Semiparametric Bayesian Classification with Longitudinal Markers

Rolando De la Cruz-Mesía and Fernando A. Quintana

Departamento de Estadística, Facultad de Matemáticas,

Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, CHILE.

Peter Müller

Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center,

Box 447, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.

Summary. We analyze data from a study involving 173 pregnant women. The data are observed values of the β-HCG hormone measured during the first 80 days of gestational age, including from 1 up to 6 longitudinal responses for each woman. The main objective in this study is to predict normal versus abnormal pregnancy outcomes from data available at the early stages of pregnancy. We achieve the desired classification with a semiparametric hierarchical model. Specifically, we consider a Dirichlet process mixture (MDP) prior for the distribution of the random effects in each group. The unknown random effects distributions are allowed to vary across groups, but are made dependent by using a design vector to select different features of a single underlying random probability measure. The resulting model is an extension of the dependent Dirichlet process model, with an additional probability model for group classification. The model is shown to perform better than an alternative model based on independent Dirichlet processes for the groups. Relevant posterior distributions are summarized using Markov chain Monte Carlo methods.

1. Introduction

We develop a semi-parametric Bayesian approach for classification based on longitudinal markers. We define a suitable extension of hierarchical models to allow such classification. We introduce a new class of models building on the dependent Dirichlet process (DDP) models proposed in MacEachern (1999). In a motivating example we compare performance of the proposed model with parametric Bayesian inference and with traditional maximum likelihood based classification.

In many disease areas longitudinal markers allow early detection of a specific disease. A typical example is the use of prostate specific antigen (PSA) profiles over time as marker for prostate cancer (Morrell et al. 1995; Inoue et al. 2004). A common feature of inference related to such data is the need for classification rules that allow coherent and easy

sequential updating as the data for a new patient accrue over time. In this paper we propose a model-based semi-parametric Bayesian approach to classification that facilitates such sequential updating. The motivating application concerns the classification of pregnancies into normal and abnormal. To detect a number of complications during pregnancy, a variety of quantities are measured at the antenatal examinations. One of these clinical variables is the beta subunit of human chorionic gonadotropin (β-HCG) which shows dramatic changes in women during pregnancy. It has been established that values of the β-HCG are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes (France *et al.*, 1996). This association has made it possible to classify, with some degree of uncertainty, the outcome of pregnancy. The inference problem is formally described as a discriminant analysis based on the longitudinal β-HCG outcome.

Classical linear discriminant analysis classifies subjects into one of g groups or populations using multivariate observations. Usually, these vector-valued observations are obtained from cross-sectional studies and represent different subject characteristics such as age, gender or other relevant factors. In general, a common and unrestricted covariance matrix is assumed in the g different groups. Modifications of this method have also been used to classify subjects when the vector of multivariate observations represents repeated measures collected in a longitudinal study. Azen and Afifi (1972) introduced a two-stage model in which a discriminant function is obtained at each time point. In a second stage, the coefficients enter a linear regression versus time to obtain a slope and intercept. These slopes and intercepts are then used as input for a final linear discriminant function. This method is limited by the fact that multiple observations per subject are required to allocate a subject to one of g groups at any point in time.

Albert (1983) extended the classical concepts of discriminant analysis to multivariate response curves observed over fixed time intervals. Using interpolation or curve fitting procedures, a time-varying distance measure between the individual curve and group-specific curves is used to allocate a subject to a group. This methodology requires that the response curves in the training sample are fully observed over the considered time interval.

Albert and Kshirsagar (1993) proposed an exploratory method based on a growth curve structure embedded in a canonical variate analysis to achieve dimension reduction in a discriminant analysis framework. The authors suggested this approach for classification but did not apply it in that setting. No longitudinal data structures other than growth curves were considered.

An important limitation in the use of linear discriminant analysis for longitudinal data is that the method is only applicable for essentially balanced data, an increasingly exceptional situation in longitudinal studies. Therefore, an approach is needed that does not rely

on complete observations over time. In recent years some work has been done regarding discriminant analysis for longitudinal data using both linear and nonlinear random effects models. Tomasko *et al.* (1999) modified linear discriminant analysis using the mixed model MANOVA for the estimation of fixed effects and for a determination of various structures of covariance matrices, including unstructured, compound symmetry, and autoregressive of order 1. Brown *et al.* (2001) discussed Bayesian methods in discriminant analysis using linear random effects models. Marshall and Barón (2000) considered nonlinear random effects models to describe profiles in different groups and state the optimal allocation rule. Fieuws *et al.* (2002) used linear as well as nonlinear random effects models for the description of group-specific profiles. Recently, De la Cruz-Mesía and Quintana (2006) give a Bayesian version to the classification problem for longitudinal data.

All these approaches consider parametric models for the random effects. Unrelated to the classification problem, several recent references generalize restrictive parametric models for longitudinal data by placing a nonparametric prior on the random effects distribution. The literature includes, among many others, Bush and MacEachern (1996), Davidian and Gallant (1993), Ishwaran and Takahara (2002), Kleinman and Ibrahim (1998), Mentré and Mallet (1994), Müller, Quintana and Rosner (2004), Müller and Rosner (1997), Walker and Wakefield (1998), and Zhang and Davidian (2001). In this paper we develop a variation of these semi-parametric Bayesian longitudinal data models suitable for sequential classification as desired for inference with longitudinal markers. Specifically, we use an ANOVA-DDP model (De Iorio et al. 2004) to introduce semi-parametric random effects models that include dependence across the subpopulations of women with normal and abnormal pregnancies. We complete the model by adding a probability for group indicators. The augmented model for the repeated measurements and group indicators allows us to formalize the desired classification.

The paper is organized as follows. We first give a brief description of the dataset in Section 2. In Section 3, we extend the framework of traditional classification methods to the longitudinal hierarchical setting. Section 4 provides a discussion of nonparametric models based on the Dirichlet process, including methods for introducing dependence across related random probability measures. In Section 5 we illustrate the proposed longitudinal classification method using data on β-HCG measured in women with normal and abnormal pregnancy outcomes. An appropriate posterior simulation scheme based on the Gibbs sampling algorithm is described. Lastly, Section 6 concludes with a final discussion.

2. Data

We consider a data set reporting repeated measurements on β-HCG for $n = 173$ young women, representing 173 different pregnancies over a period of two years in a private obstetrics clinic in Santiago, Chile. The values of β-HCG were measured during the first 80 days of gestational age. The women were classified as normal pregnancies if they had a normal delivery, or as abnormal pregnancies if they had any complication resulting in a non-terminal delivery and loss of the fetus. The 173 women altogether contribute a total of 375 observations. Each woman is measured from 1 up to 6 times. These data were originally presented in Marshall and Barón (2000). Approximately 30 percent of the women had one β-HCG measurement, 31 percent had two, 33 percent had three, and 6 percent had four or more measurements.

Figure 1 presents the subject-specific log β-HCG profiles for normal and abnormal women. The two populations appear clearly distinct when considering the ensemble of profiles. However, for any one of the profiles the classification into one or the other subpopulation is far less certain, in particular when considering series of partial responses. The main inference goal in analyzing these data is to provide a rule to classify a new patient. The rule should allow sequential updating as data accrues for the new patient. The classification will critically hinge upon the implied inference on the distribution of profiles for each of the two subpopulations. The proposed semi-parametric model defines a richer class of random effects distributions than other models.

3. Classification Using Hierarchical Models

We use an augmentation scheme of semi-parametric longitudinal data models to develop the desired model-based classification for longitudinal marker data.

Let $y_i = (y_{i1}, \ldots, y_{in_i})'$ represent the observed response vector for the i-th patient, recorded at known times $t_i' = (t_{i1}, t_{i2}, \ldots, t_{in_i})$. Here n_i is the number of repeated measurements recorded for patient i. Let $x_i \in \{0, 1, \ldots, g-1\}$ denote the known class label for the ith patient. In our application $g = 2$, with $x_i = 0$ and $x_i = 1$, indicating normal and abnormal pregnancy, respectively. The label x_i is known for women with already reported delivery, but unknown for women with partial data before delivery. Without loss of generality we assume that x_i, $i = 1, \ldots, m$, is known, and x_{m+1} is unknown. Also without loss of generality we assume that $x_i \in \{0, 1\}$ takes only two possible values. Let $y^m = (y_1, \ldots, y_m, x_1, \ldots, x_m)$ denote all data, including the recorded class memberships x_i, up to the m-th patient. The classification problem is formalized as reporting $p(x_{m+1} \mid y^m, y_{m+1})$. Here y_{m+1} is the currently available partial response vector for the new patient $m + 1$. We now construct a probability model to allow evaluation of the desired classification probabilities.

