with differing covariate values.

1.3.2 Non-parametric regression

Lin and Zhang (1999) proposed the generalized additive mixed model (GAMM), where the linear predictor is the sum of combination of linear functions of the covariates, non-parametric smooth functions of the covariates, and random effects. These models allow for flexible covariate effects, but the random effects have an additive effect on the linear predictor.

Wavelets are orthogonal families of basis functions that can be used to approximate another function. (Clyde et al., 1998) Modern wavelet theory was brought about through its applicability to signal and image processing (Akay, 2003) and computer graphics (Schroder, 1996). Specifically, wavelets are frequently used to remove noise from signal data and pictures. The ability of wavelets to eliminate noise makes them broadly applicable in statistical analysis, where the primary goal is to eliminate random error and estimate an underlying process. Morris et al. (2003) describe nonparametric wavelet regression in a model with a hierarchical dependence structure, which was accounted for through random effects. This model shows promise in the current multiple-reference-point problem, as the dependence structure could be expanded to allow for spatial association among observations.

Ray and Mallick (2003) propose a wavelet model in the context of a MRF. Their discussion in framed in the context of image analysis. In summary, they partition the image and allow the wavelet transformation to vary across partitions. A pairwise difference prior constrains the transformations to be more similar in neighboring regions than in non-neighboring regions.

1.3.3 Non-smooth functions

There are times when a function can be expected to have discontinuities in space. In the context of longitudinal studies, the occurrence of some event (i.e. disease onset,

ovulation) may be known to cause a discontinuity in the outcome of interest. In the multiple reference point format, if this event was used as a reference point, the model would need to be flexible enough to allow for a 'jump' at that reference point. For example, there is a rapid drop in estrogen levels at ovulation (Alliende, 2003). In building a spatial model for this quantity, the best model would relax the degree of association (i.e. reduce the level of smoothing) among neighboring coordinates at points of discontinuity. This may be a consideration in choosing a modeling strategy for the multiple-reference point problem.

1.4 Conclusion of initial literature review

The literature review summarizes methods in spatial analysis as well as flexible regression procedures that may be adapted to a spatial framework. These techniques may be appropriate in the development of methods for multiple reference point data. Bayesian methods will be used to implement models with broad assumptions on the spatial structure and covariate effects. Upon preparing the rest of the document, several new research questions were encountered. Citations relevant to these questions are found throughout the ensuing chapters.

1.5 Choosing an approach

Our inital work in building a flexible model for this hormone data focused on the extension of Markov random field (MRF) theories. The motivation for this type of model structure was the time-dependence seen in menstrual hormone data, leading naturally to modeling the correlation between subsequent measurements. After building a very flexible MRF model for these data, we realized that this approach was limited in both computational efficiency and interpretability. We then moved focus from the

MRF approach to an approach using splines, which capture the changes in time as slopes rather than modeling a point-by-point dependence structure. Section 1.6 outlines the MRF we applied initially, highlighting some of its successes and limitations. It also provides some further insight into multiple reference point data and the spatial nature of regression.

1.6 Preliminary work in MRFs

1.6.1 Data

At the time of preliminary analyses, the EPS data was not yet available so we used data from Brumback and Rice (1998) to investigate the applicability of Markov random fields to menstrual data. The data consisted of menstrual cycles truncated to the eight days preceding through the 15 days following ovulation. To simulate complete cycles of varied lengths, I added randomly generated numbers of days and corresponding measurements onto the beginnings and ends of each cycle. Since the cycle lengths were simulated, these preliminary analyses allow for no inference about cycle length and progesterone. However, they serve to illustrate how we applied the MRF paradigm to multiple reference point data.

Daily measurements of progesterone were available, over 69 total cycles. Consider the set of measurements from one cycle. For each value, we know the day relative to the start of the cycle (r) and the day relative to ovulation (s). The measurement location can be defined by the coordinate pair (r, s). Clearly, there are only certain values of (r, s) that can occur within the natural constraints of the menstrual cycle. Thus I define the region of interest in the coordinate plane to be the smallest parallelogram-shaped region that contains all points at which data were observed. The full model specification is given in Section 1.6.2. In summary, a HMM was fit to the data, incorporating random

effects to account for within-cycle dependence. The model still needs to be expanded to include covariate effects and to account for multiple cycles contributed from the same woman.

1.6.2 Model specification

A MRF was used to build a model for the mean log-progesterone level at each coordinate. Let μ_{rs} be the unknown mean level at (r, s), and let $\boldsymbol{\mu}$ be the $n \times 1$ vector of unknown means. The MRF framework states that, given its neighbors, μ_{rs} is independent of all other elements of $\boldsymbol{\mu}$. We define the neighbors of (r, s) to be the eight elements that entirely surround the site. Specifically, the neighbors of (r, s) are $(r-1, s)$, $(r+1, s)$, $(r, s-1)$, $(r, s+1)$, $(r-1, s-1)$, $(r+1, s+1)$, $(r+1, s-1)$, and $(r-1, s+1)$. For simplicity, this initial model assumes equal unit weights among all pairs of neighbors. Note that since we are working in a finite lattice, not all sites will have all eight neighbors. We denote the number of neighbors of (r, s) by p_{rs}, and $uv \sim jk$ is the set of all pairs of neighboring sites. The precision parameter, δ, controls the degree to which neighboring sites are similar. The desired neighborhood structure can be induced through the following pairwise difference prior:

$$p(\boldsymbol{\mu}) \propto \delta^{n/2} exp\Big\{ -\frac{\delta}{2} \sum_{uv \sim jk} (\mu_{uv} - \mu_{jk}^2) \Big\} \tag{1.7}$$

Letting ∂_{rs} be the set of all sites that neighbor (r, s) and $\boldsymbol{\mu_{-(rs)}}$ contain all elements of $\boldsymbol{\mu}$ except for μ_{rs}, the corresponding prior conditional structure is:

$$\mu_{rs} \mid \boldsymbol{\mu_{-(rs)}} \sim N\left(\frac{\sum_{jk \in \partial_{rs}} \mu_{jk}}{p_{rs}}, \delta^{-1} p_{rs} \right) \tag{1.8}$$

Let $\mathbf{y_i}$, be the vector of all log-progesterone levels for cycle i. Let R_i be the set of coordinate pairs observed in cycle i. The elements of $\mathbf{y_i}$ are assumed to be independent

given μ_i and the random effects, λ_i and ν_i. We also assume that, marginalizing over the random effects, a measurement at (r, s) is normally distributed with mean μ_{rs}. To illustrate the incorporation of the random effects, suppose (r, s) is in R_i, and the measurement at (r, s) from cycle i is called $y_{i,rs}$. We then assume the following distribution of the data:

$$y_{i,rs} \mid \mu_{rs}, \lambda_i, \nu_i \sim N\big(\lambda_i \mu_{rs} + \nu_i, \tau^{-1}\big) \tag{1.9}$$

In this case, λ_i and ν_i are the cycle-specific random effects. We assume that, given the random effects, observations within a cycle are independent. The prior distributions of λ_i and ν_i are normal, centered at 1 and 0 respectively. All elements of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are a priori independent.

This specification implies joint normality of all data given $\boldsymbol{\mu}$ and the random effects. If $\mathbf{y_{rs}}$ is the $n_{rs} \times 1$ vector of measurements at (r, s) and $\boldsymbol{\lambda_{rs}}$ and $\boldsymbol{\nu_{rs}}$ are the appropriately ordered vectors of random effects for all cycles with measurements at (r, s), the likelihood for the data from site (r, s) is:

$$\pi(\mathbf{y_{rs}} \mid \mu_{rs}, \boldsymbol{\lambda_{rs}}, \boldsymbol{\nu_{rs}}) = N\big(\boldsymbol{\lambda_{rs}} \mu_{rs} + \boldsymbol{\nu_i}, \tau^{-1} I_{n_{rs}}\big) \tag{1.10}$$

The joint data likelihood is normal and can be written as:

$$\pi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \tau) = \prod_{rs} \pi(\mathbf{y_{rs}} \mid \mu_{rs}, \boldsymbol{\lambda_{rs}}, \boldsymbol{\nu_{rs}}) \tag{1.11}$$

Since we are not observing direct realizations of the MRF, this is a hidden Markov model. One attractive effect of this particular specification is that, even though the MRF structure is present in $\boldsymbol{\mu}$, the data at (r, s), given μ_{rs} are independent of all other elements of $\boldsymbol{\mu}$. The distributional assumptions are given here.

Likelihood:

$$\pi(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \tau) \propto \prod_{i=1}^{m} \prod_{uv \in R_i} \tau^{1/2} exp(-\frac{\tau}{2}(y_{i,uv} - (\lambda_i \mu_{uv} + \nu_i))^2) \tag{1.12}$$

Priors:

$$\begin{aligned}
\pi(\tau) &\propto \tau^{a-1} e^{-b\tau} \\
\pi(\delta) &\propto \delta^{c-1} e^{-d\delta} \\
\pi(\boldsymbol{\lambda}) &\propto \prod_{i=1}^{m} exp\Big\{ -\frac{(\lambda_i - 1)^2}{2\sigma_l^2} \Big\} \\
\pi(\boldsymbol{\nu}) &\propto \prod_{i=1}^{m} exp\Big\{ -\frac{\nu_i^2}{2\sigma_n^2} \Big\} \\
\pi(\boldsymbol{\mu}) &\propto \delta^{n/2} exp\Big\{ -\frac{\delta}{2} \sum_{uv \sim jk} (\mu_{uv} - \mu_{jk})^2) \Big\}
\end{aligned}$$

where a, b, c, d, σ_l, and σ_n are specified hyperparameters.

1.6.3 Sampling algorithm

A Gibbs sampling algorithm was implemented to obtain samples from the posterior distributions. The full conditionals below were used to directly sample from the posterior distributions of $\boldsymbol{\mu}$, τ, and δ.

$$\begin{aligned}
\mu_{rs} | \tau, \delta, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{y} &\sim N\left(\frac{\delta \sum_{jk \sim rs} \mu_{jk} + \tau(\mathbf{y_{rs}}' \boldsymbol{\lambda_{rs}} - \tau \boldsymbol{\lambda}'_{\boldsymbol{rs}} \boldsymbol{\nu_{rs}})}{\delta p_{rs} + \tau \boldsymbol{\lambda}'_{\boldsymbol{rs}} \boldsymbol{\lambda_{rs}}}, (\delta p_{rs} + \tau \boldsymbol{\lambda}'_{\boldsymbol{rs}} \boldsymbol{\lambda_{rs}})^{-1} \right) \\
\tau | \boldsymbol{\mu}, \delta, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{y} &\sim Gamma\left(a + \frac{m}{2}, b + \frac{1}{2} \sum_{rs} (\mathbf{y_{rs}} - \mu_{rs} \boldsymbol{\lambda_{rs}} - \boldsymbol{\nu_{rs}})' (\mathbf{y_{rs}} - \mu_{rs} \boldsymbol{\lambda_{rs}} - \boldsymbol{\nu_{rs}}) \right) \\
\delta | \boldsymbol{\mu}, \tau, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{y} &\sim Gamma\left(c + \frac{n}{2}, d + \frac{1}{2} \sum_{uv \sim jk} (\mu_{uv} - \mu_{jk})^2 \right)
\end{aligned}$$

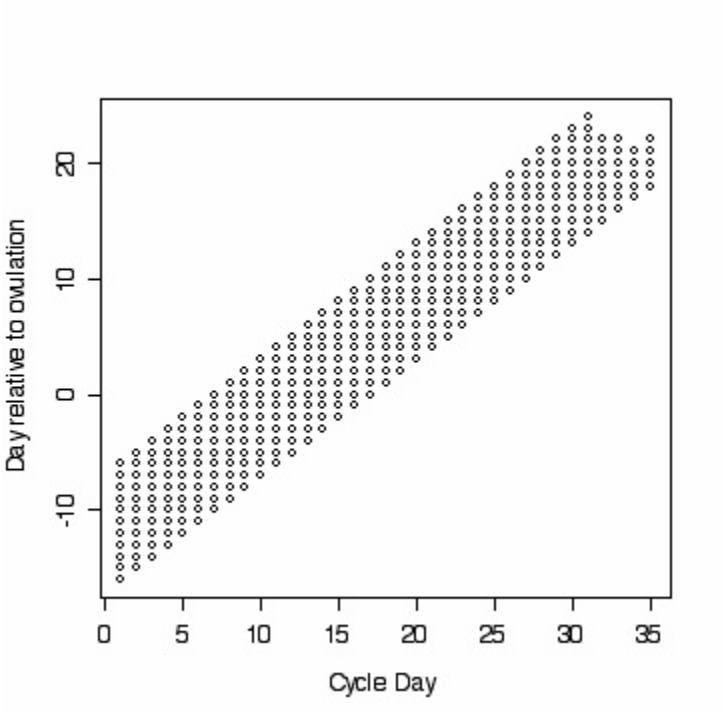

FIGURE 1.1: The region of coordinates covered by the simulated data.

Metropolis steps allowed for sampling from the posterior distributions of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. The algorithm was implemented in Matlab. The number of samples collected was 8000, 2000 of which were discarded as burn-in. Convergence was apparent for all parameters, with slower mixing for the random effect parameters.

1.6.4 Results

The first step was to examine the coordinate region covered by the data. The two reference points were day relative to cycle start and day relative to ovulation. Figure 1.1 shows the range of coordinates that were defined by the data. The smallest parallelogram containing all these points was the region over which the field was modeled.

The next step was to determine how the data varied over the region of interest.

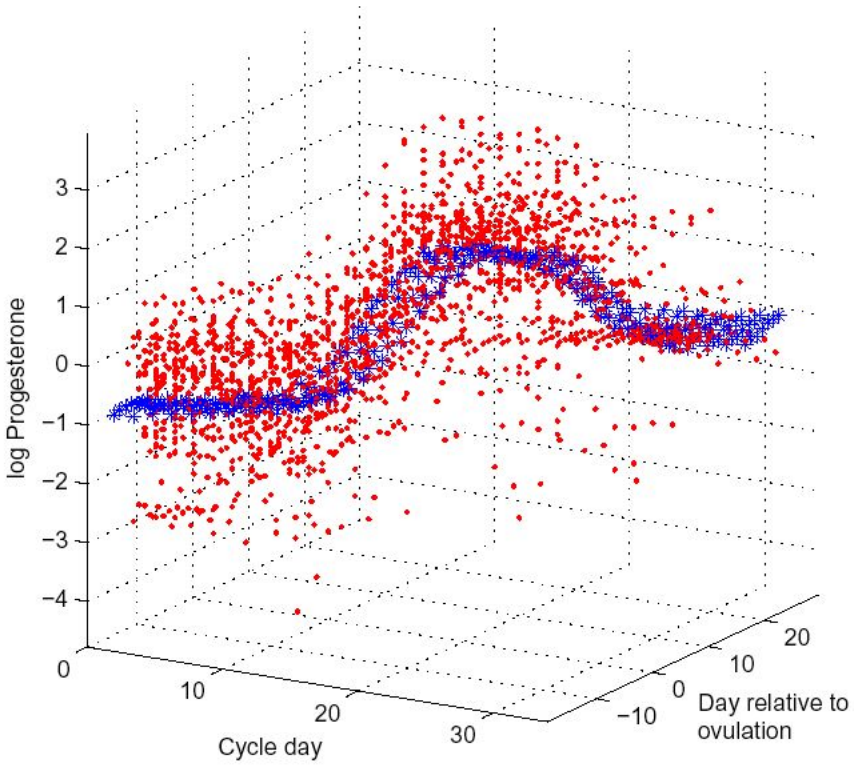

FIGURE 1.2: Progesterone over the selected region. Red points are the data, and blue points are the posterior estimate of $\boldsymbol{\mu}$.

Figure 1.2 shows all of the progesterone data, plotted in red according to coordinates. The posterior estimate of $\boldsymbol{\mu}$ is plotted in blue. The versatility of the random effects structure is illustrated in Figure 1.3. For four different cycles, these plots illustrate the data, the mean progesterone according to the coordinates of the cycle days coordinates, and the estimate of the cycle-specific trajectory, accounting for the random effects. These show that the random effects structure allows for flexible trajectory shapes.

1.6.5 Discussion of the MRF method

This approach has yielded the desired flexible regression model, but it has several limitations. Firstly, as we are modeling each data point individually, there is little

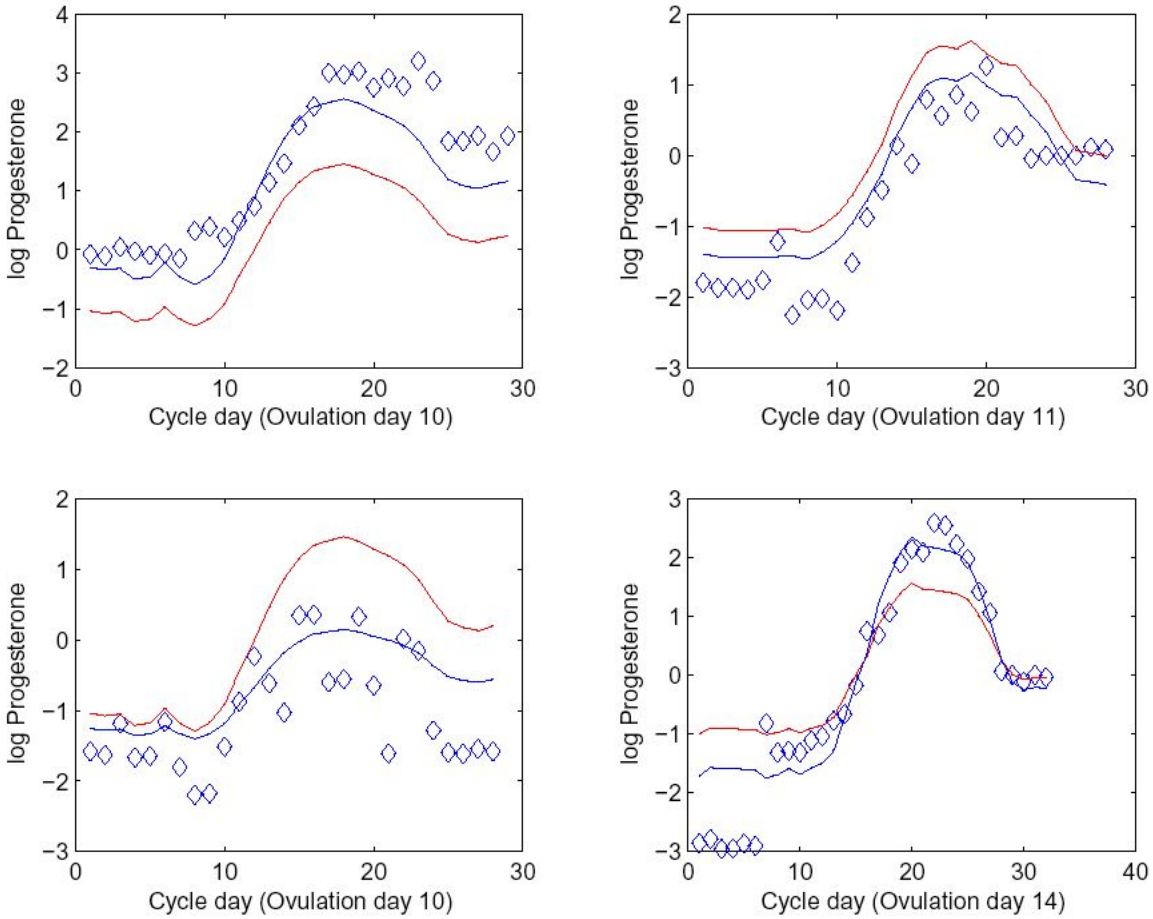

FIGURE 1.3: Progesterone for four cycles. On each plot, the diamonds are one cycle of data. The red line is the elements of $\boldsymbol{\mu}$ corresponding to that cycle. The blue line is the cycle-specific trajectory, which is the elements of $\boldsymbol{\mu}$ adjusted according to the cycle random effects.

informations directly available about slopes, peaks, etc. Posterior information about these things could be obtained through careful analysis of the samples. Given our interest in the trajectories as a whole, however, we prefer a method that focuses less on the point-to-point correlation structure and more on the shape of the curve.

1.7 This Dissertation

To flexibly model the progesterone curves while dealing with some of the issues that arose in the development of the MRF model, we propose a flexible spline model with random coefficients. Each of Chapters 2 and 3 is a self-contained article describing some methodological innovation and demonstrating its application to progesterone data. Chapter 2 describes a novel nonparametric regression model and demonstrates its ability to flexibly characterize curves. Chapter 3 outlines a method for introducing a nonparametric distribution of the coefficients describing the curves and describes a way to cluster trajectories into groups according to shape. Chapter 4 describes the joint modeling of curves and outcome variables. Chapter 5 describes some of the implications and future directions of these methodological developments.

CHAPTER 2

BAYESIAN ADAPTIVE REGRESSION SPLINES FOR HIERARCHICAL DATA

This chapter considers methodology for hierarchical functional data analysis, motivated by studies of reproductive hormone profiles in the menstrual cycle. Current methods standardize the cycle lengths and ignore the timing of ovulation within the cycle, both of which are biologically informative. Methods are needed that avoid standardization, while flexibly incorporating information on covariates and the timing of reference events, such as ovulation and onset of menses. In addition, it is necessary to account for within-woman dependency when data are collected for multiple cycles. We propose an approach based on a hierarchical generalization of Bayesian multivariate adaptive regression splines. Our formulation allows for an unknown set of basis functions characterizing the population-averaged and woman-specific trajectories in relation to covariates. A reversible jump Markov chain Monte Carlo algorithm is developed for posterior computation. Applying the methods to data from the North Carolina Early Pregnancy Study, we investigate differences in progesterone profiles between conception

and non-conception cycles.

2.1 Introduction

In many longitudinal studies, each subject contributes a set of data points that can be considered error-prone realizations of a function of time. Although it is standard practice to model the longitudinal trajectory relative to a single reference point in time, such as birth or the start of treatment, there may be several reference points that are informative about a subject's response at a given time. One example of reference points is disease onset, start of treatment, and death in a longitudinal study of quality of life. The current project uses onset of menses and ovulation as reference points in a study of reproductive hormones.

Our research was motivated by progesterone data from the North Carolina Early Pregnancy Study (NCEPS) (Wilcox et al., 1988). Daily measurements of urinary pregnanediol-3-glucuronide (PdG), a progesterone metabolite, were available for 262 complete menstrual cycles and 199 partial mid-cycle segments from a total of 173 women. It is of special interest to examine the differences in progesterone profiles between conception and non-conception cycles. The onset of menses marks the start of the follicular phase of the menstrual cycle, which ends at ovulation. The luteal phase begins at ovulation and, if no conception occurs, ends at the start of the next menses. In general, progesterone begins to rise in the follicular phase until several days into the luteal phase, when it decreases in preparation for the next cycle or, if conception has occurred, continues to rise. Figure 2.1 displays log-PdG data from one subject for a non-conception and subsequent conception cycle.

The most common way to examine hormone data within the menstrual cycle is to restrict attention to a fixed window surrounding ovulation (see Baird et al., 1997;

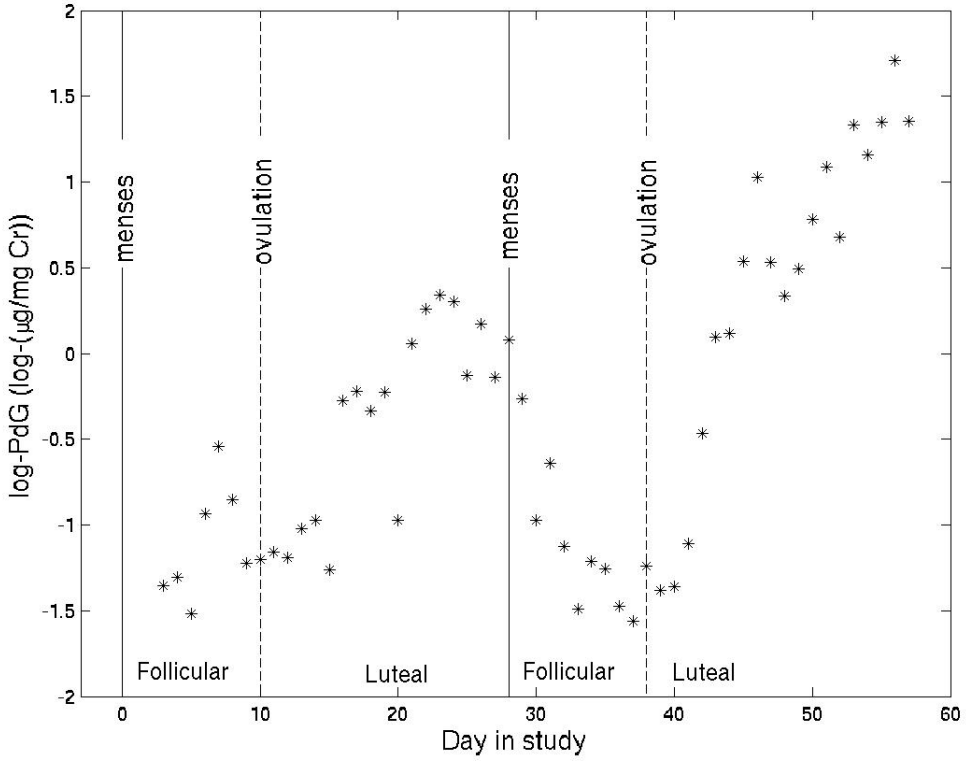

FIGURE 2.1: log(PdG) for a non-conception followed by a conception cycle from one subject. Solid lines indicate first day of each cycle, and dashed lines indicate ovulation days.

Brumback and Rice, 1998; Massafra et al., 1999; Dunson et al., 2003 for examples). This is desirable for ease of modeling, but fails to use all data by discarding days outside the window. In addition, it ignores cycle length and the relative timing of ovulation within a cycle. Another approach is to standardize all cycles to a common length. Zhang et al. (1998) modeled progesterone with smoothing splines after standardizing cycles to 28 days. Standardization discards biologically important information on the timing of ovulation, obscuring its well known relationship with hormone trajectories. van Zonneveld et al. (2003) indicated that both the onset of menses and the day of ovulation are related to hormone levels within a cycle and implemented separate analyses for windows around each of these reference points. Ideally, a single model would allow the response to vary flexibly relative to multiple reference points, while accommodating covariates and within-woman dependency.

