Fast and General Best Subset Selection using Data Depth and Resampling

Subhabrata Majumdar and Snigdhansu Chatterjee

University of Florida and University of Minnesota Twin Cities

December 21, 2017

Summary

Consider the linear model:

$$Y = X\beta + \epsilon; \quad Y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p \tag{1}$$

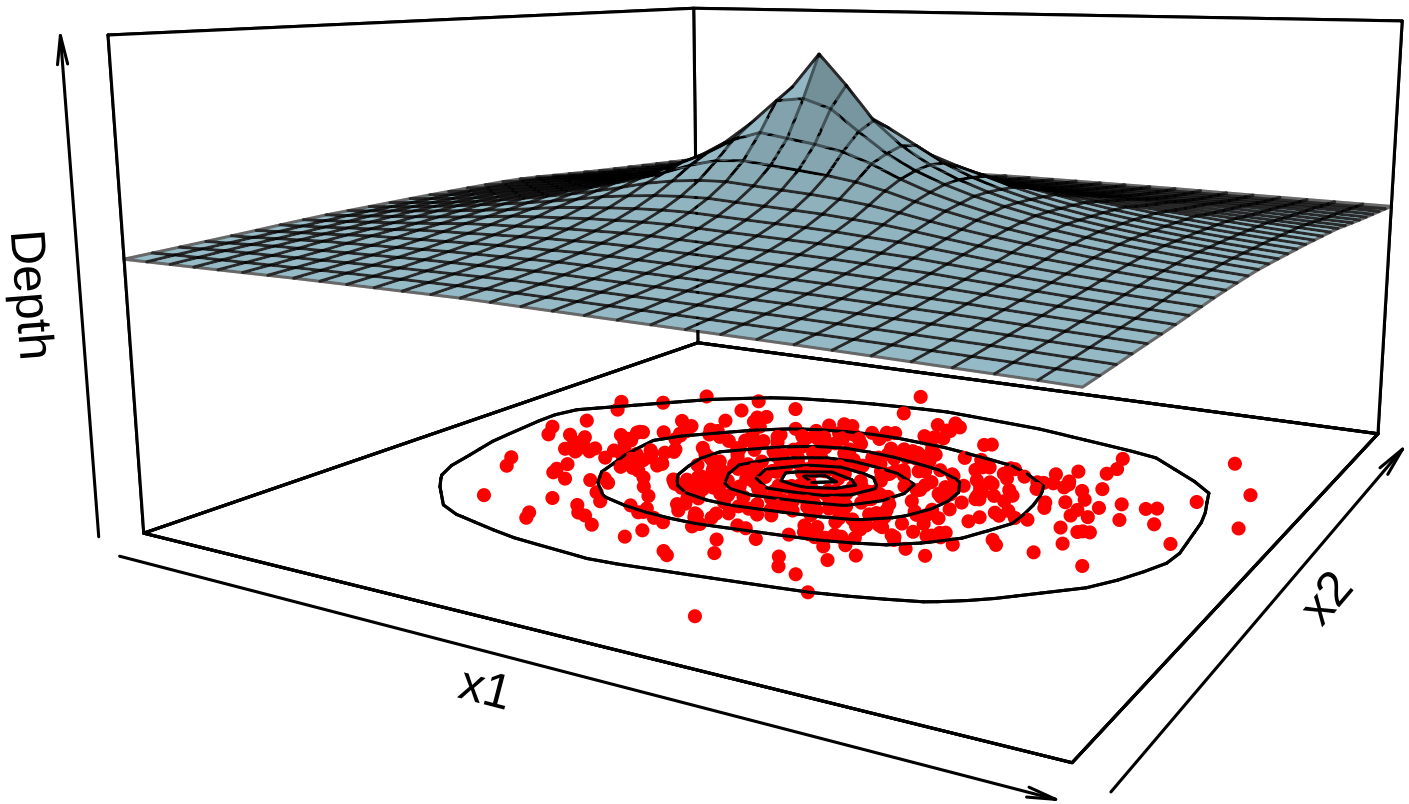
- Variable selection in (1) is a fundamental problem in statistics. There are two ways to do this- sparse penalized regression and best subset selection.
- Sparse methods have inferential and algorithmic issues. Best subset selection is computationally demanding.
- Best subset selection in other models with dependent or structured ϵ, e.g. mixed effect models is poorly explored.
- *Our method:* We propose a fast and general algorithm for best subset selection in a wide range of statistical models. Our method trains only a single model, and evaluates a model selection criterion at $p + 1$ models (p = no. of predictors) to come up with a selected set of variables.

Table of contents

Outline

Data Depth

Example: 500 points from $\mathcal{N}_2((0,0)^T, \text{diag}(2,1))$.



Data depth is a **scalar measure of how much inside a point is with respect to a data cloud**. Denote it by $D(\mathbf{x}, F)$.

Our model selection criterion

- Parameters are estimated based on a finite random sample from sample space. So the estimates have their own probability distributions- these are called *sampling distributions*;
  e.g. in linear regression, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1}) \equiv [\hat{\beta}]$.

