A New Bayesian Variable Selection Method: The Bayesian Lasso with Pseudo Variables

Qi Tang
(Joint work with Kam-Wah Tsui and Sijian Wang)

Department of Statistics
University of Wisconsin-Madison

Feb. 8, 2010

Outline

- Introduction of Bayesian Lasso
- Bayesian Lasso with Pseudo Variables
- Bayesian Group Lasso with Pseudo Variables
- Conclusions and Future Work

Variable Selection

- Why?
    - **Interpretation**: principle of parsimony.
    - **Prediction**: bias and variance tradeoff.
- What if number of variables is greater than number of observations ($p > n$)?
- Shrinkage.
    - Frequentist: loss + penalty. Examples: Ridge regression(Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996).
    - Bayesian: Likelihood $\times$ Shrinkage prior. Griffin and Brown (2005), Park and Casella (2008).

Notation

- Consider a data set with one response variable, p predictors and n observations.
- Focus on linear models: $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I_n)$.
- y is the centered response; X_is, columns of X, are standardized to have mean 0 and unit L_2 norm.

Bayesian Interpretation of Lasso

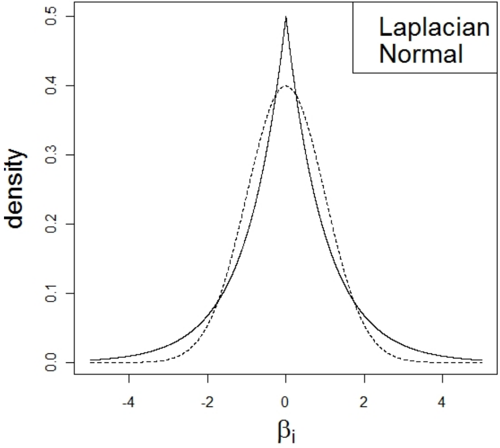
- Lasso (Tibshirani 1996):

$$\min_{\beta}\{\|y - X\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|\} \tag{1}$$

- Bayesian interpretation:
  - Consider the Bayesian model $y \sim N(X\beta, I_n)$ and $\beta_i \sim \frac{\lambda}{2} e^{-\lambda|\beta_i|}$ (Laplacian prior).
  - The solution of (1) can be interpreted as the posterior mode of β.

Laplacian Priors

- The Laplacian prior is more sparsity promoting than the normal prior.



The Bayesian Lasso (Park and Casella, 2008)

- Model $y \mid (X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I_n)$.
- Propose the conditional Laplacian prior

$$\beta_i \mid (\sigma^2, \lambda^2) \sim \frac{\lambda}{2\sigma} e^{-\lambda|\beta_i|/\sigma},$$

- Rewrite the laplacian prior into a mixture of

$$\beta_i \mid (\sigma^2, \gamma_i^2) \sim N(0, \sigma^2\gamma_i^2) \text{ and } \gamma_i^2 \mid \sigma^2 \sim Exp(\lambda^2/2). \quad (2)$$

- Empirical Bayesian treatment of λ:
  - Estimate λ by the marginal maximum likelihood $\hat{\lambda}$.
  - Assign a hyperprior that places high density at $\hat{\lambda}$.
- Estimate β_i by its posterior median.
- Limitation: heavy computation load and sparsity NOT achieved.

Outline

- Introduction of Bayesian Lasso
- **Bayesian Lasso with Pseudo Variables**
- Bayesian Group Lasso with Pseudo Variables
- Conclusions and Future Work

Benefit of Our Method

- Avoid the computation burden of finding the marginal maximum likelihood estimate.
  - Assign a prior to λ^2 that does not depend on the data.
- Achieve sparsity.

Intuition for Achieving Sparsity

- Find an unimportant pseudo variable Z as the benchmark.
- Augment the model:

$$y = \beta_z Z + X\beta + \epsilon.$$

- Criteria:
  - Orthogonal with y (true value of β_z is 0).
  - Orthogonal with X_is (keep the data structure).

Benchmark: Intercept!

- $Z_{int} = (\underbrace{1/\sqrt{n}, \ldots, 1/\sqrt{n}}_{n})^T$.

- Orthogonal with y and all the X_is.

- Does NOT depend on the specific observations.

