A nonparametric Bayesian technique for high-dimensional regression

Subharup Guha

Department of Statistics
307D Middlebush Hall
Columbia, MO 65211-6100
e-mail: GuhaSu@missouri.edu

and

Veerabhadran Baladandayuthapani

Department of Biostatistics
1400 Pressler Street
Houston, Texas 77030
e-mail: veera@mdanderson.org

Abstract: This paper proposes a nonparametric Bayesian framework called **VariScan** for simultaneous clustering, variable selection, and prediction in high-throughput regression settings. Poisson-Dirichlet processes are utilized to detect lower-dimensional latent clusters of covariates. An adaptive nonlinear prediction model is constructed for the response, achieving a balance between model parsimony and flexibility. Contrary to conventional belief, cluster detection is shown to be a posteriori consistent for a general class of models as the number of covariates and subjects grows. Simulation studies and data analyses demonstrate that VariScan often outperforms several well-known statistical methods.

Keywords and phrases: Dirichlet process, local clustering, model-based clustering, nonparametric Bayes, Poisson-Dirichlet process.

1. Introduction

An increasing number of studies involve the regression analysis of p continuous covariates and continuous or discrete univariate responses on n subjects, with p being much larger than n. The development of effective clustering and sparse regression models for reliable predictions is especially challenging in these "small n, large p" problems. The goal of the analysis is often three-pronged: *(i) Cluster identification:* We wish to identify clusters of covariates with similar patterns for the subjects. For example, in biomedical studies where the covariates are gene expression levels, subsets of genes associated with distinctive between-subject patterns may correspond to different underlying biological processes; *(ii) Detection of sparse regression predictors:* From the set of p covariates, we wish to select a sparse subset of reliable predictors for the subject-specific responses

and infer the nature of their relationship with the responses. In most genomic applications, just a few of the biological processes are usually relevant to a response variable of interest, and we need reliable and parsimonious regression models; and *(iii) Response prediction:* Using the inferred regression relationship, we wish to predict the responses of $\tilde{n}$ additional subjects for whom only covariate information is available. The reliability of an inferential procedure is measured by its prediction accuracy for out-of-sample individuals.

In high-throughput regression settings with continuous covariates and continuous or discrete outcomes, we propose a nonparametric Bayesian framework called **VariScan** for simultaneous clustering, variable selection, and prediction.

1.1. Motivating applications

Our methods and computational endeavors are motivated by recent high-throughput investigations in biomedical research, especially in cancer. Advances in array-based technologies allow for simultaneous measurements of biological units (e.g. genes) on a relatively small number of subjects. Practitioners wish to select important genes involved with disease processes and develop efficient prediction models for patient-specific clinical outcomes such as continuous survival times or categorical disease subtypes. The analytical challenges posed by such data include not only high-dimensionality but also the existence of considerable gene-gene correlations induced by biological interactions. In this article, we analyze gene expression profiles assessed using microarrays in patients with diffuse large B-cell lymphoma (DLBCL) [71] and breast cancer [78]. Both datasets are publicly available and have the following general characteristics: for individuals $i = 1, \ldots, n$, the data consist of mRNA expression levels $x_{i1}, \ldots, x_{ip}$ for p genes, where $n \ll p$, along with censored survival times denoted by w_i. More details, analytic results, and gains using our methods over competing approaches are discussed in Section 6.

The scope and success of the proposed methodology and its associated theoretical results extend far beyond the examples we discuss in this paper. For instance, the technique is not restricted to biomedical studies; we have successfully applied VariScan in a variety of other high-dimensional applications and achieved higher inferential gains than those achieved by existing methodologies.

1.2. Challenges in high-dimensional predictor detection

Despite the large number of existing methods related to clustering, variable selection and prediction, researchers continue to develop new methods to meet the challenges posed by newer applications and larger datasets. Predictor detection becomes particularly problematic in big datasets due to the pervasive collinearity of the covariates.

For a simple demonstration of this fact, consider a process that independently samples n-variate covariate column vectors $\boldsymbol{x}_1 \ldots, \boldsymbol{x}_p$, so that $p = 200$
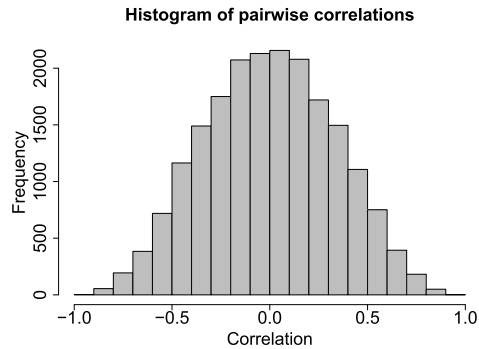
Histogram of pairwise correlations

Fig 1. *Pairwise sample correlations for $p = 200$ vectors independently generated from a multivariate normal distribution with $n = 10$ uncorrelated elements.*

vectors with $n = 10$ i.i.d. elements are generated from a common normal distribution. Although the vectors are independently generated, extreme values of the pairwise correlations are observed in the sample, as shown in the histogram of Figure 1. The proportion of extremely high or low correlations typically increases with p, and with greater correlation of the generated vectors under the true process.

Multicollinearity is common in high-dimensional problems because the n-dimensional space of the covariate columns becomes saturated with the large number of covariates. This is disadvantageous for regression because a cohort of highly correlated covariates is weakly identifiable as regression predictors. For example, imagine that the jth and k^{th} covariate columns have a sample correlation close to 1, but that neither covariate is really a predictor in a linear regression model. An alternative model in which *both* covariates are predictors with equal and opposite regression coefficients has a nearly identical joint likelihood for all regression outcomes. Consequently, an inferential procedure is often unable to choose between these competing models as the likely explanation for the data.

In the absence of strong application-specific priors to guide model selection, collinearity makes it impossible to pick the true set of predictors in high-dimensional problems. Furthermore, collinearity causes unstable inferences and erroneous test case predictions [80]. The problem is exacerbated if some of the regression outcomes are unobserved, as with categorical responses and survival applications.

1.3. *Bidirectional clustering with adaptively nonlinear functional regression and prediction*

Since the data in small n, large p regression problems are informative only about the combined effect of a cohort of highly correlated covariates, we address the issue of collinearity using clustering approaches. Specifically, VariScan utilizes

the sparsity-inducing property of Poisson-Dirichlet processes (PDPs) to first 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 selecting a suitable allocation scheme for the covariates. These partitions could provide meaningful insight into unknown biological processes (e.g., signaling pathways) represented by the latent clusters.

To flexibly capture the within-cluster pattern of the covariates, the n subjects are allowed to group differently in each cluster via a nested Dirichlet process. This feature is motivated by genomic studies [e.g., 38] that have demonstrated that subjects or biological samples often group differently under different biological processes. In essence, this modeling framework specifies a random, bidirectional (covariate, subject) nested clustering of the high-dimensional covariate matrix.

Clustering downsizes the small n, large p problem to a "small n, small q" problem, facilitating an effective stochastic search of the indices $\mathcal{S}^* \subset \{1, \ldots, q\}$ of potential *cluster predictors*. If necessary, we could then infer the indices $\mathcal{S} \subset \{1, \ldots, p\}$ of the covariate predictors. This feature differentiates the VariScan procedure from black-box nonlinear prediction methods. In addition, the technique can detect functional relationships through elements such as nonlinear functional kernels and basis functions such as splines or wavelets. An adaptive mixture of linear and nonlinear elements in the regression relationship aims to achieve a balance between model parsimony and flexibility. These aspects of VariScan define a joint model for the responses and covariates, resulting in an effective model-based clustering and variable selection procedure, improved posterior inference and accurate test case predictions.

Figure 2 illustrates the key ideas of VariScan using a toy example with $n = 10$ subjects and $p = 25$ covariates. The plot in the upper left panel represents a heatmap of the covariates. When investigators are interested in discovering a sparse prediction model for additional subjects, the posterior analysis averages over all possible realizations of two basic steps, both of which are stochastic and may be stylistically described as follows:

1. **Clustering** The column vectors are allocated in an unsupervised manner to $q = 11$ number of PDP clusters. This is plotted in the upper right panel, where the columns are grouped via bidirectional clustering to reveal the similarities in the within-cluster patterns.
2. **Variable selection and regression** One covariate is stochastically selected from each cluster and is known as the *cluster representative*. The middle right panel displays the set of representatives, $\{x_7, x_4, x_{11}, x_5, x_{24}, x_{17}, x_9, x_{12}, x_3, x_{15}, x_{14}\}$, for the 11 clusters. The regression predictors are stochastically selected from the random set of the cluster representatives. Some representatives are not associated with the response; the remaining covariates are outcome predictors and may have either a linear or nonlinear regression relationship. The linear predictors $\{x_{24}, x_{12}, x_3\}$ and non-linear
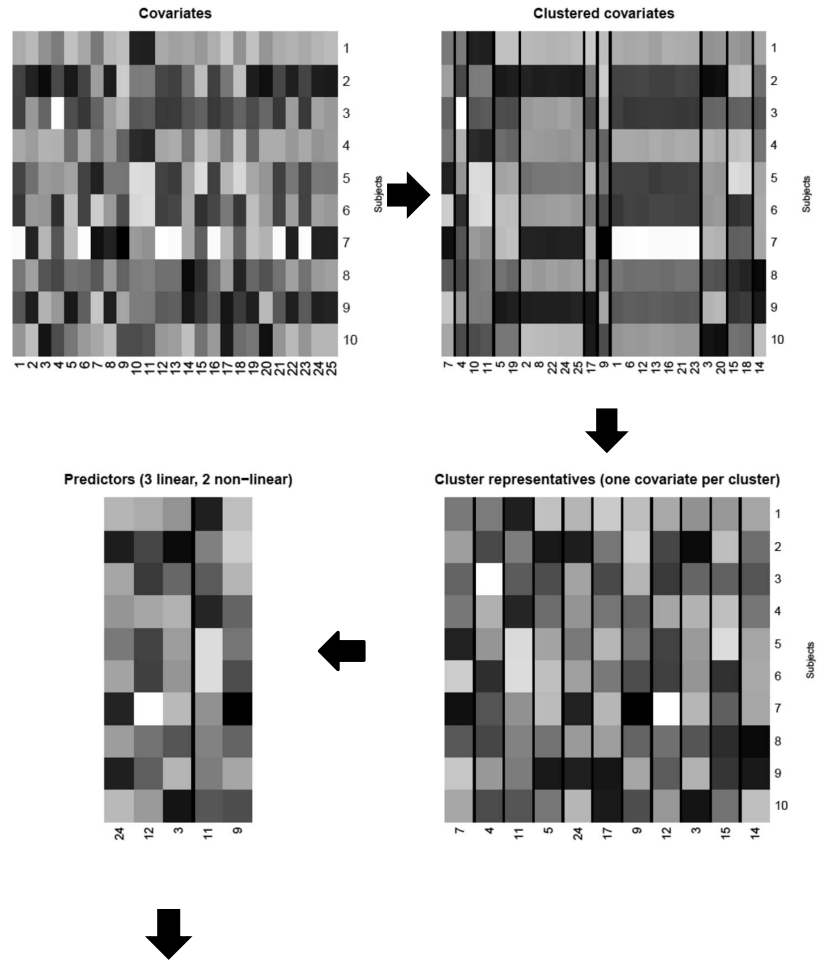
FIG 2. *Stylized example illustrating the basic methodology for reliable prediction for $n = 10$ subjects and $p = 25$ covariates allocated to $q = 11$ number of PDP clusters. The column labels represent the covariate indices. The row labels are the subjects. See the text for further explanation.*

predictors $\{\boldsymbol{x}_{11}, \boldsymbol{x}_9\}$ are shown in the middle left panel. For a nonlinear function h, the regression equation for a subject is displayed in the lower panel for a zero-mean Gaussian error, ϵ. The subscripts of the β parameters are the cluster labels, e.g., covariate $\boldsymbol{x}_{24}$ represents the fifth PDP cluster.

When out-of-the-bag prediction is not of primary interest, alternative variable selection strategies discussed in Section 2.2 may be applied.

1.4. Existing Bayesian approaches and limitations

There is vast literature on Bayesian strategies for one or more of the three inferential goals mentioned at the beginning of Section 1. A majority of Bayesian model-based clustering techniques rely on the celebrated Dirichlet process; see Müller and Mitra [61, chap. 4] for a comprehensive review. A seminal paper by [50] advocated the use of Gibbs-type priors [25, 49] for accommodating more flexible clustering mechanisms than those induced by the Dirichlet process. This work also demonstrated the practical utility of PDPs in genomic applications.

Among model-based clustering techniques based on Dirichlet processes, the approaches of [56], [13], and [60] assume that it is possible to *globally* reshuffle the rows and columns of the covariate matrix to reveal the clustering pattern. More closely related to our clustering objectives is the nonparametric Bayesian local clustering (NoB-LoC) approach of [45], which clusters the covariates *locally* using two sets of Dirichlet processes. Although some similarities exist between NoB-LoC and the clustering aspect of VariScan, there are major differences. Specifically, the VariScan framework can accommodate high-dimensional regression in addition to bidirectional clustering. Furthermore, VariScan typically produces more efficient inferences by its greater flexibility in modeling a larger class of clustering patterns via PDPs. The distinction becomes especially important for genomic datasets where PDP-based models are often preferred to Dirichlet-based models by log-Bayes factors on the order of thousands; see Section 6 for an example. Moreover, the Markov chain Monte Carlo (MCMC) implementation of VariScan explores the posterior substantially faster due to its better ability to allocate outlying covariates to singleton clusters via augmented variable Gibbs sampling. From a theoretical perspective, contrary to widely held beliefs about the non-identifiability of mixture model clusters, we discovered a remarkable theoretical property of VariScan that, as both n and p grow, a fixed set of covariates that (do not) co-cluster under the true VariScan model, also (do not) asymptotically co-cluster under its posterior.

From a regression-based Bayesian viewpoint, perhaps the most ubiquitous approaches are based on Bayesian variable selection techniques in linear and non-linear regression models. See [16] for a comprehensive review. For Gaussian responses, the common linear methods include stochastic search variable selection [23], selection-based priors [43] and shrinkage-based methods [63, 82, 26]. Some of these methods have been extended to non-linear regression contexts [74] and to generalized linear models [17, 58]. However, most of the aforementioned regression methods are based on strong parametric assumptions and do not explicitly account for the multicollinearity commonly observed in high-dimensional settings. Nonparametric approaches typically assume priors on the error residuals [28, 42] or on the regression coefficients using random effect representations [11, 53]. For nonparametric mean function estimations, they are typically based on basis function expansions such as wavelets[59] and splines [6]. We take a fundamentally different approach in this article by defining a nonparametric joint model, first on the covariates and then via an adaptive nonlinear prediction model on the responses.

The rest of this paper is organized as follows. We introduce the VariScan model in Section 2. Some theoretical results for the VariScan procedure are presented in Sections 4. Through simulations in Section 5.1 and 5.2, we demonstrate the accuracy of the clustering mechanism and compare the prediction reliability of VariScan with that of several established variable selection procedures using artificial datasets. In Section 6, we analyze the motivating gene expression microarray datasets for leukemia and breast cancer to demonstrate the effectiveness of VariScan and compare its prediction accuracy with those of competing methods. The supplementary materials contain the theorem proofs, as well as the results of additional simulation and data analyses.

2. VariScan Model

In this section, we lay out the detailed construction of the VariScan model components, which involves two major steps. First, we utilize the sparsity-inducing property of Poisson-Dirichlet processes to perform a directional nested clustering of the covariate matrix (Section 2.1). Second, we describe the choice of the cluster-specific predictors and nonlinearly relate them to Gaussian regression outcomes of the subjects (Section 2.2). Subsequently, in Section 2.3, we provide details of the model justifications and generalizations to discrete and survival outcomes.

2.1. Covariate clustering model

First, each of the p covariate matrix columns, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, is assigned to one of q latent clusters, where $q \ll p$, and where the assignments and q are unknown. That is, for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, an **allocation variable** c_j equals k if the jth covariate is assigned to the kth cluster.

We posit that the q clusters are associated with **latent vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$** of length n. The covariate columns assigned to a latent cluster are essentially contaminated versions of the cluster's latent vector and thus induce high correlations among covariates belonging to a cluster. In practice, however, a few individuals within each cluster may have highly variable covariates. We model this aspect by associating a larger error variance with those individuals. This is achieved via a Bernoulli variable, z_{ik}, for which the value $z_{ik} = 0$ indicates a high variance:

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

where τ_1^2 and τ^2 are variance parameters with inverse Gamma priors and τ_1^2 is much greater than τ^2. It is assumed that the support of τ is bounded below by a small, positive constant, τ_*. Although not necessary from a methodological perspective, this restriction guarantees the asymptotic result of Section 4. The

indicator variables for the individual–cluster combinations are a priori modeled *i.i.d* as

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

where $\xi \sim \mathrm{beta}(\iota_1, \iota_0)$. The condition $\iota_1 \gg \iota_0$ guarantees that prior probability $P(z_{ik} = 1)$ is nearly equal to 1, and so only a small proportion of the individuals have highly variable covariates within each cluster.

Allocation variables. As an appropriate model for the covariate-to-cluster allocations that accommodate a wide range of allocation patterns, we rely on the partitions induced by the *two-parameter Poisson-Dirichlet process*, $\mathbb{PDP}(\alpha_1, d)$, with discount parameter $0 \leq d < 1$ and precision or mass parameter $\alpha_1 > 0$. In genomic applications, for example, these partitions may allow for the discovery of unknown biological processes represented by the latent clusters. We defer additional details of the empirical and theoretical justifications of using PDP processes until Section 2.3.

