Feature Selection using e-values

Subhabrata Majumdar [1 2] **Snigdhansu Chatterjee** [1]

Abstract

In the context of supervised parametric models, we introduce the concept of *e-values*. An e-value is a scalar quantity that represents the proximity of the sampling distribution of parameter estimates in a model trained on a subset of features to that of the model trained on all features (i.e. the full model). Under general conditions, a rank ordering of e-values separates models that contain all essential features from those that do not.

The e-values are applicable to a wide range of parametric models. We use data depths and a fast resampling-based algorithm to implement a feature selection procedure using e-values, providing consistency results. For a p-dimensional feature space, this procedure requires fitting only the full model and evaluating $p + 1$ models, as opposed to the traditional requirement of fitting and evaluating 2^p models. Through experiments across several model settings and synthetic and real datasets, we establish that the e-values method as a promising general alternative to existing model-specific methods of feature selection.

1. Introduction

In the era of big data, feature selection in supervised statistical and machine learning (ML) models helps cut through the noise of superfluous features, provides storage and computational benefits, and gives model intepretability. Model-based feature selection can be divided into two categories (Guyon & Elisseeff, 2003): wrapper methods that evaluate models trained on multiple feature sets (Schwarz, 1978; Shao, 1996), and embedded methods that combine feature selection and training, often through sparse regularization (Tibshirani, 1996; Zou, 2006; Zou & Hastie, 2005). Both categories are extremely well-studied for independent data models, but have their own challenges. Navigating an exponentially growing feature space using wrapper methods is NP-hard (Natarajan, 1995), and case-specific search strategies like k-greedy, branch-and-bound, simulated annealing are needed. Sparse penalized embedded methods can tackle high-dimensional data, but have inferential and algorithmic issues, such as biased Lasso estimates (Zhang & Zhang, 2014) and the use of convex relaxations to compute approximate local solutions (Wang et al., 2013; Zou & Li, 2008).

Feature selection in dependent data models has received comparatively lesser attention. Existing implementations of wrapper and embedded methods have been adapted for dependent data scenarios, such as mixed effect models (Meza & Lahiri, 2005; Nguyen & Jiang, 2014; Peng & Lu, 2012) and spatial models (Huang et al., 2010; Lee & Ghosh, 2009). However these suffer from the same issues as their independent data counterparts. If anything, the higher computational overhead for model training makes implementation of wrapper methods even harder in this situation!

In this paper, we propose the framework of *e-values* as a common principle to perform best subset feature selection in a range of parametric models covering independent and dependent data settings. In essence ours is a wrapper method, although it is able to determine important features affecting the response *by training a single model*: the one with all features, or the full model. We achieve this by utilizing the information encoded in the *distribution* of model parameter estimates. Notwithstanding recent efforts (Lai et al., 2015; Singh et al., 2007), parameter distributions that are fully data-driven have generally been underutilized in data science. In our work, e-values score a candidate model to quantify the similarity between sampling distributions of parameters in that model and the full model. Sampling distribution is the distribution of a parameter estimate, based on the random data samples the estimate is calculated from. We access sampling distributions using the efficient Generalized Bootstrap technique (Chatterjee & Bose, 2005), and utilize them using data depths, which are essentially point-to-distribution inverse distance functions (Tukey, 1975; Zuo, 2003; Zuo & Serfling, 2000), to compute e-values.

How e-values perform feature selection Data depth functions quantify the inlyingness of a point in multivariate space with respect to a probability distribution or a multivariate

[1]School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN, USA [2]Currently at Splunk. Correspondence to: Subhabrata Majumdar <smajumdar@splunk.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

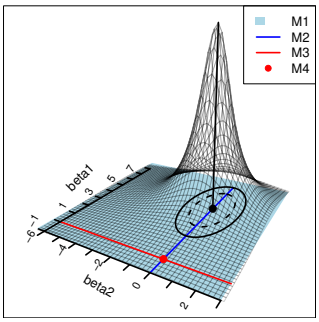


Figure 1.1. The e-value method. Solid and dotted ellipses represent Mahalanobis depth contours at some $\alpha > 0$ for sample sizes $n_1, n_2; n_2 > n_1$.

point cloud, and have seen diverse applications in the literature (Jornsten, 2004; Narisetty & Nair, 2016; Rousseeuw & Hubert, 1998). As an example, consider the *Mahalanobis depth*, defined as

$$D(x, F) = [1 + (x - \mu_F)^\top \Sigma_F^{-1}(x - \mu_F)]^{-1},$$

for $x \in \mathbb{R}^p$ and F a p-dimensional probability distribution with mean μ_F and covariance matrix Σ_F. When x is close to μ_F, $D(x, F)$ takes a high value close to 1. On the other hand, as $\|x\| \to \infty$, the depth at x approaches 0. Thus, $D(x, F)$ quantifies the proximity of x to F. Points with high depth are situated 'deep inside' the probability distribution F, while low depth points sit at the periphery of F.

Given a depth function D we define e-value as the mean depth of the finite-sample estimate of a parameter β for a candidate model $\mathcal{M}$, with respect to the sampling distribution of the full model estimate $\hat{\beta}$:

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

where $\hat{\beta}_{\mathcal{M}}$ is the estimate of β assuming model $\mathcal{M}$, $[\hat{\beta}]$ is the sampling distribution of $\hat{\beta}$, and $\mathbb{E}(\cdot)$ denotes expectation. For large sample size n, the index set $\mathcal{S}_{select}$ obtained by Algorithm 1 below elicits all non-zero features in the true parameter. We use bootstrap to obtain multiple copies of $\hat{\beta}$ and $\hat{\beta}_{-j}$ for calculating the e-values in steps 1 and 5.

Algorithm 1 Best subset selection using e-values

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

As an example, consider a linear regression with two features, Gaussian errors $\epsilon \sim N(0, 1)$, and the following candidate models (Figure 1.1). We denote by Θ_i the domain of parameters considered in model $\mathcal{M}_i; i = 1, \ldots, 4$.

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

Let $\beta_0 = (5, 0)^T$ be the true parameter. The full model sampling distribution, $\hat{\beta} \sim \mathcal{N}(\beta_0, (X^T X)^{-1})$, is more concentrated around β_0 for higher sample sizes. Thus the depths at points along the (red) line $\beta_1 = 0$, and the origin (red point), become smaller, and mean depths approach 0 for $\mathcal{M}_3$ and $\mathcal{M}_4$. On the other hand, since depth functions are affine invariant (Zuo & Serfling, 2000), mean depths calculated over parameter spaces for $\mathcal{M}_1$ (blue line) and $\mathcal{M}_2$ (blue surface) remain the same, and do not vanish as $n \to \infty$. Thus, e-values of the 'good' models $\mathcal{M}_1, \mathcal{M}_2$—models the parameter spaces of which contain β_0—separate from those of the 'bad' models $\mathcal{M}_3, \mathcal{M}_4$ more and more as n grows. Algorithm 1 is the result of this separation, and a rank ordering of the 'good' models based on how parsimonious they are (Theorem 5.1).

2. Related work

Feature selection is a widely studied area in statistics and ML. The vast amount of literature on this topic includes classics like the Bayesian Information Criterion (Schwarz, 1978, BIC) or Lasso (Tibshirani, 1996), and recent advances such as Mixed Integer Optimization (Bertsimas et al., 2016, MIO). To deal with the increasing scale and complexity of data in recent applications, newer streams of work have also materialized—such as nonlinear and model-independent methods (Song et al., 2012), model averaging (Fragoso et al., 2018), and dimension reduction techniques (Ma & Zhu, 2013). We refer the reader to Guyon & Elisseeff (2003) and the literature review in Bertsimas et al. (2016) for a broader overview of feature selection methods, and focus on the three lines of work most related to our formulation.

Shao (1996) first proposed using bootstrap for feature selection, with the squared residual averaged over a large number of resampled parameter estimates from a m-out-of-n bootstrap as selection criterion. Leave-One-Covariate-Out (Lei et al., 2018, LOCO) is based on the idea of sample-splitting. The full model and leave-one-covariate-out models are trained using one part of the data, and the rest of the data are used to calculate the importance of a feature as median difference in predictive performance between the full model and a LOCO model. Finally, Barber & Candes (2015) introduced the powerful idea of Knockoff filters, where a 'knockoff' version of the original input dataset imitating its correlation structure is constructed. This method is explicitly able to control False Discovery Rate.

Our framework of e-values has some similarities with the above methods, such as the use of bootstrap (similar to Shao (1996)), feature-specific statistics (similar to LOCO and Knockoffs), and evaluation at $p+1$ models (similar to LOCO). However, e-values are also significantly different. They do not suffer from the high computational burden of model refitting over multiple bootstrap samples (unlike Shao (1996)), are not conditional on data splitting (unlike LOCO), and have a very general theoretical formulation (unlike Knockoffs). Most importantly, unlike all the three methods discussed above, e-values require fitting only a single model, and work for dependent data models.

Previously, VanderWeele & Ding (2017) have used the term 'E-Value' in the context of sensitivity analysis. However, our and their definition of e-values are quite different. We see our e-values as a means to *evaluate* a feature with reference to a model. There are some parallels to the well known p-values used for hypothesis testing, but we see the e-value as a more general quantity that covers dependent data situations, as well as general estimation and hypothesis testing problems. While the current paper is an application of e-values for feature selection, we plan to build up on the framework introduced here on other applications, including group feature selection, hypothesis testing, and multiple testing problems.

3. Preliminaries

For a positive integer $n \geq 1$, Let $\mathcal{Z}_n = \{Z_1, \ldots, Z_n\}$ be an observable array of random variables that are not necessarily independent. We parametrize $\mathcal{Z}_n$ using a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$, and energy functions $\{\psi_i(\theta, Z_i) : 1 \leq i \leq n\}$. We assume that there is a true unknown vector of parameters θ_0, which is the unique minimizer of $\Psi_n(\theta) = \mathbb{E}\sum_{i=1}^n \psi_i(\theta, Z_i)$.

Models and their characterization Let $\mathcal{S}_* = \cup_{\theta\in\Theta} \text{supp}(\theta)$ be the common non-zero support of all estimable parameters θ (denoted by $\text{supp}(\theta)$). In this setup, we associate a model $\mathcal{M}$ with two quantities (a) The set of indices $\mathcal{S} \subseteq \mathcal{S}_*$ with unknown and estimable parameter values, and (b) an ordered vector of known constants $C = (C_j : j \notin \mathcal{S})$ at indices not in $\mathcal{S}$. Thus a generic parameter vector for the model $\mathcal{M} \equiv (\mathcal{S}, C)$, denoted by $\theta_m \in \Theta_m \subseteq \Theta = \prod_j \Theta_j$, consists of unknown indices and known constants

$$\theta_{mj} = \begin{cases} \text{unknown } \theta_{mj} \in \Theta_j & \text{for } j \in \mathcal{S}, \\ \text{known } C_j \in \Theta_j & \text{for } j \notin \mathcal{S}. \end{cases} \tag{3.1}$$

Given the above formulation, we characterize a model.

Definition 3.1. A model $\mathcal{M}$ is called *adequate* if $\sum_{j\notin\mathcal{S}} |C_j - \theta_{0j}| = 0$. A model that is not adequate, is called an *inadequate* model.

By definition the full model is always adequate, as is the model corresponding to the singleton set containing the true parameter. Thus the set of adequate models is non-empty by construction.

Another important notion is the one of *nested models*.

Definition 3.2. We consider a model $\mathcal{M}_1$ to be nested in $\mathcal{M}_2$, notationally $\mathcal{M}_1 \prec \mathcal{M}_2$, if $\mathcal{S}_1 \subset \mathcal{S}_2$ and C_2 is a subvector of C_1.

