The Annals of Statistics
2020, Vol. 48, No. 5, 2953–2981
https://doi.org/10.1214/19-AOS1917

VALID POST-SELECTION INFERENCE IN MODEL-FREE LINEAR REGRESSION

BY ARUN K. KUCHIBHOTLA*, LAWRENCE D. BROWN, ANDREAS BUJA, JUNHUI CAI, EDWARD I. GEORGE AND LINDA H. ZHAO

*Department of Statistics, The Wharton School, University of Pennsylvania, *arunku@wharton.upenn.edu*

Modern data-driven approaches to modeling make extensive use of covariate/model selection. Such selection incurs a cost: it invalidates classical statistical inference. A conservative remedy to the problem was proposed by Berk et al. (*Ann. Statist.* **41** (2013) 802–837) and further extended by Bachoc, Preinerstorfer and Steinberger (2016). These proposals, labeled "PoSI methods," provide valid inference after arbitrary model selection. They are computationally NP-hard and have limitations in their theoretical justifications. We therefore propose computationally efficient confidence regions, named "UPoSI"[1] and prove large-p asymptotics for them. We do this for linear OLS regression allowing misspecification of the normal linear model, for both fixed and random covariates, and for independent as well as some types of dependent data. We start by proving a general equivalence result for the post-selection inference problem and a simultaneous inference problem in a setting that strips inessential features still present in a related result of Berk et al. (*Ann. Statist.* **41** (2013) 802–837). We then construct valid PoSI confidence regions that are the first to have vastly improved computational efficiency in that the required computation times grow only quadratically rather than exponentially with the total number p of covariates. These are also the first PoSI confidence regions with guaranteed asymptotic validity when the total number of covariates p diverges (almost exponentially) with the sample size n. Under standard tail assumptions, we only require $(\log p)^7 = o(n)$ and $k = o(\sqrt{n/\log p})$ where k ($\leq p$) is the largest number of covariates (model size) considered for selection. We study various properties of these confidence regions, including their Lebesgue measures, and compare them theoretically with those proposed previously.

1. Introduction and motivation.

1.1. *Motivation of the problem.* In recent times, there has been a crisis in the sciences because too many published research results are found to lack reproducibility. Some of this crisis has been attributed to a failure of statistical methods to account for data-dependent exploration and modeling that precedes statistical inference. Data-dependent actions such as selection of subsets of cases, of covariates, of responses, of transformations and of model types has been aptly named "researcher degrees of freedom" (Simmons, Nelson and Simonsohn (2011)), and these may well be significant contributing factors in the current crisis. Classical statistics does not account for them because it is built on a framework where all modeling decisions are made *independently of the data on which inference is based*. But if the data are in fact used to this end prior to statistical inference, then such inference loses its justifications and the ensuing validity conferred on it by classical theories. It is therefore

Received October 2018; revised September 2019.
MSC2020 subject classifications. 62J05, 62F40, 62F25, 62F12.
Key words and phrases. Simultaneous inference, multiplier bootstrap, uniform consistency, high-dimensional linear regression, concentration inequalities, Orlicz norms, model selection.

[1] "U" is for "uniform" or "universal."

critical that the theory of statistical inference be brought up to date to account for data-driven modeling. Updating the theory that justifies statistical inferences usually requires modifying the procedures of inference such as hypothesis tests and confidence intervals. As a consequence, the new procedures may lose some power relative to the previously stipulated but illusionary power derived from classical theories. This is a necessary price to be paid for better justification of statistical inference in the context of the pre-inferential liberties taken in today's data-analytic practice. While updating of statistical theories and inference procedures will not solve all problems underlying the current crisis, it is a necessary step as it may help mitigate at least some aspects of the crisis. In what follows, we refer to all data-analytic decisions that are made using the data prior to inference as "data-driven modeling."

A second issue with theories of classical statistical inference is that many of them rely on the assumption that the data have been correctly modeled in a probabilistic sense. This means the theories tend to assume that the probability model used for the data correctly captures the observable features of the data generating process. Justifications of statistical inferences derived from such theories may therefore be invalid if the model is incorrect or (using the technical term) "misspecified." With the proliferation of data-analytic approaches in science and business, it is becoming ever more unrealistic to assume that all statistical models are correctly specified and inferences are made only after carefully vetting the model for correct specification, for example, using model diagnostics. Such vetting may never have been realistic in the first place, and it should also be said that pre-inferential diagnostics should be counted among "researcher degrees of freedom" as they may result in data-driven modeling decisions. It is therefore a mandate of realism to look for so-called "model-robust" methods of statistical inference, and for statistical theory to provide their justifications. In matters of misspecification, the situation is somewhat less dire than data-driven modeling as there exists a rich literature on the study of inference when models are misspecified. We will naturally draw on extant proposals for misspecification-robust or (using the technical term) "model-robust" inference and adapt them to our purposes.

To summarize, there exist at least two ways in which inference methods derived from classical mathematical statistics can be invalidated, namely,

(P1) data-driven modeling prior to statistical inference, and
(P2) model misspecification.

In light of the reproducibility crisis in the sciences, it is of considerable interest, even urgency, to develop methods of statistical inference and associated theoretical justifications that account for both **(P1)** and **(P2)**. Even though these problems are manifest in almost all statistical procedures used in practice, it is no simple task to provide methods of valid statistical inference that address these problems in greater generality. For this reason, the present article puts forth specifically a method of valid inference for the case that the fitting procedure is ordinary least squares (OLS) linear regression. Here, there exists a literature that documents the drastic effects of ignoring **(P1)** and **(P2)**; see, for example, Buehler and Feddersen (1963), Olshen (1973), Rencher and Pun (1980) and Freedman (1983). We will address one particular form of problem **(P1)**, namely, data-driven selection of regressor variables/covariates, and we will deal with several forms of problem **(P2)**.

Some of the earliest work that studies estimators under data-dependent modeling **(P1)** include Hjort and Claeskens (2003) and Claeskens and Carroll (2007). Although these articles deal with a general class of statistical procedures, a major limitation, in view of the current article, is that the data-dependent modeling is restricted to a very narrow class of principled variable selection methods such as optimization of AIC or some other information criterion. The fact is, however, that few data analysts will confine themselves to a strict protocol of data-driven modeling. To address broader aspects of "researcher degrees of freedom," there

have more recently emerged proposals that provide validity of statistical inference in the case of arbitrary data-driven selection of covariates. The first such proposal was by Berk et al. (2013) who solve the problem allowing misspecified response means but retaining the classical assumptions of homoskedastic and normally distributed errors. We refer to Berk et al. (2013) for many other prior works related to problem **(P1)** where data-driven modeling consists of selection of covariates. A more recent article that expands on Berk et al. (2013) is by Bachoc, Preinerstorfer and Steinberger (2016). An alternative approach is by Lee et al. (2016), Tibshirani et al. (2016), Tian, Bi and Taylor (2016) (for example). Similar to Hjort and Claeskens (2003), these proposals do not insure validity of inference against arbitrary covariate selection but against specific selection methods such as the lasso or stepwise forward selection. This type of post-selection inference is conditional on the selected model and dependent on distributional assumptions, thereby not addressing problem **(P2)**.

The present article is close in spirit to Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016) and lends their approaches a considerable degree of generality by covering both fixed covariates (as in these references) and (newly) random covariates. Bachoc, Preinerstorfer and Steinberger (2016) is the only work we know of that provides valid statistical inference under arbitrary data-dependent covariate selection and general misspecification of the regression models. Their framework assumes a situation where the set of submodels is finite and of fixed cardinality independent of the sample size. Their method of statistical inference is NP-hard, hence requires computational heuristics. To overcome these limitations, we propose here a simplified procedure, called "UPoSI," with the following properties: (1) it is comparatively computationally efficient with at most polynomial complexity in the total number of covariates, and (2) it allows the set of submodels to grow almost exponentially as a function of the sample size. Thus the procedure is also in the spirit high-dimensional statistics where the total number of covariates is allowed to be much larger than the sample size.

1.2. *Overview.* In what follows, the term "model selection" will always mean arbitrary data-driven selection of covariates, which is the only aspect of problem **(P1)** that will be addressed in this article. Furthermore, the only fitting method considered here is OLS linear regression. This limitation is for expository purposes, and results for more general types of regressions will be given elsewhere. Problem **(P2)** will be addressed by the complete absence of modeling assumptions. In particular, it will *not* be assumed that the conditional response means behave linearly in the covariates and equally it will *not* be assumed that the errors are homoskedastic and normally distributed. The goal of the UPoSI approach described here is to provide confidence regions for linear regression coefficients that are valid after model selection and allowing complete misspecification. In the process, we will prove simple but powerful results about linear regression that lend themselves to proving the validity of confidence regions. The main contributions of the current paper are as follows:

1. We treat OLS linear regression as a fitting method for linear equations while treating the associated Gaussian linear model merely as a working model that is *not* assumed to be correctly specified. We consider the case where the observations are random vectors comprised of a response variable and one or more covariates, allowing the latter to be random rather than fixed. Note that fixed covariates are assumed in the settings of Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016). Random covariates require us to interpret and understand what is being estimated more carefully. See Buja (2019) for an explanation why under misspecification the treatment of random covariates as fixed is not justified.

2. Following Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016), we decouple the inference problem from model selection, meaning that the inferences proposed

here are valid no matter how the model selection was done. This feature has pluses and minuses. On the plus side, inferences will be valid even in the presence of ad-hoc and informal selection decisions made by the data analyst, including, for example, visual diagnostics based on residual plots. On the minus side, decoupling implies that inferences cannot take into account any properties of the model selection procedure when in fact only one such procedure was used. A strong argument by Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016) in favor of decoupling, however, is that in reality data analysts will rarely limit themselves to one and only one formal selection method if it produces unsatisfactory results on the data at hand. Therefore, in order to truly contribute to solving the reproducibility crisis in the sciences, unreported informal selection should be assumed and accounted for. Decoupling of model selection and inference has a further benefit: It solves the circularity problem by permitting selection to start over and over as often as the data analyst pleases. Inferences in all selected models will be valid, whether they are found satisfactory or unsatisfactory for whatever reasons.

3. Our theory provides validity of post-selection inferences even when model selection is applied to a very large number of covariates—almost exponential in the sample size. Thus the theory is in the spirit of contemporary high-dimensional statistics which is interested in problems where the number of variables is larger than the sample size. We require, of course, model selection to produce models of size smaller than the sample size in order to avoid trivial collinearity when the number of covariates exceeds the sample size.

4. We mostly focus on one simple strategy for valid post-selection inference that has the advantage of great simplicity, both in theory and in computation—its computational cost being proportional to the number p of covariates. This is surprising as the computational complexity of Berk et al. (2013) is exponential in p due to searching through all coefficients in all submodels. The drawback of the present strategy is that its confidence regions are not aligned with the coordinate axes in covariate space, hence do not immediately provide confidence intervals for the slope parameters of the form "estimate $\pm$ half-width."

5. Most of the present results are based on deterministic inequalities that justify valid post-selection inference even when the observations are structurally dependent. These proof techniques may not produce best possible rates in some contexts, but the resulting inferences will be more robust to violations of the independence assumption.

As a caveat, it should be stated that we do not address the question of when linear regression is appropriate in a given data analytic situation when misspecification is present. We consider it a reality that many if not most linear regressions are fitted in the presence of various degrees of misspecification, and reporting results for interpretation should be accompanied by statistical inference just the same. Our goal is therefore limited to providing asymptotic justification of inference in the presence of misspecification *and* after data-driven model selection.

1.3. *Alternative approaches*. An "obvious" approach to valid post-selection inference is based on sample splitting, as examined by Rinaldo et al. (2016): Split the data randomly into two disjoint parts, then use one part for selecting a model $\hat{M}$ and the other part for inference in the selected model $\hat{M}$. If the two parts of the data are stochastically independent of each other, post-selection inference is valid. For independent observations, Rinaldo et al. (2016) were able to provide very general and powerful results. Sample splitting has considerable appeal due to its universal applicability under independence of the two parts: it "works" for any type of model selection, formal or informal, as well as for any type of model being fitted. It has some drawbacks, too, an obvious one being the reduced sample sizes of the two parts, which increase the sampling variability both at the model selection stage and at the inference stage. Another drawback is the requirement of independence of the two parts, which makes it less obvious how to generalize sample splitting to dependent data. To the customers

of statistical inference, it may also be disconcerting that the splitting procedure could have produced different results in the hands of another data analyst who would have used another random split. This potential lack of reproducibility is not a welcome feature in the context of the reproducibility crisis. It also raises concerns over potential abuse whereby data analysts sift through multiple random splits until they see results they like, making use of another "researcher degree of freedom." On the other hand, even scrupulous data analysts should be expected to look at multiple random splits if only to learn about the stability of their model selection and subsequent inferences. This is a valid concern because experience shows that for most regression data all-subset searches reveal large numbers of submodels with nearly identical performance. In summary, sample splitting has great appeal and could be highly informative, but it could also open up another Pandora's box that defeats the solution to the problem it was meant to solve.

A different type of post-selection guarantees are available from the selective inference approach of Lee et al. (2016), Tibshirani et al. (2016), Tian, Bi and Taylor (2016) and Fithian, Sun and Taylor (2014) when model selection is of a pre-specified form such as lasso selection or stepwise forward selection. The inference guarantees they provide are conditional on the selected model. Their approach is ingeniously tailored to these specific formal selection methods and takes advantage of their properties. It is, however, a model-trusting approach that relies much on the correctness of the assumed model as being finite-sample correct under a Gaussian or other exponential linear model with fixed covariates. For this reason and because so much conditioning is performed, it is unlikely that this approach enjoys much robustness to misspecification (see, e.g., Section A.20 of Tibshirani et al. (2018)). By comparison, we strive here for model robustness by limiting ourselves to asymptotically correct coverage that is marginal rather than conditional, and by allowing covariates to be treated as random rather than fixed.

The larger point to be reiterated here is that tailoring post-selection inference to a specific formal selection method such as the lasso does not address the issue that data analysts may not limit themselves to just one formal selection method and nothing else. It may be more realistic to assume, as we do here, that they exercise broader liberties that include meta-selection among multiple formal selection methods as well as informal selection using exploratory and diagnostics tools. Post-selection inference that casts a wider net on selection methods may have a better chance of making an at least partial contribution to solving the reproducibility crisis in the sciences.

