Submitted to the Annals of Statistics

GENERAL MODEL DISCOVERY USING STATISTICAL EVALUATION MAPS

By Subhabrata Majumdar and Snigdhansu Chatterjee

University of Minnesota

On any given dataset, a very wide variety of statistical models may be applicable, based on experience and tradition, robustness and sensitivity requirements, algorithmic, computational or philosophical considerations, risk and eventual usage. We propose a technique to compare such models in a very general framework. We establish that under general conditions, statistical models that adequately explain properties of the data can be well separated from those that do not. Our resampling-based approach achieves concurrent ranking of models and consistent approximation of sampling distribution of parameter estimators under any model, thus enabling inference within each model. Consequently, our proposal is one of simultaneous model discovery and inference. For traditional covariate selection problems where there are p covariates, our proposal results in a fast and parallel algorithm that fits only a single model and evaluates $p+1$ models, 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 or competitively with currently used methods for model selection. We use our procedure to elicit climatic drivers of Indian monsoon precipitation.

MSC 2010 subject classifications: Primary 62G09, 62G20; secondary 62F12, 62F40s

Keywords and phrases: Model discovery, Simultaneous model scoring and inference, Resampling, Variable selection, Data depth, Indian monsoon

CONTENTS

1. Introduction. In a typical statistical or data science exercise, both *data* and a *statistical model* is involved. While there is often little or no ambiguity about data, there can be many alternatives about how to analyze such data, and how to interpret the results, which broadly constitute the realm of statistical models. In this paper, we interpret the term *statistical model* very broadly. We recognize various possible transformations of the data, different model fitting algorithms, practical safeguards put in place to ensure robustness and sensitivity balance in the results, different methods of data analysis, different statistical paradigms of interpretation of results, as all equally deserving to be considered as crucial components of a statistical model. The example below illustrates this idea.

EXAMPLE 1.1 (Tree data). Consider the data contained in `data(trees)` in the statistical software `R`. There are 31 observations, on tree volume, height, girth. The observed data is (X_{i1}, X_{i2}, Y_i) denoting the vector of tree girth, height and volume, for $i = 1, \ldots, n$. We denote $p = 2$ for the two explanatory variables *tree girth* and *height*, used to explain the properties of the response variable *volume*.

Define the Box-Cox transformation (Box and Cox, 1964) on the response variable as $C(y, \lambda) = \log(y)\mathcal{I}_{\{\lambda=0\}} + y^{\lambda}\mathcal{I}_{\{\lambda\neq 0\}}$. We assume that Y_i's in the data are related to the other variables according to the statistical relation

$$C(Y_i, \lambda) = \beta_0 + \beta \log(X_{i1}) + \beta \log(X_{i2}) + e_i. \tag{1.1}$$

Here $\{e_i\}$ is a sequence of random variables, and we assume that $\mathbb{E}e_i = 0$ and $\mathbb{E}e_i^2 = \sigma_i^2 < \infty$. The parameters here are $\theta = (\lambda, \beta_0, \beta_1, \beta_2, \sigma_1^2, \ldots, \sigma_n^2) \in \mathbb{R}^{p_{nmax}}$ where $p_{nmax} = n + 4$.

Even in this rather simple framework, we can imagine several *statistical models* as being *per se* equally interesting or important. These include (i) the Gauss-Markov linear regression model with $\lambda = 0$ where the errors $\{e_i\}$ are independently and identically distributed (i.i.d. hereafter) with a Gaussian distribution with mean zero and variance σ^2, that is, $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, (ii) the Gauss-Markov linear regression model with any other fixed, non-random value of λ, (iii) a model where λ is estimated form data but then a Gauss-Markov linear regression model used for the rest of the analysis ignoring the randomness in the estimated λ, (iv) using a fixed λ value like 0 or 1, then using *ordinary least squares* (OLS) method to estimate regression parameters, followed by inference based on the residual bootstrap (see Efron (1979); Efron and Tibshirani (1993); Shao and Tu (1995)) that assumes homoscedasticity, (v) using robustness-driven M-estimation techniques for simultaneous estimation of $(\lambda, \beta_0, \beta_1, \beta_2)$, followed by a *wild bootstrap* resampling scheme

for statistical inference (Wu, 1986; Mammen, 1993), which provides robustness against heteroscedasticity. More general frameworks, for example, where $\mathbb{E}[Y_i|X_{i1}, X_{i2}] = \mu(X_{i1}, X_{i2})$ where the functional form of $\mu : \mathbb{R}^2 \to \mathbb{R}$ is unknown, or where a Bayesian framework is adopted and different choices of priors are made, are other examples of more general statistical models.

We submit that these are all plausible models, important from one or more considerations. Some like (iii) reflect tradition, others like (v) reflect desirable caution coupled with modern computational power. This list is far from exhaustive, for example, in (iv) each alternative resampling scheme may be called a separate model, and each choice of prior in a Bayesian framework is also a separate model.

□

The above list of possible models is far from exhaustive, but serves to illustrate the fact that statistical models arise in most of the standard procedures of data analysis, be it from classical Statistics, robustness considerations, Bayesian paradigm, risk management perspective, Occam's razor, or combinations thereof. Such models typically differ from each other in many ways, and not just in the number of covariates, or number of parameters to estimate. Often, as in the case of the heteroscedastic model coupled with resampling-based inference above, a very classical approach towards modeling or model selection, or a selection based only on a superficial reading of parsimony, can lead to leaving out greatly versatile models on both robustness and efficiency counts. In this paper, we address the problem of elicitation of suitable models for analyzing data in a very general framework. We consider candiate models that need not be nested, or philosophically or otherwise compatible with each other.

In Section 2 we provide details of the general framework that we use in this paper. Our primary goal is a clear separation of the candidate models into two groups: those that adequately explain some user-defined characteristics exhibited in the data, which we designate *adequate models*, and those that do not (inadequate models), coupled with simultaneous ranking of models and valid inference within each model. Section 3 contains a technical definition of model adequacy, as well as a generic description of a baseline model, which we call the *preferred model*. The preferred model may be the most complex candidate model (e.g. the model with all covariates in a regression problem), a model in popular or current use, a hypothesized model, or a model with known parsimony or computational advantage. Each candidate model has its own set of unknown parameters, which are estimated using a model-specific optimization framework. Since our notion of a statistical

model is quite general, we map all models to a common Euclidean reference frame $\mathbb{R}^{d_n}$, using user-defined functions. This map may be just a listing of parameters (estimated or treated as constant) when the models are nested, predictions from the models for in-sample or out-of-sample cases, or other characteristics that may be of interest, that can be high-dimensional in nature. Thus, there is scope of evaluating models based on domain-knowledge preference, potential risks of various kinds in its usage, or standard statistical measures of skill.

After mapping each candidate model to $\mathbb{R}^{d_n}$, we propose using a function called the *evaluation map*, which compares each candidate model against the *preferred model*. An evaluation map typically compares the distribution of estimated characteristics of interest from any candidate model and the preferred model, and data-depth functions (Zuo and Serfling (2000)) are special cases of the kind of functions that may act as an evaluation map.

After this we introduce a quantity called the ***e-value***, which we define as a non-negative summary statistic of the distribution of the evaluation map of a candidate model. The *e-value* of a model is a measure of how well a candidate model explains the interesting features of the data, as defined by the user-specified functions that bring it to the common Euclidean space $\mathbb{R}^{d_n}$, where the preferred model acts as a gold standard. Under very general conditions, the *e-values* for inadequate models asymptotically tends to zero, while for adequate models they tend to be bounded away from zero, thus separating "good" models from "bad" ones. Moreover, under suitable conditions and methodological choices, the *e-values* of adequate models can rank the models in terms of parsimony. Thus, if the data generating process is one of the candiate models, its *e-value* is established to be highest asymptotically under general conditions. However, we also allow the possibility that none of the candidate models, including the preferred model, adequately explain the properties of the data at hand. In such cases, only the preferred model will have a high score. Our proposal thus includes the provision for triggering a re-evaluation of models and data based on scientific caution, when only the preferred model achieves a significantly non-zero score.

Since our models can be quite general, in Section 2.5 we discuss how parameters are defined and estimated in our framework. Resampling in such general framework is proposed, and in Section 4 we establish (i) consistency and asymptotic normality of our parameter estimators, (ii) consistency of the proposed resampling procedure, (iii) consistent estimation of *e-values*. Our resampling technique is extremely parallelizable, and multiple models may be considered simultaneously. Thus, we formulate a unified system where resampling elicits both the *e-value* of any model, along with the joint

sampling distribution of all its parameter estimators.

In recent times, there is a growing concern about statistical inference after the implementation of a model selection step. Discussions and several interesting results relating to this matter may be found in Yang (2005); Leeb and Pötscher (2005); Chang, Huang and Ing (2014); Tibshirani et al. (2015, 2016) and several references therein. In this paper, we propose obtaining consistent resampling-based distributions of the estimators of *all* parameters from *all* candidate models. Thus in our framework, statistical inference is not the usual two-step procedure where the first step involves selection of a model, and the second step of actual inference somehow adjusts for the uncertainties of the first step. Our proposal is one of a *joint selection and inference* procedure, where the consistent resampling-based approximations of the sampling distributions of any collection of models are simultaneously used for inference, as well as establishing an *e-value* of a model, which may be used to preferentially treat a subset of models.

Naturally, the traditional model selection problems of identifying necessary covariates in linear regression or choosing the lag-order in autoregression, are special cases of our framework. In such problems, there is a maximum number of parameters p_{nmax}, and various candidate models consider subsets of a common set of p_{nmax} parameters. The candidate models can be arranged in a lattice, with the supremum being the *least parsimonious* or complete model that involves all p_{nmax} parameters. There are $2^{p_{nmax}}$ such models, and a full evaluation ought to consider all of these, which is an NP-Hard problem (Natarajan, 1995). Owing to the computational challenge involved, especially in older environments and paradigms of computing, various algorithms to reduce computations have been proposed, which compromise optimality and other properties of the model selection procedure. In this traditional model selection context, in Section 5 we propose a fast and parallel model selection algorithm, where only the least parsimonious model, which we consider as the preferred model in this case, is estimated from the data, and only a total of $p_{nmax}+1$ models are evaluated. One of the evaluated models is the preferred model, while the other p_{nmax} models are those where only a single explanatory variable is dropped. These latter p_{nmax} model evaluations can be implemented by parallel computations extremely quickly. The final step of the proposed algorithm is simple, if the model where the j-th parameter is dropped achieves an *e-value* that is lower than that of the preferred model, it is included in the finally selected model. We establish that the model selected using this algorithm is the most parsimonious model that fits the data, termed as the "true model" in many studies, with probability tending to one. We do necessarily advocate selecting the

most parsimonious model that fits the data irrespective of other considerations, and urge caution and evaluation of domain scientific principles and purpose before selecting any model.

Throughout this paper, we allow several quantities, like the number of parameters in each candidate model (denoted by p_{sn} for model $\mathcal{S}_n$) or the number of characteristics of interest (d_n) on which the evaluation map is computed, to tend to infinity with sample size. This *dimension asymptotics* ($p_{sn} \to \infty$ as $n \to \infty$) approach allows any candidate model to have increasing parameter dimensionality, which imitates the reality of the scientific discovery process where additional data is often used in conjunction with more fine-tuned or insightful models. Similarly, allowing the number of characteristics used for comparing models to grow with the sample size ($p_{sn} \to \infty$) reflects the scientific process. Throughout this paper, for theoretical purposes we adopt a framework involving a triangular array of models and parameters, where various parameter values and dimensions and even estimation and model evaluation procedures are allowed to change with sample size. This is partially for the same reason of being in tune with the reality of scientific discovery process, but also for additional theoretical advantages that such a framework offers, and for the purpose of being inclusive of techniques like local asymptotics, uniform convergence and several others that will form part of our future work.

Then, in Section 6 we present two illustrative examples on how our fast model selection algorithm is implemented, and its relative performance in covariate selection problems. This is followed in Section 7 with a study on the possible determinants of precipitation in the Indian sub-continent, during the summer monsoon months. The nature and drivers of Indian monsoon precipitation is one of the very challenging problems in climate studies. In view of the very large number of human beings and other species that depend on these monsoon precipitation for their livelihood, food, water and energy supply, and general survival and well-being, understanding the physical systems that contribute to this precipitation is crucial, especially in the context of a changing climate regime. Our study contributes to this goal by identifying several important climate factors. In Section 8 we discuss several caveats, future research plans and including some concluding comments. In Section A we collect the proofs of several of our theoretical results.

1.1. *Notations and conventions.* In the rest of this paper, we use the notations $|a|$ to denote the Euclidean norm of a vector a, a^T to denote the transpose of the (vector or matrix) a. All vectors are column vectors in this paper. The notation $\lambda(A)$ is used to denote a generic eigenvalue of a real,

symmetric matrix A, and similarly $\lambda_{max}(A)$ and $\lambda_{min}(A)$ are respectively used to denote the maximum and minimum eigenvalue of A. When A and B are square matrices of identical dimensions, the notation $B < A$ implies that the matrix $A - B$ is positive definite. We use the notation $\mathbb{I}_d$ for the $d \times d$ identity matrix for any positive integer d. The notation $\mathcal{I}_{\mathcal{A}}$ is the indicator function of statement $\mathcal{A}$, that is, it takes the value 1 if $\mathcal{A}$ is true and zero otherwise. The notation C, with or without subscripts and in both upper and lower case, will be used as generic for constants, without any implication that such constants are identical in all instances they occur. The notation $\xrightarrow{p}$ and $\xRightarrow{\mathcal{D}}$ denote convergence in probability and convergence in distribution. The notation $a_n \asymp b_n$ implies that $a_n = O(b_n)$ as well as $b_n = O(a_n)$. We define the Hilbert space $\ell_2 = \{\{x_n, n = 1, 2, \ldots\} : x_n \in \mathbb{R}, \sum_{n\geq 1} x_n^2 < \infty\}$, and embed finite dimensional real Euclidean space $\mathbb{R}^q$ in it as and when necessary, as the first q elements of ℓ_2. We define $\tilde{\ell}_2$ as a collection of probability measures on ℓ_2. For any metric space $\mathcal{A}$, given a point $x \in \mathcal{A}$ and a set $A \subseteq \mathcal{A}$, the distance between x and A is given by $d(x, A) := \inf_{a\in A} |x - a|$. For any function h of the parameters in any model, we will often simplify notations by using $h \equiv h_{sn} \equiv h(\theta_{sn})$, $\hat{h} \equiv \hat{h}_{sn} \equiv h(\hat{\theta}_{sn})$, $\hat{h}_r \equiv \hat{h}_{rsn} \equiv h(\hat{\theta}_{rsn})$.

