The Annals of Statistics
2019, Vol. 47, No. 6, 3438–3469
https://doi.org/10.1214/18-AOS1784

BOOTSTRAPPING AND SAMPLE SPLITTING FOR HIGH-DIMENSIONAL, ASSUMPTION-LEAN INFERENCE

BY ALESSANDRO RINALDO, LARRY WASSERMAN AND MAX G'SELL

Carnegie Mellon University

Several new methods have been recently proposed for performing valid inference after model selection. An older method is sample splitting: use part of the data for model selection and the rest for inference. In this paper, we revisit sample splitting combined with the bootstrap (or the Normal approximation). We show that this leads to a simple, assumption-lean approach to inference and we establish results on the accuracy of the method. In fact, we find new bounds on the accuracy of the bootstrap and the Normal approximation for general nonlinear parameters with increasing dimension which we then use to assess the accuracy of regression inference. We define new parameters that measure variable importance and that can be inferred with greater accuracy than the usual regression coefficients. Finally, we elucidate an inference-prediction trade-off: splitting increases the accuracy and robustness of inference but can decrease the accuracy of the predictions.

"Investigators who use [regression] are not paying adequate attention to the connection—if any—between the models and the phenomena they are studying. ...By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian..."

—David Freedman

1. Introduction. We consider the problem of carrying out assumption-lean statistical inference after model selection for high-dimensional linear regression. This a very broad and important topic in the statistical literature for which many approaches have been proposed under different settings—an overview of a subset of these can be found for instance in [24]; we defer a more detailed discussion of the literature and a list of references to Section 1.3. We are concerned with developing ways to assess the importance of the set of covariates returned by a generic model selection procedure, whereby importance is defined in terms of changes in predictive power. Our goal is to derive statistical guarantees for various measures of variable importance, imposing minimal assumptions on the data generating distribution and the model selection methodology, and allowing for increasing dimensions. In particular, though we will mainly use linear models, we do not assume that the true regression function is linear. We show the following:

Received April 2018; revised November 2018.
MSC2010 subject classifications. Primary 62F40, 62F35; secondary 62J05, 62G09, 62G20.
Key words and phrases. Sample splitting, bootstrap, regression, assumption-lean.

1. Inference based on sample splitting followed by the bootstrap (or a Normal approximation) may give assumption-lean, robust confidence intervals under weak assumptions.

2. The usual linear regression or projection parameters, providing the best linear predictor, are not necessarily a good target for inference in the assumption-lean framework. Instead, we propose new measure of variable importance, called LOCO (Leave-Out-COvariates) parameters, that are interpretable, are compatible with arbitrary model selection rules and can be estimated accurately.

3. We provide novel bounds on the accuracy of the Normal approximation and the bootstrap to the distribution of the projection parameters when the dimension increases and the linear model is misspecified. In fact, we give new general bounds on Normal approximations for nonlinear parameters of increasing dimension. Our results offer new insights on the accuracy of inference in high-dimensional situations and suggest, in particular, that the accuracy of the Normal approximation for the standard regression parameters is very poor.

4. We show that the law of the projection parameters cannot be estimated consistently based solely on training data. In contrast, sample splitting provides a simple way to obtain such an estimator.

5. We exhibit an interesting trade-off between prediction accuracy and inferential accuracy in sample splitting.

1.1. *Variable importance.* We consider a distribution-free regression framework, where the random pair $Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$ of d-dimensional covariates and response variable has an unknown distribution belonging to a large nonparametric class $\mathcal{Q}$ of probability distributions on $\mathbb{R}^{d+1}$. We impose minimal assumptions on the regression function $x \in \mathbb{R}^d \mapsto \mu(x) = \mathbb{E}[Y|X = x]$ describing the relationship between the vector of covariates and the expected value of the response variable. In particular, we do not require it to be linear. We observe data $\mathcal{D}_n = (Z_1, \ldots, Z_n)$, an i.i.d. sample of size n from some P in $\mathcal{Q}$, where $Z_i = (X_i, Y_i) \in \mathbb{R}^{d+1}$, for $i = 1, \ldots, n$ and the class $\mathcal{Q} = \mathcal{Q}_n$, to be specified later, may depend on the sample size. We apply to the data a procedure w_n, which returns both a subset of the covariates and an estimator of the regression function over the selected covariates. Formally,

$$\mathcal{D}_n \mapsto w_n(\mathcal{D}_n) = (\widehat{S}, \widehat{\mu}_{\widehat{S}}),$$

where $\widehat{S}$, the selected model, is a random, nonempty subset of $\{1, \ldots, d\}$ and $\widehat{\mu}_{\widehat{S}}$ is an estimator of the regression function $x \in \mathbb{R}^d \mapsto \mathbb{E}[Y|X_{\widehat{S}} = x_{\widehat{S}}]$ restricted to $\widehat{S}$, where $(X, Y) \sim P$ independent of $\mathcal{D}_n$ and, for a vector $x = (x(1), \ldots, x(d)) \in \mathbb{R}^d$, we set $x_{\widehat{S}} = (x(j), j \in \widehat{S})$.

The only assumption we impose on w_n is that the maximum size of the selected model be under our control; that is, $1 \leq |\widehat{S}| \leq k$, for a predefined positive integer $k \leq d$, where k and d can both increase with the sample size. The selected

model $\widehat{S}$ need not be a good approximation of any optimal model; however, optimality may be defined. Furthermore, $\widehat{\mu}_{\widehat{S}}$ need not be a consistent estimator of the regression function restricted to $\widehat{S}$. Although our framework allows for arbitrary procedures, we will be focusing on linear estimators: $x \mapsto \widehat{\mu}_{\widehat{S}}(x) = \widehat{\beta}_{\widehat{S}}^{\top} x_{\widehat{S}}$, where $\widehat{\beta}_{\widehat{S}}$ is any estimator of the linear regression coefficients for the selected variables, for instance ordinary least squares. In particular, $\widehat{\beta}_{\widehat{S}}$ may arise from fitting a sparse linear model, with, for example, the lasso or stepwise-forward regression.

Our goal is to provide statistical guarantees for various measure of variable importance applied to the covariates in $\widehat{S}$, uniformly over the choice of w_n and over all the distributions $P \in \mathcal{Q}_n$. We will accomplish this goal by producing confidence sets for four *random* parameters taking values in $\mathbb{R}^{\widehat{S}}$, each providing a different assessment of the level of statistical significance of the variables in $\widehat{S}$ from a purely *predictive* standpoint. All of the random parameters under consideration are functions of the data generating distribution P, of the sample $\mathcal{D}_n$ and its size n, and, importantly, of the model selection and the estimation procedure associated with w_n.

The projection parameter $\beta_{\widehat{S}}$. The linear projection parameter $\beta_{\widehat{S}}$ is defined to be the vector of coefficients of the best linear predictor of Y using $X_{\widehat{S}}$:

$$\beta_{\widehat{S}} = \operatorname*{argmin}_{\beta \in \mathbb{R}^{\widehat{S}}} \mathbb{E}_{X,Y}\big(Y - \beta^{\top} X_{\widehat{S}}\big)^2,$$

where $\mathbb{E}_{X,Y}$ denote the expectation with respect to the distribution of (X, Y). (Here, we are implicitly assuming that both Y and the coordinates of $X_{\widehat{S}}$ have finite second moments and that the covariance of $X_{\widehat{S}}$ is invertible.) The terminology projection parameters refers to the fact that $X^{\top}\beta_{\widehat{S}}$ is the L_2 projection of Y into the linear space of all random variables that can be obtained as linear functions of $X_{\widehat{S}}$. For a through discussion and an analysis of the properties of such parameters, see [15] as well as [12, 34, 63, 66]. The projection parameter is well-defined even though the true regression function μ is not linear. Indeed, it is immediate that

$$\beta_{\widehat{S}} = \Sigma_{\widehat{S}}^{-1} \alpha_{\widehat{S}}, \tag{1}$$

where $\alpha_{\widehat{S}} = (\alpha_{\widehat{S}}(j),\, j \in \widehat{S}), \alpha_{\widehat{S}}(j) = \mathbb{E}_{X,Y}[Y X_{\widehat{S}}(j) | \mathcal{D}_n]$ and $\Sigma_{\widehat{S}} = \mathbb{E}_X[X_{\widehat{S}} X_{\widehat{S}}^{\top} | \mathcal{D}_n]$.

The LOCO parameters $\gamma_{\widehat{S}}$ *and* $\phi_{\widehat{S}}$. Often, statisticians are interested in $\beta_{\widehat{S}}$ as a measure of the importance of the selected covariates. But there are of course other ways to quantify variable significance. Toward that end, we will consider two parameters of variable importance, which we refer to as *Leave Out COvariate Inference—or LOCO—parameters*. They were originally defined in [36] and are similar to the variable importance measures used in random forests. The first LOCO parameter is $\gamma_{\widehat{S}} = (\gamma_{\widehat{S}}(j) : j \in \widehat{S})$, where

$$\gamma_{\widehat{S}}(j) = \mathbb{E}_{X,Y}\big[|Y - \widehat{\beta}_{\widehat{S}(j)}^{\top} X_{\widehat{S}(j)}| - |Y - \widehat{\beta}_{\widehat{S}}^{\top} X_{\widehat{S}}| | \mathcal{D}_n\big]. \tag{2}$$

In the last expression, $\widehat{\beta}_{\widehat{S}}$ is any estimator of the projection parameter $\beta_{\widehat{S}}$ and $\widehat{S}(j)$ and $\widehat{\beta}_{\widehat{S}(j)}$ are obtained by rerunning the model selection and estimation procedure after removing the jth covariate. To be clear, for each $j \in \widehat{S}$, $\widehat{S}(j)$ is a subset of size at most k of $\{1, \dots, d\} \setminus \{j\}$. Notice that the selected model can be different when the jth covariate is held out from the data, so that the intersection between $\widehat{S}(j)$ and $\widehat{S}$ can be smaller than $k-1$. The interpretation of $\gamma_{\widehat{S}}(j)$ is simple: it is the increase in prediction error by not having access to the jth covariate. It is easy to extend the definition of this parameter by leaving out several variables from $\widehat{S}$ at once without additional conceptual difficulties.

The parameter $\gamma_{\widehat{S}}$ has advantages over the projection parameter $\beta_{\widehat{S}}$: it is more interpretable since it refers directly to prediction error and, as we will see, the accuracy of the Normal approximation and the bootstrap is much higher. The second type of LOCO parameters that we consider are the median LOCO parameters $\phi_{\widehat{S}} = (\phi_{\widehat{S}}(j), j \in \widehat{S})$ with

$$\phi_{\widehat{S}}(j) = \operatorname{median}\big[|Y - \widehat{\beta}_{\widehat{S}(j)}^{\top} X_{\widehat{S}}| - |Y - \widehat{\beta}_{\widehat{S}}^{\top} X_{\widehat{S}}| \mathcal{D}_n\big]. \tag{3}$$

As with $\gamma_{\widehat{S}}$, we may leave out multiple covariates at the same time.

The prediction parameter $\rho_{\widehat{S}}$. It may also be of interest to obtain an omnibus parameter that measures how well the selected model will predict future observations. To this end, we define the future predictive error as

$$\rho_{\widehat{S}} = \mathbb{E}_{X,Y}\big[|Y - \widehat{\beta}_{\widehat{S}}^{\top} X_{\widehat{S}}| | \mathcal{D}_n\big], \tag{4}$$

where $\widehat{\beta}_{\widehat{S}}$ is computed based on $\mathcal{D}_n$.

Some additional remarks on these parameter choices can be found in Supplement D.

1.2. *Goals and assumptions.* Our main goal is to provide statistical guarantees for each of the four random parameters of variable significance introduced above, under an assumption-lean framework. For notational convenience, in this section we let $\theta_{\widehat{S}}$ be any of the parameters of interest: $\beta_{\widehat{S}}$, $\gamma_{\widehat{S}}$, $\phi_{\widehat{S}}$ or $\rho_{\widehat{S}}$.