Variable Selection

- Regression model: $Y = \beta_{int} Z_{int} + X\beta + \epsilon$.
- Assign hierarchical priors and obtain posterior distributions of β_{int} and β_is.
- Measure the importance of X_i by $d_i = P(|\beta_i| > |\beta_{int}| \mid y, X)$.
  - If X_i is orthogonal with y and other variables, then $d_i = 0.5$.
  - The X_i will be selected as an important variable, if $d_i > c$, where $c > 0.5$.

Some Thoughts on Tuning c

1. Choose the c such that the false discovery rate is controlled.

2. Find the $\lim_{n\to\infty} d_i$ for the X_i that is unimportant but weakly correlated with the important variables. Use it as a guideline to choose c.

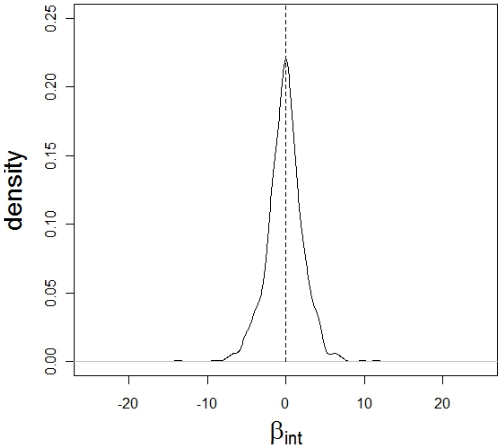
Illustration

Consider the following simulation setting (Tibshirani, 1996):

- $y = X\beta + \epsilon$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$.
- $X = (X_1, \dots, X_p)$, $X_i \sim N(0, 1)$, $cor(X_i, X_j) = 0.5^{|i-j|}$.
- $\sigma^2 = 9$.
- $n = 20$.

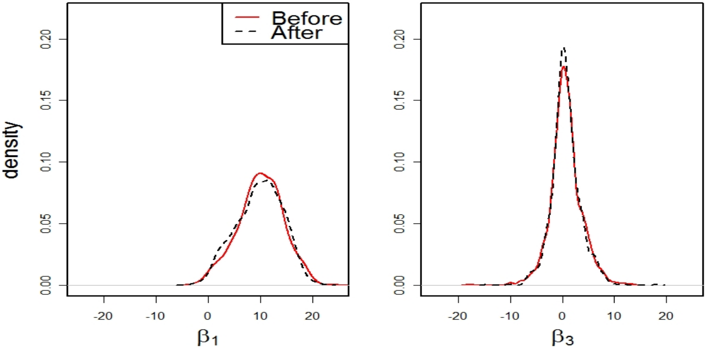
Posterior Distribution of β_{int}

- Z_{int} is a good benchmark for unimportant variables.



Changes on Posterior Distributions of β_is

- For each β_i, the estimated posterior densities are almost unaffected by adding Z_{int}.



Estimated d_i

- $\hat{d}_i$: proportion of (β_i, β_{int}) satisfying $|\beta_i| > |\beta_{int}|$.
- $\hat{\beta}_{i,PC}$: posterior median of β_i by Park and Casella's Bayesian Lasso method.
- All unimportant variables have $\hat{d}_i \leq 0.61$.
- Posterior medians do NOT yield sparsity.

β_i	3	1.5	0	0	2	0	0	0
$\hat{d}_i$	**0.96**	**0.64**	0.56	0.54	**0.78**	0.61	0.57	0.52
$\hat{\beta}_{i,PC}$	11.84	2.92	1.64	1.50	5.67	2.32	1.92	1.61

Variable Selection Result

- Empirically, $c = 0.9$ yields good sparsity.
- When $c = 0.6$, the result is almost the same as Lasso.

Table: Frequencies that each variable is selected.

β_i	3	1.5	0	0	2	0	0	0
$c = 0.9$	94	43	1	1	43	0	1	0
$c = 0.7$	100	87	19	26	93	9	14	14
$c = 0.6$	100	98	47	51	99	52	40	44
Lasso	100	96	47	51	99	48	43	46

Posteriors of β_is after Adding Z_{int}

Lemma

Consider regression model

$$y = X\beta + \epsilon \text{ with } \epsilon \sim N(0, \sigma^2 I_n)$$

and priors

$$\beta_i \mid (\sigma^2, \lambda^2) \sim \lambda/(2\sigma) e^{-\lambda|\beta_i|/\sigma}$$

for $i = 1, \ldots, p$. Let π_1 and π_2 be the joint posteriors of β_is conditional on σ^2 and λ^2 before and after adding Z_{int}, respectively. Then we have,

$$\pi_1 = \pi_2.$$

Outline

- Introduction of Bayesian Lasso
- Bayesian Lasso with Pseudo Variables
- **Bayesian Group Lasso with Pseudo Variables**
- Conclusions and Future Work

Motivation of Group Selection Method

- Assayed genes or proteins are naturally grouped by biological roles or biological pathways.
- It is desired to first select important pathways (group selection), and then select important genes (within group selection).
- Correlated important variables in the same group should all be selected.
  - Lasso tends to pick only a few of them.

