Consistent high-dimensional Bayesian variable selection via penalized credible regions

Howard Bondell

bondell@stat.ncsu.edu

NC STATE UNIVERSITY

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 $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$
  - n observations and p predictor variables
  - y_i: response for observation i
  - $\mathbf{x}_i$: (column) vector of p predictors for observation i
  - $\boldsymbol{\beta}$: (column) vector of p regression parameters
  - ϵ_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

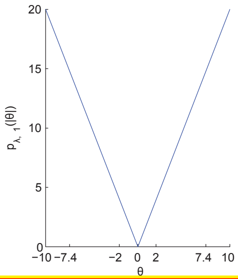
- Minimize:

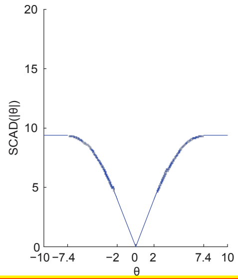
$$\|\mathrm{y} - X\beta\|^2 + \lambda J(\beta)$$

- 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, ...
- λ 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 $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)^T$

$$\delta_j = \begin{cases} 1 & \text{if } \mathbf{x}_j \text{ is included in the model,} \\ 0 & \text{if } \mathbf{x}_j \text{ is excluded from the model.} \end{cases}$$

- $p(\boldsymbol{\delta})$ is prior over model space
- Most common $p(\boldsymbol{\delta}) \propto \pi^{p_\delta}(1-\pi)^{p-p_\delta}$
  - $p_{\boldsymbol{\delta}} = \sum_{j=1}^{p} \delta_j$ - number of predictors
  - π is prior inclusion probability for each
  - Uniform prior over model space $\Leftrightarrow \pi = 1/2$
  - π set to apriori guess of proportion of important predictors
  - Put prior on π - Beta (a, b)

NC STATE UNIVERSITY

Bayesian Variable Selection

- Given $\boldsymbol{\delta}$, we have $\Pi(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma^2, \tau)$
  - Typically, σ^2 gets diffuse prior (Inverse Gamma)
  - τ are other hyperparameters needed
- Most common $\Pi(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma^2, \tau) = N\left(0, \frac{\sigma^2}{\tau}V\right)$
  - $V = I_{p_{\boldsymbol{\delta}}}$ or $V = (X_{\boldsymbol{\delta}}^T X_{\boldsymbol{\delta}})^{-1}$
  - But $p_{\boldsymbol{\delta}} > n \Rightarrow X_{\boldsymbol{\delta}}^T X_{\boldsymbol{\delta}}$ not invertible
  - Focus on $V = I$
- τ either fixed, or given Gamma prior
- Equivalent to Spike-and-Slab, i.e. $\boldsymbol{\beta}$ is mixture of mass at zero and Normal