In the proofs, the notation R, typically with various subscripts like R_n, R_{sn}, R_{rsn} and so, are used as generic for remainder terms, which contribute asymptotically negligible terms in our results. While we sometimes include algebraic details, often the tedious algebra behind moment calculations and probabilistic bound computations is omitted to contain this paper to a reasonable length and preserve clarity. However, our technical conditions are always comprehensive and explicit, and such algebraic computations can be easily carried out without much intellectual effort. In designing the technical conditions for the theoretical properties in this paper, we have striven for simplicity and not on minimal requirements. Thus, the various assumptions made in this paper are often sufficient conditions, rather than necessary ones, for the theoretical results.

2. The general framework.

2.1. *The frame of models.* In any statistical model, each parameter in each model has an assigned role. A parameter may be a constant related to the scientific process, or a tuning constant related to a computational procedure or a prediction algorithm, or may perform some other function. Examples of the former in Example 1.1 are the regression slope parameters β_1 and β_2, which quantify how the volume of wood in a tree changes with its height or girth. An example of the latter in the same context can be the parameter λ, or a tolerance or iteration limits of an iterative model fitting

procedure. Parameters can have similar roles in many models, for example, the regression coefficients β_1 and β_2 in Example 1.1 are used in all the listed models in that example. We use these general facts to describe the *frame of models* that we use in this paper.

We consider a context where the union of all parameters from all candidate models forms a countable set. Naturally, problems where the number of parameters are finite, as in a majority of statistical applications, are included in our framework. We exclude all constants that are invariant across candidate models from this count, or any unknown quantity that is not estimated in any model and is not used subsequently. The parameters across all models are laid out in any arbitrary but fixed fashion indexed by the set of integers $\{1, 2, \ldots\}$.

EXAMPLE 2.1 (Example 1.1 continued). Here, we may consider $p_{nmax} = n + 4$ as the maximum number of parameters in the system, and we may denote the p_{nmax}-dimensional vector of parameters with the generic notation

$$\begin{aligned} \theta_n &= (\lambda, \beta_0, \beta_1, \beta_2, \sigma_1^2, \ldots, \sigma_n^2) \\ &= \quad (\theta_{n1}, \theta_{n2}, \ldots, \theta_{np_{nmax}}) \text{ notationally.} \end{aligned}$$

□

We now associate a candidate model $\mathcal{M}_n$, either from a scientific discovery process or a hypothesis testing process, with two quantities:

(a) The set $\mathcal{S}_n = \{j_1, \ldots, j_{p_{sn}}\} \subseteq \{1, 2, \ldots\}$ of indices where the parameter values are unknown and estimated from the data; and
(b) An ordered vector of known constants $C_{sn} = (C_{snj} : j \notin \mathcal{S}_n)$ for parameters not indexed by $\mathcal{S}_n$.

For any n the sets $\mathcal{S}_n$ are finite, thus each model may include only a finite number of unknown real-valued constants.

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

$$\theta_{mnj} = \begin{cases} \text{Unknown } \theta_{mnj} \in \Theta_{mnj} & \text{for } j \in \mathcal{S}_n \\ \text{Known } C_{snj} \in \Theta_{mnj} & \text{for } j \notin \mathcal{S}_n. \end{cases} \tag{2.1}$$

Each $\Theta_{mnj} \subseteq \mathbb{R}$, thus all parameters are real-valued. It may be noted that in most cases, simple re-parametrization can be used to define models in a way such that the known constants in C_{sn} are all zero.

We assume that at stage n, we have a *preferred model*, which we denote by $\mathcal{M}_{*n}$, identified with the set of indices $\mathcal{S}_{*n} = \{j_{*1}, \ldots, j_{p_{*n}}\} \subseteq \{1, 2, \ldots\}$ having p_{*n} elements and known constants C_{*n}. We denote the unknown parameters and constants for this preferred model with notations similar to those of the generic model $\mathcal{S}_n$, but by using the subscript $_*$ in place of the subscript $_s$. Thus, in the preferred model we consider the p_{*n}-dimensional vector of unknown constants $\widetilde{\theta}_{*n} = (\theta_{*nj_1}, \theta_{*nj_2}, \ldots, \theta_{*nj_{psn}}) \in \prod_{j \in \mathcal{S}_{*n}} \Theta_{nj}$, and the ordered vector of known constants $C_{*n} = (C_{*nj} : j \notin \mathcal{S}_{*n})$. These are arranged in the generic parameter vector $\theta_{*n} \in \Theta_n$ given by

$$\theta_{*nj} = \begin{cases} \text{Unknown } \theta_{*nj} & \text{for } j \in \mathcal{S}_{*n} = \{i_1, \ldots, i_{p_{*n}}\}, \\ \text{Known } C_{*nj} & \text{for } j \notin \mathcal{S}_{*n}. \end{cases}$$

Depending on the context, the preferred model may relate to a hypothesized model, or the most complex or the most simple model, or relate to the current state of the art, or a "gold standard", or be "preferred" by some other criteria. Note that the *preferred model* is just one of the candidate models, and its usage will be clear later on in this paper.

2.2. *Transformation to common platform.* Suppose $G_{mn} : \Theta_{mn} \to \mathbb{R}^{d_n}$ is a known transformation to map parameters from model $\mathcal{M}_n$ to $\mathbb{R}^{d_n}$. While the candidate models may be very diverse and may relate to different physical realities, theories or hypotheses, computational or data analytic choices, the Euclidean space $\mathbb{R}^{d_n}$ is a common ground where all models may be compared. We use the notation G_{*n} for the transformation of the preferred model. The functions G_{mn} may involve random elements, however in such cases their distributions are restricted to functions involving model $\mathcal{S}_n$ and $\mathcal{S}_{*n}$ only. In principle, each G_{mn} can also be designed to map to some set $\mathcal{G}_n \neq \mathbb{R}^{d_n}$. However, in such cases we would have to address technical issues relating to topological, measure-theoretic and geometric or algebraic properties of $\mathcal{G}_n$ while studying theoretical results, which may be considered avoidable since the statistical gets to choose the maps G_{mn}. Consequently, we assume that the co-domain of each map G_{mn} is $\mathbb{R}^{d_n}$ in this paper, and avoid unnecessary mathematical complications.

The choice of the function G_{mn} may depend on the purpose for building the scientific model. This transformation allows us to consider the *science case* where the actual parameter values and their interpretation is subject to scrutiny, or *use cases* like prediction and classification problems.

Example 2.2 (Example 1.1 continued). In Example 1.1 we may be interested in *covariate selection*, where we consider which subset of regressors

(X_i's) have an influence on the response (Y_i's in both example). Generally when there are p-regressors, there are 2^p possible models on covariate choices alone. In Example 1.1 where there are three regression parameters $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$, we have 8 possible models, even when we assume all other properties of the data as given, say in the Gauss-Markov structure. Suppose we consider the model where the entire regression coefficient vector $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$ is estimated as the preferred model. We may use $G_{*n}(x) = x$ that takes $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$ to itself. For the submodel with no intercept term, we use $G_{mn}((x_0, x_1, x_2)) = ((0, x_1, x_2)) \in \mathbb{R}^3$, and so on. □

EXAMPLE 2.3 (Example 1.1 continued). In the context of the same problem, we may also consider two alternative regression strategies:

1. Semiparametric regression model:

$$C(Y_i, \lambda) = X_{i1}\beta_1 + g(X_{i2}) + \epsilon_i,$$

   for an unknown function g;
2. Semiparametric single index model:

$$C(Y_i, \lambda) = h(X_{i1}\beta_1 + X_{i2}\beta_2) + \epsilon_i,$$

   for some unknown function h.

When these two kinds of models above are considered, in addition to the linear regression models, comparing and choosing between them becomes tricky. In this case it is possible consider all modelling methods as special cases of a general model: $C(Y_i, \lambda) = h(X_{i1}\beta_1 + g(X_{i2})) + \epsilon_i$, such a representation is unintuitive, prohibitively difficult for computations and may require very strong assumptions for theoretical justification. A more interpretable platform in this scenario can be based on the predicted value of responses, and one can simply take as G_{mn} the vector of fitted values obtained in each method. □

We use the notations $G_{mn}(\hat{\theta}_{mn}) = \hat{G}_{mn}$ and $G_{*n}(\hat{\theta}_{*n}) = \hat{G}_{*n}$. Let $\tilde{\mathcal{G}}_{mn}$ denote the set of probability measures on $\mathbb{R}^{d_n}$, and we denote the probability measure extended by a generic $\hat{G}_{mn}$ by $[\hat{G}_{mn}]$. Thus for example, the expected value of $\hat{G}_{mn}$ is given by $\int t d[\hat{G}_{mn}(t)]$.

2.3. *Statistical evaluation maps.* We now introduce another function, the *statistical evaluation function* $E_n : \mathbb{R}^{d_n} \times \tilde{\mathcal{G}}_n \to [0, \infty)$, which takes as arguments a point from $\mathbb{R}^{d_n}$ and a probability measure from $\tilde{\mathcal{G}}_n$, and maps that pair into non-negative real numbers. Roughly, the quantity $E_n(y, [Y])$ is a

measure of where exactly does the point y sit with respect to the distribution of the random variable $Y \in \mathbb{R}^{d_n}$. The exact nature of the evaluation function, which will make this rough notion precise, depends on the context. Good examples of evaluation functions are probabilities of convex sets under suitable distributions, unimodal probability density functions that uniformly decrease away from the mode in any direction, and various *data-depth* functions (Zuo and Serfling, 2000). In fact, the latter is a very rich collection of relevant functions: although their properties are somewhat more restrictive than those required of an evaluation map, and performance less satisfactory in simulation studies compared to less restrictive evaluation maps.

2.4. *The e-value of models.* We now associate with each model $\mathcal{M}_n$ a functional of the evaluation map E_n: which we call the e-value. An example of e-value is the mean evaluation map function:

$$e_n(\mathcal{M}_n) = \mathbb{E}E_n\left(\hat{G}_{mn}, [\hat{G}_{*n}]\right) \tag{2.2}$$

which we concentrate on for the rest of the paper. However, any other functional of $E_n(\hat{G}_{mn}, [\hat{G}_{*n}])$ may also be used here, and a large proportion of our theoretical discussion in the rest of the paper is applicable to any smooth functional of the distribution of $E_n(\hat{G}_{mn}, [\hat{G}_{*n}])$. Furthermore, the distribution of $E_n(\hat{G}_{mn}, [\hat{G}_{*n}])$ 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.

REMARK 2.1. From a hypothesis testing perspective, e-values generalize the concept of p-values. For example, suppose the hypothesis test uses the test statistic T_{0n}, whose realized value in the data at hand is t_n, and whose distribution under the null is given by $[T_{0n}]$. Also suppose the test rejects the null for high values of T_{0n}. Define the evaluation map $E_n(y, [Y]) = \mathcal{I}_{Y>y\}}$. Notice that both the e-value and the p-value are given by $\mathbb{E}_0 E_n(t_n, [T_{0n}])$, where $\mathbb{E}_0$ stands for distribution under the null distribution. Notice in this case, the null model is the preferred model. Further discussion of the relationships between *e-values* and hypothesis tests are postponed to a future paper. □

There are two random quantities involved in the expression of $e_n(\mathcal{M}_n)$ above, namely $\hat{G}_{mn}$ and $\hat{G}_{*n}$. Typically, the distribution of either of these random quantities are not known, and have to be elicited from data. We shall use resampling methods for this purpose, the details of which are discussed below. Also, note from the expression of $e_n(\mathcal{M}_n)$ above that we *do*

not require the joint distribution of these random variables, but only their marginals. However, our methodology outlined below is adequate to obtain joint distributions of unknown parameters from several models simultaneously.

2.5. *Method of parameter estimation.* Since some or all the parameter values are unknown in a typical scientific problem, they have to be *estimated* from empirical observations. Suppose at stage n, the empirical data we have at hand is denoted by the set $\mathbf{Y} = \{Y_{n1}, \ldots, Y_{nk_n}\}$, where we do not restrict either the dimension of any of the Y_{ni}'s, or declare any properties or restrictions on them. In particular, each Y_{ni} may be infinite dimensional element, or a finite dimensional vector. The size of $\mathbf{Y}$, which we call the *sample size* and denote by k_n is assumed to be a non-decreasing sequence of integers that tends to infinity as $n \to \infty$.

In model $\mathcal{M}_n$, only the unknown elements of the parameters θ_{mn} are estimated from data. Hence for ease of exposition, we designate the subvector of θ_{mn} at indices $\mathcal{S}_n$ by θ_{sn}. In the generic model $\mathcal{M}_n$, the vector of unknown parameters θ_{sn} is the unique minimizer of

$$\Psi_{sn}(\theta) = \mathbb{E}\sum_{i=1}^{k_n} \Psi_{sni}(\theta, Y_{ni}), \tag{2.3}$$

where $\Psi_{sni}(\cdot)$ are a known triangular array of functions, for which we borrow the terminology *energy function* from optimization and related literature. Such functions have also been called *contrast functions*, see Pfanzagl (1969); Michel and Pfanzagl (1971); Bose and Chatterjee (2003). The estimator $\hat{\theta}_{sn}$ of θ_{sn} is obtained as a minimizer of the sample analog of the above, thus, $\hat{\theta}_{sn}$ minimizes

$$\hat{\Psi}_{sn}(\theta) = \sum_{i=1}^{k_n} \Psi_{sni}(\theta, Y_{ni}). \tag{2.4}$$

The *preferred model* is described in an identical way, thus the vector of unknown parameters θ_{*n} is the unique minimizer of

$$\Psi_{*n}(\theta) = \mathbb{E}\sum_{i=1}^{k_n} \Psi_{*ni}(\theta, Y_{ni}), \tag{2.5}$$

where $\Psi_{*ni}(\cdot)$ are a known triangular array of functions. The estimator $\hat{\theta}_{*n}$ of θ_{*n} is obtained as a minimizer of the sample analog of the above, thus,

$\hat{\theta}_{*n}$ minimizes

$$\hat{\Psi}_{*n}(\theta) = \sum_{i=1}^{k_n} \Psi_{*ni}(\theta, Y_{ni}). \tag{2.6}$$

We allow for the possibility that a candidate model has no unknown parameters, that is, all its parameters are known and do not need to be empirically estimated.

