Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2010, Article ID 201489, 14 pages
doi:10.1155/2010/201489

Research Article

Spiked Dirichlet Process Priors for Gaussian Process Models

Terrance Savitsky[1,2] and Marina Vannucci[1]

[1] *Department of Statistics, Rice University, Houston, TX 77030, USA*
[2] *Statistics group, RAND Corporation, Santa Monica, CA 90407, USA*

Correspondence should be addressed to Marina Vannucci, marina@rice.edu

Received 27 December 2009; Revised 19 August 2010; Accepted 5 October 2010

Academic Editor: Ishwar Basawa

We expand a framework for Bayesian variable selection for Gaussian process (GP) models by employing spiked Dirichlet process (DP) prior constructions over set partitions containing covariates. Our approach results in a nonparametric treatment of the distribution of the covariance parameters of the GP covariance matrix that in turn induces a clustering of the covariates. We evaluate two prior constructions: the first one employs a mixture of a point-mass and a continuous distribution as the centering distribution for the DP prior, therefore, clustering all covariates. The second one employs a mixture of a spike and a DP prior with a continuous distribution as the centering distribution, which induces clustering of the selected covariates only. DP models borrow information across covariates through model-based clustering. Our simulation results, in particular, show a reduction in posterior sampling variability and, in turn, enhanced prediction performances. In our model formulations, we accomplish posterior inference by employing novel combinations and extensions of existing algorithms for inference with DP prior models and compare performances under the two prior constructions.

1. Introduction

In this paper, we expand a framework for Bayesian variable selection for Gaussian process (GP) models by employing spiked Dirichlet process (DP) prior constructions over set partitions containing covariates. Savitsky et al. [1] incorporate Gaussian processes in the generalized linear model framework of McCullagh and Nelder [2] by expanding the flexibility for the response surface to lie in the space of continuous functions. Their modeling approach results in a class of nonparametric regression models where the covariance matrix depends on the predictors. GP models, in particular, accommodate high-dimensional heterogenous covariate spaces where covariates possess different degrees of linear and nonlinear association to the response, Rasmussen and Williams [3].

In this paper, we investigate mixture prior models that induce a nonparametric treatment of the distribution of the covariance parameters of the GP covariance matrix that, in turn, induces a clustering of the covariates. Mixture priors that employ a spike at zero are now routinely used for variable selection—see for example, George and McCulloch [4] and Brown et al. [5] for univariate and multivariate regression settings, respectively, and Sha et al. [6] for probit models—and have been particularly successful in applications to high-dimensional settings. These approaches employ mixture prior formulations for the regression coefficients that impose an a priori multiplicity penalty, as argued by Scott and Berger [7]. More recently, MacLehose et al. [8] have proposed a Bayesian nonparametric approach to the univariate linear model that incorporates mixture priors containing a Dirac measure component into the DP construction of Ferguson [9] and Antoniak [10]. Dunson et al. [11] use a similar spiked centering distribution in a logistic regression. As noted by these authors, DP models borrow information across covariates through model-based clustering, possibly induced through a spatial or temporal correlation, and can achieve improved variable selection and prediction performances with respect to models that use mixture priors alone. Within the modeling settings of MacLehose et al. [8] and Dunson et al. [11], the clustering induced by the Dirichlet process is on the univariate regression coefficients and strength is borrowed across covariates. Kim et al. [12] have adapted the DP modeling framework to provide meaningful posterior probabilities of sharp hypotheses on the coefficients of a random effects model. Their goal is not necessarily variable selection, but rather the more general problem of testing hypotheses on the model parameters. Their model formulation does not share information across covariates but rather clusters vectors of regression coefficients across observations.

While the prior constructions described above all use a mixture of a point mass and a continuous distribution as the centering distribution of the DP prior, in this paper we also investigate constructions that employ a mixture of a spike and a DP prior with a continuous distribution as the centering distribution. The former approach clusters all covariates, the latter induces clustering of the selected covariates only. The prior formulations we adopt show a reduction in posterior sampling variability with enhanced prediction performances in cases of covariates that express nearly the same association to the response.

In our model formulations, we accomplish posterior inference by employing novel combinations and extensions of existing algorithms for inference with DP prior models and variable selection. Unlike prior constructions for linear models, which are able to marginalize over the model space indicators and directly sample the model coefficients a posteriori, our non-linear modeling frameworks employ nonconjugate priors. We achieve robust selection results by using set partitions on which we impose a DP prior to enclose both the model and the associated parameter spaces. We optimize performances of posterior sampling with a modification of the auxiliary Gibbs algorithm of Neal [13] that accounts for a trivial cluster containing nuisance covariates. We investigate performances on simulated data under the two prior constructions. Our DP prior model constructions represent generalized nonconjugate formulations with associated posterior sampling algorithms that, while specific to GP models, may be applied to other nonconjugate settings.

The remainder of the paper is structured as follows: GP models and covariance matrix formulations are reviewed in Section 2. Section 3 introduces our spiked DP prior formulations, including separate models to cluster all covariates and only selected covariates. Sampling schemes for posterior inference are described in Section 4. Simulations are conducted in Section 5, where we compare the clustering construction to the mixture priors model and also compare the two DP prior model formulations. A benchmark dataset is analysed in Section 6. Concluding remarks are offered in Section 7.