Consider a generic semi-parametric hierarchical model of the form

$$(y_i \mid \theta_i) \sim p(y_i \mid \theta_i), \qquad (\theta_i \mid x_i, \phi, G_0, G_1) \sim G_{x_i}(\theta_i \mid \phi), \qquad (G_0, G_1 \mid \psi) \sim F_\psi. \tag{1}$$

In words, data y_i for the ith experimental unit is sampled from a probability model parameterized by a random effects vector θ_i. The θ_i are generated from a random effects distribution G_x, with $x = x_i$. The random effects distribution depends on a covariate x_i specific to the i-th sampling unit and possibly additional hyperparameters ϕ. In general, the parameter vector θ_i might be partitioned into common fixed effects θ^F and unit-specific random effects θ_i^R. Fixed effects are in common to all patients, and have no patient index i. In our example we use this partition. The model is completed by assuming a prior model for the unknown G_x. If G_x were indexed by a finite dimensional vector of hyperparameters, for example, normal moments, then the model would reduce to a traditional parametric hierarchical model. In contrast, in a non-parametric Bayesian approach G_x is assumed to be a random probability measure with an appropriate prior probability model F_ψ for the unknown distribution. In other words, F_ψ is a distribution on distributions. Here ψ indicates hyperparameters in the definition of F_ψ. A popular approach is to assume that each G_x arises from a Dirichlet process prior, independently across x, conditional on ψ. The random measures could be linked at the level of the hyperparameters. We discuss more details of this construction and the proposed alternative model in the next section.

For the top-level sampling model $p(y_i \mid \theta_i)$ in (1) we assume a nonlinear regression

$$y_{ij} = f(\theta_i; t_{ij}) + \epsilon_{ij}, \tag{2}$$

with a mean function $f(\theta; \cdot)$ parameterized by θ and evaluated at known times t_{ij}, $j = 1, \ldots, n_i$. The residual term ϵ_{ij} is assumed to be normally distributed with mean 0 and variance σ^2.

Model (1) specifies a joint probability model

$$p(y_1, \ldots, y_m \mid \phi, x_1, \ldots, x_m, \psi),$$

after marginalizing with respect to G_0, G_1 and θ_i, $i = 1, \ldots, m$. To facilitate classification we augment the model with a marginal probability for x_i:

$$Pr(x_i = x) = \pi_x. \tag{3}$$

The augmented model implies the desired classification as a conditional probability $p(x_{m+1} \mid y^m, y_{m+1})$, marginalizing with respect to the unknown θ_i, G_x and other possibly unknown hyperparameters.

In maximum likelihood classification theory, the probability that a future unit y_{m+1} belongs to group or population x, is estimated as

$$p(x_{m+1} = x \mid y_{m+1}, y^m, \hat{\Theta}) \propto \pi_x \, p(y_{m+1} \mid x_{m+1} = x, \hat{\Theta})$$

where $\hat{\Theta}$ indicates the maximum likelihood estimate of the fixed-effect parameters that remain after integrating out all the random effects. The unit is then classified in that group for which the highest probability is attained.

From a Bayesian viewpoint the classification probabilities are obtained by weighting with the posterior distributions of the parameters. Let $\Theta = (\phi, \psi, \theta_1, \ldots, \theta_m, \theta_{m+1})$ denote the vector of all parameters in the model, including those for the new $(m+1)$-st patient. Using Bayes' rule we find the probability that a new unit y_{m+1} belongs to group x as

$$\begin{aligned} P(x_{m+1} = x | y_{m+1}, y^m) &= \int p(x_{m+1} = x | y_{m+1}, y^m, \Theta) p(\Theta | y_{m+1}, y^m)\, d\Theta \\ &\propto \int \pi_x p(y_{m+1} | \Theta, x_{m+1} = x) p(\Theta | y_m)\, d\Theta. \end{aligned} \tag{4}$$

To verify (4) use $p(x_{m+1} = x | y_{m+1}, y^m, \Theta) = \pi_x p(y_{m+1} | \Theta, x_{m+1} = x) / p(y_{m+1} | \Theta)$, and $p(\Theta | y_{m+1}, y^m) \propto p(y_{m+1} | \Theta) p(\Theta | y_m)$. The integration is usually analytically intractable. Therefore, we shall construct a set of Markov chain Monte Carlo (MCMC) samples $\{\Theta^{(b)}, b = 1, \ldots, B\}$ from the posterior distribution and use the Rao-Blackwellization

$$\hat{p}_x \equiv \frac{1}{B} \sum_{b=1}^{B} \pi_x \; p(y_{m+1} \mid \Theta^{(b)}, x_{m+1} = x) \tag{5}$$

to approximate (4). If the prevalences π_x are unknown hyperparameters as well, then (5) would use the imputed values $\pi_x^{(b)}$.

Using a *percentage correctly classified* loss function (McLachlan, 2004), the Bayes classification of a future y_{m+1} is given by

$$\hat{x}_{m+1} = \arg\max_x \{p(x_{m+1} = x \mid y_{m+1}, y^m)\}.$$

The unit is classified in that group for which the highest posterior probability is attained.

4. Semi-parametric Models for Longitudinal Classification

We now discuss specific choices for the random probability measure F_ψ in (1). We start with a review of the Dirichlet process (DP) and some extensions.

The DP is a probability measure on the space of distributions functions defined on some space $\mathcal{X}$ (equipped with a σ-field $\mathcal{B}$). We use $\mathrm{DP}(M, G^\star)$ to denote the DP, where $M > 0$ is a scalar (precision parameter) and $G^\star$ is a specified baseline distribution defined on $(\mathcal{X}, \mathcal{B})$. A random distribution function G on $(\mathcal{X}, \mathcal{B})$ generated from $\mathrm{DP}(M, G^\star)$ is almost surely discrete and admits the following representation. Letting δ_a denote a Dirac measure at a we have

$$G = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\mu_\ell}. \tag{6}$$

The weights ω_l and locations μ_ℓ are generated by the following stick-breaking scheme: $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1-z_r)$, $l = 2, 3, \ldots$, with $z_\ell \overset{\text{iid}}{\sim} \text{Be}(1, M)$, and $\mu_\ell \overset{\text{iid}}{\sim} G^\star$, independently of the ω_ℓ (Sethuraman, 1994).

The use of DPs to model random distributions entails some limitations. In particular, the random probability measure G is almost surely discrete. A commonly used extension to mitigate this limitation is the DP mixture model (Antoniak, 1974). DP mixture (DPM) models avoid the discreteness by introducing an additional convolution with a continuous kernel. This model has become popular in applied Bayesian nonparametric work. The typical DPM model assumes

$$\theta_1, \ldots, \theta_m \overset{\text{iid}}{\sim} G^M(\theta), \qquad \text{with} \quad G^M(\theta) = \int f(\theta \mid \mu)\, dG(\mu), \quad \text{and} \quad G \sim \text{DP}(M, G^\star), \quad (7)$$

that is, a mixture with a DP prior on the random mixing measure G. We use G^M to denote the random mixture model with mixing measure G. The mixture model (7) can be equivalently written as a hierarchical model by introducing latent variables μ_i and breaking the mixture as $\theta_i|\mu_i \sim f(\theta \mid \mu)$ and $\mu_i \overset{\text{iid}}{\sim} G$, $i = 1, \ldots, m$, and finally $G \sim \text{DP}(M, G^\star)$. One of the attractive features of DPM models is the straightforward nature of posterior MCMC simulation. The computational effort is, in principle, independent of the dimensionality of μ_i. Efficient MCMC simulation for general DPM models is discussed, among others, in Bush and MacEachern (1996), Escobar and West (1995), MacEachern and Müller (1998), Neal (2000) and Jain and Neal (2004).

Several papers have considered extensions of DP and DPM models to hierarchical models over related random distributions, as needed to model the joint prior on (G_0, G_1) in (1). Some of the earliest developments of dependent DP models appeared in Cifarelli and Regazzini (1978), who defined dependence across related random measures $\{G_x\}$ by introducing a regression for the baseline measure $G_x^\star$ of marginally DP distributed random measures, $G_x \sim DP(M, G_x^\star)$. The model is used, for example, in Muliere and Petrone (1993), who define dependent nonparametric models $G_x \sim \text{DP}(M, G_x^\star)$ by assuming a regression in the baseline measure $G_x^\star = N(\beta x, \tau^2)$. Comparing with the notation in (1), the hyperparameters here are $\psi = (M, \beta, \tau)$. Similar models are discussed in Mira and Petrone (1996) and Giudici, Mezzetti and Muliere (2003).