- We compare sampling distributions of parameter estimates in a candidate model with that of a baseline model using a generic quantity called the ***e*-value**.

- Given some depth function $D(.,.)$ we define the e-value as

$$e(\mathcal{M}) = \mathbb{E}D(\hat{\beta}_{\mathcal{M}}, [\hat{\beta}])$$

  i.e. the expected depth of the model estimates with respect to the sampling distribution of $\hat{\beta}$.

The fast variable selection 'algorithm'

1. Obtain e-value for the full model:

$$e(\mathcal{M}_{full}) = \mathbb{E}D(\hat{\beta}, [\hat{\beta}])$$

2. Set $\mathcal{S}_{select} = \phi$.

3. For $j = 1, 2, \ldots, p$
   Replace j^{th} index of $\hat{\beta}$ by 0, name it $\hat{\beta}_{-j}$.
   Obtain $e(\mathcal{M}_{-j}) = \mathbb{E}D(\hat{\beta}_{-j}, [\hat{\beta}])$.
   If $e(\mathcal{M}_{-j}) < e(\mathcal{M}_{full})$
      Set $\mathcal{S}_{select} \leftarrow \{\mathcal{S}_{select}, j\}$.

For large enough n, $\mathcal{S}_{select}$ is the set of non-zero indices in the true parameter vector.

	DroppedVar	Cn
1	**- x2**	**0.2356008**
2	**- x3**	**0.2428004**
3	**- x4**	**0.2448785**
4	**- x1**	**0.2473548**
5	**- x5**	**0.2486610**
6	**- x20**	**0.2503475**
7	<none>	0.2505000
8	- x9	0.2522873
9	- x21	0.2538186
10	- x22	0.2547132
11	- x14	0.2548410
12	- x17	0.2554293
13	- x13	0.2559990
14	- x10	0.2564211
15	- x24	0.2566334
16	- x19	0.2568725
17	- x25	0.2573902
18	- x8	0.2578656
19	- x16	0.2588032
20	- x12	0.2590218
21	- x6	0.2595048
22	- x23	0.2598039
23	- x15	0.2605307
24	- x11	0.2606763
25	- x18	0.2610460
26	- x7	0.2613168

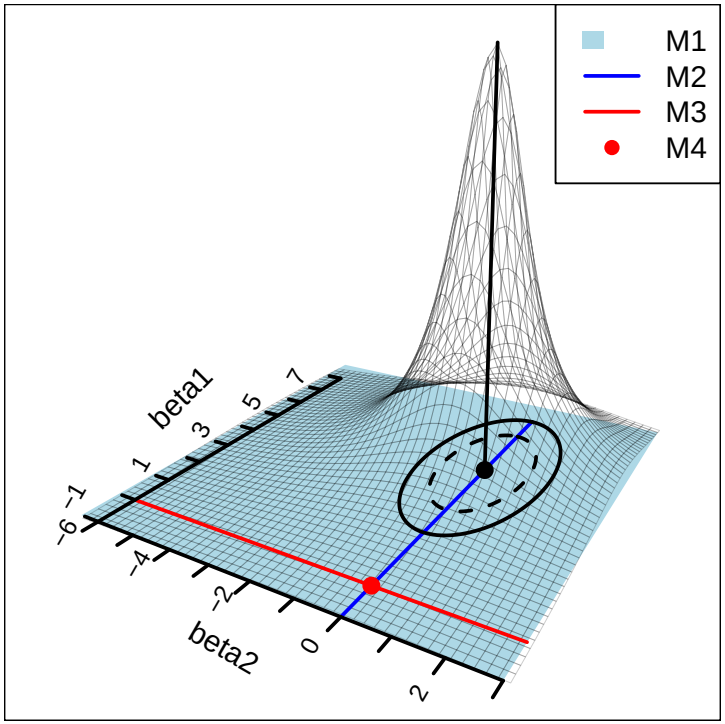
The idea- an example

Consider a linear model with $p = 2$, and true coefficient vector $\beta_0 = (5, 0)^T$. Here we have the following choice of models:

$$\begin{aligned}
\mathcal{M}_1 : &\quad Y = X_1\beta_1 + X_2\beta_2 + \epsilon; &\quad \Theta_1 &= \mathbb{R}^2 \\
\mathcal{M}_2 : &\quad Y = X_1\beta_1 + \epsilon; &\quad \Theta_2 &= \mathbb{R} \times \{0\} \\
\mathcal{M}_3 : &\quad Y = X_2\beta_2 + \epsilon; &\quad \Theta_3 &= \{0\} \times \mathbb{R} \\
\mathcal{M}_4 : &\quad Y = \epsilon; &\quad \Theta_4 &= (0, 0)^T
\end{aligned}$$

Some are good models, some are bad models.

The idea- an example



As n grows, the full model sampling distribution concentrates around β_0, so mean depths at $\mathcal{M}_3$ and $\mathcal{M}_4$ vanish.
However, both depth and density contours scale down by the same multiple as n goes down, so that mean depths at $\mathcal{M}_1$ and $\mathcal{M}_2$ remain constant.