2. Generalized Gaussian Process Models

Savitsky et al. [1] incorporate GP models within the generalized linear model framework of McCullagh and Nelder [2] by employing the relation

$$g(\eta_i) = z(\mathbf{x}_i), \quad i = 1, \ldots, n \tag{2.1}$$

for link function $g(\cdot)$, where η_i is the (possibly latent) canonical parameter for the ith observation. A Gaussian process prior is then specified on the $n \times 1$ latent vector

$$\mathbf{z}(\mathbf{X}) = ((z(\mathbf{x}_1), \ldots, z(\mathbf{x}_n))' \sim N(\mathbf{0}, \mathbf{C}), \tag{2.2}$$

with the $n \times n$ covariance matrix $\mathbf{C}$ being an arbitrarily complex function of the predictors. This general construction extends to latent regression models used for continuous, categorical, and count data, obtained by choosing the appropriate link in (2.1), and incorporates in particular the univariate regression and classification contributions of Neal [14] and Linkletter et al. [15]. Continuous data regression models, for example, are obtained by choosing the identity link function in (2.1) to obtain

$$\mathbf{y} = \mathbf{z}(\mathbf{X}) + \epsilon \tag{2.3}$$

with being $\mathbf{y}$ the $n \times 1$ observed response vector and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, (1/r)\mathbf{I}_n)$ with being r a precision parameter. For inference $\mathbf{z}(\mathbf{X})$ can be integrated out to work with a marginalized likelihood.

Savitsky et al. [1] extend GP models to also include the class of proportional hazard models of Cox [16] by defining the hazard rate as $h(t_i \mid z(\mathbf{x}_i)) = h_0(t_i) \exp(z(\mathbf{x}_i))$ where $h_0(\cdot)$ is the baseline hazard function, t is the failure time, and $z(\mathbf{x}_i)$ is defined as in (2.2). Let the triples $(t_1, \mathbf{x}_1, d_1), \ldots, (t_n, \mathbf{x}_n, d_n)$ indicate the data, with censoring index $d_i = 0$ if the observation is right censored and $d_i = 1$ if the associated failure time, t_i, is observed. Suppose that there are no ties among event/failure times and let $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ be the $D \leq n$ distinct noncensored failure times. In this paper, we use the partial likelihood formulation, defined as

$$\pi(\mathbf{t}, \mathbf{d} \mid \mathbf{z}(\mathbf{X})) = \prod_{i=1}^{D} \frac{\exp(z(\mathbf{x}_{(i)}))}{A_{(i)}} = \prod_{j=1}^{n} \left[\frac{\exp(z(\mathbf{x}_j))}{A_j} \right]^{d_j}, \tag{2.4}$$

where $A_j = \sum_{l \in R(t_j)} \exp(z(x_l) + \epsilon_l)$, with $R(t_j)$ being the set of individuals at risk right before t_j and $A_{(i)}$ the A_j evaluated at the ith failure time. The use of the partial likelihood conveniently avoids prior specification of the baseline hazard.

2.1. Covariance Formulation

A common choice for $\mathbf{C}$ in (2.2) is a covariance function that includes a constant term and an exponential term, that is,

$$\mathbf{C} = \mathrm{Cov}(\mathbf{z}(\mathbf{X})) = \frac{1}{\lambda_a}\mathbf{J}_n + \frac{1}{\lambda_z}\exp(-\mathbf{G}), \tag{2.5}$$

with $\mathbf{J}_n$ being an $n \times n$ matrix of 1's and $\mathbf{G}$ a matrix with elements $g_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{P} (\mathbf{x}_i - \mathbf{x}_j)$ and $\mathbf{P} = \text{diag}(-\log(\rho_1, \ldots, \rho_p))$, with $\rho_k \in [0, 1]$ associated to variable x_k, $k = 1, \ldots, p$. This single-term exponential covariance provides a parsimonious representation that enables a broad class of linear and non-linear response surfaces. In particular, Rasmussen and Williams [3] show how the exponential form (2.5) can be derived from a linear construction by expanding the inputs, x_j's, into an infinite basis. The chosen parametrization allows simple prior specifications and it is also used by Linkletter et al. [15] as a transformation of the exponential term used by Neal [14] and Sacks et al. [17] in their covariance matrix formulations. This construction is sensitive to scaling and we find best results by normalizing the predictors to lie in the unit cube $[0, 1]$. Other choices of $\mathbf{C}$, such as exponential constructions that include a second exponential term and Matern constructions, are reviewed in Savitsky et al. [1].

3. Spiked Dirichlet Process Prior Models for Variable Selection

Variable selection can be achieved in the GP modeling framework with covariance matrix of type (2.5) by imposing mixture priors on the covariance parameters, that is,

$$\pi(\rho_k \mid \gamma_k) = \gamma_k \mathbb{I}[0 \leq \rho_k \leq 1] + (1 - \gamma_k) \delta_1(\rho_k), \tag{3.1}$$

for ρ_k, $k = 1, \ldots, p$, which employs a $\mathcal{U}(0, 1)$ prior for $\rho_k \mid \gamma_k = 1$ and a $\delta_1(\cdot)$, that is, a point mass distribution at one, for $\gamma_k = 0$. This formulation is similar in spirit to the use of mixture priors employed for variable selection in linear regression models, as, for example, in George and McCulloch [4] and Brown et al. [5] for univariate and multivariate regression settings, respectively.

In this paper, we embed mixture priors for variable selection into Dirichlet process prior models that cluster covariates to strengthen selection. The Dirichlet process (DP) construction of Ferguson [9] and Antoniak [10] is a typical choice for a prior on an unknown distribution, G. In particular, given a set of *a priori* i.i.d. parameters, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, with $\phi_i \sim G$, we define the DP prior on $G \sim \text{DP}(\alpha, G_0)$, where G_0 is the parametric base measure defining the prior mean, $\text{E}(G) = G_0$. The concentration parameter, α, expresses the prior confidence in the base measure. Draws from G are discrete a.s., implying a positive probability of ties to instantiate randomly generated partitions. Indeed, many contributions in nonparametric Bayesian inference are formulated in terms of random partition models, that is, probability models that cluster the set of experimental units. See Quintana [18] for a nice review of nonparametric Bayesian models.

