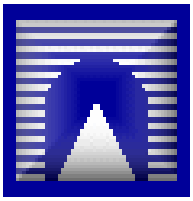
Università degli Studi Roma Tre
FACOLTÁ DI SCIENZE MATEMATICHE FISICHE E NATURALI



Graduation Thesis in Mathemathics by
Nazareno Maroni

Nonparametric Bayesian Binary Regression: an approach based on Gaussian Processes

Supervisor

Prof. Brunero Liseo

The Candidate The Supervisor

ACADEMIC YEAR 2006 - 2007
MAY 2008

AMS Classification: primary 62G08; secondary 62J12.
Key Words: Probit regression, Logistic regression, Gibbs Sampling.

Synthesis

Bayesian Analysis

Classical statistics uses only informations deriving from the likelihood. If $X_1, \ldots, X_n$ are independent random variables arise from the density $p(x, \theta)$ with $\theta \in \Theta$ a parameter, the likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} p(x_i, \theta). \tag{1}$$

One important estimator for θ used in this case is the maximum likelihood estimation that is defined as

$$\hat{\theta} \in \Theta \text{ such that } L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta). \tag{2}$$

Recently, the Bayesian method has had a big developing. The Monte Carlo Markov Chain methods allow us to generate a Markov chain that converges to the target distribution of interest. We use a Gibbs Sampling algorithm, which is a special case of MCMC algorithm. A simple case is the following. Let $\theta = (\theta_1, \theta_2) \in \Theta \subset \mathbb{R}^2$. We indicate the target distribution as $\pi(\theta_1, \theta_2)$. If we cannot sample from π but the following distributions (called full conditionals)

$$\pi_1(\theta_1|\theta_2) \qquad\qquad \pi_2(\theta_2|\theta_1)$$

are known, we can sample from these ones: we start with a value $\theta_0 = (\theta_1^{(0)}, \theta_2^{(0)})$ and we generate T values of θ: $\theta_1^{(1)} \sim \pi_1(\theta_1|\theta_2^{(0)})$, $\theta_2^{(1)} \sim \pi_2(\theta_2|\theta_1^{(1)})$,

then $\theta_1^{(2)} \sim \pi_1(\theta_1|\theta_2^{(1)})$ and so on. The sequence of these values converges to a sample from the target distribution π as T goes to ∞.

The Bayesian analysis is based, as well as on likelihood, on prior probability that are assigned to the parameter we want to estimate. Thus, the information is summarized by the posterior distribution of the parameter, and this is obtained through the Bayes theorem.

Theorem 1. Let A be an event such that $A \subset [B_1 \cup \cdots \cup B_n]$, where $B_1, \ldots, B_n$ are pairwise disjointed. Then for each i it follows that

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)\Pr(B_i)}{\sum_{j=1}^{n} \Pr(A|B_j)\Pr(B_j)}. \tag{3}$$

We consider θ as a random variable with density $\pi(\theta)$ that is the prior density. This distribution depends on the informations we have and so it is subjective. If we denote with x the vector $(x_1, \ldots, x_n)$ the expression of the Bayes theorem for the densities is, under some conditions on the existence of the denominator and on the existence of Radon-Nikodim derivatives,

$$\pi(\theta|x) = \frac{\pi(\theta)L(\theta, x)}{\int_\Theta \pi(\theta)L(\theta, x)\ \mathrm{d}\theta}. \tag{4}$$

The denominator of (4) is a normalisation constant and it is the marginal distribution of the random vector x. Thus, $\pi(\theta|x)$ is the posterior distribution for θ and to detect it we can only consider the result of the prior distribution for the likelihood.

In the nonparametric case, the "parameter" we want to estimate is a function, generally a distribution or a density function, thus the problem has a higher complexity and the MCMC (Monte Carlo Markov chain) algorithms are useful. We work on a binary regression problem.

We have a binary response variable Y and a corresponding covariate value X belonging to a covariate space $\mathcal{X}$. Our purpose is to estimate the response probability function $p(x) = \Pr(Y = 1|X = x)$. We consider a regression function that is $p(x) = H(\eta(x))$. In order to provide flexibility to the model, we do not choose an explicit form for η. Rather, we assume that $\eta(x)$ follows a Gaussian Process. Note that H plays the role of mapping $\mathbb{R} \to (0, 1)$.

Gaussian Process

We describe some basics of random fields. We start with the definition of a general random field that is:

Definition 2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and T be a parameter set. A random field is a finite or real valued function $X(t, \omega)$ which, $\forall\ t \in T$ fixed, is a measurable function of $\omega \in \Omega$, that is a random variable.

For a fixed $\omega \in \Omega$, the function $X(t, \omega)$ is a non-random function of t. This function is called a sample path or a realization of the random field and it is indicated with x_t. The parameter t is said the position. A random field can be described by its finite-dimensional distributions:

$$F_{t_1,\dots,t_k}(x_1, \dots, x_k) = \mathbb{P}(X_{t_1} \leqslant x_1, \dots, X_{t_k} \leqslant x_k)$$

that are right-continuous and nondecreasing and

$$F_{t_1,\dots,t_k}(x_1, \dots, -\infty, \dots, x_k) = 0, \qquad F_{t_1,\dots,t_k}(+\infty, \dots, +\infty) = 1.$$

The Kolmogorov's Existence Theorem establishes the existence of a random field.

Theorem 3. If a system of finite-dimensional distributions, $F_{t_1,\dots,t_k}$, satisfies the symmetry and the compatibility conditions, then there exists on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ a random field $\{X_t, t \in T\}$ having $F_{t_1,\dots,t_k}$ as its finite-dimensional distributions.

The symmetry condition is defined as following. If π is a permutation of the index set $\{1, \dots, k\}$, then

$$F_{t_1,\dots,t_k}(x_1, \dots, x_k) = F_{t_{\pi 1},\dots,t_{\pi k}}(x_{\pi 1}, \dots, x_{\pi k}),$$

provided that the events $(X_{t_1} \leqslant x_1, \dots, X_{t_k} \leqslant x_k)$ and $(X_{t_{\pi 1}} \leqslant x_{\pi 1}, \dots, X_{t_{\pi k}} \leqslant x_{\pi k})$ are identical.

