Bayesian nonparametric ROC regression modeling with application to diabetes diagnosis

Vanda Inácio[1]*,
Alejandro Jara[2], Timothy E. Hanson[3], and Miguel de Carvalho[4]
[1]Department of Statistics and Operations Research, Universidade de Lisboa
[2]Department of Statistics, Pontificia Universidade Católica de Chile
[3]Department of Statistics, University of South Carolina
[4]Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne

Abstract

The receiver operating characteristic (ROC) curve is the most widely used measure for evaluating the discriminatory performance of a continuous marker. It is well known that, in a large number of situations, additional information is available in the form of covariates and several ROC regression methods have been proposed to incorporate covariates in the ROC framework. Most of these methods assume that covariate effects have parametric forms, but this can lead to misleading conclusions if the model is misspecified. To overcome this drawback, we propose a Bayesian nonparametric estimator of the conditional ROC curve based on dependent Dirichlet processes. Our simulation study reveals a good performance of the proposed estimator. Methods are applied to real data concerning diagnosis of diabetes.

Key words: Area under the curve; Bayesian nonparametrics; Diabetes; Glucose; Linear dependent Dirichlet process; ROC curve

1 Introduction

The development and statistical evaluation of diagnostic and screening procedures, such as biomarkers and imaging technologies, are of great importance in public health and medical research. The receiver operating characteristic (ROC) curve is a popular method for evaluating the performance of continuous markers and its presence has become essential in medical studies. Based on the concept of using a threshold to classify subjects as healthy or diseased, the ROC curve is a plot of the true positive rate (TPR, the probability that a diseased subject has a positive test) versus the false positive rate (FPR, the probability that a healthy subject has a positive test), across all possible threshold values, say k. That is, the ROC curve represents the plot $\{(\text{FPR}(k), \text{TPR}(k)) = (1 - F_0(k), 1 - F_1(k)), -\infty < k < \infty\}$, where marker's results of healthy (diseased) individuals in the population are distributed according to distribution function F_0 (F_1). For $0 < u < 1$, the ROC curve is given by $\text{ROC}(u) = 1 - F_1(F_0^{-1}(1 - u))$. Related to the ROC

*Corresponding author e-mail: vanda.kinets@gmail.com

curve, several measures, such as the area under the curve (AUC) or the Youden index, are considered as summaries of the discriminatory accuracy of the marker. The AUC is most common and it is given by $\int_0^1 \mathrm{ROC}(u)\mathrm{d}u$, which can be interpreted as the probability that the marker value of a randomly selected diseased individual exceeds that for a randomly selected nondiseased individual. The AUC takes values between 0.5 (useless marker that correctly classifies disease state no better than chance) and 1 (marker with perfect discriminatory ability). The partial area under the curve can be used to focus attention on clinically important values of the test's specificity.

In a large number of situations, additional information is available in the form of covariates (e.g., age or gender of the patients) which are known to influence the accuracy of the marker. In such situations, the ROC curve, or the AUC, may be misleading if covariates are ignored. One example is the use of fingerstick post-prandial blood glucose as a marker for diagnosis diabetes, where age is known to be an important covariate for this marker. Diabetes mellitus, often simply referred to as diabetes, is a group of metabolic diseases in which a person has high blood sugar concentration, either because the body does not produce enough insulin (a hormone produced by the pancreas), or because cells do not respond to the insulin that is produced. It is believed that the aging process may be associated with relative insulin deficiency or resistance among persons who are healthy (Smith and Thompson, 1996). Diabetes double the risk of cardiovascular disease (Sarwar et al., 2010). In 2000, according to the World Health Organization, at least 171 million people worldwide suffered from diabetes, which corresponds to 2.8% of the World population. Its incidence is increasing rapidly, and it is estimated by 2030, this number will almost double (Wild et al., 2004). In our study we are interested in examining how age influences the discriminatory power of the glucose to accurately detect diabetes.

To assess possible covariate effects on the ROC curve, various methods have been proposed. Induced methodology is based on using separate regression models for the healthy and diseased populations and then computing the induced form of the ROC curve (Pepe, 1998; Faraggi, 2003; González-Manteiga et al., 2011; Rodríguez-Álvarez et al., 2011). Alternatively, direct methodology assumes a regression model for the ROC curve itself, with the effects of the covariates being directly evaluated on the ROC curve (Alonzo and Pepe, 2002; Pepe, 2003; Cai, 2004).

A crucial point when applying such methodologies is how to model, parametrically, the continuous effect of covariates on the ROC curve. Misleading results may be obtained if the effects are incorrectly specified. Although the literature is replete with nonparametric approaches to the estimation of ROC curves without covariates (Hsieh and Turnbull, 1996; Zou et al., 1997; Lloyd, 1998; Zhou and Harezlak, 2002; Peng and Zhou, 2004), in the case of conditional ROC curves few contributions have been made in the nonparametric framework.

In this work, we model the conditional ROC curve, within the induced context, using Bayesian nonparametric (BNP) techniques; specifically, we use prior distributions for collections of related probability distributions: the dependent Dirichlet process. This dependent process represents a generalization of Bayesian methods which place a prior distribution over the space of distribution functions. Bayesian nonparametric techniques allow for broadening the class of models under consideration and hence for the development of a widely applicable approach that can be used for practically any population and for a large number of diseases. Recent applications of BNP models in ROC analysis can be found in Erkanli et al. (2006), Branscum et al. (2008), Hanson et al. (2008a), Hanson et al. (2008b), and Inácio et al. (2011).

The current approaches to ROC regression, within the induced context, are based on homocedastic linear models with normal distributed errors (Faraggi, 2003), unspecified error distributions (Pepe, 1998), and heterocedastic nonparametric models based on kernel-type regression methods

(González-Manteiga et al., 2011; Rodríguez-Álvarez et al., 2011).

Unlike the kernel-based approaches which accuracy relies upon asymptotics, which needs large samples, the accuracy of our BNP approach is assessed *a posteriori* by combining a high-dimensional likelihood function with the prior distribution of the model and tuning parameters irrespective of the sample size. Prior distributions can take into account subjective beliefs about the accuracy of a marker based on expert opinion or historical information. In contrast, kernel methods cannot use any form of prior information. We also point out that the current kernel approaches to ROC regression suffers from the limitation of solely being able to address a single continuous covariate. In turn, our BNP estimator of the conditional ROC curve can either handle multiple and categorical covariates.

This paper is organized as follows. In Section 2 we provide background material on the Dirichlet process and on the dependent Dirichlet process. The model framework to estimate the conditional ROC curve as well as some computational issues are presented in Section 3. In Section 4, a simulation study is carried out to assess the performance of the proposed estimator. In Section 5, we apply our method to assess the age impact on the glucose performance to accurately diagnose diabetes. Concluding remarks are given in Section 6.

2 Background

In this section we provide background on Dirichlet processes and on dependent Dirichlet processes.

2.1 Dirichlet processes

The Dirichlet process (DP) was introduced by Ferguson (1973) as a (prior) probability model for distributions G. The DP is characterized by two parameters: a base distribution G^* (the centre of the process) and a positive scalar parameter α, which can be interpreted as a precision parameter; larger values of α result in realizations G that are closer to G^*. We write $G \sim \mathrm{DP}(\alpha, G^*)$ to indicate that a DP prior is used for the random distribution G. An important constructive definition of the DP was given by Sethuraman (1994). Letting $\delta_{\cdot}$ denote the Dirac measure, we have

$$G = \sum_{l=1}^{\infty} \omega_l \delta_{\theta_l}. \tag{1}$$

The weights ω_l and locations θ_l are generated by the following stick-breaking scheme: $\omega_1 = z_1$, $\omega_l = z_l \prod_{r=1}^{l-1}(1 - z_r)$, $l = 2, 3, \ldots$, with $z_l \overset{i.i.d.}{\sim} \mathrm{Beta}(1, \alpha)$ and $\theta_l \overset{i.i.d.}{\sim} G^*$, independently of the ω_l. According to this definition, the DP generates (almost surely) discrete distributions. 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. The typical DPM model assumes

$$y_i | G \overset{ind.}{\sim} \int f(y_i|\theta) \mathrm{d}G(\theta), \quad G \sim \mathrm{DP}(\alpha, G^*)$$

that is, a mixture with a DP prior on the random measure G. The model is, typically, completed with hyperpriors for the parameters of G^* and α. One of the main attractions of DPM models is their computational simplicity; posterior simulation for these models is well understood (MacEachern and Müller, 1998; Neal, 2000). DPMs generalize finite mixture models, offering practical advantages in modeling and inference for data that arise from non-standard distributions.