The goal of the analysis is to characterize differences in progesterone profiles between conception and non-conception cycles. When conception occurs, PdG rises in response to implantation of the conceptus, which usually occurs around the eighth or ninth day after ovulation (Baird et al., 1997). We are also interested in differences before implantation because they may predict the fertility of the cycle. Researchers have studied conception differences in midluteal (5-6 days after ovulation) and baseline (preovulatory) PdG. Studies of have shown that conception cycles have elevated midluteal PdG over paired non-conception cycles with well-timed intercourse or artificial insemination (Stewart et al., 1993; Baird et al., 1997), but one study (Lipson and Ellison, 1996) found no difference. None of these three studies found a relationship between baseline PdG and conception. However, limiting analysis to cycles with well-timed exposure to semen is biased to include non-conception cycles of inherently low fertility, failing to represent the true difference between conception and non-conception cycles. In addition, requiring paired cycles selects against couples of very high or very

low fertility and fails to use all data from women with more or less than two cycles. In a previous analysis of the NCEPS data which included cycles without well-timed intercourse, Baird et al. (1999) found that cycles with very low midluteal PdG were unlikely to be conception cycles. Although midluteal PdG did not monotonically affect the odds of conception, increased baseline PdG was associated with decreased odds of conception.

Hormone data are a special case of hierarchical functional data. The daily measurements are subject to assay errors, yielding a noisy realization of the true trajectory of urinary PdG. The hierarchy results from the multiple cycles contributed by each woman. Methods for hierarchical functional data typically require that all curves are observed over or standardized to fall in the same region (Brumback and Rice, 1998; Morris et al., 2003; Brumback and Lindstrom, 2004). To accommodate the dependence structure without cycle standardization, we propose a Bayesian method based on a hierarchical generalization of multivariate adaptive regression splines.

Our approach is related to methods for nonlinear regression and smoothing for longitudinal and correlated data. Lin and Zhang (1999) proposed the generalized additive mixed model (GAMM), where the linear predictor is the sum of linear functions of the covariates, non-parametric smooth functions of the covariates, and random effects. The GAMM allows for flexible covariate effects, and could potentially be modified to accommodate reference points. However, they are designed so that random effects vary linearly with the covariates, which may not be flexible enough to describe the differences among women in hormone profiles. Fahrmeir et al. (2004) propose a generalization of the GAMM for space-time data. They model the response using p-splines with a fixed number of knots, and each covariate enters the model independently. This successfully accounts for dependence among observations, but further extensions are needed to allow for a random number of knots and flexible interactions among covariates. Guo

(2002) introduces a non-parametric flexible model for a population trajectory and the random effects, but does not allow for the introduction of additional non-parametric covariate effects.

Motivated by applications to speech data, Brumback and Lindstrom (2004) recently proposed the use of random time transformations to align times or data features (i.e. reference points) within subjects. Inference about covariate effects is based on a comparison of the estimated transformations. Ratcliffe et al. (2002) give an example of functional regression using a spline basis, where all the functions are observed over the same region of time. These methods requires standardization of the time scale as in Zhang et al. (1998). In addition, Ratcliffe et al. (2002) allows for only one curve per subject and Brumback and Lindstrom (2004) requires that all subjects contribute a given number of curves under each covariate condition, both of which are unrealistic in menstrual studies.

James et al. (2000) describe a method for modeling sparsely-sampled growth curve data. After choosing a spline basis to represent the curves, they employ reduced rank principal components analysis to estimate the population mean function. Though their approach doesn't require standardization of time, we wish to estimate both the population curve and the trajectories themselves.

Holmes and Mallick (2001) proposed Bayesian regression with multivariate linear splines to flexibly characterize the relationship between covariates and a scalar response from independent sampling units. The number of knots and their locations are random, and smooth prediction curves are obtained by averaging over MCMC sampled models. A extension of this method yielded a generalized nonlinear regression model for a vector response (Holmes and Mallick, 2003). Our goal is to develop a new hierarchical adaptive regression splines approach to accommodate clustered functional data, potentially having unequal numbers and locations of observations per subject, a common compli-

cation in longitudinal studies. We incorporate reference point information by including time relative to each of the reference points as covariates in the regression model.

A popular method for analyzing multivariate response data with spline bases is seemingly unrelated regression (SUR), in which each subject is allowed a unique set of basis functions, but the basis coefficients are common to all subjects (Percy, 1992). We instead use one set of unknown basis functions, allowing the basis coefficients to vary from subject to subject. To estimate the population regression function, we treat the subject-specific basis coefficients as random, centered around the population mean basis coefficients. The resulting model is extremely flexible, and can be used to capture a wide variety of covariate effects and heterogeneity structures.

In section 2.2, we describe the model, prior structure and a reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995) algorithm for posterior computation. In section 2.3, we illustrate the performance of the approach for a simulation example. Section 2.4 applies the method to progesterone data from the NC-EPS, and section 2.5 discusses the results.

2.2 Methods

2.2.1 Prior specification

Typically, the number and locations of knots in a piecewise linear spline are unknown. By allowing for uncertainty in the knot locations and averaging across the resulting posterior, one can obtain smoothed regression functions. We follow previous authors (Green, 1995; Holmes and Mallick, 2001) in using the RJMCMC algorithm to move among candidate models of varying dimension. Our final predictions are constructed from averages over all sampled models. We assume a priori that all models are equally probable, so our prior on the model space is uniform.

Each piecewise linear model, M, is defined by its basis functions $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$, where $\boldsymbol{\mu}_l$ is $p \times 1$. Consider y_{ij}, the j^{th} PdG measurement for subject i. Under model M, the true relationship between y_{ij} and its covariates $\mathbf{x}'_{ij} = (1, x_{ij2}, \ldots, x_{ijp})$ can be approximated by the piecewise linear model:

$$y_{ij} = \sum_{l=1}^{k} b_{il}(\mathbf{x}'_{ij}\boldsymbol{\mu}_l)_+ + \epsilon_{ij}, \tag{2.1}$$

where $\epsilon_{ij} \overset{iid}{\sim} N(0, \tau^{-1})$. The value of the j^{th} response of subject i is approximated by a linear combination of the positive portion (denoted by the + subscript) of the inner products of the basis functions with the covariate vector, $\mathbf{x}_{ij}$. We require that each model contain an intercept basis, so we define $(\mathbf{x}'_{ij}\boldsymbol{\mu}_1)_+ \equiv 1$ for all i, j. We extend previous methods by allowing the spline coefficients, $\mathbf{b}_i$ to be subject-specific, assuming that observations within subject i are conditionally independent given $\mathbf{b}_i$.

Each piecewise linear model is linear in the basis function transformations of the covariate vectors:

$$\mathbf{y}_i = \boldsymbol{\theta}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{2.2}$$

where $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_i$ are the $n_i \times 1$ vectors of responses and random errors and $\mathbf{b}_i$ is the $k \times 1$ vector of subject specific basis coefficients for subject i. The $n_i \times k$ design matrix, $\boldsymbol{\theta}_i$, contains the basis function transformations of the covariate vectors for subject i:

$$\boldsymbol{\theta}_i = \begin{pmatrix} 1 & (\mathbf{x}'_{i1}\boldsymbol{\mu}_2)_+ & \ldots & (\mathbf{x}'_{i1}\boldsymbol{\mu}_k)_+ \\ 1 & (\mathbf{x}'_{i2}\boldsymbol{\mu}_2)_+ & \ldots & (\mathbf{x}'_{i2}\boldsymbol{\mu}_k)_+ \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (\mathbf{x}'_{in_i}\boldsymbol{\mu}_2)_+ & \ldots & (\mathbf{x}'_{in_i}\boldsymbol{\mu}_k)_+ \end{pmatrix}$$

Since we use only the positive portion of each linear spline, it is possible that a basis function does not contribute to the model for a given subject (i.e. $\boldsymbol{\theta}_i$ contains

a column of zeros, which is non-informative about the corresponding element of $\mathbf{b}_i$). To address this problem, we standardize each column of the population design matrix, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_m')'$, to have mean 0 and variance 1. Assuming independent subjects, this model specification yields the likelihood:

$$L(\mathbf{y}|\boldsymbol{b}, \tau, M) \propto \prod_{i=1}^{m} \tau^{\frac{n_i}{2}} exp\big[-\frac{\tau}{2}(\mathbf{y}_i - \boldsymbol{\theta}_i\mathbf{b}_i)'(\mathbf{y}_i - \boldsymbol{\theta}_i\mathbf{b}_i)\big] \tag{2.3}$$

This likelihood is defined conditionally on the subject-specific basis coefficients, but we wish to make inferences also on population parameters. Treating the subject-specific coefficients as random slopes, we specify a Bayesian random effects model where the subject-specific coefficients are centered around the population coefficients, $\boldsymbol{\beta}$. Under model M of dimension k, the relationship between the population and subject-specific coefficients is specified through the hierarchical structure:

$$\begin{aligned} \mathbf{b}_i|k &\sim N_k(\boldsymbol{\beta}, \tau^{-1}\boldsymbol{\Delta}^{-1}) \quad \forall i \\ \boldsymbol{\beta}|k &\sim N_k(\mathbf{0}, \tau^{-1}\lambda^{-1}\mathbf{I}_k) \end{aligned} \tag{2.4}$$

To avoid over-parameterization of an already flexible model, we assume independence among the elements of $\mathbf{b}_i$. Thus $\boldsymbol{\Delta} = diag(\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a $k \times 1$ vector. The elements of $\boldsymbol{\delta}$ and the scalars λ and τ are given independent gamma priors:

$$\pi(\tau, \lambda, \boldsymbol{\delta}) \propto \tau^{a_\tau - 1} exp(-b_\tau \tau) \lambda^{a_\lambda - 1} exp(-b_\lambda \lambda) \prod_{l=1}^{k} \big(\delta_l^{a_\delta - 1} exp(-b_\delta \delta_l) \big),$$

where $a_\tau,\ b_\tau,\ a_\lambda,\ b_\lambda,\ a_\delta$ and b_δ are pre-specified hyperparameters. Each of the $k-1$ non-intercept basis functions contains a non-zero intercept and linear effect for at least one covariate. Including multiple covariate effects in a single basis allows the covariates to dependently affect the response (i.e. allows for interactions). The number of non-zero

covariate effects in a particular basis is called the interaction level of the basis.

Under one piecewise linear model, an observation y with covariates x has the following mean and variance:

$$E(y) = \beta_1 + \sum_{l=2}^{k} \beta_l (\mathbf{x}'\boldsymbol{\mu}_l)_+$$

$$V(y) = \delta_1^{-1} + \sum_{l=2}^{k} \delta_l^{-1} (\mathbf{x}'\boldsymbol{\mu}_l)_+^2 + \tau^{-1}$$

The mean and variance can vary flexibly with the covariates and relative to each other. The elements of $\boldsymbol{\beta}$ can be positive or negative, large or small, and the elements of $\boldsymbol{\delta}$ can also be large or small. A given basis could contribute substantially to the mean and negligibly to the variance (i.e. β_l and δ_l are both large), or vice versa, so that the mean and variance of the response at a given set of covariates are not constrained to vary together.

2.2.2 Posterior computation

At each iteration, we obtain a piecewise linear model for which the parameters can be sampled directly from their full conditionals as derived from the priors and the likelihood following standard algebraic routes. Omitting details, we obtain the following full conditional posterior distributions:

$$\boldsymbol{\beta}|\mathbf{b}, \boldsymbol{\delta}, \lambda, \tau, \mathbf{D} \sim N_k\Big([\lambda \mathbf{I}_k + m\boldsymbol{\Delta}]^{-1}\boldsymbol{\Delta} \sum_{i=1}^{m} \mathbf{b}_i, \tau^{-1}[\lambda \mathbf{I}_k + m\boldsymbol{\Delta}]^{-1}\Big)$$

$$\mathbf{b}_i|\boldsymbol{\beta}, \boldsymbol{\delta}, \lambda, \tau \sim N_{k_i}\Big([\boldsymbol{\theta}_i'\boldsymbol{\theta}_i + \boldsymbol{\Delta}]^{-1}[\boldsymbol{\theta}_i'\mathbf{y}_i + \boldsymbol{\Delta}\boldsymbol{\beta}], \tau^{-1}[\boldsymbol{\theta}_i'\boldsymbol{\theta}_i + \boldsymbol{\Delta}]^{-1}\Big) \qquad i = 1, \ldots, m$$

$$\tau|\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}, \lambda \sim Gamma\Big(a_\tau + \frac{(m+1)k + n}{2},$$

$$b_\tau + \frac{1}{2}\sum_{i=1}^{m}[(\mathbf{b}_i - \boldsymbol{\beta}_i)'\boldsymbol{\Delta}(\mathbf{b}_i - \boldsymbol{\beta}_i) + (\mathbf{y}_i - \boldsymbol{\theta}_i\mathbf{b}_i)'(\mathbf{y}_i - \boldsymbol{\theta}_i\mathbf{b}_i)] + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}\Big)$$

$$\lambda|\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}, \tau \sim Gamma\Big(a_\lambda + \frac{k}{2}, b_\lambda + \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2}\Big)$$

$$\delta_l|\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}_{-l}, \lambda, \tau \sim Gamma\Big(a_\delta + \frac{m}{2}, b_\delta + \frac{\tau}{2}\sum_{i=1}^{m}(\mathbf{b}_{il} - \beta_l)^2\Big) \qquad l = 0, \ldots, (k-1)$$

where a $Gamma(a, b)$ random variable is parameterized to have expected value a/b and variance a/b^2.

The following is a description of the RJMCMC algorithm we employed:

Step 0: Initialize the model to the intercept-only basis function, where $k = 1$.

Step 1 : Propose with equal probability either to add, alter or remove a basis function. If $k = 1$ in the current model, then we cannot remove or change a basis, so we choose either to add a basis function or to skip to step 2 and redraw the parameters for the intercept basis.

ADD Generate a new basis function as follows: Draw the interaction level of the basis uniformly from $(1, \ldots, p-1)$ and randomly select the corresponding number of covariates. Set basis parameters for all other covariates equal to zero. Sample selected basis function parameters from $N(0, 1)$, then normalize to get $(\mu_{l1}, \ldots, \mu_{lp})$, the non-intercept basis parameters. Randomly select one data point, y_{ij}, and let $\mu_{l0} = \mathbf{x}'_{ij,-1}\mu_{l,-1}$. Add the new basis function to the proposed model.

ALTER: Randomly select a basis in the current model. Generate a new basis function as described above. Replace the selected basis function with the new one

REMOVE: Randomly select a basis in the current model. Delete the selected basis from the proposed model.

Step 2: Accept the proposed model with appropriate probability (described below).

Step 3: If a proposal to add or remove has been accepted, the dimension of the model has changed. In order to update the parameters from their full conditionals, all vector parameters must have dimension k^* of the new model. It suffices to adjust the dimension of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, as we can then sample $\{\mathbf{b}_i\}$ from the full conditionals. If we've added a basis, initialize β_{k^*}, the new element of $\boldsymbol{\beta}$, to a pre-determined initial value and initialize δ_{k^*} to the mean of $\boldsymbol{\delta}$ from the previous model. If a basis has been removed, delete the corresponding elements of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

Step 4: Update $\{\mathbf{b}_i\}$, $\boldsymbol{\beta}$, τ, $\boldsymbol{\delta}$, and λ from their full conditionals.

Repeat steps 1-4 for a large number of iterations, collecting samples after a burn-in to allow convergence.

A challenging aspect of the algorithm is comparing models in the RJMCMC sampler. Our prior assigns equal probability to all piecewise linear models and model proposal is based on generation of discrete random variables. Under this scenario, the probability, Q, of accepting a proposed model, M^*, is the Bayes factor comparing it to the current model, M (Holmes and Mallick, 2003, Denison et al., 2002). The Bayes factor is the ratio of the marginal likelihoods of the data under the two models:

$$Q = min\Big[1, \frac{p(\mathbf{y}|M^*)}{p(\mathbf{y}|M)}\Big].$$

The marginal likelihoods and thus the Bayes factor for this hierarchical model have no closed form. Consider instead the following marginal likelihood under model M.

$$p(\mathbf{y}|M, \boldsymbol{\delta}, \lambda) = \int\int\int L(\mathbf{y}|\mathbf{b}, \tau, \lambda, M) p(\mathbf{b}, \tau, \boldsymbol{\beta}|\boldsymbol{\delta}, \lambda, M)\, d\mathbf{b}\, d\boldsymbol{\beta}\, d\tau,$$

where $p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b}, \tau, \boldsymbol{\delta}, \lambda, M)$ is the data likelihood under model M, and $p(\mathbf{b}, \tau, \boldsymbol{\beta}|\boldsymbol{\delta}, \lambda, M)$ is the joint prior of $\mathbf{b}$, $\boldsymbol{\beta}$, and τ under model M. This integral has a closed form, so

that the likelihood can be written:

$$p(\mathbf{y}|M,\boldsymbol{\delta},\lambda) = C(\lambda,k)|\mathbf{R}|^{-\frac{1}{2}}(b_\tau + \frac{\alpha}{2})^{-(\frac{n}{2}+a_\tau)}\prod_{l=1}^{k}\delta_l^{\frac{m}{2}}\prod_{i=1}^{m}|\mathbf{U}_i|^{\frac{1}{2}} \tag{2.5}$$

where

$$\mathbf{U}_i = [\Delta + \boldsymbol{\theta}_i'\boldsymbol{\theta}_i]^{-1}$$

$$\mathbf{R} = \lambda\mathbf{I}_k + m\boldsymbol{\Delta} - \boldsymbol{\Delta}(\sum_{i=1}^{m}\mathbf{U}_i)\boldsymbol{\Delta}$$

$$\alpha = \mathbf{y}'\mathbf{y} - \sum_{i=1}^{m}\mathbf{y}_i'\boldsymbol{\theta}_i\mathbf{U}_i\boldsymbol{\theta}_i'\mathbf{y}_i - (\sum_{i=1}^{m}\mathbf{U}_i\boldsymbol{\theta}_i'\mathbf{y}_i)'\boldsymbol{\Delta}\mathbf{R}^{-1}\boldsymbol{\Delta}(\sum_{i=1}^{m}\mathbf{U}_i\boldsymbol{\theta}_i'\mathbf{y}_i)$$

$$C(\lambda,k) = \frac{{b_\tau}^{a_\tau}\lambda^{\frac{k}{2}}\Gamma(\frac{n}{2}+a_\tau)}{\Gamma(a_\tau)(2\pi)^{\frac{n}{2}}}$$

In a similar fashion, we can write the marginal likelihood for a proposed model M^* of dimension k^*.

$$p(\mathbf{y}|M^*,\boldsymbol{\delta}^*,\lambda^*) = C(\lambda^*,k^*)|\mathbf{R}^*|^{-\frac{1}{2}}(b_\tau + \frac{\alpha^*}{2})^{-(\frac{n}{2}+a_\tau)}\prod_{l=1}^{k*}{\delta_l^*}^{\frac{m}{2}}\prod_{i=1}^{m}|\mathbf{U^*}_i|^{\frac{1}{2}} \tag{2.6}$$

Suppose we propose a move from model M of dimension k to model M^* of dimension k^*. If we let the acceptance probability be the ratio of the two marginal likelihoods, then it depends on λ and $\boldsymbol{\delta}$. It also depends on λ^* and $\boldsymbol{\delta}^*$, for which we do not have estimates. Since we wish to accept or reject a model based only on its set of basis functions, we want to minimize the effects of these variance components on the acceptance probability. Specifically, we assume λ=λ^* at the current sampled value. Since $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ may be of different dimensions, we cannot assume that they are equal. Instead, we assume that they are equal in the elements corresponding to bases common to both models and condition only on those elements.

Consider a proposal to add a basis to the current model. The current model is nested in the proposed model, and the proposed model has exactly one more basis than the current model. The acceptance probability is:

$$Q = min\Big[1, \frac{p(\mathbf{y}|M^*, \lambda, \boldsymbol{\delta})}{p(\mathbf{y}|M, \lambda, \boldsymbol{\delta})}\Big]$$

The denominator has closed form, as we've shown above, and the numerator can be derived as follows, where $\boldsymbol{\delta}^* = (\boldsymbol{\delta}, \delta_{k^*})$.

$$p(\mathbf{y}|M^*, \lambda, \boldsymbol{\delta}) = \int p(\mathbf{y}, \delta_{k^*}|M^*, \boldsymbol{\delta}, \lambda) d\delta_{k^*} = \int p(\mathbf{y}|M^*, \boldsymbol{\delta}^*, \lambda)\pi(\delta_{k^*}) d\delta_{k^*}$$

$$= \frac{C(\lambda, k^*)}{\Gamma(a_\delta)} \prod_{l=1}^{k} \delta_l^{\frac{m}{2}} \int_0^\infty |\mathbf{R}^*|^{-\frac{1}{2}} (b_\tau + \frac{\alpha^*}{2})^{-(\frac{n}{2}+a_\tau)} {\delta_{k^*}}^{a_\delta + \frac{m}{2}} exp(-b_\delta \delta_{k^*}) {b_\delta}^{a_\delta} \prod_{i=1}^{m} |{\mathbf{U}^*}_i|^{\frac{1}{2}} d\delta_{k^*} \tag{2.7}$$

This integral is complicated, and we approximate it using the Laplace method. This involves fitting a scaled normal density to the integrand. Specifically, if we wish to evaluate $\int h(\theta) d\theta$, we assume that $h(\theta) \approx h(\hat{\theta}) exp(\frac{-(\theta-\bar{\theta})^2}{2\sigma^2})$, where $\bar{\theta}$ is the mode of $h(\theta)$ and $\hat{\sigma}^2$ is the estimate of the variance of the normal density. A good estimate of the mode, $\hat{\theta}$, can be obtained with a numerical search algorithm. The variance can be estimated by noting that $\frac{h(\hat{\theta})}{h(\hat{\theta}+\epsilon)} \approx exp(\epsilon^2 2\sigma^2)$. We evaluate h at $(\hat{\theta} + \epsilon)$ and $(\hat{\theta} - \epsilon)$ and average the two resulting estimates of σ^2 to get $\hat{\sigma}^2$. The integral is then approximated by $(2\pi)^{\frac{1}{2}} (\hat{\sigma})^{\frac{1}{2}} h(\hat{\theta})$. For additional information on the Laplace method and other methods for Bayes factor approximation, see DiCiccio et al. (1997).

Since the integral we want to approximate is defined over $\Re^+$ and the normal distribution is defined over the entire real line, we will transform δ_{k^*}. Simulations show that this has the added benefit of making the integrand more symmetric. Let $\omega = log(\delta_{k^*})$

and note that the prior on ω_{k^*} is:

$$\pi(\omega_{k^*}) = \frac{exp(a_\delta \omega - b_\delta[exp(\omega)]){b_\delta}^{a_\delta}}{\Gamma(a_\delta)}$$

The integral in (2.7) can be written:

$$p(\mathbf{y}|M^*, \boldsymbol{\delta}, \lambda) = \int p(\mathbf{y}, \omega|M^*, \boldsymbol{\delta}, \lambda)d\omega = \int_{-\infty}^{\infty} p(\mathbf{y}|M^*, \boldsymbol{\delta}, \omega, \lambda)\pi(\omega)d\omega$$

$$= \frac{C(\lambda, k^*)}{\Gamma(a_\delta)} \prod_{l=1}^{k} \delta_l^{\frac{m}{2}} \int exp(\omega(a_\delta + \frac{m}{2}) - b_\delta[exp(\omega)])|\mathbf{R}^*|^{-\frac{1}{2}}(b_\tau + \frac{\alpha^*}{2})^{-(\frac{n}{2}+a_\tau)} \prod_{i=1}^{m} |\mathbf{U}_i^*|^{\frac{1}{2}} d\omega$$

Similarly, a basis removal proposal involves integrating out the element of $\boldsymbol{\delta}$ corresponding to the basis proposed for removal. A proposal to alter a basis involves integrating out the element of $\boldsymbol{\delta}$ corresponding to that basis in both the numerator and the denominator.

2.2.3 Computation

In implementing the RJMCMC algorithm described above, we run a burn-in period of several thousand iterations until convergence is apparent. Convergence is evidenced by the stationarity of the distribution of the marginal likelihood in (2.5) and the distribution of k, the dimension of sampled models. Then the sampler is run for an additional period, during which each selected piecewise linear function is saved. Final estimates of the population regression function are based on averages over all the saved models, and credible intervals for the response can be calculated for any set of covariate values. In addition, the subject-specific coefficients are saved at each step, so that the individual regression function can be estimated and individual credible intervals can be calculated.

The analysis is conducted using Matlab version 7.0.1. The method is computationally intensive, especially for large datasets. However, the rates of convergence and

mixing are good enough that it can be practically implemented even in complex settings, such as that described in the data example.

2.3 Simulated data example

The simulated data do not mimic longitudinal data with reference points. Rather, we illustrate the broad applicability of the method by simulating clustered data with a covariate-dependent random effect. We simulated data for 200 subjects, with each subject contributing 30 observations from the following distribution:

$$(y_{ij}|\mathbf{x}_{ij}) \sim N\Big(x_{1ij} - x_{2ij}^2 + x_{1ij}x_{2ij} + b_i\sqrt{2|x_{1ij}|},\ 2\Big)$$

where the covariates x_{1ij} and x_{2ij} for the j^th observation from subject i are randomly generated integers between -4 and 4, and b_i is a $N(0,1)$ random term for subject i. Note that the random effect varies non-linearly with x_1. We want the method to be able to detect this variation. In addition, the model-estimated population mean, subject-specific means, and random effects should be consistent with the simulated data.

We ran the RJMCMC algorithm for 50,000 iterations, discarding the first 10,000 as burn-in. In the first chain, the hyperparameters a_τ, b_τ, a_λ, b_λ, a_δ and b_δ were all set to 0.05, yielding vague priors for the variance components. When proposals were accepted, new elements of $\boldsymbol{\beta}$ were initialized to 0. Sensitivity to hyperparameters and initial values was assessed through an additional chain where $a_\tau, a_\lambda, a_\delta = 1$, $b_\tau, b_\lambda, b_\delta = 0.5$, and the new elements of $\boldsymbol{\beta}$ were initialized to 1. The two chains yielded virtually identical results. This suggests that the method is not overly sensitive to specification of initial values and hyperparameters.

We calculated subject-specific estimates for each data point as well as population predictions over the covariate space. Figure 2.2 illustrates the model's ability to discern

features of the data. Figure 2.2a shows a scatterplot of the population mean values estimated under the algorithm against the true mean values for each covariate combination. This indicates that the model was able to distinguish the underlying population mean structure from the random effects. The empirical estimates of the random effects were calculated by subtracting the model-predicted population mean from the subject-specific posterior mean for each data point. As shown in Figure 2.2b, the empirical estimates of the random effects were generally accurate estimates of the true values of the random effects, $\{x_1^2 b_i\}$. At each iteration, the estimated variance under the current model for each set of covariate values was calculated:

$$V_e(y|x_1, x_2) = \delta_0^{-1} + \sum_{l=1}^{k-1} \delta_l^{-1} (\mathbf{x}'\boldsymbol{\mu}_l)_+^2 + \tau^{-1}$$

where $\boldsymbol{\delta}$ and τ are the estimates of the variance components under the current k-dimensional model. The empirical variance estimate can be compared to the true variance:

$$V(y|x_1, x_2) = |x_1| + 2$$

Figure 2.2c shows the average over all samples of the empirical variance at each covariate pair plotted against the true variance. The model-estimated values pick up the general trend of the true values, but there seems to be a tendency toward slight underestimation.