The compatibility condition requires that

$$F_{t_1,\dots,t_{k-1}}(x_1, \dots, x_{k-1}) = F_{t_1,\dots,t_{k-1},t_k}(x_1, \dots, x_{k-1}, +\infty).$$

Now, we introduce Gaussian processes with the definition

Definition 4. A Gaussian random field is a random field where all the finite-dimensional distributions $F_{t_1,\ldots,t_k}$ are multivariate normal distributions $\forall\ k$ and $t_1, \ldots, t_k \in \mathbb{R}$.

A Gaussian process can be specified by giving its mean and covariance kernel functions and the finite-dimensional distributions are multivariate normal distributions. The multinormal probability densities are of the form

$$p_{t_1,\ldots,t_k}(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-m)^t \Sigma^{-1}(x-m)\Big\},$$

Both symmetry and compatibility conditions hold for a Gaussian process, thus for Theorem (3) there exists a Gaussian random field with probability densities well defined and the covariance matrix positive definite.

We now give the definitions of continuity with probability one, almost surely and mean square continuity.

Definition 5. A random field X has continuous sample paths with probability one if

$$\mathbb{P}(\omega : |X(t_n, \omega) - X(t, \omega)| \to_n 0,\ \forall\ t \in T) = 1$$

$\forall\ t_n$ such that $t_n \to_n t$.
A random field X is almost surely continuous if

$$\mathbb{P}(\omega : |X(t_n, \omega) - X(t, \omega)| \to_n 0) = 1$$

$\forall\ t_n$ such that $t_n \to_n t$ and $\forall\ t \in T$.
A random field X is mean square continuous if

$$\mathbb{E}(|X(t_n) - X(t)|^2) \to_n 0$$

$\forall\ t_n$ such that $t_n \to_n t$ and $\forall\ t \in T$.

A theorem holds for mean square continuity of a process and a relation for continuous sample paths with probability one that, for a Gaussian process, can be stated as follows.

Theorem 6. Let $X(t)$ be a Gaussian process with mean function $m(t) = 0$. Let covariance function $C(t, s)$ be continuous. If

$$\mathbb{E}(|X(t) - X(s)|^2) \leqslant \frac{c}{|\log \tau|^{1+\epsilon}},$$

for $c > 0$, $\epsilon > 0$ and $\forall\ \tau < 1$, then $X(t)$ has continuous sample paths with probability one.

Mean square differentiability is a necessary condition to have differentiable sample paths and if the covariance function $C(t, s)$ is such that $\dfrac{\partial^2 C(t,s)}{\partial t_i \partial s_i}$ exists and is finite $\forall\ i = 1, \ldots, n$ at (t, t), then $X(t)$ is mean square differentiable at t.

Now, we introduce the notion of Reproducing Kernel Hilbert space. We indicate with S the space of function $f(t)$ with $t \in I$.

Definition 7. Let S be a space with a inner product. If S is a Banach space under the norm induced by the inner product, then S is a Hilbert space.

We proceed with the definition of a reproducing kernel Hilbert space:

Definition 8. A Hilbert space S of functions on I is called a reproducing kernel Hilbert space if there exists a doubly indexed function $R(t, s)$ on $I \otimes I$ which satisfies the following conditions:

(i) $R(t, \cdot) \in S \quad \forall\ t \in I$,

(ii) $\langle f, R(t, \cdot)\rangle_S = f(t) \quad \forall\ f \in S,\ t \in I$.

The function R is called the reproducing kernel of S.

In the case of a Gaussian process with mean function $m(t)$ and covariance kernel $C(t, s)$, if we assume that the covariance function is of the form

$$C(t, s) = \tau^{-1} C_0(\lambda t, \lambda s),$$

with C_0 a nonsingular covariance kernel, $\tau > 0$ and $\lambda > 0$ and if we define the set $\mathcal{A}$ as

$$\mathcal{A} = \left\{ X(t) = \sum_{i=1}^{k} a_i C_0(\lambda t, \lambda s_i),\ a_1, \ldots, a_k \in \mathbb{R},\ s_1, \ldots, s_k \in I,\ k \geqslant 1 \right\},$$

then $\bar{\mathcal{A}}$, the closure of $\mathcal{A}$ in the supremum metric, is called the reproducing kernel Hilbert space (RKHS) of C_0.

Mercer's theorem (see [11]) allows us to rewrite a Gaussian process as sum of eigenfunctions of the covariance kernel, if this is positive definite, because the series

$$C(t,s) = \sum_{i=1}^{+\infty} \lambda_i \phi_i(t) \phi_i(s),$$

absolutely converges for each (t,s) and uniformly on each compact subset of the closed index set.

Nonparametric Bayesian analysis

In the nonparametric analysis the parameter of interest is infinite dimensional and so what we estimate is the entire function distribution which the data come from. For example, a possible prior on the space of probability measures on a measurable space is the Dirichlet process.

Definition 9. Let A be a positive constant and let G be a probability measure on the space $(\mathcal{X}, \mathcal{B})$. A Dirichlet process on $(\mathcal{X}, \mathcal{B})$ wih parameters (A, G) is a random probability measure P, which assigns the value $P(B)$ to all $B \in \mathcal{B}$, such that $P(B)$ is a measurable random variable, each its realization is a probability measure on $(\mathcal{X}, \mathcal{B})$ and for $\{B_1, \ldots, B_k\}$ the joint distribution of $(P(B_1), \ldots, P(B_k))$ has Dirichlet distribution with parameters $(k, AG(B_1), \ldots, AG(B_k))$.

Rather, if we are interested to a survival distribution function, we may use an independent increment process as prior. In our case, we try to estimate the response probability function $p(x)$ and so we use a Gaussian process that is a prior used, for example, in density estimation or regression function estimation problems.

A condition that must be verified is the consistence. We have a sequence of statistical experiments $\{\mathcal{X}^{(n)}, B^{(n)}, P_\theta^{(n)} : \theta \in \Theta\}$ and $X^{(n)}$ is the observation of the n-th experiment, Θ is a topological space. Let $\mathcal{B}$ be the Borel

sigma-field on Θ and Π_n be a probability measure on $(\Theta, \mathcal{B})$ and it may depend on n. The posterior distribution is denoted by $\Pi_n(\cdot, X^{(n)})$ and is said to be a version of the conditional probability of θ given $X^{(n)}$. The consistence is defined as:

Definition 10. Let $\theta_0 \in \Theta$. The posterior distribution $\Pi_n(\cdot, X^{(n)})$ is consistent at θ_0, with respect to the given topology on Θ if $\Pi_n(\cdot, X^{(n)})$ converges weakly to δ_{θ_0} as $n \to +\infty$ under $P_{\theta_0}^{(n)}$-probability, or a.s. under the distribution induced by θ_0.