1.4. *Organization.* The remainder of the paper is organized as follows. Section 2 provides the necessary notation for a rigorous formulation of the problem of valid post-selection inference. In Section 3, the problem of post-selection inference is shown to be equivalent to a problem of simultaneous inference. In Section 4, we present a strategy for valid post-selection inference along with its main features. In Section 5, we study the rate of convergence of the Lebesgue measure of the confidence regions under independence. In Section 6, we specialize our regions for the case of fixed covariates and compare them with those in Berk et al. (2013). Section 8 describes an implementation method based on the multiplier bootstrap. Section 9 presents the simulation study comparing our method to Berk et al. (2013) and selective inference of Tibshirani et al. (2016). In Section 10, we discuss various advantages and disadvantages of the approach presented. Finally, Section 11 summarizes the results.

The Supplement (Kuchibhotla et al. (2020)) is organized as follows. Section S.1 provides further simulation comparisons. Section S.3 provides a simple generalization to regression data with problems such as missing values and/or outliers, in which case one may want to modify the estimators of the relevant second moment matrices and vectors. Section S.4 points out an interesting connection between the post-selection confidence regions proposed

here and the estimators proposed in the high-dimensional linear regression literature. Many of the proofs are deferred to Sections S.5–S.9. Most of the discussion in the paper is based on the assumption of independent random vectors, although comments about applicability to dependent random vectors are given in appropriate places. Section S.10 provides theoretical background about a high-dimensional central limit theorem and the consistency of the multiplier bootstrap. These results are required for the computation of joint quantiles for the proposed confidence regions. Section S.11 describes the functional dependence setting where the computation of required quantiles is similar to that in the independence setting.

2. Notation and problem formulation.

2.1. *Notation related to vectors, matrices and norms.* For any vector $v \in \mathbb{R}^q$ and $1 \le j \le q$, $v(j)$ denotes the jth coordinate of v. For any nonempty subset $M \subseteq \{1, 2, \ldots, q\}$, $v(M)$ denotes the subvector of v with indices in M. For instance, if $M = \{2, 4\}$ and $q \ge 4$, then $v(M) = (v(2), v(4))$. If $M = \{j\}$ is a singleton, then $v(j)$ is used instead of $v(\{j\})$. Therefore, $v(M) \in \mathbb{R}^{|M|}$ where $|M|$ denotes the cardinality of M.

For any symmetric matrix $A \in \mathbb{R}^{q\times q}$ and $M \subseteq \{1, 2, \ldots, q\}$, let $A(M)$ denote the submatrix of A with indices in $M \times M$ and for $1 \le j, k \le q$, let $A(j, k)$ denote the value at the jth row and kth column of A.

Define the r-norm of a vector $v \in \mathbb{R}^q$ for $1 \le r \le \infty$ as usual by

$$\|v\|_r := \left(\sum_{j=1}^{q} \big|v(j)\big|^r\right)^{1/r} \quad \text{for } 1 \le r < \infty, \quad \text{and} \quad \|v\|_\infty := \max_{1\le j\le q} \big|v(j)\big|.$$

Let $\|v\|_0$ denote the number of nonzero entries in v (note this is not a norm). For any symmetric matrix A, let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A. Also, let the elementwise maximum and the operator norm be defined, respectively, as

$$\|A\|_\infty := \max_{1\le j,k\le q} \big|A(j, k)\big| \quad \text{and} \quad \|A\|_{\text{op}} := \sup_{\|\delta\|_2\le 1} \|A\delta\|_2.$$

The following inequalities will be useful:

$$(1) \quad \|v\|_1 \le \|v\|_0^{1/2}\|v\|_2, \qquad \|Av\|_\infty \le \|A\|_\infty \|v\|_1, \quad \text{and} \quad \big|u^\top Av\big| \le \|A\|_\infty \|u\|_1\|v\|_1,$$

where $A \in \mathbb{R}^{q\times q}$ and $u, v \in \mathbb{R}^q$.

2.2. *Notation related to regression data and OLS.* Let $(X_i^\top, Y_i)^\top \in \mathbb{R}^p \times \mathbb{R}$ $(1 \le i \le n)$ represent a sample of n observations. The covariate vectors $X_i \in \mathbb{R}^p$ are column vectors. It is common to include an intercept term when fitting the linear regression. To avoid extra notation, we assume that all covariates under consideration are included in the vectors X_i, so the data analyst may take the first coordinate of X_i to be 1. In case that the number p of covariates varies with n, this should be interpreted as a triangular array.

Throughout, the term "model" refers to the subset of covariates present in the regression. There will be no assumption that any linear model is true for any choice of covariates. In order to describe models as subsets of covariates, we use nonempty index sets $M \subseteq \{1, 2, \ldots, p\}$ as in Section 2.1 and write $X_i(M)$ for the covariate vectors in the submodel M. For any $1 \le k \le p$, define the set of all nonempty models of size no larger than k by

$$\mathcal{M}_p(k) := \big\{M : M \subseteq \{1, 2, \ldots, p\}, 1 \le |M| \le k\big\},$$

so that $\mathcal{M}_p(p)$ is the power set of $\{1, 2, \ldots, p\}$ excluding the empty set.

To proceed further, we assume that the observations are *independent but possibly nonidentically distributed*. This assumption includes as special cases the settings of (i) independent

and identically distributed observations and (ii) of fixed (nonrandom) covariates (by defining the distribution of X_i to be a point mass at the observed X_i). Our setting is more general in that it allows some covariates to be fixed and others to be random.

For any model $M \subseteq \{1, 2, \ldots, p\}$ and $\theta \in \mathbb{R}^{|M|}$, define the OLS empirical and expected risk functions, respectively, as

$$\hat{R}_n(\theta; M) := \frac{1}{n} \sum_{i=1}^{n} \{Y_i - X_i^\top(M)\theta\}^2, \tag{2}$$

$$R_n(\theta; M) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[\{Y_i - X_i^\top(M)\theta\}^2\big]. \tag{3}$$

The notation $\mathbb{E}$ and $\mathbb{P}$ refer to expectations and probabilities with respect to all the randomness involved. Define the OLS estimator and the corresponding target for model M as

$$\hat{\beta}_{n,M} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} \hat{R}_n(\theta; M), \quad \text{and} \quad \beta_{n,M} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} R_n(\theta; M). \tag{4}$$

Thus $\hat{\beta}_{n,M}, \beta_{n,M} \in \mathbb{R}^{|M|}$. Note, however, that for $M \subsetneqq M'$ the estimate vector $\hat{\beta}_{n,M}$ is *not* a subvector of $\hat{\beta}_{n,M'}$ and the target vector $\beta_{n,M}$ is *not* a subvector of $\beta_{n,M'}$. The reason is nonorthogonality (partial collinearity) between covariates $j \in M$, $j' \in M' \setminus M$ in the sense that generally $\sum_{1 \le i \le n} X_i(j)X_i(j') \neq 0$ in the case of estimates and $\sum_{1 \le i \le n} \mathbb{E}[X_i(j)X_i(j')] \neq 0$ in the case of targets. This is why we must write M as a subscript of $\hat{\beta}_{n,M}$ and $\beta_{n,M}$ and not in parentheses. (See Section 3.1 of Berk et al. (2013) for a related discussion.)

Note also that the target $\beta_{n,M}$ is "dynamic," that is, it is permitted to change with the number of observations n. This is a consequence of allowing nonidentically distributed observations. In a framework of i.i.d. observations, the target $\beta_{n,M}$ would not depend on n.

Next, define associated second order matrices and vectors *in the full model* as follows:

$$\begin{aligned} \hat{\Sigma}_n &:= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \hat{\Gamma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \in \mathbb{R}^p, \\ \Sigma_n &:= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[X_i X_i^\top\big] \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \Gamma_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i Y_i] \in \mathbb{R}^p. \end{aligned} \tag{5}$$

Importantly, for these quantities there is no need to define separate versions in submodels M because they are just the submatrices $\hat{\Sigma}_n(M)$ and $\Sigma_n(M)$ of $\hat{\Sigma}_n$ and Σ_n, respectively, and subvectors $\hat{\Gamma}_n(M)$ and $\Gamma_n(M)$ of $\hat{\Gamma}_n$ and Γ_n, respectively. The OLS estimate of the slope vector and its target in the submodel M satisfy the following normal equations:

$$\hat{\Sigma}_n(M)\hat{\beta}_{n,M} = \hat{\Gamma}_n(M) \quad \text{and} \quad \Sigma_n(M)\beta_{n,M} = \Gamma_n(M). \tag{6}$$

REMARK 2.1. We do not solve the equations (6) on purpose because the confidence regions to be constructed below will accommodate exact collinearity by including subspaces of degeneracy. Minimizers of the objective functions $\hat{R}_n(\theta; M)$ and $R_n(\theta; M)$ defined in (2) and (3) always exist, even if they are not unique. Estimates $\hat{\beta}_{n,M}$ can only be unique when $|M| \le n$ because $\hat{\Sigma}_n(M)$ has rank at most $\min\{|M|, n\}$. Targets $\beta_{n,M}$, on the other hand, can be unique without a constraint on n because they are based on expectations rather than finite averages, so Σ_n and $\Sigma_n(M)$ can be strictly positive definite and $R_n(\theta; M)$ strictly convex with a unique minimizer even when $|M| > n$.

2.3. *Problem formulation.* Under very mild assumptions, $\hat{\beta}_{n,M} - \beta_{n,M}$ converges to zero as n tends to infinity for any fixed, nonrandom model M (see Kuchibhotla, Brown and Buja (2018)). This fact justifies calling $\hat{\beta}_{n,M}$ an estimator of $\beta_{n,M}$ or, equivalently, $\beta_{n,M}$ the target of estimation of $\hat{\beta}_{n,M}$. Also, for a fixed M, $\hat{\beta}_{n,M}$ has an asymptotic normal distribution, that is,

$$n^{1/2}(\hat{\beta}_{n,M} - \beta_{n,M}) \xrightarrow{\mathcal{L}} N(0, AV_M) \quad \big(0 \in \mathbb{R}^{|M|}, AV_M \in \mathbb{R}^{|M|\times|M|}\big),$$

for some positive definite matrix AV_M that depends on M and some moments of (X, Y); see the linear representation in Kuchibhotla, Brown and Buja (2018). The notation $\xrightarrow{\mathcal{L}}$ denotes convergence in law/distribution. Asymptotic normality lends itself for the construction of $(1-\alpha)$-confidence region $\hat{\mathcal{R}}_{n,M}$ such that

$$\liminf_{n\to\infty} \mathbb{P}(\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}) \geq 1 - \alpha$$

for any fixed $\alpha \in [0, 1]$. We approach statistical inference using confidence regions rather than statistical tests, but this is a technical rather than a conceptual choice due to the duality between confidence regions and tests: a confidence region with coverage at least $1-\alpha$ is a set of parameter values that could not be rejected at level α if used as point null hypotheses.

The problem of valid post model selection inference is to construct for given nonrandom sets of models $\mathcal{M}_p$ a set of confidence regions $\{\hat{\mathcal{R}}_{n,M} : M \in \mathcal{M}_p\}$ such that for any *random model* $\hat{M}$ depending (usually) on the same data satisfying $\mathbb{P}(\hat{M} \in \mathcal{M}_p) = 1$, we have

$$\liminf_{n\to\infty} \mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}) \geq 1 - \alpha. \tag{7}$$

The guarantee (7) is asymptotic because we strive for a theory that requires few assumptions, whereas finite sample confidence guarantees require strong assumptions.

The notation $\hat{M}$ for random models requires an elaboration of the sources of randomness envisioned here. With the reproducibility crisis in mind, we expand our view of the sources of model randomness by adopting a broad frequentist perspective that includes not only datasets but data analysts as well. Conventional frequentism can be conceived as capturing the random nature of an observed dataset in the actual world by embedding it in a universe of possible worlds with datasets characterized by a joint probability distribution of the observations. The broader frequentism proposed here is conceived as pairing the random datasets with random data analysts who have varying data analytic preferences and backgrounds. This variability among data analysts' preferences may be called "random researcher degrees of freedom," a term that refers to liberties exercised by practitioners when analyzing data in general, and when selecting regression covariates in particular. Some of the latter freedoms have been described and classified by Berk et al. ((2013), Section 1): (1) formal selection methods such as stepwise forward or backward selection, lasso-based selection using a criterion to select the penalty parameter, or all-subset search using a criterion such as C_p, AIC, BIC, RIC, etc.; (2) informal selection steps such as examination of residual plots or influence measures to judge acceptability of models; (3) post hoc selection such as making substantive trade-offs of predictive viability versus cost of data collection. The waters get further muddied even in the case of formal selection methods (1) when "informal meta-selection" is exercised: trying out multiple formal selection methods, comparing them and favoring some over others based on the results produced on the data at hand. This list of "researcher degrees of freedom" in model selection should make it evident that these freedoms are indeed exercised in practice, and that they are based on personal background, experience and motivations, as well as historic and institutional contexts. For these reasons, it may be impossible to capture in a stochastic model the randomness contributed by data analysts' exercise of their freedoms.

Following Berk et al. (2013), this infeasibility can be bypassed by adding a quantifier "for all $\hat{M}$" to the requirement (7), thereby capturing all possible ways in which selection may be performed. The benefit reaped from this step is that the requirement (7) permits a reduction to a problem of simultaneous inference. Simultaneity according to Berk et al. (2013) is over all regression slopes in all submodels, resulting in a NP-hard search problem. We will approach the simultaneity problem in a different way that lends itself to relatively inexpensive computations. A cost incurred by asserting ignorance about the selection process $\hat{M}$ is that conditioning on the selected model $\hat{M} = M$ is not possible, hence the approach of Lee et al. (2016), Tibshirani et al. (2016), Tian, Bi and Taylor (2016) and Fithian, Sun and Taylor (2014) is infeasible.

Inference, if conceived with a "for all $\hat{M}$" quantifier, requires certain limits on the freedom of model selection. The set of potential covariates must be prespecified before examining the data. For example, it is not permissible to initially declare the covariates $X_i(1), \ldots, X_i(p)$ to be the universe for searching submodels, only to decide after looking at the data to search among product interactions $X_i(j)X_i(k)$ as well. The decision to include interactions in data-driven selection must be made before looking at the data. Thus data-driven expansion of the universe of covariates for selection is not currently covered by our framework.