The PDP was introduced by [65] and later investigated by [68] and [69]. Refer to [48] for a detailed discussion of different classes of Bayesian nonparametric models, including Gibbs-type priors [25, 49] such as Dirichlet processes and PDPs. [50] were the first to implement Gibbs-type priors for clustering mechanisms that are more flexible than Dirichlet process partitions.

The PDP-based allocation variables are a priori exchangeable and evolve as follows. Since the cluster allocation labels are arbitrary, we may assume without loss of generality that $c_1 = 1$, i.e., the first covariate is assigned to the first cluster. Subsequently, for covariates $j = 2, \ldots, p$, suppose there exist $q^{(j-1)}$ distinct clusters among $c_1, \ldots, c_{j-1}$, with the kth cluster containing $n_k^{(j-1)}$ number of covariates. The predictive probability that the jth covariate is assigned to the kth 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 $c_j = q^{(j-1)} + 1$ in the second line corresponds to the jth covariate opening a new cluster. When $d = 0$, we obtain the well-known Pòlya urn scheme for Dirichlet processes [21].

In general, exchangeability holds for all product partition models [7, 70] and species sampling models [36], of which PDPs are a special case. The number of clusters, q, stochastically increases as α_1 and d increase. For d fixed, the p covariates are each assigned to p singleton clusters in the limit as $\alpha_1 \to \infty$.

A PDP achieves dimension reduction in the number of covariates because q, the random number of clusters, is asymptotically equivalent to

$$\begin{cases} \alpha_1 \cdot \log p & \text{if } d = 0 \\ T_{d,\alpha_1} \cdot p^d & \text{if } 0 < d < 1 \end{cases} \tag{2.1}$$

for a random variable $T_{d,\alpha_1} > 0$ as $p \to \infty$. This implies that the number of Dirichlet process clusters (i.e., when $d = 0$) is asymptotically of a smaller

order than the number of PDP clusters when $d > 0$. This property was effectively utilized by [50] in species prediction problems and applied to gene discovery settings. The use of Dirichlet processes to achieve dimension reduction has precedence in the literature; see [57], [40], [19] and [18].

The PDP discount parameter d is given the mixture prior $\frac{1}{2}\delta_0 + \frac{1}{2}U(0,1)$, where δ_0 denotes the point mass at 0. Posterior inference of this parameter allows us to flexibly choose between Dirichlet processes and more general PDPs for the best-fitting clustering mechanism.

Latent vectors. The hierarchical prior for the covariates is completed by specifying a *base distribution* $G^{(n)}$ in $\mathcal{R}^n$ for the latent vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$. Consistent with our goal of developing a flexible and scalable inferential procedure capable of fitting large datasets, we impose additional lower-dimensional structure on the n-variate base distribution. Specifically, since the n subjects are exchangeable, base distribution $G^{(n)}$ is assumed to be the n-fold product measure of a univariate distribution, G. This allows the individuals and clusters to communicate through the nq number of latent vector elements:

$$v_{ik} \overset{iid}{\sim} G \qquad \text{for } i = 1, \ldots, n, \text{ and } k = 1, \ldots, q. \tag{2.2}$$

The unknown, univariate distribution, G, is itself given a nonparametric Dirichlet process prior, allowing the latent vectors to flexibly capture the within-covariate patterns of the subjects:

$$G \sim \mathcal{DP}(\alpha_2) \tag{2.3}$$

for mass parameter $\alpha_2 > 0$ and univariate base distribution, $N(\mu_2, \tau_2^2)$. Being a realization of a Dirichlet process, distribution G is discrete and allows the subjects to group differently in different PDP clusters. The number of distinct values among the v_{ik}'s is asymptotically equivalent to $\alpha_2 \cdot \log nq$, facilitating further dimension reduction and scalability of inference as n approaches hundreds or thousands of individuals, as commonly encountered in genomic datasets.

2.2. Prediction and regression model

Suppose there are n_k covariates allocated to the kth cluster. We posit that each cluster elects from among its member covariates a *representative*, denoted by $\boldsymbol{u}_k$. A subset of the q cluster representatives, rather than the covariates, feature in an additive regression model that can accommodate nonlinear functional relationships. The cluster representatives may be chosen in several different ways depending on the application. Possible options include the following:

(i) Select with a priori equal probability one of the n_k covariates belonging to the kth cluster as the representative. Let s_k denote the index of the covariate chosen as the representative, so that $c_{s_k} = k$ and $\boldsymbol{u}_k = \boldsymbol{x}_{s_k}$.

(ii) Set latent vector $\boldsymbol{v}_k$ of Section 2.1 as the cluster representative.

Option *(i)* is preferable when practitioners are mainly interested in identifying the effects of individual regressors, as in gene selection applications in cancer survival times (as noted in the Introduction). Option *(ii)* is preferable when the emphasis is less on covariate selection and more on identifying clusters of candidate variables (e.g., genomic pathways) that are jointly associated with the responses.

The regression predictors are selected from among the q cluster representatives, with their parent clusters constituting the set of *cluster predictors*, $\mathcal{S}^* \subset \{1, \ldots, q\}$. Extensions of the spike-and-slab approaches [23, 43, 8] are applied to relate the regression outcomes to the cluster representatives as

$$y_i \overset{indep}{\sim} N\left(\eta_i, \sigma_i^2\right), \quad \text{where}$$

$$\eta_i = \beta_0 + \sum_{k=1}^{q} \gamma_k^{(1)} \beta_k^{(1)} u_{ik} + \sum_{k=1}^{q} \gamma_k^{(2)} h(u_{ik}, \boldsymbol{\beta}_k^{(2)}) \tag{2.4}$$

and h is a nonlinear function. Possible options for the nonlinear function h in equation (2.4) include reproducible kernel Hilbert spaces [54], nonlinear basis smoothing splines [20], and wavelets. Alternatively, due to their interpretability as a linear model, order-r splines with m number of knots [14, 31, 15] are especially attractive and computationally tractable.

The linear predictor η_i in expression (2.4) implicitly relies on a vector of cluster-specific indicators, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)$, where the triplet of indicators, $\boldsymbol{\gamma}_k = (\gamma_k^{(0)}, \gamma_k^{(1)}, \gamma_k^{(2)})$, add to 1 for each cluster k. If $\gamma_k^{(0)} = 1$, the cluster representative and none of the covariates belonging to cluster k are associated with the responses. If $\gamma_k^{(1)} = 1$, the cluster representative appears as a simple linear regressor in equation (2.4); $\gamma_k^{(2)} = 1$ implies a nonlinear regressor.

The number of linear predictors, nonlinear predictors, and non-predictors are respectively, $q_1 = \sum_{j=1}^{q} \gamma_j^{(1)}$, $q_2 = \sum_{j=1}^{q} \gamma_j^{(2)}$, and $q_0 = q - q_1 - q_2$. For a simple illustration of this concept, consider again the toy example of Figure 2, where one covariate is nominated from each cluster as the representative. Of the $q = 11$ cluster representatives, $q_1 = 3$ are linear predictors, $q_2 = 2$ are nonlinear predictors, and the remaining $q_0 = 6$ representatives are non-predictors.

For nonlinear functions h having a linear representation (e.g., splines), let $\boldsymbol{U_\gamma}$ be a matrix of n rows consisting of the intercept column and the independent regressors based on the cluster representatives. For example, if we use order-r splines with m number of knots in equation (2.4), then the number of columns, $\mathrm{col}(\boldsymbol{U_\gamma}) = q_1 + (m + r) \cdot q_2 + 1$. With $[\cdot]$ denoting densities of random variables, the prior,

$$[\boldsymbol{\gamma}] \propto \omega_0^{q_0} \omega_1^{q_1} \omega_2^{q_2} \cdot \mathcal{I}\left(\mathrm{col}(\boldsymbol{U_\gamma}) < n\right), \tag{2.5}$$

where the probabilities $\boldsymbol{\omega} = (\omega_0, \omega_1, \omega_2)$ are given by the Dirichlet distribution prior, $\boldsymbol{\omega} \sim \mathcal{D}_3(1, 1, 1)$. The truncated prior for $\boldsymbol{\gamma}$ is designed to ensure model sparsity and prevent overfitting, as explained below. Conditional on the variances of the regression outcomes in equation (2.4), we postulate a weighted

g prior for the regression coefficients:

$$\boldsymbol{\beta_\gamma}|\boldsymbol{\Sigma} \sim N_{|\mathcal{S}^*|+1}\left(\mathbf{0}, \sigma_\beta^2(\boldsymbol{U_\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{U_\gamma})^{-1}\right), \tag{2.6}$$

where matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

A schematic representation of the entire hierarchical model involving both the clustering and prediction components is shown in Figure 3.

2.3. Model justification and generalizations

In this section, we discuss the justification, consequences, and generalizations of different aspects of the VariScan model. In particular, we investigate the appropriateness of PDPs in this application as a tool for covariate clustering. We also discuss the choice of basis functions for the nonlinear prediction model and consider generalizations to discrete and survival outcomes.

Empirical justification of PDPs. We conducted an exploratory data analysis (EDA) of the gene expression levels in the DLBCL dataset of Rosenwald et al. [71]. Randomly selecting a set of $p = 500$ probes for $n = 100$ randomly chosen individuals, we iteratively applied the k-means procedure until the covariates were grouped into fairly concordant clusters with a small overall value of τ^2. The allocation pattern depicted in Figure 4 is atypical of Dirichlet processes which, as is well known among practitioners, are usually associated with a relatively small number of clusters and exponentially decaying cluster sizes. Instead, the large number of clusters ($\hat{q} = 161$) and the predominance of small clusters suggest a non-Dirichlet model for the covariate-cluster assignments. More specifically, a PDP is favored due to the slower, power law decay in the cluster sizes typically associated with these models.

Theoretical justifications for a PDP model. [73] derived the *stick-breaking representation* for a Dirichlet process, and then [68] extended it to PDPs. These stick-breaking representations have the following consequences for the induced partitions. Let $\mathbb{N}$ be the set of natural numbers. Subject to a one-to-one mapping of the PDP cluster labels into set $\mathbb{N}$, the allocation variables $c_1, \ldots, c_p$ may be regarded as 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 $V_h \stackrel{indep}{\sim} \mathrm{beta}(1 - d, \alpha_1 + hd)$. 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 π_{h_k} for some distinct integers $h_1, \ldots, h_q$.

As previously mentioned, the VariScan model assumes that the base distribution $G^{(n)}$ of the PDP is the n-fold product measure of a univariate distribution, G, which follows a Dirichlet process with mass parameter α_2. This bidirectional clustering structure has some interesting consequences. Let $\{\phi_h\}_{h=1}^\infty$ be the stick-breaking probabilities associated with this nested Dirichlet process. For two or
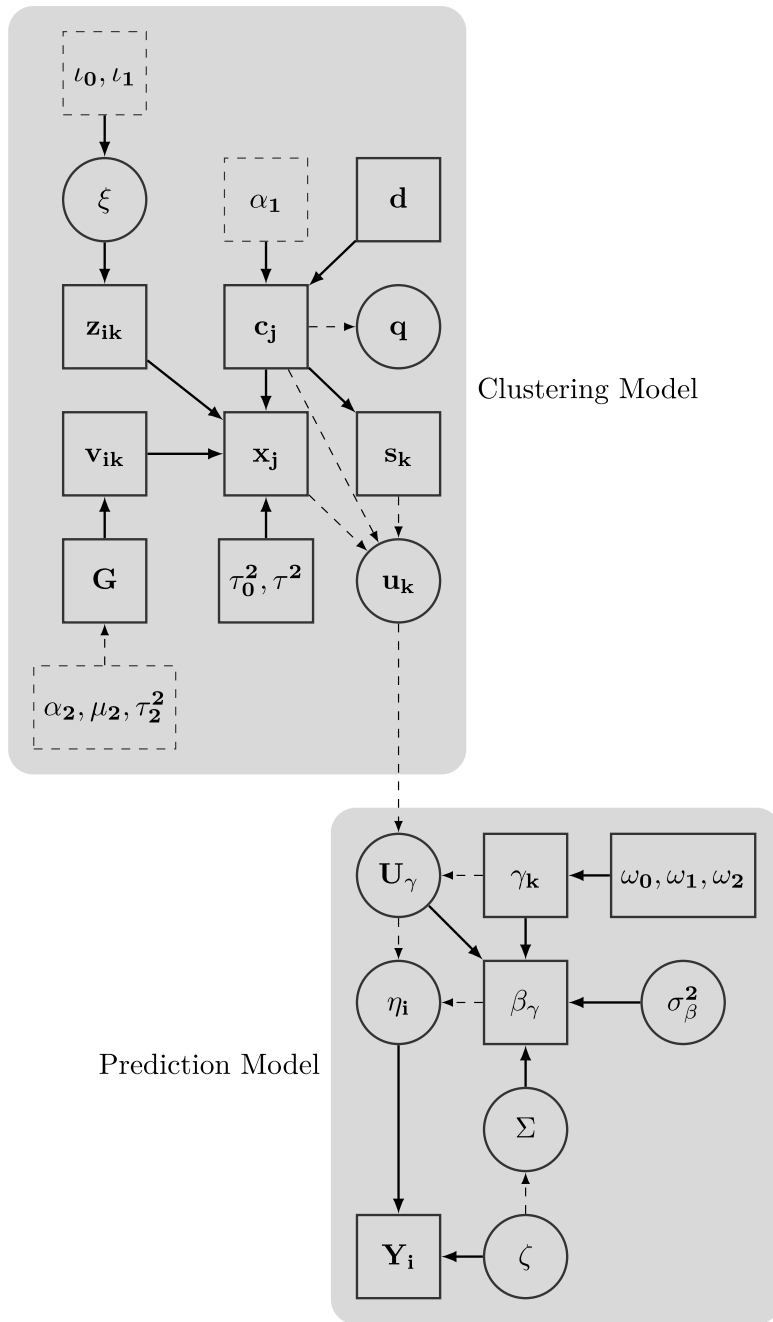
FIG 3. *Directed acyclic graph of the VariScan model in which the cluster representatives are chosen from the set of co-clustered covariates. Circles represent stochastic model parameters, solid rectangles represent data and deterministic variables, and dashed rectangles represent model constants. Solid (dashed) arrows represent stochastic (deterministic) relationships.*
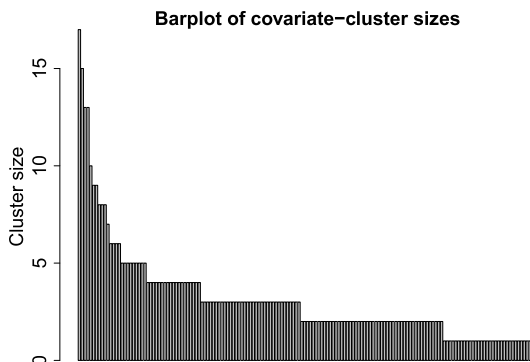
Barplot of covariate−cluster sizes

FIG 4. *Barchart of cluster sizes obtained by exploratory data analysis.*

more of the q PDP clusters, the latent vectors are identical, with a probability bounded above by $\binom{q}{2} \cdot \left(\sum_{h=1}^{\infty} \phi_h^2\right)^n$. Applying the asymptotic relationship of p and q given in expression (2.1), we find that the upper bound tends to 0 as the dataset grows, provided p grows at a slower-than-exponential rate as n grows. In fact, for n as small as 50 and p as small as 250, in simulations as well as in data analyses, we found all the latent vectors associated with the PDP clusters to be distinct. Consequently, from a practical standpoint, the VariScan allocations may be regarded as clusters with unique characteristics in even moderate-sized datasets.

Theorem 2.1 below provides formal expressions for the first and second moments of the random log-probabilities of the discrete distribution $F_{\alpha_1,d}$. In conjunction with equation (2.1), this result justifies the use of PDPs when the observed number of clusters is large or the cluster sizes decay slowly. Part 2c 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 1c suggests that in PDPs with $d > 0$ (i.e., non-Dirichlet realizations), there is a slower, power law decay of the cluster sizes as d increases. Part 3 indicates that for every α_1 and $d > 0$, a PDP realization $F_{\alpha_1,d}$ has a thicker tail compared to that of a Dirichlet process realization, $F_{\alpha_1,0}$. See Section A.1 of the Appendix for the proof.

It should be noted that the differential allocation patterns of PDPs and Dirichlet processes are well known, and have been previously emphasized in several papers, including [49] and [50]. However, it is difficult to come across a formal proof for this differential behavior. Although the theorem is primarily of interest when the base measure is non-atomic, it is relevant in this application because of the empirically observed uniqueness of the latent vectors in high-dimensional settings due to VariScan's nested structure.

Theorem 2.1. *Consider the process* $\mathbb{PDP}(\alpha_1, d)$ *with mass parameter* $\alpha_1 > 0$ *and discount parameter* $0 \le d < 1$. *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 $0 < d < 1$, the distribution $F_{\alpha_1,d} \in \mathbb{N}$ is a realization of a PDP with stick-breaking probabilities π_h, where $h \in \mathbb{N}$. However, $F_{\alpha_1,d}$ is not a Dirichlet process realization because $d \neq 0$. Then

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

   (b) $Var(\log \pi_h) = \psi_1(1-d) - \psi_1(\alpha_1) + \frac{1}{d^2}\big(\psi_1(\alpha_1/d) - \psi_1(\alpha_1/d + h)\big)$. 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.

2. For $d = 0$, the distribution $F_{\alpha_1,0} \in \mathbb{N}$ is a Dirichlet process realization with stick-breaking probabilities π_h^* based on $V_h^* \overset{iid}{\sim} beta(1, \alpha_1)$ 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) $Var(\log \pi_h^*) = \psi_1(1) - \psi_1(\alpha_1) + h/\alpha_1^2$. Thus, $\lim_{h\to\infty} Var(\log \pi_h^*) = \infty$.

