Fast and General Best Subset Selection using Data Depth and Resampling

Subhabrata Majumdar*

University of Florida Informatics Institute

and

Snigdhansu Chatterjee†

School of Statistics, University of Minnesota

Abstract

We present a technique using data depth functions and resampling to perform best subset variable selection for a wide range of statistical models. We do this by assigning a score, called an *e*-value, to a candidate model, and use a fast bootstrap method to approximate sample versions of these *e*-values. We establish that under general conditions, *e*-values can separate statistical models that adequately explain properties of the data from those that do not. This results in a fast and parallel algorithm that fits only a single model and evaluates $p+1$ models, p being the number of predictors under considerations, as opposed to the traditional requirement of fitting and evaluating 2^p models. We illustrate in simulation experiments that our proposed method typically performs better than an array of currently used methods for variable selection, in linear models as well as fixed effect selection in linear mixed models. As a real data application, we use our procedure to elicit climatic drivers of Indian summer monsoon precipitation.

Keywords: Model selection, best subset selection, linear mixed model, generalized bootstrap, Indian monsoon.

1 Introduction

Consider the linear model

$$Y = X\beta + \epsilon \tag{1.1}$$

where $Y \in \mathbb{R}^n$ is the vector of responses, $X \in \mathbb{R}^{n \times p}$ is the predictor matrix, $\beta \in \mathbb{R}^p$ is the unknown vector of coefficients, and $\epsilon \in \mathbb{R}^n$ is a vector of independent and identically distributed random errors. Many important applications of this model demand the detection of important predictors affecting the response and thus obtaining a parsimonious fit to the data in hand. Because of this, variable selection in linear models has been a well-studied topic, including classics (Akaike, 1970; Schwarz, 1978; Shao, 1996; Tibshirani, 1996) and more recent advances (Bertsimas et al., 2016; Lai et al., 2015; Zou, 2006). The two major avenues variable selection has been explored through are sparse penalized regression and best subset selection. Sparse penalized regression has received abundant attention in the past two decades due to its relative computational simplicity and ability to tackle high-dimensional data. However, there are several inferential and algorithmic issues associated with such methods: for example the bias and confidence intervals of Lasso estimates (Zhang and Zhang, 2014) and use of convex relaxations to compute approximate local solutions of nonconvex penalized regression problems (Wang et al., 2013; Zou and Li, 2008). A detailed discussion of the caveats of penalized regression procedures can be found in Bertsimas et al. (2016).

On the other hand, the major challenge of performing best subset selection by checking multiple model fits is computational. In the setup of (1.1) there are 2^p potential models, and navigating through this model space is an NP-Hard problem (Natarajan, 1995). Greedy methods like forward selection or backward deletion give no theoretical guarantee to select the correct model, and popular algorithms like leaps-and-bounds or branch-and-bounds are not computationally scalable to even moderate data dimensions ($n \simeq 100, p \simeq 50$). A recent paper by Bertsimas *et al* Bertsimas et al. (2016) proposed a tractable solution to the best subset problem using techniques from the optimization literature. It requires specifying the desired level of sparsity in the final model beforehand and works well in practice when the true model has a small number of non-zero coefficients. However, this is a largely unverifiable assumption, and there are more computational issues associated with it in high-dimensional situations, which we shall discuss later in the paper.

Although there has been some progress in best subset model selection for linear regression, it has received lesser attention for more complex models due to its perceived computational intractability and lack of intelligent search strategies. The fence methods (Jiang et al., 2008; Nguyen and Jiang, 2014) are popular in performing model selection in mixed effect models. Model selection criteria that expand the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Mallow's C_p to mixed models (Meza and Lahiri, 2005; Vaida and Blanchard, 2005) and robust regression (Ronchetti, 2000) have also been proposed. However, they still depend on multiple model fits and

suffer from the same scalability issues as linear regression best subset selection.

Our method In this paper, we propose a unified framework that is able to perform best subset selection in the coefficient vector β for a range of statistical models much broader than (1.1) by traversing a very small proportion of the space of candidate models. Even though our proposal is not based on penalization, we are able to determine important variables in a model *by training a single model*: the one with all covariates under consideration. We do this by obtaining a score for a candidate model $\mathcal{M}$ based on the amount of similarity between the *sampling distribution* of the model coefficient estimate $\hat{\beta}_{\mathcal{M}}$ and that of the least parsimonious estimate, say $\hat{\beta}$. We use data depths, which are functions that provides center-outward ranking for points in multivariate space with respect to a probability distribution (Tukey, 1975; Zuo, 2003; Zuo and Serfling, 2000) to compute this score, which we call the *e*-**value** of the candidate model.

Specifically, 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 $\hat{\beta}$, which is denoted by $[\hat{\beta}]$. In this setup, for large enough n the index set $\mathcal{S}_{select} \subseteq \{1, \ldots, p\}$ obtained using Algorithm 1 elicits all non-zero covariates in the true parameter vector.

Algorithm 1. (Best subset selection using *e*-values)
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 in $1 : p$
    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 a size-n finite sample, we use resampling to generate multiple copies of $\hat{\beta}$ and subsequently estimate $e(\mathcal{M})$ for each model.

The example in Figure 1.1 illustrates our motivation of using mean depth as a model selection criterion. Suppose $p = 2$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \sigma > 0, i = 1, \ldots, n$ in (1.1), and consider the following choice of models:

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

where Θ_i denotes the domain of parameter values corresponding to model $\mathcal{M}_i; i = 1, ..., 4$. Now assume that the true value of the coefficient vector is $\beta_0 = (5,0)^T$, a^T denoting
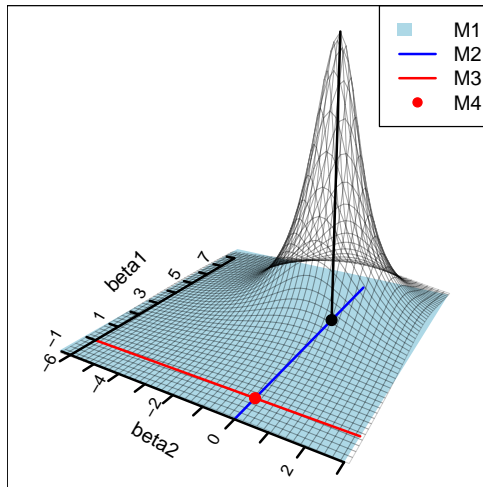
3

Figure 1.1: Illustration of the *e*-value method. The depth surface is obtained using mahalanobis depth.

the transpose of a vector or matrix-valued a. The full model estimate has sampling distribution $\mathcal{N}(\beta_0, \sigma^2(X^TX)^{-1})$. As n grows, this distribution concentrates around β_0, and the depths at points along the (red) line $\beta_2 = 0$, as well as at the (red) point $(0,0)$ become smaller and smaller. Hence, the expected depths go to 0 for $\mathcal{M}_3$ and $\mathcal{M}_4$. However, owing to affine invariance of depth functions, the range of depth values for points in $\mathcal{M}_1$ (blue line) and $\mathcal{M}_2$ (blue surface) stays the same. Moreover, depth contours coincide with density contours for elliptical distributions (Liu and Singh, 1993), thus $e(\mathcal{M}_1)$ and $e(\mathcal{M}_2)$ remain positive and do not converge to 0 as $n \to \infty$.

For the purpose of covariate selection, the space of all possible models can be divided into two classes: models with non-zero support that contains the non-zero support of the true coefficient vector, and models that do not satisfy this. Because of the above reasoning, *e*-values for models in the first class become clearly separated from *e*-values of models in the second class. We further leverage the properties of depth functions to prove that within the first class of 'good' models, a rank ordering of *e*-values of the models exists based on the amount of parsimony in them: after which Algorithm 1 becomes immediate.

Structure of paper   The paper is organized as follows. Section 2 provides background for the development of our framework. We define *e*-values formally in Section 3, present details about their estimation through resampling and their implementation through our fast algorithm. Theoretical results are presented in Section 4. In Section 5 we present illustrative examples on how our fast model selection algorithm is implemented, and its performance relative to state-of-art methods in both classical $n > p$ and high-dimensional situations. This is followed in Section 6 with a study on the possible determinants of precipitation in the Indian sub-continent during the summer monsoon months. In Section 7 we discuss the potentials of our proposed method, directions of future research, as well as some caveats and concluding comments. Proofs of all theoretical results are given in the

appendix.

Notations We denote the $(i,j)^{\text{th}}$ entry of a matrix $M \in \mathbb{R}^{A \times B}$ by $(M)_{ij}$. The notations I_A and 0_A indicate the identity matrix of dimension $A \times A$ and the vector of zeros of dimension A, respectively. For a vector $v = (v_1, \ldots, v_A)^T \in \mathbb{R}^A$ we denote its euclidean norm by $\|v\| = \sqrt{\sum_{i=1}^A v_i^2}$, and l_0-norm by $\|v\|_0 = \sum_{i=1}^A \mathbb{I}(v_i \neq 0)$: $\mathbb{I}(.)$ being the indicator function. We denote the non-zero support of v by $\text{supp}(v)$, i.e. $\text{supp}(v) = \{i : v_i \neq 0\}$. For random variables $R, R_n, n = 1, 2, \ldots$, the notation $R_n \rightsquigarrow R$ means convergence in distribution, and $R_n \overset{P}{\to} R$ or $\text{plim}_{n \to \infty} R_n = R$ means convergence in probability. Expectations are denoted by $\mathbb{E}(\cdot)$, and the variance of a scalar-valued random variable or the covariance matrix of a vector-valued random variable is denoted by $\mathbb{V}(\cdot)$.