Here we introduce probability distributions on set partitions with a particular focus on clustering the p covariates (through the *a priori* i.i.d. covariance parameters, $\boldsymbol{\phi}$), rather than the usual choice of n i.i.d. observations. Let $\boldsymbol{\phi}^* = (\phi_1^*, \ldots, \phi_M^*)'$, for $M \leq p$, define the unique values of $\boldsymbol{\phi}$, and let us define the clusters as $S_\ell = \{k : \phi_k = \phi_\ell^*\}$. Let $\mathcal{F}$ indicate the space of all possible partitions of the p covariates. The partition $\nu_p = \{S_1, \ldots, S_M\} \in \mathcal{F}$ captures a particular disjoint clustering of the covariates, with $S_k \cap S_m = \emptyset$ for $k \neq m$, such that we recover the full set of covariates in the disjoint union, $\bigcup_{k=1}^{p} S_k = S_0 = \{1, \ldots, p\}$. The DP provides the Pólya urn scheme of Blackwell and MacQueen [19] by marginalizing over G to define a joint

prior construction for a particular partition,

$$\pi(\nu_p) = \frac{\prod_{S\in\nu_p}\alpha\Gamma(p_S)}{\prod_{k=1}^{p}(\alpha+k-1)}, \tag{3.2}$$

where $\Gamma(x)$ is the gamma function and p_S the number of covariates in cluster S. Higher values of α tend to produce a larger number of clusters. This is evident if we factorize the joint prior as

$$\pi(s_k = s \mid \mathbf{s}_{-k}) = \begin{cases} \dfrac{p_{-k,s}}{p-1+\alpha} & \text{if } 1 \le s \le M^{-}, \\ \dfrac{\alpha}{p-1+\alpha} & \text{if } s = M^{-}+1, \end{cases} \tag{3.3}$$

where we introduce cluster indicators, $s_k = \ell \Rightarrow k \in S_\ell$, $k = 1,\ldots,p$ and employ exchangeability for ϕ_k to treat covariate k as the last one added. Here $p_{-k,s\neq s_t}$ indicates the number of covariates, excluding covariate k, allocated to the nontrivial cluster S. Similarly, M^{-}, captures the total number of clusters when excluding covariate k. In particular, this construction of the conditional prior reveals that the probability for covariate k to be clustered with m is uniform for all k or $\pi(s_k = s_m \mid \mathbf{s}_{-k}) \propto 1$ for $m = 1,\ldots,k-1,k+1,\ldots,p$. We complete the prior specification with $\alpha \sim \mathcal{G}(a_\alpha, b_\alpha)$ to allow the data to update the concentration parameter for a fully Bayesian approach. It is important to note that our prior construction is over set partitions that contain covariates and that all the observations are in every cluster. We next develop two specific and alternative prior formulations, the first permits clustering on all—trivial and selected—covariates and the second one focuses on clustering only the selected covariates.

3.1. Clustering All Covariates

The first prior construction we consider employs the mixture prior as the centering distribution for the DP prior model, therefore, clustering all covariates. Let us consider the covariance parameters, $\phi = (\phi_1,\ldots,\phi_p)$ with $\boldsymbol{\phi}_k = (\gamma_k, \rho_k)$, for $k = 1,\ldots,p$. We proceed with the usual DP prior construction

$$\begin{aligned} \phi_1,\ldots,\phi_p \mid G &\sim G, \\ G &\sim \mathrm{DP}(\alpha, G_0), \\ G_0 &= \left[\gamma_k \mathcal{U}(0,1) + (1-\gamma_k)\delta_1(\rho_k)\right] \times \mathrm{Bern}(w) \end{aligned} \tag{3.4}$$

which encloses the mixture prior on $\rho_k \mid \gamma_k$ and the Bernoulli prior on γ_k in the base distribution, and where w is the prior probability of covariate inclusion. A further Beta prior can be placed on w to introduce additional variation. Under model sparsity, we *a priori* expect most covariates to be excluded from the model space, which we accomplish by allocating the associated ρ_k for a nuisance covariate to the Dirac measure component of the conditional mixture prior under the setting $(\gamma_k = 0, \rho_k = 1)$, effectively reducing the dimensionality of

the parameter space. Our clustering model (3.4) therefore strengthens exclusion of nuisance covariates with a prior construction that co-clusters nuisance covariates. Let us define the trivial cluster as $S_t = \{k : \phi_k = [\phi_t^* = (\gamma_t^* = 0, \rho_t^* = 1)]\}$, with the star symbol indicating unique values. The trivial cluster can be extracted into a separate line in the conditional prior formulation over the set partitions from (3.3),

$$\pi(s_k = s \mid \mathbf{s}_{-k}) = \begin{cases} \dfrac{p_{-k,s \neq s_t}}{p - 1 + \alpha} & \text{if } 1 \leq s \neq s_t \leq M^-, \\ \dfrac{p_{-k,s=s_t}}{p - 1 + \alpha} & \text{if } s = s_t, \\ \dfrac{\alpha}{p - 1 + \alpha} & \text{if } s = M^- + 1, \end{cases} \tag{3.5}$$

Notice how prior (3.5) strengthens selection by aggregating all trivial covariates into a single cluster. Later in the paper we will employ a data augmentation approach to conduct posterior samples from this nonconjugate model formulation.

