A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles

Rolando De la Cruz-Mesía Fernando A. Quintana *

May, 2006

Abstract

This paper discusses Bayesian statistical methods for the classification of observations into two or more groups based on hierarchical models for nonlinear longitudinal profiles. Parameter estimation for a discriminant model that classifies individuals into distinct predefined groups or populations is done via appropriate posterior simulation schemes. The methods are illustrated with data coming from a study involving 173 pregnant women. The main objective in this study is to predict normal versus abnormal pregnancy outcomes from beta human chorionic gonadotropin (β-hCG) data available at early stages of pregnancy.

keywords: Discriminant Analysis; Longitudinal Data; Nonlinear hierarchical models.

*Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, CHILE. e-mail: rolando@mat.puc.cl

1 INTRODUCTION

In different biomedical situations markers are needed to detect the onset of a specific disease as early as possible, taking into account any longitudinal information that becomes available. One such example concerns pregnant women. To detect a number of complications arising during pregnancy, a variety of quantities or characteristics are measured at antenatal examinations (Yamashita *et al.* 1989; Witt *et al.* 1990; Hahlin *et al.* 1991). One of these is the Beta Human Chorionic Gonadotropin (β-hCG), also known as the "pregnant hormone" or "announcer of pregnancy", which keeps the corpus luteum ("yellow body") producing progesterone and estrogen during the early stages of pregnancy. β-hCG is a hormone that is produced by the placenta. It is detectable in the blood and urine within 10 days of fertilization. After the fertilized egg implants or attaches to the inside of the uterus or other structure inside the mother, the levels of β-hCG rise rapidly, frequently exceeding 100 milli-international units per milliliter (mIU/ml). The levels continue to increase throughout the first trimester of pregnancy and reach a peak 60 to 80 days after the fertilized egg implants. The exact level of β-hCG in the blood can be measured using standard tests. These measurements can help give a rough estimate of the age of the fetus. They can also help to determine if the pregnancy is progressing normally. Levels that are abnormally low or high may be signs that an abnormal medical condition is present. This would suggest the need for further evaluation and testing.

There is a great variation in β-hCG levels. However, it is not the absolute value that matters, but their relative changes. In a normal pregnancy, the level of this hormone approximately doubles every 1.5 days up to 5 weeks after the last menstrual period, and then every 3.5 days from the 7th week on (Frits and Guo, 1987). After the first trimester, levels should gradually decrease over time and in fact quickly decrease to zero after the pregnancy is ended. However, abnormally large levels of β-hCG may indicate choriocarcicoma of the uterus, Down syndrome in the fetus, hydatidiform mole of the fetus, or ovarian cancer. Lower than normal β-hCG levels may indicate

ectopic pregnancy, a miscarriage, or spontaneous abortion. In any case, a failure to exhibit normal growth patterns in β-hCG levels should be usually interpreted as a complicated pregnancy.

The main objective of this article is to explore a classification technique for predicting the outcome of pregnancy on the basis of the results of certain diagnostic tests administered to pregnant women. The inference problem is formally described as a discriminant analysis based on longitudinal β-hCG outcomes. Our approach is fully Bayesian and provides the posterior (or predictive) probability of outcome of pregnancy in women based on the longitudinal marker. Physicians can then make decisions on the basis of these probabilities. We consider a hierarchical structure that allows dealing with the nonlinear profiles and classification for individuals with very few or just one observation. This is specially relevant in the context of our motivating example, where more than one third of the women had two or less β-hCG measurements.

Extensions of classical discriminant analysis to multivariate response curves observed over fixed time intervals have been considered in, e.g. Albert (1983). Some extensions to the case of unbalanced data have also been proposed, typically using linear or nonlinear random effects models to describe the longitudinal profiles in each group. Verbeke and Lesaffre (1996) proposed the so-called heterogeneity model, i.e. a linear mixed effects model with random effects sampled from a mixture of normal distributions. Verbeke and Molenberghs (2000) indicated that the classification rule implied by the heterogeneity model is equivalent to the discriminant function proposed by Tomasko *et al.* (1999). Further developments along this direction have been discussed in Wernecke *et al.* (2004). Marshall and Barón (2000) considered nonlinear random-effects models to describe evolutions in different groups and state the optimal allocation rule. Brown *et al.* (2001) discussed discriminant analysis using linear random-effects models from a Bayesian viewpoint.

