Statistics Surveys
Vol. 16 (2022) 86–136
ISSN: 1935-7516
https://doi.org/10.1214/22-SS135

Post-model-selection inference in linear regression models: An integrated review*

Dongliang Zhang†

Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA
e-mail: dzhang69@jhu.edu

Abbas Khalili

Department of Mathematics and Statistics, McGill University, Montréal, QC, Canada
e-mail: abbas.khalili@mcgill.ca

Masoud Asgharian

Department of Mathematics and Statistics, McGill University, Montréal, QC, Canada
e-mail: masoud.asgharian2@mcgill.ca

Abstract: The research on statistical inference after data-driven model selection can be traced as far back as Koopmans (1949). The intensive research on modern model selection methods for high-dimensional data over the past three decades revived the interest in statistical inference after model selection. In recent years, there has been a surge of articles on statistical inference after model selection and now a rather vast literature exists on this topic. Our manuscript aims at presenting a holistic review of post-model-selection inference in linear regression models, while also incorporating perspectives from high-dimensional inference in these models. We first give a simulated example motivating the necessity for valid statistical inference after model selection. We then provide theoretical insights explaining the phenomena observed in the example. This is done through a literature survey on the post-selection sampling distribution of regression parameter estimators and properties of coverage probabilities of naïve confidence intervals. Categorized according to two types of estimation targets, namely the population- and projection-based regression coefficients, we present a review of recent uncertainty assessment methods. We also discuss possible pros and cons for the confidence intervals constructed by different methods.

MSC2020 subject classifications: Primary 62F25; secondary 62J07.
Keywords and phrases: High-dimensional linear models, model selection, population- and projection-based regression coefficients, post-selection inference.

Received April 2021.

*Abbas Khalili is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-2020-05011), and Masoud Asgharian is supported by the Natural Science and Engineering Research Council of Canada (NSERC RGPIN-2018-05618).
†Corresponding author

Contents

1 Introduction 87
2 Notation and terminology 91
3 Motivating example 93
4 Post-selection sampling distribution of parameter estimators and properties of coverage probabilities of naïve confidence intervals 96
5 Statistical estimation targets 98
5.1 Population-based target 98
5.2 Projection-based target 99
6 Post-selection inferential methodologies 100
6.1 Universally valid post-selection inference (PoSI) 100
6.2 Exact post-selection inference (EPoSI) 103
6.3 EPoSI for sequential regression 107
6.4 Recent work 109
7 Statistical inference for population-based target: bias correction and sampling techniques 113
7.1 Bias-correction 113
7.2 Re-sampling 116
7.3 Optimization perspective 120
8 Simulation study 121
8.1 Data generation 121
8.2 Assessing metrics 122
8.3 Method, code and implementation 123
8.4 Analysis of simulation results 123
9 Conclusion 127
Acknowledgments 129
Supplementary Material 129
References 129

1. Introduction

In one of his earliest work, praised for being "arguably the most influential article" on statistical inference, Fisher describes three fundamental issues in data reduction (Stigler 2005, Fisher 1922), which are the problems of specification, estimation and distribution. According to his illustration, these three problems respectively refer to (1) a "choice of the mathematical form of the population", where we know "what parameters are required to specify the population from which the sample is drawn"; (2) estimation of "values of parameters of the hypothetical population"; (3) determination of "distribution of statistics derived in the sample" for precision assessment.

To solve the second and third problems, respectively of parameter estimation and statistical distribution assessment, Fisher elucidates the renowned *maximum likelihood method* which is the principal ingredient of most classical and modern model selection and estimation procedures. Fisher's likelihood theory,

however, relies on one important assumption pertinent to the first issue on model specification, that is a correct model which accurately characterizes the true data generating mechanism is known, except for certain parameter values, before valid parameter estimation and uncertainty assessment is executed. The implication is that the empirical data shall not be involved in the selection of a final model from a set of candidate models, regardless of whether the selected model is correct or wrong. Instead, the data shall be solely used to perform point estimation and construct relevant uncertainty measures. Moreover, under this assumption, to provide an explicit answer to the third problem, Fisher advocates the use of the well-known *p-value*, a concept introduced by Pearson (1900), as an assessment benchmark for evaluating statistical significance of result(s) of a scientific experiment. In addition, Fisher (1935) implies that the criterion that p-value less than a pre-specified value may indicate that "a phenomenon is experimentally demonstrable". In other words, p-value could be considered as an indicator for replicability of results, which is commonly regarded as a golden standard in science (Shapin & Schaffer 1985).