Doob obtained a result on posterior consistency with the prior Π fixed and i.i.d. observations. Under measurable conditions on the sample space and model identifiability, he showed that the set of $\theta \in \Theta$ where consistency does not hold is a Π-null set. This follows by the convergence of the martingale $\mathbb{E}(\mathbb{1}(\theta \in B|X_1, \ldots, X_n))$ to $\mathbb{E}(\mathbb{1}(\theta \in B|X_1, X_2, \ldots)) = \mathbb{1}(\theta \in B)$.

A general result on consistency has been obtained by Schwartz. To introduce it, we must define the Kullback-Leibler divergence.

Definition 11. The Kullback-Leibler divergence, denoted as $K(\theta_1, \theta_2)$, is defined as

$$\int_{\mathcal{X}} p(x, \theta_1) \log \frac{p(x, \theta_1)}{p(x, \theta_2)} \, \mathrm{d}\mu(x). \tag{5}$$

We say that $\theta_0 \in \Theta$ is in the Kullback-Leibler support of Π, and we write $\theta_0 \in KL(\Pi)$, if $\forall\, \epsilon > 0$, $\Pi(\theta : K(\theta_0, \theta) < \epsilon) > 0$.

Schwartz's theorem is:

Theorem 12. Let $\theta_0 \in U \subset \Theta$. If there exists $m \geqslant 1$, a test function $\phi(X_1, \ldots, X_m)$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in U^c$ with the property that $\inf(\mathbb{E}_\theta(\phi(X_1, \ldots, X_m)), \theta \in U^c) > \mathbb{E}_{\theta_0}(\phi(X_1, \ldots, X_m))$ and $\theta_0 \in KL(\Pi)$, then $\Pi(\theta \in U^c|X_1, \ldots, X_n) \to 0$ a.s. P_{θ_0}.

For the first condition, it is necessary the existence of an unbiased test for testing the null hypothesis against the alternative one, and this sequence of tests $\phi_n(X_1, \ldots, X_n)$ is such that the probability of the type I error $\mathbb{E}_{\theta_0}(\phi_n(X_1, \ldots, X_n))$ and the probability of type II error $\sup_{\theta \in U^c} \mathbb{E}_\theta(1 - \phi_n(X_1, \ldots, X_n))$ converge to zero exponentially fast.

Binary Regression

We consider the case of binary regression. We have a binary response variable Y and a d-dimensional covariate x belonging to a compact subset $\mathcal{X} \in \mathbb{R}^d$. We want to estimate the response probability $p(x) = \mathbb{P}(Y = 1|x)$. The relation $p(x) = H(\eta(x))$ induces a prior for the function $p(x)$. We consider for η a Gaussian process and H a known cumulative distribution function on $\mathbb{R}$, that is strictly increasing and Lipschitz continuous. The posterior distribution is said to be consistent (see definition 10) if posterior probability of any small neighborhood containing the true parameter value converges to one.

We choose the covariance kernel of the form

$$\sigma(x, x') = \tau^{-1}\sigma_0(\lambda x, \lambda x'), \tag{6}$$

where $\sigma_0(\cdot, \cdot)$ is a nonsingular covariance kernel and $\tau > 0$ and $\lambda > 0$ are two hyper-parameters. We indicate the hyper-priors on τ and λ, respectively, with Π_τ and Π_λ, two absolutely continuous probability measures on $\mathbb{R}^+$.

The space of response probability functions used is

$$\Theta_{n,\alpha} = \{p(\cdot)|p(x) = H(\eta(x)), ||D^w\eta||_\infty < M_n, |w| \leqslant \alpha\}, \tag{7}$$

$D^w\eta$ stands for $(\partial^{|w|}/\partial^{w_1}t_1 \dots \partial^{w_d}t_d)\eta(t_1, \dots, t_d)$, $|w| = \sum_{i=1}^d w_i$, α is a positive integer and M_n is a sequence of real numbers. λ_n and τ_n are chosen such that $\Pi_\lambda(\lambda > \lambda_n) = e^{-cn}$ and $\Pi_\tau(\tau < \tau_n) = e^{-cn}$, with $c > 0$. The possible priors on η are chosen from the reproducing kernel Hilbert space of σ_0 $\mathcal{A}$.

We consider the following assumptions.

A 1. $\forall\, x \in \mathcal{X}$, the covariance function $\sigma_0(x, \cdot)$ has continuous partial derivatives up to order $2\alpha + 2$. α will be properly chosen.

The mean function $\mu(x)$ belongs to $\bar{\mathcal{A}}$.

The support of Π_λ is $\mathbb{R}^+$.

A 2. The covariate space $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$.

A 3. The transformed true response function η_0 belongs to $\bar{\mathcal{A}}$.

A 4. $\forall\, b_1 > 0$ and $b_2 > 0$, there exist sequences M_n, λ_n and τ_n such that

$$M_n^2 \tau_n \lambda_n^{-2\alpha} \geqslant b_1 n \qquad \text{and} \qquad M_n^{d/\alpha} \leqslant b_2 n.$$

We enunciate three theorems about consistency. The first is for the case of covariates arising from a distribution Q on the space of covariates $\mathcal{X}$.

Theorem 13. The random covariate X is sampled from the distribution Q. We assume that Assumptions 1, 2, 3 and 4 hold. Then, $\forall\, \epsilon > 0$,

$$\Pi\left(p : \int_{\mathcal{X}} |p(x) - p_0(x)| \, \mathrm{d}Q(x) > \epsilon \middle| Y_1, \ldots, Y_n, X_1, \ldots, X_n \right) \to 0$$

in P_0^n-probability.

If Q is unknown, we must give a prior on Q but if we assume that p is unrelated to Q and so with independent priors, posterior distributions of p and Q will be independent and we do not need to specify a prior for Q.

The second theorem is for the case of fixed design covariates:

Theorem 14. We assume that Assumptions 1, 2, 3 and 4 hold. Then, $\forall\, \epsilon > 0$,

$$\Pi\left(p : \int_{\mathcal{X}} |p(x) - p_0(x)| \, \mathrm{d}Q_n(x) > \epsilon \middle| Y_1, \ldots, Y_n \right) \to 0$$

in P_0^n-probability.

