Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2020 June 10.

Published in final edited form as:

J Am Stat Assoc. 2009 ; 104(485): 26–36. doi:10.1198/jasa.2009.0001.

Bayesian semiparametric joint models for functional predictors

Jamie L. Bigelow[1], **David B. Dunson**[1,2]

[1]Department of Statistical Science, Duke University

[2]National Institute of Environmental Health Sciences, Biostatistics Branch, MD A3-03, P.O. Box 12233, Research Triangle Park, NC 27709, USA

Summary.

Motivated by the need to understand and predict early pregnancy loss using hormonal indicators of pregnancy health, this paper proposes a semiparametric Bayes approach for assessing the relationship between functional predictors and a response. A multivariate adaptive spline model is used to describe the functional predictors, and a generalized linear model with a random intercept describes the response. 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 response model for each cluster. This very flexible method allows for the incorporation of covariates in the models for both the response and the trajectory. We apply the method to post-ovulatory progesterone data from the Early Pregnancy Study and find that the model successfully predicts early pregnancy loss.

Keywords

Bayesian clustering; Dirichlet process; Joint modeling; Functional predictor; Progesterone; Early pregnancy loss

1. Introduction

1.1. Hormonal Predictors of Pregnancy Loss

Our motivation is drawn from studies of the relationship between hormone trajectories and the occurrence of early pregnancy loss (EPL). EPL is the loss of a pregnancy within six weeks of the last menstrual period, so early that many women experience them without even knowing they had conceived. Pregnancies and losses can be detected through the examination of hormone profiles.

Data are drawn from the North Carolina Early Pregnancy Study (Baird et al., 1997), a prospective study of 221 women who collected daily urine samples while trying to conceive. These samples were assayed for metabolites of progestrone and human chorionic gonadotropin. Human chorionic gonadotropin (hCG) is a hormone released by the embryo upon implantation. hCG is usually not present in any detectable level unless a woman is pregnant. Based on examination of profiles of urinary hCG, the study investigators classified

jbigelow@stat.duke.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

cycles that did not result in clinical pregnancy (i.e. a pregnancy lasting beyond six weeks) 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. The data consisted of 165 conception cycles, 47 of which resulted in early losses. Based on these data, Wilcox et al. (1988) reported that two-thirds of losses occurred before the pregnancy was likely to be 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).

Progesterone is produced in a woman's body even when she is not pregnant. It is well known that progesterone levels rise at conception, remain high in ongoing pregnancies and decrease 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. Motivated by this hypothesis, there has been an ongoing debate among clinicians about whether treatment with exogenous progesterone may be beneficial in terms of reducing the risk of EPL.

In the recent literature, there has been considerable interest in assessing hormonal predictors of EPL. In a prospective study of Chinese textile workers, Venners et al. (2006) measured estrogen and progesterone metabolites in daily urine samples, assaying hCG to identify early pregnancy losses. Using generalized estimating equations to adjust for within-woman dependence, they assessed differences between hormone levels within clinical pregnancies (pregnancies lasting more than six weeks beyond the last menstrual period) and EPLs through stratifying on phase of the cycle. Although the hormone trajectory is most naturally viewed as a random curve, it is common practice in the epidemiologic and clinical literature to simplify the analysis by dividing time into bins. In a population-based prospective study in thirty rural Bolivian communities, Vitzthum et al. (2006) used a related approach to assess their hypothesis that EPLs are associated with (1) inadequate post-ovulatory progesterone and (2) elevated pre-ovulatory progesterone. Although their sample size was modest, they found significant evidence of (2).

The typical strategy of dividing time into bins and testing for differences within each bin has a number of drawbacks, including sensitivity to the chosen bins, issues in multiple testing if the number of bins is chosen to be high, and focus on a mean difference. Since EPL is not always caused the mother's inability to sustain the pregnancy, but can also be the result of problems with the fetus, it is likely that the progesterone trajectory close to ovulation may be predictive only in certain cycles, suggesting that analyses based on a mean shift may be overly-conservative and produce misleading results.

Our goal to to use the post-ovulatory progesterone trajectory curve as a functional predictor in a flexible joint model for EPL, then use the resulting model to predict EPL in a set of reserved cycles. A joint model relating progesterone and early loss should provide new insight into possible mechanisms of EPL. 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to developmental problems in the embryo. Progesterone levels and EPL are likely to be causally associated.

1.2. Relevant Statistical Literature

In assessing relationships between functional predictors and a scalar response, a challenging issue is interpretation of the results. A common strategy in applications is to define several summaries of the function, such as the rate of change, average value across a region, etc. These summaries can then be plugged in as predictors in a generalized linear model (GLM) for the response. Unfortunately, it is typically not clear how to best choose summaries of the function, and multicolinearities can arise if many summaries are chosen. In addition, the plug-in approach cannot easily accommodate missing data or differences among subjects in locations of measurements along the function. Such problems can be solved by fitting a joint model of the function and scalar response.

James (2002) extends GLMs to incorporate functional predictors, using functional principal components for interpretation. His method relies on modeling the function with a cubic spline and then characterizing the relationship with the response by using a weighted integral of the function. James and Silverman (2005) later proposed a more general class of models, which extends GLMs, generalized additive models and projection pursuit regression to handle functional predictors. Muller and Stadtmuller (2005) proposed an alternative generalized functional linear model framework, developing functional estimating equations and associated theory. Brown et al. (2001) instead use a Bayes wavelet-based approach to include a discretized functional predictor in a linear regression model. 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, using the model to predict high-risk pregnancies. Escabias et al. (2005) proposed a functional principal component logistic regression method motivated by environmental data.

Although functional principal components are useful in interpreting results, they may be counter-intuitive for clinicians. An alternative strategy is to develop a flexible clustering methodology, in which clusters are defined by the functional predictor and level of response. For example, progesterone trajectories in early pregnancy could be clustered jointly with an indicator of early pregnancy loss. Careful examination of the resulting clusters could identify trajectories in progesterone that are predictive of EPL, with such trajectories being of substantial clinical interest. For example, women could potentially monitor progesterone levels in urine using a home device, which recommends a visit to the woman's clinician if the trajectory shape is similar to one of the clusters indicative of EPL. Early identification of pregnancies at risk of EPL would aid in the development and testing of interventions, such as treatment with exogenous progesterone.

A variety of methods have been proposed for clustering functional data. James and Sugar (2003) propose a flexible model-based approach for clustering sparsely sampled functional data. Zhou and Wakefield (2006) and Heard et al. (2006) applied Bayesian methods for clustering of time course gene expression data. Ray and Mallick (2006) proposed a nonparametric Bayes wavelet model for clustering functional data, relying on a Dirichlet process (DP) prior (Ferguson (1973), Ferguson (1974)) for the distribution of the wavelet

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

coefficients. Bigelow and Dunson (2005) proposed an approach that allows both the basis functions and their coefficient distributions to be unknown. These methods are related to the common strategy of clustering functions through clustering of the coefficients in a basis function expansion. However, simple methods for clustering the basis coefficients, such as k-means, tend to suffer from a lack of robustness (Garcia-Escudero and Gordaliza (2005)). In addition, the Bayesian nonparametric approach generates a full posterior distribution for the clusters instead of a single set of clusters. This posterior distribution can be used to assess uncertainty in the clustering process.

Our interest focuses on clustering the progesterone trajectories and treating them as functional predictors in a model for early pregnancy loss. To address this problem, we propose a semiparametric Bayes approach. The functional predictor is characterized using an adaptive spline model. We then model the joint distribution of the basis coefficients and a random intercept in the outcome model as unknown using a DP prior, which also induces clustering. A number of authors have used DP priors for unknown random effects distributions (Bush and MacEachern (1996), Mukhopadhyay and Gelfand (1997), Kleinman and Ibrahim (1998), Brown and Ibrahim (2003) among others), but only Bigelow and Dunson (2005) consider cases in which the number of basis functions is unknown, so that the random effects have varying dimension. Here, we extend Bigelow and Dunson (2005) to allow joint modeling of functional data with a response variable.