If a model is adequate, then any model it is nested in is also adequate. In the context of feature selection this obtains a rank ordering: the most parsimonious adequate model, with $\mathcal{S} = \text{supp}(\beta_0)$ and $C = 0_{p-|\mathcal{S}|}$, is nested in all other adequate models. All models nested *within* it are inadequate.

Estimators The estimator $\hat{\theta}_*$ of θ_0 is obtained as a minimizer of the sample analog of $\Psi_n(\theta)$:

$$\hat{\theta}_* = \arg\min_\theta \Psi_n(\theta) = \arg\min_\theta \sum_{i=1}^n \psi_i\big(\theta, Z_i\big). \tag{3.2}$$

Under standard regularity conditions on the energy functions, $a_n(\hat{\theta}_* - \theta_*)$ converges to a p-dimensional Gaussian distribution as $n \to \infty$, where $a_n \uparrow \infty$ is a sequence of positive real numbers (Appendix A).

Remark 3.3. For generic estimation methods based on likelihood and estimating equations, the above holds with $a_n \equiv n^{1/2}$, resulting in the standard 'root-n' asymptotics.

The estimator in (3.2) corresponds to the *full model* $\mathcal{M}_*$, i.e. the model where all indices in $\mathcal{S}_*$ are estimated. For any other model $\mathcal{M}$, we simply augment entries of $\hat{\theta}_*$ at indices in $\mathcal{S}$ with elements of C elsewhere to obtain a model-specific coefficient estimate $\hat{\theta}_m$:

$$\hat{\theta}_{mj} = \begin{cases} \hat{\theta}_{*j} & \text{for } j \in \mathcal{S}, \\ C_j & \text{for } j \notin \mathcal{S}. \end{cases} \tag{3.3}$$

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

Depth functions Data depth functions (Zuo & Serfling, 2000) quantify the closeness of a point in multivariate space to a probability distribution or data cloud. Formally, let $\mathcal{G}$ denote the set of non-degenerate probability measures on $\mathbb{R}^p$. We consider $D : \mathbb{R}^p \times \mathcal{G} \to [0, \infty)$ to be a data depth function if it satisfies the following properties:

(B1) *Affine invariance*: 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) *Lipschitz continuity*: For any $\mathbb{G} \in \mathcal{G}$, 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) Consider random variables $Y_n \in \mathbb{R}^p$ such that $[Y_n] \rightsquigarrow \mathbb{Y} \in \mathcal{G}$. Then $D(y,[Y_n])$ converges uniformly to $D(y,\mathbb{Y})$. So that, if $Y \sim \mathbb{Y}$, then $\lim_{n\to\infty} \mathbb{E}D(Y_n,[Y_n]) = \mathbb{E}D(Y,\mathbb{Y}) < \infty$.

(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 $\mu(\mathbb{G}) \in \mathbb{R}^p$, $D(.,\mathbb{G})$ is maximum at $\mu(\mathbb{G})$:

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

Depth decreases along any ray between $\mu(\mathbb{G})$ to x, i.e. for $t \in (0,1), x \in \mathbb{R}^p$,

$$\begin{aligned} D(x,\mathbb{G}) &< D(\mu(\mathbb{G}) + t(x-\mu(\mathbb{G})),\mathbb{G}) \\ &< D(\mu(\mathbb{G}),\mathbb{G}). \end{aligned}$$

Conditions (B1), (B4) and (B5) are integral to the formal definition of data depth (Zuo & Serfling, 2000), while (B2) and (B3) implicitly arise for several depth functions (Mosler, 2013). We require only a subset of (B1)-(B5) for the theory in this paper, but use data depths throughout for simplicity.

4. The e-values framework

The *e-value* of model $\mathcal{M}$ is the mean depth of $\hat{\theta}_\mathcal{M}$ with respect to $[\hat{\theta}]$: $e_n(\mathcal{M}) = \mathbb{E}D(\hat{\theta}_m, [\hat{\theta}_*])$.

4.1. Resampling approximation of e-values

Typically, the distributions of either of the random quantities $\hat{\theta}_m$ and $\hat{\theta}_*$, are unknown, and have to be elicited from the data. Because of the plugin method in (3.3), only $[\hat{\theta}_*]$ needs to be approximated. We do this using Generalized Bootstrap (Chatterjee & Bose, 2005, GBS). For an exchangeable array of non-negative random variables independent of the data as resampling weights: $\mathcal{W}_r = (\mathbb{W}_{r1},\ldots,\mathbb{W}_{rn})^T \in \mathbb{R}^n$, the GBS estimator $\hat{\theta}_{r*}$ is the minimizer of

$$\Psi_{rn}(\theta) = \sum_{i=1}^n \mathbb{W}_{ri}\psi_i\big(\theta, Z_i\big). \tag{4.1}$$

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

$$\begin{aligned} &\mathbb{E}\mathbb{W}_{r1} = 1, \mathbb{V}\mathbb{W}_{r1} = \tau_n^2 = o(a_n^2) \uparrow \infty, \mathbb{E}W_{r1}^4 < \infty, \\ &\mathbb{E}W_{r1}W_{r2} = O(n^{-1}), \mathbb{E}W_{r1}^2W_{r2}^2 \to 1. \end{aligned} \tag{4.2}$$

Many resampling schemes can be described as GBS, such as the m-out-of-n bootstrap (Bickel & Sakov, 2008) and scale-enhanced wild bootstrap (Chatterjee, 2019). Under fairly weak regularity conditions on the first two derivatives ψ_i' and ψ_i'' of ψ_i, $(a_n/\tau_n)(\hat{\theta}_{r*} - \hat{\theta}_*)$ and $a_n(\hat{\theta}_* - \theta_0)$ converge to the same weak limit in probability (See Appendix A).

Instead of repeatedly solving (4.1), we use model quantities computed while calculating $\hat{\theta}_*$ to obtain a first-order approximation of $\hat{\theta}_{r*}$.

$$\begin{aligned} \hat{\theta}_{r*} = \hat{\theta}_* - \frac{\tau_n}{a_n}\left[\sum_{i=1}^n \psi_i''(\hat{\theta}_*, Z_i)\right]^{-1/2} \sum_{i=1}^n W_{ri}\psi_i'(\hat{\theta}_*, Z_i) \\ + R_r, \end{aligned} \tag{4.3}$$

where $\mathbb{E}_r\|R_r\|^2 = o_P(1)$, and $W_{ri} = (\mathbb{W}_{ri} - 1)/\tau_n$. Thus only Monte Carlo sampling is required to obtain the resamples. Being an embarrassingly parallel procedure, this results in significant computational speed gains.

To estimate $e_n(\mathcal{M})$ we obtain two independent sets of weights $\{\mathcal{W}_r; r = 1,\ldots,R\}$ and $\{\mathcal{W}_{r_1}; r_1 = 1,\ldots,R_1\}$ for large integers R, R_1. We use the first set of resamples to obtain the distribution $[\hat{\theta}_{r*}]$ to approximate $[\hat{\theta}_*]$, and the second set of resamples to get the plugin estimate $\hat{\theta}_{r_1m}$:

$$\hat{\theta}_{r_1mj} = \begin{cases} \hat{\theta}_{r_1*j} & \text{for } j \in \mathcal{S}, \\ C_j & \text{for } j \notin \mathcal{S}. \end{cases} \tag{4.4}$$

Finally, the resampling estimate of a model e-value is: $e_{rn}(\mathcal{M}) = \mathbb{E}_{r_1} D\big(\hat{\theta}_{r_1m}, [\hat{\theta}_{r*}]\big)$, where $\mathbb{E}_{r_1}$ is expectation, conditional on the data, computed using the resampling r_1.

4.2. Fast algorithm for best subset selection

For best subset selection, we restrict to the model class $\mathbb{M}_0 = \{\mathcal{M} : C_j = 0 \quad \forall \quad j \notin \mathcal{S}\}$ that only allows zero constant terms. 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 (c) collect covariates dropping which causes the e-value to go down.

To fit the full model, we need to determine the estimable index set $\mathcal{S}_*$. When $n > p$, the choice is simple: $\mathcal{S}_* = \{1,\ldots,p\}$. 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. Similar to past works on feature selection (Lai et al., 2015), we use a feature screening step before model fitting to achieve this (Section 5).

After obtaining $\mathcal{S}_*$ and $\hat{\theta}_*$, for each of the $p'+1$ models under consideration: the full model and all drop-1-feature models, we follow the recipe in Section 4.1 to get bootstrap e-values. This gives us all the components for a sample version of Algorithm 1, which we present as Algorithm 2.

Algorithm 2 Best subset feature selection using e-values and GBS

1. Fix resampling standard deviation τ_n.
2. Obtain GBS samples: $\mathcal{T} = \{\hat{\theta}_{1*}, ..., \hat{\theta}_{R*}\}$, and $\mathcal{T}_1 = \{\hat{\theta}_{1*}, ..., \hat{\theta}_{R_1*}\}$ using (4.3).
3. Calculate $\hat{e}_{rn}(\mathcal{M}_*) = \frac{1}{R_1}\sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1*}, [\mathcal{T}])$.
4. Set $\hat{\mathcal{S}}_0 = \phi$.
5. **for** j in $1 : p$
6. **for** r_1 in $1 : R_1$
7. Replace j^{th} index of $\hat{\theta}_{*r_1}$ by 0 to get $\hat{\theta}_{r_1,-j}$.
8. Calculate $\hat{e}_{rn}(\mathcal{M}_{-j}) = \frac{1}{R_1}\sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1,-j}, [\mathcal{T}])$.
9. **if** $\hat{e}_{rn}(\mathcal{M}_{-j}) < \hat{e}_{rn}(\mathcal{M}_*)$
10. Set $\hat{\mathcal{S}}_0 \leftarrow \{\hat{\mathcal{S}}_0, j\}$.

Choice of τ_n The intermediate rate of divergence for the bootstrap standard deviation τ_n: $\tau_n \to \infty, \tau_n/a_n \to 0$, is a necessary and sufficient condition for the consistency of GBS (Chatterjee & Bose, 2005), and that of the bootstrap approximation of population e-values (Theorem 5.4). We select τ_n using the following quantity, which we call Generalized Bootstrap Information Criterion (GBIC):

$$\text{GBIC}(\tau_n) = \sum_{i=1}^{n} \psi_i\left(\hat{\theta}(\hat{\mathcal{S}}_0, \tau_n), Z_i\right) + \frac{\tau_n}{2}\left|\text{supp}(\hat{\theta}(\hat{\mathcal{S}}_0, \tau_n))\right|, \tag{4.5}$$

where $\hat{\theta}(\hat{\mathcal{S}}_0, \tau_n)$ is the refitted parameter vector using the index set $\hat{\mathcal{S}}_0$ selected by running Algorithm 2 with τ_n. We repeat this for a range of τ_n values, and choose the index set corresponding to the τ_n that gives the smallest GBIC(τ_n). For our synthetic data experiments we take $\tau_n = \log(n)$ and use GBIC, while for one real data example, we use validation on a test set to select the optimal τ_n, both with favorable results.

Detection threshold for finite samples In practice—especially for smaller sample sizes—due to sampling uncertainty it may be difficult to ensure that the full model e-value exactly separates the null and non-null e-values, and small amounts of false positives or negatives may occur in the selected feature set $\hat{\mathcal{S}}_0$. One way to tackle this is by shifting the detection threshold in Algorithm 2 by a small δ:

$$\hat{e}^{\delta}_{rn}(\mathcal{M}_*) = (1+\delta)\hat{e}_{rn}(\mathcal{M}_*).$$

To prioritize true positives or true negatives, δ can be set to be positive or negative respectively. In one of our experiments (Section 6.3), setting $\delta = 0$ results in a small amount of false positives due to a few non-null features having e-values close to the full model e-value. Setting δ to small positive values gradually eliminates these false positives.

