

Bayesian semiparametric clustering of functional predictors

Jamie Lynn Bigelow and David B. Dunson

Biostatistics Branch

National Institute of Environmental Health Sciences

Research Triangle Park, NC 27709

January 6, 2006

SUMMARY. This article proposes a new method for the joint clustering of functional predictors with some outcome of interest. A multivariate adaptive spline model is used to describe the functions, and 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 and according to 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.

KEY WORDS: Bayesian clustering; Dirichlet process; Joint modeling

1. Introduction

The joint modeling literature has generally focused on models 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

email: bigelow@niehs.nih.gov

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 regression function for the longitudinal 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 curve and some outcome random variable. This outcome need not be the time of some event. 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 a longitudinal trajectory. His method relies on first modeling the trajectory with a cubic spline and then using an EM algorithm to estimate a function to describe the weighted relationship between the response and the predictor as integrated over time. It requires that all functions be observed over the same region of time and are to be 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 demonstrate their method using a set of Fourier basis functions to model fetal heart-rate traces, then using the model to predict high-risk pregnancies. We are interested in modeling the joint relationship between a longitudinal response and some outcome, where not all longitudinal responses are observed over the same region of time, nor is time required to be the only covariate flexibly affecting the response function.

We extend the multivariate linear spline model with random coefficients to the case where each trajectory 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 observations. Though demonstrated 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 preg-

nancy 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 (Elish et al., 1999; Zinaman et al., 1996; Wang et al., 2003).

Given the high incidence, there has been discussion on potential mechanisms of EPL. The current project examines progesterone post-ovulation, comparing EPL 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.

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 Bigelow and Dunson (2005b). 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.

2.1 Multivariate linear splines

Bigelow and Dunson (2005b) describe 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, $\boldsymbol{\mu}_M = (\boldsymbol{\mu}_{M1}, \dots, \boldsymbol{\mu}_{Mk_M})$. When y_{ij} is the j^{th} response from subject i , $i = 1, \dots, N$; $j = 1, \dots, n_i$, 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:

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

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

$$\mathbf{y}_i = \mathbf{H}_{Mi} \mathbf{b}_{Mi} + \boldsymbol{\epsilon}_{Mi}, \quad M \in \mathcal{M} \quad (2)$$

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:

$$p(\mathbf{y}|\mathbf{b}_M, \tau_M, M) \propto \tau_M^{\frac{n}{2}} \prod_{i=1}^N \exp\left[-\frac{\tau_M}{2}(\mathbf{y}_i - \mathbf{H}_{Mi} \mathbf{b}_{Mi})'(\mathbf{y}_i - \mathbf{H}_{Mi} \mathbf{b}_{Mi})\right] \quad M \in \mathcal{M} \quad (3)$$

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

$$\mathbf{b}_{Mi} \stackrel{iid}{\sim} G_M, \quad i = 1, \dots, N; \quad M \in \mathcal{M} \quad (4)$$

The distribution G_M could be given some parametric form for all $M \in \mathcal{M}$. Bigelow and Dunson (2005a) 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 Bigelow and Dunson (2005b) for details.)

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.

$$p(z|\xi, \phi) = \exp(a(\phi)^{-1}(z\xi - B(\xi)) + c(z, \phi)) \quad (5)$$

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 $\{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, where $\mathbf{z}_i = \{z_{i1}, \dots, z_{in_i}\}$ for $i = 1, \dots, N$, and the likelihood of one observation z_{ij} is an exponential family density.

$$p(z_{ij}|\xi_{ij}, \phi_{ij}) = \exp(\phi_{ij}^{-1}(z_{ij}\xi_{ij} - B(\xi_{ij})) + c(z_{ij}, \phi_{ij})) \quad (6)$$

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:

$$p(\mathbf{z}|\xi_{ij}, \phi_{ij}) = \prod_{i=1}^N \prod_{j=1}^{n_i} \exp(\phi_i^{-1}(z_{ij}(\mathbf{u}'_{ij}\boldsymbol{\gamma}_i) - B(\mathbf{u}'_{ij}\boldsymbol{\gamma}_i)) + c(z_{ij}, \phi_i)) \quad (7)$$

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:

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

where the gamma hyperparameters are pre-specified. Alternatively, a Wishart distribution and its hyperparameters could be specified for the prior precision of γ_i . The random effects $\{\gamma_i\}$ can be sampled through the use of a rejection algorithm, and γ_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.