While we assume here that the parameters θ_{sn} and their estimators have a bijective map to a subset of an Euclidean space, with little or no modifications, many parts of the developments below extend to general metric spaces or to infinite-dimensional vector spaces, though we do not explore such generalizations here. We assume in particular, that a distance metric exists on any parameter space under consideration, and we use the notation $d(x, y)$ to denote the distance between two elements x and y on such space.

We now state the conditions we assume on the energy functions for the rest of this paper. Note that several sets of alternative conditions can be developed, to address various parameter estimations techniques prevalent in Statistics. However, in order to contain the length of this paper and to preserve clarity, we only address the case where the energy function $\Psi_{sni}(\cdot, \cdot)$ is smooth in the first argument. This case covers a vast number of models routinely considered in statistics.

Henceforth, we occasionally drop the subscript $_s$ and $_*$ when there is no scope for confusion for notational simplicity, since the developments presented in the rest of this section are applicable to any model. We often drop the second argument from estimating functionals, thus for example $\Psi_{ni}(\theta) \equiv \Psi_{sni}(\theta, Y_{ni})$. Other notational simplifications in various contexts of this section, will be presented as related contexts arise.

In a neighborhood of θ_{sn}, the functions Ψ_{sni} are thrice continuously differentiable in the first argument, with the successive derivatives being denoted by Ψ_{ksni}, $k = 0, 1, 2$. That is, there exists a $\delta > 0$ such that for any $\theta = \theta_{sn} + t$ satisfying $d(0, t) < \delta$ we have

$$\frac{d}{d\theta}\Psi_{sni}(\theta) = \Psi_{0sni}(\theta) \in \mathbb{R}^{p_{sn}},$$

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

$$\begin{aligned}\Psi_{0sni(a)}(\theta) = \Psi_{0sni(a)}(\theta_{sn}) + \Psi_{1sni(a)}(\theta_{sn})t \\ + 2^{-1}t^T\Psi_{2sni(a)}(\theta_{sn} + ct)t, \text{ for } a = 1, \ldots p_{sn},\end{aligned}$$

for some $c \in (0, 1)$ possibly depending on a.

We assume that for each s and n, there is a sequence of σ-fields $\mathcal{F}_{sn1} \subset \mathcal{F}_{sn2} \ldots \mathcal{F}_{snk_n}$ such that $\{\sum_{i=1}^{j} \Psi_{0sni}(\theta_{sn}), \mathcal{F}_{snj}\}$ is a martingale.

The spectral decomposition of $\Gamma_{0sn} = \sum_{i=1}^{k_n} \mathbb{E}\Psi_{0sni}(\theta_{sn})\Psi_{0sni}^T(\theta_{sn})$ is given by

$$\Gamma_{0sn} = P_{0sn}\Lambda_{0sn}P_{0sn}^T,$$

where $P_{0sn} \in \mathbb{R}^{p_{sn}} \times \mathbb{R}^{p_{sn}}$ is an orthogonal matrix whose columns contain the eigenvectors, and Λ_{0sn} is a diagonal matrix contining the eigenvalues of Γ_{0sn}. We assume that Γ_{0sn} is positive definite, that is, all the diagonal entries of Λ_{0sn} are positive numbers. We assume that there is a constant $\delta_{0s} > 0$ such that $\lambda_{min}(\Gamma_{0sn}) > \delta_{0s}$ for all sufficiently large n. The matrices Λ_{0sn}^c for various real numbers c are defined in the obvious way, that is, these are diagonal matrices where the j-th diagonal entry is raised to the power c.

Let $\Gamma_{1sni}(\theta_{sn})$ be the $p_{sn} \times p_{sn}$ matrix whose a-th row is $\mathbb{E}\Psi_{1sni(a)}(\theta_{sn})$; we assume this expectation exists. Define

$$\Gamma_{1sn}(\theta_{sn}) = \sum_{i=1}^{k_n} \Gamma_{1sni}(\theta_{sn}).$$

We assume that $\Gamma_{1sn} \equiv \Gamma_{1sn}(\theta_{sn})$ is nonsingular for each s and n. Suppose the singular value decomposition of Γ_{1sn} is given by

$$\Gamma_{1sn} = P_{1sn}\Lambda_{1sn}Q_{1sn}^T,$$

where $P_{1sn}, Q_{1sn} \in \mathbb{R}^{p_{sn}} \times \mathbb{R}^{p_{sn}}$ are orthogonal matrices, and Λ_{1sn} is a diagonal matrix. We assume that the diagonal entries of Λ_{1sn} are all positive, which implies that *in the population, at the true value of the parameter* the energy functional $\sum \Psi_{sn}$ actually achieves a minimal value. We define the matrices Λ_{1sn}^c for various real numbers c as diagonal matrices where the j-th diagonal entry is raised to the power c. Correspondingly, we define $\Gamma_{1sn}^c = P_{1sn}\Lambda_{1sn}^c Q_{1sn}^T$. We assume that there is a constant $\delta_{1s} > 0$ such that $\lambda_{min}(\Gamma_{1sn}^T\Gamma_{1sn}) > \delta_{1s}$ for all sufficiently large n.

We define the matrix

$$A_{sn} = \Gamma_{0sn}^{-1/2}\Gamma_{1sn}.$$

We assume the following conditions:

(A1): The minimum eigenvalue of $A_{sn}^T A_{sn}$ tends to infinity. That is, there is a sequence $a_{sn} \uparrow \infty$ as $n \to \infty$ such that

$$\lambda_{min}(\Gamma_{1sn}\Gamma_{0sn}^{-1}\Gamma_{1sn}^T) \asymp a_{sn}^2. \tag{2.7}$$

(A2): For any model $\mathcal{M}_n$, there exists a sequence of positive reals $\{\gamma_{sn}\}$ that is bounded away from zero, such that

$$\lambda_{max}(\Gamma_{1sn}^{-1}\Gamma_{0sn}^{2}\Gamma_{1sn}^{-T}) = o(\gamma_{sn}^{-2}) \tag{2.8}$$

as $n \to \infty$ for any s.

(A3):

$$\mathbb{E}\Big|A_{sn}^{-1}(\sum_{i=1}^{k_n}\Psi_{1sni} - \Gamma_{1sn})A_{sn}^{-1}\Big|_F^2 = o(p_{sn}\gamma_{sn}^{-2}). \tag{2.9}$$

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

(A4): For the symmetric matrix $\Psi_{2sni(a)}(\theta)$ and for some $\delta_0 > 0$, there exists a symmetric matrix $M_{2sni(a)}$ such that

$$\sup_{|\theta-\theta_{sn}|<\delta_0} \Psi_{2sni(a)}(\theta) < M_{2sni(a)},$$

satisfying

$$\sum_{a=1}^{p_{sn}}\sum_{i=1}^{k_n}\mathbb{E}\lambda_{max}^2(M_{2sni(a)}) = o(a_{sn}^6 n^{-1} p_{sn}\gamma_{sn}^{-2}) \tag{2.10}$$

(A5): For any vector $c \in \mathbb{R}^{p_{sn}}$ with $|c| = 1$, we define the random variable $Z_{sni} = -c^T\Gamma_{0sn}^{-1/2}\Psi_{0sni}$ for $i = 1, \dots k_n$. We assume that

$$\sum_{i=1}^{k_n} Z_{sni}^2 \xrightarrow{p} 1, \text{ and } \mathbb{E}[\max_i |Z_{sni}|] \to 0. \tag{2.11}$$

(A6): Assume that

$$\lambda_{max}(\Gamma_{1sn}\Gamma_{0sn}^{-1}\Gamma_{1sn}^{T}) \asymp a_{sn}^2. \tag{2.12}$$

We use the notation $\hat{\Psi}_{ksni}$ for $\Psi_{ksni}(\hat{\theta}_{sn})$, for $k = 0, 1, 2$.

THEOREM 2.1. *Assume conditions (*A1 - *(*A5*) and that $p_{sn}^2 k_n^{-1} \to 0$ as $n \to \infty$. Then $\hat{\theta}_{sn}$ is a consistent estimator of θ_{sn}), and $A_{sn}(\hat{\theta}_{sn} - \theta_{sn})$ converges weakly to the standard Normal distribution in p_{sn}-dimension.*

*Under the additional condition (*A6*), we have that $a_{sn}(\hat{\theta}_{sn} - \theta_{sn})$ converges weakly to a Normal distribution in p_{sn}-dimension.*

The technical conditions (A1) - (A5) are very broad within the framework of smooth energy functions, and allows for different rates of convergence of different parameter estimators, depending on the matrix A_{sn}. The additional condition (2.12) in (A6) is a natural condition that, coupled with (A1), ensures identical rate of convergence a_{sn} for all the paramater estimators in a model. Classical regularity conditions on estimating functions ensure assumptions (A1) - (A6) hold with $a_{sn} \equiv k_n^{1/2}$.

3. Model adequacy and its relation to *e-values*.

3.1. *Model adequacy.* We now define an importance concept for use in the rest of this paper. Recall that in the model $\mathcal{M}_n$, estimators of θ_{sn} are obtained by minimizing (2.4). We assume the technical conditions (A1) - (A6) in this section, although the concepts and results presented in this section can be developed with only (A1) - (A5) .

DEFINITION 3.1. *The model $\mathcal{M}_n$ is called $(G, *)$-inadequate or just inadequate in short, if*

$$\lim_{n\to\infty} \min\{a_{sn}, a_{*n}\}(G_{mn}(\theta_{mn}) - G_{*n}(\theta_{*n})) \neq 0. \tag{3.1}$$

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

For completeness, and compatibility with considerable part of literature on model selection, we also consider a stronger version of inadequacy below.

DEFINITION 3.2. *The model $\mathcal{M}_n$ is called strictly $(G, *)$-inadequate or just strictly inadequate in short, if*

$$\liminf_{n\to\infty} |(G_{mn}(\theta_{mn}) - G_{*n}(\theta_{*n}))| \neq 0. \tag{3.2}$$

Note that this notion of adequacy of a model depends on the choice of the preferred model, as well as the transformation maps $\{G_{sn}\}$. Models having the same functional structure, but employing different methods of handling data or parameter estimation, may not be both adequate (or inadequate) depending on their relative efficiency and robustness properties. For example, suppose $Y_i \overset{i.i.d.}{=} N(\mu, 1)$ and the preferred model is based on the likelihood. Here, the models that use estimates of $\mu \in \mathbb{R}$ based on the sample median or a trimmed least squares criterion with $o(n)$ observations deleted are adequate models, while ones that use a heavy trimming or only a finite number of observations are inadequate.

Note that owing to the consistency results presented in Section 2.5, the preferred model is always adequate, so the set of adequate models is non-empty by construction. Since the notion of parsimony is important in this context, we define the *minimal adequate* model as the adequate model that has the smallest number of parameters estimated from the data. Note that our framework ensures that there is always a minimal adequate model, though in general, the uniqueness of the minimal adequate model is not guaranteed.

In classical covariate selection problems, as in linear regression where in $\mathcal{S}_n$ a subset of covariates X_s is used in fitting the expression $Y = X_s\beta_s + e_s$, the above notion of model inadequacy captures the standard notions of model "incorrectness". For example, easily constructed examples show that for obvious choices of $\{G_{mn}\}$ (3.1) reduces to $\lim_{n\to\infty} n^{1/2}(\mathbb{E}Y - X_s\beta_s) \neq 0$. Note that this is weaker than the traditional condition $(\mathbb{E}Y - X_s\beta_s) \neq 0$ about model "incorrectness" implicit in many studies, which is captured in our condition for a model being *strictly inadequate*. The concept of the minimal adequate model merges with that of a "true model" used in many studies. In this paper we only consider models that are either adequate or strictly inadequate, other cases will be considered in future work.

Our definition of model adequacy allows asymptotically similar models to be clubbed together, and has flexibility to accomodate local asymptotic characteristics. For instance, suppose $Y_{ni} = X_{1i}\beta_{01} + X_{2i}\delta_n + e_i$ for some $\beta_{01} \in \mathbb{R}, \delta_n = o(n^{-0.6})$, $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $i = 1, \ldots, k_n$. Suppose the preferred model uses $\Theta_{*n} = \{(\beta_1, \beta_2, \sigma^2) : \beta_1, \beta_2 \in \mathbb{R}, \sigma^2 > 0\}$. Here, the (constant) sequence of models $\mathcal{M}_n$ with $\Theta_{mn} = \{(\beta_1, 0, \sigma^2)^T : \beta_1 \in \mathbb{R}, \sigma^2 > 0\}$ will be an adequate model. Such models can arise from prior choices in Bayesian variable selection techniques, for example see Narisetty and He (2014); Ročkovà and George (2016).

3.2. *Model adequacy and e-values.* We now present our first result on the model elicitation process, which separates the inadequate models from the adequate models. We assume for the rest of this paper that each function G_{mn} is smooth, but alternative sets of conditions to ensure theoretical results can be easily conceived. The functions G_{mn} are designed and selected by the statistician, so assumptions to ensure tractable and nice theoretical properties are not unreasonable.

Note that the j-th element of the function G_{mn}, denoted by $G_{mnj}(\cdot) \equiv G_j(\cdot)$, is a map from a subset of $\mathbb{R}^{p_{mn}}$ to $\mathbb{R}$, for $j = 1, \ldots, d_n$.

(B1): We assume that such functions $G_j(\cdot)$ are smooth functions in a neighborhood of $\theta_{mn} \equiv \theta$. Specifically, there exists a $\delta > 0$ such that for

$x = \theta + t$ with $|t| < \delta$, we have the following expansion

$$G_j(x) = G_j(\theta) + G_{1j}^T(\theta)t + 2^{-1}t^T R_j(\theta + ct)t,$$

for some $c \in (0, 1)$.