5. Theoretical results

We now investigate theoretical properties of e-values. Our first result separates inadequate models from adequate models at the population level, and gives a rank ordering of adequate models using their population e-values.

Theorem 5.1. *Under conditions B1-B5, for a finite sequence of adequate models $\mathcal{M}_1 \prec \ldots \prec \mathcal{M}_k$ and any inadequate models $\mathcal{M}_{k+1}, \ldots, \mathcal{M}_K$, we have for large n*

$$e_n(\mathcal{M}_1) > \ldots > e_n(\mathcal{M}_k) > \max_{j\in\{k+1,\ldots K\}} e_n(\mathcal{M}_j).$$

As $n \to \infty$, $e_n(\mathcal{M}_i) \to \mathbb{E}D(Y, [Y]) < \infty$ with Y having an elliptic distribution if $i \le K$, else $e_n(\mathcal{M}_i) \to 0$.

We define the *data generating model* as $\mathcal{M}_0 \prec \mathcal{M}_*$ with estimable indices $\mathcal{S}_0 = \text{supp}(\theta_0)$ and constants $C_0 = 0_{p-|\mathcal{S}_0|}$. Then we have the following.

Corollary 5.2. *Assume the conditions of Theorem 5.1. Consider the restricted class of candidate models $\mathbb{M}_0$ in Section 4.2, where all known coefficients are fixed at 0. Then for large enough n, $\mathcal{M}_0 = \arg\max_{\mathcal{M}\in\mathbb{M}_0}[e_n(\mathcal{M})]$.*

Thus, when only the models with known parameters set at 0 (the model set $\mathbb{M}_0$) are considered, e-value indeed maximizes at the true model at the *population level*. However there are still 2^p possible models. This is where the advantage of using e-values—their one-step nature—comes through.

Corollary 5.3. *Assume the conditions of Corollary 5.2. 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 for large n we have $e_n(\mathcal{M}_{-j}) < e_n(\mathcal{M}_*)$.*

Dropping an essential feature from the full model makes the model inadequate, which has very small e-value for large enough n, whereas dropping a non-essential feature increases the e-value (Theorem 5.1). Thus, *simply collecting features dropping which cause a decrease in the e-value suffices for feature selection*.

Following the above results, we establish model selection consistency of Algorithm 2 at the sample level. This means that the probability that the one-step procedure is able to select the true model feature set goes to 1, when the procedure is repeated for a large number of randomly drawn datasets from the data-generating distribution.

Theorem 5.4. *Consider two sets of generalized 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-values as:*

$$\hat{e}_{rn}(\mathcal{M}) = \frac{1}{R_1}\sum_{r_1=1}^{R_1} D(\hat{\theta}_{r_1 m}, [\mathcal{T}]), \tag{5.1}$$

where $[\mathcal{T}]$ is the empirical distribution of the corresponding bootstrap samples, and $\hat{\theta}_{r_1 m}$ are obtained as in Section 4.1. Consider the feature set $\hat{\mathcal{S}}_0 = \{\hat{e}_{rn}(\mathcal{M}_{-j}) < \hat{e}_{rn}(\mathcal{M}_*)\}$. Then as $n, R, R_1 \to \infty$, $P_2(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \to 1$, with P_2 as probability conditional on data and bootstrap samples.

Theorem 5.4 is contingent on the fact that the the true model $\mathcal{M}_0$ is a member of $\mathbb{M}_0$, that is $\mathcal{S}_0 \subseteq \mathcal{S}_*$. This is ensured trivially when $n \geq p$. If $p > n$, $\mathbb{M}_0$ is the set of all possible models on the feature set selected by an appropriate screening procedure. In high-dimensional linear models, we use Sure Independent Screening (Fan & Lv, 2008, SIS) for this purpose. Given that $\mathcal{S}_*$ is selected using SIS, Fan & Lv (2008) proved that, for constants $C > 0$ and κ that depend on the minimum signal in θ_0,

$$P(\mathcal{M}_0 \in \mathbb{M}_0) \geq 1 - O\left(\frac{\exp[-Cn^{1-2\kappa}]}{\log n}\right). \qquad (5.2)$$

For more complex models, model-free filter methods (Koller & Sahami, 1996; Zhu et al., 2011) can be used to obtain $\mathcal{S}_*$. For example, under mild conditions on the design matrix, the method of Zhu et al. (2011) is consistent:

$$P(|\mathcal{S}_* \cap \mathcal{S}_0^c| \geq r) \leq \left(1 - \frac{r}{p+d}\right)^d, \qquad (5.3)$$

with positive integers r and d: $d = p$ being a good practical choice (Zhu et al., 2011, Theorem 3). Combining (5.2) or (5.3) with Corollary 5.4 as needed establishes asymptotically accurate recovery of $\mathcal{S}_0$ through Algorithm 2.

6. Numerical experiments

We implement e-values using a GBS with scaled resample weights $W_{ri} \sim \text{Gamma}(1,1) - 1$, and resample sizes $R = R_1 = 1000$. We use Mahalanobis depth for all depth calculations. Mahalanobis depth is much less computation-intensive than other depth functions (Dyckerhoff & Mozharovskyi, 2016; Liu & Zuo, 2014), but is not usually preferred in applications due to its non-robustness. However, we do not use any robustness properties of data depth, so are able to use it without any concern. For each replication for each data setting and method, we compute performance metrics on test datasets of the same dimensions as the respective training dataset. All our results are based on 1000 such replications.

6.1. Feature selection in linear regression

Given a true coefficient vector β_0, we use the model $Z \equiv (Y, X), Y = X\beta_0 + \epsilon$, with $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, and $n = 500$ and $p = 100$. We generate the rows of X independently from $\mathcal{N}_p(0, \Sigma_X)$, where Σ_X follows an equicorrelation structure having correlation coefficient ρ: $(\Sigma_X)_{ij} = \rho^{|i-j|}$. Under this basic setup, we consider the following settings

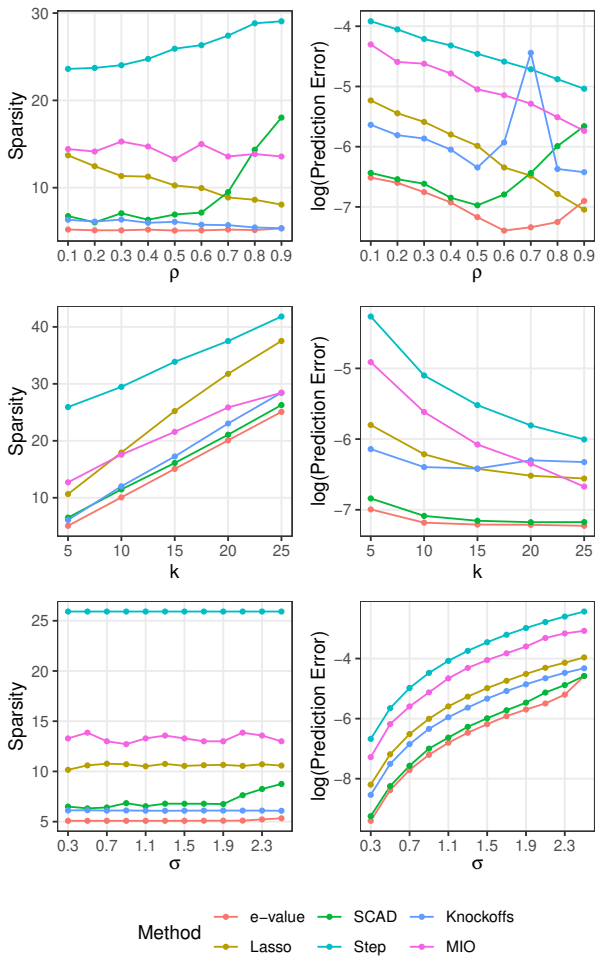


Figure 6.1. Performance metrics for $n = 500, p = 100$: top, middle and bottom rows indicate simulation settings 1, 2 and 3, respectively. Competing methods include LASSO and SCAD-penalized regression, Stepwise regression using BIC and backward deletion, Knockoff filters (Barber & Candes, 2015), and Mixed Integer Optimization (Bertsimas et al., 2016, MIO).

to evaluate the effect of different magnitudes of feature correlation in X, sparsity level in β_0 and error variance σ.

- Setting 1 (effect of feature correlation): We repeat the above setup for $\rho = 0.1, \ldots, 0.9$, fixing $\beta_0 = (1.5, 0.5, 1, 1.5, 1, 0_{p-5})$, $\sigma = 1$;
- Setting 2 (effect of sparsity level): We consider $\beta_0 = (1_k, 0_{p-k})$, with varying degrees of the sparsity level $k = 5, 10, 15, 20, 25$. We fix $\rho = 0.5, \sigma = 1$;
- Setting 3 (effect of noise level): We consider different values of noise level (equivalent to having different signal-to-noise ratios) by testing for $\sigma = 0.3, 0.5, \ldots, 2.3$, fixing $\beta_0 = (1_5, 0_{p-5})$, $\rho = 0.5$.

To implement e-values, we repeat Algorithm 2 for $\tau_n \in \{0.2, 0.6, 1, 1.4, 1.8\}\log(n)\sqrt{\log(p)}$, and select the model having the lowest GBIC(τ_n).

Figure 6.1 summarizes the comparison results. Across all three settings, our method consistently produces the most predictive model, i.e. the model with the lowest prediction error. It also produces the sparsest model almost always. Among the competitors, SCAD performs the closest to e-values in setting 2 for both the metrics. However, SCAD seems to be more severely affected by high noise level (i.e. high σ) or high feature correlation (i.e. high ρ). Lasso and Step tend to select models with many null variables, and have higher prediction errors. Performance of the other two methods (MIO, knockoffs) is middling.

Method	e-value	Lasso	SCAD	Step	Knockoff	MIO
Time	6.3	0.4	0.9	20.1	1.9	127.2

Table 6.1. Mean computation time (in seconds) for all methods.

We present computation times for Setting 1 averaged across all values of ρ in Table 6.1. All computations were performed on a Windows desktop with an 8-core Intel Core-i7 6700K 4GHz CPU and 16GB RAM. For e-values, using a smaller number of bootstrap samples or running computations over bootstrap samples in parallel greatly reduces computation time with little to no loss of performance. However, we report its computation time over a single core similar to other methods. Sparse methods like Lasso and SCAD are much faster as expected. Our method has lower computation time than Step and MIO, and much better performance.

6.2. High-dimensional linear regression ($p > n$)

We generate data from the same setup as Section 6.1, but with $n = 100, p = 500$. We perform an initial screening of features using SIS, then apply the e-values and other methods on this SIS-selected predictor set. Figure 6.2 summarizes the results. In addition to competing methods, we report metrics corresponding to the original SIS-selected predictor set as a baseline. Sparsity-wise, e-values produce the most improvement over the SIS baseline among all methods, and tracks the true sparsity level closely. Both e-values and Knockoffs produce sparser estimates as ρ grows higher (Setting 1). However, unlike the Knockoffs our method maintains good prediction performance even at high feature correlation. In general e-values maintain good prediction performance, although this difference is less obvious than the low-dimensional case. Note that for $k = 25$ in setting 2, SIS screening produces overly sparse feature sets, affecting prediction errors for all methods.