Extra Notation

- g: Number of groups
- k: index of groups; j: index of variables inside groups. For example, $X_{k,j}$ is the jth variable in group k.
- p_k: number of variables in group k. Assume there is no overlap, that is, $p = \sum_{k=1}^{g} p_k$.

Current Lasso Type Methods for Group Selection

- Frequentist approach
  - Designed for group selection only: Yuan & Lin (2006)
  - Designed for both group selection and within group selection: Ma & Huang (2007); Huang at el. (2009); Wang at el. (2009).
- Bayesian approach. Raman et al.(2009).

Hierarchical Priors with Group Structure

- Model: $y = \beta_{int} Z_{int} + \sum_{k=1}^{g} \sum_{j=1}^{p_k} \beta_{k,j} X_{k,j} + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I_n)$.
- $\beta_{k,j} \sim N(0, \gamma_k^2 \sigma^2 / p_k)$.
  - Variables in the same group are shrunk simultaneously.
  - γ_k^2 measures the total variations of p_k variables: $\beta_{k,1}, \ldots, \beta_{k,p_k}$.
- $\gamma_k^2 \sim Exp\left(\frac{\lambda^2}{2p_k}\right)$.
  - Treat $\beta_{k,j}$ equally across groups. $E(\beta_{k,j})$ and $V(\beta_{k,j})$ do not depend on k or j.

Group Selection

- Definition of important group: groups have at least one important variable.

- Selection of important groups: the kth group is selected, if $\max_j\{P(|\beta_{k,j}| > |\beta_{int}|)\} > c$.

Within Group Selection: More Benchmarks

- Limitation of variable selection by β_{int}:
  unimportant variables in the important groups are less likely to be removed.
- Solution: find a benchmark $Z_{k,ben}$ for group with $p_k > 1$.
- The regression model becomes

$$y = \beta_{int} Z_{int} + \sum_{k=1}^{m} \left(\beta_{k,ben} Z_{k,ben} + \sum_{j=1}^{p_k} \beta_{k,j} X_{k,j} \right) + \sum_{k=m+1}^{g} \beta_{k,1} X_{k,1} + \epsilon,$$

  where m is the number of groups with size greater than 1.
- How to make $Z_{k,ben}$ orthogonal with other variables and benchmarks?
- Data augmentation!

An Example: Construction of Two More Benchmarks

- Data: 7 observations, two groups, $\{X_{1,1}, X_{1,2}\}$ and $\{X_{2,1}, X_{2,2}\}$.
- Z_{int} is orthogonal to y and $X_{k,j}$s.

Obs.	y	$X_{1,1}$	$X_{1,2}$	$X_{2,1}$	$X_{2,2}$	Z_{int}
1						$1/\sqrt{7}$
2						$1/\sqrt{7}$
3						$1/\sqrt{7}$
4			Data			$1/\sqrt{7}$
5						$1/\sqrt{7}$
6						$1/\sqrt{7}$
7						$1/\sqrt{7}$

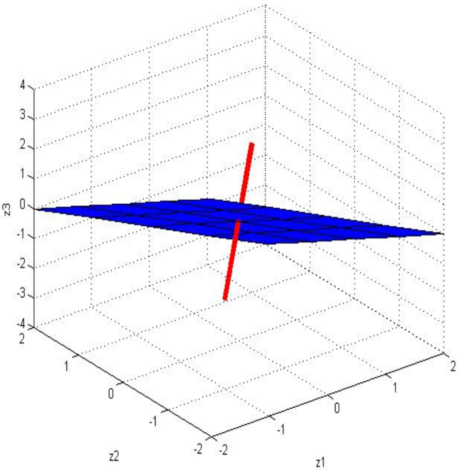
Data Augmentation

Obs.	y	$X_{1,1}$	$X_{1,2}$	$X_{2,1}$	$X_{2,2}$	$Z_{1,ben}$	$Z_{2,ben}$	Z_{int}
1						a_1	a_2	$1/\sqrt{9}$
2						a_1	a_2	$1/\sqrt{9}$
3						a_1	a_2	$1/\sqrt{9}$
4			Data			a_1	a_2	$1/\sqrt{9}$
5						a_1	a_2	$1/\sqrt{9}$
6						a_1	a_2	$1/\sqrt{9}$
7						a_1	a_2	$1/\sqrt{9}$
8	0	0	0	0	0	$-a_1 b_1$	a_2	$1/\sqrt{9}$
9	0	0	0	0	0	0	$-a_2 b_2$	$1/\sqrt{9}$