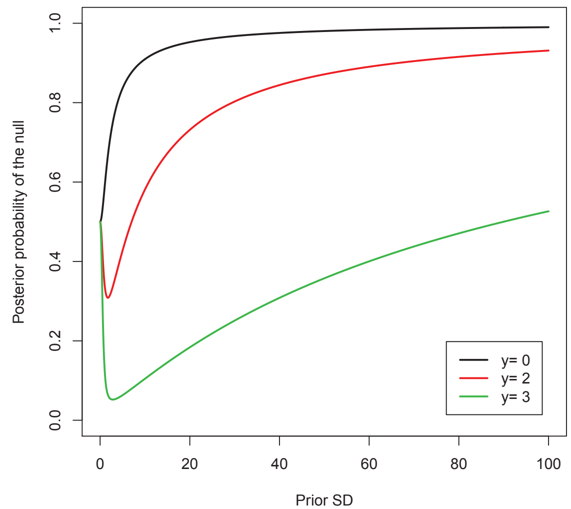
NC STATE UNIVERSITY

Bayesian Variable Selection

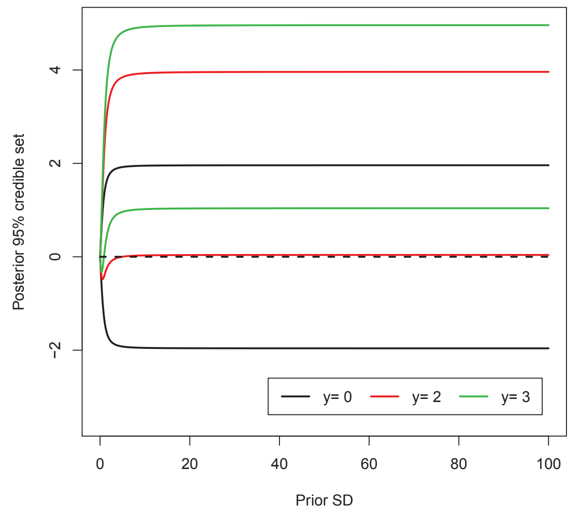
- Crank out Bayes' rule and get posterior probability for each configuration of δ
- Instead, use stochastic search (SSVS) to visit models with MCMC chain
  - Estimate posterior probabilities by proportion of times visited
- Highest posterior model $\Leftrightarrow$ comparing Bayes Factors
- Alternative: 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

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

$$p(\sigma^2) = IG(0.01, 0.01)$$

- $\mathcal{C}_\alpha$ is $(1-\alpha)\times 100\%$ credible region
- For fixed hyperparameter, τ, get elliptical regions

$$\mathcal{C}_\alpha = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \Sigma^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \le C_\alpha\}, \text{ for some } C_\alpha$$

- $\hat{\boldsymbol{\beta}}$, Σ - posterior mean, variance
  - Closed form if τ fixed —- $\hat{\boldsymbol{\beta}} = \left(X^T X + \tau I\right)^{-1} X^T y$
  - Otherwise, simple short MCMC run used
- Prior on $\tau \Rightarrow$ 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

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{\beta}||_0$$

subject to

$$\boldsymbol{\beta} \in \mathcal{C}_\alpha$$

- Chosen model for given α defined by set of indices, $\mathcal{A}_n^\alpha = \{j \ : \ \tilde{\beta}_j \neq 0\}$.

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)

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

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

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

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

- 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 $\rho_a'(|\hat{\beta}_j|) = \frac{a(a+1)}{\left(a+|\hat{\beta}_j|\right)^2}$

- $\hat{\boldsymbol{\beta}}$ is posterior mean
- Using Lagrangian gives

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

- Constant absorbed into λ_α
- One-to-one correspondence between λ_α and α

Computation

- Optimization becomes

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

- For $a \to 0$,

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

- Adaptive Lasso form
  - LARS algorithm gives full path as vary α

Selection Consistency

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

THEOREM 1. *Under general conditions, if $C_n \to \infty$ and $n^{-1} C_n \to 0$, then the credible set method is consistent in variable selection, i.e.* $P(\mathcal{A}_n^{\alpha_n} = \mathcal{A}) \to 1$

- Also holds for $p \to \infty$, but $p/n \to 0$

Selection Consistency

- What about $p >> n$?
- Asymptotics with $p/n \to 0$ not entirely relevant
- Posterior mean - Ridge Regression form
- $\hat{\boldsymbol{\beta}} = \left(X^T X + \tau I\right)^{-1} X^T y$
- If $p/n \nrightarrow 0$, can show that $\hat{\boldsymbol{\beta}}$ not mean square consistent

$$E\left\{\left(\hat{\beta} - \beta^0\right)^T \left(\hat{\beta} - \beta^0\right)\right\} \nrightarrow 0$$

Selection Consistency

- Consider rectangular credible regions - not elliptical
- Just use diagonal elements of Σ 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 τ