We will rely on sample splitting: assuming for notational convenience that the sample size is $2n$, we randomly split the data $\mathcal{D}_{2n}$ into two halves, $\mathcal{D}_{1,n}$ and $\mathcal{D}_{2,n}$. Next, we run the model selection and estimation procedure w_n on $\mathcal{D}_{1,n}$, obtaining both $\widehat{S}$ and $\widehat{\mu}_{\widehat{S}}$. We then use the second half of the sample $\mathcal{D}_{2,n}$ to construct an estimator $\widehat{\theta}_{\widehat{S}}$ and a confidence set $\widehat{C}_{\widehat{S}}$ for $\theta_{\widehat{S}}$ satisfying the following properties:

Concentration:

$$\limsup_{n\to\infty} \sup_{w_n \in \mathcal{W}_n} \sup_{P \in \mathcal{Q}_n} \mathbb{P}\big(\|\widehat{\theta}_{\widehat{S}} - \theta_{\widehat{S}}\| > r_n\big) \to 0, \tag{5}$$

Coverage validity (honesty):

$$\liminf_{n\to\infty} \inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{Q}_n} \mathbb{P}(\theta_{\widehat{S}} \in \widehat{C}_{\widehat{S}}) \geq 1 - \alpha, \tag{6}$$

Accuracy:

$$\limsup_{n\to\infty} \sup_{w_n\in\mathcal{W}_n} \sup_{P\in\mathcal{Q}_n} \mathbb{P}\big(\nu(\widehat{C}_{\widehat{S}}) > \epsilon_n\big) \to 0, \tag{7}$$

where $\alpha \in (0, 1)$ is a prespecified level of significance, $\mathcal{W}_n$ is the set of all the model selection and estimation procedures on samples of size n, r_n and ϵ_n both vanish as $n \to \infty$ and ν is the volume (Lebesgue measure) of the set. The probability statements above take into account both the randomness in the sample $\mathcal{D}_n$ and the randomness associated to splitting it into halves.

REMARK. The property that the coverage of $\widehat{C}_{\widehat{S}}$ is guaranteed uniformly over the entire class $\mathcal{Q}_n$ is known as (asymptotic) honesty [37]. Note that the confidence sets are for random parameters (based on half the data) but the uniform coverage, accuracy and concentration guarantees hold with the respect to the distribution of the entire sample and the randomness associated to splitting.

The statistical guarantees listed above ensure that both $\widehat{\theta}_{\widehat{S}}$ and $\widehat{C}_{\widehat{S}}$ are robust with respect to the choice of w_n. We seek validity over all model selection and estimation rules because, in realistic data analysis, the procedure w_n can be very complex. In particular, the choice of model can involve: plotting, outlier removal, transformations, choosing among various competing models, etc. Thus, unless we have validity over all w_n, there will be room for unconscious biases to enter.

1.3. *Related work.* The problem of inference after model selection has received much attention lately. Much of the work falls broadly into three categories: inference uniformly over selection procedure, inference with regard to a particular debiased or desparsified model and inference conditional on model selection. We discuss these approaches in more detail in Section B.

The uniform approach includes POSI [12], which constructs valid inferential procedures regardless of the model selection procedure by maximizing over all possible model selections. This method assumes Normality and a fixed, known variance, and is computationally expensive. The idea is further extended in [5, 6] by considering more general parameters of interest and by allowing for heteroskedasticity, nonnormality and model misspecification.

Most other approaches focus on a particular model selection procedure and conduct inference for selections made by that procedure. This includes the literature on debiased or desparsified regularized models; see, for example, [13, 14, 25, 33, 50, 64, 68, 69]. This line of work aims at constructing confidence intervals for parameters in high-dimensional regression and can be used for the selected model if a Bonferroni correction is applied. However, these results typically rely on the assumption that the linear model is correct as well as on various other regularity assumptions on the design matrix and the error distribution.

A separate literature on selective inference has focused on inference with respect to the selected model, conditional on the event of that a particular model is

selected. This began with [38], but was developed more fully in [28, 34] and [63]. Further works in this area include [39–41, 61–63]. In the simplest version, the distribution of $\sqrt{n}(\widehat{\beta}(j) - \beta(j))$ conditional on the selected model has a truncated Gaussian distribution, if the errors are Normal and the covariates are fixed. The cumulative distribution function of the truncated Gaussian is used as a pivot to obtain tests and confidence intervals. This approach requires Normality, and a fixed, known variance.

There have been several additional approaches to this problem that do not fall in any of these broad categories. While this is a larger literature than can be addressed completely here, it includes early work on model selection [32] and model averaging interpretations [30]; the impossibility results of [35] and [15] on random X and model misspecification; methods based on resampling or sample splitting [17, 18, 26, 44, 67]; stability selection [43, 57]; the conformal inference approach of [36]; goodness-of-fit tests of [56]; moment-constraint-based uniform confidence sets [4]; [42] on inference about groups of variables under general designs; [9] in the instrumental variable setting; [10] on post-selection inference for Z-estimators and the knockoffs approach of [7] and later [16]. Although they are not directed at linear models, [65] and [45] address similar problems for random forests.

Sample splitting. Perhaps the oldest method for inference after model selection is sample splitting: half the data $\mathcal{D}_{1,n}$ are used for model fitting and the other half $\mathcal{D}_{2,n}$ are used for inference.[1]

The earliest references for sample splitting include [8, 23, 27, 29], page 13 of [46, 47], page 37 of [48] and [51]. To quote Barnard: "... the simple idea of splitting a sample in two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics..."

To the best of our knowledge there are only two methods that achieve asymptotically honest coverage: sample splitting and uniform inference. Uniform inference is based on estimating the distribution of the parameter estimators over all possible model selections. In general, this is infeasible. But we compare sample splitting and uniform inference in a restricted model in Section 3.

1.4. *Outline.* In Section 2, we provide results in the concentration, coverage validity and accuracy for the estimators described above. In Section 3, we compare sample splitting to nonsplitting strategies. In Section 4, we establish Berry–Esseen bounds for nonlinear statistics with possibly increasing dimension. Section 5 contains concluding remarks. Extra results, proofs and a discussion of another version of the bootstrap, are relegated to the Supplement [55], including numerical examples in Supplement A and comments on other methods in Supplement B.

[1]For simplicity, we assume that the data are split into two parts of equal size. The problem of determining the optimal size of the split is not considered in this paper. Some results on this issue are contained in [58].

1.5. *Notation.* Let $Z = (X, Y) \sim P$ where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$. We write $X = (X(1), \ldots, X(d))$ to denote the components of the vector X. If X is a random quantity with distribution P, we will write the expectation with respect to P as $\mathbb{E}_X[\cdot]$, $\mathbb{E}_P[\cdot]$ or, when it is clear from the context, simply $\mathbb{E}[\cdot]$. Define $\Sigma = \mathbb{E}[XX^\top]$ and $\alpha = (\alpha(1), \ldots, \alpha(d))$ where $\alpha(j) = \mathbb{E}[YX(j)]$. Let $\sigma = \mathrm{vec}(\Sigma)$, where vec is the operator that stacks a matrix into one large vector, and $\psi \equiv \psi(P) = (\sigma, \alpha)$. Similarly, define $\Omega = \Sigma^{-1}$ and $\omega = \mathrm{vec}(\Omega)$. The regression function is $\mu(x) = \mathbb{E}_{Y|X=x}[Y|X = x]$. We use ν to denote Lebesgue measure. We write $a_n \preceq b_n$ to mean that there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all large n. For a nonempty subset $S \subset \{1, \ldots, d\}$ of the covariates X_S or $X(S)$ denotes the corresponding elements of X: $(X(j) : j \in S)$. Similarly, $\Sigma_S = \mathbb{E}[X_S X_S^\top]$ and $\alpha_S = \mathbb{E}[YX_S]$. Integer index subscripts X_i will always be used to index a collection $X_1, X_2, \ldots$, and not to reference individual elements of X. $A \otimes B$ denotes the Kronecker product of matrices. The commutation matrix $K_{m,n}$ is the $mn \times mn$ matrix defined by $K_{m,n}\,\mathrm{vec}(A) = \mathrm{vec}(A^\top)$. For any $k \times k$ matrix A, $\mathrm{vech}(A)$ denotes the column vector of dimension $k(k+1)/2$ obtained by vectorizing only the lower triangular part of $k \times k$ matrix A.

2. Main results. Recall that we rely on data splitting: we randomly split the $2n$ data points into two halves $\mathcal{D}_{1,n}$ and $\mathcal{D}_{2,n}$. Then, for a given choice of the model selection and estimation rule w_n, we use $\mathcal{D}_{1,n}$ to select a nonempty set of variables $\widehat{S} \subset \{1, \ldots, d\}$ where $k = |\widehat{S}| < n$. For the LOCO and prediction parameters, based on $\mathcal{D}_{1,n}$, we also compute $\widehat{\beta}_{\widehat{S}}$, any estimator of the projection parameters restricted to $\widehat{S}$. In addition, for each $j \in \widehat{S}$, we further compute, still using $\mathcal{D}_{1,n}$ and the rule w_n, $\widehat{\beta}_{\widehat{S}(j)}$, the estimator of the projection parameters over the set $\widehat{S}(j)$. Also, for $l = 1, 2$, we denote with $\mathcal{I}_{l,n}$ random subset of $\{1, \ldots, 2n\}$ containing the indexes for the data points in $\mathcal{D}_{l,n}$.

2.1. *Projection parameters.* In his section, we will derive various statistical guarantees for the projection parameters, defined in (1). We will first define the class of data generating distributions on $\mathbb{R}^{d+1}$ for which our results hold. In the definition below, S denotes a nonempty subset of $\{1, \ldots, d\}$ and $W_S = (\mathrm{vech}(X_S X_S^\top), X_S Y)$.

DEFINITION 1. Let $\mathcal{P}_n^{\mathrm{OLS}}$ be the set of all probability distributions P on $\mathbb{R}^{d+1}$ with zero mean, Lebesgue density and such that, for some positive quantities A, a, u, U, v and $\overline{v}$:

1. the support of P is contained in $[-A, A]^{d+1}$;
2. $\min_{\{S:\,|S|\leq k\}} \lambda_{\min}(\Sigma_S) \geq u$ and $\max_{\{S:\,|S|\leq k\}} \lambda_{\max}(\Sigma_S) \leq U$, where $\Sigma_S = \mathbb{E}_P[X_S X_S^\top]$;
3. $\min_{\{S:\,|S|\leq k\}} \lambda_{\min}(\mathrm{Var}_P(W_S)) \geq v$ and $\max_{\{S:\,|S|\leq k\}} \lambda_{\max}(\mathrm{Var}_P(W_S)) \leq \overline{v}$;
4. $\min\{U, \overline{v}\} \geq \eta$, for a fixed $\eta > 0$.

The first compactness assumption is made out of convenience, and may be easily relaxed by assuming instead that Y and the coordinates of X are sub-Gaussian. In particular, it is weaker than assuming that the entire vector X is sub-Gaussian. The bound on the smallest eigenvalue of Σ_S, uniformly over all subsets S is natural: the projection parameter is only well-defined provided that Σ_S is invertible for all the subsets S under consideration. The quantities v and $\overline{v}$ in part 3. are akin to fourth moment conditions. In particular, one can always take $\overline{v} \le A^2k^2$ in the very worst case. Finally, the assumption of zero mean is imposed out of convenience and to simplify our derivations, so that we need not to be concerned with an intercept term. As remarked above, in all of our results we have kept track of the dependence on the constants a, u, U, v and $\overline{v}$, so that we may in fact allow all of these quantities to change with n (but we do treat A as fixed and, therefore, have incorporated it into the constants).

We will be studying the ordinary least squares estimator of the random projection parameter $\beta_{\widehat{S}}$ defined in (1). This estimator is

$$\widehat{\beta}_{\widehat{S}} = \widehat{\Sigma}_{\widehat{S}}^{-1}\widehat{\alpha}_{\widehat{S}}, \tag{8}$$

where, for any nonempty subset S of $\{1, \dots, d\}$,

$$\widehat{\alpha}_S = \frac{1}{n}\sum_{i\in\mathcal{I}_{2,n}} Y_i X_i(S) \quad \text{and} \quad \widehat{\Sigma}_S = \frac{1}{n}\sum_{i\in\mathcal{I}_{2,n}} X_i(\widehat{S})X_i(S)^\top. \tag{9}$$

Note that $\beta_{\widehat{S}}$ depends on $\mathcal{D}_{1,n}$ through $\widehat{S}$, and $\widehat{\beta}_{\widehat{S}}$ is based on the sub-sample $\mathcal{D}_{2,n}$ and restricted to the coordinates $\widehat{S}$. Since each $P \in \mathcal{P}_n^{\mathrm{OLS}}$ has a Lebesgue density, $\widehat{\Sigma}_{\widehat{S}}$ is invertible almost surely as long as $n \ge k \ge |\widehat{S}|$.