Figure 2.2d is a traceplot of the model marginal likelihood (2.5) over the sampled iterations. The distribution of this quantity, and of the associated predictions, appears to be stationary, so we find no evidence against convergence of the MCMC algorithm.

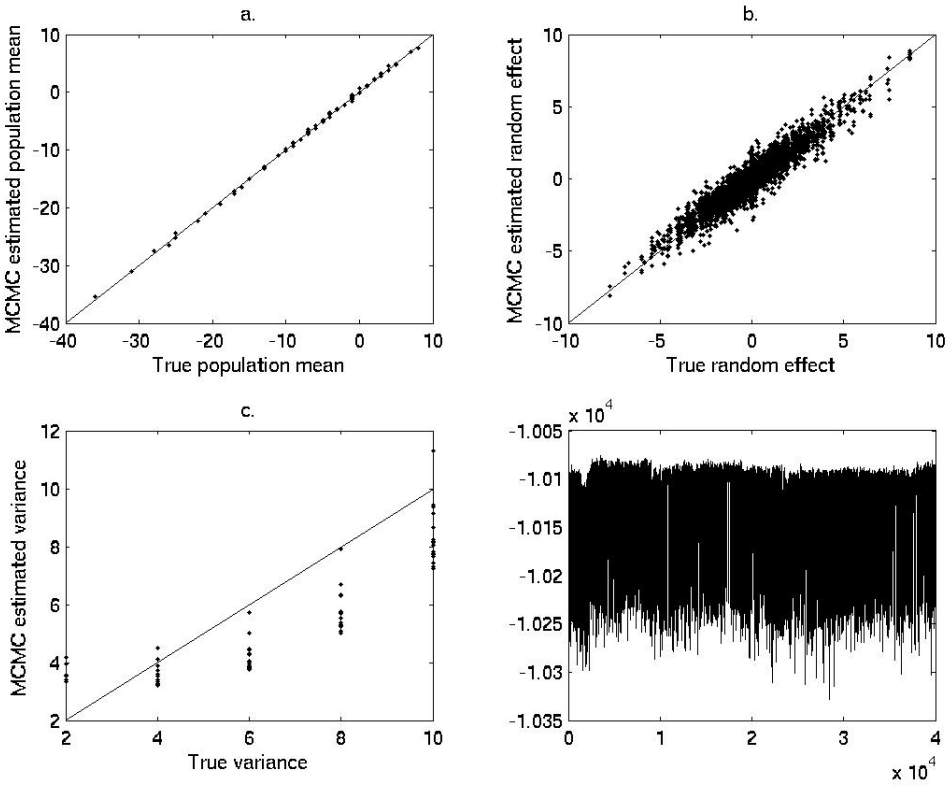

FIGURE 2.2: Plots used to evaluate of algorithm performance using simulated data

2.4 Progesterone example

2.4.1 Estimation

We applied these methods to the progesterone data from the NCEPS described in Section 2.1 with the goal of assessing differences in PdG profiles between conception and non-conception cycles. We were particularly interested in examining differences prior to implantation, since these may indicate hormonal effects on fecundability and conception probabilities.

We apply the approach described in 2.2 with three covariates and an intercept. The first covariate is an indicator of whether the cycle during which the measurement was taken resulted in conception. The final two covariates contain the reference point information. They are number of days since cycle start (onset of menses) and number of days relative to ovulation in current cycle. So if response y_{ij} was observed on the third day of a non-conception cycle where ovulation occurred on day 14, then $\mathbf{x}_{ij} = (1, 0, 3, -11)'$.

Vague priors on the variance components were achieved by setting the hyperparameters to 0.05. We collected 40,000 MCMC samples after a 20,000 iteration burn-in.

2.4.2 Inference

We can use model estimates to assess the relationship among progesterone, cycle conception status, and the two reference points. The main analysis goal was to gain a better understanding of the differences in progesterone between conception and non-conceptions cycles. At each iteration, we calculate several summary variables for each cycle based on the trajectories estimated by the subject-specific coefficients. Early follicular PdG was the average over the first 5 days of the cycle, baseline PdG was the average from 6 days until 2 days before ovulation, and midluteal PdG was the average

on days 5 and 6 after ovulation. The early luteal PdG rise was the change in PdG from 1 day to 5 days after ovulation. We record the mean of each of these variables for conception and non-conception cycles at each iteration, using these samples to create overall means and credible intervals.

Baird et al. (1999) suggested that conception was less likely in cycles with low midluteal PdG. To test this, we find the 10^{th} percentile of midluteal PdG over all cycles at each iteration and record the proportion of cycles that are conceptions both under and over this threshold. In the process, we obtain posterior means and 95% pointwise credible intervals for the population PdG trajectory for conception and non-conception cycles at any location relative to the reference points.

2.5 Results

Convergence was deemed adequate, as the distribution of the marginal likelihood appeared to be stationary. In addition, the distribution of the dimension of sampled models was stationary. Sample collection took approximately 72 hours.

Figure 2.3 displays data from a single subject, the fitted PdG curve, and the predicted population mean log-PdG given the woman's covariates. The subject-specific curve captures the subject's data more closely than the population curve, illustrating the potential for a shape difference between the population-mean and subject-specific curves.

Table 2.1 gives the estimated differences in log-PdG between conception and non-conception cycles for the intervals of interest. Early follicular PdG over the first five days of the cycle tended to be higher in non-conception cycles. In addition, non-conception cycles tended to have higher baseline and slightly higher midluteal PdG than conception cycles. There was a larger average early luteal PdG rise in conception

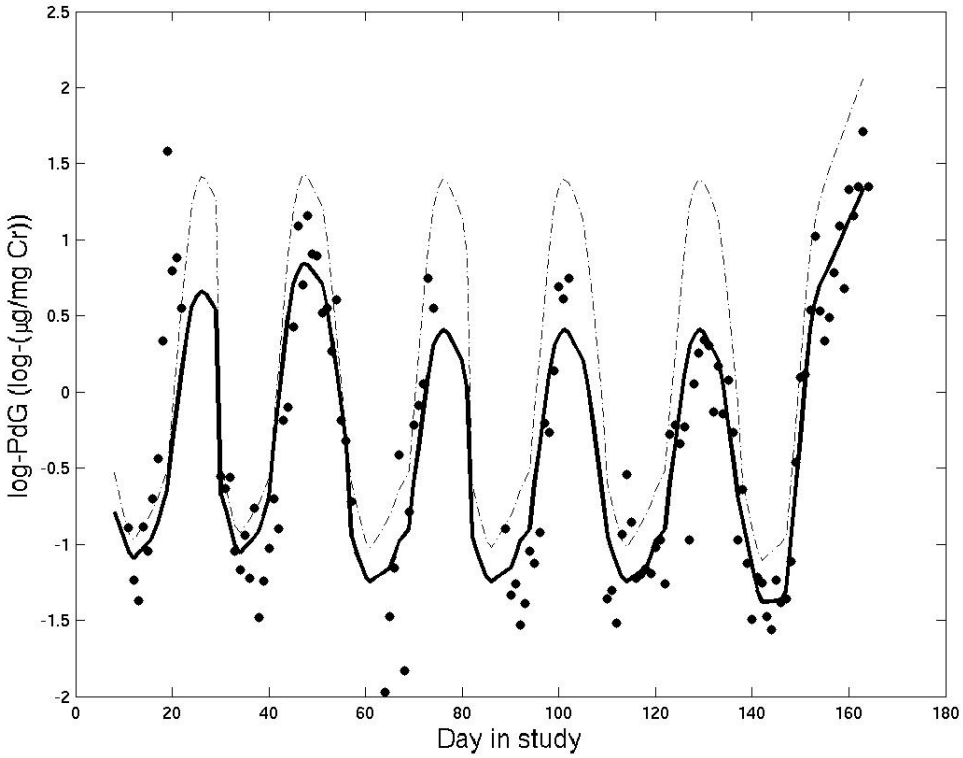

FIGURE 2.3: log(PdG) data (points) and estimated log(PdG) (solid line) for a single woman. The dashed line is the estimated population mean log-PdG given her covariates.

TABLE 2.1: Summary variables describing the conception vs. non-conception difference in log(PdG) across the menstrual cycle with 95% credible intervals. Estimates are based on an average of subject-specific trajectories at each iteration.

	log-PdG		
	Conception	Non-conception	Difference
Early follicular	-0.94 [-0.98, -0.90]	-0.64 [-0.67, -0.61]	-0.30 [-0.35,-0.25]
Baseline	-0.94 [-0.97, -0.91]	-0.78 [-0.80, -0.77]	-0.16 [-0.19,-0.12]
Midluteal	1.19 [1.14, 1.24]	1.31 [1.28, 1.35]	-0.13 [-0.18,-0.07]
Early luteal rise	1.18 [1.10, 1.26]	1.12 [1.06, 1.18]	0.07 [-0.05,0.15]

TABLE 2.2: Probability of conception in cycles with very low vs. normal/high midluteal (days 5-6 after ovulation) PdG, with 95% credible intervals.

	Estimate, 95% CI
Probability of conception, midluteal PdG $< 10^{th}$ percentile	0.144 [0.098,0.195]
Probability of conception, midluteal PdG $\geq 10^{th}$ percentile	0.217 [0.211,0.222]
Difference in conception probabilities	0.073 [0.016,0.124]

cycles, though the 95% credible interval for the difference includes zero.

Table 2.2 summarizes the relationship between conception status and low midluteal progesterone, with 95% credible intervals. Those cycles with estimated midluteal PdG in the lowest 10% were less likely to be conception cycles than those with higher PdG, although in Table 2.1 we saw that non-conception cycles had higher midluteal PdG on average.

These results have been based on the subject-specific basis coefficients only, and we now examine population progesterone curves. Figure 2.4 displays the predicted population mean log-progesterone for the first 28 days of a conception and non-conception cycle with ovulation on day 14. It is apparent from this figure, and from examination of similar figures with a range of alternative ovulation days, that progesterone rises following ovulation in conception cycles, but peaks and then drops in non-conception cycles. This result is consistent with the biological role of progesterone and with previous findings of Baird et al. (1997). In addition, these population curves support our findings that conception cycles tend to have lower pre-ovulatory progesterone.

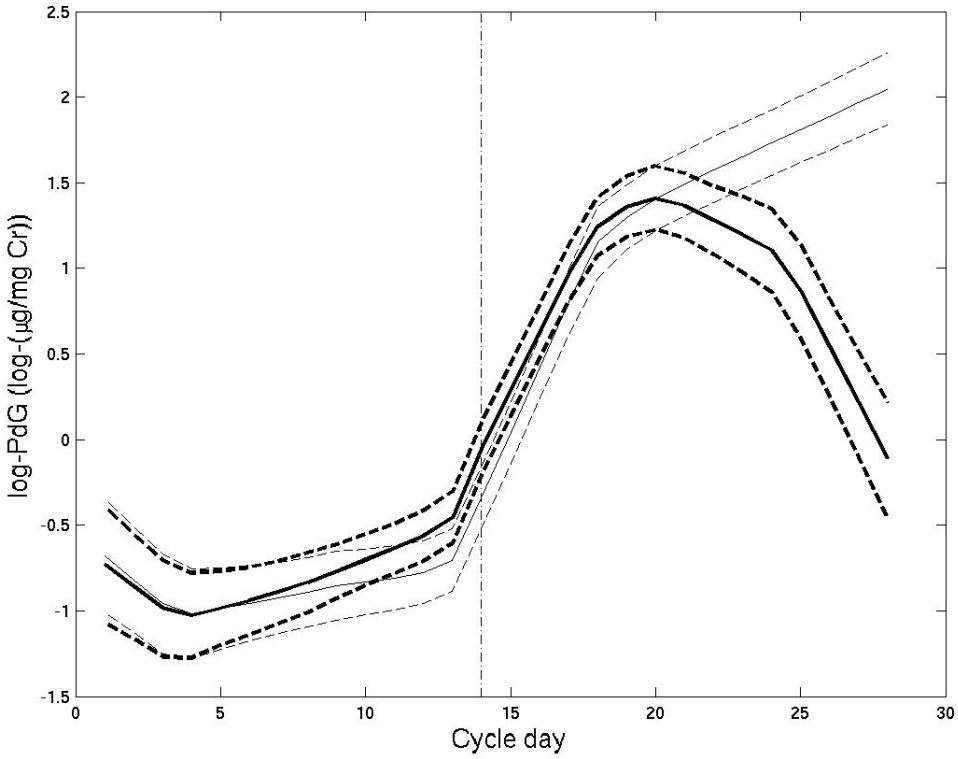

FIGURE 2.4: Estimated population mean log(PdG) for a conception (thin line) and non-conception (heavy line) cycle with ovulation on day 14. Pointwise 95% credible intervals are given by the dashed lines.

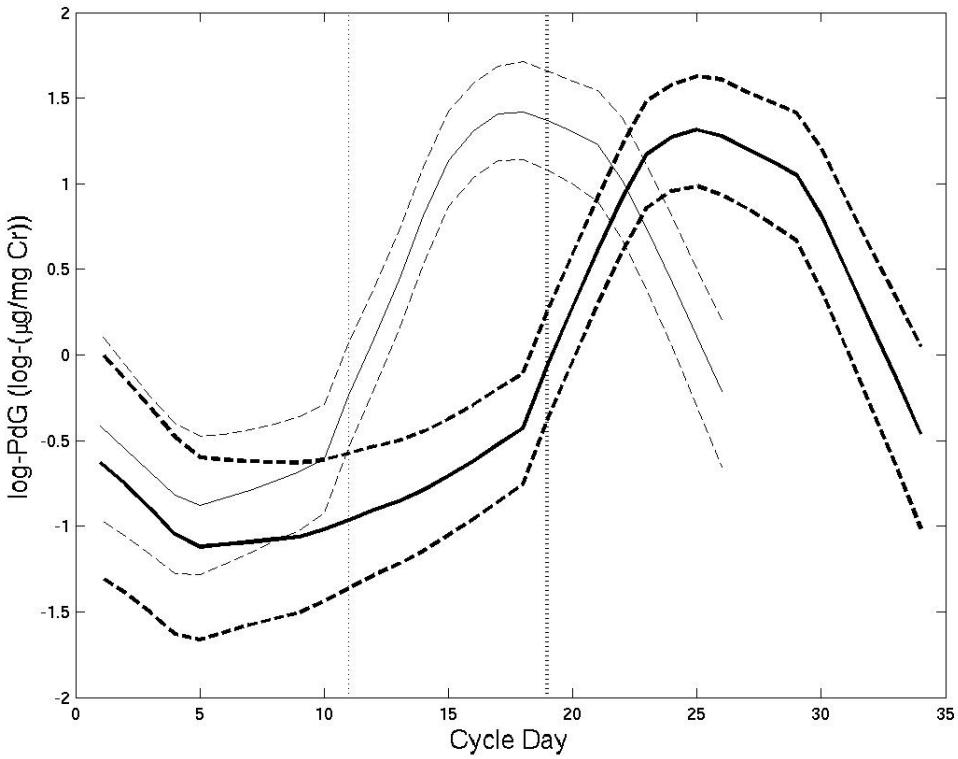

FIGURE 2.5: Estimated population mean log(PdG) for non-conception cycles with ovulation on day 10 (thin line) and day 18 (heavy line). Pointwise 95% credible intervals are given by the dashed lines. Vertical lines indicate ovulation days.

Figure 2.5 shows the population-average progesterone curves for non-conception cycles when ovulation occurred on the 10^{th} day of the cycle (early) and on the 18^{th} day of the cycle (late). The estimated curves are different, indicating that the model was adequate in discerning the effect of the timing of ovulation on progesterone. The fact that the peak occurs earlier when ovulation occurs earlier is consistent with previous findings about the relationship between progesterone and ovulation (Baird et al., 1997).

Finally, we examined the adequacy of the Laplace approximation to the marginal likelihood. Twenty model proposals were selected at equally-spaced intervals through-

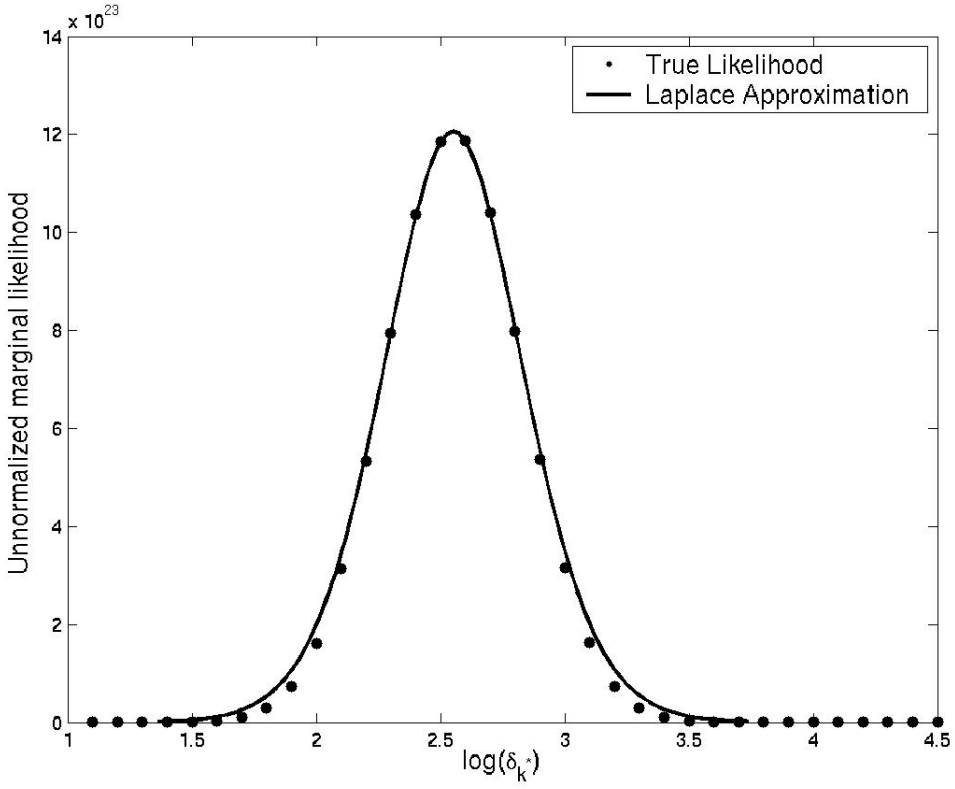

FIGURE 2.6: The unnormalized marginal likelihood for a proposed model, $p(\mathbf{y}|\lambda, \boldsymbol{\delta}, M^*)$ and its corresponding Laplace approximation.

out the sampling period, and a plot of each true unnormalized marginal likelihood was compared to the scaled normal approximation. The fit was found to be quite good, although in general the approximation tends to have slightly fatter tails than the true likelihood. Figure 2.6 displays the likelihood from a birth proposal and its Laplace approximation.

2.6 Discussion

We proposed Bayesian regression with multivariate linear splines for hierarchical data. This is an extension of the method for independent responses (Holmes and Mallick,

2001) to include subject-specific basis coefficients assumed to be centered around the population coefficients for each of the sampled models.

A different method was developed independently by Verzilli et al. (2005) for predicting the functional consequences of amino acid polymorphisms. Their approach also relies on Bayesian multivariate adaptive regression splines, though they accommodate within-cluster dependency using a simple cluster-specific random intercept. A random intercept is not flexible enough to accommodate the variability in hormone trajectories, motivating our use of a general hierarchical structure for the basis coefficients.

Analysis of the NCEPS data has yielded new insight about the relationship between progesterone and cycle conception status. It has been speculated that very low midluteal PdG may be indicative of a low fertility cycle, and also that signals from the conceptus may promote a pre-implantation increase in progesterone. Our results support both of these hypotheses, as we found evidence for a slightly steeper post-ovulatory PdG rise in conception cycles.

Previous analyses of these and other data (Baird et al., 1997; Stewart et al., 1993) found that non-conception cycles have lower midluteal progesterone than conception cycles, but we found the opposite. However, these previous studies examined non-conception cycles from women of known fertility who were exposed to sperm during a potentially fertile phase of the cycle (either through intercourse or artificial insemination). These non-conception cycles were therefore likely to be of low fertility. A previous analysis of these data using non-conception cycles regardless of intercourse timing found that cycles with low midluteal progesterone were unlikely to be conception cycles (Baird et al., 1999). We found the same, but we also found that, on average, midluteal progesterone was higher in non-conception cycles.

In light of these previous results, our findings suggest that those cycles with very low progesterone are of low fertility, but that high pre-ovulatory progesterone does not

imply an increased probability of conception. Intercourse timing was not used as a covariate here, but it may be informative in future analyses to explicitly differentiate non-conception cycles that were due to lack of intercourse from those that were of low fertility.

The method was applied to longitudinal data, but it could be used in any hierarchical regression problem where the functional form of the relationship between the covariates and the response is unknown. The NCEPS data lends itself readily to a discussion of the incorporation of reference points, but this method is also appropriate for regression when there are no reference points of interest. In this sense, the regression model is widely applicable.

In addition, reference points are not unique to longitudinal data. Brumback and Lindstrom (2004) use reference points to line up features of speech pattern data. Functional data can also occur over space (Morris et al., 2003), in which case the reference points are spatial rather than temporal locations. Rice (2004) discusses the similarities among modeling longitudinal and other types of functional data. Often, the analysis goals are the same, and methods designed for one tend to apply to both. This method is readily applicable to hierarchical functional data such as that studied by Morris et al. (2003).

The Bayesian RJMCMC paradigm allowed estimation of a smooth function based on piecewise-linear models with unknown knots and estimation of the population regression function based on the subject-specific basis coefficients. Although we used piecewise linear splines for their interpretability, this methods could be applied with other basis sets (see Denison et al., 2002 for a discussion).

CHAPTER 3

BAYESIAN SEMIPARAMETRIC CLASSIFICATION OF FUNCTIONAL DATA

Motivated by the problem of classifying hormone trajectories, we propose a flexible semiparametric Bayesian methodology for hierarchical functional data. The approach is based on a hierarchical spline model, with the number and location of knots and the distribution of the random spline coefficients treated as unknown. Assuming a discrete distribution for the spline coefficients, we obtain a procedure that clusters trajectories into classes, with the class-specific trajectories, the number of classes, and the allocation of subjects to classes unknown. The procedure relies on a generalization of the Dirichlet process to a collection of unknown distributions having varying dimension. We develop an efficient reversible jump Markov chain Monte Carlo algorithm by constructing dependence within this collection of distributions. The methods are illustrated using progesterone trajectories through the human menstrual cycle.

3.1 Introduction

Latent trajectory models (LTMs) have wide applications in biology and the social sciences (Legler et al., 2004; see Curran and Hussong, 2003 for a review). Most commonly applied to longitudinal studies, these models treat the data as noisy realizations from some underlying trajectory. Rather than modeling a distinct trajectory for each sampling unit, sampling units can be classified into groups based on trajectory shape. By using a nonparametric Bayesian method for the assignment of subjects to clusters, while also allowing cluster-specific trajectories to be unknown, we obtain a flexible methodology for nonparametric modeling and clustering of hierarchical functional data. By using multivariate adaptive regression splines, with unknown numbers and locations of knots, the method can also deal flexibly with covariates having complex interactions.

Our method is motivated by reproductive hormone data from the North Carolina Early Pregnancy Study (EPS) (Wilcox et al., 1988; Baird et al., 1997). The data consist of daily progesterone measurements in women who are trying to become pregnant. These data are presented as trajectories, one for each menstrual cycle. Progesterone and other reproductive hormones vary across the menstrual cycle, and modeling the underlying trajectory is complex. Cycles tend to vary in length and in the timing of ovulation, which is often closely related to the trajectory shape. We wish to identify typical clusters of hormone trajectories and determine which, if any, tend to be predictive of high or low fertility. This example is one special case of hierarchical functional data. Methods for hierarchical functional data typically require that all curves are observed over or standardized to fall in the same region of time (Brumback and Rice, 1998; Morris et al., 2003; Brumback and Lindstrom, 2004).

The dependency structure of these data is complex, as there is a longitudinal dependence within a trajectory (daily measurements from each cycle) in addition to multiple trajectories (cycles) from each independent sampling unit. We develop a method where

the subject-specific random effects are clustered into groups. For ease of presentation, we restrict attention to one cycle per woman. Thus, clustering the women is equivalent to clustering curves, and a graphical representation is straightforward.

To accommodate the dependency structure of hierarchical functional data without requiring truncation of data or standardization of time, we developed a hierarchical generalization of multivariate linear splines in Chapter 2. We used a Bayesian reversible jump MCMC (RJMCMC) algorithm to collect a large number of plausible piecewise-linear spline models with varied numbers and locations of knots. The models were summarized into one smooth regression function through Bayesian model averaging. Between-subject variation was accounted for by assigning each subject a set of random spline coefficients which are normally distributed around some population mean. Subjects with similar trajectories should tend to have similar random coefficients, so classification in terms of these random effects may be appropriate.

Verbeke and Lesaffre (1996) demonstrate that accurate estimation of random effects can be hindered by incorrect specification of the random effects distribution. Clustering based on random effects relies on good estimates of these effects, so specifying an overly restrictive distribution may result in misleading clustering. In addition, the often-specified normal distribution assumes no identical values, thus doesn't produce clusters of identical curves. An approach is needed in which the number of distinct curves is smaller than the sample size. The normal distribution is also not biologically motivated in this case, because menstrual cycle hormone curves tend not to vary smoothly and regularly about a mean function.

In general, the field of cluster analysis examines data for groups of observations that are similar to one another and dissimilar to the others. Here, our observations of interest are trajectories. Identifying clusters is a common goal in data mining (Huang, 1998), especially in large microarray datasets (Medvedovic and Sivaganesan, 2002; Tseng and

Wong, 2005). Comprehensive reviews of clustering methods are given in Fraley and Raftery (2002) and Denison et al. (2002).

There have been several methods proposed for clustering curve data. James and Sugar (2003) use a fixed set of basis functions to describe curves with random effects for the coefficients, fitting the model with an EM algorithm. They use a method developed in Sugar and James (2003) to find the most appropriate number of clusters based on a "distortion function". This approach is not designed for the case where the underlying spline model varies. Ma et al. (2005) examine gene expression profiles over a fixed time interval, dividing the profiles up into three regions and fitting a polynomial spline in each region. Then they use an EM algorithm to fit the model and BIC to compare models and determine the best number of clusters.