The third theorem is for one-dimensional covariate and we use assumption 1 and we replace assumptions 2, 3, 4 with the foolowing. Let $x_{i,n}$ be the covariate values in ascending order and let $S_{i,n} = x_{i+1,n} - x_{i,n}$ be the distance between two consecutive covariate values.

A 5. Let $\delta > 0$, there exist K_1 and an integer N such that, for $n > N$, we have that

$$\sum_{i : S_{i,n} > k_1 n^{-1}} S_{i,n} \leqslant \delta.$$

Theorem 15. Suppose that Assumption 5 holds and that $\mathcal{X}$ is a bounded interval of $\mathbb{R}$. Assume that for the prior Assumption 1 holds. We assume that $\eta_0(x)$ and the mean function $\mu(x)$ have continous derivatives on $\mathcal{X}$ up to the second order, that the covariance kernel has continous partial derivatives up to the six order and that Π_λ and Π_τ are such that $\tau_n^{-1}\lambda_n^4 = O(n)$. Then, $\forall\, \epsilon > 0$,

$$\Pi\left(p : \int_{\mathcal{X}} |p(x) - p_0(x)|\, \mathrm{d}x > \epsilon \middle| Y_1, \ldots, Y_n\right) \to 0$$

in P_0^n-probability.

For the proof of Theorem 13 we must use an upper bound for the ϵ-covering number $N(\epsilon, \Theta_n, d_V)$ (defined as the number of ϵ-balls required to cover the space of density $\mathcal{P}$ with respect to the metric d_V) and an estimation of the probability of the complement of the sieve Θ_n defined in (7). The result is obtained by verifying prior positivity and entropy conditions of the general results of Ghosal, Ghosh and Ramamoorthi [9].

We consider a binary response variable Y corresponding to a vector valued covariate X. We want to estimate the response probability function

$$p(x) = \mathbb{P}(Y = 1|X = x) \tag{8}$$

for every x belonging to the covariate space, assuming indipendent observations. Our approach is to consider the model $p(x) = H(\eta(x, \beta))$, where H is a cumulative distribution function called link function and $\eta(x, \beta)$ is generally a nonlinear function depending on the covariate values and an unknown parameter β.

We induce a prior probability on $p(x)$ using a Gaussian process $\eta(x)$ through the relation $p(x) = H(\eta(x))$ where the link function $H : \mathbb{R} \to [0, 1]$ is a smooth cdf. The process mean function indicates where the prior probability is concentrated, while the covariance kernel gives the smoothness of the sample paths of the process. In particularly, we need sample paths to be a dense subset of the space of all continuous function $f : X \to \mathbb{R}$. The link function H maps the image of the Gaussian process into the unit

interval to make it possible to think in probability terms. For H, we use the probit link function and we introduce (see [2]) some latent variables to give a partial conjugacy to our model in order to simplify calculations.

Let $Y = (Y_1, \ldots, Y_n)^t$ be the random vector of our binary response observations corresponding to the covariate value $X = (X_1, \ldots, X_n)^t$, where each X_i has d components. If the observed value of X is $x = (x_1, \ldots, x_n)^t$, then, conditional on X, we assume Y_i's are independent random variables with success probability $p(x_i)$ for a smooth function $p(x)$ that is the function we want to estimate. Let $\mathcal{X}$ be the set of all covariate values. We assume $\mathcal{X}$ is compact.

We define, as prior, a Gaussian process with mean function $\mu(x)$ and covariance kernel $\sigma(x, y)$. If $x_1, \ldots, x_n$ are the covariate values corresponding to the observed variables Y_i's, we define $x'_1, \ldots, x'_k$ as the distinct covariate values and d_i as the number of x'_i's that are present in the observed vector $(x_1, \ldots, x_n)^t$, $\tilde{x} = (x'_1, \ldots, x'_k)^t$, $\tilde{\eta} = (\eta(x'_1), \ldots, \eta(x'_k))^t$, $\tilde{\mu} = (\mu(x'_1), \ldots, \mu(x'_k))^t$ and $\Sigma = (\sigma_{ij})$ the matrix such that $\sigma_{ij} = \sigma(x'_i, x'_j)$. For some x that is different from all x_i's, we have the following theorem:

Theorem 16. The conditional distribution of $\eta(x)$ given $\tilde{\eta}$ is normal with mean $\mu(x) + \sigma(x, \tilde{x})^t \Sigma^{-1}(\tilde{\eta} - \tilde{\mu})$ and variance $\sigma(x, x) - \sigma(x, \tilde{x})^t \Sigma^{-1} \sigma(x, \tilde{x})$ where $\sigma(x, \tilde{x}) = (\sigma(x, x'_1), \ldots, \sigma(x, x'_k))^t$.

Now, let H be the standard normal cdf Φ. We introduce some latent variables as in Albert and Chib [2]. Let $Z = (Z_1, \ldots, Z_n)^t$ be these unobservable latent variables such that, conditional on η, Z_i's are independent normal variables with mean $\eta(x_i)$ and variance 1. We assume Y_i's are functions of these variables, and $Y_i = \mathbb{1}(Z_i > 0)$. Thus, Y_i's, conditional on η, are independent Bernoulli random variables with success probability $\Phi(\eta(x_i))$.

We define $U = (U_1, \ldots, U_k)$ where $U_i = \frac{1}{d_i} \sum_j Z_j$ and Z_j's are such that their corresponding covariate value is x'_i. If D is the diagonal matrix with the i-th diagonal element equal to d_i, then we obtain the following theorem:

Theorem 17. The conditional distribution of $\tilde{\eta}$ given (Z, Y) is a k-variate normal with mean vector $\mu^* = \Sigma^* D(U - \tilde{\mu}) + \tilde{\mu}$ and covariance matrix $\Sigma^* =$

$(D + \Sigma^{-1})^{-1}$, i.e.

$$\tilde{\eta}|(Z, Y) \sim N_k(\mu^*, \Sigma^*). \tag{9}$$

The conditional distributions of the latent variables Z_i's given $(\tilde{\eta}, Y)$ are

$$Z_i|\tilde{\eta}, Y \overset{ind.}{\sim} \begin{cases} N(\eta(x_i), 1)|Z_i < 0 & \text{if } Y_i = 0 \\ N(\eta(x_i), 1)|Z_i > 0 & \text{if } Y_i = 1 \end{cases} \tag{10}$$

In an eventual Gibbs sampler algorithm we can use the conditional distributions (9) e (10) to sample from the joint distribution of $(\tilde{\eta}, Z|Y)$, discarding Z that is not useful for our purpose.

