Nonparametric Variable Selection, Clustering and Prediction for High-Dimensional Regression

Subharup Guha     Veerabhadran Baladandayuthapani *

July 22, 2014

*Subharup Guha is Associate Professor, Department of Statistics, University of Missouri; Veerabhadran Baladandayuthapani is Associate Professor, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, and Adjunct Associate Professor, Department of Biostatistics, Rice University (email addresses: GuhaSu@missouri.edu, Veera@mdanderson.org). This work was supported by the National Science Foundation under award DMS-0906734 to SG and the National Institutes of Health under grant R01 CA160736 to VB. The authors thank Professor Steven N. MacEachern for helpful comments, Ganiraju Manyam for help with biological interpretations of the results, and participants of the SAMSI Summer Workshop on “Semiparametric Bayesian Inference: Applications in Pharmacokinetics and Pharmacodynamics,” organized by Professors Peter Müller and Gary Rosner in July, 2010, for stimulating discussions. We are grateful to the two referees and the Associate Editor for invaluable suggestions that significantly improved the paper.
Abstract

The development of parsimonious models for reliable inference and prediction of responses in high-dimensional regression settings is often challenging due to relatively small sample sizes and the presence of complex interaction patterns between a large number of covariates. We propose an efficient, nonparametric framework for simultaneous variable selection, clustering and prediction in high-throughput regression settings with continuous or discrete outcomes, called VariScan.

The VariScan model utilizes the sparsity induced by Poisson-Dirichlet processes (PDPs) to group the covariates into lower-dimensional latent clusters consisting of covariates with similar patterns among the samples. The data are permitted to direct the choice of a suitable cluster allocation scheme, choosing between PDPs and their special case, a Dirichlet process. Subsequently, the latent clusters are used to build a nonlinear prediction model for the responses using an adaptive mixture of linear and nonlinear elements, thus achieving a balance between model parsimony and flexibility.

We investigate theoretical properties of the VariScan procedure that differentiate the allocations patterns of PDPs and Dirichlet processes both in terms of the number and relative sizes of their clusters. Additional theoretical results guarantee the high accuracy of the model-based clustering procedure, and establish model selection and prediction consistency. Through simulation studies and analyses of benchmark data sets, we demonstrate the reliability of VariScan’s clustering mechanism and show that the technique compares favorably to, and often outperforms, existing methodologies in terms of the prediction accuracies of the subject-specific responses.

Keywords: Bayesian semiparametric models; Dirichlet process; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Model-based clustering; Model selection consistency; Prediction consistency; Nonlinear functional relationships; Poisson-Dirichlet process; Small $n$, large $p$ problems.
1 Introduction

Suppose the available data in an investigation consist of continuous or discrete responses and \( p \) continuous covariates on \( n \) subjects, arranged in an \( n \) by \( p \) matrix. We assume that only a subset of the covariates are statistically associated with the responses, i.e., for subjects \( i = 1, \ldots, n \), the responses \( w_i \in \mathcal{R} \) are assumed to be associated with an unknown subset of the covariates \( x_{i1}, \ldots, x_{ip} \). The goal of the analysis is two-pronged. First, we wish to infer a common, sparse set of predictor indices for all the subjects, i.e., a subset \( S \subset \{1, \ldots, p\} \) of dimension \( q << p \) consisting of the indices of the covariates that are significantly associated with the responses. Second, we wish to predict the responses of \( \tilde{n} \) additional subjects for whom only covariate information is available. The development of parsimonious regression models that can be used for reliable predictions is challenging. This is especially true of “small \( n \), large \( p \)” regression problems arising in many areas such as high-throughput genomics, imaging and environmental applications.

Several innovative strategies have been developed to meet these challenges in various contexts, with reasonable degrees of success. Most (if not all) of these approaches can be classified into three broad categories based on their basic construction: (a) linear variable selection methods, (b) regression methods using low-dimensional projections of the covariate space, and (c) nonlinear prediction methods. The linear variable selection methods include stepwise selection \( \text{[Peduzzi et al.] 1980} \), penalized regression approaches such as lasso (and its variants) \( \text{[Tibshirani] 1997} \), and non-concave penalized likelihood approaches \( \text{[Fan and Li] 2002} \). Bayesian linear variable selection approaches include spike and slab mixture priors \( \text{[Mitchell and Beauchamp] 1988} \), stochastic search variable selection \( \text{[George and McCulloch] 1993} \), Gibbs-based variable selection \( \text{[Delaportas et al.] 1982} \), Bayesian model averaging \( \text{[Madigan and Raftery] 1994} \), \( \text{[Volinsky et al.] 1997} \) and indicator priors \( \text{[Kuo and Mallick] 1997} \). The stochastic search variable selection approach of George and McCulloch \( \text{[1993]} \) has been extended to multivariate settings by Brown et al. \( \text{[1998]} \) and to generalized linear mixed models by Cai and Dunson \( \text{[2006]} \). Effective variable selection methods have also been developed for multinomial
probit models by Sha et al. (2004), and for microarray data with censored outcomes by Lee and Mallick (2004) and Sha et al. (2006). Work related to the method we present is the product partition model on covariates proposed by Müller et al. (2011). Methods based on regression using low-dimensional projections of the covariate space include partial least squares (Nguyen and Rocke 2002; Li and Gui 2004) and (supervised) principal components methods (Bair and Tibshirani 2004). Non-linear prediction methods include statistical and machine learning techniques such as support vector machines (Cristianini and Shawe-Taylor 2000), L
2
-boosting (Hothorn and Buhlmann 2006) and ensemble methods such as random forests (Ishwaran et al. 2010) and Bonato et al. (2010).

Our motivating application arises from a high-throughput genomics setting where microarray-based expression levels of genes (usually thousands) are available for a limited number of patient samples (tens or hundreds). We wish to select important genes (variables) as well as develop efficient prediction models for patient-specific clinical outcomes such as survival times or disease subtypes. Refer to Sinha and Dey (1997) for a review of semiparametric Bayesian methods for survival outcomes. To illustrate our method, we use an accelerated failure time (AFT) model (Buckley and James 1979; Cox and Oakes, 1984) to analyze the diffuse large B-cell lymphoma (DLBCL) dataset of Rosenwald et al. (2002) and the breast cancer dataset of van’t Veer et al. (2002), both of which have the following general structure. For individuals \(i = 1, \ldots, n\), the data consist of (i) the survival time \(w_i > 0\), (ii) failure status \(\delta_i = 0\) if \(w_i\) is right-censored and \(\delta_i = 1\) if \(w_i\) is uncensored, and (iii) expression levels \(x_{i1}, \ldots, x_{ip}\) for \(p\) genes, with \(p\) being much larger than \(n\). Thus, the log–failure-time \(y_i\) equals \(\log(w_i)\) if \(\delta_i = 1\), and \(y_i\) is latent but exceeds \(\log(w_i)\) if \(\delta_i = 0\).

In a regression setting, we refer to \(y_1, \ldots, y_n\) as the regression outcomes, and fit the model:

\[
y_i \overset{\text{indep}}{\sim} N(\eta_i, \sigma_i^2),
\]

where the regression mean \(\eta_i = \beta_0 + \sum_{j \in S} \beta_j x_{ij}\). George and McCulloch (1993), Kuo and Mallick (1997), and Brown, Vannucci, and Fearn (1998) have proposed the use of
latent indicator variables to identify the covariate matrix columns that are associated with the regression outcomes: \( \eta_i = \beta_0 + \sum_{j=1}^p \gamma_j \beta_j x_{ij} \), where \( \gamma_j \) is an indicator that corresponds to the \( j^{th} \) covariate column being a predictor. The \( \gamma_j \)'s are assumed to be i.i.d. Bernoulli(\( \omega \)). The number of model predictors is then \( |S| = \sum_{j=1}^p \gamma_j \). With \( X_\gamma \) denoting the \( n \) by \( (|S|+1) \) predictor matrix including the intercept column consisting of all ones, and defining \( \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) \), a weighted version of the g prior [Zellner, 1986] is assumed for the regression coefficients: \( \beta_\gamma | \Sigma \sim N_{|S|+1}(0, \sigma_\beta^2 (X_\gamma' \Sigma^{-1} X_\gamma)^{-1}) \) for an unknown \( \sigma_\beta^2 > 0 \).

**Difficulties with predictor detection in small \( n \), large \( p \) regression problems.**

The regression predictors are typically difficult to detect when \( n \ll p \). It is often observed that the \( n \)-dimensional space of the covariate columns becomes “saturated” due to the large number of covariates. The high sample collinearity between the covariates causes them to become weakly identifiable as predictors. For a simple example, imagine that the \( j^{th} \) and \( k^{th} \) covariate columns have a sample correlation close to 1, but that neither covariate is a predictor in the “true” regression model. It is easy to see that an alternative model having both covariates as predictors and \( \beta_j \approx -\beta_k \), but otherwise having the same set of remaining predictors and regression coefficients as the true model, has a nearly identical joint likelihood for all possible regression outcomes.

In general, without any strong, application-specific priors to guide model selection, collinearity makes it difficult to distinguish between competing models on the basis of their likelihood functions and pick the “true” predictors. Furthermore, it is well established that high collinearity causes unstable inferences and erroneous test case predictions [Weisberg, 1985]. This problem is exacerbated if some of the regression outcomes are unobserved, as in survival applications.

Due to these inherent challenges of small \( n \), large \( p \) regression problems, we propose **VariScan**, a nonparametric technique for clustering, variable selection and prediction in high-dimensional regression settings with continuous or discrete outcomes. Since the data are informative regarding the joint effects of correlated covariates rather than the
individual covariates, VariScan utilizes the sparsity-inducing property of Poisson-Dirichlet processes (PDPs) to group the \( p \) columns of the covariate matrix into \( q \) latent clusters, where \( q \ll p \), with each cluster consisting of columns with similar patterns across the subjects. The data are allowed to direct the choice between a class of PDPs and their special case, a Dirichlet process, for a suitable allocation scheme for the covariates. Theoretical results differentiate the allocations patterns of PDPs and Dirichlet processes in terms of the number and relative sizes of their clusters. The within-cluster patterns, common to all the members of the clusters, are flexibly modeled using Dirichlet processes, as opposed to linear projections such as principal components and partial least squares. In contrast to existing mixture model-based clustering techniques, the remarkable theoretical property of VariScan that a fixed set of covariates that (do not) co-cluster under the true model, also (do not) asymptotically co-cluster under the posterior, is established.