NC STATE UNIVERSITY

Simulation Study

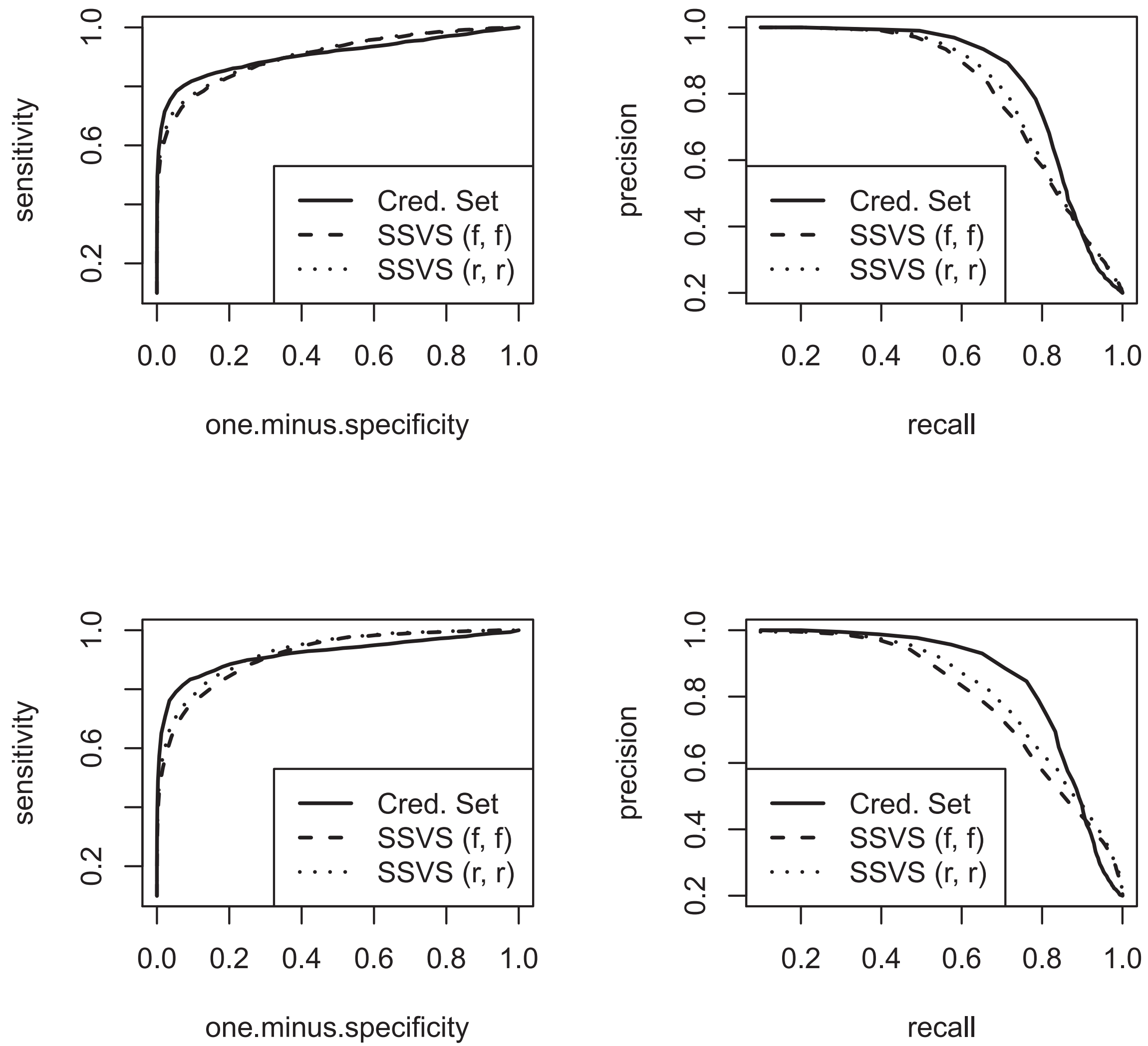
- Linear Regression Model with $N(0, 1)$ errors
- $n = 60$ observations (same as real data example)
- $p \in \{50, 500, 2000\}$ also $N(0, 1)$ with $AR(1), \rho \in \{0.5, 0.9\}$
- Results based on 200 datasets for each of the 6 setups

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)