If H is the cdf of a smooth unimodal symmetric density on the entire real line, H can be represented as the scale mixture of standard normal cdf with respect to a cdf G taking values on $(0, +\infty)$. We can introduce two sets of unobservable latent variables: $Z = (Z_1, \ldots, Z_n)^t$ and $V = (V_1, \ldots, V_n)^t$. These variables are such that

$$\begin{aligned} V_i|\tilde{\eta} &\overset{i.i.d.}{\sim} G, \\ Z_i|\tilde{\eta}, V &\overset{ind.}{\sim} N(\eta(x_i), V_i^{-1}) \end{aligned}$$

where, as above, $Y_i = \mathbb{1}(Z_i > 0)$. Defining D_V as the diagonal matrix with i-th diagonal element equals to $d_i v_i$ and if G has Lebesgue density g, we have the following theorem:

Theorem 18. Let μ_V^* and Σ_V^* be, respectively, equal to $\Sigma_V^* D_V(U - \tilde{\mu}) + \tilde{\mu}$ and $(D_V + \Sigma^{-1})^{-1}$. Then

$$\tilde{\eta}|Z, V, Y \sim N_k(\mu_V^*, \Sigma_V^*), \tag{11}$$

$$V_i|Z, \tilde{\eta}, Y \overset{ind.}{\sim} g_i(v) \propto \phi\big((Z_i - \eta(x_i))\sqrt{v}\big)\big)g(v), \tag{12}$$

$$Z_i|V, \tilde{\eta}, Y \sim \begin{cases} N(\eta(x_i), V_i^{-1})|Z_i < 0 & \text{if } Y_i = 0 \\ N(\eta(x_i), V_i^{-1})|Z_i > 0 & \text{if } Y_i = 1 \end{cases} \tag{13}$$

Now, we present the hierarchical model we use for the algorithm. For the mean function we use the parametric form

$$\mu(x, \beta) = \beta_1 \mu_1(x) + \cdots + \beta_m \mu_m(x) \tag{14}$$

where m is fixed, $\beta = (\beta_1, \ldots, \beta_m)^t$ is unknown and $\mu_1, \ldots, \mu_m$ are known functions on $\mathcal{X}$. For the covariance kernel we choose $\sigma(x, x') = \frac{1}{\tau}\,\sigma_0(x, x'; \lambda)$ with σ_0 a known kernel and $\tau > 0,\ \lambda$ two unknown hyper-parameters. Then, the hierarchical model is the following:

$$\begin{aligned}
\tau &\sim Gamma(a, b),\\
\beta|\tau &\sim N_m(\beta_0, \Gamma),\\
\eta|\beta, \tau &\sim Gaussian\ Process(\mu(\cdot, \beta), \sigma_0(x, x'; \lambda)/\tau),\\
Y_i|\tilde{\eta}, \beta, \tau &\sim Bernoulli(\Phi(\eta_i)) \text{ (independent)}.
\end{aligned}$$

We introduce latent variables Z_i's as above and we define $\tilde{\Sigma}_0$ as the $k \times k$ matrix with (i, j)-th element equals to $\sigma_0(x_i', x_j'; \lambda)$ and M as the $k \times m$ matrix with (i, j)-th element equals to $\mu_j(x_i')$.

The conditional distributions $(\tilde{\eta}|\beta, \tau, Z, Y)$ and $(Z|\beta, \tau, \tilde{\eta}, Y)$ are similar to (9) and (10), while the conditional distributions $(\beta|\tau, \tilde{\eta}, Z, Y)$ and $(\tau|\beta, \tilde{\eta}, Z, Y)$ are giving from the following

Theorem 19. Let Γ^*, β_0^*, a^* and b^* be, respectively, equal to $(\tau M^t \tilde{\Sigma}_0^{-1} M + \Gamma^{-1})^{-1}$, $\tau \Gamma^* M^t \tilde{\Sigma}_0^{-1}(\tilde{\eta} - M\beta_0) + \beta_0$, $a + \frac{k}{2}$ and $b + \frac{1}{2}(\tilde{\eta} - M\beta)^t \tilde{\Sigma}_0^{-1}(\tilde{\eta} - M\beta)$. Then

$$\beta|\tau, \tilde{\eta}, Z, Y \sim N_m(\beta_0^*, \Gamma^*), \tag{15}$$

$$\tau|\beta, \tilde{\eta}, Z, Y \sim Gamma(a^*, b^*). \tag{16}$$

We use the conditional distributions (9), (10), (15) and (16) in our algorithm to generate from $(\beta, \tau, \tilde{\eta}, Z|Y)$.

The non-informative choice for the hyper-parameters, that means to choice the matrix Γ^{-1} as the zero matrix and $a = b = 0$, simplifies the parameters β_0^*, a^* and b^* as it follows:

$$\begin{aligned}
\beta_0^* =& (M^t \tilde{\Sigma}_0^{-1} M)^{-1} M^t \tilde{\Sigma}_0^{-1}(\tilde{\eta} - M\beta_0) + \beta_0 =\\
& (M^t \tilde{\Sigma}_0^{-1} M)^{-1} M^t \tilde{\Sigma}_0^{-1} M M^t (M M^t)^{-1} \tilde{\eta} = M^t (M M^t)^{-1} \tilde{\eta},\\
a^* =& \frac{k}{2}\\
b^* =& \frac{1}{2}(\tilde{\eta} - M\beta)^t \tilde{\Sigma}_0^{-1}(\tilde{\eta} - M\beta)
\end{aligned}$$

Since β_0 does not appear we may arbitrary choose it.

For an implementation we consider a dataset on the Donner party that is a group of wagon train emigrants. The dataset donner is included in the LearnBayes package of R and contains the age, the gender and the survival status for 45 members of the party age 15 to 65. We analyze this data set in three cases.

The first is the classical case and we use the regression model $\mathbb{P}(Y_i = 1) = \Phi(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i})$, where $x_{1,i}$ stands for the age of the i-th member and $x_{2,i}$ stands for the gender (1 if male, 0 if female). We estimate the parameters β_0, β_1 and β_2 with the maximum likelihood. We use the R function *glm* with the probit link.