Again following Berk et al. (2013), a curious aspect of the target of estimation has to be noted: $\beta_{n,\hat{M}}$ has become a random quantity with a random dimension $|\hat{M}|$, whereas for a fixed M the target $\beta_{n,M}$ is a constant. After data-driven modeling, the selected target $\beta_{n,\hat{M}}$ has become random due to data-driven selection $\hat{M}$. This, however, is the only randomness present: among all possible targets $\{\beta_{n,M} : M \in \mathcal{M}_p\}$, one is randomly selected, namely, $\beta_{n,\hat{M}}$. The associated estimate $\hat{\beta}_{n,\hat{M}}$ in the random model $\hat{M}$, in addition to its intrinsic variability, also incurs the randomness due to selection.

3. Equivalence of post-selection and simultaneous inference. The first step toward achieving the goal of constructing a set of confidence regions $\{\hat{\mathcal{R}}_{n,M} : M \in \mathcal{M}_p\}$ satisfying (7) is to convert the post-selection inference problem into a simultaneous inference problem. This conversion is provided by Theorem 3.1, which parallels Berk et al. (2013) but offers the generality needed here. The theorem is proved for a finite number n of observations, but a version using "$\liminf_{n\to\infty}$" follows readily.

THEOREM 3.1. *For any set of confidence regions* $\{\hat{\mathcal{R}}_{n,M} : M \in \mathcal{M}_p\}$ *and* $\alpha \in [0, 1]$, *the following two statements are equivalent*:

(1) *The post-selection inference problem is solved, that is,*

$$\mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}) \geq 1 - \alpha$$

for all data-dependent model selections $\hat{M}$ *satisfying* $\hat{M} \in \mathcal{M}_p$.

(2) *The simultaneous inference problem over* $M \in \mathcal{M}_p$ *is solved, that is,*

$$\mathbb{P}\Big(\bigcap_{M\in\mathcal{M}_p} \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}\}\Big) \geq 1 - \alpha.$$

PROOF. For fixed $M \in \mathcal{M}_p$, let $\mathcal{A}_M = \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}\}$ be one coverage event inside (2), and similarly for random $\hat{M}$ let $\mathcal{A}_{\hat{M}} = \{\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}\}$ be the coverage event in (1). Note that both are random events due to the randomness of $\hat{\mathcal{R}}_{n,M}$.

(2) $\Rightarrow$ (1): Trivially, $\mathcal{A}_{\hat{M}} \supseteq \bigcap_{M\in\mathcal{M}_p} \mathcal{A}_M$ because $\hat{M} \in \mathcal{M}_p$, implying (1).

(1) $\Rightarrow$ (2): To prove this implication, it is sufficient to construct a data-driven (hence random) selection procedure $\hat{M}$ that satisfies

$$\mathcal{A}_{\hat{M}} = \bigcap_{M \in \mathcal{M}_p} \mathcal{A}_M. \tag{8}$$

This is achieved by letting $\hat{M}$ be any selection procedure that satisfies

$$\hat{M} \in \arg\min_{M \in \mathcal{M}_p} \mathbb{1}\{\mathcal{A}_M\},$$

where $\mathbb{1}\{A\}$ denotes the indicator of event A. It follows that $\mathbb{1}\{\mathcal{A}_{\hat{M}}\} = \min_{M \in \mathcal{M}_p} \mathbb{1}\{\mathcal{A}_M\}$, which is equivalent to (8). This completes the proof of (1) $\Rightarrow$ (2). $\square$

REMARK 3.1. The proof makes no use of the regression context at all; it is merely about indexed random sets/events $\mathcal{A}_M$ and random selections $\hat{M}$ of the indexes M. The second part of the proof shows the existence of adversarial random selection procedures $\hat{M}$ that require simultaneous coverage over all M. To achieve this, $\hat{M}$ only has to pick any model M for which $\hat{\mathcal{R}}_{n,M}$ fails to cover $\beta_{n,M}$ if such M exist; else, if $\hat{\mathcal{R}}_{n,M}$ covers $\beta_{n,M}$ for all M, $\hat{M}$ can be any model M. This provides an existence proof for $\hat{M}$ that satisfies (8) but it does not provide an actionable selection procedure for real data analysts because it depends on the unknown true $\beta_{n,M}$.

REMARK 3.2. The theorem establishes the equivalence of family-wise simultaneous coverage and post-selection coverage allowing for arbitrary random (data-driven) selection. The argument, because it makes no use of the regression context, applies to any type of regression. It also applies universally to any type of confidence procedure $\hat{\mathcal{R}}_{n,M}$.

REMARK 3.3. The analog of Theorem 3.1 above in Berk et al. (2013) is their Lemma 4.1 ("Significant triviality bound"), and their adversarial selection method is "p-value hunting" described in their Section 4.9. This selection method, however, does not correspond to the worst-case selection methods of Theorem 3.1. The difference is that "p-value hunting" is worst-case under a global null hypothesis and not for coverage events of arbitrary unknown $\beta_{n,M}$. Under a global null hypothesis, "p-value hunting" becomes actionable and leads to worst case selection under this null. For the UPoSI procedure to be proposed below, we will give a test-based actionable version of worst-case selections in Section 6.3. At this point, we have not even specified a confidence procedure yet. Theorem 3.1 states the equivalence of post-selection and simultaneous inference in utmost generality for any confidence procedure, whereas the setup of Berk et al. (2013) provides the equivalence for their PoSI procedure only.

REMARK 3.4 (Inherent high-dimensionality). Returning to regression, note that in view of Theorem 3.1, valid post-selection inference is inherently a high-dimensional problem in the sense that the number of parameters subject to estimation and inference is large, indeed, often larger than the sample size. For illustration, consider a common regression setting where the number of covariates is $p = 10$ and the sample of size $n = 500$. Estimation and testing of the slopes in the full model seems unproblematic because there are 50 observations per parameter. Now, for the post-selection inference problem with all nonempty submodels, there are $2^p - 1 = 1023$ vector parameters of varying dimensions, adding up to a total of $p2^{p-1} = 5120$ parameters in the various submodels, exceeding the sample size $n = 500$ by a factor of ten, thus constituting an inference problem in the high-dimensional category.

Theorem 3.1 shows that in order to achieve post-selection inference that is asymptotically valid across all data-driven selection procedures $\hat{M}$, it is necessary and sufficient to construct a set of confidence regions $\hat{\mathcal{R}}_{n,M}$ such that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\bigcap_{M\in\mathcal{M}_p} \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}\} \Big) \geq 1-\alpha. \tag{9}$$

The solutions to the post-selection inference problem proposed here will satisfy (9).

4. The UPoSI approach to post-selection inference.

4.1. *Valid confidence regions.* Equipped with the required notation, we proceed to constructing confidence regions $\hat{\mathcal{R}}_{n,M}$ for linear regression. Define the estimation errors of $\hat{\Sigma}_n$ and $\hat{\Gamma}_n$ from (5) as follows:

$$\begin{aligned} \mathcal{D}_n^{\Sigma} &:= \|\hat{\Sigma}_n - \Sigma_n\|_\infty = \max_{M\in\mathcal{M}_p(2)} \big\|\hat{\Sigma}_n(M) - \Sigma_n(M)\big\|_\infty, \\ \mathcal{D}_n^{\Gamma} &:= \|\hat{\Gamma}_n - \Gamma_n\|_\infty = \max_{M\in\mathcal{M}_p(1)} \big\|\hat{\Gamma}_n(M) - \Gamma_n(M)\big\|_\infty. \end{aligned} \tag{10}$$

The equalities on the right-hand side are useful trivialities given here for later use: $\mathcal{M}_p(2)$ and $\mathcal{M}_p(1)$ are the sets of all models of sizes bounded by 2 and 1, respectively, where size 1 is sufficient for "max" to reach all elements of the Γ vectors, but size 2 is needed for "max" to reach all off-diagonal elements of the Σ matrices as well. Importantly, neither $\mathcal{D}_n^{\Sigma}$ nor $\mathcal{D}_n^{\Gamma}$ is a function of submodels M.

The quantities $\mathcal{D}_n^{\Sigma}$ and $\mathcal{D}_n^{\Gamma}$ are statistics whose quantiles will play an essential role in the construction of the confidence regions to be defined next. In each submodel $M \in \mathcal{M}_p(p)$, we will construct for the parameter vector $\beta_{n,M}$ two confidence regions: The first satisfies *finite sample* guarantees at the cost of lesser transparency, whereas the second satisfies *asymptotic* guarantees with the benefit of greater simplicity. The motivations for the particular forms of these regions will become clear in the course of the elementary proofs of the theorems to follow. With these remarks in mind, we define two types of "UPoSI" confidence regions:

$$\hat{\mathcal{R}}_{n,M} := \big\{\theta \in \mathbb{R}^{|M|} : \big\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \theta\}\big\|_\infty \leq C_n^{\Gamma}(\alpha) + C_n^{\Sigma}(\alpha)\|\theta\|_1\big\}, \tag{11}$$

$$\hat{\mathcal{R}}^{\dagger}_{n,M} := \big\{\theta \in \mathbb{R}^{|M|} : \big\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \theta\}\big\|_\infty \leq C_n^{\Gamma}(\alpha) + C_n^{\Sigma}(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}, \tag{12}$$

where $C_n^{\Gamma}(\alpha)$ and $C_n^{\Sigma}(\alpha)$ are bivariate joint upper α quantiles of $\mathcal{D}_n^{\Gamma}$ and $\mathcal{D}_n^{\Sigma}$ in (10):

$$\mathbb{P}\big(\mathcal{D}_n^{\Gamma} \leq C_n^{\Gamma}(\alpha) \text{ and } \mathcal{D}_n^{\Sigma} \leq C_n^{\Sigma}(\alpha)\big) \geq 1-\alpha. \tag{13}$$

REMARK 4.1 (Restriction of models for selection). The confidence regions defined in (11) and (12) do not take advantage of restricted model universes such as "sparse model selection" where $\hat{M} \in \mathcal{M}_p(k)$ searches only models of sizes up to k ($< p$). It might, however, be of practical interest to consider the post-selection inference problem when the set of models used in selection is indeed a strict subset of the set $\mathcal{M}_p(p)$ of all models. This can be accommodated with an obvious tweak whereby

$$\mathcal{D}_n^{\Gamma}(\mathcal{M}) := \sup_{M\in\mathcal{M}} \big\|\hat{\Gamma}_n(M) - \Gamma_n(M)\big\|_\infty \quad \text{and} \quad \mathcal{D}_n^{\Sigma}(\mathcal{M}) := \sup_{M\in\mathcal{M}} \big\|\hat{\Sigma}_n(M) - \Sigma_n(M)\big\|_\infty$$

become functions of the restricted model universe $\mathcal{M}$ ($\subsetneq \mathcal{M}_p(p)$). Note, however, that according to (10) we have $\mathcal{D}_n^{\Gamma}(\mathcal{M}) = \mathcal{D}_n^{\Gamma}$ as long as the model universe $\mathcal{M}$ includes all models of size one, and $\mathcal{D}_n^{\Sigma}(\mathcal{M}) = \mathcal{D}_n^{\Sigma}$ as long as $\mathcal{M}$ includes all models of size two. This is the case, for example, when "sparse model selection" is used, meaning $\mathcal{M} = \mathcal{M}_p(k)$ for $k < p$.

Thus confidence regions of the form (11) do not gain from "sparse model selection." This is so because the regions depend effectively only on marginal and bivariate properties of the observations (X_i, Y_i) and their distributions through Γ_n, $\hat{\Gamma}_n$, Σ_n and $\hat{\Sigma}_n$.

Further observations on $(\mathcal{D}_n^\Gamma, \mathcal{D}_n^\Sigma)$ and $(C_n^\Gamma(\alpha), C_n^\Sigma(\alpha))$:

- Bivariate quantiles are not unique: one may marginally increase one and decrease the other suitably, maintaining the bivariate coverage probability $1-\alpha$. Allowed is any choice of $C_n^\Gamma(\alpha)$ and $C_n^\Sigma(\alpha)$ that satisfies (13).
- These quantiles are not known and must be estimated from the data. A bootstrap procedure is described in Section 8.
- The estimation errors $\mathcal{D}_n^\Gamma$ and $\mathcal{D}_n^\Sigma$, being based on averages of quantities of dimensions $p \times 1$ and $p \times p$, respectively, converge by the law of large numbers to zero as $n \to \infty$ under mild conditions (see Lemma 5.1). Hence, $\max\{C_n^\Gamma(\alpha), C_n^\Sigma(\alpha)\} = o(1)$ as $n \to \infty$.

4.2. *Validity of the confidence regions $\hat{\mathcal{R}}_{n,M}$.* We next prove validity of the simultaneous inference guarantee (9). This will be done in Theorem 4.1 for the confidence regions $\hat{\mathcal{R}}_{n,M}$ where $M \in \mathcal{M}_p(p)$, and in Theorem 4.2 for the confidence regions $\hat{\mathcal{R}}_{n,M}^\dagger$ where $M \in \mathcal{M}_p(k)$ for some $k \le p$.

THEOREM 4.1. *The UPoSI confidence regions $\{\hat{\mathcal{R}}_{n,M} : M \in \mathcal{M}_p(p)\}$ defined in (11) satisfy*

$$\mathbb{P}\Big(\bigcap_{M \in \mathcal{M}_p(p)} \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}\}\Big) \ge 1-\alpha. \tag{14}$$

Furthermore, for any random model $\hat{M}$ with $\hat{M} \in \mathcal{M}_p(p)$, we have

$$\mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}) \ge 1-\alpha. \tag{15}$$

As mentioned earlier, this theorem is nonasymptotic as it provides guarantees for finite samples. It is not directly actionable because the bivariate quantiles used in the construction of the confidence regions need to be estimated. Hence actionable versions of these regions end up having only asymptotic guarantees as well. In addition, if the observations are not identically distributed, any bootstrap procedure used for estimation will not be able to account for the differences in the distributions, but it will be asymptotically conservative. See Liu and Singh (1995) and Section 8 below.