The foundation on which Fisher's maximum likelihood method and the subsequent adoption of p-values for precision assessment rests, however, is seldom satisfied in practice, which is "a quiet scandal in the statistical community" as phrased by Breiman (1992). In fact, the exact mathematical form of population is often unidentified beforehand. Usually, upon assuming a specific class of true data generating mechanism such as a linear or a generalized linear regression model, and identifying a set of potential predictor variables, a model for statistical inference is usually selected from a set of candidate models through either a well-defined, ill-defined or opaque data-dependent fashion (Berk et al. 2013). A data-driven selection of tuning parameters in obtaining penalized least-squares estimators further invalidates that assumption. The final model so selected, on which parameter estimation and uncertainty assessment is based, is not fixed as required by Fisher's view, but indeed random. In other words, even with the same model selection procedure, different realizations of the true data generating mechanism may well lead to different selected models which may deviate from the true model with non-negligible probability. Therefore, by integrating additional uncertainty from model selection into inferential process, the validity of statistical inference guaranteed by Fisher's likelihood theory becomes suspicious.

In addition to the aforementioned violation of Fisher's fundamental assumption, negligence of the multiple testing problems in the era of mass production of scientific publications further deepen the concern with respect to the validity of statistical inference in ensuring the replicability of scientific experiment, particularly on the legitimacy of p-values. This consideration has been highlighted across several scientific communities, for example behavioral genetics (Mann 1994) and genomics research (Lander & Kruglyak 1995), which culminated in the thought-provoking article by Ioannidis (2005), who boldly assert that "most claimed research findings are false". As such, to ensure the replicability of scientific discoveries, several initiatives have been launched within the last decade involving appeal for more cautious use or even abandonment of p-values. Dis-

missing these movements as "misguided attacks", Benjamini (2020) argue that (i) failure to ascertain the correct level of variability, and (ii) ignorance of the effect of selection on statistical inference are two crucial yet often neglected causes for the irreplicability of scientific results. He concludes that we do not need to refrain from counting on classic measures, such as p-values and confidence intervals, for assessing statistical significance. Rather, these assessment benchmarks have to be adjusted for valid statistical inference.

In view of the foregoing reasoning and perspectives from Ioannidis (2005) and Benjamini (2020) on replicability of results, given that the presumed condition of Fisher's likelihood method is infringed since one data set is simultaneously used for model selection and statistical inference, one would naturally ask what detrimental effects this violation might provoke on parameter estimation and precision assessment so that the resulting measures (without adjustments), such as p-values or confidence intervals, represent sources of concern across scientific communities. To answer this question, several important contributions have been accomplished over the past two decades, Pötscher (1991, 1995); Pötscher & NovÁk (1998); Leeb & Pötscher (2003, 2005, 2006, 2008); Kabaila (1995, 1998, 2005, 2009); Kabaila & Leeb (2006); Kabaila & Giri (2009) and Berk et al. (2009), among others. The aforementioned research recognize and explain the undesirable outcomes of distorted sampling distribution of regression parameter estimators obtained after model selection and reduced coverage probability of confidence regions that are constructed in a naïve fashion where the data-driven nature of a selected model is ignored. Furthermore, via the lens of replicability of results, the confidence regions so constructed may contain zero after necessary adjustment by accounting for model selection randomness, whereas they may not include zero without such adjustments. Therefore, classic uncertainty quantification measures formulated without considering additional layer of randomness from selection indeed contribute to the irreplicability of scientific results. As such, it is imperative to devise reliable uncertainty quantification measures to account for randomness inherited from data-dependent model selection procedures.

Recently, various proposals for constructing asymptotically valid p-values as well as marginal and simultaneous confidence intervals for the so-called population-based regression coefficients (see equation (1) for definition) in high-dimensional linear regression models have been suggested (Zhang & Zhang 2014, van der Geer et al. 2014, Javanmard & Montanari 2014*a*, Bühlmann 2013, Neykov et al. 2018, Ning & Liu 2017, 2014, Belloni et al. 2015, 2014, Belloni, Chernozhukov & Wei 2013, Wasserman & Roeder 2009, Meinshausen et al. 2009, Meinshausen & Bühlmann 2010, Shah & Samworth 2013, Khalili & Vidyashankar 2018, Chernozhukov et al. 2018, Belloni, Chernozhukov & Hansen 2013, Chatterjee & Lahiri 2011, 2013, Liu & Yu 2013, Dezeure et al. 2017*a*, Minnier et al. 2011, Dezeure et al. 2017*b*). Dezeure et al. (2015) provides a comprehensive review and numerical comparison of a wide spectrum of statistical inferential methods specifically designed for the population-based targets. They also introduced an `R` package `hdi` (*h*igh-*d*imensional *i*nference).