2 Preliminaries

We introduce the theoretical framework necessary for our methodology in this section. This includes specifications for the following three things: (1) an appropriate method of estimation, (2) categorization of candidate models, and (3) the choice of a depth function.

Suppose at stage n there is a triangular array of functions

$$\{\Psi_{ni}(\theta_n, Z_{ni}) : 1 \leq i \leq k_n, n \geq 1\} \tag{2.1}$$

where $\mathcal{Z}_n = \{Z_{n1}, \ldots, Z_{nk_n}\}$ is an observable array of random variables, and $\theta_n \in \Theta_n \subseteq \mathbb{R}^p$. Functions like $\Psi_{ni}(\cdot)$ have been referred to as *energy functions* in optimization and related literature, and have also been called *contrast functions* (Bose and Chatterjee, 2003; Michel and Pfanzagl, 1971; Pfanzagl, 1969). This framework naturally includes different kinds of regressions, mixed effect models, times series and other traditional topics where covariate or model selection has been historically studied. Note that we use a triangular array-based framework here, where the energy functions and the true parameter may depend on n for precise theoretical study, and to capture the scientific reality that data, models and goals of research evolve over time.

We assume that there is a true unknown vector of parameters θ_{0n}, which is the unique minimizer of

$$\Psi_n(\theta_n) = \mathbb{E} \sum_{i=1}^{k_n} \Psi_{ni}(\theta_n, Z_{ni}) \tag{2.2}$$

Also denote the common non-zero support of all estimable parameters as $\mathcal{S}_{*n} = \cup_{\theta_n \in \Theta_n} \text{supp}(\theta_n)$. 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

$$\theta_{mnj} = \begin{cases} \text{Unknown } \theta_{mnj} \in \Theta_{nj} & \text{for } j \in \mathcal{S}_n \\ \text{Known } C_{nj} \in \Theta_{nj} & \text{for } j \notin \mathcal{S}_n \end{cases} \quad (2.3)$$

So that in each model a part of the parameter vector consists of known constants, and the other part is unknown.

2.1 Method of estimation

The estimator $\hat{\theta}_{*n}$ of θ_{0n} is obtained as a minimizer of the sample analog of (2.2). Thus

$$\hat{\theta}_{*n} = \arg\min_{\theta_n} \hat{\Psi}_n(\theta_n) = \arg\min_{\theta_n} \sum_{i=1}^{k_n} \Psi_{ni}\left(\theta_n, Y_{ni}\right) \quad (2.4)$$

We assume an elliptical asymptotic distribution for $\hat{\theta}_{*n}$. Following Fang et al. (1990), the density function of an elliptically distributed random variable takes the form:

$$h(x; \mu, \Sigma) = |\Sigma|^{-1/2} g((x - \mu)^T \Sigma^{-1}(x - \mu))$$

where $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p \times p}$ is positive semi-definite, and g is a non-negative scalar-valued density function that is continuous and strictly increasing, and is called the *density generator* of the elliptical distribution. We denote such an elliptical distribution by $\mathcal{E}(\mu, \Sigma, g)$. For the asymptotic parameter distribution we also assume the following 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 positive-definite 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 $\text{plim}_{n \to \infty} V_n = V$.

In practice we mostly deal with Gaussian limiting distributions, where (A1)-(A2) are naturally satisfied.

The estimator in (2.4) corresponds to the *full model* $\mathcal{M}_{*n}$, i.e. the model where all indices in $\mathcal{S}_{*n}$ are estimated. For any other model $\mathcal{M}_n$, we simply augment entries of $\hat{\theta}_{*n}$ at indices in $\mathcal{S}_n$ with elements of C_n elsewhere to obtain a model-specific coefficient estimate:

$$\hat{\theta}_{mnj} = \begin{cases} \hat{\theta}_{*nj} & \text{for } j \in \mathcal{S}_n \\ C_{nj} & \text{for } j \notin \mathcal{S}_n \end{cases} \quad (2.5)$$

The logic behind this plug-in estimate is simple: for a candidate model $\mathcal{M}_n$, a joint distribution of its estimated parameters, i.e. $[\hat{\theta}_{sn}]$, can actually be obtained from $[\hat{\theta}_{*n}]$ by marginalizing at indices $\mathcal{S}_n$.

6

2.2 Characterization of models

Given the formulation of models in (2.3), any model can be classified into two categories: *adequate* and *inadequate* models.

Definition 2.1. *A model $\mathcal{M}_n$ is called adequate if*

$$\lim_{n\to\infty} \sum_{j\notin\mathcal{S}_n} |C_{nj} - \theta_{0nj}| = 0 \qquad (2.6)$$

A model that is not adequate, will be called an inadequate model.

We also consider a stronger version of adequacy below.

Definition 2.2. *The model $\mathcal{M}_n$ is called strictly adequate if, for all n and $j \notin \mathcal{S}_n$, $C_{nj} = \theta_{0nj}$.*

By definition the full model is always adequate, as is the model corresponding to the singleton set containing the true parameter, i.e. $\{\theta_{0n}\}$ (because $\text{supp}(\theta_{0n}) \subseteq \mathcal{S}_{*n}$). Thus the set of adequate models is non-empty by construction.

In classical variable selection where all elements of the constant terms C_n are set to 0 for all models, as in linear regression where a subset of covariates X_s is used in fitting the expression $Y = X_s\beta_s + \epsilon_s$, the above notion of model adequacy captures standard notions of model 'correctness'. Specifically, (2.6) reduces to $\lim_{n\to\infty} \mathbb{E}Y - X_s\beta_s = 0$. This is weaker than the traditional condition $\mathbb{E}Y - X_s\beta_s = 0$ about model 'correctness' implicit in many studies, which is captured in our condition for a model being *strictly adequate*. In addition, our definition allows asymptotically similar models to be clubbed together, and has the flexibility to accommodate local asymptotic characteristics. For instance, consider the data-generating model $Y_{ni} = X_{1i}\beta_{01} + X_{2i}\delta_n + \epsilon_i$ for some $\beta_{01} \in \mathbb{R}, \delta_n = o(1), \epsilon_i \overset{i.i.d.}{=} N(0, \sigma_0^2)$ with $\sigma_0 > 0$ and $i = 1, \ldots, k_n$. Evidently, here $\theta_{0n} = (\beta_{01}, \delta_n, \sigma_0)^T$. With our formulation, the (constant) sequence of models $\mathcal{M}_n$ with $\Theta_{mn} = \{(\beta_1, 0, \sigma^2)^T : \beta_1 \in \mathbb{R}, \sigma > 0\}$ will be adequate. Such situations can arise from prior choices in Bayesian variable selection techniques, for example see Narisetty and He (2014); Ročková and George (2016).

Besides model adequacy, a notion that will be important in our framrwork is the one of *nested models*.

Definition 2.3. *We consider a model $\mathcal{M}_{1n}$ to be nested in $\mathcal{M}_{2n}$, notationally $\mathcal{M}_{1n} \prec \mathcal{M}_{2n}$, if $\mathcal{S}_{1n} \subset \mathcal{S}_{2n}$ and C_{2n} is a subvector of C_{1n}.*

If a model $\mathcal{M}_n^{ad}$ is adequate, then any model $\mathcal{M}_n$ such that $\mathcal{M}_n^{ad} \prec \mathcal{M}_n$ is also adequate. In the context of covariate selection in the linear model of (1.1), this obtains a linear ordering, with the most parsimonious adequate model having $\mathcal{S}_n = \text{supp}(\beta_0)$ and $C_n = 0_{p-|\mathcal{S}|}$ nested in all other adequate models. However all models nested *within* it shall be inadequate.

2.3 Depth functions

Let $\mathcal{G}$ denote the set of probability measures on $\mathbb{R}^p$, and let $D : \mathbb{R}^p \times \mathcal{G} \to [0, \infty)$ be a data depth function (Zuo and Serfling, 2000). We consider data depth functions satisfying the following properties:

(B1) The map D is invariant to affine transformations, i.e. for any non-singular matrix $A \in \mathbb{R}^{p \times p}$, and $b \in \mathbb{R}^p$ and random variable Y having distribution $\mathbb{G} \in \mathcal{G}$,

$$D(x, \mathbb{G}) = D(Ax + b, [AY + b])$$

(B2) The map $D(., \mathbb{G})$ is Lipschitz continuous, where $\mathbb{G} \in \mathcal{G}$ is a fixed, non-degenerate probability distribution. That is, there exists $\delta > 0$ and $\alpha \in (0, 1)$, possibly depending on $\mathbb{G}$ such that whenever $|x - y| < \delta$, we have

$$|D(x, \mathbb{G}) - D(y, \mathbb{G})| < |x - y|^\alpha$$

(B3) Assume that $Y_n \in \mathbb{R}^p$ is a sequence of random variables converging in distribution to some $\mathbb{Y} \in \mathcal{G}$. Then $D(y, [Y_n])$ converges uniformly to $D(y, \mathbb{Y})$. In particular, if $Y \sim \mathbb{Y}$, then

$$\lim_{n \to \infty} \mathbb{E}D(Y_n, [Y_n]) = \mathbb{E}D(Y, \mathbb{Y})$$

exists and is finite.

(B4) For any $\mathbb{G} \in \mathcal{G}$, $\lim_{\|x\| \to \infty} D(x, \mathbb{G}) = 0$.