2.2 Dependent Dirichlet processes

The problem of defining priors over related random probability distributions has received increased attention over the past few years. MacEachern (1999) proposed the dependent DP (DDP) as an approach to define a dependent prior model for a set of random measures $\{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p\}$ indexed by continuous covariates $\mathbf{x}$, with $G_{\mathbf{x}} \sim \text{DP}(\alpha, G^*)$ marginally. Recall Sethuraman's stick-breaking representation (1) for the DP random measure, $G_{\mathbf{x}} = \sum_{l=1}^{\infty} \omega_{\mathbf{x}l}\delta_{\theta_{\mathbf{x}l}}$. The key idea behind the DDP is to introduce dependence across the measures $G_{\mathbf{x}}$ by assuming the distributions of the point masses $\theta_{\mathbf{x}l}$ to be dependent across different levels of $\mathbf{x}$, but independent across l. In the basic version of the DDP the weights are assumed to be the same across $\mathbf{x}$, that is, $\omega_{\mathbf{x}l} = \omega_l$. For computational reasons, in what follows, we fix the weights $\omega_{\mathbf{x}l}$ across covariates and introduce the dependence through the point mass locations $\theta_{\mathbf{x}l}$. This approach has been successfully applied to ANOVA (De Iorio et al., 2004), survival analysis (De Iorio et al., 2009), spatial modeling (Gelfand et al., 2005), functional data (Dunson and Herring, 2006), time series (Caron et al., 2008), and discriminant analysis (De La Cruz et al., 2007).

We build our proposal on the construction introduced by De Iorio et al. (2004) and De Iorio et al. (2009). They proposed a particular version of the DDP where the component of the atoms defining the location in a DDP mixture model follows a linear regression model, i.e., $\theta_{\mathbf{x}l} = (\mathbf{x}'\boldsymbol{\beta}_l, \sigma_l^2)$, where $\mathbf{x}$ is the design vector and the $\boldsymbol{\beta}_l$ are the vectors of regression coefficients, which are i.i.d. from the distribution G^*, $\boldsymbol{\beta}_l \overset{i.i.d.}{\sim} G^*$. An advantage of this model for related random probability measures, referred to as the linear DDP (LDDP), is that it can be represented as DPM of linear (in the coefficients) regression models, when the model is convolved with a normal kernel. Hereafter, we use the normal distribution as the kernel. More specifically, letting $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ be the regression data, the model reduces to

$$y_i | G \overset{ind.}{\sim} \int \phi(y_i; \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \mathrm{d}G(\boldsymbol{\beta}, \sigma^2),$$

where $\phi(\cdot; \mu, \sigma^2)$ stands for the probability density function of the normal distribution with mean μ and variance σ^2. Note that the linear specification is highly flexible and can include standard nonlinear transformations of the predictors, for example, additive models based on B-splines (Lang and Brezger, 2004), linear forms in the continuous predictors themselves and categorical predictors. This model includes a wide variety of regression and mixture models as special cases, including normal linear regression, linear regression with the residual density modeled as finite or infinite mixtures of Gaussian distributions and finite mixtures of Gaussian linear regressions.

3 Models and methods

Let $\{(y_{0i}, \mathbf{x}_{0i})\}_{i=1}^{n_0}$, $\{(y_{1j}, \mathbf{x}_{1j})\}_{j=1}^{n_1}$, $F_0(y|\mathbf{x})$ and $F_1(y|\mathbf{x})$ be the data and the cumulative distribution functions from the healthy and diseased populations, respectively, where y stands for the marker result and $\mathbf{x}$ for the covariate value. We propose a BNP dependent prior probability model for the conditional ROC curve within the induced context. In particular, we consider a modeling approach based on LDDP priors.

3.1 Induced BNP ROC regression model

Instead of specifying a location-scale regression model for the marker values in each population (Pepe, 1998; González-Manteiga et al., 2011; Rodríguez-Álvarez et al., 2011), we model directly the conditional densities and distributions functions of each group. Specifically, to model the marker values as a function of covariates, we consider a mixture model

$$y_{0i}|G_0 \overset{ind.}{\sim} \int \phi(y_{0i}; \mathbf{x}'_{0i}\boldsymbol{\beta}_0, \sigma_0^2)\mathrm{d}G_0(\boldsymbol{\beta}_0, \sigma_0^2),$$

$$G_0|\alpha_0, G_0^* \sim \mathrm{DP}(\alpha_0, G_0^*),$$

and

$$y_{1j}|G_1 \overset{ind.}{\sim} \int \phi(y_{1j}; \mathbf{x}'_{1j}\boldsymbol{\beta}_1, \sigma_1^2)\mathrm{d}G_1(\boldsymbol{\beta}_1, \sigma_1^2),$$

$$G_1|\alpha_1, G_1^* \sim \mathrm{DP}(\alpha_1, G_1^*).$$

We assume a conjugate normal–gamma distribution for the baseline measure, i.e., we take

$$G_0^* = N_p(\boldsymbol{\beta}_0|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\Gamma(\sigma_0^{-2}|\tau_{01}/2, \tau_{02}/2)$$

as well as

$$G_1^* = N_p(\boldsymbol{\beta}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\Gamma(\sigma_1^{-2}|\tau_{11}/2, \tau_{12}/2).$$

The corresponding conditional cumulative distribution functions are given by

$$F_0(y|\mathbf{x}) = \sum_{l=1}^{\infty} \omega_{0l}\Phi(y|\mathbf{x}'\boldsymbol{\beta}_{0l}, \sigma_{0l}^2)$$

and

$$F_1(y|\mathbf{x}) = \sum_{l=1}^{\infty} \omega_{1l}\Phi(y|\mathbf{x}'\boldsymbol{\beta}_{1l}, \sigma_{1l}^2),$$

where $\Phi(\cdot; \mu, \sigma^2)$ is the normal distribution function with mean μ and variance σ^2. For a given value $\mathbf{x}$ of the covariate, the conditional ROC curve is defined by, for $0 < u < 1$,

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - F_1(F_0^{-1}(1-u|\mathbf{x})|\mathbf{x}),$$

and where F^{-1} stands for the conditional quantile function. The corresponding AUC is

$$\mathrm{AUC}_{\mathbf{x}} = \int_0^1 \mathrm{ROC}_{\mathbf{x}}(u)\mathrm{d}u.$$

3.2 Prior information and posterior simulation

To complete the model specification we assume the following independent hyperpriors:

$$\begin{aligned} \alpha_0|a_0, b_0 &\sim \Gamma(a_0, b_0), \quad \tau_{02}|\tau_{s_{01}}, \tau_{s_{02}} \sim \Gamma(\tau_{s_{01}}/2, \tau_{s_{02}}/2), \\ \mu_0|\mathbf{m}_0, \mathbf{S}_0 &\sim N_p(\mathbf{m}_0, \mathbf{S}_0), \quad \boldsymbol{\Sigma}_0|\nu_0\psi_0 \sim IW_p(\nu_0, \psi_0), \end{aligned}$$

and

$$\begin{aligned}\alpha_1|a_1,b_1 &\sim \Gamma(a_1,b_1), \quad \tau_{12}|\tau_{s_{11}},\tau_{s_{12}} \sim \Gamma(\tau_{s_{11}}/2,\tau_{s_{12}}/2),\\ \mu_1|\mathbf{m}_1,\mathbf{S}_1 &\sim N_p(\mathbf{m}_1,\mathbf{S}_1), \quad \mathbf{\Sigma}_1|\nu_1\psi_1 \sim IW_p(\nu_1,\psi_1),\end{aligned}$$

all with fixed hyperparameters. Here, $IW_p(\nu,\psi)$ denotes a p-dimensional inverted-Wishart distribution with ν degrees of freedom and scale matrix ψ.

