Available online at www.sciencedirect.com

ScienceDirect

Statistical Methodology 4 (2007) 227–243

Statistical Methodology
Official Journal of the International Indian Statistical Association

www.elsevier.com/locate/stamet

Nonparametric binary regression using a Gaussian process prior

Nidhan Choudhuri[a], Subhashis Ghosal[b], Anindya Roy[c,*]

[a] Spatial Data Analytics Corporation, 1950 Old Gallows Road, Suite 300, Vienna, VA 22182-3990, United States
[b] Department of Statistics, North Carolina State University, 2501 Founders Drive, Raleigh, NC 27695-8203, United States
[c] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, United States

Received 21 July 2005; received in revised form 23 July 2006; accepted 24 July 2006

Abstract

The article describes a nonparametric Bayesian approach to estimating the regression function for binary response data measured with multiple covariates. A multiparameter Gaussian process, after some transformation, is used as a prior on the regression function. Such a prior does not require any assumptions like monotonicity or additivity of the covariate effects. However, additivity, if desired, may be imposed through the selection of appropriate parameters of the prior. By introducing some latent variables, the conditional distributions in the posterior may be shown to be conjugate, and thus an efficient Gibbs sampler to compute the posterior distribution may be developed. A hierarchical scheme to construct a prior around a parametric family is described. A robustification technique to protect the resulting Bayes estimator against miscoded observations is also designed. A detailed simulation study is conducted to investigate the performance of the proposed methods. We also analyze some real data using the methods developed in this article.

Keywords: Gibbs sampler; Latent variable; Link function; Response probability; Robustification

1. Introduction

Consider a random binary response Y measured with a vector valued covariate X. The problem is to estimate the response probability function

$$p(x) = P(Y = 1 \mid X = x) \tag{1.1}$$

* Corresponding author.
E-mail addresses: nidhan.choudhuri@spadac.com (N. Choudhuri), ghosal@stat.ncsu.edu (S. Ghosal), anindya@math.umbc.edu (A. Roy).

1572-3127/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2006.07.003

over the entire covariate space based on independent observations. The problem commonly occurs in many fields of application, such as medical and spatial statistics. Traditionally, a parametric approach to the specification of the function p is taken using the model $p(x) = H(\alpha + \beta^{\mathrm{T}} x)$, where α and β are unknown parameters and H is a cumulative distribution function (cdf), called the link function. Logistic regression, which is one of the most commonly used statistical methods, chooses the standard logistic cdf as the link function. A close alternative is the probit regression, where H is chosen to be the standard normal cdf Φ. More generally, the linear function could be replaced by some other parametric function $\eta(x, \beta)$, such as a polynomial.

Despite the wide use of logistic regression and other parametric methods, the inference is sensitive to the choices of H and η. Hence it is sensible to avoid these assumptions by taking a nonparametric approach. If multiple responses are available at some x, then $p(x)$ may be estimated by the observed proportion of positive responses. However, sufficiently many responses may not be available at a single x, and thus one needs to "borrow strength" from responses at other x's by imposing either a global condition or a local smoothness condition. Moreover, such conditions are necessary to estimate p at an unobserved x.

A common semiparametric approach to the problem is to specify $\eta(x)$ as a linear function, but to estimate the link function H from the data. Both frequentist and Bayesian methods have been proposed in the literature using this approach. In the Bayesian context, Gelfand and Kuo [6], and Newton et al. [15] used a Dirichlet process prior for H. To obtain a smoother estimate, Mallick and Gelfand [14] modeled H as a mixture of beta cdf's with a prior probability on the mixture weights. Basu and Mukhopadhyay [3] modeled the link function as a Dirichlet scale mixture of truncated normal cdf's. Although, these semiparametric approaches provide a flexible model for selecting the link function, the linearity of η is often restrictive in that the resulting probability function is always monotone and has some specific shape. In many applications, such as when the response to a dose of a drug is studied, monotonicity is a reasonable assumption. However, in some applications, monotonicity may not hold. Even for dose response studies, if toxicity can result from the drug, the response probability can exhibit non-monotonicity. Higher order polynomial terms need to be included in order to incorporate non-monotonicity and interaction effects, which increases the complexity and subjectivity of the modeling.

A completely nonparametric estimate may be obtained by keeping the link function fixed, but modeling $\eta(x)$ as an arbitrary function. This kind of flexible shape modeling can produce any useful shape, especially for spatial data. For instance, this approach can produce arbitrarily shaped equal probability contours. In contrast, the approach of varying H and keeping the form of $\eta(x)$ fixed can produce only specifically shaped equal probability contours. It may be noted that once $\eta(x)$ is allowed to be arbitrary, keeping H fixed leads to no loss of generality in shape. O'Sullivan et al. [17] proposed a penalized likelihood approach, where the estimator for $\eta(x)$ turns out to be a polynomial spline. Gu [9] provided a data driven smoothing parameter for the penalized likelihood estimator via a generalized cross-validation technique. Hastie and Tibshirani [10] proposed a local scoring method for additive models. Tibshirani and Hastie [19] introduced the local likelihood approach which was extended to locally weighted likelihood approach by Staniswalis [18].

In the Bayesian context, Wood and Kohn [22] modeled $\eta(x)$ as an affine transform of the integrated Brownian motion while using the standard normal cdf for the link function. The resulting Bayes estimate of η was found to be a cubic smoothing spline. Posterior expectations were computed by an extended state-space representation of η, and then by providing a Gibbs sampling algorithm for sampling from the posterior distribution of this state-space model. DiMatteo et al. [5] described a free-knot spline approach for likelihood based models, where a piecewise cubic polynomial was fitted to the function of interest, while putting priors on the number of knots, the location of the knots, and the coefficients of the polynomials. Sampling from the posterior distribution was performed by a reversible-jump Markov chain Monte Carlo (MCMC) algorithm. These methods were applied to some real and simulated data, but their application is limited to additive models.

This article describes a Bayesian approach to the nonparametric estimation of $p(x)$. A prior probability on $p(x)$ is induced by using a Gaussian process $\eta(x)$ by the relation $p(x) = H(\eta(x))$, where the link function $H : \mathbb{R} \to [0, 1]$ is a known smooth cdf. Earlier, Leonard [12] and Lenk [11] used Gaussian processes to construct priors in the context of Bayesian density estimation on a bounded interval. Recall that a Gaussian process is a stochastic process such that each finite dimensional distribution is multivariate normal. Thus the Gaussian process $\eta(x)$ is specified by its mean function and covariance kernel. The smoothness of the covariance kernel essentially controls the smoothness of the sample paths of $\eta(x)$. The mean function of the process determines the location of concentration of prior probabilities. The roles of these two functions have some similarities with the roles of the precision parameter and the center measure of a Dirichlet process, which determine the concentration of probabilities and the location of the concentration respectively in that case. However, the Gaussian process is conceptually different from the Dirichlet process in that the former is a prior on the space of real valued functions while the Dirichlet process is a prior on cdfs. In the particular case of binary regression, the Dirichlet process is used as a prior on H while the Gaussian process is used as prior on $\eta(x)$. Consequently, the Dirichlet process and its variations are used for monotone binary regression, while the Gaussian process is used when the shape of the response probability is totally unrestricted. Under certain conditions, Ghosal and Roy [7] have recently established the posterior consistency of the Gaussian process prior for binary regression.