- $a_1 = 1/\sqrt{56}$; $b_1 = 7$
- $a_2 = 1/\sqrt{72}$; $b_2 = 8$
- $Z_{1,ben}$, $Z_{2,ben}$ and Z_{int} are pairwise orthogonal and also orthogonal with response and predictors.

Geometry Interpretation of Data Augmentation

- Adding one zero observation brings the original data to $n+1$ dimensional space.



Steps of Constructing Benchmarks

1. Add m zero observations to the original data, where m is the number of groups with $p_k > 1$.

2. Let $Z_{k,ben} = \{\underbrace{a_k, \ldots, a_k}_{n+k-1}, -a_k b_k, \underbrace{0, \ldots, 0}_{m-k}\}$, where $a_k = [(n+k-1)(n+k)]^{-1/2}$ and $b_k = n+k-1$.

3. Let $Z_{int} = \{\underbrace{(m+n)^{-1/2}, \ldots, (m+n)^{-1/2}}_{m+n}\}$.

Group Selection and Within Group Selection

- Regression model:

$$y = \beta_{int} Z_{int} + \sum_{k=1}^{m} \left(\beta_{k,ben} Z_{k,ben} + \sum_{j=1}^{p_k} \beta_{k,j} X_{k,j} \right) + \sum_{k=m+1}^{g} \beta_{k,1} X_{k,1} + \epsilon.$$

- Assign hierarchical priors with group structure and obtain posterior distributions of the coefficients.
- Group selection: the kth group is selected, if $\max_j \{P(|\beta_{k,j}| > |\beta_{int}|)\} > c$.
- Within group selection: suppose group k is selected, then $X_{k,j}$ is selected if $P(|\beta_{k,j}| > |\beta_{k,ben}|) > c$.

Illustration

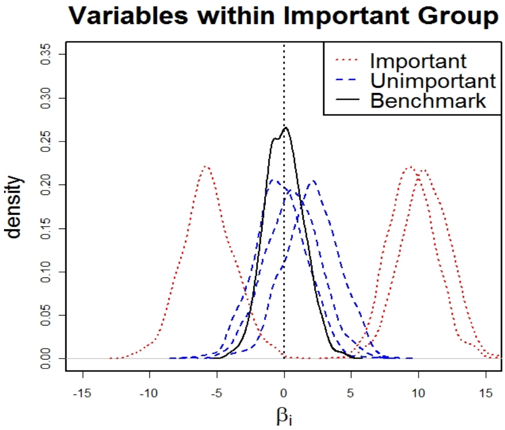
- Consider that $p = 20$, $g = 6$, and

$$\begin{aligned}\beta &= [\underbrace{1.5, -0.8, 0, 0, 0, 1.2}_{G1}, \underbrace{0, 0, 0.8, 0}_{G2}, \\ &\quad \underbrace{1.2, 0, 0, 0, 0}_{G3}, \underbrace{0, 0, 0}_{G4}, \underbrace{0}_{G5}, \underbrace{0.8}_{G6}]^T.\end{aligned}$$

- $y = X\beta + \epsilon$; $X_{k,j} \sim N(0,1)$, $cov(X_{k,i}, X_{k,j}) = 0.5^{|i-j|}$ for $k = 1, 2$; $cov(X_{k,j}, X_{k,l}) = 0$ for $k = 3, 4$ and $j \neq l$.
- Signal to noise ratio is 3.
- $n = 100$
- Let $c = 0.9$.