(B2): Let $G_1 \in \mathbb{R}^{p_{mn}} \times \mathbb{R}^{d_{mn}}$ be the matrix whose j-th column in $G_{1j}(\theta)$. We assume that there is a sequence of positive definite matrices $\{M_{1n}\}$ such that

$$G_1 G_1^T < M_{1n}, \qquad \sup_n \lambda_{max}(M_{1n}) < \infty.$$

(B3): We assume that there is a sequence of positive definite matrices $\{M_{2nj}\}$ such that

$$\sup_{t:|t|<\delta} R_j(\theta + ct) < M_{2nj}, \qquad \sup_n \lambda_{max}(M_{2nj}) < \infty.$$

We now state the technical conditions we assume on the sequence of evaluation maps.

(C1): Each E_n is invariant to location and scale transformations. That is, for any $a \in \mathbb{R}, b \in \mathbb{R}^{d_n}$ and random variable G having distribution $\mathbb{G} \in \tilde{\mathcal{G}}_n$,

$$E_n(x, \mathbb{G}) = E_n(ax + b, [aG + b]) \tag{3.3}$$

(C2): We assume that each E_n is Lipschitz continuous in the first argument. That is, there exists $\delta > 0$ and $\alpha \in (0, 1)$, possibly depending on the (non-degenerate) measure $\mathbb{G} \in \tilde{\mathcal{G}}_n$ such that whenever $|x - y| < \delta$, we have

$$|E_n(x, \mathbb{G}) - E_n(y, \mathbb{G})| < |x - y|^\alpha. \tag{3.4}$$

(C3): Suppose $\{\mathbb{G}_n\}$ is a tight sequence of probability measures on ℓ_2, with weak limit $\mathbb{G}_\infty$. Further assume that $Y_n \in \mathbb{R}^{d_n}$ is a random variable that follows the marginal distribution of the first d_n co-ordinates under $\mathbb{G}_n$. Also suppose $E_\infty : \ell_2 \times \tilde{\ell}_2 \to [0, \infty)$ be a map such that when restricted to the first d_n co-ordinates, E_∞ matches E_n. Then we assume that there exists an element Y_∞, possibly random, in ℓ_2 such that

$$\lim_{n\to\infty} \mathbb{E}E_n(Y_n, [Y_n]) = \mathbb{E}E_\infty(Y_\infty, \mathbb{G}_\infty). \tag{3.5}$$

Also, we assume that there exists a $\mu(\mathbb{G}_\infty) \in \ell_2$ such that

$$E_\infty(\mu(\mathbb{G}_\infty), \mathbb{G}_\infty) = \sup_{x\in\ell_2} E_\infty(x, \mathbb{G}_\infty) < \infty. \tag{3.6}$$

(C4): Suppose $\{Y_n\}$, $\{Z_n\}$ are two sequences of random variables on $\mathbb{R}^{d_n}$ such that there exists an increasing sequence of positive reals $\{C_n\}$ and a sequence $\{D_n \in \mathbb{R}^{d_n}\}$, for which

$$C_n(Y_n - D_n) \text{ has a limiting distribution in } \tilde{\ell}_2, \text{ and}$$
$$C_n|Z_n - D_n| \xrightarrow{p} \infty \text{ as } n \to \infty,$$

we have that

$$\lim_{n\to\infty} \mathbb{E}E_n(Z_n, [Y_n]) = 0. \tag{3.7}$$

On the other hand, if $Z_n - D_n \xrightarrow{p} 0$ and $C_n|Z_n - D_n| = O_P(1)$, then

$$\lim_{n\to\infty} \mathbb{E}|E_n(Z_n, [Y_n]) - E_n(Y_n, [Y_n])| = 0. \tag{3.8}$$

We are now at a stage to present our population-level result that forms the foundation of all the following analysis.

THEOREM 3.1. *Assume conditions (*A1*) - (*A4*), (*A6*) and that* $\mathbb{E}|\hat{\theta}_{sn}|^4 < \infty$*. For a sequence of transformation maps* $\{G_{mn}\}$ *satisfying conditions (*B1*) - (*B3*) and a sequence of evaluation functions* $\{E_n\}$ *satisfying (*C1*) - (*C4*) the following hold as* $n \to \infty$*:*

1. *the preferred model* $\mathcal{M}_{*n}$ *achieves the limit defined in* (3.5) *and in particular,* $e_n(\mathcal{M}_{*n}) < \infty$*, where* $Y_n = a_{*n}(\hat{G}_{*n} - G_{*n})$*,*
2. *for an adequate model* $\mathcal{M}_n$*,* $|e_n(\mathcal{M}_n) - e_n(\mathcal{M}_{*n})| \to 0$*,*
3. *for a strictly inadequate model* $\mathcal{M}_n$*,* $e_n(\mathcal{M}_n) \to 0$*.*

This result 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. Note that we do not need (A5) for this result, which is used to obtain a limiting distribution.

4. Resampling for simultaneous model selection and inference. Recall that in (2.4) and (2.6) we obtain the estimator $\hat{\theta}_{sn}$ by minimizing

$$\hat{\Psi}_{sn}(\theta) = \sum_{i=1}^{k_n} \Psi_{sni}(\theta, Y_{ni}).$$

The parameter θ_{sn} is the unique minimizer of the expectation of the above.

We propose using the resampling strategy of Chatterjee and Bose (2005) to obtain approximations of the sampling distributions of $\hat{\theta}_{sn}$ for any s and n. Thus, we obtain the resampling estimator $\hat{\theta}_{rsn}$ as the minimizer of

$$\hat{\Psi}_{rsn}(\theta) = \sum_{i=1}^{k_n} \mathbb{W}_{rsni} \Psi_{sni}(\theta, Y_{ni}). \tag{4.1}$$

Here, $\{\mathbb{W}_{rsni}\}$ are *resampling weights*, which are random variables that are independent of the data. Several resampling schemes can be described in the above format and are discussed in Chatterjee and Bose (2005). However, while such schemes can in general be used to obtain estimator of the distribution of $\hat{\theta}_{sn}$, they will not necessarily perform well for the model elicitation task we set out in this paper.

We now state the conditions on the resampling weights $\mathbb{W}_{rsni}$, which for any n may be collected together in the vector $\mathcal{W}_{rsn} = (\mathbb{W}_{rsn1}, \ldots, \mathbb{W}_{rsnk_n})^T \in \mathbb{R}^{k_n}$. We assume that this is an exchangeable array of non-negative random variables, independent of the data. The index r denotes that these are related to the resampling procedure. The actual implementation of the resampling procedure is carried out by generating independent copies $\mathcal{W}_{1sn}, \ldots, \mathcal{W}_{Rsn}$ for some sufficiently large integer R, and using them in a Monte Carlo procedure, where for any $r = 1, \ldots, R$, we minimize

$$\sum_{i=1}^{k_n} \mathbb{W}_{rsni} \Psi_{sni}(\theta, Y_{ni}) \tag{4.2}$$

to obtain the resampling version of the estimator $\hat{\theta}_{rsn} \in \mathbb{R}^{p_{sn}}$.

We assume that for each $i = 1, \ldots, k_n$, $\mathbb{E}\mathbb{W}_{rsni} = \mu_{sn}$ and $\mathbb{V}\mathbb{W}_{rsni} = \tau_{sn}^2$, consequently we write the centered and scaled resampling weights as

$$W_{rsni} = \tau_{sn}^{-1}(\mathbb{W}_{rsni} - \mu_{sn}),$$

thus $\mathbb{W}_{rsni} = \mu_{sn} + \tau_{sn} W_{rsni}$.

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

$$\mathbb{E}\mathbb{W}_{rsni} = \mu_{sn}, \tag{4.3}$$

$$\mathbb{V}\mathbb{W}_{rsni} = \tau_{sn}^2 \uparrow \infty, \tag{4.4}$$

$$\tau_{sn}^2 = o(a_{sn}^2), \tag{4.5}$$

$$\mathbb{E}W_{rsn1} W_{rsn2} = O(k_n^{-1}), \tag{4.6}$$

$$\mathbb{E}W_{rsn1}^2 W_{rsn2}^2 \to 1, \tag{4.7}$$

$$\mathbb{E}W_{rsn1}^4 < \infty. \tag{4.8}$$

Since $\mathbb{W}_{rsni} \geq 0$ almost surely and is non-degenerate, we have $\mu_{sn} > 0$. We assume that $\mu_{sn} + \tau_{sn}^2 = O(\tau_{sn}^2)$. Our analysis below suggests that the properties of the resampling procedure depend only on the *coefficient of variation* ratio τ_{sn}/μ_{sn}, and without loss of generality we can set $\mu_{sn} = 1$ for all s and n. We assume that $\gamma_{sn} \leq \tau_{sn}$ from (2.8).

EXAMPLE 4.1 (The m-out-of-n (moon) bootstrap:). In our framework, the *moon*-bootstrap is identified with $\mathcal{W}_{rsn}$ having a Multinomial distribution with parameters m and probabilities $k_n^{-1}(1, \ldots, 1) \in \mathbb{R}^{k_n}$, by a factor of k_n/m. Thus we have $\mathbb{E}\mathbb{W}_{rsni} = \mu_{sn} = (m^{-1}k_n)(m/k_n) = 1$, and $\mathbb{V}\mathbb{W}_{rsni} = \tau_{sn}^2 = (m^{-1}k_n)^2(mk_n^{-1}(1 - k_n^{-1}) = O(m^{-1}k_n)$. In typical applications of the *moon*-bootstrap, as in its application in this paper, we require that $m \to \infty$ and $m/k_n \to 0$ as $n \to \infty$. Thus we have $\tau_{sn}^2 \to \infty$ as $n \to \infty$, thus the *scale* factor of the resampling weights $\mathbb{W}_{rsni}$ tend to infinity with n. We use the term *scale-enhanced* resampling for schemes like the *moon*-bootstrap where the variance of (properly centered) resampling weights tend to infinity with n.

There are numerous problems where the moon-bootstrap provides consistent approximation to the distribution of statistics of interest, and all such cases are included in our framework. Since such cases are too numerous to list and review of resampling consistency is not central to this paper, we only demonstrate the properties of our resampling procedure in one interesting framework below. □

EXAMPLE 4.2 (The scale-enhanced Bayesian bootstrap:). A version of Bayesian bootstrap may be constructed by choosing $\mathbb{W}_{rsni}$ to be independent and identically distributed Gamma random variables, with mean $\mu_{sn} = 1$ and variance $\tau_{sn}^2 \to \infty$ 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. □

We now have the result on consistency of the above resampling scheme.

THEOREM 4.1. *Assume conditions (*A1 - (A5*) and that* $p_{sn}^2 k_n^{-1} \to 0$ *as* $n \to \infty$. *Additionally, assume that the resampling weights* $\mathbb{W}_{rsni}$ *are exchangeable random variables satisfying the conditions* (4.3)-(4.8).

Define $\hat{B}_{sn} := \mu_{sn}\tau_{sn}^{-1}\hat{\Gamma}_{0sn}^{1/2}\hat{\Gamma}_{1sn}^{-1}$, *where* $\hat{\Gamma}_{0sn}$ *and* $\hat{\Gamma}_{1sn}$ *are sample equivalents of* Γ_{0sn} *and* Γ_{1sn}, *respectively. Conditional on the data,* $\hat{B}_{sn}(\hat{\theta}_{rsn} - \hat{\theta}_{sn})$ *also converges weakly to the standard Normal distribution in probability.*

*Under the additional condition (*A6*), defining* $b_{sn} = \mu_{sn}\tau_{sn}^{-1}a_{sn}$, *we have that the distributions of* $a_{sn}(\hat{\theta}_{sn} - \theta_{sn})$ *and* $b_{sn}(\hat{\theta}_{rsn} - \hat{\theta}_{sn})$ *converge to the*

same weak limit in probability.

We now discuss how to estimate the *e-value* of any model using resampling. We implement two independent resampling procedures, one from model $\mathcal{M}_n$ and another from $\mathcal{M}_*$ respectively indexed by r and r_1, both satisfying the conditions stated above. We use the first set of samples to generate coefficient vectors $\hat{\theta}_{rmn}$ corresponding to the model $\mathcal{M}_n$:

$$\hat{\theta}_{rmnj} = \begin{cases} \hat{\theta}_{rsnj} & \text{for } j \in \mathcal{S}_n; \\ C_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases} \tag{4.9}$$

We use $\hat{\theta}_{r_1*n}$ to obtain the distribution $[\hat{G}_{r_1*n}]$ conditional on the data. The resampling estimate of a model *e-value* is defined as

$$e_{rn}(\mathcal{M}_n) = \mathbb{E}_r E_n(\hat{G}_{rmn}, [\hat{G}_{r_1*n}]),$$

where $\mathbb{E}_r$ is resampling expectation conditional on the data. This is the usual resampling version of the *e-value*. In this section, we establish how $e_{rn}(\mathcal{M}_n)$ may be used to sort adequate and strictly inadequate models, using results established thus far in this paper.

THEOREM 4.2. *Assume conditions (*A1 - *(*A6*), (*B1*) - (*B3*), (*C1*) - (*C4*), $p_{sn}^2 k_n^{-1} \to 0$ as $n \to \infty$, and that $\mathbb{E}|\hat{\theta}_{sn}|^8 < \infty$. Additionally, assume that the resampling weights $\mathbb{W}_{rsni}$ are exchangeable random variables satisfying the conditions* (4.3)-(4.8).

Assume that $b_{*n} = o(\min\{a_{*n}, a_{sn}\})$, and $b_{sn} \asymp b_{*n}$. Then $e_{rn}(\mathcal{M}_n)$ converges to zero if $\mathcal{M}_n$ is a sequence of strictly inadequate models, and to the limit defined in (3.5) *if $\mathcal{M}_n$ is a sequence of adequate models.*

The above theorem is one of several possible alternatives: multiple variants of resampling strategies and technical conditions can be devised to achieve similar results. For example, we may obtain both the independent resamples from $\mathcal{M}_*$; an option we explore in the context of Section 5 later.

5. Fast variable selection using data depth. The traditional application domain for statistical model selection has been in *covariate selection*: for regression, mixed effect models, time series and other problems. Also, in many instances, the number of parameters does not grow significantly faster than the sample size. In such situations, it is feasible to consider the least parsimonious model as the preferred model. This is routinely done in practice, for example in classical model selection techniques (Konishi and

Kitagawa, 1996; Claeskens and Hjort, 2008), and the fence method (Jiang et al., 2008).