Linking the related nonparametric models through a regression on the parameters of the nonparametric models limits the nature of the dependence to the structure of this regression. MacEachern (1999) proposes the dependent DP (DDP) as an alternative approach to define a dependent prior model for a set of random measures $\{G_x\}$, with $G_x \sim$ DP marginally. Recall Sethuraman's stick-breaking representation (6) for the DP random measure, $G_x =$

$\sum_h \omega_{xh}\delta_{\mu_{xh}}$. The key idea behind the DDP is to introduce dependence across the measures G_x by assuming the distribution of the point masses μ_{xh} to be dependent across different levels of x, but still independent across h. In the basic version of the DDP the weights are assumed to be the same across x, that is, $\omega_{xh} = \omega_h$. To introduce dependence of μ_{xh} across x MacEachern (1999) uses a Gaussian process. An application to spatial modeling is further developed in Gelfand et al. (2005) by allowing the locations θ to be drawn from a random field (a Gaussian process). The same method to induce dependence is used in De Iorio et al. (2004) to achieve an analysis of variance (ANOVA)-type structure on μ_{xh} across x. Griffin and Steel (2006) introduce dependence in nonparametric distributions by making the weights in the Sethuraman representation dependent on the covariates. We chose to fix the weights w_{xh} across covariates and introduce the dependence through the point mass locations μ_{xh}, mainly because of computational simplicity.

The construction introduced in De Iorio et al. (2004) is a natural approach to introduce dependent DP measures to implement (1). Specifically, let $d_i' = (1, 0)$ if $x_i = 0$ and $d_i' = (1, 1)$ if $x_i = 1$. We assume:

$$\theta_i \sim G_x^M(\theta_i), \quad \text{with} \quad G_x^M(\theta) = \int \mathcal{N}(\theta|\alpha d_i, \tau^2)\, dG(\alpha), \quad G \sim \text{DP}(M, G_\psi^\star). \tag{8}$$

Here ψ indicates hyperparameters in the definition of the base measure. In words, the trick to construct dependent random measures G_x^M is to start with a random measure on the coefficients α. Depending on x_i, a design vector d_i selects a linear function of the α. Finally, using an additional convolution with a normal kernel we define continuous and dependent random measures G_x^M. Introducing latent variables α_i, model (8) can be equivalently rewritten as a hierarchical model:

$$\theta_i = \alpha_i d_i + \eta_i, \quad \alpha_i \sim G, \quad G \sim \text{DP}(M, G_\psi^\star), \tag{9}$$

with $\eta_i \sim \mathcal{N}(0, \tau^2)$. Let p denote the dimension of θ_i. The latent variable α_i is a $(p \times 2)$ random matrix. The first column, α_{i0}, is the random effects vector for a patient from group $x = 0$. The second column, α_{i1} is the offset to generate a random effect for a patient from group $x = 1$. The proposed modeling strategy implies that the α_{i0} parameters are estimated from data coming from both groups. At the same time, we can learn about possible dependencies between α_{i0} and α_{i1} that may be group-specific. Learning about such features is not possible with alternative models involving a priori independent nonparametric models, e.g. two independent DPs. Under a model with two independent DPs we would only learn about α_{i0} for patients from the group $x_i = 0$, and about $\alpha_{i0} + \alpha_{i1}$ for patients from the group $x_i = 1$. Inference about the dependence of α_{i0} and α_{i1} for a future patients would not be possible. Later, in Section 5.2 and Table 2, we will show how in the example the increased borrowing of strength in the dependent model leads to a small improvement

of the misclassification rate, from 21.9% to 19.6%. The model is completed with a sampling model for y_i, $y_i \sim p(y_i \mid \theta_i)$, a marginal prior, $Pr(x_i = 1) = \pi_1$, for x_i, and hyperpriors on unknown hyperparameters, including τ^2, M, ψ and π_1. See Section 5 for an example of specific choices in an application.

The equivalent hierarchical model (9) highlights the nature of the model as a DP mixture model, allowing the use of any of the posterior MCMC simulation methods proposed for such models. Compared to MCMC for DP mixture models, as summarized, for example, in MacEachern and Müller (2000), the only additional step is the imputation of the latent group indicators x_i. We briefly summarize key features of the MCMC implementation. The discrete nature of the DP random measure G implies a positive probability for ties among the latent quantities α_i in (9). The configuration of ties determines many details of the MCMC. Let k denote the number of distinct values among $\{\alpha_i,\ i = 1, \ldots, m\}$ and let $\{\alpha_j^*,\ j = 1, \ldots, k\}$ denote such values. Recall from the discussion after (6) that $\alpha_j^* \sim G^\star$, i.i.d. We define configuration indicators s_i with $s_i = j$ if and only if $\alpha_i = \alpha_j^*$. The unique values $(\alpha_1^*, \ldots, \alpha_k^*)$ and configuration indicators s together provide an alternative representation of $(\alpha_1, \ldots, \alpha_m)$. The marginal prior for $(s_1, \ldots, s_m \mid G^\star, M)$, marginalizing in particular with respect to the random probability measure G can easily be described. It is known as the Polya urn scheme (Blackwell and MacQueen, 1973). This fact greatly simplifies posterior MCMC simulation for the DP mixture models, such as (9) together with (3) and the sampling model (2). We outline the transition probabilities used in the MCMC implementation. We use notation $[x \mid y, z]$ to indicate that the parameter x is updated conditional on currently imputed values for y and z. We use Y to generically denote all data, θ to indicate the set of all θ_i, and $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_m)$, etc. Also, $\phi = (\tau^2, \beta_{1x}, \beta_{2x}, \sigma_x^2,\ x = 0, 1)$ and $\psi = (\xi, R, M)$. Each iteration of the MCMC consists of the transition probabilities $[\alpha_j^* \mid s, \theta, \psi, \phi]$, $[\alpha_i \mid \theta_i, \alpha_{-i}, \psi, \phi]$, $[\theta_i \mid \alpha_i, \phi, Y]$, $[\beta \mid \theta, \phi, Y]$, $[x_{m+1} \mid \ldots]$, and transition probabilities to change the remaining parameters σ_x^2, τ^2, ξ, R, and M.

Note that changes in the sampling model (2) would not impact the transition probabilities for α_i and the parameters specific to the DP model. Updating θ_i proceeds like inference in a fully parametric model with sampling model (2) and normal prior $\theta_i \sim N(\alpha_i d_i, \tau^2)$. In other words, the computational effort related to the longitudinal model is the same as in a fully parametric model.

5. Application

5.1. Model specification

We apply the proposed model to the analysis of the longitudinal β-HCG data. Mean values of the log β-HCG for the 173 women show a nonlinear relationship with days of pregnancy. Figure 1 shows time profiles for normal and abnormal pregnancies. The analysis in Marshall and Barón (2000) suggests that woman-to-woman variation is adequately accounted for by the introduction of random effects to model the asymptotic behavior of the log β-HCG level (θ_i below). They proposed the following nonlinear random effects model. Recall that $y_i = (y_{i1}, \ldots, y_{in_i})'$ are the observed log β-HCG measurements at occasions $t_i = (t_{i1}, \ldots, t_{in_i})'$ for woman $i = 1, \ldots, m = 173$, and $x = 0, 1$ indicate, respectively, normal and abnormal pregnancy groups.

$$y_i|(x_i = x) \sim \mathcal{N}(\mu_{ix}, \sigma_x^2 \mathbf{I}_{n_i}), \qquad \text{with}$$
$$\mu_{ix} = \theta_i \left[1 + \exp\{-(t_i - \beta_{1x})/\beta_{2x}\}\right]^{-1}. \tag{10}$$

Here θ_i is a scalar subject-specific random effect, and $\beta_x = (\beta_{1x}, \beta_{2x})$, $x = 0, 1$, are bivariate fixed effects for the abnormal and the normal group, respectively. In model (10), the vector $(\theta_i, \beta_{1x}, \beta_{2x})$ characterizes the profile for the ith woman in group x. Marshall and Barón (2000) and De la Cruz-Mesía and Quintana (2006) assumed $\theta_i \sim \mathcal{N}(\theta_x, \tau_x^2)$.