In our approach the classes or groups are predefined and the task is to understand

the basis for the classification from a set of labeled subjects (training dataset). This information is then used to classify future subjects. In the case that the classes or groups are unknown a priori and need to be estimated from the data, latent class models offer a fruitful approach. See additional developments along this direction in Muthén and Shedden (1999), Lin *et al.* (2000), McCulloch *et al.* (2002), Muthén *et al.* (2002), and references therein.

The rest of this article 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. In Section 4 we illustrate the proposed longitudinal method using data from Santiago, Chile on the β-hCG measured in women with normal and abnormal pregnancy outcomes. An appropriate posterior simulation scheme based on the Gibbs sampling algorithm is described. Finally, Section 5 discusses the results.

2 PREGNANT WOMEN DATA

There is an increased risk of early pregnancy loss after assisted reproduction. Various studies have addressed the question about early β-hCG values and their relationship to pregnancy outcome after in-vitro fertilization (e.g., Yamashita *et al.* 1989). We consider a dataset with a total of 173 young women, representing different pregnancies over a period of 2 years in a private fertilization obstetrics clinic in Santiago, Chile. β-hCG concentrations for the 173 women were measured during the first 80 days of gestational age. One of the main targets of the study was to evaluate these concentrations at early stages of pregnancy, with the purpose of identifying women with a high risk of loss. Consequently, pregnancy outcomes were divided into two groups: normal and abnormal. 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 resulting dataset thus

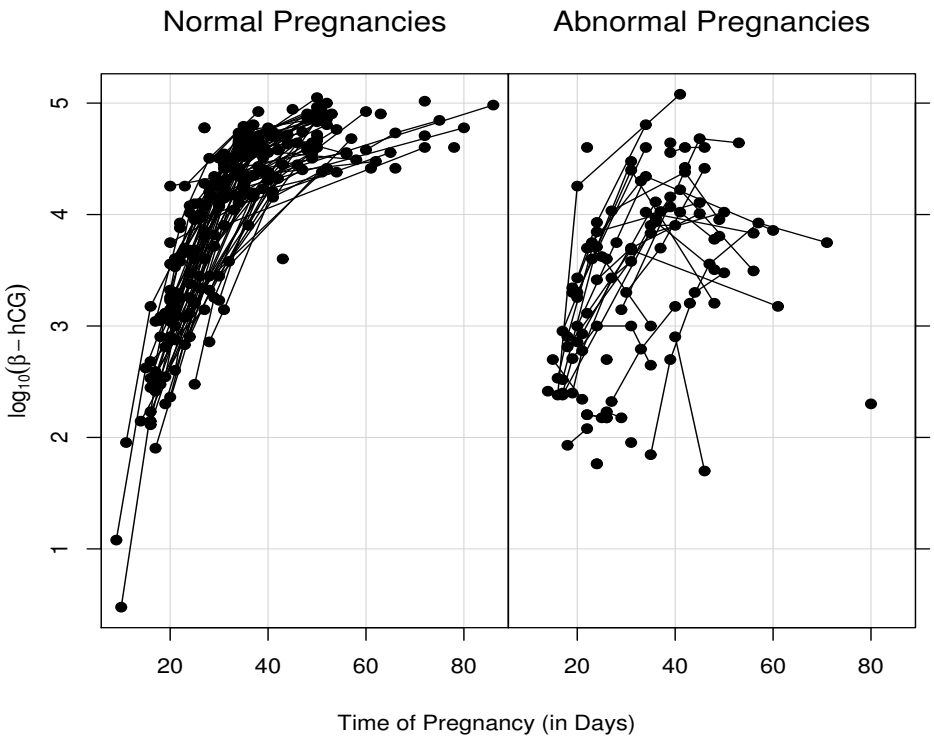

Figure 1: Observed profiles of β-hCG for all 173 women.

consists of 124 patients diagnosed with normal pregnancy and 49 patients with abnormal pregnancy. The 173 women altogether contribute a total of 375 observations, where the number of samples per woman ranged from 1–6 (median 2). These data were originally presented in Marshall and Barón (2000).

The outcome that we analyze are the vectors of time-varying β-hCG measurements for the 173 women. Approximately 30 per cent of the 173 women had one β-hCG measurement, 31 per cent had two, 33 per cent had three, and 6 per cent had four or more measurements. Figure 1 presents the subject-specific $\log_{10}$ β-hCG profiles for both groups.

