Bayesian selection of best subsets in high-dimensional regression

Shiqiang Jin

Department of Statistics
Kansas State University, Manhattan, KS

Joint work with

Gyuhyeong Goh

Kansas State University, Manhattan, KS

July 31, 2019

Bayesian linear regression model in High-dimensional Data

- Consider a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is a response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ is a model matrix, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a coefficient vector and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$.

- We assume $p > n$, i.e. High-dimensional data.

- We assume only a few number of predictors are associated with the response, i.e. β is sparse.

Bayesian linear regression model in High-dimensional Data

- To better explain the sparsity of β, we introduce a latent index set $\gamma \subset \{1, \ldots, p\}$ so that $\boldsymbol{X}_\gamma$ represents a sub-matrix of $\boldsymbol{X}$ containing $\boldsymbol{x}_j$, $j \in \gamma$.
- e.g. $\gamma = \{1, 3, 4\} \Rightarrow \boldsymbol{X}_\gamma = (\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_4)$.
- The full model in (1) can be reduced to

$$\boldsymbol{y} = \boldsymbol{X}_\gamma \beta_\gamma + \boldsymbol{\epsilon}. \tag{2}$$

Priors and marginal posterior distribution

- Our Research Goals are to obtain:

(i) k most important predictors out of $\binom{p}{k}$ candidate models;

(ii) a single best model from among 2^p candidate models.

- We consider

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

where $|\gamma|$ is number of elements in γ.

Priors and marginal posterior distribution

- Our Research Goals are to obtain:

(i) k most important predictors out of $\binom{p}{k}$ candidate models;

(ii) a single best model from among 2^p candidate models.

- We consider

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

where $|\gamma|$ is number of elements in γ.

Priors and marginal posterior distribution

- Given k, it can be shown that

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

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

- Hence, $g(\boldsymbol{\gamma})$ is our model selection criterion.

Best subset selection Algorithm

- Note that our goals are to obtain (i) k most important predictors out of $\binom{p}{k}$ models; (ii) a single best model from among 2^p candidate models.

- Hence, this can be fallen into the description of best subset selection as follows:

(i) Fixed size: for $k = 1, 2, \ldots, p$, select the best subset model by

$$\mathcal{M}_k = \arg_\gamma \max_{|\gamma|=k} g(\gamma)$$

from $\binom{p}{k}$ candidate models.

(ii) Varying size: Pick the single best model from $\mathcal{M}_1, \ldots, \mathcal{M}_p$.

- Challenge of best subset selection:
  e.g. (i) $p = 1000, k = 5$, $\binom{1000}{5} \approx 8 \times 10^{12}$; (ii) $p = 40$, $2^{40} \approx 10^{12}$.

Best subset selection Algorithm

- Note that our goals are to obtain (i) k most important predictors out of $\binom{p}{k}$ models; (ii) a single best model from among 2^p candidate models.

- Hence, this can be fallen into the description of best subset selection as follows:

(i) Fixed size: for $k = 1, 2, \ldots, p$, select the best subset model by

$$\mathcal{M}_k = \arg_\gamma \max_{|\gamma|=k} g(\gamma)$$

from $\binom{p}{k}$ candidate models.

(ii) Varying size: Pick the single best model from $\mathcal{M}_1, \ldots, \mathcal{M}_p$.

- Challenge of best subset selection:
e.g. (i) $p = 1000, k = 5$, $\binom{1000}{5} \approx 8 \times 10^{12}$; (ii) $p = 40$, $2^{40} \approx 10^{12}$.

Best subset selection Algorithm

- Note that our goals are to obtain (i) k most important predictors out of $\binom{p}{k}$ models; (ii) a single best model from among 2^p candidate models.

- Hence, this can be fallen into the description of best subset selection as follows:

(i) Fixed size: for $k = 1, 2, \ldots, p$, select the best subset model by

$$\mathcal{M}_k = \arg_{\gamma} \max_{|\gamma|=k} g(\gamma)$$

from $\binom{p}{k}$ candidate models.