2. Semiparametric Joint Models

2.1. Basic Structure

For subject i, ($i = 1,\ldots,N$), let z_i denote the response, let $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{i,n_i}\right)'$ denote a vector of n_i error-prone observations of the functional predictor, $f_i(\cdot)$, and let $\mathbf{x}_i = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right)'$ denote covariates describing the *locations* of these observations. For example, in a simple case, $f_i : \mathbb{R}^+ \rightarrow \mathbb{R}$ is a time-varying predictor and $\mathbf{x}_{ij}$ is the time of the j^{th} observation of this predictor for subject i. We focus on the general case in which $\mathbf{x}_{ij}$ is a vector of covariates.

We assume the following basic modeling structure:

$$z_i \sim g(\xi_i, \phi) \text{ and } y_{ij} \sim N\left(f_i(\mathbf{x}_{ij}), \tau^{-1}\right), \tag{1}$$

where $g(\xi_i, \phi)$ is an exponential-family density, with canonical parameter ξ_i and scale parameter ϕ. In order to allow the functional predictor for subject i, $f_i(\cdot)$, to be predictive of the response variable, z_i, we consider flexible models for the joint probability measure of $\{f_i, \xi_i\}$.

Because f_i is a random function defined at infinitely-many points, modeling is facilitated by considering a finite representation of f_i in terms of a basis function expansion:

$$f_i\left(\mathbf{x}_{ij}\right) = \sum_{h=1}^{k_M} b_{M_i h} \mu_{M_h}\left(\mathbf{x}_{ij}\right) = \mathbf{H}_{M_i} \mathbf{b}_{M_i}, \tag{2}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\boldsymbol{\mu}_M = \{\mu_{M_1}, \ldots, \mu_{M,k_M}\}$ denotes a set of k_M basis functions specific to model $M \in \mathcal{M}$, with $\mathcal{M}$ denoting the model space. In addition, $\mathbf{b}_{M_i} = \left(b_{M_i 1}, \ldots, b_{M_i k_M}\right)'$ are basis coefficients specific to subject i and model M. The space of models, $\mathcal{M}$, corresponds to a chosen class of basis functions, while a particular model, M, corresponds to a selection of k_M basis functions in this class. For example, we focus on $\mathcal{M}$ corresponding to the class of multivariate linear splines, with the different models in this class selecting different numbers and locations of knots. Using model averaging to allow uncertainty in knot selection, multivariate piecewise linear splines can approximate any smooth surface. However, the methods are general and different classes of basis functions can be incorporated without complications. In order to flexibly characterize the hormone trajectories without incorporating large numbers of basis functions, we allow uncertainty in basis function selection.

2.2. Joint Models and Clustering

To allow the response distribution to vary systematically according to the shape of the functional predictor, it is necessary to specify a joint model for f_i and z_i. For sake of interpretability and flexibility, it is appealing to group the random functions into trajectory clusters, with the cluster identifier included as a categorical predictor in the response model. For example, in the hormone application, the progesterone trajectories in early pregnancy could be grouped into k clusters, with the early loss probability varying systematically across these clusters, adjusting for covariates within a generalized linear model (GLM).

This is accomplished by first specifying a GLM for the conditional distribution of the response:

$$h\{\mathrm{E}(z_i | a_i, \boldsymbol{\gamma}, \mathbf{u}_i)\} = a_i + \mathbf{u}_i'\boldsymbol{\gamma}, \tag{3}$$

where $h(\cdot)$ is a monotone link function, a_i is a random effect for subject i, $\mathbf{u}_i$ is a vector of known covariates, and $\boldsymbol{\gamma}$ is a vector of fixed effects. Usually $h()$ is the canonical link so that $a_i + \mathbf{u}_i'\boldsymbol{\gamma} = \xi_i$, the canonical parameter. For example, in the early pregnancy loss application, expression (3) will be chosen as a logistic regression model, where a_i will characterize the baseline log odds of early loss for woman i adjusting for covariates.

We then induce a joint model on the functional predictor, f_i, and response, z_i, through a hierarchical model for the joint distribution of the random effect, a_i, and subject-specific basis coefficients, $\mathbf{b}_{M_i}$:

$$\left(\mathbf{b}_{M_i}, a_i\right) \sim G_M = \sum_{h=1}^{\infty} p_h \delta_{\Theta_{Mh}},\ p_h = V_h \prod_{l<h} (1 - V_l),\ \forall M \in \mathcal{M}, \tag{4}$$

where G_M is the joint distribution (or, more formally, the joint probability measure) under model M, $\{p_h\}_{h=1}^{\infty}$ is an infinite sequence of probability weights, δ_Θ is a probability

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

measure concentrated at Θ, and $V_h \overset{iid}{\sim} \text{beta}(1, \alpha)$ are stick-breaking random probabilities that are mutually independent of the atoms $\Theta_{Mh} = \{\beta_{Mh}, \psi_h\} \overset{iid}{\sim} G_{0M}$, for $h = 1,\ldots,\infty$.

Note that (4) is the Sethuraman (1994) stick-breaking representation of the DP, implying that $G_M \sim DP(\alpha G_{0M})$ within each model $M \in \mathcal{M}$. The vector Θ_{Mh} corresponds to the value of the k_M basis coefficients and random effect under model M for cluster h, with $h = 1,\ldots,\infty$ across the infinitely many clusters represented in the population. As is clear from the stick-breaking representation, the probability that a randomly selected subject is allocated to cluster h decreases stochastically as h increases, with the rate of this decrease higher for smaller α. This induces a prior on the random partition of the n subjects into r clusters, with the pairwise prior clustering probability being

$$\Pr\left\{(\mathbf{b}_{M_i}, a_i) = (\mathbf{b}_{M_{i'}}, a_{i'})\right\} = \frac{1}{1+\alpha}, \text{ for } i \neq i'.$$

We let $S_{Mi} = h$ denote that subject i belongs to cluster h under model M, which implies that $f_i(\cdot) = \sum_{l=1}^{k_M} \theta_{Mhl}\mu_{M_l}(\cdot)$. In addition, all subjects with $S_{Mi} = h$ have intercept $a_i = \psi_{Mh}$ in the GLM (3) for the response variable. Suppose that z_i is a binary response indicating occurrence of an adverse event, such as early pregnancy loss. Then, clusters having trajectory shapes predictive of an increased risk of the event will tend to have high values of ψ_{Mh} relative to other clusters.

Note that this Dirichlet process-based joint clustering approach has important advantages over clustering methods based on finite mixture models. First, allowing infinitely many clusters in the general population, the DP prior allows new clusters to be introduced as a subjects are added to the sample, though allocation to existing clusters is increasingly favored as the sample size increases. This allows incoming women to have rare conditions not yet observed in the sample. Second, the approach avoids the possibility of over-fitting through generating the cluster-specific parameters from a parametric base distribution G_{0M} a priori. As clusters are added and the number of subjects per cluster decreases, the fit will increasingly resemble that of a parametric hierarchical model. Hence, to allow deviations from the parametric model, it is necessary to favor fewer clusters.

2.3. Uncertainty in Basis Functions

Without uncertainty in the basis functions, the model specified in subsections 2.1 and 2.2 is quite similar to previous Bayesian hierarchical models that have incorporated DP priors to allow random effects distributions to be unknown (refer to Section 1 for references). In the setting of parametric models for a single function, there is a rich literature on methods that allow uncertainty in basis function selection. Refer, for example, to Biller (2000), Lindstrom (2002) Hansen and Kooperberg (2002), Wood et al. (2002), and Holmes and Mallick (2003). Posterior computation in such models typically relies on one of two Markov chain Monte Carlo (MCMC) strategies: (1) reversible jump MCMC (RJMCMC) (Green (1995)); or (2) embedding all the models in a single encompassing model having very many basis

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

functions, with a subset of the bases selected by allowing zero basis coefficients (Smith and Kohn (1996)).

In our setting, efficient posterior computation is much more challenging, since we have a nonparametric prior on a collection of curves that are characterized in terms of an unknown set of basis functions. Fortunately, the specification in (4) facilitates efficient posterior computation by inducing a prior that can be factorized as a prior on the allocation of subjects to clusters multiplied by a prior on the coefficients within each cluster. Bigelow and Dunson (2005) provide details on this factorization and propose an efficient Metropolis-Hastings algorithm for moving between models, with a nested Gibbs sampler for updating the model parameters. The Bigelow and Dunson (2005) paper was motivated by the problem of functional clustering, but did not consider joint modeling with a response variable. In Section 3, we outline an approach to generalize the Bigelow and Dunson (2005) MH-nested Gibbs algorithm to the joint modeling setting.

