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

Subha Guha  
|University of Missouri  
GuhaSu@missouri.edu

Joint work with Veera Baladandayuthapani  
M. D. Anderson Cancer Center

June 24, 2015
Motivating application

- Microarray-based gene expression levels and survival outcomes for limited number of patients
- Diffuse large B-cell lymphoma (DLBCL) dataset of Rosenwald et al. (2002) and breast cancer dataset of van’t Veer et al. (2002)
- For individuals $i = 1, \ldots, n$:
  - survival time $w_i > 0$
  - AFT model: $y_i = \log w_i$
  - failure status $\delta_i = 0$ if $w_i$ is right-censored and $\delta_i = 1$ if $w_i$ is uncensored
  - expression levels $x_{i1}, \ldots, x_{ip}$ for $p$ genes, with $p \gg n$.
- In general, continuous covariates arranged in an $n$ by $p$ matrix, where $p \gg n$, and $n$ discrete/continuous responses
Motivating application

Goals of analysis:

- Infer sparse set of predictor indices: subset $S \subset \{1, \ldots, p\}$ of dimension $q << p$, consisting of indices of covariates significantly associated with responses
- Predict responses of $\tilde{n}$ additional subjects for whom only covariate information available

Development of parsimonious models for reliable predictions is challenging

Especially challenging for “small $n$, large $p$” regression problems arising in many areas
Existing variable selection methods I

Linear variable selection methods

- Stepwise selection (Peduzzi et al., 1980), penalized regression approaches such as lasso (and its variants) (Tibshirani, 1997), and non-concave penalized likelihood approaches (Fan and Li, 2002)

- Bayesian approaches – spike and slab mixture priors (Mitchell and Beauchamp, 1988), stochastic search variable selection (George and McCulloch, 1993), Gibbs-based variable selection (Dellaportas et al., 1982), Bayesian model averaging (Madigan and Raftery, 1994; Volinsky et al., 1997) and indicator priors (Kuo and Mallick, 1997)

- George and McCulloch (1993) approach extended to multivariate settings by Brown et al. (1998) and to generalized linear mixed models by Cai and Dunson (2006)

- Methods for multinomial probit models by Sha et al. (2004), and for microarray data with censored outcomes by Lee and Mallick (2004) and Sha et al. (2006)
Existing variable selection methods II

Regression using low-dimensional projections of covariate space

- Partial least squares (Nguyen and Rocke, 2002; Li and Gui, 2004)
- Supervised principal components (Bair and Tibshirani, 2004)

Non-linear prediction methods

- Statistical and machine learning techniques: support vector machines (Cristianini and Shawe-Taylor, 2000), $L_2$-boosting (Hothorn and Buhlmann, 2006)
- Ensemble methods: random forests (Ishwaran et al., 2010) and Bonato et al. (2010)
Difficulties in small \( n \), large \( p \) problems

- Regression predictors hard to detect
- Collinearity because \( n \)-dimensional space quickly saturated because \( p \gg n \)
- Highly correlated covariates are weakly identifiable as predictors
Difficulties in small $n$, large $p$ problems

- Regression predictors hard to detect
- Collinearity because $n$-dimensional space quickly saturated because $p \gg n$
- Highly correlated covariates are weakly identifiable as predictors

**Example**

- $j^{th}$ and $k^{th}$ covariate have sample correlation close to 1, but that neither covariate is a “true” predictor
- Alternative model: *both* covariates are predictors and $\beta_j \approx -\beta_k$
- Almost identical likelihood

- Collinearity $\Rightarrow$ unstable inferences and erroneous test case predictions (Weisberg, 1985)
VariScan

- Exploits sparsity-inducing property of Poisson-Dirichlet processes (PDPs) to group $p$ columns of covariate matrix into $q$ latent clusters, where $q \ll p$
- Clusters consist of covariates with similar patterns for subjects
- Latent clusters used to build nonlinear prediction model for responses
- Small $n$, large $p \rightarrow$ small $n$, small $q$
- Adaptive mixture of linear and nonlinear elements in regression – balance between model parsimony and flexibility
Modeling the covariates and latent clusters