In this section, for simplicity, we assume that the least parsimonious model has a fixed $p \equiv p_{nmax}$ parameters for all n. Our methodology and results below go through even for cases where $p_{nmax}^2 k_n^{-1} \to 0$ as $n \to \infty$. We will occasionally drop the subscripts $_n$ or $_{nmax}$, as well as $_*$ in all subscripts corresponding to the preferred model. Although we still keep the subscript in e_n because it is calculated based on the estimators $\hat{\theta}_m$ that depends on a size n-sample. We shall consider this least parsimonious model as the preferred model, and often refer to it as the 'full model' from now on. For this section, all candidate models are sub-models of this model, thus all vectors C of constants (2.1) are set to zero. An example is that of linear regression with total p covariates, and different candidate models are obtained by setting subsets of regression coefficients to zero. In such models, obtaining the most parsimonious model that fits the data, for example by using the Bayesian Information Criterion (BIC) (Schwarz, 1978), a full-scale analysis would require analyzing all 2^p possible candidate models. This is an NP-Hard problem (Natarajan, 1995), and becomes computationally intractable even for moderate data dimensions ($n \simeq 100, p \simeq 50$). Several *ad-hoc* techniques that are in use do not guarantee, in the absence of stringent conditions, that the probability of selecting the most parsimonious model that fits the data tends to one as sample size increases. In this section we propose a fast and scalable algorithm to tackle this problem, i.e. detect variables with non-zero coefficients, through implementing our general e-values framework.

5.1. *A plugin parameter estimate.* This subsection applies to all problems where the preferred model is the least parsimonious model, as in the context of this section. In such cases, we first obtain the consistent estimator $\hat{\theta}_{*n} = (\hat{\theta}_{*n1}, \ldots, \hat{\theta}_{*np_{nmax}})^T$. Then, for a general model $\mathcal{M}_n$ specified by the set $\mathcal{S}_n = \{j_{n1}, \ldots, j_{np_{sn}}\} \subseteq \{1, 2, \ldots, p_{nmax}\}$ and the vector of potentially non-zero constants C_n, we define the parameter estimates to be

$$\hat{\theta}_{mnj} = \begin{cases} \hat{\theta}_{*nj} & \text{for } j \in \mathcal{S}_n; \\ C_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases} \tag{5.1}$$

Thus, we do not fit the model $\mathcal{M}$ separately, but simply plug-in estimators from the preferred model at the indices in $\mathcal{S}$. Resampling version of parameter estimators are obtained as

$$\hat{\theta}_{rmnj} = \begin{cases} \hat{\theta}_{r*nj} & \text{for } j \in \mathcal{S}_n; \\ C_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases} \tag{5.2}$$

The logic behind this is simple: for a candidate model $\mathcal{M}_n$, a joint distribution of the estimator of its parameters, i.e. $[\hat{\theta}_{sn}]$, can actually be obtained from $[\hat{\theta}_{*n}]$ by marginalizing at indices $\mathcal{S}_n$. This makes it easy to guarantee that the distribution of parameter estimates for any selected model is consistently approximated through the corresponding sampling distributions by our method. We conjecture that this logic may be applied in the context of several other model selection methods also, but do not pursue that line of study in this paper.

The above plug-in step has two additional major advantages. We do not separately analyze each candidate model, thus saving massively in computations. Additionally, this approach leads to an easier comparison of any candidate model to the preferred model.

5.2. *Simplifications.* At this stage we make a few simplifying assumptions that will allow us to obtain specialized results relevant in the context. First of all we assume G_{mn} to be the identity function, i.e. $G_m(\theta) = \theta$ for any $\mathcal{M}$ and $\theta \in \Theta$, thus $d_n \equiv p_n \equiv p$. This simplifies the geometry of the model space as a lattice: we now consider a model $\mathcal{M}_1$ to be contained 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 $\mathcal{M}_a$ is adequate, then any model $\mathcal{M}$ such that $\mathcal{M}_a \prec \mathcal{M}$ is also adequate. In the context of covariate selection, this obtains a linear ordering, with the most parsimonious adequate model being the minimal adequate model, and all models strictly contained in it being inadequate.

For the evaluation functions, we take a single map $E : \mathbb{R}^p \times \tilde{\mathbb{R}}^p \to [0, \infty)$ for all n that satisfies the following properties:

(D1): The map E 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 $\mathbb{G} \in \tilde{\mathbb{R}}^p$, the set of probability measures on $\mathbb{R}^p$,

$$E(x, \mathbb{G}) = E(Ax + b, [AY + b]) \tag{5.3}$$

(D2): The map E_n is Lipschitz continuous in the first argument. That is, there exists $\delta > 0$ and $\alpha \in (0, 1)$, possibly depending on the measure $\mathbb{G} \in \tilde{\mathcal{G}}_n$ such that whenever $|x - y| < \delta$, we have

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

(D3): Assume that $Y_n \in \mathbb{R}^p$ is a sequence of random variables converging in distribution to some $\mathbb{Y} \in \tilde{\mathbb{R}}^p$. Then $E(y, [Y_n])$ converges uniformly to $E(y, \mathbb{Y})$.

(D4): For any $\mathbb{G} \in \tilde{\mathbb{R}}^p$, $\lim_{\|x\|\to\infty} E(x, \mathbb{G}) = 0$.

(D5): For any $\mathbb{G} \in \mathbb{R}^p$ with a point of symmetry $\mu(\mathbb{G}) \in \mathbb{R}^p$, we have for any $t \in (0,1)$ and any $x \in \mathbb{R}^p$

$$E(x,\mathbb{G}) < E(\mu(\mathbb{G}) + t(x - \mu(\mathbb{G})), \mathbb{G}) < E(\mu(\mathbb{G}), \mathbb{G}) = \sup_{x \in \mathbb{R}^p} E(x,\mathbb{G}) < \infty \tag{5.4}$$

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

It can be seen that (D1)- (D5) are either equivalent or marginally stronger versions of (C1)- (C5). In the framework of this section, these are easily verifiable and practical conditions, that are satisfied by most data depth functions (Zuo and Serfling, 2000; Mosler, 2013) that may be used as evaluation functions. Using the slightly stronger conditions (D1)-(D5) results in additional refinements in the behavior of *e-values* as stated below.

We shall assume elliptical asymptotic distributions for full model estimators $\hat{\theta} = \hat{\theta}_*$. Following Fang, Kotz and Ng (1990), elliptical distributions can be formally defined using their characteristic function:

DEFINITION 5.1. *A p-dimensional random vector X is said to elliptically distributed if and only if there exist a vector $\mu \in \mathbb{R}^p$, a positive semi-definite matrix $\Omega \equiv \Sigma^{-1} \in \mathbb{R}^{p \times p}$ and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $t \mapsto \phi_{X-\mu}(t)$ of $X - \mu$ corresponds to $t \mapsto \phi(t^T \Sigma t), t \in \mathbb{R}^p$.*

The density function of an elliptically distributed random variable takes the form:

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

where 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:

(E1): The limiting distribution $\mathbb{T}$ of the full model estimate, i.e. $a_n(\hat{\theta} - \theta_0), (a_{sn} \equiv a_n)$ is distributed as $\mathcal{E}(0_p, V, g)$, for some positive-definite matrix V and density generator function g;

(E2): For almost every data sequence $\mathbf{Y}$, There exists a sequence of positive definite matrices V_n such that $\text{plim}_{n\to\infty} V_n = V$.

In practice we mostly deal with Gaussian limiting distributions, where (E1)-(E2) are naturally satisfied.

5.3. *Derivation of the algorithm.* We are now at a stage to present a result that forms the foundation of our fast algorithm.

THEOREM 5.1. *Assume conditions (*A1*) -(*A6*), (*D1*) -(*D5*), and (*E1*) -(*E2*). Then, for a finite sequence of adequate models* $\mathcal{M}_1 \prec \ldots \prec \mathcal{M}_k$ *and any finite collection of inadequate models* $\mathcal{M}_{k+1}, \ldots, \mathcal{M}_K$*, we have*

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

for large enough n*.*

Recall that nowhere above have we used the actual data generating process as a candidate model. When that condition is available, we have the following result following easily from Theorem 5.1.

COROLLARY 5.1. *Consider the collection of candidate models* $\mathbb{M}_0 = \{\mathcal{M} : c_j = 0 \quad \forall \quad j \notin \mathcal{S}\}$*. Suppose* $\mathcal{M}_0 \in \mathbb{M}_0$ *is an adequate model such that its associated index set* $\mathcal{S}_0 = \{j : \theta_{0j} \neq 0\}$*, i.e. it estimates all non-zero indices in the preferred coefficient vector* θ_0*. Then there exists a positive integer* N *so that for all* $n_1 > N$*,*

$$\mathcal{M}_0 = \arg \max_{\mathcal{M} \in \mathbb{M}_0} [e_{n_1}(\mathcal{M})] \tag{5.5}$$

At this point the total number of candidate models being considered is 2^p. However, in the present framework of covariate selection with the full model being the preferred one, to determine the minimal adequate model $\mathcal{S}_0$ one does not need to sift through all possible subsets or employ *ad-hoc* search strategies like forward selection/ backward deletion. We show that checking e-values at only p marginal models is sufficient for this purpose. In order to do this, we further restrict our attention to those candidate models where only a single parameter set to zero. That is, for such models $p_s = p - 1$. This collection of marginal sub-models can be studied in parallel: e.g. computations for these can be done on separate processors or computers.

The following result offers an alternate representation of the minimal adequate model using this much smaller set of models, after which the fast selection algorithm will be immediate.

COROLLARY 5.2. *Consider the models* $\mathcal{S}_{-j} = \{1, \ldots, p\} \setminus \{j\}$ *for* $j = 1, \ldots, p$*. Then* $\mathcal{S}_{-j}$ *is an inadequate model, that is, covariate* j *is a necessary component of a minimal adequate model, for sufficiently large* n *if and only if*

$$e_n(\mathcal{S}_{-j}) < e_n(\mathcal{S}_*) \tag{5.6}$$

In short, this happens because dropping an essential predictor from the full model makes the model inadequate, which has very small e-value for large enough sample size, whereas dropping a non-essential predictor increases the e-value: thus simply collecting those predictors that cause decrease in the e-value on dropping them from the model suffices for variable selection.

Thus, our fast algorithm for the evaluation of models consists of only 3 steps: (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. A safer version of this recipe can be to keep on dropping covariates until no sub-model achieves a lower e-value. In numeric studies we conducted we did not find substantial difference between selecting covariates directly based on whether $e_n(\mathcal{S}_{-j}) < e_n(\mathcal{S}_*)$, and this backward deletion method. Also in an empirical data-analytic setup, the performance of our algorithm is dependent on several factors, like sample size, signal-to-noise ratio, the estimation model and the resampling technique used: although we later show that our method in general performs better than the state-of-the-art across multiple modelling situations that take the above into account.

5.4. *Bootstrap implementation.* A sample version of the above variable selection recipe that incorporates bootstrap to estimate the sampling distributions $[\hat{\theta}]$, $[\hat{\theta}_s]$ is the following:

1. Generate two independent set of bootstrap weights, of size R and R_1, and obtain the corresponding approximations to the full model sampling distribution, say $[\hat{\theta}_r]$ and $[\hat{\theta}_{r_1}]$;
2. For $j = 1, 2, \ldots p$, estimate the e-value of $\mathcal{S}_{-j}$ as

$$\hat{e}_n(\mathcal{S}_{-j}) = \mathbb{E}_r E(\hat{\theta}_{r,-j}, [\hat{\theta}_{r_1}]) \tag{5.7}$$

   with $\hat{\theta}_{r,-j}$ obtained from $\hat{\theta}_r$ by replacing the j-th coordinate with 0;
3. Estimate the set of non-zero covariates as $\hat{\mathcal{S}}_0 = \{j : \hat{e}_n(\mathcal{S}_{-j}) < \hat{e}_n(\mathcal{S}_*)\}$

To make the sample e-values appropriately mimic the population level quantities, the bootstrap method used must adhere to the guidelines in Section 4.

6. Simulation studies. We now present the results of two simulation studies to compare the performance of our proposed fast variable selection method using model e-values, with the model selection procedures obtained from backward deletion and all subset regression versions that aim to minimize the Akaike Information Criterion (AIC: Akaike (1970)) or the BIC

for linear model, and sparse regularization-based methods for linear mixed models. In both examples below, we assume that the expectation of the response Y is a linear function of a few covariates, and the model selection problem is the classical one of identifying the set of covariates which have a non-zero effect on $\mathbb{E}Y$. Finally we use halfspace depth (Tukey, 1975) as the evaluation function.

6.1. *Selecting covariates in linear regression.* We use the first $p = 10$ columns of a simulated dataset from Prof. Charles Geyer's website (`http://www.stat.umn.edu/geyer/5102/data/ex6-8.txt`) and $n = 100$ randomly chosen rows, and arrange then in a $n \times p$ covariate matrix X. Each non-zero regression slope parameter takes the value 1, and we add independent standard Normal noise to generate the response vector, thus obtaining the framework $Y = X\beta + \epsilon$.

We generate data under different choices of the size of the minimal adequate model:b by first selecting $k \in \{2, 4, 6, 8\}$, then setting the first k coefficients of the regression slope β to be 1, and the rest $p - k$ slope parameters to be zero. The values of $\tau = \tau_n/\sqrt{n}$, the standard deviation of the resampling weights scaled by $\sqrt{n}$, is selected on a grid between 1 and 10 in 0.1 length intervals. We use a resampling Monte Carlo size $R = R_1 = 1000$ for use in (5.7), Finally the entire exercise is repeated 1000 times independently. We report here the results on the proportion of times out of this 1000 replications of the study when the minimal adequate model is selected. This is the numeric approximation of the "probability of selecting the true model".

We use the backward deletion and all-subset regression search strategies while using AIC and BIC as the model selection criterion. We use the leaps-and-bound algorithm, implemented in the `R` package `leaps`, for all-subset search. We display the results of this study in Figure 6.1 for the *moon*-bootstrap and in Figure 6.2 for the gamma bootstrap. As can be seen from these figures, the proposed *e-value*-based method performs very well when τ_n^2 is neither too small or too large, with the scale-enhances Gamma-weights resampling being better. We experimented with other choices of n, p, R_1, R_2, and it seems considering $\tau \in (4, 8)$ in this problem ensures exact minimal adequate model selection with high chance, and typically better performance than BIC in this regard. As long as R and R_1 are of the order of a few hundreds or higher, the variation from the resampling Monte Carlo step seems ignorable.