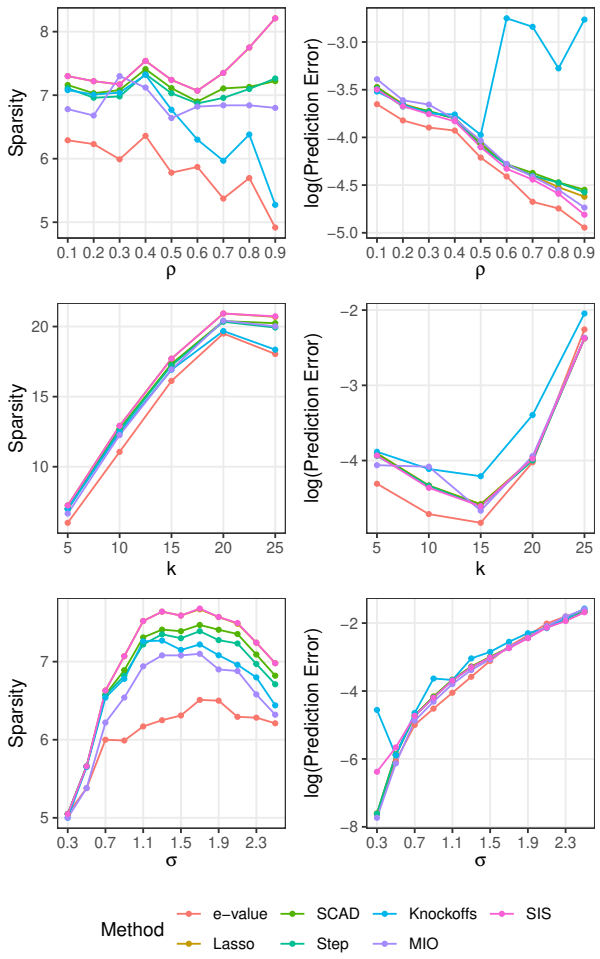


Figure 6.2. Performance metrics for $n = 100, p = 500$.

6.3. Mixed effect model

We use the simulation setup from Peng & Lu (2012): $Y = X\beta + RU + \epsilon$, where the data $Z \equiv (Y, X, R)$ consists of m independent groups of observations with multiple (n_i) dependent observations in each group, R being the within-group random effects design matrix. We consider 9 fixed effects and 4 random effects, with the true fixed effect coefficient $\beta_0 = (1_2, 0_7)$ and random effect coefficient U drawn from $\mathcal{N}_4(0, \Delta)$. The random effect covariance matrix Δ has elements $(\Delta)_{11} = 9$, $(\Delta)_{21} = 4.8$, $(\Delta)_{22} = 4$, $(\Delta)_{31} = 0.6$, $(\Delta)_{32} = (\Delta)_{33} = 1$, $(\Delta)_{4j} = 0; j = 1, \ldots, 4$, and the error variance of ϵ is set at $\sigma^2 = 1$. The goal here is to select covariates of the fixed effect. We use two scenarios for our study: Setting 1, where the number of groups (m) considered is 30, and the number of observations in the i^{th} group, $i = 1, \ldots, m$, is $n_i = 5$, and Setting 2, where $m = 60, n_i = 10$. Finally, we generate 100 independent

Method		Setting 1: $n_i = 5, m = 30$				Setting 2: $n_i = 10, m = 60$			
		FPR	FNR	Acc	MS	FPR	FNR	Acc	MS
e-value	$\delta = 0$	9.4	0.0	76	2.33	0.0	0.0	100	2.00
	$\delta = 0.01$	6.7	0.0	82	2.22	0.0	0.0	100	2.00
	$\delta = 0.05$	1.0	0.0	97	2.03	0.0	0.0	100	2.00
	$\delta = 0.1$	0.3	0.0	99	2.01	0.0	0.0	100	2.00
	$\delta = 0.15$	0.0	0.0	100	2.00	0.0	0.0	100	2.00
SCAD (Peng & Lu, 2012)	BIC	21.5	9.9	49	2.26	1.5	1.9	86	2.10
	AIC	17	11.0	46	2.43	1.5	3.3	77	2.20
	GCV	20.5	6	49	2.30	1.5	3	79	2.18
	$\sqrt{\log n/n}$	21	15.6	33	2.67	1.5	4.1	72	2.26
M-ALASSO (Bondell et al., 2010)		-	-	73	-	-	-	83	-
SCAD-P (Fan & Li, 2012)		-	-	90	-	-	-	100	-
rPQL (Hui et al., 2017)		-	-	98	-	-	-	99	-

Table 6.2. Performance comparison for mixed effect models. We compare e-values with a number of sparse penalized methods: (a) Peng & Lu (2012) that uses SCAD penalty and different methods of selecting regularization tuning parameters, (b) The adaptive lasso-based method of Bondell et al. (2010), (c) The SCAD-P method Fan & Li (2012), and (d) regularized Penalized Quasi-Likelihood Hui et al. (2017, rPQL). For comparison with Peng & Lu (2012), we present mean false positive (FPR) and false negative (FNR) rates, Accuracy (Acc), and Model Size (MS), i.e. the number of non-zero fixed effects estimated. To compare with other methods we only use Acc, since they did not report the rest of the metrics.

datasets for each setting. To implement e-values by minimizing GBIC(τ_n), we consider $\tau_n \in \{1, 1.2, \ldots, 4.8, 5\}$. To tackle small sample issues in Setting 1 (Section 4.2), we repeat the model selection procedure using the shifted e-values $\hat{e}^{\delta}_{rn}(\cdot)$ for $\delta \in \{0, 0.01, 0.05, 0.1, 0.15\}$.

Without shifted thresholds, e-values perform commendably in both settings. For Setting 2, it reaches the best possible performance across all metrics. However, we observed that in a few replications of setting 1, a small number of null input features had e-values only slightly lower than the full model e-value, resulting in increased FPR and model size. We experimented with lowered detection thresholds to mitigate this. As seen in Table 6.2, increasing δ gradually eliminates the null features, and e-values reach perfect performance across all metrics for $\delta = 0.15$.

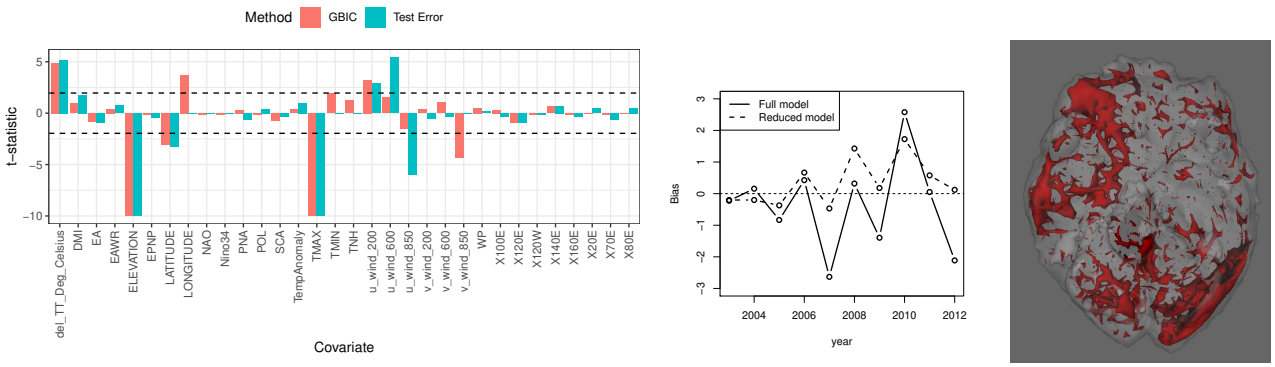
7. Real data examples

Indian monsoon data To identify the driving factors behind precipitation during the Indian monsoon season using e-values, we obtain data on 35 potential covariates (see Appendix D) from National Climatic Data Center (NCDC) and National Oceanic and Atmospheric Administration (NOAA) repositories for 1978–2012. We consider annual medians of covariates as fixed effects, log yearly rainfall at a weather station as output feature, and include year-specific random intercepts. To implement e-values, we use projection depth (Zuo, 2003) and GBS resample sizes $R = R_1 = 1000$. We train our model on data from the years 1978-2002, run e-values best subset selection for tuning parameters $\tau_n \in \{0.05, 0.1, \ldots, 1\}$. We consider two methods to select the best refitted model: (a) minimizing GBIC(τ_n), and (b) minimizing forecasting errors on samples from 2003–2012.

Figure 7.1a plots the t-statistics for features from the best refitted models obtained by the above two methods. Minimizing for GBIC and test error selects 32 and 26 covariates, respectively. The largest contributors are maximum temperature and elevation, which are related to precipitation based on the Clausius-Clapeyron relation (Li et al., 2017; Singleton & Toumi, 2012). All other selected covariates have documented effects on Indian monsoon (Krishnamurthy & Kinter III, 2003; Moon et al., 2013). Reduced model forecasts obtained from a rolling validation scheme (i.e. i.e. use data from 1978–2002 for 2003, 1979-2003 for 2004 and so on) have less bias across testing years (Fig. 7.1b).

Spatio-temporal dependence analysis in fMRI data This dataset is due to Wakeman & Henson (2015), where each of the 19 subjects go through 9 runs of consecutive visual tasks. Blood oxygen level readings are recorded across time as 3D images made of $64 \times 64 \times 33$ (total 135,168) voxels. Here we use the data from a single run and task on subject 1, and aim to estimate dependence patterns of readings across 210 time points and areas of the brain.

We fit separate regressions at each voxel (Appendix E), with second order autoregressive terms, neighboring voxel readings and one-hot encoded visual task categories in the design matrix. After applying the e-value feature selection, we compute the F-statistic at each voxel using selected coefficients only, and obtain their p-values. Fig. 7.1c highlights voxels with p-values < 0.05. Left and right visual cortex areas show high spatial dependence, with more dependence



(a) (b) (c)

Figure 7.1. (a) Plot of t-statitics for Indian Monsoon data for features selected by at least one of the methods. Dashed lines indicate standard Gaussian quantiles at tail probabilities 0.025 and 0.975, (b) Full vs. reduced model rolling forecast comparisons, and (c) fMRI data: smoothed surface obtained from p-values show high spatial dependence in right optic nerve, auditory nerves and auditory cortex (top left), left visual cortex (bottom right) and cerebellum (lower middle).

on the left side. Signals from the right visual field obtained by both eyes are processed by the left visual cortex. The lopsided dependence pattern suggests that visual signals from the right side led to a higher degree of processing in our subject's brain. We also see activity in the cerebellum, the role of which in visual perception is well-known (Calhoun et al., 2010; Kirschen et al., 2010).

8. Conclusion

In this work, we introduced a new paradigm of feature selection through *e-values*. The e-values can be particularly helpful in situations where model training is costly and potentially distributed across multiple servers (i.e. federated learning), so that a brute force parallelized approach of training and evaluating multiple models is not practical or even possible.

There are three immediate extensions of the framework presented in this paper. Firstly, grouped e-values are of interest to leverage prior structural information on the predictor set. There are no conceptual difficulties in evaluating overlapping *and* non-overlapping groups of predictors using e-values in place of individual predictors. However technical conditions may be required to ensure a rigorous implementation. Secondly, our current formulation of e-values essentially relies upon the number of samples being more than the effective number of predictors for a unique covariance matrix of the parameter estimate to asymptotically exist. When $p > n$, using a sure screening method to filter out unnecessary variables works well empirically. However this needs theoretical validation. Thirdly, instead of using mean depth, other functionals of the (empirical) depth distribution—such as quantiles—can be used as e-values. Similar to stability selection (Meinshausen & Buhlmann, 2010), it may be possible to use the intersection of predictor sets obtained by using a number of such functionals in Algorithm 2 as the final selected set of important predictors.

An effective implementation of e-values hinges on the choice of bootstrap method and tuning parameter τ_n. To this end, we see opportunity for enriching the research on empirical process methods in complex overparametrized models, such as Deep Neural Nets (DNN), which the e-values framework can build up on. Given the current push for interpretability and trustworthiness of DNN-based decision making systems, there is potential for tools to be developed within the general framework of e-values that provide local and global explanations of large-scale deployed model outputs in an efficient manner.

