Bayesian Approaches for High-Dimensional Data Analysis

Gyuhyeong Goh

Department of Statistics
Kansas State University, Manhattan, KS

Joint work with

Shiqiang Jin

Kansas State University, Manhattan, KS

Introduction

Data structure

- In practice, we frequently observe the following data structure:

Y	X_1	X_2	$\cdots$	X_p
y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

  where Y is the response and $X_1, X_2, \ldots, X_p$ are the predictors.

- We are interested in modeling the relationship between $X_1, \ldots, X_p$ and Y.

Introduction

Statistical models

- Regression models play an important role in many application domains for analyzing or predicting a response based on multiple predictors.
  - Linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  - Logistic regression:

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

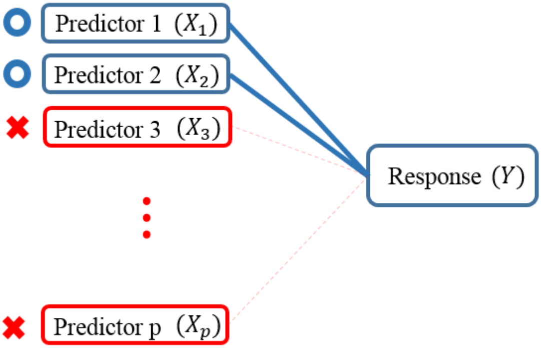
  - Cox regression:

$$h(Y) = h_0(Y) \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

Introduction

Variable selection

- *Variable selection* is the process of selecting a subset of relevant predictors in regression analysis. (e.g., BIC, AIC, CV, R_a^2, C_p, ...)



Introduction

Challenges of high-dimensional variable selection

- There are 2^p candidate models

  e.g., for $p = 3$

$$
\begin{aligned}
\texttt{Model 1}: \quad Y &= \beta_0 + \epsilon \\
\texttt{Model 2}: \quad Y &= \beta_0 + \beta_1 X_1 + \epsilon \\
\texttt{Model 3}: \quad Y &= \beta_0 + \beta_2 X_2 + \epsilon \\
\texttt{Model 4}: \quad Y &= \beta_0 + \beta_3 X_3 + \epsilon \\
\texttt{Model 5}: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\
\texttt{Model 6}: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon \\
\texttt{Model 7}: \quad Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \\
\texttt{Model 8}: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon
\end{aligned}
$$

- When p is large, it is challenging to find the best subset.

  e.g., $2^{40} \approx 1,000,000,000,000.$

Introduction

Challenges of high-dimensional variable selection

- There are 2^p candidate models

  e.g., for $p = 3$

$$\begin{aligned}
\texttt{Model 1}: \quad Y &= \beta_0 + \epsilon \\
\texttt{Model 2}: \quad Y &= \beta_0 + \beta_1 X_1 + \epsilon \\
\texttt{Model 3}: \quad Y &= \beta_0 + \beta_2 X_2 + \epsilon \\
\texttt{Model 4}: \quad Y &= \beta_0 + \beta_3 X_3 + \epsilon \\
\texttt{Model 5}: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\
\texttt{Model 6}: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon \\
\texttt{Model 7}: \quad Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \\
\texttt{Model 8}: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon
\end{aligned}$$

- When p is large, it is challenging to find the best subset.

  e.g., $2^{40} \approx 1,000,000,000,000$.

Introduction

Connection between sparse estimation and variable selection

- Variable selection is equivalent to estimating sparse coefficients.
  - Linear regression:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ &\Updownarrow \\ Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + 0 X_3 + \cdots + 0 X_p + \epsilon \end{aligned}$$

  - Logistic regression:

$$\begin{aligned} P(Y=1) &= \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)} \\ &\Updownarrow \\ P(Y=1) &= \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + 0 X_3 + \cdots + 0 X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + 0 X_3 + \cdots + 0 X_p)} \end{aligned}$$

- Hence, variable selection can be done by producing sparse estimator of coefficients.

High-dimensional linear regression models

- Define

$$\left[\begin{array}{c|cccc} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array}\right] = \left[\begin{array}{c|cccc} \boldsymbol{y} & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_p \end{array}\right]$$

- Consider

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$.