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 Bigelow and Dunson (2005b) 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.

3.1 Prior specification

The data consist of N trajectory/outcome pairs. We model the trajectory according to the methods given in Bigelow and Dunson (2005b), 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.

$$L(\mathbf{y}|\boldsymbol{\theta}_M, \tau_M, M, \mathbf{S}_M) \propto \tau_M^{n/2} \exp\left[-\frac{\tau_M}{2} \sum_{j=1}^{r_M} \sum_{i \in I_{Mj}} (\mathbf{y}_i - \mathbf{H}_i \boldsymbol{\theta}_{Mj})' (\mathbf{y}_i - \mathbf{H}_i \boldsymbol{\theta}_{Mj})\right] \quad (8)$$

$$L(\mathbf{Z}|\boldsymbol{\xi}, \phi) \propto \prod_{i=1}^N \prod_{j=1}^{n_i} \exp(\phi_i^{-1}(z_{ij} \xi_{ij} - B(\xi_{ij})) + c(z_{ij}, \phi_i)) \quad (9)$$

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:

$$\boldsymbol{\xi}_i = \boldsymbol{\eta}_i \equiv \mathbf{U}_i \boldsymbol{\gamma}_i + \mathbf{J}_{n_i} a_i \quad (10)$$

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 .

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

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 (10) are empty. To include covariates in the GLMM for the outcome, we specify the following additional priors:

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

where a_ω , b_ω , a_ψ 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 $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_r\}$. When we exclude subject i there are $r^{(i)} \leq r$ distinct values, denoted $\boldsymbol{\theta}^{(i)} = \{\boldsymbol{\theta}_1^{(i)}, \dots, \boldsymbol{\theta}_{r^{(i)}}^{(i)}\}$

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:

$$q_{i,j} \propto \begin{cases} \alpha h_i(\mathbf{y}_i, \mathbf{z}_i) & \text{if } j = 0 \\ n_j^{(i)} f_j(\mathbf{y}_i, \mathbf{z}_i | \boldsymbol{\theta}_j) & \text{if } j > 0 \end{cases} \quad (11)$$

$$h_i(\mathbf{y}_i, \mathbf{z}_i) = \int f_i(\mathbf{y}_i, \mathbf{z}_i | \mathbf{b}_i, a_i) dG_0(\mathbf{b}_i, a_i) \quad (12)$$

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:

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

$S_i = 0$ 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. Posterior computation

Computation is similar to that for the trajectory-only model in Bigelow and Dunson (2005b). 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 3.1. At each iteration, we have a current spline model, M , and current values of the parameters $\{\boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\beta}, \tau, \lambda, \boldsymbol{\delta}, \boldsymbol{\gamma}, \gamma_0, \omega, \nu, \boldsymbol{\psi}\}$. 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; Bigelow and Dunson, 2005a):

$$Q(M', M) = \min \left[1, \frac{p(\mathbf{y}|M')}{p(\mathbf{y}|M)} \times R \right] \quad (13)$$

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 (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. Bigelow and Dunson (2005b) outline conditions under which alternative acceptance probabilities are valid, and per those results we use:

$$Q(M', M) = \min \left[1, \frac{p(\mathbf{y}|M', \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)}{p(\mathbf{y}|M, \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)} \times R \right] \quad (14)$$

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 $\boldsymbol{\delta}_M$ with number of elements equal to the minimum of the dimensions of M and M' . The reasoning behind this is presented in detail in Bigelow and Dunson (2005b). 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 can be found in Bigelow and Dunson (2005a).

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:

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

where $q_{i,j}$ is given in (11) and depends on the likelihood and $h_i(\mathbf{y}_i, \mathbf{z}_i)$ is given in (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:

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

Bigelow and Dunson (2005b) show 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.

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

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, $\boldsymbol{\theta}_{j,k+1}$, 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 $j = 1, \dots, r$.