In order to relate $\widehat{\beta}_{\widehat{S}}$ to $\beta_{\widehat{S}}$, it will first be convenient to condition on $\widehat{S}$ and thus regard $\beta_{\widehat{S}}$ as a k-dimensional deterministic vector of parameters, which depends on some unknown $P \in \mathcal{P}_n^{\mathrm{OLS}}$. Then $\widehat{\beta}_{\widehat{S}}$ is an estimator of a fixed parameter $\beta_{\widehat{S}} = \beta_{\widehat{S}}(P)$ computed using an i.i.d. sample $\mathcal{D}_{2,n}$ from the same distribution $P \in \mathcal{P}_n^{\mathrm{OLS}}$. Notice that $\widehat{\beta}_{\widehat{S}}$ is not an unbiased estimator of $\beta_{\widehat{S}}$, conditionally or unconditionally on $\mathcal{D}_{2,n}$.

For each $P \in \mathcal{P}_n^{\mathrm{OLS}}$, we can represent the parameters $\Sigma_{\widehat{S}} = \Sigma_{\widehat{S}}(P)$ and $\alpha_{\widehat{S}} = \alpha_{\widehat{S}}(P)$ in vectorized form as

$$\psi = \psi_{\widehat{S}} = \psi(\widehat{S}, P) = \begin{bmatrix} \mathrm{vech}(\Sigma_{\widehat{S}}) \\ \alpha_{\widehat{S}} \end{bmatrix} \in \mathbb{R}^b, \tag{10}$$

where $b = \frac{k^2+3k}{2}$. Similarly, based on the subsample $\mathcal{D}_{2,n}$ we define the n random vectors

$$W_i = \begin{bmatrix} \mathrm{vech}\big(X_i(\widehat{S})X_i(\widehat{S})^\top\big) \\ Y_i \cdot X_i(\widehat{S}) \end{bmatrix} \in \mathbb{R}^b, \qquad i \in \mathcal{I}_{2,n},$$

and their average

$$\widehat{\psi} = \widehat{\psi}_{\widehat{S}} = \frac{1}{n} \sum_{i \in \mathcal{I}_{2,n}} W_i. \tag{11}$$

It is immediate to see that $\mathbb{E}_P[\widehat{\psi}] = \psi$, for all $P \in \mathcal{P}_n^{\mathrm{OLS}}$.

We express both the projection parameter $\beta_{\widehat{S}}$ and the least squares estimator $\widehat{\beta}_{\widehat{S}}$ as nonlinear functions of ψ and $\widehat{\psi}$, respectively, in the following way. Let $g : \mathbb{R}^b \to \mathbb{R}^k$ be given by

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \big(\mathrm{math}(x_1)\big)^{-1} x_2, \tag{12}$$

where x_1 and x_2 correspond to the first $k(k+1)/2$ and the last k coordinates of x, respectively, and math is the inverse mapping of vech, that is, $\mathrm{math}(x) = A$ if and only if $\mathrm{vech}(A) = x$. Notice that g is well-defined over the convex set

$$\left\{ \begin{bmatrix} \mathrm{vech}(\Sigma) \\ x \end{bmatrix} : \Sigma \in \mathcal{C}_k^+, x \in \mathbb{R}^k \right\},$$

where $\mathcal{C}_k^+$ is the cone of positive definite matrices of dimension k. It follows from our assumptions that, for each $P \in \mathcal{P}_n^{\mathrm{OLS}}$, ψ is in the domain of g and, as long as $n \geq d$, so is $\widehat{\psi}$, almost surely. Thus, we may write

$$\beta_{\widehat{S}} = g(\psi_{\widehat{S}}) \quad \text{and} \quad \widehat{\beta}_{\widehat{S}} = g(\widehat{\psi}_{\widehat{S}}).$$

Concentration of $\widehat{\beta}_{\widehat{S}}$. We begin by deriving high probability concentration bounds for $\widehat{\beta}_{\widehat{S}}$ around $\beta_{\widehat{S}}$. When there is no model selection nor sample splitting—so that $\widehat{S}$ is deterministic and equal to $\{1, \ldots, d\}$—our results yield consistency rates for the ordinary least squares estimator of the projection parameters, under increasing dimensions and a misspecified model. An analogous result was established in [31] for linear and ridge regression without model selection, where the approximation error $\mu(x) - x^\top \beta$ is accounted for explicitly.

THEOREM 1. *Let*

$$B_n = C \frac{1}{u^3} \sqrt{U^3 k \frac{\log k + \log n}{n}}$$

and assume that $\max\{B_n, u B_n\} \to 0$ *as* $n \to \infty$. *Then, there exists a constant* $C > 0$, *dependent on* A *and* η *only, such that, for all* n *large enough,*

$$\sup_{w_n \in \mathcal{W}_n} \sup_{P \in \mathcal{P}_n^{\mathrm{OLS}}} \|\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}}\| \leq C B_n, \tag{13}$$

with probability at least $1 - \frac{2}{n}$ *with respect to joint distribution of the entire sample and of the splitting process.*

Theorem 1 can be easily generalized to cover the case in which the model selection and the computation of the projection parameters are performed on the entire data set and not on separate, independent splits. In this situation, we seek a high probability bound for the quantity

$$\max_S \|\beta_S - \widehat{\beta}_S\|,$$

where the maximum is over all nonempty subsets of $\{1, \ldots, d\}$ of size at most k and $\widehat{\beta}_S = \widehat{\Sigma}_S^{-1}\widehat{\alpha}_S$ (see equation (9)). Since there are less than $(\frac{ed}{k})^k$ such subsets, a union bound argument will yield a rate of consistency for the projection parameter under arbitrary model selection rules without relying on sample splitting. We omit the details.

Confidence sets for the projection parameters: *Normal approximations*. We will now derive confidence intervals for the projection parameters using on a high-dimensional Normal approximation to $\widehat{\beta}_{\widehat{S}}$. The construction of such confidence sets entails approximating the dominant linear term in the Taylor series expansion of $\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}}$ by a centered Gaussian vector in $\mathbb{R}^{\widehat{S}}$ with the same covariance matrix $\Gamma_{\widehat{S}}$ (see (54) in Section 4). The coverage properties of the resulting confidence sets depend crucially on the ability to estimate such covariances. For that purpose, we use a plug-in estimator, given by

$$\widehat{\Gamma}_{\widehat{S}} = \widehat{G}_{\widehat{S}}\widehat{V}_{\widehat{S}}\widehat{G}_{\widehat{S}}^{\top}, \tag{14}$$

where $\widehat{V}_{\widehat{S}} = \frac{1}{n}\sum_{i=1}^n [(W_i - \widehat{\psi})(W_i - \widehat{\psi})^{\top}]$ is the $b \times b$ empirical covariance matrix of the W_i's and the $k \times b$ matrix $\widehat{G}_{\widehat{S}}$ is the Jacobian of the mapping g, given explicitly in (106) [55], evaluated at $\widehat{\psi}$.

The first confidence set for the projection parameter based on the Normal approximation that we propose is an L_∞ ball of appropriate radius centered at $\widehat{\beta}_{\widehat{S}}$:

$$\widehat{C}_{\widehat{S}} = \left\{\beta \in \mathbb{R}^k : \|\beta - \widehat{\beta}_{\widehat{S}}\|_\infty \le \frac{\widehat{t}_\alpha}{\sqrt{n}}\right\}, \tag{15}$$

where $\widehat{t}_\alpha$ is a random radius (dependent on $\mathcal{D}_{2,n}$) such that

$$\mathbb{P}\big(\big\|\widehat{\Gamma}_{\widehat{S}}^{1/2} Q\big\|_\infty \le \widehat{t}_\alpha\big) = \alpha, \tag{16}$$

with Q a random vector having the k-dimensional standard Gaussian distribution and independent of the data.

In addition to the L_∞ ball given in (15), we also construct a confidence set for $\beta_{\widehat{S}}$ to be a hyperrectangle, with sides of different lengths in order to account for different variances in the covariates. This can be done using the set

$$\widetilde{C}_{\widehat{S}} = \bigotimes_{j \in \widehat{S}} \widetilde{C}(j), \tag{17}$$

where

$$\widetilde{C}(j) = \left[\widehat{\beta}_{\widehat{S}}(j) - z_{\alpha/(2k)}\sqrt{\frac{\widehat{\Gamma}_{\widehat{S}}(j,j)}{n}}, \widehat{\beta}_{\widehat{S}}(j) + z_{\alpha/(2k)}\sqrt{\frac{\widehat{\Gamma}_{\widehat{S}}(j,j)}{n}}\right],$$

with $\widehat{\Gamma}_{\widehat{S}}$ given by (14) and $z_{\alpha/(2k)}$ the upper $1-\alpha/(2k)$ quantile of a standard Normal variate. Notice that we use a Bonferroni correction to guarantee a nominal coverage of $1-\alpha$.

THEOREM 2. *Let $\widehat{C}_{\widehat{S}}$ and $\widetilde{C}_{\widehat{S}}$ the confidence sets defined in* (15) *and* (17), *respectively. Let*

$$u_n = u - K_{2,n} \quad \text{and} \quad U_n = U + K_{2,n}, \tag{18}$$

where

$$K_{2,n} = CA\sqrt{kU\frac{\log k + \log n}{n}},$$

with $C = C(\eta) > 0$ the universal constant in (101). *Assume, in addition, that n is large enough so that u_n is positive. Then, for a $C > 0$ dependent on A only,*

$$\inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n^{\mathrm{OLS}}} \mathbb{P}(\beta \in \widehat{C}_{\widehat{S}}) \geq 1 - \alpha - C(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3}) \tag{19}$$

and

$$\inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n^{\mathrm{OLS}}} \mathbb{P}(\beta \in \widetilde{C}_{\widehat{S}}) \geq 1 - \alpha - C(\Delta_{n,1} + \Delta_{n,2} + \widetilde{\Delta}_{n,3}), \tag{20}$$

where

$$\Delta_{n,1} = \frac{1}{\sqrt{v}}\left(\frac{\overline{v}^2 k^2 (\log kn)^7)}{n}\right)^{1/6},$$

$$\Delta_{n,2} = \sqrt{\frac{k^5 \overline{v} (\log n)^2 \log k}{n} \max\left\{\frac{U_n^2}{u_n^7}, \frac{1}{u_n^4}\right\}},$$

$$\Delta_{n,3} = \left(\frac{U^2}{v}\right)^{1/3}\left(\overline{v}^2 k^3 \frac{\log n}{n}\log^4 k\right)^{1/6} (\mathrm{EIG}_n)^{1/3} \quad \text{and}$$

$$\widetilde{\Delta}_{n,3} = \min\left\{\Delta_{n,3}, \frac{U^2}{v}\overline{v}\frac{k^{5/2}}{u_n^3 u^2}\frac{\log n}{n}\log k\right\}$$

with

$$\mathrm{EIG}_n = \left(\frac{U}{u^{5/2}}\max\left\{\frac{U_n}{u_n^{7/2}}, \frac{1}{u_n^2}\right\}\right). \tag{21}$$

A few remarks are in order:

1. The coverage probability is affected by three factors: the term $\Delta_{n,1}$, which bounds the approximation error stemming from the high dimensional Berry–Esseen theorem (see Theorem 27); the term $\Delta_{n,2}$, which is a high probability bound on the size of the remainder term in the Taylor series expansion of $\beta_{\widehat{S}}$ around $\widehat{\beta}_{\widehat{S}}$ and can therefore be thought of as the price for the nonlinearity of the projection parameter, and the terms $\Delta_{n,3}$ and $\widetilde{\Delta}_{n,3}$, which are due to the fact that the covariance of the estimator is unknown and needs to be also estimated, leading to another source of error (the bootstrap procedure, described below, implicitly estimates this covariance).

2. In terms of dependence of k on n, all other things being equal, the remainder term $\Delta_{2,n}$ exhibit the worst rate, as it constrains k to be of smaller order than $n^{1/5}$ in order to guarantee asymptotic coverage of $\widehat{C}_{\widehat{S}}$. This same term also contains the worst dependence on u, the uniform bound on the smallest eigenvalue of all covariance matrices of the form Σ_S, for $S \subset \{1, \ldots, d\}$ with $0 < S \leq k$. On the other hand, under mild moment assumptions, the term k^2 in $\Delta_{n,1}$ can be eliminated. However, the dependence of the rates on the dimension and on the minimal eigenvalue is overall quite poor. While this phenomenon is, to an extent, unavoidable, we do not make any claim as to the sharpness of our bounds.