(B5) For any $\mathbb{G} \in \mathcal{G}$ with a point of symmetry $\mu(\mathbb{G}) \in \mathbb{R}^p$, we have for any $t \in (0, 1)$ and any $x \in \mathbb{R}^p$

$$D(x, \mathbb{G}) < D(\mu(\mathbb{G}) + t(x - \mu(\mathbb{G})), \mathbb{G}) < D(\mu(\mathbb{G}), \mathbb{G}) = \sup_{x \in \mathbb{R}^p} D(x, \mathbb{G}) < \infty$$

That is, the depth function takes a maximum value at $\mu(\mathbb{G})$, and is strictly decreasing along any ray connecting $\mu(\mathbb{G})$ to any point $x \in \mathbb{R}^p$.

Conditions (B1), (B4) and (B5) are integral to the formal definition of data depth (Zuo and Serfling, 2000), while (B2) and (B3) implicitly arise for several depth functions (Mosler, 2013). For most of the theoretical analysis in this paper we require only a subset of these properties. However, we shall continue to use data depths throughout for simplicity.

3 The e-values methodology

We now associate with each model $\mathcal{M}_n$ a functional of the depth of the model estimate with respect to the full model sampling distribution. We call this the **e-value** of the model. An example of e-value is the mean depth:

$$e_n(\mathcal{M}_n) = \mathbb{E}D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}]) \tag{3.1}$$

8

which we concentrate on for the rest of the paper. However, any other functional of $D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}])$ may also be used here, and a large proportion of our theoretical discussion is applicable to any smooth functional of the distribution of $D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}])$. Furthermore, the distribution of $D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}])$ is itself informative, and has an important role to play in the study of uniform convergence. We defer all this discussion and analysis to future research.

3.1 Resampling approximation of e-values

There are two random quantities involved in (3.1), namely $\hat{\theta}_{mn}$ and $\hat{\theta}_{*n}$. Typically, the distribution of either of them is not known, and have to be elicited from the data. We use resampling for this purpose. Because of the plugin method in (2.5), only $[\hat{\theta}_{*n}]$ needs to be approximated, and we propose using Generalized Bootstrap (Chatterjee and Bose, 2005) for this. We take an exchangeable array of non-negative random variables, independent of the data, as resampling weights: $\mathcal{W}_{rn} = (\mathbb{W}_{rn1}, \dots, \mathbb{W}_{rnk_n})^T \in \mathbb{R}^{k_n}$, and obtain the resampling estimator $\hat{\theta}_{r*n}$ as the minimizer of

$$\hat{\Psi}_{rn}(\theta_n) = \sum_{i=1}^{k_n} \mathbb{W}_{rni} \Psi_{ni}\Big(\theta_n, Z_{ni}\Big). \tag{3.2}$$

We assume the following conditions on the resampling weights as $n \to \infty$:

$$\mathbb{E}\mathbb{W}_{rn1} = 1, \tag{3.3}$$
$$\mathbb{V}\mathbb{W}_{rn1} = \tau_n^2 \uparrow \infty, \tag{3.4}$$
$$\tau_n^2 = o(a_n^2), \tag{3.5}$$
$$\mathbb{E}W_{rn1}W_{rn2} = O(k_n^{-1}), \tag{3.6}$$
$$\mathbb{E}W_{rn1}^2 W_{rn2}^2 \to 1, \tag{3.7}$$
$$\mathbb{E}W_{rn1}^4 < \infty. \tag{3.8}$$

Several resampling schemes can be described in the above format and are discussed in Chatterjee and Bose (2005). Here we provide two examples.

Example 3.1 (The m-out-of-n or *moon*-bootstrap)**.** In our framework, the *moon*-bootstrap is identified with $\mathcal{W}_{rn}$ having a Multinomial distribution with parameters m and probabilities $k_n^{-1}(1, \dots, 1) \in \mathbb{R}^{k_n}$, by a factor of k_n/m. Thus we have $\mathbb{E}\mathbb{W}_{rni} = (m^{-1}k_n)(m/k_n) = 1$, and $\mathbb{V}\mathbb{W}_{rni} = \tau_n^2 = (m^{-1}k_n)^2(mk_n^{-1}(1 - k_n^{-1}) = O(m^{-1}k_n)$. In typical applications of the *moon*-bootstrap, we require that $m \to \infty$ and $m/k_n \to 0$ as $n \to \infty$. Thus we have $\tau_n^2 \to \infty$ as $n \to \infty$, that is, the scale factor of the resampling weights $\mathbb{W}_{rni}$ tending to infinity with n. $\square$

Example 3.2 (The scale-enhanced Bayesian bootstrap)**.** A version of Bayesian bootstrap may be constructed by choosing $\mathbb{W}_{rni}$ to be independent and identically distributed

Gamma random variables, with mean 1 and variance $\tau_n^2 \to \infty$ as $n \to \infty$. The functionality of this resampling scheme and Bayesian interpretation remain similar to the standard Bayesian bootstrap, however some convenient properties like conjugacy are lost. □

Under fairly weak regularity conditions on the first two derivatives of $\Psi_{ni}(\cdot)$, Chatterjee and Bose (2005) proved that $(a_n/\tau_n)(\hat{\theta}_{r*n} - \hat{\theta}_{*n})$ converges to the same weak limit in probability as $a_n(\hat{\theta}_{*n} - \theta_{0n})$. To estimate the e-value of any model we obtain two independent copies of the solution to (3.2), which we index by r and r_1. We use the first set of resamples to obtain the distribution $[\hat{\theta}_{r*n}]$ conditional on the data that approximates $[\hat{\theta}_{*n}]$, and the second set of resamples to obtain $\hat{\theta}_{r_1mn}$ corresponding to the model $\mathcal{M}_n$:

$$\hat{\theta}_{r_1mnj} = \begin{cases} \hat{\theta}_{r_1*nj} & \text{for } j \in \mathcal{S}_n; \\ C_{nj} & \text{for } j \notin \mathcal{S}_n \end{cases} \tag{3.9}$$

Consequently, the resampling estimate of a model e-value is defined as

$$e_{rn}(\mathcal{M}_n) = \mathbb{E}_{r_1} D\left(\hat{\theta}_{r_1mn}, [\hat{\theta}_{r*n}]\right), \tag{3.10}$$

where $\mathbb{E}_{r_1}$ is expectation, conditional on the data, of the resampling indexed by r_1.

The actual implementation of the resampling procedure is carried out by generating independent copies $\mathcal{W}_{1n}, \ldots, \mathcal{W}_{Rn}$ for some sufficiently large integer R. We do not need to solve (3.2) explicitly to calculate $\hat{\theta}_{r*n}$. Rather, one can use already computed model quantities in the following approximation to get them:

$$\hat{\theta}_{r*n} = \hat{\theta}_{*n} - \frac{\tau_n}{a_n} \left[\sum_{i=1}^{n} W_i \Psi_i''(\hat{\theta}_{*n}, Z_{ni}) \right]^{-1} \sum_{i=1}^{n} W_i \Psi_i'(\hat{\theta}_{*n}, Z_{ni}) + R_{rn}; \tag{3.11}$$

$$\mathbb{E}_r \|R_{rn}\|^2 = o_P(1)$$

This is immediate from Theorem 3.5 in Chatterjee and Bose (2005). Consequently only Monte Carlo sampling is required to obtain the resamples, resulting in significant computational speed gains.

3.2 Fast algorithm for best subset selection

For best subset selection we consider the class of candidate models

$$\mathbb{M}_0 = \{\mathcal{M} : C_j = 0 \quad \forall \quad j \notin \mathcal{S}\} \tag{3.12}$$

and drop n in the subscripts of $\mathcal{M}_n, C_n, \mathcal{S}_n, \theta_n$ (and their variants). Note that all adequate models are strictly adequate here, and their non-zero support contains $\text{supp}(\theta_0)$. In this setup our fast selection algorithm consists of only three stages:

(a) Fit the full model and estimate its e-value,

(b) Replace each covariate by 0 and compute e-value of all such reduced models, and

10

(c) Collect covariates dropping which causes the *e*-value to go down.

To compute the full model, we need to determine the estimable set of predictor indices $\mathcal{S}_*$. In traditional $n > p$ situations, the choice is simple: $\mathcal{S}_* = \{1, \ldots, p\}$. In high-dimnensional cases when $p > n$, we need to ensure that $p' = |\mathcal{S}_*| < n$, so that $\hat{\theta}_*$ (properly centered and scaled) has a unique asymptotic distribution, i.e. satisfies condition (A1). We use fast screening methods available in the literature to achieve this, as implemented previously by Lai et al. (2015). For high-dimensional linear regression we use the Sure Independent Screening (SIS) procedure (Fan and Lv, 2008), while for more complex cases like mixed models, single index models etc., we propose using the model-free screening method of Zhu et al. (2011).

After obtaining $\mathcal{S}_*$ and the full model estimate, for each of the $p' + 1$ models under consideration: the full model and all dropped-predictor models, we generate two sets of resampling weights, use them in (3.11) to obtain bootstrapped coefficient estimates, and finally calculate finite sample approximations of the bootstrap *e*-values in (3.10). This gives us all the components for a sample version of the population level algorithm 1. We present this as Algorithm 2.