Outline

Definition of a model

At stage n there is a triangular array of functions

$$\{\Psi_{ni}(\theta_n, Z_{ni}) : 1 \leq i \leq k_n, n \geq 1\}$$

where $\mathcal{Z}_n = \{Z_{n1}, \ldots, Z_{nk_n}\}$ is an observable array of random variables, and $\theta_n \in \Theta_n \subseteq \mathbb{R}^p$.
The true unknown vector of parameters θ_{0n} is the unique minimizer of

$$\Psi_n(\theta_n) = \mathbb{E} \sum_{i=1}^{k_n} \Psi_{ni}(\theta_n, Z_{ni})$$

The common non-zero support of all estimable parameters is $\mathcal{S}_{*n} = \cup_{\theta_n \in \Theta_n} \text{support}(\theta_n)$.

In this general setup, we associate a candidate model $\mathcal{M}_n$ with two quantities:

(a) The set of indices $\mathcal{S}_n \subseteq \mathcal{S}_{*n}$ where the parameter values are unknown and **estimated from the data**, and

(b) an ordered vector of **known constants** $C_n = (C_{nj} : j \notin \mathcal{S}_n)$ for parameters not indexed by $\mathcal{S}_n$.

The generic parameter vector corresponding to this model, denoted by $\theta_{mn} \in \Theta_{mn} \subseteq \Theta_n := \prod_j \Theta_{nj}$, will thus have the structure

$$\theta_{mnj} = \begin{cases} \text{unknown } \theta_{mnj} \in \Theta_{nj} & \text{for } j \in \mathcal{S}_n, \\ \text{known } C_{nj} \in \Theta_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases}$$

The estimator $\hat{\theta}_{*n}$ of θ_{0n} is obtained as

$$\hat{\theta}_{*n} = \underset{\theta_n}{\operatorname{argmin}}\, \hat{\Psi}_n(\theta_n) = \underset{\theta_n}{\operatorname{argmin}} \sum_{i=1}^{k_n} \Psi_{ni}\left(\theta_n, Y_{ni}\right)$$

We assume an elliptical asymptotic distribution for $\hat{\theta}_{*n}$ with the conditions-

(A1) There exists a sequence of positive reals $a_n \uparrow \infty$ such that $a_n(\hat{\theta}_{*n} - \theta_{0n}) \rightsquigarrow \mathcal{E}(0_p, V, g)$, for some p.d. matrix $V \in \mathbb{R}^{p\times p}$ and density generator function g;

(A2) For almost every data sequence $\mathcal{Z}_n$, There exists a sequence of positive definite matrices $V_n \in \mathbb{R}^{p\times p}$ such that $\text{plim}_{n\to\infty} V_n = V$.

For any other model $\mathcal{M}_n$, we use the plugin estimate:

$$\hat{\theta}_{mnj} = \begin{cases} \hat{\theta}_{*nj} & \text{for } j \in \mathcal{S}_n, \\ C_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases}$$

‘Good’ and ‘bad’ models

- A model $\mathcal{M}_n$ is called *adequate* if

$$\lim_{n\to\infty} \sum_{j\notin \mathcal{S}_n} |C_{nj} - \theta_{0nj}| = 0$$

  A model that is not adequate, will be called an *inadequate* model.
- The model $\mathcal{M}_n$ is called *strictly adequate* if, for all n and $j \notin \mathcal{S}_n$, $C_{nj} = \theta_{0nj}$.

Covers obvious cases:

$$
\begin{aligned}
(1, 2, 3, 0, -1) &\quad - \quad \text{preferred parameter vector} \\
(*, 0, *, *, *) &\quad - \quad \text{inadequate model} \\
(*, *, *, *, *) &\quad - \quad \text{full model} \\
(*, *, *, 0, *) &\quad - \quad \text{strictly adequate model}
\end{aligned}
$$

Covers limiting cases:, e.g. $(*, *, *, \delta_n), \delta_n = o(1)$ will be an adequate model in our framework.

Such data generating models, e.g.

$$Y_{ni} = X_{1i}\beta_{01} + X_{2i}\delta_n + \epsilon; \quad \beta_{01} \in \mathbb{R}, \delta_n = o(1)$$

for linear regression, frequently arise from prior choices in bayesian variable selection techniques.

We now use a *data depth function*:

$$D : \mathbb{R}^p \times \tilde{\mathbb{R}}^p \mapsto [0, \infty)$$

to quantify the relative position of θ_{mn} with respect to the preferred model estimate distribution. Denote this by

$$D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}])$$

***e*-value** is simply a functional of the distribution of this random evaluation function. Denote this by $e_n(\mathcal{M}_n)$.