PROOF. The proof is surprisingly elementary, involving only simple manipulations of the estimating equations, and is free of stochastic assumptions. We start by differencing the normal equations of estimates and targets; see (6). This holds for all $M \in \mathcal{M}_p(p)$:

$$\hat{\Sigma}_n(M)\hat{\beta}_{n,M} - \Sigma_n(M)\beta_{n,M} = \hat{\Gamma}_n(M) - \Gamma_n(M).$$

Telescope the left-hand side by subtracting and adding $\hat{\Sigma}_n(M)\beta_{n,M}$:

$$\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M}) + \big(\hat{\Sigma}_n(M) - \Sigma_n(M)\big)\beta_{n,M} = \hat{\Gamma}_n(M) - \Gamma_n(M),$$

Move the second summand on the left to the right-hand side of the equality, take the sup norm and apply the triangle inequality on the right-hand side:

$$\big\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \beta_{n,M}\}\big\|_\infty \le \big\|\hat{\Gamma}_n(M) - \Gamma_n(M)\big\|_\infty + \big\|\{\hat{\Sigma}_n(M) - \Sigma_n(M)\}\beta_{n,M}\big\|_\infty.$$

Applying the second inequality in (1) to the last term it follows that

$$\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \beta_{n,M}\}\|_\infty \le \|\hat{\Gamma}_n(M) - \Gamma_n(M)\|_\infty + \|\hat{\Sigma}_n(M) - \Sigma_n(M)\|_\infty \|\beta_{n,M}\|_1.$$

Because $\hat{\Gamma}_n(M) - \Gamma_n(M)$ and $\hat{\Sigma}_n(M) - \Sigma_n(M)$ are just a subvector and a submatrix of $\hat{\Gamma}_n - \Gamma_n$ and $\hat{\Sigma}_n - \Sigma_n$, respectively, we get

$$\|\hat{\Sigma}_n(M)\{\beta_{n,M} - \hat{\beta}_{n,M}\}\|_\infty \le \|\hat{\Gamma}_n - \Gamma_n\|_\infty + \|\hat{\Sigma}_n - \Sigma_n\|_\infty \|\beta_{n,M}\|_1. \tag{16}$$

This inequality is deterministic and holds for any sample and for all $M \in \mathcal{M}_p(p)$. These facts allow us to take the intersection of the events (16) over all submodels M and transform it into a "probability one" statement. Using $\mathcal{D}_n^\Gamma$ and $\mathcal{D}_n^\Sigma$ defined in (10), we have

$$\mathbb{P}\Big(\bigcap_{M \in \mathcal{M}_p(p)} \{\|\Sigma_n(M)\{\beta_{n,M} - \hat{\beta}_{n,M}\}\|_\infty \le \mathcal{D}_n^\Gamma + \mathcal{D}_n^\Sigma \|\beta_{n,M}\|_1\}\Big) = 1. \tag{17}$$

From the definitions of $C_n^\Gamma(\alpha)$ and $C_n^\Sigma(\alpha)$ in (13) follows the required result (14). The second result (15) for random models follows by an application of Theorem 3.1. □

REMARK 4.2 (Validity guarantee for large p). The guarantee (14) in Theorem 4.1 is valid for every sample size n and any number of covariates p. In particular, $p \gg n$ and $p = \infty$ are covered by the theorem even though $\hat{\Sigma}_n(M)$ is necessarily singular for $|M| > n$. In this case, the confidence region $\hat{\mathcal{R}}_{n,M}$ simply contains a nontrivial affine subspace of $\mathbb{R}^p$.

REMARK 4.3 (Estimation of bivariate quantiles). The finite sample guarantee (14) requires the bivariate quantiles $C_n^\Gamma(\alpha)$ and $C_n^\Sigma(\alpha)$ of $\mathcal{D}_n^\Gamma$ and $\mathcal{D}_n^\Sigma$, respectively, to satisfy (13) for all $p, n \ge 1$. In general, these bivariate quantiles can only be estimated consistently in the asymptotic sense as explained in Section 8.

REMARK 4.4 (Independence not used). Earlier we assumed for concreteness independent observations (X_i, Y_i), $1 \le i \le n$. Theorem 4.1, however, holds without this assumption as the proof made no use of independence. Validity of the post-selection guarantee holds as long as $C_n^\Gamma(\alpha)$ and $C_n^\Sigma(\alpha)$ are valid bivariate quantiles in the sense of (13).

4.3. *Asymptotic validity of the confidence regions* $\hat{\mathcal{R}}_{n,M}^\dagger$. The confidence region $\hat{\mathcal{R}}_{n,M}$ is difficult to analyze in terms of its shape and its Lebesgue measure. Because of these difficulties, we also prove asymptotic validity of the more intuitive confidence regions of the form $\hat{\mathcal{R}}_{n,M}^\dagger$ defined in (12). As these regions depend on estimates $\hat{\beta}_{n,M}$ whose variability explodes with increasing collinearity, we need to control the minimum eigenvalue of $\Sigma_n(M)$ for models up to size k. We therefore define

$$\Lambda_n(k) := \min_{M \in \mathcal{M}_p(k)} \lambda_{\min}\big(\Sigma_n(M)\big),$$

and make use of the following assumption:

(**A1**)(k) The estimation error $\mathcal{D}_n^\Sigma$ satisfies $k\mathcal{D}_n^\Sigma = o_p(\Lambda_n(k))$ as $n \to \infty$.

This assumption is used for uniform consistency of the least squares estimator in $\|\cdot\|_1$-norm as in Lemma 4.1 below. The rate of convergence of $\mathcal{D}_n^\Sigma$ to zero implies a rate constraint on the maximal submodel size k. Here, as before, $k = k_n$ is allowed to be a sequence depending on n. As can be expected, the dependence structure between the observations (X_i, Y_i), $1 \le i \le n$ and their moments determine the rate at which $\mathcal{D}_n^\Sigma$ converges to zero. Under certain tail assumptions on the observations as well as independence, k can grow at the rate $o(\sqrt{n/\log p})$

if $\Lambda_n(k)$ is bounded away from zero; see Lemma 5.1 for more details. The theorem is stated with this high-level assumption so that it is more widely applicable, in particular to various structural dependencies on the observations; see Theorem S.11.1 for a specific result.

Assumption (**A1**)(k) allows exact collinearity of the full set of covariates but prohibits exact collinearity of any k-subset of covariates. In particular, the minimum eigenvalue of Σ_n can converge to zero or even be zero as $n \to \infty$ if $p = p_n$ changes with n.

Before stating a theorem about asymptotic validity of the post-selection confidence regions $\hat{\mathcal{R}}^{\dagger}_{n,M}$, we give conditions for uniform-in-model consistency of $\hat{\beta}_{n,M}$ to $\beta_{n,M}$. See Section S.5 for a proof, and also Kuchibhotla et al. (2018) for more results of this flavor.

LEMMA 4.1. *For all $k \geq 1$ satisfying $k\mathcal{D}_n^{\Sigma} \leq \Lambda_n(k)$ and for all $M \in \mathcal{M}_p(k)$,*

$$\|\hat{\beta}_{n,M} - \beta_{n,M}\|_1 \leq \frac{|M|(\mathcal{D}_n^{\Gamma} + \mathcal{D}_n^{\Sigma}\|\beta_{n,M}\|_1)}{\Lambda_n(k) - k\mathcal{D}_n^{\Sigma}}. \tag{18}$$

The following theorem states the validity of the simultaneous inference guarantee for $\hat{\mathcal{R}}^{\dagger}_{n,M}$.

THEOREM 4.2. *For every $1 \leq k \leq p$ that satisfies* (**A1**)(k), *the confidence regions $\hat{\mathcal{R}}^{\dagger}_{n,M}$ defined in* (12) *satisfy*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\bigcap_{M\in\mathcal{M}_p(k)} \{\beta_{n,M} \in \hat{\mathcal{R}}^{\dagger}_{n,M}\}\Big) \geq 1 - \alpha.$$

See Section S.6 of the supplement for a proof.

4.4. *Further remarks on the confidence regions $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}^{\dagger}_{n,M}$.*

1. *Standardizing covariates:* The confidence regions $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}^{\dagger}_{n,M}$ are not equivariant with respect to linear transformation of covariates or the response. A simple way to obtain equivariance with respect to diagonal linear transformations (changes of units) is to standardize the covariates to have sample mean 0 and sample variance 1. Because the validity of the confidence regions does not require independence (Remark 4.4), data-driven standardization will not affect the post-selection guarantee as long as marginal means and variances are estimated consistently. This may affect the volume of the confidence regions, not in terms of rate but in terms of constants, because the intercept is no longer needed in $\|\beta_{n,M}\|_1$; see Section 10 for more details. (Note that traditional fixed covariate theory of linear models achieves equivariance under unit changes through the use of t-statistics which are free of units.)

2. *Comparison of $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}^{\dagger}_{n,M}$ in testing:* As mentioned earlier, the shape of the confidence region $\hat{R}_{n,M}$ is not easily described. However, $\hat{\mathcal{R}}_{n,M}$ has advantages over $\hat{\mathcal{R}}^{\dagger}_{n,M}$ for significance testing of null hypotheses $H_{0,M} : \beta_{n,M} = 0$. The level α test based on $\hat{\mathcal{R}}_{n,M}$ rejects $H_{0,M}$ if

$$\|\hat{\Sigma}_n(M)\hat{\beta}_{n,M}\|_{\infty} \geq C_n^{\Gamma}(\alpha).$$

By comparison, the level α test based on the confidence region $\hat{\mathcal{R}}^{\dagger}_{n,M}$ rejects $H_{0,M}$ if

$$\|\hat{\Sigma}_n(M)\hat{\beta}_{n,M}\|_{\infty} \geq C_n^{\Gamma}(\alpha) + C_n^{\Sigma}(\alpha)\|\hat{\beta}_{n,M}\|_1.$$

Thus $\hat{\mathcal{R}}_{n,M}$ results in more rejections and hence greater power than $\hat{\mathcal{R}}^{\dagger}_{n,M}$ at the same level α. A similar argument holds even if the null hypothesis is changed to $H_{0,M} : \beta_{n,M} = \theta_0 \in \mathbb{R}^{|M|}$ for some sparse θ_0, for example.

3. *Shape of the confidence regions:* The region $\hat{\mathcal{R}}_{n,M}$ has a complex shape that is not easy to describe. The confidence region $\hat{\mathcal{R}}^{\dagger}_{n,M}$ is a parallelepiped because it can be described by $2|M|$ linear inequalities that come in pairs of parallel constraints (with random coefficients). The Lebesgue measure of this confidence region is easier to study than that of the region $\hat{\mathcal{R}}_{n,M}$ (see Proposition 5.1 below).

4. *Implied "prediction" regions:* Following Berk et al. (2013), Bachoc, Leeb and Pötscher (2019) considered the problem of post-selection prediction regions. Mathematically, the problem is to cover $x_0(\hat{M})^{\top}\beta_{n,\hat{M}}$ for a randomly selected model $\hat{M}$ and a given fixed vector $x_0 \in \mathbb{R}^p$. Bachoc, Leeb and Pötscher (2019) consider different versions of the problem depending on whether one observes x_0 or only $x_0(\hat{M})$. These issues do not arise here because our UPoSI confidence regions are not based on a maximization over different models. Based on the confidence regions $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}^{\dagger}_{n,M}$, we may construct post-selection regions for the best linear prediction surface. The functional $\beta_{n,\hat{M}}$ (as defined in (4)) provides the best linear prediction $x^{\top}(M)\beta_{n,M}$ to the response based on the covariates $X(M) = x(M)$. For any random model $\hat{M}$, we have $\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}}$ with probability at least $1-\alpha$. Hence, for *any* set $\mathcal{X} \subseteq \mathbb{R}^p$,

$$\mathbb{P}\big(x^{\top}(\hat{M})\beta_{n,\hat{M}} \in \{x^{\top}(\hat{M})\theta : \theta \in \hat{\mathcal{R}}_{n,\hat{M}}\} \text{ for all } x \in \mathcal{X}\big) \geq 1-\alpha.$$

Similarly, for any random model $\hat{M} \in \mathcal{M}_p(k)$, we get

$$\liminf_{n\to\infty} \mathbb{P}\big(x^{\top}(\hat{M})\beta_{n,\hat{M}} \in \{x^{\top}(\hat{M})\theta : \theta \in \hat{\mathcal{R}}^{\dagger}_{n,\hat{M}}\} \text{ for all } x \in \mathcal{X}\big) \geq 1-\alpha.$$

5. Rate bounds on $\mathcal{D}_n^{\Gamma}$, $\mathcal{D}_n^{\Sigma}$ and Lebesgue measure of the regions. Before proceeding further with the study of the confidence regions, we examine the rates at which $\mathcal{D}_n^{\Gamma}$ and $\mathcal{D}_n^{\Sigma}$ converge to zero under assumptions on the tail behavior as well as independence. As mentioned in Remark 4.4, the validity of the post-selection coverage guarantee does not require independence, hence rate results under "functional dependence" are possible also (see Section S.11 of the supplement). Let $Z_i = (X_i^{\top}, Y_i)^{\top}$ for $1 \leq i \leq n$ and define

$$\hat{\Omega}_n := \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^{\top}, \quad \text{and} \quad \Omega_n := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[Z_i Z_i^{\top}\big] \in \mathbb{R}^{(p+1)\times(p+1)}.$$

Observe that $\max\{\mathcal{D}_n^{\Gamma}, \mathcal{D}_n^{\Sigma}\} \leq \|\hat{\Omega}_n - \Omega_n\|_{\infty}$. The following lemma from Kuchibhotla and Chakrabortty (2018) proves a finite sample bound for the expected value of the maximum absolute value of $\hat{\Omega}_n - \Omega_n$. For this result, set for $\gamma > 0$ and any random variable W,

$$\|W\|_{\psi_\gamma} := \inf\big\{C > 0 : \mathbb{E}\big[\psi_\gamma\big(|W|/C\big)\big] \leq 1\big\}, \tag{19}$$

where $\psi_\gamma(x) := \exp(x^\gamma) - 1$ for $x \geq 0$. For $0 < \gamma < 1$, $\|\cdot\|_{\psi_\gamma}$ is not a norm but a quasi-norm. A random variable W satisfying $\|W\|_{\psi_\gamma} < \infty$ is called a sub-Weibull random variable of order γ. The special cases $\gamma = 1$ and $\gamma = 2$ correspond to the well-known classes of subexponential and sub-Gaussian random variables.