- Assume that $p > n$.

- Our aim is to obtain a sparse estimator for β.

Sparse estimation with L_0-penalty

- The sparse estimator can be obtained by minimizing

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_0,$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$ and $\lambda \geq 0$ controls degrees of sparsity.

- e.g., $\lambda = 2$ (AIC), $\lambda = \log n$ (BIC), $\lambda = \log p$ (RIC), ...

- However, the use of L_0-penalty leads to a non-convex optimization problem, which is computationally intractable in high-dimensional settings.

Sparse estimation with L_0-penalty

- The sparse estimator can be obtained by minimizing

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_0,$$

  where $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$ and $\lambda \geq 0$ controls degrees of sparsity.

- e.g., $\lambda = 2$ (AIC), $\lambda = \log n$ (BIC), $\lambda = \log p$ (RIC), ...

- However, the use of L_0-penalty leads to a non-convex optimization problem, which is computationally intractable in high-dimensional settings.

Sparse estimation with convex penalties

- Many convex penalty functions have been proposed.
  e.g., lasso, adaptive lasso, elastic net, MCP, ...

- The lasso estimator (Tibshirani, 1996) can be obtained by minimizing

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

  where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.

- However, optimal λ selection is required. (CV, GCV, BIC, ...)

- In addition, convex penalties generate the shrinkage bias on the resulting estimator of β.

Sparse estimation with convex penalties

- Many convex penalty functions have been proposed.
  e.g., lasso, adaptive lasso, elastic net, MCP, ...

- The lasso estimator (Tibshirani, 1996) can be obtained by minimizing

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

  where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.

- However, optimal λ selection is required. (CV, GCV, BIC, ...)

- In addition, convex penalties generate the shrinkage bias on the resulting estimator of $\boldsymbol{\beta}$.

L_0-penalty vs Convex penalties

	L_0-penalty	Convex penalties
Computation	hard	easy
Bias	unbiased	biased
λ selection	easy	hard
$var(\hat{\beta})$	available	mostly n/a

- We propose a Bayesian method to overcome challenges in non-convex optimization.

Reduced models

- Let γ be an index set, $\gamma \subset \{1, \ldots, p\}$.
- Let $\boldsymbol{X}_\gamma$ be a sub-matrix of $\boldsymbol{X}$ containing $\boldsymbol{x}_j$ $j \in \gamma$.
  e.g. $\gamma = \{1, 2, 3\} \Rightarrow \boldsymbol{X}_\gamma = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$.
- Given γ, our model reduces to

$$\boldsymbol{y} = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}_\gamma$ is a sub-vector of $\boldsymbol{\beta}$ corresponding to γ.

Bayesian best subset selection

- Suppose that we are interested in finding a best subset of size k.
- In a Bayesian framework, best subset selection can be done by estimating $\boldsymbol{\gamma}$.
- The Bayesian estimator of $\boldsymbol{\gamma}$ is obtained by maximizing

$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \int f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2)\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2)d(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2),$$

where $f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \gamma, \sigma^2)$ is the likelihood and $\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \gamma, \sigma^2)$ is the prior.

Prior specification & posterior distribution

- For computational convenience, we consider

$$\begin{aligned}\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2,\boldsymbol{\gamma} &\sim \text{Normal}(0,\tau\sigma^2\boldsymbol{I}_{|\boldsymbol{\gamma}|}),\\ \sigma^2 &\sim \text{Inverse-Gamma}(a_\sigma/2, b_\sigma/2),\\ \pi(\boldsymbol{\gamma}) &\propto \mathbb{I}(|\boldsymbol{\gamma}|=k),\end{aligned}$$

  where $|\gamma|$ denotes the number of elements in γ.

- Given k, it can be shown that

$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \frac{(\tau^{-1})^{\frac{|\boldsymbol{\gamma}|}{2}}}{|\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + \tau^{-1}\boldsymbol{I}_{|\gamma|}|^{\frac{1}{2}}\,(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}_\gamma\boldsymbol{y}+b_\sigma)^{\frac{a_\sigma+n}{2}}}\mathbb{I}(|\boldsymbol{\gamma}|=k),$$

  where $\boldsymbol{H}_\gamma = \boldsymbol{I}_n - \boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + \tau^{-1}\boldsymbol{I}_{|\gamma|})^{-1}\boldsymbol{X}_\gamma^{\mathrm{T}}$.