We shall now elaborate on the following choice of the *e*-value:

$$e_n(\mathcal{M}_n) = \mathbb{E} D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}])$$

Resampling approximation of *e*-values

- We use resampling to approximate the distributions of the random quantities $\hat{\theta}_{mn}$ and $\hat{\theta}_{*n}$. The resampling estimate of a model *e*-value is defined using two bootstrap samples:

$$e_{rn}(\mathcal{M}_n) = \mathbb{E}_{r_1} D\big(\hat{\theta}_{r_1 mn}, [\hat{\theta}_{r*n}]\big)$$

- For $\hat{\theta}_{*n}$, we use *Generalized bootstrap* (Chatterjee and Bose, 2005) that is based on the following approximation:

$$\hat{\theta}_{r*n} = \hat{\boldsymbol{\theta}}_{*n} - \frac{\tau_n}{a_n}\left[\sum_{i=1}^{n} W_i \Psi''_{ni}(\hat{\boldsymbol{\theta}}_{*n}, Z_{ni})\right]^{-1} \sum_{i=1}^{n} W_i \Psi'_{ni}(\hat{\boldsymbol{\theta}}_{*n}, Z_{ni}) + R_{rn};$$

$$\mathbb{E}_r \|R_{rn}\|^2 = o_P(1), \tau_n \to \infty, \tau_n = o(a_n)$$

- For $\hat{\theta}_{mn}$ we use the plugin estimate

$$\hat{\theta}_{r_1 mnj} = \begin{cases} \hat{\theta}_{r_1 * nj} & \text{for } j \in \mathcal{S}_n; \\ C_{nj} & \text{for } j \notin \mathcal{S}_n \end{cases}$$

Fast algorithm for variable selection

1. (Dropping n in all subscripts except e) Fix resampling standard deviation τ.
2. Obtain bootstrap samples: $\mathcal{T} = \{\hat{\theta}_{1*}, ..., \hat{\theta}_{R*}\}$, and $\mathcal{T}_1 = \{\hat{\theta}_{1*}, ..., \hat{\theta}_{R_1*}\}$.
3. Calculate $\hat{e}_{rn}(\mathcal{M}_*) = \frac{1}{R_1} \sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1*}, [\mathcal{T}_1])$.
4. Set $\hat{\mathcal{S}}_0 = \phi$.
5. For j in $1 : p$
   For r_1 in $1 : R_1$
      Replace j^{th} index of $\hat{\theta}_{*r_1}$ by 0 to get $\hat{\theta}_{r_1,-j}$.
   Calculate $\hat{e}_{rn}(\mathcal{M}_{-j}) = \frac{1}{R_1} \sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1,-j}, [\mathcal{T}_1])$.
   If $\hat{e}_{rn}(\mathcal{M}_{-j}) < \hat{e}_{rn}(\mathcal{M}_*)$
      Set $\hat{\mathcal{S}}_0 \leftarrow \{\hat{\mathcal{S}}_0, j\}$.

Corollary

Consider two sets of bootstrap estimates of $\hat{\theta}_*$*:* $\mathcal{T} = \{\hat{\theta}_{r*} : r = 1, \ldots, R\}$ *and* $\mathcal{T}_1 = \{\hat{\theta}_{r_1*} : r_1 = 1, \ldots, R_1\}$*. Obtain sample e-value estimates as*

$$\hat{e}_{rn}(\mathcal{M}) = \frac{1}{R_1} \sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1 m}, [\mathcal{T}])$$

where $[\mathcal{T}]$ *is the empirical distribution of the corresponding bootstrap samples. Consider the set of predictors* $\hat{\mathcal{S}}_0 = \{\hat{e}_{rn}(\mathcal{M}_{-j}) < \hat{e}_{rn}(\mathcal{M}_*)\}$*. Then as* $n, R, R_1 \to \infty$*,*

$$P_2(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \to 1$$

where P_2 *is the probability conditional on the data and bootstrap samples, and* $\mathcal{S}_0$ *is the true index set of non-zero predictors.*

Outline

Simulation: linear models

Methods compared:

(a) Mixed Integer Optimization (MIO) (Bertsimas et al., 2016)
(b) Lasso penalized regression
(c) SCAD penalized regression
(d) BIC forward selection

Comparison metrics:

$$\text{Sparsity}(\hat{\beta}) = \|\hat{\beta}\|_0; \quad \text{PE}(\hat{\beta}) = \frac{\|X_{\text{test}}\hat{\beta} - X_{\text{test}}\beta_0\|}{\|X_{\text{test}}\beta_0\|}$$

