A Model-Selection Criterion for Regression Estimators Based on Data Depth

Subho Majumdar
Snigdhansu Chatterjee

University of Minnesota, School of Statistics

University of Minnesota
Driven to Discover℠

Solve model selection!

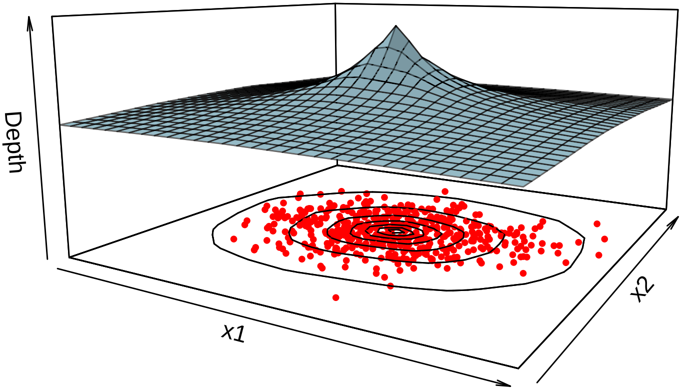
We Provide a bootstrap-based algorithm that:

1. Is **model-selection consistent**, i.e. for large enough sample size, P(variables selected by method equals actual set of important variables) $\rightarrow 1$;
2. **Works for a general class of regression models**;
3. Has **linear time complexity** with respect to number of parameters in the model.

Outlines

- **Preliminaries**: data depth, conditional models;
- **The algorithm**: population-level derivation and large sample properties
- **Results**: simulations and real data analysis

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



A scalar measure of how much inside a point is with respect to a data cloud

For any multivariate distribution $F = F_{\mathbf{X}}$, the depth of a point $\mathbf{x} \in \mathbb{R}^p$, say $D(\mathbf{x}, F_{\mathbf{X}})$ is any real-valued function that provides a 'center outward ordering' of $\mathbf{x}$ with respect to F (Zuo and Serfling, 2000).

Desirable properties (Liu, 1990)

(P1) *Affine invariance*: $D(A\mathbf{x} + \mathbf{b}, F_{A\mathbf{X}+\mathbf{b}}) = D(\mathbf{x}, F_{\mathbf{X}})$

(P2) *Maximality at center*: $D(\theta, F_{\mathbf{X}}) = \sup_{\mathbf{x} \in \mathbb{R}^p} D(\mathbf{x}, F_{\mathbf{X}})$ for $F_{\mathbf{X}}$ with center of symmetry θ, the *deepest point* of $F_{\mathbf{X}}$.

(P3) *Monotonicity w.r.t. deepest point*: $D(\mathbf{x}; F_{\mathbf{X}}) \leq D(\theta + a(\mathbf{x} - \theta), F_{\mathbf{X}})$

(P4) *Vanishing at infinity*: $D(\mathbf{x}; F_{\mathbf{X}}) \to \mathbf{0}$ as $\|\mathbf{x}\| \to \infty$.

Examples: Projection depth, Halfspace depth, Mahalanobis depth.

Conditional models

Consider the general estimation problem

$$y_i = h(\mathbf{x}_i^T \boldsymbol{\beta}) + \epsilon_i; \quad i = 1, ..., n$$

Assume $h(.)$ to be known, and ϵ having an arbitrary error distribution.

In this setup, a candidate model can be uniquely identified a $\boldsymbol{\beta}$ whose some indices are fixed (at values $\boldsymbol{\beta}_c$) and others (indices α) are unknown.

We say this combination is a **conditional model**:

$$\mu = (\alpha, \boldsymbol{\beta}_c)$$

$$\beta = \begin{array}{cl} \left[\begin{array}{c} ? \\ ? \\ ? \\ ? \\ ? \\ \hdashline 1 \\ 2 \\ 0 \\ \text{-}1 \end{array}\right] & \begin{array}{l} \\ \\ \text{estimated} \\ \text{in } \alpha \\ \\ \\ \\ \text{Fixed in } \beta_c \\ \\ \\ \end{array} \end{array}$$

- The set of all possible conditional models is:

$$\mathcal{M}_c = \left\{ (\alpha, \boldsymbol{\beta}_c) : \alpha \subseteq \mathcal{I}_p, \boldsymbol{\beta}_c \in \mathbb{R}^{|\mathcal{I}_p \setminus \alpha|} \right\}$$

  with $\mathcal{I}_p = \{1, 2, ..., p\}$.