LEMMA 5.1. *Fix $n, p \geq 2$. Suppose the random vectors Z_i, $1 \leq i \leq n$ are independent and satisfy for some $0 < \gamma \leq 2$,*

$$\max_{1\leq i\leq n} \max_{1\leq j\leq p+1} \big\|Z_i(j)\big\|_{\psi_\gamma} \leq K_{n,p}, \tag{20}$$

for some positive constant $K_{n,p}$. Then

$$\mathbb{E}\big[\sqrt{n}\|\hat{\Omega}_n - \Omega_n\|_{\infty}\big] \leq C_\gamma\big\{A_{n,p}\sqrt{\log p} + K_{n,p}^2(\log p \log n)^{2/\gamma} n^{-1/2}\big\},$$

and for all $\alpha \in (0, 1]$,

$$\max\{C_n^{\Gamma}(\alpha), C_n^{\Sigma}(\alpha)\} \le 7A_{n,p}\sqrt{\frac{\log(\frac{3}{\alpha}) + 2\log p}{n}} + \frac{C_\gamma K_{n,p}^2 (\log(2n))^{2/\gamma}(\log(\frac{3}{\alpha}) + 2\log p)^{2/\gamma}}{n},$$

where C_γ *is a positive universal constant that grows at the rate of* $(1/\gamma)^{1/\gamma}$ *as* $\gamma \downarrow 0$ *and* $A_{n,p}^2 := \max_{1\le j\le k\le p+1} n^{-1}\sum_{i=1}^n \mathrm{Var}(Z_i(j)Z_i(k))$.

PROOF. See Theorem 4.1 of Kuchibhotla and Chakrabortty (2018). A similar result holds for $\gamma > 2$, that is, random variables with tails lighter than the Gaussian. See Theorem 3.4 of Kuchibhotla and Chakrabortty (2018) for a result along these lines. □

The confidence regions $\hat{\mathcal{R}}_{n,M}^{\dagger}$ are simple parallelepipeds and can be seen as linear transformations of $\|\cdot\|_\infty$-norm balls. Hence, their Lebesgue measures can be computed exactly. Since the confidence regions are valid over a large number of models, we give a relative Lebesgue measure result uniform over a set of models. For a measurable set $A \subseteq \mathbb{R}^q$ with $q \ge 1$, let $\mathbf{Leb}(A)$ denote the Lebesgue measure of A in $\mathbb{R}^q$. For convenience, we do not use different notations for the Lebesgue measure for different $q \ge 1$.

PROPOSITION 5.1. *For any* $k \ge 1$ *such that assumption* (**A1**)(k) *are satisfied, the uniform relative Lebesgue measure result holds*:

$$\sup_{M\in\mathcal{M}_p(k)} \frac{\mathbf{Leb}(\hat{\mathcal{R}}_{n,M}^{\dagger})\Lambda_n^{|M|}(k)}{(2C_n^{\Gamma}(\alpha) + 2C_n^{\Sigma}(\alpha)\|\beta_{n,M}\|_1)^{|M|}} = O_p(1).$$

Moreover, if $\Lambda_n^{-1}(k) = O(1)$, *then, in the setting of Lemma* 5.1,

$$\mathbf{Leb}(\hat{\mathcal{R}}_{n,M}^{\dagger}) = O_p\left(\sqrt{\frac{|M|\log p}{n}}\right)^{|M|} \quad \textit{uniformly for } M \in \mathcal{M}_p(k), \tag{21}$$

if p *and* n *satisfy* $(\log p)^{2/\alpha}(\log n)^{2/\alpha - 1/2} = o(n^{1/2})$.

PROOF. See Section S.7 of the supplement for a detailed proof. □

REMARK 5.1 (The question of optimality of rates). Even though the problem of post-selection inference is studied from various perspectives as discussed in Section 2.3, we do not know of a result regarding the optimal size of confidence regions in the post-selection problem. The volume rates derived in Proposition 5.1, however, seem sharp and are better than other existing post-selection confidence regions. A comparison of volumes with the PoSI regions of Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016) is in Section 6 for fixed covariates.

For intuition, consider simultaneous confidence intervals for the mean of a multivariate normal random vector in $\mathbb{R}^q$ with i.i.d. $N(0, 1)$ coordinates: The cube formed from them has volume of order $(\sqrt{\log q/n})^q$ (in the worst case). Thus inference for q parameters that can be estimated by asymptotically normal estimators would have worst case volume of order $(\sqrt{\log q/n})^q$. Toward the sharpness of our volume rates in Proposition 5.1, we note that if an "oracle" informs us that the selected model $\hat{M}$ will have cardinality s, then for valid post-selection inference we need inference for $\binom{p}{s}$ parameters, $\beta_{n,M}$, $|M| = s$. Because these are

estimable by $\hat{\beta}_{n,M}$ which are asymptotically normal, the worst case confidence regions would have volume of order $(\sqrt{s\log(ep/s)/n})^s$. By comparison, our proposition leads to volumes of order $(\sqrt{s\log p/n})^s$ (for model $\hat{M}$ with $|\hat{M}| = s$) even without knowledge of an oracle. These arguments are still mostly heuristic, and even the meaning of optimality in this context is not entirely clear. We hope to produce a rigorous optimality framework in the future.

6. Confidence regions for fixed covariates.

6.1. *Simplifications for fixed covariates.* Because most of the post-selection inference literature as reviewed in Section 2.1 assumes fixed covariates, it is of particular interest to understand how the UPoSI confidence regions behave in this case. We can interpret fixed covariates as having point mass distributions at the observed value X_i, hence the two second moment matrices for the covariates coincide:

$$\Sigma_n = n^{-1}\sum_{i=1}^n \mathbb{E}\big[X_i X_i^\top\big] = n^{-1}\sum_{i=1}^n X_i X_i^\top = \hat{\Sigma}_n.$$

It follows that $\mathcal{D}_n^\Sigma = \|\hat{\Sigma}_n - \Sigma_n\|_\infty = 0$, hence its quantiles vanish, $C_n^\Sigma(\alpha) = 0$ for all $\alpha \in [0,1]$. Furthermore, the target specializes to

$$\beta_{n,M} = \left(\frac{1}{n}\sum_{i=1}^n X_i(M)X_i^\top(M)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i(M)\mathbb{E}[Y_i]\right),$$

and the two types of UPoSI confidence regions (11) and (12) are identical:

$$\hat{\mathcal{R}}_{n,M} = \hat{\mathcal{R}}_{n,M}^\dagger = \big\{\theta \in \mathbb{R}^{|M|} : \big\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \theta\}\big\|_\infty \le C_n^\Gamma(\alpha)\big\}.$$

Finally, for fixed covariates the assumption (**A1**)(k), which was needed for $\hat{\mathcal{R}}_{n,M}^\dagger$, is trivially satisfied because $\mathcal{D}_n^\Sigma = 0$. Thus, by Theorem 4.1 (or 4.2), finite sample valid post-selection inference holds for all model sizes under no model or distributional assumptions, as were still needed in Berk et al. (2013).

6.2. *Worst-case selection for UPoSI with fixed covariates.* We showed in Section 3 that for all confidence regions there exist worst-case or adversarial selection procedures $\hat{M}$ that require simultaneous control of coverage for all submodels M that $\hat{M}$ is permitted to search. We will now characterize such worst-case selection for the UPoSI confidence regions $\hat{\mathcal{R}}_{n,M}$ in the fixed covariate setting. We start by noting

$$\begin{aligned}\big\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \beta_{n,M}\}\big\|_\infty &= \left\|\frac{1}{n}\sum_{i=1}^n X_i(M)\big(Y_i - \mathbb{E}[Y_i]\big)\right\|_\infty \\ &= \max_{j\in M}\left|\frac{1}{n}\sum_{i=1}^n X_i(j)\big(Y_i - \mathbb{E}[Y_i]\big)\right|.\end{aligned}$$

Choose a single data-dependent (hence random) coordinate $\hat{J} \in \{1,2,\dots,p\}$ such that

$$\hat{J} \in \operatorname*{arg\,max}_{1\le j\le p}\left|n^{-1}\sum_{i=1}^n X_i(j)\big(Y_i - \mathbb{E}[Y_i]\big)\right|.$$

It follows that for any model M that contains $\hat{J}$ we have

$$\big\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \beta_{n,M}\}\big\|_\infty = \max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^n X_i(j)\big(Y_i - \mathbb{E}[Y_i]\big)\right| = \mathcal{D}_n^\Gamma.$$

Hence, for any random model $\hat{M}$ for which $\mathbb{P}(\hat{J} \in \hat{M}) = 1$, the coverage of $\hat{\mathcal{R}}_{n,\hat{M}}$ is exactly equal to $(1-\alpha)$.

An interesting consequence is the following: The number of models that contain $\hat{J}$ is 2^{p-1}, or half of all models. This fact suggests intuitively that the UPoSI confidence regions will often be sharp, that is, not conservative.

6.3. *Comparison of UPoSI worst-case selection with PoSI of Berk et al. (2013).* In Remark 3.3, we noted that the worst-case selection procedure for PoSI described in Berk et al. (2013) as "p-value hunting" is not directly comparable with the worst-case selection procedures designed here for confidence procedures. The reason is that "p-value hunting" derives from testing null hypotheses $\beta_{n,M}(j) = 0$. To achieve comparability of worst-case procedures between UPoSI and PoSI, we need to turn the UPoSI confidence regions into tests of the null hypotheses $\beta_{n,M} = 0\ (\in \mathbb{R}^{|M|})$. This is done by considering the events that the UPoSI confidence regions $\hat{\mathcal{R}}_{n,M}$ cover $\beta_{n,M} = 0$ assuming that this is the truth. The test statistic implied by the region $\hat{\mathcal{R}}_{n,M}$ is $\|\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \beta_{n,M}\}\|_\infty$, which for $\beta_{n,M} = 0$ simplifies to $\|\hat{\Sigma}_n(M)\hat{\beta}_{n,M}\|_\infty = \|\hat{\Gamma}_n(M)\|_\infty$. Under simultaneity for all M, the overall test statistic becomes $\|\hat{\Gamma}_n\|_\infty = \max_{1\le j\le p} |\frac{1}{n}\sum_{1\le i\le n} X_i(j)Y_i|$. Up to an irrelevant factor, this amounts to a maximum over individual test statistics of the form $|\sum_i X_i(j)Y_i|$.

If the covariates are standardized to remove units and place them on comparable scales for scale equivariance, the test statistics $\hat{\Gamma}_n(j) = \sum_i X_i(j)Y_i$ essentially become t-statistics for the univariate regressions of the response on each covariate separately. A worst-case selection is therefore any model M that has a nonempty intersection with $\arg\max_{1\le j\le p} |\sum_i X_i(j)Y_i|$. Thus, in the fixed covariate setting and for testing null hypotheses $\beta_{n,M} = 0$, the worst-case selection procedure for UPoSI turns out to be "significance hunting" as well, but over only p marginal t-tests, compared to PoSI of Berk et al. (2013) whose worst-case selection hunts over $p2^{p-1}$ t-tests.

If UPoSI is related to marginal screening and simple regressions, one may ask how it can be the basis for simultaneous inference in multiple regression and for testing composite null hypotheses such as $\beta_{n,M} = 0\ (\in \mathbb{R}^{|M|})$. The reason is that $\sum_i X_i(j)\mathbb{E}[Y_i] = 0$ for $j \in M$ and $\Sigma_n(M)$ nonsingular entails $\beta_{n,M} = 0$.

6.4. *A comparison of Lebesgue measures of UPoSI regions with Berk et al. (2013).* The rate bound (21) for $\mathbf{Leb}(\hat{\mathcal{R}}^\dagger_{n,M})$ in Proposition 5.1 is written explicitly for general random covariates. For fixed covariates, we showed that $C_n^\Sigma(\alpha) = 0$ and $\hat{\mathcal{R}}_{n,M} = \hat{\mathcal{R}}^\dagger_{n,M}$. From the proof of Proposition 5.1, we get

$$\mathbf{Leb}(\hat{\mathcal{R}}^\dagger_{n,M}) \le |\Sigma_n(M)|^{-1}\big(2C_n^\Gamma(\alpha)\big)^{|M|} \quad \text{for all } M \in \mathcal{M}_p(p).$$

Using the assumptions of Lemma 5.1, we further get

$$\mathbf{Leb}(\hat{\mathcal{R}}^\dagger_{n,M}) = O_p\big(|\Sigma_n(M)|^{-1}\big)\big(\sqrt{n^{-1}\log p}\big)^{|M|}. \tag{22}$$

This is much smaller than the size shown for general random covariates in (21) of Proposition 5.1. An explanation for this discrepancy between fixed and random covariates is as follows: The confidence regions $\hat{\mathcal{R}}_{n,M}$ (11) and $\hat{\mathcal{R}}^\dagger_{n,M}$ (12) are written in terms of $\hat{\Sigma}_n(M)\{\hat{\beta}_{n,M} - \beta_{n,M}\}$, but for fixed covariates we have $\hat{\Sigma}_n(M)\beta_{n,M} = \Sigma_n(M)\beta_{n,M} = \Gamma_n(M)$. While the confidence regions are written for $\beta_{n,M}$, they are equivalent to confidence regions for the population functionals $\Gamma_n(M)$. These are just subvectors of the p-dimensional vector Γ, for which one can construct a confidence region with length $\sqrt{\log p/n}$ on each coordinate, explaining the smaller size of (22). For random covariates, this reasoning does not

hold because $\hat{\Sigma}_n$ and Σ_n are not identical, injecting sampling variability caused by the random covariates.

