Consistent high-dimensional Bayesian variable selection via penalized credible regions

Howard Bondell

bondell@stat.ncsu.edu

Joint work with Brian Reich
Outline

- High-Dimensional Variable Selection
- Bayesian Variable Selection
- Selection via Credible Sets
  - Joint / Marginal
- Asymptotic Properties
- Examples
- Conclusion
Variable Selection Setup

- Linear regression model \( y_i = x_i^T \beta + \epsilon_i \)
- \( n \) observations and \( p \) predictor variables
- \( y_i \): response for observation \( i \)
- \( x_i \): (column) vector of \( p \) predictors for observation \( i \)
- \( \beta \): (column) vector of \( p \) regression parameters
- \( \epsilon_i \): iid errors - mean zero, constant variance

- Ultra-high dimensional data, \( p >> n \)
- Only subset of predictors are relevant
- If \( \beta_j = 0 \) then variable \( j \) is effectively removed from the model
Variable Selection Methods

- All Subsets - $2^p$ !!!!
- Forward Selection
- Backward Elimination - Not possible for $p > n$
- Stepwise
- Penalization Methods can be effective
- Bayesian Methods
  - Exhaustive Search - $2^p$ !!!!
  - Stochastic Search
Penalization Methods

- Minimize:
  \[ \| y - X \beta \|^2 + \lambda J(\beta) \]

- LASSO: \( J(\beta) = \sum_{j=1}^{p} |\beta_j| \)

- Elastic Net: \( J(\beta) = (1 - c) \sum_{j=1}^{p} \beta_j^2 + c \sum_{j=1}^{p} |\beta_j| \)

- Adaptive LASSO, SCAD, MCP, OSCAR, ...

- \( \lambda \) and \( c \) chosen by AIC, BIC, Cross-Val, GCV

Shrinkage creates bias
- Reduces variance
- Achieves selection by setting exact zeros
Ultra High-Dimensional Data

- When $p >> n$, before performing penalization methods, common to screen down first
- Sure Independence Screening
  - Rank by marginal correlations
  - Reduce typically to $p < n$
- Perform forward selection sequence
  - Again reduce to $p < n$
- Then perform penalized regression
  - SCAD (Smoothly Clipped Absolute Deviation) typical
Bayesian Variable Selection

Each candidate model indexed by \( \delta = (\delta_1, \cdots, \delta_p)^T \)

\[
\delta_j = \begin{cases} 
1 & \text{if } x_j \text{ is included in the model,} \\
0 & \text{if } x_j \text{ is excluded from the model.}
\end{cases}
\]

\( p(\delta) \) is prior over model space

Most common \( p(\delta) \propto \pi^{p_\delta} (1 - \pi)^{p-p_\delta} \)

\( p_\delta = \sum_{j=1}^{p} \delta_j \) - number of predictors

\( \pi \) is prior inclusion probability for each

Uniform prior over model space \( \Leftrightarrow \pi = 1/2 \)

\( \pi \) set to apriori guess of proportion of important predictors

Put prior on \( \pi \) - Beta \((a, b)\)
Bayesian Variable Selection

- Given $\delta$, we have $\Pi(\beta|\delta, \sigma^2, \tau)$
  - Typically, $\sigma^2$ gets diffuse prior (Inverse Gamma)
  - $\tau$ are other hyperparameters needed

- Most common $\Pi(\beta|\delta, \sigma^2, \tau) = N\left(0, \frac{\sigma^2}{\tau}V\right)$

  - $V = I_{p\delta}$ or $V = (X_{\delta}^T X_{\delta})^{-1}$

  - But $p_{\delta} > n \Rightarrow X_{\delta}^T X_{\delta}$ not invertible

  - Focus on $V = I$

- $\tau$ either fixed, or given Gamma prior

- Equivalent to Spike-and-Slab, i.e. $\beta$ is mixture of mass at zero and Normal
Bayesian Variable Selection

- Crank out Bayes’ rule and get posterior probability for each configuration of $\delta$
- Instead, use stochastic search (SSVS) to visit models with MCMC chain
  - Estimate posterior probabilities by proportion of times visited
- Highest posterior model $\Leftrightarrow$ comparing Bayes Factors
- Alternative: Use marginal posterior for each variable
  - Include variable in final model if $P(\delta_j = 1|X, y) > t$ for some threshold
  - Median probability model (Barbieri and Berger, 2004) use $t = 1/2$
  - Optimal predictive model under certain conditions
Lindley’s Paradox