Acknowledgements

This work is part of the first author (SM)'s PhD thesis (Majumdar, 2017). He acknowledges the support of the University of Minnesota Interdisciplinary Doctoral Fellowship during his PhD. The research of SC is partially supported by the US National Science Foundation grants 1737918, 1939916 and 1939956 and a grant from Cisco Systems.

Reproducibility

Code and data for the experiments in this paper are available at `https://github.com/shubhobm/e-values`.

References

Barber, R. F. and Candes, E. J. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43:2055–2085, 2015.

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

Bickel, P. and Sakov, A. On the Choice of m in the m Out of n Bootstrap and its Application to Confidence Bounds for Extrema. *Stat. Sinica*, 18:967–985, 2008.

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

Calhoun, V. D., Adali, T., and Mcginty, V. B. fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *Neuroimage*, 14:1080–1088, 2010.

Chatterjee, S. The scale enhanced wild bootstrap method for evaluating climate models using wavelets. *Statist. Prob. Lett.*, 144:69–73, 2019.

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

Dietz, L. R. and Chatterjee, S. Logit-normal mixed model for Indian monsoon precipitation. *Nonlin. Processes Geophys.*, 21:939–953, sep 2014.

Dietz, L. R. and Chatterjee, S. Investigation of Precipitation Thresholds in the Indian Monsoon Using Logit-Normal Mixed Models. In Lakshmanan, V., Gilleland, E., McGovern, A., and Tingley, M. (eds.), *Machine Learning and Data Mining Approaches to Climate Science*, pp. 239–246. Springer, 2015.

Dyckerhoff, R. and Mozharovskyi, P. Exact computation of the halfspace depth. *Comput. Stat. Data Anal.*, 98:19–30, 2016.

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

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

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

Fragoso, T. M., Bertoli, W., and Louzada, F. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1): 1–28, 2018.

Ghosh, S., Vittal, H., Sharma, T., et al. Indian Summer Monsoon Rainfall: Implications of Contrasting Trends in the Spatial Variability of Means and Extremes. *PLoS ONE*, 11:e0158670, 2016.

Gosswami, B. N. *The Global Monsoon System: Research and Forecast*, chapter The Asian Monsoon: Interdecadal Variability. World Scientific, 2005.

Guyon, I. and Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

Huang, H.-C., Hsu, N.-J., Theobald, D. M., and Breidt, F. J. Spatial Lasso With Applications to GIS Model Selection. *J. Comput. Graph. Stat.*, 19:963–983, 2010.

Hui, F. K. C., Muller, S., and Welsh, A. H. Joint Selection in Mixed Models using Regularized PQL. *J. Amer. Statist. Assoc.*, 112:1323–1333, 2017.

Jornsten, R. Clustering and classification based on the L_1 data depth. *J. Mult. Anal.*, 90:67–89, 2004.

Kirschen, M. P., Chen, S. H., and Desmond, J. E. Modality specific cerebro-cerebellar activations in verbal working memory: an fMRI study. *Behav. Neurol.*, 23:51–63, 2010.

Knutti, R., Furrer, R., Tebaldi, C., et al. Challenges in Combining Projections from Multiple Climate Models. *J. Clim.*, 23:2739–2758, 2010.

Koller, D. and Sahami, M. Toward optimal feature selection. In *13th International Conference on Machine Learning*, pp. 284–292, 1996.

Krishnamurthy, C. K. B., Lall, U., and Kwon, H.-H. Changing Frequency and Intensity of Rainfall Extremes over India from 1951 to 2003. *J. Clim.*, 22:4737–4746, 2009.

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

Lahiri, S. N. Bootstrapping M-estimators of a multiple linear regression parameter. *Ann. Statist.*, 20(‘):1548–1570, 1992.

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

Lee, H. and Ghosh, S. Performance of Information Criteria for Spatial Models. *J. Stat. Comput. Simul.*, 79:93–106, 2009.

Lei, J. et al. Distribution-Free Predictive Inference for Regression. *J. Amer. Statist. Assoc.*, 113:1094–1111, 2018.

Li, X., Wang, L., Guo, X., and Chen, D. Does summer precipitation trend over and around the Tibetan Plateau depend on elevation? *Int. J. Climatol.*, 37:1278–1284, 2017.

Liu, X. and Zuo, Y. Computing halfspace depth and regression depth. *Commun. Stat. Simul. Comput.*, 43(5): 969–985, 2014.

Ma, Y. and Zhu, L. A review on dimension reduction. *International Statistical Review*, 81(1):134–150, 2013.

Majumdar, S. *An Inferential Perspective on Data Depth*. PhD thesis, University of Minnesota, 2017.

Meinshausen, N. and Buhlmann, P. Stability selection. *J. R. Stat. Soc. B*, 72:417–473, 2010.

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

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

Mosler, K. Depth statistics. In Becker, C., Fried, R., and Kuhnt, S. (eds.), *Robustness and Complex Data Structures*, pp. 17–34. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-35493-9.

Narisetty, N. N. and Nair, V. N. Extremal Depth for Functional Data and Applications. *J. Amer. Statist. Assoc.*, 111:1705–1714, 2016.

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

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

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

Rousseeuw, P. J. and Hubert, M. Regression Depth. *J. Amer. Statist. Assoc.*, 94:388–402, 1998.

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

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

Singh, K., Xie, M., and Strawderman, W. E. Confidence distribution (CD) – distribution estimator of a parameter. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, volume 54 of *Lecture Notes-Monograph Series*, pp. 132–150. IMS, 2007.

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

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012.

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

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

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

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

VanderWeele, T. and Ding, P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med.*, 167(4):268–274, 2017.

Wakeman, D. G. and Henson, R. N. A multi-subject, multimodal human neuroimaging dataset. *Scientif. Data*, 2: article 15001, 2015.

Wang, B., Ding, Q., Fu, X., et al. Fundamental challenge in simulation and prediction of summermonsoon rainfall. *Geophys. Res. Lett.*, 32:L15711, 2005.

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

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

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

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

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

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

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

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

Appendix

A. Consistency of Generalized Bootstrap

We first state a number of conditions on the energy functions $\psi_i(\cdot,\cdot)$, under which we state and prove two results to ensure consistency of the estimation and bootstrap approximation procedures. In the context of the main paper, the conditions and results here ensure that the full model parameter estimate $\hat{\theta}_*$ follows a Gaussian sampling distribution, and Generalized Bootstrap (GBS) can be used to approximate this sampling distribution.

A.1. Technical conditions

Note that several sets of alternative conditions can be developed (Chatterjee & Bose, 2005; Lahiri, 1992), many of which are amenable to our results. However, for the sake of brevity and clarity, we only address the case where the energy function $\psi_i(\cdot,\cdot)$ is smooth in the first argument. This case covers a vast number of models routinely considered in statistics and machine learning.

We often drop the second argument from energy function, thus for example $\psi_i(\theta) \equiv \psi_i(\theta, Z_i)$, and use the notation $\hat{\psi}_{ki}$ for $\psi_{ki}(\hat{\theta}_*)$, for $k = 0, 1, 2$. Also, for any function $h(\theta)$ evaluated at the true parameter value θ_*, we use the notation $h \equiv h(\theta_*)$. When A and B are square matrices of identical dimensions, the notation $B < A$ implies that the matrix $A - B$ is positive definite.

In a neighborhood of θ_*, we assume the functions ψ_i are thrice continuously differentiable in the first argument, with the successive derivatives denoted by ψ_{ki}, $k = 0, 1, 2$. That is, there exists a $\delta > 0$ such that for any $\theta = \theta_* + t$ satisfying $\|t\| < \delta$ we have

$$\frac{d}{d\theta}\psi_i(\theta) := \psi_{0i}(\theta) \in \mathbb{R}^p,$$

and for the a-th element of $\psi_{0i}(\theta)$, denoted by $\psi_{0i(a)}(\theta)$, we have

$$\psi_{0i(a)}(\theta) = \psi_{0i(a)}(\theta_*) + \psi_{1i(a)}(\theta_*)t + 2^{-1}t^T\psi_{2i(a)}(\theta_* + ct)t,$$

for $a = 1, \ldots p$, and some $c \in (0, 1)$ possibly depending on a. We assume that for each n, there is a sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \ldots \subset \mathcal{F}_n$ such that $\{\sum_{i=1}^{j} \psi_{0i}(\theta_*), \mathcal{F}_j\}$ is a martingale.

The spectral decomposition of $\Gamma_{0n} := \sum_{i=1}^{n} \mathbb{E}\psi_{0i}\psi_{0i}^T$ is given by $\Gamma_{0n} = P_{0n}\Lambda_{0n}P_{0n}^T$, where $P_{0n} \in \mathbb{R}^p \times \mathbb{R}^p$ is an orthogonal matrix whose columns contain the eigenvectors, and Λ_{0n} is a diagonal matrix containing the eigenvalues of Γ_{0n}. We assume that Γ_{0n} is positive definite, that is, all the diagonal entries of Λ_{0n} are positive numbers. We assume that there is a constant $\delta_0 > 0$ such that $\lambda_{min}(\Gamma_{0n}) > \delta_0$ for sufficiently large n.

Let $\Gamma_{1i}(\theta_*)$ be the $p \times p$ matrix whose a-th row is $\mathbb{E}\psi_{1i(a)}$; we assume this expectation exists. Define $\Gamma_{1n} = \sum_{i=1}^{n} \Gamma_{1i}(\theta_*)$. We assume that Γ_{1n} is nonsingular for each n. The singular value decomposition of Γ_{1n} is given by $\Gamma_{1n} = P_{1n}\Lambda_{1n}Q_{1n}^T$, where $P_{1n}, Q_{1n} \in \mathbb{R}^p \times \mathbb{R}^p$ are orthogonal matrices, and Λ_{1n} is a diagonal matrix. We assume that the diagonal entries of Λ_{1n} are all positive, which implies that *in the population, at the true value of the parameter* the energy functional $\Psi_n(\theta_*)$ actually achieves a minimal value. We define Λ_{kn}^c for various real numbers c as diagonal matrices where the j-th diagonal entry of Λ_{kn} is raised to the power c, for $k = 0, 1$. Correspondingly, we define $\Gamma_{1n}^c = P_{1n}\Lambda_{1n}^c Q_{1n}^T$. We assume that there is a constant $\delta_1 > 0$ such that $\lambda_{max}(\Gamma_{1n}^T\Gamma_{1n}) < \delta_1$ for all sufficiently large n. Define the matrix $A_n = \Gamma_{0n}^{-1/2}\Gamma_{1n}$. We assume the following conditions:

(C1) The minimum eigenvalue of $A_n^T A_n$ tends to infinity. That is, there is a sequence $a_n \uparrow \infty$ as $n \to \infty$ such that

$$\lambda_{min}\big(\Gamma_{1n}\Gamma_{0n}^{-1}\Gamma_{1n}^T\big) \asymp a_n^2. \tag{C.1}$$

(C2) There exists a sequence of positive reals $\{\gamma_n\}$ that is bounded away from zero, such that

$$\lambda_{max}\big(\Gamma_{1n}^{-1}\Gamma_{0n}^2\Gamma_{1n}^{-T}\big) = o(\gamma_n^{-2}) \quad \text{as } n \to \infty. \tag{C.2}$$

(C3)

$$\mathbb{E}\left\|A_n^{-1}\Big(\sum_{i=1}^{n}\psi_{1i} - \Gamma_{1n}\Big)A_n^{-1}\right\|_F^2 = o(p\gamma_n^{-2}). \tag{C.3}$$

where $\|A\|_F$ denotes the Frobenius norm of matrix A.