6.2. *Model selection in the presence of random effects.* Here we use the repeated measures simulation setup from Peng and Lu (2012). This is a

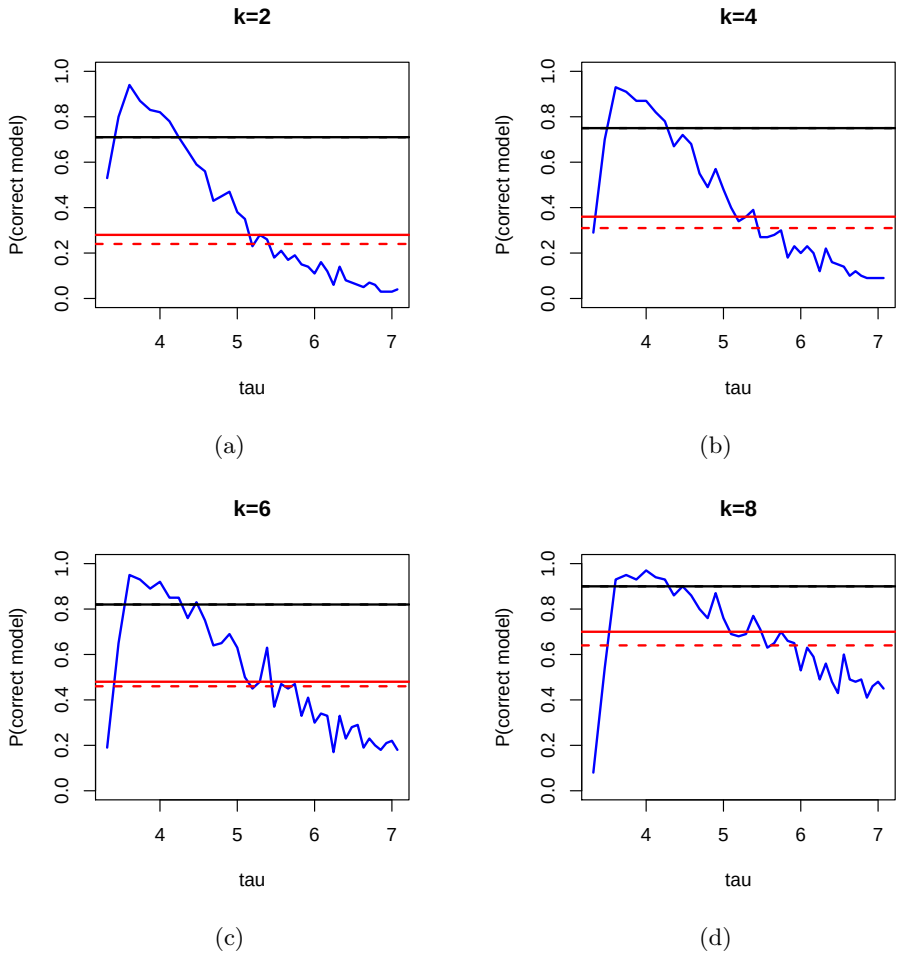
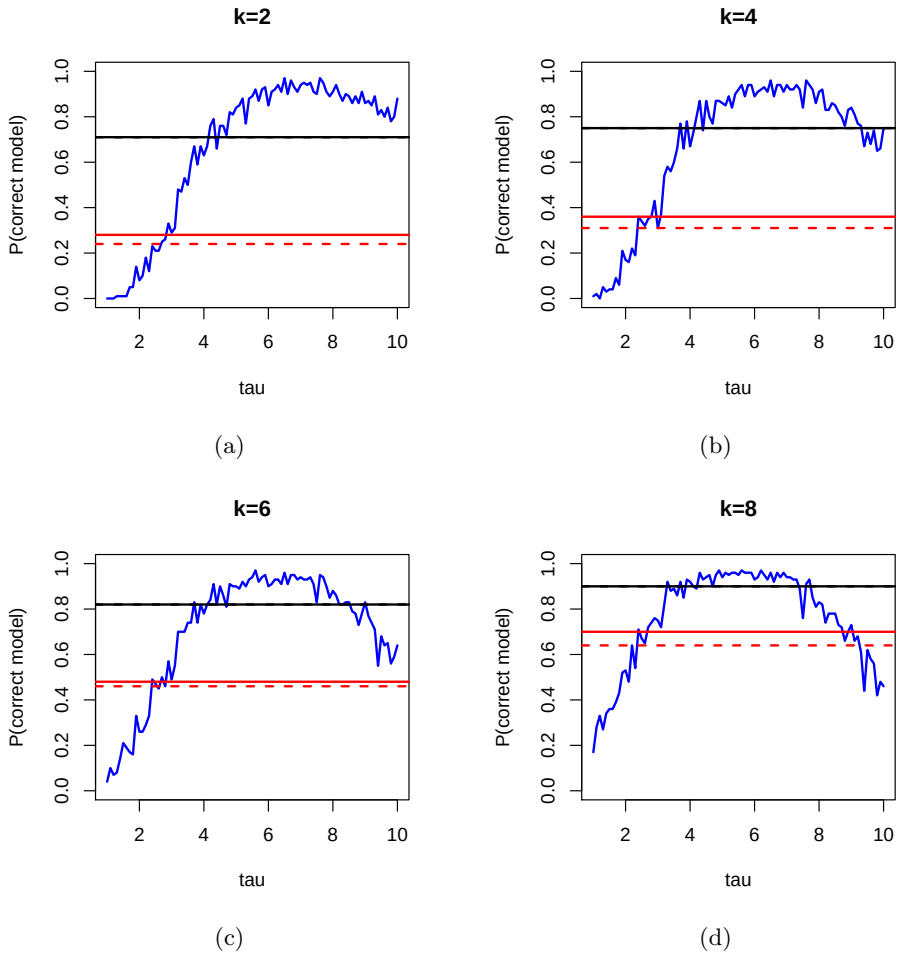


FIG 6.1. *Empirical probabilities of selecting the correct model through moon bootstrap for several levels of sparsity: The e-values method- blue solid, AIC backward deletion- red dotted, AIC all subset- red solid, BIC backward deletion- black dotted, BIC all subset- black solid*



(a) (b)

(c) (d)

FIG 6.2. *Empirical probabilities of selecting the correct model through gamma bootstrap for several levels of sparsity: The e-values method- blue solid, AIC backward deletion- red dotted, AIC all subset- red solid, BIC backward deletion- black dotted, BIC all subset- black solid*

random intercept-only model:

$$Y = X\beta + ZU + \epsilon; \quad Y \in \mathbb{R}^{n_i}$$

with $U \sim N(0, \Delta), \Delta \in \mathbb{R}^{n_i \times n_i}$ being positive definite, and $\epsilon \sim N(0, \sigma^2 \mathbb{I}_{n_i}), \sigma > 0$. The data consists of several (say m) independent groups of observations with multiple (say n_i) observations in each groups, with Z being the within-group random effects design matrix. We consider 9 fixed effects and 4 random effects, with true $\beta = (0, 1, 1, 0, 0, 0, 0, 0, 0)$ and random effect covariance matrix:

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

The error variance σ^2 is set at 1. The goal is to select the covariates of the fixed effect, thus essentially identify the covariates corresponding to the entries where β is non-zero. We use two scenarios for our study: one where the number of subjects considered is 30, and the number of observations per subject is 5, and another where the number of subjects considered is 60, and the number of observations per subject is 10.

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

$$\hat{\beta}_r = \hat{\beta} + \frac{\tau_n}{\sqrt{n}} (X^T \hat{V}^{-1} X)^{-1} W_r X^T \hat{V}^{-1} (y - X\hat{\beta}) + R_{rn} \tag{6.1}$$

with $\mathbb{E}_r |R_{rn}|^2 = o_P(1), W_r = \text{diag}(\mathbb{W}_{r1}\mathbb{I}_4, \ldots, \mathbb{W}_{r9}\mathbb{I}_4)$ and $\hat{V} = \hat{\sigma}^2 \mathbb{I}_p + Z\hat{\Delta}Z^T$. This is immediate from theorem 3.2 in Chatterjee and Bose (2005).

We consider $\tau = \tau_n/\sqrt{n} \in \{1, \ldots, 15\}$ here, and independent Gamma random variables as the resampling weights $\mathbb{W}_{rj}$, $j = 1, \ldots, 9$. We consider multiple characteristics of the model that obtains the highest e-value, including the number of parameters it involves, the proportion of times the minimal adequate model is obtained, the proportion of times a zero-valued (non-zero-valued) element of beta was identified as non-zero (zero), that is, the proportion of false positives (negatives), and so on.

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. Our results are presented in Table 6.1. It can be seen the *e-value*-based method handsomely outperforms the method

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		$n_i = 5, m = 30$			$n_i = 10, m = 60$		
e-value based	$\tau = 1$	59.9	0.0	5.61	44.3	0.0	4.43
	$\tau = 2$	33.0	0.0	3.45	15.5	0.0	2.54
	$\tau = 3$	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
	$\tau = 9$	1.0	0.0	2.03	0.3	0.0	2.01
	$\tau = 10$	0.7	0.0	2.02	0.3	0.0	2.01
	$\tau = 12$	0.7	0.0	2.02	0.0	0.0	2.00
	$\tau = 15$	0.7	0.0	2.02	0.0	0.0	2.00
Peng and Lu (2012)	BIC	21.5	9.9	2.26	1.5	1.9	2.10
	AIC	17	11.0	2.43	1.5	3.3	2.20
	GCV	20.5	6	2.30	1.5	3	2.18
	$\sqrt{\log n/n}$	21	15.6	2.67	1.5	4.1	2.26

TABLE 6.1

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

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 6.2, along with the corresponding results from two other papers. The proposed e-value based procedure performs best here for $\tau \geq 5$ for the smaller sample setting, and for $\tau \geq 7$ for larger sample setting.

7. Application: Linear mixed effect model for 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 et al., 2003; Wang et al., 2005; Trenberth, 2011). 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 $10^\circ \times 12^\circ$ grid, but a similar study on a $1^\circ \times 1^\circ$ gridded data by Ghosh, Luniya and Gupta (2009) suggested that there are both increasing and decreasing trends of extreme rainfall events, depending on the location. Additionally, Krishnamurty, Lall and Kwon (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. There are significant spatial and temporal variabilities at various scales discovered by Dietz and Chatterjee (2014) and Dietz and Chatterjee (2015).

Method		Setting 1	Setting 2
e-value based	$\tau = 1$	3	14
	$\tau = 2$	30	60
	$\tau = 3$	61	86
	$\tau = 4$	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
	$\tau = 12$	98	100
	$\tau = 15$	98	100
Bondell, Krishna and Ghosh (2010)		73	83
Peng and Lu (2012)		49	86
Fan and Li (2012)		90	100

TABLE 6.2

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

Here we attempt to identify the driving factors behind precipitation during the Indian monsoon season using our *e-value*-based model selection criterion. 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 on 35 potential covariates 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;
- 10 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

	$e_{rn}\mathcal{S}_{-j}$
- 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 7.1
Ordered values of $e_{rn}\mathcal{S}_{-j}$ when dropping one variable at a time in the Indian summer precipitation data

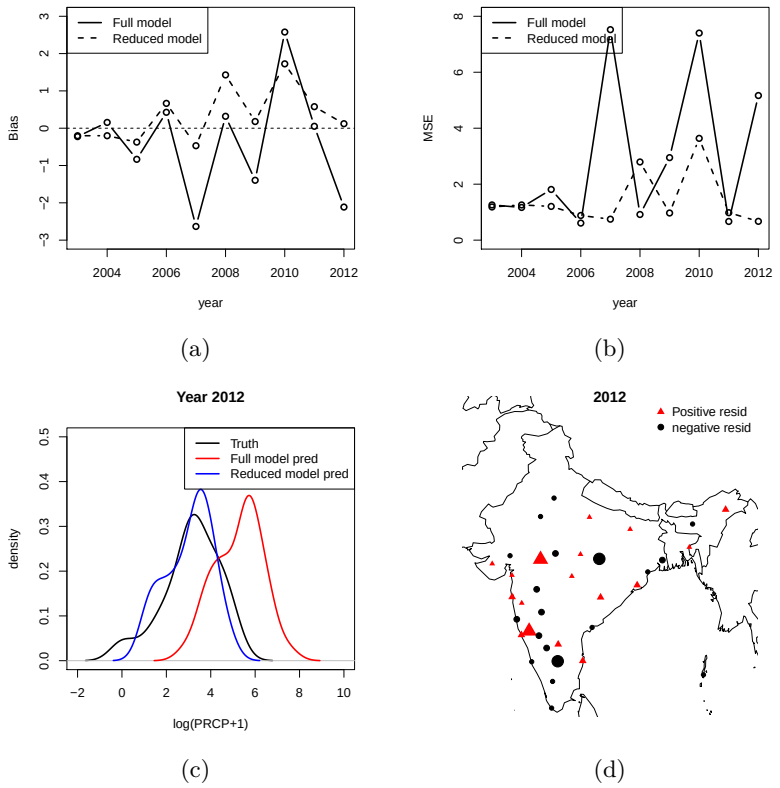


FIG 7.1. *Comparing full model rolling predictions with reduced models: (a) Bias across years, (b) MSE across years, (c) density plots for 2012, (d) stationwise residuals for 2012*

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 sciences.

As a modeling step, we consider the annual medians of all the above covariates as fixed effects, the log yearly rainfall at a weather station as response variable, and include year-specific random intercepts. Table 7.1 lists the estimated *e-values* in increasing order for the full model as well as all 35 models where a single variable is dropped. We use resample Monte Carlo sizes $R_1 = R_2 = 1000$. The variables that are listed before *none* appears in Table 7.1 are considered relevant by our *e-value* criterion.

All the 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 is based on the Clausius-Clapeyron relation is now classical knowledge in Physics. 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.

To check out-of-sample prediction performance of the estimated minimal adequate, we use a rolling validation scheme. For each of the 10 test years: 2003–2012, we select important variables from the model built on past 25 year's 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 7.1 summarizes results obtained through this process. Across all testing years, reduced model predictions have less bias as well as are more stable (panels a and b, respectively). The better approximations of truth by reduced models is also evident from the density plot for 2012 in panel c, and there does not seem to be any spatial patterns in its residuals as well (panel d).