Results: $n = 1000, p = 60$

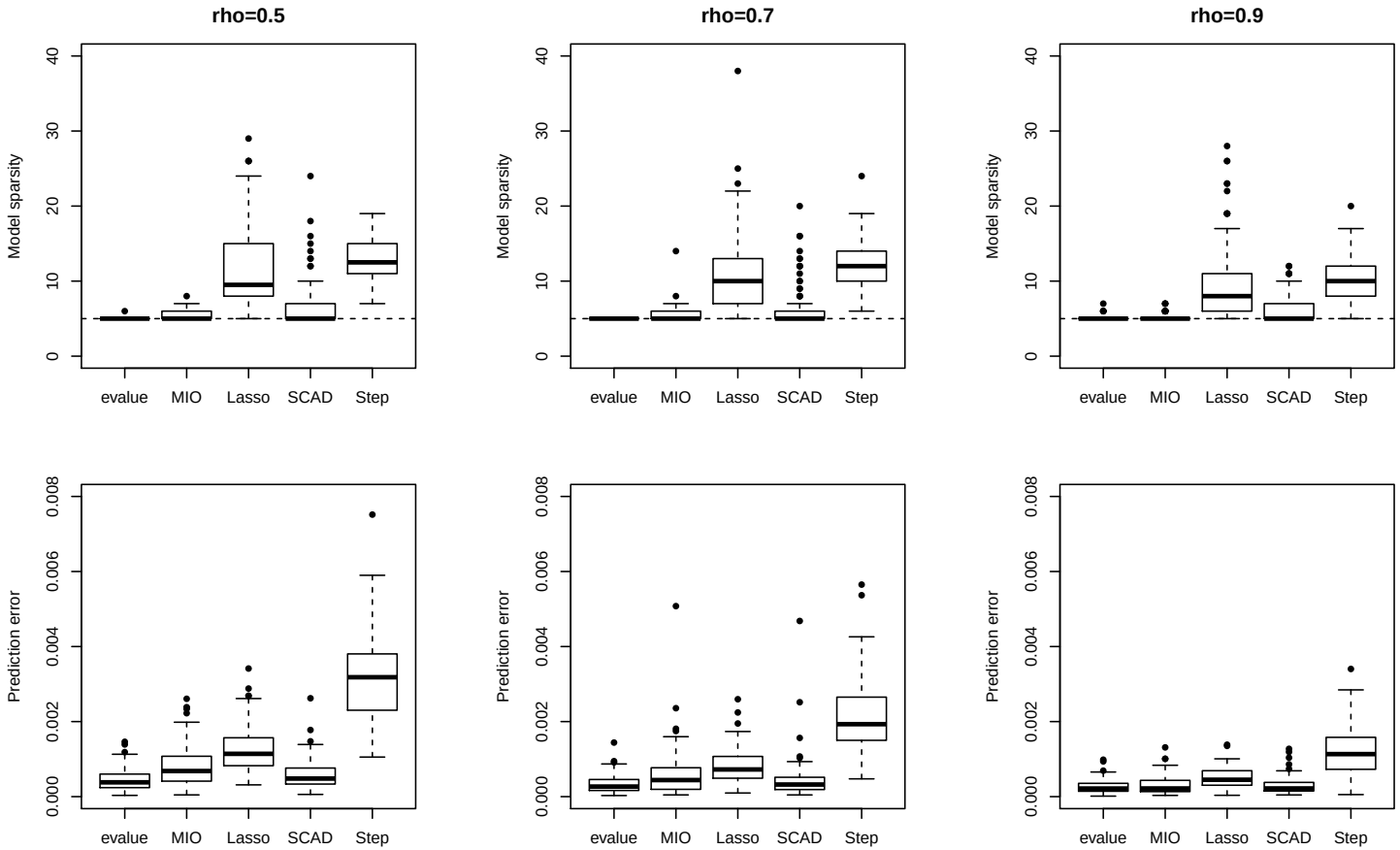


Figure: Model sparsity (top row) and prediction performance (bottom row) for all methods in $n = 60, p = 1000$ setup.