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

sub-population is far less certain, in particular when considering series of partial responses. Thus, our main inference goal in analyzing these data is to provide a classification rule for a new patient. The rule should allow sequential updating as data accrues for the new patient.

3 CLASSIFICATION USING NONLINEAR HIERARCHICAL MODELS

Suppose we are given a training dataset comprising m units $\{(y_i, z_i), i = 1, \ldots, m\}$. Here $y_i = (y_{i1}, \ldots, y_{in_i})' \in \Re^{n_i}$ represents the observed response vector for the ith unit, taken at arbitrary times $t' = (t_1, t_2, \ldots, t_{n_i})$. Let $z_i \in \{1, 2, \ldots, g\}$ denote the known group label for the ith unit. In our application, $m = 173$, $g = 2$, with $z_i = 1$ and $z_i = 2$ indicating normal and abnormal pregnancy, respectively. The label z_i is known for some women with already reported delivery, but unknown for women with partial data before delivery. Without loss of generality we assume that z_i, $i = 1, \ldots, m$, is known, and z_{m+1} is unknown. Let $y^m = (y_1, \ldots, y_m, z_1, \ldots, z_m)$ denote all data, including the recorded class memberships z_i, up to the m-th patient.

Classification can be done either on a within-sample or a predictive basis. For the former, the posterior probabilities $\{p(z_i = k|y^m); k = 1, \ldots, g\}$ are appropriate, and these can be directly estimated as empirical averages in their MCMC run. The predictive classification problem assumes that a future unit y_{m+1} with unknown label z_{m+1} is recorded, so interest focuses on inference about z_{m+1}, i.e., we are in principle interested in $\{p(z_{m+1}|y^m, y_{m+1})\}$. Here y_{m+1} is the currently available partial response vector for the new patient $m + 1$.

We consider for group k a generic hierarchical model of the form

$$y_{ik}|\theta_{ik} \sim p(y_{ik}|\theta_{ik}), \qquad \theta_{ik} \sim G(\theta_{ik}|\phi_k). \tag{1}$$

In words, data y_{ik} for the ith sampling unit in group k are sampled from a prob-

ability model parameterized by a vector θ_{ik}. We will assume the parameter vector θ_{ik} to be partitioned into common fixed effects θ_k^F and unit-specific random effects θ_{ik}^R. The θ_{ik}^R are assumed to be generated from a distribution parametrized by a d-dimensional hyperparameter vector ϕ_k, thus implying a parametric model for random effects. Bayesian inference under parametric assumptions for random effects has been discussed by Bennet *et al.* (1995), Racine-Poon (1985), Wakefield *et al.* (1994), Wakefield (1996), Wakefield and Bennet (1996), and Ziegelmann and Brown (2001), among others.

For the top-level sampling model $p(y_{ik}|\theta_{ik})$ in (1) we assume a nonlinear regression

$$y_{ijk} = f(\theta_{ik}; t_{ijk}) + \epsilon_{ijk},$$

with a mean function $f(\theta_{ik}; \cdot)$ parameterized by θ_{ik} and evaluated at known times t_{ijk}, $j = 1, \ldots, n_i$, $k = 1, \ldots, g$. The residual term ϵ_{ijk} is assumed to be normally distributed with mean 0 and variance σ_k^2. To facilitate classification we augment the model with a marginal probability for z_i:

$$\Pr(z_i = k) = \pi_k,$$

with $k = 1, \ldots, g$. Specific choices for our motivating example will be discussed in the next Section. Note, however, that the augmented model implies the desired classification as a conditional probability $p(z_{m+1}|y^m, y_{m+1})$, marginalizing with respect to the unknown θ_i, and other possibly unknown hyperparameters.