(C4) For the symmetric matrix $\psi_{2i(a)}(\theta)$ and for some $\delta_2 > 0$, there exists a symmetric matrix $M_{2i(a)}$ such that

$$\sup_{\|\theta-\theta_*\|<\delta_2} \psi_{2i(a)}(\theta) < M_{2i(a)},$$

satisfying

$$\sum_{a=1}^{p}\sum_{i=1}^{n} \mathbb{E}\lambda_{max}^2\big(M_{2i(a)}\big) = o\big(a_n^6 n^{-1} p\gamma_n^{-2}\big). \tag{C.4}$$

(C5) For any vector $c \in \mathbb{R}^p$ with $\|c\| = 1$, we define the random variable $Z_{ni} = -c^T\Gamma_{0n}^{-1/2}\psi_i$ for $i = 1, \ldots n$. We assume that

$$\sum_{i=1}^{n} Z_{ni}^2 \xrightarrow{p} 1, \text{ and } \mathbb{E}\big[\max_i \|Z_{ni}\|\big] \to 0. \tag{C.5}$$

(C6) Assume that

$$\lambda_{max}\big(\Gamma_{1n}\Gamma_{0n}^{-1}\Gamma_{1n}^T\big) \asymp a_n^2. \tag{C.6}$$

The technical conditions (C1)-(C5) are extremely broad, and allow for different rates of convergence of different parameter estimators. The additional condition (C6) is a natural condition that, coupled with (C1), ensures identical rate of convergence a_n for all the parameter estimators in a model.

Standard regularity conditions on likelihood and estimating functions that have been routinely assumed in the literature are special cases of the framework above. In such cases, (C1)-(C6) hold with $a_n \equiv n^{1/2}$, resulting in the standard "root-n" asymptotics.

A.2. Results

We first present the consistency and asymptotic normality of the estimation process in Theorem A.1 below.

Theorem A.1. *Assume conditions (C1)-(C5). Then $\hat{\theta}_*$ is a consistent estimator of θ_*, and $A_n(\hat{\theta}_* - \theta_*)$ converges weakly to the p-dimensional standard Gaussian distribution. Under the additional condition (C6), we have that $a_n(\hat{\theta}_* - \theta_*)$ converges weakly to a Gaussian distribution in p-dimension.*

Proof of Theorem A.1. We consider a generic point $\theta = \theta_* + A_n^{-1}t$. From the Taylor series expansion, we have

$$\psi_{0i(a)}(\theta) = \psi_{0i(a)} + \psi_{1i(a)}A_n^{-1}t + 2^{-1}t^T A_n^{-T}\psi_{2i(a)}(\tilde{\theta}_*)A_n^{-1}t,$$

for $a = 1, \ldots p$, and $\tilde{\theta}_* = \theta_* + cA_n^{-1}t$ for some $c \in (0, 1)$.

Recall our convention that for any function $h(\theta)$ evaluated at the true parameter value θ_*, we use the notation $h \equiv h(\theta_*)$. Also define the p-dimensional vector $R_n(\tilde{\theta}_*, t)$ whose a-th element is given by

$$R_{n(a)}(\tilde{\theta}_*, t) = t^T A_n^{-T}\sum_{i=1}^{n}\psi_{2i(a)}(\tilde{\theta}_*)A_n^{-1}t.$$

Thus we have

$$\begin{aligned}
p^{-1/2}A_n^{-1}\sum_{i=1}^{n}\psi_{0i}(\theta_* + A_n^{-1}t) &= p^{-1/2}A_n^{-1}\sum_{i=1}^{n}\psi_{0i} + p^{-1/2}A_n^{-1}\sum_{i=1}^{n}\psi_{1i}A_n^{-1}t + 2^{-1}p^{-1/2}A_n^{-1}R_n(\tilde{\theta}_*, t)\\
&= p^{-1/2}A_n^{-1}\sum_{i=1}^{n}\psi_{0i} + p^{-1/2}A_n^{-1}\Gamma_{1n}A_n^{-1}t\\
&\quad + p^{-1/2}A_n^{-1}\big(\sum_{i=1}^{n}\psi_{1i} - \Gamma_{1n}\big)A_n^{-1}t\\
&\quad + 2^{-1}p^{-1/2}A_n^{-1}R_n(\tilde{\theta}_*, t).
\end{aligned}$$

Fix $\epsilon > 0$. We first show that there exists a $C_0 > 0$ such that

$$\mathbb{P}\Big[\big\|p^{-1/2}A_n^{-1}\sum_{i=1}^{n}\psi_{0i}\big\| > C_0\Big] < \epsilon/2. \tag{A.7}$$

For this, we compute

$$\begin{aligned} p^{-1}\mathbb{E}\big\|A_n^{-1}\sum_{i=1}^{n}\psi_{0i}\big\|^2 &= p^{-1}\mathbb{E}\sum_{i,j=1}^{n}\psi_{0i}^T A_n^{-T}A_n^{-1}\psi_{0j} \\ &= p^{-1}\mathrm{tr}\left(A_n^{-T}A_n^{-1}\mathbb{E}\sum_{i=1}^{n}\psi_{0i}\psi_{0i}^T\right) \\ &= p^{-1}\mathrm{tr}\left(A_n^{-T}A_n^{-1}\Gamma_{0n}\right) \\ &= O(1) \end{aligned}$$

from assumption (C.2).

Define

$$S_n(t) = p^{-1/2}A_n^{-1}\big(\sum_{i=1}^{n}\psi_{0i}(\theta_* + A_n^{-1}t) - \sum_{i=1}^{n}\psi_{0i}\big) - p^{-1/2}\Gamma_{1n}^{-1}\Gamma_{0n}t.$$

We next show that for any $C > 0$, for all sufficiently large n, we have

$$\mathbb{E}\Big[\sup_{\|t\|\le C}\big\|S_n(t)\big\|\Big]^2 = o(1). \tag{A.8}$$

This follows from (C.3) and (C.4). Note that

$$S_n(t) = p^{-1/2}A_n^{-1}\big(\sum_{i=1}^{n}\psi_{1i} - \Gamma_{1n}\big)A_n^{-1}t + 2^{-1}p^{-1/2}A_n^{-1}R_n(\tilde{\theta}_n, t).$$

Thus,

$$\sup_{\|t\|\le C}\big\|S_n(t)\big\| \le p^{-1/2}\sup_{\|t\|\le C}\big\|A_n^{-1}\big(\sum_{i=1}^{n}\psi_{1i} - \Gamma_{1n}\big)A_n^{-1}t\big\| + 2^{-1}p^{-1/2}\sup_{\|t\|\le C}\big\|A_n^{-1}R_n(\tilde{\theta}_*, t)\big\|.$$

We consider each of these terms separately.

For any matrix $M \in \mathbb{R}^p \times \mathbb{R}^p$, we have

$$\sup_{\|t\|\le C}\big\|Mt\big\| = \sup_{\|t\|\le C}\Big[\sum_{i=1}^{p}\big(\sum_{j=1}^{p}M_{ij}t_j\big)^2\Big]^{1/2} \le \sup_{\|t\|\le C}\Big[\sum_{i=1}^{p}\sum_{j=1}^{p}M_{ij}^2\sum_{j=1}^{p}t_j^2\Big]^{1/2} = \big\|M\big\|_F\sup_{\|t\|\le C}\|t\| = C\big\|M\big\|_F.$$

Using $M = A_n^{-1}\big(\sum_{i=1}^{n}\psi_{1i} - \Gamma_{1n}\big)A_n^{-1}$ and (C.3), we get one part of the result.

For the other term, we similarly have

$$\begin{aligned} \Big[\sup_{\|t\|\le C}\big\|p^{-1/2}A_n^{-1}R_n(\tilde{\theta}_*, t)\big\|\Big]^2 &= p^{-1}\sup_{\|t\|\le C}\big\|A_n^{-1}R_n(\tilde{\theta}_*, t)\big\|^2 \\ &\le p^{-1}\lambda_{max}\big(A_n^{-T}A_n^{-1}\big)\sup_{\|t\|\le C}\big\|R_n(\tilde{\theta}_*, t)\big\|^2 \\ &\le p^{-1}\lambda_{max}\big(A_n^{-1}A_n^{-T}\big)\sup_{\|t\|\le C}\big\|R_n(\tilde{\theta}_*, t)\big\|^2 \\ &\le p^{-1}a_n^{-2}\sup_{\|t\|\le C}\big\|R_n(\tilde{\theta}_*, t)\big\|^2. \end{aligned}$$

Note that

$$\big(\sup_{\|t\|\le C}\big\|R_n(\tilde\theta_*,t)\big\|\big)^2=\sup_{\|t\|\le C}\big\|R_n(\tilde\theta_*,t)\big\|^2.$$

Now

$$\begin{aligned}\big\|R_n(\tilde\theta_*,t)\big\|^2&=\sum_{a=1}^p\big(R_{*n(a)}(\tilde\theta_*,t)\big)^2\\&=\sum_{a=1}^p\big(t^TA_n^{-T}\sum_{i=1}^n\psi_{2i(a)}(\tilde\theta_*)A_n^{-1}t\big)^2\\&=\sum_{a=1}^p\sum_{i,j=1}^n t^TA_n^{-T}\psi_{2i(a)}(\tilde\theta_*)A_n^{-1}tt^TA_n^{-T}\psi_{2j(a)}(\tilde\theta_*)A_n^{-1}t.\end{aligned}$$

Based on this, we have

$$\begin{aligned}\sup_{\|t\|\le C}\big\|R_n(\tilde\theta_*,t)\big\|^2&=\sup_{\|t\|\le C}\sum_{a=1}^p\sum_{i,j=1}^n t^TA_n^{-T}\psi_{2i(a)}(\tilde\theta_*)A_n^{-1}tt^TA_n^{-T}\psi_{2*nj(a)}(\tilde\theta_*)A_n^{-1}t\\&\le\sup_{\|t\|\le C}\sum_{a=1}^p\sum_{i,j=1}^n t^TA_n^{-T}M_{2i(a)}A_n^{-1}tt^TA_n^{-T}M_{2j(a)}A_n^{-1}t\\&\le\sup_{\|t\|\le C}\big\|A_n^{-1}t\big\|^4\sum_{a=1}^p\Big(\sum_{i=1}^n\lambda_{max}\big(M_{2i(a)}\big)\Big)^2\\&\le C^4n\lambda_{max}^2\big(A_n^{-T}A_n^{-1}\big)\sum_{a=1}^p\sum_{i=1}^n\lambda_{max}^2\big(M_{2i(a)}\big).\end{aligned}$$

Putting all these together, we have

$$\begin{aligned}\mathbb{E}\Big[\sup_{\|t\|\le C}\big\|p^{-1/2}A_n^{-1}R_n(\tilde\theta_*,t)\big\|\Big]^2&=p^{-1}\mathbb{E}\Big[\sup_{\|t\|\le C}A_n^{-1}R_n(\tilde\theta_*,t)\Big]^2\\&\le p^{-1}a_n^{-2}\mathbb{E}\Big[\sup_{\|t\|\le C}\big\|R_n(\tilde\theta_*,t)\big\|\Big]^2\\&=O\big(p^{-1}a_n^{-2}\big)\mathbb{E}\Big[\sup_{\|t\|\le C}\big\|R_n(\tilde\theta_*,t)\big\|\Big]^2\\&=O\big(p^{-1}na_n^{-6}\big)\sum_{a=1}^p\sum_{i=1}^n\mathbb{E}\lambda_{max}^2\big(M_{2ni(a)}\big)\\&=o(1),\end{aligned}$$

using (C.4).