3. Posterior Computation

3.1. Model and Prior Specification

We focus on multivariate linear basis functions, with the prior over the model space depending only on the number of basis functions in each model, $p(M) \propto p(k_M) = \binom{T}{k_M - 1}^{-1} K^{-1}$, where K is the maximum number of basis functions allowed in a model and T is very large. From (2), we have

$$\mathbf{y}_i = \mathbf{H}_{Mi}\mathbf{b}_{Mi} + \boldsymbol{\epsilon}_{Mi},\ M \in \mathscr{M}, \tag{5}$$

where $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_{Mi}$ are the $n_i \times 1$ vectors of trajectories 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 under model M.

The data consist of n trajectory × response pairs, $\{\mathbf{y}_i, z_i\}_{i=1}^n$. Following the specification described in Section 2, the joint likelihood can be factored as a product of:

$$\begin{aligned} L(\mathbf{y}|\mathbf{b}_M, \tau_M, M, \mathbf{S}_M) &\propto \tau_M^{N/2}\exp\left\{-\frac{\tau_M}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{H}_i\mathbf{b}_{Mi})'(\mathbf{y}_i - \mathbf{H}_i\mathbf{b}_{Mi})\right\} \\ L(\mathbf{z}|\mathbf{a}_M, \boldsymbol{\gamma}_M, \phi_M, \mathbf{U}) &\propto \prod_{i=1}^{n} g(z_i; \xi_i, \phi), \end{aligned} \tag{6}$$

Where $N = \sum_{i=1}^{n} n_i$ and a canonical link is used so that $\xi_i = a_i + \mathbf{u}_i'\boldsymbol{\gamma}$. The prior under model M is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

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. The prior specification is completed with a prior on γ, which can follow as for a typical GLM.

3.2. Metropolis-Hastings with Nested Gibbs Algorithm

The approach described in Bigelow and Dunson (2005) considered the model specified as above, but without the response component. It involves a Metropolis-Hastings (Hastings, 1970) step for moving between models, with a nested Gibbs sampler used to update the model parameters. In particular, the algorithm alternates between the following steps: (1) propose a move from model $M \rightarrow M'$ according to the proposal density $T(M', M)$; (2) accept this move with probability $Q(M', M)$; and (3) update the unknowns within the current model, including the clustering configuration, $\mathbf{S}_M$, the unique values, $\boldsymbol{\theta}_M$, the parameters within G_{M0}, the residual precision, τ_M, and the regression coefficients, γ. Conditionally on the selected basis functions implied by M, posterior computation for the model unknowns can proceed via previously proposed MCMC algorithms for DP mixture models. Hence, we leave out the details on step (3), relying on the collapsed Gibbs sampler (Bush and MacEachern, 1996) in the implementation.

In step (1), we propose to change M by either adding, removing or altering a basis function with equal probability. The key to our approach is the choice of acceptance probability, $Q(M', M)$, which we take as:

$$Q(M', M) = \min\left[1, \frac{p(\mathbf{y}, \mathbf{z} \mid M', \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)}{p(\mathbf{y}, \mathbf{z} \mid M, \mathbf{S}_M, \boldsymbol{\delta}_{adj}, \lambda_M)} \times R\right] \quad (7)$$

where (7) is conditional on the current values $\mathbf{S}_M$, the cluster configuration, and λ_M, a component of the prior precision of $\boldsymbol{\beta}_M$. Recalling that the dimension of $\boldsymbol{\delta}_M$ is equal to the dimension of 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 acceptance probability in the joint modeling setting simplifies to that in the trajectory-only model in Bigelow and Dunson (2005). Details and an analytic form are given in the appendix.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3.3. Clustering and Prediction

The proposed model allocates women to clusters that are defined in terms of the woman's progesterone trajectory and risk of EPL. Because the approach allows uncertainty in the number of clusters and the allocation of women to clusters, one encounters a label-switching problem in which the meaning of the clusters varies across the posterior samples. This label switching does not present a problem in obtaining model-averaged estimates of a woman's trajectory and predictions of risk of EPL. Indeed, accounting for uncertainty in the choice of basis functions, number of clusters and allocation to clusters in performing prediction is a major advantage of our proposed approach.

However, in interpreting the results it is appealing to obtain a single estimate of the cluster-specific trajectories and risk of EPL. Several approaches have been proposed for estimating clusters based on the MCMC output in DP mixture models. Dahl (2006) and Lau and Green (2007) proposed approaches for estimating an optimal partition of clusters. We instead implement an earlier approach proposed by Medvedovic and Sivaganesan (2002), which estimates clusters by thresholding the pairwise posterior probabilities that two subjects are grouped together.

We used a posterior probability of 0.40 as the threshold. The threshold choice is necessarily subjective, with high values leading to inappropriate grouping of scientifically distinct trajectories and low values producing too many groups. We find 0.40 to be a good compromize based on simulations and examination of clusters produced in data analyses.

The DP precision parameter, α is a key hyperparameter, since the number of clusters is proportional to $\alpha \log(N)$. Following common practice for DP mixture models, we set $\alpha = 1$ to encourage clusters to be introduced slowly with increasing sample size. This favors a sparser representation of the data, though the data play a strong role in the determination of the posterior distribution of the number of clusters.

4. Simulated Data Example

Data were simulated as six groups of 25 trajectories, where each trajectory consisted of noisy observations around a smooth mean curve and was paired with a 0/1 response. Figure 1 shows the simulated data, the mean trajectories, and the 0/1 response for each of the 6 groups. There were 10 measurements along each curve collected at varied timepoints. Of the 25 trajectories in each group, 5 were truncated to contain only 5 measurements in the first half of the time scale.

We collected 50,000 samples after a 5,000 iteration burn-in. Examination of trace plots of precision parameters and the number of basis functions showed no evidence against convergence. The samples were first used to assess whether the algorithm had correctly clustered the trajectories based on their mean functions and accurately estimated the probability of success. We clustered observations that appeared in the same cluster in 40% or more of the samples. Based on the posterior distributions for subjects within each of these groups, we estimated the trajectory for each cluster and the probability that a member of each cluster had response equal to 1. Figure 2 illustrates the performance of the model in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

correctly identifying clusters. In addition, it provides means and 95% credible intervals for the response probabilities.

The approach correctly identified trajectories of similar shapes and clustered them together. The clusters were identified as simulated, with two exceptions. The trajectories in groups 1 and 2 were classified together. This misclassification is understandable, as groups 1 and 2 had rather high variability around their mean function at the start of the trajectory, and a similar shape to group 6 in the latter part. Four of the most unusual trajectories from group 4 were not classified with the others in group 4. A look at the estimated means of each group in the final plot in Figure 2 shows that the algorithm sorted out four trajectories where the trajectory decreased at the start then rose steeply to a peak from those where the response began to rise immediately.

Within the MCMC samples, observations were misclassified very rarely. This simulation shows that the model discriminated among trajectories of different shapes and provided an accurate estimate of the response probability for each cluster. Of specific interest is the first cluster, where the model did not distinguish the similar mean functions of the first two groups from the sixth group, but the success probability for that cluster was correctly estimated to be about 1/3.

Our primary goal is prediction. When a new trajectory is presented, we wish to be able classify it and predict the response probability. A new set of 150 trajectories was simulated in the same way as the first set. The new trajectories were then classified according to the MCMC samples reserved from the original run, and the probability of success was estimated as well. In summary, the trajectories from groups 1 and 2, which had a true success probability of 1/2 in the original data, had a predicted success probability of 0.47, with a 95% credible interval [0.25, 0.71]. The trajectories from groups 3 and 5, which had a success probability of 1, had an predicted probability of success of 0.96 [0.88, 1.00]. Groups 4 and 6 had a true success probability of zero, and the predicted success probability was 0.05 [0.00, 0.15].

5. Early Pregnancy Study

With the goal of distinguishing between cycles with high and low probability of early pregnancy loss, we applied our approach to conception cycles from the NC-Early Pregnancy study. In those cycles labeled clinical pregnancies, the investigators, based on hCG, concluded that the embryo had 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 hCG assay was highly accurate, and hCG does not typically occur in measurable levels outside of pregnancy, so there should be very little error in the identification of EPL. The data consisted of 165 conception cycles, 47 of which resulted in early losses. We randomly selected 10 early loss and 10 clinical pregnancy cycles to reserve for prediction, then fit the joint model to the remaining 145 cycles.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To illustrate the joint model for a trajectory and a response, we apply it to progesterone data for the early losses and the clinical pregnancies, with early loss status serving as the binary response for each cycle. The trajectory was defined to begin at the hormonally-determined day of ovulation and to last for up to 40 days. In most cases data collection was truncated before 40 days due to either the start of a menstrual period (in early losses) or a clinically identified pregnancy. To simplify presentation of results to clinicians, we purposely avoided adjusting for covariates in our analysis. Figure 3 presents some of the data from early losses and clinical pregnancies. It shows how PdG tends to rise when conception occurs and then drop off if the pregnancy is lost.