The surprisingly small rate of (22) calls for a comparison with the Lebesgue volumes of the fixed covariate PoSI confidence regions of Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016). They are both based on the quantiles of the statistic

$$\max_{M\in\mathcal{M}_p(k)} \max_{j\in M} \bigl|\sqrt{n}\bigl(\hat{\beta}_{n,M}(j) - \beta_{n,M}(j)\bigr)/\sigma_{n,M}(j)\bigr|, \tag{23}$$

where the denominators are some form of standard errors or estimates thereof. (Their choice differs between these works; for simplicity, we assume them to be known.) Based on this "max-$|t|$" statistic (23), a confidence region for $\beta_{n,M}$ is

$$\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M} := \Bigl\{\theta\in\mathbb{R}^{|M|} : \max_{1\le j\le |M|} \bigl|\sqrt{n}\bigl(\hat{\beta}_{n,M}(j) - \theta(j)\bigr)/\sigma_{n,M}(j)\bigr| \le C_{n,k}(\alpha)\Bigr\}, \tag{24}$$

where $C_{n,k}(\alpha)$ is the upper α quantile of the max-$|t|$ statistic so as to achieve coverage $(1-\alpha)$. For fixed covariates and Gaussian response, $\sqrt{n}(\hat{\beta}_{n,M} - \beta_{n,M})$ is normally distributed, and from Berk et al. ((2013), Theorem 6.2) it is known that the max-$|t|$ statistic (23) can be of the order $\sqrt{k\log(ep/k)}$; see Bachoc, Blanchard and Neuvial (2018) for sharpness of this rate. This implies that $C_{n,k}(\alpha)$ can be of the order $\sqrt{k\log(ep/k)}$, and hence the Lebesgue measure of the confidence region $\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}$ satisfies

$$\mathbf{Leb}\bigl(\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}\bigr) = O_p(1)\biggl(\sqrt{\frac{k\log p}{n}}\biggr)^{|M|} \quad \text{uniformly over all } M\in\mathcal{M}_p(k). \tag{25}$$

This shows that the confidence region $\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}$ is less favorable than $\hat{\mathcal{R}}^{\dagger}_{n,M}$ $(= \hat{\mathcal{R}}_{n,M})$ in at least two aspects. First, the size of the confidence region has an additional factor $\sqrt{k}$ that makes the region huge in comparison. Second, the Lebesgue measure does not scale with model size $|M|$. For example, after searching over the set of models $\mathcal{M}_p(k)$, if the analyst settles on a (random) model of size 1, then the post-selection confidence region $\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}$ has a size that still scales with k. In sharp contrast, the UPoSI confidence region $\hat{\mathcal{R}}^{\dagger}_{n,M}$ (even for random covariates) has Lebesgue measure scaling only with the size $|M|$ of the selected model $M=\hat{M}$ and does not depend on the size k of the largest models in the search space $\mathcal{M}_p(k)$ that is accessible to a data-driven selection procedure $\hat{M}$.

6.5. *Fixed covariates with the Restricted Isometry Property (RIP).* The rate bound (25) above is derived using the fact that $C_{n,k}(\alpha)$ can in general be of order $\sqrt{k\log(ep/k)}$. For orthogonal designs ($\hat{\Sigma}_n = I_p$, the identity matrix in $\mathbb{R}^{p\times p}$), Berk et al. (2013) proved that $C_{n,k}(\alpha) = O(\sqrt{\log p})$, hence the size of the region $\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}$ matches that of our confidence region. Because the construction of Berk et al. (2013) is based on normality, the exact size of the confidence region $\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}$ could be better (i.e., smaller) than the size of $\hat{\mathcal{R}}^{\dagger}_{n,M}$. It is also interesting to note that, for orthogonal designs, $\hat{\mathcal{R}}^{\dagger}_{n,M}$ provides a rectangle with sides parallel to the coordinate axis, hence is of the same shape as that of $\hat{\mathcal{R}}^{\texttt{max-}|t|}_{n,M}$. Recently, Bachoc, Blanchard and Neuvial (2018) showed that the orthogonal design restriction can be relaxed to RIP. A symmetric matrix $A\in\mathbb{R}^{p\times p}$ is said to satisfy RIP of order k with RIP constant δ if for all $M\in\mathcal{M}_p(k)$ and for all $\theta\in\mathbb{R}^{|M|}$, $(1-\delta)\|\theta\|^2 \le \theta^\top A(M)\theta \le (1+\delta)\|\theta\|^2$. This is equivalent to

$$\max_{|M|\le k} \bigl\|A(M) - I_{|M|}\bigr\|_{\mathrm{op}} \le \delta, \tag{26}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. Thus if $\hat{\Sigma}_n$ satisfies RIP of order k, all covariate subsets of size k are nearly orthogonal. Theorem 3.3 of Bachoc, Blanchard and Neuvial (2018) proves that for fixed covariates, a Gaussian response, and $\hat{\Sigma}_n$ RIP of order k with constant δ, we have

$$C_{n,k}(\alpha) = O\Bigg(\sqrt{\frac{\log p}{n}} + \delta c(\delta)\sqrt{\frac{k\log(ep/k)}{n}}\Bigg).$$

Here, $c(\delta)$ is an increasing nonnegative function satisfying $c(\delta) \to 1$ as $\delta \to 0$. Hence under the RIP condition (26) with $\delta\sqrt{k} \to 0$, the Lebesgue measure of the confidence region $\hat{\mathcal{R}}^{\texttt{max-|t|}}_{n,M}$ matches again that of our confidence region $\hat{\mathcal{R}}^{\dagger}_{n,M}$. Interestingly, under the RIP condition for $\hat{\Sigma}_n$ with $\delta \to 0$, the confidence region $\hat{\mathcal{R}}^{\dagger}_{n,M}$ provides a parallelepiped with sides nearly parallel to the coordinate axis. Even more strikingly, the following result holds for fixed covariates (but not requiring Gaussianity).

PROPOSITION 6.1. *Define the confidence region*

$$\hat{\mathcal{R}}^{\texttt{RIP}}_{n,M} := \{\theta \in \mathbb{R}^{|M|} : \|\hat{\beta}_{n,M} - \theta\|_\infty \le C^{\Gamma}_n(\alpha)\}.$$

If, for any $1 \le k \le p$, the matrix $\hat{\Sigma}_n$ satisfies the RIP condition of order k with RIP constant δ and $\delta\sqrt{k} = o(1)$ as $n \to \infty$, then

$$\liminf_{n\to\infty} \mathbb{P}\Bigg(\bigcap_{M\in\mathcal{M}_p(k)} \{\beta_{n,M} \in \hat{\mathcal{R}}^{\texttt{RIP}}_{n,M}\}\Bigg) \ge 1 - \alpha.$$

Furthermore, for all $M \in \mathcal{M}_p(k)$,

$$d_H\big(\hat{\mathcal{R}}_{n,M}, \hat{\mathcal{R}}^{\texttt{RIP}}_{n,M}\big) \le C^{\Gamma}_n(\alpha)\big(\delta|M|^{1/2}/(1-\delta)\big), \tag{27}$$

where $d_H(A, B)$ for two sets A, B is the Hausdorff distance between A and B in terms of the Euclidean distance, that is,

$$d_H(A, B) = \max\Big\{\sup_{a\in A}\inf_{b\in B}\|a - b\|_2, \sup_{b\in B}\inf_{a\in A}\|a - b\|_2\Big\}.$$

PROOF. See Section S.8 of the supplement for a proof. □

Inequality (27) implies that the distance between the modified confidence region $\hat{\mathcal{R}}^{\texttt{RIP}}_{n,M}$ and the original confidence region $\hat{\mathcal{R}}_{n,M}$ converges to zero at the rate of $C^{\Gamma}_n(\alpha)\delta\sqrt{k}$ as $\delta = o(1)$. This rate should not be taken in an absolute sense because a "good" confidence region should converge to a singleton set. In our case, it is easy to see that the "Hausdorff radii" satisfy the following inequality:

$$\max\big\{d_H\big(\hat{\mathcal{R}}^{\texttt{RIP}}_{n,M}, \{\hat{\beta}_{n,M}\}\big), (1-\delta)d_H\big(\hat{\mathcal{R}}_{n,M}, \{\hat{\beta}_{n,M}\}\big)\big\} \le C^{\Gamma}_n(\alpha)|M|^{1/2}.$$

Hence inequality (27) implies that the confidence regions $\hat{\mathcal{R}}^{\texttt{RIP}}_{n,M}$ and $\hat{\mathcal{R}}_{n,M}$ get closer to each other faster than they get close to a singleton if $\delta \to 0$.

RIP is a well-known condition in the high-dimensional linear regression literature, but it is also known to be a very restrictive. It implies a requirement of near orthogonal covariate subsets, which is often not justified in practice.

REMARK 6.1 (Generalization of Theorem 3.3 of Bachoc, Blanchard and Neuvial (2018)). This result gives a bound on the expectation of $\sup\{\|\hat{\beta}_{n,M} - \beta_{n,M}\|_\infty : M \in \mathcal{M}_p(k)\}$ for fixed

covariates and a Gaussian response. The following inequality in the proof of Proposition 6.1 provides a deterministic bound on this supremum quantity:

$$\sup_{M\in\mathcal{M}_p(k)} \|\hat{\beta}_{n,M} - \beta_{n,M}\|_\infty \le \mathcal{D}_n^\Gamma\big[1 + \delta\sqrt{k}/(1-\delta)\big].$$

This, along with Lemma 5.1, proves the rate bound in a more general setting.

7. Implied confidence regions and comparisons. For practical purposes, it would be easier to interpret/understand confidence regions in a submodel M if they were rectangles formed from intervals for the coefficients in the submodel M. An obvious conservative construction is to form the smallest confidence rectangle that encloses $\hat{\mathcal{R}}^\dagger_{n,M}$. Note that $\hat{\mathcal{R}}^\dagger_{n,M}$ can be written as

$$\hat{\mathcal{R}}^\dagger_{n,M} = \big\{\hat{\beta}_{n,M} + \big(\hat{\Sigma}_n(M)\big)^{-1}\delta : \|\delta\|_\infty \le C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1, \delta\in\mathbb{R}^{|M|}\big\}.$$

We will now show that the smallest rectangle containing $\hat{\mathcal{R}}^\dagger_{n,M}$ can be obtained analytically from this reformulation of the confidence region. The projection of $\hat{\mathcal{R}}^\dagger_{n,M}$ onto the first coordinate axis is given by $[\hat{L}^\dagger_{n,M}(1), \hat{U}^\dagger_{n,M}(1)]$ where

$$\begin{aligned}
\hat{L}^\dagger_{n,M}(1) &:= \inf\big\{e_1^\top\big(\hat{\beta}_{n,M} + \big(\hat{\Sigma}_n(M)\big)^{-1}\delta\big) : \|\delta\|_\infty \le C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}\\
&= \hat{\beta}_{n,M}(1) - \sup\big\{e_1^\top\big(\hat{\Sigma}_n(M)\big)^{-1}\delta : \|\delta\|_\infty \le C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\},\\
\hat{U}^\dagger_{n,M}(1) &:= \sup\big\{e_1^\top\big(\hat{\beta}_{n,M} + \big(\hat{\Sigma}_n(M)\big)^{-1}\delta\big) : \|\delta\|_\infty \le C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}\\
&= \hat{\beta}_{n,M}(1) + \sup\big\{e_1^\top\big(\hat{\Sigma}_n(M)\big)^{-1}\delta : \|\delta\|_\infty \le C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\},
\end{aligned} \tag{28}$$

where $e_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{|M|}$ is the first basis vector. The second equality for $\hat{L}^\dagger_{n,M}(1)$ in (28) follows from the symmetric of δ. By the duality of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ norms, we get

$$\begin{aligned}
&\sup\big\{e_1^\top\big(\hat{\Sigma}_n(M)\big)^{-1}\delta : \|\delta\|_\infty \le C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}\\
&\quad= \big\|e_1^\top\big(\hat{\Sigma}_n(M)\big)^{-1}\big\|_1\big\{C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}.
\end{aligned}$$

Hence the smallest rectangle enclosing $\hat{\mathcal{R}}^\dagger_{n,M}$ is given by

$$\hat{\mathcal{B}}^\dagger_{n,M} := \prod_{j\in M}[\hat{L}^\dagger_{n,M}(j), \hat{U}^\dagger_{n,M}(j)], \tag{29}$$

where $\hat{L}^\dagger_{n,M}(j)$ and $\hat{U}^\dagger_{n,M}(j)$ satisfy

$$\begin{aligned}
\hat{\beta}_{n,M}(j) - \hat{L}^\dagger_{n,M}(j) &= \hat{U}^\dagger_{n,M}(j) - \hat{\beta}_{n,M}(j)\\
&= \big\|e_j^\top\big(\hat{\Sigma}_n(M)\big)^{-1}\big\|_1\big\{C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}.
\end{aligned}$$

One can also define an enclosing rectangle $\hat{\mathcal{B}}_{n,M}$ for $\hat{\mathcal{R}}_{n,M}$. In this case, an analytical form of $\hat{\mathcal{B}}_{n,M}$ does not exist but can be obtained by solving a linear programming problem (Belloni, Rosenbaum and Tsybakov ((2017), equation (42) of the supplement)). The region $\hat{\mathcal{B}}^\dagger_{n,M}$ inherits symmetry around $\hat{\beta}_{n,M}$ from $\hat{\mathcal{R}}^\dagger_{n,M}$. The same is not the case for $\hat{\mathcal{R}}_{n,M}$.

We next examine by how much the Lebesgue measure of $\hat{\mathcal{B}}^\dagger_{n,M}$ could inflate in relation to $\hat{\mathcal{R}}^\dagger_{n,M}$, whose Lebesgue measure we studied in Section 5. From the definition of $\hat{L}^\dagger_{n,M}(j)$ and $\hat{U}^\dagger_{n,M}(j)$ leading to (29), the width of the rectangle on coordinate j is given by

$$|\hat{U}^\dagger_{n,M}(j) - \hat{L}^\dagger_{n,M}(j)| = 2\big\|e_j^\top\big(\hat{\Sigma}_n(M)\big)^{-1}\big\|_1\big\{C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\big\}. \tag{30}$$

The quantity parenthesized above is bounded in Proposition 5.1, and thus under the conditions of Proposition 5.1 (including $(\Lambda_n(k))^{-1} = O(1)$), we get

$$|\hat{U}^{\dagger}_{n,M}(j) - \hat{L}^{\dagger}_{n,M}(j)| = \begin{cases} O_p\big(\sqrt{|M| \log p/n}\big) & \text{if the covariates are fixed,} \\ O_p\big(\sqrt{|M|^2 \log p/n}\big) & \text{if the covariates are random.} \end{cases} \tag{31}$$

However, assuming the conditions ($\delta_{\mathtt{RIP}}\sqrt{k} = o(1)$) of Proposition 6.1, we get

$$|\hat{U}^{\dagger}_{n,M}(j) - \hat{L}^{\dagger}_{n,M}(j)| = \begin{cases} O_p\big(\sqrt{\log p/n}\big) & \text{if the covariates are fixed,} \\ O_p\big(\sqrt{|M| \log p/n}\big) & \text{if the covariates are not fixed.} \end{cases} \tag{32}$$

See the Appendix for a proof. In summary, the (projected) rectangular confidence regions can have an inflation of order $\sqrt{|M|}$ and no inflation (rate-wise) if the RIP condition is satisfied. It should be noted that the bounds (31) and (32) are worst case bounds.