The Gibbs sampling scheme for the LDDP model is based on the non-conjugate updating scheme of Algorithm 8 in Neal (2000), which modifies the auxiliary variable approach of MacEachern (1999). Pertinent details, including relevant full conditional distributions are provided in the online supplementary materials for De Iorio et al. (2009) and Jara et al. (2010). A function for fitting the two LDDP models and extracting ROC curves for any covariate $\mathbf{x}$ has been included in the `R` package `DPpackage`. The function `LDDProc(.)` implements the proposed BNP ROC regression model.

3.3 Inference

Inferences are based on Markov chain Monte Carlo (MCMC) iterates $\{(F_{0\mathbf{x}}^{(t)}, F_{1\mathbf{x}}^{(t)}), t = 1,\ldots,T\}$. Each iteration of the MCMC algorithm is used to obtain $\text{ROC}_{\mathbf{x}}(u) = 1 - F_1^{(t)}(F_0^{-1(t)}(1-u|\mathbf{x})|\mathbf{x})$ and $\text{AUC}_{\mathbf{x}}^{(t)} = \int_0^1 \text{ROC}_{\mathbf{x}}^{(t)}(u)\text{d}u$, where Monte Carlo integration is used to evaluate the integral.

The ROC curve is estimated as $\text{ROC}_{\mathbf{x}}(u) = (1/T)\sum_{t=1}^{T}\text{ROC}_{\mathbf{x}}^{(t)}(u)$. Point estimates of the AUC can be based on the MCMC mean (or median) and 95% credible intervals can also be obtained based on the MCMC iterates.

4 Simulation study

To evaluate the practical performance of our estimator, we conducted simulations under three different scenarios, namely: (a) a linear scenario, (b) a nonlinear scenario, and (c) a mixture model scenario. In each of the three scenarios we ran a total of 100 replications, because simulations are computationally time-consuming, and we considered the same sample size with $n = n_0 = n_1 = 50, 100, 200$. The efficiency and robustness of our estimator was studied by comparing it with its main competitors, namely the semiparametric model of Pepe (1998) and the nonparametric kernel estimator of González-Manteiga et al. (2011) and Rodríguez-Álvarez et al. (2011). The main difference between these latter two models is the order of the local polynomial kernel smoothers used for estimating the regression functions; whereas González-Manteiga et al. (2011) uses a local constant fit (order 0), Rodríguez-Álvarez et al. (2011) uses a linear fit (order 1). Since local constant regression suffers from boundary–bias problems (Fan and Gijbels, 1996), we followed Rodríguez-Álvarez et al. (2011). Given that these kernel estimators are only developed for univariate covariates, we restricted the simulation study to this framework.

To implement our BNP estimator, we fitted a mixture of B-splines models with $\mathbf{x}'\beta = \beta_0 + \sum_{j=1}^{3}\Upsilon_j(x)\beta_j$, where $\Upsilon_k(x)$ corresponds to the kth B-spline basis evaluated at x. We also used the Zellner's g-prior (Zellner, 1983), with $g = 10^2$. The following values for the hyperparameters were considered: $a_0 = a_1 = 5$, $b_0 = b_1 = 1$, $\mathbf{m}_0 = \mathbf{m}_1 = (0,0,0,0)$, $\mathbf{S}_0 = \mathbf{S}_1 = 10^2 \times \mathbf{I}_4$, $\nu_0 = \nu_1 = 6$, $\psi_0 = \psi_1 = \mathbf{I}_4$, $\tau_{s_{01}} = \tau_{s_{11}} = 6.01$, $\tau_{s_{02}} = \tau_{s_{12}} = 2.01$, and $\tau_{01} = \tau_{11} = 6.01$. These values lead to noninformative prior distributions. In all cases, we considered 2000 iterations of the MCMC algorithm, after a burn-in period of 2000 iterates. For the implementation of the kernel estimator,

the needed regression and variance functions were estimated using local linear and local constant fits, respectively. The Gaussian kernel $K(u) = (1/\sqrt{2\pi})\exp(-u^2/2)$ was chosen and generalized cross-validation was used to select the optimal bandwidth. More details on existing methods are given in Appendix.

The discrepancy between each estimator of the ROC curve and the true ROC curve was measured in terms of the empirical version of the global mean squared error

$$\text{MSE} = \frac{1}{n_x}\sum_{l=1}^{n_x}\frac{1}{n_u}\sum_{r=1}^{n_u}(\widehat{\text{ROC}}_{x_l}(u_r) - \text{ROC}_{x_l}(u_r))^2,$$

where $n_x = 25$ and $n_u = 100$, $u \in (0, 1)$. Table 1 summarizes the average MSE along with the standard deviation for each scenario, method and sample size.

4.1 Scenario 1—linear scenario

The purpose of including a linear scenario is to ascertain the loss of efficiency of our estimator when the parametric assumption holds, i.e., when the effect of the covariate is linear. We generated data from:

$$\begin{aligned} y_0|x_0 &\sim N(0.5 + x_0, 2.25), \\ y_1|x_1 &\sim N(2 + 4x_1, 4), \end{aligned}$$

where the covariates x_0 and x_1 are independently generated from $U(-1, 1)$. Figure 1 depicts the estimated AUC along the 2.5% and 97.5% simulation quantiles. All the three methods recover the functional form of the true AUC successfully. As expected, the semiparametric estimator has the better performance. The kernel and BNP approaches have similar performances with a slight better performance of the kernel estimator. The variance of all estimates decreases as sample size increases and this happens also in the following scenarios.

4.2 Scenario 2—nonlinear scenario

In this scenario the effect of the covariate is far from linear. We generated data from:

$$\begin{aligned} y_0|x_0 &\sim N(\sin(\pi(x_0 + 1)), 0.25), \\ y_1|x_1 &\sim N(0.5 + x_1^2, 1), \end{aligned}$$

where the covariates x_0 and x_1 are independently generated from $U(-1, 1)$. In this scenario, as can be seen in Figure 2, the BNP and kernel estimators successfully recover the functional form of the true AUC; as expected, the estimates of the semiparametric model are clearly unsuitable. The results in terms of the MSEs are shown in Figure 4 (b). These boxplots show that the BNP estimator is competitive with the kernel estimator, providing even slightly better results for small ($n = 50$) and moderate ($n = 100$) sample sizes.

4.3 Scenario 3—mixture model scenario

As a challenging scenario, we simulated data for the diseased group from a mixture model, with the mixture weights depending on the covariate, different error variances and a nonlinear mean

function for the second component. For the healthy group, we simulated data with a nonlinear mean function and with the variance depending on the covariate. In short, we simulated data from:

$$\begin{aligned} y_0|x_0 &\sim N(\sin(\pi x_0), 0.2 + 0.5\exp(x_0)), \\ y_1|x_1 &\sim \exp(x_1)/(1+\exp(x_1))N(x_1, 0.25) + (1 - (\exp(x_1)/(1+\exp(x_1))))N(x_1^3, 1), \end{aligned}$$

where x_0 and x_1 are independently generated from $U(-1, 1)$. Figure 3 depicts the estimated AUC along the 2.5% and 97.5% and from this figure we can see that the nonparametric estimators recover the true AUC successfully, whereas the semiparametric estimator is unable to recover it. Figure 4 (c) shows the boxplot of the MSEs for this scenario. As can be seen, the MSEs produced by the semiparametric estimator are much larger than the ones produced by the nonparametric estimators. The BNP estimator, under this scenario, clearly outperforms the kernel estimator, providing smaller MSEs (with also smaller variance), for all sample sizes considered.

5 Application to diabetes diagnosis

5.1 Data description and motivation

Our data are from a population-based pilot survey of diabetes in Cairo, Egypt, in which postprandial blood glucose measurements were obtained from a fingerstick on 286 subjects. The gold standard for diagnosing diabetes, according to the World Health Organization criteria, consists of a fasting plasma glucose value $\geqslant$ 140 mg/dl or a 2 hour plasma glucose value $\geqslant$ 200 mg/dl following a 75g oral glucose challenge. Based on these criteria 88 subjects were classified as diseased and 198 as healthy. This data has also been analyzed in Smith and Thompson (1996), Faraggi (2003), and in González-Manteiga et al. (2011).