The paper is organized as follows. In the next section, a more precise description of the Gaussian process prior is given. The covariance kernel of the Gaussian process can be chosen so that all possible sample paths of $\eta(x)$ form a dense subset in the space of all real valued continuous functions on the domain space, and hence the prior charges all possible continuous response probability functions. Note that the role of the link function H is to simply map the range space $\mathbb{R}$ of $\eta(x)$ to the unit interval pointwise. In Section 3, we describe Markov chain Monte-Carlo (MCMC) methods based on Gibbs sampling to compute the posterior distribution. We specifically work with the probit link function, because in this case, partial conjugacy in the model is obtained by introducing some latent variables as in Albert and Chib [2]. At the end Section 3, we describe extensions of the Gibbs sampling algorithm to handle a wider variety of link functions. Although completely general sample paths can arise, additive or partially additive models may be obtained as a special case through appropriate prior specification; details are given in Section 4. Through a hierarchical scheme, the prior may be built around any parametric model, simply by choosing a Gaussian process prior with mean function equal to the target parametric family and by putting further priors on the unspecified parameters; details are given in Section 5. In Section 6, we indicate how to make the suggested procedures more robust against possible miscoding of observations. Section 7 describes the findings from a simulation study. In Section 8, we use the proposed method to estimate the response probability as a function of stimulus level in a psychological experiment.

2. Gaussian process prior

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be the random binary observations measured along with the corresponding covariate values $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$, where each X_i has d component variables.

Let the observed values of X be $\boldsymbol{x} = (x_1, \ldots, x_n)$. Conditional on $\boldsymbol{X}$, the Y_i's are independent with probability of success $p(x_i)$ for some smooth function $p(x)$. Let $\mathfrak{X}$ denote the range space of the covariate. Then, the function of interest, $p(x)$, is a smooth map from $\mathfrak{X}$ to [0, 1]. We shall work with the case when $\mathfrak{X}$ is compact. The covariate values, $\boldsymbol{X}$, may arise from a fixed design or may arise randomly from a joint distribution Q, under which the X_i's may possibly be dependent. In the latter case, Q may be viewed as a nuisance parameter. Under the natural assumption that the regression function is unrelated to the distribution of the covariate, the likelihood for p can be separated from that of Q. Thus, with independent priors on p and Q, the posterior distribution of p will be free of Q. As a result, posterior distribution of p may be obtained without even specifying a prior on Q, and the computation will be the same as in the case of fixed covariates.

For nonparametric models, a traditional subjective elicitation of the prior is not feasible due to the vastness of the parameter space. Priors are usually constructed from considerations of mathematical tractability, the feasibility of computation (usually with the help of MCMC methods), and good large sample behavior. The form of the prior is chosen according to some default mechanism, while the key parameters of the prior are chosen to reflect the prior beliefs. A well accepted criterion for the choice of a nonparametric prior is that the prior has a large or full topological support. A Gaussian process on domain $\mathfrak{X}$ is a random real valued function $\eta(x)$ such that all possible finite dimensional distributions $(\eta(x_1), \ldots, \eta(x_n))^{\mathrm{T}}$ are multivariate normal, with $E(\eta(x_i)) = \mu(x_i)$ and $\operatorname{cov}(\eta(x_i), \eta(x_j)) = \sigma(x_i, x_j)$, $i, j = 1, \ldots, k$. The fixed real valued function $\mu(x)$ is known as the mean function, and the function of two variables $\sigma(x, x')$, is known as the covariance kernel, and must satisfy the condition that for every $k \geq 1$ and $x_1, \ldots, x_k \in \mathfrak{X}$, the $k \times k$ matrix $((\sigma(x_i, x_j)))_{i,j=1,\ldots,k}$ is positive definite. Gaussian processes enjoy many interesting properties; see Adler [1], Cramér and Leadbetter [4, Chapter 9], and van der Vaart and Wellner [20, Appendix A.2]. For an appropriate choice of covariance kernel, a Gaussian process has a large support in the space of all smooth functions. A wide variety of functions can arise as the sample paths of the Gaussian process $\eta(x)$. More precisely, the support of a Gaussian process is the reproducing kernel Hilbert space generated by the covariance kernel with a shift by the mean function; see Ghosal and Roy [7, Theorem 4]. For example, the eigenfunctions of the univariate covariance kernel $\sigma(x, x') = \mathrm{e}^{-\gamma(x-x')^2}/\tau$ spans the space of all smooth functions if γ is allowed to vary freely. Large topological support for the prior can be obtained by either putting a prior on γ over the positive half line, or by choosing a relatively large value of γ. Since the range of $\eta(x)$ for any fixed x is the entire real line $\mathbb{R}$, we must use a link function H to map it into the unit interval. Thus, we induce a prior on $p(x)$ by assuming that $\{\eta(x) : x \in \mathfrak{X}\}$ is a Gaussian process with mean function $\mu(x)$ and covariance kernel $\sigma(x, x')$.

There are two functional hyper-parameters in the Gaussian process prior that could be chosen to reflect any prior belief. The mean function $\mu(x)$ reflects the prior guess about $\eta(x)$. The covariance kernel is composed of two components, the variance component $\sigma(x, x)$, and the correlation kernel $r(x, x') = \sigma(x, x')/\sqrt{\sigma(x, x)\sigma(x', x')}$. The variance component may be chosen to be non-constant to reflect different degrees of faith on $\mu(x)$ at different x. The most important parameter is the correlation kernel that controls the local smoothness of the sample paths of $\eta(x)$ and thus determines the extent of strength borrowed from different neighbors in obtaining the posterior process. If sharp changes are expected in some region of the covariate space, then $r(x, x')$ should rapidly decrease to zero as x' moves away from x. If $p(x)$ is expected to be flat in some region, the correlation should decrease slowly. Thus the prior can be locally adaptive. If it is known beforehand that fewer covariate values are expected in some region, the correlation should drop slowly so that strength could be borrowed form distant neighbors. Thus the prior is also covariate adaptive.

Any cumulative distribution function (cdf) H that preserves the smoothness of η may be considered a link function. Although the prior distribution of p depends on the choice of the link function, the posterior distribution of p remains quite insensitive to the particular choice of link function, especially for moderately large and large samples. Nonparametric modeling of η makes the model robust against the choice of the link function. However, the posterior distribution of η depends on H. This is because the posterior distribution of p targets the true response function p_0, and thus the posterior distribution of η targets $H^{-1}(p_0)$, which depends on H. The link function should not be estimated from the data by putting a prior on H, as this, along with an arbitrary specification of η, will lead to identifiability problems. Besides, a fixed link function does not reduce the flexibility of the nonparametric estimation of p. We prefer to use the probit link for certain computational advantages, as discussed later.

Intuitively, in order to recover the entire function $p(x)$, the covariate values must fill up the whole of $\mathfrak{X}$ gradually with increasing sample size. This will happen automatically if the covariate is a random variable arising from a non-singular distribution. For a non-random covariate, the condition appears as a non-trivial restriction unless one measures accuracy by the average performance at the observed covariate values only; see Ghosal and Roy [7] for more details.

3. MCMC algorithm for posterior computation

The posterior distribution of η is not analytically tractable, and thus an MCMC procedure will be used to compute the posterior distribution. As the link functions are nonlinear, the posterior expectation of $\eta(x)$ may not be plugged into the link function to obtain the posterior expectation of $p(x)$. The Monte-Carlo averages of $\eta(x)$ and $p(x)$ have to be calculated separately. One may also compute the pointwise posterior variance and credible bands from these Monte Carlo samples.

We shall describe an MCMC procedure to sample from the joint posterior distribution of η evaluated only at the observed covariate values. The sample from the posterior distribution of the entire function η may then be generated via data augmentation. Since the computation in the case of random covariates remains same as that of fixed covariates, for notational simplicity, we shall omit $\boldsymbol{X}$ from the conditioning variables.