Using the methods applied in Section 4, the 145 women were grouped into clusters. There were 31 of these clusters, though 18 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 and 5 show the data for each of the 31 clusters, pointwise means and credible intervals for the trajectories in each cluster, and the model-estimated probabilities and credible intervals that a cycle in a given cluster was an early loss.

With the exception of the 8th cluster, which contained one early loss and two clinical pregnancies (though both clinical pregnancies had no PdG measurements after day 16, before the early loss occurred), every cluster was homogeneous with respect to early loss status. The early loss status of each cluster is apparent from the point estimate of the early loss probability below each graph. The second cluster was the largest, consisting of 59 clinical pregnancies, so that 55% of all clinical pregnancies were grouped together. While both the early loss and clinical groups had outliers, the early losses were more spread out among several clusters. Ignoring the 10 early loses and eight clinical pregnancies classified alone, the remaining 27 early losses were spread out among 9 clusters, whereas the 100 remaining clinical pregnancies were in only 5 clusters.

The 20 trajectories reserved for prediction were then analyzed. We wish to be able to predict the early loss status of a future observed trajectory. It would be most desirable to predict the early loss before it occurs, based on a short sequence of PdG measurements, so that any appropriate action (e.g. progesterone supplementation) could be taken. In order to examine the utility of PdG trajectories for estimating the potential for early loss, we truncated each of the reserved cycles, estimated the probability of early loss, then added additional days and examined how the probability changed. The estimated probabilities as a function of the number of days used to predict are given in Figure 6. With a few exceptions, the model assigned high probabilities of early loss to the early loss cycles, and low probabilities to the clinical pregnancies.

Figure 7 shows the average predicted probabilities of early loss for the reserved cycles, comparing early loses to clinical pregnancies. The embryo implants 6–10 days after ovulation, which is when the curves begin to diverge. Shortly after implantation, the model accurately discriminated between EPL and non-EPL cycles.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

6. Discussion

This article has applied a flexible semiparametric Bayes Dirichlet-proces based joint modeling approach, with the goal of using post-ovulatory progesterone measurements to the predict early pregnancy loss. Progesterone acts to maintain the pregnancy, so there is thought among clinicians that treatment with exogenous progesterone may prevent certain early losses. In order to implement such an intervention, it is necessary to accurately identify women at risk of loss before the loss is irreversible. Our results suggest that is should be possible to identify at-risk pregnancies a few days after implantation. Affordable home devices are now available to measure the progesterone metabolite PdG, so a clinical trial could potentially be designed in which the devices are programmed to predict impending loss and signify the need for treatment.

The trajectory shapes in Figures 4 and 5 are quite interesting to reproductive biologists hypothesizing mechanisms of pregnancy loss, as well as to clinicians considering use of exogenous progesterone treatments to prevent fetal death. It is certainly the case that substantial heterogeneity exists in the shapes of the trajectories among both early losses and surviving pregnancies. This is in contrast to the idealized trajectories presented in medical training, and has an important impact on the use of progesterone data in targeted interventions.

Our proposed approach jointly grouped subjects into clusters based on both the predictor and response data. For functional predictors and a univariate response, there tends to be much more data in the predictor component, so that the subject's response data has minimal impact on the posterior distribution of the cluster allocation. This is appealing from an interpretability standpoint, since it tends to be counter-intuitive to epidemiologists to have the functional clusters dependent on the response. However, if we are interested in allowing the response data to inform more about the clusters, a possible strategy is to weight the response likelihood in allocating subjects to clusters. For example, a weight of zero would correspond to unsupervised clustering, a weight of one to supervised clustering (our focus), and weights greater than one to response-driven clustering. This is an interesting area for future work.

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. This research was supported by the Intramural Research Program of the NIH, and NIEHS.

Appendix A

Acceptance Probability

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 $b_K = d_1 = 0$.

We specified the prior on model M as a function of the dimension of model M, where $p(M) \propto p(k_M) = \binom{T}{k_M - 1}^{-1} K^{-1}$.

The following acceptance probability is discussed in Bigelow and Dunson (2005) for the trajectory-only model:

$$Q(M', M) = \min\left\{[1, \frac{p(\mathbf{y}|M', \mathscr{V})p(\mathscr{V}|M')p(M')S(M, M')}{p(\mathbf{y}|M, \mathscr{V})p(\mathscr{V}|M)p(M)S(M', M)}\right\} \quad (A.1)$$

where $\mathscr{V}$ is a set of model parameters and $S(M', M)$ is the proposal density for model M' conditional on the current model, M. A MCMC sampler with this acceptance probability and the proposal and prior structure described above has limiting distribution $p(M|\mathscr{V}, \mathbf{y})$.

The term $R = \frac{p(M')S(M, M')}{p(M)S(M', M)}$ simplifies to a known constant. Let $\mathscr{T}$ 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 randomly sampling from $\mathscr{T}$. Since the proposal distribution is discrete, the need for a Jacobian in the acceptance probability is eliminated. If T, the number of elements in $\mathscr{T}$, 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 proposal of new basis functions through a continuous process.

The acceptance probability for the joint modeling example is therefore,

$$Q(M', M) = \min\left\{[1, \frac{p(\mathbf{y}, \mathbf{z}|M', \mathscr{V})p(\mathscr{V}|M')}{p(\mathbf{y}, \mathbf{z}|M, \mathscr{V})p(\mathscr{V}|M)} \times R\right\} \quad (A.2)$$

To maximize sampler efficiency, we would prefer that $\mathscr{V}$ be empty so that the chain converges to give us samples from $p(M|\mathbf{y}, \mathbf{z})$, the posterior over the model space. In the model described in this paper, choosing $\mathscr{V}$ to be empty renders Q intractable, so we resort to conditioning on the current values of a few model parameters and sampling from the conditional posterior distribution over the model space. If we alternate updating the elements of $\mathscr{V}$ from their full conditionals in a Gibbs sampler and updating the model using the acceptance probability Q, we obtain samples over the marginal posterior distribution of the model space.

The quantity $p(\mathbf{y}, \mathbf{z}|M, \mathbf{S}, \boldsymbol{\delta}, \lambda)$ can be computed directly, where all parameters are model-dependent but the model indicator M is suppressed for notational simplicity, so $p(\mathbf{y}, \mathbf{z}|M, \mathbf{S}, \boldsymbol{\delta}, \lambda) = \int p(\mathbf{z}, \mathbf{y}|\mathbf{a}, \boldsymbol{\gamma}, \mathbf{S}, \mathbf{b}, \tau, \lambda, M)p(\mathbf{a}, \boldsymbol{\gamma}, \mathbf{b}, \tau, \boldsymbol{\beta}|\mathbf{S}, \boldsymbol{\delta}, \lambda, M)\, d\mathbf{b}\, d\boldsymbol{\beta}\, d\tau\, d\mathbf{a}\, d\boldsymbol{\gamma}, = \int p(\mathbf{z}|\mathbf{a}, \boldsymbol{\gamma}, \mathbf{S}, M)p(\mathbf{a}, \boldsymbol{\gamma}|\mathbf{S}, M)d\mathbf{a}\, d\boldsymbol{\gamma} \times \int p(\mathbf{y}|\mathbf{S}, \mathbf{b}, \tau, \lambda, M)p(\mathbf{b}, \tau, \boldsymbol{\beta}|\mathbf{S}, \boldsymbol{\delta}, \lambda, M)\, d\mathbf{b}\, d\boldsymbol{\beta}\, d\tau$ and the first integral is the same for all $M \in \mathscr{M}$, thus for the purposes of model comparison

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$p(\mathbf{y}, \mathbf{z} \mid M, \mathbf{S}, \boldsymbol{\delta}, \lambda) \propto \int p(\mathbf{y} \mid \mathbf{S}, \mathbf{b}, \tau, \lambda, M) p(\mathbf{b}, \tau, \boldsymbol{\beta} \mid \mathbf{S}, \boldsymbol{\delta}, \lambda, M)\, d\mathbf{b}\, d\boldsymbol{\beta}\, d\tau = p(\mathbf{y} \mid M, \mathbf{S}, \boldsymbol{\delta}, \lambda)$$

where $p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b}, \tau, \boldsymbol{\delta}, \lambda, M)$ is the trajectory 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. Note that the acceptance probability depends on the trajectory data alone and not on the response data. This is useful in that the acceptance probability need not be derived separately for various functional forms of the likelihood of $\mathbf{z}$.