From a Bayesian viewpoint the classification probabilities are obtained by weighting the posterior distributions of the parameters. Using Bayes' rule and some algebraic manipulations, the classification probability that a new unit y_{m+1} belongs to the kth group is

$$\begin{aligned} P(z_{m+1} = k|y_{m+1}, y^m) &= \int p(z_{m+1} = k|y_{m+1}, y^m, \Theta) p(\Theta|y_{m+1}, y^m)\, d\Theta \\ &= \int p(z_{m+1} = k|y_{m+1}, \Theta) p(\Theta|y_{m+1}, y^m)\, d\Theta \end{aligned}$$

$$
\begin{aligned}
&\propto \int p(z_{m+1} = k | y_{m+1}, \Theta) p(\Theta | y^m)\, d\Theta \\
&= \int \frac{\pi_k p(y_{m+1} | \Theta_k)}{\sum_\ell \pi_\ell p(y_{m+1} | \Theta_\ell)} p(\Theta | y^m)\, d\Theta. \qquad (2)
\end{aligned}
$$

However, the integration is usually analytically intractable. Therefore, we shall construct a set of stationary samples $\{\Theta^{(b)}, b = 1, \ldots, B\}$ from their posterior distribution and use

$$
\frac{1}{B} \sum_{b=1}^{B} \frac{\pi_k p(y_{m+1} | \Theta_k^{(b)})}{\sum_\ell \pi_\ell p(y_{m+1} | \Theta_\ell^{(b)})} \qquad (3)
$$

to approximate (2). If the prevalences π_k are unknown hyperparameters as well, then (3) would use the imputed values $\pi_k^{(b)}$.

Under the *zero-one* loss function (Hastie *et al.* 2001), the Bayes classification of an existing unit y_i and a future one y_{m+1} are respectively given by

$$
\hat{z}_i = \arg\max_k \{p(z_i = k | y^m)\} \quad \text{and} \quad \hat{z}_{m+1} = \arg\max_k \{p(z_{m+1} = k | y^m, y_{m+1})\}. \qquad (4)
$$

In other words, the unit is classified in that group for which the highest posterior probability is attained, thus minimizing the expected misclassification rate. Of course there are other loss functions that could be used, depending on whether it is more important to avoid false positives or false negatives cases (see Hastie *et al.* 2001).

4 APPLICATION

We apply the proposed model to the analysis of the longitudinal β-hCG data discussed in Section 2.

4.1 Model specification

Mean values of the $\log_{10}$ β-hCG for these 173 women show a nonlinear relationship with days of pregnancy. Figure 1 shows time profiles for normal and abnormal subjects. 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 $\log_{10}$ β-hCG level (θ_{ik} below). Concretely, they proposed the following nonlinear random-effects model:

$$y_{ijk} = \frac{\theta_{ik}}{1 + \exp\{-(t_{ijk} - \theta_{k2})/\theta_{k3}\}} + e_{ijk}, \tag{5}$$

where y_{ijk} represents the jth $\log_{10}$ β-hCG measurement on the ith woman in group k, where $k = 1, 2$ represent, respectively, normal and abnormal pregnancy groups, t_{ijk} is the day at which the y_{ijk} measurement was recorded, and the θ_{ik}, $i = 1, \ldots, 173$, are assumed $N(\theta_{k1}, \tau_k^2)$ and independent of the e_{ijk}. The parameters $\boldsymbol{\theta}_1' = (\theta_{11}, \theta_{12}, \theta_{13})$ and $\boldsymbol{\theta}_2' = (\theta_{21}, \theta_{22}, \theta_{23})$ represent the population parameters of the logistic curve for the normal and abnormal pregnancy groups, respectively.

After integrating out random effects in the kth group, the likelihood in the kth group is still described by means of a normal distribution:

$$y_{ik}|\boldsymbol{\theta}_k \sim \mathcal{N}_{n_i}\left(\boldsymbol{\mu}_k(t_{ik}), \tau_k^2 \boldsymbol{v}_{ik}\boldsymbol{v}_{ik}' + \sigma_k^2 \mathbf{I}_i\right),$$

where the mean vector $\boldsymbol{\mu}_k(t_{ik})$ has elements

$$\mu_k(t_{ijk}) = \frac{\theta_{k1}}{1 + \exp\{-(t_{ijk} - \theta_{k2})/\theta_{k3}\}},$$

and represents the population curve at time t_{ijk}. Furthermore, $\mathbf{I}$ is an $n_i \times n_i$ identity matrix, and the vector $\boldsymbol{v}_{ik}$ has elements

$$v_{ijk} = \frac{1}{1 + \exp\{-(t_{ijk} - \theta_{k2})/\theta_{k3}\}},$$

which depend on the values of the unknown population parameters θ_{k2} and θ_{k3}. We complete the Bayesian formulation of model (5) assuming prior independence for parameters, with distributions specified as