Let $x'_1, \ldots, x'_k$ be the distinct covariate values and d_j be the repetition of the value x'_j. Let $\tilde{\boldsymbol{x}} = (x'_1, \ldots, x'_k)$, $\boldsymbol{\eta} = (\eta(x'_1), \ldots, \eta(x'_k))^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu(x'_1), \ldots, \mu(x'_k))^{\mathrm{T}}$, and $\boldsymbol{\Sigma}$ be the matrix with (i, j)-th element equals to $\sigma(x'_i, x'_j)$. Then for some x that is different from the observed x_i's, the conditional distribution of $\eta(x)$ given $\boldsymbol{\eta}$ is univariate normal with mean $\mu(x) - \sigma(x, \tilde{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu})$ and variance $\sigma(x, x) - \sigma(x, \tilde{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sigma(x, \tilde{\boldsymbol{x}})$, where $\sigma(x, \tilde{\boldsymbol{x}}) = (\sigma(x, x'_1), \ldots, \sigma(x, x'_k))^{\mathrm{T}}$. Since the likelihood involves only $\boldsymbol{\eta}$, the conditional distribution of $\eta(x)$ given $(\boldsymbol{\eta}, \boldsymbol{Y})$ will be the same. Thus, samples from the posterior distribution of $\eta(x)$ may be generated by sampling from this univariate conditional distribution in each step of the Markov chain for $\boldsymbol{\eta}$. This approach may be extended to generate samples from the posterior distribution of the entire function η.

3.1. Probit link

Consider the standard normal cdf Φ as the link function. We shall introduce latent variables as in Albert and Chib [2] to obtain partial conjugacy in the model. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ be some unobservable latent variables such that conditional on η, the Z_i's are independent normal random variables with mean $\eta(x_i)$ and variance 1. Assume that the observations Y_i's are functions of these latent variables defined as $Y_i = I(Z_i > 0)$. Then, conditional on η, the Y_i's are independent Bernoulli random variables with success probability $\Phi(\eta(x_i))$ and thus lead to the probit link model. Had we observed the Z_i's, the posterior distribution of η could be computed analytically by virtue of conjugacy in the Gaussian observation and the Gaussian prior for the mean. Since $\boldsymbol{Z}$ is unobservable, we shall sample from the joint posterior distribution of $(\boldsymbol{Z}, \boldsymbol{\eta})$ via a Gibbs sampler and then discard $\boldsymbol{Z}$.

The prior distribution of $\boldsymbol{\eta}$ is k-variate normal with mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$. Given $\boldsymbol{\eta}$, the Z_i's are independent normal random variables with variance 1 and mean $\eta(x_i)$. Let U_j be the average of the latent variables z_i's, for which the covariate value equals to x'_j and $\boldsymbol{U} = (U_1, \dots, U_k)$. Let $\boldsymbol{D}$ be the diagonal matrix with j-th diagonal element equals to d_j. Then the conditional distribution of $\boldsymbol{\eta}$ given $\boldsymbol{Z}$ is k-variate normal with dispersion matrix $\boldsymbol{\Sigma}^* = (\boldsymbol{D} + \boldsymbol{\Sigma}^{-1})^{-1}$ and mean vector $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*\boldsymbol{D}(\boldsymbol{U} - \boldsymbol{\mu}) + \boldsymbol{\mu}$. Since the Y_i's are deterministic function of the Z_i's, the conditional distribution of $\boldsymbol{\eta}$ given $(\boldsymbol{Z}, \boldsymbol{Y})$ will be the same as that given only $\boldsymbol{Z}$. Thus

$$\boldsymbol{\eta} \mid \boldsymbol{Z}, \boldsymbol{Y} \sim N_k(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*). \tag{3.1}$$

The value of Y_i indicates whether Z_i is negative or positive, and thus

$$Z_i \mid \boldsymbol{\eta}, \boldsymbol{Y} \sim \text{ independent} \begin{cases} N(\eta(x_i), 1) \mid Z_i > 0, & \text{if } Y_i = 1, \\ N(\eta(x_i), 1) \mid Z_i < 0, & \text{if } Y_i = 0. \end{cases} \tag{3.2}$$

Thus the two conditional distributions in (3.1) and (3.2) may be used in a Gibbs sampler to sample from the distribution of $(\boldsymbol{Z}, \boldsymbol{\eta} \mid \boldsymbol{Y})$. We shall start the chain at $\boldsymbol{\eta} = \boldsymbol{\mu}$.

While computing $\boldsymbol{\Sigma}^*$ for large n, the computation of $\boldsymbol{\Sigma}^{-1}$ must be avoided to prevent system overflow, since $\boldsymbol{\Sigma}$ will be a near-singular matrix. This near-singularity problem may also arise in small samples if two or more covariate values are close to each other. A spectral decomposition of the form $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathrm{T}}$ will be helpful here. If the number of observations for each distinct covariate values is equal to c, then $\boldsymbol{\Sigma}^* = \boldsymbol{P}(c\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathrm{T}}$. For unequal numbers of observations in distinct covariate values, we need to simultaneously diagonalise $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$. Consider a spectral decomposition of the matrix $\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{P}\boldsymbol{\Lambda}^{1/2}$ that is of the form $\tilde{\boldsymbol{P}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{P}}^{\mathrm{T}}$. Then $\boldsymbol{\Sigma}^* = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\tilde{\boldsymbol{P}}[\boldsymbol{I} + \tilde{\boldsymbol{\Lambda}}]^{-1}\tilde{\boldsymbol{P}}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^{\mathrm{T}}$.

A direct Gibbs sampler algorithm of componentwise updating $\boldsymbol{\eta}$, that does not introduce latent variables, leads to very slow movements in the Markov chain as the components of $\boldsymbol{\eta}$ are highly correlated. Besides, the conditional posterior distribution of one component given the others does not have any simple form, and thus may require a Metropolis–Hastings type algorithm in each Gibbs loop. However the choice of the proposal distribution becomes difficult due to extremely small conditional variance and high dependence on the conditioning variables. The advantage of this latent variable approach is that the conditional distribution of $\boldsymbol{\eta}$ given $(\boldsymbol{Z}, \boldsymbol{Y})$ is analytically tractable, while sampling the latent variables conditional on $(\boldsymbol{\eta}, \boldsymbol{Y})$ is simple because of their independence and known form.

3.2. Arbitrary unimodal symmetric link

Let the link function H be the cdf having a smooth unimodal symmetric density on the real line. Then H may be represented as the scale mixture of mean zero normal cdf, and hence,

$$H(t) = \int_0^\infty \Phi(t\sqrt{v})\mathrm{d}G(v),$$

for some known cdf G on $(0, \infty)$. Introduce two sets of unobservable latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ and $\boldsymbol{V} = (V_1, \ldots, V_n)^{\mathrm{T}}$ such that

$$\begin{aligned} V_i \mid \boldsymbol{\eta} &\sim \text{ i.i.d. } G, \\ Z_i \mid \boldsymbol{\eta}, \boldsymbol{V} &\sim \text{ independent } N(\eta_i, V_i^{-1}) \end{aligned}$$

and $Y_i = I(Z_i > 0)$. Then, conditional on η, the Y_i's are independent Bernoulli random variables with success probability $H(\eta_i)$.