James and Hastie (2001) apply linear discriminant analysis to classify functional data, where the number of classes is pre-determined. de la Cruz and Quintana (2005) use Bayesian methods and Marshall and Barón (2000) developed a mixed effects model for classification of hormone trajectories into pre-defined groups. With the goal of identifying Olympic athletes who use growth hormone injections, Brown et al. (2001) developed a Bayesian method which defines trajectory classes based on a training dataset with known classification. Muthén and Shedden (1999) use an EM algorithm to identify trajectories in young adult drinking behavior that are likely to lead to alcohol dependence. Unlike these methods, we wish to allow for both non-parametric specification of random effects and the underlying spline model and an unspecified number of clusters.

Our approach relies on a hierarchical spline model. The spline model is defined by a set of basis functions, and within-woman dependence is achieved by allowing each woman her own set of random spline coefficients. The distribution of the random spline coefficients is treated as unknown and discrete. This structure automatically groups subjects into clusters having identical basis coefficients, and hence similar trajectories.

If the spline basis functions were known, then a Dirichlet process prior (DPP) (Ferguson, 1973; 1974) could be used for the unknown distribution of the basis coefficients. Unfortunately, even though the number of clusters and the allocation of subjects to clusters would be flexible, the cluster-specific trajectories would be limited in flexibility by the particular basis functions chosen. Hence, it would be very difficult to capture the great variety of trajectories in hormones that are observed. It is therefore desirable to allow uncertainty in the spline basis functions, including the numbers and locations of knots. We are then faced with the problem of nonparametric modeling of the distribution of the spline coefficients when the set of basis functions is unknown and of unknown dimension.

Motivated by this problem, we develop an approach for modeling a collection of unknown distributions having varied dimension. Starting with a DPP for one particular distribution in the collection, we build dependence by sharing the prior on the clustering behavior across distributions. The dependence is such that each distribution in the collection marginally follows a DPP. This construction is conceptually similar to the dependent DP of MacEachern (1999; 2001) and De Iorio et al. (2004), though they did not consider the case in which the dimension varies. Our approach greatly facilitates posterior computation, and we develop an efficient Markov chain Monte Carlo (MCMC) algorithm that combines reversible jump (Green, 1995) with the Pólya urn Gibbs sampler (West et al., 1994; MacEachern, 1994; Ishwaran and James, 2001).

3.2 Methods

3.2.1 Multivariate linear splines

Suppose we have vector observations from N trajectories that we wish to classify according to shape. The ability to do this first relies on a flexible and accurate regression

model for each trajectory. Holmes and Mallick (2001) propose Bayesian regression with multivariate linear splines (see also Hansen and Kooperberg, 2002; Wood et al., 2002; Holmes and Mallick, 2002) for scalar response data. Chapter 2 generalized this approach to hierarchical functional data.

Although the method we propose could be used with a variety of spline bases, we choose to model the true relationship between covariates and response as piecewise linear (see Holmes and Mallick, 2001 for a discussion of multivariate linear spline models). We have no knowledge of the most appropriate piecewise linear model to use. In fact, the underlying trajectories may be smooth curves that are unlikely to be well-represented by any one piecewise linear model. Instead, we wish to consider models from some model space $\mathcal{M}$, with varying numbers and locations of knots. We use the RJMCMC algorithm of Green (1995) to add and remove knots, sampling models from $\mathcal{M}$ having high posterior probability. The final curve estimates are weighted averages over all sampled models.

A single multivariate piecewise linear model, $M \in \mathcal{M}$, is defined by a set of k_M basis functions, $\boldsymbol{\mu}_M = (\boldsymbol{\mu}_{M1}, \ldots, \boldsymbol{\mu}_{Mk_M})$. We consider a regression setting where y_{ij} is the j^{th} response from subject i, $i = 1, \ldots, N$; $j = 1, \ldots n_i$. The relationship between y_{ij} and its $(p \times 1)$ set of covariates $\mathbf{x}_{ij}$ can be approximated by a linear combination of the positive portions (denoted by the + subscript) of the inner products of the basis functions with the covariate vector:

$$y_{ij} = \sum_{l=1}^{k_M} b_{Mil}(\mathbf{x}'_{ij}\boldsymbol{\mu}_{Ml})_+ + \epsilon_{Mij}, \quad M \in \mathcal{M} \tag{3.1}$$

where ϵ_{Mij} is a random error. More transparently, each piecewise linear model is linear in the basis function transformations of the covariate vectors:

$$\mathbf{y}_i = \mathbf{H}_{Mi}\mathbf{b}_{Mi} + \boldsymbol{\epsilon}_{Mi}, \quad M \in \mathcal{M} \tag{3.2}$$

where $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_{Mi}$ are the $n_i \times 1$ vectors of responses and random errors, $\mathbf{b}_{Mi}$ is the $k_M \times 1$ vector of basis coefficients for subject i, and the design matrix $\mathbf{H}_{Mi}$ contains the basis function transformations of the covariate vectors for subject i.

Assuming conditional independence of the elements of $\mathbf{y}_i$ given $\mathbf{b}_{Mi}$, and $N(0, \tau_M^{-1})$ errors, the conditional likelihood under model M is:

$$p(\mathbf{y}|\boldsymbol{b}_M, \tau_M, M) \propto \tau_M^{\frac{n}{2}} \prod_{i=1}^{N} exp\big[-\frac{\tau_M}{2}(\mathbf{y}_i - \mathbf{H}_{Mi}\mathbf{b}_{Mi})'(\mathbf{y}_i - \mathbf{H}_{Mi}\mathbf{b}_{Mi})\big] \;\; M \in \mathcal{M} \quad (3.3)$$

where $n = \sum_{i=1}^{N} n_i$. Continuing Bayesian specification of the model, we put a prior on $\boldsymbol{b}_M = (\mathbf{b}_{M1}, \ldots, \mathbf{b}_{MN})$:

$$\mathbf{b}_{Mi} \overset{iid}{\sim} G_M, \;\; i = 1, \ldots, N; \;\; M \in \mathcal{M} \quad (3.4)$$

The distribution G_M could be given some parametric form for all $M \in \mathcal{M}$. For example, in Chapter 2 we specified G_M to be Gaussian, which implies that all subject-specific coefficients are normally distributed around some population mean. In the quest to uncover clusters of similar trajectories, this normality assumption makes little sense. Instead, we treat G_M as an unknown distribution having a DPP, for all $M \in \mathcal{M}$.

3.2.2 Dirichlet process

Thie Dirichlet process has been used to relax distributional assumptions on random effects (Bush and MacEachern, 1996; Mukhopadhyay and Gelfand, 1997; Kleinman and Ibrahim, 1998; Ishwaran and Takahara, 2002). We develop a method for the use of the DP to relax distributional assumptions when the random effect is of varying dimension, taking advantage of the natural clustering behavior of the DP in our analysis.

From the Sethuraman (1994) stick-breaking representation of the DP, $G_M \sim DP(\alpha_M G_{M0})$

can be written as an infinite mixture of point masses:

$$G_M = \sum_{h=1}^{\infty} w_{Mh}\delta_{\theta_{Mh}} \tag{3.5}$$

where δ_θ is the degenerate distribution placing all its mass at θ, $\{w_{Mh} : h = 1, \ldots, \infty\}$ is an infinite sequence of random weights generated from a stick-breaking process, $\frac{w_{Mh}}{\prod_{l=1}^{h-1}(1-w_{Ml})} \sim \text{beta}(1, \alpha_M)$, and $\{\theta_{Mh}, h = 1, \ldots, \infty\}$ is a corresponding sequence of random atoms generated independently from the base distribution G_{M0}.

The stick-breaking representation shows that if G_M follows a Dirichlet process, then G_M is almost surely discrete. This discreteness implies that, since $\mathbf{b}_{Mi} \overset{iid}{\sim} G_M$, the subjects $i = 1, \ldots, N$ will be clustered into $r_M \leq N$ groups (Antoniak, 1974; Escobar, 1994). The clustering behavior is more transparent from the Blackwell and MacQueen (1973) Pólya urn scheme, which shows that the specified prior structure corresponds to the following set of conditional priors for the elements of $\mathbf{b}_M = \{\mathbf{b}_{M1}, \ldots, \mathbf{b}_{MN}\}$:

$$\mathbf{b}_{Mi} \,|\, \mathbf{b}_{M(i)}, M, G_{M0} \sim \left(\frac{\alpha}{\alpha + N - 1}\right) G_{M0} + \left(\frac{1}{\alpha + N - 1}\right) \sum_{i' \neq i} \delta_{\mathbf{b}_{Mi'}}, \quad i = 1, \ldots, N \tag{3.6}$$

where $\mathbf{b}_{M(i)}$ is the set of random coefficient vectors for all but the i^{th} subject. The infinite-dimensional G_M has been integrated out, and the induced prior on the random effects for subject i conditional on the random effects for other subjects is a mixture of the base distribution and a discrete uniform distribution on the other subjects' values. This process tends to group subjects into clusters, occasionally assigning a subject to a new cluster with value sampled from the base distribution. We denote the r_M distinct values of the random coefficients $\boldsymbol{\theta}_M = (\boldsymbol{\theta}_{M1}, \ldots, \boldsymbol{\theta}_{Mr_M})$. We let $\mathbf{S}_M$ be the $N \times 1$ partition vector indicating the cluster membership of each subject. The likelihood of the data conditional on the allocation of subjects to r_M clusters and on $\boldsymbol{\theta}_M$ follows the

form:

$$L(\mathbf{y}|\boldsymbol{\theta}_M, \tau_M, M, \mathbf{S}_M) \propto \tau_M^{n/2} exp\big[-\frac{\tau_M}{2} \sum_{j=1}^{r_M} \sum_{i \in I_{Mj}} (\mathbf{y}_i - \mathbf{H}_i \boldsymbol{\theta}_{Mj})'(\mathbf{y}_i - \mathbf{H}_i \boldsymbol{\theta}_{Mj}) \big] \quad (3.7)$$

where I_{Mj} is the set of subjects in cluster j. This specification results from a DPP for G_M for all $M \in \mathcal{M}$. However, we have not specified any dependency structure between G_M and $G_{M'}$ for $M, M' \in \mathcal{M}$. Note that due to the changing dimension between the different models in $\mathcal{M}$ it is necessary to implement a RJMCMC algorithm for posterior computation. The next section illustrates the implementation of the RJMCMC. As we will see later, RJMCMC will be impractical without defining dependence between G_M and $G_{M'}$ for all $M, M' \in \mathcal{M}$.

3.2.3 Reversible jump MCMC sampler

The RJMCMC sampler was proposed by Green (1995) as a generalization of the Metropolis-Hastings algorithm (Hastings, 1970) for a parameter of varying dimension. Here, we use it to update the model, where the dimension varies because the number of basis functions can change. At each iteration, a proposal is made to change the current model, M to an new model, M'. The proposal is accepted with probability $Q(M', M)$, which must meet certain regularity conditions in order to sample from the target distribution of interest.

The goal of the RJ step is to sample from the posterior distribution of the elements of $\mathcal{M}$ given the data, $\mathbf{y}$. A sufficient condition for attaining the correct target distribution is that the Markov chain satisfies detailed balance under that distribution (Green, 1995). If the following equality holds for any $M, M' \in \mathcal{M}$ for a unique distribution π, then detailed balance is achieved with π as the limiting distribution of the sampler

(Brémaud, 1999):

$$\pi(M)T(M', M) = \pi(M')T(M, M') \tag{3.8}$$

where $T(M', M)$ is the probability of transitioning to model M' given the current model is M. Given our goal of sampling from the posterior of the model space, we wish to construct the acceptance probability so that $\pi(M) = p(M|\mathbf{y})$. Note that $T(M', M) = S(M', M)Q(M', M)$, where $S(M', M)$ is the probability of proposing a transition from the current model M to M', and $Q(M', M)$ is the probability of accepting that proposal.

To minimize sample autocorrelation, it is optimal to make the acceptance probability as large as possible subject to (3.8) (Percy, 1973). The optimal probability for the RJ sampler takes the form (Green, 1995; Denison et al., 2002):

$$Q(M', M) = min\Big[1, \frac{p(M'|\mathbf{y})S(M, M')}{p(M|\mathbf{y})S(M', M)}\Big] \tag{3.9}$$

which can be rewritten as:

$$Q(M', M) = min\Big[1, \frac{p(\mathbf{y}|M')p(M')S(M, M')}{p(\mathbf{y}|M)p(M)S(M', M)}\Big] \tag{3.10}$$

where $p(\mathbf{y}|M)$ is the marginal likelihood under model M. Thus, the acceptance probability is the product of the Bayes factor, the ratio of the priors on the models, and the ratio of the proposal densities governing the moves between models. When the models have different dimensions, the proposal ratio may contain a Jacobian term (Green, 1995). However, the need for a Jacobian can be avoided through careful specification of the prior on the model space and generation of proposals (Holmes and Mallick, 2000; Denison et al., 2002).

We need to choose prior probabilities for each model in $\mathcal{M}$. The number of basis functions composing model M, denoted k_M, is allowed to range from 1 to some maxi-

mum, K. We specify the following prior for k_M, for all $M \in \mathcal{M}$: $p(k_M) = \binom{T}{k_M - 1}^{-1} K^{-1}$, where T is very large. A priori, all basis functions are presumed equally likely, so the prior probability of any model M is $p(M) = p(k_M)$. The prior on the model space is discussed in more detail in Appendix A.

At each iteration we propose either to add a basis to the model, remove a basis from the model, or alter a basis in the model. All moves are proposed with equal probability, except that adding a basis is unacceptable when $k_M = K$ and removing a basis is unacceptable when $k_M = 1$. If we propose to remove a basis, that basis is randomly chosen. If we propose to add a basis, a new basis is generated as described by Holmes and Mallick (2001). If we propose to alter a basis, one is randomly chosen, and a new basis is generated to replace it.

This prior on the model space and move proposal structure leads to the acceptance probability:

$$Q(M', M) = min\Big[1, \frac{p(\mathbf{y}|M')}{p(\mathbf{y}|M)} \times R\Big] \tag{3.11}$$

where, as shown in Appendix A, $R = \frac{p(M')S(M,M')}{p(M)S(M',M)}$ simplifies to the ratio of the probability of proposing the selected move type to the probability of proposing the reverse move type if we were starting at model M'. Thus, the acceptance probability is simply the Bayes factor comparing the proposed model to the current model, multiplied by a known constant (Holmes and Mallick, 2000, Denison et al., 2002).

Unfortunately, $p(\mathbf{y}|M)$ does not have closed form here and the Bayes factor is intractable. However, the likelihood $p(\mathbf{y}|M, \mathcal{V})$, where $\mathcal{V}$ is some set of additional parameters, can be calculated directly. If we choose $\mathcal{V}$ to be parameters common to both M and M', straightforward algebra shows that the expression in (3.10) can be written:

$$Q(M', M) = min\Big[1, \frac{p(\mathbf{y}|M', \mathcal{V})p(\mathcal{V}|M')p(\mathcal{V}|M, \mathbf{y})}{p(\mathbf{y}|M, \mathcal{V})p(\mathcal{V}|M)p(\mathcal{V}|M', \mathbf{y})} \times R\Big] \tag{3.12}$$

The ratio $\frac{p(\mathcal{V}|M')p(\mathcal{V}|M,\mathbf{y})}{p(\mathcal{V}|M)p(\mathcal{V}|M',\mathbf{y})}$ may be difficult to calculate. A common approach (Holmes and Mallick, 2001; Denison et al., 2002) is to instead use the acceptance probability:

$$Q_2(M', M) = min\Big[1, \frac{p(\mathbf{y}|M', \mathcal{V})}{p(\mathbf{y}|M, \mathcal{V})} \times R\Big] \tag{3.13}$$

At each iteration, the model is first updated through a reversible jump step and an MCMC algorithm is used to draw $\mathcal{V}$ from $p(\mathcal{V}|M, \mathbf{y})$ under the current model. As long as the reversible jump step appropriately samples from $p(M|\mathcal{V}, \mathbf{y})$, then a large number of iterations of this process will obtain a sample from $p(M|\mathbf{y})$.

Theorem 1

The acceptance probability Q_2 in a reversible jump algorithm yields convergence to $p(M|\mathcal{V}, \mathbf{y})$ if $p(\mathcal{V}|M)$ is constant across all $M \in \mathcal{M}$.

Proof. Without loss of generality, for two models M and M' in $\mathcal{M}$:

$$Q_2(M', M) = \frac{p(\mathbf{y}|M', \mathcal{V})}{p(\mathbf{y}|M, \mathcal{V})} \times R = \frac{p(\mathbf{y}|M', \mathcal{V})p(M')S(M, M')}{p(\mathbf{y}|M, \mathcal{V})p(M)S(M', M)}$$

$$Q_2(M, M') = 1$$

Using the detailed balance equations in (3.8), we get:

$$\pi(M)S(M', M)\frac{p(\mathbf{y}|M', \mathcal{V})p(M')S(M, M')}{p(\mathbf{y}|M, \mathcal{V})p(M)S(M', M)} = \pi(M')S(M, M') \quad \forall M \in \mathcal{M}$$

$$\frac{\pi(M)}{p(\mathbf{y}|M, \mathcal{V})p(M)} = \frac{\pi(M')}{p(\mathbf{y}|M', \mathcal{V})p(M')}$$

Solving these equations for $\pi(M)$, the limiting distribution of the reversible jump step, if $\mathcal{V}$ is known, is:

$$\pi(M) = \frac{p(M|\mathbf{y}, \mathcal{V})}{p(\mathcal{V}|M)} \quad \forall M \in \mathcal{M}$$

If $p(\mathcal{V}|M)$ is constant across all $M \in \mathcal{M}$, then $\pi(M) = p(M|\mathbf{y}, \mathcal{V})$ is the solution to the detailed balance equations. So the acceptance probability Q_2 is appropriate when $p(\mathcal{V}|M)$ is constant across all $M \in \mathcal{M}$. □

We propose a more general form of the acceptance probability, where the condition of equal priors on $\mathcal{V}$ over the model space need not be met. Theorem 2 is easily verified by modeling the proof of Theorem 1.

Theorem 2

If the ratio $\frac{p(\mathcal{V}|M')}{p(\mathcal{V}|M)}$ is known for any $\{M, M'\}$ in $\mathcal{M}$, then the following acceptance probability satisfies the detailed balance equations for $\pi(M) = p(M|\mathbf{y}, \mathcal{V})$:

$$Q_3(M', M) = min\Big[1, \frac{p(\mathbf{y}|M', \mathcal{V})p(\mathcal{V}|M')}{p(\mathbf{y}|M, \mathcal{V})p(\mathcal{V}|M)} \times R\Big] \tag{3.14}$$

□

The acceptance probability in Q_3 provides an alternative to Q_1 when the Bayes factor comparing the two models can not be computed. It has the advantage over Q_2 that it requires less stringent assumptions about the prior on $\mathcal{V}$.

Embedding this reversible jump step within an MCMC algorithm, we can alternate between updating the model using the acceptance probability Q_3 given in (3.14) and using an MCMC sampler to update $\mathcal{V}$ from its full conditional under the selected model. After convergence, the sample of models obtained through this MCMC algorithm is then from the target distribution $p(M|\mathbf{y})$.

3.3 Model specification & Implementation

3.3.1 Prior specification

Recalling the data likelihood is given in (3.7), we complete the Bayesian specification of the model described in Section 3.2.1 by putting priors on all parameters. Beginning with the Dirichlet process on the distribution of the random coefficients, we assign the following prior structure for model $M \in \mathcal{M}$:

$$\begin{aligned}
\mathbf{b}_{Mi} &\overset{iid}{\sim} G_M, \;\; i = 1, \ldots, N \\
G_M &\sim DP(\alpha_M G_{M0}) \\
G_{M0} &= N_{k_M}(\boldsymbol{\beta}_M, \tau_M^{-1}\boldsymbol{\Delta}_M^{-1}) \\
\boldsymbol{\beta}_M &\sim N_{k_M}(\mathbf{0}, \tau_M^{-1}\lambda_M^{-1}\mathbf{I}_{k_M}) \\
\pi(\tau_M, \lambda_M, \boldsymbol{\delta}_M) &\propto \tau_M^{a_\tau - 1} exp(-b_\tau \tau_M) \lambda_M^{a_\lambda - 1} exp(-b_\lambda \lambda_M) \prod_{l=1}^{k_M} \left(\delta_{Ml}^{a_\delta - 1} exp(-b_\delta \delta_{Ml})\right)
\end{aligned} \tag{3.15}$$

where $\boldsymbol{\Delta}_M = \text{diag}(\boldsymbol{\delta}_M)$ and $a_\tau,\ b_\tau,\ a_\lambda,\ b_\lambda,\ a_\delta$ and b_δ are pre-specified hyperparameters constant across models. The Dirichlet precision parameter α_M is assumed to be known for simplicity, though modifications to place a gamma hyperprior on α_M are straightforward (Escobar and West, 1995).

3.3.2 Reversible Jump

Since we wish to allow the basis functions to change, we implement RJMCMC to move among models in $\mathcal{M}$. The reversible jump acceptance probability as given in (3.11) is not appropriate here, as the hierarchical model based on the likelihood in (3.7) and the priors in (eq.priors) makes $\frac{p(M'|\mathbf{y})}{p(M|\mathbf{y})}$ impossible to calculate. However, through careful

integration, we find:

$$p(\mathbf{y}|M,\boldsymbol{\delta}_M,\lambda_M,\mathbf{S}_M) = C(\lambda_M,k_M)|\mathbf{R}_M|^{-\frac{1}{2}}(b_\tau + \frac{A_M}{2})^{-(\frac{N}{2}+a_\tau)}\prod_{l=1}^{k_M}\delta_{Ml}^{r_M/2}\prod_{j=1}^{r_M}|\mathbf{U}_{Mj}|^{-\frac{1}{2}} \tag{3.16}$$

where

$$\begin{aligned}
\mathbf{U}_{Mj} &= \sum_{i\in\mathcal{I}_j}(\Delta_M + \mathbf{H}'_{Mj}\mathbf{H}_{Mj}) \text{ for } j = 1,\ldots,r_M \\
\mathbf{R}_M &= \lambda_M\mathbf{I}_{k_M} + r_M\boldsymbol{\Delta}_M - \boldsymbol{\Delta}_M(\sum_{j=1}^{r}\mathbf{U}_{Mj}^{-1})\boldsymbol{\Delta}_M \\
A_M &= \mathbf{y}'\mathbf{y} - \sum_{j=1}^{r}\langle\sum_{i\in\mathcal{I}_{Mj}}\mathbf{H}'_{Mi}\mathbf{y}_i, \mathbf{U}_{Mj}^{-1}\rangle - \langle\boldsymbol{\Delta}_M\sum_{j=1}^{r_M}\mathbf{U}_{Mj}^{-1}\mathbf{H}'_{Mi}\mathbf{y}_i, \mathbf{R}_M^{-1}\rangle \\
C(\lambda_M,k_M) &= \frac{{b_\tau}^{a_\tau}\lambda_M^{\frac{k_M}{2}}\Gamma(\frac{N}{2}+a_\tau)}{\Gamma(a_\tau)(2\pi)^{\frac{N}{2}}}
\end{aligned}$$

where $\mathcal{I}_{Mj}$ is again the set of subjects in the j^{th} cluster, $j = 1,\ldots,r_M$ and $\langle A,B\rangle$ denotes the quadratic form $A'BA$. Because we have a closed form for this conditional likelihood, we would like to use the acceptance probability Q_3 given in (3.14). At the point of model comparison, we have sampled values of all parameters under the current model. Let $\mathcal{V} = \{\mathbf{S}_M,\boldsymbol{\delta}_M,\lambda_M\}$ be the set of current parameter values, which were sampled under model M. From Theorem 2, the acceptance probability comparing M to M' can be written:

$$Q_3(M',M) = min\Big[1, \frac{p(\mathbf{y}|M',\mathbf{S}_M,\boldsymbol{\delta}_M,\lambda_M)p(\mathbf{S}_M,\boldsymbol{\delta}_M,\lambda_M|M')}{p(\mathbf{y}|M,\mathbf{S}_M,\boldsymbol{\delta}_M,\lambda_M)p(\mathbf{S}_M,\boldsymbol{\delta}_M,\lambda_M|M)} \times R\Big] \tag{3.17}$$

This is a ratio of conditional likelihood times a prior ratio times a known constant. We consider first the conditional likelihoods. For the result in Theorem 2 to hold, $\{\mathbf{S}_M,\boldsymbol{\delta}_M,\lambda_M\}$ must be parameters that are defined under both models. The scalar parameter λ_M and the N-dimensional vector $\mathbf{S}_M$ are in all models in $\mathcal{M}$. The parameter

$\boldsymbol{\delta}_M$ is a vector that is present in all models, but whose dimension is equal to the number of basis functions, which may vary from model to model.

When the proposal is to alter a basis function in M, $k_M = k_{M'}$ and we can use (3.16) to calculate $p(\mathbf{y}|M', \mathbf{S}_M, \boldsymbol{\delta}_M, \lambda_M)$. However, when we propose to add or remove a basis function, $k_M \neq k_{M'}$ and this quantity can not be directly calculated. We define $\boldsymbol{\delta}_{adj}$ to be the subset of $\boldsymbol{\delta}_M$ corresponding to basis functions common to both M and M'. Consequently, the dimension of $\boldsymbol{\delta}_{adj}$ is $\min(k_M, k_{M'})$. Letting $\boldsymbol{\delta}_{adj} = \boldsymbol{\delta}_M$ when we're proposing to alter a basis function, we accept a proposal with the following probability:

$$Q_4(M', M) = min\Big[1, \frac{p(\mathbf{y}|M', \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)p(\mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M|M')}{p(\mathbf{y}|M, \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)p(\mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M|M)} \times R\Big] \tag{3.18}$$

While $p(\mathbf{y}|M, \mathbf{S}_M, \boldsymbol{\delta}_M, \lambda_M)$ has closed form, $p(\mathbf{y}|M, \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)$ may not. If we propose to add a basis to the current model, the current piecewise linear model is nested in the proposed model and is one basis smaller, and $\boldsymbol{\delta}_{adj} = \boldsymbol{\delta}_M$. The denominator of the acceptance probability has closed form, as shown above, and we use a Laplace approximation to estimate the one-dimensional integral in the numerator. Similarly, if the proposal is to remove a basis from M, then $\boldsymbol{\delta}_{adj}$ is a $(k_M - 1)$-dimensional subvector of $\boldsymbol{\delta}_M$ and we use the Laplace approximation in the numerator. Further details on the approximation were provided in Chapter 2.

We have defined a closed form for the ratio of the conditional likelihoods, and it is necessary to derive an expression for the ratio of priors $\frac{p(\mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M|M')}{p(\mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M|M)}$. Note that this ratio does not depend on the data. Under our prior structure, $p(\mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M|M) = p(\mathbf{S}_M|M)p(\boldsymbol{\delta}_{adj}|M)p(\lambda_M|M)$. In addition, the gamma priors on the elements of δ_{adj}, and λ_M do not depend on the model, so that $p(\boldsymbol{\delta}_{adj}|M) = p(\boldsymbol{\delta}_{adj})$ and $p(\lambda_M|M) = p(\lambda_M)$ for all $M \in \mathcal{M}$.