$$\boldsymbol{\theta}_k \sim \mathcal{N}_3(\boldsymbol{\theta}_0, \boldsymbol{D}_0), \tag{6}$$

$$\sigma_k^2 \sim \mathcal{IG}(a_1, c_1), \tag{7}$$

$$\tau_k^2 \sim \mathcal{IG}(a_2, c_2), \tag{8}$$

for $k = 1, 2$. Here, $\mathcal{IG}$ denote the inverse gamma distribution. In practice the specification of hyperparameters $\boldsymbol{\theta}_0$, $\boldsymbol{D}_0$, a_1, c_1, a_2, and c_2 may be difficult, so we choose in the next subsection hyperpameter values defining vague priors.

4.2 Posterior computation

Posterior distributions of the parameters were estimated using the Gibbs sampling algorithm. The value of the hyperparameters in (6), (7), and (8) were taken as $\boldsymbol{\alpha}_0 = (0,0,0)^T$, $\boldsymbol{D}_0 = 1\,000\,\mathbf{I}_3$, $a_1 = a_2 = 3$, and $c_1 = c_2 = 0.01$, which gives the prior variance of σ^2 and τ^2 to be 2 500. The resulting prior densities are proper, but vague and hence relatively uninformative. Prior probabilities of group membership where 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.

The full conditionals for implementing the Gibbs sampler for α_{k1}, $\boldsymbol{b}_{ik}$, σ_k^2, and τ_k^2, are straightforwardly derived as normal, normal, inverse gamma, and inverse gamma distribution, respectively. The full conditionals for α_{k2} and α_{k3} are not available in closed-form. A separate random walk Metropolis algorithm was used to simulate each $\alpha_{k\ell}$ ($\ell = 2, 3$). For this, we use a normal proposal distribution centered at the current value of $\alpha_{k\ell}$ with variance given by a constant c (we chose $c = 2$) times the inverse information matrix.

To perform the Gibbs sampling, we chose starting points in a neighborhood of the MLEs of model parameters. For this we use the library `NLME` of Pinheiro and Bates (2000) to fit the models. The algorithm was implemented using computer code written in C. We generated 250 000 iterations. After 10 000 iterations, samples were collected, at a spacing of 240 iterations, to obtain independent samples. Finally we totaled $B = 1\,000$ samples for calculating posterior quantities of interest.

We checked convergence for each monitored variable, using the `CODA` software of Best *et al.* (1995) to produce convergence diagnostics. Application of Geweke's (1995) convergence criterion separately to each of the model parameters showed no evidence against convergence, as the absolute value of the Z-statistic was less than 1.7 in all cases. Also, all sequences passed the stationarity test of Heidelberger and Welch (1983).

4.3 Results

The mean gestational ages (days) in women with normal and abnormal pregnancies, were 34.5, and 32.9 days, respectively; thus, no significant difference in gestational age was found among the groups. We found that concentrations of β-hCG were significantly lower in women with an abnormal pregnancy than in those with a normal pregnancy.

Table 1: Summary of model fitting

	Normal Pregnancies ($k = 1$)		Abnormal Pregnancies ($k = 2$)	
Parameter	mean	sd	mean	sd
θ_{k1}	4.739	0.047	3.788	0.156
θ_{k2}	15.610	0.337	11.740	2.038
θ_{k3}	7.349	0.427	7.738	2.231
σ_k^2	0.094	0.012	0.177	0.036
τ_k^2	0.030	0.015	0.587	0.171

Table 1 presents posterior summaries of the parameters. Observed differences in the group-specific estimates of the variance components models suggest that there is more between-subject variability in the abnormal group than in the normal group.

As part of the analysis we estimated individual $\log_{10}$ β-hCG profiles and standard errors. Fitted profiles with ± 2 SD curves are displayed for six selected patients in Figure 2. Three of them in the normal group (patients 3, 66, 24) and the remaining three in the abnormal group (patients 20, 27, 45). We note that the inference captures the varying observation error between subjects. An important feature of the Bayesian approach is that the credible intervals take into account the variability of all parameters, including unknown hyperparameters. We note here that all these patients were correctly classified according to our proposed approach.