Let $\boldsymbol{D}_V$ be the diagonal matrix with j-th diagonal element equal to $d_j v_j$, and $\boldsymbol{\Sigma}_V^* = (\boldsymbol{D}_V + \boldsymbol{\Sigma}^{-1})^{-1}$ and $\boldsymbol{\mu}_V^* = \boldsymbol{\Sigma}_V^* \boldsymbol{D}_V(\boldsymbol{U} - \boldsymbol{\mu}) + \boldsymbol{\mu}$. Suppose G has a Lebesgue density or probability mass function g. Then,

$$\boldsymbol{\eta} \mid \boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{Y} \sim N_k(\boldsymbol{\mu}_V^*, \boldsymbol{\Sigma}_V^*), \tag{3.3}$$

$$V_i \mid \boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{Y} \sim g_i(v) \propto \phi([Z_i - \eta_i]\sqrt{v}) g(v), \tag{3.4}$$

$$Z_i \mid \boldsymbol{V}, \boldsymbol{\eta}, \boldsymbol{Y} \sim \begin{cases} N(\eta_i, V_i^{-1}) \mid (Z_i > 0), & \text{if } Y_i = 1, \\ N(\eta_i, V_i^{-1}) \mid (Z_i < 0), & \text{if } Y_i = 0. \end{cases} \tag{3.5}$$

Thus the three conditional distributions in (3.3)–(3.5) may be used in a Gibbs sampler to sample from the distribution of $(\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{\eta} \mid \boldsymbol{Y})$. We shall start the chain with $\boldsymbol{\eta} = \boldsymbol{\mu}$ and $\boldsymbol{V} = (1, \ldots, 1)^{\mathrm{T}}$ and update $\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{V}$ in an order.

If the mixing distribution G is a gamma distribution with parameters α and β, as in the t-link, then the conditional distribution in (3.4) is also a gamma distribution with parameters $(\alpha + 1/2)$ and $[2\beta + (Z_i - \eta_i)^2]/2$. If G is a discrete distribution, then too an exact sample may be drawn from the conditional distribution in (3.4). Even if G is neither discrete nor absolutely continuous, the conditional distribution of $(V_i \mid \boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{Y})$ is absolutely continuous with respect to G, and the corresponding Radon–Nikodym derivative is proportional to $\phi([Z_i - \eta_i]\sqrt{v})$, and thus samples may be drawn via acceptance–rejection sampling.

4. Additive models

Additive regression models are often preferred over non-additive models because the effects of covariates can be examined one at a time. As the combined effect of the covariates is assumed to be the sum of their individual effects, the regression function, after a suitable link transformation, breaks up into a sum of functions of the individual covariates. For an additive binary regression model, $\eta(x)$ is thus assumed to be of the form

$$\eta(x) = \alpha + \eta^{(1)}(x^{(1)}) + \cdots + \eta^{(d)}(x^{(d)}), \tag{4.1}$$

where $x^{(j)}$ is the j-th component of the vector valued covariate, and the $\eta^{(j)}$'s are smooth functions of one variable and α is a scalar. Thus the estimation of η reduces to that of the component functions $\eta^{(j)}$'s and α. However, the $\eta^{(j)}$'s in (4.1) are identifiable up to a constant and restrictions are needed for identifiability. A commonly used restriction is

$$\eta^{(j)}(x_0^{(j)}) = 0, \quad j = 1, \ldots, d, \tag{4.2}$$

for some arbitrary fixed point $x_0 = (x_0^{(1)}, \ldots, x_0^{(d)})$.

The Gaussian process prior described here can incorporate additive models. If the mean function and the covariance kernel of the prior Gaussian process are additive, that is, $\mu(x) = \mu^{(1)}(x^{(1)}) + \cdots + \mu^{(d)}(x^{(d)})$ and $\sigma(x, s) = \sigma^{(1)}(x^{(1)}, s^{(1)}) + \cdots + \sigma^{(d)}(x^{(d)}, s^{(d)})$ for some functions $\mu^{(1)}, \ldots, \mu^{(d)}$ of one variable and some covariance kernel $\sigma^{(1)}, \ldots, \sigma^{(d)}$ on scalar fields. Then the Gaussian process η on $\mathbb{R}^d$ may be represented as

$$\eta(x) = \eta^{(1)}(x^{(1)}) + \cdots + \eta^{(d)}(x^{(d)}) \tag{4.3}$$

for some independent Gaussian processes $(\eta^{(1)}, \ldots, \eta^{(d)})$ on the scalar field. The mean function and the covariance kernel of $\eta^{(j)}$ are $\mu^{(j)}$ and $\sigma^{(j)}$, respectively. Hence the prior sample paths of η are additive. While, the additivity of μ is necessary for additivity in η, the additivity of σ is just a sufficient condition. The process η may be additive even if σ is not, however, then the components functions $\eta^{(j)}$'s will no longer be independent. Note that the space of all continuous additive functions is a closed subspace in the space of all multivariate continuous functions. Thus, a prior concentrated on this subspace will lead to a posterior that is also concentrated in this subspace, and so is the posterior mean. Thus additive choice of the hyper-parameters leads to the desired additive models. One may similarly obtain a partially additive model.

Computations for this additive model may be carried out by estimating η through the posterior mean $\hat{\eta}$ via the Gibbs sampler described in Section 3. Since the posterior mean is an additive function of its component variables, the parameters in (4.1) are estimated as $\hat{\alpha} = \hat{\eta}(x_0)$ and $\hat{\eta}^{(j)}(x^{(j)}) = \hat{\eta}(x_{0,j}) - \hat{\eta}(x_0)$, where $x_{0,j}$ is the d-variate vector obtained by replacing the j-th component of x_0 by $x^{(j)}$. If one desires to obtain the posterior variance or credible bands for these parameters, then MCMC samples are needed from their posterior distributions. This may be done by computing α and the $\eta^{(j)}$'s from η in each Gibbs cycle.

The subjective choice of x_0 makes the identifiability condition in (4.2) artificial. Although the posterior distribution of η is free from the choice of x_0, the component functions $\eta^{(j)}$'s, and hence their estimates, depend on this choice. Besides, the prior variance and hence the posterior variance of $\eta^{(j)}(x^{(j)})$ converge to 0 as $x^{(j)}$ approaches $x_0^{(j)}$ and hence any posterior credible band for $\eta^{(j)}$ shrinks to 0 as we approach this point. Since α is the combined effect of all component covariates at x_0, its credible band does not provide any information about individual covariate effects at that point.

If the covariate space is of the form $\mathfrak{X} = \prod_{j=1}^{d} \mathfrak{X}^{(j)}$, where $\mathfrak{X}^{(j)}$ is the range space of the j-th covariate, then identifiability may also be achieved through a global restriction

$$\int_{\mathfrak{X}^{(j)}} \eta^{(j)}(x^{(j)}) \mathrm{d}x^{(j)} = 0, \quad j = 1, \ldots, d. \tag{4.4}$$

In this case, the scalar parameter $\alpha = |\mathfrak{X}|^{-1} \int_{\mathfrak{X}} \eta$, where $|\mathfrak{X}|$ stands for the Lebesgue measure of $\mathfrak{X}$, represents the overall effect of all the covariates over their entire ranges. The component functions and α may also be calculated from η in each Gibbs cycle, at the expense of additional computation.

A commonly used Bayesian approach to the additive model with restriction (4.2) is to put priors directly on the $\eta^{(j)}$'s and assume that the $\eta^{(j)}$'s and α are a priori independent. Commonly, α is given a normal prior with large variance and $\eta^{(j)}$ is given a Wiener process or integrated Wiener process starting at $x_0^{(j)}$ as its prior. Conditional on the $\eta^{(j)}$'s and $\boldsymbol{Z}$, α is normally distributed. The distribution of $\eta^{(j)}$, conditional on α, $\boldsymbol{Z}$ and the other $\eta^{(l)}$'s, is another Gaussian process similar to (3.1). Thus, a Gibbs sampler may be applied for componentwise updating $(\alpha, \eta^{(1)}, \ldots, \eta^{(d)}, \boldsymbol{Z} \mid \boldsymbol{Y})$. However, this approach cannot be easily applied to the additive models with identifiability restrictions (4.4), since it is hard to define process priors with restrictions on integrals like (4.4). Besides, this approach does not extend easily to partially additive models.