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

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:

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

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:

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

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

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

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 3.1.

$$\begin{aligned} p(\boldsymbol{\beta} | \boldsymbol{\theta}, \mathbf{S}, \boldsymbol{\delta}, \lambda) &= N\left((\lambda \mathbf{I} + r \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta} \sum_{j=1}^r \boldsymbol{\theta}_{j,1:k}, \tau^{-1} (\lambda \mathbf{I} + r \boldsymbol{\Delta})^{-1} \right) \\ \lambda | \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}, \tau &\sim \text{Gamma}\left(a_\lambda + \frac{k}{2}, b_\lambda + \frac{\tau \boldsymbol{\beta}' \boldsymbol{\beta}}{2} \right) \\ \delta_l | \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\delta}_{-l}, \lambda, \tau &\sim \text{Gamma}\left(a_\delta + \frac{r}{2}, b_\delta + \frac{\tau}{2} \sum_{j=1}^r (\theta_{jl} - \beta_l)^2 \right) \quad l = 1, \dots, k \\ \tau | \mathbf{y}, \boldsymbol{\theta}, \mathbf{S} \boldsymbol{\delta}, \lambda, \nu &\sim \text{Gamma}\left(a_\tau + \frac{N + (r+1)k}{2}, b_\tau + \frac{1}{2} \left(\sum_{i=1}^N (\mathbf{y}_i - \mathbf{H}_i \mathbf{b}_i)' (\mathbf{y}_i - \mathbf{H}_i \mathbf{b}_i) \right. \right. \\ &\quad \left. \left. + \lambda \boldsymbol{\beta}' \boldsymbol{\beta} + \sum_{j=1}^r (\boldsymbol{\theta}_{j,1:k} - \boldsymbol{\beta})' \boldsymbol{\Delta} (\boldsymbol{\theta}_{j,1:k} - \boldsymbol{\beta}) \right) \right) \end{aligned} \tag{15}$$

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.

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

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 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.

[Figure 1 about here.]

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 all samples of the logit of the outcome model's random intercept, a_i . Figure 2 gives the mean trajectory for the three final clusters as well as the data from each.

[Figure 2 about here.]

We estimated the outcome probabilities and compared them to the underlying population probabilities. The estimates and 95% credible intervals are given in Table 1. The credible interval contained the true value in all cases.

[Table 1 about here.]

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 trajectories 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.

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 (3) and (4) illustrate some of the data from early losses and clinical pregnancies. They show how PdG tends to rise when conception occurs and then drop off if the pregnancy is lost.

[Figure 3 about here.]

[Figure 4 about here.]

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 5 and 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.

[Figure 5 about here.]

[Figure 6 about here.]

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 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 Bigelow and Dunson (2005a) 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.

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.

ACKNOWLEDGEMENTS

We would like to thank Allen Wilcox, Donna Baird and Clare Weinberg for generously providing the data and for their helpful comments on the approach.

REFERENCES

- Baird, D., Wilcox, A., Weinberg, C., Kamel, F., McConaughey, 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.
- Bigelow, J. and Dunson, D. (2005a). Bayesian adaptive regression splines for hierarchical data. *Duke University ISDS Discussion paper 05-06* .
- Bigelow, J. and Dunson, D. (2005b). Semiparametric classification in hierarchical functional data analysis. *Duke University ISDS Discussion paper 05-18* .
- 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.
- Chib, S. and Hamilton, B. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Biometrics* **61**, 64–73.
- 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.
- 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.
- Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- 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. (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society, Series B* **63**, 3–17.
- James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* **64**, 411–432.
- 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.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, England, 2nd edition.
- Percy, D. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.
- 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.
- Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834.
- 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.
- 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.
- Winter, E., Wang, J., Davies, M. and Norman, R. (2002). Early pregnancy loss following assisted reproductive technology treatment. *Human Reproduction* **17**, 3220–3223.
- 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.