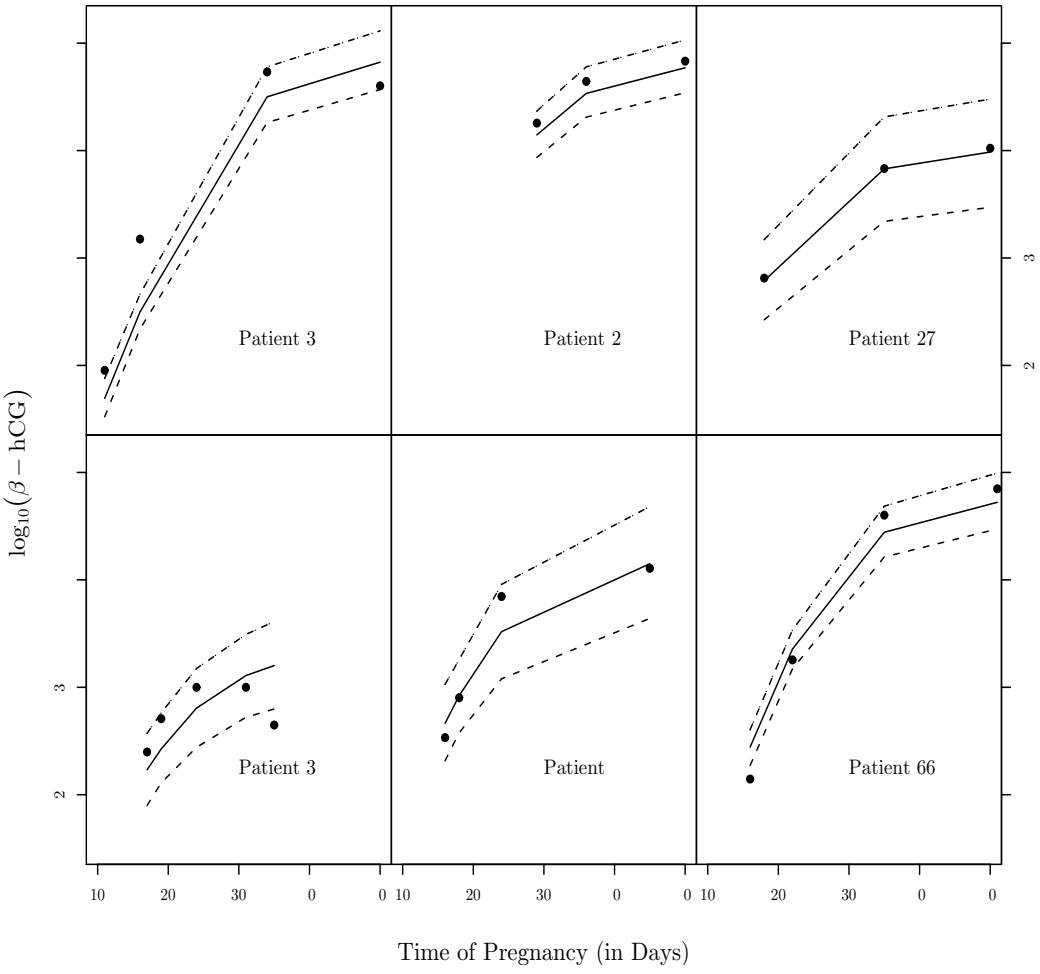

Figure 2: Fitted curves for three patients in the normal group (patients 3, 66, 24), and three in the abnormal group (patients 20, 27, 45). The points are the actual observations. The solid lines represent the fitted curves and the dashed lines represent fitted curve $\pm$ two posterior standard deviation.

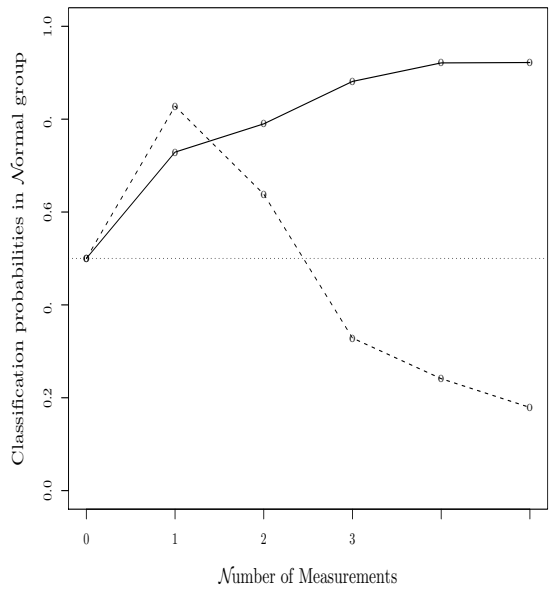

Figure 3: Evolution of classification probabilities for one normal (solid line) and one abnormal (dashed line) future patient as a function of the number of observations.