3. The coverage rates obtained for the LOCO and prediction parameters below in Section 2.2 are significantly faster then the ones for the projection parameters, and hold under less restrictions on the class of data generating distributions. We regard this as another reason to prefer the LOCO parameters.

4. As a function of sample size, there is a term of order $n^{-1/6}$ in $\Delta_{1,n}$ and $\Delta_{3,n}$. The exponent $1/6$ comes from the Berry–Esseen bound in Section 3. Chernozhukov et al. [22] conjecture that this rate is optimal for high-dimensional central limit theorems. Their conjecture is based on the lower bound result in [11]. If their conjecture is true, then this is best rate that can be hoped for in general.

5. The rates are slower than the rate obtained in the central limit theorem given in [53] for robust regression estimators. A reason for such discrepancy is that [53] assumes, among the other things, that the linear model is correct. In this case, the least squares estimators is conditionally unbiased. Without the assumption of model correctness, there is a substantial bias.

We now consider the accuracy of the confidence set given by the hyper-rectangle $\widetilde{C}_{\widehat{S}}$ from equation (17) by deriving an upper bound on the length of the largest side of $\max_{j \in \widehat{S}} \widetilde{C}(j)$. Similar rates can be obtained for the length of the sides of the hypercube confidence set $\widehat{C}_{\widehat{S}}$ given in (15).

COROLLARY 3. *With probability at least* $1 - \frac{2}{n}$, *the maximal length of the sides of the hyperrectangle* $\widetilde{C}_{\widehat{S}}$ *is bounded by*

$$C \sqrt{\frac{\log k}{n} \left(\mathrm{EIG}_n \overline{v} \sqrt{\frac{k^3 \log n}{n}} + \frac{U^2}{u^5} \overline{v} \right)}$$

for a constant $C > 0$ depending on A only, uniformly over all $P \in \mathcal{P}_n^{\mathrm{OLS}}$, where EIG_n *is as in* (21).

Confidence sets for the projection parameters: *The bootstrap*. The confidence set in (15) based on the Normal approximation require the evaluation of both the matrix $\widehat{\Gamma}_{\widehat{S}}$ and the quantile $\widehat{t}_\alpha$ in (16), which may be computationally inconvenient. Similarly, the hyperrectangle (17) requires computing the diagonal entries in $\widehat{\Gamma}_{\widehat{S}}$. Below we show that the paired bootstrap can be deployed to construct analogous confidence sets, centered at $\widehat{\beta}_{\widehat{S}}$, without knowledge of $\widehat{\Gamma}_{\widehat{S}}$. Throughout, by the bootstrap distribution we mean the empirical probability measure associated to the subsample $\mathcal{D}_{2,n}$ and conditionally on $\mathcal{D}_{1,n}$ and the outcome of the sample splitting procedure. We let $\widehat{\beta}^*_{\widehat{S}}$ denote the estimator of the projection parameters $\beta_{\widehat{S}}$ arising from an i.i.d. sample of size n drawn from the bootstrap distribution.

For a given $\alpha \in (0, 1)$, let $\widehat{t}^*_\alpha$ be the smallest positive number such that

$$\mathbb{P}\big(\sqrt{n}\|\widehat{\beta}^*_{\widehat{S}} - \widehat{\beta}_{\widehat{S}}\| \le \widehat{t}^*_\alpha | \mathcal{D}_{2,n}\big) \ge 1 - \alpha.$$

Next, let $(\widetilde{t}^*_j,\ j \in \widehat{S})$ be such that

$$\mathbb{P}\big(\sqrt{n}|\widehat{\beta}^*_{\widehat{S}}(j) - \widehat{\beta}_{\widehat{S}}(j) \le \widetilde{t}^*_j, \forall j | \mathcal{D}_{2,n}\big) \ge 1 - \alpha.$$

By the union bound, each $\widetilde{t}^*_j$ can be chosen to be the largest positive number such that

$$\mathbb{P}\big(\sqrt{n}|\widehat{\beta}^*_{\widehat{S}}(j) - \widehat{\beta}_{\widehat{S}}(j) > \widetilde{t}^*_j, |\mathcal{D}_{2,n}\big) \le \frac{\alpha}{k}.$$

Consider the following two bootstrap confidence sets:

$$\begin{aligned} \widehat{C}^*_{\widehat{S}} &= \left\{\beta \in \mathbb{R}^{\widehat{S}} \colon \|\beta - \widehat{\beta}_{\widehat{S}}\|_\infty \le \frac{\widehat{t}^*_\alpha}{\sqrt{n}}\right\}, \\ \widetilde{C}^*_{\widehat{S}} &= \left\{\beta \in \mathbb{R}^{\widehat{S}} \colon |\beta(j) - \widehat{\beta}_{\widehat{S}}(j)| \le \frac{\widetilde{t}^*_j}{\sqrt{n}}, \forall j \in \widehat{S}\right\}. \end{aligned} \tag{22}$$

It is immediate that $\widehat{C}^*_{\widehat{S}}$ and $\widetilde{C}^*_{\widehat{S}}$ are just the bootstrap equivalent of the confidence sets of (15) and (17), respectively.

THEOREM 4. *Let*

$$v_n = v - K_{1,n}, \qquad \overline{v}_n = \overline{v} + K_{1,n}, \qquad u_n = u - K_{2,n} \quad \textit{and} \quad U_n = U + K_{2,n},$$

where

$$K_{1,n} = CA^2\sqrt{b\overline{v}\frac{\log b + \log n}{n}} \quad \textit{and} \quad K_{2,n} = CA\sqrt{kU\frac{\log k + \log n}{n}},$$

with $C = C(\eta) > 0$ the constant in (101). *Assume that n is large enough so that $v_n = v - K_{1,n}$ and $u_n = u - K_{2,n}$ are both positive. Then, for a constant $C = C(A) > 0$,*

$$\inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n^{\mathrm{OLS}}} \mathbb{P}\big(\beta_{\widehat{S}} \in C_{\widehat{S}}^*\big) \geq 1 - \alpha - C\big(\Delta_{n,1}^* + \Delta_{n,2}^* + \Delta_{n,3}\big), \tag{23}$$

where $C_{\widehat{S}}^*$ is either one of the bootstrap confidence sets in (22),

$$\Delta_{n,1}^* = \frac{1}{\sqrt{v_n}} \left(\frac{k^2 \overline{v}_n^2 (\log kn)^7)}{n} \right)^{1/6},$$

$$\Delta_{n,2}^* = \frac{U_n}{\sqrt{v_n}} \sqrt{\frac{k^5 \overline{v}_n (\log n)^2 \log k}{n} \max\left\{ \frac{U_n^2}{u_n^7}, \frac{1}{u_n^4} \right\}}$$

and $\Delta_{n,3}$ is as in Theorem 2.

The sparse case. Now we briefly discuss the case of sparse fitting where $k = O(1)$ so that the size of the selected model is not allowed to increase with n. In this case, things simplify considerably. The standard central limit theorem shows that

$$\sqrt{n}(\widehat{\beta} - \beta) \rightsquigarrow N(0, \Gamma),$$

where $\Gamma = \Sigma^{-1} \mathbb{E}[(Y - \beta^\top X)^2] \Sigma^{-1}$. Furthermore, Γ can be consistently estimated by the sandwich estimator $\widehat{\Gamma} = \widehat{\Sigma}^{-1} A \widehat{\Sigma}^{-1}$ where $A = n^{-1} \mathbb{X}^\top R \mathbb{X}$, $\mathbb{X}_{ij} = X_i(j)$, R is the $k \times k$ diagonal matrix with $R_{ii} = (Y_i - X_i^\top \widehat{\beta})^2$. By Slutsky's theorem, valid asymptotic confidence sets can be based on the Normal distribution with $\widehat{\Gamma}$ in place of Γ [15].

Nonetheless, even if k is kept bounded, the effect of d increasing may be nontrivial and the results of the previous section would be more appropriate. For instance, the parameter $u = \min_{|S| \leq k} \lambda_{\min}(\Sigma_S)$ may decrease as d grows even when k is fixed. Hence, the usual fixed k asymptotics may be misleading.

2.2. *LOCO parameters.* Now we turn to the LOCO parameter $\gamma_{\widehat{S}} \in \mathbb{R}^{\widehat{S}}$ of (2), where $\widehat{S}$ is the model selected on the first half of the data. In order to derive confidence sets for $\gamma_{\widehat{S}}$, we will assume that the data generating distribution belongs to the class $\mathcal{P}_n^{\mathrm{LOCO}}$ of all distributions on $\mathbb{R}^{d+1}$ supported on $[-A, A]^{d+1}$, for some fixed constant $A > 0$. Clearly, the class $\mathcal{P}_n^{\mathrm{LOCO}}$ is significantly larger then the class $\mathcal{P}_n^{\mathrm{OLS}}$ considered for the projection parameters.

Next, we wrote the vector of LOCO parameters as $\widehat{\gamma}_{\widehat{S}} = \frac{1}{n} \sum_{i \in \mathcal{I}_{2,n}} \delta_i$, where $(\delta_i, i \in \mathcal{I}_{2,n})$ are independent and identically distributed random vectors such that, for any $i \in \mathcal{I}_{2,n}$ and $j \in \widehat{S}$,

$$\delta_i(j) = \big|Y_i - \widehat{\beta}_{\widehat{S}(j)}^\top X_i\big(\widehat{S}(j)\big)\big| - \big|Y_i - \widehat{\beta}_{\widehat{S}}^\top X_i(\widehat{S})\big|, \tag{24}$$

with $X_i(\widehat{S})$ the subvector of X_i consisting only of the coordinates in $\widehat{S}$.

For technical reasons detailed in Supplement E, we redefine the vector of LOCO parameters $\gamma_{\widehat{S}}$ so that its jth coordinate is

$$\gamma_{\widehat{S}}(j) = \mathbb{E}_{X,Y,\xi_j}\big[\big|Y - t_\tau\big(\widehat{\beta}^\top_{\widehat{S}(j)} X_{\widehat{S}(j)}\big)\big| - \big|Y - t_\tau\big(\widehat{\beta}^\top_{\widehat{S}} X_{\widehat{S}}\big)\big| + \epsilon\xi(j)\big], \tag{25}$$

where $\epsilon > 0$ is a prespecified small number, $\xi = (\xi(j),\, j \in \widehat{S})$ is a random vector comprised of independent Uniform$(-1, 1)$, independent of the data, and t_τ is a threshold function

$$x \in \mathbb{R} \mapsto t_\tau(x) = \begin{cases} x & \text{if } |x| \le \tau, \\ \operatorname{sign}(x) & \text{otherwise.} \end{cases} \tag{26}$$

Accordingly, we redefine the estimator $\widehat{\gamma}_{\widehat{S}}$ of this modified LOCO parameters as

$$\widehat{\gamma}_{\widehat{S}} = \frac{1}{n} \sum_{i \in \mathcal{I}_{2,n}} \delta_i, \tag{27}$$

where the δ_is are random vector in $\mathbb{R}^{\widehat{S}}$ such that the jth coordinate of δ_i is

$$\big|Y_i - t_\tau\big(\widehat{\beta}^\top_{\widehat{S}(j)} X_i\big(\widehat{S}(j)\big)\big)\big| - \big|Y_i - t_\tau\big(Y_i - \widehat{\beta}^\top_{\widehat{S}} X_i(\widehat{S})\big)\big| + \epsilon\xi_i(j), \qquad j \in \widehat{S}.$$

For simplicity, we take ϵ and τ to be fixed but we will keep explicit track of these quantities in the constants.

We first establish a simple concentration bound of the modified LOCO parameters.

LEMMA 5.

$$\sup_{w_n \in \mathcal{W}_n} \sup_{P \in \mathcal{P}_n^{\mathrm{LOCO}}} \mathbb{P}\left(\|\widehat{\gamma}_{\widehat{S}} - \gamma_{\widehat{S}}\|_\infty \le \big(2(A+\tau) + \epsilon\big)\sqrt{2\frac{\log k + \log n}{n}} \right) \ge 1 - \frac{1}{n}.$$

We now construct confidence sets for $\gamma_{\widehat{S}}$. Just like we did with the projection parameters, we consider two types of methods: one based on Normal approximations and the other on the bootstrap.