This reduces the small \( n \), large \( p \) problem to a “small \( n \), small \( q \)” problem, facilitating an effective stochastic search of the indices \( S^* \subset \{1, \ldots, q\} \) of the cluster predictors, from which we may infer the indices \( S \subset \{1, \ldots, p\} \) of the covariate predictors associated with the responses, as opposed to the typical “black-box” nonlinear prediction methods mentioned before. In addition, the technique is capable of detecting nonlinear functional relationships through elements such as nonlinear functional kernels and basis functions such as splines or wavelets. The adaptive mixture of linear and nonlinear elements in the regression relationship aims to achieve a balance between model parsimony and flexibility. In essence, the technique specifies a random, bidirectional nested clustering of the high-dimensional covariate matrix and builds a nonlinear prediction model for the responses using the latent clusters as covariates. Together, these components of VariScan define a joint model for the responses and covariates that results in an effective model-based clustering and variable selection procedure, improved posterior inferences and accurate test case predictions, which we demonstrate via theoretical guarantees, simulations and real data analyses.

The rest of the paper is organized as follows. We develop the VariScan model and
corresponding theoretical justifications in Section 2. In Section 3 we describe a posterior inference strategy based on Markov chain Monte Carlo (MCMC) techniques. Consistency results for the VariScan procedure are presented in Sections 4. Through simulations in Section 5 and 6 we demonstrate the accuracy of the clustering mechanism and compare the performance of VariScan with those of several existing variable selection procedures for survival outcomes. In Section 7 we analyze the motivating gene expression microarray datasets in leukemia and breast cancer to demonstrate the effectiveness of VariScan as a model-based clustering procedure, and compare the prediction accuracy of VariScan with those of competing methods. Additional supplementary materials contain all technical proofs, additional simulation and real data analyses results.

2 VariScan Model Construction

We model the responses and covariates in a hierarchical manner. Section 2.1 details the models for the covariates and their allocation to the latent clusters. Section 2.2 describes the choice of the cluster-specific predictors and nonlinearly relates them to the possibly latent, subject-specific Gaussian regression outcomes. Section 2.3 links the regression outcomes with the observed responses, which may be either continuous or discrete. Together, these components define a coherent model that could be used for both inference and prediction.

2.1 Modeling the Covariates and Latent Clusters

For the columns \( x_1, \ldots, x_p \) of the (continuous) covariate matrix, suppose each column vector belongs to exactly one of \( q \ll p \) clusters, where the cluster memberships and \( q \) are unknown. For the covariate (column) \( j = 1, \ldots, p \), the covariate-to-cluster assignment is determined by an allocation variable \( c_j \) that equals \( k \) if the \( j^{th} \) covariate belongs to the \( k^{th} \) cluster, where \( k = 1, \ldots, q \).

Furthermore, the clusters are associated with latent vectors \( v_1, \ldots, v_q \), each of length
n. Typically, the covariates are noisy versions of the latent vector components, resulting in high correlations among covariates that belong to a cluster. However, within each cluster, the covariates of a few individuals may be highly variable. To account for this greater heterogeneity, we model the covariates of these individuals with a larger variance. Specifically, for the $j^{th}$ covariate, given that the allocation variable $c_j$ equals $k$ and given an indicator variable $z_{ik}$, we assume for $i = 1, \ldots, n$ that

$$x_{ij} \mid z_{ik}, c_j = k \overset{indep}{\sim} \begin{cases} N(v_{ik}, \tau_1^2) & \text{if } z_{ik} = 0 \\ N(v_{ik}, \tau^2) & \text{if } z_{ik} = 1 \end{cases}$$

where $\tau_1^2$ and $\tau^2$ are component-specific parameters with inverse Gamma priors such that $\tau_1^2 \gg \tau^2$. The value $z_{ik} = 0$ indicates that the covariates of subject $i$ belonging to the $k^{th}$ cluster have an unusually high variance. The indicator variables for the (individual, cluster) combinations are apriori distributed as

$$z_{ik} \overset{iid}{\sim} \text{Ber}(\xi), \quad i = 1, \ldots, n \text{ and } k = 1, \ldots, q,$$

where $\xi \sim \text{beta}(\iota_0, \iota_1)$ with $\iota_0 \ll \iota_1$, so that $P(z_{ik} = 1)$ is high and only a small proportion of covariates have a large variance.

**Allocation variables.** To gain an intuitive understanding of an appropriate model for the covariate-to-cluster allocation, we performed an exploratory data analysis (EDA) of an actual gene expression dataset. The DLBCL data set of Rosenwald et al. (2002) consists of gene expression levels for 240 patients on 7,399 microarray elements (probes), representing approximately 4,128 genes. Eliminating the data for 5 individuals with a survival time of zero and imputing the small number of missing expression values with their probe-specific means, we randomly selected $p = 500$ probes and $n = 100$ individuals, iteratively applying the k-means procedure to group the covariates into clusters.

The iterations were terminated when the following conditions were satisfied: (i) all
within-cluster pairwise correlations of the covariates exceeded 0.3, and (ii) the allocation $R^2$ exceeded 0.7. Under the assumption that all the $z_{ik}$’s are equal to 1, the stopping conditions encourage within-cluster concordance and a small value of $\tau^2$. Figure 1 displays a barchart of the cluster sizes. The pattern we observe is uncharacteristic of a Dirichlet process, which is usually dominated by a small number of clusters with exponentially decreasing sizes. Specifically, for $p = 500$, the large number of clusters ($\hat{q} = 161$) and the predominance of relatively small clusters are strongly suggestive of a non-Dirichlet type of allocation for the covariate-cluster assignments.

The aforementioned EDA suggests the need for a wider range of allocation patterns, such as that provided by a class of generalizations of a Dirichlet process called the two-parameter PDP, introduced by [Perman et al. (1992)] and further studied by [Pitman (1995)] and [Pitman and Yor (1997)]. The allocation variables are apriori exchangeable for PDPs, and more generally, for product partition models [Barry and Hartigan (1993); Quintana and Iglesias (2003)] and species sampling models [Ishwaran and James (2003)]. We assume the following prior for the allocation variables of the covariates:

$$c_1, \ldots, c_p \sim \text{PDP}(\alpha_1, d)$$

where the discount parameter $0 \leq d < 1$ and mass parameter $\alpha_1 > 0$. The number of distinct clusters, $q$, is stochastically increasing in $\alpha_1$ and $d$. For a fixed $d$, all the covariates
are assigned to separate clusters (i.e., $q = p$) as $\alpha_1 \to \infty$. For a fixed $\alpha_1$, setting $d = 0$ yields a Dirichlet process with mass parameter $\alpha_1$.

Conditional on the parameters $\alpha_1$ and $d$, the allocation variables of a PDP evolve as follows. We may assume without loss of generality that $c_1 = 1$. Subsequently, for $j = 2, \ldots, p$, suppose there are $q^{(j-1)}$ distinct clusters among $c_1, \ldots, c_{j-1}$, with the $k^{th}$ cluster containing $n_k^{(j-1)}$ number of covariates, where $k = 1, \ldots, q^{(j-1)}$. The predictive probability that the $j^{th}$ covariate belongs to the $k^{th}$ cluster is then

$$P(c_j = k \mid c_1, \ldots, c_{j-1}) \propto \begin{cases} n_k^{(j-1)} - d & \text{if } k = 1, \ldots, q^{(j-1)} \\ \alpha_1 + q^{(j-1)} \cdot d & \text{if } k = q^{(j-1)} + 1 \end{cases}$$

where the event $c_j = q^{(j-1)} + 1$ corresponds to the $j^{th}$ covariate opening a new cluster. When $d = 0$, we obtain the well known Pólya urn scheme for Dirichlet processes (Ferguson, 1973). Refer to Lijoi and Prünster (2010) for a detailed discussion of Bayesian nonparametric models, including Dirichlet processes and PDPs.

The use of PDPs in this setting achieves dimension reduction for the covariate clusters because the random number of clusters, $q = q(p)$, is asymptotically equivalent to

$$\begin{cases} \alpha_1 \cdot \log p & \text{if } d = 0 \quad (\text{Dirichlet process}) \\ T_{d,\alpha_1} \cdot p^d & \text{if } 0 < d < 1 \end{cases}$$

for a random variable $T_{d,\alpha_1} > 0$. This implies that, as $p \to \infty$, the number of clusters of a Dirichlet process is of smaller order than that of a PDP with discount parameter $d > 0$. Dirichlet processes have been previously utilized for dimension reduction; for example, see Medvedovic et al. (2004), Kim et al. (2006), Dunson et al. (2008) and Dunson and Park (2008). In essence, this provides an effective dimension reduction clustering technique for regression settings that we exploit in our model.

Sethuraman (1994) derived the stick-breaking representation for a Dirichlet process, and then Pitman (1995) extended it to PDPs as follows. Let $\mathbb{N}$ be the set of natural
numbers. Subject to a one-to-one transformation of the first $q$ natural numbers into $\mathbb{N}$, the allocation variables $c_1, \ldots, c_p$ are i.i.d. samples from a discrete distribution $F_{\alpha_1,d}$ on $\mathbb{N}$ with stick-breaking probabilities $\pi_1 = V_1$ and $\pi_h = V_h \prod_{t=1}^{h-1} (1 - V_t)$ for $h = 2, 3, \ldots,$ where $V_h \overset{\text{indep}}{\sim} \text{beta}(1 - d, \alpha_1 + hd)$. This implies that for large values of $p$ and for clusters $k = 1, \ldots, q$, the frequencies $n_k^{(p)} / p$ are approximately equal to $\pi_{h_k}$ for some distinct integers $h_1, \ldots, h_q$.

The following theorem provides expressions for the first and second moments of the random log-probabilities of $F_{\alpha_1,d}$. Part 1c provides an explanation for the fact that Dirichlet process allocations typically consist of a small number of clusters, only a few of which are large, with exponential decay in the cluster sizes. Part 2c suggests that for PDPs with $d > 0$ (i.e., non-Dirichlet process realizations), there is a slower, power law decay of the cluster sizes as $d$ increases. Part 3 indicates that for every $\alpha_1$ and $d > 0$, a PDP realization $F_{\alpha_1,d}$ is thicker tailed compared to a Dirichlet process realization, $F_{\alpha_1,0}$. In conjunction with equation (3) above, Theorem 2.1 essentially justifies the use of PDPs when the observed number of clusters is large or the cluster sizes decay slowly. See Supplementary Materials Section ?? for the proof.

**Theorem 2.1** Consider the PDP model (2). Let $\psi(x) = d \log \Gamma(x) / dx$ denote the digamma function and $\psi_1(x) = d^2 \log \Gamma(x) / dx^2$ denote the trigamma function.