In Section S.2 of the supplement, we consider another type of implied confidence region related to "max-$|t|$" intervals.

8. Computation by multiplier bootstrap. All of the confidence regions defined in the previous sections depend only on the available data except for the (joint) quantiles $C^{\Gamma}_n(\alpha)$ and $C^{\Sigma}_n(\alpha)$ which depend on the true data distribution. Computation and estimation of the joint bivariate quantiles $C^{\Gamma}_n(\alpha)$ and $C^{\Sigma}_n(\alpha)$ is the most important component of an application of the UPoSI approach to valid post-selection inference. In this section, we apply a high-dimensional central limit theorem to justify the use of the multiplier bootstrap to estimate these quantiles, but in the setting of Lemma 5.1 the classical resampling bootstrap works as well (see Chernozhukov, Chetverikov and Kato (2017) and Zhang and Cheng (2014) for a detailed discussion). For simplicity, we will only discuss the multiplier bootstrap approach in the case of independent random vectors using the results proved in Section S.10 of the supplement, and refer to Zhang and Cheng (2014) for the case of functional dependence described in Section S.11.

Define vectors W_i that contain the contribution of case i to $\hat{\Gamma}_n$ and $\hat{\Sigma}_n$, that is,

$$W_i := \big(\{X_i(j)Y_i\}_{1\le j\le p}, \{X_i(l)X_i(m)\}_{1\le l\le m\le p}\big). \tag{33}$$

The dimension of W_i is q given by $q = 2p + p(p-1)/2 = O(p^2)$. To construct bivariate quantiles for $\mathcal{D}^{\Gamma}_n$ and $\mathcal{D}^{\Sigma}_n$, consider the event $\{\mathcal{D}^{\Gamma}_n \le t_1, \mathcal{D}^{\Sigma}_n \le t_2\}$. As shown in equation (E.13) in Section S.10 of the supplement, this event for any $t_1, t_2 \ge 0$ can be written as a (symmetric) rectangle in terms of

$$S^W_n := n^{-1/2} \sum_{i=1}^{n} \{W_i - \mathbb{E}[W_i]\}.$$

For present purposes, we will assume independent but possibly nonidentically distributed observations (X_i, Y_i). As a consequence, $\mathbb{E}[W_i]$ may not all be equal, in which case it is not possible to estimate these expectations consistently. However, the following procedure will provide conservative inference (recall the remark following Theorem 4.1).

Let $e_1, e_2, \ldots, e_n$ be independent standard normal random variables and define

$$S^{e,W}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i(W_i - \overline{W}_n) \quad \text{where } \overline{W}_n := \frac{1}{n}\sum_{i=1}^{n} W_i.$$

Write $S^{e,W}_n(\mathbf{I})$ for the first p coordinates of $S^{e,W}_n$ (corresponding to $\hat{\Gamma}_n$) and $S^{e,W}_n(\mathbf{II})$ for the remaining coordinates of $S^{e,W}_n$ (corresponding to $\hat{\Sigma}_n$). The following algorithm gives the pseudo-program for implementing the multiplier bootstrap:

1. Generate B_n random vectors from a standard normal distribution of dimension n. Let these be denoted by $\{e_{i,j} : 1 \le i \le n, 1 \le j \le B_n\}$.
2. Compute the jth replicate of $S_n^{e,W}$ as

$$S_{n,j}^{\star} := \frac{1}{n}\sum_{i=1}^{n} e_{i,j}(W_i - \overline{W}_n) \in \mathbb{R}^q \quad \text{for } 1 \le j \le B_n.$$

3. Find any two numbers $(\hat{C}_{1n}^{\Gamma}(\alpha), \hat{C}_{2n}^{\Sigma}(\alpha))$ such that

$$\frac{1}{B_n}\sum_{i=1}^{B_n} \mathbb{1}\{\|S_{n,j}^{\star}(\mathbf{I})\|_{\infty} \le \hat{C}_{1n}^{\Gamma}(\alpha), \|S_{n,j}^{\star}(\mathbf{II})\|_{\infty} \le \hat{C}_{2n}^{\Sigma}(\alpha)\} \ge 1-\alpha.$$

Here, $\mathbb{1}\{A\}$ is the indicator function of a set A.

The following theorem proves the validity of the multiplier bootstrap under assumption (20) of Lemma 5.1. Recall the definition of W_i from (33). Note that we only prove asymptotic conservativeness instead of consistency (which does not hold without further assumptions, such as identical distributions); see Remark 8.1 below and Remark S.10.1 in Section S.10 of the supplement. This can be easily understood by noting that $\mathbb{E}[W_i]$ is replaced by the average $\overline{W}_n$ which is not a consistent estimator. For the following result, define

$$L_{n,p} := \max_{1\le j\le q} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[|W_i(j) - \mathbb{E}[W_i(j)]|^3].$$

Also, recall the sub-Weibull norm defined in (19).

THEOREM 8.1. *Let $(X_i^{\top}, Y_i)^{\top}$, $1 \le i \le n$ be independent random vectors satisfying*

$$\min_{1\le j\le q} n^{-1}\sum_{i=1}^{n} \operatorname{Var}(W_i(j)) \ge B > 0,$$

and for $1 \le i \le n$,

$$\max\Big\{\max_{1\le j\le p} \|X_i(j)\|_{\psi_\gamma}, \|Y_i\|_{\psi_\gamma}\Big\} \le K_{n,p}. \tag{34}$$

Further if $n, p \ge 1$ are such that

$$\max\{L_{n,p}^{-1}K_{n,p}(\log p)^{1+6/\gamma}, L_{n,p}^2 \log^7 p, K_{n,p}^6 \log p, K_{n,q}^2(\log p \log n)^{4/\gamma}\} = o(n), \tag{35}$$

then the multiplier bootstrap described above provides conservative inference in the sense that with probability converging to one, we have

$$\inf_{t_1,t_2\ge 0}\{\mathbb{P}(\mathcal{D}_n^{\Gamma} \le t_1, \mathcal{D}_n^{\Sigma} \le t_2) - \mathbb{P}(\|S_{n,j}^{eW}(\mathbf{I})\|_{\infty} \le t_1, \|S_{n,j}^{eW}(\mathbf{II})\|_{\infty} \le t_2 \mid \mathcal{Z}_n)\} \ge 0,$$

where $\mathcal{Z}_n := \{(X_i^{\top}, Y_i)^{\top} : 1 \le i \le n\}$.

PROOF. Theorems S.10.1 and S.10.2 (stated in Section S.10 of the supplement) apply in the setting above because under assumption (34) we have

$$\max_{1\le i\le n}\max_{1\le j\le q}\|W_i(j)\|_{\psi_{\gamma/2}} \le \max_{1\le i\le n}\max\Big\{\max_{1\le j\le p}\|X_i(j)\|_{\psi_\gamma}, \|Y_i\|_{\psi_\gamma}\Big\}^2 \le K_{n,p}^2.$$

The rate restriction (35) on n and p ensures that the bounds in Theorem S.10.1 and S.10.2 both converge to zero. See again Remark S.10.1 for the conservative nature of the result. □

Condition (35) essentially means that $\log p$ is (very) small compared to n; it can grow with n but slowly. A relaxation of condition (35) can be found in Deng and Zhang (2017). By Theorem 8.1, the estimates $(\hat{C}_{1n}^{\Gamma}(\alpha), \hat{C}_{2n}^{\Sigma}(\alpha))$ are consistent for some quantities that can conservatively replace the quantiles $(C_n^{\Gamma}(\alpha), C_n^{\Sigma}(\alpha))$ of $(\mathcal{D}_n^{\Gamma}, \mathcal{D}_n^{\Sigma})$ in (13). This conservativeness always exists in the case of fixed covariates.

REMARK 8.1 (Consistency under identical distributions). In a framework of merely independent but not necessarily identically distributed random vectors, one can show that it is impossible to prove consistency (for a proof see Kuchibhotla, Brown and Buja (2018) in a simpler setting which, however, applies to the current setting as well). If in addition we assume identically distributed random vectors, the results of Section S.10 prove that the multiplier bootstrap described above is in fact consistent under the assumptions of Theorem 8.1.

9. Simulation study. In this section, we compare our confidence regions with those of the extant literature. Because this literature assumes fixed covariates, our comparisons will be limited to this setting. The only other methods with a universal post-selection guarantee are those of Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016). Additionally, we compare our method with selective inference as proposed by Tibshirani et al. (2016), even though this comparison is not quite fair because their method is specifically designed for certain selection strategies (LARS and forward stepwise selection). We consider the data generating model

$$Y_i = X_i^{\top}\beta_0 + \varepsilon_i, \quad 1 \le i \le n \text{ with } \beta_0 = \mathbf{0}_p, \text{ and } \varepsilon_i \overset{\text{iid}}{\sim} N(0,1).$$

The reason for taking $\beta_0 = \mathbf{0}_p$ is to avoid the effect of conservativeness discussed in the previous section; see also Section S.1 for more details. In computing our confidence regions, we do *not* use the normality of errors in the model. Code and more details are available at https://github.com/post-selection-inference/R. The following three settings of covariates will be considered:

1. *Setting A (orthogonal design):* X_i are such that $\hat{\Sigma}_n = \sum_{i=1}^n X_i X_i^{\top}/n = I_p$, the identity matrix in p dimensions. The data is generated by starting with a random matrix with i.i.d. Gaussian entries and applying Gram–Schmidt to satisfy $\hat{\Sigma}_n = I_p$.
2. *Setting B (exchangeable design):* X_i are chosen such that $\hat{\Sigma}_n = I_p + \alpha \mathbf{1}_p \mathbf{1}_p^{\top}$ with $\alpha = -1/(p+2)$, which is close to the degenerate case attained for $\alpha = -1/p$. The data is first generated as in Setting A and then multiplied by $\hat{\Sigma}_n^{1/2}$.
3. *Setting C (worst-case design):* X_i are chosen such that

$$\hat{\Sigma}_n := \begin{bmatrix} I_{p-1} & c\mathbf{1}_{p-1} \\ \mathbf{0}_{p-1}^{\top} & \sqrt{1-(p-1)c^2} \end{bmatrix} \quad \text{where } c^2 = \frac{1}{2(p-1)},$$

Settings A and B lead to the best rate for the "max-$|t|$" approach (23), while Setting C leads to the worst rate. See Berk et al. ((2013), Sections 6.1 and 6.2) for results in these three settings. We take $n = 200$, $p = 15$ to compare our method with Berk et al. (2013) and $n = 1000$, $p = 500$ to compare our method with selective inference for steps one to five in forward stepwise selection and in LARS. The results (from 100 replications) are summarized in Figure 1 while the exact numbers are given in the supplement.

In Setting C, the collection of all submodels naturally fall into two categories: those containing the last covariate and those that do not. Because the last covariate is highly correlated with the other covariates, the volume of the Berk et al. (2013) regions and our projected regions are both large for the first category of models and small for the second. This division in volumes can be seen from the two dots for each model size in Figure 1.

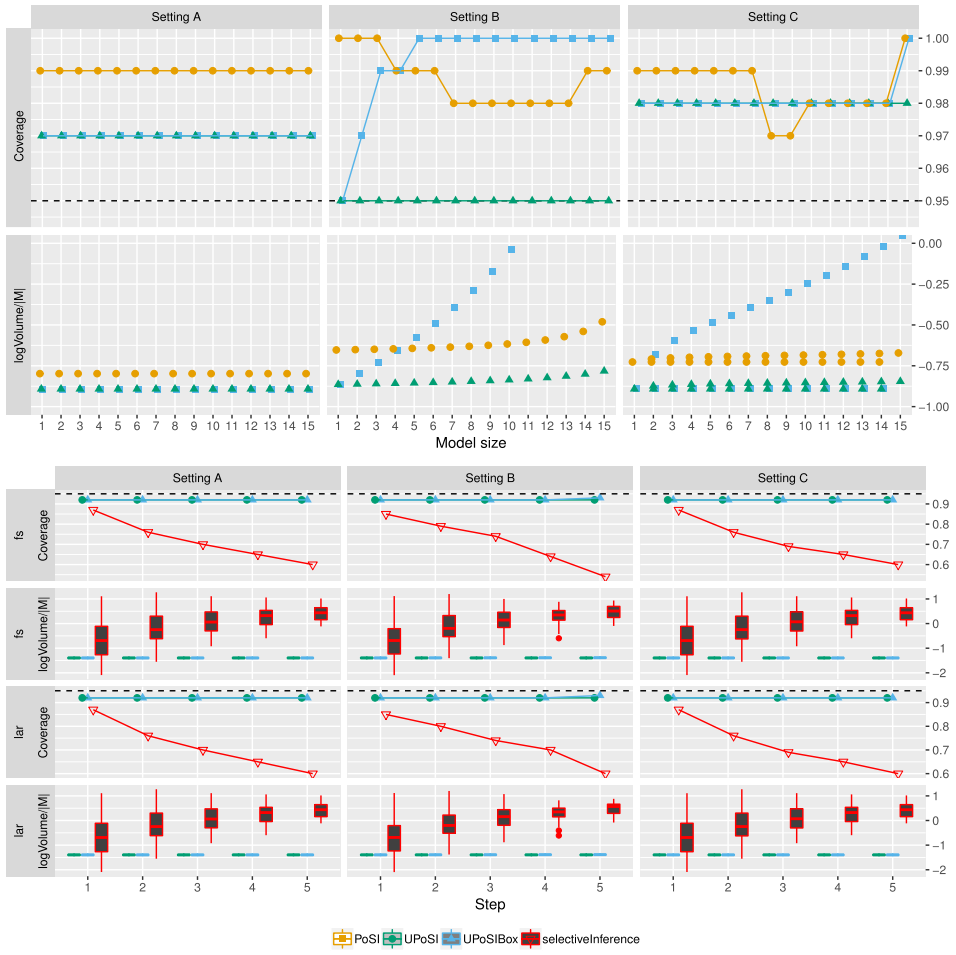

FIG. 1. *Comparison of "UPoSI" with "PoSI" (Berk et al. (2013)) and "selective Inference" (Tibshirani et al. (2016)). Included are the "UPoSI" confidence regions* $\hat{\mathcal{R}}^{\dagger}_{n,M}$ (12) *and the larger "UPoSIBox" regions* $\hat{\mathcal{B}}^{\dagger}_{n,M}$ (29). *The first two plots provide comparisons with the "PoSI" regions* (24) *of Berk et al. (2013). The next four plots show comparisons with "selective inference." Rather than providing overall simultaneous coverage, we show simultaneous coverage for different model sizes separately*: $1 \le |M| \le 15$ *for comparison with "PoSI" and* $1 \le |M| \le 5$ *for comparison with "selective inference." Because the volume of a region in* $|M|$ *dimensions scales like* $C^{|M|}$ *for some constant* C, *we plot, for example,* $\log(\mathrm{Leb}(\hat{\mathcal{R}}^{\dagger}_{n,M}))/|M|$, *which allows comparison across different model sizes. Recall that in Setting C models fall into two groups*: *those that contain the last covariate, and those that do not. This is the reason for showing two dots for each model size in Setting C. The size of dots indicates the proportion of models in each group. The dashed lines in the coverage plots show the nominal confidence level* 0.95.