In the prior specification above, we have a distinct and independent G_M for each $M \in \mathcal{M}$. We need to calculate $\frac{p(\mathbf{S}_M|M')}{p(\mathbf{S}_M|M)}$. The parameter $\mathbf{S}_M$ is discrete with a large

number of possible values, making $p(\mathbf{S}_M|M)$ very complicated. The stick-breaking representation of the Dirichlet process is informative about the complicated distribution of $\mathbf{S}_M$. Specifically, we can note that the prior on the cluster allocation under model M depends only on the stick-breaking weights. These weights are random observations from distributions dependent only on α_M.

Theorem 3

If $\alpha_M \equiv \alpha$ for all $M \in \mathcal{M}$, then for some cluster allocation $\mathbf{S}$, $p(\mathbf{S}|M) = p(\mathbf{S}|M')$ for all $\{M, M'\}$ in $\mathcal{M}$

Proof. First note that the prior on the $(N \times 1)$ vector $\mathbf{S}$ can be written: $p(\mathbf{S}|M, \alpha_M) = p(S_1|M, \alpha_M)p(S_2|S_1, M, \alpha_M) \dots p(S_N|S_1, \dots, S_{N-1}, M, \alpha_M)$. Without loss of generality, we can choose S_1 arbitrarily, yielding the following set of conditional priors:

$$\begin{aligned}
p(S_1|M, \alpha_M) &= 1 \\
p(S_2|S_1, M, \alpha_M) &= \left(\frac{1}{\alpha_M + 1}\right)^{1(S_2 = S_1)} \left(\frac{\alpha_M}{\alpha_M + 1}\right)^{1(S_2 \neq S_1)} \\
p(S_i|S_1, \dots, S_{i-1}, M, \alpha_M) &= \left(\frac{\alpha_M}{\alpha_M + i - 1}\right)^{1(S_i \neq S_h,\, h<i)} \prod_{h=1}^{i-1} \left(\frac{1}{\alpha_M + i - 1}\right)^{1(S_i = S_h)} \\
&\quad \text{for } \; i = 3, \dots, N
\end{aligned}$$

If $\alpha_M \equiv \alpha$ for all $M \in \mathcal{M}$, then these conditional priors and therefore the prior on $\mathbf{S}$ do not depend on the model, so $p(\mathbf{S}|M) = p(\mathbf{S}|M')$ for all $\{M, M'\}$ in $\mathcal{M}$. □

As a consequence of Theorem 3, if we specify that the DP precision parameters are equivalent for all models in $\mathcal{M}$, then the acceptance probability becomes:

$$Q_4(M', M) = min\Big[1, \frac{p(\mathbf{y}|M', \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)}{p(\mathbf{y}|M, \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)} \times R\Big] \tag{3.19}$$

As shown in Section 3.2.3, since $\boldsymbol{\delta}$ and λ are unspecified, the validity of this acceptance probability relies on alternating between updating the model and then updating $\mathbf{S}, \boldsymbol{\delta}$

and λ under the model, which can be done directly by Gibbs sampling from the full conditionals.

3.3.3 Pólya urn Gibbs sampling

At each iteration, we use a reversible jump step to sample a model from $\mathcal{M}$. All other parameters (including the DP cluster allocation) are then updated through Gibbs sampling under the accepted model, and the process is repeated. Suppose, through a reversible jump step, we have selected model M, which may not be equal to the previous model, M_p. We have values for all model parameters as sampled under M_p, and we desire to sample all parameters under M.

We then initialize the parameters $\{\lambda_M, \tau_M, \mathbf{S}_M\}$ to $\{\lambda_{M_p}, \tau_{M_p}, \mathbf{S}_{M_p}\}$. The other model parameters depend on the dimension. If $k_M \neq k_{M_p}$, then we cannot simply initialize all the parameters to their values under model M_p. We need to define $\boldsymbol{\delta}_M, \boldsymbol{\theta}_M$, and $\boldsymbol{\beta}_M$ of the appropriate dimension in order to appropriately sample from the full conditionals. We also have $\boldsymbol{\delta}_{M_p}, \boldsymbol{\theta}_{M_p}$, and $\boldsymbol{\beta}_{M_p}$ from the previous model. If $k_M = k_{M_p}+1$, we initialize the new element of $\boldsymbol{\delta}_M$ to the mean of $\boldsymbol{\delta}_{M_p}$ and we initialize the new element of θ_j to 0 for $j = 1, \ldots, r_M$. If $k_M = k_{M_p} - 1$, we simply create $\boldsymbol{\delta}_M, \boldsymbol{\theta}_M$, and $\boldsymbol{\beta}_M$ by removing the elements of $\boldsymbol{\delta}_{M_p}, \boldsymbol{\theta}_{M_p}$, and $\boldsymbol{\beta}_{M_p}$ corresponding the the basis function that is present in M_p but not in M.

Consequently, the number of clusters r_M remains unchanged from the previous model. We then update all model parameters from their full conditionals under M. For notational simplicity, we suppress the model indicator subscript on the parameters so that $\{k_M, \boldsymbol{\delta}_M, \boldsymbol{\theta}_M, \lambda_M, \tau_M, r_M, \mathbf{b}_M, \boldsymbol{\beta}_M, G_{M0}\} \equiv \{k, \boldsymbol{\delta}, \boldsymbol{\theta}, \lambda, \tau, r, \mathbf{b}, \boldsymbol{\beta}, G_0\}$.

The updating of the random effects, $\mathbf{b}$, is based on the Pólya urn Gibbs sampling algorithm (MacEachern, 1994; West et al., 1994; Escobar, 1994; Escobar and West, 1995). Under a given model M, recall that the r distinct values of the random effects

are denoted $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r)$. Excluding subject i, there are $r^{(i)} \leq r$ distinct values of the random effects, denoted $(\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{r^{(i)}}^{(i)})$

As explained by West et al. (1994), the full conditional of $\mathbf{b}_i$ is a mixture the base prior updated with subject i's data and point masses at all other random effects. The mixing weights are given by:

$$q_{i,j} = \begin{cases} \alpha h_i(\mathbf{y}_i) & \text{if } j = 0 \\ n_j^{(i)} f_j(\mathbf{y}_i|\boldsymbol{\theta}_j) & \text{if } j > 0 \end{cases}$$

where $q_{i,0}$ is the weight given to the posterior under the base prior, $q_{i,j}$ is the weight given to a point mass at the j^{th} random effect, and $n_j^{(i)}$ is the size of the j^{th} cluster excluding subject i.

$$h_i(\mathbf{y}_i) = \int f_i(\mathbf{y}_i|\mathbf{b}_i) dG_0(\mathbf{b}_i)$$

The hierarchical model given in Section 3.3.1 is conjugate under the base prior, so we can obtain a closed form for $h_i(\mathbf{y}_i)$ and $G_{i,0}$.

$$h_i(\mathbf{y}_i) = \frac{\tau^{\frac{n_i}{2}} |\boldsymbol{\Delta}|^{\frac{1}{2}} (2\pi)^{-\frac{n_i}{2}}}{|\mathbf{H}_i'\mathbf{H}_i + \boldsymbol{\Delta}|^{\frac{1}{2}}} exp\Big(\frac{\tau}{2}[(\mathbf{H}_i'\mathbf{y}_i + \boldsymbol{\Delta}\boldsymbol{\beta})(\mathbf{H}_i'\mathbf{H}_i + \boldsymbol{\Delta})^{-1}(\mathbf{H}_i'\mathbf{y}_i + \boldsymbol{\Delta}\boldsymbol{\beta}) - (\mathbf{y}_i'\mathbf{y}_i + \boldsymbol{\beta}'\boldsymbol{\Delta}\boldsymbol{\beta})] \quad (3.20)$$

$$G_{i,0} = N_k\Big((\mathbf{H}_i'\mathbf{H}_i + \boldsymbol{\Delta})^{-1}(\mathbf{H}_i'\mathbf{y}_i + \boldsymbol{\Delta}\boldsymbol{\beta}), (\mathbf{H}_i'\mathbf{H}_i + \boldsymbol{\Delta})^{-1}\Big) \quad (3.21)$$

We could sample $\{\mathbf{b}_i\}$ from its full conditional for each subject, but computational efficiency is improved by exploiting the clustering behavior of the DP and instead sampling $\mathbf{S}$ and then $\boldsymbol{\theta}$ from their full conditional distributions.

The cluster indicator for subject i has the following full conditional posterior distribution:

$$p(S_i = j|\mathbf{y}, \mathbf{b}^{(i)}, \mathbf{S}^{(i)}, r^{(i)}) = q_{i,j} \quad \text{for } i = 1, \ldots, N$$

$S_i = 0$ implies the creation of a new cluster containing only subject i. A corresponding new value of $\mathbf{b}_i$ is drawn from $G_{i,0}$, the base prior updated with subject i's

random coefficients, and $\boldsymbol{\theta}$ and r are updated appropriately.

$$
\begin{aligned}
p(\boldsymbol{\theta}_j|\mathbf{y},\mathbf{S},r) &= N\Big((\boldsymbol{\Delta}+\sum_{i\in\mathcal{I}_j}\mathbf{H}_i'\mathbf{y}_i)^{-1}(\boldsymbol{\Delta}\boldsymbol{\beta}+\sum_{i\in\mathcal{I}_j}\mathbf{H}_i'\mathbf{H}_i),\tau^{-1}(\boldsymbol{\Delta}+\sum_{i\in\mathcal{I}_j}\mathbf{H}_i'\mathbf{y}_i)^{-1}\Big) \\
p(\boldsymbol{\beta}|\boldsymbol{\theta},\mathbf{S},\boldsymbol{\delta},\lambda) &= N\Big((\lambda\mathbf{I}+r\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}\sum_{j=1}^{r}\boldsymbol{\theta}_j,\tau^{-1}(\lambda\mathbf{I}+r\boldsymbol{\Delta})^{-1}\Big)
\end{aligned}
$$

Our prior distributions for the precision parameters yield the following Gamma full conditional posterior distributions:

$$
\begin{aligned}
\lambda|\boldsymbol{\beta},\mathbf{b},\boldsymbol{\delta},\tau &\sim Gamma\Big(a_\lambda+\frac{k}{2},b_\lambda+\frac{\tau\boldsymbol{\beta}'\boldsymbol{\beta}}{2}\Big) \\
\delta_l|\boldsymbol{\beta},\mathbf{b},\lambda,\tau,\delta_h\ h\neq l &\sim Gamma\Big(a_\delta+\frac{N}{2},b_\delta+\frac{\tau}{2}\sum_{j=1}^{r}(\theta_{jl}-\beta_l)^2\Big) \qquad l=1,\ldots,k \\
\tau|\mathbf{y},\mathbf{b},\mathbf{c},\boldsymbol{\delta},\lambda,\nu &\sim Gamma\Big(a_\tau+\frac{N+(r+1)k}{2},b_\tau+\frac{1}{2}\big(\sum_{i=1}^{N}(\mathbf{y}_i-\boldsymbol{H}_i\mathbf{b}_i)'(\mathbf{y}_i-\boldsymbol{H}_i\mathbf{b}_i) \\
&\qquad +\lambda\boldsymbol{\beta}'\boldsymbol{\beta}+\sum_{j=1}^{r}(\boldsymbol{\theta}_j-\boldsymbol{\beta})'\boldsymbol{\Delta}(\boldsymbol{\theta}_j-\boldsymbol{\beta})\big)\Big)
\end{aligned}
$$

3.4 Simulations

3.4.1 Simulated data

To illustrate the method, we simulated data from four clusters, with 25 subjects per cluster and 10 observations per subject. There were two covariates. One could be thought of as time and consisted of randomly generated continuous values from 0 to 5. The other was a time-dependent covariate, which ranged from -10 to 10 and increased linearly with time. Each subject's trajectory is a linear combination of this continuous covariate and one of four smooth functions of time. These four smooth functions, along with the 250 data points generated for each cluster, are shown in Figure 3.1. Within-subject dependence was induced by assigning each subject an additive random effect

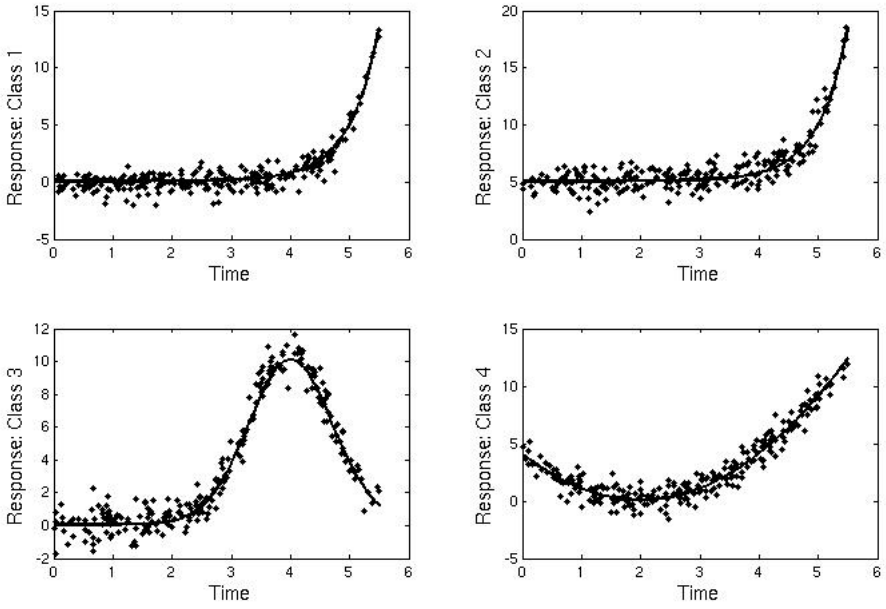

FIGURE 3.1: Underlying population curves (lines) and data (points) for each of the four clusters. Each cluster had 250 observations from 25 subjects

drawn from a $N(0, 1/4)$ distribution, and the additive residual error for each data point was drawn independently from $N(0, 1/16)$.

We require a priori specification of α, the DP precision parameter. Escobar (1994) demonstrates that high values of α yield large numbers of clusters. For interpretability, we wish to have only a small number of clusters. Based on the guidelines given in Escobar (1994), we choose $\alpha = 0.5$. The hyperparameters for the gamma priors on the precision parameters are all set to 0.05, yielding somewhat vague priors centered at 1.

We ran the algorithm in Matlab for 20,000 iterations after a 5,000 iteration burn-in period. Sample collection took approximately 10 hours. At each iteration, we collected the cluster membership for each subject, the individual cluster trajectories, and the population mean trajectory.

To identify a single set of latent classes from our set of 20,000 sampled cluster structures, we use a hierarchical clustering algorithm. Following Medvedovic and Siva-

ganesan (2002), who searched for clusters of co-expressed genes in microarray data, we define the total distance between two vector observations to be the proportion of MCMC samples at which two observations were put in different clusters. Observations separated by small distances are often sampled in the same cluster and thus are more likely to truly belong to the same class.

To use these pairwise distances for classification, we require that any observations separated by less than some threshold distance be in the same class. This threshold thus becomes the minimum distance allowed between two observations in separate classes. A high threshold yields few classes, and a low threshold will result in many smaller classes. The choice of the threshold may be largely driven by the application. Hierarchical clustering lends itself readily to the creations of a cluster tree, a visual representation of how tight the clusters are and how the choice of threshold affects the number of clusters. Fraley and Raftery (2002) explore an EM algorithm and other methods for estimating the appropriate threshold or number of classes to choose. This hierarchical clustering algorithm represents an alternative to the commonly-used k means clustering algorithm, which is not applicable to the categorical output of the DP clustering procedure, and the k-modes approach of Huang (1998), which is applicable but not easily applied.

3.4.2 Simulation results

The hierarchical clustering structure can be displayed in a highly informative tree form, where the observations are plotted on the x-axis and the y-axis gives the maximum distance within a cluster. Recall that distance is the proportion of MCMC samples in which the observations were placed in separate clusters. Figure 3.2 shows the top of the hierarchical tree. The four clusters are correctly distinguished. We notice that observations from the same cluster may sometimes have low probability of being observed in the same cluster. However, observations from different true clusters were virtually

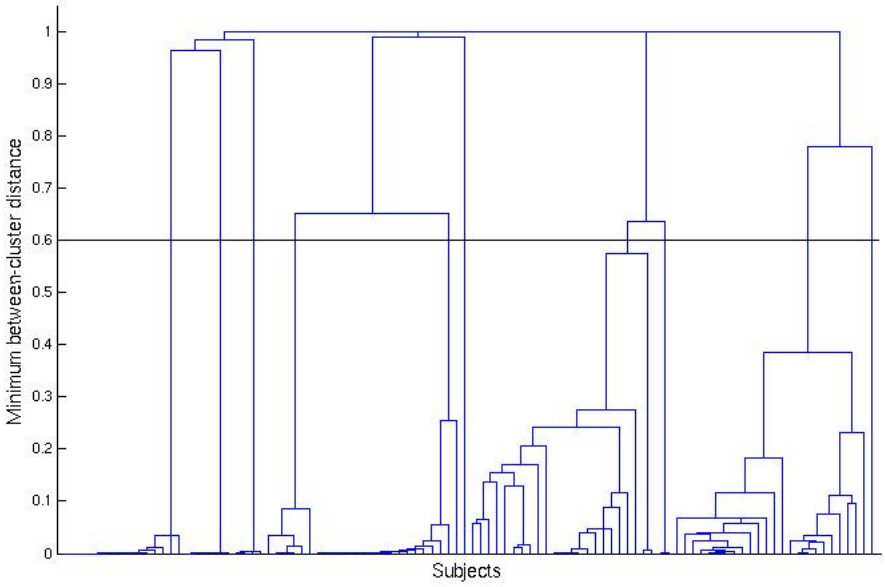

FIGURE 3.2: Hierarchical class tree. The between-class distance is given on the y-axis, where the distance is the posterior probability that the two classes are distinct. The four true clusters are the four classes at the top of the tree, and more classes are seen as the minimum distance decreases. The solid horizontal line is the threshold at which our classes were created.

never assigned to the same cluster.

We classified the observations so that the minimum between-group distance is 60%, meaning that observations were grouped together if they appeared in the same cluster more than 40% of the time. Ten classes resulted. Each of the four true clusters was split into one large class, ranging in size from 15-24 observations, and one to three smaller classes ranging in size from 1-6 observations.

Figure 3.3 shows the fitted curve for each of the four true clusters as well as the fitted curve for each of the ten classes. The model fit the population trajectories closely. Examination of the raw data indicates that the model correctly identified women in each of the four groups with unusual random effects. We assessed convergence by examining the stability of the distributions of the number of basis functions in the

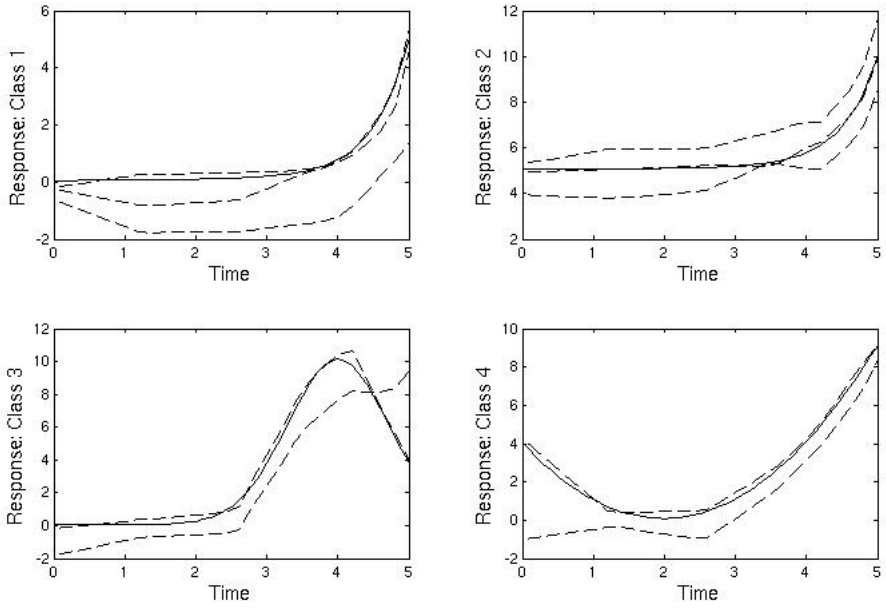

FIGURE 3.3: Four true clusters (solid lines) and the ten classes (dashed lines)

model, number of clusters, and the precision parameters τ and λ. Traceplots of these parameters provided no evidence against convergence.

3.5 Progesterone example

3.5.1 Data

The progesterone data are from the North Carolina Early Pregnancy Study (EPS; Wilcox et al., 1988; Baird et al., 1997). Women in this study provided daily urine specimens. We consider daily urinary progesterone measurements from the onset of menstruation up until two days after ovulation. Implantation of a conceptus is known to affect progesterone, and this region is designed to precede implantation. We randomly selected one menstrual cycle from each of 172 women to use in our analysis. If a woman had both conception and non-conception cycles available, we randomly selected first

the cycle type and then the cycle. Conception cycles made up 65 (38%) of the 172 cycles.

We used Bayesian multivariate linear splines to model the trajectories, allowing for a nonparametric distribution of the random woman-specific basis coefficients around the population value. Our goal is to identify trajectory clusters and to describe any clusters that are predictive of high or low conception probability. Covariates include day since menstruation and day relative to ovulation, both of which may be related to hormone trajectories (van Zonneveld et al., 2003). We ran the algorithm for 50,000 iterations after a 10,000 iteration burn-in period.

3.5.2 Progesterone results

Analysis of the EPS data yielded several interesting results. Unlike the more ideal simulated data case, there were no completely disjoint classes to examine. As in the simulation, we selected a minimum between-group distance of 60%. This yielded one very large class with 133 observations and eight small classes with 1-14 observations each. Figure 3.4 shows the estimated progesterone trajectories for the dominant class and the eight smaller classes. In general, the trajectory in the dominant class appears to contain mid-range values and be relatively flat. The other classes tend to have sharp decreases at the start or end of the window, or to have exceptionally high or low values. These smaller clusters likely represent less common hormone patterns, and these trajectories may be clinically important. Perhaps more informative is Figure 3.5, which displays the raw data for each of the nine final classes. The method has appropriately grouped similar observations together. In addition, the presence of sets of very similar curves and very different shapes for each cluster support the decision to relax the parametric assumption that all women are normally distributed around a trajectory described by some mean set of coefficients and instead describe it through a

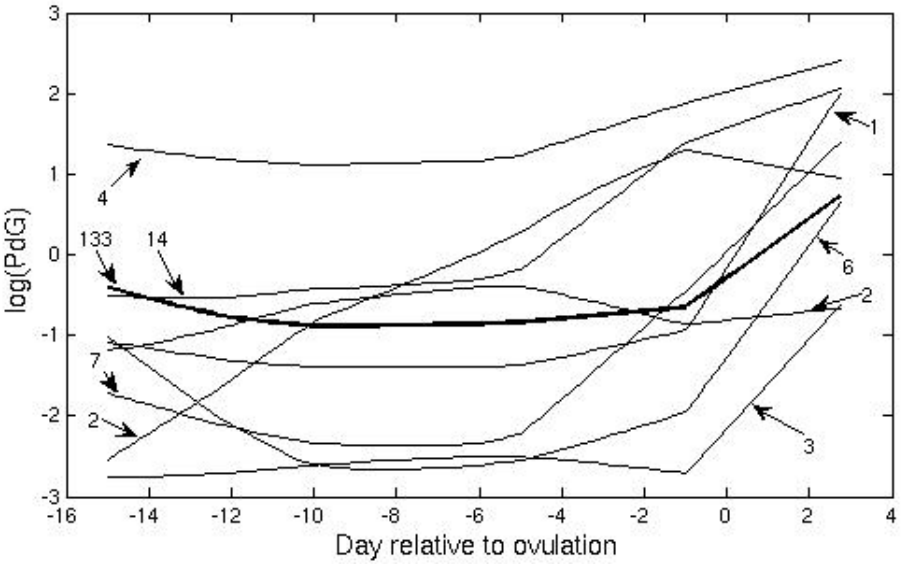

FIGURE 3.4: The dark line is the estimated trajectory for the large final class. The lighter lines are the estimated trajectories for the smaller classes. Arrows indicate the number of cycles in each cluster.

Dirichlet process.

3.5.3 Sensitivity Analysis

Adequate mixing and convergence of cluster allocation is sometimes of concern in DPP models, as the algorithm has a tendency to become trapped in a local mode (Jain and Neal, 2004). To combat this, we collected a large number of samples. In addition, we found that parallel chains from different initial cluster allocations yielded nearly identical results, and the number of clusters appeared to mix adequately across iterations. This led us to conclude that the mixing was adequate to avoid local mode problems. We also examined the implications of using different values of α. Large values of α tended to increase the average number of clusters at each iteration. In the post-processing of the samples, this meant that distances between observations tended to be larger. The same final clusters were found, although requiring a larger threshold between-group

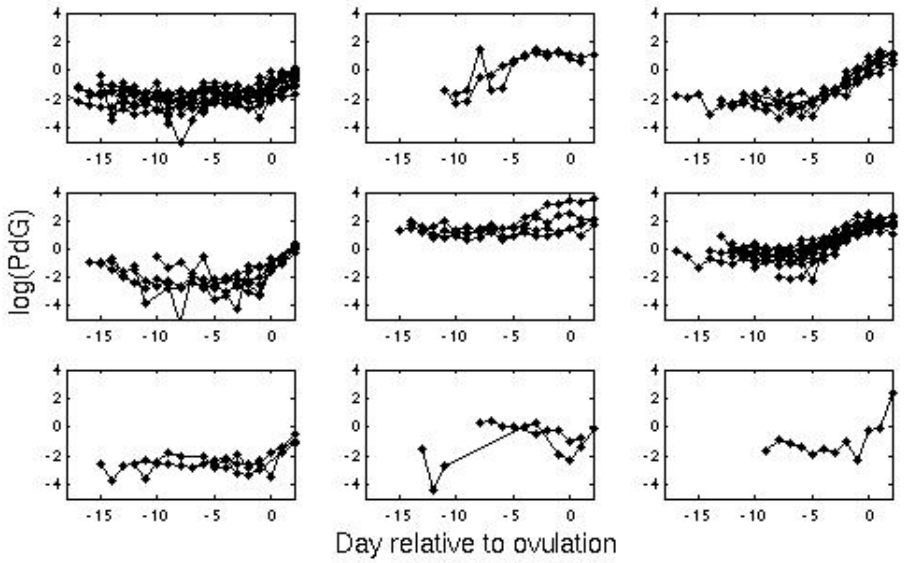

FIGURE 3.5: These are data from all nine final classes. There are twenty trajectories presented for the dominant cluster (first plot), and the other eight clusters are entirely displayed.

distance. Small values of α had the expected opposite effect.