A simple parametric model with a normal random effects distribution is adequate to describe subject-specific profiles and to fit smooth profiles to observed data. However, detailed features of the random effects model can critically change the predictive classification probabilities for patients with random effects that are imputed away from the center of the estimated random effects distributions. This leads us to consider the semi-parametric model (8), or its equivalent version (9). We will later compare the proposed non-parametric inference with a comparable parametric model and show how the nonparametric extension changes critical predictions.

In (9) we assume for the baseline distribution $G^\star$ a 2-dimensional normal distribution. Specifically, we take $G^\star = \mathcal{N}_2(\xi, R)$. To complete the model specification, we assume independent hyperpriors

$$\beta_x \sim \mathcal{N}_2(\beta_{0x}, B_{0x}), \quad \sigma_x^2 \sim \mathcal{IG}(a_{0x}, b_{0x}), \quad \tau^2 \sim \mathcal{IG}(c_0, d_0),$$
$$\xi \sim \mathcal{N}_2(\xi_0, \Sigma_\xi), \quad R^{-1} \sim \text{Wishart}_2\left(q, [qR_0]^{-1}\right) \tag{11}$$

Here, $\mathcal{IG}(a, b)$ denote the inverse gamma distribution, parameterized to have mean $1/(b(a-1))$. The Wishart prior on R^{-1} is parameterized such that $E(R^{-1}) = R_0^{-1}$. The first parameter of the Wishart distribution is the scalar parameter; the second is the matrix parameter.

The implementation of model (9) requires adopting specific values for M, a_{0x}, b_{0x}, c_0, d_0, q, β_{0x}, ξ_0, B_{0x}, Σ_ξ and R_0. The parameter M of the DP prior $\mathrm{DP}(M, G^\star)$ controls how close a realization of the process is to the baseline distribution $G^\star$. Additionally, in the DP mixture model, M controls the distribution of the number of distinct elements of the vector $(\alpha_1, \ldots, \alpha_m)$ and hence the number of distinct components of the mixture (see Antoniak, 1974, and De Iorio et al., 2004, for more details). Treating M as an unknown hyperparameter and assuming a gamma prior, $M \sim \mathcal{G}(a_M, b_M)$, Escobar and West (1995) derive an efficient posterior sampling scheme for M. We follow this approach, using $a_M = b_M = 1$.

The values of the other hyperparameters in (11) were taken as $\beta_{00} = \beta_{01} = \xi_0 = (0, 0)^T$, $B_{00} = B_{01} = \Sigma_\xi = 10000 \times \mathbf{I}_2$, $q = 3$, $R_0 = \mathbf{I}_2$, $a_{00} = a_{01} = c_0 = 3$, and $b_{00} = b_{01} = d_0 = 0.01$. These choices imply a prior mean variance of σ_k^2 and τ^2 equal to 2,500. Here $\mathbf{I}_2$ is the 2×2 identity matrix. Prior probabilities of group membership were assumed proportional to the size of the groups in the training sample. We also performed the analysis with different hyperparameter values, obtaining very similar results. This suggests robustness to the hyperparameter choices.

Updating the latent mixture parameters α_i and the hyperparameters β_x, σ_x^2, τ^2 and M proceeds with standard posterior simulation methods for DP mixtures. See, for example, MacEachern and Müller (1998) and De Iorio et al. (2004) for a full description of the Gibbs sampling scheme.

The full conditionals for implementing the Gibbs sampler, are not available in closed-form for β_{1x} and β_{2x}. To update β_{1x} and β_{2x} we thus use a Metropolis-Hastings step with a normal approximation to the full conditional as the candidate distribution. Resampling M is done by introducing a latent beta-distributed variable, as described by Escobar and West (1995), based on West (1992).

To perform the Gibbs sampling, we chose starting points in a neighborhood of the MLEs of model parameters. In theory the Markov chain convergence and ergodic properties are independent of the initial values. In practice, however, a good choice of starting points shortens the number of iterations required until practical convergence. We generated 100,000 iterations. After 10,000 iterations, samples were collected, at a spacing of 90 iterations, to obtain approximately independent samples, leaving us with a total of $B = 1,000$ posterior Monte Carlo samples for calculating posterior quantities of interest.

To diagnose convergence, we used methods available in the BOA package (Smith, 2004). Because of the high dimensional parameter vector, we prefer to use diagnostics, such as those proposed by Geweke (1992), which do not require multiple parallel chains.

5.2. Results

Panel (a) in Figure 2 shows histograms of the subject-specific parameters θ_i estimated using the empirical Bayes methods as implemented in the SAS System. Specifically, it shows the posterior means of θ_i conditional on all ther other hyperparameters being evaluated at their MLE's, for model (10) with normally distributed random effects (see, e.g. Vonesh and Chinchilli, 1997). Panel (b) shows posterior predictive draws under the Bayesian semiparametric approach (BSP) for both abnormal and normal groups, respectively. In panel (b), a smooth curve shows the posterior estimated random effects distribution $\overline{G^M}(\theta) = E(G^M(\theta) \mid y^m)$. For comparison, panel (a) shows a kernel density estimate based on the histogram of the corresponding estimates. To evaluate the posterior mean $\overline{G^M}$ we exploit the identity $\overline{G^M}(\theta) = p(\theta_{m+1} \mid y^m)$, which follows from

$$p(\theta_{m+1} \mid y^m, x_{m+1} = x) = \int p(\theta_{m+1} \mid G_x^M) dp(G_x^M \mid y^m) = \int G_x^M(\theta_{m+1}) dp(G_x^M) = \overline{G^M}.$$

We can, therefore, approximate $\overline{G^M}$ by a kernel density estimate of posterior predictive draws, $\theta_{m+1} \sim p(\theta_{m+1} \mid y^m)$. The MLE estimates show asymmetry in the normal group and bimodality in the abnormal group. A nonparametric specification of the distribution of the random-effects allows for the flexibility to estimate such features. See Figure 2b.

The parameter M induces a distribution on the number of clusters into which the observations fall. Recall the definition of configuration indicators s_i in the discussion following (9). We refer to sets of observations with equal configuration indicators, i.e., a common value α_i, as clusters. The DDP model that we use to implement inference in this article relies on a single mass parameter, M. For this model, clusters of observations occur both within and across groups. The number of clusters is stochastically increasing with the number of observations (see De Iorio et al. 2004). Recall that k was defined as the number of clusters. Let k_x denote the number of clusters of observations in group x. We find the posterior mean $E(k_x \mid y^m)$ (standard deviation $\mathrm{SD}(k_x \mid y^m)$) to be 5.9 (1.8), and 5.2 (1.5) for $x = 0$ and $x = 1$, respectively. The posterior mean $E(k \mid y^m)$ (standard deviation $\mathrm{SD}(k \mid y^m)$) is 6.3 (1.9).

As part of the analysis we estimated individual β-HCG profiles and standard errors. These profiles can be used to assess goodness of fit. Fitted profiles with ± 2 posterior standard deviations curves are displayed for six selected patients in Figure 3. Three of them in the normal group (patients 2, 66, 75) and the remaining three in the abnormal group (patients 15, 29, 45). Based on these plots we informally assess the goodness of fit of the model to the data. The posterior inference captures the varying observation error between subjects.

Next, we consider the problem of evaluating the classification rule. This is naturally carried out through an estimate of the associated misclassification rates. At this point we

could apply the rule to the observed data and count the (relative) frequencies of misclassified observations. In doing so, we conclude that the BSP yields the best results (data not shown). However, it can be argued that this yields overly optimistic misclassification error rates as the same observations are used to determine and to evaluate the classification rule (McLachlan, 2004). Another traditional approach is cross-validation (Lachenbruch and Mickey, 1968). It computes the classification rule by leaving out one subject at a time and records whether this observation is correctly classified or not. Here, we classify an observation as abnormal if the posterior probability for $x_i = 0$ is greater than one half. Table 1 presents the results using cross-validation based on the method described in Section 4, the Bayesian parametric approach developed by De la Cruz-Mesía and Quintana (2006) and a frequentist method developed by Marshall and Barón (2000). We found interesting differences between the three approaches. The misclassification rate under the BSP model is 14.5% (25/173), which is less than under the Bayesian parametric (BP) model and the MLE-based method, 17.3% (30/173) and 18.5% (32/173), respectively. A traditional way to summarize the above results is a Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate for the different possible cutpoints of the classification rule (0.5 was used when calculating the results displayed in Table 1). Figure 4 shows this curve for both Bayesian models. We see how the BSP model improves upon the BP method (higher area under the curve). To further understand the corresponding classification, Figure 5 shows estimated classification probabilities for all 173 women, arranged by true x_i, and within each group sorted in decreasing order. We see how the BSP model dominates the BP model for most of the range, in the sense of implying higher and lower probabilities for normal and abnormal pregnancies, respectively. The most noticeable exception is the rightmost part of the abnormal cases (lowest classification probabilities), where this trend is reversed. But this is of little concern, as at that range of values for the probabilities, almost any rule would classify these women as abnormal.