Once the posterior probabilities $p(z_i = k|y^m)$ are estimated, several quantities of interest can be readily evaluated. Among these, "sensitivity" and "specificity" are popular ways to summarize the classification results. Letting $A = \{$ the patient is classified as abnormal $\}$, and $E = \{$ the patient actually belongs to the abnormal group $\}$, then sensitivity and specificity are defined respectively as $\Pr[A|E]$ and $\Pr\left[\bar{A}|\bar{E}\right]$. These are simply estimated as $\sum_{i=1}^{m} I\{\hat{z}_i = z_i = 2\}/\sum_{i=1}^{m} I\{z_i = 2\}$ and $\sum_{i=1}^{m} I\{\hat{z}_i = z_i = 1\}/\sum_{i=1}^{m} I\{z_i = 1\}$, respectively, where $\hat{z}_i$ was defined in (4). From our results we found 57% sensitivity and 95% specificity.

We also computed the misclassification error rate, which was found to be 16% for this dataset. However, it is well known that the error rate obtained by applying the classifier to the same data from which it has been formed tends to be biased downward as an estimate of the true error rate (Hastie *et al.* 2001). Several methods are available to solve this problem. For moderately large datasets, we could consider

a series of randomly chose divisions of the data into two components, one reserved for deriving the classification rule (the training sample) and the other to assess this rule (the test sample). Under this method, the estimated error rate is the average error rate over all the generated divisions. For smaller datasets like the one at hand, a cross-validation (CV) technique can be used to compensate for the lack of data. Concretely, we consider the leave-one-out CV (looCV) method, which consists of splitting the m samples into a training sample of size $m-1$ and a test sample of size 1. The estimated error rate is then the average number of times the test sample was misclassified over the m possible divisions. A critical advantage of this method is that all the data can be used for training, without the need of setting subjects aside purely for testing. The looCV error rate is an important statistical estimator of the performance of a classification rule when the sample size is small. It is frequently used by many researchers. A direct generalization of the above procedure is the k-fold CV, where the data are divided randomly into k components of approximately equal sizes. Next, $k-1$ components are used to compute the classification rule, which is then tested on the omitted component. This process is repeated k times, until all components are used to assess the classification rule, and the error rates averaged.

We performed both, looCV and 5-fold CV, obtaining misclassification error rates of 17% and 17.3%, respectively. Both error rates are very similar, which suggests that these methods provide good estimates that are reasonably consistent. See further properties and discussion of these methods in Hastie *et al.* (2001).

In the above analysis we have used all the available information. However, it is interesting to assess the predictive power of our model. A possible way of doing this is to study how the classification probabilities change with the number of available observations, for a given patient. Thus we generate from the corresponding posterior predictive distributions, one future patient for each group and evaluate (2) for up to five possible observations. Time points were chosen from the empirical distribution of observed times within each group. Figure 3 shows the evolution of these probabilities

for each future patient. For the normal patient, we observe a steady growth of the probabilities. For the abnormal patient, however, 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 but a decrease in some of the patients. Thus the classification probabilities for abnormal patients require a few more observations than the abnormal ones to reflect the correct outcome.

5 DISCUSSION

This paper proposes a general Bayesian framework for the classification of longitudinal profiles where the underlying models in each group or population are given by nonlinear hierarchical models. The approach has the advantage of being appropriate for classifying longitudinal profiles of datasets with an unbalanced data structure. It also has the ability to use all the 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.

This 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. In this context, the approach we have presented solves one important problem.