Considering randomness in estimation of regression coefficient(s) after model

selection, Berk et al. (2013) argue that the population-based regression coefficient(s) (referred to as target(s) from now on) may have interpretation drawbacks in some applications. They therefore present an alternative estimation target(s), known as the projection-based regression coefficient(s). Focusing on such targets, Berk et al. (2013) and Lee et al. (2016) propose two methods for constructing valid post-selection confidence intervals. They are respectively known as the post-selection inference (PoSI) and the exact post-selection inference (EPoSI) methods.

To the best of our knowledge, Leeb et al. (2015) is the only paper that provides a comparative study on methods of constructing post-selection confidence intervals for the projection-based targets. Specifically, these authors compared the naïve confidence intervals with those constructed using the PoSI approach proposed by Berk et al. (2013), where AIC, BIC and LASSO (Tibshirani 1996) are used for model selection. Their simulations indicate that the PoSI confidence intervals do deliver at least the desired minimal coverage probability for the projected coefficients, while naïve confidence intervals in general fail to do so. Interestingly, dramatic under-coverage is observed for both types of confidence intervals if the projection-based targets are replaced by the population-based regression coefficients. Such under-coverage could be explained by the fact that the PoSI confidence intervals are not designed for the population-based target.

Yet, the existing literature on post-model-selection (PMS) inference stops short on three different fronts. First, there is no integrated review of existing literature in post-selection inference. It is therefore an arduous task to grasp the essence of post-selection inference in a succinct and unified fashion. Second, the empirical studies by Leeb et al. (2015) fall short in: (1) comparing confidence intervals constructed using the Scheffé's (Scheffé 1959) and the EPoSI (Lee et al. 2016) approaches; (2) considering other important indicators such as the length of confidence intervals. Third, a collection of recent methodological advancement has not been presented in a unified fashion.

This manuscript aims at filling the aforementioned gaps and presents an integrated guide for the contemporary PMS inference. Particularly, using simulations, we compare naïve confidence intervals against those constructed by the Scheffé's method, the PoSI and the EPoSI approaches. In terms of the average coverage probabilities for the projection-based regression coefficients, we find that when conditioning on the selected submodel only, a mix of under- and satisfactory coverage is observed for the EPoSI confidence intervals while the PoSI confidence intervals consistently achieve over-coverage. In terms of the average length of confidence intervals, those constructed by the EPoSI method are in general shorter than the ones by the PoSI approach, reinforcing the conservative nature underpinning the PoSI method. On the other hand, when conditioning on both the selected submodel and corresponding sign vector of point estimator(s), similar phenomenon with respect to the average coverage probabilities is seen for the EPoSI and PoSI confidence intervals. However, in this case, the EPoSI confidence intervals can sometimes be much wider. We have observed through simulations that the EPoSI confidence intervals, obtained via conditioning on the selected submodel or both the selected submodel and corresponding sign

vector of parameter estimator(s), sometimes have infinite lengths. This has also been analytically discussed in the recent manuscript Kivaranovic & Leeb (2021).

The rest of the manuscript is organized as follows: in Section 2, we introduce notation and terminology. In Section 3, we provide a motivating example based on the linear regression model to give some tangible hints to the pitfalls of ignoring data-dependent model selection on statistical inference. In Section 4, we present a summary of key results in finite- and large-sample distributional properties of the PMS estimators as well as properties of coverage probabilities of naïve confidence intervals, serving as theoretical insights into the phenomena observed in Section 3. In Section 5, we introduce and discuss some merits and demerits of both population- and projection-based estimation targets. Focusing on the projection-based targets, we then provide a selective review of existing post-selection methods for uncertainty assessment in Section 6. Particular emphasis is placed on the construction of valid post-selection confidence intervals and p-values. This is followed by a concise presentation of various recent methodological development in post-model-selection inference. In Section 7, we provide an overview of statistical inference methods designed specifically for the population-based targets, even though they may not be considered as post-selection methods in the most strict sense. In Section 8, focusing on methods designed for projection-based regression coefficients, we present comparative simulations to evaluate different constructions of post-selection confidence intervals in linear regression models. Conclusions are given in Section 9.