- Covariate columns: $x_1, \ldots, x_p$ (length $n$)
- Each covariate belongs to exactly one of $q \ll p$ clusters, $q$ unknown
- For covariate $j = 1, \ldots, p$, allocation variable:
  \[ c_j = k \quad \text{if } j^{th} \text{ covariate belongs to } k^{th} \text{ cluster} \]
  for cluster $k = 1, \ldots, q$
- Covariates are noisy versions of latent vectors $v_1, \ldots, v_q$ of length $n$ associated with clusters
- EDA suggests non-Dirichlet allocation
EDA: covariate-to-cluster allocation

DLBCL dataset – gene expression data for 240 patients on 7,399 microarray elements (probes) representing approximately 4,128 genes

- Randomly selected $p = 500$ genes of $n = 100$ individuals
- Applied k-means to get within-cluster correlations $> .3$ and $R^2 \approx 70\%$
EDA: covariate-to-cluster allocation

DLBCL dataset – gene expression data for 240 patients on 7,399 microarray elements (probes) representing approximately 4,128 genes

- Randomly selected $p = 500$ genes of $n = 100$ individuals
- Applied k-means to get within-cluster correlations $> .3$ and $R^2 \approx 70\%$

- Pattern and $\hat{q} = 161$ strongly suggest non-DP allocation
Poisson-Dirichlet processes

- Introduced by Perman et al. (1992) and further studied by Pitman (1995) and Pitman and Yor (1997)
- Apriori exchangeable $c_1, \ldots, c_p \sim \text{PDP}(\alpha_1, d)$
  - Mass parameter $\alpha_1 > 0$, and discount parameter $0 \leq d < 1$
  - For fixed $\alpha_1$, $d = 0$ yields Dirichlet process with mass parameter $\alpha_1$
  - Number of distinct clusters, $q$, stochastically increasing in $\alpha_1$ and $d$

- **Dimension reduction**
  Random number of clusters, $q$, is asymptotically equivalent to

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

for a random variable $T_{d, \alpha_1} > 0$

- In VariScan model, data decides: $d \sim \frac{1}{2} \delta_0 + \frac{1}{2} U(0, 1)$
Modeling the latent vector elements

- Prior distribution in $\mathcal{R}^n$ for latent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_q$
- Clusters communicate
  \[ v_{ik} \sim^{	ext{iid}} G \quad i = 1, \ldots, n, \text{ and } k = 1, \ldots, q \]
- Flexibly capture patterns of subjects
  \[ G \sim \text{DP} \left( \alpha_2 ; \mathcal{N}(\mu_2, \tau_2^2) \right) \]
  \[ \Rightarrow G \text{ is discrete and latent vector elements are shared} \]
- NoB-LoC approach of Lee et al. (2013)
Predictor choices and regression outcomes

- All $n_k$ covariates in $k^{th}$ cluster have apriori equal chance of being chosen as cluster representative $u_k$
- Regression predictors are chosen from $q$ cluster representatives
- Cluster representatives feature in additive regression model

$$y_i \overset{indep}{\sim} N \left( \beta_0 + \sum_{k=1}^{q} \gamma_k^{(1)} \beta_k^{(1)} u_{ik} + \sum_{k=1}^{q} \gamma_k^{(2)} h(u_{ik}, \beta_k^{(2)}), \sigma^2 \right)$$

for a nonlinear function $h$

- Options for $h$ include order-$r$ splines with $m$ number of knots (de Boor, 1978; Hastie and Tibshirani, 1990; Denison et al., 1998)
Simulation study I

DLBCL dataset

- Randomly selected $p = 500$ genes and $n = 100$ individuals
- 15 independent replications of
  - Selected 10 genes with pairwise correlations less than 0.5 as true predictor set $S \subset \{1, \ldots, 500\}$, so that $|S| = 10$
  - For $\beta^* \in \{0.2, 0.6, 1.0\}$
    - Generated individuals’ failure times: $t_i \sim \mathcal{E}_i$, exponential with mean $\exp(\beta^* \sum_{j \in S} x_{ij})$, for $i = 1, \ldots, 100$
    - For 20% of individuals, generated censoring time $u_i \sim \mathcal{E}_i \cdot (-\infty, t_i)$. Set outcomes $y_i = \log u_i$ and failure status $\delta_i = 0$
    - For remaining individuals, set $y_i = \log t_i$ and $\delta_i = 1$
- Randomly assigned 67 individuals to training set
Simulation study