Algorithm 2. (Best subset selection using *e*-values and bootstrap)
1. Fix resampling standard deviation τ_n.
2. Obtain bootstrap samples: $\mathcal{T} = \{\hat{\theta}_{1*}, \ldots, \hat{\theta}_{R*}\}$, and $\mathcal{T}_1 = \{\hat{\theta}_{1*}, \ldots, \hat{\theta}_{R_1*}\}$.
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 $1 : p$
$\quad$ For r_1 in $1 : R_1$
$\quad\quad$ Replace j^{th} index of $\hat{\theta}_{*r_1}$ by 0 to get $\hat{\theta}_{r_1, -j}$.
$\quad\quad$ Calculate $\hat{e}_{rn}(\mathcal{M}_{-j}) = \frac{1}{R_1} \sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1, -j}, [\mathcal{T}_1])$.
$\quad\quad$ If $\hat{e}_{rn}(\mathcal{M}_{-j}) < \hat{e}_{rn}(\mathcal{M}_*)$
$\quad\quad\quad$ Set $\hat{\mathcal{S}}_0 \leftarrow \{\hat{\mathcal{S}}_0, j\}$.

Remark Choice of the bootstrap standard deviation τ_n plays a central role in the performance of Algorithm 2. Its intermediate rate of divergence, i.e. $\tau_n \to \infty, \tau_n/a_n \to 0$, is a necessary and sufficient condition for the consistency of the generalized bootstrap procedure (see Chatterjee and Bose (2005) for details), as well as that of the bootstrap approximation of population *e*-values (Theorem 4.2). Variable selection literature has several examples of such quantities in action that result in model selection consistency: for example the penalty term of BIC, or the tuning parameters of adaptive lasso (Zou, 2006) and Dantzig selector (Candes and Tao, 2007). We demonstrate the performance of Algorithm 2 with several choices of τ_n in our simulations, and use validation on a test set to select the optimal τ_n in our data example.

11

4 Theoretical results

We investigate theoretical properties of e-values in this section. Our first results are in the most general setup of adequate and inadequate models. We then focus on strictly adequate models, and finally on $\mathbb{M}_0$, as a consequence deriving our fast algorithm.

4.1 Separation of adequate models

Our first result on the model elicitation process separates inadequate models from adequate models at the population level.

Theorem 4.1. *Assume the conditions (A1)-(A2) on $\hat{\theta}_{*n}$, and that $\mathbb{E}\|\hat{\theta}_{*n}\|^4 < \infty$. For any data-depth function satisfying (B1)-(B4), the following hold as $n \to \infty$:*

1. *For the full model $\mathcal{M}_{*n}$, $e_n(\mathcal{M}_{*n}) \to e_0 < \infty$, where $e_0 = \mathbb{E}D(Y, [Y])$ with $Y \sim \mathcal{E}(0_p, V, G)$,*

2. *For an adequate model $\mathcal{M}_n$, $|e_n(\mathcal{M}_n) - e_n(\mathcal{M}_{*n})| \to 0$,*

3. *For inadequate model $\mathcal{M}_n$, $e_n(\mathcal{M}_n) \to 0$.*

This ensures that for large enough n, it is possible to find some threshold $\epsilon_n \leq e_n(\mathcal{M}_{*n})$ such that all inadequate models have e-values less than the threshold, while e-values for all adequate models fall above it.

For generalized bootstrap approximations of population e-values through (3.10), we have the following equivalent result:

Theorem 4.2. *Assume conditions of Theorem 4.1, and $\mathbb{E}\|\hat{\theta}_{*n}\|^8 < \infty$. Additionally, assume that the resampling weights $\mathbb{W}_{rni}$ are exchangeable random variables satisfying the conditions (3.3)-(3.8). Then, conditional on the data, $e_{rn}(\mathcal{M}_n)$ converges to zero in probability if $\mathcal{M}_n$ is a sequence of inadequate models, and to e_0 if it is a sequence of adequate models.*

Remark An examination of the proofs of the above two results, which follow similar lines, reveals that they do not use all properties of data depths. Specifically, only location and scale invariance is needed in place of the much stronger affine invariance in (B1), and condition (B5) is not needed at all. We also do not use condition (A2) for $\hat{\theta}_{*n}$. Consequently, theorems 4.1 and 4.2 continue to hold for depth-like functions in a diverging-p setup under relaxed versions of conditions (B1)-(B5) and some technical conditions on the energy functions, and possibly have broader implications than variable selection only. We plan to explore this in detail in future work.

4.2 Derivation of algorithm

Our next result is concerned with strictly adequate models, and forms the foundation for the fast model selection algorithm.

Theorem 4.3. *Assume conditions (A1)-(A2), (B1)-(B5). Then, for a finite sequence of strictly adequate models $\mathcal{M}_{1n} \prec \ldots \prec \mathcal{M}_{kn}$ and any finite collection of inadequate models $\mathcal{M}_{k+1,n}, \ldots, \mathcal{M}_{Kn}$, we have*

$$e_n(\mathcal{M}_{1n}) > \ldots > e_n(\mathcal{M}_{kn}) > \max_{j \in \{k+1,\ldots K\}} e_n(\mathcal{M}_{jn})$$

for large enough n.

We set the constant terms C_n to 0 now, and drop the subscript $_n$ from $\mathcal{M}_n, \mathcal{S}_n, C_n, \theta_n$ and their variants. Recall that we have not yet used the actual data generating process as a candidate model. Let us define the *data generating model* as $\mathcal{M}_0 \prec \mathcal{M}_*$ with associated estimable index set $\mathcal{S}_0 = \mathrm{supp}(\theta_0)$ and $C_0 = 0_{p-|\mathcal{S}_0|}$ for $j \notin \mathcal{S}_0$. Then we have the following result.

Corollary 4.1. *Assume the conditions of Theorem 4.3. Consider the collection of candidate models $\mathbb{M}_0 = \{\mathcal{M} : C_j = 0 \quad \forall \quad j \notin \mathcal{S}\}$. Then for large enough n,*

$$\mathcal{M}_0 = \arg \max_{\mathcal{M} \in \mathbb{M}_0} [e_n(\mathcal{M})] \tag{4.1}$$

Thus, within the restricted set of models $\mathbb{M}_0$, the *e*-value maximizes at the true model. However there are still 2^p possible models. This is where the true advantage of using *e*-values comes through. It starts with the following observation.

Corollary 4.2. *Assume the conditions of Corollary 4.1. Consider the models $\mathcal{M}_{-j} \in \mathbb{M}_0$ with $\mathcal{S}_{-j} = \{1, \ldots, p\} \setminus \{j\}$ for $j = 1, \ldots, p$. Then covariate j is a necessary component of $\mathcal{M}_0$, i.e. $\mathcal{M}_{-j}$ is an inadequate model, if and only if*

$$e_n(\mathcal{M}_{-j}) < e_n(\mathcal{M}_*) \tag{4.2}$$

for sufficiently large n.

Dropping an essential predictor from the full model makes the model inadequate, which has very small *e*-value for large enough n (Theorem 4.1), whereas dropping a non-essential predictor increases the *e*-value owing to Theorem 4.3. Thus, *simply collecting those predictors that cause decrease in the e-value on dropping them from the model suffices for variable selection.*

4.3 Model selection consistency

Consistency of bootstrap analogues of all model e-values follows from Theorem 4.2. This, in conjunction with the results from Section 4.2 establishes model selection consistency of the finite sample bootstrap procedure.

Corollary 4.3. *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*

$$\hat{e}_{rn}(\mathcal{M}) = \frac{1}{R_1} \sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1 m}, [\mathcal{T}]) \tag{4.3}$$

where $[\mathcal{T}]$ is the empirical distribution of the corresponding bootstrap samples, and $\hat{\theta}_{r_1 m}$ are obtained using (3.10). Consider the set of predictors $\hat{\mathcal{S}}_0 = \{\hat{e}_{rn}(\mathcal{M}_{-j}) < \hat{e}_{rn}(\mathcal{M}_*)\}$. Then as $n, R, R_1 \to \infty$,

$$P_2(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \to 1 \tag{4.4}$$

where P_2 is the probability conditional on the data and bootstrap samples.

Equation (4.4) is contingent on the fact that the the data generating model $\mathcal{M}_0$ is indeed a member of $\mathbb{M}_0$, i.e. $\mathcal{S}_0 \subseteq \mathcal{S}_*$. This is ensured trivially in classical $n > p$ situations. If p is larger than n, $\mathbb{M}_0$ is the set of all possible models on the predictor set selected by an appropriate predictor screening procedure.

In high-dimensional linear models, we use the SIS procedure for this purpose. Given that $\mathbb{M}_0$ is taken as the set of variables screened by SIS, Fan and Lv (2008) proved under fairly weak conditions that

$$P(\mathcal{M}_0 \in \mathbb{M}_0) \geq 1 - O\left(\frac{\exp[-Cn^{1-2\kappa}]}{\log n}\right) \tag{4.5}$$

where $C > 0$ and κ are constants that depend on the minimum signal in θ_0. For more complex models we use the model-free screening method of Zhu et al. (2011) to obtain the estimable index set $\mathcal{S}_*$. Under mild conditions on the predictor matrix, Theorem 3 in Zhu et al. (2011) ensures the validity of their procedure. In our notations, the result therein states as

$$P(|\mathcal{S}_* \cap \mathcal{S}_0^c| \geq r) \leq \left(1 - \frac{r}{p+d}\right)^d \tag{4.6}$$

for positive integers r and d: $d = p$ being a good practical choice as per the numerical studies in Zhu et al. (2011).