- Problem with Bayes Factors (posterior probabilities)
- Diffuse prior typical in practice
- Simple case
  - Sample of size 1, from $N(\mu, 1)$
  - $\mu = 0$ vs. $\mu \neq 0$ - More diffuse prior $\Rightarrow$ Prob of $H_0 \rightarrow 1$

(a) Posterior Probability in favor of Null for various prior standard deviations.
(b) 95% Posterior Credible Set for various prior standard deviations.
Other Drawbacks

- Typical methods, such as SSVS, require:
  - Proper prior distribution
  - Choice of prior on model space (inclusion probabilities)
  - Posterior threshold choice
  - MCMC chains to estimate posterior probabilities (often need very long runs)

- Results can be sensitive to each choice

- Marginal inclusion probabilities may be poor under high correlation
  - Highly correlated predictors may each show up equally often
  - But each only a small number of times
Joint Credible Regions

- Specify prior only on parameters in full model

\[ \Pi(\beta | \sigma^2, \tau) = N \left( 0, \frac{\sigma^2}{\tau} I \right) \]

\[ p(\sigma^2) = IG(0.01, 0.01) \]

- \( C_\alpha \) is \((1 - \alpha) \times 100\%\) credible region
- For fixed hyperparameter, \( \tau \), get elliptical regions

\[ C_\alpha = \{ \beta : (\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) \leq C_\alpha \}, \text{ for some } C_\alpha \]

- \( \hat{\beta}, \Sigma \) - posterior mean, variance
- Closed form if \( \tau \) fixed —— \( \hat{\beta} = (X^T X + \tau I)^{-1} X^T y \)
- Otherwise, simple short MCMC run used
- Prior on \( \tau \) ⇒ elliptical contours still valid credible sets
Joint Credible Regions

- All points within region may be feasible parameter values
- Among these, we seek a sparse solution
- Search within the region for the ‘sparsest’ point

\[ \tilde{\beta} = \arg \min_{\beta} ||\beta||_0 \]
subject to
\[ \beta \in C_\alpha \]

- Chosen model for given \( \alpha \) defined by set of indices,
\[ A_\alpha = \{ j : \tilde{\beta}_j \neq 0 \}. \]
Joint Credible Regions

- Problems with searching for sparsest solution
  - High dimensional region - combinatorial search
  - Also Non-unique

- Replace $L_0$ by smooth bridge between $L_0$ and $L_1$ (Lv and Fan, 2009)

\[
\sum_{j=1}^{p} \rho_a(|\beta_j|),
\]

\[
\rho_a(t) = \frac{(a+1)t}{a+t} = \left(\frac{t}{a+t}\right) I(t \neq 0) + \left(\frac{a}{a+t}\right) t, \quad t \in [0, \infty),
\]

\[
\rho_0(t) = \lim_{a \to 0^+} \rho_a(t) = I(t \neq 0)
\]

\[
\rho_\infty(t) = \lim_{a \to \infty} \rho_a(t) = t
\]

- Interest on $\rho_a(t)$ for $a \approx 0$. 
Computation

- Non-convex penalty function
- Local linear approximation to penalty
  \[
  \rho_a(|\beta_j|) \approx \rho_a(|\hat{\beta}_j|) + \rho'_a(|\hat{\beta}_j|) \left( |\beta_j| - |\hat{\beta}_j| \right),
  \]
  with \[\rho'_a(|\hat{\beta}_j|) = \frac{a(a+1)}{(a+|\hat{\beta}_j|)^2}\]
- \[\hat{\beta}\] is posterior mean
- Using Lagrangian gives
  \[
  \tilde{\beta} = \arg\min \left\{ (\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) + \lambda \alpha \sum_{j=1}^p \frac{|\beta_j|}{(a+|\hat{\beta}_j|)^2} \right\}
  \]
- Constant absorbed into \[\lambda \alpha\]
- One-to-one correspondence between \[\lambda \alpha\] and \[\alpha\]
Computation

- Optimization becomes

\[ \tilde{\beta} = \arg\min \left\{ (\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) + \lambda \alpha \sum_{j=1}^{p} \frac{|\beta_j|}{(a + |\hat{\beta}_j|)^2} \right\} \]