Results: $n = 60, p = 1000$

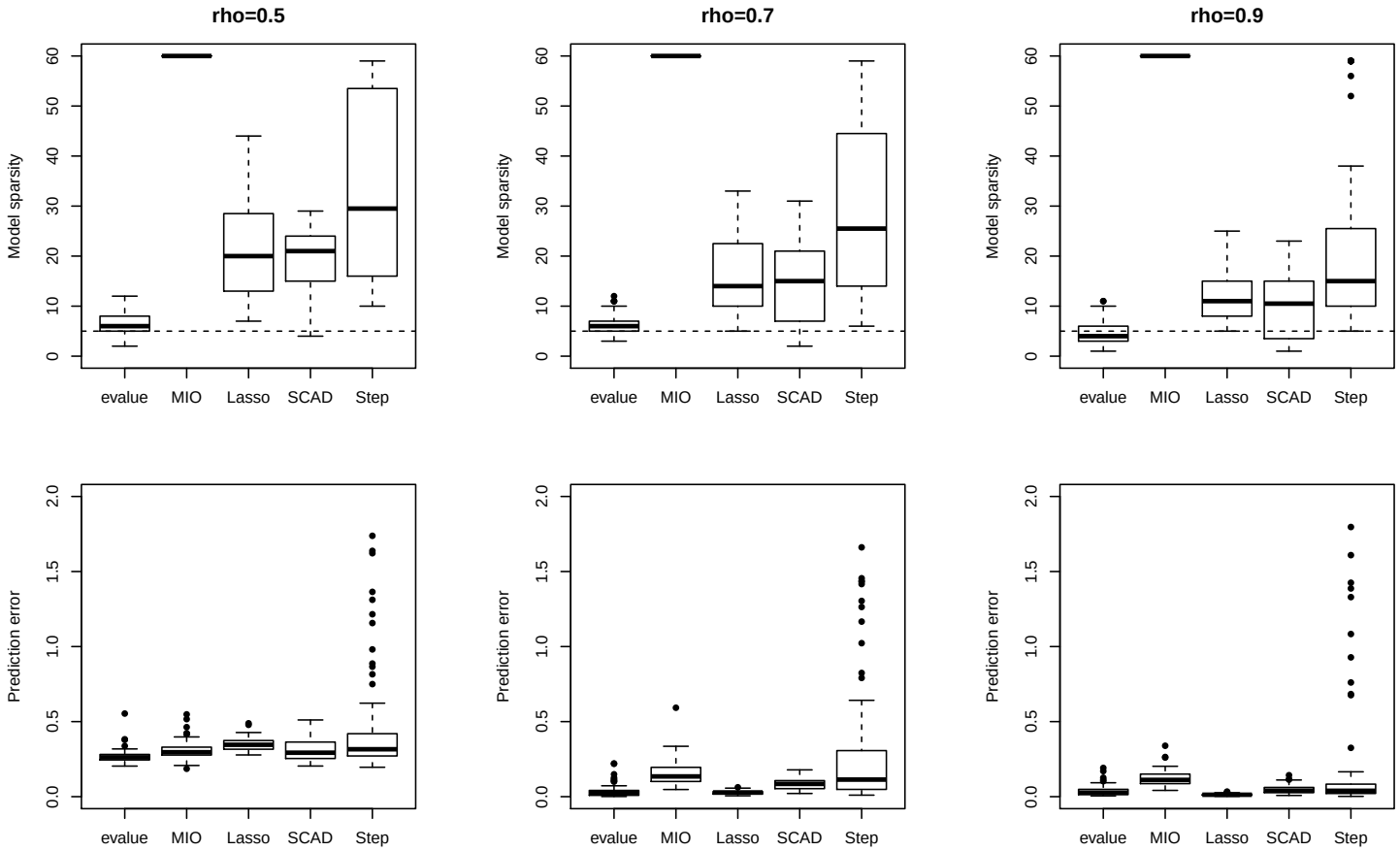


Figure: Model sparsity (top row) and prediction performance (bottom row) for all methods in $n = 60, p = 1000$ setup.

Results: effect of tuning parameter

Setting 1: $n = 1000, p = 60$						
Choice of τ_n	$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
	Sparsity	PE ($\times 10^{-4}$)	Sparsity	PE ($\times 10^{-4}$)	Sparsity	PE ($\times 10^{-4}$)
$\tau_n = \log n$	5.01	4.5	5.00	3.3	5.06	2.6
$\tau_n = n^{0.1}$	16.16	33.6	16.85	23.3	17.89	16.3
$\tau_n = n^{0.2}$	5.74	8.1	6.03	5.7	6.75	4.8
$\tau_n = n^{0.3}$	5.01	4.5	5.01	3.3	4.96	5.0
$\tau_n = n^{0.4}$	5.00	4.4	5.00	3.3	3.14	633.6
Setting 2: $n = 60, p = 1000$						
Choice of τ_n	$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
	Sparsity	PE ($\times 10^{-2}$)	Sparsity	PE ($\times 10^{-2}$)	Sparsity	PE ($\times 10^{-2}$)
$\tau_n = \log n$	6.63	4.5	6.34	3.4	4.79	3.7
$\tau_n = n^{0.1}$	7.57	4.6	7.24	3.0	7.16	2.1
$\tau_n = n^{0.2}$	7.38	4.6	7.23	3.0	6.61	2.3
$\tau_n = n^{0.3}$	6.94	4.6	6.58	3.0	5.44	3.0
$\tau_n = n^{0.4}$	6.08	4.1	5.66	3.9	3.90	6.0

Table: Model sparsity and prediction errors for different choices of τ

Simulation: linear mixed models

$$Y_i = X_i\beta + \epsilon \in \mathbb{R}^{n_i}$$
$$epsilon \sim N(0, V_i); \quad V_i = \sigma^2 I_{n_i} + Z_i \Delta Z_i^T$$

- m subjects, n_i observations per subject, $n = m \times n_i$ total observations;
- $p = 9, \beta = (1, 1, 0, 0, 0, 0, 0, 0, 0)^T$;
- Elements of $X_1, \ldots, X_m$ chosen from Unif$(-2, 2)$, random effect design matrix Z_i is first 4 columns of X_i;
- $$\Delta = \begin{pmatrix} 9 & & & \\ 4.8 & 4 & & \\ 0.6 & 1 & 1 & \\ 0 & 0 & 0 & 0 \end{pmatrix}$$
- Two settings: (i) $m = 30, n_i = 5$, (ii) $m = 60, n_i = 10$;
- We use i.i.d. draws of Gamma(1,1) as bootstrap weights $W_i + 1$.