5. Hierarchical models and hyper-priors

The choices of the hyper-parameters μ and σ are critical in the prior elicitation. Hyper-priors may be put on these functional parameters in order to reduce subjectivity. One may consider some parametric forms for them while putting priors on these hyper-parameters. Although one may estimate these hyper-parameters from the data, the parameter of interest is still η.

To build the prior around a parametric family, consider

$$\mu(x; \boldsymbol{\beta}) = \beta_1 \mu^{(1)}(x) + \cdots + \beta_m \mu^{(m)}(x) \tag{5.1}$$

where m is a fixed integer, $(\mu_1, \ldots, \mu_m)$ are known functions on $\mathfrak{X}$ and the scalar hyper-parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta)^{\mathrm{T}}$ are unknown. Such a class of parametric family covers a wide variety of functions with the appropriate choice of the μ_j's. For example, $\mu(x; \boldsymbol{\beta})$ could be a polynomial in x. The nonparametric component comes from the covariance kernel. Consider the simplest parametric form $\sigma(x, x') = \sigma_0(x, x'; \lambda)/\tau$ for some known kernel σ_0 and unknown hyper-parameter $\tau > 0$ and λ. Note that, the posterior mean of $p(x)$ almost interpolates the data as $\tau \to 0$ while the posterior distribution is concentrated near the prior mean function as $\tau \to \infty$. Thus, a hyper-prior on τ helps the data choose the degree of smoothing.

A gamma distribution on τ and an independent m-variate normal distribution on $\boldsymbol{\beta}$ seems to be appropriate as these two lead to partially conjugate posteriors. Thus, we consider the following hierarchical model

$$\begin{aligned}
&\tau \sim \text{ Gamma}(a, b),\\
&\boldsymbol{\beta} \mid \tau \sim N_m(\boldsymbol{\beta}_0, \Gamma),\\
&\eta \mid \boldsymbol{\beta}, \tau \sim \text{ Gaussian process } (\mu(\cdot; \boldsymbol{\beta}), \sigma_0(x, x'; \lambda)/\tau),\\
&Y_i \mid \boldsymbol{\eta}, \boldsymbol{\beta}, \tau \sim \text{ independent Bernoulli } (\Phi(\eta_i)).
\end{aligned}$$

Let us introduce the latent variables $\boldsymbol{Z}$ as in Section 3.1. Let $\boldsymbol{\eta}$ be the vector defined in Section 3. The conditional distributions of $(\boldsymbol{\eta} \mid \boldsymbol{\beta}, \tau, \boldsymbol{Z}, \boldsymbol{Y})$ and $(\boldsymbol{Z} \mid \boldsymbol{\beta}, \tau, \boldsymbol{\eta}, \boldsymbol{Y})$ are similar to (3.1) and (3.2). Let $\boldsymbol{\Sigma}_0$ be the $k \times k$ matrix with (i, j)-th elements equal to $\sigma_0(x'_i, x'_j; \lambda)$ and $\boldsymbol{M}$ be the $k \times m$ matrix with (i, j)-th elements equal to $\mu_j(x'_i)$. Then $\boldsymbol{\eta} \mid \boldsymbol{\beta}, \tau \sim N_k(\boldsymbol{M}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{\Sigma}_0)$. Since $(\boldsymbol{Z}, \boldsymbol{Y})$ affects the distribution of $\boldsymbol{\beta}, \tau$ only through $\boldsymbol{\eta}$, we have

$$\boldsymbol{\beta} \mid \tau, \boldsymbol{\eta}, \boldsymbol{Z}, \boldsymbol{Y} \sim N_m(\boldsymbol{\beta}_0^*, \boldsymbol{\Gamma}^*), \tag{5.2}$$

$$\tau \mid \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{Z}, \boldsymbol{Y} \sim \text{ Gamma}(a^*, b^*), \tag{5.3}$$

where $\boldsymbol{\Gamma}^* = [\tau \boldsymbol{M}^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{M} + \boldsymbol{\Gamma}^{-1}]^{-1}$, $\boldsymbol{\beta}_0^* = \tau \boldsymbol{\Gamma}^* \boldsymbol{M}^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\eta} - \boldsymbol{M}\boldsymbol{\beta}_0) + \boldsymbol{\beta}_0$, $a^* = a + k/2$ and $b^* = b + (\boldsymbol{\eta} - \boldsymbol{M}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\eta} - \boldsymbol{M}\boldsymbol{\beta})/2$. Hence the conditional distributions in (3.1), (3.2), (5.2) and (5.3) may be used in a Gibbs sampler to generate from the distribution of $(\boldsymbol{\beta}, \tau, \boldsymbol{\eta}, \boldsymbol{Z} \mid \boldsymbol{Y})$. One may also use the technique given in Section 3.2 if the link function is different from the normal cdf.

In practice, analysis with a non-informative choice of the prior, which is also a member of the conjugate family in a limiting sense, is most likely to be carried out, since subjective information at the second stage of hierarchy is usually not available. This corresponds to the choice $\boldsymbol{\Gamma}^{-1} = \boldsymbol{0}$, the zero matrix, leading to the improper uniform prior for $\boldsymbol{\beta}$, and $a = b = 0$, which is the

Jeffreys prior τ^{-1} for the rate parameter τ. In this case, the posterior updating formula simplifies to $\boldsymbol{\beta}_0^* \boldsymbol{M}^{-1}\boldsymbol{\eta}$, $a^* = k/2$ and $b^* = (\boldsymbol{\eta} - \boldsymbol{M\beta})^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\eta} - \boldsymbol{M\beta})/2$. Note that the choice of $\boldsymbol{\beta}_0$ does not matter (and hence it can be taken to be the zero vector for definiteness). In all the simulations we perform in Section 7 and the real data analysis in Section 8, we use this non-informative choice.

Although we considered the simplest hierarchical model for the covariance kernel with only one hyper-parameter, the method extends to richer models. Consider a multi-parameter representation of the form

$$\sigma(x, x') = \tau_1^{-1}\sigma_1(x, x') + \cdots + \tau_l^{-1}\sigma_l(x, x') \tag{5.4}$$

for some fixed integer l and some known kernels $\sigma_1, \ldots, \sigma_l$ such that the reproducing kernel Hilbert space generated by them are pair-wise orthogonal. This is a very useful hierarchical representation in an additive model, where σ_j is a covariance kernel for the component function of the j-th covariate. However the model (5.4) is more general. In such cases, the functional parameter η may be represented as

$$\eta(x) = \mu(x; \boldsymbol{\beta}) + \eta_1(x) + \cdots + \eta_l(x)$$

for some independent Gaussian process $\eta_1, \ldots, \eta_l$ with mean function equal to 0 and covariance kernel equal to σ_j/τ_j. Put independent gamma priors on the τ_j's. Then the conditional distribution of η_j, given the other η_t's, τ_j, $\boldsymbol{Z}$, $\boldsymbol{\beta}$ and $\boldsymbol{Y}$, will be another Gaussian process, while the conditional distribution of τ_j will depend only on η_j.

6. Robustification

In this section, we describe how to robustify our procedure against miscoding of the response variable. Following an idea of Verdinelli and Wasserman [21], we introduce indicator variables $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ such that $\psi_i = 1$ indicates that Y_i is miscoded and $\psi_i = 0$ indicates that Y_i is correctly coded. Had we observed these indicator variables, the Bayesian analysis could be carried through by adjusting the Y_i's, that is, by changing Y_i to $(1 - Y_i)$ for the miscoded data. Since these variables are unobservable, we treat them as unknown parameters and put priors on them. The joint posterior distribution of $(\boldsymbol{\psi}, \eta)$ is then used to obtain a robust estimation of η, and also to identify the miscoded observations. Note that this approach treats any outlier in the response variable as a miscoded data point and thus is also robust against outliers. However, the procedure cannot identify whether an observation is miscoded or is an outlier.