Stochastic search via MCMC

- A simplest way is to generate a random sample from $\pi(\gamma|\boldsymbol{y})$:

$$\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(T)} \overset{iid}{\sim} \pi(\gamma|\boldsymbol{y}) \quad \Rightarrow \quad \hat{\gamma} = \arg \max_{1 \le t \le T} \pi(\gamma^{(t)}|\boldsymbol{y}),$$

  but this is impossible due to the complexity of $\pi(\gamma|\boldsymbol{y})$.

- As an alternative, we can generate a Markov chain using Markov chain Monte Carlo (MCMC) computation.

- However, MCMC algorithms are too slow and often fail when p is large.

Shotgun Stochastic Search (SSS)

- Hans et al. (2007) propose SSS using the idea of parallel computation.
- Let $\gamma^{(t)}$ be a current state of γ and $\hat{\gamma}$ be a current best subset.
- Define a neighborhood of $\gamma^{(t)}$ as

$$\mathcal{N}(\gamma^{(t)}) = \{\gamma^{(t)}\} \cup \{\gamma^{(t)} \cup \{j\} : j \notin \gamma^{(t)}\} \cup \{\gamma^{(t)} \setminus \{j'\} : j' \in \gamma^{(t)}\}.$$

SSS algorithm

- Update $\hat{\gamma}$ by iterating the following steps:

  Step1. Compute $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ for each $\boldsymbol{\gamma} \in \mathcal{N}(\boldsymbol{\gamma}^{(t)})$ in parallel (parallel computing).

  Step2. If $\pi(\hat{\boldsymbol{\gamma}}|\boldsymbol{y}) < \max_{\boldsymbol{\gamma} \in \mathcal{N}(\boldsymbol{\gamma}^{(t)})} \pi(\boldsymbol{\gamma}|\boldsymbol{y})$, then update

  $$\hat{\boldsymbol{\gamma}} = \arg \max_{\boldsymbol{\gamma} \in \mathcal{N}(\boldsymbol{\gamma}^{(t)})} \pi(\boldsymbol{\gamma}|\boldsymbol{y})$$

  Step3. Update $\boldsymbol{\gamma}^{(t+1)}$ by generating a sample from $\mathcal{N}(\boldsymbol{\gamma}^{(t)})$ with probabilities proportional to $\pi(\boldsymbol{\gamma}|\boldsymbol{y})\mathbb{I}\{\boldsymbol{\gamma} \in \mathcal{N}(\boldsymbol{\gamma}^{(t)})\}$.

Limitations of SSS

- When k is fixed, SSS is not applicable.
- When parallel computing is not available, SSS is inefficient.
- Stochastic search algorithm requires a “burn-in” period.

Hybrid best subset search with a fixed k

```
1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. Repeat    #deterministic search
       Update $\tilde{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_+(\hat{\gamma})} \pi(\gamma|\boldsymbol{y})$ ;    # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$
       Update $\hat{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_-(\tilde{\gamma})} \pi(\gamma|\boldsymbol{y})$;    # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$
   until convergence.
3. Set $\gamma^{(0)} = \hat{\gamma}$.
4. Repeat for $t = 1, \ldots, T$:    #stochastic search
       Generate $\gamma^* \sim \pi_\alpha(\gamma|\boldsymbol{y}) \propto \{\pi(\gamma|\boldsymbol{y})\}^\alpha \mathbb{I}\{\gamma \in \mathcal{N}_+(\gamma^{(t-1)})\}$;    # $\alpha \in [0,1]$
       Generate $\gamma^{(t)} \sim \pi_\alpha(\gamma|\boldsymbol{y}) \propto \{\pi(\gamma|\boldsymbol{y})\}^\alpha \mathbb{I}\{\gamma \in \mathcal{N}_-(\gamma^*)\}$;
       If $\pi(\hat{\gamma}|\boldsymbol{y}) < \pi(\gamma^{(t)}|\boldsymbol{y})$, then update $\hat{\gamma} = \gamma^{(t)}$, break the loop, and go to 2.
5. Return $\hat{\gamma}$.
```