1. For $d = 0$, the distribution $F_{\alpha_1,0} \in \mathbb{N}$ is a Dirichlet process realization with stick-breaking probabilities $\pi^*_h$ based on $V_h \overset{\text{iid}}{\sim} \text{beta}(1, \alpha_1)$ for $h \in \mathbb{N}$. Then
   
   (a) $E(\log \pi^*_h) = \psi(1) - \psi(\alpha_1) - h / \alpha_1$. Thus, $\lim_{h \to \infty} E(\log \pi^*_h) = -\infty$.
   
   (b) $\text{Var}(\log \pi^*_h) = \psi_1(1) - \psi_1(\alpha_1) + h / \alpha_1^2$. Thus, $\lim_{h \to \infty} \text{Var}(\log \pi^*_h) = \infty$.
   
   (c) As $h \to \infty$, $\sqrt{h} \left( \frac{1}{h} \log(\pi^*_h) + 1 / \alpha_1 \right) \overset{L}{\to} N(0, 1 / \alpha_1^2)$. This implies that as $h \to \infty$, the random stick-breaking Dirichlet process probabilities, $\pi^*_h$, are stochastically equivalent to $e^{-h/\alpha_1}$.

2. For $0 < d < 1$, the distribution $F_{\alpha_1,d} \in \mathbb{N}$ is a realization of a PDP with stick-breaking probabilities $\pi_h$, where $h \in \mathbb{N}$. However, $F_{\alpha_1,d}$ is not a Dirichlet process.
realization because \(d \neq 0\). Then

(a) \(E(\log \pi_h) = \psi(1-d) - \psi(\alpha_1) + \frac{1}{d}(\psi(\alpha_1/d) - \psi(\alpha_1/d + h))\). This implies that \(\lim_{h \to \infty} E(\log \pi_h) = -\infty\).

(b) \(Var(\log \pi_h) = \psi_1(1-d) - \psi_1(\alpha_1) + \frac{1}{d^2}(\psi_1(\alpha_1/d) - \psi_1(\alpha_1/d + h))\). Unlike a Dirichlet process realization, \(\lim_{h \to \infty} Var(\log \pi_h)\) is finite regardless of \(d > 0\).

(c) For any \(\alpha_1 > 0\) and as \(h \to \infty\), \(\log \pi_h / \log h^{-1/d} \xrightarrow{p} 1\) for non-Dirichlet process realizations.

3. As \(h \to \infty\), \(\sqrt{h} \left(\frac{1}{h} \log(\pi^*_h / \pi_h) + 1/\alpha_1\right) \xrightarrow{L} N(0, 1/\alpha_1^2)\). That is, as \(h \to \infty\), the ratios of the Dirichlet process and non-Dirichlet process stick-breaking random probabilities, \(\pi^*_h / \pi_h\), are stochastically equivalent to \(e^{-h/\alpha_1}\) for every \(d > 0\).

Remark

By Lemma 1 of Ishwaran and James (2003), \(\lim_{h \to \infty} E(\log \pi^*_h) = -\infty\) in Part 1a of Theorem 2.1 is equivalent to \(\sum_{h=1}^\infty \pi^*_h = 1\) almost surely for a Dirichlet process. A similar comment applies in Part 2a for a PDP.

In the VariScan model, the parameter \(d\) in the PDP model, equation (2), is given the mixture prior \(\frac{1}{2} \delta_0 + \frac{1}{2} U(0, 1)\), where \(\delta_0\) denotes a point mass at 0. This allows the mixture prior to flexibly choose between a Dirichlet process and a more general PDP for a suitable clustering mechanism of the covariates.

Latent vector elements. The PDP prior specification is completed by a base distribution in \(\mathcal{R}^n\) for the i.i.d. latent vectors. The \(nq\) number of components of the latent vectors \(v_1, \ldots, v_q\) are assumed to have the following distribution:

\[ v_{ik} \overset{iid}{\sim} G \quad i = 1, \ldots, n, \text{ and } k = 1, \ldots, q, \]  

(4)

allowing the clusters to communicate through shared latent vector elements. Furthermore, the real-valued distribution \(G\) is given a nonparametric Dirichlet process prior, which
allows the latent vectors to flexibly capture the within-covariate patterns of the subjects:

$$G \sim DP(\alpha_2; N(\mu_2, \tau_2^2))$$  \hspace{1cm} (5)

with mass parameter $\alpha_2 > 0$ and base distribution $N(\mu_2, \tau_2^2)$. This implies that $G$ is discrete and that the number of distinct values among the $v_{ik}$'s is asymptotically equivalent to $\alpha_2 \cdot \log nq$. In Section 3, we demonstrate that this allocation scheme for the latent vector elements is validated by the real DLBCL dataset.

In essence, the afore-mentioned probability model specifies a random, bidirectional nested clustering of the $n$ by $p$ covariate matrix. Unlike the model based clustering approaches of Fraley and Raftery (2002), Quintana (2006) and Freudenberg et al. (2010), VariScan does not assume that it is possible to globally reshuffle the rows and columns of the covariate matrix to reveal a clustering pattern. Instead, somewhat similarly to the nonparametric Bayesian local clustering (NoB-LoC) approach of Lee et al. (2013), VariScan clusters the covariates locally using two sets of product partition models (Hartigan, 1990; Barry and Hartigan, 1993; Crowley, 1997). However, there are significant differences between NoB-LoC and the clustering aspect of VariScan, in that VariScan is primarily motivated by high-dimensional regression problems rather than bi-clustering, which is the emphasis of NoB-LoC. In addition, NoB-LoC relies solely on Dirichlet processes for clustering; whereas VariScan permits a mixture of Dirichlet processes and PDPs.

2.2 Modeling the Predictor Choices and Regression Outcomes

For $k = 1, \ldots, q$, let $n_k$ be the number of covariates belonging to the $k^{th}$ cluster, so that $n_k = \sum_{j=1}^{p} I(c_j = k)$ and $\sum_{k=1}^{q} n_k = p$. To gain an intuitive understanding, imagine that each cluster nominates from its covariate members a representative $u_k$, and that all $n_k$ covariates have an apriori equal chance of being nominated. Let $s_k$ denote the index of the covariate belonging to the $k^{th}$ cluster that is chosen as its representative, so that $c_{s_k} = k$ and $u_k = x_{s_k}$. In accordance with our cluster-based strategy for dimension
reduction, the responses are directly related to the cluster representatives rather than the individual covariates. The regression predictors are then chosen from the set of \( q \) cluster representatives, and the indices of their clusters constitute the set of cluster predictors, \( S^* \subset \{1, \ldots, q\} \). We emphasize that the latent vectors \( \mathbf{v}_k \) of Section 2.1 determine the allocation of the covariates to the clusters, and so indirectly but significantly influence the choice of the influence of the cluster representatives. As an alternative modeling strategy, we could also choose the latent vectors themselves as the cluster representatives. The former approach is more interpretable because practitioners often think in terms of individual regressors and their corresponding effects on the outcome.

The nominated cluster representatives are featured in an additive regression model that can accommodate nonlinear functional relationships. Specifically, the regression outcomes are assumed to have the distribution

\[
y_i \overset{\text{indep}}{\sim} N(\eta_i, \sigma_i^2), \quad \text{where} \quad \eta_i = \beta_0 + \sum_{k=1}^q \gamma_k^{(1)} \beta_k^{(1)} u_{ik} + \sum_{k=1}^q \gamma_k^{(2)} h(u_{ik}, \beta_k^{(2)}) \tag{6}
\]

for a nonlinear function \( h \). The expression for \( \eta_i \) implicitly relies on the triplet of cluster-specific indicators, \( \gamma_k = (\gamma_k^{(0)}, \gamma_k^{(1)}, \gamma_k^{(2)}) \), where \( \gamma_k^{(0)} + \gamma_k^{(1)} + \gamma_k^{(2)} = 1 \). The value \( \gamma_k^{(0)} = 1 \) corresponds to the cluster representative \( u_k \) not appearing in equation (6) and none of the covariates in latent cluster \( k \) being associated with the responses. The value \( \gamma_k^{(1)} = 1 \) corresponds to \( u_k \) appearing as a simple linear regressor in equation (6), and \( \gamma_k^{(2)} = 1 \) corresponds to its occurrence in a nonlinear form. This adaptive mixture of linear and nonlinear elements aims to achieve a balance between model parsimony and flexibility.

Possible options for the function \( h \) in equation (6) include nonlinear function kernels such as those based on reproducible kernel Hilbert spaces (Mallick et al., 2005), nonlinear basis smoothing splines (Eubank, 1999), and wavelets. Especially attractive due to their ease of construction and interpretability as a linear model are order-\( r \) splines with \( m \)
number of knots (de Boor, 1978; Hastie and Tibshirani, 1990; Denison et al., 1998a):

\[ h_{rm}(u_{ik}, \beta_k^{(2)} | \kappa_{sk}) = \beta_{k,1} u_{ik} + \cdots + \beta_{k,r} u_{ik} + \sum_{t=1}^{m} \beta_{k,r+t} (u_{ik} - \kappa_{sk,t})^r. \]

where \( a_+^r = (\max\{0, a\})^r \) and \( \kappa_{sk} \) denotes the vector of \( m \) knots associated with the \( s_k^{th} \) covariate. This construction allows one to capture the linear dependencies, and perhaps more crucially, the nonlinear functional structures between the covariates and responses. This formulation can be viewed as a special case (without interactions) of multivariate adaptive regression splines, proposed by Friedman (1991) and extended in the Bayesian framework by Denison et al. (1998b) and Baladandayuthapani et al. (2006).

The set of covariate predictors is then \( S = \{ s_k : \gamma_k^{(1)} + \gamma_k^{(2)} > 0, k = 1, \ldots, q \} \) and it is a subset of \( \{ 1, \ldots, p \} \). The number of cluster predictors that appear as simple linear regressors in equation (6) is \( q_1 = \sum_{j=1}^{q} \gamma_j^{(1)} \), and the number that appear as nonlinear predictors is \( q_2 = \sum_{j=1}^{q} \gamma_j^{(2)} \). The number of cluster representatives that are non-predictors is \( q_0 = q - q_1 - q_2 \). The total number of cluster predictors is \( |S^*| = q_1 + q_2 \), which equals the number of covariate predictors, \( |S| \).