The second case is the parametric Bayesian case. We analyze the data set with a non-informative prior for β. The regression model is the same as in the classical case. The R function *bayes.probit*, contained in the LearnBayes pack, gives a simulated sample from the joint posterior distribution of the regression vector β.

The third is the nonparametric case. The regression model is the hierarchical model previously presented. We use, for β, a multivariate normal prior with mean $(0, 0, 0)$ and covariance matrix $5 \cdot 10^4 I_3$. We use, for τ, a gamma prior with shape parameter 60 and rate parameter 50. We consider, for the process, the mean function and the covariance kernel

$$\mu(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \text{ where } x_2 = \{0, 1\},$$
$$\sigma(x^{(1)}, x^{(2)}) = \frac{1}{\tau} \frac{10}{|x_1^{(1)} - x_1^{(2)}| + |x_2^{(1)} - x_2^{(2)}| + 2}.$$

The Gibbs sampling algorithm samples from the full conditional densities and, descarding the first 1000 values of β, we obtain the two curves of survival probabilities plotted by age. The code is the following:

```
> n<-length(donner$age)
> X<-donner
> X.ord<-X
> for (i in 1:(n-1)){
+ for (j in 1:(n-i)){
+ a<-X.ord[j,]
+ b<-X.ord[j+1,]
+ if (a[1]>b[1]){
+ X.ord[j,]<-b
+ X.ord[j+1,]<-a}
+ if (a[1]==b[1]){
+ if (a[2]<b[2]){
+ X.ord[j,]<-b
+ X.ord[j+1,]<-a}}
+ }}
> d<-c()
> B<-X.ord
> B[1,]<-X.ord[1,]
> i<-2
> y<-matrix(ncol=2,nrow=n)
> k<-n
> for (h in 1:n){
+ d[h]<-1
+ y[h,1]<-i-1
+ while(i<=n && (B[h,1]==X.ord[i,1]) && (B[h,2]==X.ord[i,2])){
+ d[h]<-d[h]+1
+ i<-i+1
+ k<-k-1
+ }
+ y[h,2]<-i-1
+ B[h+1,]<-X.ord[i,]
```

```
+ i<-i+1}
> y<-y[1:k,]
> B1<-matrix(nrow=k,ncol=2)
> for(i in 1:k){
+ for(j in 1:2){
+ B1[i,j]<-B[i,j]}}
> M<-matrix(ncol=3,nrow=k)
> for(i in 1:k){
+ M[i,1]<-1
+ M[i,2]<-B1[i,1]
+ M[i,3]<-B1[i,2]
+ }
> d<-d[1:k]
> D<-diag(d)
> Y<-c()
> Y<-X.ord[,3]
>
> beta0<-c(0,0,0)
> a<-60
> b<-50
> beta.in<-beta.hat2
>
> gamma<-c(.0005,.0005,.0005)
> inv.gamma<-diag(gamma)
>
> sigma0<-function(x1,x2){
> a<-(x1[1]-x2[1])
> a1<-abs(a)
> b<-(x1[2]-x2[2])
> b1<-abs(b)
> return(10/(a1+b1+2))
```

```
> }
>
> sigma0.tilde<-matrix(nrow=k,ncol=k)
> for (i in 1:k){
+ for (j in 1:k){
+ sigma0.tilde[i,j]<-sigma0(B1[i,],B1[j,])
+ }}
> inv.sigma0.tilde<-solve(sigma0.tilde)
>
> mu.tilde.gen<-function(beta,M){
+ mu.tilde<-c()
+ mu.tilde<-beta%*%t(M)
+ return(mu.tilde)}
>
> mu.tilde.in<-mu.tilde.gen(beta.in,M)
> mu0<-mu.tilde.gen(beta0,M)
> eta.tilde.in<-mu.tilde.in
>
> eta.gen<-function(eta.tilde,y){
+ k<-length(eta.tilde)
+ eta<-c()
+ for(i in 1:k){
+ for(j in y[i,1]:y[i,2]){
+ eta[j]<-eta.tilde[i]
+ }}
+ return(eta)}
>
> Z.gen<-function(Y,eta.tilde,n,y){
+ Z<-c()
+ eta<-eta.gen(eta.tilde,y)
+ bb<-pnorm(-eta)
```

```
+ tt=(bb*(1-Y)+(1-bb)*Y)*runif(n)+bb*Y
+ Z<-qnorm(tt)+eta
+ return(Z)}
>
> Z.in<-Z.gen(Y,mu.tilde.in,n,y)
>
> U.gen<-function(Z,d,y){
+ k<-length(d)
+ U<-rep(0,times=k)
+ for(i in 1:k){
+ for(j in y[i,1]:y[i,2]){
+ U[i]<-U[i]+Z[j]}
+ U[i]<-U[i]/d[i]}
+ return(U)}
>
> U.in<-U.gen(Z.in,d,y)
>
> eta.tilde.gen<-function(D,inv.sigma0.tilde,tau,U,mu.tilde){
+ eta.tilde<-mvrnorm(n=1,mu=solve(D+tau*inv.sigma0.tilde)%*%
+ D%*%t(U-mu.tilde),Sigma=solve(D+tau*inv.sigma0.tilde))
+ return(eta.tilde)}
>
> beta.gen<-function(eta.tilde,tau,inv.gamma,beta0,M,
+ inv.sigma0.tilde,mu0){
+ gamma.ast<-solve(tau*t(M)%*%inv.sigma0.tilde%*%M+inv.gamma)
+ beta0.ast<-(tau*gamma.ast%*%t(M)%*%inv.sigma0.tilde%*%