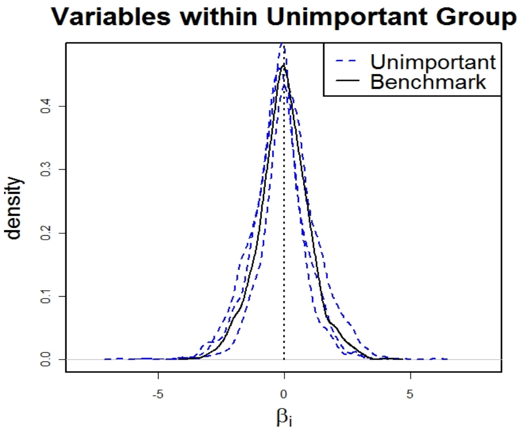
Posteriors of Variables in An Important Group

- Important variables deviate further from the benchmark.



Posteriors of Variables in An Unimportant Group

- All the unimportant variables are very close to the benchmark.



Group Selection Result

Table: The frequency each group is selected in 100 simulations.

Group	1	2	3	4	5	6
Size	6	4	5	3	1	1
Important	Y	Y	Y	N	N	Y
Selected	100	94	100	1	1	99

Within Group Selection Result

- Average false discovery rate is 0.053 (0.011); average false negative rate is 0.013 (0.003).
- Average number of selected variables is 6.13 (0.08). (True number is 6)

Table: Number of times each variable is selected in 100 simulations.

Variable	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$
True Coef.	1.5	-0.8	0	0	0	1.2
Selected	100	90	4	5	5	100

A Big p Small n Example

- Let $p = 200$ and $n = 100$. There are 40 groups and each group consisted of 5 variables.
- $\beta_{1,j}$s in group 1 : $(1.2, 0.8, 0, 0, 1.6)$
- $\beta_{2,j}$s in group 2 : $(1, -0.9, -1.1, -1.3, 0.8)$
- $\beta_{3,j}$s in group 3: $(0.8, 0, 0, 0, 0)$
- $\beta_{k,j}$s in group 4 to 8 are all zero.
- The above 8 groups form a block and is replicated 5 times to yield the coefficients of 240 variables in total.
- There are 45 important variables and 255 unimportant variables.

Covariance Structure

- The $X_{k,j}$s in the each block are generated from multivariate normal with mean 0 and covariance structure:

$$cov(X_{k,i}, X_{m,j}) = 1/3(0.5)^{|k-m|}.$$

- Variables in different blocks are uncorrelated.
- Signal to noise ratio is 10.

Group Selection Result

- When $c = 0.7$, unimportant groups are effectively removed.
- When $c = 0.7$, false discovery rate is 5.1% (0.8%) and the group false negative rate is 24.0% (0.4%).
- When $c = 0.6$, false discovery rate is 23.9% (0.9%) and the group false negative rate is 16.1% (0.5%).

Table: Frequencies that first 8 groups are selected based on 100 simulations.

Group	1	2	3	4	5	6	7	8
Important	Y	Y	Y	N	N	N	N	N
Selected($c = 0.7$)	91	39	5	2	2	4	0	1
Selected($c = 0.6$)	99	78	38	10	17	19	9	15

Within Group Selection Result

- Unimportant variables in group 1 (important group) are effectively removed when $c = 0.7$.
- When $c = 0.6$, average false discovery rate over all groups is 33% (0.8%) and average false negative rate is 12% (0.2%).
- When $c = 0.7$, average false discovery rate over all groups is 11% (0.8%) and average false negative rate is 17% (0.2%).

Table: Frequencies that 5 variables in the first group are selected based on 100 simulations.

(k, j)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
True $\beta_{k,j}$	1.2	0.8	0.0	0.0	1.6
Selected($c = 0.7$)	68	41	14	12	77
Selected($c = 0.6$)	91	73	42	40	98

Outline

- Introduction of Bayesian Lasso
- Bayesian Lasso with Pseudo Variables
- Bayesian Group Lasso with Pseudo Variables
- **Conclusions and Future Work**

Conclusions

- Intercept is a good benchmark for unimportant variables.
- Bayesian Lasso with pseudo variables achieve the sparsity.
- Bayesian Group Lasso with pseudo variables achieve both good group selection and within group selection results.

Future Work

- Optimize the threshold.
- More numerical comparisons with other variable selection methods.
- Real data analysis.

Other Work

- Shao, J. & **Tang, Q.**, Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation. (Submitted)

- **Tang, Q.** & Qian, P.Z.G., Enhancing the Sample Average Approximation method with U designs. (In revision)

Acknowledgement (Alphabetic)

- Jun Shao, University of Wisconsin-Madison
- Kam-Wah Tsui, University of Wisconsin-Madison
- Peter Qian, University of Wisconsin-Madison
- Sijian Wang, University of Wisconsin-Madison

Thank you!