Simulation results

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		$n_i = 5, m = 30$			$n_i = 10, m = 60$		
e-value based	$\tau_n/\sqrt{n} = 1$	57.4	0.0	5.24	43.8	0.0	4.03
	2	30.4	0.0	3.32	12.3	0.0	2.42
	3	15.6	0.0	2.54	3.2	0.0	2.10
	4	7.3	0.0	2.24	1.0	0.0	2.03
	5	3.0	0.0	2.09	0.7	0.0	2.02
	6	1.7	0.0	2.05	0.3	0.0	2.01
	7	1.0	0.0	2.03	0.0	0.0	2.00
	8	0.7	0.0	2.02	0.0	0.0	2.00
	9	0.0	0.0	2.00	0.0	0.0	2.00
	10	0.0	0.0	2.00	0.0	0.0	2.00
Peng and Lu (2012)	BIC	21.5	9.9	2.26	1.5	1.9	2.10
	AIC	17	11.0	2.43	1.5	3.3	2.20
	GCV	20.5	6	2.30	1.5	3	2.18
	$\sqrt{\log n/n}$	21	15.6	2.67	1.5	4.1	2.26

Comparison between our method and that proposed by Peng and Lu (2012) through average false positive percentage, false negative percentage and model size

Simulation results

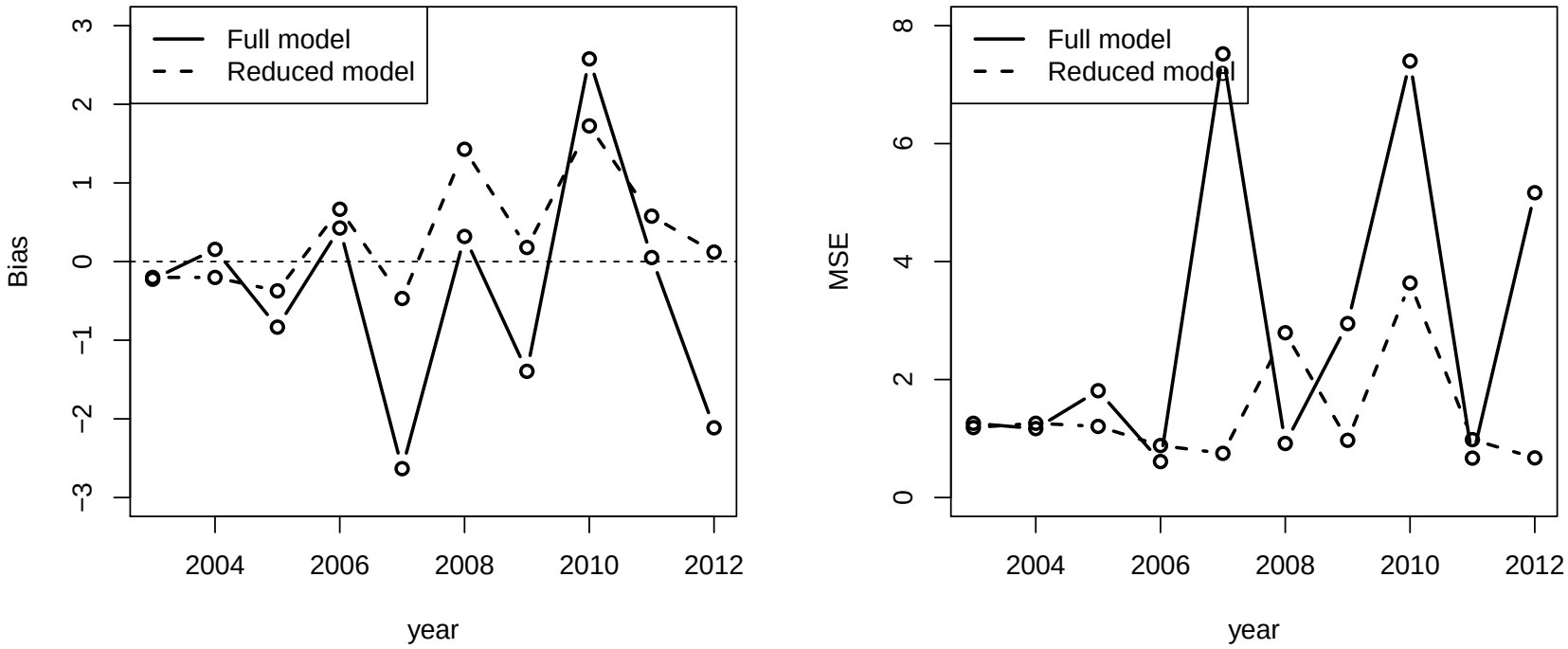
Method	$\tau_n/\sqrt{n}$	Setting 1	Setting 2
e-value based	1	2	16
	2	36	67
	3	60	91
	4	80	97
	5	91	98
	6	95	99
	7	97	100
	7	98	100
	8	100	100
	10	100	100
Bondell et al. (2010)		73	83
Peng and Lu (2012)		49	86
Fan and Li (2012)		90	100

Comparison of our method and three sparsity-based methods of mixed effect model selection through accuracy of selecting correct fixed effects

Application: selection of important predictors in Indian monsoon

- Annual median observations for 1978-2012;
- Local measurements across 36 weather stations (e.g. elevation, latitude, longitude), as well as global variables (e.g. El-Nino, tropospheric temperature variations) : total 35 predictors;
- Aim is two-fold: (i) Selecting important predictors, (ii) providing good predictions using the reduced model.