In all three settings, the UPoSI regions have smaller volume with coverage close to the required confidence of 0.95. In Setting A, the projected confidence region $\hat{\mathcal{B}}^{\dagger}_{n,M}$ is the same as $\hat{\mathcal{R}}^{\dagger}_{n,M}$, and in Settings B and C the projected confidence regions are wider for larger model sizes and are cut-off for comparison purposes. For selective inference, the coverage $\mathbb{P}(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}^{\texttt{SelInf}}_{n,\hat{M}})$ and the volume of the region get worse as the selection steps proceed. Correct coverage deteriorates in the null case simulated here. Apparently, inclusion of incorrect regressors has detrimental effects on the coverage provided by selective inference. (We used the `R` functions `fsInf`, `larInf` with `type='active'` from package `selectiveInference`.)

10. Discussion of the UPoSI approach. We list various advantages and disadvantages of the UPoSI approach proposed in this article. We start with the advantages.

- Most fundamentally, the UPoSI confidence regions are asymptotically valid for post-selection inference under quite arbitrary misspecification and after arbitrary model selection. As such, they represent the first proposal that provides valid and strong post-selection inference in this generality.
- The UPoSI confidence regions are computationally inexpensive. For any selected model M, they depend only on the least squares linear regression estimator $\hat{\beta}_{n,M}$ in that model and model-independent upper joint quantiles $C_n^{\Gamma}(\alpha)$, $C_n^{\Sigma}(\alpha)$ (defined in (13) and (10)). Computation of the latter requires a number of operations that grows no more than a linear function of p (Section 8). This computationally low burden is in sharp contrast to the valid post-selection inference method proposed by Berk et al. (2013) or Bachoc, Preinerstorfer and Steinberger (2016) which requires solving for $\hat{\beta}_{n,M}$ in all the submodels M considered for selection, amounting generally to an NP-hard problem.
- The size of the confidence regions $\hat{\mathcal{R}}_{n,M}^{\dagger}$ in terms of Lebesgue measure converges to zero at a rate that is minimax in the high-dimensional linear regression literature. We therefore suspect this might be the optimal rate here, too, but at present we do not have a proof of, or even a framework for, optimality. An issue is that the volume of the confidence region for model M is computed with respect to Lebesgue measure in $\mathbb{R}^{|M|}$.
- There is one more advantage which might not seem like one at first glance. The UPoSI confidence region for $\beta_{n,M}$ for a particular model does not require information on how many models are being used for model selection. The volume of the confidence region for $\beta_{n,M}$ depends only on features of the selected model M. This implies that the confidence regions $\hat{\mathcal{R}}_{n,M}^{\dagger}$, $M \in \mathcal{M}_p(k)$ can often have much smaller volumes than the ones produced using the approach of Berk et al. (2013) or Bachoc, Preinerstorfer and Steinberger (2016).

There are some issues associated with the UPoSI approach.

- First, our confidence regions are not equivariant under linear transformations of the covariates as noted in Section 4.4. This lack of equivariance is shared with most methods for high-dimensional linear regression that induce sparsity or group sparsity. A more limited but practically more critical form of equivariance would be under changes of units of the variables, that is, under diagonal linear transformations of the observations. A commonly suggested method to attain such equivariance is to standardize all the variables with a dispersion measure, usually the standard deviation (as well as center at a location measure, usually the mean). After standardization, the observations are no longer independent. This is one of the reasons why we did not assume independence in Theorems 4.1, 4.2 and S.3.1. For an application of these results, one needs to prove the rates for the error norms $\mathcal{D}_n^{\Gamma\star}$ and $\mathcal{D}_n^{\Sigma\star}$. We leave it to the reader to verify that the rates are exactly those obtained in Lemma 5.1 (using a Slutsky-type argument). A similar derivation was used by Cui, Leng and Sun (2016).
- Another issue with the current approach is that the proposed confidence regions are motivated by, and justified by, deterministic inequalities that can be loose in some cases. More looseness derives from allowing nonidentical distributions of the observations (see Theorem 8.1).
- Finally, the UPoSI confidence regions are for the full vectors $\beta_{n,M}$ which may entail further looseness for coordinate-wise inference. In spite of these concerns, we obtain asymptotic rates that may well be close to the best possible.

We emphasize before ending this section that the main focus of the current approach is validity and better computational complexity, not optimality. However, optimality holds for our confidence regions as mentioned in Remark 6 for fixed covariates.

11. Conclusions and future directions. In this paper, we have considered a computationally efficient approach to valid post-selection inference in linear regression under arbitrary data-driven method of variable selection. The approach here is very different from the other methodologies available in the literature and is based on the estimating equation of linear regression. Since our confidence regions are based on deterministic inequalities, our results provide valid post-selection inference even under dependence and non-identically distributed random vectors. For this reason, the setting of the current work is the most general available in the literature of post-selection inference.

In addition to providing several valid confidence regions, we compare the Lebesgue measure of our confidence regions with the ones from Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016). This comparison shows that our confidence regions are much smaller (in terms of volume) in case of fixed (non-stochastic) covariates. In general, the volume of our confidence regions scales with the cardinality of model $\hat{M}$ chosen. This is a feature not available from the works of Berk et al. (2013) and Bachoc, Preinerstorfer and Steinberger (2016). Note that the confidence intervals from the selective/conditional inference literature have infinite expected length as shown in Kivaranovic and Leeb (2018).

APPENDIX: PROOFS OF RESULTS IN SECTION 7

Following the inequalities (30), if the covariates are fixed, then $C_n^{\Sigma}(\alpha) = 0$. Hence

$$\begin{aligned}|\hat{U}^{\dagger}_{n,M}(j) - \hat{L}^{\dagger}_{n,M}(j)| &\le \|e_j^{\top}(\hat{\Sigma}_n(M))^{-1}\|_1 C_n^{\Gamma}(\alpha) \\ &= \|e_j^{\top}(\hat{\Sigma}_n(M))^{-1}\|_1 O(\sqrt{\log p/n}),\end{aligned}$$

under the assumptions of Proposition 5.1. Hence

$$\begin{aligned}|\hat{U}^{\dagger}_{n,M}(j) - \hat{L}^{\dagger}_{n,M}(j)| &\le \sqrt{|M|}\|e_j^{\top}(\hat{\Sigma}_n(M))^{-1}\|_2 O\left(\sqrt{\frac{\log p}{n}}\right) \\ &= O_p\left(\sqrt{\frac{|M|\log p}{n}}\right),\end{aligned}$$

if $(\Lambda_n(k))^{-1} = O(1)$. If the covariates are random (or just not fixed), then

$$\begin{aligned}|\hat{U}^{\dagger}_{n,M}(j) - \hat{L}^{\dagger}_{n,M}(j)| &\le \sqrt{|M|}\|e_j^{\top}(\hat{\Sigma}_n(M))^{-1}\|_2 O\left(\sqrt{\frac{|M|\log p}{n}}\right) \\ &= O_p\left(\sqrt{\frac{|M|^2\log p}{n}}\right).\end{aligned}$$

This proves (31). To prove (32), first note that if $\hat{\Sigma}_n$ satisfies the RIP condition of order k with constant $\delta_{\mathtt{RIP}}$, then

$$\begin{aligned}\|e_j^{\top}(\hat{\Sigma}_n(M))^{-1}\|_1 &\le 1 + \|e_j^{\top}(I_{|M|} - (\hat{\Sigma}_n(M))^{-1})\|_1 \\ &\le 1 + \sqrt{|M|}\|I_{|M|} - (\hat{\Sigma}_n(M))^{-1}\|_{\mathrm{op}} \le 1 + \frac{\delta_{\mathtt{RIP}}\sqrt{|M|}}{1-\delta_{\mathtt{RIP}}}.\end{aligned}$$

The assumption $\delta_{\mathtt{RIP}}\sqrt{k} = o(1)$ implies (32).

Acknowledgments. The authors would like to thank Bikram Karmakar and Abhishek Chakrabortty for helpful discussions.

The second, third and sixth authors were supported in part by NSF Grants DMS-10-07657, DMS-1310795 and NSF567018. The fifth author was supported in part by NSF Grant DMS-14-06563.

SUPPLEMENTARY MATERIAL

Supplement to "Valid post-selection inference in model-free linear regression" (DOI: 10.1214/19-AOS1917SUPP; .pdf). See the last paragraph of Section 1.4 for an outline of the supplement.

REFERENCES

BACHOC, F., BLANCHARD, G. and NEUVIAL, P. (2018). On the post selection inference constant under restricted isometry properties. *Electron. J. Stat.* **12** 3736–3757. MR3878579 https://doi.org/10.1214/18-ejs1490

BACHOC, F., LEEB, H. and PÖTSCHER, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *Ann. Statist.* **47** 1475–1504. MR3911119 https://doi.org/10.1214/18-AOS1721

BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2016). Uniformly valid confidence intervals post-model-selection. Preprint. Available at arXiv:1611.01043.

BELLONI, A., ROSENBAUM, M. and TSYBAKOV, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 939–956. MR3641415 https://doi.org/10.1111/rssb.12196

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

BUEHLER, R. J. and FEDDERSEN, A. P. (1963). Note on a conditional property of Student's t. *Ann. Math. Stat.* **34** 1098–1100. MR0150864 https://doi.org/10.1214/aoms/1177704034

BUJA, A., BROWN, L. D., BERK, R. A., GEORGE, E. I., TRASKIN, M., PITKIN, E., ZHAO, L. H. and ZHANG, K. (2019). Models as approximations, part I: Consequences illustrated with linear regression. *Statist. Sci*. To appear. Available at arXiv:1404.1578.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963 https://doi.org/10.1214/16-AOP1113

CLAESKENS, G. and CARROLL, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94** 249–265. MR2331485 https://doi.org/10.1093/biomet/asm034

CUI, Y., LENG, C. and SUN, D. (2016). Sparse estimation of high-dimensional correlation matrices. *Comput. Statist. Data Anal.* **93** 390–403. MR3406221 https://doi.org/10.1016/j.csda.2014.10.001

DENG, H. and ZHANG, C.-H. (2017). Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. Preprint. Available at arXiv:1705.09528.

FITHIAN, W., SUN, D. L. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

FREEDMAN, D. A. (1983). A note on screening regression equations. *Amer. Statist.* **37** 152–155. MR0702208 https://doi.org/10.2307/2685877

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. MR2041481 https://doi.org/10.1198/016214503000000828

KIVARANOVIC, D. and LEEB, H. (2018). Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. Preprint. Available at arXiv:1803.01665.

KUCHIBHOTLA, A. K., BROWN, L. D. and BUJA, A. (2018). Model-free study of ordinary least squares linear regression. Preprint. Available at arXiv:1809.10538.

KUCHIBHOTLA, A. K. and CHAKRABORTTY, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Preprint. Available at arXiv:1804.02605.

KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A., GEORGE, E. I. and ZHAO, L. (2018). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. Preprint. Available at arXiv:1802.05801.

KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A., CAI, J., GEORGE, E. I. and ZHAO, L. H. (2020). Supplement to "Valid post-selection inference in model-free linear regression." https://doi.org/10.1214/19-AOS1917SUPP.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

LIU, R. Y. and SINGH, K. (1995). Using i.i.d. bootstrap inference for general non-i.i.d. models. *J. Statist. Plann. Inference* **43** 67–75. MR1314128 https://doi.org/10.1016/0378-3758(94)00008-J

OLSHEN, R. A. (1973). The conditional level of the F-test. *J. Amer. Statist. Assoc.* **68** 692–698. MR0359198

RENCHER, A. C. and PUN, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics* **22** 49–53.

RINALDO, A., WASSERMAN, L., G'SELL, M.,and LEI, J. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *Ann. Statist.* To appear. Available at arXiv:1611.05401.

SIMMONS, J. P., NELSON, L. D. and SIMONSOHN, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22** 1359–1366.

TIAN, X., BI, N. and TAYLOR, J. (2016). MAGIC: A general, powerful and tractable method for selective inference. Preprint. Available at arXiv:1607.02630.

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689 https://doi.org/10.1080/01621459.2015.1108848

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.* **46** 1255–1287. MR3798003 https://doi.org/10.1214/17-AOS1584

ZHANG, X. and CHENG, G. (2014). Bootstrapping high dimensional time series. Preprint. Available at arXiv:1406.1037.