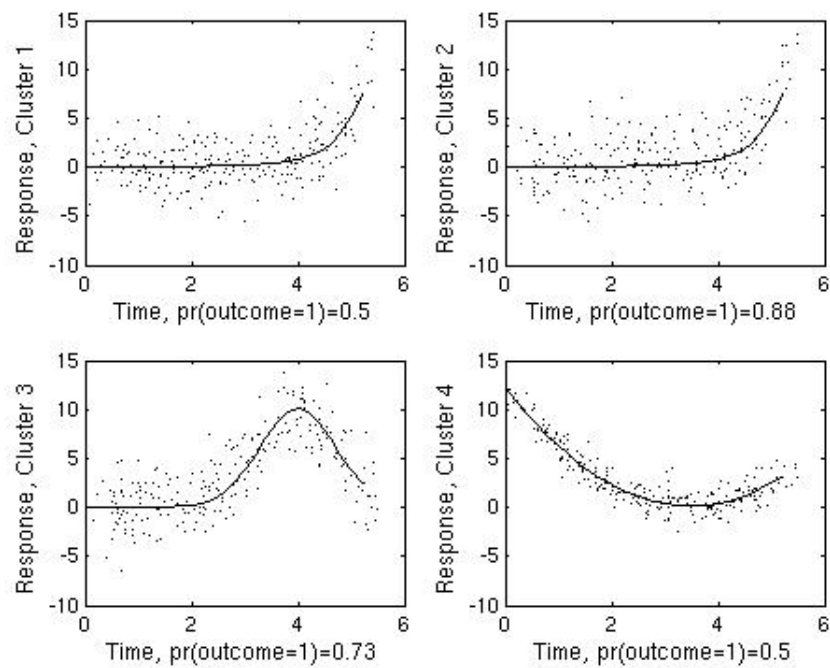


Figure 1. Underlying population curves (lines) and data (points) for each of the four clusters along with outcome probabilities. The top two plots have the same underlying trajectory with different outcome probabilities.

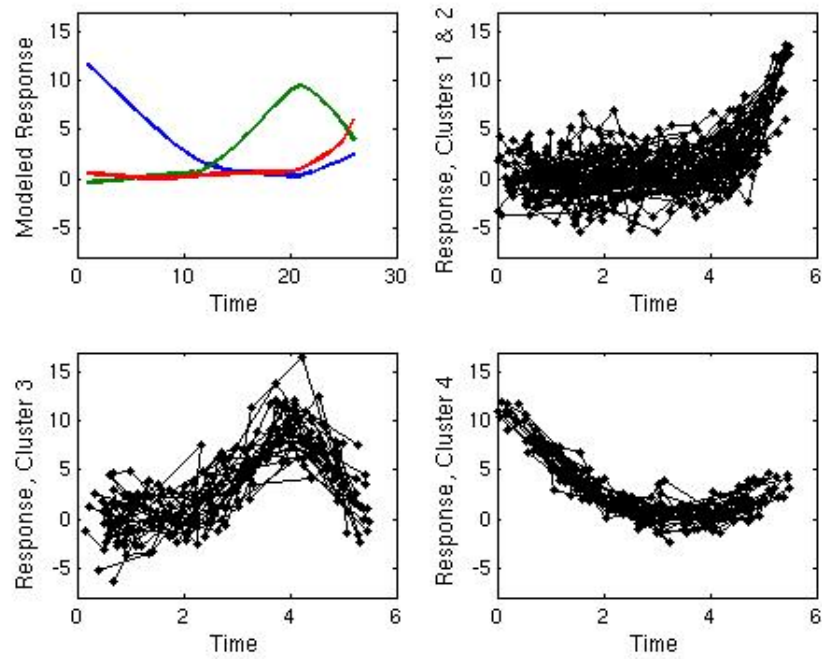


Figure 2. The plot in the first quadrant is of the mean trajectory for the three final clusters. The three remaining plots contain all trajectory data for the three final clusters.