We applied the ROC methodology proposed in this paper to this diabetes study, with the aim of using glucose levels to detect patients having a higher risk of diabetes problems, and to assess the effect of age on the accuracy of this marker.

5.2 BNP ROC modeling

The data analysis is divided into two parts. First, we examine the glucose performance as a marker to diagnose diabetes and then we conducted our ROC regression analysis which takes into account the effect of age.

5.2.1 ROC analysis of the discriminatory ability of glucose

We carried out an initial analysis to evaluate the discriminatory capacity of glucose to detect diabetes, ignoring the age effect. Figure 5 (a) and (b) shows the DPM of normals estimated densities of the glucose levels in the healthy and diseased populations. As expected, disease subjects tend to have more probability mass for higher values of glucose. In Figure 5 (c) is presented the corresponding estimator of the ROC curve. The curves lies well above the diagonal line, indicating a good discriminatory performance of the glucose to distinguish subjects with diabetes from those who are healthy. This can also be seen from the AUC, which is 0.858.

5.2.2 Induced Bayesian nonparametric ROC regression analysis

After analyzing the discriminatory capacity of glucose, we conducted our ROC regression analysis, which takes into account the affect of age. We used the same prior specification as in the simulation study.

Figure 6 (a) shows the scatterplot of the data for the healthy and diseased populations, and here we can see that the relationship between age and glucose is nonlinear in the diseased population. In Figure 6 (b) are presented the estimated mean regression functions, along with 95% credible intervals, which suggest that older healthy subjects tend to have high values of glucose. This agrees with Smith and Thompson (1996) who suggest that the aging process is associated with relative insulin deficiency or resistance among people who are healthy. The different behavior between the diseased and nondiseased populations is evident, especially for younger ages, and this is reinforced by Figure 7, where we plot the conditional densities of the glucose levels at different ages. We can understand from this figure how the probability mass changes over different age profiles. A more complete view is depicted in Figure 8, where we show a surface of conditional densities, with each profile corresponding to a conditional density analogous to the ones shown in Figure 7.

The estimated covariate specific ROC surface is shown in Figure 9. The estimated ROC curves, in Figure 5, appear to indicate that for 20 and 40 years old ages, the discriminatory capacity of the marker is greater than that see in Figure 5 (c), where as for an age of 80 years old it is substantially lower. This can also be seen from the AUCs, which is 0.974 $(0.931, 0.996)$ for an age of 20, and 0.877 $(0.797, 0.936)$, and 0.758 $(0.644, 0.866)$ for ages of 40 and 80 years, respectively.

To examine the age effect further, Figure 11 provide the estimate for the AUC. A pointwise 95% credible interval is also presented to give an indicator of the variability of the estimator process. This figure clearly show the effect of age. The discrimination between the diseased and the healthy by the glucose is much better for young subjects than for old subjects. The AUC= 0.858 that was obtained above, while ignoring age, can be seen from Figure 11 to correspond to an age of 46 years, which is close to the average age of the healthy subjects (48 years). Thus, ignoring age will result in an underestimated area for subjects under the age of 46 years, and an overestimate for those older than 46 years. Owing to the strong falling off in the AUC for old subjects, it is questionable whether this marker should be used for screening such older people. Observing the pointwise credible intervals, one can see that the lower bound for younger ages exceeds the upper bound for old ages. This indicates a substantial effect of age on AUC. Further, the width of the confidence interval for younger ages is much shorter than that for old ages, and hence, inference for young subjects will be more precise.

6 Conclusions

We have presented a flexible BNP model to estimate the ROC curve in the presence of covariates based on LDDP priors. Since the conditional distributions of the marker results were modeled nonparametrically, our approach allows for great flexibility in terms of the shapes of these distributions. In fact, because a nonparametric approach was used, the presented model is applicable to a wide range of markers and for a wide range of diseases.

Semiparametric and nonparametric kernel approaches were considered for estimating the ROC curve and the AUC. Simulation results indicate that even when the linearity assumption holds, our BNP estimator gives relatively accurate results. The performance of our estimator is competitive with, and in some cases significantly better than, the kernel-based estimator. It is, however, rela-

tively easy to obtain nonparametric estimates of conditional ROC curves utilizing the kernel based approaches. Our BNP approach depend on intensive computations and can be time-consuming. The advantage, however, is that the posterior uncertainty about the conditional ROC curve can be assessed probabilistically using the posterior predictive distribution of the conditional ROC, whereas the kernel approaches relies upon large sample frequentist coverage (or confidence intervals) which do not have a probabilistic interpretation.

No rule of thumb exists to say whether a parametric or nonparametric model should be used, the choice of models is data dependent. If the parametric assumption holds, a natural option is to use a parametric model. However, given that the linear specification of the LDDP is highly flexible, including, for example, linear and nonlinear transformations of the predictors, different models can be fitted and model selection can be based on the log pseudo marginal likelihood statistics (Geisser and Eddy, 1979).

In our application we evaluated the effect of age on the glucose discriminatory capacity to detect diabetes. We found out that older ages reduce the accuracy of the glucose to distinguish between healthy and diseased subjects. This observation should be taken into account on the use of this marker in clinical diagnosis of diabetes.

Acknowledgements

VI thanks Antónia Amaral Turkman for support and Wenceslao González-Manteiga for having shared his expertise on kernel techniques with her. The research of VI is funded by the Portuguese Foundation for Science and Technology through the PhD grant SFRH/BD/47742/2008.

Appendix

Details on existing methods

Semiparametric model (Pepe, 1998)

This method is based on specifying a homocedastic linear regression model for the healthy and diseased groups, i.e.,

$$\begin{aligned} y_0 &= \tilde{\mathbf{x}}'\boldsymbol{\beta}_0 + \sigma_0\varepsilon_0 \\ y_1 &= \tilde{\mathbf{x}}'\boldsymbol{\beta}_1 + \sigma_1\varepsilon_1, \end{aligned}$$

where $\tilde{\mathbf{x}}' = (1, \mathbf{x}')'$, $\boldsymbol{\beta}_0 = (\beta_{00}, \ldots, \beta_{0p})$ and $\boldsymbol{\beta}_1 = (\beta_{10}, \ldots, \beta_{1p})$ are $(p+1)$-dimensional vectors of unknown parameters, and ε_0 and ε_1 are i.i.d. $N(0,1)$, with distribution functions F_0 and F_1, respectively. The estimation procedure consists of the following steps:

1. estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ by ordinary least squares, on the basis of samples $\{(y_{0i}, \mathbf{x}_{0i})\}_{i=1}^{n_0}$ and $\{(y_{1j}, \mathbf{x}_{1j})\}_{j=1}^{n_1}$;
2. estimate σ_0^2 and σ_1^2 as
$$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n_0}(y_{0i} - \tilde{\mathbf{x}}_{0i}'\hat{\boldsymbol{\beta}}_0)^2}{n_0 - p - 1} \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \tilde{\mathbf{x}}_{1j}'\hat{\boldsymbol{\beta}}_1)^2}{n_1 - p - 1};$$
3. estimate the cumulative distribution functions F_0 and F_1 on the basis of the empirical distributions of the standardized residuals
$$\hat{F}_0(y) = \frac{1}{n_0}\sum_{i=1}^{n_0} I\left[\frac{y_{0i} - \tilde{\mathbf{x}}_{0i}'\hat{\boldsymbol{\beta}}_0}{\hat{\sigma}_0} \leqslant y\right] \quad \text{and} \quad \hat{F}_1(y) = \frac{1}{n_1}\sum_{j=1}^{n_1} I\left[\frac{y_{1j} - \tilde{\mathbf{x}}_{1j}'\hat{\boldsymbol{\beta}}_1}{\hat{\sigma}_1} \leqslant y\right];$$
4. For a given value of the covariate $\mathbf{x}$, calculate the covariate specific ROC curve
$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(u) = 1 - \hat{F}_1(\tilde{\mathbf{x}}'\hat{\boldsymbol{\beta}} + \hat{\alpha}\hat{F}_0^{-1}(1-u)), \quad 0 < u < 1,$$
where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_1)/\hat{\sigma}_1$ and $\hat{\alpha} = \hat{\sigma}_0/\hat{\sigma}_1$.