+ t(eta.tilde-mu0))+beta0
+ beta<-mvrnorm(n=1,mu=beta0.ast,Sigma=gamma.ast)
+ return(beta)}
>
> tau.gen<-function(k,eta.tilde,M,beta,inv.sigma0.tilde,a,b){
```

```
+ tau<-rgamma(n=1,shape=a+k/2,rate=b+0.5*((eta.tilde-beta%*%
+ t(M))%*%inv.sigma0.tilde%*%t(eta.tilde-beta%*%t(M))))
+ return(tau)}
>
> tau.in<-a/b
> tau<-tau.in
> mu.tilde<-mu.tilde.in
> U<-U.in
> Z<-Z.in
> beta2<-matrix(nrow=10000,ncol=3)
> for(i in 1:10000){
+ eta.tilde<-c()
+ eta.tilde<-eta.tilde.gen(D,inv.sigma0.tilde,tau,U,mu.tilde)
+ Z<-Z.gen(Y,eta.tilde,n,y)
+ U<-U.gen(Z,d,y)
+ beta2[i,]<-beta.gen(eta.tilde,tau,inv.gamma,beta0,M,
+ inv.sigma0.tilde,mu0)
+ mu.tilde<-mu.tilde.gen(beta2[i,],M)
+ tau<-tau.gen(k,eta.tilde,M,beta2[i,],inv.sigma0.tilde,a,b)}
> summary(beta2)

       V1                 V2                  V3
 Min.   :-10.2786   Min.   :-0.23390   Min.   :-3.8752
 1st Qu.: -1.0704   1st Qu.:-0.06247   1st Qu.:-0.9363
 Median :  0.8371   Median :-0.01960   Median :-0.2791
 Mean   :  0.8149   Mean   :-0.01991   Mean   :-0.2837
 3rd Qu.:  2.6678   3rd Qu.: 0.02250   3rd Qu.: 0.3672
 Max.   : 12.0784   Max.   : 0.21215   Max.   : 3.4516

> par(mfcol=(c(3,2)))
> plot(beta2[,1],type="l",xlab="Iterations",ylab="",
```

```
+ main="Trace of (Intercept)")
> plot(beta2[,2],type="l",xlab="Iterations",ylab="",
+ main="Trace of age")
> plot(beta2[,3],type="l",xlab="Iterations",ylab="",
+ main="Trace of male")
> plot(density(beta2[,1]),ylab="",main="Density of (Intercept)")
> points(beta2[,1],rep(0,times=10000),pch=20)
> plot(density(beta2[,2]),ylab="",main="Density of age")
> points(beta2[,2],rep(0,times=10000),pch=20)
> plot(density(beta2[,3]),ylab="",main="Density of male")
> points(beta2[,3],rep(0,times=10000),pch=20)
>
> beta3<-matrix(ncol=3,nrow=9000)
> for(i in 1:9000){
+ beta3[i,]<-beta2[1000+i,]}
> beta.hat3<-c()
> beta.hat3[1]<-sum(beta3[,1])/9000
> beta.hat3[2]<-sum(beta3[,2])/9000
> beta.hat3[3]<-sum(beta3[,3])/9000
> beta.hat3
[1]  0.81163233 -0.01985297 -0.28357137
> windows()
> curve(pnorm(beta.hat3[1]+beta.hat3[2]*x+beta.hat3[3]),
+ from=15,to=65,ylim=c(0,1),col="blue",xlab="Age",
+ ylab="P(Y=1|X=x)",main="Survival probability")
> curve(pnorm(beta.hat3[1]+beta.hat3[2]*x),from=15,to=65,
+ add=TRUE,ylim=c(0,1),col="red")
> legend(50,0.9,c("male","female"),lty=1,col=c("blue","red"))
>
> quant<-matrix(nrow=3,ncol=2)
> quant2[1,]<-quantile(beta3[,1],probs=c(0.05,0.95))
```

```
> quant2[2,]<-quantile(beta3[,2],probs=c(0.05,0.95))
> quant2[3,]<-quantile(beta3[,3],probs=c(0.05,0.95))
>
> windows()
> curve(pnorm(beta.hat3[1]+beta.hat3[2]*x+beta.hat3[3]),
+ from=15,to=65,ylim=c(0,1),xlab="Age",ylab="P(Y=1|X=x)",
+ col="blue",main="Male survival probability")
> curve(pnorm(quant2[1,1]+quant2[2,1]*x+quant2[3,1]),lty=2,
+ from=15,to=65,col="blue",add=TRUE,ylim=c(0,1))
> curve(pnorm(quant2[1,2]+quant2[2,2]*x+quant2[3,2]),lty=4,
+ col="blue",from=15,to=65,add=TRUE,ylim=c(0,1))
> legend(48,0.8,c("5% quantile","95% quantile","mean"),
+ lty=c(2,4,1),col=c("blue","blue","blue"))
>
> windows()
> curve(pnorm(beta.hat3[1]+beta.hat3[2]*x),from=15,to=65,
+ ylim=c(0,1),xlab="Age",ylab="P(Y=1|X=x)",
+ col="red",main="Female survival probability")
> curve(pnorm(quant2[1,1]+quant2[2,1]*x),lty=2,from=15,
+ to=65,col="red",add=TRUE,ylim=c(0,1))
> curve(pnorm(quant2[1,2]+quant2[2,2]*x),lty=4,from=15,
+ to=65,col="red",add=TRUE,ylim=c(0,1))
> legend(48,0.8,c("5% quantile","95% quantile","mean"),
+ lty=c(2,4,1),col=c("red","red","red"))
```