We assume that a priori each observation has equal probability of being miscoded or being an outlier, is independent of other observations being miscoded, and is also independent of η. Let r be our a priori guess for the probability of an observation being miscoded. If no prior information is available, we set r to be a small number between 0.01 and 0.1.

Observe that, conditional on $(\boldsymbol{\psi}, \eta)$, the Y_i's are independent Bernoulli random variables with probability of success $[\{1 - \psi_i\}H(\eta(x_i)) + \psi_i\{1 - H(\eta(x_i))\}]$. Hence, conditional on $(\boldsymbol{Y}, \eta)$, the ψ_i's are independent with

$$P(\psi_i = 1 \mid \boldsymbol{Y}, \eta) = \begin{cases} \dfrac{r[1 - H(\eta(x_i))]}{r[1 - H(\eta(x_i))] + (1 - r)H(\eta(x_i))}, & \text{if } Y_i = 1, \\ \dfrac{rH(\eta(x_i))}{rH(\eta(x_i)) + (1 - r)[1 - H(\eta(x_i))]}, & \text{if } Y_i = 0. \end{cases} \tag{6.1}$$

First consider the probit link without any hyper-prior. Introduce latent variables $\boldsymbol{Z}$ as in Section 3.1 with the adjustment for miscoding, that is, $Y_i = 1$ if $\{Z_i > 0, \psi_i = 0\}$ or

$\{Z_i < 0, \psi_i = 1\}$. Then

$$Z_i \mid \boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{Y} \sim \begin{cases} N(\eta_i, 1)I(Z_i > 0), & \text{if } Y_i + \psi_i = 1, \\ N(\eta_i, 1)I(Z_i < 0), & \text{if } Y_i + \psi_i \neq 1. \end{cases} \tag{6.2}$$

Thus samples from the joint distribution of $(\psi_i, Z_i \mid \eta, \boldsymbol{Y})$ may be drawn by first sampling ψ_i using (6.1) and then sampling Z_i using (6.2). The distribution of $\boldsymbol{\eta}$ given $\boldsymbol{Z}$ does not depend on $\boldsymbol{Y}$ or $\boldsymbol{\psi}$. Thus $(\boldsymbol{\eta} \mid \boldsymbol{Z}, \boldsymbol{\psi}, \boldsymbol{Y})$ is the same as in (3.1). Hence a Gibbs sampler for sampling from the joint posterior distribution of $(\boldsymbol{\psi}, \boldsymbol{Z}, \eta \mid \boldsymbol{Y})$ may be described by using the conditional distributions $(\boldsymbol{\psi}, \boldsymbol{Z} \mid \boldsymbol{\eta}, \boldsymbol{Y})$ and $(\boldsymbol{\eta} \mid \boldsymbol{\psi}, \boldsymbol{Z}, \boldsymbol{Y})$. In the next section, we shall also present a robustified version of one of the simulation studies carried out there.

The algorithm may be extended similarly for an arbitrary symmetric link by introducing the latent variables $\boldsymbol{V}$ as in Section 3.2. For a hierarchical model as in Section 5, the conditional distributions of the hyper-parameters given the latent variables, do not depend on $\boldsymbol{\psi}$ as well as the joint distribution of $\boldsymbol{\psi}$, and the latent variables given the functional parameter η do not depend on the hyper-parameters. Hence this robust approach easily extends to hierarchical models.

7. Simulation study

Some simulation studies are performed to evaluate the performance of the proposed estimator. Priors are described through the hierarchical models in Section 5. The proposed method is then compared with the local likelihood estimator (LLE) [13, Chapter 4]. The `C` programming language is used to write the code for the Bayesian method while the `S+` function `locfit()` is used for LLE. The bandwidth parameter in LLE is automatically chosen by `locfit()` using a cross validation method.

In the case of a single covariate, two different response probability functions are considered:

$$p(x) = \frac{e^{6x-2}}{1 + e^{6x-2}}, \tag{7.1}$$

$$p(x) = 3.6x(1 - x) + 0.05. \tag{7.2}$$

The range space of the covariate is the unit interval [0, 1]. The response probability is monotone in (7.1), and is unimodal in (7.2). Four different sample sizes, $n = 50$, 100, 200 and 500 are considered for both the models. For each sample size, we consider both equally spaced fixed design points for the covariate values and covariate values arising from a random uniform distribution. We consider 1000 Monte-Carlo replicates for each of these 16 situations.

We consider two prior distributions for η, and they differ only in modeling the mean functionals of the Gaussian process. In the first case, we consider a linear expression of the form

$$\mu(x; \beta) = \beta_1 + \beta_2 x, \tag{7.3}$$

while in the other, we consider a quadratic expression of the form

$$\mu(x; \beta) = \beta_1 + \beta_2 x + \beta_3 x^2. \tag{7.4}$$

A priori the β_l's are given the improper uniform prior. We choose a covariance kernel of the form $\sigma(x, x') = \exp\{-10(x - x')^2\}/\tau$ for both cases, while τ is a priori distributed with improper density τ^{-1} and is independent of the β_l's. The probit link is used in both cases. For a given sample, the posterior mean is used as an estimate of the regression function p, while the posterior

Table 1
Mean L_1-error from 1000 Monte Carlo replicates for a single covariate

	Fixed covariates				Random covariates			
	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
Monotone $p(x)$								
BE1	0.0814	0.0516	0.0397	0.0253	0.0821	0.0559	0.0402	0.0255
BE2	0.0817	0.0517	0.0397	0.0253	0.0824	0.0560	0.0402	0.0255
LLE	0.0892	0.0560	0.0413	0.0261	0.0899	0.0601	0.0421	0.0262
Unimodal $p(x)$								
BE1	0.1143	0.0718	0.0499	0.0315	0.1139	0.0720	0.0501	0.0318
BE2	0.0944	0.0645	0.0425	0.0305	0.0951	0.0652	0.0473	0.0310
LLE	0.1021	0.0682	0.0484	0.0313	0.1023	0.0691	0.0490	0.0317

BE1 = Bayes estimator with a linear model for μ, BE2 = Bayes estimator with a quadratic model for μ, LLE = local likelihood estimator.

mean is computed through 20 000 MCMC samples collected after a burn-in period of 4000. The estimate is evaluated at 101 grid points equally spaced between [0, 1]. If the program needs to be coded in `WinBUGS`, one will need to approximate improper priors by proper ones, since `WinBUGS` does not allow the former. One may, for instance, put $N(0, 10^6)$ prior on the β_l's with small a and b, say $a = 10^{-3}$, $b = 10^{-6}$. We actually ran some trial simulations with this diffuse proper prior, but did not detect any difference with the results corresponding to the improper prior.

To measure the overall error in estimation for an estimator $\hat{p}$, we considered the integrated absolute error (IAE) or L_1-error defined as $\int_{\mathfrak{X}} |\hat{p}(x) - p(x)| \mathrm{d}x$. For each of the Monte Carlo replicates, we computed two different Bayes estimators using the two priors and then computed the L_1-error. The average L_1-errors from these 1000 replicates are reported in Table 1. The average L_1-errors for the local likelihood estimator (LLE) are also computed using 1000 Monte-Carlo replicates, and are presented in Table 1.

For the monotone regression function in (7.1), the two Bayes estimators both outperform the LLE for all sample sizes and for both the random and fixed design schemes. For the unimodal regression function in (7.2), the Bayes estimator with a quadratic model for μ (BE2) outperforms the LLE, while the other Bayes estimator with only a linear expression for μ (BE1) is only slightly inferior to the LLE. The true regression function, after the probit transformation, is not quadratic although it is unimodal. Thus, a quadratic model for μ, along with the nonparametric component in the Gaussian process prior, is expected to perform well. Even though BE1 models μ only by a linear term, it also performs quite well. Thus the proposed Bayes procedure is robust with respect to the specification of the mean function of the Gaussian process in a hierarchical model. However, some prior knowledge in this regard may enhance the performance of the resulting estimators.