   (c) As $h \to \infty$, $\sqrt{h}\left(\frac{1}{h}\log(\pi_h^*) + 1/\alpha_1\right) \xrightarrow{L} N(0, 1/\alpha_1^2)$. This implies that as $h \to \infty$, the random stick-breaking Dirichlet process probabilities, π_h^*, are stochastically equivalent to e^{-h/α_1}.

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, π_h^*/π_h, are stochastically equivalent to e^{-h/α_1} for every $d > 0$.

Remark. By Lemma 1 of [36], $\lim_{h\to\infty} E(\log \pi_h^*) = -\infty$ in Part 2a of Theorem 2.1 implies that $\sum_{h=1}^{\infty} \pi_h^* = 1$ almost surely for a Dirichlet process. A similar comment applies in Part 1a for a PDP.

Empirical justification of nested Dirichlet process model for the latent vector elements. For the DLBCL dataset, Figure 5 presents a summary of the VariScan model estimates for the 14,000 latent vector elements with estimated Bernoulli indicators $\hat{z}_{ik} = 1$. More than 87% of the $n\hat{q} = 16,500$ latent vector elements were estimated to 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. Further details about the inferential procedure are provided in Section 3. In Figure 5, the small number of clusters corresponding to the large number of latent vector elements and the sharp decline in the cluster sizes compared with Figure 4 are consistent with Dirichlet process allocation patterns. Similar results were obtained for the breast cancer data and for other genomic datasets that we have analyzed.

Choice of basis functions: model parsimony versus flexibility. The reliability of inference and prediction rapidly deteriorates as the number of cluster predictors and additive nonlinear components in equation (2.4) increases beyond
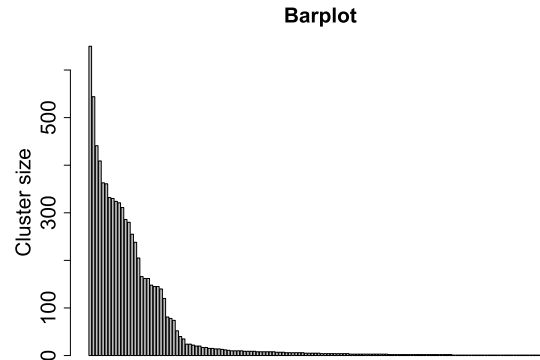
Barplot

FIG 5. *For the DLBCL dataset, least-squares Dirichlet process configuration of the more than 14,000 latent vector elements with Bernoulli indicators equal to 1.*

a threshold value and approaches the number of subjects, n. The restriction in the prior (2.5) prevents over-fitting. It ensures that the matrix U_{γ}, 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 β_{γ} in equation (2.4). In spline-based models, the relatively small number of subjects also puts a constraint on the order of the splines, often necessitating the use of linear splines with $m = 1$ knot per cluster in equation (2.4). In the applications presented in this paper, we fixed the knot for each covariate at the sample median.

Unusually small values of σ_i^2 in equation (2.4) 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, linear regression assumes that $\sigma_i^2 = \sigma^2$. We have found that non-informative priors for σ^2 do not work well because the optimal model sizes for variable selection are unknown. Additionally, we have found that it is helpful to restrict the range of σ^2 based on reasonable goals for inferential precision. In the examples discussed in this paper, we assigned the following truncated prior: $\sigma^{-2} \sim \chi_{\nu}^2 \cdot \mathcal{I}\left(0.95^{-1}/\mathrm{Var}(\hat{y}) < \sigma^{-2} < 0.5^{-1}/\mathrm{Var}(\hat{y})\right)$, where the degrees of freedom ν 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 σ^{-2} approximately corresponds to the constraint, $0.5 < R^2 < 0.95$, quantifying the effectiveness of regression. As Sections 5.2 and 6 demonstrate, the aforementioned strategies often result in high reliability of the response predictions.

Generalizations for discrete or survival outcomes. In a general investigation, the subject-specific responses may be discrete or continuous, and/or may be censored. In such cases, the responses, denoted by $w_1, \ldots, w_n$, can be modeled as deterministic transformations of random variables R_i from an exponential family distribution. The Laplace approximation [30] transforms each

R_i into a Gaussian *regression outcome*, y_i, that can be modeled using our VariScan model proposed above. The details of the calculation are as follows. For a set of functions f_i, we assume that $w_i = f_i(R_i)$ and density function $[R_i \mid \varrho_i, \varsigma] = r(R_i, \varsigma) \cdot \exp\left(\frac{R_i \, \varrho_i - b(\varrho_i)}{a(\varsigma)}\right)$, where $r(\cdot)$ is a non-negative function, ς is a dispersion parameter, ϱ_i is the canonical parameter, and $[\cdot]$ represents densities with respect to a dominating measure. The Laplace approximation relates the R_i's to Gaussian regression outcomes: $y_i = \eta_i + \frac{\partial \eta_i}{\partial \mu_i} \cdot (R_i - \mu_i)$ is approximately $N\left(\eta_i, \sigma_i^2\right)$ with precision $\sigma_i^{-2} = \{b''(\mu_i)\}^{-1} \left(\partial \mu_i / \partial \eta_i\right)^2$. For an appropriate link function $g(\cdot)$, the mean η_i equals $g(\mu_i)$. Gaussian, Poisson, and binary responses are applicable in this setting. Accelerated failure time (AFT) censored outcomes [10, 12] also fall into this modeling framework.

The idea of using a Laplace-type approximation for exponential families is not new. Some early examples in Bayesian settings include [83], [2], and [3]. 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 can be filtered through a Metropolis-Hastings step to obtain samples from the target posterior. Alternatively, inferential 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, [1] used truncated normal sampling to obtain a probit model for binary responses, and [32] utilized a scale mixture representation of the normal distribution [4, 81] to implement logistic regression using latent variables.

3. Posterior inference

Starting with an initial configuration obtained by a naïve, preliminary analysis, the model parameters are iteratively updated by MCMC methods. Due to the intensive nature of the posterior inference, the analysis is performed in two stages, with cluster detection followed by predictor discovery:

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

Stage 1a The allocation variables, latent vector elements, and binary indicators are iteratively updated 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 [13], these pairwise probabilities are used to compute a point estimate, called the *least-squares allocation*, for the allocation variables. Further details of the MCMC procedure are provided in Sections C.1 and C.2 of the Appendix.

Stage 1b Conditional on the least-squares allocation being the true clustering of the covariates, a second MCMC sample of the latent vector elements and binary indicators is generated. Again applying the technique of [13], we compute a point estimate, called the *least-squares configuration*, for the latent vector elements and binary indicators.

Stage 2 Conditional on the least-squares allocation and least-squares configuration, and focusing on the responses, the cluster predictors and latent regression outcomes, if any, are generated to obtain a third MCMC sample. The MCMC sample is post-processed to predict the responses for out-of-the-bag test set individuals. The interested reader is referred to Sections C.3, C.4 and C.5 of the Appendix for details.

As a further benefit of a coherent model for the covariates, VariScan is able to perform model-based imputations of any missing subject-specific covariates as part of the MCMC procedure.

4. Clustering consistency

As mentioned in Section 2.1, the latent vectors associated with two or more PDP clusters may be identical under the VariScan model, but this probability becomes vanishingly small as n grows. Consequently, for practical purposes, the VariScan allocations may be interpreted as distinct, identifiable clusters in even moderately large datasets. In order to study the reliability of VariScan's clustering procedure in our targeted big data applications, we make large-sample theoretical comparisons between the VariScan model's cluster allocations and the true allocations of a hypothetical covariate generating process.

In the 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 22]. Some partial solutions are available in the Bayesian literature. For example, in finite mixture models, rather than assuming exchangeability of the mixture component parameters, [66] regard them as draws from a repulsive process, leading to fewer, better separated and more interpretable clusters. [72] 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 to the non-identifiability of the detected clusters in fixed n settings, Theorem 4.1 establishes the interesting fact that, when p and n are both 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. On the other hand, if n is allowed to grow with p, then objects in $\mathcal{R}^n$ 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 5.1, we exhibit the high accuracy of VariScan's clustering-related inferences.

The true model. The VariScan model's exchangeability assumption for the p covariates stems from our belief in the existence of a true, unknown de Finetti density in $\mathcal{R}^n$ from which the column vectors arise as a random sample. In particular, for any given n and p, we make the following assumptions about the true covariate-generating process:

(a) The column vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are a random sample of size p from an n-variate distribution $P_0^{(n)}$ convolved with n-variate, independent-component Gaussian errors.

(b) The true distribution $P_0^{(n)}$ is discrete in the space $\mathcal{R}^n$. Let the n-dimensional atoms of $P_0^{(n)}$ be denoted by $\boldsymbol{v}_t^{(0)} = (v_{1t}^{(0)}, \ldots, v_{nt}^{(0)})'$ for positive integers t.

(c) Due to the discreteness of distribution $P_0^{(n)}$, there exist true allocation variables, $c_1^{(0)}, \ldots, c_p^{(0)}$, mapping the p covariates to distinct atoms of $P_0^{(n)}$. For subjects $i = 1, \ldots, n$, and columns $j = 1, \ldots, p$, the covariates are then distributed as

$$x_{ij} \mid c_j^{(0)} \overset{indep}{\sim} N(v_{i\,c_j^{(0)}}^{(0)}, \tau_0^2), \tag{4.1}$$

(d) The n-variate atoms of distribution $P_0^{(n)}$ are i.i.d. realizations of the n-fold product measure of a univariate distribution, G_0. Consequently, the atom elements are $v_{it}^{(0)} \overset{i.i.d.}{\sim} G_0$ for $i = 1, \ldots, n$, and $j = 1, \ldots, p$.

(e) The true distribution G_0 is non-atomic and has compact support on the real line.

Let $\mathcal{L} = \{j_1, \ldots, j_L\} \subset \{1, \ldots, p\}$ be a fixed subset of L covariate indexes. Given a vector of inferred allocations $\boldsymbol{c} = (c_1, \ldots, c_p)$, we quantify the inferential accuracy by the *proportion of correctly clustered covariate pairs*:

$$\varkappa_{\mathcal{L}}(\boldsymbol{c}) = \frac{1}{\binom{L}{2}} \sum_{j_1 \neq j_2 \in \mathcal{L}} \mathcal{I}\left(\mathcal{I}(c_{j_1} = c_{j_2}) = \mathcal{I}(c_{j_1}^{(0)} = c_{j_2}^{(0)}) \right). \tag{4.2}$$

A value near 1 indicates high accuracy of inferred allocations $\boldsymbol{c}$ for the set $\mathcal{L}$. Notice that the measure $\varkappa_{\mathcal{L}}(\boldsymbol{c})$ is invariant to permutations of the clusters labels. This is desirable because the labels are arbitrary.

Theorem 4.1. *Denote the covariate matrix by $\boldsymbol{X}_{np}$. In addition to assumptions (a)–(e) about the true covariate-generating process, suppose that the true standard deviation τ_0 in equation (4.1) is bounded below by τ_*, the small, positive constant postulated in Section 2.1 as a lower bound for the VariScan model parameters, τ_1 and τ.*

Let $\mathcal{L} = \{j_1, \ldots, j_L\} \subset \{1, \ldots, p\}$ be a fixed subset of L covariate indexes. Then there exists an increasing sequence of numbers $\{p_n\}$ such that, as n grows and provided $p > p_n$, the VariScan clustering inferences for the covariate subset $\mathcal{L}$ are a posteriori consistent. That is,

$$\lim_{\substack{n \to \infty \\ p > p_n}} P\left[\varkappa_{\mathcal{L}}(\boldsymbol{c}) = 1 \mid \boldsymbol{X}_{np} \right] \to 1.$$
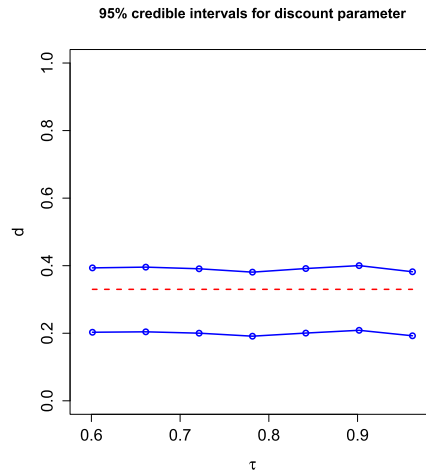
95% credible intervals for discount parameter

Fig 6. *95% posterior credible intervals for the discount parameter, d for different values of τ_0. The true value, d_0, is shown by the red dashed line.*

See Section E of the Appendix for the proof. The results rely on non-trivial extensions, in several directions, of the important theoretical insights provided by [24]. Specifically, we extend Theorem 3 of [24] to densities on $\mathcal{R}^n$ arising as convolutions of vector locations with errors distributed as zero-mean finite normal mixtures.

5. Simulation studies

5.1. Cluster-related inferences

We investigated VariScan's accuracy as a clustering procedure using artificial datasets for which the true clustering pattern is known. We simulated the co-variates for $n = 50$ subjects and $p = 250$ genes from a discrete distribution convolved with Gaussian noise, and compared the co-clustering posterior prob-abilities of the p covariates with the truth. The parameters of the true model were chosen to approximately match the corresponding estimates for the DL-BCL dataset of [71]. Specifically, for each of 25 synthetic datasets, and for the true model's parameter τ_0 in Theorem 4.1 belonging to the range $[0.60, 0.96]$, we generated the following quantities to obtain the matrix $\boldsymbol{X}$ in step 3 below.

1. **True allocation variables:** We generated $c_1^{(0)}, \ldots, c_p^{(0)}$ as the partitions induced by a PDP with true discount parameter $d^{(0)} = 0.33$ and mass pa-rameter $\alpha_1 = 20$. 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 $G_0 \sim \mathcal{DP}(\alpha_2)$, with mass $\alpha_2 = 10$ and uniform base distribution U_0 on the interval $[1.4, 2.6]$.

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

No responses were generated in this study. Applying the general technique of [13] developed for Dirichlet process models, we computed a point estimate for the allocations, called the *least-squares configuration*, and denoted by $\hat{c}_1, \ldots, \hat{c}_p$. For the full set of covariates, we estimated the accuracy of the least-squares allocation by the *estimated proportion of correctly clustered covariate pairs*,

$$\hat{\varkappa} = \frac{1}{\binom{p}{2}} \sum_{j_1 \neq j_2 \in \{1, \ldots, p\}} \mathcal{I}\bigg(\mathcal{I}(\hat{c}_{j_1} = \hat{c}_{j_2}) = \mathcal{I}(c_{j_1}^{(0)} = c_{j_2}^{(0)}) \bigg).$$

A high value of $\hat{\varkappa}$ is indicative of VariScan's high clustering accuracy for all p covariates.

For each value of τ_0, the second column of Table 1 displays the percentage $\hat{\varkappa}$ averaged over the 25 independent replications. We find that, for each τ_0, significantly less than 5 pairs were incorrectly clustered out of the $\binom{250}{2} = 31{,}125$ different covariate pairs, and so $\hat{\varkappa}$ was significantly greater than 0.999. The posterior inferences appear to be robust to large noise levels, i.e., large values of τ_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. Recall that the non-atomicity of true distribution G_0 is a sufficient condition of Theorem 4.1. Although the condition is not satisfied in this setting, we nevertheless obtained highly accurate clustering-related inferences for the full set of $p = 250$ covariates.

Accurate inferences were also obtained for the PDP discount parameter, $d \in [0, 1)$. Figure 6 plots the 95% posterior credible intervals for d against different values of τ_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 evaluated using the log-Bayes factor, $\log(P[d > 0|\boldsymbol{X}]/P[d = 0|\boldsymbol{X}])$. With Θ^* representing all the parameters except d, and applying Jensen's inequality, the log-Bayes factor exceeds

TABLE 1
For different values of simulation parameter τ_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.

True τ_0	**Percent $\hat{\varkappa}$**	**95% C.I. for lower bound of log-BF**
0.60	99.984 (0.000)	(11.05, 11.10)
0.66	99.978 (0.000)	(11.17, 11.25)
0.72	99.976 (0.000)	(10.89, 10.98)
0.78	99.973 (0.001)	(10.23, 10.31)
0.84	99.971 (0.000)	(10.86, 10.93)
0.90	99.960 (0.000)	(11.88, 11.94)
0.96	99.941 (0.001)	(10.49, 10.56)

$E(\log(\frac{P[d>0|\boldsymbol{X},\Theta^*]}{p[d=0|\boldsymbol{X},\Theta^*]}) \mid \boldsymbol{X})$, which (unlike the log-Bayes factor) can be estimated using just the post–burn-in MCMC sample. For each τ_0, the third column of Table 1 displays 95% posterior credible intervals for this lower bound. 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.