For models with nonlinear functions \( h \) that can be interpreted as a linear model, let \( \gamma = (\gamma_1, \ldots, \gamma_q) \) and \( U_\gamma \) be a matrix of \( n \) rows consisting of the intercept column and the independent regression variables based on the cluster representatives. Let \( \text{col}(U_\gamma) \) denote the number of columns of \( U_\gamma \). For example, if we use order-\( r \) splines with \( m \) number of knots in equation (6), then \( \text{col}(U_\gamma) = q_1 + (m + r) \cdot q_2 + 1 \). With the symbol \([\cdot] \) representing densities, the prior for \( \gamma \) is

\[ [\gamma] \propto \omega_0^q \omega_1^q \omega_2^q \cdot I(\text{col}(U_\gamma) < n) \]  

(7)

where \( \omega_0 + \omega_1 + \omega_2 = 1 \), and \( (\omega_0, \omega_1, \omega_2) \sim D_3(1,1,1) \), a Dirichlet distribution. The restricted support of \( \gamma \) induces model sparsity, as discussed below. Conditional on \( \Sigma = \ldots \)
diag(\(\sigma_1^2, \ldots, \sigma_n^2\)), as before, a weighted g prior is assumed for the regression coefficients:

\[
\beta_{\gamma} | \Sigma \sim N_{|S^*|+1} \left(0, \sigma_{\beta}^2 (U_{\gamma}' \Sigma^{-1} U_{\gamma})^{-1}\right).
\] (8)

An advantage of the VariScan procedure is its ability to quantify nonlinear functional relationships between the responses and covariates. The nonlinearity measure \(N \in [0, 1]\) is defined as the posterior expectation,

\[
N = E(\frac{\omega_2}{\omega_1 + \omega_2} | w, X).
\] (9)

The nonlinearity measure can be interpreted as the posterior predictive probability that a hypothetical, additional cluster appears as a predictor in equation (6) in a nonlinear form, rather than as a simple linear regressor. That is, \(N\) is the posterior probability that \(\gamma_{q+1}^{(2)} = 1\). A value of \(N\) close to 0 (1) corresponds to linear (nonlinear) associations between the response and a majority of the predictors.

**Model parsimony versus flexibility.** Although the model assumptions guarantee that the number of clusters, \(q\), is much smaller than the number of covariates, \(p\), it is frequently observed that \(q\) exceeds the number of subjects, \(n\); examples include the DLBCL (Rosenwald et al., 2002) and breast cancer (van’t Veer et al., 2002) datasets. The reliability of inferences and future predictions then rapidly deteriorates as the number of cluster predictors and the number of additive nonlinear components in equation (6) increase. In spline-based models, this puts a constraint on the order of the splines, often necessitating the use of linear splines with \(m = 1\) knot per cluster in equation (6). In the applications presented in this paper, we fixed the knot for each covariate at the sample median. The restriction in the prior (7) also prevents over-fitting. It ensures that the matrix \(U_{\gamma}\), consisting of the independent regression variables, has fewer columns than rows, and is a sufficient condition for the existence of \((U_{\gamma}' \Sigma^{-1} U_{\gamma})^{-1}\) and the least-squares estimate of \(\beta_{\gamma}\) in equation (6).
Furthermore, unusually small values of $\sigma_i^2$ in equation (6) correspond to over-fitted models, whereas unusually large values correspond to under-fitted models. Any parameters that determine $\sigma_1^2, \ldots, \sigma_n^2$ are key, and their priors must be carefully chosen. For instance, simple linear regression (e.g. AFT survival analysis for the DLBCL and breast cancer datasets) assumes that $\sigma_i^2 = \sigma^2$. We have found that non-informative priors for $\sigma^2$ do not work well because the optimal model sizes for variable selection are unknown. This is especially true when a large proportion of regression outcomes are censored. Additionally, we have found that it is helpful to restrict the range of $\sigma^2$ based on reasonable goals for inference precision. In the survival examples discussed in this paper, we assigned the following truncated prior: $\sigma^{-2} \sim \chi_\nu^2 \cdot \mathcal{I}(0.95^{-1}/\text{Var}(\hat{y}) < \sigma^{-2} < 0.5^{-1}/\text{Var}(\hat{y}))$, where the degrees of freedom $\nu$ were appropriately chosen and the vector $\hat{y}$ relied on EDA estimates of latent regression outcomes from a previous study or the training set individuals. The support for $\sigma^{-2}$ was chosen to approximately correspond to the constraint, $0.5 < R^2 < 0.95$, quantifying the effectiveness of regression. As Sections 6 and 7 demonstrate, the aforementioned strategies often result in high reliability of predictions in survival applications.

### 2.3 Modeling the Responses

Lastly, we model the relationship between the observed responses $w_i$ and regression outcomes $y_i$. The $w_i$’s, which may be either continuous or discrete, are assumed to be deterministic transformations of independent variables $R_i$ having exponential family distributions. That is, for a set of functions $f_i$, we assume that $w_i = f_i(R_i)$ and

$$\left[R_i \mid \varrho_i, \varsigma\right] = r(R_i, \varsigma) \cdot \exp \left( \frac{R_i \varrho_i - b(\varrho_i)}{a(\varsigma)} \right)$$

where $r(\cdot)$ is a non-negative function, $\varsigma$ is a dispersion parameter, and $\varrho_i$ is the canonical parameter. The mean $\mu_i = E[R_i|\varrho_i, \varsigma]$ equals $b'(\varrho_i)$ and $\text{Var}[R_i|\varrho_i, \varsigma]$ equals $b''(\varrho_i)a(\varsigma)$. For an appropriate link function $g(\cdot)$, the regression mean $\eta_i$, defined in equation (6), equals
$g(\mu_i)$. Gaussian regression is a special case of this setting for a normal density, identity link, and dispersion parameter $\varsigma = \sigma^2$. Poisson regression corresponds to a Poisson density, log link, and $\varsigma = 1$. Logistic (probit) regression corresponds to a Bernoulli density, logit (probit) link, and $\varsigma = 1$. Survival analysis with AFT models also fits into this framework: for Gaussian $R_i$’s and an independent set of censoring times $C^\dagger_1, \ldots, C^\dagger_n$, we have $\log w_i = \min(R_i, \log C^\dagger_i)$ and $\delta_i = I(R_i \leq \log C^\dagger_i)$.

**Relationship with regression outcomes.** The Laplace approximation (Harville, 1977) relates the $R_i$’s to the regression outcomes:

$$y_i = \eta_i + \frac{\partial \eta_i}{\partial \mu_i} \cdot (R_i - \mu_i)$$

$$\sim N(\eta_i, \sigma^2_i) \quad (11)$$

with precision $\sigma_i^{-2} = \{b''(\mu_i)\}^{-1} (\partial \mu_i / \partial \eta_i)^2$. The idea of using a Laplace-type approximation to infer the model parameters of exponential families has precedence in the literature; some examples include Zeger and Karim (1991), Albert and Chib (1994), and Albert et al. (1998). For linear regression, the approximation is exact with $y_i = R_i$. The Laplace approximation is not restrictive even when it is approximate; for example, MCMC proposals for the model parameters made using equation (11) can be filtered through a Metropolis-Hastings step to obtain samples from the target posterior. Alternatively, inference strategies relying on normal mixture representations through auxiliary variables could be used to relate the $R_i$’s to the $y_i$’s. For instance, Albert and Chib (1993) used truncated normal sampling to obtain a probit model for binary responses, and Holmes and Held (2006) utilized a scale mixture representation of the normal distribution (Andrews and Mallows, 1974; West, 1987) to implement logistic regression using latent variables.

The schematic architecture of the VariScan model is shown in Figure 2 using a directed acyclic graph.
Figure 2: Directed acyclic graph of the VariScan model in which the cluster representatives are chosen from the set of co-clustered covariates. Circles represent stochastic model parameters, solid rectangles represent data and deterministic variables, and dashed rectangles represent model constants. Solid (dashed) arrows represent stochastic (deterministic) relationships.
3 Posterior inference

Starting with an initial configuration obtained by a naïve, preliminary analysis, the model parameters are iteratively updated by MCMC methods. Section 3.1 describes the generation of the allocation variables. Section 3.2 describes the updates of the latent vector elements and their binary indicators. Sections 3.3 and 3.4 respectively describe the MCMC updates of the cluster predictors and any latent regression outcomes. Section 3.5 discusses the prediction of responses for individuals with only covariates available.

Due to the intensive nature of the posterior inference, the analysis can be done in two stages, with cluster detection followed by predictor discovery:

Stage 1 Focusing on only the covariates and ignoring the responses:

Stage 1a The procedures of Sections 3.1 and 3.2 are iteratively performed until the MCMC chain converges. Monte Carlo estimates are computed for the posterior probability of clustering for each pair of covariates. Applying the technique of Dahl (2006), these pairwise probabilities are used to compute a point estimate for the allocation variables, which is called the least-squares allocation.

Stage 1b Conditional on the least-squares allocation as the true clustering of the covariates, a second MCMC sample is generated using the procedure described in Section 3.2. Again applying the technique of Dahl (2006), we compute a point estimate, called the least-squares configuration, for the set of latent vector elements \{v_{ik}\} and indicators \{z_{ik}\}.

Stage 2 Conditional on the least-squares allocation and least-squares configuration, and focussing on the responses, a third MCMC sample is generated using the strategies of Sections 3.3 and 3.4. The sample is post-processed to obtain posterior inferences for the predictors. As described in Section 3.5, the sample can also be used to predict the outcomes of subjects with unknown responses.

As a further benefit of having a well-defined model for the covariates, as part of the
MCMC procedure, VariScan performs model-based imputations of any missing covariate values.

### 3.1 Covariate-to-cluster Allocation

For \( j = 1, \ldots, p \), the full conditional distribution of allocation variable \( c_j \) is not available in closed form. Nevertheless, we borrow ideas from sequential importance sampling (refer to Liu 2008, chap. 3) to devise a Gibbs sampler. The details of this MCMC procedure are provided in Section ?? of the Supplementary Materials. Applying this strategy, new clusters were successfully opened 8.25% of the time for the DLBCL dataset and 9.47% of the time for the breast cancer dataset. Key to its success is the assumption that the clusters borrow strength through a common distribution \( G \) for their latent vector elements.

For the DLBCL data, the upper left panel of Figure 3 displays the estimated posterior density of the PDP’s discount parameter \( d \). The estimated posterior probability of the event \( [d = 0] \) is exactly zero, implying that a non-Dirichlet process clustering mechanism is strongly favored by the data, as suggested earlier by the EDA. The upper right panel of Figure 3 plots the estimated posterior density of the number of clusters. The a posteriori large number of clusters (for \( p = 500 \) covariates) is suggestive of a PDP model with \( d > 0 \) (i.e. a non-Dirichlet process model). The lower left panel of Figure 3 summarizes the cluster sizes of the least-squares allocation (Dahl 2006). The large number of clusters (\( \hat{q} = 165 \)) and the multiplicity of small clusters are very unusual for a Dirichlet process, justifying the use of the more general PDP model.