We investigated the convergence properties of the MCMC chains as well as that of the procedure. To see the effect of the initial values for the hyper-parameters a and b, we started 4 MCMC chains with different initial values corresponding to 4 different combinations of high and low values of a and b, i.e., ($a = 0.1, b = 0.1$), ($a = 0.1, b = 100$), ($a = 10, b = 0.1$) and ($a = 10, b = 100$). We tracked the value of the estimated probability at $x = 0.5$ for the monotone model (7.1) with fixed covariates. Fig. 1 shows the plot of the logit of the estimated value of $p(0.5)$ over 6000 posterior runs for the 4 different chains. The chains seem to have smaller variability after reaching stationarity and seems to mix well.

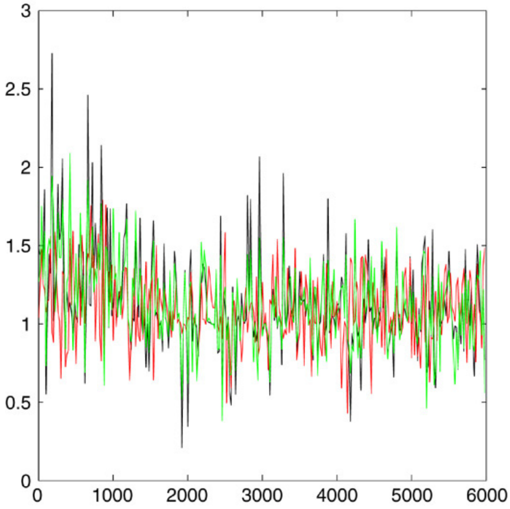


Fig. 1. Mixing properties of MCMC chains for different starting values of hyperparameters.

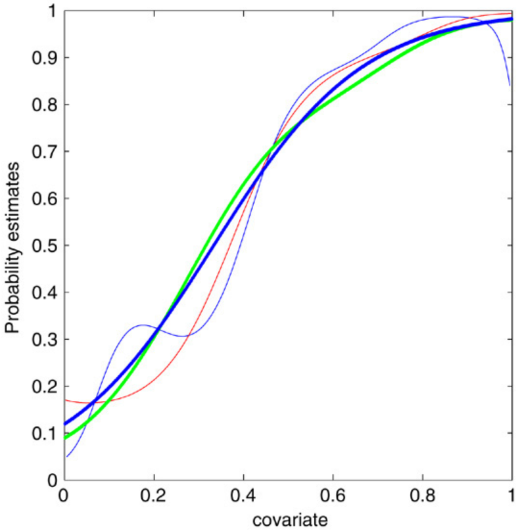


Fig. 2. Convergence of estimator to the true function with increasing sample sizes; $n = 100$, 200 and 1000. True function is in bold.

We also investigated convergence of the nonparametric Bayesian estimator to the true function as the number of design points increased. We used the monotone model (7.1) with fixed design covariates. Fig. 2 shows the estimated graph for $n = 100$, 200 and $n = 1000$. The true function is one of the bold graphs, and the other bold graph is for $n = 1000$. The figure indicates that the nonparametric estimator can be expected to be consistent, a result proven in Ghosal and Roy [7].

Table 2
Mean L_1-error from 1000 Monte Carlo replicates with three covariates

	Fixed covariates				Random covariates			
	$n = 64$	$n = 125$	$n = 216$	$n = 512$	$n = 64$	$n = 125$	$n = 216$	$n = 512$
Additive $p(x)$								
BE	0.2176	0.1271	0.0913	0.0576	0.2281	0.1344	0.0957	0.0596
LLE	0.2258	0.1306	0.0939	0.0577	0.2361	0.1373	0.0976	0.0596
Nonadditive $p(x)$								
BE	0.2401	0.1465	0.1076	0.0675	0.2510	0.1532	0.1103	0.0693
LLE	0.2497	0.1517	0.1114	0.0698	0.2603	0.1591	0.1150	0.0717

BE = Bayes estimator, LLE = local likelihood estimator.

We also consider a simulation study with three covariates. Two different regression functions are considered:

$$p(x) = \Phi(3x^{(1)} + 2x^{(2)}x^{(2)} + 4x^{(3)}(1 - x^{(3)}) - 3) \tag{7.5}$$

$$p(x) = \frac{1}{2}x^{(1)}x^{(2)}x^{(2)} + x^{(3)}(1 - x^{(3)}) + \frac{1}{8}. \tag{7.6}$$

The range space of the covariate is the unit cube $[0, 1]^3$. The regression function in (7.5) is additive in the probit model, while the one in (7.6) is non-additive. Four different sample sizes, $n = 64$, 125, 216 and 512 are considered, and for each sample size, we consider both fixed design and random covariate. The random covariate scheme consists of i.i.d. samples from the uniform distribution on the unit cube, while the fixed design is chosen on an equally spaced k^3 grid. Sample sizes are thus chosen as k^3. We obtain 1000 Monte Carlo replicates for each of these 16 situations.

For the prior on η, we consider a mixed quadratic expression for μ as

$$\mu(x; \boldsymbol{\beta}) = \beta_0 + \sum_j \beta_j x^{(j)} + \sum_{j \le l} \beta_{j,l} x^{(j)} x^{(l)}.$$

As before, the prior distribution on the β's is taken to be the improper Lebesgue measure. We choose a covariance kernel of the form $\sigma(x, x') = \exp\{-10\|x - x'\|^2\}/\tau$, while τ is again given an improper prior density τ^{-1}, independent of the β's. The probit link is used in both cases. The posterior is computed at 20^3 equally spaced grids through 20 000 MCMC samples, collected after a burn-in period of 4000.

The average L_1-errors based on 1000 replicates are presented for the Bayes estimator (BE) and the LLE in Table 2. In all the situations, the Bayes estimator (BE) outperforms the LLE. Note that we do not need to make any assumption on the additivity of the covariate effects even in this case as the true regression function is additive in the probit model. However, any such belief may be reflected through the additive models described in Section 4.

Below, we study how the robustification technique works for the first simulation scheme presented in this section. Table 3 gives the integrated L_1-error for the monotone response model (7.1), and for fixed equi-spaced design covariates under different combinations of sample size and probability of miscoding r. The Bayesian estimator, denoted as BE, uses a linear parametric form of the mean function $\mu(x)$, where each coefficient of the polynomial is given a diffuse prior; see the next section for further description. The robust Bayesian estimator, denoted by RBE, uses the same prior at each step and a robustification step in addition. The degree of miscoding is varied

Table 3
Mean L_1-error from 1000 Monte Carlo replicates for a single covariate

	$n = 100$			$n = 200$		
	$r = 0.00$	$r = 0.05$	$r = 0.20$	$r = 0.00$	$r = 0.05$	$r = 0.20$
BE1	0.0516	0.0751	0.1467	0.0397	0.0605	0.1317
RBE	0.0687	0.0790	0.1402	0.0577	0.0601	0.1185

BE = Bayes estimator, RBE = Robustified version of BE.