Nonparametric model (González-Manteiga et al., 2011 and Rodríguez-Álvarez et al., 2011)

In this method a nonparametric heterocedastic regression model is assumed for the test result

$$\begin{aligned} y_0 &= \mu_0(x) + \sigma_0(x)\varepsilon_0 \\ y_1 &= \mu_1(x) + \sigma_1(x)\varepsilon_1, \end{aligned}$$

where x is a continuous covariate, μ_0 and μ_1 are the regression functions, and σ_0 and σ_1 are the variance functions. ε_0, ε_1 i.i.d. $\sim N(0,1)$ and are assumed to be independent of the covariate x, and with distribution functions F_0 and F_1, respectively. The proposed estimation procedure is as follows:

1. For a given value x of the covariate, estimate the regression functions μ_0 and μ_1 as

$$\begin{aligned} \hat{\mu}_0(x) &= \hat{\psi}(x, \{(y_{0i}, x_{0i})\}_{i=1}^{n_0}, h_0, p_0), \\ \hat{\mu}_1(x) &= \hat{\psi}(x, \{(y_{1j}, x_{1j})\}_{j=1}^{n_1}, h_1, p_1), \end{aligned}$$

   where $\hat{\phi}$ is the local polynomial kernel estimator (Fan and Gijbels, 1996), h_0 and h_1 are the smoothing parameters or bandwidths, and p_0 and p_1 are the order of polynomials, in the healthy and diseased populations, respectively;

2. estimate the variance functions σ_0^2 and σ_1^2 in a similar fashion

$$\begin{aligned} \hat{\sigma}_0^2(x) &= \hat{\psi}(x, \{(z_{0i}, x_{0i})\}_{i=1}^{n_0}, g_0, q_0), \\ \hat{\sigma}_1^2(x) &= \hat{\psi}(x, \{(z_{1j}, x_{1j})\}_{j=1}^{n_1}, g_1, q_1), \end{aligned}$$

   where $z_{0i} = (y_{0i} - \hat{\mu}_0(x_{0i}))^2$, $z_{1j} = (y_{1j} - \hat{\mu}_1(x_{1j}))^2$, g_0 and g_1 are the bandwidths and q_0 and q_1 are the order of the polynomials;

3. estimate the cumulative distribution functions F_0 and F_1 on the basis of the empirical distributions of the standardized residuals

$$\hat{F}_0(y) = \frac{1}{n_0}\sum_{i=1}^{n_0} I\left[\frac{y_{0i} - \hat{\mu}_0(x_{0i})}{\hat{\sigma}_0(x_{0i})} \leqslant y\right] \quad \text{and} \quad \hat{F}_1(y) = \frac{1}{n_1}\sum_{j=1}^{n_1} I\left[\frac{y_{1j} - \hat{\mu}_1(x_{1j})}{\hat{\sigma}_1(x_{1j})} \leqslant y\right];$$

4. compute the covariate specific ROC curve as follows:

$$\widehat{\mathrm{ROC}}_x(u) = 1 - \hat{F}_1\left(\frac{\hat{\mu}_0(x) - \hat{\mu}_1(x)}{\hat{\sigma}_1(x)} + \frac{\hat{\sigma}_0(x)}{\hat{\sigma}_1(x)}\hat{F}_0^{-1}(1-u)\right), \quad 0 < u < 1.$$

References

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

Alonzo, T.A., Pepe. M.S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, **3**, 421–432.

Branscum, A.J., Johnson, W.O., Hanson, T.E., Gardner, I.A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine*, **27**, 2474–2496.

Cai, T. (2004). Semiparametric ROC regression analysis with placement values. *Biostatistics*, **5**, 45–60.

Caron, F., Davy, M., Doucet, A., Duflos, E., Vanheeghe, P. (2008) Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Transactions on Signal Processing*, **56**, 71–84.

De Iorio, M., Johnson, W.O., Müller, P., Rosner, G.L. (2009). Bayesian nonparametric nonproportional hazards survival modelling. *Biometrics*, **65**, 762–771.

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

De La Cruz, R., Quintana, F.A., Müller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *Journal of the Royal Statistical Society: Series C*, **56**, 119–137.

Dunson, D.B., Herring, A.H. (2006). Semiparametric Bayesian latent trajectory models. Technical report, ISDS Discussion Paper 16, Duke University, Durham, NC, USA.

Erkanli, A., Sung, M., Costello, E.J., Angold, A. (2006). Bayesian semiparametric ROC analysis. *Statistics in Medicine*, **25**, 3905–3928.

Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician*, **52**, 179–192.

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

Freedman, D.A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Annals of the Institute of Statistical Mathematics*, **34**, 1194–1216.

Geisser, S., Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand, A.E., Kottas, A., MacEachern, S.N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, **100**, 1021–1035.

González-Manteiga, W., Pardo-Fernandéz, J.C., Van Keilegon, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics*, **38**, 169–184.

Hanson, T.E., Branscum, A., Gardner, I. (2008). Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling*, **8**, 81–96.

Hanson, T.E., Kottas, A., Branscum, A.J. (2008). Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. *Journal of the Royal Statistical Society: Series C*, **57**, 207–225.

Hsieh, F., Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, **24**, 24–40.

Inácio, V., Turkman, A.A., Nakas, C.T., Alonzo, T.A. (2011). Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal*, **53**, 1011–1024.

Jara, A., Lesaffre, E., De Iorio, M., Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly–interval–censored data. *Annals of Applied Statistics*, **4**, 2126–2149.

Lang, S., Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Lloyd, C.J. (1998). Using smooth receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93**, 1356–1364.

MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the section on Bayesian Statistical Science, Alexandria, VA*. American Statistical Association.

MacEachern, S.N., Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

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

Peng, L., Zhou, X.H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*, **118**, 129–143.

Pepe, M.S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, **54**, 124–135.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, New York.

Rodríguez-Álvarez, M.X., Roca-Pardiñas, J., Cadarso-Suárez, C. (2011). ROC curve and covariates: extending the induced methodology to the non-parametric framework. *Statistics and Computing*, **21**, 483–495.

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

Sarwar, N., Gao, P., Seshasai, S.R., Gobin, R., Kaptoge, S., Di Angelantonio, E., Ingelsson, E., Lawlor, D.A., Selvin, E., Stampfer, M., Stehouwer, C.D., Lewington, S., Pennells, L., Thompson, A., Sattar, N., White, I.R., Ray, K.K., Danesh, J. (2010). Diabetes mellitus fasting blood glucose concentration and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet*, **375**, 2215–2222.

Smith, P.J., Thompson, T.J. (1996). Correcting for confounding in analyzing receiver operating characteristic curves. *Biometrical Journal*, **7**, 857–863.

Wild., S., Roghic, G., Green, A., Sicree, R., King, H. (2004). Global prevalence of diabetes: estimates for 2000 and projection for 2030. *Diabetes Care*, **27**, 1047–1053.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *Journal of the Royal Statistical Society: Series D*, **32**, 23–34.

Zhou, X.H., Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine*, **21**, 2045–2055.

Zou, K.H., Hall, W.J., Shapiro, D.E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, **16**, 2143–2156.

Table 1: Averages (and standard deviations) of the estimated mean squared errors obtained from the 100 datasets simulated according to Scenarios 1, 2 and 3, for different sample sizes and different approaches.