5.2. *Prediction accuracy*

We evaluate the operating characteristics of our methods using a simulation study based on the DLBCL dataset of [71]. 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, $\mathcal{S} \subset \{1, \ldots, 500\}$, so that $|\mathcal{S}| = 10$.
2. For each value of $\beta^* \in \{0.2, 0.6, 1.0\}$:
   (a) For subjects $i = 1, \ldots, 100$, generate failure times t_i from distribution $\mathcal{E}_i$, denoting the exponential distribution with mean $\exp(\beta^* \sum_{j \in \mathcal{S}} x_{ij})$. Note that the model used to generate the outcomes differs from VariScan assumption (2.4) for the log-failure times.
   (b) For 20% of individuals, generate their censoring times as follows: $u_i \sim \mathcal{E}_i \cdot \mathcal{I}(u_i < t_i)$. 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 67 individuals to the training set and 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 as described 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 [76], adaptive lasso [84], elastic net [85], L_2-boosting [33], random survival forests [37], and supervised principal components [5], which have been implemented in the R packages glmnet, mboost, randomSurvivalForest, and superpc. The "RSF-VH" version of the random survival forest 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 [29]. Let the set of "usable" pairs of subjects be $\mathcal{U} = \{(i, j) : w_i < w_j, \delta_i = 1\} \cup$
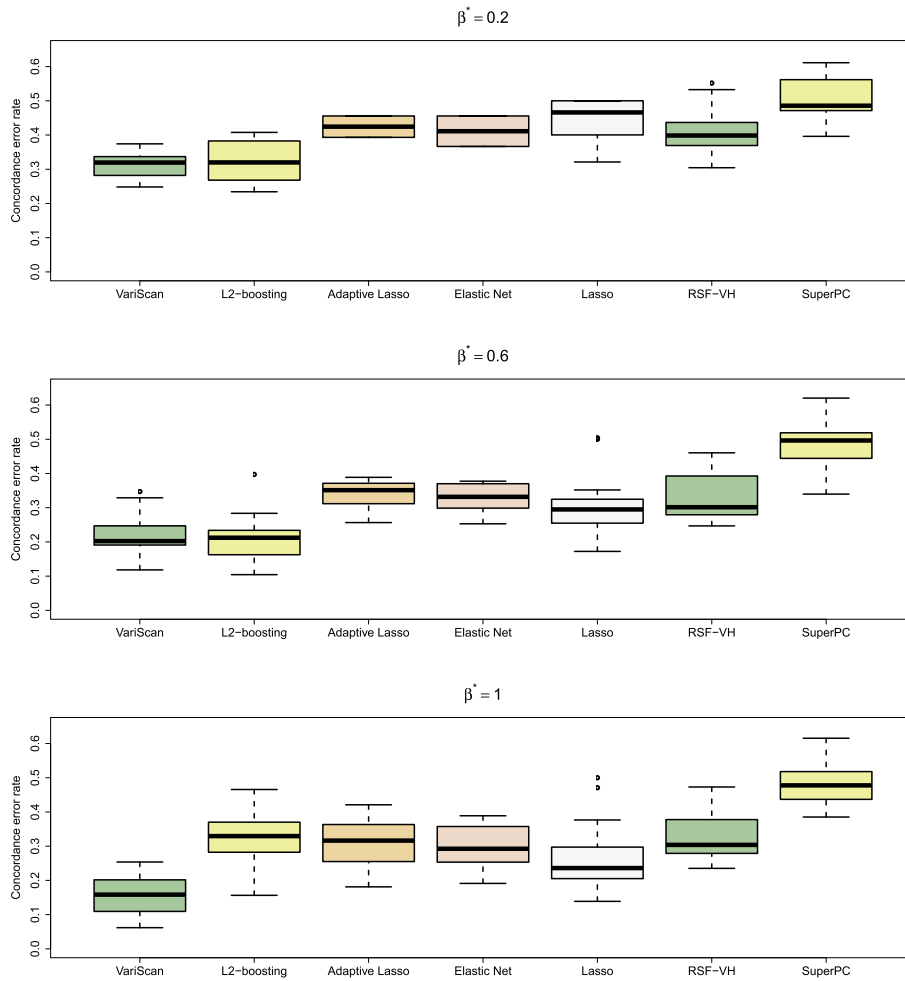
FIG 7. *Side-by-side boxplots comparing the percentage concordance error rates of the different techniques in the simulation study.*

$\{(i,j) : w_i = w_j, \delta_i \neq \delta_j\}$. The concordance error rate of a procedure is [55] $1 - C = \frac{1}{|\mathcal{U}|} \sum_{(i,j) \in \mathcal{U}} \mathcal{I}(\tilde{w}_i \geq \tilde{w}_j) - \frac{1}{2|\mathcal{U}|} \sum_{(i,j) \in \mathcal{U}} \mathcal{I}(\tilde{w}_i = \tilde{w}_j)$, where $\tilde{w}_i$ 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 $\tilde{y}_i$ are the predicted regression outcomes.

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 7 depicts boxplots of the concordance error rates of the procedures sorted
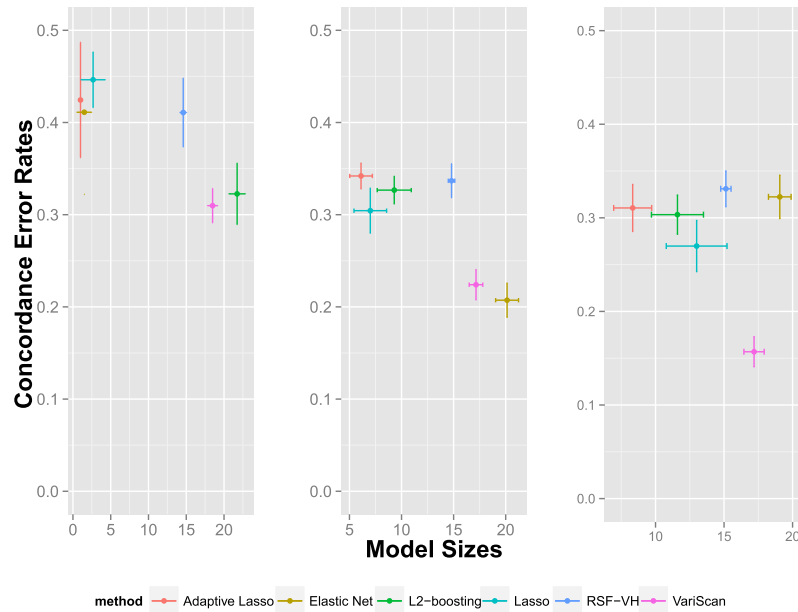
FIG 8. *Plot of concordance error rates versus model sizes for the competing methods along with the standard errors (shown by whiskers). The left, middle and right plots respectively correspond to effect size β^* equal to 0.2, 0.6, and 1.*

by increasing order of prediction accuracy. We find that as β^* increases, the concordance error rates progressively decrease for most procedures, including VariScan. For larger β^*, 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 8 the error rates versus model sizes for the different methods, thereby providing a joint examination of model parsimony and prediction. To aid 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 those from L_2-boosting, which typically produces the largest models among the depicted methods. The three panels correspond to increasing effect size, β^*. A few facts are evident from the plots. VariScan seems to balance sparsity and prediction the best for all values of β^*, with its performance increasing appreciably with β^*. 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 β^*, we observe substantial variability between the simulation runs for the penalization approaches, as reflected by the large standard errors. Further simulation study comparisons of VariScan and the competing approaches are presented in Section F.1 of the Appendix.

Nonlinearity measure. Unlike some existing approaches, VariScan is able to measure the degree of nonlinearity in the relationships between the responses and covariates. For example, we could define *nonlinearity measure* $\mathcal{N}$ as the posterior expectation,

$$\mathcal{N} = E\Big(\frac{\omega_2}{\omega_1 + \omega_2}\Big|\boldsymbol{w}, \boldsymbol{X}\Big). \tag{5.1}$$

This represents the posterior odds that a hypothetical, new cluster is a nonlinear predictor in equation (2.4) rather than a simple linear regressor. A value of $\mathcal{N}$ close to 1 corresponds to predominantly nonlinear associations between the responses and their predictors.

Averaging over the 15 independent replications of the simulation, as β^* varied over the set $\{0.2, 0.6, 1.0\}$, the estimates of the nonlinearity measure $\mathcal{N}$ defined in equation (5.1), 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 β^*, increases. Such interpretable measures of nonlinearity are not provided by the competing methods.

6. 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. [71], 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. [78], 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 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.

The numbers of clusters for the least-squares allocation of covariates, $\hat{q}$, computed in stage 1a of the analysis, were 165 and 117, respectively, for the DLBCL and breast cancer datasets. The nonlinearity measure $\mathcal{N}$ estimates were 0.97 and 0.75, respectively, with small standard errors. This indicates 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 $d = 0$) was exactly 0 for both datasets, justifying the PDP-based allocation scheme.

For the DLBCL data, the upper panel of Figure 9 displays the estimated posterior density of the PDP's discount parameter d. The estimated posterior

Posterior density of PDP parameter d

Histogram of number of clusters q

Barplot
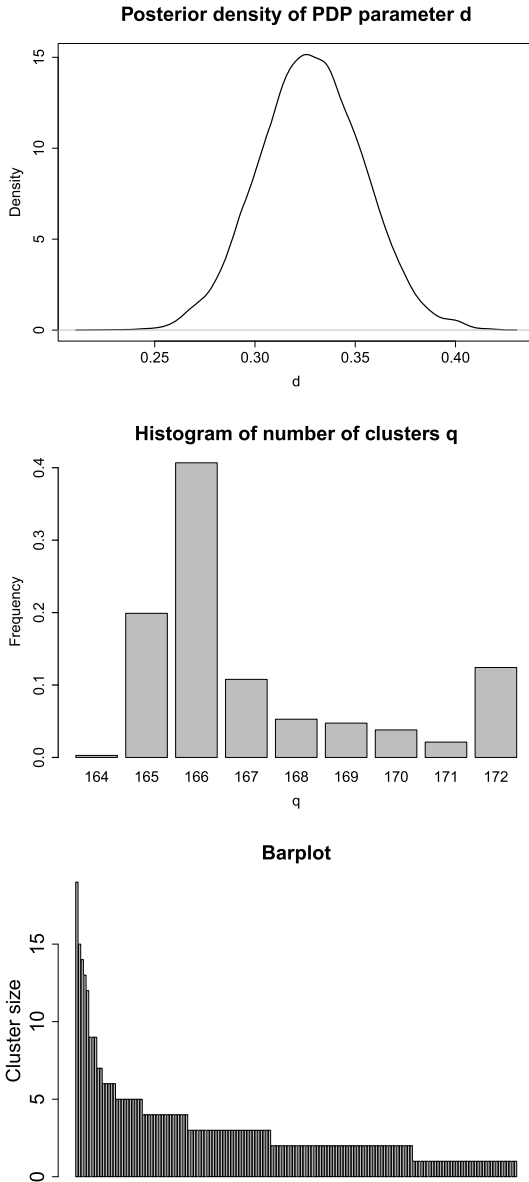
FIG 9. *Posterior summaries for the DLBCL dataset. The top panels and the lower panel summarize the least-squares covariate-to-cluster PDP allocation of the 500 genes.*

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 middle panel of Figure 9 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
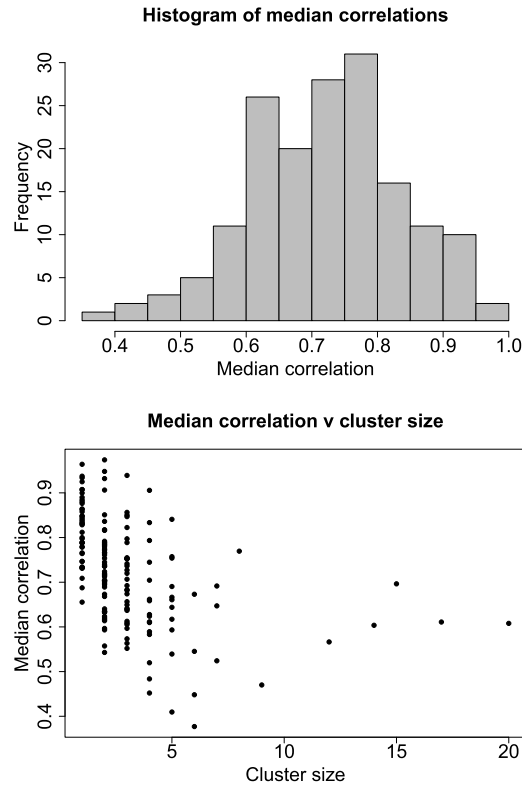
FIG 10. *For the DLBCL dataset, median pairwise correlations for the $\hat{q} = 165$ PDP clusters in the least-squares allocation of stage 1a.*

process model). The lower panel of Figure 9 summarizes the cluster sizes of the least-squares allocation [13]. 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.

The effectiveness of VariScan as a model-based clustering procedure can be shown as follows. 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 10. The plots reveal fairly good within-cluster concordance regardless of the cluster size. Figure 11 displays heatmaps for the DLBCL covariates that were allocated to column clusters having more than 10 members. The panels display the covariates before and after bidirectional clustering of the subjects and probes, with the lower panel of Figure 11 illustrating the within-cluster patterns revealed 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 $\hat{z}_{ik} = 0$).
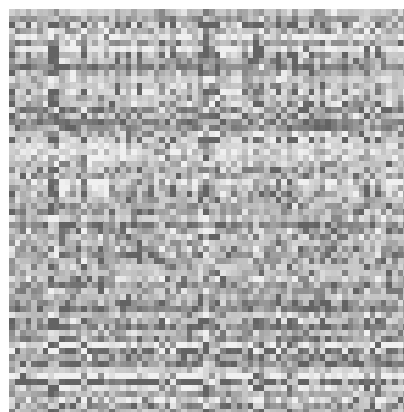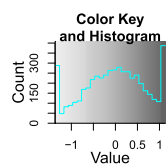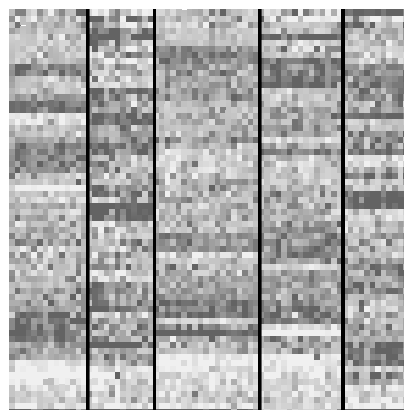
FIG 11. *Heatmaps of DLBCL covariates that were assigned to latent column clusters with more than 10 members. The panels display the covariates before and after bidirectional local clustering by VariScan. The vertical lines in the bottom panel mark the covariate clusters. The color key for both panels is displayed at the top of the plot.*
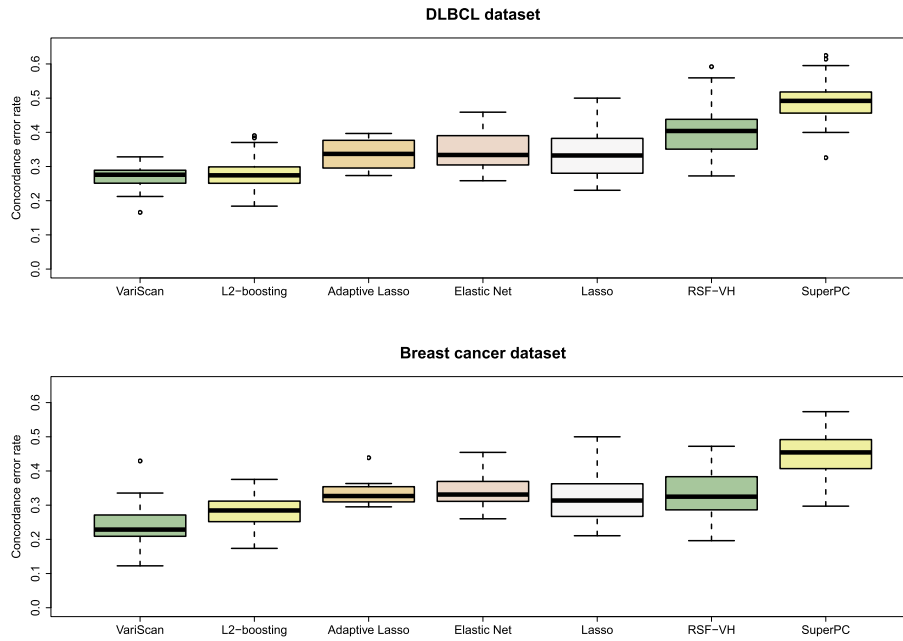
FIG 12. *Side-by-side boxplots of percentage concordance error rates for the benchmark datasets.*

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 12. 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. Further results of the data analysis and comparisons are available in Section G of the Appendix.

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 existing literature. For the breast cancer dataset, our survey indicated several prominent genes related to breast cancer development and progression, such as TGF-B2 [9], ABCC3, which is known to be up-regulated in primary breast cancers, and LAPTM4B, which is related to breast carcinoma relapse with metastasis [47]. 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 [52], and BNIP3, which is down-regulated in DLBCL and is a known marker associated with positive survival [67]. A detailed functional/mechanistic analysis of the main set of genes for both datasets is provided in Section G of the Appendix.

7. Conclusions

Utilizing the sparsity-inducing property of PDPs, VariScan offers an efficient technique for clustering, variable selection, and prediction in high-dimensional regression problems. The covariates are grouped into a smaller number of clusters consisting of covariates with similar across-subject patterns. We theoretically demonstrate how a PDP allocation can be differentiated from a Dirichlet process allocation in terms of the relative sizes of the latent clusters. We provide a theoretical explanation for the impressive ability of VariScan to a posteriori detect the true covariate clusters for a general class of models.

In simulations and real data analysis, we show that VariScan makes highly accurate cluster-related inferences. The technique consistently outperforms established methodologies such as random survival forests, L_2-boosting, and supervised principal components, in terms of prediction accuracy. In the analyses of benchmark microarray datasets, we identified several genes having known implications in cancer development and progression, which further engenders our hypothesis.

The VariScan methodology focuses on continuous covariates as a proof of concept, achieving simultaneous clustering, variable selection, and prediction in high-throughput regression settings and possessing appealing theoretical and empirical properties. Generalization to count, categorical, and ordinal covariates is possible. It is important to investigate the dependence structures and theoretical properties associated with the more general framework. This will be the focus of our group's future research.

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 graphical processing units [75]. We plan to make the software available as an R package for general use in the near future. 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.