Hybrid best subset search with a fixed k

```
1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. Repeat   #deterministic search
        Update $\tilde{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_+(\hat{\gamma})} \pi(\gamma|\boldsymbol{y})$ ;   # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$
        Update $\hat{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_-(\tilde{\gamma})} \pi(\gamma|\boldsymbol{y})$;   # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$
   until convergence.
3. Set $\boldsymbol{\gamma}^{(0)} = \hat{\gamma}$.
4. Repeat for $t = 1, \ldots, T$:   #stochastic search
        Generate $\gamma^* \sim \pi_\alpha(\gamma|\boldsymbol{y}) \propto \{\pi(\gamma|\boldsymbol{y})\}^\alpha \mathbb{I}\{\gamma \in \mathcal{N}_+(\boldsymbol{\gamma}^{(t-1)})\}$;   # $\alpha \in [0, 1]$
        Generate $\gamma^{(t)} \sim \pi_\alpha(\gamma|\boldsymbol{y}) \propto \{\pi(\gamma|\boldsymbol{y})\}^\alpha \mathbb{I}\{\gamma \in \mathcal{N}_-(\boldsymbol{\gamma}^*)\}$;
        If $\pi(\hat{\gamma}|\boldsymbol{y}) < \pi(\boldsymbol{\gamma}^{(t)}|\boldsymbol{y})$, then update $\hat{\gamma} = \gamma^{(t)}$, break the loop, and go to 2.
5. Return $\hat{\gamma}$.
```

Key features of proposed algorithm

- In Steps 2, computing $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ for all $\boldsymbol{\gamma} \in \mathcal{N}_{+}(\hat{\gamma})$ (or all $\boldsymbol{\gamma} \in \mathcal{N}_{-}(\tilde{\gamma})$) can be done simultaneously in a single computation.

- In Step 4, the idea of escort distribution (used in statistical physics and thermodynamics) is introduced to stimulate the movement of Markov chain.

- An escort distribution of $p(x)$ is given as

$$p_{\alpha}(x) = \frac{\{p(x)\}^{\alpha}}{\sum_{x \in \mathcal{X}} \{p(x)\}^{\alpha}}.$$

Escort distributions

Let

$$p(x) = \begin{cases} 0.1 & x = 1 \\ 0.85 & x = 2 \\ 0.05 & x = 3 \end{cases}.$$

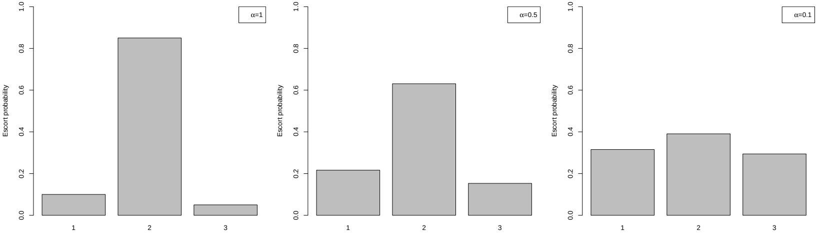


Figure: Escort distributions of $p(x)$.

Best subset selection with unknown k

- We extend the proposed method to best subset selection with varying $k(< K)$, where K is a pre-specified upper bound, $K < n$.
- In our Bayesian framework, this extension can be easily done by assigning a prior for k.
- Note that the uniform prior, $k \sim \text{Uniform}\{1, \ldots, K\}$, tends to assign larger probability to a larger subset.
- We define

$$\pi(k) \propto 1/\binom{p}{k}\mathbb{I}(k \le K).$$

Hybrid best subset search with varying k

- Bayesian best subset selection can be done by maximizing

$$\pi(\gamma, k|\boldsymbol{y}) \propto f(\boldsymbol{y}|\gamma, k)\pi(\gamma|k)\pi(k)$$

  over (γ, k).