Combining (4.5) or (4.6) with Corollary 4.3 as needed establishes asymptotically accurate recovery of the support set $\mathcal{S}_0$ through Algorithm 2 in Section 3.2. For classical $n > p$ cases this holds for any statistical model satisfying the conditions (A1) and (A2), whereas in high-dimensional situations it is contingent upon additional conditions required by the screening method used.

5 Simulation studies

We now present three simulation studies to examine the empirical performance of our proposed method using model e-values, along with a host of other variable selection methods for linear and linear mixed models. In terms of n and p, we consider both the classical $n > p$ case (Sections 5.1 and 5.3) and the high-dimensional case where $p >> n$ (Section 5.2). For all depth calculations, we use halfspace depth (Tukey, 1975).

5.1 Selecting covariates in linear regression $(n > p)$

We use the model $Y = X\beta_0 + \epsilon$, with $\epsilon \sim \mathcal{N}_n(0, I_n)$, $n = 1000$ and $p = 60$. We generate the rows of X independently from $\mathcal{N}_p(0, \Sigma_X)$, with $(\Sigma_X)_{ij} = \rho^{|i-j|}$. In the true coefficient vector β_0, we set the first 5 coefficients to have non-zero values equal to 1, and the other slope parameters to be zero. We repeat this setup for $\rho \in \{0.5, 0.7, 0.9\}$, and generate 100 independent datasets for each ρ. We use the scale-enhanced Bayesian bootstrap of Example 3.2 with $\tau_n \in \{\log n, n^{0.1}, n^{0.2}, n^{0.3}, n^{0.4}\}$, and resampling Monte Carlo size $R = R_1 = 1000$ for the calculation of bootstrap e-values using (4.3).

Competing methods For comparative evaluation our technique, we consider methods that come from different domains in the model selection literature: (a) The Mixed Integer Optimization (MIO) method of Bertsimas et al. (2016), which is based on advanced discrete optimization methods; (b) LASSO and (c) SCAD-penalized linear regression, which represent the domain of penalized regressions; and (d) Stepwise regression using the BIC and forward selection, which is a conventional favorite.

MIO has two versions: all subset and forward selection. The all subset method was much more computationally intensive than other methods, so we used the forward selection MIO. This returns one model for each of the $p + 1$ possible sparsity levels of the coefficient vector, following which we select the model with lowest BIC. For the penalized regression methods we use ten-fold cross validation to select the optimal tuning parameter.

Performance measures We use model sparsity and prediction error on an independently generated dataset $(y_{\text{test}}, X_{\text{test}})$ using the same setup as the training data to evaluate each of the methods under consideration. For any estimate $\hat{\beta}$, these metrics are defined as

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

These are plausible metrics of model evaluation when the features are highly correlated (Bertsimas et al., 2016).

Choice of tuning parameter The top few rows of Table 5.1 give the average sparsity and PE values of our e-value based method for different values of the bootstrap standard deviation τ_n. Both the metrics take lowest values for $\rho = 0.7$ and $\rho = 0.9$ when $\tau_n = \log n$.

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 5.1: Model sparsity and prediction errors for different choices of τ_n

For $\rho = 0.5$, $\tau_n = n^{0.4}$ has marginally lower values, however it performs much worse than $\tau_n = \log n$ when $\rho = 0.9$.

Comparison with other methods   Figure 5.1 presents boxplots of these performance measures, with $\tau_n = \log n$ assumed for our e-value based method. Our method gives the sparsest models and lowest prediction errors among all methods across all values of ρ. The second sparsest models are given by MIO, although its prediction errors are higher. On the other hand, prediction errors of SCAD are similar to our method, but our models are more sparse. For all methods, higher values of ρ result in slightly sparser models and higher prediction errors.

5.2   High-dimensional linear regression ($p >> n$)

Here we generate the data from the same setup as Section 5.1, but with $n = 60, p = 1000$. To implement the e-values method, we first perform an initial screening of variables using SIS, then apply our best subset selection procedure.

Among several choices of the tuning parameter, $\tau_n = \log n$ and $\tau_n = n^{0.1}$ are good choices (see second half of Table 5.1). The first choice provides sparser models, while the second choice has lower prediction errors as the correlation among features increases.

To compare with the $n > p$ case, we report the results corresponding to $\tau_n = \log n$ in the boxplots of Figure 5.2. As compared to the low-dimensional case in Section 5.1, the e-values method provides much sparser estimates than other methods. MIO always ends up selecting the least parsimonious model with 60 predictors. Our method has good performance in terms of prediction error as well. Stepwise regression is the most unstable among all five methods across all values of ρ and both performance measures used.
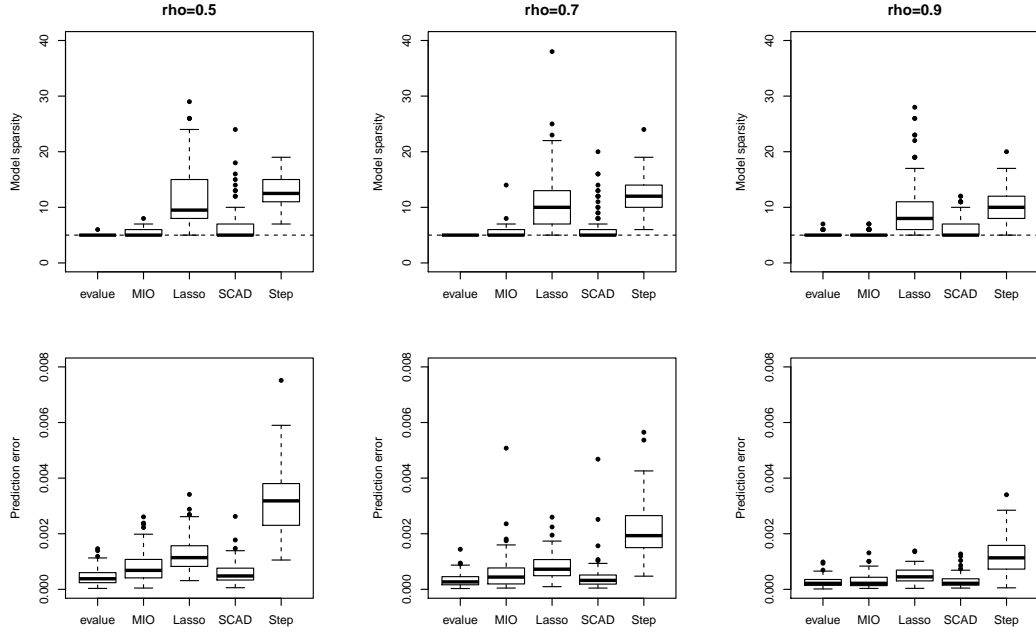
Figure 5.1: Model sparsity (top row) and prediction performance (bottom row) for all methods in $n = 1000, p = 60$ setup. Dotted line in top row indicates true sparsity level.
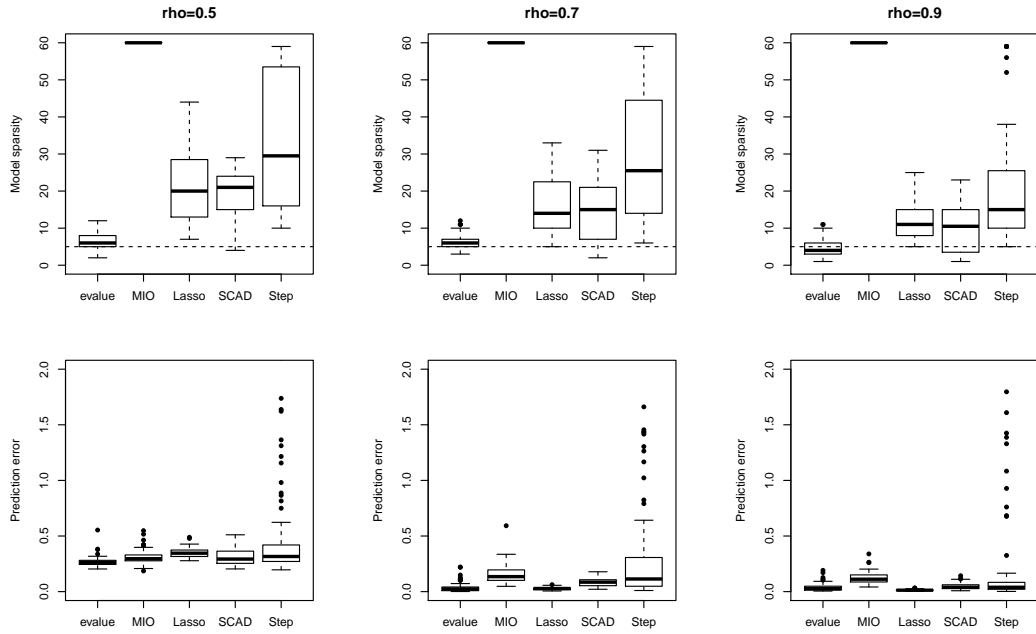

Figure 5.2: Model sparsity (top row) and prediction performance (bottom row) for all methods in $n = 60, p = 1000$ setup.