Appendix A: Details of theoretical calculations and results

A.1. Proof of Theorem 2.1

1. (a) Suppose $h \geq 2$. Because V_h and $(1 - V_t)$ have beta distributions, it can be shown that $E(\log V_h) = \psi(1 - d) - \psi(\alpha_1 + 1 + (h - 1)d)$ and $E(\log(1 - V_t)) = \psi(\alpha_1 + td) - \psi(\alpha_1 + 1 + (t - 1)d)$. Now,

$$E(\log \pi_h) - E(\log V_h)$$
$$= \sum_{t=1}^{h-1} E(\log(1 - V_t))$$

$$= \sum_{t=1}^{h-1}\{\psi(\alpha_1+td)-\psi(\alpha_1+1+(t-1)d)\}$$

$$= \sum_{t=1}^{h-1}\{\psi(\alpha_1+td)-\overline{\psi(\alpha_1+(t-1)d)+\frac{1}{\alpha_1+(t-1)d}}\}$$

$$\text{(recurrence relation for } \psi(x))$$

$$= \sum_{t=1}^{h-1}\{\psi(\alpha_1+td)-\psi(\alpha_1+(t-1)d)\}-\sum_{t=1}^{h-1}\frac{1}{\alpha_1+(t-1)d}$$

$$= \psi(\alpha_1+(h-1)d)-\psi(\alpha_1)-\sum_{t=1}^{h-1}\frac{1}{\alpha_1+(t-1)d}$$

$$= \psi(\alpha_1+(h-1)d)-\psi(\alpha_1)-\frac{1}{d}\sum_{t=1}^{h-1}\frac{1}{\alpha_1/d+(t-1)}$$

$$= \psi(\alpha_1+(h-1)d)-\psi(\alpha_1)$$
$$-\frac{1}{d}\left\{\psi(\alpha_1/d)+\sum_{t=1}^{h-1}\frac{1}{\alpha_1/d+(t-1)}-\psi(\alpha_1/d)\right\}$$

$$= \psi(\alpha_1+(h-1)d)-\psi(\alpha_1)-\frac{1}{d}\big(\psi(\alpha_1/d+h-1)-\psi(\alpha_1/d)\big).$$

Since $E(\log V_h)=\psi(1-d)-\psi(\alpha_1+1+(h-1)d)$, therefore,

$$E(\log \pi_h)$$
$$= E(\log V_h)+\psi(\alpha_1+(h-1)d)-\psi(\alpha_1)$$
$$-\frac{1}{d}\left(\psi(\alpha_1/d+h)-\psi(\alpha_1/d)\right)$$
$$= \psi(1-d)-\overline{\psi(\alpha_1+1+(h-1)d)-\psi(\alpha_1+(h-1)d)}-\psi(\alpha_1)$$
$$-\frac{1}{d}\left(\psi(\alpha_1/d+h-1)-\psi(\alpha_1/d)\right)$$
$$= \psi(1-d)-\frac{1}{\alpha_1+(h-1)d}-\psi(\alpha_1)$$
$$-\frac{1}{d}\left(\psi(\alpha_1/d+h-1)-\psi(\alpha_1/d)\right)$$
$$= \psi(1-d)-\psi(\alpha_1)$$
$$-\frac{1}{d}\left(\psi(\alpha_1/d+h-1)+\frac{1}{\alpha_1/d+h-1}-\psi(\alpha_1/d)\right)$$
$$= \psi(1-d)-\psi(\alpha_1)-\frac{1}{d}\left(\psi(\alpha_1/d+h)-\psi(\alpha_1/d)\right).$$

It can be verified that the same formula is applicable when $h=1$. Since $\lim_{x\to\infty}\psi(x)/\log(x)=1$, we have $\lim_{h\to\infty}E(\log \pi_h)=-\infty$.

(b) By a similar calculation, we obtain the expression for $\mathrm{Var}(\log \pi_h)$.

Since $\lim_{x \to \infty} \psi_1(x) = 0$, $\lim_{h \to \infty} \mathrm{Var}(\log \pi_h) = \psi_1(1-d) - \psi_1(\alpha_1) + \frac{1}{d^2} \psi_1(\alpha_1/d)$, which is finite.

(c) Applying Chebychev's inequality, for any $\epsilon > 0$,

$$
\begin{aligned}
P\left(\left| \frac{\log \pi_h}{\log h} + \frac{1}{d} \right| \geq \epsilon \right) &\leq \frac{E(\log \pi_h + \frac{1}{d} \log h)^2}{(\log h)^2 \epsilon^2} \\
&= \frac{\mathrm{Var}(\log \pi_h)}{(\log h)^2 \epsilon^2} + \frac{(E \log \pi_h + \frac{1}{d} \log h)^2}{(\log h)^2 \epsilon^2}.
\end{aligned}
\tag{A.1}
$$

From Part 1b, we have that $\lim_{h \to \infty} \mathrm{Var}(\log \pi_h)/(\log h)^2 = 0$. Moreover,

$$
\begin{aligned}
\lim_{h \to \infty} \frac{E \log \pi_h}{\log h} &= \lim_{h \to \infty} \frac{\psi(1-d) - \psi(\alpha_1) + \frac{1}{d}\left(\psi(\alpha_1/d) - \psi(\alpha_1/d + h) \right)}{\log h} \\
&= -\frac{1}{d} \lim_{h \to \infty} \frac{\psi(\alpha_1/d + h)}{\log h} = -\frac{1}{d}.
\end{aligned}
$$

Therefore, the second term on the right hand side of expression (A.1) is also 0, and $\log \pi_h / \log h \xrightarrow{p} -1/d$.

2. The expressions for DP models can be easily computed because the V_h^* are i.i.d. beta variables. The asymptotic normal distribution in Part 2c follows from the Central Limit Theorem because

$$
\sqrt{h} \left(\frac{1}{h} \log(\pi_h^*) - \frac{1}{h} \sum_{t=1}^{h-1} \log(1 - V_t^*) \right) \xrightarrow{L} 0.
$$

3. Part 1c implies that $\sqrt{h} \left(\frac{1}{h} \log(\pi_h) \right) \xrightarrow{L} 0$. The stated result follows from Part 2c and Slutsky's theorem.

Appendix B: A basic lemma for MCMC updates

Lemma B.1. *Suppose we are given the cluster-predictors $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$ in assumption (2.4). Suppose $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ and σ_β^2 are also known, so that $\Theta = (\sigma_\beta^2, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is known. Upon integrating out the regression coefficients in assumption (2.4) with respect to Zellner's weighted g prior (2.6), we have that*

1. *The regression outcomes are marginally distributed as*

$$
\boldsymbol{y} \mid \Theta \sim N_n \left(\boldsymbol{0}, \boldsymbol{\Sigma} + \sigma_\beta^2 \boldsymbol{X}_{\boldsymbol{\gamma}} (\boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1} \boldsymbol{X}_{\boldsymbol{\gamma}}' \right).
$$

2. *The density of the regression outcomes has the simplified form:*

$$
[\boldsymbol{y} \mid \Theta] = (2\pi)^{-n/2} \left(1 + \sigma_\beta^2 \right)^{-r/2} \prod_{i=1}^{n} \sigma_i^{-1} \exp\left(\frac{-\|\boldsymbol{\phi}\|^2}{2} \right) \exp\left(\frac{\|\boldsymbol{H\phi}\|^2}{2(1 + \sigma_\beta^{-2})} \right)
$$

where vector $\boldsymbol{\phi} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$, matrix $\boldsymbol{H} = \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'$ with $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X_\gamma}$, and $r = rank(\boldsymbol{H})$.

Proof. Rewriting assumption (2.4) as $\boldsymbol{y} \mid \boldsymbol{\beta_\gamma}, \Theta \sim N_n(\boldsymbol{X_\gamma}\boldsymbol{\beta_\gamma}, \boldsymbol{\Sigma})$,

1. Follows from elementary properties of multivariate normal distributions.
2. Part 1 implies that $\boldsymbol{\phi} \mid \Theta \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n + \sigma_\beta^2 \boldsymbol{H})$. Writing $\boldsymbol{\Delta} = \boldsymbol{I}_n + \sigma_\beta^2 \boldsymbol{H}$, the density, $[\boldsymbol{\phi} \mid \Theta] = \frac{|\boldsymbol{\Delta}|^{-1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\boldsymbol{\phi}'\boldsymbol{\Delta}^{-1}\boldsymbol{\phi}\right)$. Applying a well-known matrix identity, we have that $\boldsymbol{\Delta}^{-1} = \boldsymbol{I}_n - 1/(1 + \sigma_\beta^{-2})\boldsymbol{H}$. Since matrix $\boldsymbol{H}$ is idempotent with eigenvalues that are either 0 or 1, its rank r is the sum of its eigenvalues. Simultaneously diagonalizing $\boldsymbol{I}_n$ and $\boldsymbol{H}$, we get $|\boldsymbol{\Delta}| = \left(1 + \sigma_\beta^2\right)^r$ and

$$[\boldsymbol{\phi} \mid \Theta] = (2\pi)^{-n/2}\left(1 + \sigma_\beta^2\right)^{-r/2}\exp\left(\frac{-\|\boldsymbol{\phi}\|^2}{2}\right)\exp\left(\frac{\|\boldsymbol{H}\boldsymbol{\phi}\|^2}{2(1 + \sigma_\beta^{-2})}\right),$$

where $\|\boldsymbol{H}\boldsymbol{\phi}\|^2 = \boldsymbol{\phi}'\boldsymbol{H}\boldsymbol{\phi}$ is the squared projection of $\boldsymbol{\phi}$ on the column space of $\boldsymbol{H}$. Transforming back to $\boldsymbol{y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\phi}$, we obtain the desired expression, since the Jacobian equals $\prod_{i=1}^{n} \sigma_i^{-1}$. □

Appendix C: MCMC procedure

C.1. Covariate-to-cluster Allocation

For $j = 1, \ldots, p$, the full conditional of allocation variable c_j is available in closed form given the distinct number of latent vector elements. However, this calculation becomes very intensive since n and p are large. To accommodate the updates of the large number of covariate-related parameters, we apply the fast and accurate Metropolis-Hastings algorithm developed by [27] for Pölya urn schemes.

C.2. Latent Vectors and Indicators

Among the allocation variables $c_1, \ldots, c_p$, suppose there are q clusters, with cluster k consisting of $n_k = \sum_{j=1}^{p} \mathcal{I}(c_j = k)$ covariates for $k = 1, \ldots, q$. As $i = 1, \ldots, n$ and $k = 1, \ldots, q$ vary, the sufficient statistics $\bar{x}_{ik} = \sum_{j=1}^{p} x_{ij} \cdot \mathcal{I}(c_j = k)/n_k$ are independently distributed as $N(0, \tau_1^2/n_k)$ if $z_{ik} = 0$, and as $N(v_{ik}, \tau^2/n_k)$ if $z_{ik} = 1$. Dirichlet process prior (2.3) 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. To accommodate the large number of latent vector elements, we apply the fast data squashing algorithm of [27].

C.3. Cluster Predictors and Cluster Representatives

The choice of basis functions such as splines and wavelets for the nonlinear functionals h in (2.4) 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 $\boldsymbol{\beta_\gamma}$ to iteratively update the vector of indicators $\boldsymbol{\gamma}_k = (\gamma_k^{(0)}, \gamma_k^{(1)}, \gamma_k^{(2)})$, for clusters $k = 1, \ldots, q$. Given the cluster representative $\boldsymbol{u}_k$ 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 following theorem 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 $\boldsymbol{u}_k$ and $\boldsymbol{\gamma}_k$. After a cycle of updates of q indicators and cluster representatives has been completed, the regression coefficients $\boldsymbol{\beta_\gamma}$ may be jointly generated from the full conditional if necessary.

Theorem C.1. *For $t = 1, 2$, consider the models $M_t : \boldsymbol{y} \sim N_n\big(\boldsymbol{X}_t\boldsymbol{\beta}_t, \boldsymbol{\Sigma}\big)$ with $\boldsymbol{\beta}_t$ distributed according to Zellner's weighted g prior (2.6). Suppose $\boldsymbol{X}_2$ has m additional columns relative to $\boldsymbol{X}_1$ so that $M_1 \subset M_2$. Let $\boldsymbol{V}_t = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_t$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$, where $\phi_i = y_i/\sigma_i$. Project the m additional columns of $\boldsymbol{V}_2$ orthogonal to the column space of $\boldsymbol{V}_1$. Mutually orthogonalizing the resultant columns, rescaling the non-zero vectors to unit length, and discarding any zero vectors, we obtain the basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m'}\}$ where $m' \leq m$. Then*

$$\frac{[\boldsymbol{y} \mid M_2, \boldsymbol{\Sigma}, \sigma_\beta]}{[\boldsymbol{y} \mid M_1, \boldsymbol{\Sigma}, \sigma_\beta]} = \big(1 + \sigma_\beta^2\big)^{-m'/2} \cdot \exp\bigg\{\frac{1}{2(1 + \sigma_\beta^{-2})} \sum_{s=1}^{m'} (\boldsymbol{\phi}'\boldsymbol{e}_s)^2\bigg\}.$$

Proof. From Lemma B.1, under models M_1 and M_2:

$$[\boldsymbol{y} \mid M_t, \boldsymbol{\Sigma}, \sigma_\beta] = (2\pi)^{-n/2}\big(1 + \sigma_\beta^2\big)^{-r_t/2}\prod_{i=1}^{n}\sigma_i^{-1} \exp\left(\frac{-\|\boldsymbol{\phi}\|^2}{2}\right) \exp\left(\frac{\|\boldsymbol{H}_t\boldsymbol{\phi}\|^2}{2(1 + \sigma_\beta^{-2})}\right),$$

$$t = 1, 2,$$

where $\boldsymbol{H}_t = \boldsymbol{V}_t(\boldsymbol{V}_t'\boldsymbol{V}_t)^{-1}\boldsymbol{V}_t'$ with $\boldsymbol{V}_t$ defined as in Theorem C.1, and $r_t = \text{rank}(\boldsymbol{H}_t)$. The nested models $M_1 \subset M_2$ imply that their respective projection matrices $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are such that

$$\boldsymbol{H}_2 = \boldsymbol{H}_1 + \sum_{s=1}^{m'} \boldsymbol{H}_{\boldsymbol{e}_s}$$

$$= \boldsymbol{H}_1 + \sum_{s=1}^{m'} \boldsymbol{e}_s\boldsymbol{e}_s'$$

because $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m'}\}$ is an orthonormal basis and $r_2 - r_1 = m' \leq m$. Since $\|\boldsymbol{H}_t\boldsymbol{\phi}\|^2 = \boldsymbol{\phi}'\boldsymbol{H}_t\boldsymbol{\phi}$, we obtain the stated expression for the ratio of densities. $\square$

C.4. Latent Regression Outcomes

Suppose the regression outcomes y_i are latent, but the responses w_i are observed for some subjects. The latent y_i's can be iteratively sampled as follows. Let $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}_{\boldsymbol{\gamma}}$. Compute the symmetric *projection* or *hat* matrix of order n: $\boldsymbol{H} = ((h_{it})) = \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'$. The prior distribution of y_i given the remaining regression outcomes is

$$y_i \mid \boldsymbol{y}_{-i} \sim N\left(\frac{\sigma_i \sum_{t \neq i} h_{it}y_t/\sigma_t}{\varphi^2 - h_{ii}}, \ \frac{\sigma_i^2\,\varphi^2}{\varphi^2 - h_{ii}} \right) \tag{C.1}$$

where $\varphi^2 = 1 + \sigma_{\beta}^{-2}$. The proof is given below. The conditional prior can then be updated using the response w_i to generate aposteriori, $y_i \sim [y_i \mid \boldsymbol{y}_{-i}, w_i]$.

Proof. The conditional prior in equation (C.1) is obtained as follows. Let $\boldsymbol{h}_{-i}$ be the column vector consisting of the remaining $(n-1)$ elements on the ith column of $\boldsymbol{H}$ after excluding h_{ii}, the ith diagonal element of $\boldsymbol{H}$. Let $\boldsymbol{\Sigma}_{-i}$ denote the submatrix of order $(n-1)$ obtained by dropping the ith row and i^{th} column of $\boldsymbol{\Sigma}$. Since matrix $\boldsymbol{\Sigma}$ is diagonal, $y_i = \sigma_i\phi_i$ and $\boldsymbol{y}_{-i} = \boldsymbol{\Sigma}_{-i}\boldsymbol{\phi}_{-i}$. In particular, y_i and $\boldsymbol{y}_{-i}$ involve mutually exclusive components of the vector $\boldsymbol{\phi}$. In Part 2 of Lemma B.1, the quadratic term

$$\begin{aligned}
\|\boldsymbol{H}\boldsymbol{\phi}\|^2 &= \boldsymbol{\phi}'\boldsymbol{H}\boldsymbol{\phi} \\
&= h_{ii}\,\phi_i^2 + 2(\boldsymbol{h}_{-i}'\boldsymbol{\phi}_{-i})\phi_i + k_i \\
&= h_{ii}\,\phi_i^2 + 2(\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i})\phi_i + k_i
\end{aligned}$$

where k_i does not involve ϕ_i. Since y_i is functionally related to ϕ_i but not to $\boldsymbol{\phi}_{-i}$, we isolate the terms involving ϕ_i in Part 2 of Lemma B.1:

$$\begin{aligned}
[y_i \mid \boldsymbol{y}_{-i}, \Theta] &\propto \exp\left(-\frac{\phi_i^2}{2}\right)\exp\left(\frac{h_{ii}\,\phi_i^2}{2\varphi^2}\right)\exp\left(\frac{2(\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i})\phi_i}{2\varphi^2}\right) \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\varphi^2 - h_{ii}}{\varphi^2}\right)\left(\phi_i^2 - \frac{2(\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i})}{\varphi^2 - h_{ii}}\phi_i\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{\varphi^2 - h_{ii}}{\varphi^2}\right)\left(\phi_i - \frac{\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i}}{\varphi^2 - h_{ii}}\right)^2\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\varphi^2 - h_{ii}}{\varphi^2}\right)\left(\frac{y_i}{\sigma_i} - \frac{\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i}}{\varphi^2 - h_{ii}}\right)^2\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\varphi^2 - h_{ii}}{\sigma_i^2\,\varphi^2}\right)\left(y_i - \frac{\sigma_i\,\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i}}{\varphi^2 - h_{ii}}\right)^2\right\}
\end{aligned}$$

Since $\boldsymbol{h}_{-i}'\boldsymbol{\Sigma}_{-i}^{-1/2}\boldsymbol{y}_{-i} = \sum_{t \neq i} h_{it}y_t/\sigma_t$, we obtained the conditional prior (C.1).
$\square$

C.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 $\tilde{n} \times \text{col}(\boldsymbol{U}_{\boldsymbol{\gamma}})$ matrix. Given the set of predictors $\boldsymbol{\gamma}$ and variances $\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_{\tilde{n}}^2$ in relation (2.4), the following theorem provides expressions for the posterior predictions of the regression outcomes, $\tilde{\boldsymbol{y}}$.