- Our algorithm proceeds as follows:
  1. **Repeat** for $k = 1, \ldots, K$:
     a. Given k, implement the hybrid search algorithm.
     b. Set $\hat{\gamma}_k = \hat{\gamma}$.
  2. Find $\hat{\gamma}_{k_*}$ such that

$$\log f(\boldsymbol{y}|\hat{\gamma}_{k_*}, k_*) - \log \binom{p}{k_*} \geq \log f(\boldsymbol{y}|\hat{\gamma}_k, k) - \log \binom{p}{k}, \quad \text{any } k \leq K.$$

Connection to extended BIC

- It can be shown that our posterior criterion is asymptotically equivalent to the extended BIC (EBIC) of Chen and Chen (2008).

- EBIC corresponds to the L_0-penalty sparse estimation with

$$\lambda = \log(n) + \frac{2}{k} \log \binom{p}{k}.$$

Model selection consistency

- The proposed Bayesian approach possesses the model selection consistency in the high-dimensional setting with $p = p_n = O(n^{\xi})$ for $\xi \geq 1$.
- Hence, as $n \to \infty$, our variable selection procedure identifies the true model with probability tending to one.

Simulation study

Setup

- For given $n = 100$, we generate the data from

$$y_i \overset{ind}{\sim} \text{Normal}\left(\sum_{j=1}^{p} \beta_j x_{ij}, 1\right),$$

where

- $(x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \overset{iid}{\sim} \text{Normal}(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p}$ and $\Sigma_{ij} = \rho^{|i-j|}$,
- $\beta_j \overset{iid}{\sim} \text{Uniform}\{-1, -2, 1, 2\}$ if $j \in \gamma$ and $\beta_j = 0$ if $j \notin \gamma$.
- γ is an index set of size 4 randomly selected from $\{1, 2, \ldots, p\}$.
- We consider four scenarios for p and ρ:
  (i) $p = 200$, $\rho = 0.1$, (ii) $p = 200$, $\rho = 0.9$,
  (iii) $p = 1000$, $\rho = 0.1$, (iv) $p = 1000$, $\rho = 0.9$.

Simulation study

Results (high-dimensional scenarios)

Table: MC size=2,000; FDR (false discovery rate), TRUE% (percentage of the true model detected), SIZE (selected model size), HAM (Hamming distance).

Scenario	Method	FDR (s.e.)	TRUE% (s.e.)	SIZE (s.e.)	HAM (s.e.)
i	Proposed	0.006 (0.001)	96.900 (0.388)	4.032 (0.004)	0.032 (0.004)
	SCAD	0.034 (0.002)	85.200 (0.794)	4.188 (0.011)	0.188 (0.011)
	MCP	0.035 (0.002)	84.750 (0.804)	4.191 (0.011)	0.191 (0.011)
	ENET	0.016 (0.001)	92.700 (0.582)	4.087 (0.007)	0.087 (0.007)
	LASSO	0.020 (0.002)	91.350 (0.629)	4.109 (0.009)	0.109 (0.009)
ii	Proposed	0.023 (0.002)	88.750 (0.707)	3.985 (0.006)	0.203 (0.014)
	SCAD	0.059 (0.003)	74.150 (0.979)	4.107 (0.015)	0.480 (0.022)
	MCP	0.137 (0.004)	55.400 (1.112)	4.264 (0.020)	1.098 (0.034)
	ENET	0.501 (0.004)	0.300 (0.122)	7.716 (0.072)	5.018 (0.052)
	LASSO	0.276 (0.004)	15.550 (0.811)	5.308 (0.033)	2.038 (0.034)

Simulation study

Results (ultra high-dimensional scenarios)

Table: MC size$=2,000$; FDR (false discovery rate), TRUE% (percentage of the true model detected), SIZE (selected model size), HAM (Hamming distance).