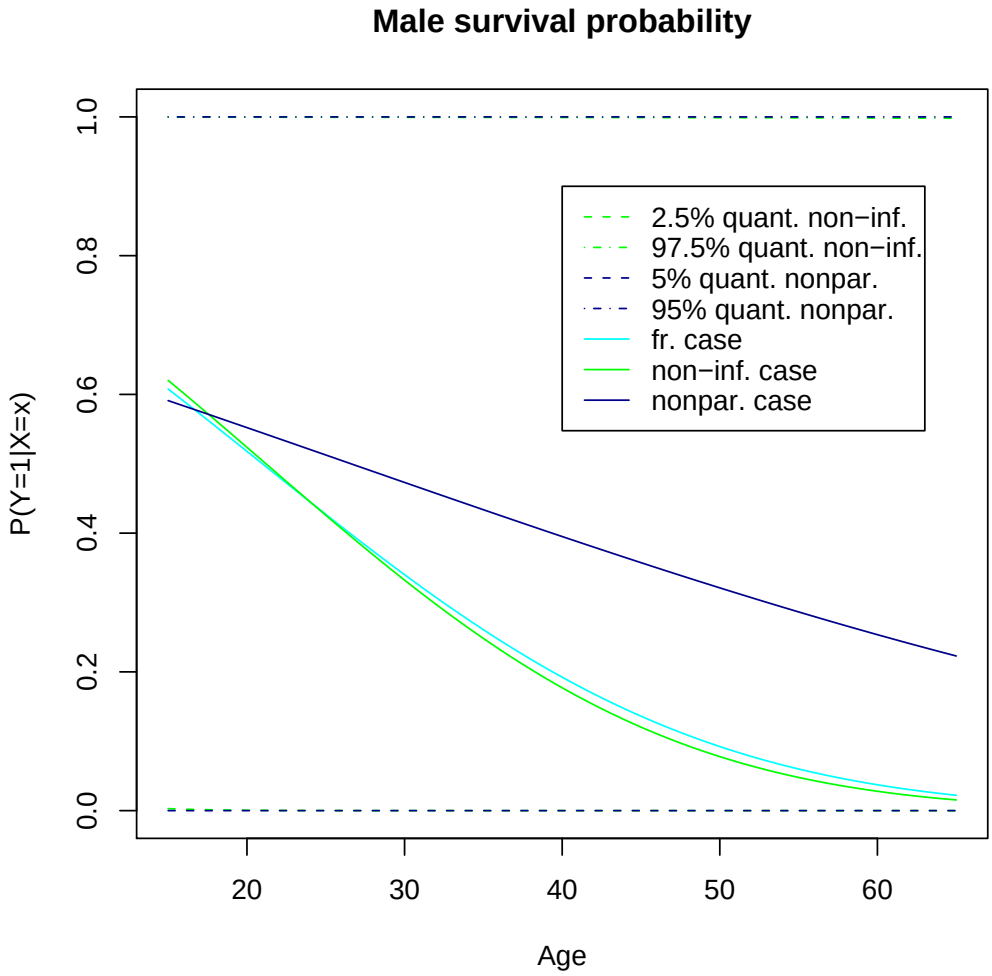


Figure 1: There are the curves for male survival probability obtained with all the algorithms.

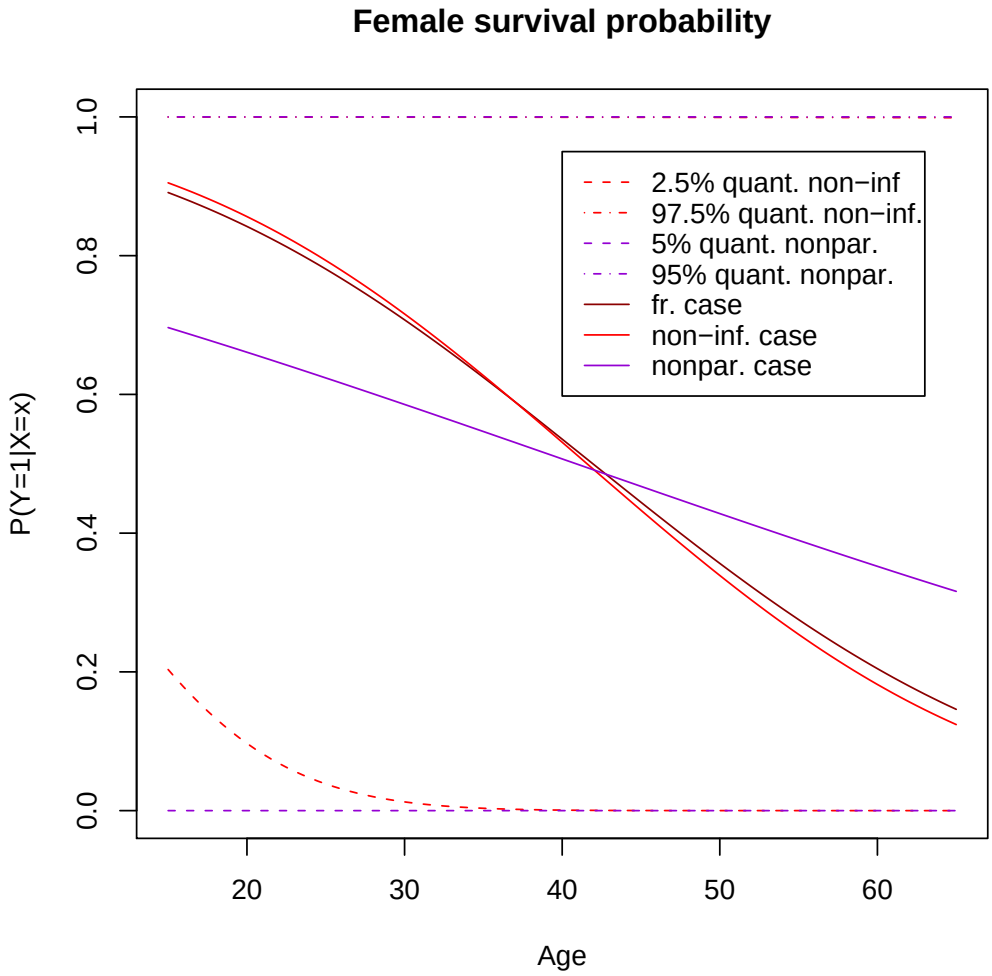


Figure 2: There are the curves for female survival probability obtained with all the algorithms.

Figure 1 shows the male survival probability curves obtained with the three different approaches and the quantile curves for the non-informative case and the nonparametric case. Figure 2 shows the female survival probability curves and the quantile curves for the non-informative case and the nonparametric case. With our choices of covariance kernel and prior distributions, the result obtained with the nonparametric approach differs from both classical case and parametric non-informative case.

Bibliography

[1] P. Abrahamsen. A Review of Gaussian Random Fields and Correlation Functions. Second Edition, April 1997.

[2] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 1993.

[3] A. R. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density function. *Unpublished manuscript*, 1988.

[4] A. R. Barron. New approaches to Bayesian consistency. *The Annals of Statistics*, 2004.

[5] N. Chopin. Jim Albert: Bayesian computation with R. *Statistics and Computing*. Springer, 2007.

[6] N. Choudhuri, S. Ghosal, and A. Roy. Bayesian methods for function estimation. *Handbook of Statistics*, 2005.

[7] N. Choudhuri, S. Ghosal, and A. Roy. Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology*, 2007.

[8] P. Diaconis and D. A. Freedman. Nonparametric binary regression: a Bayesian approach. *The Annals of Statistics*, 1993.

[9] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 1999.

[10] S. Ghosal and A. Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 2006.

[11] H. Q. Minh, P. Niyogi, and Y. Yao. *Mercer's Theorem, Feauture Maps, and Smoothing*, volume 4005/2006 of *Lecture Notes in Computer Science*, pages 154–168. Springer, 2006.

[12] L. Schwartz. On Bayes procedure. *Probability Theory and Related Fields*, 1965.

[13] S. T. Tokdar and J. K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 2007.