Theorem C.2. *Given the set of predictors $\boldsymbol{\gamma}$ in model (2.4), let $\tilde{\boldsymbol{U}}_{\boldsymbol{\gamma}}$ be the matrix of cluster representative elements consisting of $\tilde{n}$ rows and $\text{col}(\boldsymbol{U}_{\boldsymbol{\gamma}})$ number of columns. Define $\tilde{\boldsymbol{\Sigma}} = diag(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_{\tilde{n}}^2)$. Then $\tilde{\boldsymbol{y}} \mid \boldsymbol{y} \sim N_{\tilde{n}}(\frac{1}{1+\sigma_\beta^{-2}}\tilde{\boldsymbol{y}}_{lse}, \tilde{\boldsymbol{\Sigma}} + \frac{1}{1+\sigma_\beta^{-2}}\tilde{\boldsymbol{H}})$, where $\tilde{\boldsymbol{y}}_{lse} = \tilde{\boldsymbol{U}}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{lse}$ with the vector of the least-squares estimates, $\hat{\boldsymbol{\beta}}_{lse} = (\boldsymbol{U}_{\boldsymbol{\gamma}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{U}_{\boldsymbol{\gamma}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$, and where $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{U}}_{\boldsymbol{\gamma}}(\boldsymbol{U}_{\boldsymbol{\gamma}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_{\boldsymbol{\gamma}})^{-1}\tilde{\boldsymbol{U}}_{\boldsymbol{\gamma}}'$. Therefore, under a squared error loss, the vector of the predicted regression outcomes for the $\tilde{n}$ subjects is $E[\tilde{\boldsymbol{y}} \mid \boldsymbol{y}] = \frac{1}{1+\sigma_\beta^{-2}}\tilde{\boldsymbol{y}}_{lse}$.*

Proof. Using the notation of Lemma B.1, since $\boldsymbol{\phi} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \Theta \sim N_n(\boldsymbol{V}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{I}_n)$, the posterior corresponding to the prior (2.6) is

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\phi}, \Theta \sim N_{|\mathcal{S}^*|+1}\left(\frac{\hat{\boldsymbol{\beta}}_{lse}}{1+\sigma_\beta^{-2}}, \frac{1}{1+\sigma_\beta^{-2}}(\boldsymbol{V}'\boldsymbol{V})^{-1}\right).$$

For the $\tilde{n}$ additional subjects, define $\tilde{\boldsymbol{\phi}} = \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{V}} = \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{U}}_{\boldsymbol{\gamma}}$. Then $\tilde{\boldsymbol{\phi}} \mid \boldsymbol{\phi}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \Theta \sim N_n(\tilde{\boldsymbol{V}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{I}_n)$, which is independent of $\boldsymbol{\phi}$. Marginalizing over the above posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, we get

$$\tilde{\boldsymbol{\phi}} \mid \boldsymbol{\phi}, \Theta \sim N_{\tilde{n}}\left(\frac{1}{1+\sigma_\beta^{-2}}\tilde{\boldsymbol{V}}\hat{\boldsymbol{\beta}}_{lse}, \boldsymbol{I}_n + \frac{1}{1+\sigma_\beta^{-2}}\tilde{\boldsymbol{V}}(\boldsymbol{V}'\boldsymbol{V})^{-1}\tilde{\boldsymbol{V}}'\right)$$

Since $\boldsymbol{\phi}$ is a 1-1 mapping of $\boldsymbol{y}$, conditioning on $\boldsymbol{\phi}$ is equivalent to conditioning on $\boldsymbol{y}$. Re-expressing in terms of $\tilde{\boldsymbol{y}}$, $\boldsymbol{X}_{\boldsymbol{\gamma}}$, and $\tilde{\boldsymbol{U}}_{\boldsymbol{\gamma}}$, we get the result. $\square$

Appendix D: Some general results on the posterior consistency of multivariate density estimation

The following lemma extends Theorem 3 of [24] to densities on $\mathcal{R}^n$ arising as convolutions of vector locations with errors distributed as zero-mean finite normal mixtures. The lemma is a key argument in the proof of the paper's Theorem 4.1 presented in Section E of the Appendix.

Lemma D.1. *For $t \in \mathcal{R}$ and a positive integer K, let $\boldsymbol{\zeta} = (p_1, \ldots, p_K, \tau_1, \ldots, \tau_K)$, where $p_r \geq 0$ and $\sum_{r=1}^K p_r = 1$. Define the finite mixture of univariate Gaussian densities, $\phi_{\boldsymbol{\zeta}}(t) = \sum_{r=1}^K p_r N(t|0, \tau_r^2)$. For $\boldsymbol{x} = (x_1, \ldots, x_n)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$, consider density functions on $\mathcal{R}^n$ of the form $f_{\boldsymbol{\zeta}, P^{(n)}}(\boldsymbol{x}) =$*

$\int_{\mathcal{R}^n} \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) \, dP^{(n)}(\boldsymbol{\theta})$, where $\phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) = \prod_{i=1}^n \phi_{\boldsymbol{\zeta}}(x_i - \theta_i)$, and $P^{(n)}$ is an n-variate distribution. Let $\boldsymbol{\zeta}$ and $P^{(n)}$ have the priors μ and $\Pi^{(n)}$ respectively. Let Π_n^* be the distribution induced by $\mu \times \Pi^{(n)}$ on the set of density functions $\{f_{\boldsymbol{\zeta}, P^{(n)}}\}$.

Suppose $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ are a random sample from a true density $f_0(\boldsymbol{x}) = f_{\boldsymbol{\zeta}_0, P_0^{(n)}}(\boldsymbol{x})$. Assume that (i) $P_0^{(n)}$ is compactly supported in $\mathcal{R}^n$ and belongs to the weak support of the prior $\Pi^{(n)}$, and (ii) $\boldsymbol{\zeta}_0$ belongs to the weak support of the prior μ. Then, for n fixed:

1. Conditional on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ and as $p \to \infty$, the distribution Π_n^* is weakly consistent at f_0.
2. Let Ω_n be the set of densities on $\mathcal{R}^n$ and the density function $f \sim \Pi_n^*$. Suppose $g : \Omega_n \to \mathcal{R}$ is a continuous mapping. Conditional on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ and as $p \to \infty$, the random variable $g(f)$ weakly converges to $g(f_0)$.

Proof.

1. The proof extends the arguments of [24]. For $k > 1$, suppose $S = \{\boldsymbol{x} : |\boldsymbol{x}| < k\} \subset \mathcal{R}^n$ is such that $P_0^{(n)}(S) = 1$. It is easy to see that f_0 has moments of all orders. For $\eta > 0$, choose $k' > k$ such that $\int_{|\boldsymbol{x}| > k'} |\boldsymbol{x}|^2 f_0(\boldsymbol{x}) \, d\boldsymbol{x} < \eta$. Since $k' > 1$, this implies that both $\int_{|\boldsymbol{x}| > k'} f_0(\boldsymbol{x}) \, d\boldsymbol{x}$ and $\int_{|\boldsymbol{x}| > k'} |\boldsymbol{x}| f_0(\boldsymbol{x}) \, d\boldsymbol{x}$ are less than η. Now

$$\int_{\mathcal{R}^n} f_0 \log \left(\frac{f_{\boldsymbol{\zeta}, P_0^{(n)}}}{f_{\boldsymbol{\zeta}, P^{(n)}}} \right) = \int_{A_-} f_0 \log \left(\frac{f_{\boldsymbol{\zeta}, P_0^{(n)}}}{f_{\boldsymbol{\zeta}, P^{(n)}}} \right) + \int_{A_0} f_0 \log \left(\frac{f_{\boldsymbol{\zeta}, P_0^{(n)}}}{f_{\boldsymbol{\zeta}, P^{(n)}}} \right)$$
$$+ \int_{A_+} f_0 \log \left(\frac{f_{\boldsymbol{\zeta}, P_0^{(n)}}}{f_{\boldsymbol{\zeta}, P^{(n)}}} \right) \tag{D.1}$$

where $A_0 = \{\boldsymbol{x} : |\boldsymbol{x}| < k'\}$, $A_+ = \{\boldsymbol{x} : x_1 > 0, |\boldsymbol{x}| > k'\}$, and $A_- = \{\boldsymbol{x} : x_1 < 0, |\boldsymbol{x}| > k'\}$.

Let $\tau_{\min} := \min_r \tau_r$ and $\tau_{\max} := \max_r \tau_r$. With $\boldsymbol{0}_{n-1}$ representing the vector of $(n-1)$ zeros, for an arbitrary set of probabilities $\{p_j\}_{j=1}^\infty$ that sum to 1, define the vectors $\boldsymbol{\zeta}_{\min} := (p_1, , \dots, p_K, \tau_{\min}, \dots \tau_{\min})$, $\boldsymbol{\zeta}_{\max} := (p_1, , \dots, p_K, \tau_{\max}, \dots \tau_{\max})$, $\boldsymbol{x}_{\min} := (|\boldsymbol{x}| - k, \boldsymbol{0}_{n-1})'$, and $\boldsymbol{x}_{\max} := (|\boldsymbol{x}| + k, \boldsymbol{0}_{n-1})'$. Also define $\tau_-^2 = 1/\tau_{\min}^2 - 1/\tau_{\max}^2$ and $\tau_+^2 = 1/\tau_{\min}^2 + 1/\tau_{\max}^2$. Since $P_0^{(n)}$ belongs to the weak support of prior $\Pi^{(n)}$, it follows that $\Pi^{(n)}\{P^{(n)} : P^{(n)}(S) > 1/2\} > 0$. If $P^{(n)}(S) > 1/2$, then

$$\int_{A_-} f_0 \log \left(\frac{f_{\boldsymbol{\zeta}, P_0^{(n)}}}{f_{\boldsymbol{\zeta}, P^{(n)}}} \right) \leq \int_{A_-} f_0 \log \left(\frac{\int_S \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) \, dP_0^{(n)}(\boldsymbol{\theta})}{\int_S \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) \, dP^{(n)}(\boldsymbol{\theta})} \right)$$
$$\leq \int_{A_-} f_0 \log \left(\frac{\phi_{\boldsymbol{\zeta}_{\min}}^{(n)}(\boldsymbol{x}_{\min})}{\phi_{\boldsymbol{\zeta}_{\max}}^{(n)}(\boldsymbol{x}_{\max}) \cdot \frac{1}{2}} \right)$$
$$= \frac{-\tau_-^2}{2} \int_{A_-} |\boldsymbol{x}|^2 f_0(\boldsymbol{x}) d\boldsymbol{x} + k\tau_+^2 \int_{A_-} |\boldsymbol{x}| f_0(\boldsymbol{x}) d\boldsymbol{x}$$

$$+ \left(\frac{-\tau_-^2 k^2}{2} + \log 2 \right) \int_{A_-} f_0(\boldsymbol{x}) d\boldsymbol{x}$$

$$< \left(\frac{-\tau_-^2}{2} + k\tau_+^2 \frac{-\tau_-^2 k^2}{2} + \log 2 \right) \eta.$$

Similarly, we obtain an identical bound for the third term in (D.1).

Clearly, $c := \inf_{|\boldsymbol{x}| \leq k'} \inf_{|\boldsymbol{\theta}| \leq k} \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) > 0$. The family of functions $\{\phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{x} \in A_0\}$, viewed as a set of functions of $\boldsymbol{\theta} \in S$, is uniformly equicontinuous. By the Arzela-Ascoli theorem applied to compact metric spaces, there exist finitely many points $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*$ such that for any $\boldsymbol{x} \in A_0$, there exists an i with

$$\sup_{\boldsymbol{\theta} \in S} | \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) - \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x}_i^* - \boldsymbol{\theta}) | < c\,\delta. \tag{D.2}$$

Let

$$E = \left\{ P^{(n)} : \left| \int \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x}_i^* - \boldsymbol{\theta})\, dP_0^{(n)}(\boldsymbol{\theta}) - \int \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x}_i^* - \boldsymbol{\theta})\, dP^{(n)}(\boldsymbol{\theta}) \right| < c\,\delta; \right.$$

$$\left. i = 1, \ldots, m \right\}$$

Since E is a weak neighborhood of $P_0^{(n)}$, and since $P_0^{(n)}$ belongs to the weak support of $\Pi^{(n)}$, we have $\Pi^{(n)}(E) > 0$. Let $P^{(n)} \in E$. Then for any $\boldsymbol{x} \in A_0$, choosing an appropriate $\boldsymbol{x}_i^*$ from (D.2) and using a triangulation argument, we get

$$\left| \frac{\int \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta})\, dP^{(n)}(\boldsymbol{\theta})}{\int \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta})\, dP_0^{(n)}(\boldsymbol{\theta})} - 1 \right| < 3\delta$$

and therefore

$$\left| \frac{\int \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta})\, dP_0^{(n)}(\boldsymbol{\theta})}{\int \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta})\, dP^{(n)}(\boldsymbol{\theta})} - 1 \right| < \frac{3\delta}{1 - 3\delta}$$

provided $\delta < 1/3$. Thus, for any fixed $\boldsymbol{\zeta}$, for $P^{(n)}$ in a set of positive $\Pi^{(n)}$-probability, we have

$$\int_{\mathcal{R}^n} f_0 \log \left(\frac{f_{\boldsymbol{\zeta}, P_0^{(n)}}}{f_{\boldsymbol{\zeta}, P^{(n)}}} \right) < 2 \left(\frac{-\tau_-^2}{2} + k\tau_+^2 \frac{-\tau_-^2 k^2}{2} + \log 2 \right) \eta + \frac{3\delta}{1 - 3\delta}. \tag{D.3}$$

For any $\boldsymbol{\zeta}$,

$$\int_{\mathcal{R}^n} f_0 \log \left(f_0 / f_{\boldsymbol{\zeta}, P^{(n)}} \right) = \int_{\mathcal{R}^n} f_0 \log \left(f_0 / f_{\boldsymbol{\zeta}, P_0^{(n)}} \right)$$

$$+ \int_{\mathcal{R}^n} f_0 \log \left(f_{\boldsymbol{\zeta}, P_0^{(n)}} / f_{\boldsymbol{\zeta}, P^{(n)}} \right). \tag{D.4}$$

We observe that the first term on the right hand side (RHS) of (D.4) converges to 0 as $\boldsymbol{\zeta} \to \boldsymbol{\zeta}_0$.

Given any $\epsilon > 0$, choose a neighborhood N of $\boldsymbol{\zeta}_0$ such that $\tau_r > 0$ for all r and such that whenever $\boldsymbol{\zeta} \in N$, the first term on the RHS of (D.4) is less than $\epsilon/2$. Next choose η and δ so that for any $\boldsymbol{\zeta} \in N$, the RHS of (D.3) is less than $\epsilon/2$. Since $\boldsymbol{\zeta}_0$ belongs to the weak support of μ, the result follows.

2. Part 1 establishes the fact that the set of densities Ω_n is equipped with the weak convergence metric. The result follows from the Continuous Mapping Theorem [refer to 77, pp. 258–259].

$\square$

The following Lemma establishes the posterior consistency of covariate co-clustering for a general class of models as p and n both tend to ∞. The result's applicability extends well beyond the high-dimensional regression problems investigated by VariScan. For example, if $\min(n, p)$ is large, the posterior consistency of co-clustering is obtained for the NoB-LoC approach of [44].

Lemma D.2. *Let the column vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ of covariate matrix $\boldsymbol{X}_{np}$ be a random sample from the true density $f_{\boldsymbol{\zeta}_0, P_0^{(n)}}$ defined in Lemma D.1. In addition to the Lemma D.1 assumptions, suppose that the prior μ for parameter vector $\boldsymbol{\zeta} = (p_1, \ldots, p_K, \tau_1, \ldots, \tau_K)$ allows a strictly positive lower bound for standard deviations $\tau_1, \ldots, \tau_K$. In other words, there exists a constant $\tau_* > 0$ such that $\tau_* \leq \min_r \tau_r$ with probability 1. Suppose also that:*

1. *The true n-dimensional distribution $P_0^{(n)}$ is discrete. Consequently, there exists a set of true allocation variables, $c_1^{(0)}, \ldots, c_p^{(0)}$, mapping the p co-variates to the distinct atoms of distribution $P_0^{(n)}$.*

2. *The n-variate atoms of distribution $P_0^{(n)}$ are i.i.d. realizations of the n-fold product measure of a non-atomic real distribution, G_0. The distribution G_0 has compact support.*

Let $\mathcal{L} = \{j_1, \ldots, j_L\} \subset \{1, \ldots, p\}$ be a fixed subset of L covariate indexes. Then there exists an increasing sequence of numbers $\{p_n\}$ such that, as n grows and provided $p > p_n$, the clustering inferences for the covariate subset $\mathcal{L}$ are aposteriori consistent. That is,

$$\lim_{\substack{n \to \infty \\ p > p_n}} P\big[\varkappa_{\mathcal{L}}(\boldsymbol{c}) = 1 \mid \boldsymbol{X}_{np}\big] \to 1$$

where $\varkappa_{\mathcal{L}}(\boldsymbol{c})$, the proportion of correctly clustered covariate pairs, is defined in equation (4.2) of the paper.