- **Correct conditional models** are conditional models such that $\boldsymbol{\beta}_c$ is a subvector of $\boldsymbol{\beta}$, made from its elements at indices NOT in α, i.e. $\mathcal{I}_p \setminus \alpha$;

- **Wrong conditional models** are conditional models such that
  - At least one element of $\boldsymbol{\beta}_c$ is not in $\boldsymbol{\beta}$;
  - Or $\boldsymbol{\beta}_c$ is a subvector of $\boldsymbol{\beta}$, but not at indices $\mathcal{I}_p \setminus \alpha$.

$$\beta = \begin{array}{c} \left[\begin{array}{c} ? \\ ? \\ ? \\ ? \\ ? \\ \hdashline 1 \\ 2 \\ 0 \\ -1 \end{array}\right] \end{array} \begin{array}{l} \text{estimated} \\ \text{in } \alpha \\ \\ \text{Fixed in } \beta_c \end{array}$$

Given a conditional model μ, estimate $\boldsymbol{\beta}$ at indices α and append that by $\boldsymbol{\beta}_c$ to obtain a p-dimensional estimate of $\boldsymbol{\beta}$: part fixed, part random. Denote this by $\tilde{\boldsymbol{\beta}}_n(\mu)$.

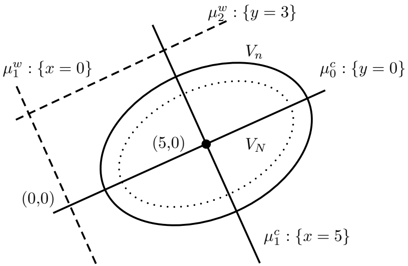
F_n is the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ satisfying standard regularity conditions.

Then our selection criterion is defined as:

$$C_n(\mu) = \mathbb{E}_{F_n|\mu} \left[D \left(\tilde{\boldsymbol{\beta}}_n(\mu), F_n \right) \right]$$

$$\beta = \begin{array}{cc} \left[\begin{array}{c} ? \\ ? \\ ? \\ ? \\ ? \end{array}\right] & \begin{array}{c} \text{estimated} \\ \text{in } \alpha \end{array} \\ \hline \left[\begin{array}{c} 1 \\ 2 \\ 0 \\ \text{-}1 \end{array}\right] & \text{Fixed in } \beta_c \end{array}$$

- Under a regression setup, any candidate model is nothing but a subset of this space of coefficients;
- Each point in the possible space of coefficients has a depth;
- We choose the model with **largest expected depth** among all candidate models, and show that this is indeed the correct model.



The algorithm

1. For large enough n, Calculate C_n for full model;
2. Drop a predictor, calculate C_n for the reduced model;
3. Repeat for all p predictors;
4. Collect predictors dropping which causes C_n to decrease. These are the predictors in the smallest correct model.

	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

Estimating C_n

$$C_n(\mu) = \mathbb{E}_{F_n|\mu}\left[D\left(\tilde{\boldsymbol{\beta}}_n(\mu), F_n \right) \right]$$

In a sample setup we neither know multiple instances of $\tilde{\beta}_n(\mu)$, nor of $\hat{\beta}_n$ (for getting hold of F_n).

Solution: use bootstrap!

$$\hat{C}_n(\mu) = \mathbb{E}_b\left[D\left(\tilde{\beta}_n^b(\mu), F_n^{b_1} \right) \right]$$

b and b_1 denote the collections of random sample weights for two *independent* wild bootstrap samples.

$$C_n(\mu) = \mathbb{E}_{F_n|\mu} \left[D \left(\tilde{\boldsymbol{\beta}}_n(\mu), F_n \right) \right]$$

In a sample setup we neither know multiple instances of $\tilde{\boldsymbol{\beta}}_n(\mu)$, nor of $\hat{\boldsymbol{\beta}}_n$ (for getting hold of F_n).

Solution: use bootstrap!

$$\hat{C}_n(\mu) = \mathbb{E}_b \left[D \left(\tilde{\beta}_n^b(\mu), F_n^{b_1} \right) \right]$$

b and b_1 denote the collections of random sample weights for two *independent* wild bootstrap samples.

$$C_n(\mu) = \mathbb{E}_{F_n|\mu} \left[D \left(\tilde{\boldsymbol{\beta}}_n(\mu), F_n \right) \right]$$

In a sample setup we neither know multiple instances of $\tilde{\boldsymbol{\beta}}_n(\mu)$, nor of $\hat{\boldsymbol{\beta}}_n$ (for getting hold of F_n).

Solution: use bootstrap!

$$\hat{C}_n(\mu) = \mathbb{E}_b \left[D \left(\tilde{\boldsymbol{\beta}}_n^b(\mu), F_n^{b_1} \right) \right]$$

b and b_1 denote the collections of random sample weights for two *independent* wild bootstrap samples.