### 3.2 Latent Vectors and Indicators

Among the allocation variables \( c_1, \ldots, c_p \), suppose there are \( q \) clusters, with cluster \( k \) consisting of \( n_k = \sum_{j=1}^{p} I(c_j = k) \) covariates for \( k = 1, \ldots, q \). As \( i = 1, \ldots, n \) and \( k = 1, \ldots, q \) vary, the sufficient statistics \( \bar{x}_{ik} = \sum_{j=1}^{p} x_{ij} \cdot I(c_j = k)/n_k \) are independently
Figure 3: Posterior summaries for the DLBCL dataset. The top panels and the lower left panel summarize the least-squares covariate-to-cluster PDP allocation of the 500 genes. The lower right panel depicts the least-squares Dirichlet process configuration of the more than 14,000 latent vector elements with binary indicators equal to 1.
Figure 4: For the DLBCL dataset, median pairwise correlations for the \( \hat{q} = 165 \) PDP clusters in the least-squares allocation of Stage 1a.

distributed as \( N(0, \tau_1^2/n_k) \) if \( z_{ik} = 0 \), and as \( N(v_{ik}, \tau_2^2/n_k) \) if \( z_{ik} = 1 \). Dirichlet process prior (5) is conjugate to the above distribution and to the sampling distribution of the \( z_{ik} \)'s. For \( i = 1, \ldots, n \), and \( k = 1, \ldots, q \), we can therefore update the bivariate vector \((v_{ik}, z_{ik})\) by Gibbs sampling.

In Stage 1b of the two-stage analysis, we computed the least-squares configuration of the latent vector elements for the DLBCL sample. More than 87\% of the \( n\hat{q} = 16,500 \) latent vector elements have \( \hat{z}_{ik} = 1 \), implying that a relatively small proportion of covariate values for the DLBCL dataset can be regarded as random noise having no clustering structure. The lower right panel of Figure 3 presents a summary of the least-squares configuration for the latent vector elements with \( \hat{z}_{ik} = 1 \). For the more than 14,000 latent vector elements with \( \hat{z}_{ik} = 1 \), there are only 157 distinct values representing the estimated point masses of the distribution \( G \). The configuration has mainly large clusters and closely resembles the typical configuration for a Dirichlet process model, justifying assumption (5).

For each of the \( \hat{q} = 165 \) clusters in the least-squares allocation of Stage 1a, we computed the correlations between its member covariates and the latent vector for individuals with \( \hat{z}_{ik} = 1 \). The cluster-wise median correlations are plotted in Figure 4. The plots reveal fairly good within-cluster concordance regardless of the cluster size.
3.3 Cluster Predictors and Cluster Representatives

The choice of basis functions such as splines and wavelets for the nonlinear functionals $h$ in (3) result in non-linear terms that are additive in analytic (e.g., polynomial or periodic) functions of the cluster representatives. In such cases, it is possible to integrate out the regression coefficients $\beta_{\gamma}$ to iteratively update the vector of indicators $\gamma_k = (\gamma_k^{(0)}, \gamma_k^{(1)}, \gamma_k^{(2)})$, for clusters $k = 1, \ldots, q$. Given the cluster representative $u_k$ and the set of indicators for the remaining $(q-1)$ clusters, the sub-models corresponding to $\gamma_k^{(0)} = 1$, $\gamma_k^{(1)} = 1$, and $\gamma_k^{(2)} = 1$, are then progressively nested.

The general result of Theorem ?? of the Supplementary Materials is exploited to quickly compute, up to a multiplicative constant, the likelihood functions for these three sub-models. This makes it possible to easily perform joint updates for $u_k$ and $\gamma_k$. After a cycle of updates of $q$ indicators and cluster representatives has been completed, the regression coefficients $\beta_{\gamma}$ may be jointly generated from the full conditional if necessary.

3.4 Latent Regression Outcomes

Suppose the regression outcomes $y_i$ are latent, but the responses $w_i$ are observed for some subjects. For example, for right-censored survival times under the AFT model, the regression outcome $y_i$ is latent for individuals with $\delta_i = 0$, although the survival time $w_i$ is observed for these individuals and it is known that $\log w_i < y_i$.

The latent $y_i$’s can be iteratively sampled as follows. Let $V = \Sigma^{-1/2}U_\gamma$. Compute the symmetric projection or hat matrix of order $n$: $H = ((h_{it})) = V(V'V)^{-1}V'$. As proved in Section ?? of the Supplementary Materials, the prior distribution of $y_i$ given the remaining regression outcomes is $y_i \mid y_{-i} \sim N\left(\frac{\alpha_i \sum_{t \neq i} y_t h_{it}/\sigma_t}{\varphi^2 h_{ii}}, \frac{\sigma^2 \varphi^2}{\varphi^2 h_{ii}}\right)$ where $\varphi^2 = 1 + \sigma_{\beta}^{-2}$.

The conditional prior can be updated using the response $w_i$ to generate a posteriori, $y_i \sim [y_i \mid y_{-i}, w_i]$. For right-censored outcomes under the AFT model, this corresponds to a truncated normal full conditional, yielding a Gibbs sampler.
3.5 Predictions

Suppose there are \( \tilde{n} \) additional individuals with unobserved responses but with available covariates \( \tilde{x}_{i1}, \ldots, \tilde{x}_{ip} \) for \( i = 1, \ldots, \tilde{n} \). As with the training set, we arrange the cluster representative elements for the test cases in an \( \tilde{n} \times \text{col}(U_\gamma) \) matrix. Given the set of predictors \( \gamma \) and variances \( \tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_\tilde{n}^2 \) in relation [6], the following theorem provides expressions for the posterior predictions of the regression outcomes, \( \tilde{y} \). See Supplementary Materials Section ?? for the proof.

**Theorem 3.1** Given the set of predictors \( \gamma \) in model [6], let \( \tilde{U}_\gamma \) be the matrix of cluster representative elements consisting of \( \tilde{n} \) rows and \( \text{col}(U_\gamma) \) number of columns. Define \( \tilde{\Sigma} = \text{diag}(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_\tilde{n}^2) \). Then \( \tilde{y} \mid y \sim N_{\tilde{n}} \left( \frac{1}{1+\sigma_\beta^2} \tilde{y}_{lse} ; \tilde{\Sigma} + \frac{1}{1+\sigma_\beta^2} \tilde{H} \right) \), where \( \tilde{y}_{lse} = \tilde{U}_\gamma \hat{\beta}_{lse} \) with the vector of the least-squares estimates, \( \tilde{\beta}_{lse} = (U_\gamma'\Sigma^{-1}U_\gamma)^{-1}U_\gamma'\Sigma^{-1}y \), and where \( \tilde{H} = \tilde{U}_\gamma(U_\gamma'\Sigma^{-1}U_\gamma)^{-1}\tilde{U}_\gamma' \). Therefore, under a squared error loss, the vector of the predicted regression outcomes for the \( \tilde{n} \) subjects is \( E[\tilde{y} \mid y] = \frac{1}{1+\sigma_\beta} \tilde{y}_{lse} \).

4 Consistency results

The first part of the following theorem explores the reliability of VariScan’s assignment of the covariate matrix columns to the PDP clusters. In the more general problem of using mixture models to allocate \( p \) objects to an unknown number of clusters, the problem of non-identifiability and redundancy of the detected clusters has been extensively documented in Bayesian and frequentist applications (e.g., see [Frühwirth-Schnatter 2006]). Some partial solutions are available in the Bayesian literature. For example, in finite mixture models, rather than assuming exchangeability of the mixture component parameters, [Petralia et al. 2012] regard them as draws from a repulsive process, leading to fewer, better separated and more interpretable clusters. [Rousseau and Mengersen 2011] show that a carefully chosen prior leads to asymptotic emptying of the redundant components in over-fitted finite mixture models. The underlying strategy of these procedures is that
they focus on detecting the correct number of clusters rather than the correct allocation of the \( p \) objects.

In contrast, Part 1 of Theorem 4.1 establishes the interesting fact that, if \( p > n \) and \( n \) is large, a fixed set of covariates that (do not) co-cluster under the true process, also (do not) asymptotically co-cluster under the posterior. The key intuition is that, as with most mixture model applications, when \( n \)-dimensional objects are clustered and \( n \) is small, it is possible for the clusters to be erroneously placed too close together even if \( p \) is large. However, if \( n \) is also allowed to grow, then objects in \( \mathbb{R}^n \) eventually become well separated. Consequently, for \( n \) and \( p \) large enough, the VariScan method is able to infer the true clustering for a fixed subset of the \( p \) covariate columns. In the sequel, using synthetic datasets in Section 5, we exhibit the high accuracy of the clustering-related inferences.

In investigations where the nonlinear function \( h \) appearing in (6) has a linear representation, the maximum number of latent clusters is finite, and conditional on the true allocation of the \( p \) covariates to the clusters, Part 2 of the theorem establishes model selection and prediction consistency for the VariScan procedure, guaranteeing reliable inferences in large datasets. The proof is given in Section ?? of the Supplementary Materials.

**Theorem 4.1** To facilitate the asymptotic results, denote the covariate matrix by \( \mathbf{X}_{np} \) and the regression outcome vector by \( \mathbf{y}_n = (y_1, \ldots, y_n)' \). Suppose that under the true model, the \( p \) columns of the matrix \( \mathbf{X}_{np} \) are iid realizations of an \( n \)-variate discrete distribution \( P_0^{(n)} \) convolved with Gaussian noise. Specifically, let the \( n \)-dimensional atoms of \( P_0^{(n)} \) be denoted by \( \mathbf{v}_t^{(0)} = (v_{1t}^{(0)}, \ldots, v_{nt}^{(0)})' \) for positive integers \( t \). Since \( P_0^{(n)} \) is discrete, there exist true allocation variables \( c_1^{(0)}, \ldots, c_p^{(0)} \) mapping the covariate columns to the atoms. The covariates are then distributed as \( x_{ij} \mid c_j^{(0)} \overset{\text{indep}}{\sim} N(v_{ij}^{(0)}, \tau_0^2) \), for subjects \( i = 1, \ldots, n \), and columns \( j = 1, \ldots, p \). Suppose that the atom elements \( v_{jt}^{(0)} \) are iid \( G_0 \) having compact support on the real line.