Normal approximation. Obtaining high-dimensional Berry–Esseen bounds for $\widehat{\gamma}_{\widehat{S}}$ is nearly straightforward since, conditionally on $\mathcal{D}_{1,n}$ and the splitting, $\widehat{\gamma}_{\widehat{S}}$ is just a vector of averages of bounded and independent variables with nonvanishing variances. Thus, there is no need for a Taylor approximation and we can apply directly the results in [22]. In addition, we find that the accuracy of the confidence sets for this LOCO parameter is higher than for the projection parameters.

Similar to what we did in Section 2.1, we derive two approximate confidence sets: one is an L_∞ ball and the other is a hyperrectangle whose jth side length is proportional to the standard deviation of the jth coordinate of $\widehat{\gamma}_{\widehat{S}}$. Both sets are centered at $\widehat{\gamma}_{\widehat{S}}$.

Below, we let $\alpha \in (0, 1)$ be fixed and let

$$\widehat{\Sigma}_{\widehat{S}} = \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \widehat{\gamma}_{\widehat{S}})(\delta_i - \widehat{\gamma}_{\widehat{S}})^{\top}, \tag{28}$$

be the empirical covariance matrix of the δ_is. The first confidence set is the L_∞ ball

$$\widehat{D}_{\widehat{S}} = \{\gamma \in \mathbb{R}^k : \|\gamma - \widehat{\gamma}_{\widehat{S}}\|_\infty \le \widehat{t}_\alpha\}, \tag{29}$$

where $\widehat{t}_\alpha$ is such that

$$\mathbb{P}\big(\|Z_n\|_\infty \le \widehat{t}_\alpha\big) = 1 - \alpha,$$

with $Z_n \sim N(0, \widehat{\Sigma}_{\widehat{S}})$. The second confidence set we construct is instead the hyperrectangle

$$\widetilde{D}_{\widehat{S}} = \bigotimes_{j \in \widehat{S}} \widehat{D}(j), \tag{30}$$

where, for any $j \in \widehat{S}$, $\widehat{D}(j) = [\widehat{\gamma}_{\widehat{S}}(j) - \widehat{t}_{j,\alpha}, \widehat{\gamma}_{\widehat{S}}(j) + \widehat{t}_{j,\alpha}]$, with $\widehat{t}_{j,\alpha} = z_{\alpha/2k}\sqrt{\frac{\widehat{\Sigma}_{\widehat{S}}(j,j)}{n}}$.

The above confidence sets have the same form as the confidence sets for the projection parameters (63), (66). The key difference is that for the projection parameters we use the estimated covariance of the linear approximation to $\widehat{\beta}_{\widehat{S}}$, while for the LOCO parameter $\widehat{\gamma}_{\widehat{S}}$ we rely on the empirical covariance (28), which is a much simpler estimator to compute.

In the next result, we derive coverage rates for both confidence sets.

THEOREM 6. *There exists a universal constant $C > 0$ such that*

$$\inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n^{\mathrm{LOCO}}} \mathbb{P}(\gamma_{\widehat{S}} \in \widehat{D}_{\widehat{S}}) \ge 1 - \alpha - C(\mathrm{E}_{1,n} + \mathrm{E}_{2,n}) - \frac{1}{n} \tag{31}$$

and

$$\inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n^{\mathrm{LOCO}}} \mathbb{P}(\gamma_{\widehat{S}} \in \widetilde{D}_{\widehat{S}}) \ge 1 - \alpha - C(\mathrm{E}_{1,n} + \widetilde{\mathrm{E}}_{2,n}) - \frac{1}{n}, \tag{32}$$

where

$$\mathrm{E}_{1,n} = \frac{2(A+\tau)+\epsilon}{\epsilon}\left(\frac{(\log nk)^7}{n}\right)^{1/6}, \tag{33}$$

$$\mathrm{E}_{2,n} = \frac{N_n^{1/3}(2\log 2k)^{2/3}}{\underline{\epsilon}^{2/3}}, \tag{34}$$

$$\widetilde{\mathrm{E}}_{2,n} = \min\left\{\mathrm{E}_{2,n}, \frac{N_n z_{\alpha/(2k)}}{\epsilon^2}\big(\sqrt{2+\log(2k)}+2\big)\right\} \tag{35}$$

and

$$N_n = \big(2(A+\tau)+\epsilon\big)^2 \sqrt{\frac{4\log k + 2\log n}{n}} \tag{36}$$

and $\epsilon_n = \sqrt{\epsilon^2 - N_n}$.

REMARK. The term $\mathrm{E}_{1,n}$ quantifies the error in applying the high-dimensional normal approximation to $\widehat{\gamma}_{\widehat{S}} - \gamma_{\widehat{S}}$, given in [22]. The second error term $\mathrm{E}_{2,n}$ is due to the fact that $\Sigma_{\widehat{S}}$ is unknown and has to be estimated using the empirical covariance matrix $\widehat{\Sigma}_{\widehat{S}}$. To establish $\mathrm{E}_{2,n}$, we use the Gaussian comparison Theorem 28. We point out that the dependence in ϵ displayed in the term $\mathrm{E}_{2,n}$ above does not follow directly from Theorem 2.1 in [22]. It can be obtained by tracking constants and using Nazarov's inequality Theorem 27 in the Supplementary Material, Section J, for details.

COROLLARY 7. *With probability at least* $1 - \frac{1}{n}$, *the maximal length of the sides of the hyperrectangle* $\widetilde{C}_n$ *is bounded by*

$$C\big(2(A+\tau)+\epsilon\big)\sqrt{\frac{\log k}{n}\bigg(1 + \frac{(4\log k + 2\log n)^{1/2}}{n^{1/2}}\bigg)},$$

for a universal constant $C > 0$, *uniformly over all* $P \in \mathcal{P}_n^{\mathrm{LOCO}}$.

The bootstrap. We now demonstrate the coverage of the paired bootstrap version of the confidence set for $\gamma_{\widehat{S}}$ given above in (29). The bootstrap distribution is the empirical measure associated to the n triplets $\{(X_i, Y_i, \xi_i), i \in \mathcal{I}_{2,n}\}$ and conditionally on $\mathcal{D}_{1,n}$. Let $\widehat{\gamma}^*_{\widehat{S}}$ denote the estimator of the LOCO parameters (25) of the form (27) computed from an i.i.d. sample of size n drawn from the bootstrap distribution. Notice that $\mathbb{E}[\widehat{\gamma}^*_{\widehat{S}} | (X_i, Y_i, \xi_i), i \in \mathcal{I}_{2,n}] = \widehat{\gamma}_{\widehat{S}}$. For a given $\alpha \in (0,1)$, let $\widehat{t}^*_\alpha$ be the smallest positive number such that

$$\mathbb{P}\big(\sqrt{n}\|\widehat{\gamma}^*_{\widehat{S}} - \widehat{\gamma}_{\widehat{S}}\| \le \widehat{t}^*_\alpha | (X_i, Y_i, \xi_i), i \in \mathcal{I}_{2,n}\big) \ge 1 - \alpha.$$

Next, let $(\widetilde{t}^*_j,\ j \in \widehat{S})$ be such that

$$\mathbb{P}\big(\sqrt{n}|\widehat{\gamma}^*_{\widehat{S}}(j) - \widehat{\gamma}_{\widehat{S}}(j)| \le \widetilde{t}^*_j, \forall j | (X_i, Y_i, \xi_i), i \in \mathcal{I}_{2,n}\big) \ge 1 - \alpha.$$

In particular, using the union bound, each $\widetilde{t}^*_j$ can be chosen to be the largest positive number such that

$$\mathbb{P}\big(\sqrt{n}|\widehat{\gamma}^*_{\widehat{S}}(j) - \widehat{\gamma}_{\widehat{S}}(j)| > \widetilde{t}^*_j | (X_i, Y_i, \xi_i), i \in \mathcal{I}_{2,n}\big) \le \frac{\alpha}{k}.$$

Consider the following two bootstrap confidence sets:

$$\begin{aligned} \widehat{D}^*_{\widehat{S}} &= \bigg\{\gamma \in \mathbb{R}^{\widehat{S}} \colon \|\gamma - \widehat{\gamma}_{\widehat{S}}\|_\infty \le \frac{\widehat{t}^*_\alpha}{\sqrt{n}}\bigg\} \quad \text{and} \\ \widetilde{D}^*_{\widehat{S}} &= \bigg\{\gamma \in \mathbb{R}^{\widehat{S}} \colon |\gamma_j - \widehat{\gamma}_{\widehat{S}}(j)| \le \frac{\widetilde{t}^*_j}{\sqrt{n}}, \forall j\bigg\}. \end{aligned} \tag{37}$$

THEOREM 8. *Using the same notation as in Theorem 6, assume that n is large enough so that* $\epsilon_n = \sqrt{\epsilon^2 - N_n}$ *is positive. Then there exists a universal constant* $C > 0$ *such that the coverage of both confidence sets in (37) is at least*

$$1 - \alpha - C\left(\mathrm{E}_{1,n}^* + \mathrm{E}_{2,n} + \frac{1}{n}\right),$$

where

$$\mathrm{E}_{1,n}^* = \frac{2(A+\tau)+\epsilon_n}{\epsilon_n}\left(\frac{(\log nk)^7}{n}\right)^{1/6}.$$

2.3. *Median LOCO parameters.* For the median loco parameters $(\phi_{\widehat{S}}(j), j \in \widehat{S})$ given in (3), finite sample inference is relatively straightforward. In detail, for each $j \in \widehat{S}$ and $i \in \mathcal{I}_{2,n}$, recall the definition of $\delta_i(j)$ in (24) and let $\delta_{(1)}(j) \le \cdots \le \delta_{(n)}(j)$ be the corresponding order statistics. We will not impose any restrictions on the data generating distribution. In particular, for each $j \in \widehat{S}$, the median of $\delta_i(j)$ needs not be unique. Consider the interval

$$E_j = \big[\delta_{(l)}(j), \delta_{(u)}(j)\big],$$

where

$$l = \left\lceil \frac{n}{2} - \sqrt{\frac{n}{2}\log\left(\frac{2k}{\alpha}\right)} \right\rceil \quad \text{and} \quad u = \left\lfloor \frac{n}{2} + \sqrt{\frac{n}{2}\log\left(\frac{2k}{\alpha}\right)} \right\rfloor, \tag{38}$$

and construct the hypercube

$$\widehat{E}_{\widehat{S}} = \bigotimes_{j \in \widehat{S}}^{n} E_j. \tag{39}$$

Then standard results about confidence sets for medians using order statistics, along with a union bound, imply that $\widehat{E}_{\widehat{S}}$ is a $1 - \alpha$ confidence set for the median LOCO parameters, uniformly over the class $\mathcal{P}_n$ of all distributions for (X, Y).

PROPOSITION 9. *For every n,*

$$\inf_{w_n \in \mathcal{W}_n} \inf_{P \in \mathcal{P}_n} \mathbb{P}(\phi_{\widehat{S}} \in \widehat{E}_{\widehat{S}}) \ge 1 - \alpha. \tag{40}$$

Of course, if the median of $\delta_i(j)$ is not unique, the length of the corresponding confidence interval does not shrink as n increases. But if the median is unique for each $j \in \widehat{S}$, and under addition smoothness conditions, we obtain that the maximal length the side of the confidence rectangle $\widehat{E}_{\widehat{S}}$ is of order $O(\sqrt{\frac{\log k + \log n}{n}})$, with high probability.

THEOREM 10. *Suppose that there exists positive numbers M and η such that, for each $j \in \widehat{S}$, the cumulative distribution function of each $\delta_i(j)$ is differentiable with derivative no smaller than M at all points at a distance no larger than η from its (unique) median. Then, for all n for which*

$$\frac{1}{n} + \sqrt{\frac{1}{2n}\log\left(\frac{2k}{\alpha}\right)} + \sqrt{\frac{\log 2kn}{2n}} \leq \eta M,$$

the sides of $\widehat{E}_{\widehat{S}}$ have length uniformly bounded by

$$\frac{2}{M}\left(\frac{1}{n} + \sqrt{\frac{1}{2n}\log\left(\frac{2k}{\alpha}\right)} + \sqrt{\frac{\log 2kn}{2n}}\right),$$

with probability at least $1 - \frac{1}{n}$.