Since we have defined

$$S_n(t)=p^{-1/2}A_n^{-1}\big(\sum_{i=1}^n\psi_{0i}(\theta_*+A_n^{-1}t)-\sum_{i=1}^n\psi_{0i}\big)-p^{-1/2}\Gamma_{1n}^{-1}\Gamma_{0n}t,$$

we have

$$p^{-1/2}A_n^{-1}\sum_{i=1}^n\psi_{0i}(\theta_*+p^{1/2}A_n^{-1}t)=S_n(t)+p^{-1/2}A_n^{-1}\sum_{i=1}^n\psi_{0i}+A_n^{-1}\Gamma_{1n}A_n^{-1}t.$$

Hence

$$\begin{aligned}
&\inf_{\|t\|=C}\Big\{p^{-1/2}t^T\Gamma_{1n}A_n^{-1}\sum_{i=1}^n\psi_{0i}(\theta_*+p^{1/2}A_n^{-1}t)\Big\}\\
&=\inf_{\|t\|=C}\Big\{t^T\Gamma_{1n}S_n(t)+p^{-1/2}t^T\Gamma_{1n}A_n^{-1}\sum_{i=1}^n\psi_{0i}+t^T\Gamma_{1n}A_n^{-1}\Gamma_{1n}A_n^{-1}t\Big\}\\
&\geq \inf_{\|t\|=C}t^T\Gamma_{1n}S_n(t)+p^{-1/2}\inf_{\|t\|=C}t^T\Gamma_{1n}A_n^{-1}\sum_{i=1}^n\psi_{0i}+\inf_{\|t\|=C}t^T\Gamma_{1n}A_n^{-1}\Gamma_{1n}A_n^{-1}t\\
&\geq -C\delta_1^{1/2}\sup_{|t|=C}|S_n(t)|-C\delta_1^{1/2}p^{-1/2}\|A_n^{-1}\sum_{i=1}^n\psi_{0i}\|+C^2\delta_0.
\end{aligned}$$

The last step above utilizes facts like $a^Tb\geq-\|a\|\|b\|$. Consequently, defining $C_1=C\delta_0/\delta_1^{1/2}$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\inf_{\|t\|=C}\Big\{t^T\Gamma_{1n}A_n^{-1}\sum_{i=1}^n\psi_{0i}(\theta_*+p^{1/2}A_n^{-1}t)\Big\}<0\Big]\\
&\leq\mathbb{P}\Big[\sup_{\|t\|=C}|S_n(t)|+|A_n^{-1}\sum_{i=1}^n\psi_{0i}|>C_1\Big]\\
&\leq\mathbb{P}\Big[\sup_{\|t\|=C}\big\|S_n(t)\big\|>C_1/2\Big]+\mathbb{P}\Big[\big\|A_n^{-1}\sum_{i=1}^n\psi_{0i}\big\|>C_1/2\Big]<\epsilon,
\end{aligned}$$

for all sufficiently large n, using (A.7) and (A.8). This implies that with a probability greater than $1-\epsilon$ there is a root T_n of the equations $\sum_{i=1}^n\psi_{0i}(\theta_*+A_n^{-1}t)$ in the ball $\{\|t\|<C\}$, for some $C>0$ and all sufficiently large n. Defining $\hat{\theta}_*=\theta_*+A_n^{-1}T_n$, we obtain the desired result. Issues like dependence on ϵ and other technical details are handled using standard arguments, see Chatterjee & Bose (2005) for related arguments.

Since we have $\sup_{\|t\|<C}\|S_n(t)\|=o_P(1)$, and T_n lies in the set $\|t\|<C$, define $-R_n=S_nT_n=o_P(1)$. Consequently

$$\begin{aligned}
-R_n&=S_nT_n\\
&=p^{-1/2}A_n^{-1}\big(\sum_{i=1}^n\psi_{0i}(\theta_*+A_n^{-1}T_n)-\sum_{i=1}^n\psi_{0i}\big)-p^{-1/2}\Gamma_{1n}^{-1}\Gamma_{0n}T_n\\
&=p^{-1/2}A_n^{-1}\sum_{i=1}^n\psi_{0i}-p^{-1/2}\Gamma_{1n}^{-1}\Gamma_{0n}T_n.
\end{aligned}$$

Thus,

$$T_n=-\Gamma_{0n}^{-1}\Gamma_{1n}A_n^{-1}\sum_{i=1}^n\psi_{0i}+p^{1/2}\Gamma_{0n}^{-1}\Gamma_{1n}R_n=-\Gamma_{0n}^{-1/2}\sum_{i=1}^n\Psi_{0i}+p^{1/2}\Gamma_{0n}^{-1}\Gamma_{1n}R_n.$$

Note that our conditions imply that for any c with $\|c\|=1$, we have that c^TT_n has two terms, where $\mathbb{V}\big(-c^T\Gamma_{0n}^{-1/2}\sum_{i=1}^n\psi_{0i}\big)=1$ and

$$\mathbb{E}\big[p^{1/2}c^T\Gamma_{0n}^{-1}\Gamma_{1n}R_n\big]^2=O(1),$$

using (C.2). Using (C.5) we also have that for any c with $\|c\|=1$, c^TT_n converges in distribution to $N(0,1)$. This completes the proof. □

We now have a parallel result on consistency of the GBS resampling scheme. The essence of this theorem is that under the same set of conditions, several resampling schemes are consistent resampling procedures to implement the e-values framework.

Theorem A.2. *Assume conditions (C1)-(C5). Additionally, assume that the resampling weights $\mathbb{W}_{rni}$ are exchangeable random variables satisfying the conditions (4.2). Define $\hat{B}_n = \mu_n \tau_n^{-1} \hat{\Gamma}_{0n}^{1/2} \hat{\Gamma}_{1n}^{-1}$, where $\hat{\Gamma}_{0n}$ and $\hat{\Gamma}_{1n}$ are sample equivalents of Γ_{0n} and Γ_{1n}, respectively. Conditional on the data, $\hat{B}_n(\hat{\theta}_{r*} - \hat{\theta}_*)$ converges weakly to the p-dimensional standard Gaussian distribution in probability.*

Under the additional condition (C6), defining $b_n = \mu_n \tau_n^{-1} a_n$, the distributions of $a_n(\hat{\theta}_* - \theta_*)$ and $b_n(\hat{\theta}_{r*} - \hat{\theta}_*)$ converge to the same weak limit in probability.

Proof of Theorem A.2. This proof has steps similar to that of the proof of Theorem A.1, apart from several additional technicalities. We omit the details. □

B. Theoretical proofs

The results in Section 5 hold under more general conditions than in the main paper–specifically, when the asymptotic distribution of $\hat{\beta}_*$ comes from a general class of *elliptic* distributions, rather than simply being multivariate Gaussian.

Following Fang et al. (1990), the density function of an elliptically distributed random variable $\mathcal{E}(\mu, \Sigma, g)$ 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 density function that is non-negative, scalar-valued, continuous and strictly increasing. For the asymptotic distribution of $\hat{\beta}_*$, we assume the following conditions:

(A1) There exist a sequence of positive reals $a_n \uparrow \infty$, positive-definite (PD) matrix $V \in \mathbb{R}^{p \times p}$ and density g such that $a_n(\hat{\theta}_* - \theta_0)$ converges to $\mathcal{E}(0_p, V, g)$ in distribution, denoted by $a_n(\hat{\theta}_* - \theta_0) \rightsquigarrow \mathcal{E}(0_p, V, g)$;
(A2) For almost every dataset $\mathcal{Z}_n$, There exist PD matrices $V_n \in \mathbb{R}^{p \times p}$ such that $\text{plim}_{n \to \infty} V_n = V$.

These conditions are naturally satisfied for a Gaussian limiting distribution.

B.1. Proofs of main results

Proof of Theorem 5.1. We divide the proof into four parts:

1. ordering of adequate model e-values,
2. convergence of all adequate model e-values to a common limit,
3. convergence of inadequate model e-values to 0,
4. comparison of adequate and inadequate model e-values,

Proof of part 1. Since we are dealing with a finite sequence of nested models, it is enough to prove that $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$ for large enough n, when both $\mathcal{M}_1$ and $\mathcal{M}_2$ are adequate models and $\mathcal{M}_1 \prec \mathcal{M}_2$.

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 $\mathcal{M}_1^0, \mathcal{M}_2^0$ that have 0 in all indices outside $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Take some $\theta_{10} \in \Theta_1^0$, which is the parameter space corresponding to $\mathcal{M}_1^0$, and replace its (zero) entries at indices $j \in \mathcal{S}_2 \setminus \mathcal{S}_1$ by some non-zero $\delta \in \mathbb{R}^{p - |\mathcal{S}_2 \setminus \mathcal{S}_1|}$. Denote it by $\theta_{1\delta}$. Then we shall have

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

where $\mathbb{E}_{s1}$ denotes the expectation taken over the marginal of the distributional argument $\mathbb{T}_0$ at indices $\mathcal{S}_1$. Notice now that by construction $\theta_{1\delta} \in \Theta_2^0$, the parameter space corresponding to $\mathcal{M}_2^0$, and since the above holds for all possible δ, we can take expectation over indices $\mathcal{S}_2 \setminus \mathcal{S}_1$ 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 $a_n V_n^{-1/2}(\hat{\theta}_* - \theta_0) \rightsquigarrow \mathbb{T}_0$. Denote $\mathbb{T}_n = [a_n V_n^{-1/2}(\hat{\theta}_* - \theta_0)]$, 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 $t(\theta) = V_n^{1/2}\theta/a_n + \theta_0$ to both arguments of the depth function above. This will keep the depths constant following affine invariance, i.e. $D(t(\theta_{10}), [\hat{\theta}_*]) = D(\theta_{10}, \mathbb{T}_n)$ and $D(t(\theta_{20}), [\hat{\theta}_*]) = D(\theta_{20}, \mathbb{T}_n)$. Since this transformation maps Θ_1^0 to Θ_1, the parameter space corresonding to $\mathcal{M}_1$, we get $\mathbb{E}_{s1} D(t(\theta_{10}), [\hat{\theta}_*]) > \mathbb{E}_{s2} D(t(\theta_{20}), [\hat{\theta}_*])$, i.e. $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$.

Proof of part 2. For the full model $\mathcal{M}_*$, $a_n(\hat{\theta}_* - \theta_0) \rightsquigarrow \mathcal{E}_p(0, V, g)$ by (A1). It follows now from a direct application of condition (B3) that $e_n(\mathcal{M}_*) \to \mathbb{E}(Y, [Y]$ where $Y \sim \mathcal{E}(0_p, V, g)$.

For any other adequate model $\mathcal{M}$, we use (B 1) property of D:

$$D(\hat{\theta}_m, [\hat{\theta}_*]) = D\Big(\hat{\theta}_m - \theta_0, \big[\hat{\theta}_0 - \theta_0\big]\Big), \tag{B.1}$$

and decompose the first argument

$$\hat{\theta}_m - \theta_0 = (\hat{\theta}_m - \hat{\theta}_*) + (\hat{\theta}_* - \theta_0). \tag{B.2}$$

We now have

$$\hat{\theta}_m = \theta_m + a_n^{-1} T_{mn},$$
$$\hat{\theta}_* = \theta_* + a_n^{-1} T_{*n},$$

where T_{mn} is non-degenerate at the indices $\mathcal{S}$, and $T_{*n} \rightsquigarrow \mathcal{E}(0_p, V, g)$. For the first summand of the right-hand side in (B.2) we then have

$$\hat{\theta}_m - \hat{\theta}_* = \theta_m - \theta_0 + R_n, \tag{B.3}$$

where $\mathbb{E}\|R_n^2\| = O(a_n^{-2})$, and θ_{mj} equals θ_{0j} in indices $j \in \mathcal{S}$ and C_j elsewhere. Since $\mathcal{M}$ is adequate, $\theta_m = \theta_0$. Thus, substituting the right-hand side in (B.2) we get

$$\left| D\left(\hat{\theta}_m - \theta_0, \left[\hat{\theta}_* - \theta_0\right]\right) - D\left(\hat{\theta}_* - \theta_0, \left[\hat{\theta}_* - \theta_0\right]\right)\right| \leq \|R_n\|^\alpha, \tag{B.4}$$

from Lipschitz continuity of $D(\cdot)$ given in (B2). Part 2 now follows.