This integral has a closed form, so that the likelihood can be written:

$$p(\mathbf{y} \mid M, \boldsymbol{\delta}, \lambda) = C(\lambda, k) |\mathbf{R}|^{-\frac{1}{2}} \left(b_\tau + \frac{\alpha}{2}\right)^{-\left(\frac{n}{2} + a_\tau\right)} \prod_{l=1}^{k} \delta_l^{\frac{m}{2}} \prod_{i=1}^{m} |\mathbf{U}_i|^{\frac{1}{2}} \tag{A.3}$$

where

$$\mathbf{U}_i = (\Delta + \mathbf{H}_i' \mathbf{H}_i)^{-1}$$

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

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

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

The final adjustment to the acceptance probability is to note that the appropriate dimension of δ varies among elements of $\mathcal{M}$. We define $\boldsymbol{\delta}_{\text{adj}}$ to be the subvector of $\boldsymbol{\delta}$ corresponding to the basis functions common to both M and M' and use the acceptance probability

$$Q(M', M) = \min\left\{1, \frac{p\left(\mathbf{y} \mid M', \mathbf{S}, \boldsymbol{\delta}_{\text{adj}}, \lambda\right)}{p\left(\mathbf{y} \mid M, \mathbf{S}, \boldsymbol{\delta}_{\text{adj}}, \lambda\right)} \times R\right\} \tag{A.4}$$

Since M and M' differ by at most one basis function, finding the marginal likelihood under each model will require either using (A.3) directly or integrating out one element of $\boldsymbol{\delta}$ numerically (we used the Laplace method).

References

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. [PubMed: 9455822]

Bigelow J and Dunson D (2005). Posterior simulation across nonparametric models for functional clustering. Duke University ISDS Discussion paper 05–18.

Biller C (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. Journal of Computational and Graphical Statistics 9, 122–140.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Brown E and Ibrahim J (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. Biometrics 59, 221–228. [PubMed: 12926706]

Brown P, Fearn T and Vannucci M (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. Journal of the American Statistical Association 96, 398–408.

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

Dahl D (2006). Model-based clustering for expression data via a Dirichlet process mixture model In Do K, Muller P and Vannucci M, editors, Bayesian Inference for gene expression and proteomics. Cambridge University Press.

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.

Escabias M, Aguilera A and Valderrama M (2005). Modeling environmental data by functional principal components logistic regression. Environmetrics 16, 95–107.

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.

Garcia-Escudero L and Gordaliza A (2005). A proposal for robust curve clustering. Journal of Classification 22, 185–201.

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

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

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

Heard N, Holmes C and Stephens D (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. Journal of the American Statistical Association 101, 18–29.

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

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

James G and Silverman B (2005). Functional adaptive model estimation. Journal of the American Statistical Association 100, 565–576.

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

Kleinman K and Ibrahim J (1998). A semi-parametric Bayesian approach to generalized linear mixed models. Statistics in Medicine 17, 2579–2596. [PubMed: 9839349]

Lau J and Green P (2007). Bayesian model-based clustering procedures. Journal of Computational and Graphical Statistics 16, 526–558.

Lindstrom M (2002). Bayesian estimation of free-knot splines using reversible jumps. Computational Statistics & Data Analysis 41, 255–269.

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. [PubMed: 1379267]

Medvedovic M and Sivaganesan S (2002). Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 18, 1194–1206. [PubMed: 12217911]

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

Muller H and Stadtmuller U (2005). Generalized functional linear models. Annals of Statistics 33, 774–805.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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. [PubMed: 11933037]

Ray S and Mallick B (2006). Functional clustering by Bayesian wavelet methods. Journal of the Royal Statistical Society, Series B 68, 305–332.

Smith M and Kohn R (1996). Nonparametric regression using Bayesian variable selection. Journal of Econometrics 75, 317–343.

Venners S, Liu X, Perry M, Korrick S, Zhiping L, Yang F, Yang J, Lasley B, Xu X and Wang X (2006). Urinary estrogen and progesterone metabolite concentrations in menstrual cycles of fertile women with non-conception, early pregnancy loss or clinical pregnancy. Human Reproduction 21, 2272–2280. [PubMed: 16798842]

Vitzthum V, Spielvogel H, Thornburg J and West B (2006). Prospective study of early pregnancy loss in humans. Fertility and Sterility 86, 373–379. [PubMed: 16806213]

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. [PubMed: 12620443]

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. [PubMed: 3393170]

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

Zhou C and Wakefield J (2006). A Bayesian mixture model for partitioning gene expression data. Biometrics 62, 515–525. [PubMed: 16918916]

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. [PubMed: 8774277]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

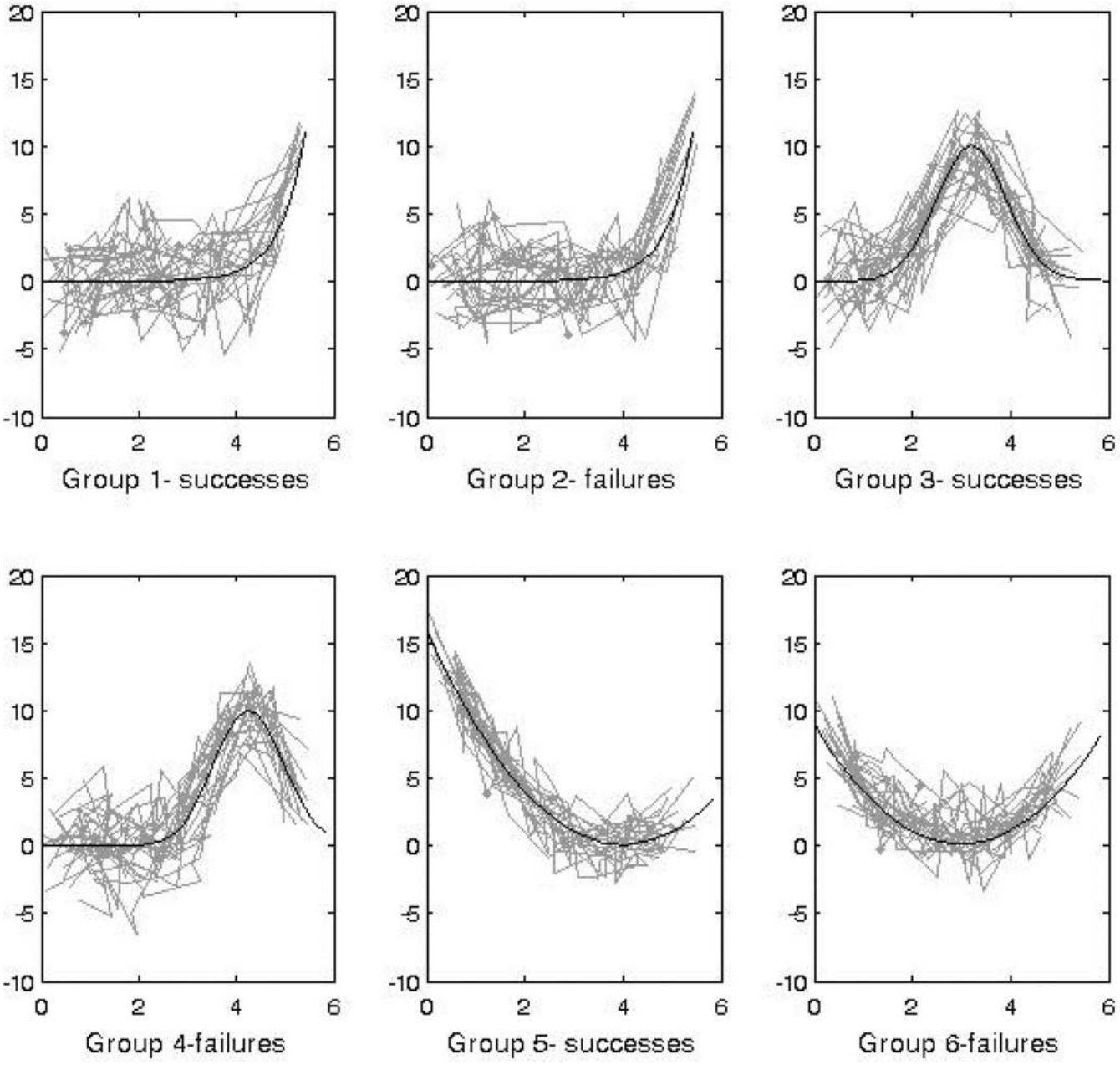

Figure 1.