2.4. *Prediction error.* To construct a confidence interval for the future prediction error parameter $\rho_{\widehat{S}}$, consider the set

$$\widehat{F}_{\widehat{S}} = [\widehat{\rho}_S - z_{\alpha/2}s_n/\sqrt{n}, \widehat{\rho}_S + z_{\alpha/2}s_n/\sqrt{n}], \tag{41}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ upper quantile of a standard normal distribution,

$$\widehat{\rho}_{\widehat{S}} = \frac{1}{n}\sum_{i\in\mathcal{I}_2} A_i, \qquad s_n^2 = \frac{1}{n}\sum_{i\in\mathcal{I}_2}(A_i - \widehat{\rho}_{\widehat{S}})^2 \quad \text{and}$$

$$A_i = \left|Y_i - \widehat{\beta}_{\widehat{S}}^{\top} X_i(\widehat{S})\right| \qquad \forall i \in \mathcal{I}_{2,n}.$$

For any P, let $\sigma_n^2 = \sigma_n^2(P) = \mathrm{Var}_P(A_1)$ and $\mu_{3,n} = \mu_{3,n}(P) = \mathbb{E}_P[|A_1 - \mathbb{E}_P[A_1]|^3]$. Then one may hope that a direct application of the one-dimensional Berry–Esseen theorem (see, e.g., Theorem 1.1, Chapter 11, in [60]), would yield that

$$\inf_{w_n\in\mathcal{W}_n} \mathbb{P}(\rho_{\widehat{S}} \in \widehat{F}_{\widehat{S}}) \geq 1 - \alpha - O\left(\frac{\mu_{3,n}}{\sigma_n\sqrt{n}}\right).$$

The asymptotic behavior of the last term on the right-hand side depends on how σ_n and $\mu_{3,n}$ scale with n (assuming they are well-defined), and in general cannot be controlled uniformly well over data generating distributions and procedures. In order to obtain uniform coverage guarantees, we slightly modify our target for inference, similarly to the way we dealt with the LOCO parameters in Section 2.2. Thus, we redefine the prediction parameter to be

$$\rho_{\widehat{S}} = \mathbb{E}\left[\left|Y - t_\tau\left(\widehat{\beta}_{\widehat{S}}^{\top} X(\widehat{S})\right)\right| + \epsilon\xi\right],$$

where t_τ is the the threshold function in (26) and ξ an independent noise variate uniformly distributed on $[-1, 1]$. The positive parameters τ and ϵ are chosen to ensure that the variance of the A_i's does not vanish and that their third moment

does not explode as n grows. Indeed, with this modification, we can ensure that $\sigma_n^2 \geq \epsilon^2$ and $\mu_{3,n} \leq (A+\tau+\epsilon)^3$ uniformly in n (and $s_n \leq 4(A+\tau+\epsilon)^2$, almost surely). Of course, we may let τ and ϵ change with n in a controlled manner. But for fixed choices of τ and ϵ, it is easy to see that the coverage of the interval (41) is

$$\inf_{w_n \in \mathcal{W}_n} \mathbb{P}(\rho_{\widehat{S}} \in \widehat{F}_{\widehat{S}}) \geq 1 - \alpha - C\left(\frac{1}{\sqrt{n}}\right),$$

for all data generating distributions, where C is a constant dependent only on A, τ and ϵ. Furthermore, the length of the confidence interval has parametric rate $O(\frac{1}{\sqrt{n}})$.

3. Prediction/accuracy tradeoff: Comparing splitting to uniform inference. There is a price to pay for sample splitting: the selected model may be less accurate because only part of the data are used to select the model. Thus, splitting creates gains in accuracy and robustness for inference but with some loss of prediction accuracy. We call this the *inference-prediction tradeoff*. In this section, we study this phenomenon by comparing splitting with uniform inference (defined below). We use uniform inference for the comparison since this is the any other method we know of that achieves (7). We study this use with a simple model where it is feasible to compare splitting with uniform inference. We will focus on the *many means problem* which is similar to regression with a balanced, orthogonal design. The data are $Y_1, \ldots, Y_{2n} \sim P$ where $Y_i \in \mathbb{R}^D$. Let $\beta = (\beta(1), \ldots, \beta(D))$ where $\beta(j) = \mathbb{E}[Y_i(j)]$. In this section, the model $\mathcal{P}_n$ is the set of probability distributions on $\mathbb{R}^D$ such that $\max_j \mathbb{E}|Y(j)|^3 < C$ and $\min_j \mathrm{Var}(Y(j)) > c$ for some positive C and c, which do not change with n or D (these assumptions could of course be easily relaxed). Below, we will only track the dependence on D and n and will use the notation $\preceq$ to denote inequality up to constants.

To mimic forward stepwise regression—where we would choose a covariate to maximize correlation with the outcome—we consider choosing j to maximize the mean. Specifically, we take

$$\widehat{S} \equiv w(Y_1, \ldots, Y_{2n}) = \operatorname*{argmax}_j \overline{Y}(j), \tag{42}$$

where $\overline{Y}(j) = (1/2n)\sum_{i=1}^{2n} Y_i(j)$. Our goal is to infer the random parameter $\beta_{\widehat{S}}$. The number of models is D. In forward stepwise regression with k steps and d covariates, the number of models is $D = d^k$. So the reader is invited to think of D as being very large. We will compare splitting versus nonsplitting with respect to three goals: estimation, inference and prediction accuracy.

Splitting: In this case, we take Let $\mathcal{D}_{1,n} = \{i : 1 \leq i \leq n\}$ and $\mathcal{D}_{2,n} = \{i : n+1 \leq i \leq 2n\}$. Then

$$\widehat{S} \equiv w(Y_1, \ldots, Y_n) = \operatorname*{argmax}_j \overline{Y}(j), \tag{43}$$

where $\overline{Y}(j) = (1/n)\sum_{i=1}^{n} Y_i(j)$. The point estimate and confidence interval for the random parameter $\beta_{\widehat{S}}$ are

$$\widehat{\beta}_{\widehat{S}} = \frac{1}{n}\sum_{i=n+1}^{2n} Y_i(\widehat{S})$$

and

$$\widehat{C}_{\widehat{S}} = [\widehat{\beta}_{\widehat{S}} - sz_{\alpha/2}/\sqrt{n}, \widehat{\beta}_{\widehat{S}} + sz_{\alpha/2}/\sqrt{n}],$$

where $s^2 = n^{-1}\sum_{i=n+1}^{2n}(Y_i(\widehat{S}) - \widehat{\beta}_{\widehat{S}})^2$.

Uniform inference (nonsplitting). By "nonsplitting," we mean that the selection rule and estimator are invariant under permutations of the data. In particular, we consider uniform inference which is defined as follows. Let $\widehat{\beta}(s) = (2n)^{-1}\sum_i Y_i(s)$ be the average over all the observations. Let $\widehat{S} = \operatorname{argmax}_s \widehat{\beta}(s)$. Our point estimate is $\widehat{\beta}_{\widehat{S}} \equiv \widehat{\beta}(\widehat{S})$. Now define

$$F_n(t) = \mathbb{P}\Big(\sup_s \sqrt{2n}|\widehat{\beta}(s) - \beta(s)| \le t\Big).$$

We can consistently estimate F_n by the bootstrap:

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sup_s \sqrt{2n}|\widehat{\beta}^*(s) - \widehat{\beta}(s)| \le t | Y_1, \ldots, Y_{2n}\Big).$$

A valid confidence set for β is $R = \{\beta : \|\beta - \widehat{\beta}\|_\infty \le t/\sqrt{2n}\}$ where $t = \widehat{F}_n^{-1}(1-\alpha)$. Because this is uniform over all possible models (i.e., over all s), it also defines a valid confidence interval for a randomly selected coordinate. In particular, we can define

$$\widehat{C}_{\widehat{S}} = [\widehat{\beta}_{\widehat{S}} - t/\sqrt{2n}, \widehat{\beta}_{\widehat{S}} + t/\sqrt{2n}].$$

Both confidence intervals satisfy (7).

We now compare $\widehat{\beta}_{\widehat{S}}$ and $\widehat{C}_{\widehat{S}}$ for both the splitting and nonsplitting procedures. The reader should keep in mind that, in general, $\widehat{S}$ might be different between the two procedures, and hence $\beta_{\widehat{S}}$ may be different.

Estimation. First, we consider estimation accuracy.

LEMMA 11. *For the splitting estimator,*

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}|\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}}| \preceq n^{-1/2}. \tag{44}$$

For nonsplitting, we have

$$\inf_{\widehat{\beta}} \sup_{P\in\mathcal{P}_n} \mathbb{E}|\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}}| \succeq \frac{(\log D)^{1/4}}{\sqrt{n}}. \tag{45}$$

Thus, the splitting estimator converges at a $n^{-1/2}$ rate. In fact, this bound holds for any selection rule. In contrast, nonsplitting estimators have a slower rate, even with the added assumption of Normality.

Inference. Now we turn to inference. For splitting, we use the usual Normal interval $\widehat{C}_{\widehat{S}} = [\widehat{\beta}_{\widehat{S}} - z_\alpha s/\sqrt{n}, \widehat{\beta}_{\widehat{S}} + z_\alpha s/\sqrt{n}]$ where s^2 is the sample variance from $\mathcal{D}_{2,n}$. We then have, as a direct application of the one-dimensional Berry–Esseen theorem the following.

LEMMA 12. *Let $\widehat{C}_{\widehat{S}}$ be the splitting-based confidence set. Then*

$$\inf_{P\in\mathcal{P}_n} \mathbb{P}(\beta_{\widehat{S}} \in \widehat{C}_{\widehat{S}}) = 1 - \alpha - \frac{c}{\sqrt{n}} \tag{46}$$

for some c. Also,

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}\big[\nu(\widehat{C}_{\widehat{S}})\big] \preceq n^{-1/2}, \tag{47}$$

where ν is Lebesgue measure. More generally,

$$\inf_{w\in\mathcal{W}_n} \inf_{P\in\mathcal{P}_n} \mathbb{P}(\beta_{\widehat{S}} \in \widehat{C}_{\widehat{S}}) = 1 - \alpha - \frac{c}{\sqrt{n}} \tag{48}$$

for some fixed $c > 0$, and

$$\sup_{w\in\mathcal{W}_n} \sup_{P\in\mathcal{P}_n} \mathbb{E}\big[\nu(\widehat{C}_{\widehat{S}})\big] \preceq n^{-1/2}. \tag{49}$$

LEMMA 13. *Let $\widehat{C}_{\widehat{S}}$ be the uniform confidence set. Then*

$$\inf_{P\in\mathcal{P}_n} \mathbb{P}(\beta_{\widehat{S}} \in \widehat{C}_{\widehat{S}}) = 1 - \alpha - \left(\frac{c(\log D)^7}{n}\right)^{1/6} \tag{50}$$

for some fixed > 0. Also,

$$\sup_{P\in\mathcal{P}_{2n}} \mathbb{E}\big[\nu(\widehat{C}_{\widehat{S}})\big] \succeq \sqrt{\frac{\log D}{n}}. \tag{51}$$

The proof is a straightforward application of results in [20, 22]. We thus see that the splitting method has better coverage and narrower intervals.

Can we estimate the law of $\widehat{\beta}(\widehat{S})$? An alternative nonsplitting method to uniform inference is to estimate the law F_{2n} of $\sqrt{2n}(\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}})$. Here, we show that the law of $\sqrt{2n}(\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}})$ cannot be consistently estimated even if we assume that the data are Normally distributed and even if D is fixed (not growing with n). This was shown for fixed population parameters in [35]. We adapt their proof to the random parameter case in the following lemma.

LEMMA 14. *Suppose that* $Y_1, \ldots, Y_{2n} \overset{i.i.d.}{\sim} N(\beta, I_D)$, *for some* $\beta \in \mathbb{R}^D$, *and let* $\psi_n(\beta) = \mathbb{P}(\sqrt{2n}(\widehat{\beta}_{\widehat{S}} - \beta_{\widehat{S}}) \leq t)$. *There is no locally uniformly consistent estimator of* $\psi_n(\beta)$.