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		Setting 1: $n_i = 5, m = 30$			Setting 2: $n_i = 10, m = 60$		
e-value based	$\tau = 3$	15.9	0.0	2.59	5.2	0.0	2.17
	$\tau = 4$	8.0	0.0	2.28	2.8	0.0	2.09
	$\tau = 5$	5.2	0.0	2.18	2.0	0.0	2.06
	$\tau = 6$	2.7	0.0	2.09	0.7	0.0	2.02
	$\tau = 7$	2.2	0.0	2.07	0.3	0.0	2.01
	$\tau = 8$	1.5	0.0	2.05	0.3	0.0	2.01
Peng and Lu 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 5.2: Comparison between our method and that proposed by Peng and Lu (2012) through average false positive percentage, false negative percentage and model size

5.3 Best subset selection in presence of random effects

Here we use the simulation setup from Peng and Lu (2012). This is a random intercept-only model: $Y = X\beta + ZU + \epsilon$. The data consists of independent groups of observations with multiple observations in each groups, Z being the within-group random effects design matrix. We consider 9 fixed effects and 4 random effects, with true $\beta_0 = (0, 1, 1, 0, 0, 0, 0, 0, 0)$ and random effect covariance matrix Δ with elements $(\Delta)_{11} = 9$, $(\Delta)_{21} = 4.8$, $(\Delta)_{22} = 4$, $(\Delta)_{31} = 0.6$, $(\Delta)_{32} = (\Delta)_{33} = 1$, $(\Delta)_{4j} = 0; j = 1, \ldots, 4$. The error variance of ϵ is set at $\sigma^2 = 1$. The goal is to select covariates of the fixed effect. We use two scenarios for our study: one where the number of subjects (m) considered is 30, and the number of observations in the ith subject, $i = 1, \ldots, n$, is $n_i = 5$, and another where $m = 60, n_i = 10$.

Given the original restricted maximum likelihood estimates $\hat{\beta}, \hat{\sigma}^2, \hat{\Delta}$, for the resampling step we use the computational approximation

$$\hat{\beta}_r \simeq \hat{\beta} + \frac{\tau_n}{\sqrt{n}}(X^T \hat{V}^{-1} X)^{-1} W_r X^T \hat{V}^{-1}(y - X\hat{\beta})$$

with $W_r = \text{diag}(W_{r1} I_4, \ldots, W_{r9} I_4)$ and $\hat{V} = \hat{\sigma}^2 I_p + Z\hat{\Delta}Z^T$. This adapts equation (3.11) in the mixed model setup. We consider $\tau = \tau_n/\sqrt{n} \in \{1, \ldots, 15\}$ here, and independent Gamma(1,1) random variables as the resampling weights $\{W_{r1} + 1, \ldots, W_{r9} + 1\}$.

We compare our results with those obtained using the mixed effect model selection method of Peng and Lu Peng and Lu (2012), as well as two other methods, all of which are based on sparse estimation of the fixed effect coefficients. In the method proposed by Peng and Lu (2012), the tuning parameter can be selected using several different criteria. We present the false positive percentage (FPR%), false negative percentage (FNR%) and model sizes corresponding to four such criteria in Table 5.2. Our e-value based method handsomely outperforms the method proposed by Peng and Lu (2012), especially in smaller sample sizes, as long as $\tau \geq 4$. We also compare the percentages of times the correct model was identified, and these results are presented in Table 5.3, along with the corresponding results from two other papers. The proposed e-value based

Method		Setting 1	Setting 2
e-value based	$\tau = 4$	79	92
	$\tau = 5$	87	94
	$\tau = 6$	93	98
	$\tau = 7$	94	99
	$\tau = 8$	96	99
	$\tau = 9$	97	99
	$\tau = 10$	98	99
Bondell *et al* Bondell et al. (2010)		73	83
Peng and Lu Peng and Lu (2012)		49	86
Fan and Li Fan and Li (2012)		90	100

Table 5.3: Comparison of our method and three sparsity-based methods of mixed effect model selection through accuracy of selecting correct fixed effects

procedure performs best here for $\tau \geq 5$ for the smaller sample setting, and for $\tau \geq 7$ for larger sample setting.

6   Eliciting drivers of Indian monsoon precipitation

Various studies indicate that our knowledge about the physical drivers of precipitation in India is incomplete; this is in addition to the known difficulties in modeling precipitation itself (Knutti et al., 2010; Trenberth, 2011; Trenberth et al., 2003; Wang et al., 2005). For example, Goswami et al. (2006) discovered an upward trend in frequency and magnitude of extreme rain events, using daily central Indian rainfall data on a $10° \times 12°$ grid, but a similar study on a $1° \times 1°$ gridded data by Ghosh et al. (2009) suggested that there are both increasing and decreasing trends of extreme rainfall events, depending on the location. Additionally, Krishnamurty et al. (2009) reported increasing trends in exceedances of the 99th percentile of daily rainfall; however, there is also a decreasing trend for exceedances of the 90th percentile data in many parts of India. Significant spatial and temporal variabilities at various scales have also been discovered by Dietz and Chatterjee (2014) and Dietz and Chatterjee (2015).

Here we attempt to identify the driving factors behind precipitation during the Indian monsoon season using our e-value based model selection criterion. Data are obtained from the repositories of the National Climatic Data Center (NCDC) and National Oceanic and Atmospheric Administration (NOAA), for the years 1978–2012. We obtained data on 35 potential covariates of the Indian summer precipitation based on existing knowledge and conjectures from the actual physics driving Indian summer precipitations. Details are provided in the supplementary material.

We consider annual medians of all covariates as fixed effects, the log yearly rainfall at a weather station as response variable, and include year-specific random intercepts. We use projection depth (Zuo, 2003) and resample Monte Carlo sizes $R = R_1 = 1000$ to calculate bootstrap equivalents of the e-values. We train our model on data from the years 1978-2002, run best subset selection for tuning parameter values $\tau_n \in n^k; k \in \{0.01, 0.02, \ldots, 0.16\}$, and select the predictor set corresponding to the τ_n that minimizes

| Predictor dropped	e-value of $\mathcal{M}_{-j}$
- TMAX	0.1490772
- X120W	0.2190159
- ELEVATION	0.2288938
- X120E	0.2290021
- del_TT_Deg_Celsius	0.2371846
- X80E	0.2449195
- LATITUDE	0.2468698
- TNH	0.2538924
- Nino34	0.2541503
- X10W	0.2558397
- LONGITUDE	0.2563105
- X100E	0.2565388
- EAWR	0.2565687
- X70E	0.2596766
- v_wind_850	0.2604214
- X140E	0.2609039
- X40W	0.261159
- SolarFlux	0.2624313
- X160E	0.2626321
- EPNP	0.2630901
- TempAnomaly	0.2633658
- u_wind_850	0.2649837
- WP	0.2660394
none	0.2663496
- POL	0.2677756
- TMIN	0.268231
- X20E	0.2687891
- EA	0.2690791
- u_wind_200	0.2692731
- u_wind_600	0.2695297
- SCA	0.2700276
- DMI	0.2700579
- PNA	0.2715089
- v_wind_200	0.2731708
- v_wind_600	0.2748239
- NAO	0.2764488
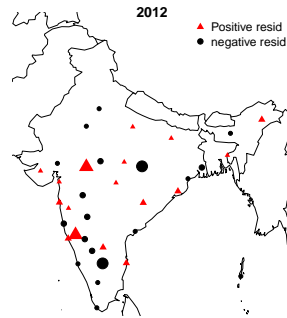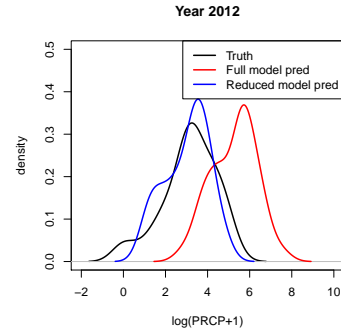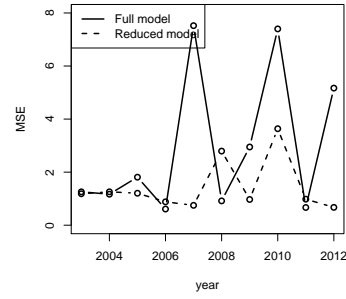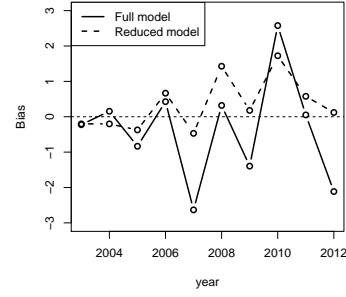
Table 6.1: Ordered e-values when dropping one variable at a time in the Indian summer precipitation data

Figure 6.1: Comparing full model rolling predictions with reduced models: (Top to bottom) Bias across years, Mean Squared Error (MSE) across years, density plots for 2012, stationwise residuals for 2012

fixed effect prediction errors on the samples from 2003–2012. Table 6.1 lists the estimated *e*-values in increasing order for this full model, as well as all 35 models where a single variable is dropped. Variables listed in the rows above *none* in Table 6.1 are considered relevant by our *e*-value criterion.

All variables selected by our procedure have documented effects on Indian monsoon (Krishnamurthy and Kinter III, 2003; Moon et al., 2013). The single largest contributor is the *maximum temperature* variable, whose relation to precipitation based on the Clausius-Clapeyron relation is now classical knowledge in physics (Singleton and Toumi, 2012). It seems that wind velocities high up in the atmosphere are not significant contributors, and the fact that many covariates are selected in the process highlights the complexity of the system.

We checked out-of-sample prediction performance of the estimated model using a rolling validation scheme. For each of the 10 test years: 2003–2012, we select important variables from the model built on past 25 years' data (i.e. use data from 1978–2002 for 2003, 1979-2003 for 2004 and so on), build a model using them and compare predictions on test year obtained from this model with those from the full model. Figure 6.1 summarizes results obtained through this process. Across all testing years, reduced model predictions have less bias as well as are more stable (top two panels). The better approximations of truth by reduced models is also evident from the density plot for 2012 in the third panel, and the residuals do not show any spatial patterns (bottom panel).