Proof. Let true distribution $P_0^{(n)} = \sum_{t=1}^{\infty} \pi_t \, \delta_{\boldsymbol{v}_t^{(0)}}$ where $\{\pi_t\}_{t=1}^{\infty}$ are probabilities that sum to 1, $\delta_{\boldsymbol{v}_t^{(0)}}$ is the point mass at the n-dimensional atom, $\boldsymbol{v}_t^{(0)} = (v_{1t}^{(0)}, \ldots, v_{nt}^{(0)})'$, where $v_{it}^{(0)} \overset{i.i.d.}{\sim} G_0$. Since distribution G_0 is non-atomic, the n-dimensional atoms $\{\boldsymbol{v}_t^{(0)}\}_{t=1}^{\infty}$ of distribution $P_0^{(n)}$ are all distinct. Let j_1 and

j_2 be two covariate indices belonging to the set $\mathcal{I}_L$. Due to the exchangeability of the covariates and the allocation variables under the true model, and their exchangeability also under the inference procedure, we may assume without loss of generality that $j_1 = 1$ and $j_2 = 2$. Let $\boldsymbol{X}_{np}^* = \{x_j : j \geq 3\}$ be the matrix of the remaining $(p-2)$ covariates.

Case 1. Suppose that, in reality, $c_1^{(0)} = c_2^{(0)}$. That is, for some positive integer t, covariate $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are each equal to the atom $\boldsymbol{v}_t^{(0)}$ plus an n-vector of i.i.d. errors. We will prove that $\lim_{\substack{n\to\infty \\ p > p_n}} P\left(c_1 = c_2 \mid c_1^{(0)} = c_2^{(0)}, \boldsymbol{X}_{np}\right) = 1$.

Consider posterior inferences for the hypotheses, $H_0 : c_1 = c_2$ versus $H_1 : c_1 \neq c_2$. Given the matrix of the remaining $(p-2)$ covariates, $\boldsymbol{X}_{np}^*$, we define the *conditional log-likelihood ratio of H_0 versus H_1* as

$$L_{np} := \log\left(\frac{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{X}_{np}^*]}{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_1, \boldsymbol{X}_{np}^*]}\right), \tag{D.5}$$

where the symbol $[\cdot]$ represents densities of random variables.

Recall that covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have common density $f_{\boldsymbol{\zeta}, P^{(n)}}(\boldsymbol{x})$ under the inference procedure. Therefore, the joint marginal $[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{X}_{np}^*] = E^{\Pi_n^*}\left([\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{\zeta}, P^{(n)}] \mid \boldsymbol{X}_{np}^*\right)$, where the distribution Π_n^* on the densities $\{f_{\boldsymbol{\zeta}, P^{(n)}}\}$ is defined in the statement of Lemma D.1, and the expectation is conditional on the matrix $\boldsymbol{X}_{np}^*$. Applying Parts 1 and 2 of Lemma D.1, we find that, for n fixed and as $p \to \infty$, the integrand $[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{\zeta}, P^{(n)}] \overset{L}{\to} [\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]$. Now, being a Gaussian mixture, the integrand $[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{\zeta}, P^{(n)}] \leq (2\pi\tau_*^2)^{-n}$, where τ_* is a positive constant. Hence, by the dominated convergence theorem, the joint marginal $[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{X}_{np}^*] \overset{L}{\to} [\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]$. An identical argument applies when we condition on the alternative hypothesis. For fixed n and as $p \to \infty$, we therefore have

$$L_{np} \overset{L}{\to} g_0^{(n)} := \log\left(\frac{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]}\right).$$

In other words, given $\epsilon > 0$, there exists an increasing sequence of integers $\{p_n\}$ such that

$$|L_{np} - g_0^{(n)}| < \epsilon \quad \text{for every } n \text{ and whenever } p > p_n. \tag{D.6}$$

Now, denoting by c_1^* the latent component of true distribution $P_0^{(n)}$ from which covariate $\boldsymbol{x}_1$ arises,

$$g_0^{(n)} = \log\left(\frac{E^{c_1^*}[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{E^{c_1^*}[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]}\right)$$

$$\text{where } \Pr[c_1^* = t \mid \boldsymbol{\zeta}_0, P_0^{(n)}] = \pi_t \text{ for } t \in \mathbb{N}$$

$$\geq \log E^{c_1^*} \left(\frac{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]} \right)$$

$$\text{(by Jensen's inequality: } \frac{EW_1}{EW_2} \geq E(\frac{W_1}{W_2}))$$

$$\geq E^{c_1^*} \log \left(\frac{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]} \right)$$

$$\text{(by Jensen's inequality)}$$

$$= \sum_{c_1^*=1}^{\infty} \pi_{c_1^*} \log \left(\frac{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{[\boldsymbol{x}_1, \boldsymbol{x}_2 \mid c_1^*, H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]} \right)$$

$$= \sum_{c_1^*=1}^{\infty} \pi_{c_1^*} \log \left(\frac{[\boldsymbol{x}_1 \mid c_1^*, \boldsymbol{\zeta}_0, P_0^{(n)}] \cdot [\boldsymbol{x}_2 \mid c_1^*, H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{[\boldsymbol{x}_1 \mid c_1^*, \boldsymbol{\zeta}_0, P_0^{(n)}] \cdot [\boldsymbol{x}_2 \mid c_1^*, H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]} \right)$$

$$= \sum_{t=1}^{\infty} \pi_t \log \left(\frac{[\boldsymbol{x}_2 \mid c_1^* = t, H_0, \boldsymbol{\zeta}_0, P_0^{(n)}]}{[\boldsymbol{x}_2 \mid c_1^* = t, H_1, \boldsymbol{\zeta}_0, P_0^{(n)}]} \right). \tag{D.7}$$

Under the hypothesis H_0, and conditional on $c_1^* = t$, $[\boldsymbol{x}_2 \mid c_1^* = t, \boldsymbol{x}_1, \boldsymbol{\zeta}_0, P_0^{(n)}] = \phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_t^{(0)})$. Similarly, under the hypothesis H_1, and conditional on $c_1^* = t$, $[\boldsymbol{x}_2 \mid c_1^* = t, \boldsymbol{\zeta}_0, P_0^{(n)}] = \frac{1}{(1-\pi_t)} \sum_{u \neq t} \pi_u \phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_u^{(0)})$. Therefore, the right hand side of (D.7) equals

$$\sum_{t=1}^{\infty} \pi_t \log \left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_t^{(0)})}{\sum_{u \neq t} \frac{\pi_u}{(1-\pi_t)} \phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_u^{(0)})} \right)$$

$$= \sum_{t=1}^{\infty} \pi_t \log \left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_t^{(0)})}{E^U \phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_U^{(0)})} \right),$$

$$\text{where } \Pr[U = u] = \frac{\pi_u}{(1-\pi_t)} \text{ for } u \in \mathbb{N} - \{t\}$$

$$\geq \sum_{t=1}^{\infty} \pi_t \log E^U \left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_t^{(0)})}{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_U^{(0)})} \right)$$

$$\text{(by Jensen's inequality: } \frac{EW_1}{EW_2} \geq E(\frac{W_1}{W_2}))$$

$$\geq \sum_{t=1}^{\infty} \pi_t E^U \log \left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_t^{(0)})}{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_U^{(0)})} \right)$$

$$\text{(by Jensen's inequality)}$$

$$= \sum_{t:\pi_t>0} \pi_t \sum_{u \neq t} \frac{\pi_u}{(1-\pi_t)} \log \left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_t^{(0)})}{\phi_{\boldsymbol{\zeta}_0}^{(n)}(\boldsymbol{x}_2 - \boldsymbol{v}_u^{(0)})} \right)$$

$$= n \cdot \sum_{t:\pi_t > 0} \pi_t \sum_{u \neq t} \frac{\pi_u}{(1 - \pi_t)} \bar{W}_{tu}^{(n)} \tag{D.8}$$

where $\bar{W}_{tu}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \log(\phi_{\boldsymbol{\zeta}_0}^{(n)}(x_{i2} - v_{it}^{(0)})/\phi_{\boldsymbol{\zeta}_0}^{(n)}(x_{i2} - v_{iu}^{(0)}))$ for natural numbers $u \neq t$. Since hypothesis H_0 is true by assumption, as $n \to \infty$, we obtain on a set of probability 1:

$$\bar{W}_{tu}^{(n)} \to \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \log\left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(x - v_t)}{\phi_{\boldsymbol{\zeta}_0}^{(n)}(x - v_u)}\right) \phi_{\boldsymbol{\zeta}_0}^{(n)}(x - v_t) \cdot dx \cdot G_0(dv_t) \cdot G_0(dv_u) \tag{D.9}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(v_t, v_u) \cdot G_0(dv_t) \cdot G_0(dv_u), \quad \text{where we define}$$

$$K(v_t, v_u) = \int_{-\infty}^{\infty} \log\left(\frac{\phi_{\boldsymbol{\zeta}_0}^{(n)}(x - v_t)}{\phi_{\boldsymbol{\zeta}_0}^{(n)}(x - v_u)}\right) \phi_{\boldsymbol{\zeta}_0}^{(n)}(x - v_t) \, dx$$

Being a KL divergence, the function $K(v_t, v_u)$ is non-negative. Furthermore, it is strictly positive whenever $v_t \neq v_u$, which is a certain event for non-atomic distribution $G_0 \times G_0$. This implies that the random variable $\bar{W}_{tu}^{(n)}$ in equation (D.9) converges almost surely to a positive constant whenever indexes $u \neq t$. Combining expressions (D.7), (D.8) and (D.9), $\lim_n g_0^{(n)} = \infty$ on a set of probability 1. In conjunction with relation (D.6), this implies that the conditional log-likelihood ratio satisfies on a set of probability 1:

$$\lim_{\substack{n \to \infty \\ p > p_n}} L_{np} = \infty. \tag{D.10}$$

The "conditional prior" probability, $P[c_1 = c_2 \mid \boldsymbol{X}_{np}^*] = E^{\boldsymbol{\zeta}, P^{(n)}} P[c_1 = c_2 \mid \boldsymbol{\zeta}, P^{(n)}, \boldsymbol{X}_{np}^*]$. Parts 1 and 2 of Lemma D.1 and the dominated convergence theorem give us $\lim_{p \to \infty} P[c_1 = c_2 \mid \boldsymbol{X}_{np}^*] = P[c_1 = c_2 \mid \boldsymbol{\zeta}_0, P_0^{(n)}] = \sum_{t=1}^{\infty} \pi_t^2 \in (0, 1)$. Therefore, the "conditional prior" odds ratio of the hypotheses has the limit:

$$\lim_{p \to \infty} \frac{[H_0 \mid \boldsymbol{X}_{np}^*]}{[H_1 \mid \boldsymbol{X}_{np}^*]} \xrightarrow{L} \frac{\sum_{t=1}^{\infty} \pi_t^2}{1 - \sum_{t=1}^{\infty} \pi_t^2} > 0. \tag{D.11}$$

Applying results (D.10) and (D.11), the limiting posterior odds ratio of the hypotheses is computed as

$$\frac{[H_0 \mid \boldsymbol{X}_{np}]}{[H_1 \mid \boldsymbol{X}_{np}]} = \exp(L_{np}) \cdot \frac{[H_0 \mid \boldsymbol{X}_{np}^*]}{[H_1 \mid \boldsymbol{X}_{np}^*]} \xrightarrow{L} \infty$$

as $n \to \infty$ and provided $p > p_n$. In other words,

$$\lim_{\substack{n \to \infty \\ p > p_n}} P\left(c_1 = c_2 \mid c_1^{(0)} = c_2^{(0)}, \boldsymbol{X}_{np}\right) = 1.$$

Case 2. Suppose that $c_1^{(0)} \neq c_2^{(0)}$ in reality. That is, for positive integers $u \neq t$, the covariate $\boldsymbol{x}_1$ is equal to the atom $\boldsymbol{v}_t^{(0)}$ plus an n-vector of i.i.d. errors and covariate $\boldsymbol{x}_2$ is equal to the atom $\boldsymbol{v}_u^{(0)}$ plus an n-vector of i.i.d. errors. Since the distribution G_0 is non-atomic by assumption, $\boldsymbol{v}_t^{(0)} \neq \boldsymbol{v}_u^{(0)}$.

By an identical argument as in Case 1, it can be shown that the conditional log-likelihood ratio of H_0 versus H_1, defined in equation (D.5), satisfies on a set of probability 1:

$$\lim_{\substack{n \to \infty \\ p > p_n}} (-L_{np}) = \infty$$

and hence, that

$$\lim_{\substack{n \to \infty \\ p > p_n}} P\left(c_1 \neq c_2 \mid c_1^{(0)} \neq c_2^{(0)}, \boldsymbol{X}_{np}\right) = 1.$$

Applying Slutsky's theorem, the result is extended to all covariate pairs belonging to the finite subset $\mathcal{L}$. Since the correct identification of the cluster allocations for the set $\mathcal{L}$ is equivalent to the event $\varkappa_{\mathcal{L}}(\boldsymbol{c}) = 1$, the lemma follows. $\square$

Appendix E: Proof of Theorem 4.1

The sufficient conditions *(i)* and *(ii)* of Lemma D.1 of the Appendix are satisfied by the true model assumed in Theorem 4.1. The VariScan prior for fitting $\boldsymbol{X}_{np}$ can be cast in the framework of Lemma D.1. Specifically, the p columns of $\boldsymbol{X}_{np}$ are distributed as $\boldsymbol{x}_j \overset{i.i.d.}{\sim} f_{\boldsymbol{\zeta}, P^{(n)}}(\boldsymbol{x})$ for $j = 1, \dots, p$, where the common density on $\mathcal{R}^n$ has the form $f_{\boldsymbol{\zeta}, P^{(n)}}(\boldsymbol{x}) = \int_{\mathcal{R}^n} \phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) \, dP^{(n)}(\boldsymbol{\theta})$ with $\phi_{\boldsymbol{\zeta}}^{(n)}(\boldsymbol{x} - \boldsymbol{\theta}) = \prod_{i=1}^{n} \phi_{\boldsymbol{\zeta}}(x_i - \theta_i)$ for the finite mixture, $\phi_{\boldsymbol{\zeta}}(t) = \xi \cdot N(t|0, \tau^2) + (1 - \xi) \cdot N(t|0, \tau_1^2)$, so that $K = 2$ and $\boldsymbol{\zeta} = (\xi, 1 - \xi, \tau, \tau_1)$. As mentioned in Section 2.1, the VariScan prior assumes that $\tau_1 \gg \tau > \tau_*$ for a small, positive constant, τ_*. The distribution $P^{(n)}$ for the vector $\boldsymbol{\theta}$ follows a PDP with parameters (d, α_1). This partitions the p columns of $\boldsymbol{X}_{np}$ into q clusters sharing the latent vectors as their common value of $\boldsymbol{\theta}$. The individual latent vector elements are i.i.d. from distribution G, which itself has the Dirichlet process prior (2.3) and is discrete. By well-known properties of the Dirichlet process, the non-atomic true distribution G_0 belongs to its weak support. Lemma D.2 above is therefore applicable, and the stated posterior consistency follows.

Appendix F: Simulation studies: Further results

F.1. Prediction accuracy

Table 2 displays the number of covariate predictors ($|\hat{\mathcal{S}}|$) for VariScan along with the model sizes for the other methods. The displayed values are averages
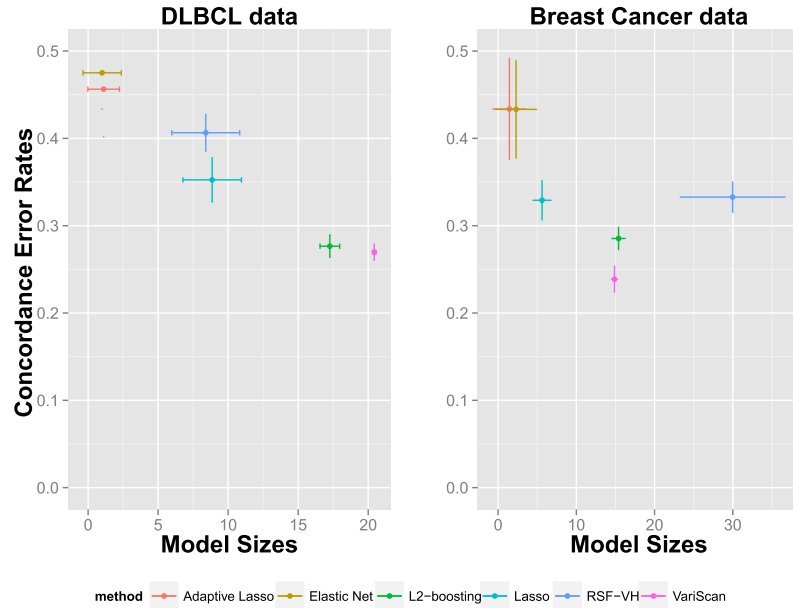
FIG 13. *Plot of concordance error rates versus model sizes for the competing methods along with the standard errors (shown by whiskers). The left, middle and right respectively correspond to effect size β^* equal to 0.2, 0.6, and 1.*

TABLE 2
Comparison of model sizes, averaged over the 15 independent replications, for the simulated data. The standard errors are displayed in parentheses.