3.2. Clustering Selected Covariates

Alternatively, we can use prior models that employ a mixture of a spike and a DP prior with a continuous distribution as the centering distribution, therefore, inducing clustering of the selected covariates only. We construct this model as

$$\begin{aligned} \rho_k \mid \gamma_k &\sim \gamma_k G + (1 - \gamma_k)\delta_1(\cdot), \\ G &\sim \mathrm{DP}(\alpha, G_0), \\ G_0 &\sim \mathcal{U}(0, 1), \end{aligned} \tag{3.6}$$

which may be written more intuitively as $(\rho_k \mid \gamma_k = 1, G) \sim G$. This formulation confines the set partitions to cluster only the selected covariates. While the dimension of the selected covariates, p_γ, will change at every iteration of the MCMC algorithm for posterior inference, we may still marginalize over G, given p_γ, to produce the Pólya urn prior formulation (3.3) where we set $\boldsymbol{\phi} = \{\rho_k \mid \gamma_k = 1\}$,

$$\pi(s_k = s \mid \mathbf{s}_{-k}) = \begin{cases} \dfrac{p_{-k,s}}{p_\gamma - 1 + \alpha} & \text{if } 1 \leq s \leq M^-, \\ \dfrac{\alpha}{p_\gamma - 1 + \alpha} & \text{if } s = M^- + 1. \end{cases} \tag{3.7}$$

We note that the normalizing expression in the denominator now uses p_γ, rather than p, to account for our reduced clustering set. Trivial covariates are not clustered together so that we *a priori* expect reduced efficacy to remove trivial covariates from the model space. In other words, this prior construction produces a relatively flatter prior for assignment to nontrivial clusters under model sparsity as compared to (3.5). Yet, we expect improvement in computational speed as we are only clustering $p_\gamma \leq p$ covariates.

4. Markov Chain Monte Carlo Methods

We accomplish posterior inference by employing novel combinations and extensions of existing algorithms for inference with DP models and variable selection. In particular, we adapt the auxiliary Gibbs algorithm of Neal [13] to data augmentation schemes that draw posterior samples for our nonconjugate model formulations.

First we extend our notation and use $\boldsymbol{\rho}_\gamma$ to indicate the $\boldsymbol{\rho}$ vector where $\rho_k = 1$ when $\gamma_k = 0$, for $k = 1, \ldots, p$. For model (3.5), we collect the covariance parameters in $\boldsymbol{\Theta} = (\mathbf{s}, \boldsymbol{\phi}^*, \lambda_a, \lambda_z)$ where $\boldsymbol{\phi}^* = (\gamma^*, \rho^*_{\gamma^*})$ indicates the M unique cluster values of $\boldsymbol{\phi} = (\gamma, \rho_\gamma)$. For model (3.7), we have $\boldsymbol{\Theta} = (\mathbf{s}, \gamma, \boldsymbol{\rho}^*_\gamma, \lambda_a, \lambda_z)$ to reflect the fact that covariate selection is done separately from the clustering.

We recall the augmented data likelihood of Savitsky et al. [1] employed as a generalized formulation for GP models to construct the joint posterior distribution over model parameters,

$$\pi(\mathbf{D} \mid \boldsymbol{\Theta}, \mathbf{h}), \tag{4.1}$$

with $D_i \in \{y_i, \{t_i, d_i, z(\mathbf{x}_i)\}\}$ and $\mathbf{D} := \{D_1, \ldots, D_n\}$ to capture the observed data *augmented* by the unobserved GP variate, $\mathbf{z}(\mathbf{X})$, in the case of latent responses, such as for the survival model (2.4). Note that $\mathbf{D}$ depends on the concentration parameter for clustering covariates, α, through the prior on $\mathbf{s}$. We collect all other model-specific parameters in $\mathbf{h}$; for example, $\mathbf{h} = r$ for the univariate regression model (2.3).

4.1. Clustering All Covariates

We first define our sampling algorithm using the covariate clustering construction (3.5) which includes all covariates—both trivial and selected. Our MCMC algorithm sequentially samples $\mathbf{s}, \boldsymbol{\phi}^*, \alpha, \lambda_a, \lambda_z, \mathbf{h}$ in a Gibbs-type fashion. We improve efficiency of the auxiliary Gibbs algorithm of Neal [13] used to sample $\mathbf{s}$ by making a modification that avoids duplicate draws of the trivial cluster. The sampling scheme we propose is as follows.

(1) Update **s:** The auxiliary Gibbs algorithm of Neal [13] achieves sampling of the cluster indicators by introducing temporary auxiliary parameters typically generated from the base distribution (3.4). While multiple draws of nontrivial cluster are almost surely unique, repeated draws of the trivial cluster are entirely duplicative. We make a modification to the auxiliary Gibbs algorithm by ensuring that our state space always contains the trivial cluster, therefore, avoiding duplicate generations. The algorithm employs a tuning parameter, ω, as the number of temporary auxiliary parameters to be generated from the prior to facilitate sampling each s_k at every MCMC iteration. We begin by drawing the ω auxiliary parameters from the conditional prior given the current state space values. If $\nexists \ell \in \{1, \ldots, p\} : s_l = s_t$, then one of possibly multiple auxiliary parameters has a connection to the trivial state. We thus sample $\phi^*_{M^-+1} \sim \delta_0(\gamma^*_k)\delta_1(\rho^*_k)$, which draws this value as the trivial cluster. If, however, $\exists \ell \in \{1, \ldots, p\} : s_l = s_t$, then the auxiliary parameters are independent of the trivial state and are sampled as nontrivial clusters from $\delta_1(\gamma_k)\mathbb{I}[0 \leq \rho_k \leq 1]$, as in the original auxiliary Gibbs algorithm. Next, we draw the cluster indicator, s_k, in a Gibbs Step from the

conditional posterior over the set partitions with a state that includes our auxiliary parameters,

$$\pi(s_k = s \mid \mathbf{s}_{-k}, \mathbf{D}) \propto \begin{cases} \dfrac{p_{-k,s\neq s_t}}{p-1+\alpha}\pi(\mathbf{D} \mid \boldsymbol{\Theta}, \mathbf{h}) & \text{if } 1 \leq s \neq s_t \leq M^-, \\ \dfrac{p_{-k,s=s_t}}{p-1+\alpha}\pi(\mathbf{D} \mid \boldsymbol{\Theta}, \mathbf{h}) & \text{if } s = s_t, \\ \dfrac{[\alpha^*/w]}{p-1+\alpha}\pi(\mathbf{D} \mid \boldsymbol{\Theta}, \mathbf{h}) & \text{if } M^- < s \leq M^- + w, \end{cases} \tag{4.2}$$

where we abbreviate (4.1) with $\pi(\mathbf{D} \mid \phi_s^*)$ with $\phi_s^* \in \boldsymbol{\phi}^*$, the unique parameter associated to cluster index, s. In the examples below, we use $\omega = 3$, and therefore a probability to assign a covariate to each of the new clusters as proportional to $[\alpha/w]$. Neal [13] notes that larger values of ω produce posterior draws of lower autocorrelation.