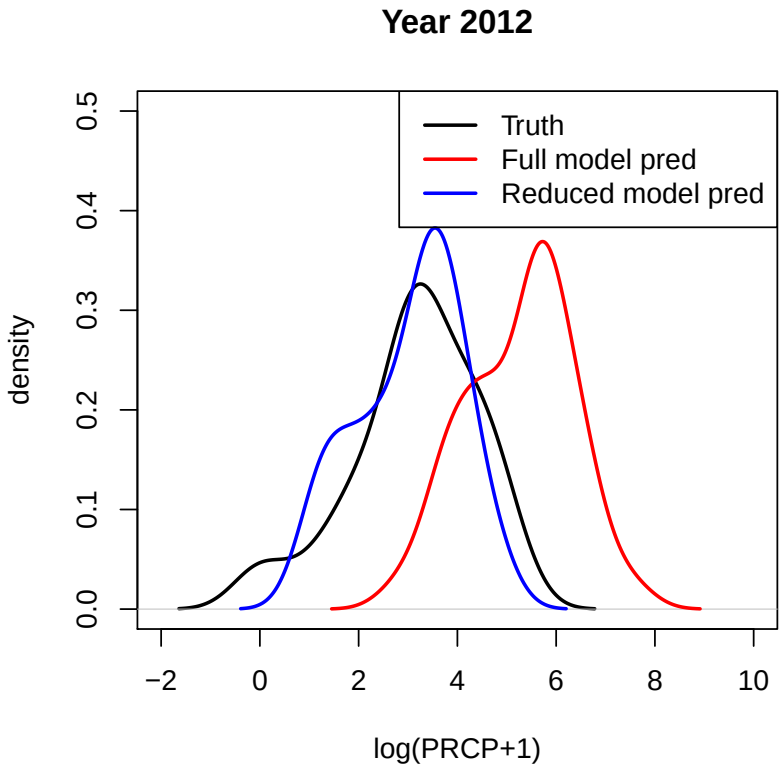
Application: results



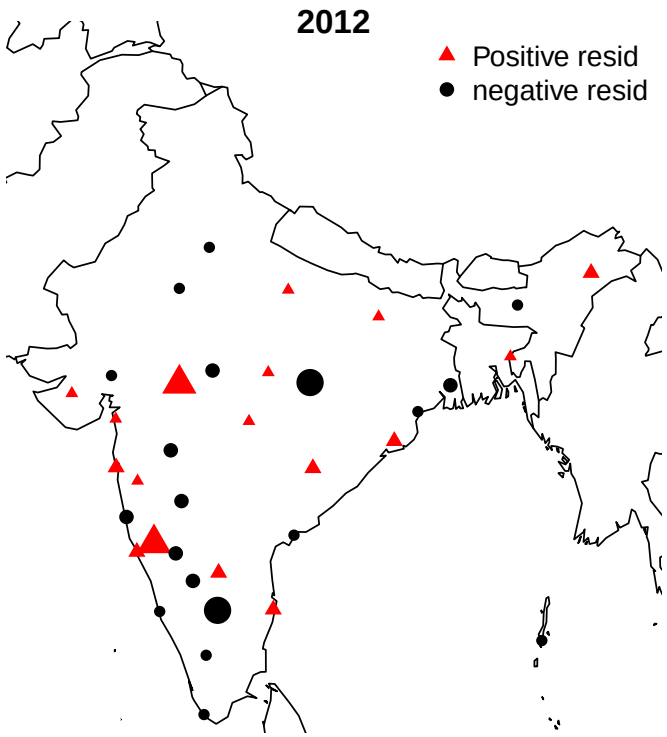
(a) (b)

Figure: Comparing full model rolling predictions with reduced models: (a) Bias across years, (b) MSE across years

Application: results



(a)



(b)

Figure: Comparing full model rolling predictions with reduced models: (a) density plots for 2012, (b) stationwise residuals for 2012

References

D. Bertsimas, A. King, and R. Mazumder. Best Subset Selection via a Modern Optimization Lens. *Ann. Statist.*, 44(2):813–852, 2016.

H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077, 2010.

S. Chatterjee and A. Bose. Generalized bootstrap for estimating equations. *Ann. Statist.*, 33:414–436, 2005.

Y. Fan and R. Li. Variable selection in linear mixed effect models. *Ann. Statist.*, 40(4):2043–2068, 2012.

H. Peng and Y. Lu. Model selection in linear mixed effect models. *J. Multivariate Anal.*, 109:109–129, 2012.

THANK YOU!

https://arxiv.org/abs/1706.02429

Questions?