8. Caveats, conclusions. We present above an expansive framework and principle, where the definition of a statistical model is very broad, estimation procedures very general, resampling algorithms broad and general. In such a scenario, we propose a scheme of simultaneous model selection and resampling-based inference, using the newly defined *e-value*. An extremely fast algorithm obtains consistent true model selection with probability tending to one by fitting and using only a singl model. Simulation results show that our algorithm performs better than traditional methods in two illustrative examples, and a case study on Indian summer precipitation identifies several important physical drivers of monsoon precipitation. Theoretical consistency results of multiple kinds are provided.

While the above framework is extremely broad-based, multiple details require cautious approach and more detailed studies. The choice of the resampling algorithm, the tuning parameter τ_n associated with it, should be subject to further scrutiny. Our results suggest excellent *asymptotic* properties and seem to be borne out in our simulation experiments, but finite-sample performance of our procedure needs further study. Higher order correctness is a possibility in our context since we use resampling methods, which deserves further study. We have remarked earlier that uniform convergence, local asymptotics and more deep asymptotic studies are needed to understand the workings of our proposal more thoroughly. The current framework includes *dimension asymptotics* where the parameter dimensions are allowed to grow with the sample size, but we do not include extremely

high-dimensional parameters in our study. The sensitivity of the results to the choice of the evaluation maps, and the way $E_n(y, [Y])$ is summarized to obtain the *e-value* deserve further study. A further, perhaps philosophical, issue is the sensitivity of the results to the choice of the preferred model. While in practice this may not matter much, the choice of the preferred model reflects a choice of paradigms and scientific principles.

Acknowledgments:. This research is partially supported by the National Science Foundation (NSF) under grants # IIS-1029711 and # DMS-1622483 and by the National Aeronautics and Space Administration (NASA). The first author also acknowledges the University of Minnesota Interdisciplinary Doctoral Fellowship program.

APPENDIX A: PROOFS

Proof of Theorem 2.1: We consider a generic point $\theta = \theta_{sn} + A_{sn}^{-1}t$. From the Taylor series expansion, we have

$$\Psi_{0sni(a)}(\theta) = \Psi_{0sni(a)}(\theta_{sn}) + \Psi_{1sni(a)}(\theta_{sn})A_{sn}^{-1}t + 2^{-1}t^T A_{sn}^{-T}\Psi_{2sni(a)}(\tilde{\theta}_{sn})A_{sn}^{-1}t, \text{ for } a = 1, \dots p_{sn},$$

and $\tilde{\theta}_{sn} = \theta_{sn} + cA_{sn}^{-1}t$ for some $c \in (0, 1)$.

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

$$R_{sn(a)}(\tilde{\theta}_{sn}, t) = t^T A_{sn}^{-T} \sum_{i=1}^{k_n} \Psi_{2sni(a)}(\tilde{\theta}_{sn})A_{sn}^{-1}t.$$

Thus we have

$$\begin{aligned}
&p_{sn}^{-1/2}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\theta_{sn} + A_{sn}^{-1}t) \\
&= p_{sn}^{-1/2}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} + p_{sn}^{-1/2}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{1sni}A_{sn}^{-1}t + 2^{-1}p_{sn}^{-1/2}A_{sn}^{-1}R_{sn}(\tilde{\theta}_{sn}, t) \\
&= p_{sn}^{-1/2}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} + p_{sn}^{-1/2}A_{sn}^{-1}\Gamma_{1sn}A_{sn}^{-1}t \\
&\quad + p_{sn}^{-1/2}A_{sn}^{-1}(\sum_{i=1}^{k_n}\Psi_{1sni} - \Gamma_{1sn})A_{sn}^{-1}t \\
&\quad + 2^{-1}p_{sn}^{-1/2}A_{sn}^{-1}R_{sn}(\tilde{\theta}_{sn}, t).
\end{aligned}$$

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

$$\mathbb{P}\Big[\big|p_{sn}^{-1/2}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\big| > C_0\Big] < \epsilon/2. \tag{A.1}$$

For this, we compute

$$\begin{aligned} & p_{sn}^{-1}\mathbb{E}|A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}|^2 \\ &= p_{sn}^{-1}\mathbb{E}\sum_{i,j=1}^{k_n}\Psi_{0sni}^T A_{sn}^{-T}A_{sn}^{-1}\Psi_{0snj} \\ &= p_{sn}^{-1}\mathrm{tr}A_{sn}^{-T}A_{sn}^{-1}\mathbb{E}\sum_{i=1}^{k_n}\Psi_{0sni}\Psi_{0sni}^T \\ &= p_{sn}^{-1}\mathrm{tr}A_{sn}^{-T}A_{sn}^{-1}\Gamma_{0sn} \\ &= O(1) \end{aligned}$$

from assumption (2.8).

Define

$$S_{sn}(t) = p_{sn}^{-1/2}A_{sn}^{-1}\Big(\sum_{i=1}^{k_n}\Psi_{0sni}(\theta_{sn}+A_{sn}^{-1}t) - \sum_{i=1}^{k_n}\Psi_{0sni}\Big) - p_{sn}^{-1/2}\Gamma_{1sn}^{-1}\Gamma_{0sn}t,$$

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

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

This follows from (2.9) and (2.10).

Note that

$$S_{sn}(t) = p_{sn}^{-1/2}A_{sn}^{-1}\Big(\sum_{i=1}^{k_n}\Psi_{1sni}-\Gamma_{1sn}\Big)A_{sn}^{-1}t + 2^{-1}p_{sn}^{-1/2}A_{sn}^{-1}R_{sn}(\tilde{\theta}_{sn},t).$$

Thus,

$$\begin{aligned} \sup_{|t|\le C}|S_{sn}(t)| \le\ & p_{sn}^{-1/2}\sup_{|t|\le C}\Big|A_{sn}^{-1}\Big(\sum_{i=1}^{k_n}\Psi_{1sni}-\Gamma_{1sn}\Big)A_{sn}^{-1}t\Big| \\ & + 2^{-1}p_{sn}^{-1/2}\sup_{|t|\le C}|A_{sn}^{-1}R_{sn}(\tilde{\theta}_{sn},t)|. \end{aligned}$$

We consider each of these terms separately.

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

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

Using $M = A_{sn}^{-1}(\sum_{i=1}^{k_n} \Psi_{1sni} - \Gamma_{1sn})A_{sn}^{-1}$ and (2.9), we get one part of the result.

For the other term, we similarly have

$$\begin{aligned}&\Big[\sup_{|t|\le C} |p_{sn}^{-1/2} A_{sn}^{-1} R_{sn}(\tilde\theta_{sn}, t)|\Big]^2 \\ &= p_{sn}^{-1} \sup_{|t|\le C} \Big[|A_{sn}^{-1} R_{sn}(\tilde\theta_{sn}, t)|\Big]^2 \\ &\le p_{sn}^{-1} \lambda_{max}(A_{sn}^{-T} A_{sn}^{-1}) \sup_{|t|\le C} |R_{sn}(\tilde\theta_{sn}, t)|^2 \\ &\le p_{sn}^{-1} \lambda_{max}(A_{sn}^{-1} A_{sn}^{-T}) \sup_{|t|\le C} |R_{sn}(\tilde\theta_{sn}, t)|^2 \\ &\le p_{sn}^{-1} a_{sn}^{-2} \sup_{|t|\le C} |R_{sn}(\tilde\theta_{sn}, t)|^2\end{aligned}$$

Note that

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

Now

$$\begin{aligned}&|R_{sn}(\tilde\theta_{sn}, t)|^2 \\ &= \sum_{a=1}^{p_{sn}} \big(R_{sn(a)}(\tilde\theta_{sn}, t)\big)^2 \\ &= \sum_{a=1}^{p_{sn}} \big(t^T A_{sn}^{-T} \sum_{i=1}^{k_n} \Psi_{2sni(a)}(\tilde\theta_{sn}) A_{sn}^{-1} t\big)^2 \\ &= \sum_{a=1}^{p_{sn}} \sum_{i,j=1}^{k_n} t^T A_{sn}^{-T} \Psi_{2sni(a)}(\tilde\theta_{sn}) A_{sn}^{-1} t t^T A_{sn}^{-T} \Psi_{2snj(a)}(\tilde\theta_{sn}) A_{sn}^{-1} t\end{aligned}$$

Based on this, we have

$$\begin{aligned}
&\sup_{|t|\le C} \big|R_{sn}(\tilde\theta_{sn},t)\big|^2 \\
&= \sup_{|t|\le C} \sum_{a=1}^{p_{sn}} \sum_{i,j=1}^{k_n} t^T A_{sn}^{-T} \Psi_{2sni(a)}(\tilde\theta_{sn}) A_{sn}^{-1} t t^T A_{sn}^{-T} \Psi_{2snj(a)}(\tilde\theta_{sn}) A_{sn}^{-1} t \\
&\le \sup_{|t|\le C} \sum_{a=1}^{p_{sn}} \sum_{i,j=1}^{k_n} t^T A_{sn}^{-T} M_{2sni(a)} A_{sn}^{-1} t t^T A_{sn}^{-T} M_{2snj(a)} A_{sn}^{-1} t \\
&\le \sup_{|t|\le C} \big|A_{sn}^{-1}t\big|^4 \sum_{a=1}^{p_{sn}} \Big(\sum_{i=1}^{k_n} \lambda_{max}\big(M_{2sni(a)}\big)\Big)^2 \\
&\le C^4 n \lambda_{max}^2(A_{sn}^{-T}A_{sn}^{-1}) \sum_{a=1}^{p_{sn}} \sum_{i=1}^{k_n} \lambda_{max}^2(M_{2sni(a)}).
\end{aligned}$$

Putting all these together, we have

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{|t|\le C} |p_{sn}^{-1/2} A_{sn}^{-1} R_{sn}(\tilde\theta_{sn},t)|\Big]^2 \\
&= p_{sn}^{-1}\mathbb{E}\Big[\sup_{|t|\le C} A_{sn}^{-1} R_{sn}(\tilde\theta_{sn},t)\Big]^2 \\
&\le p_{sn}^{-1} a_{sn}^{-2}\mathbb{E}\Big[\sup_{|t|\le C} \big|R_{sn}(\tilde\theta_{sn},t)\big|\Big]^2 \\
&= O(p_{sn}^{-1}a_{sn}^{-2})\mathbb{E}\Big[\sup_{|t|\le C} \big|R_{sn}(\tilde\theta_{sn},t)\big|\Big]^2 \\
&= O(p_{sn}^{-1} n a_{sn}^{-6}) \sum_{a=1}^{p_{sn}} \sum_{i=1}^{k_n} \mathbb{E}\lambda_{max}^2\big(M_{2sni(a)}\big) \\
&= o(1)
\end{aligned}$$

using (2.10).

Define

$$S_{sn}(t) = p_{sn}^{-1/2} A_{sn}^{-1} \Big(\sum_{i=1}^{k_n} \Psi_{0sni}(\theta_{sn} + A_{sn}^{-1}t) - \sum_{i=1}^{k_n} \Psi_{0sni}\Big) - p_{sn}^{-1/2}\Gamma_{1sn}^{-1}\Gamma_{0sn} t,$$

hence

$$\begin{aligned}
&p_{sn}^{-1/2} A_{sn}^{-1} \sum_{i=1}^{k_n} \Psi_{0sni}(\theta_{sn} + p_{sn}^{1/2} A_{sn}^{-1} t) \\
&\qquad = S_{sn}(t) + p_{sn}^{-1/2} A_{sn}^{-1} \sum_{i=1}^{k_n} \Psi_{0sni} + A_{sn}^{-1}\Gamma_{1sn} A_{sn}^{-1} t.
\end{aligned}$$

Hence we have

$$\begin{aligned}
&\inf_{|t|=C}\Big\{p_{sn}^{-1/2}t^T\Gamma_{1sn}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\theta_{sn}+p_{sn}^{1/2}A_{sn}^{-1}t)\Big\}\\
&=\inf_{|t|=C}\Big\{t^T\Gamma_{1sn}S_{sn}(t)+p_{sn}^{-1/2}t^T\Gamma_{1sn}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\\
&\qquad+t^T\Gamma_{1sn}A_{sn}^{-1}\Gamma_{1sn}A_{sn}^{-1}t\Big\}\\
&\geq\inf_{|t|=C}t^T\Gamma_{1sn}S_{sn}(t)+p_{sn}^{-1/2}\inf_{|t|=C}t^T\Gamma_{1sn}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\\
&\qquad+\inf_{|t|=C}t^T\Gamma_{1sn}A_{sn}^{-1}\Gamma_{1sn}A_{sn}^{-1}t\\
&\geq-C\delta_{1s}\sup_{|t|=C}|S_{sn}(t)|-C\delta_{1s}p_{sn}^{-1/2}|A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}|+C^2\delta_{0s}
\end{aligned}$$

The last step above utilizes facts like $a^Tb\geq-|a||b|$.

Consequently, defining $C_1=C\delta_{0s}/\delta_{1s}$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\inf_{|t|=C}\Big\{t^T\Gamma_{1sn}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\theta_{sn}+p_{sn}^{1/2}A_{sn}^{-1}t)\Big\}<0\Big]\\
\leq&\mathbb{P}\Big[\sup_{|t|=C}|S_{sn}(t)|+|A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}|>C_1\Big]\\
\leq&\mathbb{P}\Big[\sup_{|t|=C}|S_{sn}(t)|>C_1/2\Big]+\mathbb{P}\Big[|A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}|>C_1/2\Big]\\
&<\epsilon,
\end{aligned}$$

for all sufficiently large n, using (A.1) and (A.2).