Scenario	Method	FDR (s.e.)	TRUE% (s.e.)	SIZE (s.e.)	HAM (s.e.)
iii	Proposed	0.004 (0.001)	98.100 (0.305)	4.020 (0.003)	0.020 (0.003)
	SCAD	0.027 (0.002)	87.900 (0.729)	4.145 (0.010)	0.145 (0.010)
	MCP	0.031 (0.002)	86.550 (0.763)	4.172 (0.013)	0.172 (0.013)
	ENET	0.035 (0.002)	84.850 (0.802)	4.181 (0.013)	0.206 (0.012)
	LASSO	0.014 (0.001)	93.850 (0.537)	4.073 (0.007)	0.073 (0.007)
iv	Proposed	0.023(0.002)	89.850 (0.675)	4.005 (0.005)	0.190 (0.013)
	SCAD	0.068 (0.003)	74.250 (0.978)	4.196 (0.014)	0.493 (0.023)
	MCP	0.152 (0.004)	53.750 (1.115)	4.226 (0.017)	1.202 (0.035)
	ENET	0.417 (0.005)	0.150 (0.087)	6.228 (0.068)	4.089 (0.043)
	LASSO	0.265 (0.004)	19.500 (0.886)	5.139 (0.029)	1.909 (0.035)

Real data application

Data description

- We apply the proposed method to Breast Invasive Carcinoma (BRCA) data generated by The Cancer Genome Atlas (TCGA) Research Network `http://cancergenome.nih.gov`.
- The data set is available at the `R` package `curatedTCGAData`.
- The data set contains $17,814$ gene expression measurements (recorded on the log scale) of 526 patients with primary solid tumor.
- BRCA1 is a tumor suppressor gene and its mutations predispose women to breast cancer (Findlay et al., 2018).
- Our goal here is to identify the best fitting model for estimating an association between BRCA1 (response variable) and the other genes (independent variables).

Real data application

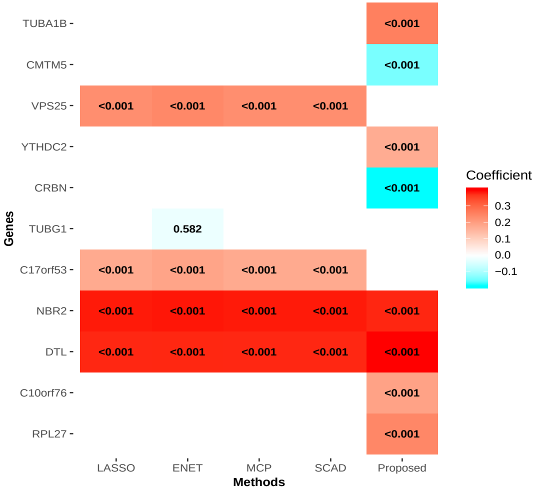
Results (based on $4,000$ genes)

Table: Model comparison

	# of selected	PMSE	BIC	EBIC
Proposed	8	0.60	984.45	1099.50
SCAD	4	0.68	1104.69	1166.47
MCP	4	0.68	1104.69	1166.47
ENET	5	0.68	1110.65	1186.25
LASSO	4	0.68	1104.69	1166.47

Real data application

Results (cont.)



Genes	LASSO	ENET	MCP	SCAD	Proposed
TUBA1B					<0.001
CMTM5					<0.001
VPS25	<0.001	<0.001	<0.001	<0.001	
YTHDC2					<0.001
CRBN					<0.001
TUBG1		0.582			
C17orf53	<0.001	<0.001	<0.001	<0.001	
NBR2	<0.001	<0.001	<0.001	<0.001	<0.001
DTL	<0.001	<0.001	<0.001	<0.001	<0.001
C10orf76					<0.001
RPL27					<0.001

Figure: Except C10orf76, 7 genes are documented as *diseases-related genes*

Concluding remarks

- Parallel computing is applicable to our algorithm with varying k.
- The proposed method can be extended to multivariate linear regression models, binary regression models, and multivariate mixed responses models (in progress).

REFERENCES

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association 102*(478), 507–516.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Findlay, G. M., R. M. Daza, B. Martin, M. D. Zhang, A. P. Leith, M. Gasperini, J. D. Janizek, X. Huang, L. M. Starita, and J. Shendure (2018). Accurate classification of brca1 variants with saturation genome editing. *Nature 562*(7726), 217.

Q/A

THANK YOU