(ii) Varying size: Pick the single best model from $\mathcal{M}_1, \ldots, \mathcal{M}_p$.

- Challenge of best subset selection:
e.g. (i) $p = 1000, k = 5$, $\binom{1000}{5} \approx 8 \times 10^{12}$; (ii) $p = 40$, $2^{40} \approx 10^{12}$.

Neighborhood Search

- To avoid the exhaustive computation, we resort to the idea of Neighborhood Search proposed by Madigan et al. (1995) and Hans et al. (2007).
- Let $\gamma^{(t)}$ be a current state of γ, $|\gamma^{(t)}| = k$ is model size.
- Addition neighbor: $\mathcal{N}_+(\gamma^{(t)}) = \{\gamma^{(t)} \cup \{j\} : j \notin \gamma^{(t)}\}$; model size?
- Deletion neighbor: $\mathcal{N}_-(\gamma^{(t)}) = \{\gamma^{(t)} \setminus \{j'\} : j' \in \gamma^{(t)}\}$; model size?
- e.g. Suppose $p = 4, k = 2$. Let $\gamma^{(t)} = \{1, 2\}$, then

$$\begin{aligned} \mathcal{N}_+(\gamma^{(t)}) &= \{\{1,2,3\},\{1,2,4\}\}, \\ \mathcal{N}_-(\gamma^{(t)}) &= \{\{1\},\{2\}\}. \end{aligned}$$

Hybrid best subset search with a fixed k

- Note our Goal (i) is to find $\hat{\gamma} = \arg_{\gamma} \max_{|\hat{\gamma}|=k} g(\gamma)$.

1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. **Repeat** `#deterministic search:local optimum`

   Update $\tilde{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_+(\hat{\gamma})} g(\gamma)$; # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$

   Update $\hat{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_-(\tilde{\gamma})} g(\gamma)$; # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$

   until convergence.

- In Step 2, we have $p+1$ many candidate models in all $\gamma \in \mathcal{N}_+(\hat{\gamma})$ and all $\gamma \in \mathcal{N}_-(\tilde{\gamma})$) in each iteration.
- compute $g(\gamma)$ $p+1$ times in each iteration.

Hybrid best subset search with a fixed k

- Note our Goal (i) is to find $\hat{\gamma} = \arg_{\gamma} \max_{|\hat{\gamma}|=k} g(\gamma)$.

1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. **Repeat** `#deterministic search:local optimum`

   Update $\tilde{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_+(\hat{\gamma})} g(\gamma)$; # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$

   Update $\hat{\gamma} \leftarrow \arg\max_{\gamma \in \mathcal{N}_-(\tilde{\gamma})} g(\gamma)$; # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$

   until convergence.

- In Step 2, we have $p+1$ many candidate models in all $\gamma \in \mathcal{N}_+(\hat{\gamma})$ and all $\gamma \in \mathcal{N}_-(\tilde{\gamma}))$ in each iteration.
- compute $g(\gamma)$ $p+1$ times in each iteration.

1st Key features of proposed algorithm

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

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

- compute inverse matrix and determinant $p+1$ times.
- We propose the following formula and we can show that evaluating all candidate models in addition neighbor can be done simultaneously in this single computation.

$$\begin{aligned}\boldsymbol{g}(\mathcal{N}_+(\hat{\gamma})) \quad = \quad & \left\{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}_{\hat{\gamma}}\boldsymbol{y} + b_\sigma)\mathbf{1}_p - \frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}_{\hat{\gamma}}\boldsymbol{y})^2}{\tau^{-1}\mathbf{1}_p + \mathrm{diag}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}_{\hat{\gamma}}\boldsymbol{X})}\right\}^{-\frac{a_\sigma+n}{2}} \times \\ & \left\{\tau^{-1}\mathbf{1}_p + \mathrm{diag}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}_{\hat{\gamma}}\boldsymbol{X})\right\}^{-1/2}, \qquad (3)\end{aligned}$$

- Similarly for $\boldsymbol{g}(\mathcal{N}_-(\hat{\gamma}))$.