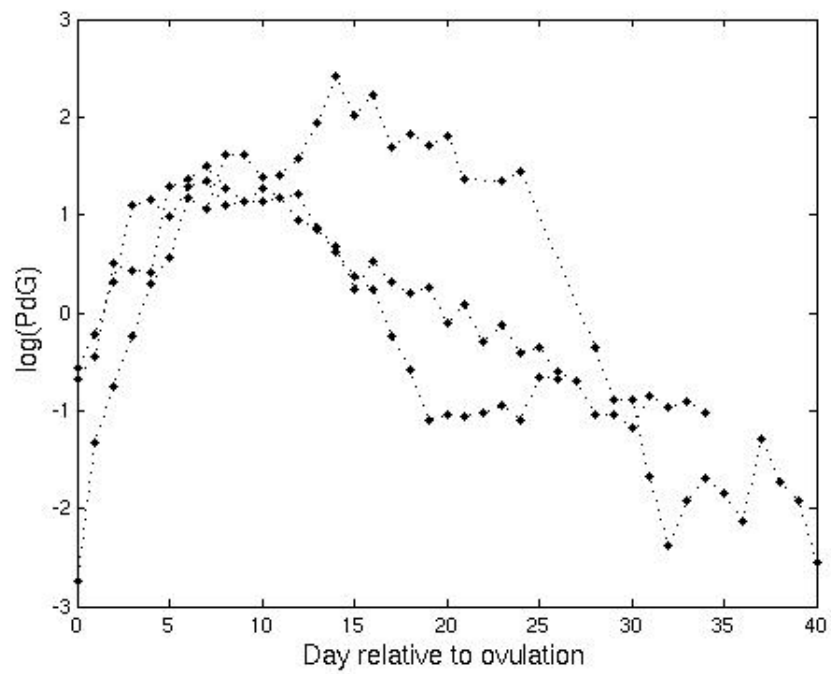


Figure 3. Progesterone data beginning at the estimated day of ovulation for three early losses.

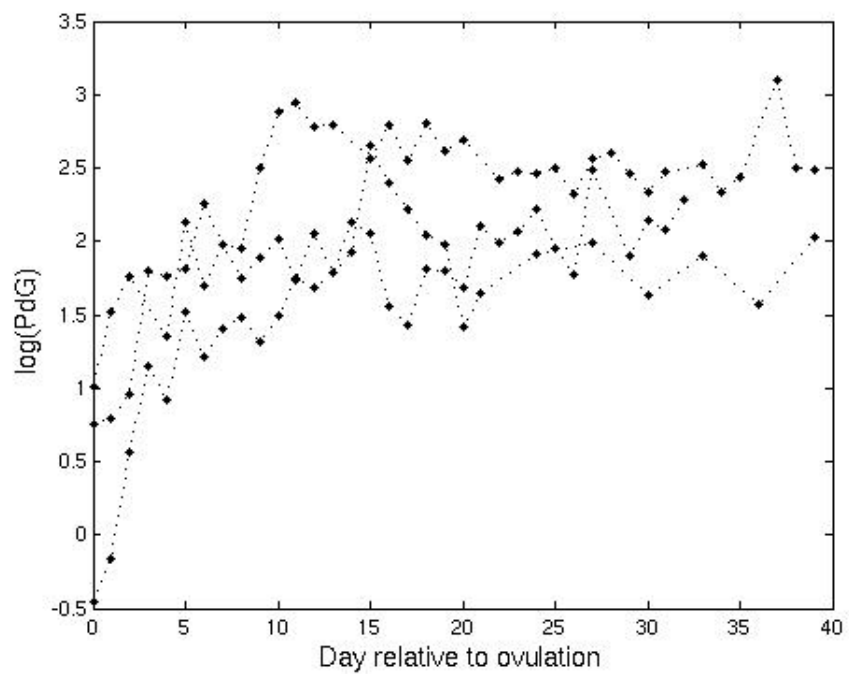


Figure 4. Progesterone data beginning at the estimated day of ovulation for three clinical pregnancies.