Method	$\beta^* = 0.2$	$\beta^* = 0.6$	$\beta^* = 1$
VariScan	18.49 (0.35)	17.15 (0.65)	17.19 (0.74)
L_2-boosting	21.73 (0.56)	20.13 (1.10)	19.07 (0.83)
Adaptive Lasso	1.00 (0.00)	6.11 (1.09)	8.33 (1.38)
Elastic Net	1.50 (0.50)	9.30 (1.63)	11.60 (1.9)
Lasso	2.67 (0.81)	7.00 (1.56)	13.00 (2.22)
RSF-VH	14.47 (0.27)	15.47 (0.26)	15.13 (0.32)
SuperPC	86.27 (35.3)	31.40 (6.91)	43.53 (13.32)

computed over the 15 independent replications. For the methods VariScan, L_2-boosting, and RSF-VH, the model sizes were approximately unchanged as β^* varied. The plots of sparsity versus prediction error rates are provided in Figure 13. For both the datasets, the plots demonstrate the effectiveness of VariScan in producing highly predictive models with small model sizes.

Appendix G: Analysis of benchmark data sets: Further results

The number of VariScan covariate predictors, $|\hat{\mathcal{S}}|$, along with the model sizes for the other methods, are presented in Table 3. Numerical summaries of these
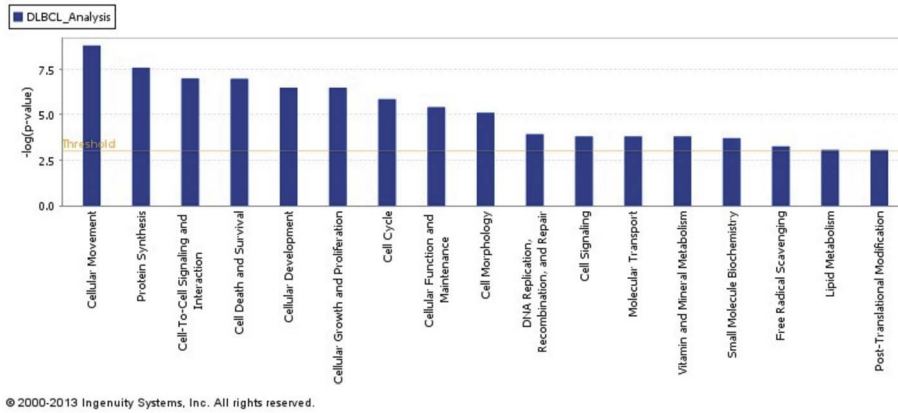
FIG 14. *Molecular and cellular functions associated with the genes selected by the VariScan model.*

TABLE 3
Comparison of the model sizes, averaged over the 50 independent replications, for the benchmark datasets. The standard errors are displayed in parentheses.

Method	DLBCL dataset	Breast cancer dataset
VariScan	20.43 (0.06)	14.88 (0.20)
L_2-boosting	17.26 (0.35)	15.40 (0.44)
Adaptive Lasso	4.67 (1.12)	4.12 (1.34)
Elastic Net	5.27 (1.46)	7.36 (1.63)
Lasso	8.86 (1.05)	5.62 (0.60)
RSF-VH	8.40 (1.21)	29.98 (3.37)
SuperPC	53.00 (13.23)	85.66 (15.08)

TABLE 4
Comparison of the percentage concordance error rates, averaged over the 50 independent replications, for the benchmark datasets. The standard errors are displayed in parentheses.

Method	DLBCL dataset	Breast cancer dataset
VariScan	26.96 (0.46)	23.87 (0.76)
L_2-boosting	27.65 (0.64)	28.53 (0.64)
Adaptive Lasso	33.72 (1.56)	33.93 (1.62)
Elastic Net	34.71 (1.90)	34.36 (1.62)
Lasso	35.25 (1.26)	32.91 (1.13)
RSF-VH	40.64 (1.04)	33.28 (0.88)
SuperPC	48.71 (0.77)	45.02 (0.85)

error rates are computed in Table 4. Figure 14 illustrates the molecular and cellular functions associated with the genes selected by the VariScan model using Ingenuity Pathway Analysis software.

Biological interpretations. For subsequent biological interpretations, for both datasets we selected genes that have high posterior probability of being selected (0.75 in our case) for prediction in the models, which were then analyzed for their role in cancer progression by cross-referencing with existing lit-

erature. For the breast cancer genes our survey indicated that prominent genes for breast cancer like TGF-B2 are included in this gene set [9]. The functional roles for various genes of this set in the context of breast cancer are as follows. ABCC3 is up-regulated in primary breast cancer and known to mediate the resistance of Taxane in HER2 subtype [64, 62]. The cell adhesion molecule ALCAM is previously established as a marker of breast cancer to predict the adjuvant chemotherapy response [35]. MAPK signaling is enhanced by CAMP (cathelicidin antimicrobial protein), which promotes the metastasis of breast cancer [79]. Overexpression in non-invasive ductal carcinoma compared to invasive ductal carcinoma of CHGB (chromogranin B), advocates a role of early marker for CHB in ductal carcinoma [41]. Lysosomal transmembrane protein LC27 (LAPTM4B) promotes the resistance to anthracyclines and consequently causes the relapse of breast carcinoma with metastasis [47]. Also, an allele of LC27 (LAPTM4B*2) is predictive of breast cancer susceptibility and metastasis [46]. The chemokine ligand CCL21 with significant up-regulation acts as a marker for breast cancer cells with lymph node metastasis [51]. Iron transporter gene SLC11A2 is over-expressed in breast cancer cells to compensate the enhanced demand for iron in proliferating cells [39]. Transactivation of epidermal growth factor (EGF) enhances the expression of SLC37A1 in breast cancer cells [34].

For the DLBCL data there are 25 genes with high posterior probability of more than 0.75. Cytokine interaction is dominant in this gene-set with the presence multiple chemokine ligands (CXCL9 and CCL18) and interleukin receptors of IL2 and IL5. Functional analysis performed on these genes by Ingenuity Pathway Analysis (IPA) software reinforced this notion. IPA is web-based functional analysis tool that identifies the most relevant pathways, molecular networks, and biological functions for gene lists. Figure 1 illustrates the molecular and cellular functions associated with the genes of interest. The significant enrichment of ontology terms cellular movement, cell-to-cell signaling and interaction might specify their association with cytokine-cytokine receptor interaction. The enrichment of the cellular functions such as growth, development, morphology, cell cycle and protein synthesis indicate the generic proliferative nature associated with the genes. CCND1, a member of the Cyclin D family which is previously known to be associated with DLBCL is also a predictor with high probability [52]. Another interesting gene predictor BNIP3 (BCL2/adenovirus E1B 19kDa interacting protein 3) which is down-regulated in DLBCL with hyper methylation of the proximal CpG island, can be a marker associated with positive survival [67].

Acknowledgements

References

[1] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993. MR1224394

[2] J.H. Albert and S. Chib. Bayes inference in regression models with arma(p, q) errors. *Journal of Econometrics*, 64:183–206, 1994. MR1310523

[3] J.H. Albert, S. Chib, and R. Winkelmann. Posterior simulation and bayes factors in panel count data models. *Journal of Econometrics*, 86:33–54, 1998.

[4] D.F. Andrews and C.L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, 36:99–102, 1974. MR0359122

[5] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2:511–522, 2004.

[6] V. Baladandayuthapani, B.K. Mallick, and R.J. Carroll. Spatially adaptive bayesian penalized regression splines (p-splines). *Journal of Computational and Graphical Statistics*, 14(2):378–394, 2005. MR2160820

[7] D. Barry and J.A. Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88:309–319, 1993. MR1212493

[8] P.J. Brown, M. Vannucci, and T. Fearn. Multivariate bayesian variable selection and prediction. *J. R. Stat. Soc. Series B*, 60:627–641, 1998. MR1626005

[9] M.B. Buck and C. Knabbe. TGF-beta signaling in breast cancer. *Ann. N. Y. Acad. Sci.*, 1089:119–126, Nov 2006.

[10] J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66:429–436, 1979.

[11] C.A. Bush and S.N. MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

[12] D. Cox and D. Oakes. *Analysis of survival data*. London: Chapman and Hall, 1984. MR0751780

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

[14] C. de Boor. *A Practical Guide to Splines*. New York: Springer Verlag, 1978. MR0507062

[15] D.G.T. Denison, B.K. Mallick, and A.F.M. Smith. Automatic bayesian curve fitting. *Journal of the Royal Statistical Society, Series B*, 60:333–350, 1998. MR1616029

[16] G.T.D. Denison, C.C. Holmes, B.K. Mallick, and A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley, 2002. ISBN 9780471490364. URL https://books.google.com/books?id=SIlDWySNuXgC. MR1962778

[17] D.K. Dey, S.K. Ghosh, and B.K. Mallick. *Generalized Linear Models: A Bayesian Perspective*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis, 2000. ISBN 9780824790349. URL https://books.google.com/books?id=Y5AARf7oTNkC. MR1893779

[18] D.B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008. MR2521586

[19] 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*, 103:534–546, 2008. MR2523991

[20] R. Eubank. *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker., 1999. MR1680784

[21] T.S. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–223, 1973. MR0350949

[22] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. New York: Springer, 2006. MR2265601

[23] E. George and R. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

[24] S. Ghosal, J.K. Ghosh, and R.V. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics*, 27:143–158, 1999. MR1701105

[25] A. Gnedin and J. Pitman. Regenerative composition structures. *Annals of Probability*, 33:445–479, 2005. MR2122798

[26] J.E. Griffin, P.J. Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010. MR2596440

[27] S. Guha. Posterior simulation in countable mixture models for large datasets. *Journal of the American Statistical Association*, 105:775–786, 2010. MR2724860

[28] T. Hanson and W.O. Johnson. Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association*, 97(460), 2002. MR1951256

[29] F. Harrell, R. Califf, D. Pryor, K. Lee, and R. Rosati. Evaluating the yield of medical tests. *J. Amer. Med. Assoc.*, 247:2543–2546, 1982.

[30] D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–340, 1977. MR0451550

[31] T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990. ISBN 0412343908. MR1082147

[32] C.C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168, 2006. MR2227368

[33] T. Hothorn and P. Buhlmann. Model-based boosting in high dimensions. *Bioinformatics*, 22:2828–2829, 2006.

[34] D. Iacopetta, R. Lappano, A.R. Cappello, M. Madeo, E.M. De Francesco, A. Santoro, R. Curcio, L. Capobianco, V. Pezzi, M. Maggiolini, and V. Dolce. SLC37A1 gene expression is up-regulated by epidermal growth factor in breast cancer cells. *Breast Cancer Res. Treat.*, 122(3):755–764, Aug 2010.

[35] M. Ihnen, V. Muller, R.M. Wirtz, C. Schroder, S. Krenkel, I. Witzel, B.W. Lisboa, F. Janicke, and K. Milde-Langosch. Predictive impact of activated leukocyte cell adhesion molecule (ALCAM/CD166) in breast cancer. *Breast Cancer Res. Treat.*, 112(3):419–427, Dec 2008.

[36] H. Ishwaran and L.F. James. Generalized weighted chinese restaurant processes for species sampling mixture models. *Statist. Sinica*, 13:1211–1235, 2003. MR2026070

[37] H. Ishwaran, U.B. Kogalur, et al. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105:205–217, 2010. MR2757200

[38] D. Jiang, C. Tang, and A. Zhang. Clustering analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16:1370–1386, 2004.

[39] X.P. Jiang, R.L. Elliott, and J.F. Head. Manipulation of iron transporter genes results in the suppression of human and mouse mammary adenocarcinomas. *Anticancer Res.*, 30(3):759–765, Mar 2010.

[40] S. Kim, M.G. Tadesse, and M. Vannucci. Variable selection in clustering via dirichlet process mixture models. *Biometrika*, 93:877–893, 2006. MR2285077

[41] N. Kimura, R. Yoshida, S. Shiraishi, M. Pilichowska, and N. Ohuchi. Chromogranin A and chromogranin B in noninvasive and invasive breast carcinoma. *Endocr. Pathol.*, 13(2):117–122, 2002.

[42] S. Kundu and D.B. Dunson. Bayes variable selection in semiparametric linear models. *Journal of the American Statistical Association*, 109(505):437–447, 2014. MR3180575

[43] L. Kuo and B. Mallick. Bayesian semiparametric inference for the accelerated failure time model. *Canadian J. Stat.*, 25:457–472, 1997.

[44] J. Lee, P. Müller, and Y. Ji. A nonparametric bayesian model for local clustering. Technical report, Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, 2013a.

[45] J. Lee, P. Müller, Y. Zhu, and Y. Ji. A nonparametric bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association*, 108:775–788, 2013b. MR3174662

[46] X. Li, X. Kong, X. Chen, N. Zhang, L. Jiang, T. Ma, and Q. Yang. LAPTM4B allele *2 is associated with breast cancer susceptibility and prognosis. *PLoS ONE*, 7(9):e44916, 2012.

[47] Y. Li, L. Zou, Q. Li, B. Haibe-Kains, R. Tian, Y. Li, C. Desmedt, C. Sotiriou, Z. Szallasi, J.D. Iglehart, A.L. Richardson, and Z.C. Wang. Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nat. Med.*, 16(2):214–218, Feb 2010.

[48] A. Lijoi and I. Prünster. *Models beyond the Dirichlet process*, pages 80–136. Cambridge Series in Statistical and Probabilistic Mathematics, 2010. MR2730661

[49] A. Lijoi, R.H. Mena, and I. Prünster. Controlling the reinforcement in bayesian nonparametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:715–740, 2007a. MR2370077

[50] A. Lijoi, R.H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94:769–786, 2007b. MR2416792

[51] Y. Liu, R. Ji, J. Li, Q. Gu, X. Zhao, T. Sun, J. Wang, J. Li, Q. Du, and B. Sun. Correlation effect of EGFR and CXCR4 and CCR7 chemokine receptors in predicting breast cancer metastasis and prognosis. *J. Exp. Clin. Cancer Res.*, 29:16, 2010.

[52] I.S. Lossos and D. Morgensztern. Prognostic biomarkers in diffuse large B-cell lymphoma. *J. Clin. Oncol.*, 24(6):995–1007, Feb 2006.

[53] R.F. MacLehose and D.B. Dunson. Bayesian semiparametric multiple shrinkage. *Biometrics*, 66(2):455–462, 2010. MR2758825

[54] B.K. Mallick, D. Ghosh, and M. Ghosh. Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:219–234, 2005. MR2137322

[55] M. May, P. Royston, M. Egger, A.C. Justice, and J.A.C. Sterne. 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, 2004.

[56] M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18:1194–1206, 2002.

[57] M. Medvedovic, K.Y. Yeung, and R.E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20:1222–1232, 2004.

[58] M.C. Meyer and P.W. Laud. Predictive variable selection in generalized linear models. *Journal of the American Statistical Association*, 97(459):859–871, 2002. MR1941415

[59] J.S. Morris and R.J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199, 2006. MR2188981

[60] P. Müller, F. Quintana, and G.L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20:260–278, 2011. MR2816548

[61] P. Müller and R. Mitra. Bayesian nonparametric inference–why and how. *Bayesian analysis (Online)*, 8(2), 2013. MR3066939

[62] C. O'Brien, G. Cavet, A. Pandita, X. Hu, L. Haydu, S. Mohan, K. Toy, C.S. Rivers, Z. Modrusan, L.C. Amler, and M.R. Lackner. Functional genomics identifies ABCC3 as a mediator of taxane resistance in HER2-amplified breast cancer. *Cancer Res.*, 68(13):5380–5389, Jul 2008.

[63] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008. MR2524001

[64] L. Partanen, J. Staaf, M. Tanner, V.J. Tuominen, A. Borg, and J. Isola. Amplification and overexpression of the ABCC3 (MRP3) gene in primary breast cancer. *Genes Chromosomes Cancer*, 51(9):832–840, Sep 2012.

[65] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probab. Theory Related Fields*, 92:21–39, 1992. MR1156448

[66] F. Petralia, V. Rao, and D.B. Dunson. Repulsive Mixtures. *ArXiv e-prints*, April 2012.

[67] B.L. Pike, T.C. Greiner, X. Wang, D.D. Weisenburger, Y.H. Hsu, G. Renaud, T.G. Wolfsberg, M. Kim, D.J. Weisenberger, K.D. Siegmund, W. Ye, S. Groshen, R. Mehrian-Shai, J. Delabie, W.C. Chan, P.W. Laird, and J.G. Hacia. DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status. *Leukemia*, 22(5):1035–1043, May 2008.

[68] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102:145–158, 1995. MR1337249

[69] J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900, 1997. MR1434129

[70] F.A. Quintana and P.L. Iglesias. Bayesian clustering and product partition models. *J. R. Statist. Soc. B*, 65:557–574, 2003. MR1983764

[71] A. Rosenwald et al. 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, 2002.

[72] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B*, 73:689–710, 2011. MR2867454

[73] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. MR1309433

[74] M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

[75] M.A. Suchard, Q. Wang, C. Chan, J. Frelinger, A. Cron, and M. West. Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19(2):419–438, 2010. MR2758309

[76] R. Tibshirani. The lasso method for variable selection in the cox model. *Stat. Med.*, 16:385–395, 1997.

[77] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. MR1652247

[78] L.J. van't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[79] G. Weber, C.I. Chamorro, F. Granath, A. Liljegren, S. Zreika, Z. Saidak, B. Sandstedt, S. Rotstein, R. Mentaverri, F. Sanchez, A. Pivarcsi, and M. Stahle. Human antimicrobial protein hCAP18/LL-37 promotes a metastatic phenotype in breast cancer. *Breast Cancer Res.*, 11(1):R6, 2009.

[80] S. Weisberg. *Applied Linear Regression*. J. Wiley and Sons, NY, 1985.

[81] M. West. On scale mixtures of normal distributions. *Biometrika*, 74:646–648, 1987. MR0909372

[82] X. Xu, M. Ghosh, et al. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 2015. MR3432244

[83] S.L. Zeger and M.R. Karim. Generalized linear models with random effects: A gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86, 1991. MR1137101

[84] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. MR2279469

[85] H. Zou and T. Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005. MR2137327