Hybrid best subset search with a fixed k

- Note our Goal (i) is to find $\hat{\gamma} = \arg_{\gamma} \max_{|\hat{\gamma}|=k} g(\gamma)$.

1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. **Repeat** `#deterministic search:local optimum`

   Update $\tilde{\gamma} \leftarrow \arg\max \boldsymbol{g}(\mathcal{N}_+(\hat{\gamma}))$; # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$

   Update $\hat{\gamma} \leftarrow \arg\max \boldsymbol{g}(\mathcal{N}_-(\tilde{\gamma}))$; # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$

   until convergence.
3. Set $\gamma^{(0)} = \hat{\gamma}$.
4. **Repeat** for $t = 1, \ldots, T$: `#stochastic search:global optimum`

   Sample $\gamma^* \sim \pi(\gamma|\boldsymbol{y}) = \frac{\{g(\gamma)\}}{\sum_{\gamma}\{g(\gamma)\}} \mathbb{I}\{\gamma \in \mathcal{N}_+(\gamma^{(t-1)})\}$;

   Sample $\gamma^{(t)} \sim \pi(\gamma|\boldsymbol{y}) = \frac{\{g(\gamma)\}}{\sum_{\gamma}\{g(\gamma)\}} \mathbb{I}\{\gamma \in \mathcal{N}_-(\gamma^*)\}$;

   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 $\hat{\gamma}$.

- Note that all $g(\gamma)$ are computed simultaneously in their neighbor space.

Hybrid best subset search with a fixed k

- Note our Goal (i) is to find $\hat{\gamma} = \arg_{\gamma} \max_{|\hat{\gamma}|=k} g(\gamma)$.

1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. **Repeat** `#deterministic search:local optimum`

   Update $\tilde{\gamma} \leftarrow \arg\max \boldsymbol{g}(\mathcal{N}_+(\hat{\gamma}))$; # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$

   Update $\hat{\gamma} \leftarrow \arg\max \boldsymbol{g}(\mathcal{N}_-(\tilde{\gamma}))$; # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$

   until convergence.
3. Set $\gamma^{(0)} = \hat{\gamma}$.
4. **Repeat** for $t = 1, \ldots, T$: `#stochastic search:global optimum`

   Sample $\gamma^* \sim \pi(\gamma|\boldsymbol{y}) = \frac{\{g(\gamma)\}}{\sum_{\gamma}\{g(\gamma)\}} \mathbb{I}\{\gamma \in \mathcal{N}_+(\gamma^{(t-1)})\}$;

   Sample $\gamma^{(t)} \sim \pi(\gamma|\boldsymbol{y}) = \frac{\{g(\gamma)\}}{\sum_{\gamma}\{g(\gamma)\}} \mathbb{I}\{\gamma \in \mathcal{N}_-(\gamma^*)\}$;

   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 $\hat{\gamma}$.

- Note that all $g(\gamma)$ are computed simultaneously in their neighbor space.

2nd Key features of proposed algorithm

- To avoid staying in the local optimum for a long time in step 4, we use the escort distribution.
- The idea of escort distribution (used in statistical physics and thermodynamics) is introduced to stimulate the movement of Markov chain.
- An escort distribution of $g(\gamma)$ is given as

$$\frac{\{g(\gamma)\}^{\alpha}}{\sum_{\gamma}\{g(\gamma)\}^{\alpha}}, \alpha \in [0,1]$$

Escort distributions

- e.g. Consider 3 candidate models with its probability:
- If $\alpha = 1$,

$$\frac{\{g(\gamma)\}^{\alpha}}{\sum_{\gamma}\{g(\gamma)\}^{\alpha}} = \begin{cases} 0.02 & \text{model 1} \\ 0.90 & \text{model 2} \\ 0.08 & \text{model 3} \end{cases}.$$