3.6 Discussion

We have defined a Dirichlet prior structure that can be successfully applied to parameters of varying dimension. In the case of the multivariate spline model with random knots, the prior relaxes the distributional assumptions on the varying-dimension random effects. We have demonstrated the effectiveness of this method in a reversible jump framework. In addition, we have discussed ways to interpret the clustering information that is automatically produced by the Dirichlet process. The simulated data example shows that this method is appropriate for identifying and describing clusters of longitudinal trajectories.

Many methods for functional data rely on functions observed over the same region

of time, often with observations in each function at fixed points. This method instead allows for variation in number and location of observations, making it applicable in the presence of missing data. We only examined the response as a function of time relative to two reference points, but it would be appropriate to also include other covariates or reference points. Regardless of the covariates used, the method groups together observations with a similar relationship between the covariates and response.

This Bayesian regression model is useful for semiparametric modeling and clustering of longitudinal data, but can be identically applied to spatial data or clustered data with no time/space component. This method also has applications in genetics, especially to time-course gene expression data, where clustering profiles is of interest both to improve understanding of gene expression and to reduce the dimensionality of huge microarray datasets (Luan and Li, 2003; Liang et al., 2005).

Examination of the EPS data has provided an interesting illustration of the underlying clusters, but a deeper analysis of the hormone data is warranted. Previous results have indicated that low progesterone around the time when implantation of the conceptus would occur may be indicative of low cycle fecundability (Baird et al., 1997; Stewart et al., 1993). Future work will examine this hypothesis, apply the method to other reproductive hormones, and may include extensions to joint modeling of multiple reproductive hormones or the joint modeling of trajectory and conception status.

In addition, alternative methods to post-process the MCMC samples for cluster interpretation may benefit from further exploration. We found evidence that there may be small clusters or even outlying observations that do not belong in any cluster. Tseng and Wong (2005) describe an algorithm for the identification of "tight and stable" clusters, which is designed to identify tight clusters without forcing outlying observations into clusters.

CHAPTER 4

JOINT MODELING OF FUNCTIONAL AND OUTCOME DATA

This chapter proposes a new method for the joint clustering of trajectories with some outcome of interest. The trajectories are modeled using the flexible spline model of Chapter 3. The outcome is modeled through a generalized linear model with a random intercept. Through specifying the random intercept to follow a Dirichlet process jointly with the random spline coefficients, we obtain a procedure that clusters trajectories according to shape while estimating the parameters of the outcome model for each cluster. This very flexible method allows for the incorporation of covariates in the models for both the outcome and the trajectory. We apply the method to post-ovulatory progesterone data from the Early Pregnancy Study and find that the model successfully separates clinical pregnancies from early pregnancy losses.

4.1 Introduction

The joint modeling literature has generally focused on methods for longitudinal and time-to-event data. This is useful in determining the relationship between biomarkers measured over time and the risk of disease progression, cure, or death. Brown and Ibrahim (2003) use a Bayesian method, specifying a nonparametric Dirichlet process (DP) prior on the parameters of the longitudinal trajectory, then modeling the hazard conditional on the trajectory at a given time. Brown et al. (2005) develop a Bayesian method suitable for the case when the longitudinal variable is multivariate. Tsiatis and Davidian (2004) give a review of methods for joint longitudinal and time-to-event modeling. In short, the methods tend to rely on formulating an appropriate model for the trajectory and then defining the hazard function at each time-point as some function of the trajectory value at that time-point.

The current joint modeling problem is outside this time-to-event framework, as we consider the relationship between a function and some outcome random variable. The outcome need not be the time of some event, and the function need not be longitudinal. In related work, Chib and Hamilton (2002) propose a Bayesian semiparametric model for the effect of time-varying binary treatment on a longitudinal response. Instead, we look at a single time-independent variable along with each trajectory. James (2002) proposes a generalized linear model where one of the predictors is functional. His method relies on modeling the function with a cubic spline and using an EM algorithm to describe the relationship between the response and the integral of the weighted function. It requires that all functions are observed over the same region of time and modeled in only one covariate. Ratcliffe et al. (2002) describe a logistic model for a binary response where one of the covariates is functional. They model fetal heart-rate traces using a set of Fourier basis functions, then use the model to predict high-risk pregnancies. We are interested in modeling the joint relationship between a function

and some outcome, where not all functions have the same domain, and other covariates may be of interest either in modeling the function or the outcome.

We extend a multivariate linear spline model with random coefficients for the modeling of functional data to the case where each function is observed jointly with some outcome of interest. We employ the Dirichlet process to relax distributional assumptions about the random effects and to glean information about underlying clusters of functional predictors. Though demonstrated here on longitudinal data, the multivariate adaptive spline model is appropriate for examining the joint relationship between some outcome and a regression over time, space, or any other support.

The method is applied to progesterone data from the Early Pregnancy Study (Baird et al., 1997). One of the aims of the Early Pregnancy Study was to study early pregnancy loss (EPL). Based on examination of human chorionic gonadotropin (hCG) profiles, the study investigators classified cycles that did not result in clinical pregnancy as either early loss cycles or true non-conception cycles. A detectable rise in urinary hCG signaled implantation of the conceptus, and a subsequent decline indicated that the pregnancy was lost. Based on these analyses, Wilcox et al. (1988) reported that two-thirds of losses occurred before the pregnancy was clinically detected (i.e. before 6 weeks) and that nearly a third of all conceptions resulted in EPL. Other studies have reported similar incidence of EPL (Ellish et al., 1999; Zinaman et al., 1996; Wang et al., 2003).

The current project examines post-ovulatory progesterone, comparing EPL menstrual cycles to those cycles resulting in clinical pregnancy. The most distinctive feature of progesterone in this context is that it remains high in ongoing pregnancies and decreases once the pregnancy is lost. It has also been noted that progesterone tends to be slightly lower in the early weeks of pregnancy in those cycles with EPL (Lower and Yovich, 1992). This suggests that EPL, in many cases, may be the result of a pregnancy

that was weak at the onset rather than the immediate result of some trauma.

Winter et al. (2002) note that EPL in the context of assisted reproductive technology can be financially and emotionally costly. They report a 16% EPL rate and an increase in risk with smoking and poor quality embryos, but no change in risk with age or BMI after adjusting for other factors. Henriksen et al. (2004) found that alcohol consumption during the week of conception also increased the risk of EPL. In a previous analysis of data from the Early Pregnancy Study, Wilcox et al. (1998) found evidence that a longer time between ovulation and conception led to to an increased risk of EPL. They hypothesized that this was due to deterioration of the quality of the oocyte as it aged after ovulation.

No one mechanism of early loss is known. Environmental factors, stress on the part of the mother, and poor quality of the embryo may all manifest themselves as early loss. Consequently, a joint model between progesterone and early loss makes sense. In some cases, the drop in progesterone may signal the mother's inability to continue the pregnancy. In other cases, it may be a response to the embryo's inability to survive. There is likely a direct causal relationship between progesterone and EPL, but the direction of causality may vary.

4.2 Methods

Our model relies on the incorporation of a Bayesian generalized linear model for the outcome into the flexible longitudinal trajectory model described in Chapter 3. In this section, we describe a multivariate adaptive spline model for the longitudinal trajectory and methods for Bayesian analysis of the generalized linear model. Finally, we describe the integration of these two approaches to create a joint model for a curve and an outcome.

4.2.1 Multivariate linear splines

Chapter 3 describes a flexible spline model where the distribution of the individual curves around the population mean is nonparametric. The spline model is based on a generalization of the adaptive spline method of Holmes and Mallick (2001), where the dependency within subjects is accounted for through random effects. A nonparametric distribution on the random effects naturally groups subjects with similar random effects into clusters.

We use a Bayesian model which treats the covariates and response as piecewise linear, with varying numbers and locations of knots. We use the reversible jump MCMC algorithm of Green (1995) to add and remove knots, sampling models having high posterior probability. The final curve estimates are weighted averages over all sampled models, which leads to smooth curve estimates from the non-smooth piecewise linear samples.

A single multivariate piecewise linear model, M, is defined by a set of k_M basis functions, $\boldsymbol{\mu}_M = (\boldsymbol{\mu}_{M1}, \ldots, \boldsymbol{\mu}_{Mk_M})$. When y_{ij} is the j^{th} response from subject i, $i = 1, \ldots, N$; $j = 1, \ldots n_i$, the relationship between y_{ij} and its $(p \times 1)$ set of covariates $\mathbf{x}_{ij}$ can be approximated by a linear combination of the positive portions (denoted by the + subscript) of the inner products of the basis functions with the covariate vector:

$$y_{ij} = \sum_{l=1}^{k_M} b_{Mil}(\mathbf{x}'_{ij}\boldsymbol{\mu}_{Ml})_+ + \epsilon_{Mij}, \quad M \in \mathcal{M} \tag{4.1}$$

where ϵ_{Mij} is a random error. More transparently, each piecewise linear model is linear in the basis function transformations of the covariate vectors:

$$\mathbf{y}_i = \mathbf{H}_{Mi}\mathbf{b}_{Mi} + \boldsymbol{\epsilon}_{Mi}, \quad M \in \mathcal{M} \tag{4.2}$$

where $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_{Mi}$ are the $n_i \times 1$ vectors of responses and random errors, $\mathbf{b}_{Mi}$ is the $k_M \times 1$ vector of random basis coefficients for subject i, and the design matrix $\mathbf{H}_{Mi}$ contains the basis function transformations of the covariate vectors for subject i.

Assuming conditional independence of the elements of $\mathbf{y}_i$ given $\mathbf{b}_{Mi}$, and $N(0, \tau_M^{-1})$ errors, the conditional likelihood under model M is:

$$p(\mathbf{y}|\boldsymbol{b}_M, \tau_M, M) \propto \tau_M^{\frac{n}{2}} \prod_{i=1}^{N} exp\big[-\frac{\tau_M}{2}(\mathbf{y}_i - \mathbf{H}_{Mi}\mathbf{b}_{Mi})'(\mathbf{y}_i - \mathbf{H}_{Mi}\mathbf{b}_{Mi})\big] \;\; M \in \mathcal{M} \quad (4.3)$$

where $n = \sum_{i=1}^{N} n_i$. Continuing Bayesian specification of the model, we put a prior on $\boldsymbol{b}_M = (\mathbf{b}_{M1}, \ldots, \mathbf{b}_{MN})$:

$$\mathbf{b}_{Mi} \overset{iid}{\sim} G_M, \;\; i = 1, \ldots, N; \;\; M \in \mathcal{M} \quad (4.4)$$

The distribution G_M could be given some parametric form for all $M \in \mathcal{M}$. Chapter 2 specified G_M to be Gaussian, which implies that all subject-specific coefficients are normally distributed around some population mean. In the quest to uncover clusters of similar trajectories, the normality assumption makes little sense. Instead, we treat G_M as an unknown distribution by assigning it a Dirichlet process prior, which automatically provides information about underlying classes (see Chapter 3 for details.)

4.2.2 Generalized linear models

We combine the trajectory model with a generalized linear model for the outcome. Generalized linear models are an extension of normal linear models to the case where the response may not be normal. Here, we consider the response as arising from the exponential family. The random variable Z follows an exponential family distribution

if the density of Z can be written in the following form.

$$p(z|\xi,\phi) = exp\big(a(\phi)^{-1}(z\xi - B(\xi)) + c(z,\phi)\big) \tag{4.5}$$

where the distribution is said to have canonical parameter ξ and scale parameter ϕ. $B(\cdot)$ (the cumulant function) and $c(\cdot,\cdot)$ are functions that determine the particular class of distributions within the exponential family. Many common distributions fit into this form, including the Normal, Poisson, Multinomial, and Gamma distributions. The term $a(\phi)$ is commonly equal to ϕ, and we assume that here for ease of illustration.

A random variable in this family has expected value $\mu = \partial B(\xi)/\partial\xi$, the first derivative of the cumulant function with respect to the canonical parameter. The corresponding inverse function, $\xi(\mu)$, is known as the canonical link function.

The relationship between μ and covariates is often expressed through the generalized linear model, letting $g(\mu) = \boldsymbol{\eta}$. The term $\boldsymbol{\eta} \equiv \mathbf{U}\boldsymbol{\gamma}$ is the linear predictor where $\mathbf{U}$ is the covariate matrix and $\boldsymbol{\gamma}$ is the parameter vector. In normal linear models, $g(\cdot)$ is taken to be the identity function, but the identity function makes little sense unless Z can take any value on the real line. A natural and commonly used choice for $g(\cdot)$ is the canonical link $\xi(\cdot)$ (McCullagh and Nelder, 1989).

We may need to accommodate multiple observations from independent sampling units, so we employ a generalized linear mixed model (GLMM). Suppose we have a set of independent responses $\{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$, where $\mathbf{z}_i = \{z_{i1},\ldots,z_{in_i}\}$ for $i = 1,\ldots,N$, and the likelihood of one observation z_{ij} is an exponential family density.

$$p(z_{ij}|\xi_{ij},\phi_{ij}) = exp\big(\phi_{ij}^{-1}(z_{ij}\xi_{ij} - B(\xi_{ij})) + c(z_{ij},\phi_{ij})\big) \tag{4.6}$$

Suppose each observation z_{ij} has corresponding covariates $\mathbf{u}_{ij}$, and $E(z_{ij}) = \mu_{ij}$. We use a generalized linear model with the canonical link, and let $\xi_{ij}(\mu_{ij}) = \eta_{ij} \equiv \mathbf{u}'_{ij}\boldsymbol{\gamma}_i$,

where $\boldsymbol{\gamma}_i$ is a parameter vector unique to the i^{th} subject. If we assume $\phi_{ij} \equiv \phi_i$ for all $\{i, j\}$ and that observations within subject i are independent given $\{\boldsymbol{\gamma}_i, \phi_i\}$, then the likelihood becomes:

$$p(\mathbf{z}|\xi_{ij}, \phi_{ij}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} exp\big(\phi_i^{-1}(z_{ij}(\mathbf{u}'_{ij}\boldsymbol{\gamma}_i) - B(\mathbf{u}'_{ij}\boldsymbol{\gamma}_i)) + c(z_{ij}, \phi_i)\big) \tag{4.7}$$

Bayesian methods for generalized linear mixed models have many desirable properties. The intractable integrals that plague likelihood-based inference in GLMMs are not a problem here, as we can use the Gibbs sampler to draw from the posteriors of interest. A common prior structure for the GLMM described above is:

$$\begin{aligned}
\boldsymbol{\gamma}_i &\overset{iid}{\sim} N(\gamma_0, \text{diag}(\boldsymbol{\psi})^{-1}) \\
\boldsymbol{\gamma}_0 &\sim N(\mathbf{0}, \omega^{-1} I_{p_o}) \\
\pi(\omega, \boldsymbol{\psi}) &\propto \omega^{a_\omega - 1} exp(-b_\omega \omega) \prod_{l=1}^{p_o} \big(\psi_l^{a_\psi - 1} exp(-b_\psi \psi_l)\big)
\end{aligned}$$

where the gamma hyperparameters are pre-specified. Alternatively, a Wishart distribution and its hyperparameters could be specified for the prior precision of $\boldsymbol{\gamma}_i$. The random effects $\{\boldsymbol{\gamma}_i\}$ can be sampled through the use of a rejection algorithm, and $\boldsymbol{\gamma}_0$ and the precision parameters can be updated conjugately from their full conditionals. Routine implementation of the GLMM in WinBUGS uses an adaptive rejection algorithm to sample from the random effects distribution.

4.3 Model

We describe the joint modeling of a curve and an outcome, where the likelihood of the outcome is in the exponential family. The use of the word 'outcome' does not imply a causal relationship. In fact, the method we describe is appropriate for characterizing

relationships when either the trajectory or the outcome is hypothesized to impact the other or when no causal relationship is hypothesized between the two. The relationship between progesterone and early loss is a good illustration of the case when no single causal relationship is biologically motivated.

Combining methods from Chapter 3 with Bayesian methods for generalized linear models, we obtain a model that clusters jointly the trajectory and the observed outcome. In the examples presented, the outcome model contains no covariates, but we provide the theory necessary to include covariates in the Bayesian generalized linear model. In addition, we focus on the case where each subject provides one trajectory/outcome pair but present details for the case where there are multiple pairs per subject.

4.3.1 Prior specification

The data consist of N trajectory/outcome pairs. We model the trajectory according to the methods given in Chapter 3, adding an additional nonparametric component for the outcome. The trajectory and the outcome follow multivariate normal and exponential family distributions respectively, with the likelihoods given here.

$$L(\mathbf{y}|\boldsymbol{\theta}_M, \tau_M, M, \mathbf{S}_M) \propto \tau_M^{n/2} exp\big[-\frac{\tau_M}{2}\sum_{j=1}^{r_M}\sum_{i\in I_{Mj}}(\mathbf{y}_i - \mathbf{H}_i\boldsymbol{\theta}_{Mj})'(\mathbf{y}_i - \mathbf{H}_i\boldsymbol{\theta}_{Mj})\big] \quad (4.8)$$

$$L(\mathbf{Z}|\boldsymbol{\xi}, \boldsymbol{\phi}) \propto \prod_{i=1}^{N}\prod_{j=1}^{n_i} exp\big(\phi_i^{-1}(z_{ij}\xi_{ij} - B(\xi_{ij})) + c(z_{ij}, \phi_i)\big) \quad (4.9)$$

where ϕ_i is the canonical parameter and ξ_i is the dispersion parameter for some exponential family distribution with cumulant function $B(\cdot)$. We use a generalized linear model with canonical link for the outcome, so that:

$$\boldsymbol{\xi}_i = \boldsymbol{\eta}_i \equiv \mathbf{U}_i\boldsymbol{\gamma}_i + \mathbf{J}_{n_i}a_i \quad (4.10)$$

where $\mathbf{U}_i$ is an $(n_i \times p_o)$ matrix of trajectory-specific covariates and $\mathbf{J}_{n_i}$ is an $(n_i \times 1)$ vector of ones. The $(p_o \times 1)$ vector $\boldsymbol{\gamma}_i$ describes the relationship between the covariates and the outcome, and the scalar intercept a_i is jointly modeled with the trajectory. The DP governs the joint distribution of the random coefficients and a random intercept for the outcome model. Because of this, the DP will cluster jointly based on the trajectory and the random intercept, though we expect the likelihood to be heavily dominated by the more abundant trajectory data. The following is the prior structure under model M.

$$
\begin{aligned}
\begin{pmatrix} \mathbf{b}_{Mi} \\ a_i \end{pmatrix} &\overset{iid}{\sim} G_M, \;\; i = 1, \ldots, N \\
G_M &\sim DP(\alpha \, G_{M0}) \\
G_{M0} &= N_{k_M+1}\Bigg(\begin{pmatrix} \boldsymbol{\beta}_M \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_M \boldsymbol{\Delta}_M & 0 \\ 0 & \nu \end{pmatrix}^{-1}\Bigg) \\
\boldsymbol{\Delta}_M &= \text{diag}(\boldsymbol{\delta}_M) \\
\boldsymbol{\beta}_M &\sim N_{k_M}(\mathbf{0}, \tau_M^{-1}\lambda_M^{-1}\mathbf{I}_{k_M}) \\
\pi(\tau_M, \lambda_M, \boldsymbol{\delta}_M) &\propto \tau_M^{a_\tau - 1} exp(-b_\tau \tau_M) \lambda_M^{a_\lambda - 1} exp(-b_\lambda \lambda_M) \prod_{l=1}^{k_M} \big(\delta_{Ml}^{a_\delta - 1} exp(-b_\delta \delta_{Ml})\big) \\
\pi(\nu) &\propto \nu^{a_\nu - 1} exp(-b_\nu \nu)
\end{aligned}
$$

where α, a_ν, b_ν, a_τ, b_τ, a_λ, b_λ, a_δ and b_δ are pre-specified hyperparameters constant across models. This prior structure is complete if there are no covariates in the outcome model, that is $p_o = 0$ and the matrices $\mathbf{U}_i$ in (4.10) are empty. To include covariates

in the GLMM for the outcome, we specify the following additional priors:

$$\begin{aligned}
\boldsymbol{\gamma}_i &\overset{iid}{\sim} N(\gamma_0, \text{diag}(\boldsymbol{\psi})^{-1}) \\
\boldsymbol{\gamma}_0 &\sim N(\mathbf{0}, \omega^{-1} I_{p_o}) \\
\pi(\omega, \boldsymbol{\psi}) &\propto \omega^{a_\omega - 1} exp(-b_\omega \omega) \prod_{l=1}^{p_o} \left(\psi_l^{a_\psi - 1} exp(-b_\psi \psi_l) \right)
\end{aligned}$$

where $a_\omega,\ b_\omega,\ a_\psi$ and b_ψ are pre-specified hyperparameters constant across models and $\boldsymbol{\gamma}$ is a $p_o \times 1$ vector describing the relationship between the covariates in the outcome model and the outcome.

The DP naturally clusters the observations into groups, so that there are $r \leq N$ distinct values of $\{\mathbf{b}_i, a_i\}$, which are given in the set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_r\}$. When we exclude subject i there are $r^{(i)} \leq r$ distinct values, denoted $\boldsymbol{\theta}^{(i)} = \{\boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{r^{(i)}}^{(i)}\}$

The above specification yields the following conditional joint posterior of the random effects for subject i. The model indicator is suppressed for notational simplicity.

$$\mathbf{b}_i, a_i | M, \mathbf{y}, \mathbf{z}, \boldsymbol{\beta}, \lambda, \tau, \boldsymbol{\delta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\theta}^{(i)} \sim q_{i,0} G_{i,0} + \sum_{j=1}^{r^{(i)}} q_{i,j} \delta_{\boldsymbol{\theta}_j^{(i)}}$$

where $\delta_{\boldsymbol{\theta}_j^{(i)}}$ is a point mass at $\boldsymbol{\theta}_j^{(i)}$, and $G_{i,0}$ is the full joint posterior of $(\mathbf{b}_i, a_i)$ under the base prior, G_0. In other terms, $dG_{i,0}(\mathbf{b}_i) \propto f_i(\mathbf{y}_i, \mathbf{z}_i | \mathbf{b}_i, a_i) dG_0$, where $f_i(\mathbf{y}_i, \mathbf{z}_i | \mathbf{b}_i, a_i)$ is the data likelihood for subject i. The mixing weights are given by:

$$q_{i,j} \propto \begin{cases} \alpha h_i(\mathbf{y}_i, \mathbf{z}_i) & \text{if } j = 0 \\ n_j^{(i)} f_j(\mathbf{y}_i, \mathbf{z}_i | \boldsymbol{\theta}_j) & \text{if } j > 0 \end{cases} \tag{4.11}$$

$$h_i(\mathbf{y}_i, \mathbf{z}_i) = \int f_i(\mathbf{y}_i, \mathbf{z}_i | \mathbf{b}_i, a_i) dG_0(\mathbf{b}_i, a_i) \tag{4.12}$$

where $f_i(\mathbf{y}_i, \mathbf{z}_i | \boldsymbol{\theta}_j)$ is the joint likelihood of the trajectory and the outcome for subject i. If the joint likelihood $f_i(\mathbf{y}_i, \mathbf{z}_i | \mathbf{b}_i, a_i)$ is normal, then we have conjugacy, yielding a

closed form for $G_{i,0}$ and $h_i(\mathbf{y}_i, \mathbf{z}_i)$. In the progesterone example, the outcome (early loss) is binary, so a rejection algorithm can be used to sample from the posterior.

The random effects $\{\mathbf{b}_i, a_i\}$ could be sampled directly for each subject, but computational efficiency is improved by exploiting the clustering behavior of the DP and instead sampling the cluster allocation $\mathbf{S}$ and then the distinct random effects $\boldsymbol{\theta}$ from their full conditional distributions (West et al., 1994). The cluster indicator for subject i has the following full conditional posterior distribution:

$$p(S_i = j|\mathbf{y}, \mathbf{z}, \mathbf{b}^{(i)}, \mathbf{a}^{(i)}, \mathbf{S}^{(i)}, r^{(i)}) = q_{i,j} \quad \text{for } i = 1, \ldots, N$$

$S_i = 0$ implies the creation of a new cluster containing only subject i. A corresponding new value of $\{\mathbf{b}_i, a_i\}$ is drawn from $G_{i,0}$, and $\boldsymbol{\theta}$ and r are updated appropriately.

4.4 Posterior computation

Computation is similar to that for the trajectory-only model in Chapter 3. This section contains a description of the MCMC algorithm used to update from the posterior distributions of the parameters. These steps are based on the likelihood and priors given in Section 4.3.1. At each iteration, we have a current spline model, M, and current values of the parameters $\{\boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta}, \tau, \lambda, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\gamma}_0, \omega, \nu, \boldsymbol{\psi}\}$. These steps demonstrate how to update the model and then update the parameters from their full conditionals.

Step 1: Update spline model

Propose a change to M by either adding, removing, or altering a basis function. Accept or reject this change according to the appropriate acceptance probability.

Green (1995) proposed the RJMCMC sampler as a generalization of the Metropolis-Hastings algorithm (Hastings, 1970) for a parameter of varying dimension. Here, we use it to update the spline model, where the dimension varies because the number of basis

functions can change. At each iteration, a proposal is made to change the current model, M to a new model, M'. The proposal is accepted with probability $Q(M', M)$, which must meet certain regularity conditions in order to sample from the target distribution of interest. To minimize sample autocorrelation, it is optimal to make the acceptance probability as large as possible subject to these regularity conditions (Percy, 1973). Here, the optimal probability for the RJ sampler takes the form (Green, 1995; Denison et al., 2002):

$$Q(M', M) = min\Big[1, \frac{p(\mathbf{y}|M')}{p(\mathbf{y}|M)} \times R\Big] \tag{4.13}$$

where $p(\mathbf{y}|M)$ is the marginal likelihood under model M and R is the ratio of proposing the current move type (add, remove, or alter) to the probability of proposing the reverse move type starting at M'. Thus, the acceptance probability is the product of the likelihood ratio and a known constant. However, the acceptance probability in (4.13) is not appropriate under the current model because we can not calculate the likelihood ratio. However, $p(\mathbf{y}|M, \boldsymbol{\delta}_M, \lambda_M, \mathbf{S}_M)$ does have closed form. Chapter 3 outlines conditions under which alternative acceptance probabilities are valid, and per those results we use:

$$Q(M', M) = min\Big[1, \frac{p(\mathbf{y}|M', \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)}{p(\mathbf{y}|M, \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)} \times R\Big] \tag{4.14}$$

where we're conditioning on the current values $\mathbf{S}_M$ and λ_M. Recalling that the dimension of $\boldsymbol{\delta}_M$ is equal to the dimension on the model, we let $\boldsymbol{\delta}_{adj}$ be a subvector of δ_M with number of elements equal to the minimum of the dimensions of M and M'. The reasoning behind this has been presented in detail in Chapter 3. In summary, it is nonsensical to condition on a parameter that is larger than that allowed in the model and conditioning on the exact same parameter values in the numerator and denominator leads to a valid acceptance probability. If the proposed change to the model

is addition or removal of a basis, then either the numerator or denominator (but not both) will have closed form. The other will be a one-dimensional integral, which we estimate using a Laplace approximation. Further details on the approximation were given in Chapter 1.