(2) Update ϕ^***:** We update the cluster parameters, $\boldsymbol{\phi}^* = (\phi_1^*, \ldots, \phi_M^*), M \leq p$, using a Metropolis-within-Gibbs. This scheme consists of 2 moves: a between-models move to jointly update (γ_k^*, ρ_k^*) for $k = 1, \ldots, M$ in a componentwise fashion, and a within model move to update ρ_k^* for covariates in the current model after the between-models move. We use uniform proposals for the ρ_k^*s. Under our clustering formulation, we update the M clusters, and not the p covariates, therefore, borrowing strength among coclustered covariates.

(3) Update α**:** We employ the two-step Gibbs sampling algorithm of Escobar and West [20] constructed as a posterior mixture of two Gamma distributions with the mixture component, η, drawn from a beta distribution. The algorithm is facilitated by the conditional independence of α from $\mathbf{D}$, given $\mathbf{s}$.

(4) Update $\{\lambda_a, \lambda_z, \mathrm{h}\}$: These are updated using Metropolis-Hastings moves. Proposals are generated from the Gamma distributions centered at the previously sampled values.

4.2. Clustering Selected Covariates

We describe the steps to perform updates for $\mathbf{s}, \gamma, \boldsymbol{\rho}_\gamma^*, \alpha, \lambda_a, \lambda_z, \mathbf{h}$ from (3.7).

(1) Update $\mathbf{s}$: Obtain draws in a Gibbs Step for $\mathbf{s} = (s_1, \ldots, s_{p_\gamma})$, employing the auxiliary Gibbs algorithm in the usual way according to the conditional posterior

$$\pi(s_k = s \mid \mathbf{s}_{-k}, \mathbf{D}) = \begin{cases} \dfrac{p_{-k,s}}{p_\gamma - 1 + \alpha}\pi(\mathbf{D} \mid \boldsymbol{\Theta}, \mathbf{h}) & \text{if } 1 \leq s \leq M^-, \\ \dfrac{\alpha}{p_\gamma - 1 + \alpha}\pi(\mathbf{D} \mid \boldsymbol{\Theta}, \mathbf{h}) & \text{if } M^- < s \leq M^- + w. \end{cases} \tag{4.3}$$

(2) Update $\boldsymbol{\gamma}$: We update γ componentwise by employing a Metropolis-within-Gibbs algorithm. Notice that an update that adds a covariate also adds a cluster and, similarly, the removal of a covariate also discards a cluster in the case where the cluster contains only the deleted covariate.

(3) Update $\boldsymbol{\rho}_{\gamma}^{*}$: We update the unique ρ_k's values for the previously selected p_γ covariates in a componentwise fashion using a Metropolis-within-Gibbs algorithm and uniform proposals.

(4) Update α: As in Step 3. of the previous algorithm.

(5) Update $\{\lambda_a, \lambda_z, \mathbf{h}\}$: As in Step 4. of the previous algorithm.

For both MCMC schemes, final selection of the variables is accomplished by employing a cutoff value for the marginal posterior probabilities of inclusion of single variables based on a target expected False Discovery Rate (EFDR) in a fashion similar to Newton et al. [21]. For example, let ξ_k be the posterior probability of the event $\gamma_k = 1$, that is, a significant association of the kth predictor to the response. We fix α, a prespecified false discovery rate, and select those covariates with posterior probabilities of exclusion under the null hypothesis, $1 - \xi_k$, that are below threshold, κ, that is,

$$\alpha = \mathrm{EFDR}(\kappa) = \sum_{k=1}^{p} \frac{(1-\xi_k)\mathbb{I}_{(1-\xi_k \le \kappa)}}{\mathbb{I}_{(1-\xi_k \le \kappa)}}, \tag{4.4}$$

with $\mathbb{I}_{(\cdot)}$ the indicator function. As noted by Kim et al. [12], the optimal posterior threshold, κ, may be determined as the maximum value in the set $\{\kappa : \mathrm{EFDR}(\kappa) \le \alpha\}$.

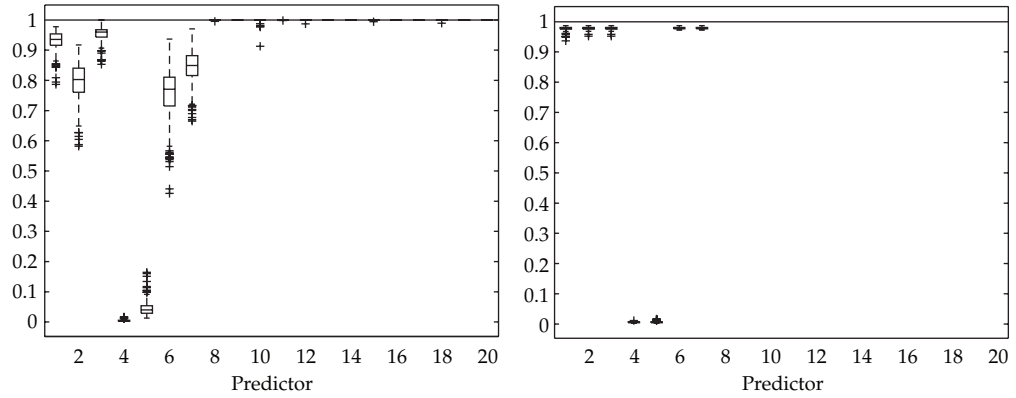
5. Simulation Study

We explore performances of the proposed models on simulated data and on a benchmark dataset. Results will show that the application of DP priors may supply a reduction in posterior sampling variability that, in turn, enhances prediction performances, for cases where there is an expected clustering among covariates. We investigate performances under the two prior constructions described above.