The reported ROC curve provides a conservative comparison in the following sense. It is based on classification of patients with complete data recorded over the first 80 days of gestational age. More important for an informed clinical treatment decision are differences in early prediction, based on early responses only. To illustrate this use, we generate from the posterior predictive distribution data for one future patient for each group and evaluate (4) for up to five possible observations. Figure 6 shows how the classification probabilities change as we accrue more data. The figure compares inference under the proposed semi-parametric model and a corresponding parametric model fixing the random probability measure G at the base measure $G^\star$. For the normal patient, we observe a steady growth of the probabilities. In contrast, for the abnormal patient this probability first increases and starts to decrease to values that leave no question about the classification. A possible explanation for this

is the rather heterogeneous patterns found for abnormal patients. Indeed, many of these show an initial increase in the $\log_{10}$ β-HCG responses (just as all the normal patients do) followed by a decrease in some of the patients. Thus the classification probabilities for abnormal patients require a few more observations than the normal ones to reflect the correct outcome. For the abnormal patient the predictive classification probabilities using the BSP model decrease more rapidly than under the BP model. After two observations we find a difference greater than 10% in predictive probabilities. From a clinical perspective, a 10% difference in predicted probabilities can be key to making the right treatment decision at this critical early time. The ROC curve shown in Figure 4 evaluates classification based on profiles over the entire observation period. As shown in Figure 6, the improvement for classification based on the first two or three observations is even larger.

To assess the model fit and compare different models, we calculate the conditional predictive ordinate (CPO) (Gelfand *et al.*, 1995) for each observation. Chen *et al.* (Chap. 10, 2000) show in detail how to obtain Monte Carlo estimates of the CPO statistics. We can compare different models using sums of log CPOs of the individual observations. Define $\widehat{\text{CPO}}_i$ to be the Monte Carlo estimates of the ith subject's CPO statistic. Greater values of $S = \left(\sum \log \widehat{\text{CPO}}_i\right)$indicate a better fit. We found $S = -117.2$ for the BSP model. For the BP model we found $S = -124.1$. The difference suggests that the BSP model provides a marginally better fit to the log β-HCG data than its parametric counterpart.

We next investigate the effect of the dependence introduced in the DDP compared with a model with two independent DP mixtures. Figure 7 displays the results of 500 posterior predictive draws from the bivariate distribution $p(\alpha_{m+1} \mid Y)$. We can identify two large clusters, each suggesting negative correlation among main effect and abnormal pregnancy offset parameters. The resulting covariance structure clearly differs among these components. Note that such findings would not be possible under a model with two independent DPs. To compare our model with that defined by two independent DP mixtures we changed (8) by using $d_i' = (1, 0)$ and $d_i' = (0, 1)$ for $x_i = 0$ and $x_i = 1$, respectively. We use iBSP to refer to the new model. For a fair comparison we use the same hyperparameter choices as before, implying in particular that the marginal probability models for the random effects distributions G_x^M, $x = 0, 1$, remain unchanged under the BSP and the iBSP models. We carried out the same inference as described in Figure 6, focusing on the classification for a future woman, $m + 1$, after the first $n_{m+1} = 2$ observations, assuming that the unknown truth is $x_{m+1} = 0$, i.e., an abnormal pregnancy. Figure 6 reports the classification probabilities $Pr(x_{m+1} = 1 \mid y_{m+1,1}, y_{m+1,2}, y^m) = 50\%$ for the proposed BSP model, and 63% for the BP model. For the iBSP model we find a probability of 55%, justifying the minor additional effort to implement the DDP model. However, this depends on a single patient, as just described. We investigated this issue further, considering the classification

of *every* patient, based only on the first two observations, and assuming the same proportion of normal pregnancies as was empirically observed. This is essentially equivalent to a cross-validation of the inference for all patients. Table 2 summarizes the classification as Normal/Abnormal under each of the competing models. The reported misclassification rates show an improvement under the dependent model compared to the independent model.

Finally, we investigate the effect of hyperparameter choices on the reported inference. Again, consider the predicted classification for a future woman (assuming the unknown truth to be $x_{m+1} = 0$) based on one or two observations. Table 3 shows these probabilities for different combinations of M, $E(\tau^2)$ and $E(\sigma_x^2)$. The corresponding probabilities do exhibit some variation, but the implied classifications remain unchanged. In fact, the estimated error rates are the same as reported in Table 1 in all cases (data not shown).

6. Discussion

We have proposed a model-based approach to classification of longitudinal profiles. The underlying models in each group or population are given by nonlinear semiparametric models. Flexibility for relaxing the distributional assumptions is introduced using a nonparametric specification on the random effects models. Dependence in the growth curves is introduced through a design vector indicating group membership and selecting appropriate features of a common underlying random probability measure. The approach is appropriate for classifying longitudinal profiles of datasets with unbalanced data structure. It uses all available information for classifying subjects over time, regardless of the number or timing of the observations. Moreover, the influence on discrimination of both the between-group and within-group components variability can be readily quantified, and the posterior simulation scheme is straightforwardly implemented. The approach is particularly appropriate for decision-making in clinical practice where the number and times of observations are often arbitrary and depend on the progression of the patient.

A key feature of our approach is the flexibility provided by the nonparametric model for random effects. A straightforward generalization of our approach could accommodate more information available. This can be done by inclusion of more covariates or by considering other markers, thus extending the framework to a multivariate one.

Limitations of the proposed model are the reliance on posterior simulation and the nature of the non-parametric generalization. Although posterior simulation is straightforward, it does require some problem-specific software development. The non-parametric modeling is on the random effects distribution only, but still requires the user to chose a parametric model for $p(y_i \mid \theta_i)$. Alternative models could use the available data from patients $i = 1, \ldots, m$, to learn about the nature of the longitudinal dependence, using, for example,

methods reviewed in Denison et al. (2002). Another possible extension is the use of problem-specific decision rules. In the reported inference we classified patients by maximum posterior predictive probability of group membership. Alternatively, one could imagine an approach that takes into account the sequential nature of the decision problem. It is conceivable that even with high probability of abnormal pregnancy a clinician might decide to wait for one more measurement, trading off the additional information with a possible loss in treatment options.

Finally model (8) and (9) allows an easy generalization to more general nonparametric priors on G. In particular, one can easily replace the DP model by a species sampling model (Pitman 1996), which allows more general prior distributions on configurations of the α parameters. See further discussion of such models in Ishwaran and James (2003) and in Quintana (2006).

Acknowledgments

We wish to thank Professor Guillermo Marshall for giving us access to his β-HCG data set. The first author thanks the Comisión Nacional de Investigación Científica y Tecnológica - CONICYT, for partially supporting his Ph.D. studies at the Pontificia Universidad Católica de Chile. Part of the research was done while the first author was visiting at Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center, under grant MECESUP PUC 0103. Research supported by NIH/NCI grant 2 R01 CA75981-04A1 and by grants FONDECYT 1020712 and 1060729. We would also like to thank the editor, associate editor and the reviewers for their constructive comments, which helped to substantially improve this manuscript.

References

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

Albert, A. (1983). Discriminant analysis based on multivariate response curve: a descriptive approach to dynamic allocation. *Statistics in Medicine*, **2**, 95–106.

Albert, J. M., and Kshirsagar, A. M. (1993). The reduced-rank growth curve model for discriminant analysis of longitudinal data. *Australian Journal of Statistics*, **35**, 345–357.

Azen, S. P., and Afifi, A. A. (1972). Two models for assessing prognosis on the basis of successive observations. *Mathematical Biosciences*, **14**, 169–176.

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

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

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

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

Cifarelli, D. M., and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability. The use of associative means. (in Italian) *Annali del Instituto di Matematica Finianziara dell Universitá di Torino, Serie III*, 12, 1–36.

Davidian, M., and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, **80**, 475–488.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, **99**, 205–215.

De la Cruz-Mesía, R., and Quintana, F. (2006). A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles. *Biostatistics*. In press.

Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002), *Bayesian Methods for Nonlinear Classification and Regression.* New York: Wiley.