Scenario 1	Semiparametric	Kernel	BNP
$n = 50$	0.0084 (0.0057)	0.0131 (0.0073)	0.0138 (0.0080)
$n = 100$	0.0045 (0.0026)	0.0074 (0.0043)	0.0075 (0.0045)
$n = 200$	0.0022 (0.0014)	0.0036 (0.0020)	0.0040 (0.0024)
Scenario 2	Semiparametric	Kernel	BNP
$n = 50$	0.0385 (0.0056)	0.0130 (0.0064)	0.0106 (0.0056)
$n = 100$	0.0364 (0.0037)	0.0079 (0.0041)	0.0070 (0.0035)
$n = 200$	0.0345 (0.0022)	0.0042 (0.0017)	0.0049 (0.0032)
Scenario 3	Semiparametric	Kernel	BNP
$n = 50$	0.0534 (0.0090)	0.0302 (0.0156)	0.0160 (0.0093)
$n = 100$	0.0499 (0.0057)	0.0155 (0.0064)	0.0087 (0.0047)
$n = 200$	0.0470 (0.0036)	0.0098 (0.0041)	0.0056 (0.0028)

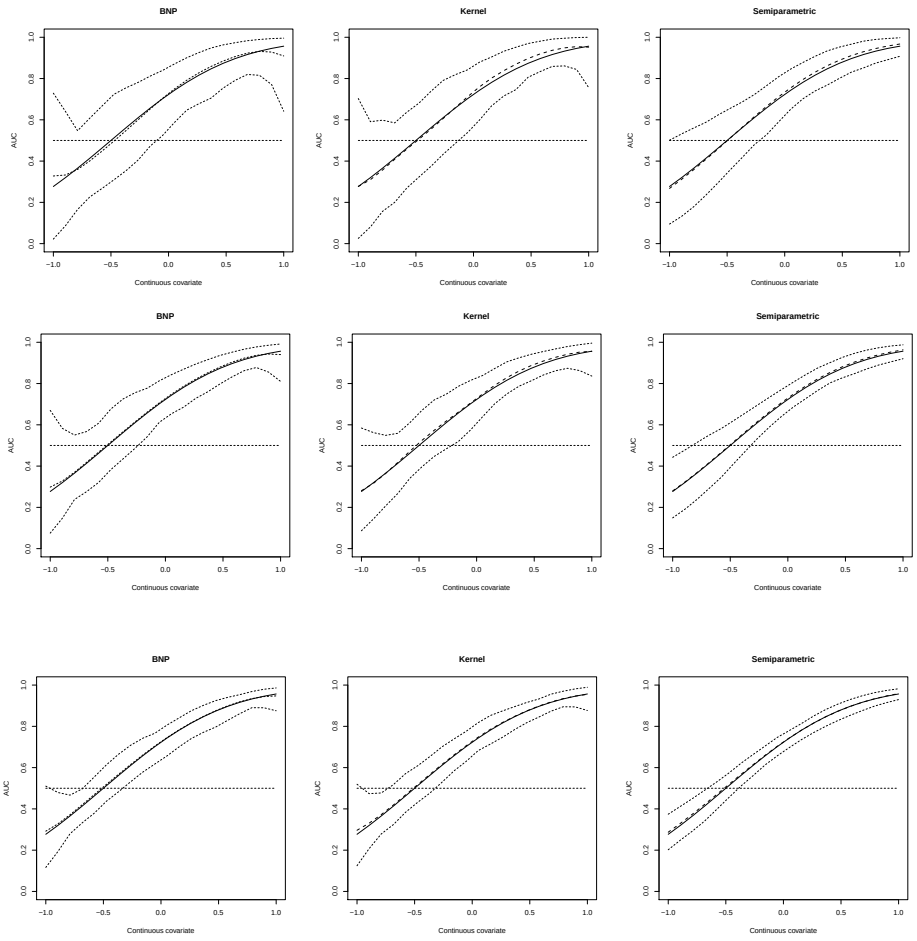


Figure 1: True AUC (solid line) versus the average of simulated AUCs, along with 2.5% and 97.5% simulation quantiles (dashed line), for Scenario 1. Row 1: $n = 50$, Row 2: $n = 100$, Row 3: $n = 200$.

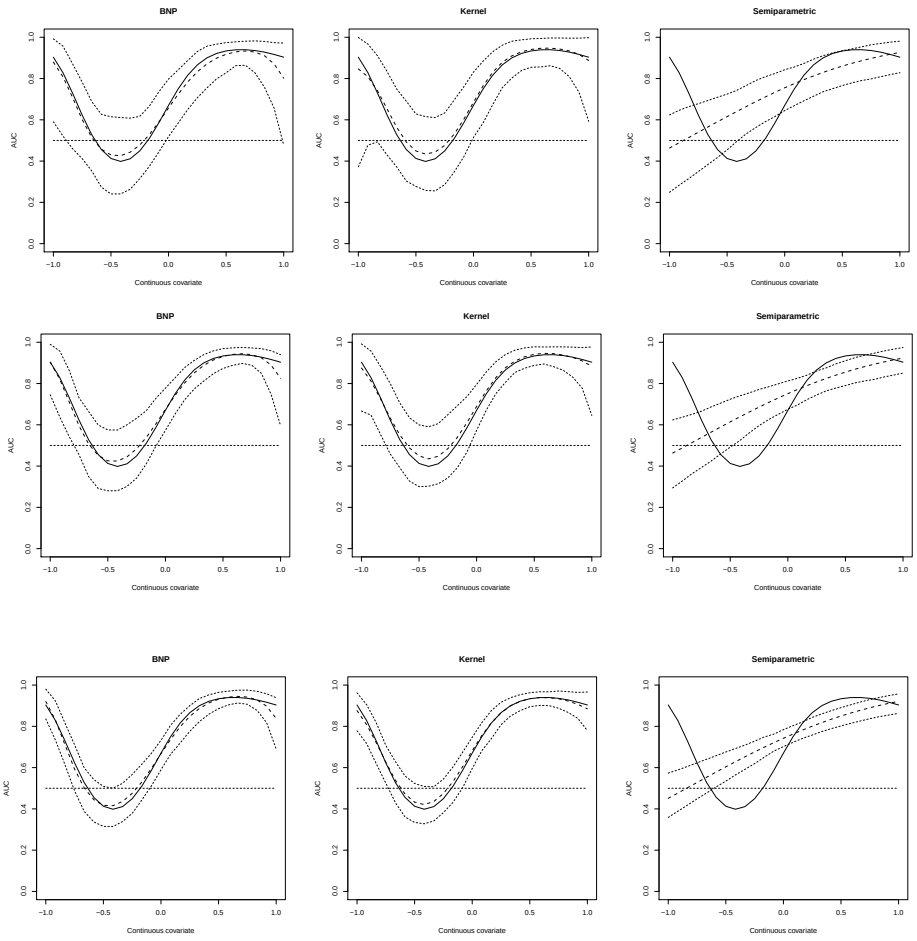


Figure 2: True AUC (solid line) versus the average of simulated AUCs, along with 2.5% and 97.5% simulation quantiles (dashed line), for Scenario 2. Row 1: $n = 50$, Row 2: $n = 100$, Row 3: $n = 200$.