5.1. Hyperparameters Setting

In all examples below, we generally follow the guidelines for hyperparameter settings given in Savitsky et al. [1] for prior settings related to the mixture prior construction of Section 3 and to specific data models. In particular, we employ $\mathbf{G}(1,1)$ priors on λ_a, λ_z. In addition, we center and normalize the response and transform the design matrix to lie in $[0,1]^p$ to produce a small intercept term, which in turn supplies a better conditioned GP covariance matrix. Savitsky et al. [1] note little sensitivity of the results to the choice of w, the prior expectation of covariate inclusion. Here we set $w = 0.025$ in all examples below. In the univariate regression model (2.3), the parameters of the prior on the precision error term, $r \sim \mathbf{G}(a_r, b_r)$, should be set to estimate the *a priori* expected residual variance. We choose $(a_r, b_r) = (2, 0.1)$.

As for the DP priors, we choose $\alpha \sim \mathbf{G}(1,1)$, a setting that produces a prior expected number of clusters of about 7.5 for $p = 1000$ covariates. We briefly discuss sensitivity to the choice of these hyperparameter settings in the simulation results.



(a) Posterior Samples of ρ_k (b) Posterior Samples of ρ_k

Figure 1: Effect of covariate clustering employing prior model (3.5): Univariate regression model ($n = 130, p = 1000$). Box plots of posterior samples for the ρ_k's; (a) shows results without covariate clustering; (b) shows results with covariate clustering.

5.2. *Clustering versus No Clustering*

We first consider the univariate regression model (2.3) and compare performance of covariate clustering under (3.5) with the original GP construction (3.1) of Savitsky et al. [1]. With the latter approach, we employ their MCMC-scheme 2 algorithm to accomplish posterior inference. Results we report were obtained by using a response kernel that includes both linear and non-linear associations and where subgroups of covariates share the same functional form, to induce clustering,

$$y = x_1 + x_2 + x_3 + \sin(9x_4) + \sin(9x_5) + 1.3x_6 + 1.3x_7 + \epsilon, \tag{5.1}$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = .05$, and with covariates simulated from a $\mathcal{U}(0,1)$. We use $p = 1000$ covariates, with the response kernel constructed from the first 7. We do 10,000 MCMC iterations, discarding half as burn-in. Results are presented in Figure 1; plots (a), (b) present box plots for posterior samples of the ρ_k's without clustering and under the clustering model (3.5), respectively. Only the first 20 covariates are displayed, to help visualization. One readily notes both the reduced spread *between* covariates sharing the same functional form and *within* covariate sampled values (of ρ_k) for all covariates under application of our clustering model. Such reduction in within and between spreads of the posterior sampling affects, in turn, prediction performances. In particular, we assessed prediction accuracy by looking at the mean-squared prediction error (MSPE) normalized by the variance of the randomly selected test set, that we term "normalized MSPE". The normalized MSPE declines from 0.12 to 0.02 under application of our clustering model. We further applied the least-squares posterior clustering algorithm of Dahl [22] that chooses among the sampled partitions, post-burn-in, those that are closest to the empirical pairwise clustering probabilities obtained from averaging over posterior samples. Our model returned the correct 3 clusters.

5.3. Clustering All versus Selected Covariates

Next we compare performances for the two prior models (3.5) and (3.7), clustering all covariates and selected covariates only, respectively. We conduct this comparison under the Cox survival model (2.4). The latent response kernel is constructed as

$$y = 3.5x_1 + 3.5x_2 + 3.5x_3 - 1.5\sin(5x_4) - 1.5\sin(9x_5) - 2.5x_6 - 2.5x_7 + \epsilon, \tag{5.2}$$