- For \( a \to 0 \),

\[ \tilde{\beta} \approx \arg\min \left\{ (\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) + \lambda \alpha \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j|^2} \right\} \]

- Adaptive Lasso form
  - LARS algorithm gives full path as vary \( \alpha \)
Selection Consistency

- Sequence of credible sets \((\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) \leq C_n\)
- Sequence of models \(A_{\alpha_n}\)
- One-to-one correspondence between \(\alpha_n\) and \(C_n\)
- True model \(A\)

**Theorem 1.** Under general conditions, if \(C_n \rightarrow \infty\) and \(n^{-1} C_n \rightarrow 0\), then the credible set method is consistent in variable selection, i.e.

\[ P(A_{\alpha_n} = A) \rightarrow 1 \]

- Also holds for \(p \rightarrow \infty\), but \(p/n \rightarrow 0\)
Selection Consistency

- What about $p >> n$?
- Asymptotics with $p/n \to 0$ not entirely relevant
- Posterior mean - Ridge Regression form

$$\hat{\beta} = \left( X^T X + \tau I \right)^{-1} X^T y$$

- If $p/n \not\to 0$, can show that $\hat{\beta}$ not mean square consistent

$$E \left\{ (\hat{\beta} - \beta^0)^T (\hat{\beta} - \beta^0) \right\} \not\to 0$$
Selection Consistency

- Consider rectangular credible regions - not elliptical
- Just use diagonal elements of $\Sigma$ ignoring covariances
- Construct credible sets separately for each parameter
- Simple componentwise thresholding on posterior mean (t-statistics)

**Theorem 2.** Let $\tau \to \infty$ and $\tau = O \left( \left( n^2 \log p \right)^{1/3} \right)$ then the posterior thresholding approach is consistent in selection when the dimension $p$ satisfies $\log p = O \left( n^c \right)$ for some $0 \leq c < 1$.

- Selection consistency for exponential growing dimension, $\log p = o(n)$
- Also applies to ridge regression with ridge parameter $\tau$
Simulation Study

- Linear Regression Model with $N(0, 1)$ errors
- $n = 60$ observations (same as real data example)
- $p \in \{50, 500, 2000\}$ also $N(0, 1)$ with $AR(1)$, $\rho \in \{0.5, 0.9\}$
- Results based on 200 datasets for each of the 6 setups
Simulation Study

Consider ordering of predictors induced by:
- Joint credible regions
- Marginal posterior thresholding
- Stochastic Search (with various choices of prior)
- LASSO

To measure reliability of ordering:
- ROC curve - measures sensitivity vs. specificity - related to type I error
- PRC (Precision-Recall) curve - related to False Discovery rate
Simulation Study

$p = 50$, $n = 60$ \( \rho = 0.5 \) (Top) and \( \rho = 0.9 \) (Bottom)
Simulation Study

- $p = 500$, $n = 60$
- Area under ROC and PRC curves

<table>
<thead>
<tr>
<th></th>
<th>ROC Area $\rho = 0.5$</th>
<th>ROC Area $\rho = 0.8$</th>
<th>PRC Area $\rho = 0.5$</th>
<th>PRC Area $\rho = 0.8$</th>
<th>CPU Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Credible Sets</td>
<td>0.946 (0.004)</td>
<td>0.989 (0.001)</td>
<td>0.708 (0.011)</td>
<td>0.873 (0.007)</td>
<td>20.93</td>
</tr>
<tr>
<td>Marginal Credible Sets</td>
<td>0.932 (0.004)</td>
<td>0.979 (0.002)</td>
<td>0.687 (0.011)</td>
<td>0.862 (0.007)</td>
<td>20.93</td>
</tr>
<tr>
<td>SSVS (fixed, fixed)</td>
<td>0.902 (0.004)</td>
<td>0.924 (0.004)</td>
<td>0.620 (0.011)</td>
<td>0.634 (0.010)</td>
<td>1222.91</td>
</tr>
<tr>
<td>SSVS (random, fixed)</td>
<td>0.929 (0.004)</td>
<td>0.957 (0.003)</td>
<td>0.672 (0.010)</td>
<td>0.693 (0.009)</td>
<td>1222.91</td>
</tr>
<tr>
<td>SSVS (fixed, random)</td>
<td>0.897 (0.005)</td>
<td>0.924 (0.004)</td>
<td>0.615 (0.011)</td>
<td>0.656 (0.010)</td>
<td>1222.91</td>
</tr>
<tr>
<td>SSVS (random, random)</td>
<td>0.925 (0.005)</td>
<td>0.955 (0.003)</td>
<td>0.665 (0.010)</td>
<td>0.692 (0.009)</td>
<td>1222.91</td>
</tr>
</tbody>
</table>
Simulation Study