Then there exists a sequence of numbers \( p_n \geq n \) such that, as \( n \) grows and provided \( p > p_n \), we have the following results:
1. **Clustering:** The inferences are consistent for the cluster memberships of any subset, \( I_L \), consisting of \( L < \infty \) covariate indices. That is, subject to a permutation of the cluster labels,

\[
\lim_{n \to \infty} \mathbb{P}[c_j = c_j^{(0)} \text{ for all } j \in I_L | \mathbf{y}_n, \mathbf{X}_{np}] \to 1.
\]

2. Suppose that the number of atoms of \( P_0^{(n)} \) is a finite number, \( Q_0 \). Assume that the true model for the regression outcomes is \( y_i \text{ independent } \sim N(\eta_{i}^{(0)}, \sigma_0^2) \), with \( \eta_{i}^{(0)} = \lambda_0 + \sum_{t=1}^{Q_0} \theta_{t}^{(1)} \lambda_{t}^{(1)} v_{ik}^{(0)} + \sum_{t=1}^{Q_0} \theta_{t}^{(2)} h(v_{ik}^{(0)}, \lambda_{t}^{(2)}) \), where the true regression coefficients are denoted by \( \lambda_0, \lambda^{(1)} = (\lambda_{1}^{(1)}, \ldots, \lambda_{Q_0}^{(1)})' \), and \( \lambda^{(2)} = (\lambda_{1}^{(2)}, \ldots, \lambda_{Q_0}^{(2)})' \). The triplets of indicators, \( \theta_{t} = (\theta_{t}^{(0)}, \theta_{t}^{(1)}, \theta_{t}^{(2)})' \), sum to 1 for every \( t = 1, \ldots, Q_0 \). The nonlinear function \( h \) is assumed to have a linear representation: \( h(v, \lambda) = \sum_{s=1}^{m} \lambda_{s} h_{s}(v) \), for some analytic functions \( h_{1}, \ldots, h_{m} \).

Also suppose that the cluster allocation variables are correctly inferred for the \( p \) columns. The number of detected clusters is then \( q = Q_0 \) and the allocation vector, subject to a permutation of the cluster labels, is \( c_p^{(0)} = (c_1^{(0)}, \ldots, c_p^{(0)})' \). For a detected set of indicators \( \gamma = (\gamma_{1}, \ldots, \gamma_{Q_0})' \), let the model with marginalized regression coefficients be denoted by \( M_{\gamma} \) and let the matrix of predictor variables, defined in Section 2.2, be denoted by \( U_{\gamma} \). The true model with marginalized regression coefficients is \( M_{\theta} \) and the corresponding matrix of predictors is \( U_{\theta} \).

We represent the vector of non-zero true regression coefficients by \( \lambda_{\theta} \) and this includes the intercept \( \lambda_0 \). For any model \( M_{\gamma} \) that does not contain the true model \( M_{\theta} \), suppose the true regression coefficients are such that

\[
\lim_{n \to \infty} \frac{1}{n} \lambda_{\theta}^T U_{\theta}^T (I_n - P_{\gamma}) U_{\theta} \lambda_{\theta} = b_\theta \in (0, \infty)
\]

where \( P_{\gamma} \) is the projection matrix onto the span of \( U_{\gamma} \). Recall that the predicted value of \( y_n \) by VariScan is \( \eta_n \) which is defined in \[\theta\]. Then
(a) **Model selection** is consistent, in the sense that the inferred model $\mathcal{M}_\gamma$ satisfies

$$\lim_{n \to \infty} \frac{P(\mathcal{M}_\gamma = \mathcal{M}_\theta | \mathbf{c}_p^{(0)}, \mathbf{y}_n, \mathbf{X}_{np})}{P(\mathcal{M}_\gamma = \mathcal{M}_\theta | \mathbf{c}_p^{(0)}, \mathbf{y}_n, \mathbf{X}_{np})} = 1,$$

provided (i) the true model is different from the null (no predictor) model, i.e.,

$$\sum_{i=1}^{Q_0} \vartheta_i^{(0)} < Q_0,$$

and (ii) the latent vectors are chosen as the cluster representatives in Section 2.2.

(b) **Prediction** is consistent in the following sense: given the $n$ covariate columns and the regression outcomes of the first $(n-1)$ subjects,

$$\lim_{n \to \infty} \left( E[\eta_n | \mathbf{c}_p^{(0)}, \mathbf{y}_{n-1}, \mathbf{X}_{np}] - \eta_n^{(0)} \right) = 0.$$

Although these results rely on important theoretical insights provided by (Ghosal et al., 1999) and Liang et al. (2008), they are non-trivial extensions of those works in several directions. Specifically, Part 1 of the above Theorem extends Theorem 3 of Ghosal et al. (1999) to densities on $\mathbb{R}^n$ arising as convolutions of vector locations with errors distributed as zero-mean finite normal mixtures. Parts 2a and 2b extends Theorems 3 and 4 of Liang et al. (2008) to covariates that are realizations of a latent stochastic process convolved with noise.

5 Simulation study: cluster-related inferences

We investigated the validity of Theorem 4.1 and VariScan’s accuracy as a clustering procedure using artificial datasets for which the true clustering pattern is known. For this, we simulated the covariates for $n = 50$ subjects and $p = 250$ genes from a discrete distribution convolved with Gaussian noise, and compared the co-clustering posterior probabilities of the $p$ covariates with the truth. The parameters of the true model were chosen to approximate match the corresponding estimates for the DLBCL dataset of Rosenwald et al. (2002). Specifically, for each of 25 synthetic datasets, and for the true
model’s parameter $\tau_0$ in Theorem 4.1 belonging to the range $[0.60, 0.96]$, we generated the following quantities to obtain the matrix $X$ in Step 3 below:

1. **True allocation variables:** $c_1^{(0)}, \ldots, c_p^{(0)} \sim \text{PDP}(\alpha_1, d^{(0)})$, for the true discount parameter $d^{(0)} = 0.33$ and mass parameter $\alpha_1 = 20$. The true number of clusters, $Q_0$, was thereby computed for this non-Dirichlet allocation.

2. **Latent vector elements:** For $i = 1, \ldots, n$ and $k = 1, \ldots, Q_0$, elements $v_{ik}^{(0)} \overset{iid}{\sim} G_0$, where $G_0 \sim \text{DP}(\alpha_2; U_0)$, with mass $\alpha_2 = 10$ and uniform base distribution $U_0$ on the interval $[1.4, 2.6]$.

3. **Covariates:** $x_{ij} \overset{\text{indep}}{\sim} N(v_{icj}^{(0)}, \tau_0^2)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

No responses were generated in this study, and each dataset was fit using the techniques described in Stages 1a and 1b of Section 3. As mentioned there, we computed the least-squares allocation $\hat{c}_1, \ldots, \hat{c}_p$ of the covariate columns to the clusters. We then estimated the accuracy of the least-squares allocation by the proportion of correctly clustered covariate pairs, $\hat{\kappa} = \frac{1}{(p^2)} \sum_{j_1 \neq j_2 \in \{1, \ldots, p\}} \mathcal{I}(\hat{c}_{j_1} = \hat{c}_{j_2}) = \mathcal{I}(c_{j_1}^{(0)} = c_{j_2}^{(0)})$. A high value of $\hat{\kappa}$ is indicative of VariScan’s high clustering accuracy.

For each value of $\tau_0$, the second column of Table 1 displays the percentage $\hat{\kappa}$ averaged over the 25 independent replications. We find that, for each $\tau_0$, significantly less than 5 pairs were incorrectly clustered out of the $\binom{250}{2} = 31,125$ different covariate pairs, and so $\hat{\kappa}$ was significantly greater than 0.999. The posterior inferences appear to be robust to large noise levels, i.e., large values of $\tau_0$. For every dataset, $\hat{q}$, the estimated number of clusters in the least-squares allocation was exactly equal to $Q_0$, the true number of clusters.

Accurate inferences were also obtained for the PDP discount parameter, $d \in [0, 1)$. Figure 5 plots the 95% posterior credible intervals for $d$ against different values of $\tau_0$. The posterior inferences are substantially more precise than the prior and each interval contained the true value, $d_0 = 0.33$. Furthermore, in spite of being assigned a prior probability of 0.5, there is no posterior mass allocated to Dirichlet process models. The ability of VariScan to discriminate between PDP and Dirichlet process models was evalu-
Table 1: For different values of simulation parameter $\tau_0$, column 2 displays the proportion of correctly clustered covariate pairs, with the standard errors for the 25 independent replications shown in parentheses. Column 3 presents 95% posterior credible intervals for the lower bound of the log-Bayes factor of PDP models relative to Dirichlet process models. See the text for further explanation.

The Bayes factors are significantly greater than $e^{10} = 22,026.5$ and are overwhelmingly in favor of PDP allocations, i.e., the true model.

6 Simulation study: prediction accuracy

We evaluate the operating characteristics of our methods using a simulation study based on the DLBCL dataset of Rosenwald et al. (2002). To generate the simulated data, we selected $p = 500$ genes from the original gene expression dataset of 7,399 probes, as detailed below:

1. Select 10 covariates with pairwise correlations less than 0.5 as the true predictor set, $S \subset \{1, \ldots, 500\}$, so that $|S| = 10$.

2. For each value of $\beta^* \in \{0.2, 0.6, 1.0\}$:
(a) For subjects $i = 1, \ldots, 100$, generate the failure times as follows: $t_i \sim E_i$ where $E_i$ denotes the exponential distribution with mean $\exp(\beta^* \sum_{j \in S} x_{ij})$. Note that the model used to generate the outcomes differs from VariScan assumption (6) for the log-failure times.

(b) For 20% of individuals, generate their censoring times as follows: $u_i \sim E_i \cdot I(u_i < t_i)$. Set the survival times of these individuals to $w_i = \log u_i$ and their failure statuses to $\delta_i = 0$.

(c) For the remaining individuals, set $w_i = \log t_i$ and $\delta_i = 1$.

3. Randomly assign the data from 67 individuals to the training set and assign the data from the remaining 33 individuals to the test set.

4. Assuming the AFT survival model, apply the VariScan procedure with linear splines and $m = 1$ knot per spline. Choose a single covariate from each cluster as the representative in Section 2.2. Make posterior inferences using the training data and predict the outcomes for the test cases.

We analyzed the same set of simulated data using six other techniques for gene selection with survival outcomes: lasso (Tibshirani 1997), adaptive lasso (Zou 2006), elastic net (Zou and Trevor 2005), $L_2$-boosting (Hothorn and Buhlmann 2006), random survival forests (Ishwaran et al. 2010), and supervised principal components (Bair and Tibshirani 2004), which have been implemented in the R packages glmnet, mboost, randomSurvivalForest, and superpc. The “RSF-VH” version of the random survival forests procedure was chosen because of its success in high-dimensional problems. The selected techniques are excellent examples of the three categories of approaches for small $n$, large $p$ problems (variable selection, nonlinear prediction, and regression based on lower-dimensional projections) discussed in Section 1. We repeated this procedure over fifteen independent replications.