7   Conclusion

An immediate goal for future research is incorporating *dimension asymptotics* where the parameter dimension is allowed to grow with the sample size, as well as extremely high-dimensional parameters in our framework. Higher order correctness is a possibility in our context since we use resampling methods. The sensitivity of the results to the choice of the depth functions, resampling algorithm and the tuning parameter τ_n associated with it, and the way $D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}])$ is summarized to obtain the *e*-value deserve further study. Our results suggest excellent *asymptotic* properties that are borne out in the simulations, but finite-sample performance of the procedure needs further scrutiny.

To our knowledge, no other non-penalized method is available in the literature in which computational cost of the covariate selection step scales linearly with dimension of the feature space. This makes the *e*-values a perfect candidate for being an accurate and scalable model selection method in big data. Because our algorithm easily lends itself to parallel computing across different sets of bootstrap samples, it opens up the possibility to be combined with bootstrap methods on subsamples of the full dataset, like the *moon*-bootstrap, bag of little bootstraps (Kleiner et al., 2014), or subsampled double bootstrap (Sengupta et al., 2016). For ensemble methods in the machine learning literature, e.g. random forest and bagging, by extension of the same logic *e*-values present an exciting opportunity to explore the formulation of a method for consistent feature ranking.

A   Proofs of theoretical results

Proof of Theorem 4.1. Part 1 follows directly from assumption (B3).

Part 2. Assuming now that $\mathcal{M}_n$ is an adequate model, we use (B1) property of D:

$$D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}]) = D\Big(\hat{\theta}_{mn} - \theta_{0n}, \big[\hat{\theta}_{0n} - \theta_{0n}\big]\Big), \tag{A.1}$$

and decompose the first argument

$$\hat{\theta}_{mn} - \theta_{0n} = (\hat{\theta}_{mn} - \hat{\theta}_{0n}) + (\hat{\theta}_{*n} - \theta_{0n}). \tag{A.2}$$

where θ_{mnj} equals θ_{0nj} in indices $j \in \mathcal{S}_n$ and C_{nj} elsewhere.

Now we have, for any $\mathcal{M}_n$,

$$\hat{\theta}_{mn} = \theta_{mn} + a_n^{-1} T_{mn}$$

where T_{mn} is non-degenerate at the $\mathcal{S}_n$ indices. For the first summand of the right-hand side in (A.2) we get

$$\hat{\theta}_{mn} - \hat{\theta}_{*n} = \theta_{mn} - \theta_{0n} + R_n \tag{A.3}$$

where $\mathbb{E}\|R_n^2\| = O(a_n^{-2})$. Since $\mathcal{M}_n$ is an adequate model, $\theta_{mn} - \theta_{0n} = o(1)$. Thus, substituting the above right-hand side in (A.2) we get

$$\left| D\left(\hat{\theta}_{mn} - \theta_{0n}, \big[\hat{\theta}_{*n} - \theta_{0n}\big]\right) - D\left(\hat{\theta}_{*n} - \theta_{0n}, \big[\hat{\theta}_{*n} - \theta_{0n}\big]\right) \right| \leq \|R_n\|^\alpha \tag{A.4}$$

from Lipschitz continuity of $D(\cdot)$ given in (B2). The result now follows.

Part 3. Since the depth function E is invariant under location and scale transformations, we have

$$D(\hat{\theta}_{mn}, [\hat{\theta}_{*n}]) = D\Big(a_n(\hat{\theta}_{mn} - \theta_{0n}), \big[a_n(\hat{\theta}_{*n} - \theta_{0n})\big]\Big). \tag{A.5}$$

Decomposing the first argument,

$$a_n(\hat{\theta}_{mn} - \theta_{0n}) = a_n(\hat{\theta}_{mn} - \theta_{mn}) + a_n(\theta_{mn} - \theta_{0n}). \tag{A.6}$$

Since $\mathcal{M}_n$ is inadequate, given $\delta > 0$ there exists a subsequence indexed by j_n such that $|C_{nj_n} - \theta_{0j_n}| > \delta$ for all n. The result follows by application of condition (B4). $\qquad\square$

Proof of Theorem 4.2. Several derivations here are similar to those of the proof of Theorem 4.1. Consequently we skip some details.

We have

$$D\left(\hat{\theta}_{r_1 mn}, [\hat{\theta}_{r*n}]\right) = D\left(\frac{a_n}{\tau_n}\left(\hat{\theta}_{r_1 mn} - \hat{\theta}_{*n}\right), \left[\frac{a_n}{\tau_n}\left(\hat{\theta}_{r*n} - \hat{\theta}_{*n}\right)\right]\right)$$

$$= D\left(\frac{a_n}{\tau_n}\left(\hat{\theta}_{r_1 mn} - \hat{\theta}_{mn}\right) - \frac{a_n}{\tau_n}\left(\hat{\theta}_{mn} - \hat{\theta}_{*n}\right), \left[\frac{a_n}{\tau_n}\left(\hat{\theta}_{r*n} - \hat{\theta}_{*n}\right)\right]\right).$$

Conditional on the data, $(a_n/\tau_n)(\hat\theta_{r_1 mn} - \hat\theta_{mn})$ has the same weak limit as $a_n(\hat\theta_{mn} - \theta_{mn})$, and (A.3) and $\tau_n \to \infty$ combine to give

$$\frac{a_n}{\tau_n}\left(\hat\theta_{mn} - \hat\theta_{*n}\right) \xrightarrow{P} 0$$

as $n \to \infty$. The results follow directly now. □

Proof of Theorem 4.3. Since we are dealing with a finite sequence of nested models, it is enough to prove that $e_n(\mathcal{M}_{1n}) > e_n(\mathcal{M}_{2n})$ for large enough n, when both $\mathcal{M}_{1n}$ and $\mathcal{M}_{2n}$ are strictly adequate models and $\mathcal{M}_{1n} \prec \mathcal{M}_{2n}$.

Suppose $\mathbb{T}_0 = \mathcal{E}(0_p, I_p, g)$. Affine invariance implies invariant to rotational transformations, and since the evaluation functions we consider decrease along any ray from the origin because of (B5), $E(\theta, \mathbb{T}_0)$ is a monotonocally decreasing function of $\|\theta\|$ for any $\theta \in \mathbb{R}^p$. Now consider the models $\mathcal{M}_{1n}^0, \mathcal{M}_{2n}^0$ that have 0 in all indices outside $\mathcal{S}_{1n}$ and $\mathcal{S}_{2n}$, respectively. Take some $\theta_{10} \in \Theta_{1n}^0$, which is the parameter space corresponding to $\mathcal{M}_{1n}^0$, and replace its (zero) entries at indices $j \in \mathcal{S}_{2n} \setminus \mathcal{S}_{1n}$ by some non-zero $\delta \in \mathbb{R}^{p - |\mathcal{S}_{2n} \setminus \mathcal{S}_{1n}|}$. Denote it by $\theta_{1\delta}$. Then we shall have

$$\begin{aligned}
\theta_{1\delta}^T \theta_{1\delta} > \theta_{10}^T \theta_{10} &\Rightarrow& D(\theta_{10}, \mathbb{T}_0) > D(\theta_{1\delta}, \mathbb{T}_0) \\
&\Rightarrow& \mathbb{E}_{s1} D(\theta_{10}, \mathbb{T}_0) > \mathbb{E}_{s1} D(\theta_{1\delta}, \mathbb{T}_0)
\end{aligned}$$

where $\mathbb{E}_{s1}$ denotes the expectation taken over the marginal of the distributional argument $\mathbb{T}_0$ at indices $\mathcal{S}_{1n}$. Notice now that by construction $\theta_{1\delta} \in \Theta_{2n}^0$, the parameter space corresponding to $\mathcal{M}_{2n}^0$, and since the above holds for all possible δ, we can take expectation over indices $\mathcal{S}_{2n} \setminus \mathcal{S}_{1n}$ in both sides to obtain $\mathbb{E}_{s1} D(\theta_{10}, \mathbb{T}_0) > \mathbb{E}_{s2} D(\theta_{20}, \mathbb{T}_0)$, with θ_{20} denoting a general element in Θ_{20}.

Combining (A1) and (A2) we get $a_n V_n^{-1/2}(\hat\theta_{*n} - \theta_{0n}) \rightsquigarrow \mathbb{T}_0$. Denote $\mathbb{T}_n = [a_n V_n^{-1/2}(\hat\theta_{*n} - \theta_{0n})]$, and choose a positive $\epsilon < (\mathbb{E}_{s1} D(\theta_{10}, \mathbb{T}_0) - \mathbb{E}_{s2} D(\theta_{20}, \mathbb{T}_0))/2$. Then, for large enough n we shall have

$$|D(\theta_{10}, \mathbb{T}_n) - D(\theta_{10}, \mathbb{T}_0)| < \epsilon \quad \Rightarrow \quad |\mathbb{E}_{s1} D(\theta_{10}, \mathbb{T}_n) - \mathbb{E}_{s1} D(\theta_{10}, \mathbb{T}_0)| < \epsilon$$

following condition (B4). Similarly we have $|\mathbb{E}_{s2} D(\theta_{20}, \mathbb{T}_n) - \mathbb{E}_{s2} D(\theta_{20}, \mathbb{T}_0)| < \epsilon$ for the same n for which the above holds. This implies $\mathbb{E}_{s1} D(\theta_{10}, \mathbb{T}_n) > \mathbb{E}_{s2} D(\theta_{20}, \mathbb{T}_n)$.