Underlying population curves (black) and simulated trajectories (gray) for each of the six simulated groups along with the response (success or failure) for each group. The first two plots have the same underlying trajectories.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

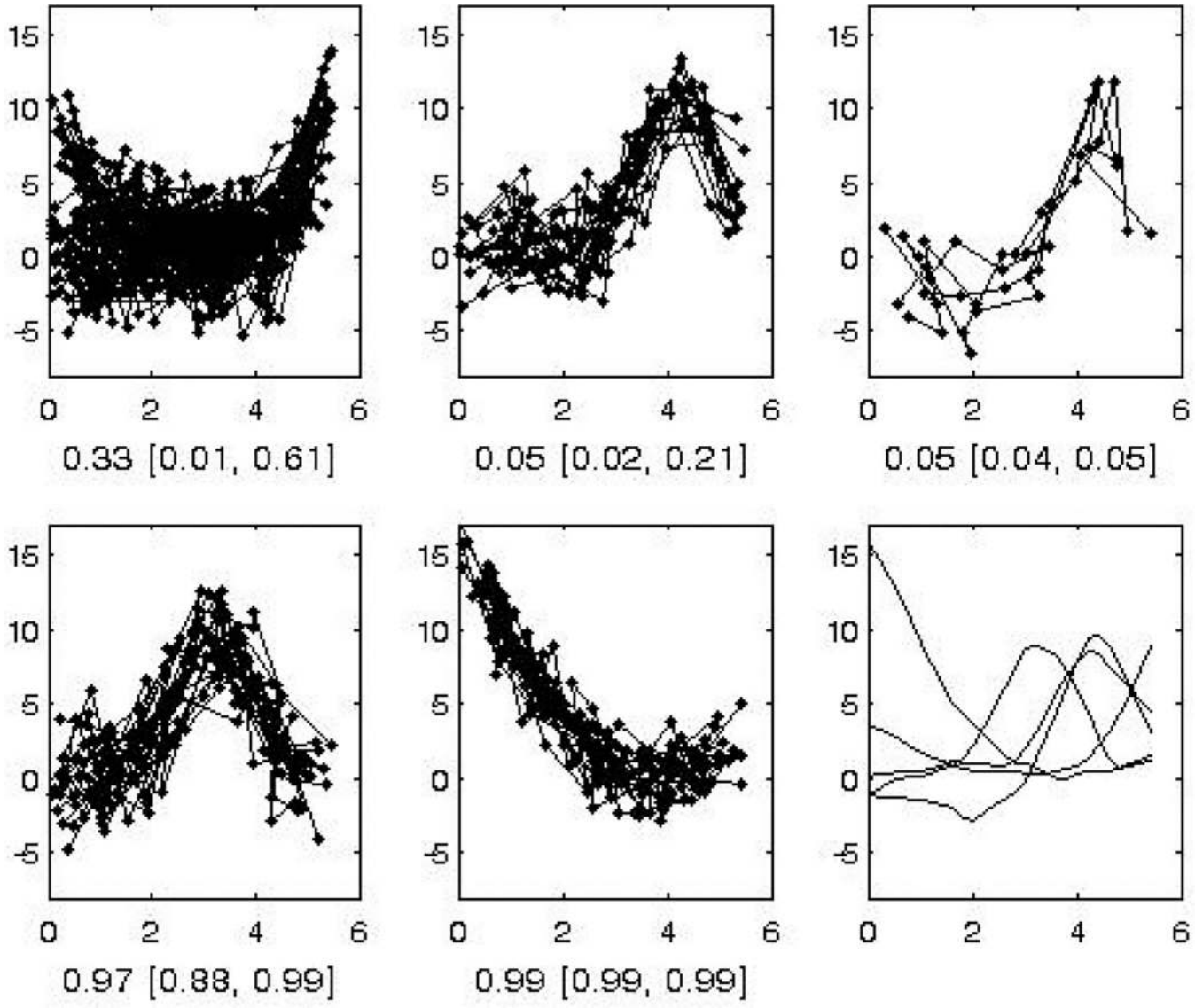

Figure 2.

The plot in the first quadrant is of the average trajectory of each of the final clusters. The five remaining plots contain all simulated data for the three final clusters. Also given are point estimates and credible intervals for the response probablitity corresponding to each of the identified clusters.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

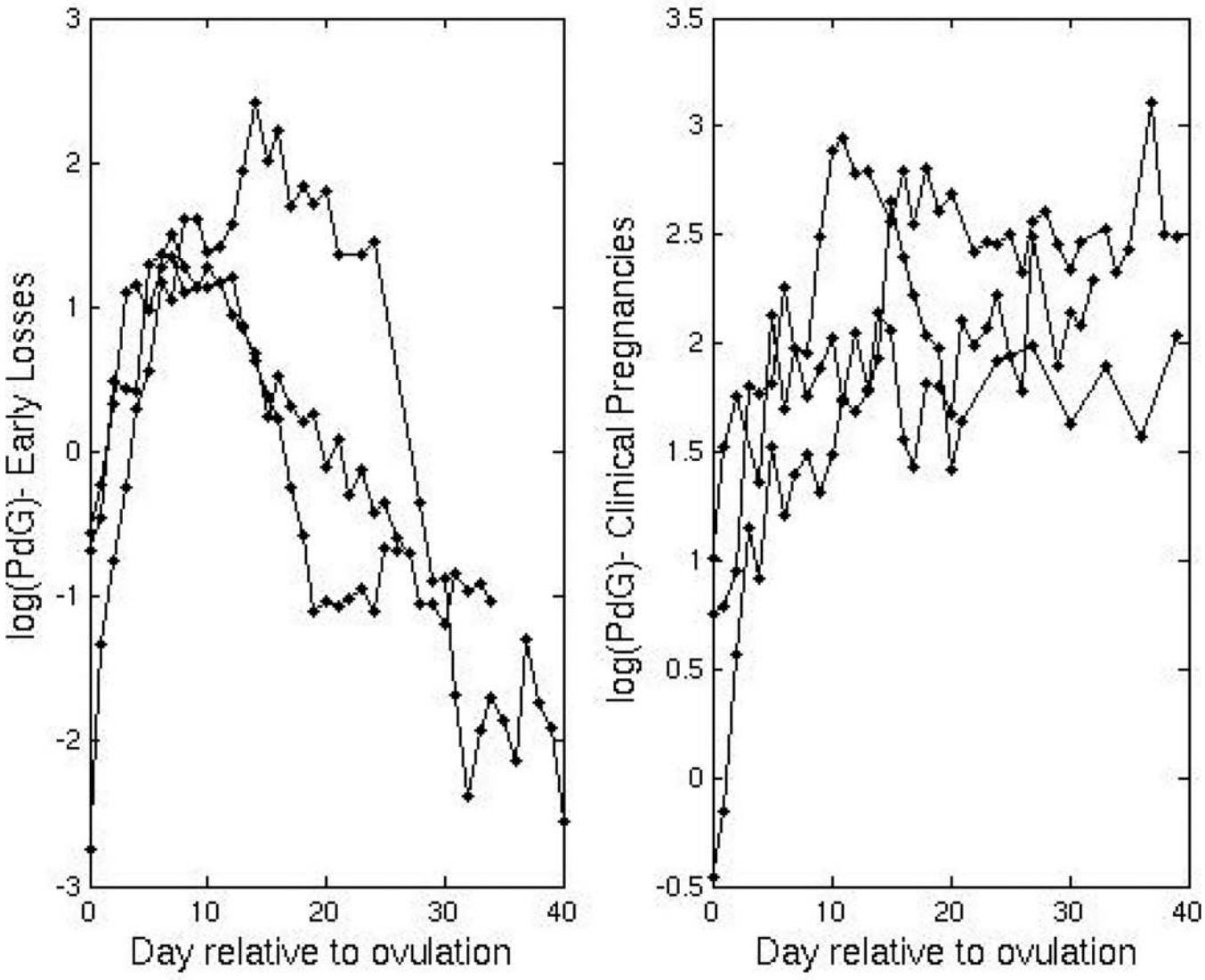

Figure 3.
Progesterone data beginning at the estimated day of ovulation for three early losses and three clinical pregnancies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