We compared the prediction errors of the methods using the concordance error rate, which is defined as $1 - C$, where $C$ denotes the c index of Harrell et al. (1982). Let the set of
“usable” pairs of subjects be \( \mathcal{U} = \{(i, j) : w_i < w_j, \delta_i = 1\} \cup \{(i, j) : w_i = w_j, \delta_i \neq \delta_j\} \). The concordance error rate of a procedure is (May et al., 2004): 
\[
1 - C = \frac{1}{|\mathcal{U}|} \sum_{(i,j) \in \mathcal{U}} I(\bar{w}_i \geq \bar{w}_j) - \frac{1}{2|\mathcal{U}|} \sum_{(i,j) \in \mathcal{U}} I(\bar{w}_i = \bar{w}_j),
\]
where \( \bar{w}_i \) is the predicted response of subject \( i \). For example, for the VariScan procedure applied to analyze AFT survival outcomes, the predicted responses are \( \tilde{w}_i = \exp(\tilde{y}_i) \), where \( \tilde{y}_i \) is computed as in Section 3.5.

The concordance error rate measures a procedure’s probability of incorrectly ranking the failure times of two randomly chosen individuals. The accuracy of a procedure is inversely related to its concordance error rate. The measure is especially useful for comparisons because it does not rely on the survivor function, which is estimable by VariScan, but not by some of the other procedures. Figure 6 depicts boxplots of the concordance error rates of the procedures sorted by increasing order of prediction accuracy. Numerical summaries of the same error rates of the procedures are presented in Table ?? of the Supplementary Materials. We find that as \( \beta^* \) increases, the concordance error rates progressively decrease for most procedures, including VariScan. For larger \( \beta^* \), the error rates for VariScan are significantly lower than the error rates for the other methods.

In order to facilitate a more systematic evaluation, we have plotted in Figure 7 the error rates versus model sizes for the different methods, thereby providing a joint examination of model parsimony and prediction. To aid a visual interpretation, we did not include the supervised principal components method, since it performs the worst in terms of prediction and detects models that are two to four fold larger than \( L_2 \)-boosting, which typically produces the largest models among the depicted methods. The three panels...
correspond to increasing effect size, $\beta^*$. A few facts are evident from the plots. VariScan seems to balance sparsity and prediction the best for all values of $\beta^*$, with its performance increasing appreciably with $\beta^*$. Penalization approaches such as lasso, adaptive lasso, and elastic net produce sparser models but have lower prediction accuracies. $L_2$-boosting is comparable to Variscan in terms of prediction accuracy, but detects larger models for the lower effect sizes (left and middle panel); Variscan is the clear winner for the largest effect size (right panel). Additionally, especially for the largest $\beta^*$, we observe substantial variability between the simulation runs for the penalization approaches, as reflected by the large standard errors.

Averaging over the 15 independent replications of the simulation, as $\beta^*$ varied over the set $\{0.2, 0.6, 1.0\}$, the estimates of the nonlinearity measure $N$ defined in equation (9), were 0.72, 0.41, and 0.25, respectively. The corresponding standard errors were 0.04, 0.07, and 0.06. This indicates that on the scale of the simulated log–failure times, simple linear regressors are increasingly preferred to linear splines as the signal-to-noise ratio, quantified by $\beta^*$, increases. Such interpretable measures of nonlinearity are not provided by the competing methods.

7 Analysis of benchmark data sets

Returning to the two publicly available datasets of Section 1, we chose $p = 500$ probes for further analysis. For the DLBCL dataset of Rosenwald et al. (2002), we randomly selected 100 out of the 235 individuals who had non-zero survival times. Of the individuals selected, 50% had censored failure times. For the breast cancer dataset of van’t Veer et al. (2002), we analyzed the 76 individuals with non-zero survival times, of which 44 individuals (57.9%) had censored failure times.

We performed 50 independent replications of the three steps that follow. (i) We randomly split the data into training and test sets in a 2:1 ratio. (ii) We analyzed the survival times and $p = 500$ gene expression levels of the training cases using the techniques
Figure 6: Side-by-side boxplots comparing the percentage concordance error rates of the different techniques in the simulation study.
Figure 7: Plot of concordance error rates versus model sizes for the competing methods along with the standard errors (shown by whiskers). The left, middle and right respectively correspond to effect size $\beta^*$ equal to 0.2, 0.6, and 1.

VariScan, lasso, adaptive lasso, elastic net, $L_2$-boosting, random survival forests, and supervised principal components. (iii) The different techniques were used to predict the test case outcomes. For the VariScan procedure, a single covariate from each cluster was chosen to be the cluster representative.

Posterior inferences for some VariScan parameters are summarized in Table 2. The number of clusters for the least-squares allocation of covariates, $\hat{q}$, computed in Stage 1a of the analysis, is considerably smaller for the breast cancer dataset. The relatively high estimates for the nonlinearity measure $\mathcal{N}$ indicate that the responses in both datasets, but especially in the DLBCL dataset, have predominantly nonlinear relationships with the predictors. In spite of being assigned a prior probability of 0.5, the estimated posterior probability of the Dirichlet process model (corresponding to discount parameter $d = 0$) is exactly 0 for both datasets, justifying the allocation scheme in equation (2).

Figure 8 displays heatmaps for the DLBCL covariates that were allocated to column clusters having more than 10 members. As shown in Figure 4 we found that the smaller
Table 2: Posterior inferences for selected VariScan parameters.

| Parameter          | DLBCL dataset | Breast cancer dataset |
|--------------------|---------------|-----------------------|
| $\hat{q}$          | 165           | 117                   |
| $\hat{N}$          | 0.97 (0.00)   | 0.75 (0.02)           |
| $\hat{P}[d = 0| \text{data}]$ | 0             | 0                     |

clusters (not shown in Figure 8) typically have much better concordance. As we mentioned, the technique we propose shuffles the covariate matrix rows and columns locally, allowing the subjects to group differently in different clusters. The panels display the covariates before and after bidirectional clustering of the subjects and probes, with the lower panel of Figure 8 illustrating the within-cluster patterns discovered by VariScan. For each column cluster in the lower panel, the uppermost rows represent the covariates of any subjects that do not follow the cluster structure and which are better modeled as random noise (i.e., covariates with $\hat{z}_{ik} = 0$). The graphs demonstrate the effectiveness of VariScan as a model-based clustering procedure.

Comparing the test case predictions with the actual survival times, boxplots of numerical summaries of the concordance error rates for all the methods are presented in Figure 9. Numerical summaries of these error rates are computed in Table ?? of the Supplementary Materials. The success of VariScan appears to be robust to the different censoring rates of survival datasets. Although $L_2$-boosting had comparable error rates for the DLBCL dataset, VariScan had the lowest error rates for both datasets. In addition, the plots of sparsity versus prediction error rates are provided in Figure ?? of the Supplementary Materials. The plots clearly show that VariScan performs the best for both the datasets in producing highly predictive models with lower model sizes. For both the datasets, the plots demonstrate the effectiveness of VariScan in producing highly predictive models with small model sizes.

For subsequent biological interpretations, we selected genes having high probability of being selected as predictors (with the upper percentile decided by the model size). We then analyzed these genes for their role in cancer progression by cross-referencing with the
Figure 8: Heatmaps of DLBCL covariates that were assigned to latent column clusters with more than 10 members. The panels display the covariates before and after bidirectional local clustering by VariScan. The vertical lines in the bottom panel mark the covariate-clusters. The color key for both panels is displayed at the top of the plot.
Figure 9: Side-by-side boxplots of percentage concordance error rates for the benchmark datasets.
existing literature. For the breast cancer dataset, our survey indicated several prominent genes related to breast cancer development and progression, such as TGF-B2 (Buck and Knabbe 2006), ABCC3, which is known to be up-regulated in primary breast cancers, and LAPTM4B, which is related to breast carcinoma relapse with metastasis (Li et al., 2010).

For the DLBCL dataset, we found several genes related to DLBCL progression, such as the presence of multiple chemokine ligands (CXCL9 and CCL18), interleukin receptors of IL2 and IL5 (Lossos and Morgensztern 2006), and BNIP3, which is down-regulated in DLBCL and is a known marker associated with positive survival (Pike et al., 2008).

A detailed functional/mechanistic analysis of the main set of genes for both datasets is provided in Section ?? of the Supplementary Materials.

8 Conclusions

In summary, VariScan offers an efficient methodology for high-dimensional clustering, variable selection, and prediction for continuous and discrete responses. The VariScan model exploits the sparsity of PDPs as dimension-reduction devices. Specifically, the covariates are grouped into lower-dimensional latent clusters consisting of covariates having similar patterns for the subjects, and are permitted to choose between PDPs and their special case, a Dirichlet process, for a suitable cluster allocation scheme. We theoretically determine how a PDP-based clustering is able to be distinguished from a Dirichlet process in terms of the number and relative sizes of their clusters. We also provide a theoretical explanation for the ability of VariScan to detect the true allocation scheme of the covariates, and demonstrate model selection and prediction consistency.

We exploit different features of the VariScan model to develop an MCMC strategy that includes Metropolis-Hastings steps and a Gibbs sampler with efficient sequential importance sampling moves for cluster allocation. In simulations and real data analysis, we show that VariScan makes highly accurate cluster-related inferences. In predictive accuracy, the technique compares favorably with several existing methodologies for survival applications, consistently outperforming nonlinear techniques such as random survival
forests and $L_2$-boosting, as well as supervised principal components. These findings make a compelling case for the use of VariScan in high-dimensional regression settings such as genomics where it is critically important to detect predictive (or prognostic) models relying on a few, but important, genes that can be further biologically validated via functional experiments. In the analyses of benchmark microarray datasets, we identified several genes having known implications in cancer development and progression, which further engenders our hypothesis.

As discussed in Section 3, due to the intensive nature of the MCMC inference, we performed these analyses in two stages, with cluster detection followed by predictor discovery. We are currently working on implementing VariScan’s MCMC procedure in a parallel computing framework using the graphical processing units of computers. This computer code will soon be available as an R package for general purpose use. The single-stage analysis will allow the regression and clustering results to be interrelated, as implied by the VariScan model. We anticipate being able to dramatically speed up the calculations by multiple orders of magnitude, which will allow for single-stage inferences of user-specified datasets on ordinary desktop and laptop computers.

References

Albert, J. H. and Chib, S. (1993), “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88, 669–679.

— (1994), “Bayes Inference in Regression Models with ARMA(p, q) Errors,” *Journal of Econometrics*, 64, 183–206.

Albert, J. H., Chib, S., and Winkelmann, R. (1998), “Posterior Simulation and Bayes Factors in Panel Count Data Models,” *Journal of Econometrics*, 86, 33–54.

Andrews, D. F. and Mallows, C. L. (1974), “Scale Mixtures of Normal Distributions,” *Journal of the Royal Statistical Society, Series B*, 36, 99–102.