Escobar, M. D., and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Fieuws, S., Verbeke, G., and Brant, L. J. (2003). Classification of Longitudinal Profiles using Nonlinear Mixed-Effects Models. *Technical Report*. Biostatistical Centre, Katholieke Universiteit Leuven, Leuven, Belgium.

France, J. T., Keelan, J., Song, L., Liddell, H., Zanderigo, A., and Knox, B. (1996). Serum concentrations of human chorionic gonadotrophin and immunoreactive inhibin in early pregnancy and recurrent miscarriage: a longitudinal study. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, **36**(3), 325–330.

Gelfand, A. E, Dey, D.K., and Chang, H. (1995). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4*. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds). Oxford: Oxford University Press.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian Nonparametric Spatial Modelling With Dirichlet Processes Mixing. *Journal of the American Statistical Association*, **100**, 1021–1035.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*. Bernardo, J. M., Berger, J. O.,

Dawid, A. P., and Smith, A. F. M., eds., Oxford University Press, 169–194.

Giudici, P., Mezzetti, M., and Muliere, P. (2003). Mixtures of Dirichlet Process Priors for Variable Selection in Survival Analysis. *Journal of Statistical Planning and Inference*, **111**, 101–115.

Griffin, J. E., and Steel, M. F. J. (2006). Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Society*, **101**, 179–194.

Inoue, L.Y.T., Etzioni, R., Slate, E.H., Morrell, C. and Penson, D.F. (2004). Combining longitudinal studies of PSA. *Biostatistics*, **5** (3), 483–500.

Ishwaran, H. and James, L. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of the American Statistical Association*, **96**, 161–173.

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

Ishwaran, H. and James, L. J. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statistica Sinica*, **13**, 1211–1235.

Jain, S., and Neal, R. M. (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics*, **13**, 158–182.

Kleinman, K. P., and Ibrahim, J. G. (1998). A Semi-parametric Bayesian approach to the random effects model. *Biometrics*, **54**, 265–278.

Lachenbruch, P. A., and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

MacEachern, S. N. (1999). Dependent nonparametric process. in *ASA Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA.

MacEachern, S. N., and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**(2), 223–238.

MacEachern, S. N and Müller, P. (2000). Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models, in *Robust Bayesian analysis*, eds. Ríos-Insua, D. and Ruggeri, F., New York: Springer, pp. 295–315.

Marshall, G., and Barón, A. E. (2000). Linear Discriminant Models for Unbalanced Longitudinal Data. *Statistics in Medicine*, **19**, 1969–1981.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.

Mentré, F., and Mallet, A. (1994). Handling Covariates in Population Pharmacokinetics. *International Journal of Biomedical Computing*, **36**, 25–33.

Mira, A., and Petrone, S. (1996). Bayesian Hierarchical Nonparametric Inference for

Change-Point Problems. in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press.

Morrell, C.H., Pearson, J.D., Carter, H.B. and Brant, L.J. (1995). Estimating unknown transition times using a piecewise nonlinear mixed effects model in men with prostate cancer. *Journal of the American Statistical Association*, **90**, 45-53.

Muliere, P., and Petrone, S. (1993). A Bayesian Predictive Approach to Sequential Search for an Optimal Dose: Parametric and Nonparametric Models. *Journal of the Italian Statistical Society*, **2**, 349-364.

Muliere, P. and Secchi, P. (1995). *A note on a proper Bayesian Bootstrap*, Technical Report No. 18, Università degli Studi di Pavia, Dipartamento di Economia Politica e Metodi Quantitativi.

Müller, P., Quintana, F. A., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Ser. B*, **66**, 735–749.

Müller, P., and Rosner, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, **92**, 1279–1292.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme, *in* Ferguson, T. S., Shapeley, L. S. and MacQueen, J. B. (eds.) *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Hayward, California: IMS Lecture Notes - Monograph Series, pp. 245–268.

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference* **136**, 2407–2429.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.

Smith, B. J. (2004). Bayesian Output Analysis Program (BOA), Version 1.1.2 for S-PLUS and R. *Available at http://www.public-health.uiowa.edu/boa.*

Tomasko, L., Helms, R. W., and Snapinn, S. M. (1999). A discriminant analysis extension to mixed models. *Statistics in Medicine*, **18**, 1249–1260.

Vonesh, E. F. and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker, Inc.

Walker, S. G., and Wakefield, J. C. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya, Series B*, **60**, 196–212.

West, M. (1992). Hyperparameter Estimation in Dirichlet Process Mixture Models. Technical Report 92–A03, Duke University, ISDS.

Zhang, D., and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795–802.

Table 1. Classification results using Bayesian parametric (BP), semiparametric (BSP) and Classical methods.

	Classification						
	Classical		**BP**		**BSP**		
Groups	n	a	n	a	n	a	
Normal (n)	113	11	115	9	117	7	124
Abnormal (a)	21	28	21	28	18	31	49
	134	39	136	37	134	39	173

Table 2. Cross-validation using the first two observations for DDP and DP

	Classification				
	DDP		**DP**		
Groups	n	a	n	a	
Normal(n)	118	6	116	8	124
Abnormal(a)	28	21	30	19	49
	146	27	146	27	173

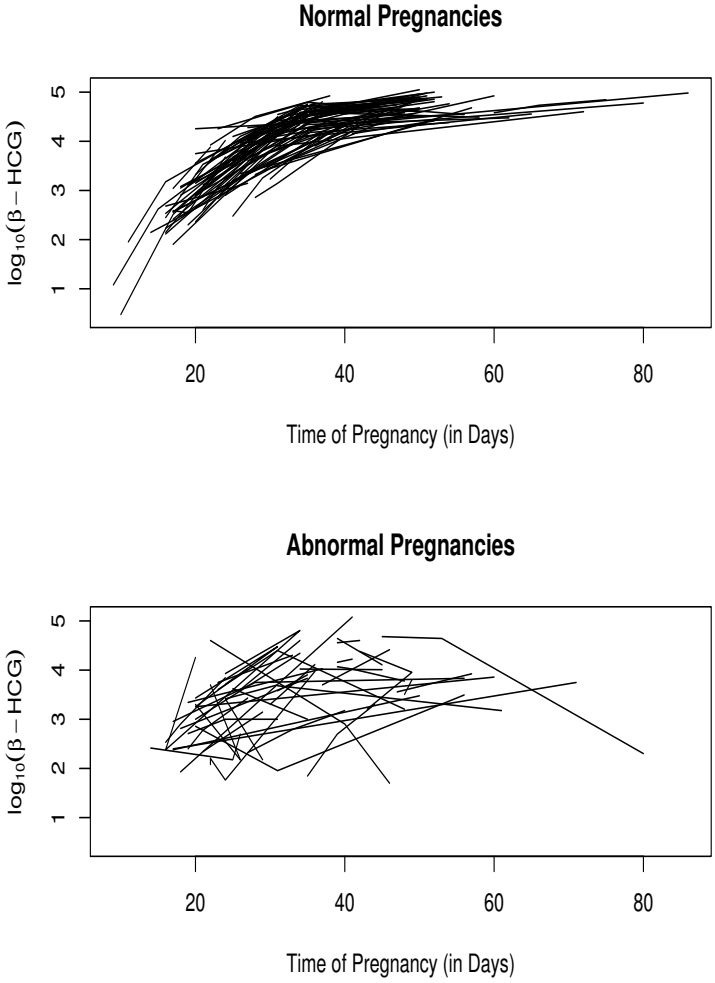

Fig. 1. Observed profiles of β-HCG for all 173 women.