NC STATE UNIVERSITY

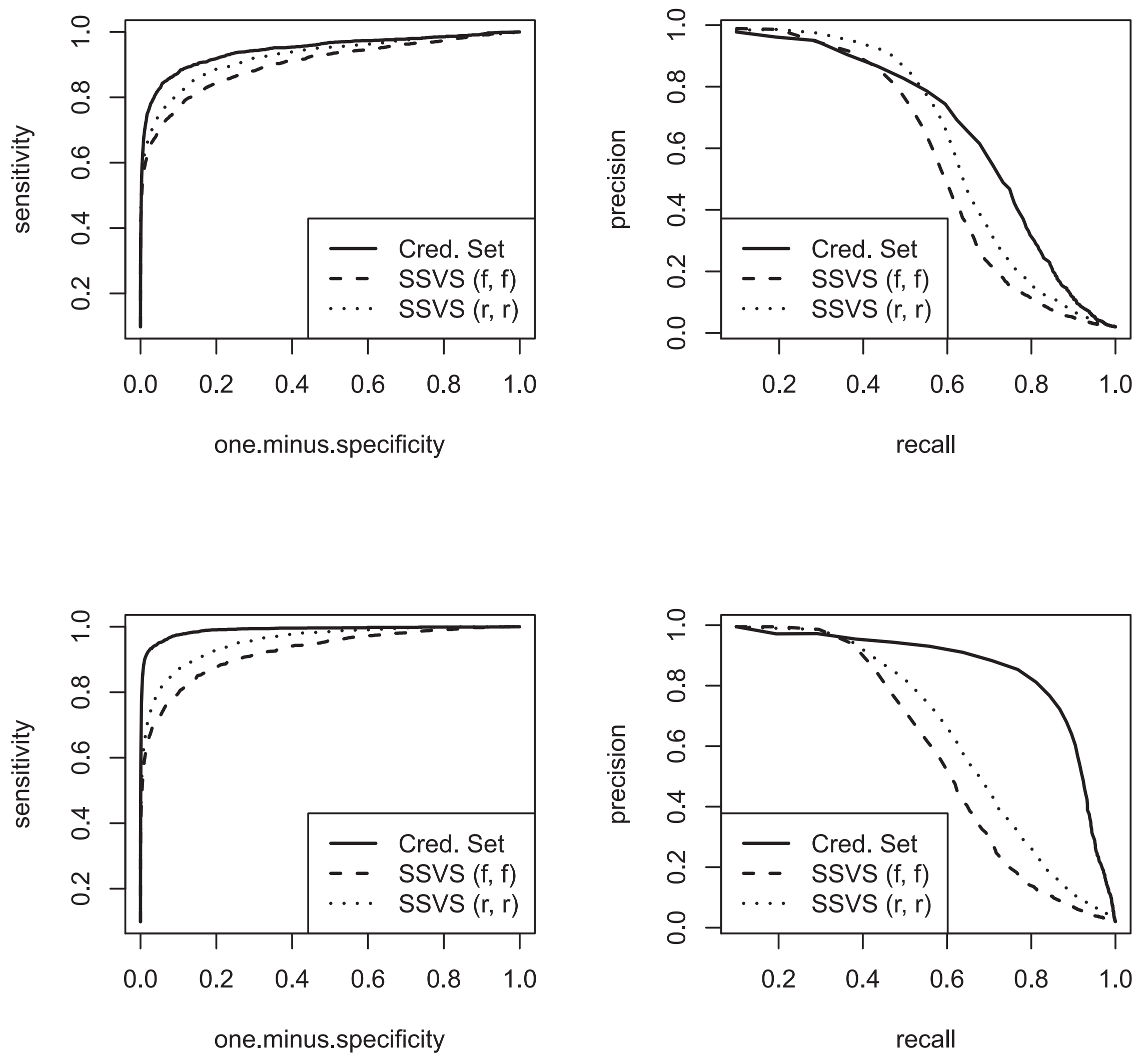
Simulation Study

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

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

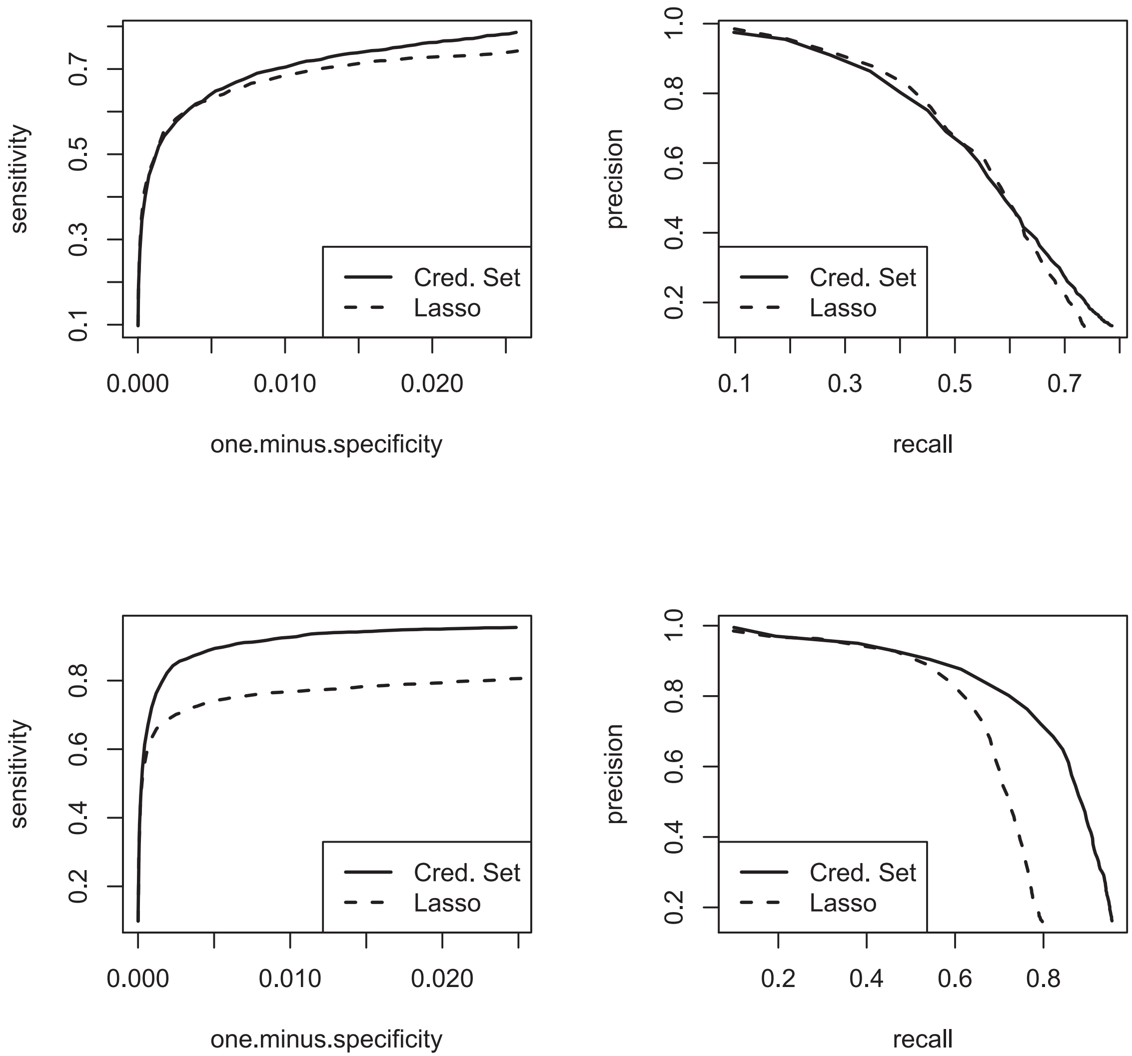
Simulation Study

- $p = 500$, $n = 60$ $\rho = 0.5$ (Top) and $\rho = 0.9$ (Bottom)



Simulation Study

- $p = 2000$, $n = 60$ $\rho = 0.5$ (Top) and $\rho = 0.9$ (Bottom)



Ultra High-Dimension

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).

	$n = 100$				$n = 200$				$n = 500$			
	CS	COV	MS	IP	CS	COV	MS	IP	CS	COV	MS	IP
Marginal Sets	9.0	31.0	3.22	2.06	24.0	47.0	3.37	2.38	39.0	54.0	3.01	2.49
SIS + SCAD	1.0	15.0	4.08	1.82	5.0	35.0	6.06	2.28	6.0	59.0	11.62	2.56

	$n = 1000$				$n = 2000$			
	CS	COV	MS	IP	CS	COV	MS	IP
Marginal Sets	45.0	61.0	2.98	2.58	62.0	74.0	2.89	2.71
SIS + SCAD	12.0	64.0	14.62	2.62	23.0	79.0	17.96	2.78

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.

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

Conclusion

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