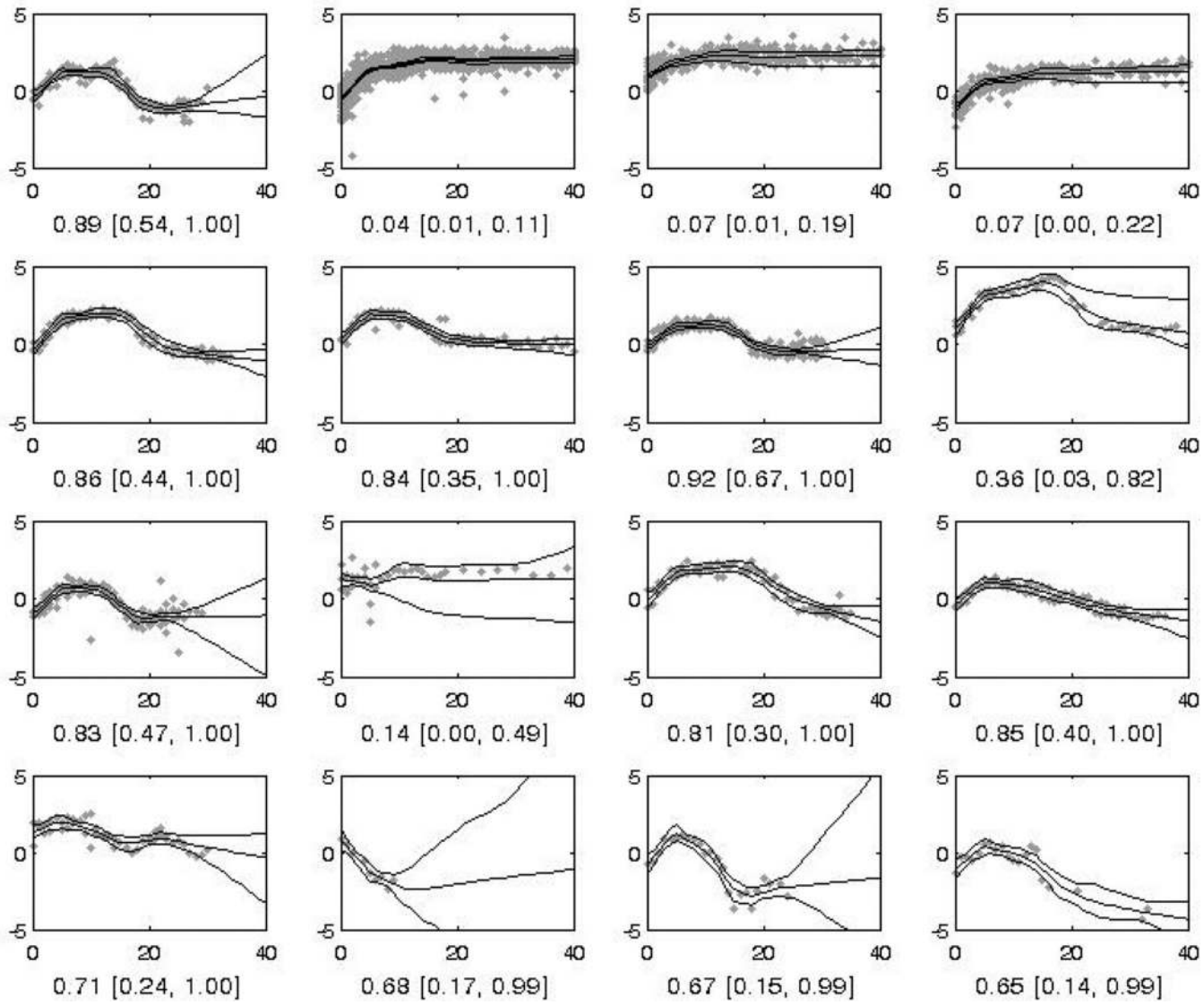

Figure 4.

PdG data from 16 of the 31 clusters, along with estimated trajectories and pointwise 99% credible intervals for ech cluster. Below each plot is the model-estimated probability of early loss for each cluster along with the 95% credible interval.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

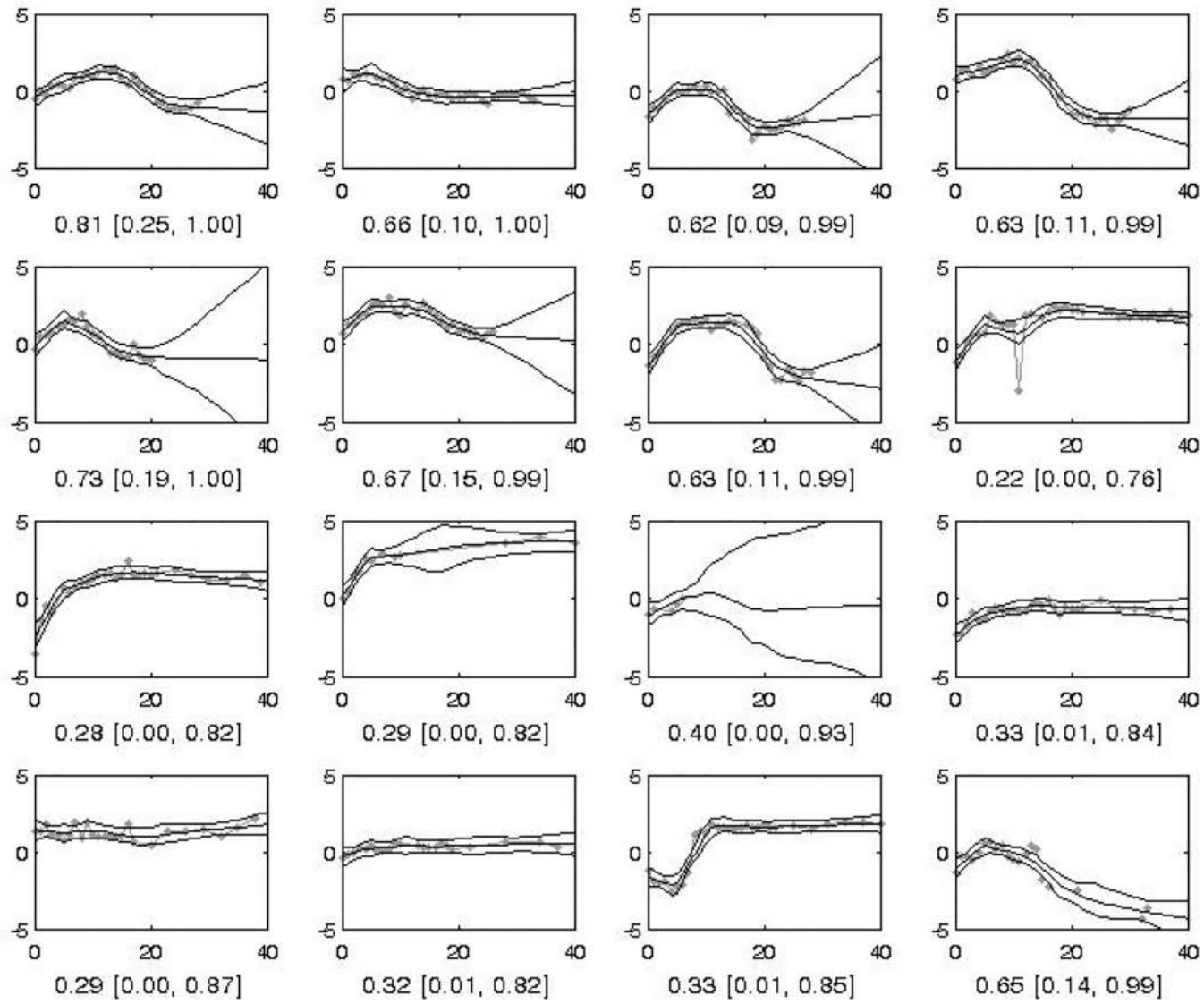

Figure 5.