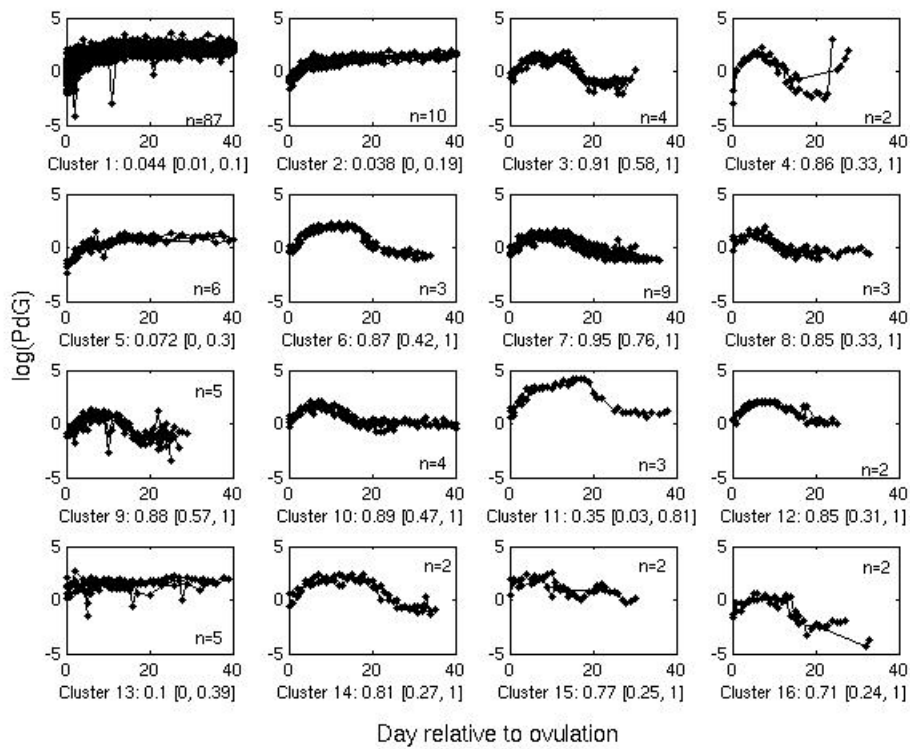


Figure 5. Data from the 16 classes containing more than one observation. Below each plot is the model-estimated probability of early loss for each cluster along with the 95% credible interval. The cluster sizes are given on the plots.

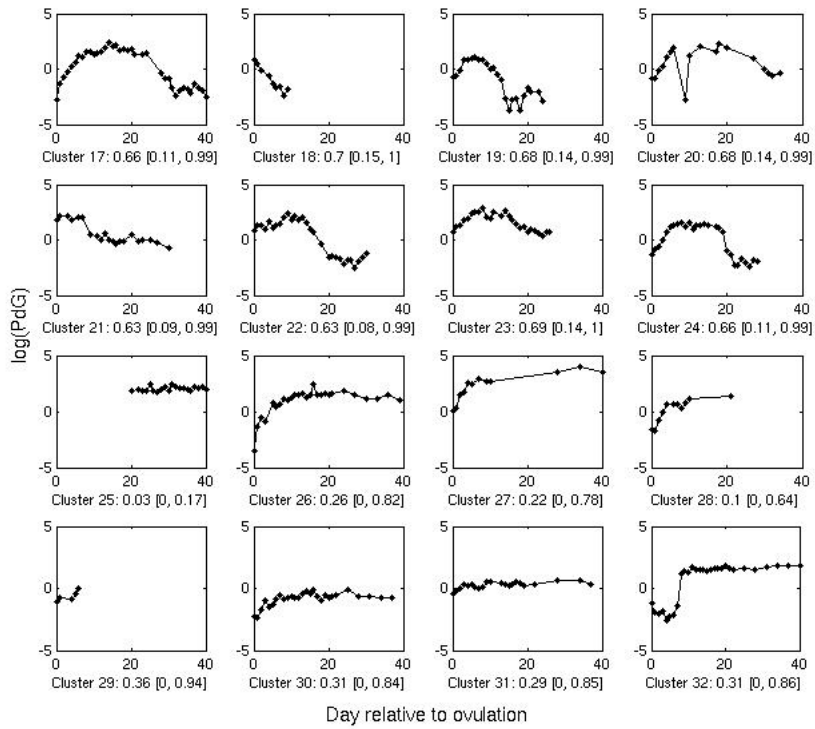


Figure 6. Data from the 16 trajectories that were not clustered with any others. Below each plot is the model-estimated probability of early loss for each trajectory along with the 95% credible interval. The first two rows of plots are EPLs, while the remaining eight are clinical pregnancies.

Table 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]