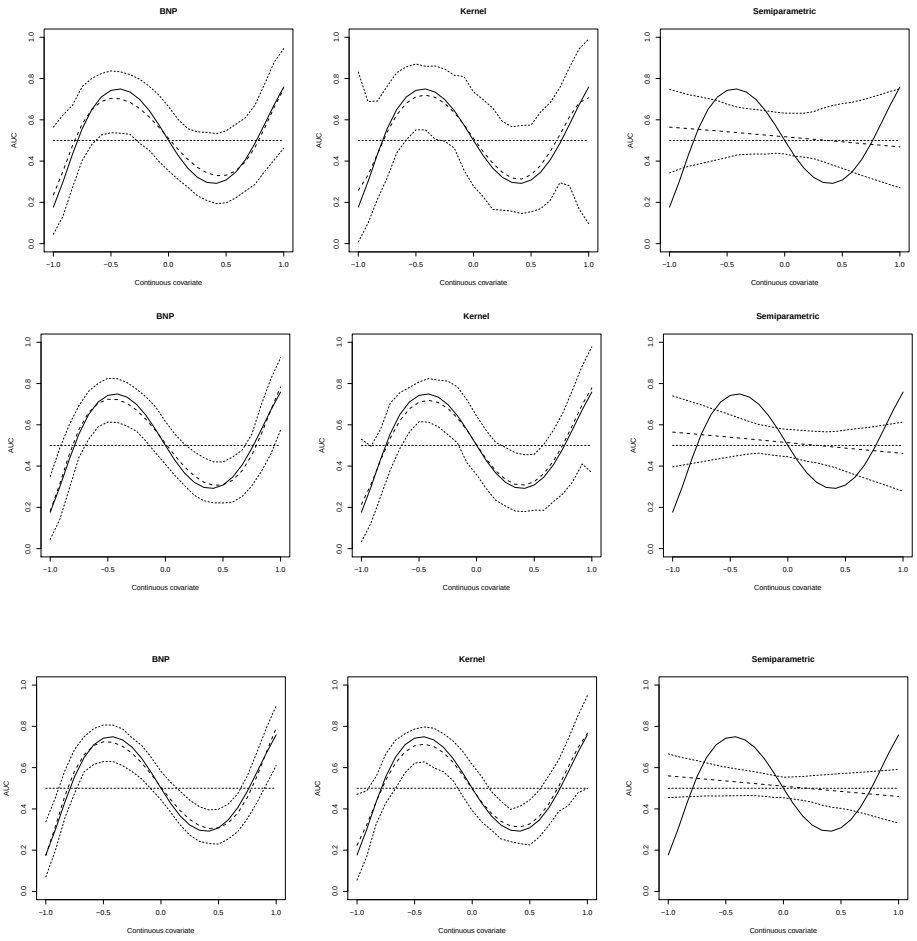


Figure 3: True AUC (solid line) versus the average of simulated AUCs, along with 2.5% and 97.5% simulation quantiles (dashed line), for Scenario 3. Row 1: $n = 50$, Row 2: $n = 100$, Row 3: $n = 200$.

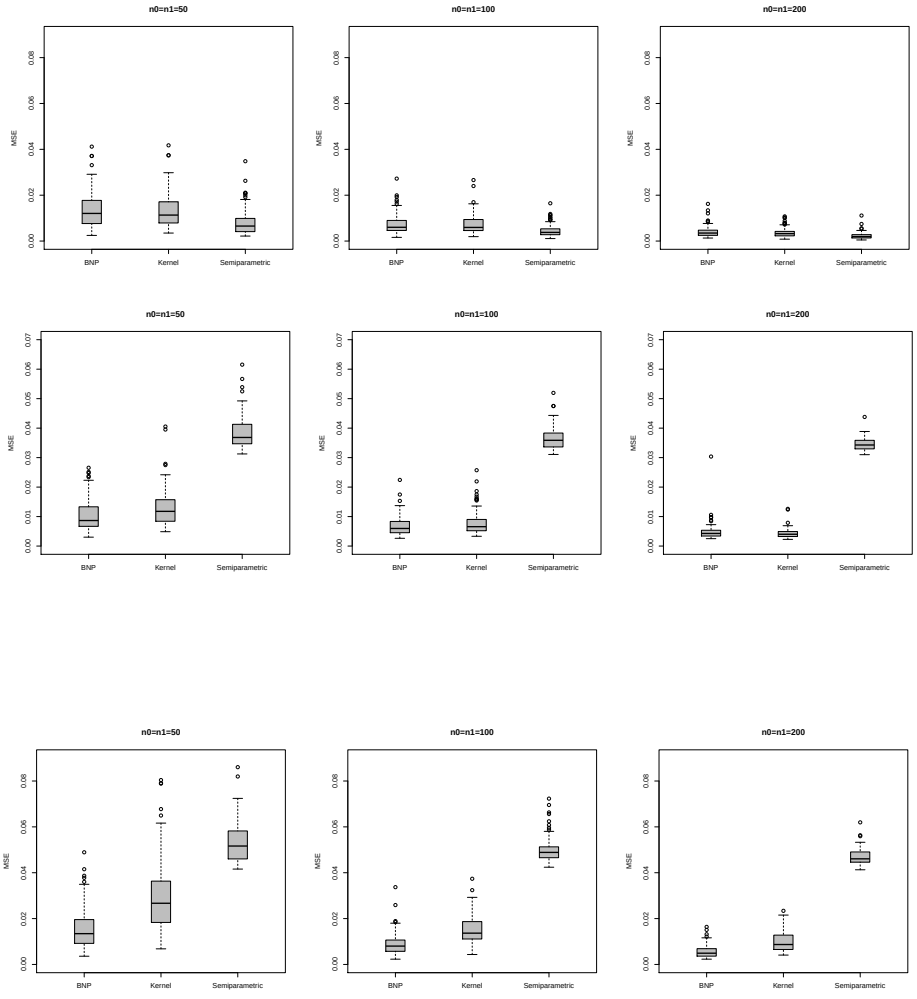
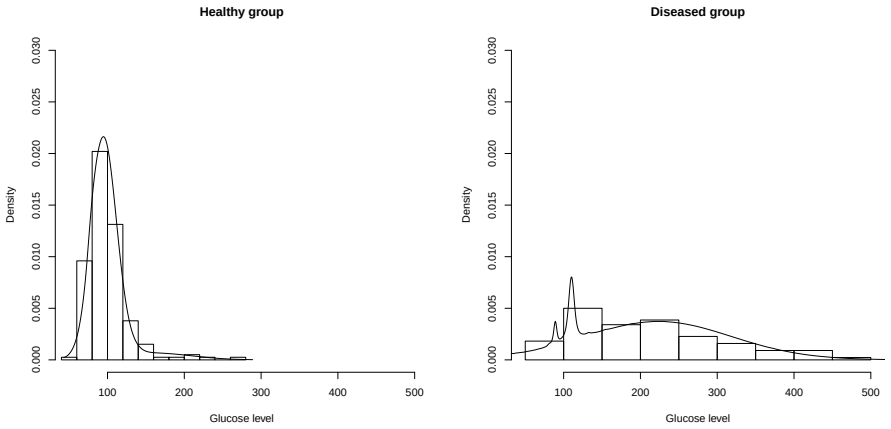
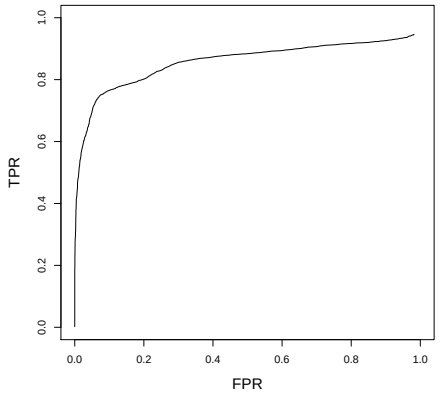


Figure 4: Row 1: MSE for simulation Scenario 1; Row 2: MSE for simulation Scenario 2, Row 3: MSE for simulation Scenario 3.



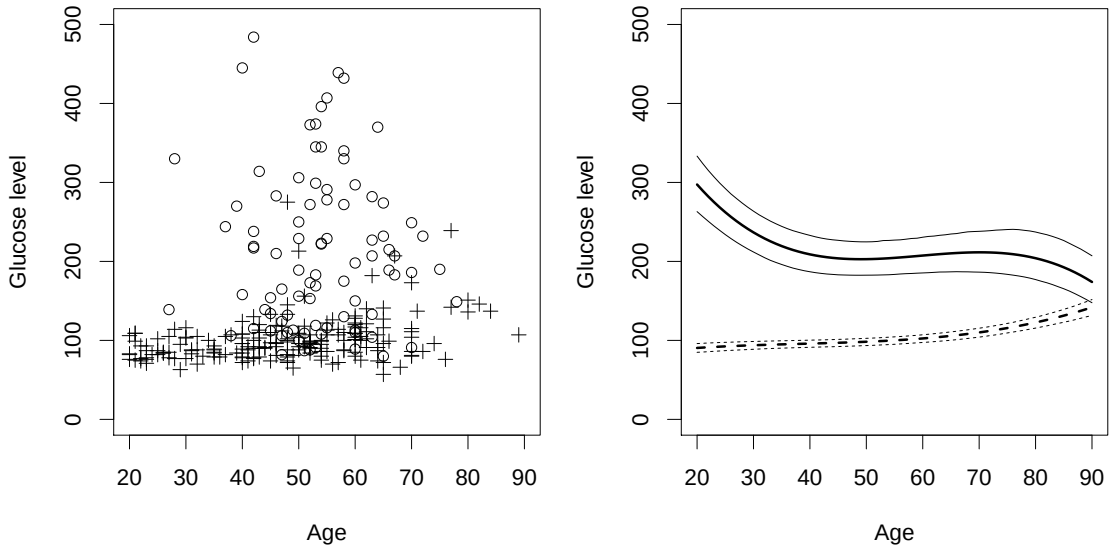
(a)

(b)



(c)

Figure 5: (a) DPM of normals estimated densities of the glucose levels in healthy and; (b) diseased populations; (c) ROC curve of the glucose levels with no age effect: Bayesian estimator using DPM of normals.