- Compared results to some existing methods for gene selection of survival outcomes:
  - **lasso**: Friedman et al. (2008)
  - **adaptive lasso**: Zou (2006)
  - **elastic net**: Zou and Trevor (2005)
  - **$L_2$-boosting**: Hothorn and Buhlmann (2006)
  - **RSF-VH**: Ishwaran et al. (2010)
  - **SuperPC**: Bair and Tibshirani (2004)

- **Concordance error rate**
  - 1 minus C-index of Harrell et al. (1982)
  - Measures procedure’s probability of incorrectly ranking failure times of two randomly chosen individuals
  - Does not rely on survivor function (not estimated by all procedures)
Simulation study

Concordance error rates I

$\beta^* = 0.2$

VariScan  L2-boosting  Adaptive Lasso  Elastic Net  Lasso  RSF-VH  SuperPC

Concordance error rate
Concordance error rates II

\[ \beta^* = 0.6 \]

VariScan, L2-boosting, Adaptive Lasso, Elastic Net, Lasso, RSF-VH, SuperPC

Concordance error rate

Subha Guha (MU)
Concordance error rates III

\( \beta^* = 1 \)
Data analysis

Motivating datasets

DLBCL dataset
- Randomly selected 100 out of 240 individuals after removing 5 with zero survival time
- \( p = 500 \) genes
- 50% of chosen individuals censored

Breast cancer dataset
- 76 individuals after removing 2 with zero survival time
- \( p = 500 \) genes
- 57.9% of individuals censored

50 independent replications of

- Randomly split data into training and test sets in 2:1 ratio
- Analyzed same set of individuals using multiple methods
- Used the different techniques to predict test case outcomes
Concordance error rates I

DLBCL dataset

VariScan L2-boosting Adaptive Lasso Elastic Net Lasso RSF-VH SuperPC

Concordance error rate

VariScan L2-boosting Adaptive Lasso Elastic Net Lasso RSF-VH SuperPC
Concordance error rates II

Breast cancer dataset

VariScan L2-boosting Adaptive Lasso Elastic Net Lasso RSF-VH SuperPC

Concordance error rate

VariScan L2-boosting Adaptive Lasso Elastic Net Lasso RSF-VH SuperPC

Breast cancer dataset

Subha Guha (MU)
VariScan can quantify nonlinear functional relationships between responses and covariates

Nonlinearity measure $\hat{N}$: posterior probability that a hypothetical, additional cluster appears as a predictor in a nonlinear form, rather than as a simple linear regressor

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DLBCL dataset</th>
<th>Breast cancer dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{N}$</td>
<td>0.97 (0.00)</td>
<td>0.75 (0.02)</td>
</tr>
<tr>
<td>$\hat{q}$</td>
<td>165</td>
<td>117</td>
</tr>
<tr>
<td>$\hat{P}[d = 0</td>
<td>w, X]$</td>
<td>0</td>
</tr>
</tbody>
</table>
Benchmark data sets

DLBCL results I

Posterior density of PDP parameter $d$

Density

Subha Guha (MU)
Median pairwise correlations for the $\hat{q} = 165$ PDP clusters in the least-squares allocation
Heatmap before clustering by VariScan
Heatmap after clustering by VariScan
Theoretical results

Different allocations patterns of PDPs and Dirichlet processes in terms of number and relative sizes of clusters.

High accuracy of clustering procedure: for $n$ and $p$ large enough, infer true clustering for any fixed subset of covariates.

Model selection consistency.

Prediction consistency.
Conclusions

- Efficient methodology for high-dimensional clustering, variable selection, and prediction
- Continuous covariates
- Censored/uncensored discrete/continuous outcomes
- Exploits sparsity of PDPs as dimension-reduction device
- Data permitted to choose between PDPs and Dirichlet processes for suitable clustering mechanism
- Theoretical guarantees for inference accuracy
- Compares favorably with and, in predictive accuracy, consistently outperforms existing methodologies for survival applications
- For benchmark datasets, identified several genes having known implications in cancer development and progression