Prediction accuracy. Now we discuss prediction accuracy which is where splitting pays a price. The idea is to identify a population quantity θ that model selection is implicitly targeting and compare splitting versus nonsplitting in terms of how well they estimate θ. The purpose of model selection in regression is to choose a model with low prediction error. So, in regression, we might take θ to be the prediction risk of the best linear model with k terms. In our many-means model, a natural analog of this is the parameter $\theta = \max_j \beta(j)$.

We have the following lower bound, which applies to all estimators, including the ones based on sample splitting.

LEMMA 15. *Let* $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\beta, I_D)$, *for some* $\beta \in \mathbb{R}^D$. *Let* $\theta = \max_j \beta(j)$. *Then*

$$\inf_{\widehat{\theta}} \sup_{\beta} E\big[(\widehat{\theta} - \theta)^2\big] \geq \frac{2 \log D}{n}.$$

To understand the implications of this result, let us write

$$\widehat{\beta}(S) - \theta = \underbrace{\widehat{\beta}(S) - \beta(S)}_{L_1} + \underbrace{\beta(S) - \theta}_{L_2}. \tag{52}$$

The first term, L_1, is the focus of most research on post-selection inference. We have seen it is small for splitting and large for nonsplitting. The second term takes into account the variability due to model selection which is often ignored. Because L_1 is of order $n^{-1/2}$ for splitting, and the because the sum is of order $\sqrt{\log D/n}$ it follows that splitting must, at least in some cases, pay a price by have L_2 large. In regression, this would correspond to the fact that, in some cases, splitting leads to models with lower predictive accuracy.

To summarize: splitting gives more precise estimates and coverage for the selected parameter than nonsplitting (uniform) inference. But the two approaches can be estimating different parameters. This manifests itself by the fact that splitting can lead to less precise estimates of the population parameter θ. In the regression setting, this would correspond to the fact that splitting the data can lead to selecting models with poorer prediction accuracy.

4. Berry–Esseen bounds for nonlinear parameters with increasing dimension. The results in this paper depend on a Berry–Esseen bound for regression with possibly increasing dimension. In this section, there is no model selection or splitting. We set $d = k$ and $S = \{1, \ldots, k\}$ where $k < n$ and k can increase with n.

These results will be applied after model selection and sample splitting. Existing Berry–Esseen results for nonlinear parameters are given in [1–3, 19, 52, 59]. Our results are in the same spirit but we keep careful track of the effect of dimension and the eigenvalues of Σ, while leveraging results from [20, 22] on high-dimensional central limit theorems for simple convex sets.

We derive a general result on the accuracy of the Normal approximation over hyperrectangles for nonlinear parameters. We make use of three findings from [21, 22] and [49]: the Gaussian anticoncentration theorem, the high-dimensional central limit theorem for sparely convex sets and the Gaussian comparison theorem, given in Supplement J as Theorems 26, 27 and 28, respectively. (We restate these results in slightly different forms than they appear in the original papers. We do this because we need to keep track of certain constants that affect our results.)

Let $W_1, \dots, W_n$ be an independent sample from a distribution P on $\mathbb{R}^b$ belonging to the class $\mathcal{P}_n$ of probability distribution supported on a subset of $[-A, A]^b$, for some fixed $A > 0$ and such that

$$v = \inf_{P \in \mathcal{P}_n} \lambda_{\min}\big(V(P)\big) \quad \text{and} \quad \overline{v} = \sup_{P \in \mathcal{P}_n} \lambda_{\max}\big(V(P)\big) \geq 1,$$

where $V(P) = \mathbb{E}_P[(W_i - \psi)(W_i - \psi)^\top]$. We allow the class $\mathcal{P}_n$ to change with n, so that b, v and $\overline{v}$—but not A—are to be regarded as functions of n, although we do not express such dependence in our notation for ease of readability. Notice that, in the worse case, $\overline{v}$ can be of order b.

Let $g = (g_1, \dots, g_s)^\top : \mathbb{R}^b \to \mathbb{R}^s$ be a twice-continuously differentiable vector-valued function defined over an open, convex subset $\mathcal{S}_n$ of $[-A, A]^b$ such that, for all $P \in \mathcal{P}_n$, $\psi = \psi(P) = \mathbb{E}[W_1] \in \mathcal{S}_n$. Let $\widehat{\psi} = \widehat{\psi}(P) = \frac{1}{n}\sum_{i=1}^n W_i$ and assume that $\widehat{\psi} \in \mathcal{S}_n$ almost surely, for all $P \in \mathcal{P}_n$. Finally, set $\theta = g(\psi)$ and $\widehat{\theta} = g(\widehat{\psi})$. For any point $\psi \in \mathcal{S}_n$ and $j \in \{1, \dots, s\}$, we will write $G_j(\psi) \in \mathbb{R}^b$ and $H_j(\psi) \in \mathbb{R}^{b \times b}$ for the gradient and Hessian of g_j at ψ, respectively. We will set $G(\psi)$ to be the $s \times b$ Jacobian matrix whose jth row is $G_j^\top(\psi)$.

To derive a high-dimensional Berry–Esseen bound on $g(\psi) - g(\widehat{\psi})$, we will study its first-order Taylor approximation. Toward that end, we will require a uniform control over the size of the gradient and Hessian of g. Thus we set

$$B = \sup_{P \in \mathcal{P}_n} \max_{j=1,\dots,s} \big\|G_j\big(\psi(P)\big)\big\| \quad \text{and} \quad \overline{H} = \sup_{\psi \in \mathcal{S}_n} \max_{j=1,\dots,s} \big\|H_j(\psi)\big\|_{\mathrm{op}}, \tag{53}$$

where $\|H_j(\psi)\|_{\mathrm{op}}$ is the operator norm.

The covariance matrix of the linear approximation of $g(\psi) - g(\widehat{\psi})$, which for any $P \in \mathcal{P}_n$, is given by

$$\Gamma = \Gamma\big(\psi(P), P\big) = G\big(\psi(P)\big)V(P)G\big(\psi(P)\big)^\top, \tag{54}$$

plays a crucial role in our analysis. In particular, our results will depend on the smallest variance of the linear approximation to $g(\psi) - g(\widehat{\psi})$:

$$\underline{\sigma}^2 = \inf_{P \in \mathcal{P}_n} \min_{j=1,\dots,s} G_j^\top\big(\psi(P)\big)V(P)G_j\big(\psi(P)\big). \tag{55}$$

With these definitions in place, we are now ready to prove the following high-dimensional Berry–Esseen bound.

THEOREM 16. *Assume that $W_1, \dots, W_n$ is an i.i.d. sample from some $P \in \mathcal{P}_n$ and let $Z_n \sim N(0, \Gamma)$. Then there exists a $C > 0$, dependent on A only, such that*

$$\sup_{P \in \mathcal{P}_n} \sup_{t>0} \left| \mathbb{P}\left(\sqrt{n}\|\widehat{\theta} - \theta\|_\infty \le t\right) - \mathbb{P}\left(\|Z_n\|_\infty \le t\right) \right| \le C(\Delta_{n,1} + \Delta_{n,2}), \tag{56}$$

where

$$\Delta_{n,1} = \frac{1}{\sqrt{v}} \left(\frac{\overline{v}^2 b (\log 2bn)^7}{n} \right)^{1/6}, \tag{57}$$

$$\Delta_{n,2} = \frac{1}{\underline{\sigma}} \sqrt{\frac{b^2 \overline{v}^2 \overline{H}^2 (\log n)^2 \log b}{n}}. \tag{58}$$

REMARK. Under the additional, very mild, moment bound condition $\mathbb{E}[(v^\top(\widehat{\psi} - \psi))^4] \le C'$ or a universal constant $C' > 0$, it is easy to show that the term b^2 drops out in the expression for $\Delta_{n,1}$, so that the dependence on b is only polylogarithmic.

Asymptotically honest confidence sets: *The normal approximation approach*. We now show how to use the high-dimensional central limit theorem Theorem 16 to construct asymptotically honest confidence sets for θ. We will first to obtain a consistent estimator of the covariance matrix $\Gamma = G(\psi)V(\psi)G(\psi)^\top$ of the linear approximation to $\widehat{\theta} - \theta$. In conventional fixed-dimension asymptotics, we would appeal to Slutzky's theorem and ignore the effect of replacing Γ with a consistent estimate. But in computing Berry–Esseen bounds with increasing dimension we may not discard the effect of estimating Γ. As we will see below, this extra step will bring an additional error term that must be accounted for. We will estimate Γ with the plug-in estimator

$$\widehat{\Gamma} = G(\widehat{\psi})\widehat{V}G(\widehat{\psi})^\top, \tag{59}$$

where $\widehat{V} = \frac{1}{n}\sum_{i=1}^n W_i W_i^\top - \widehat{\psi}\widehat{\psi}^\top$ is the empirical covariance matrix. Below, we bound the elementwise difference between Γ and $\widehat{\Gamma}$. Although this is in general a fairly weak notion of consistency in covariance matrix estimation, it is all that is needed to apply the Gaussian comparison Theorem 28, which will allow us to extend the Berry–Esseen bound established in Theorem 16 to the case when Γ is estimated.

LEMMA 17. *Let*

$$\aleph_n = \max\left\{ \overline{H} B \overline{v} \sqrt{b \frac{\log n}{n}}, B^2 \sqrt{b\overline{v} \frac{\log b + \log n}{n}} \right\}. \tag{60}$$

There exists a $C > 0$ dependent on A only such that

$$\sup_{P\in\mathcal{P}_n} \mathbb{P}\Big(\max_{j,l} |\widehat{\Gamma}(j,l) - \Gamma(j,l)| \geq C\aleph_n\Big) \leq \frac{2}{n}. \tag{61}$$

Now we construct the confidence set. Let $Q = (Q(1), \ldots, Q(s))$ be i.i.d. standard Normal variables, independent of the data. Let $\widehat{Z} = \widehat{\Gamma}^{1/2} Q$ and define $\widehat{t}_\alpha$ by

$$\mathbb{P}(\|\widehat{Z}\|_\infty > \widehat{t}_\alpha | \widehat{\Gamma}) = \alpha. \tag{62}$$

Finally, let

$$\widehat{C}_n = \left\{\theta \in \mathbb{R}^s : \|\theta - \widehat{\theta}\|_\infty \leq \frac{\widehat{t}_\alpha}{\sqrt{n}}\right\}. \tag{63}$$

THEOREM 18. *There exists a $C > 0$, dependent only on A, such that*

$$\inf_{P\in\mathcal{P}} \mathbb{P}(\theta \in \widehat{C}_n) = 1 - \alpha - C\left(\Delta_{n,1} + \Delta_{n,2} + \Delta_{n,3} + \frac{1}{n}\right), \tag{64}$$

where

$$\Delta_{n,3} = \frac{\aleph_n^{1/3}(2\log 2s)^{2/3}}{\underline{\sigma}^{2/3}}. \tag{65}$$

In addition to L_∞ balls, we can also construct hyperrectangle confidence sets, with side lengths proportional to the standard errors of the projection parameters. In detail, we define

$$\widetilde{C}_n = \bigotimes_{j\in S} C(j), \tag{66}$$

where

$$C(j) = \left[\widehat{\beta}_S(j) - z_{\alpha/(2s)}\sqrt{\frac{\widehat{\Gamma}_n(j,j)}{n}}, \widehat{\beta}_S(j) + z_{\alpha/(2s)}\sqrt{\frac{\widehat{\Gamma}_n(j,j)}{n}}\right],$$

with $\widehat{\Gamma}$ given by (14) and $z_{\alpha/(2s)}$ the upper $1-\alpha/(2s)$ quantile of a standard normal variate. Notice that we use a Bonferroni correction to guarantee a nominal coverage of $1-\alpha$. Also, note that $z_{\alpha/(2s)} = O(\sqrt{\log s})$, for each fixed α. The coverage rate for this other confidence set is derived in the next result.

THEOREM 19. *Let*

$$\widetilde{\Delta}_{n,3} = \min\left\{\Delta_{3,n}, \frac{\aleph_n z_{\alpha/(2s)}}{\underline{\sigma}^2}\left(\sqrt{2 + \log(2s)} + 2\right)\right\}. \tag{67}$$

There exists a $C > 0$, dependent only on A, such that

$$\inf_{P\in\mathcal{P}_n} \mathbb{P}(\theta \in \widetilde{C}_n) \geq (1-\alpha) - C\left(\Delta_{n,1} + \Delta_{n,2} + \widetilde{\Delta}_{n,3} + \frac{1}{n}\right).$$

Asymptotically honest confidence sets: *The bootstrap approach*. To construct the confidence set (63), one has to compute the estimator $\widehat{\Gamma}$ and the quantile $\widehat{t}_\alpha$ in (62), which may be computationally inconvenient. Similarly, the hyper-rectangle (66) requires computing the diagonal entries in $\widehat{\Gamma}$.