(a) (b)

Figure 6: (a) Scatterplot of the data. The diseased group is represented by circles and the healthy group by crosses. (b) Estimated mean functions along with 95% credible intervals. *Dashed line*: healthy group. *Solid line*: diseased group.

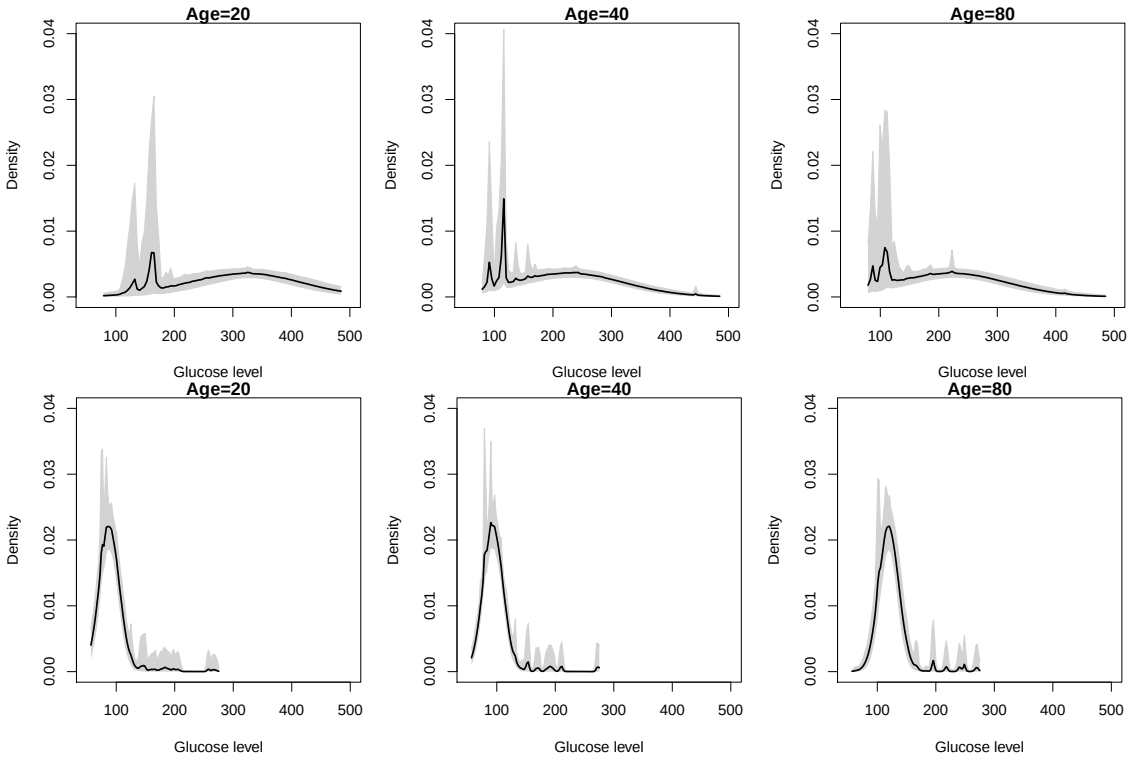
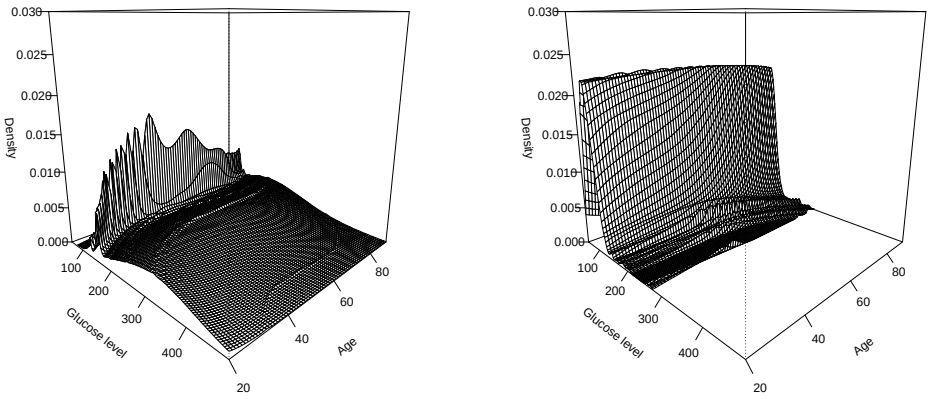


Figure 7: Top: Estimated densities for glucose conditional on a range of values of the age (ages $20, 40$, and 80) for the diseased population and 95% credible intervals. Bottom: Estimated densities for glucose conditional on a range of values of the age (ages $20, 40$, and 80) for the healthy population and 95% credible intervals.



(a) (b)

Figure 8: Estimated density surfaces for glucose conditional on age for (a) diseased population; (b) healthy population.

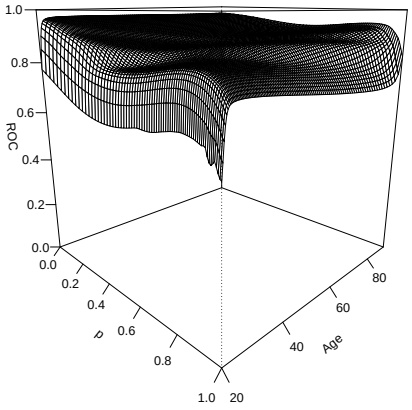


Figure 9: Estimated ROC surface.

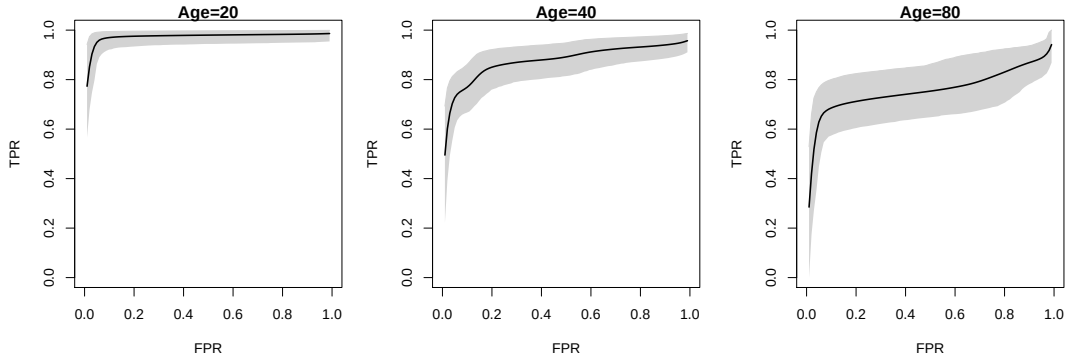


Figure 10: Estimated conditional ROC curves along 95% credible intervals for ages 20, 40, and 80.

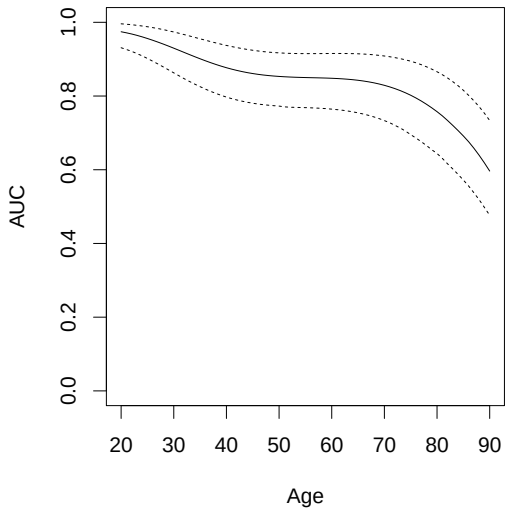


Figure 11: Estimated AUC as a function of age with a 95% credible interval.