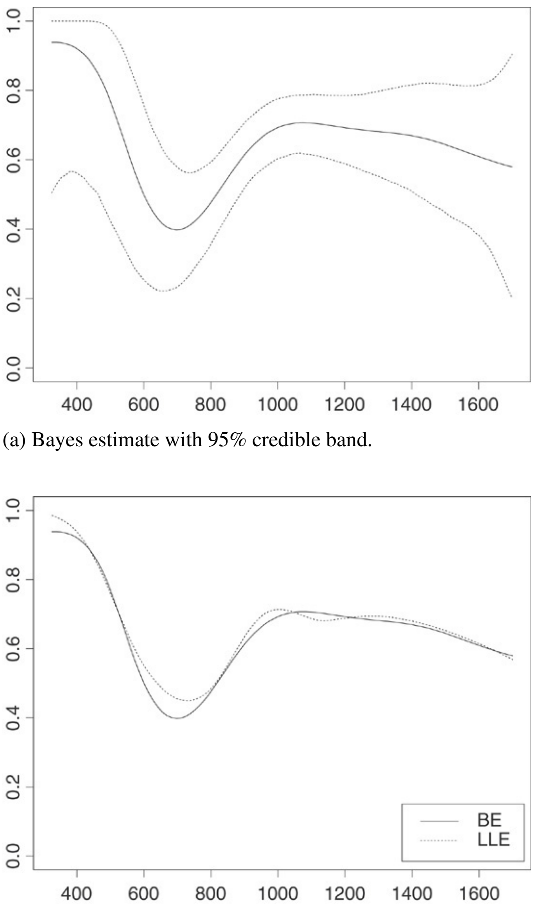
from none ($r = 0.00$) to moderate ($r = 0.05$) to severe ($r = 0.20$), and the sample sizes are $n = 100$ and $n = 200$. Robustification does not seem to have any effect positive effect when the degree of miscoding is small, but it does seem beneficial when there is a high miscoding probability. In all the examples, the prior value for r was set to 0.10.

Although most of the simulations are performed with a probit link, we also considered a small scale simulation with the logit and the t-link. The estimate for $p(x)$ for larger samples remains nearly the same for the different link functions. For smaller sample sizes, although the estimate varies from one link to another, the average performance remains very similar, and thus the simulation results are not presented here.

8. Application to a psychological experiment

We apply our methodology to the data from a psychophysical experiment presented in Grill et al. [8]. The data is on subject's ability to consciously recognize a kinesthetic stimulus. There were 11 subjects. Each subject went through a series of trials, where in each trial a kinesthetic stimulus was given 1000 ms after the beginning of the trial, and a click was generated by a computer randomly within 1–2000 ms from the start of the trial. The subjects were asked to identify the order in which the click and the stimulus arrived. The response was recorded by monitoring the movement of the metacarpal joint of the subject's finger. The binary response was whether the subject was able to correctly identify if the click came first or the stimulus. The response curve was the probability function of the level of the stimulus ranging from 250 to 2000 at intervals of 25. Rosenberger and Grill [16] describe the experiment in more detail. Some of the subjects exhibited a non-monotone response function. Non-monotone behavior can be attributed to greater variability of the response proportions at lower stimulus levels for severely neurologically impaired individuals. We use the 296 observations of subject 8 to estimate the response curve. Subject 8 is reported to be severely neurologically impaired.

We consider a hierarchical prior with a quadratic model for μ as described in (7.4), and a covariance with the scaling in x fit in the unit interval. Fig. 1(a) shows the nonparametric Bayes estimate of the response function. The pointwise 95% credible band is computed as the interval between the 2.5-th and 97.5-th percentile of the 20 000 MCMC samples and is also plotted along with the estimate. The plot shows a non-monotonic behavior of the regression function with a big dip in the response probability for moderately low values of the stimulus level, and a downturn in the response probability for higher stimulus values. Rosenberger and Grill [16] reported that the observations associated with the low stimulus values may be suspect. Thus, even though we plot estimates of the entire response function, we will use only the part associated with moderately high values of stimulus level for our illustration. The credible band is wider on the edges due to the edge-effect, as strength may only be borrowed from neighbors on one side. We also computed the LLE with the bandwidth parameter chosen automatically by `locfit()`. The plot in Fig. 3(b)



(a) Bayes estimate with 95% credible band.

(b) Local likelihood and Bayes estimates.

Fig. 3. Estimates of the response probability for the psychophysical experiment data.

presents the nonparametric Bayesian estimate along with the LLE. The estimates are similar, but the nonparametric Bayesian estimate appears to be slightly smoother.

Acknowledgement

The research of the second author is partially supported by NSF grant number DMS-0349111.

References

[1] R.J. Adler, An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes, in: IMS Lecture Notes-Monograph Series, vol. 12, Institute of Mathematical Statistics, Hayward, CA, 1990.

[2] J. Albert, S. Chib, Bayesian analysis of binary and polychotomous response. Data, J. Amer. Statist. Assoc. 88 (1993) 669–679.

[3] S. Basu, S. Mukhopadhyay, Bayesian analysis of binary regression using symmetric and asymmetric links, Sankhyā Ser. B 62 (2000) 372–387.

[4] H. Cramér, M.R. Leadbetter, Stationary and Related Stochastic Processes. Sample Function Properties and their Applications, John Wiley and Sons, New York, 1967.

[5] I. DiMatteo, C.R. Genovese, R.E. Kass, Bayesian curve-fitting with free-knot splines, Biometrika 88 (2001) 1055–1071.

[6] A.E. Gelfand, L. Kuo, Nonparametric Bayesian bioassay including ordered polytomous response, Biometrika 78 (1991) 657–666.

[7] S. Ghosal, A. Roy, Posterior consistency of Gaussian process prior for nonparametric binary regression, Ann. Statist. 34 (5) 2006 (in press).

[8] S.E. Grill, W.F. Rosenberger, K. Boyle, M. Cannon, M. Hallett, Perception of timing of kinesthetic stimuli, Neuro Report 9 (1998) 4001–4005.

[9] C. Gu, Adaptive spline smoothing in non-Gaussian regression models, J. Amer. Statist. Assoc. 85 (1990) 801–807.

[10] T.J. Hastie, R.J. Tibshirani, Generalized additive models: Some applications, J. Amer. Statist. Assoc. 82 (1987) 371–386.

[11] P.J. Lenk, The logistic normal distribution for Bayesian, nonparametric, predictive densities, J. Amer. Statist. Assoc. 83 (1988) 509–516.

[12] T. Leonard, Density estimation, stochastic processes, and prior information, J. Roy. Statist. Soc. Ser. B 40 (1978) 113–146.

[13] C. Loader, Local Regression and Likelihood, Springer-Verlag, New York, 1999.

[14] B.K. Mallick, A.E. Gelfand, Generalized linear models with unknown link functions, Biometrika 81 (1994) 237–245.

[15] M.A. Newton, C. Czado, R. Chappell, Bayesian inference for semiparametric binary regression, J. Amer. Statist. Assoc. 91 (1996) 142–153.

[16] W.F. Rosenberger, S.E. Grill, A sequential design for psychophysical experiment: An application to estimating timing of sensory events, Stat. Med. 16 (1997) 2245–2260.

[17] F. O'Sullivan, B.S. Yandell, W.J. Raynor, Automatic smoothing of regression functions in generalized linear models, J. Amer. Statist. Assoc. 81 (1986) 96–103.

[18] J.G. Staniswalis, The kernel estimate of a regression function in likelihood-based models, J. Amer. Statist. Assoc. 84 (1989) 276–283.

[19] R. Tibshirani, T. Hastie, Local likelihood estimation, J. Amer. Statist. Assoc. 82 (1987) 559–567.

[20] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer-Verlag, New York, 1996.

[21] I. Verdinelli, L. Wasserman, Bayesian analysis of outlier problems using the Gibbs sampler, Stat. Comput. 1 (1991) 105–117.

[22] S. Wood, R. Kohn, A Bayesian approach to robust binary nonparametric regression, J. Roy. Statist. Soc. Ser. B 93 (1998) 203–213.