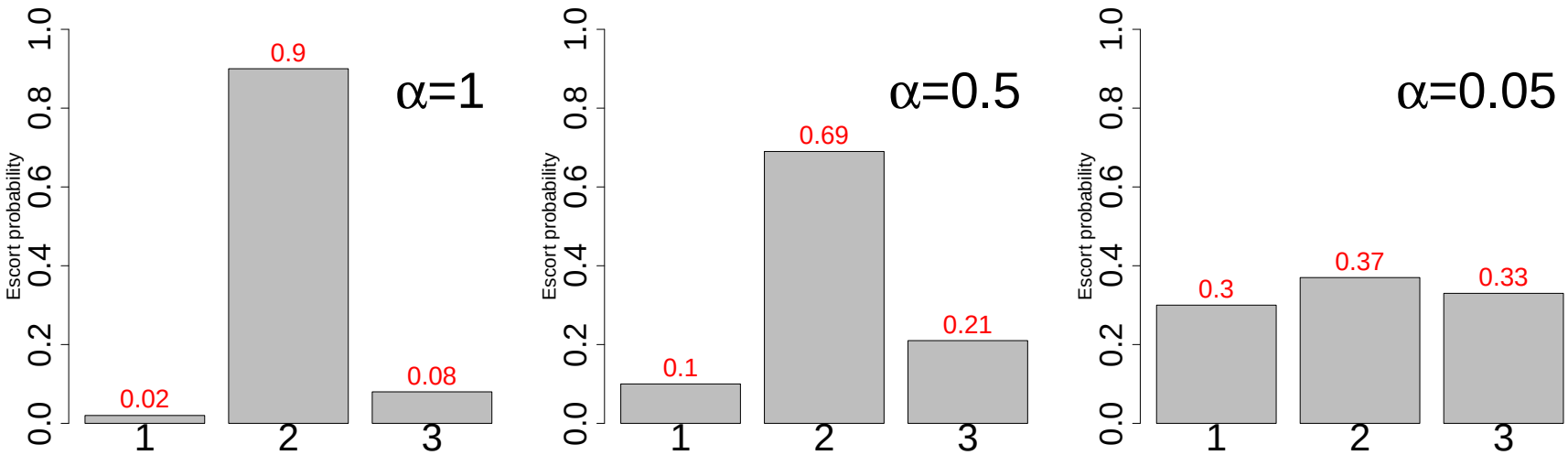

Figure: Escort distributions of $\pi_{\alpha}(\gamma|\boldsymbol{y})$.

Hybrid best subset search with a fixed k

- Note our Goal (i) is to find $\hat{\gamma} = \arg_{\gamma} \max_{|\hat{\gamma}|=k} g(\gamma)$.

1. Initialize $\hat{\gamma}$ s.t. $|\hat{\gamma}| = k$.
2. **Repeat** `#deterministic search:local optimum`

   Update $\tilde{\gamma} \leftarrow \arg\max \boldsymbol{g}(\mathcal{N}_+(\hat{\gamma}))$; # $\mathcal{N}_+(\hat{\gamma}) = \{\hat{\gamma} \cup \{j\} : j \notin \hat{\gamma}\}$

   Update $\hat{\gamma} \leftarrow \arg\max \boldsymbol{g}(\mathcal{N}_-(\tilde{\gamma}))$; # $\mathcal{N}_-(\tilde{\gamma}) = \{\tilde{\gamma} \setminus \{j\} : j \in \tilde{\gamma}\}$

   until convergence.
3. Set $\gamma^{(0)} = \hat{\gamma}$.
4. **Repeat** for $t = 1, \ldots, T$: `#stochastic search:global optimum`

   Sample $\gamma^* \sim \pi_\alpha(\gamma|\boldsymbol{y}) = \frac{\{g(\gamma)\}^\alpha}{\sum_\gamma \{g(\gamma)\}^\alpha} \mathbb{I}\{\gamma \in \mathcal{N}_+(\gamma^{(t-1)})\}$; # $\alpha \in [0, 1]$

   Sample $\gamma^{(t)} \sim \pi_\alpha(\gamma|\boldsymbol{y}) = \frac{\{g(\gamma)\}^\alpha}{\sum_\gamma \{g(\gamma)\}^\alpha} \mathbb{I}\{\gamma \in \mathcal{N}_-(\gamma^*)\}$;

   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 $\hat{\gamma}$.