Definition (Mammen, 1993)

Suppose $\hat{\epsilon}$ are the residuals obtained from with-intercept linear regression of $\mathbf{y}$ on X.

$$\mathbf{y}_b = \hat{\mathbf{y}} + U_b \hat{\epsilon}, X_b = X$$

$U_b = \text{diag}(U_{1b}, ..., U_{nb})$ with the U_{ib}-s drawn independently from a probability distribution with mean 0, variance τ_n^2.

Why use it here?

$$\hat{\boldsymbol{\beta}}_n^b = \hat{\boldsymbol{\beta}}_n + H_n^{-1} \frac{1}{n} \sum_{i=1}^n u_{ib} \mathbf{x}_i (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n); \quad H_n = \frac{1}{n} X^T X$$

(a) Correct conditional models:

$$\lim_{n\to\infty} P\left[\hat{C}_n(\mu^c) = C_n(\mu^c)\right] = 1 \quad \text{when } \tau_n \to \infty$$

(b) Drop-1 wrong conditional models:

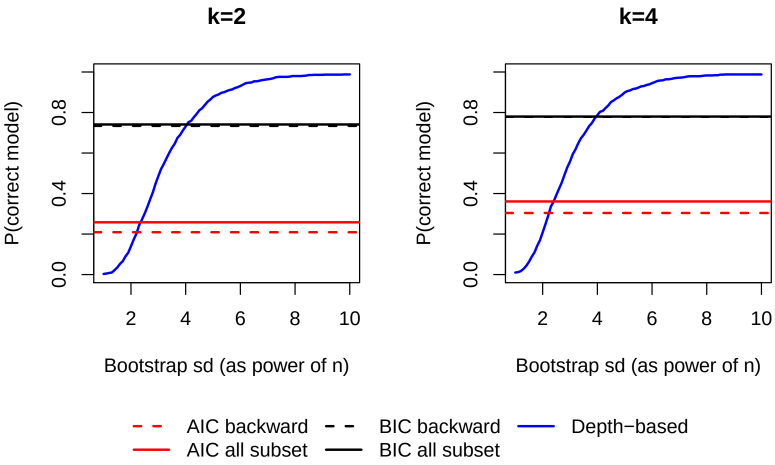
$$\begin{aligned} \lim_{n\to\infty} P\left[\hat{C}_n(\mu^w) > C_n(\mu^w)\right] &= 1 \quad \text{when } \tau_n \to \infty \\ \lim_{n\to\infty} P\left[\hat{C}_n(\mu^w) = C_n(\mu^w)\right] &= 1 \quad \text{when } \tau_n \to \infty \text{ and } \tau_n/\sqrt{n} \to 0 \end{aligned}$$

(c) Model-selection consistency:
When $\tau_n \to \infty$ and $\tau_n/\sqrt{n} \to 0$, the one-step procedure finds the correct model with probability going to 1.

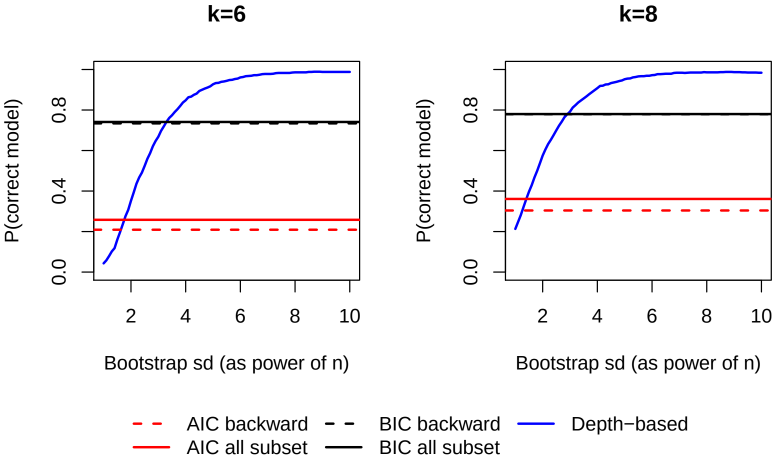
Simulation 1: setup

- $n = 100, p = 10$;
- Coefficient vector β is made of k ones and $p - k$ zeros: $k = 2, 4, 6, 8$;
- Randomly chosen 100 rows and first 10 columns in the dataset available in `http://www.stat.umn.edu/geyer/5102/data/ex6-8.txt` taken as X. Responses generated as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}_p, I_p)$;
- 1000 such samples are drawn;
- Bootstrap sample size 1000 for estimating C_n. Bootstrap standard deviation $\tau_n = \text{seq}(1, 10, \text{by} = 0.1)$;
- Compared with AIC and BIC backward deletion and all-subset regression.