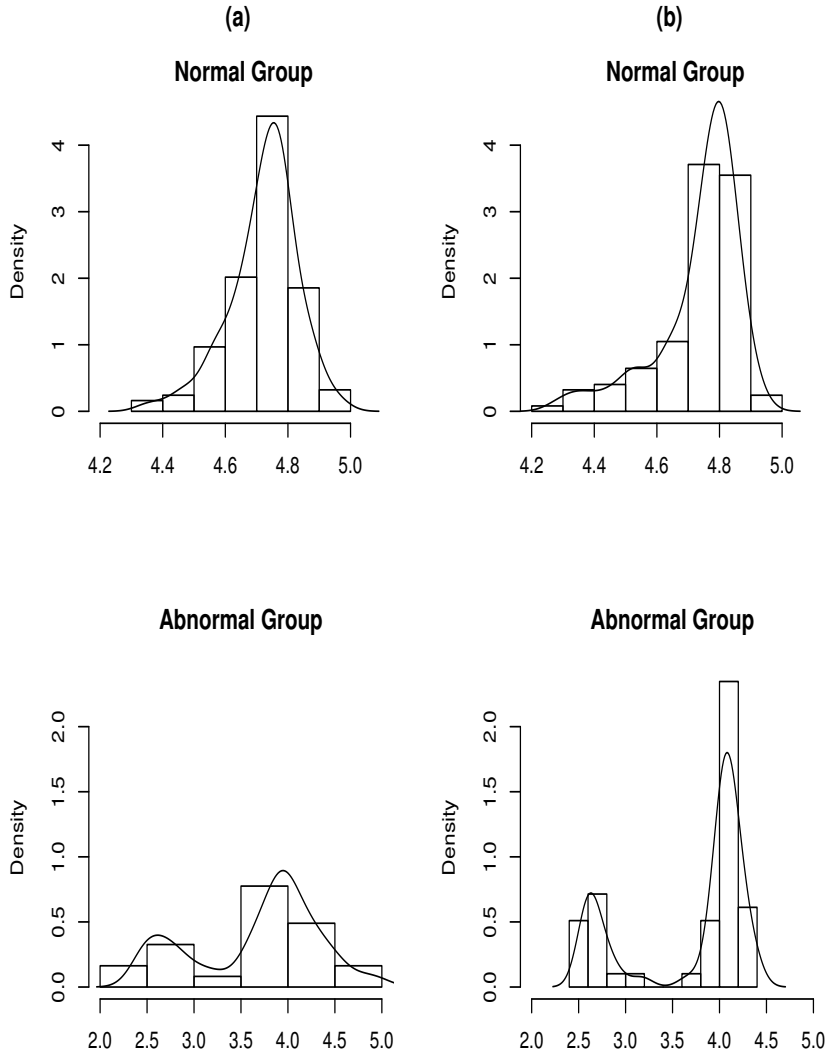

Fig. 2. Estimated Subject-specific (θ_i) parameters using (a) MLE and (b) posterior predictive draws for the BSP model with a smooth curve overlaid on each plot.

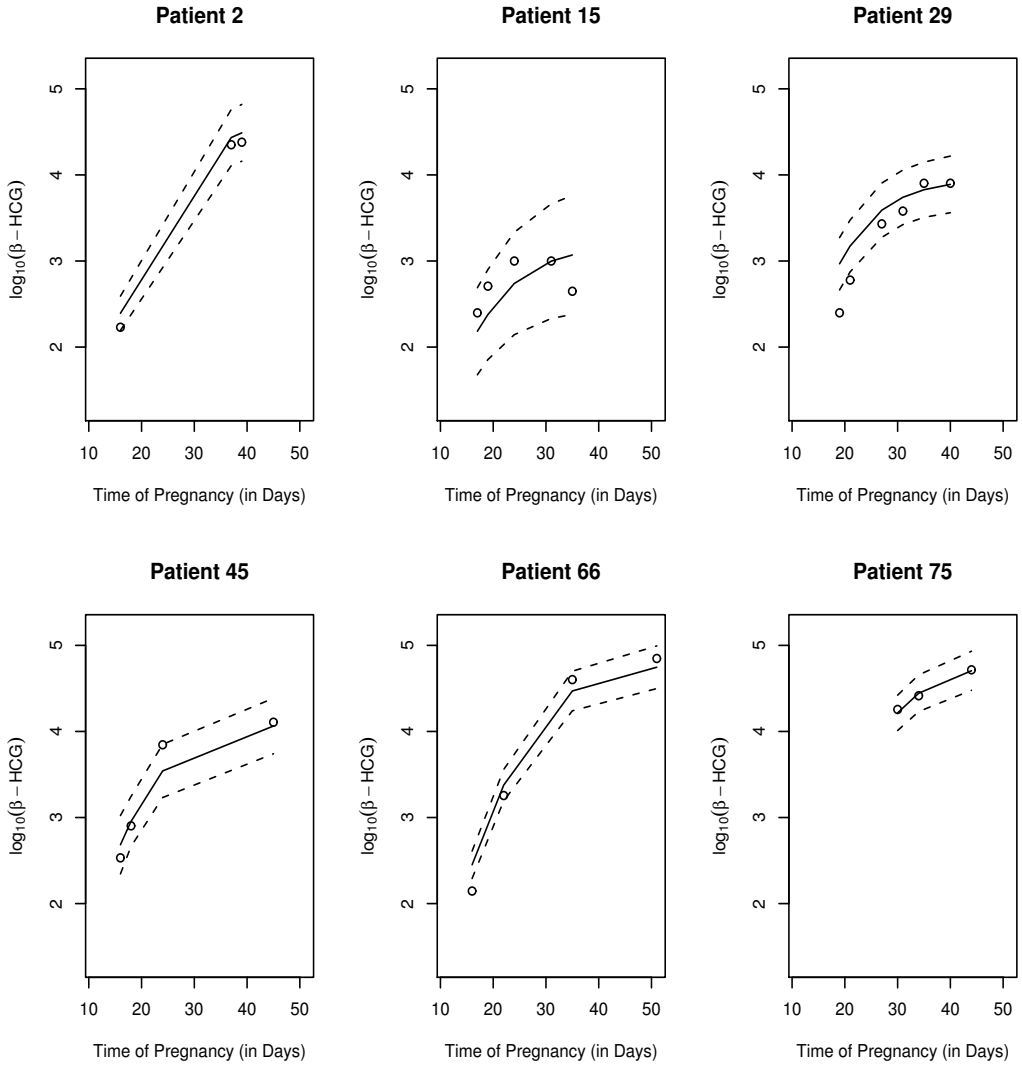

Fig. 3. Fitted Curves for three patients in the normal group (patients 2, 66, 75), and three in the abnormal group (patients 15, 29, 45). The points are the actual observations. The solid lines represent the fitted curves; the dashed lines represent fitted curve $\pm$ two posterior standard deviation.

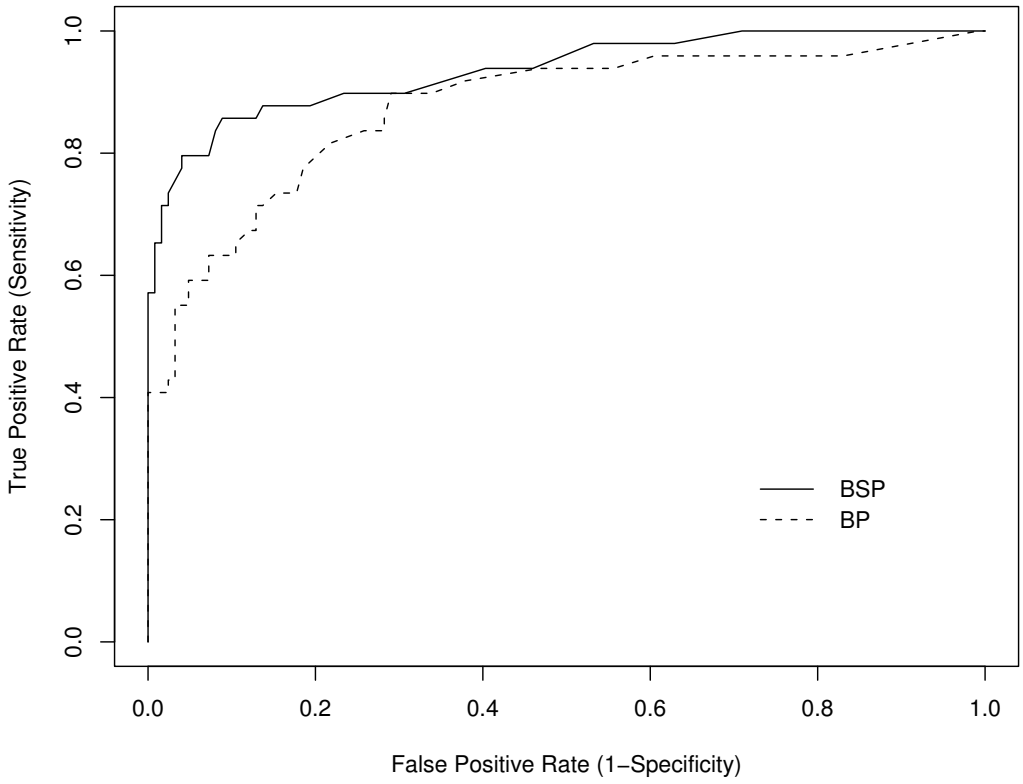

Fig. 4. ROC curves for classification under the Bayesian semiparametric model (solid line) and the Bayesian parametric model (dashed line).