In our study the group "abnormal pregnancies" implies women with any complication resulting in a non-terminal delivery and loss of the fetus. This involves a wide range of pathologies including spontaneous abortion (hard to prevent in advance) and ectopic pregnancy (an acute condition that requires surgery). Measurements of

serum β-hCG at regular examination times may help identifying women with a high probability of abnormal pregnancy and who may benefit from closer surveillance. For instance, in the case ectopic pregnancy this may help reducing the risk of tubal rupture. Our results may help both physicians and patients make informed treatment decisions on the basis of objective staging information. Indeed, a reliable and inexpensive diagnostic test to differentiate between normal pregnancies and pregnancies with early adverse outcome might reduce the psychological tension and anxiety present in many patients. It may also help reducing treatment cost by essentially making it more efficient. On the other hand, for patients judged by this test to be at high risk of unfavorable outcome, a more careful follow up might help reducing the associated risks.

Other markers, such as serum progesterone measurements, have been used to distinguish normal from abnormal (spontaneous miscarriage) pregnancies. Thus a straightforward generalization of our approach would 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.

Acknowledgments

We are grateful to Guillermo Marshall for facilitating us the 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.

References

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

Bennett, J.E. Racine-Poon, A. and Wakefield, J.C. (1995). MCMC for non-linear hierarchical models. In *Markov Chain Monte Carlo in Practice,* Gilks, W.R. Richardson, S. and Spiegelhalter, D.J. (eds). Oxford: Oxford University Press.

Best, N.G. Cowles, M.K. and Vines, S.K. (1995). *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs sampling output, Version 0.3*, MRC Biostatistics Unit, Cambridge.

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

Frits, M.A. and Guo, S.M. (1987). Doubling time of human chorionic gonadotropin (hCG) in early normal pregnancy: relationship to hCG concentration and gestational age. *Fertil Steril* **47** 584–589.

Geweke, J. (1995). 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: Oxford University Press.

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

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research* **31** 1109–1144.

Hahlin, M. Sjoblom, P. and Lindblom, B. (1991). Combined use of progesterone and human chorionic gonadotrophin determination for differential diagnosis of very early pregnancy. *Fertil Steril* **55** 492–496.

Lin, H.Q. McCulloch, C.E. Turnbull, B.W. Slate, E.H. and Clark, L.C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine* **19** 1303–1318.

Marshall, G. and Barón, A.E. (2000). Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine* **19** 1969–1981.

McCulloch, C.E. Lin, H. Slate, E.H. and Turnbull, B.W. (2002). Discovering subpopulation structure with latent class mixed models. *Statistics in Medicine* **21** 417–429.

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

Muthén, B. Brown, C.H. Masyn, K. Jo, B. Khoo, S.T. Yang, C.C. Wang, C.P. Kellam, S.G. Carlin, J.B. Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3** 459–475.

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

Racine-Poon, A. (1985). A Bayesian approach to nonlinear random effects models. *Biometrics* **41** 1015–1024.

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

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New-York.

Walker, S. (1996). An EM algorithm for nonlinear random effects models. *Biometrics* **52** 934–944.

Wakefield, J. (1996). The Bayesian analysis of population pharmacokinetic models. *Journal of American Statistical Association* **92** 62–75.

Wakefield, J. and Bennet, J. (1996). Bayesian modeling of covariates for population pharmacokinetic models. *Journal of American Statistical Association* **91** 917–927.

Wakefield, J.C. Smith, A.F.M. Racine-Poon, A. and Gelfand, A.E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics* **43** 201–221.

Wernecke, K.-D. Kalb, G. Schink, B. and Wegner, B. (2004). A mixed model approach to discriminant analysis with longitudinal data. *Biometrical Journal* **46**(2), 246–254.

Witt, B.R. Wolf, G.C. Wainwright, C.J. Johnston, P.D. and Thorneycroft, I.H. (1990). Relaxin, Ca-125, progesterone, estradiol, Schwangerschaft protein, and human chorionic gonadotrophin as predictors of outcome in threatened and not-threatened pregnancy. *Fertil Steril* **53** 1029–1036.

Yamashita, T. Okamoto, S. Thomas, A. MacLachlan, V. and Healy, D.L. (1989). Predicting pregnancy outcome after in-vitro fertilization and embryo transfer using estradiol, progesterone and human chorionic gonadotrophin β-subunit. *Fertil Steril* **51** 304–309.

Ziegelmann, P. and Brown, P.J. (2001). Bayesian approach in pharmacokinetics models. In *Bayesian Methods with Applications to Science, Policy, and Official Statistics*. George, E.I. *et al.* (eds). Selected Papers from om ISBA 2000, 583–592.