- Note that all $g(\gamma)$ are computed simultaneously in their neighbor space.

Best subset selection with varying k

Note that Goal (ii): a single best model from among 2^p candidate models.

- We extend "fixed" k to varying k by assigning a prior on k.
- Note that the uniform prior, $k \sim \text{Uniform}\{1, \ldots, K\}$, tends to assign larger probability to a larger subset (see Chen and Chen (2008)).
- We define

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

Hybrid best subset search with varying k

- Bayesian best subset selection can be done by maximizing

$$\pi(\gamma, k|\boldsymbol{y}) \propto g(\gamma)/\binom{p}{k} \tag{4}$$

over (γ, k).

- Our algorithm proceeds as follows:
  1. **Repeat** for $k = 1, \ldots, K$:
     Given k, implement the hybrid search algorithm to obtain best subset model $\hat{\gamma}_k$.
  2. Find the best model $\hat{\gamma}^*$ obtained by

$$\hat{\gamma}^* = \arg \max_{k \in \{1,\ldots,K\}} \left(g(\hat{\gamma}_k)/\binom{p}{k} \right). \tag{5}$$

Consistency of model selection criterion

Theorem

Let γ_* indicate the true model. Define $\Gamma = \{\gamma : |\gamma| \leq K, \gamma \neq \gamma_*\}$. Assume that $p = O(n^\xi)$ for $\xi \geq 1$. Under the asymptotic identifiability condition of Chen and Chen (2008), if $\tau \to \infty$ as $n \to \infty$ but $\tau = o(n)$, then the proposed Bayesian subset selection possesses the Bayesian model selection consistency, that is,

$$\pi(\gamma_*|\boldsymbol{y}) > \max_{\gamma \in \Gamma} \pi(\gamma|\boldsymbol{y}) \tag{6}$$

in probability as $n \to \infty$.

- As $n \to \infty$, the maximizer of $\pi(\gamma|\boldsymbol{y})$ is the true model based on our model selection criterion.

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

  - $(x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \overset{iid}{\sim} \text{Normal}(\mathbf{0}_p, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\Sigma_{ij})_{p \times p}$ 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: 2,000 replications; 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.)
$p = 200$	Proposed	0.006 (0.001)	96.900 (0.388)	4.032 (0.004)	0.032 (0.004)
$\&\rho = 0.1$	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)
$p = 200$	Proposed	0.023 (0.002)	88.750 (0.707)	3.985 (0.006)	0.203 (0.014)
$\&\rho = 0.9$	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: 2,000 replications; 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.)
$p = 1000$	Proposed	0.004 (0.001)	98.100 (0.305)	4.020 (0.003)	0.020 (0.003)
$\&\rho = 0.1$	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)
$p = 1000$	Proposed	0.023(0.002)	89.850 (0.675)	4.005 (0.005)	0.190 (0.013)
$\&\rho = 0.9$	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 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).

Real data application

Results (based on 4,000 genes)

- Our goal here is to identify the best fitting model for estimating an association between BRCA1 (response variable) and the other genes (independent variables).

$$\text{BRCA1} = \beta_1 * \text{NBR2} + \beta_2 * \text{DTL} + \ldots + \beta_p * \text{VPS25} + \epsilon.$$

- Results:

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.)