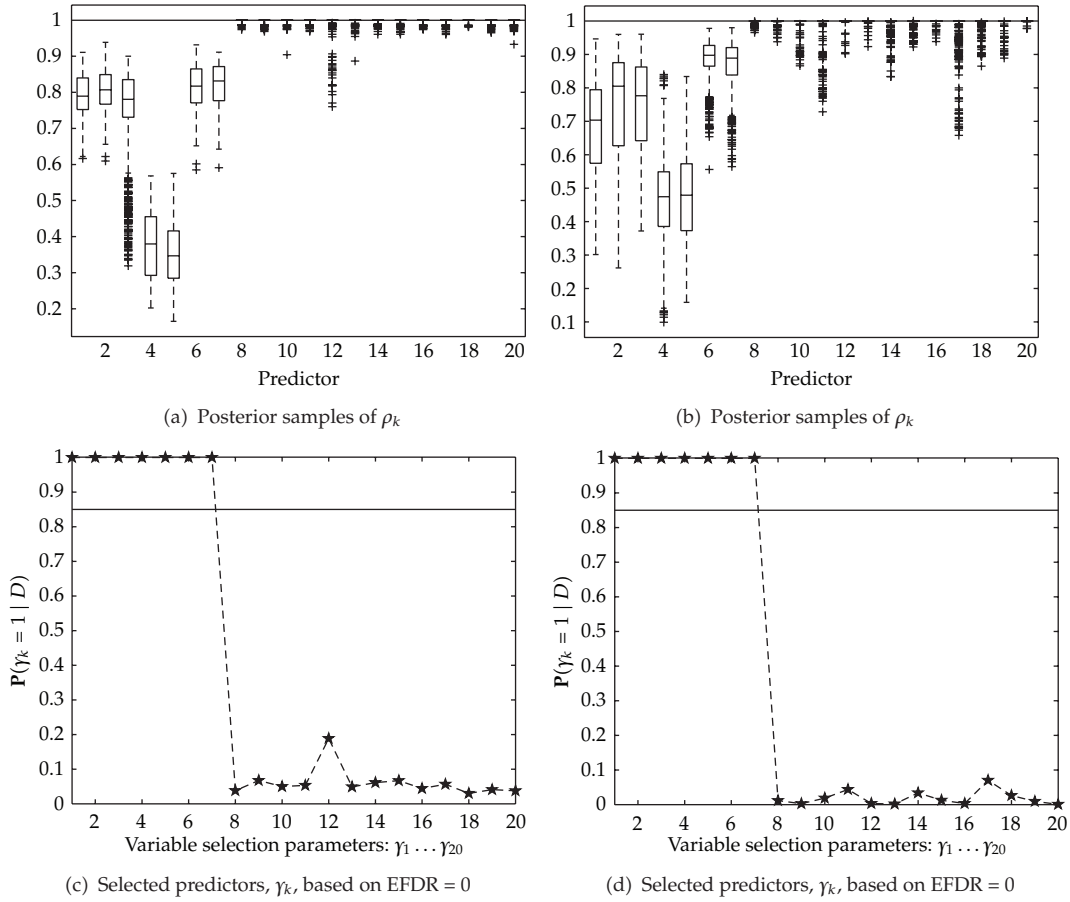
with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = .05$, and with covariates simulated from a $\mathcal{U}(0,1)$. We generate observed $t \sim \text{Exp}(1)/(\lambda \exp(y))$, where we employ $\lambda = 1$. We subsequently randomly censor 5% of our generated survival times. Figure 2 presents the results for clustering all covariates (plots (a), (c)) and only selected covariates (plots (b), (d)). Again, we see the expected clustering behavior among selected covariates in both models, with a slightly less sharp cluster separation in the latter case, indicating a reduction in borrowing of strength among coclustered covariates. We further experimented with the prior expected number of clusters by employing $\alpha \sim \mathcal{G}(a,1)$, with $a = 3 - -5$, and found a further slight reduction of within covariate sampling spread for selected covariates with increasing a, likely resulting from the greater tendency to produce more clusters.

It is worth noting that, under model sparsity, the MCMC algorithm of the model clustering selected covariates only is about 13-14 times faster than the one under the model that clusters all covariates. More precisely, results presented here for the former model were obtained with a computation of 4600 CPU-seconds as compared to 63000 CPU-seconds for the latter model clustering all covariates. This is not surprising under model sparsity, since the model formulation clustering all covariates assigns p covariates to clusters on every MCMC iteration, while the construction clustering selected covariates assigns only p_γ. Computation times do of course increase for both clustering methods proportionally to the number of true clusters. Reported CPU run times were based on utilization of Matlab with a 2.4 GHz Quad Core (Q6600) PC with 4 GB of RAM running 64-bit Windows XP.

6. Benchmark Data Application

We analyze the Boston Housing data of Breiman and Friedman [23] using the covariate clustering model (3.5) that includes all covariates. This data set relates $p = 13$ predictors to the median value of owner-occupied homes in $n = 506$ census tracks in the Boston metropolitan area; see Breiman and Friedman [23] for a detailed description of the predictors. We held out a randomly chosen validation set of 250 observations. Figure 3 compares box plots of marginal posterior samples of ρ_k for all covariates in the following two models: (a) excluding clustering (result reproduced from Savitsky et al. [1]), and (b) clustering all covariates using (3.5). The normalized MSPEs were 0.1 and 0.09, respectively. Clustering covariates therefore induces a relatively modest improvement in performances, though by itself this is not a clear indicator to prefer this formulation.

We again observe some reduction in spread in the posterior samples for the clustered covariates with respect to the formulation of Savitsky et al. [1] that does not cluster covariates, though the effect is less pronounced than what observed for our simulations. When clustering covariates posterior samples for covariates which are frequently coclustered during the MCMC tend to express greater location alignment and general distribution similarity for sampled values. Based on alignment of posterior sampled values, a simple visual inspection



(a) Posterior samples of ρ_k

(b) Posterior samples of ρ_k

(c) Selected predictors, γ_k, based on EFDR = 0

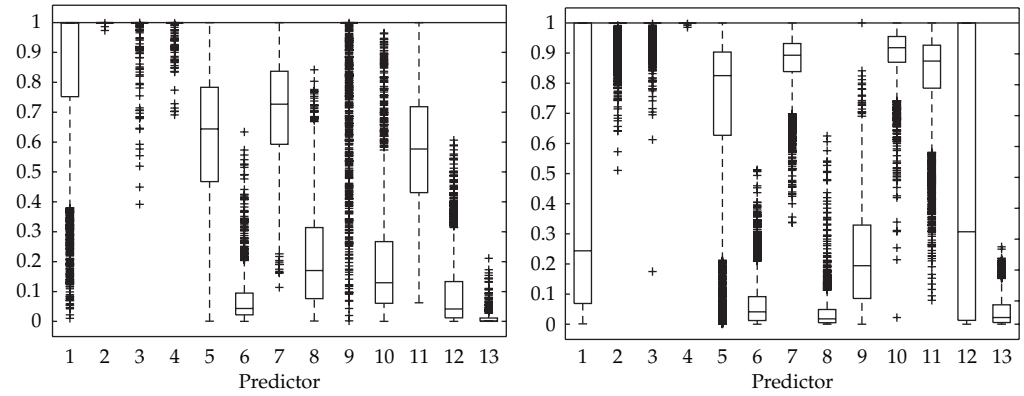
(d) Selected predictors, γ_k, based on EFDR = 0

Figure 2: Effect of covariate clustering: Survival model ($n = 150$, $p = 1000$). Box plots of posterior samples for the ρ_k's and marginal posterior probabilities for the γ_k's; (a) and (c) show results with clustering of *all* covariates; (b) and (d) show results with clustering of only *selected* covariates.

of plot (b) of Figure 3 suggests two clusters, $\phi_1^* = \{x_6, x_8, x_{13}\}, \phi_2^* = \{x_7, x_{10}, x_{11}\}$. Indeed, the posterior configuration with the minimum score suggested by the least squares clustering algorithm of Dahl [22], which provides an analytical approach for selecting clusters from among the posterior partitions, contained ϕ_1 and a separate cluster capturing $\{x_7, x_{11}\}$. The set partition with the second lowest least squares deviation score defines this second cluster as $\{x_7, x_{10}\}$. These results then generally support our subjective visual interpretation.