\( p = 500, \ n = 60 \quad \rho = 0.5 \ \text{(Top)} \ \text{and} \ \rho = 0.9 \ \text{(Bottom)} \)
Simulation Study

\( p = 2000, \ n = 60 \quad \rho = 0.5 \) (Top) and \( \rho = 0.9 \) (Bottom)
Table 1: Selection performance for $p = 10,000$ with 3 important predictors for various choices of $n$ based on 100 datasets. The entries in the table denote Correct Selection Proportion (CS), Coverage Proportion (COV), Average Model Size (MS), and Average Number of Important Predictors out of the 3 Included (IP).

<table>
<thead>
<tr>
<th></th>
<th>$n = 100$</th>
<th>$n = 200$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CS</td>
<td>COV</td>
<td>MS</td>
</tr>
<tr>
<td>Marginal Sets</td>
<td>9.0</td>
<td>31.0</td>
<td>3.22</td>
</tr>
<tr>
<td>SIS + SCAD</td>
<td>1.0</td>
<td>15.0</td>
<td>4.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$n = 1000$</th>
<th>$n = 2000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CS</td>
<td>COV</td>
</tr>
<tr>
<td>Marginal Sets</td>
<td>45.0</td>
<td>61.0</td>
</tr>
<tr>
<td>SIS + SCAD</td>
<td>12.0</td>
<td>64.0</td>
</tr>
</tbody>
</table>
Real Data Analysis

- Mouse Gene Expression (Lan et al., 2006)
- 60 arrays (31 female, 29 male mice)
- 22,575 genes + gender \( (p = 22,576) \)
- Fit with \( n = 55 \), leave out 5 for testing

Table 1: Mean squared prediction error and model size based on 100 random splits of the real data, with standard errors in parenthesis. The 3 response variables are PEPCK, GPAT, and SCD1.

<table>
<thead>
<tr>
<th></th>
<th>PEPCK</th>
<th></th>
<th>GPAT</th>
<th></th>
<th>SCD1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSPE Model Size</td>
<td>MSPE Model Size</td>
<td>MSPE Model Size</td>
<td>MSPE Model Size</td>
<td>MSPE Model Size</td>
<td>MSPE Model Size</td>
</tr>
<tr>
<td>Marginal Sets ( p = 22,576 )</td>
<td>2.14 (0.15)  7.1 (0.41)</td>
<td>4.70 (0.45)  9.3 (0.59)</td>
<td>3.54 (0.26)  7.6 (0.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIS + SCAD ( p = 22,576 )</td>
<td>2.82 (0.18)  2.3 (0.09)</td>
<td>5.88 (0.44)  2.6 (0.10)</td>
<td>3.44 (0.22)  3.2 (0.14)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joint Sets ( p = 2,000 )</td>
<td>2.03 (0.14)  9.6 (0.46)</td>
<td>3.83 (0.34)  4.2 (0.43)</td>
<td>3.04 (0.22)  22.0 (0.56)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal Sets ( p = 2,000 )</td>
<td>1.84 (0.14)  23.3 (0.67)</td>
<td>5.33 (0.41)  21.8 (0.72)</td>
<td>3.27 (0.21)  19.1 (0.71)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LASSO ( p = 2,000 )</td>
<td>3.03 (0.19)  7.7 (0.96)</td>
<td>5.03 (0.42)  3.3 (0.79)</td>
<td>3.25 (0.31)  19.7 (0.77)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- Variable selection via Bayesian Credible sets
  - Sparse solution within set
  - Elliptical regions consistent if $p/n \rightarrow 0$
  - Rectangular regions consistent if $\log p = o(n)$
- Computationally feasible even in high dimensions
- Excellent finite sample performance
- Extensions to other models