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

$$D(\hat{\theta}_m, [\hat{\theta}_*]) = D\Big(a_n(\hat{\theta}_m - \theta_0), \big[a_n(\hat{\theta}_* - \theta_0)\big]\Big). \tag{B.5}$$

Decomposing the first argument,

$$a_n(\hat{\theta}_m - \theta_0) = a_n(\hat{\theta}_m - \theta_m) + a_n(\theta_m - \theta_0). \tag{B.6}$$

Since $\mathcal{M}$ is inadequate, $\sum_{j \notin \mathcal{S}} |C_j - \theta_{0j}| > 0$, i.e. θ_m and θ_0 are not equal in at least one (fixed) index. Consequently as $n \to \infty$, $\|a_n(\theta_m - \theta_0)\| \to \infty$, thus $e_n(\mathcal{M}) \to 0$ by condition (B4).

Proof of part 4. For any inadequate model $\mathcal{M}_j, k < j \leq K$, suppose N_j is the integer such that $e_{n_1}(\mathcal{M}_{j_1}) < e_{n_1}(\mathcal{M}_*)$ for all $n_1 > N_j$. Part 3 above ensures that such an integer exists for every inadequate model. Now define $N = \max_{k<j\leq K} N_j$. Thus $e_{n_1}(\mathcal{M}_*)$ is larger than e-values of all inadequate models $\mathcal{M}_{j_1}$ for $k < j \leq K$. □

Proof of Corollary 5.2. By construction, $\mathcal{M}_0$ is nested in all other adequate models in $\mathbb{M}_0$. Hence Theorem 4.1 implies $e_n(\mathcal{M}_0) > e_n(\mathcal{M}^{ad}) > e_n(\mathcal{M}^{inad})$ for any adequate model $\mathcal{M}^{ad}$ and inadequate model $\mathcal{M}^{inad}$ in $\mathbb{M}_0$ and large enough n. □

Proof of Corollary 5.3. 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 4.1: leading to a contradiction and thus completing the proof. □

Proof of Theorem 5.4. Corollary 4.2 implies that

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

Now define $\bar{\mathcal{S}}_0 = \{j : e_{rn}(\mathcal{M}_{-j}) < e_{rn}(\mathcal{M}_*)\}$. We also use an approximation result.

Lemma B.1. *For any adequate model $\mathcal{M}$, the following holds for fixed n and an exchangeable array of GB resamples $\mathbb{W}_r$ as in the main paper:*

$$e_{rn} = e_n(\mathcal{M}) + R_r, \quad \mathbb{E}_r|R_r|^2 = o_P(1). \tag{B.7}$$

Using Lemma B.1 for $\mathcal{M}_{-j}$ and $\mathcal{M}_*$ we now have

$$e_{rn}(\mathcal{M}_{-j}) = e_n(\mathcal{M}_{-j}) + R_{rj},$$
$$e_{rn}(\mathcal{M}_*) = e_n(\mathcal{M}_*) + R_{r*},$$

such that $\mathbb{E}_r|R_{r*}|^2 = o_P(1)$ and $\mathbb{E}_r|R_{rj}|^2 = o_P(1)$ 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 $R, R_1 \to \infty$, completing the proof. □

B.2. Proofs of auxiliary results

Proof of Lemma B.1. We decompose the resampled depth function as

$$\begin{aligned} D\Big(\hat{\theta}_{r_1 m}, [\hat{\theta}_{r*}]\Big) &= D\Big(\frac{a_n}{\tau_n}(\hat{\theta}_{r_1 m} - \hat{\theta}_*), \Big[\frac{a_n}{\tau_n}(\hat{\theta}_{r*} - \hat{\theta}_*)\Big]\Big) \\ &= D\Big(\frac{a_n}{\tau_n}(\hat{\theta}_{r_1 m} - \hat{\theta}_m) - \frac{a_n}{\tau_n}(\hat{\theta}_m - \hat{\theta}_*), \Big[\frac{a_n}{\tau_n}(\hat{\theta}_{r*} - \hat{\theta}_*)\Big]\Big). \end{aligned}$$

Conditional on the data, $(a_n/\tau_n)(\hat{\theta}_{r_1 m} - \hat{\theta}_m)$ has the same weak limit as $a_n(\hat{\theta}_m - \theta_m)$, and the same holds for $(a_n/\tau_n)(\hat{\theta}_{r_1 *} - \hat{\theta}_*)$ and $a_n(\hat{\theta}_* - \theta_*)$. Now (B.3) and $\tau_n \to \infty$ combine to give

$$\frac{a_n}{\tau_n}(\hat{\theta}_m - \hat{\theta}_*) \xrightarrow{P} 0,$$

as $n \to \infty$. Lemma B.1 follows directly now. □

C. Details of experiments

Among the competing methods, for stepwise regression there is no tuning. MIO requires specification of a range of desired sparsity levels and a time limit for the MIO solver to run for each sparsity level in the beginning. We specify the sparsity range to be $\{1, 3, \ldots, 29\}$ in all settings to cover the sparsity levels across different simulation settings, and the time limit to be 10 seconds. We select the optimal MIO sparsity level as the one for which the resulting estimate gives the lowest BIC. We use BIC to select the optimal regularization parameter for Lasso and SCAD as well. The Knockoff filters come in two flavors: Knockoff and Knockoff+. We found that Knockoff+ hardly selects any features in our settings, so use the Knockoffs for evaluation, setting its false discovery rate threshold at the default value of 0.1. We shall include these details in the appendix.

D. Details of Indian Monsoon data

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 et al., 2003; Trenberth, 2011; Wang et al., 2005). For example, (Gosswami, 2005) discovered an upward trend in frequency and magnitude of extreme rain events, using daily central Indian rainfall data on a $10^\circ \times 12^\circ$ grid, but a similar study on a $1^\circ \times 1^\circ$ gridded data by (Ghosh et al., 2016) suggested that there are both increasing and decreasing trends of extreme rainfall events, depending on the location. Additionally, (Krishnamurthy et al., 2009) reported increasing trends in exceedances of the 99th percentile

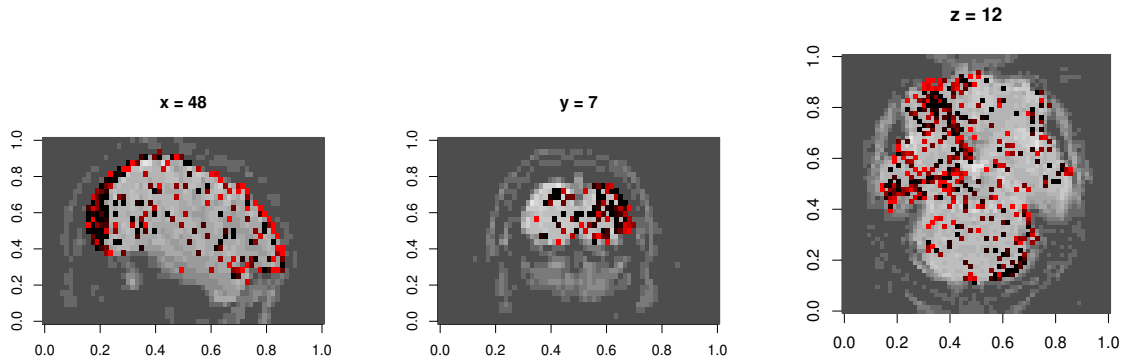


Figure E.1. (Top) Plot of significant p-values at 95% confidence level at the specified cross-sections; (bottom) a smoothed surface obtained from the p-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas

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 for Indian Monsoon (Dietz & Chatterjee, 2014; 2015).

We attempt to identify the driving factors behind precipitation during the Indian monsoon season using our e-value based feature selection technique. Data is 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 35 potential predictors of the Indian summer precipitation:

(A) Station-specific: (from 36 weather stations across India) Latitude, longitude, elevation, maximum and minimum temperature, tropospheric temperature difference (ΔTT), Indian Dipole Mode Index (DMI), Niño 3.4 anomaly;

(B) Global:

- u-wind and v-wind at 200, 600 and 850 mb;
- Ten indices of Madden-Julian Oscillations: 20E, 70E, 80E, 100E, 120E, 140E, 160E, 120W, 40W, 10W;
- Teleconnections: North Atlantic Oscillation (NAO), East Atlantic (EA), West Pacific (WP), East Pacific/North Pacific (EPNP), Pacific/North American (PNA), East Atlantic/Western Russia (EAWR), Scandinavia (SCA), Tropical/Northern Hemisphere (TNH), Polar/Eurasia (POL);
- Solar Flux;
- Land-Ocean Temperature Anomaly (TA).

These covariates are all based on existing knowledge and conjectures from the actual Physics driving Indian summer precipitations. The references provided earlier in this section, and multiple references contained therein may be used for background knowledge on the physical processes related to Indian monsoon rainfall, which after decades of study remains one of the most challenging problems in climate science.

E. Details of fMRI data implementation

Typically, the brain is divided by a grid into three-dimensional array elements called voxels, and activity is measured at each voxel. More specifically, a series of three-dimensional images are obtained by measuring Blood Oxygen Level Dependent (BOLD) signals for a time interval as the subject performs several tasks at specific time points. A single fMRI image typically consists of voxels in the order of 10^5, which makes even fitting the simplest of statistical models computationally intensive when it is repeated for all voxels to generate inference, e.g. investigating the differential activation of brain region in response to a task.

The dataset we work with comes from a recent study involving 19 test subjects and two types of visual tasks (Wakeman & Henson, 2015). Each subject went through 9 runs, in which they were showed faces or scrambled faces at specific time points. In each run 210 images were recorded in 2 second intervals, and each 3D image was of the dimension of $64 \times 64 \times 33$,

which means there were 135,168 voxels. Here we use the data from a single run on subject 1, and perform a voxelwise analysis to find out the effect of time lags and BOLD responses at neighboring voxels on the BOLD response at a voxel. Formally we consider separate models at voxel $i \in \{1, 2, ..., V\}$, with observations across time points $t \in \{1, 2, ..., T\}$.

Clubbing together the stimuli, drift, neighbor and autoregressive terms into a combined design matrix $\tilde{X} = (\tilde{x}(1)^T, ..., \tilde{x}(T)^T)^T$ and coefficient vector θ_i, we can write $y_i(t) = \tilde{x}(t)^T\theta_i + \epsilon_i(t)$. We estimate the set of non-zero coefficients in θ_i using the e-value method. Suppose this set is R_i, and its subsets containing coefficient corresponding to neighbor and non-neighbor (i.e. stimuli and drift) terms are S_i and T_i, respectively. To quantify the effect of neighbors we now calculate the corresponding F-statistic:

$$F_i = \frac{(\sum_{n \in S_i} \tilde{x}_{i,n}\hat{\theta}_{i,n})^2}{(y_i(t) - \sum_{n \in T_i} \tilde{x}_{i,n}\hat{\theta}_{i,n})^2} \frac{|n - T_i|}{|S_i|},$$

and obtain its p-value, i.e. $P(F_i \geq F_{|S_i|,|n-T_i|})$.

Figure E.1 shows plots of the voxels with a significant p-value from the above F-test, with a darker color associated with lower p-value, as opposed to the smoothed surface in the main paper. Most of the significant terms were due to the coefficients corresponding to neighboring terms. A very small proportion of voxels had any autoregressive effects selected (less than 1%), and most of them were in regions of the image that were outside the brain, indicating noise.

In future work, we aim to expand on the encouraging findings and repeat the procedure on other individuals in the study. An interesting direction here is to include subject-specific random effects and correlate their clinical outcomes (if any) to observed spatial dependency patterns in their brain.