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

Since we have

$$\sup_{|t|<C}|S_{sn}(t)|=o_P(1),$$

and T_{sn} lies in the set $|t| < C$, define $-R_{sn} = S_{sn}(T_{sn}) = o_P(1)$. We consequently have

$$\begin{aligned}-R_{sn} &= S_{sn}(T_{sn}) \\ &p_{sn}^{-1/2}A_{sn}^{-1}(\sum_{i=1}^{k_n}\Psi_{0sni}(\theta_{sn}+A_{sn}^{-1}T_{sn}) - \sum_{i=1}^{k_n}\Psi_{0sni}) - p_{sn}^{-1/2}\Gamma_{1sn}^{-1}\Gamma_{0sn}T_{sn} \\ &p_{sn}^{-1/2}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} - p_{sn}^{-1/2}\Gamma_{1sn}^{-1}T_{sn}.\end{aligned}$$

Thus,

$$\begin{aligned}T_{sn} &= -\Gamma_{0sn}^{-1}\Gamma_{1sn}A_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} + p^{1/2}\Gamma_{0sn}^{-1}\Gamma_{1sn}R_{sn} \\ &= -\Gamma_{0sn}^{-1/2}\sum_{i=1}^{k_n}\Psi_{0sni} + p^{1/2}\Gamma_{0sn}^{-1}\Gamma_{1sn}R_{sn}.\end{aligned}$$

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

$$\mathbb{E}\big[p^{1/2}c^T\Gamma_{0sn}^{-1}\Gamma_{1sn}R_{sn}\big]^2 = 0(1)$$

using (2.8).

Using (2.11) we also have that for any c with $|c| = 1$

$$c^T T_{sn} \overset{\mathcal{D}}{\Rightarrow} N(0,1).$$

□

Proof of Theorem 3.1: *Part 1* follows directly from assumption (C3).

Part 2. Assuming now that $\mathcal{M}_n$ is an adequate model, we use (C1) property of E_n:

$$E_n(\hat{G}_{mn}, [\hat{G}_{*n}]) = E_n\Big(\hat{G}_{mn} - G_{*n}, [\hat{G}_{*n} - G_{*n}]\Big), \tag{A.3}$$

and decompose the first argument

$$\hat{G}_{mn} - G_{*n} = (\hat{G}_{mn} - \hat{G}_{*n}) + (\hat{G}_{*n} - G_{*n}). \tag{A.4}$$

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

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

where T_{mn} is non-degenerate at the $\mathcal{S}_n$ indices. In terms of these, we can write the j-th element of $G_{mn}(.) \equiv G(.)$ as

$$G_j(\hat{\theta}) = G_j(\theta) + a_n^{-1} G_{1j}^T(\theta) T_n + 2a_n^{-2} T_n^T R_j(\theta + c a_n^{-1} T_n) T_n.$$

Our conditions ensure that $\mathbb{E}|T_n|^4 < \infty$, consequently we have that $a_n(\hat{G} - G) = G_1^T T_n + R_n$, with $\mathbb{E}|R_n^2| = O(a_n^{-2})$. Coming back to the first summand of the right-hand side in (A.4) we get

$$\hat{G}_{mn} - \hat{G}_{*n} = G_{mn} - G_{*n} + R_n, \tag{A.5}$$

where $\mathbb{E}|R_n^2| = O(\min\{a_{sn}, a_{*n}\}^{-2})$. Since $\mathcal{M}_n$ is an adequate model, $G_{mn} - G_{*n} = o((\min\{a_{sn}, a_{*n}\}^{-1}))$. Thus, substituting the above right-hand side in (A.4) we get

$$\begin{aligned} &\left| E_n\left(\hat{G}_{mn} - G_{*n}, \left[\hat{G}_{*n} - G_{*n}\right]\right) - E_n\left(\hat{G}_{*n} - G_{*n}, \left[\hat{G}_{*n} - G_{*n}\right]\right)\right| \\ &\qquad \leq |R_n|^{\alpha}, \end{aligned} \tag{A.6}$$

from of Lipschitz continuity of E_n given in C2, where $\mathbb{E}|R_n^2| = O(\min\{a_{sn}, a_{*n}\}^{-2})$. The result now follows.

Part 3. Since the evaluation map E_n is invariant under location and scale transformations, we have

$$E_n(\hat{G}_{mn}, [\hat{G}_{*n}]) = E_n\Big(a_{*n}(\hat{G}_{mn} - G_{*n}), \left[a_{*n}(\hat{G}_{*n} - G_{*n})\right]\Big). \tag{A.7}$$

Decomposing the first argument,

$$a_{*n}(\hat{G}_{mn} - G_{*n}) = \frac{a_{*n}}{a_{sn}} . a_{sn}(\hat{G}_{mn} - G_{mn}) + a_{*n}(G_{mn} - G_{*n}). \tag{A.8}$$

Since $\mathcal{M}_n$ is strictly inadequate, given $\delta > 0$ there exists a subsequence indexed by $\{j_n\}$ such that $a_{*n}|G_{mj_n} - G_{*j_n}| \to \infty$ as $n \to \infty$. The result now follows by an application of condition C4. □

Proof of Theorem 4.1: This proof has steps similar to that of the proof of Theorem 2.1, apart from several additional technicalities. We omit the details.

□

Proof of Theorem 4.2: Several details here are similar to those of the proof of Theorem 3.1, consequently we skip some details here.

We have

$$\begin{aligned} & E\Big(G(\hat{\theta}_{rsn}), [G(\hat{\theta}_{r*n})]\Big) \\ &= E\Big(b_{*n}(\hat{G}_{rsn} - \hat{G}_{*n}), [b_{*n}(\hat{G}_{r*n} - \hat{G}_{*n})]\Big) \\ &= E\Big(b_{*n}(\hat{G}_{rsn} - \hat{G}_{sn}) - b_{*n}(\hat{G}_{sn} - \hat{G}_{*n}), [b_{*n}(\hat{G}_{r*n} - \hat{G}_{*n})]\Big). \end{aligned}$$

Our results from previous sections ensure that $\hat{G}_{sn}$ converges to G_{sn} in probability for all models $\mathcal{M}_n$. The result now follows directly for adequate models, and from (3.7) for inadequate models.

□

Proof of theorem 5.1: 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 under suitable conditions.

Suppose $\mathbb{T}_0 = \mathcal{E}(0_p, \mathbb{I}_p, g)$. Affine invariance implies invariant to rotational transformations, and since the evaluation functions we consider decrease along any ray from the origin because of (D5), $E(\theta, \mathbb{T}_0)$ is a monotonocally decreasing function of $|\theta|$ for any $\theta \in \mathbb{R}^p$. Now consider the models $\mathcal{M}_{10}, \mathcal{M}_{20}$ that have 0 in all indices outside $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Take some $\theta_{10} \in \Theta_{10}$, which is the parameter space corresponding to $\mathcal{M}_{10}$, 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

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

where $\mathbb{E}_s$ denotes the expectation taken over the marginal of the distributional argument $\mathbb{T}_0$ at indices $\mathcal{S}_1$. Notice now that by construction $\theta_{1\delta} \in \Theta_{20}$, the parameter space corresponding to $\mathcal{M}_{20}$, 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 $\mathbb{E}_{s1} E(\theta_{10}, \mathbb{T}_0) > \mathbb{E}_{s2} E(\theta_{20}, \mathbb{T}_0)$, with θ_{20} denoting a general element in Θ_{20}.

Now combining (E1) and (E2) we get $a_n V_n^{-1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathbb{T}_0$. Suppose $\mathbb{T}_n := [a_n V_n^{-1/2}(\hat{\theta} - \theta_0)]$. Now choose a positive $\epsilon < (\mathbb{E}_{s1} E(\theta_{10}, \mathbb{T}_0) - \mathbb{E}_{s2} E(\theta_{20}, \mathbb{T}_0))/2$. Then, for large enough n we shall have

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

following condition (D4). Similarly we have $|\mathbb{E}_{s2} E(\theta_{20}, \mathbb{T}_n) - \mathbb{E}_{s2} E(\theta_{20}, \mathbb{T}_0)| < \epsilon$ for the same n for which the above holds. This implies $\mathbb{E}_{s1} E(\theta_{10}, \mathbb{T}_n) > \mathbb{E}_{s2} E(\theta_{20}, \mathbb{T}_n)$.

Now apply the affine transformation $t(\theta) = V_n^{1/2}\theta/a_n + \theta_0$ to both arguments of the evaluation function above. This will keep the depths constant following affine invariance, i.e. $E(t(\theta_{10}), [\hat{\theta}]) = E(\theta_{10}, \mathbb{T}_n)$ and $E(t(\theta_{20}), [\hat{\theta}]) = E(\theta_{20}, \mathbb{T}_n)$. Since this transformation maps Θ_{10} to Θ_1, the parameter space corresonding to $\mathcal{M}_1$, we get $\mathbb{E}_{s1}E(t(\theta_{10}), [\hat{\theta}]) > \mathbb{E}_{s2}E(t(\theta_{20}), [\hat{\theta}])$, i.e. $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$.

□

Proof of corollary 5.1: 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 5.1 implies $e_n(\mathcal{M}_0) > e_n(\mathcal{M}^c)$ for any adequate model $\mathcal{M}^c \in \mathbb{M}_0$ and large enough n.

For an inadequate model $\mathcal{M}^w$, suppose $N(\mathcal{M}^w)$ is the integer such that $e_{n_1}(\mathcal{M}^w) < e_{n_1}(\mathcal{M}_*)$ for all $n_1 > N(\mathcal{M}^w)$. Part 3 of theorem 3.1 ensures that such an integer exists for every inadequate model. Now define $N = \max_{\mathcal{M}^w \in \mathbb{M}_0} N(\mathcal{M}^w)$: we can do this since $\mathbb{M}_0$ has countably finite elements. Thus $e_{n_1}(\mathcal{M}_0)$ is larger than e-values of all inadequate models in $\mathbb{M}_0$. □

Proof of corollary 5.2: Consider $j \in \mathcal{S}_0$. Then $\theta_0 \notin \mathcal{S}_{-j}$, hence $\mathcal{S}_{-j}$ is inadequate. By choice of n_1, e-values of all inadequate models are less than that of $\mathcal{S}_*$, hence $e_{n_1}(\mathcal{S}_{-j}) < e_{n_1}(\mathcal{S}_*)$.

On the other hand, suppose there exists a j such that $e_{n_1}(\mathcal{S}_{-j}) \leq e_{n_1}(\mathcal{S}_*)$ but $j \notin \mathcal{S}_0$. Now $j \notin \mathcal{S}_0$ means that $\mathcal{S}_{-j}$ is an adequate model. Since $\mathcal{S}_{-j}$ is nested within $\mathcal{S}_*$ for any j, and the full model is always adequate, we have $e_{n_1}(\mathcal{S}_{-j}) > e_{n_1}(\mathcal{S}_*)$ by theorem 5.1: leading to a contradiction and thus completing the proof. □

REFERENCES

Akaike, H. (1970). Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics* **22** 203 – 217.

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** 1069 – 1077.

Bose, A. and Chatterjee, S. (2003). Generalized Bootstrap for Estimators of Minimizers of Convex Functions. *Journal of Statistical Planning and Inference* **117** 225 – 239.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 211 – 252.

Chang, C.-H., Huang, H.-C. and Ing, C.-K. (2014). Asymptotic Theory of Generalized Information Criterion for Geostatistical Regression Model Selection. *The Annals of Statistics* **42** 2441 – 2468.

Chatterjee, S. and Bose, A. (2005). Generalized Bootstrap for Estimating Equations. *The Annals of Statistics* **33** 414 – 436.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Univ. Press.

Dietz, L. and Chatterjee, S. (2014). Logit-Normal Mixed Model for Indian Monsoon Precipitation. *Nonlinear Processes in Geophysics* **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* 239 – 246. Springer, New York, NY, USA.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7** 1 – 26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC press, Boca Raton, USA.

Fan, Y. and Li, R. (2012). Variable Selection in Linear Mixed Effects Models. *The Annals of Statistics* **40** 2043 – 2068.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric multivariate and related distributions. Monographs on Statistics and Applied Probability* **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. *Atmospheric Science Letters* **10** 285-290.

Goswami, B. N., 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. *The Annals of Statistics* **36** 1669 – 1692.

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

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83** 875–890.

Krishnamurthy, V. and Kinter III, J. L. (2003). The Indian Monsoon and its Relation to Global Climate Variability. In *Global Climate: Current Research and Uncertainties in the Climate System* (X. Rodo and F. A. Comin, eds.) 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. *Journal of Climate* **22** 4737-4746.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and fiction.

Econometric Theory **21** 21 – 59.

Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics* **21** 255 – 285.

Michel, R. and Pfanzagl, J. (1971). The Accuracy of the Normal Approximation for Minimum Contrast Estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **18** 73 – 84.

Moon, J.-Y., Wang, B., Ha, K.-J. and Lee, J.-Y. (2013). Teleconnections Associated with Northern Hemisphere Summer Monsoon Intraseasonal Oscillation. *Climate dynamics* **40** 2761 – 2774.

Mosler, K. (2013). Depth Statistics. In *Robustness and Complex Data Structures* (C. Becker, R. Fried and S. Kuhnt, eds.) 17-34. Springer Berlin Heidelberg.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42** 789–817.

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

Peng, H. and Lu, Y. (2012). Model Selection in Linear Mixed Effect Models. *Journal of Multivariate Analysis* **109** 109 – 129.

Pfanzagl, J. (1969). On the Measurability and Consistency of Minimum Contrast Estimates. *Metrika* **14** 249 – 272.

Ročkovà, V. and George, E. I. (2016). The Spike-and-Slab LASSO. *J. Amer. Statist. Assoc.* **0** 0–0. http://dx.doi.org/10.1080/01621459.2016.1260469.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6** 461 – 464.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York, USA.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. *arXiv preprint arXiv:1506.06266*.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association* **111** 600 – 620.

Trenberth, K. E. (2011). Changes in precipitation with climate change. *Climate Research* **47** 123–138.

Trenberth, K. E., Dai, A., Rasmussen, R. M. and Parsons, D. B. (2003). The changing character of precipitation. *Bulletin of the American Meteorological Society* **84** 1205–1217.

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

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. *Geophysical Research Letters* **32**.

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics* **14** 1261 – 1295.

Yang, Y. (2005). Can the Strengths of AIC and BIC be Shared? A Conflict Between Model Identification and Regression Estimation. *Biometrika* **92** 937 – 950.

Zuo, Y. and Serfling, R. (2000). General Notions of Statistical Depth Function. *The Annals of Statistics* **28** 461 – 482.

School of Statistics,
University of Minnesota,
224 Church Street S. E., Minneapolis 55455, USA
E-mail: majum010@umn.edu
chat019@umn.edu