Simulation 1: results



Simulation 1: results



Simulation 2: Synthetic data analysis

- $p = 9, \boldsymbol{\beta} = (1, 1, 0, 0, 0, 0, 0, 0)^T$;
- Linear mixed model: m subjects, n_i observations per subject, $n = m \times n_i$ total observations;
- Elements of $X_{n \times p}$ chosen from Unif$(-2, 2)$, random effect design matrix Z is first 4 columns of X;
- $\mathbf{y}_i = X_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 I + Z_i D Z_i^T)$ with

$$D = \begin{pmatrix} 9 & & & \\ 4.8 & 4 & & \\ 0.6 & 1 & 1 & \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

- Two settings: (i) $m = 30, n_i = 5$, (ii) $m = 60, n_i = 10$;

Simulation 2: results

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		$n_i = 5, m = 30$			$n_i = 10, m = 60$		
Depth-based	$\tau = 1$	60.1	0.0	5.35	56.7	0.0	4.96
	$\tau = 2$	30.8	0.0	3.21	29.4	0.0	3.09
	$\tau = 3$	11.1	0.0	2.37	9.6	0.0	2.32
	$\tau = 4$	2.4	0.0	2.14	1.8	0.0	2.01
	$\tau = 5$	1	0.0	2.03	0.0	0.0	2.00
	$\tau = 6$	0.2	0.0	2.01	0.0	0.0	2.00
	$\tau = 7$	0.0	0.0	2.00	0.0	0.0	2.00
	$\tau = 8$	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

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

Simulation 2: results

Method		Setting 1	Setting 2
Depth-based	$\tau = 1$	1	1.5
	$\tau = 2$	29.5	29
	$\tau = 3$	70	73.5
	$\tau = 4$	93	94.5
	$\tau = 5$	97	100
	$\tau = 6$	99.5	100
	$\tau = 7$	100	100
	$\tau = 8$	100	100
Bondell et al. (2010)		73	83
Peng and Lu (2012)		49	86
Fan and Li (2012)		90	100

Table : 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.

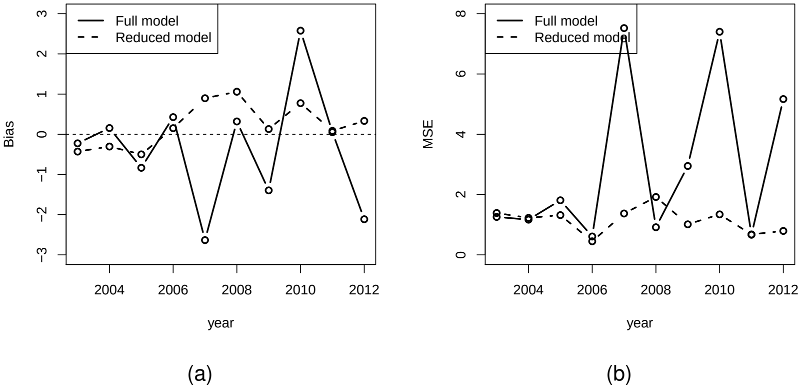
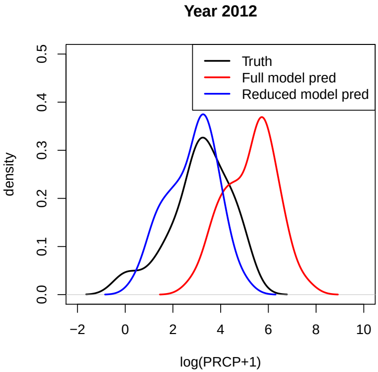
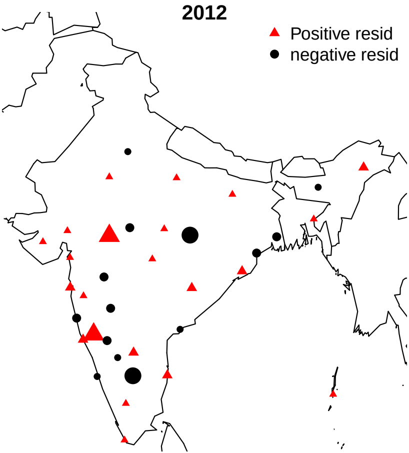


Figure : Comparing full model rolling predictions with reduced models: (a) Bias across years, (b) MSE across years

Application: results



(a)



(b)

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

Future work

1. Robust M- or MM-estimation of regression coefficients;
2. Explore its connection with existing bootstrap-based methods of variable selection;
3. Develop versions for classification problems and high-dimensional regression

References

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.

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

R.Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18:405–414, 1990.

E. Mammen. Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Ann. Statist.*, 21(1):255–285, 1993.

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

Y. Zuo and R. Serfling. General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482, 2000.

THANK YOU!

Acknowledgement: NSF Grant IIS-1029711