Bair, E. and Tibshirani, R. (2004), “Semi-Supervised Methods to Predict Patient Survival From Gene Expression Data,” *PLoS Biology*, 2, 511–522.

Baladandayuthapani, V., Holmes, C. C., Mallick, B. K., and Carroll, R. J. (2006), *Modeling Nonlinear Gene Interactions using Bayesian MARS*, In Do K. A., Mueller P. and Vannucci M. (eds.) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press.

Barry, D. and Hartigan, J. A. (1993), “A Bayesian analysis for change point problems,” *Journal of the American Statistical Association*, 88, 309–319.
Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2010), “Bayesian ensemble methods for survival prediction in gene expression data,” *Bioinformatics*.

Brown, P. J., Vannucci, M., and Fearn, T. (1998), “Multivariate Bayesian variable selection and prediction,” *J. R. Stat. Soc. Series B*, 60, 627–641.

Buck, M. B. and Knabbe, C. (2006), “TGF-beta signaling in breast cancer,” *Ann. N. Y. Acad. Sci.*, 1089, 119–126.

Buckley, J. and James, I. (1979), “Linear regression with censored data,” *Biometrika*, 66, 429–436.

Cai, B. and Dunson, D. (2006), “Bayesian Covariance Selection in Generalized Linear Mixed Models,” *Biometrics*, 62, 446–457.

Cox, D. and Oakes, D. (1984), *Analysis of survival data*, London: Chapman and Hall.

Cristianini, N. and Shawe-Taylor, J. (2000), *Introduction to Support Vector Machines (and other kernel-based learning methods)*, Cambridge University Press.

Crowley, E. M. (1997), “Product Partition Models for Normal Means,” *Journal of the American Statistical Association*, 92, 192–198.

Dahl, D. B. (2006), *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*, Cambridge University Press.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer Verlag.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (1982), *Bayesian variable selection using the Gibbs sampling*, Marcel Dekker, Inc., New York, pp. 273–286.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998a), “Automatic Bayesian curve fitting,” *Journal of the Royal Statistical Society, Series B*, 60, 333–350.

— (1998b), “Bayesian MARS,” *Statistics and Computing*, 8, 337–346.

Dunson, D. B., Herring, A. H., and Engel, S. M. (2008), “Bayesian selection and clustering of polymorphisms in functionally-related genes,” *Journal of the American Statistical Association*, 103, 534–546.

Dunson, D. B. and Park, J.-H. (2008), “Kernel stick-breaking processes,” *Biometrika*, 95, 307–323.

Eubank, R. (1999), *Nonparametric Regression and Spline Smoothing*, New York: Marcel Dekker.

Fan, J. and Li, R. (2002), “Variable selection for Cox’s proportional hazards model and frailty model,” *Ann. Stat.*, 30, 74–99.

Ferguson, T. S. (1973), “A Bayesian Analysis of Some Nonparametric Problems,” *Annals of Statistics*, 1, 209–223.

Fraley, C. and Raftery, A. E. (2002), “Model-based clustering, discriminant analysis, and density estimation,” *Journal of the American Statistical Association*, 97, 611–631.

Freudenberg, J. M., Sivaganesan, S., Wagner, M., and Medvedovic, M. (2010), “A semiparametric bayesian model for unsupervised differential co-expression analysis,” *BMC Bioinformatics*, 11, 234.
Friedman, J. H. (1991), “Multivariate Adaptive Regression Splines,” *Annals of Statistics*, 19, 1–141.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer.

George, E. and McCulloch, R. (1993), “Variable selection via Gibbs sampling,” *Journal of the American Statistical Association*, 88, 881–889.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), “Posterior consistency of Dirichlet mixtures in density estimation,” *The Annals of Statistics*, 27, 143–158.

Harrell, F., Califf, R., Pryor, D., Lee, K., and R., R. (1982), “Evaluating the yield of medical tests,” *J. Amer. Med. Assoc.*, 247, 2543–2546.

Hartigan, J. A. (1990), “Partition Models,” *Communications in Statistics, Part A Theory and Methods*, 19, 2745–2756.

Harville, D. A. (1977), “Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems,” *Journal of the American Statistical Association*, 72, 320–340.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized additive models*, London: Chapman & Hall.

Holmes, C. C. and Held, L. (2006), “Bayesian auxiliary variable models for binary and multinomial regression,” *Bayesian Analysis*, 1, 145–168.

Hothorn, T. and Buhlmann, P. (2006), “Model-Based Boosting in High Dimensions,” *Bioinformatics*, 22, 2828–2829.

Ishwaran, H. and James, L. F. (2003), “Generalized weighted Chinese restaurant processes for species sampling mixture models,” *Statist. Sinica*, 13, 1211–1235.

Ishwaran, H., Kogalur, U. B., et al. (2010), “High-Dimensional Variable Selection for Survival Data,” *Journal of the American Statistical Association*, 105, 205–217.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006), “Variable selection in clustering via Dirichlet process mixture models,” *Biometrika*, 93, 877–893.

Kuo, L. and Mallick, B. (1997), “Bayesian semiparametric inference for the accelerated failure time model,” *Canadian J. Stat.*, 25, 457–472.

Lee, J., Müller, P., and Ji, Y. (2013), “A Nonparametric Bayesian Model for Local Clustering,” Tech. rep., Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center.

Lee, K. and Mallick, B. (2004), “Bayesian methods for variable selection in survival models with application to DNA microarray data,” *Sankhya*, 66, 756–778.

Li, H. and Gui, J. (2004), “Partial Cox regression analysis for high-dimensional microarray gene expression data,” *Bioinformatics*, 20, 208–215.

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., Desmedt, C., Sotiriou, C., Szallasi, Z., Iglehart, J. D., Richardson, A. L., and Wang, Z. C. (2010), “Amplification of LAPT4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer,” *Nat. Med.*, 16, 214–218.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), “Mixtures of g-priors for Bayesian Variable Selection,” *Journal of the American Statistical Association*, 103, 410–423.
Lijoi, A. and Prünster, I. (2010), *Models beyond the Dirichlet process*, Cambridge Series in Statistical and Probabilistic Mathematics, pp. 80–136.

Liu, J. (2008), *Monte Carlo strategies in scientific computing*, Springer Verlag.

Lossos, I. S. and Morgensztern, D. (2006), “Prognostic biomarkers in diffuse large B-cell lymphoma,” *J. Clin. Oncol.*, 24, 995–1007.

Madigan, D. and Raftery, A. (1994), “Model selection and accounting for model uncertainty in graphical models using Occam’s window,” *Journal of the American Statistical Association*, 89, 1535–1546.

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005), “Bayesian classification of tumours by using gene expression data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 219–234.

May, M., Royston, P., Egger, M., Justice, A., and Sterne, J. (2004), “Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy,” *Statist. Medicine*, 23, 2375–2398.

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004), “Bayesian mixture model based clustering of replicated microarray data,” *Bioinformatics*, 20, 1222–1232.

Mitchell, T. J. and Beauchamp, J. J. (1988), “Bayesian Variable Selection in Linear Regression,” *Journal of the American Statistical Association*, 83, 1023–1036.

Müller, P., Quintana, F., and Rosner, G. L. (2011), “A Product Partition Model With Regression on Covariates,” *Journal of Computational and Graphical Statistics*, 20, 260–278.

Nguyen, D. and Rocke, D. (2002), “Partial least squares proportional hazard regression for application to DNA microarray survival data,” *Bioinformatics*, 18, 1625–1632.

Peduzzi, P. N., Hardy, R. J., and Holford, T. R. (1980), “A stepwise variable selection procedure for nonlinear regression models,” *Biometrics*, 36, 511–516.

Perman, M., Pitman, J., and Yor, M. (1992), “Size-biased sampling of Poisson point processes and excursions,” *Probab. Theory Related Fields*, 92, 21–39.

Petralia, F., Rao, V., and Dunson, D. B. (2012), “Repulsive Mixtures,” *ArXiv e-prints*.

Pike, B. L., Greiner, T. C., Wang, X., Weisenburger, D. D., Hsu, Y. H., Renaud, G., Wolfsberg, T. G., Kim, M., Weisenberger, D. J., Siegmund, K. D., Ye, W., Groshen, S., Mehrian-Shai, R., Delabie, J., Chan, W. C., Laird, P. W., and Hacia, J. G. (2008), “DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status,” *Leukemia*, 22, 1035–1043.

Pitman, J. (1995), “Exchangeable and partially exchangeable random partitions,” *Probab. Theory Related Fields*, 102, 145–158.

Pitman, J. and Yor, M. (1997), “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.” *Ann. Probab.*, 25, 855–900.

Quintana, F. A. (2006), “A predictive view of Bayesian clustering,” *Journal of Statistical Planning and Inference*, 136, 2407–2429.

Quintana, F. A. and Iglesias, P. L. (2003), “Bayesian clustering and product partition models,” *J. R. Statist. Soc. B*, 65, 557–574.
Rosenwald, A. et al. (2002), “The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large B-Cell Lymphoma,” *The New England Journal of Medicine*, 346, 1937–1947.

Rousseau, J. and Mengersen, K. (2011), “Asymptotic behaviour of the posterior distribution in overfitted mixture models,” *Journal of the Royal Statistical Society: Series B*, 73, 689–710.

Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statistica Sinica*, 4, 639–650.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006), “Bayesian variable selection for the analysis of microarray data with censored outcome,” *Bioinformatics*, 22, 2262–2268.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., et al. (2004), “Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage,” *Biometrics*, 60, 812–819.

Sinha, D. and Dey, D. K. (1997), “Semiparametric Bayesian analysis of survival data,” *Journal of the American Statistical Association*, 92, 1195–1212.

Tibshirani, R. (1997), “The lasso method for variable selection in the Cox model,” *Stat. Med.*, 16, 385–395.

van’t Veer, L. J. et al. (2002), “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, 415, 530–536.

Volinsky, C. et al. (1997), “Bayesian model averaging in proportional hazard models: Assessing the risk of stroke,” *App. Stat.*, 46, 433–448.

Weisberg, S. (1985), *Applied Linear Regression*, J. Wiley and Sons, NY.

West, M. (1987), “On Scale Mixtures of Normal Distributions,” *Biometrika*, 74, 646–648.

Zeger, S. L. and Karim, M. R. (1991), “Generalized linear models with random effects: A Gibbs sampling approach,” *Journal of the American Statistical Association*, 86, 79–86.

Zellner, A. (1986), *On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions*, New York: Elsevier, pp. 233–243.

Zou, H. (2006), “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Trevor, T. (2005), “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society, Series B*, 67, 301–320.