PdG data for the 15 remaining clusters, along with estimated trajectories and pointwise 99% credible intervals for ech cluster. Below each plot is the model-estimated probability of early loss for each cluster along with the 95% credible interval.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

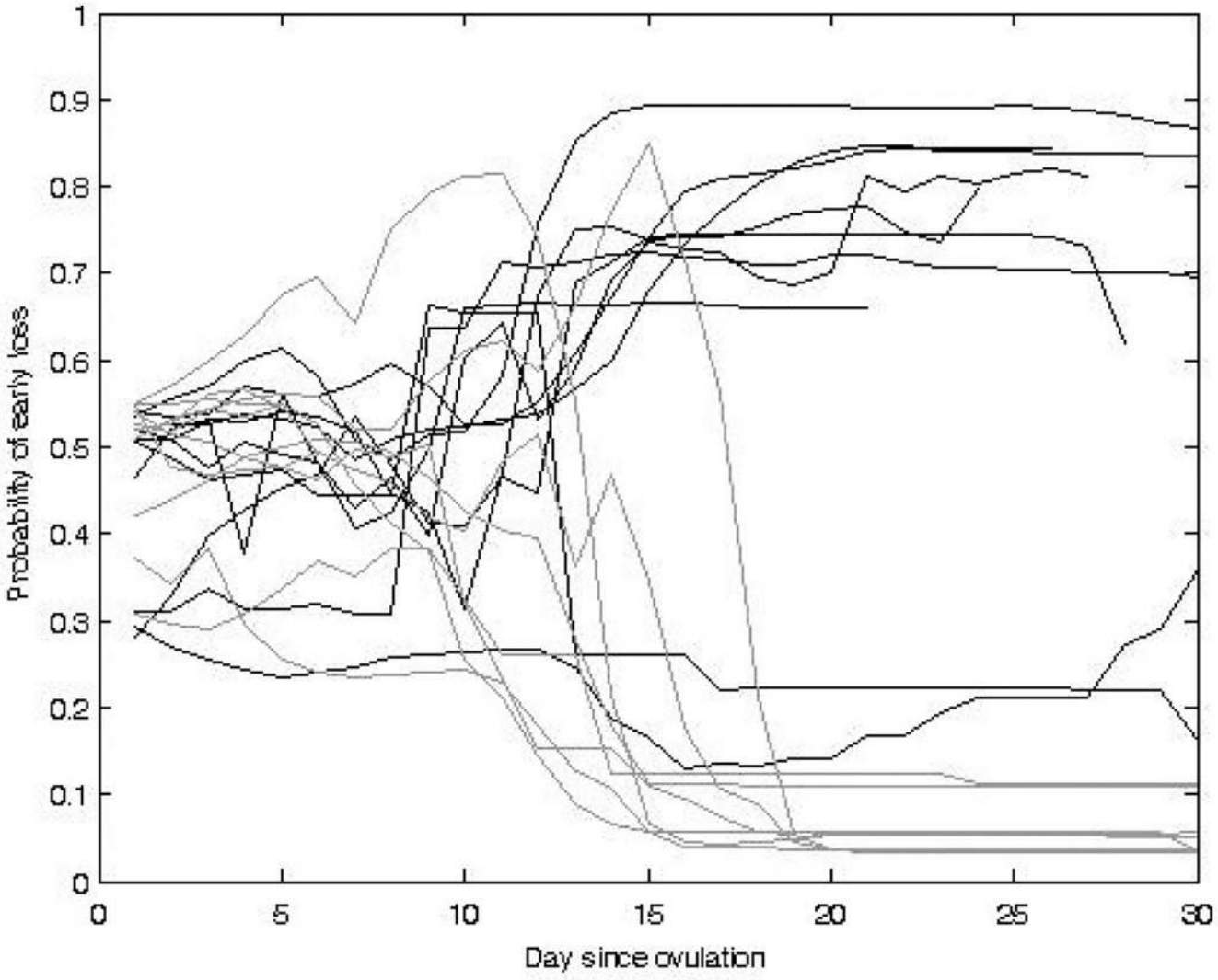

Figure 6.
Predicted probability of early loss for the reserved cycles, as a function of number of days since ovulation. Actual early loss cycles are black and clinical pregnancies are gray.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

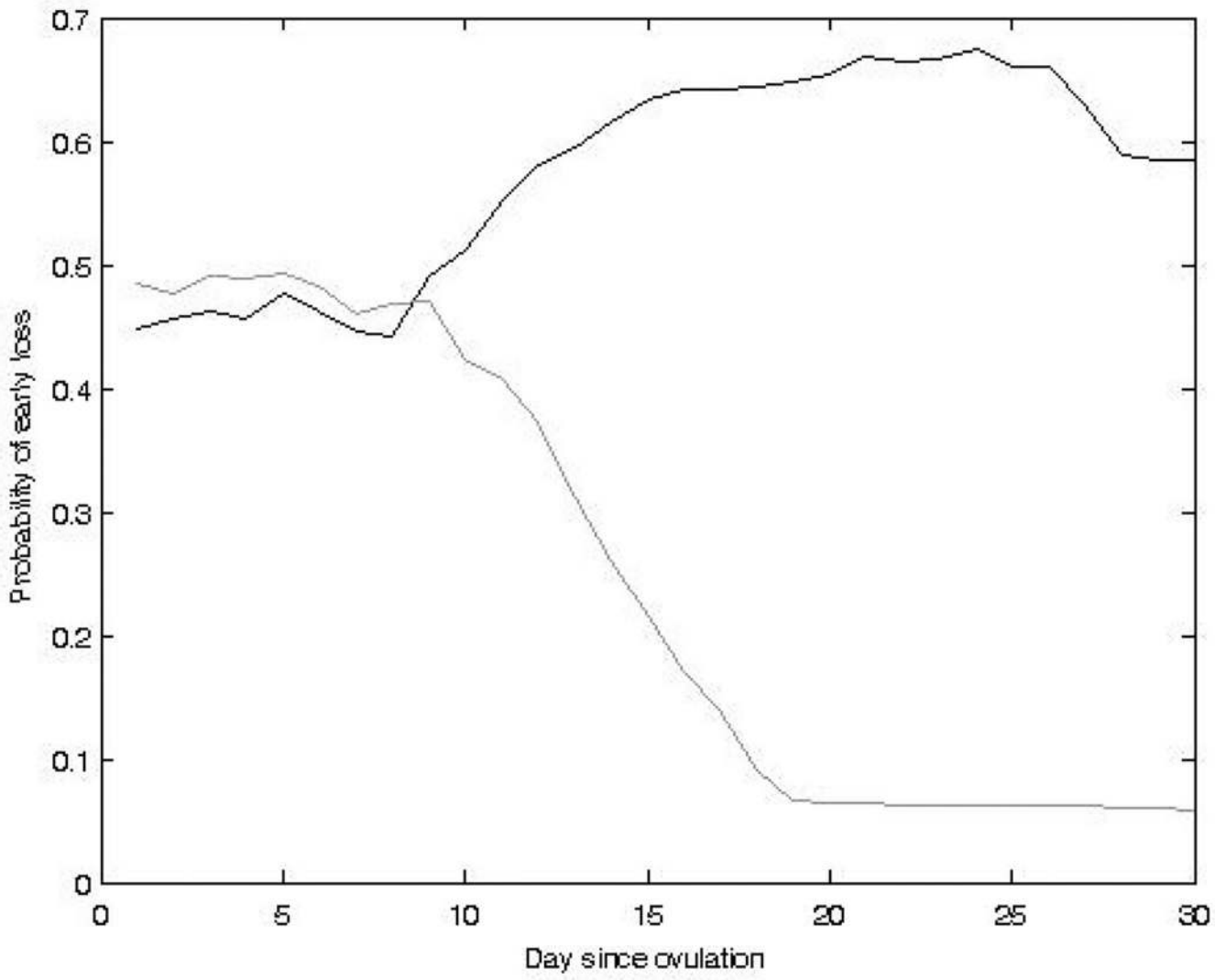

Figure 7.
Average predicted probability of early loss as a function of number of days since ovulation for the early loss (black) and clinical pregnancy (gray) cycles.