Below we rely on the bootstrap to construct analogous confidence sets, centered at $\widehat{\theta}$, which do not need knowledge of $\widehat{\Gamma}$. We let $\widehat{\psi}^*$ denote the sample average of an i.i.d. sample of size n from the bootstrap distribution, which is the empirical measure associated to the sample $(W_1, \dots, W_n)$. We also let $\widehat{\theta}^* = g(\widehat{\psi}^*)$.

For a fixed $\alpha \in (0, 1)$, let $\widehat{t}^*_\alpha$ be the smallest positive number such that

$$\mathbb{P}\big(\sqrt{n}\|\widehat{\theta}^* - \widehat{\theta}\| \le \widehat{t}^*_\alpha | (W_1, \dots, W_n)\big) \ge 1 - \alpha$$

and let $(\widetilde{t}^*_j,\ j = 1, \dots, s)$ be such that

$$\mathbb{P}\big(\sqrt{n}|\widehat{\theta}^*(j) - \widehat{\theta}(j)| \le \widetilde{t}^*_j, \forall j | (W_1, \dots, W_n)\big) \ge 1 - \alpha.$$

By the union bound, each $\widetilde{t}^*_j$ can be chosen to be the largest positive number such that

$$\mathbb{P}\big(\sqrt{n}|\widehat{\theta}^*(j) - \widehat{\beta}(j)| > \widetilde{t}^*_j | (W_1, \dots, W_n)\big) \le \frac{\alpha}{s}.$$

Consider the following two bootstrap confidence sets:

$$\begin{aligned} \widehat{C}^*_n &= \left\{\theta \in \mathbb{R}^s \colon \|\theta - \widehat{\theta}\|_\infty \le \frac{\widehat{t}^*_\alpha}{\sqrt{n}}\right\}, \\ \widetilde{C}^*_n &= \left\{\theta \in \mathbb{R}^s \colon |\theta(j) - \widehat{\theta}(j)| \le \frac{\widetilde{t}^*_j}{\sqrt{n}}, \forall j \in \widehat{S}\right\}. \end{aligned} \tag{68}$$

THEOREM 20. *Assume the same conditions of Theorem* 16 *and that* $\widehat{\psi}$ *and* $\widehat{\psi}^*$ *belong to* $\mathcal{S}_n$ *almost surely. Suppose that* n *is large enough that the quantities* $\underline{\sigma}_n^2 = \underline{\sigma}^2 - C\aleph_n > 0$ *and* $v_n = v - C\urcorner_n$ *are positive, where* C *is the larger of the two constants in* (61) *and in* (101) *and*

$$\urcorner_n = \sqrt{b\overline{v}\frac{\log b + \log n}{n}}.$$

Also set $\overline{v}_n = \overline{v} + C\urcorner_n$. *Then, for a constant* C *depending only on* A,

$$\inf_{P \in \mathcal{P}_n} \mathbb{P}\big(\theta \in \widehat{C}^*_n\big) \ge 1 - \alpha - C\left(\Delta^*_{n,1} + \Delta^*_{n,2} + \Delta_{n,3} + \frac{1}{n}\right), \tag{69}$$

where

$$\Delta^*_{n,1} = \frac{1}{\sqrt{v_n}}\left(\frac{\overline{v}_n b(\log 2bn)^7}{n}\right)^{1/6}, \qquad \Delta^*_{n,2} = \frac{1}{\underline{\sigma}_n}\sqrt{\frac{b\overline{v}_n\overline{H}^2(\log n)^2 \log b}{n}},$$

and $\Delta_{n,3}$ *is given in* (65). *Similarly,*

$$\inf_{P \in \mathcal{P}_n} \mathbb{P}\big(\theta \in \widetilde{C}^*_n\big) \ge 1 - \alpha - C\left(\Delta^*_{n,1} + \Delta^*_{n,2} + \Delta_{n,3} + \frac{1}{n}\right). \tag{70}$$

5. Conclusions. In this paper, we have taken a modern look at inference based on sample splitting. We have also investigated the accuracy of Normal and bootstrap approximations and we have suggested new parameters for variable significance in regression.

Despite the fact that sample splitting is on old idea, there remain many open questions. For example, in this paper, we focused on a single split of the data. One could split the data many times and somehow combine the confidence sets. However, for each split we are essentially estimating a different (random) parameter. So currently, it is nor clear how to combine this information.

The bounds on coverage accuracy—which are of interest beyond sample splitting—are upper bounds. An important open question is to find lower bounds. Also, it is an open question whether we can improve the bootstrap rates. For example, the remainder term in the Taylor approximation of $\sqrt{n}(\widehat{\beta}(j) - \beta(j))$ is

$$\frac{1}{2n} \int \int \delta^\top H_j\big((1-t)\psi + t\widehat{\psi}\big)\delta \, dt,$$

where $\delta = \sqrt{n}(\widehat{\psi} - \psi)$. By approximating this quadratic term, it might be possible to correct the bootstrap distribution. Pouzo [54] has results for bootstrapping quadratic forms that could be useful here. In Supplement C, we see that a modified bootstrap, which we called the image bootstrap, has very good coverage accuracy even in high dimensions. Future work is needed to compute the resulting confidence set efficiently.

Finally, we remind the reader that we have taken an assumption-lean perspective. If there are reasons to believe in some parametric model then, of course, the distribution-free, sample splitting approach used in this paper will be suboptimal.

Acknowledgments. The authors are grateful to the AE and the reviewers for comments that led to substantial improvements on the presentation and the discovery of a mistake in the original version of the manuscript. We also thank Lukas Steinberger, Peter Buhlmann, Iosif Pinelis and Arun Kumar Kuchibhotla for helpful suggestions and Jing Lei and Ryan Tibshirani for comments on early drafts.

SUPPLEMENTARY MATERIAL

Supplement to "Bootstrapping and sample splitting for high-dimensional, assumption-lean inference" (DOI: 10.1214/18-AOS1784SUPP; .pdf). This supplement provides additional material, including numerical examples, comments on other approaches, an alternative bootstrap approach, and algorithmic statements of the studied procedures. The supplement also includes proofs of many of the results stated in this paper.

REFERENCES

[1] ANASTASIOU, A. and GAUNT, R. E. (2016). Multivariate normal approximation of the maximum likelihood estimator via the delta method. Preprint. Available at arXiv:1609.03970.

[2] ANASTASIOU, A. and LEY, C. (2015). New simpler bounds to assess the asymptotic normality of the maximum likelihood estimator. Preprint. Available at arXiv:1508.04948.
[3] ANASTASIOU, A. and REINERT, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli* **23** 191–218. MR3556771
[4] ANDREWS, D. W. K. and GUGGENBERGER, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica* **77** 721–762. MR2531360
[5] BACHOC, F., LEEB, H. and PÖTSCHER, B. M. (2014). Valid confidence intervals for post-model-selection predictors. Available at arXiv:1412.4605.
[6] BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2016). Uniformly valid confidence intervals post-model-selection. Available at arXiv:1611.01043.
[7] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876
[8] BARNARD, G. A. (1974). Discussion of "Cross-validatory choice and assessment of statistical predictions," by M. Stone. *J. Roy. Statist. Soc. Ser. B* 133–135.
[9] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. B. (2013). *Inference for High-Dimensional Sparse Econometric Models*. vol. 3 245–295. Cambridge Univ. Press.
[10] BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097
[11] BENTKUS, V. Y. (1985). Lower bounds for the rate of convergence in the central limit theorem in Banach spaces. *Lith. Math. J.* **25** 312–320.
[12] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122
[13] BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549
[14] BÜHLMANN, P. and VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Stat.* **9** 1449–1473. MR3367666
[15] BUJA, A., BERK, R., BROWN, L., GEORGE, E., PITKIN, E., TRASKIN, M., ZHAO, L. and ZHANG, K. (2015). Models as approximations—A conspiracy of random regressors and model deviations against classical inference in regression. *Statist. Sci.* **1460**.
[16] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: "model-X" knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878
[17] CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974
[18] CHATTERJEE, A. and LAHIRI, S. N. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41** 1232–1259. MR3113809
[19] CHEN, L. H. Y. and SHAO, Q.-M. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli* **13** 581–599. MR2331265
[20] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448
[21] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields* **162** 47–70. MR3350040
[22] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963
[23] COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189

[24] DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840

[25] DEZEURE, R., BÜHLMANN, P. and ZHANG, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719. MR3713586

[26] EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

[27] FARAWAY, J. J. (1995). Data splitting strategies for reducing the e ect of model selection on inference. Technical report, Citeseer.

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

[29] HARTIGAN, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303–1317. MR0261737

[30] HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. MR2041481

[31] HSU, D., KAKADE, S. M. and ZHANG, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.* **14** 569–600. MR3201956

[32] HURVICH, C. M. and TSAI, C. (1990). The impact of model selection on inference in linear regression. *Amer. Statist.* **44** 214–217.

[33] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[34] 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

[35] LEEB, H. and PÖTSCHER, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24** 338–376. MR2422862

[36] LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342

[37] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135

[38] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

[39] LOFTUS, J. R. and TAYLOR, J. E. (2015). Selective inference in regression models with groups of variables. Preprint. Available at arXiv:1511.01478.

[40] MARKOVIC, J. and TAYLOR, J. (2016). Bootstrap inference after using multiple queries for model selection. Available at arXiv:1612.07811.

[41] MARKOVIC, J., XIA, L. and TAYLOR, J. (2017). Comparison of prediction errors: Adaptive p-values after cross-validation. Available at arXiv:1703.06559.

[42] MEINSHAUSEN, N. (2015). Group bound: Confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 923–945. MR3414134

[43] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

[44] MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584

[45] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26, 41. MR3491120

[46] MILLER, A. J. (1990). *Subset Selection in Regression. Monographs on Statistics and Applied Probability* **40**. CRC Press, London. MR1072361

[47] MORAN, P. A. P. (1973). Dividing a sample into two parts. A statistical dilemma. *Sankhyā Ser. A* **35** 329–333. MR0518783

[48] MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics.* Addison-Wesley Series in Behavioral Science: Quantitative Methods.

[49] NAZAROV, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*. *Lecture Notes in Math*. **1807** 169–187. Springer, Berlin. MR2083397

[50] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[51] PICARD, R. R. and BERK, K. N. (1990). Data splitting. *Amer. Statist.* **44** 140–147.

[52] PINELIS, I. and MOLZON, R. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electron. J. Stat.* **10** 1001–1063. MR3486424

[53] PORTNOY, S. (1987). A central limit theorem applicable to robust regression estimators. *J. Multivariate Anal.* **22** 24–50. MR0890880

[54] POUZO, D. (2015). Bootstrap consistency for quadratic forms of sample averages with increasing dimension. *Electron. J. Stat.* **9** 3046–3097. MR3450756

[55] RINALDO, A., WASSERMAN, L. and G'SELL, M. (2019). Supplement to "Bootstrapping and sample splitting for high-dimensional, assumption-lean inference." DOI:10.1214/18-AOS1784SUPP.

[56] SHAH, R. D. and BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 113–135. MR3744714

[57] SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

[58] SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494. MR1224373

[59] SHAO, Q.-M., ZHANG, K. and ZHOU, W.-X. (2016). Stein's method for nonlinear statistics: A brief survey and recent progress. *J. Statist. Plann. Inference* **168** 68–89. MR3412222

[60] SHORACK, G. R. (2000). *Probability for Statisticians*. *Springer Texts in Statistics*. Springer, New York. MR1762415

[61] TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381

[62] 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

[63] 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

[64] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[65] WAGER, S., HASTIE, T. and EFRON, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **15** 1625–1651. MR3225243

[66] WASSERMAN, L. (2014). Discussion: "A significance test for the lasso" [MR3210970]. *Ann. Statist.* **42** 501–508. MR3210975

[67] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

[68] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

[69] ZHANG, X. and CHENG, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768. MR3671768

DEPARTMENT OF STATISTICS AND DATA SCIENCE
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: arinaldo@cmu.edu
larry@stat.cmu.edu
mgsell@cmu.edu