From this point forward, we suppress the model indicator subscript for notational simplicity. If we have accepted a new model of different dimension than the old model, we update the current values of $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\theta}$ so that they have the correct dimension, initializing any new parameters to pre-specified values.

Step 2: Update $\mathbf{b}$ *and* $\mathbf{a}$ *from their full conditionals.*

Updating the random effects that under a DP prior is equivalent to updating the cluster allocation $\mathbf{S}$ and the set of distinct random effects $\boldsymbol{\theta}$.

First update $\mathbf{S}$, one subject at a time:

$$p(S_i = j|\mathbf{y}, \mathbf{z}, \mathbf{b}^{(i)}, \mathbf{S}^{(i)}, r^{(i)}) = q_{i,j} \quad \text{for } i = 1, \ldots, N; j = 1, \ldots, r.$$

where $q_{i,j}$ is given in (4.11) and depends on the likelihood and $h_i(\mathbf{y}_i, \mathbf{z}_i)$ is given in (4.12). Let g_0 be the density associated with the base distribution G_0. Under the base prior we've specified, the trajectory parameters and the outcome parameter are independent so that $g_0 = g_{0b}g_{0a}$, where g_{0b} is the base prior density of the random spline coefficients for the trajectory model and g_{0a} is the base prior density for the random intercept in the outcome model. We let G_{0a} and G_{0b} denote the distributions corresponding to these two densities. Because $\mathbf{y}$ and $\mathbf{z}$ are a priori independent given their subject-specific parameters, we can write:

$$h_i(\mathbf{y}_i, \mathbf{z}_i) = \int f_i(\mathbf{y}_i|\mathbf{b}_i) dG_{0\mathbf{b}}(\mathbf{b}_i) \int f_i(\mathbf{z}_i|a_i) dG_{0a}(a_i) = h_i(\mathbf{y}_i)h_i(\mathbf{z}_i)$$

Chapter 3 showed that the trajectory portion, $h_i(\mathbf{y}_i)$, has closed form. The outcome

portion, $h_i(\mathbf{z}_i)$, can be written as a one dimensional integral.

$$\begin{aligned}
h_i(\mathbf{y}_i) &= \frac{\tau^{\frac{n_i}{2}}|\boldsymbol{\Delta}|^{\frac{1}{2}}(2\pi)^{-\frac{n_i}{2}}}{|\mathbf{H}_i'\mathbf{H}_i+\boldsymbol{\Delta}|^{\frac{1}{2}}} exp\big(\frac{\tau}{2}[(\mathbf{H}_i'\mathbf{y}_i+\boldsymbol{\Delta}\boldsymbol{\beta})'(\mathbf{H}_i'\mathbf{H}_i+\boldsymbol{\Delta})^{-1}(\mathbf{H}_i'\mathbf{y}_i+\boldsymbol{\Delta}\boldsymbol{\beta})-(\mathbf{y}_i'\mathbf{y}_i+\boldsymbol{\beta}'\boldsymbol{\Delta}\boldsymbol{\beta})]\big) \\
h_i(\mathbf{z}_i) &= \int f_i(\mathbf{z}_i|a_i)dG_{0a}(a_i) \\
&= \int exp\big(\sum_{j=1}^{n_i}(z_{ij}(\mathbf{u}_{ij}\boldsymbol{\gamma}_i+a_i))-B(\mathbf{u}_{ij}\boldsymbol{\gamma}_i+a_i)+c(z_{ij})\big)\big(\frac{\nu}{2\pi}\big)^{\frac{1}{2}}exp(-\frac{\nu}{2}a_i^2)da_i
\end{aligned}$$

If the outcome likelihood is normal, then $h_i(\mathbf{z}_i)$ has closed form. For other exponential family likelihoods, we use a normal approximation to evaluate $h_i(\mathbf{z}_i)$ for the desired functions $B(\cdot)$ and $c(\cdot)$. In the Bernoulli case, $B(x) = log(1+exp(x))$ and $c(x) = 0$.

Next we update $\boldsymbol{\theta}$ given the new $\mathbf{S}$. For a given cluster j, $\boldsymbol{\theta}_j$ contains $k+1$ elements. The first k elements, $\boldsymbol{\theta}_{j,1:k}$, correspond to the random slopes used to describe the trajectory for members of cluster j. The remaining element, $\theta_{j,k+1}$, is the random intercept for the outcome model for members of cluster j. To update $\boldsymbol{\theta}_j$ from the full conditional, we sample from the base prior updated with the trajectory and outcome data for all subjects in cluster j, for $j = 1, \ldots, r$.

$$p(\boldsymbol{\theta}_j|\mathbf{S},\mathbf{y},\mathbf{z},\ldots) \propto \prod_{i\in\mathcal{I}_j} f_i(\mathbf{y}_i,\mathbf{z}_i|\boldsymbol{\theta}_j)g_0(\boldsymbol{\theta}_j)$$

where $\mathcal{I}_j$ is the set of subjects in cluster j. Since the likelihoods of $\mathbf{y}_i$ and $\mathbf{z}_i$ are independent, we have:

$$p(\boldsymbol{\theta}_j|\mathbf{S},\mathbf{y},\mathbf{z},\ldots) \propto \big(\prod_{i\in\mathcal{I}_j} f_i(\mathbf{z}_i|\boldsymbol{\theta}_{j,k+1})g_{0a}(\boldsymbol{\theta}_{j,k+1})\big)\big(\prod_{i\in\mathcal{I}_j} f_i(\mathbf{y}_i|\boldsymbol{\theta}_{j,1:k})g_{0b}(\boldsymbol{\theta}_{j,1:k})\big)$$

Thus, for each cluster j, we can sample $\boldsymbol{\theta}_{j,1:k}$ independently of $\boldsymbol{\theta}_{j,k+1}$. The normal likelihood for the trajectory data yields conjugacy with the normal base prior and can

sample $\boldsymbol{\theta}_{j,1:k}$ from the following full conditional:

$$p(\boldsymbol{\theta}_{j,1:k}|\mathbf{y},\mathbf{S},r) = N\Big((\boldsymbol{\Delta}+\sum_{i\in\mathcal{I}_j}\mathbf{H}_i'\mathbf{y}_i)^{-1}(\boldsymbol{\Delta}\boldsymbol{\beta}+\sum_{i\in\mathcal{I}_j}\mathbf{H}_i'\mathbf{H}_i), \tau^{-1}(\boldsymbol{\Delta}+\sum_{i\in\mathcal{I}_j}\mathbf{H}_i'\mathbf{y}_i)^{-1}\Big)$$

The full conditional of $\boldsymbol{\theta}_{j,k+1}$ under the base prior is:

$$p(\theta_{j,k+1}|\mathbf{S},\mathbf{y},\mathbf{z},\dots) \propto exp(-\frac{\nu}{2}\theta_{j,k+1}^2)\prod_{i\in\mathcal{I}_j}\prod_{j=1}^{n_i} exp(z_{ij}\theta_{j,k+1} - B(\mathbf{u}_{ij}\boldsymbol{\gamma}_i+\theta_{j,k+1}))$$

If the likelihood of the outcome is not normal, we may not be able to sample from this directly. Instead, a Metropolis step is used to sample from the full conditional under the appropriate $B(\cdot)$ and $c(\cdot,\cdot)$. For purposes of illustration, the full conditionals in the following steps assume $c(\cdot,\cdot)\equiv 0$, which is the case in the Bernoulli distribution. Similar calculations can be used to define a sampling scheme when $c(\cdot,\cdot)\neq 0$.

Step 3: Update hyperparameters for longitudinal trajectory

Update $\boldsymbol{\beta}$, τ, λ, and $\boldsymbol{\delta}$ from their full conditionals. These are all conjugate under the prior structure in Section 4.3.1.

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{\theta},\mathbf{S},\boldsymbol{\delta},\lambda) &= N\Big((\lambda\mathbf{I}+r\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}\sum_{j=1}^{r}\boldsymbol{\theta}_{j,1:k}, \tau^{-1}(\lambda\mathbf{I}+r\boldsymbol{\Delta})^{-1}\Big) \\
\lambda|\boldsymbol{\beta},\mathbf{b},\boldsymbol{\delta},\tau &\sim Gamma\Big(a_\lambda+\frac{k}{2}, b_\lambda+\frac{\tau\boldsymbol{\beta}'\boldsymbol{\beta}}{2}\Big) \\
\delta_l|\boldsymbol{\beta},\mathbf{b},\boldsymbol{\delta}_{-l},\lambda,\tau &\sim Gamma\Big(a_\delta+\frac{r}{2}, b_\delta+\frac{\tau}{2}\sum_{j=1}^{r}(\theta_{jl}-\beta_l)^2\Big) \qquad l=1,\dots,k \\
\tau|\mathbf{y},\boldsymbol{\theta},\mathbf{S}\boldsymbol{\delta},\lambda,\nu &\sim Gamma\Big(a_\tau+\frac{N+(r+1)k}{2}, b_\tau+\frac{1}{2}\Big(\sum_{i=1}^{N}(\mathbf{y}_i-\boldsymbol{H}_i\mathbf{b}_i)'(\mathbf{y}_i-\boldsymbol{H}_i\mathbf{b}_i) \\
&\qquad +\lambda\boldsymbol{\beta}'\boldsymbol{\beta}+\sum_{j=1}^{r}(\boldsymbol{\theta}_{j,1:k}-\boldsymbol{\beta})'\boldsymbol{\Delta}(\boldsymbol{\theta}_{j,1:k}-\boldsymbol{\beta})\Big)\Big)
\end{aligned} \tag{4.15}$$

Step 4: Update hyperparameters for outcome model

Update ν, the precision of the random intercept under the base prior. If the outcome model contains any covariates, then update $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}_0$, ω, and $\boldsymbol{\psi}$ from their full conditionals. Under the prior structure described, all can be updated conjugately except for the subject-specific slope vectors, which can be updated using Metropolis steps.

$$\begin{aligned}
\nu|\ldots &\sim Gamma\Big(a_\nu + \frac{r}{2}, b_\nu + \frac{\tau \sum_{j=1}^{r} \theta_{j,k+1}^2}{2}\Big) \\
p(\boldsymbol{\gamma}_i|\ldots) &\propto exp\big(\sum_{j=1}^{n_i} z_{ij}\mathbf{u}_{ij}\boldsymbol{\gamma}_i - \sum_{j=1}^{n_i} B(\mathbf{u}_{ij}\boldsymbol{\gamma}_i + a_i) - \frac{1}{2}(\boldsymbol{\gamma}_i'\boldsymbol{\Psi}\boldsymbol{\gamma}_i + 2\boldsymbol{\gamma}_i'\boldsymbol{\Psi}\boldsymbol{\gamma}_0)\big) \\
\psi_l|\ldots &\sim Gamma\Big(a_\psi + \frac{N}{2}, b_\psi + \frac{1}{2}\sum_{i=1}^{N}(\gamma_{il} - \gamma_{0l})^2\Big) \qquad l = 1, \ldots, p_o \\
\omega|\ldots &\sim Gamma\Big(a_\omega + \frac{k}{2}, b_\omega + \frac{\boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0}{2}\Big)
\end{aligned}$$

4.5 Simulated data example

Data were simulated from trajectories centered around one of three parametric curves. Each simulated trajectory was also assigned a binary outcome status, either 1 or 0. The data were actually simulated from four distinct groups, where two groups had the same underlying trajectory but different response probabilities. Figure 4.1 shows the simulated data, the underlying trajectories, and the probabilities that a member of each of the 4 groups will have outcome equal to 1. Each of the four groups contained 25 trajectories with 10 measurements at varied timepoints. We ran the algorithm for 25,000 iterations after 2,000 burn-in. Examination of traceplots of precision parameters and the number of basis functions showed no evidence against convergence. As in Chapter 3, we classified together observations that appeared in the same cluster in 40% or more of the samples. We calculated the mean trajectory for each cluster as well as the modeled probability that a trajectory in each cluster had outcome equal to 1. Under this logistic model, the modeled outcome probability is the mean over

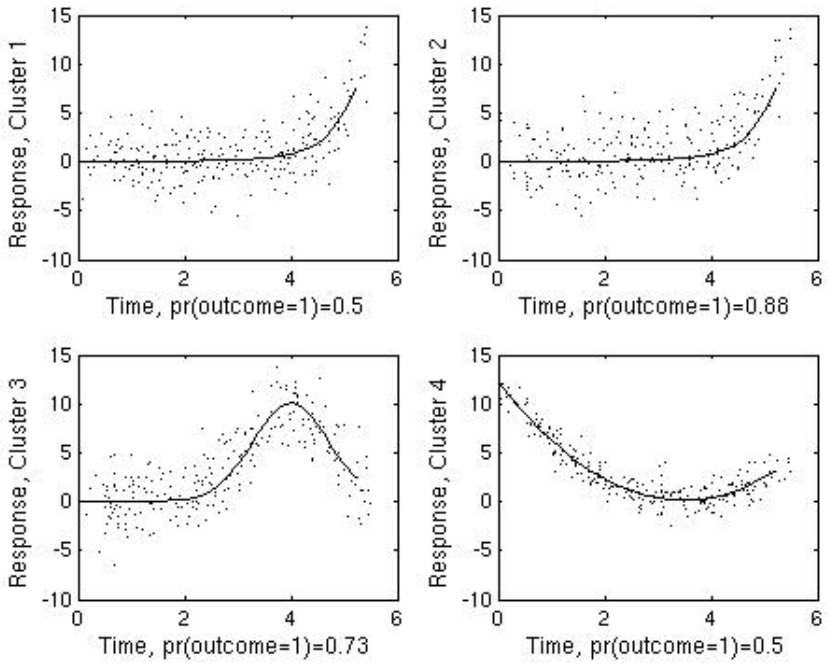

FIGURE 4.1: Underlying population curves (lines) and data (points) for each of the four clusters along with outcome probabilities. The top two plots have the same underlying trajectory with different outcome probabilities.

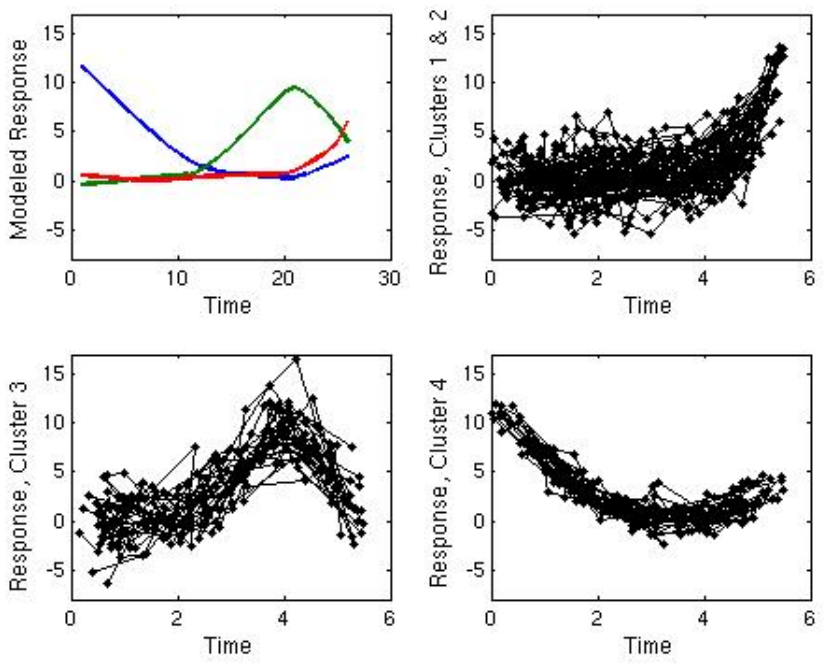

FIGURE 4.2: The plot in the first quadrant is of the mean trajectory for the three final clusters. The three remaining plots contain all trajectory data for the three final clusters.

all samples of the logit of the outcome model's random intercept, a_i. Figure 4.2 gives the mean trajectory for the three final clusters as well as the data from each. We estimated the outcome probabilities and compared them to the underlying population probabilities. The estimates and 95% credible intervals are given in Table 4.1. The credible interval contained the true value in all cases. The trajectory clusters were correctly identified. Within the MCMC samples, observations were misclassified only very rarely. This simulation shows that the model clearly discriminated among trajec-

TABLE 4.1: True population outcome probabilities and estimated probabilities with 95% credible intervals for each of the final clusters.

	n	Population probability	Mean sampled probability [95% CI]
Clusters 1 & 2	50	0.69	0.73 [0.60, 0.83]
Cluster 3	25	0.73	0.76 [0.59, 0.89]
Cluster 4	25	0.50	0.34 [0.18, 0.52]

tories of different shapes and provided an accurate estimate of the random intercept in the outcome model for each cluster. As expected, the model did not distinguish between two clusters with similar trajectories and different response probabilities. This is because there is only one observation per subject and the response was binary. As a mixture of Bernoulli distributions is also Bernoulli, two groups of people with the same underlying trajectory and different outcome probabilities appear as one large group, with an outcome probability somewhere between that of the two smaller groups. Other exponential family distributions without this property may affect clustering differently.

4.6 Early Pregnancy Study example

We applied the joint model to conception cycles from the NC-Early Pregnancy study. In those cycles labeled clinical pregnancies, the embryo appeared to the investigators, based on hCG, to have survived at least six weeks beyond the last menstrual period. Cycles in which a detectable hCG rise occurred but did not last more than six weeks beyond the last menstrual period (LMP) were labeled 'early losses'. The data consisted of 165 conception cycles, 47 of which resulted in early losses.

To illustrate the joint model for a trajectory and an outcome, we apply it to progesterone data for the early losses and the clinical pregnancies, with early loss status serving as the binary outcome for each cycle. The trajectory was defined to begin at the ratio-determined day of ovulation and to last for up to 40 days. For purposes of illustration, the only covariate we used was day relative to ovulation. We could, however, have incorporated other reference-point based covariates such as day relative to implantation of the conceptus. We could also have included non-reference point based covariates such as age or parity. Figures 4.3 and 4.4 illustrate some of the data from early losses and clinical pregnancies. They show how PdG tends to rise when concep-

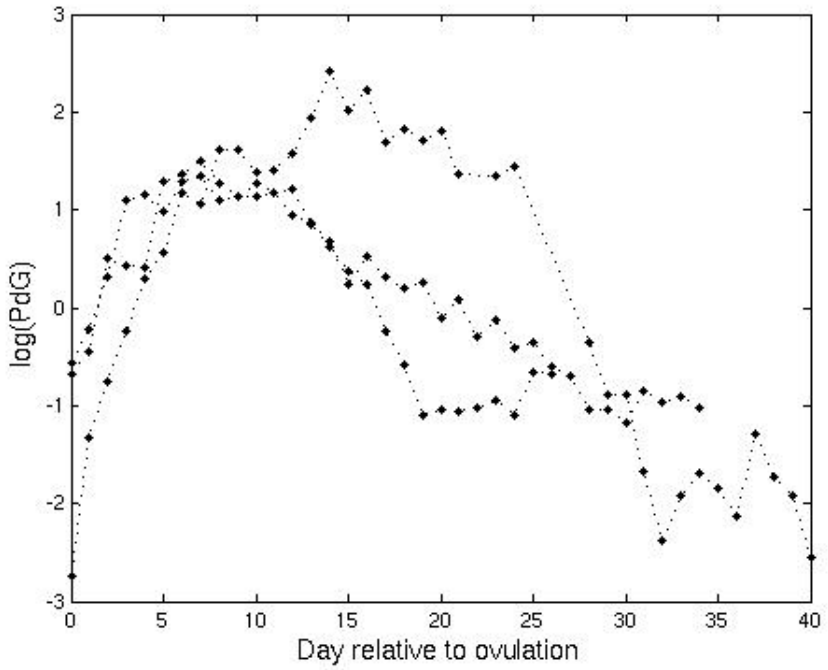

FIGURE 4.3: Progesterone data beginning at the estimated day of ovulation for three early losses.

tion occurs and then drop off if the pregnancy is lost. Using the threshold of 0.40, we sorted the 165 subjects into final clusters. There were 32 of these clusters, though 16 contained only one observation. We calculated the mean trajectory for each cluster and the mean probability that a cycle in that cluster was an early loss. Figures 4.5 and 4.6 show the data for each of the 32 clusters and the model-estimated probabilities and credible intervals that a cycle in a given cluster was an early loss. With the exception of cluster 11, which contained one early loss and two clinical pregnancies, every cluster was homogeneous with respect to early loss status. The first cluster consisted of 87 clinical pregnancies, so that 74% of all clinical pregnancies fell into one class. While both the early loss and clinical groups had outliers, the early losses were more spread out among several clusters. Ignoring the eight outliers in each group, the remaining 39 early losses were spread out among 12 clusters, whereas the 157 remaining clinical

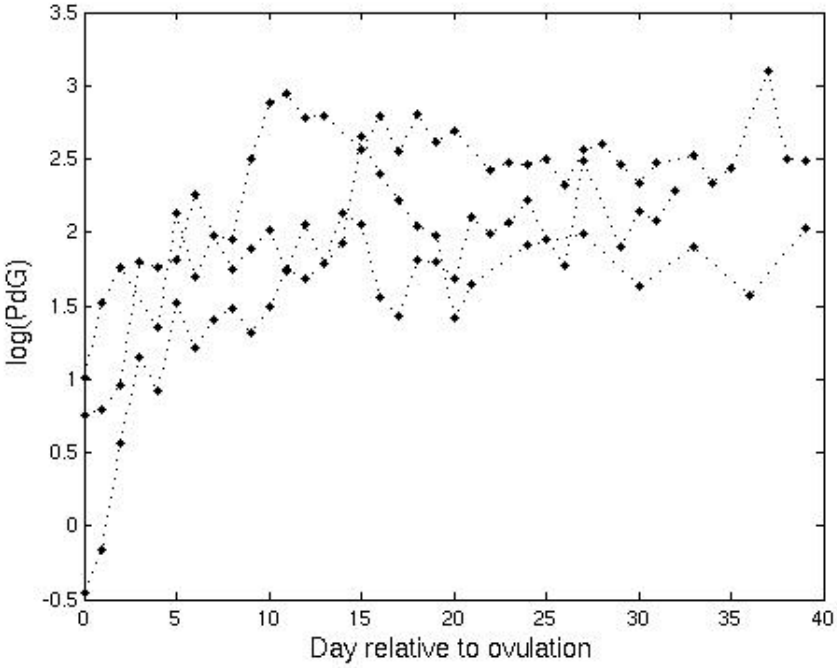

FIGURE 4.4: Progesterone data beginning at the estimated day of ovulation for three clinical pregnancies.

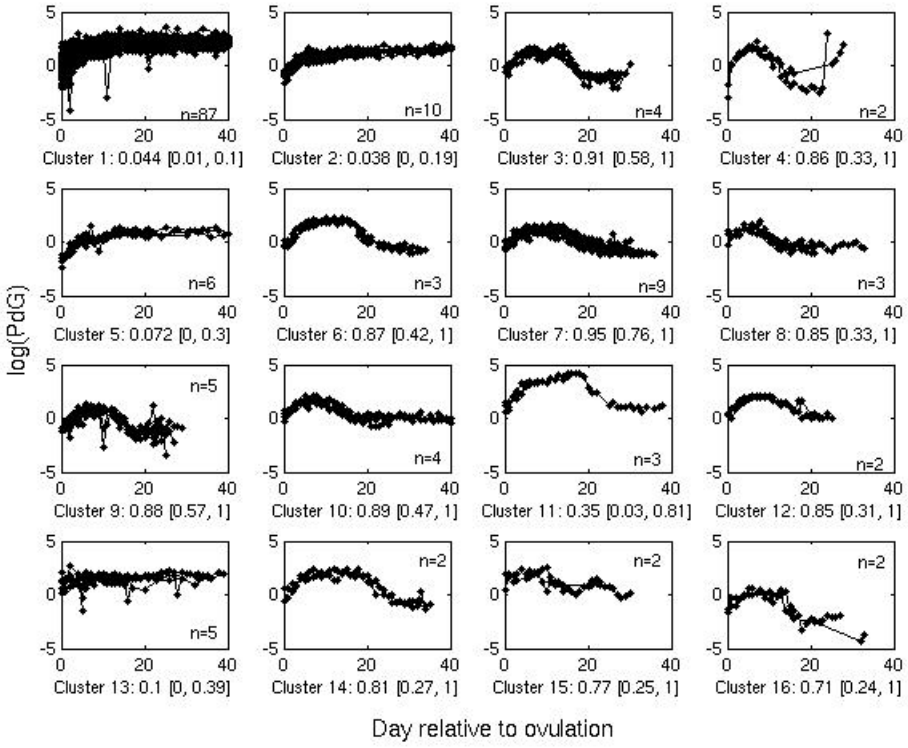

FIGURE 4.5: Data from the 16 classes containing more than one observation. Below each plot is the model-estimated probability of early loss for each cluster along with the 95% credible interval. The cluster sizes are given on the plots.

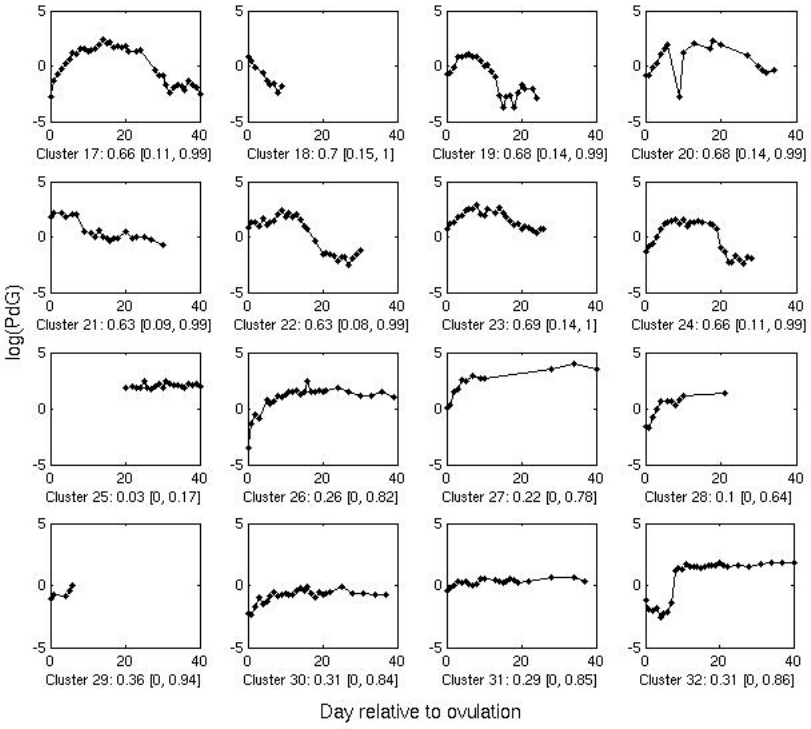

FIGURE 4.6: Data from the 16 trajectories that were not clustered with any others. Below each plot is the model-estimated probability of early loss for each trajectory along with the 95% credible interval. The first two rows of plots are EPLs, while the remaining eight are clinical pregnancies.

pregnancies were in only 5 clusters. This variation in early loss trajectories supports the hypothesis that there is no one mechanism for EPL.