7. Discussion

In this paper, we have expanded the framework for Bayesian variable selection for generalized Gaussian process (GP) models by employing spiked Dirichlet process (DP) prior constructions over set partitions containing covariates. Our approach results in a nonparametric treatment of the distribution of the covariance parameters of the GP covariance matrix that in turn induces a clustering of the covariates. We have proposed MCMC schemes for posterior inference that use modifications of the auxiliary Gibbs



(a) Posterior samples of ρ_k

(b) Posterior samples of ρ_k

Figure 3: Boston Housing Data. Posterior samples of $\boldsymbol{\rho}$. Plots (a), (b) present box plots of ρ_k. (a) without clustering models; (b) with clustering on all covariates.

algorithm of Neal [13] to facilitate posterior sampling under model sparsity avoiding the generation of duplicate trivial clusters. Our simulation results have shown a reduction in posterior sampling variability and enhanced prediction performances. In addition, we have evaluated two prior constructions: the first one employs a mixture of a point-mass and a continuous distribution as the centering distribution for the DP prior, therefore, clustering all covariates. The second one employs a mixture of a spike and a DP prior with a continuous distribution as the centering distribution, which induces clustering of the selected covariates only. While the former prior construction achieves a better clusters separation, clustering only selected covariates is computationally more efficient.

In the future, it may be interesting to extend our nonparametric covariate clustering models to hierarchical structures that impose some prior dependence among covariates. Another possible extension of our modeling framework includes augmentation with the simultaneous employment of a complementary clustering of observations in a Dirichlet mixture construction incorporating the regression error term of the model. There is no inherent conflict between these two schemes since all observations are in every covariate cluster.

Acknowledgments

M. Vannucci is partially supported by NIH-NHGRI, Grant number R01-HG003319 and by NSF-DMS, Grant number 1007871. T. Savitsky was supported under NIH-NCI Grant T32 CA096520.

References

[1] T. D. Savitsky, M. Vannucci, and N. Sha, "Variable selection for nonparametric Gaussian process priors: Models and computational strategies," *Statistical Science*. Revised.

[2] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, UK, 1983.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, Mass, USA, 2006.

[4] E. I. George and R. E. McCulloch, "Approaches for bayesian variable selection," *Statistica Sinica*, vol. 7, no. 2, pp. 339–373, 1997.
[5] P. J. Brown, M. Vannucci, and T. Fearn, "Multivariate Bayesian variable selection and prediction," *Journal of the Royal Statistical Society. Series B*, vol. 60, no. 3, pp. 627–641, 1998.
[6] N. Sha, M. Vannucci, M. G. Tadesse et al., "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage," *Biometrics*, vol. 60, no. 3, pp. 812–828, 2004.
[7] J. G. Scott and J. O. Berger, "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem," Tech. Rep., 2008.
[8] R. F. MacLehose, D. B. Dunson, A. H. Herring, and J. A. Hoppin, "Bayesian methods for highly correlated exposure data," *Epidemiology*, vol. 18, no. 2, pp. 199–207, 2007.
[9] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.
[10] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.
[11] D. B. Dunson, A. H. Herring, and S. M. Engel, "Bayesian selection and clustering of polymorphisms in functionally related genes," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 534–546, 2008.
[12] S. Kim, D. Dahl, and M. Vannucci, "Spiked dirichlet process prior for bayesian multiple hypothesis testing in random effects models," *Bayesian Analysis*, vol. 4, no. 4, pp. 707–732, 2009.
[13] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.
[14] R. M. Neal, "Regression and classification using Gaussian process priors," in *Bayesian Statistics, 6*, A. D. J. M. Bernando, J. O. Berger, and A. Smith, Eds., pp. 475–501, Oxford University Press, New York, NY, USA, 1999.
[15] C. Linkletter, D. Bingham, N. Hengartner, D. Higdon, and K. Q. Ye, "Variable selection for Gaussian process models in computer experiments," *Technometrics*, vol. 48, no. 4, pp. 478–490, 2006.
[16] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B*, vol. 34, pp. 187–220, 1972.
[17] J. Sacks, S. B. Schiller, and W. J. Welch, "Designs for computer experiments," *Technometrics*, vol. 31, no. 1, pp. 41–47, 1989.
[18] F. A. Quintana, "A predictive view of Bayesian clustering," *Journal of Statistical Planning and Inference*, vol. 136, no. 8, pp. 2407–2429, 2006.
[19] D. Blackwell and J. B. MacQueen, "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, vol. 1, pp. 353–355, 1973.
[20] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.
[21] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist, "Detecting differential gene expression with a semiparametric hierarchical mixture method," *Biostatistics*, vol. 5, no. 2, pp. 155–176, 2004.
[22] D. B. Dahl, "Model-based clustering for expression data via a dirichlet process mixture model," in *Bayesian Inference for Gene Expression and Proteomics*, K. Do, P. Müller, and M. Vannucci, Eds., pp. 201–215, Cambridge University Press, Cambridge, UK, 2006.
[23] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–619, 1985.