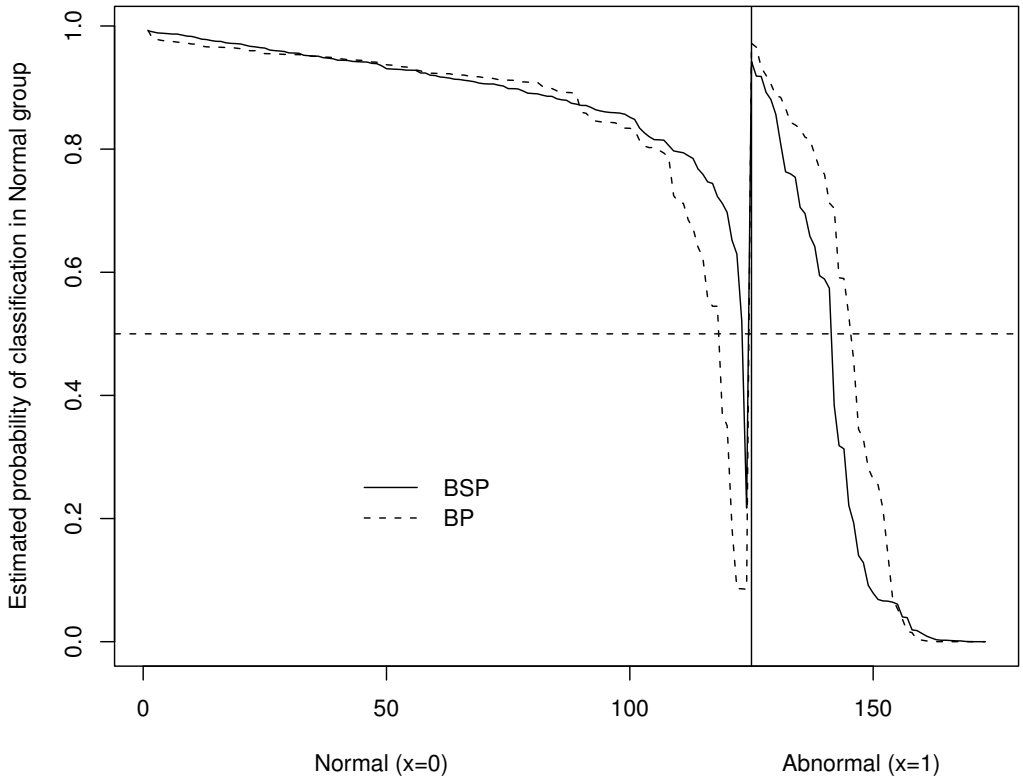

Fig. 5. Estimated probabilities of classification in the normal group for all individuals in decreasing order within normal and abnormal groups. The solid lines correspond to the Bayesian semiparametric model and the dashed lines to the Bayesian parametric model.

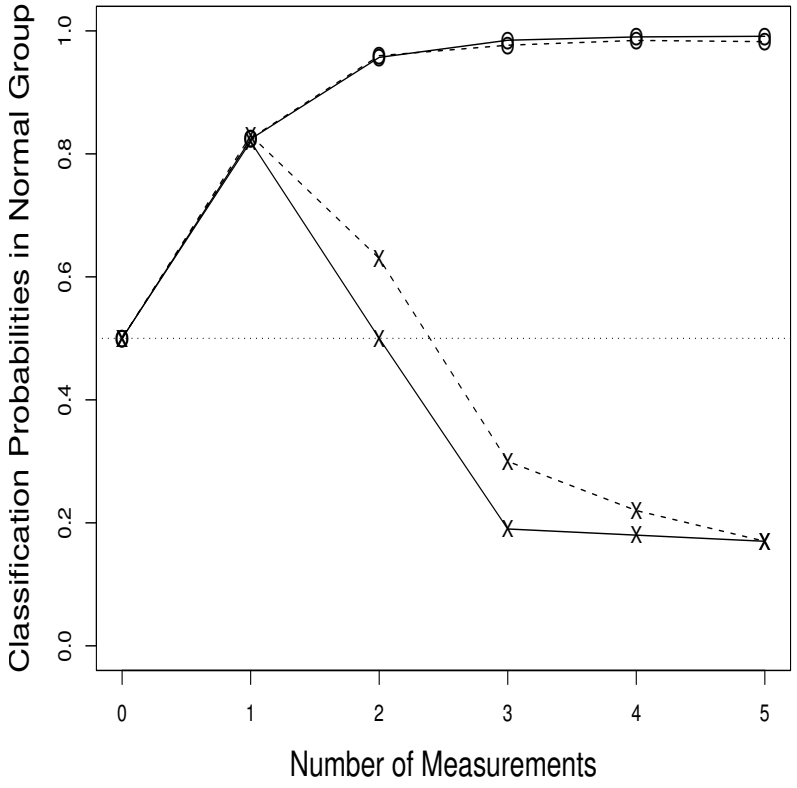

Fig. 6. Evolution of classification probabilities for one normal (upper curves) and one abnormal (lower curves) future patient as a function of the number of observations. The solid lines correspond to the Bayesian semiparametric model and the dashed lines to the Bayesian parametric model. The difference in inference is critical. For example, consider inference after the second observation. It the (unknown) truth is an abnormal pregnancy, the probability of a correct classification under the proposed model is 10% higher under a parametric model.

Table 3. The effect of hyperparameter choices on the classification probabilities. We report, for different combinations of M, $E(\sigma_x^2)$ and $E(\tau^2)$, the classification probability for a normal patient (i.e. the unknown truth is $x_{m+1} = 0$) given one (p_1) and two (p_2) observations. Here, 1^* denotes that for this case $M \sim \mathcal{G}(1, 1)$.

		M			
$E(\sigma_x^2)$	$E(\tau^2)$	1^*	5	10	
5	5	0.819	0.816	0.820	p_1
		0.954	0.950	0.950	p_2
	50	0.822	0.818	0.818	p_1
		0.955	0.964	0.955	p_2
	500	0.818	0.816	0.814	p_1
		0.955	0.961	0.956	p_2
50	5	0.822	0.821	0.817	p_1
		0.955	0.954	0.951	p_2
	50	0.824	0.818	0.814	p_1
		0.957	0.961	0.959	p_2
	500	0.819	0.819	0.814	p_1
		0.954	0.955	0.958	p_2
500	5	0.819	0.813	0.815	p_1
		0.954	0.942	0.947	p_2
	50	0.824	0.818	0.819	p_1
		0.959	0.958	0.957	p_2
	500	0.818	0.807	0.804	p_1
		0.955	0.949	0.947	p_2

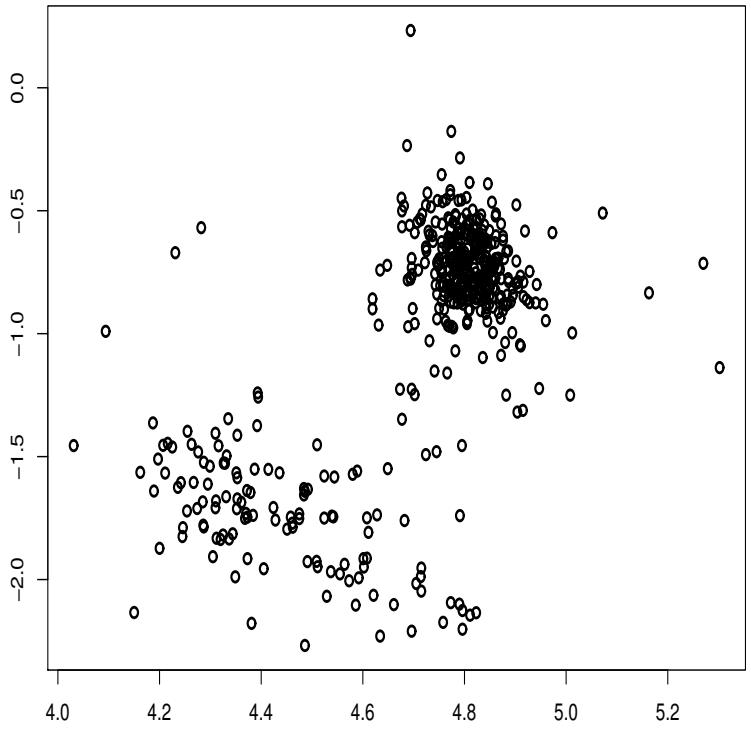

Fig. 7. Scatterplot of 500 posterior predictive draws of ANOVA coefficients α_{m+1} for a future patient.