To gain further insight into the performance of this model and the differences between early loss and clinical pregnancy trajectories, we fit the model from Chapter 3 to these data. In other words, we re-fit the model without including the outcome. As expected, since the clustering is based on the trajectory shapes and not the outcome, we found the same clusters of trajectories. Biologically, the separation of the clusters is interesting because the model has effectively separated hCG-determined early losses from conceptions resulting in clinical pregnancies based on the shape of the PdG trajectory.

4.7 Discussion

We've developed a model for joint regression of a trajectory and a univariate outcome. The post-ovulatory progesterone example illustrates the appropriateness of the model for clustering trajectories and for providing model-based estimates of outcome probabilities. The results support the hypothesis that the types of early loss are varied and may be due to several underlying biological mechanisms which are manifested in different hormone trajectories.

The examples focused on the one observation per subject Bernoulli case. Because a mixture of Bernoulli distributions is itself Bernoulli, the model would not have been able to form clusters based on the outcome. Thus, although the theory is very similar, the interpretation may differ substantially when the outcome was from a more complex exponential family distribution. Although we did not demonstrate it, the model allows for multiple trajectory/outcome pairs per subject. In that case, each subject's set of Bernoulli responses would be binomial, and there would be information available to

cluster based on outcome probability. However, it is still likely that the clustering would be dominated by the trajectory since there is a large amount of trajectory data and only a single outcome. If we were truly interested in clustering according to both trajectory and outcome, we could increase the weight given to the outcome likelihood in the Dirichlet process clustering.

Theory has been developed for the incorporation of covariates into the outcome model. Although we have not yet performed simulations in the setting where covariates are present, the model is a clear incorporation of methods for Bayesian generalized linear models into this clustering framework. However, the interpretation of the random intercept and of the nature of observations with similar random intercepts will change in the presence of covariates.

CHAPTER 5

CONCLUDING REMARKS

5.1 Summary

Motivated by reproductive hormone trajectories, this dissertation has proposed new methodology with applications widely beyond the menstrual hormone setting. Through the identification of reference points in longitudinal data, we have developed statistical methodology that is applicable beyond the longitudinal setting.

Chapter 2 described the first mixed model in the setting where the dimension of the underlying spline model varies and the random effects are modeled flexibly. In addition, this model is superior to most nonparametric curve models, which tend to require all curves be observed over the same covariate space. The generalization of longitudinal data to multiple reference point data has applications in longitudinal medical studies, and the resulting spline model is widely applicable in non-longitudinal settings.

Chapter 3 described a new method for putting a nonparametric distribution on a set of nonparametrically modeled curves, yielding a very flexible method that is useful for both describing curves and for dividing curves into clusters. Chapter 4 proposed an extension where an outcome can be modeled jointly with the curve itself.

5.2 Computational Notes

The MRF model described in Chapter 1 was implemented in R, and all other computation was done in Matlab. One average, the sample collection took about 36 hours. The model in Chapter 2 was the slowest due to the need to calculate a random effect for every subject at every iteration. The model in Chapter 3 was the fastest.

The Laplace approximation was introduced in order to maintain the theoretical integrity of the reversible jump sampler. However, I ran the model in Chapter 2 without the approximation (i.e. just using the naive acceptance probability in (3.17)) and the results were identical. Unable to come up with a theoretical justification for that acceptance probability, I continued to use the time-consuming one.

5.3 Methodology for menstrual hormone data

This section contains a summary of some issues that arose in deciding how to best develop these models to suit the needs of menstrual hormone data. The first issue that arose was the need to control for intercourse in order to best predict conception. In Chapter 2, we divided conception and non-conception cycles without controlling for intercourse patterns. This affected our interpretation of the results. We made inference about the differences between the set of non-conception and the set of conception cycles. The set of non-conception cycles probably included low fertility cycles where well-timed intercourse did not lead to conception and also normal fertility cycles where there was no well-timed intercourse.

Unable to assume that the non-conception cycles were all of low fertility, we could not make assertions about PdG differences and fertility. However, our findings were consistent with previous work comparing conception cycles and non-conception cycles with well-timed intercourse, which indicated that low fertility cycles have very low

mid-luteal progesterone (Baird et al., 1997), and we found that the cycles with the lowest mid-luteal PdG were unlikely to be conception cycles. These cycles with very low midluteal PdG likely corresponded to the subset of non-conception cycles that were of low fertility.

We expected the model averaging to produce smooth curves, but instead many of the real-data plots have changepoints the day before ovulation, although the rest of the plot appears to be a smooth curve (see Figures 2.5 and 3.5 for examples). Because the simulated data produced smooth plots, this is more likely an artifact of the ovulation estimation method than a reflection of the model. The marker of ovulation was based on a drop in the estrogen to progesterone ratio, which likely corresponds to a rise in progesterone. This does not affect the quality of the presented results, but was carefully considered when interpreting those results.

In discussing methods for clustering menstrual cycle data, we repeatedly encountered the issue of whether it is more appropriate (i.e. biologically informative) to cluster women or to cluster menstrual cycles. We chose to cluster according to cycle in Chapters 3 and 4 (although the early loss and conception data for Chapter 4 were such that very few women contributed more than one cycle, so the clustering of cycles and women were virtually the same thing).

Although we did not perform any predictive analyses, the identification of patterns of menstrual cycle hormones that are associated with certain outcomes (e.g. early loss, high or low fertility) could help to prospectively find cycles when a woman is most likely to achieve pregnancy. This cycle-specific model was the focus of our methods in Chapters 3 and 4.

The model developed in Chapter 2, however, was for woman-specific characterization of menstrual cycle patterns. Rather than modeling each cycle individually, we modeled a central tendency for all cycles from a woman. So if we ran the clustering

algorithm from Chapter 2 on the data from Chapter 1, which contained multiple cycles per woman, we would be identifying clusters of women whose cycles tended to follow the same general patterns.

It would be possible to include multiple cycles per woman while allowing for cycle-to-cycle variation. Another level can be added to the hierarchy in the Bayesian model in Chapter 2, allowing for cycle-specific coefficients to be centered around the woman-specific basis coefficients. A similar extension of the Dirichlet process model in Chapter 3 to the case with multiple cycles per woman and cycle-specific basis coefficients is straightforward if the goal is to identify clusters of women, but less so if the goal is to classify cycles.

5.4 Potential Applications & Future Work

Within the setting of menstrual hormone data, it may be interesting to see how hormone trajectories predict fertility. This is complicated, however, by the need to control for timing of intercourse. Several models are available to estimate probability of conception based on intercourse timing (Barrett and Marshall, 1969; Wilcox et al., 1995; Stanford et al., 2003; Dunson and Stanford, 2005) and to subsequently quantify the relative fertility of cycles. A joint model for conception probability due to intercourse and hormone trajectory could be developed by combining these two model types.

This entire work has focused on the case where the response (progesterone) is normally distributed given the model and the covariates. Holmes and Mallick (2003) extended the original Holmes and Mallick (2001) model to include non-normal responses. We could potentially apply that extension to these models. This would be helpful in the case of menstrual data for modeling trajectories of discrete responses or of responses that are clearly not normally distributed. It is of interest to collect daily information

about cervical mucus, because changes in mucus are known to reflect the timing of ovulation as well as control the movement of sperm through the cervix (Billing et al., 1989; WHO, 1983; Katz, 1991; Kunz et al., 1997; Bigelow et al., 2004). Mucus is often categorized based on texture and color (see Colombo and Masarotto (2000) for an example). An extension of this model could follow the trajectory of a categorical variable like mucus.

A potential application of the multiple reference point approach is found in another area of reproductive epidemiology. Gestational age at birth is a strong predictor of stillbirth and infant death. However, researchers debate over the most appropriate way to analyze gestational-age-specific mortality. For example, should an infant born at 28 weeks who survives for four weeks be considered as 28-weeks gestation? Or does he have something in common with the infant who dies at birth after 32 weeks gestation?

Cheung (2004) discuss various methods for calculating gestational-age-specific mortality, and the advantages and disadvantages of each. In a commentary on Cheung (2004), Wilcox and Weinberg (2004) point out that birth is a traumatic event and must certainly be taken into account when looking at gestational-age-specific mortality. Using the concept of multiple reference points, we could potentially examine mortality according to both gestational age and time since birth, both providing estimates of mortality and considering the question of whether both time scales are important.

We have noted the applicability of this method to functional and spatial data. Images are a type of spatial data. The ability to search through a database and identify images similar to a presented image has applications in medical imaging as well in computer-based storage of non-medical images such as photos and artwork. Goldberger et al. (2006) describe a method for searching through a database of photos to identify those most similar to one presented. In summary, they use a Gaussian mixture model to identify homogeneous regions of the image, searching through a database for those

images whose underlying mixture distributions are most similar (see also Greenspan et al. (2004)). Our method could be developed to potentially cluster images not only according to large homogeneous regions, but according to non-homogeneous regions with similar shading/color gradients.

Our joint model has exciting potential applications in curve and image classification. For example, suppose rather than modeling longitudinal curves we model the spatial variation in medical images. As we sort these images into clusters, we can also model the disease rate of each cluster. A future image of unknown disease status can be classified into a group of images with known disease status, yielding a prediction of the disease status. In general, the models that focus on clustering have many potential applications in prediction. Especially in the context of the joint model, it would be desirable to present a trajectory, assign it to a cluster, and then predict the distribution of the outcome based on the distribution on outcomes in that cluster.

APPENDIX A

Reversible Jump Acceptance Probability

Let $\mathcal{T}$ be the very large set of all basis functions we wish to consider for the piecewise linear model. Generating a new basis function corresponds to sampling from the set $\mathcal{T}$. Since this proposal process is discrete, the need for a Jacobian in the acceptance probability is eliminated. If T, the number of elements in $\mathcal{T}$, is so large compared to the number of iterations that the probability of ever proposing the same basis function twice throughout the course of the algorithm is effectively zero, then this discrete process is equivalent to random generation of a new basis function through a continuous process.

Recall we have specified the following prior for k_M, for all $M \in \mathcal{M}$: $p(k_M) = \binom{T}{k_M - 1}^{-1} K^{-1}$. A priori, all basis functions are presumed equally likely. So the prior probability of any model M is $p(M) = p(k_M) = \binom{T}{k_M - 1}^{-1} K^{-1}$.

At each iteration we propose a change to the current model. If the current model is of dimension k (suppressing the model indicator subscript for notational convenience), we propose to add a new basis (birth) with probability b_k, we propose to remove a basis (death) with probability d_k, and we propose to alter a basis with probability $1 - d_k - b_k$. All acceptable move types are assigned equal probability, so $b_k = d_k = 1/3$ for all k except that $b_1 = d_K = 1/2$, and $b_K = d_1 = 0$.

Consider, for example, a proposal to change model M to the $(k+1)$-dimensional model M' (birth proposal). The ratio of the model priors is $\frac{p(M')}{p(M)} = \frac{k+1}{T-k}$. The proposal

density, $S(M') = b_k \times \frac{1}{T-k}$, is the probability of selecting a birth move multiplied by the probability of adding the correct basis to add. The reverse proposal density, $S(M) = d_{k+1} \times \frac{1}{k+1}$, is the probability of proposing a death from model M' multiplied by the probability of removing the basis function that would yield model M.

Substituting into (3.10), we get the following acceptance probability for a birth proposal.

$$Q = min\Big[1, \frac{p(\mathbf{y}|M')d_{k+1}}{p(\mathbf{y}|M)b_k}\Big] \tag{A.1}$$

Similar calculations for proposed basis removals and alterations show that the acceptance probability has the general form given in (3.11).

REFERENCES

Akay, M. (2003). Wavelets for image-analysis. *IEEE Engineering in Medicine and Biology* **14**, 534–535.

Allcroft, D. and Glasbey, G. (2003). A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society, Series C* **14**, 534–535.

Alliende, M. (2003). Mean versus individual hormonal profiles in the menstrual cycle. *Fertility and Sterility* **78**, 90–95.

Antoniak, C. (1974). Mixtures of Dirichlet processes with application to nonparametric problems. *The Annals of Statistics* **2**, 1152–1174.

Assunção, R., Potter, J. and Cavenagh, S. (2002). A Bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine* **21**, 2057–2075.

Baird, D., Weinberg, C., Zhou, H., Kamel, F., McConnaughey, D., Kesner, J. and Wilcox, A. (1999). Preimplantation urinary hormone profiles and the probability of conception in healthy women. *Fertility and Sterility* **71**, 40–49.

Baird, D., Wilcox, A., Weinberg, C., Kamel, F., McConnaughey, D., Musey, P. and Collins, D. (1997). Preimplantation hormonal differences between the conception and non-conception menstrual cycles of 32 normal women. *Human Reproduction* **12**, 2607–2613.

Barrett, J. and Marshall, J. (1969). The risk of conception on different days of the menstrual cycle. *Population studies* **23**, 455–461.

Berthelson, K. and Moller, J. (2003). Likelihood and non-parametric Bayesian MCMC inference for spatial plot processes based on perfect simulation and path sampling. *Scandinavian Journal of Statistics* **30**, 549–564.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

Besag, J. (1986). On the statistical-analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B* **48**, 259–302.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3–41.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregression. *Biometrika* **82**, 733–746.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with 2 applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–20.

Bigelow, J., Dunson, D., Stanford, J., Ecochard, R., Gnoth, C. and Colombo, B. (2004). Mucus observations in the fertile window: a better predictor of conception than timing of intercourse. *Human Reproduction* **19**, 889–892.

Billing, E., Billings, J. and Catarinich, M. (1989). *Billings atlas of the Ovulation Method*. Ovulation Method Research and Reference centre of Australia, Melbourne, Australia.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

Brémaud, P. (1999). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer-Verlag, New York NY.

Brown, E. and Ibrahim, J. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

Brown, E., Ibrahim, J. and DeGruttola, V. (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61**, 64–73.

Brown, P., Kenward, M. and Bassett, E. (2001). Bayesian discrimination with longitudinal data. *Biostatistics* **2**, 417–432.

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

Brumback, L. and Lindstrom, M. (2004). Self modeling with flexible, random time transformations. *Biometrics* **60**, 461–470.

Bush, C. and MacEachern, S. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Cheung, Y. (2004). Gestational-age-specific mortality. *American Journal of Epidemiology* **160**, 207–210.

Chib, S. and Hamilton, B. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Biometrics* **61**, 64–73.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset

selection in wavelets. *Biometrika* **85**, 391–401.

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

Cressie, N. and Chan, N. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association* **84**, 393–401.

Curran, P. and Hussong, A. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology* **112**, 526–544.

De Iorio, M., Müller, P., Rosner, G. and MacEachern, S. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

de la Cruz, R. and Quintana, F. (2005). A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles. *http://www.mat.puc.cl/ quintana/trbp.pdf* .

Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002). *Bayesian methods for nonlinear classification and regression.* John Wiley and Sons, Chichester, West Sussex, England.

DiCiccio, T., Kass, R., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulations and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.

Dunson, D., Chen, Z. and Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59**, 521–530.

Dunson, D. and Stanford, J. (2005). Bayesian inferences on predictors of conception probabilities. *Biometrics* **61**, 126–133.

Ecochard, R., Boehringer, H., Rabilloud, M. and Marret, H. (2001). Chronological aspects of ultrasonic, hormonal, and other indirect indices of ovulation. *British Journal of Obstetrics and Gynecology* **108**, 822–829.

Ellish, N., Saboda, K., O'Connor, J., Nasca, P., Stanek, E. and Boyle, C. (1999). A prospective study of early pregnancy loss. *Human Reproduction* **11**, 406–412.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression

for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 715–745.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2**, 615–629.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Goldberger, J., Gordon, S. and Greenspan, H. (2006). An information theoretic framework for unsupervised image clustering. *IEEE Transactions on Image Processing* **15**, 449–458.

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

Greenspan, H., Dvir, G. and Rubner, Y. (2004). Context-dependent segmentation and matching in image databases. *Journal of Computer Vision and Image Understanding* **93**, 86–109.

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

Handcock, M. and Stein, M. (1993). A Bayesian analysis of kriging. *Technometrics* **35**, 403–410.

Handcock, M. and Wallis, J. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association* **89**, 368–378.

Hansen, M. and Kooperberg, C. (2002). Spline adaptation in extended linear models. *Statistical Science* **17**, 2–20.

Harlow, S.D. amd Zeger, S. (1991). An application of longitudinal methods to the analysis of menstrual diary data. *Journal of Clinical Epidemiology* **44**, 1015–1025.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Henriksen, T., Hjollund, N., Jensen, T., Bonde, J., Andersson, A., Kolstad, H., Ernst, E., Giwercman, A., Skakkebaek, N. and Olsen, J. (2004). Alcohol consumption at the time of conception and spontaneous abortion. *American Journal of Epidemiology* **160**, 661–667.

Holmes, C. and Mallick, B. (2000). Bayesian wavelet networks for nonparametric regression. *IEEE Transactions on Neural Networks* **11**, 27–35.

Holmes, C. and Mallick, B. (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society, Series B* **63**, 3–17.

Holmes, C. and Mallick, B. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *Journal of the American Statistical Association* **98**, 352–368.

Holmes, C.C. Denison, D. and Mallick, B. (2002). Accounting for model uncertainty in seemingly unrelated regressions. *Journal of Computational and Graphical Statistics* **11**, 533–551.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* **2**, 283–304.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. and Takahara, G. (2002). Independent and identically distributed monte carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association* **97**, 1154–1166.

Jain, S. and Neal, R. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics* **13**, 158–182.

James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* **64**, 411–432.

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B* **63**, 533–550.

James, G., Hastie, T. and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98**, 397–408.

Jasienska, G., Thune, I. and Ellison, P. (2000). Energetic factors, ovarian steroids and the risk of breast cancer. *European Journal of Cancer Prevention* **9**, 231–239.

Kaiser, M. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis* **73**, 199–220.

Kato, I., Toniolo, P., Koenig, K., Shore, R., Zeleniuch-Jacquotte, A., Akhmedkhanov, A. and Riboli, E. (1999). Epidemiologic correlates with menstrual cycle length in middle aged women. *European Journal of Epidemiology* **15**, 809–814.

Katz, D. (1991). Human cervical-mucus- research update. *American Journal of Obstetrics and Gynecology* **165**, 1984–1986 Part 2 Suppl S.

Kleinman, K. and Ibrahim, J. (1998). A semi-parametric Bayesian approach to the random effects model. *Biometrics* **54**, 921–938.

Knorr-Held, L., Rasser, G. and Becker, N. (2002). Disease mapping of stage-specific cancer incidence data. *Biometrics* **58**, 492–501.

Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics* **52**, 169–183.

Kunsch, H., Geman, S. and Kehagias, A. (1995). Hidden markov random fields. *Annals of Applied Probability* **5**, 577–602.

Kunz, G., Beil, D., Deiniger, H., Einspanier, A., Mall, G. and Leyendecker, G. (1997). The uterine peristaltic pump. normal and impeded sperm transport within the female genital tract. *Advances in Experimental Medicine and Biology* **424**, 267–277.

Legler, J., Davis, W., Potosky, A. and Hoffman, R. (2004). Latent variable modelling of recovery trajectories: sexual function following radical prostatectomy. *Statistics in Medicine* **23**, 2875–2893.

Liang, Y., Tayo, B., Cay, X. and Kelemen, A. (2005). Differential and trajectory methods for time course gene expression data. *Bioinformatics* **21**, 3009–3016.

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

Lipson, S. and Ellison, P. (1996). Comparison of salivary steroid profiles in naturally occurring conception and non-conception cycles. *Human Reproduction* **11**, 2090–2096.

Lower, A. and Yovich, J. (1992). The value of serum levels of oestradiol, progesterone, and β-human chorionic gonadotropin in the prediction of early pregnancy loss. *Human Reproduction* **7**, 711–717.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19**, 474–482.

Ma, P., Castillo-Davis, C., Zhong, W. and Liu, J. (2005). Curve clustering to discover patterns in time-course gene expression data. *Working paper available at: http://ilabs.inquiry.uiuc.edu/ilab/fallbiosem/ documents/2380/home/ma-et-al-2005.pdf* .

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics- Simulation and Computation* **23**, 727–

741.

MacEachern, S. (1999). *Dependent nonparametric processes*. Proceedings of the Section on Bayesian Statistical Science: American Statistical Association, Alexandria, VA.

MacEachern, S. (2001). *Decision theoretic aspects of dependent nonparametric processes*. Bayesian Methods with Applications to Science, Policy and Official Statistics, ed. E. George.

Marshall, G. and Barón, A. (2000). Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine* **19**, 1961–1981.

Massafra, C., De Felice, C., Agnusdei, D., Gioia, D. and Bagnoli, F. (1999). Androgens and osteocalcin during the menstrual cycle. *Journal of Clinical Endocrinology and Metabolism* **84**, 971–974.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, England, 2nd edition.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

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

Mukhopadhyay, S. and Gelfand, A. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association* **92**, 633–639.

Murphy, S., Bentley, G. and OHanesian, M. (1995). An analysis for menstrual data with time-varying covariates. *Statistics in Medicine* **14**, 1843–1857.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463–469.

Percy, D. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.

Percy, D. (1992). Prediction for seemingly unrelated regressions. *Journal of the Royal Statistical Society, Series B* **54**, 243–252.

Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.

Ratcliffe, S., Heller, G. and Leader, L. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine* **21**, 1115–1127.

Ratcliffe, S., Leader, L. and Heller, G. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression. *Statistics in Medicine* **21**, 1103–1114.

Ray, S. and Mallick, B. (2003). A bayesian transformation model for wavelet shrinkage. *IEEE Transactions in Image Processing* **12**, 1512–1521.

Rice, J. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* **14**, 613–629.

Romero, A., Villamayor, F., Grau, M., Sacristan, A. and Ortiz, J. (1992). Relationship between fetal weight and litter size in rats- application to reproductive toxology studies. *Reproductive Toxicology* **6**, 453–456.

Schroder, P. (1996). Wavelets in computer graphics. *Proceedings of the IEEE* **84**, 615–625.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statistica Sinica* **2**, 639–650.

Stanford, J., Smith, K. and Dunson, D. (2003). Vulvar mucus observations and the probability of pregnancy. *Obstetrics and Gynecology* **101**, 1285–1293.

Stanford, J., White, J. and Hatasaka, H. (2002). Timing intercourse to achieve pregnancy: current evidence. *Obstetrics and Gynecology* **100**, 1333–1341.

Stewart, D., Overstreet, D., Nakajima, S. and Lasley, B. (1993). Enhanced ovarian and steroid secretion before implantation in early human pregnancy. *Journal of Clinical Endocrinology and Metabolism* **76**, 1470–1476.

Sugar, C. and James, G. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association* **98**, 750–763.

Tseng, G. and Wong, W. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**, 10–16.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834.

Unzer, S., Dossantos, J., Moreira, A., Vilanova, M. and Desa, M. (1995). Alterations in plasma gonadotropin and sex steroid-levels in obese ovulatory and chronically anovulatory women. *Journal of Reproductive Medicine* **40**, 516–520.

van Zonneveld, P., Scheffer, G., Broekmans, F., Blankenstein, M., de Jong, F., Looman, C., Habbema, J. and te Velde, E. (2003). Do cycle disturbances explain the age-related decline of female fertility? Cycle characteristics of women aged over 40 years

compared with a reference population of young women. *Human Reproduction* **18**, 495–501.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Verzilli, C., Whittaker, J., Stallard, N. and Chasman, D. (2005). A hierarchical Bayesian model for predicting the functional consequences of amino-acid polymorphisms. *Journal of the Royal Statistical Society, Series C* **54**, 191–206.

Waller, L., Carlin, B., Xia, H. and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* **92**, 607–617.

Wang, X., Chen, C., Wang, L., Chen, D., Guang, W. and French, J. (2003). Conception, early pregnancy loss, and time to clinical pregnancy: a population-based prospective study. *Fertility and Sterility* **79**, 577–584.

West, M., Müller, P. and Escobar, M. (1994). Hierarchical priors and mixture models with application in regression and density estimation. In Smith, A. and Freeman, P., editors, *A Tribute to D.V. Lindley*, . Wiley, New York.

WHO (1983). A prospective multicentre trial of the ovulation method of natural family planning. III. characteristics of the menstrual cycle and of the fertile phase. *Fertility and Sterility* **40**, 773–778.

Wilcox, A. and Weinberg, C. (2004). Invited commentary: Analysis of gestational-age-specific mortalityon what biologic foundations? *American Journal of Epidemiology* **160**, 213–214.

Wilcox, A., Weinberg, C. and Baird, D. (1995). Timing of sexual intercourse in relation to ovulation effects on the probability of conception, survival of the pregnancy, and sex of the baby. *New England Journal of Medicine* **333**, 1517–1521.

Wilcox, A., Weinberg, C. and Baird, D. (1998). Post-obulatory ageing of the human oocyte and embryo failure. *Human Reproduction* **13**, 394–397.

Wilcox, A., Weinberg, C., O'Connor, J., Baird, D., Schlatterer, J., Canfield, R., Armstrong, E. and Nisula, B. (1988). Incidence of early loss of pregnancy. *New England Journal of Medicine* **319**, 189–194.

Wilcox, A., Weinberg, C., Wehmann, R., Armstrong, E., Canfield, R. and Nisula, B. (1985). Measuring early pregnancy loss: laboratory and field methods. *Fertility and Sterility* **44**, 366–374.

Winter, E., Wang, J., Davies, M. and Norman, R. (2002). Early pregnancy loss following assisted reproductive technology treatment. *Human Repoduction* **17**, 3220–3223.

Wood, S., Jiang, W. and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513–528.

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.

Zinaman, M., O'Connor, J., Clegg, E., Selevan, S. and Brown, C. (1996). Estimates of human fertility and pregnancy loss. *Fertility and Sterility* **65**, 503–509.