Now apply the affine transformation $t(\theta_n) = V_n^{1/2}\theta_n/a_n + \theta_{0n}$ to both arguments of the depth function above. This will keep the depths constant following affine invariance, i.e. $D(t(\theta_{10}), [\hat\theta_{*n}]) = D(\theta_{10}, \mathbb{T}_n)$ and $D(t(\theta_{20}), [\hat\theta_{*n}]) = D(\theta_{20}, \mathbb{T}_n)$. Since this transformation maps Θ_{1n}^0 to Θ_{1n}, the parameter space corresonding to $\mathcal{M}_{1n}$, we get $\mathbb{E}_{s1} D(t(\theta_{10}), [\hat\theta_{*n}]) > \mathbb{E}_{s2} D(t(\theta_{20}), [\hat\theta_{*n}])$, i.e. $e_n(\mathcal{M}_{1n}) > e_n(\mathcal{M}_{2n})$.

For any inadequate model $\mathcal{M}_{jn}, k < j \leq K$, suppose N_{jn} is the integer such that $e_{n_1}(\mathcal{M}_{jn_1}) < e_{n_1}(\mathcal{M}_{*n_1})$ for all $n_1 > N_{jn}$. Part 3 of Theorem 4.1 ensures that such an integer exists for every inadequate model. Now define $N = \max_{k < j \leq K} N_{jn}$. Thus $e_{n_1}(\mathcal{M}_{*n_1})$ is larger than e-values of all inadequate models $\mathcal{M}_{jn_1}$ for $k < j \leq K$. □

Proof of Corollary 4.1. By construction, $\mathcal{M}_0$ is the unique minimal adequate model in $\mathbb{M}_0$, and should be nested in all other adequate models therein. Hence Theorem 4.3 implies $e_n(\mathcal{M}_0) > e_n(\mathcal{M}^{ad})$ for any adequate model $\mathcal{M}^{ad} \in \mathbb{M}_0$ and large enough n. $\square$

Proof of Corollary 4.2. Consider $j \in \mathcal{S}_0$. Then $\theta_0 \notin \mathcal{M}_{-j}$, hence $\mathcal{M}_{-j}$ is inadequate. By choice of n_1 from Corollary 4.1, e-values of all inadequate models are less than that of $\mathcal{M}_*$, hence $e_{n_1}(\mathcal{M}_{-j}) < e_{n_1}(\mathcal{M}_*)$.

On the other hand, suppose there exists a j such that $e_{n_1}(\mathcal{M}_{-j}) \leq e_{n_1}(\mathcal{M}_*)$ but $j \notin \mathcal{S}_0$. Now $j \notin \mathcal{S}_0$ means that $\mathcal{M}_{-j}$ is an adequate model. Since $\mathcal{M}_{-j}$ is nested within $\mathcal{M}_*$ for any j, and the full model is always adequate, we have $e_{n_1}(\mathcal{M}_{-j}) > e_{n_1}(\mathcal{M}_*)$ by Theorem 4.3: leading to a contradiction and thus completing the proof. $\square$

Proof of Corollary 4.3. Corollary 4.2 implies that

$$\mathcal{S}_0 = \{j : e_n(\mathcal{M}_{-j}) < e_n(\mathcal{M}_*)\}$$

Now define $\bar{\mathcal{S}}_0 = \{j : e_{rn}(\mathcal{M}_{-j}) < e_{rn}(\mathcal{M}_*)\}$. Theorem 4.2 implies that

$$e_{rn}(\mathcal{M}_{-j}) = e_n(\mathcal{M}_{-j}) + R_{jn}$$
$$e_{rn}(\mathcal{M}_*) = e_n(\mathcal{M}_*) + R_{*n}$$

such that $\mathbb{E}_r|R_{*n}|^2 = o_P(1)$ and $\mathbb{E}_r|R_{jn}|^2 = o_P(1)$ for all j. Hence $P_1(\bar{\mathcal{S}}_0 = \mathcal{S}_0) \to 1$ as $n \to \infty$, P_1 being probability conditional on the data. Similarly one can prove that the probability conditional on the bootstrap samples that $\bar{\mathcal{S}}_0 = \hat{\mathcal{S}}_0$ holds goes to 1 as $R, R_1 \to \infty$, which completes the proof. $\square$

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22(1):203–217.

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

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Bose, A. and Chatterjee, S. (2003). Generalized bootstrap for estimators of minimizers of convex functions. *J. Stat. Plan. Inference*, 117(2):225–239.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.*, 33(1):414–436.

Dietz, L. and Chatterjee, S. (2014). Logit-normal mixed model for Indian monsoon precipitation. *Nonlin. Proc. Geophys.*, 21:934–953.

Dietz, L. and Chatterjee, S. (2015). Extreme thresholds in Indian monsoon precipitation using logit-Normal mixed models. In *Machine Learning and Data Mining Approaches to Climate Science*, pages 239–246. Springer, New York, NY, USA.

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *J. R. Statist. Soc. B*, 70:849–911.

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

Fang, K. T., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions. Monographs on Statistics and Applied Probability*, volume 36. Chapman and Hall Ltd., London.

Ghosh, S., Luniya, V., and Gupta, A. (2009). Trend analysis of indian summer monsoon rainfall at different spatial scales. *Atmos. Sci. Lett.*, 10:285–290.

Goswami, B., Venugopal, V., Sengupta, D., Madhusoodanan, M. S., and Xavier, P. K. (2006). Increasing trend of extreme rain events over india in a warming environment. *Science*, 314:1442–1445.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *Ann. Statist.*, 36(4):1669–1692.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *J. R. Statist. Soc. B*, 76(4):765–814.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *J. Clim.*, 23(10):2739–2758.

Krishnamurthy, V. and Kinter III, J. L. (2003). The Indian monsoon and its relation to global climate variability. In Rodo, X. and Comin, F. A., editors, *Global Climate: Current Research and Uncertainties in the Climate System*, pages 186–236. Springer.

Krishnamurty, C. K. B., Lall, U., and Kwon, H.-H. (2009). Changing frequencey and intensity of rainfall extremes over india from 1951 to 2003. *J. Clim.*, 22(18):4737–4746.

Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015). Generalized Fiducial Inference for Ultrahigh-Dimensional Regression. *J. Amer. Statist. Assoc.*, 110(510):760–772.

Liu, R. and Singh, K. (1993). A Quality Index Based on Data Depth and Multivariate Rank Tests. *J. Amer. Statist. Assoc.*, 88:252–260.

Meza, J. and Lahiri, P. (2005). A note on the C_p statistic under the nested error regression model. *Surv. Methodol.*, 31:105–109.

Michel, R. and Pfanzagl, J. (1971). The accuracy of the normal approximation for minimum contrast estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 18(1):73–84.

Moon, J.-Y., Wang, B., Ha, K.-J., and Lee, J.-Y. (2013). Teleconnections associated with Northern Hemisphere summer monsoon intraseasonal oscillation. *Clim. Dyn.*, 40(11-12):2761–2774.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and Complex Data Structures*, pages 17–34. Springer Berlin Heidelberg.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42:789–817.

Natarajan, B. K. (1995). Sparse Approximate Solutions to Linear Systems. *Siam. J. Comput.*, 24:227–234.

Nguyen, T. and Jiang, J. (2014). Restricted fence method for covariate selection in longitudinal data analysis. *Biostatistics*, 13(2):303–314.

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

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, 14(1):249–272.

Ronchetti, E. (2000). *Data Segmentation and Model Selection for Computer Vision: A Statistical Approach*, chapter Robust Regression Methods and Model Selection. Springer. pp 31–40.

Ročkovà, V. and George, E. I. (2016). The Spike-and-Slab LASSO. *J. Amer. Statist. Assoc.*, 0:0–0.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Sengupta, S., Volgushev, S., and Shao, X. (2016). A Subsampled Double Bootstrap for Massive Data. *J. Amer. Statist. Assoc.*, 111(515):1222–1232.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434):655–665.

Singleton, A. and Toumi, R. (2012). Super-Clausius-Clapeyron scaling of rainfall in a model squall line. *Quat. J. R. Met. Soc.*, 139(671):334–339.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(267–288).

Trenberth, K. E. (2011). Changes in precipitation with climate change. *Clim. Res.*, 47(1-2):123–138.

Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B. (2003). The changing character of precipitation. *Bull. Am. Meteorol. Soc.*, 84(9):1205–1217.

Tukey, J. W. (1975). Mathematics and picturing data. In James, R., editor, *Proceedings of the International Congress on Mathematics*, volume 2, pages 523–531.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *J. Amer. Statist. Assoc.*, 93:120–131.

Wang, B., Ding, Q., Fu, X., Kang, I.-S., Jin, K., Shukla, J., and Doblas-Reyes, F. (2005). Fundamental challenge in simulation and prediction of summer monsoon rainfall. *Geophys. Res. Lett.*, 32(15).

Wang, L., Kim, Y., and Li, R. (2013). Calibrating Nonconvex Penalized Regression in Ultra-high Dimension. *Ann. Statist.*, 41:2505–2536.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B*, 76(1):217–242.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-Free Feature Screening for Ultrahigh-Dimensional Data. *J. Amer. Statist. Assoc.*, 106(496):1464–1475.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.*, 101:1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36:1509–1533.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.*, 31:1460–1490.

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