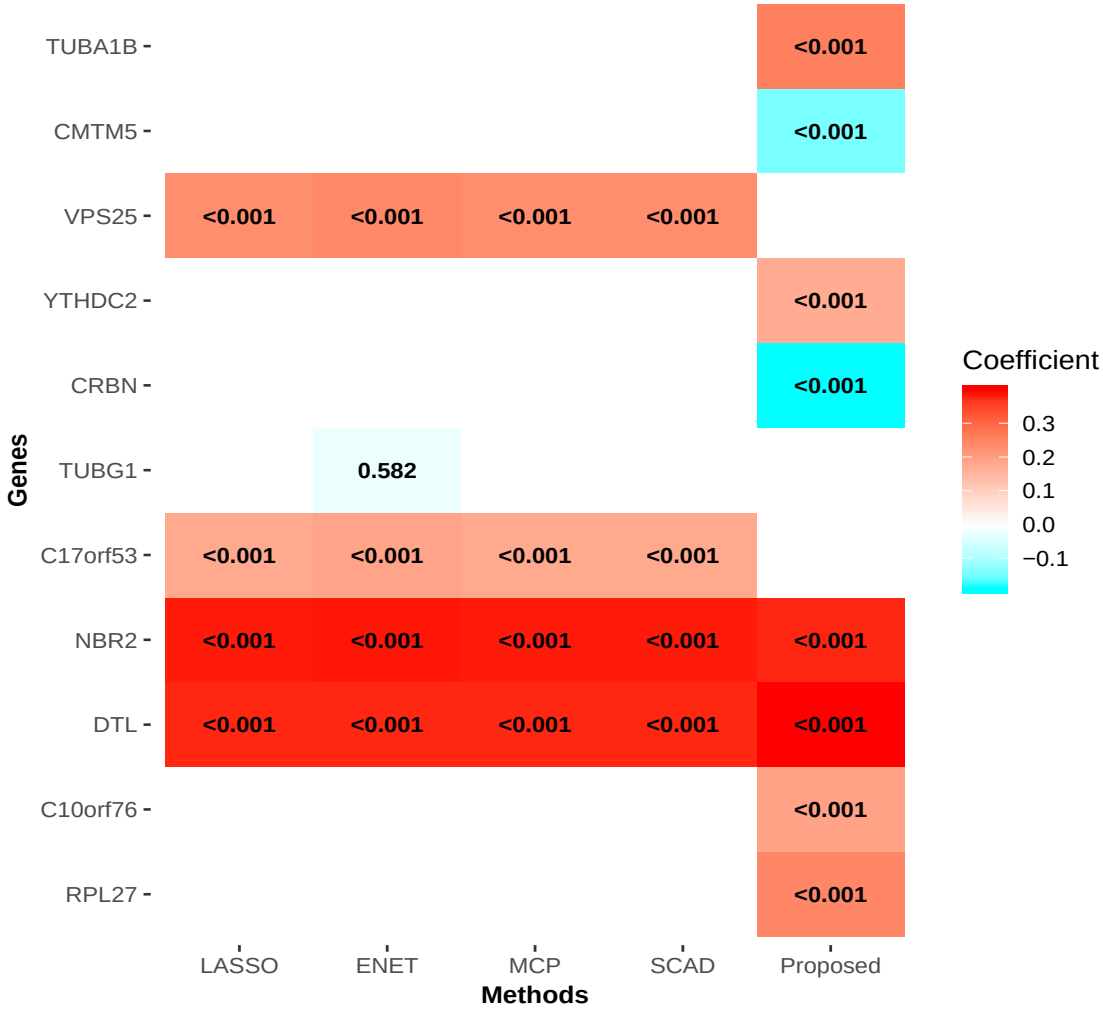

Figure: Except C10orf76, 7 genes are documented as diseases-related genes

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

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.

Madigan, David, Jeremy York, and Denis Allard (1995). Bayesian Graphical Models for Discrete Data. International Statistical Review / Revue Internationale De Statistique 63(2), 215-232.

Contact: jinsq@ksu.edu

THANK YOU