2. Notation and terminology

Let $Y \in \mathbb{R}$ be the response variable and $x_j \in \mathbb{R}, j = 1, ..., p$, be the predictors (covariates). Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top \in \mathbb{R}^n$ be the response vector and $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$, where $\boldsymbol{x}_i^\top = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$, denote the $n \times p$ design matrix. For any d-dimensional vector $\boldsymbol{v} = (v_1, ..., v_d) \in \mathbb{R}^d$, we denote its ℓ_q-norm by $||\boldsymbol{v}||_q = \left(\sum_{j=1}^d |v_j|^q\right)^{1/q}$, for $1 \leq q < \infty$. In the case of Euclidean norm $||\cdot||_2$, we omit the subscript and write $||\cdot||$. In what follows, we mainly have either $d = p$ or n. Furthermore, $||v||_\infty = \max_{j=1,...,d} |v_j|$.

A linear regression model with independently and identically distributed (*i.i.d.*) errors is defined by

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}^0$, and $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_p^0)^\top$ denotes the true *population-based regression coefficient* (*target*), $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$ is the error vector with $\epsilon_i \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$, and ϵ_i's are independent of the design matrix $\boldsymbol{X}$.

Let the index set $\boldsymbol{M} = \{j_1, ..., j_m\} \subseteq \{1, ..., p\}$ denote a linear submodel which contains explanatory variables with indices in $\boldsymbol{M}$, where $|\boldsymbol{M}| = m \leq p$ denotes the size of submodel $\boldsymbol{M}$. Then we define $\boldsymbol{X}_{\boldsymbol{M}} \in \mathbb{R}^{n \times m}$ to be the $n \times m$ submatrix of the design matrix $\boldsymbol{X}$ with columns indexed by $\boldsymbol{M}$. The linear regression model corresponding to a submodel $\boldsymbol{M}$ is given by

$$\boldsymbol{Y} = \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\beta}_{\boldsymbol{M}} + \boldsymbol{\epsilon}. \tag{2}$$

If we set $\boldsymbol{M} = \{1, ..., p\}$ in (2), that is the index set of *all* the predictors, we obtain the linear regression model (1) with the true population-based regression coefficient $\boldsymbol{\beta}^0$. Indeed, this observation consists of one rudimentary aspect of the interpretation for $\boldsymbol{\beta}^0$, which is referred to as the full model interpretation of parameters by Berk et al. (2013). More specifically, the linear regression model in (1) is considered as the true data generating mechanism for the response vector $\boldsymbol{Y}$. As such, a causal connection between the response and the predictors is implicitly embedded in this interpretation. Under this setup, coefficient estimates for de-selected predictors are assigned to be zero, and the resulting submodel comprising only the selected predictors is merely the "computational compression"(*ibid.*, Appendix, B.1) of the data. While this interpretation has merits in explaining natural and physical scientific phenomena, Berk et al. (2013) point out that: (i) the idea of having a full model exhibiting a causal interpretation is controversial, and (ii) inferential complications may arise from this interpretation as well (Leeb & Pötscher 2005).

To avoid these contentious elements, as an alternative to the full model interpretation of parameters, Berk et al. (2013) propose the *submodel interpretation of parameters*. Instead of assuming a linear structure for the true model, one would give up the proposition of having a full model all together and only assume the existence of the true mean response vector $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ in model (1). Then, the so-called *projection-based target* corresponding to a given submodel $\boldsymbol{M}$ is defined as

$$\boldsymbol{b}_{\boldsymbol{M}} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^{|\boldsymbol{M}|}} \|\boldsymbol{\mu} - \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{b}\|^2 = (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1}\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{\mu}. \tag{3}$$

We will provide more discussion about the two estimation targets $\boldsymbol{\beta}^0$ and $\boldsymbol{b}_{\boldsymbol{M}}$ in Sections 5.1 and 5.2, respectively.

Assuming the design matrix $\boldsymbol{X}$ is fixed, we define $\widehat{\boldsymbol{M}}_n$ to be a data-dependent model selection procedure, a measurable function of $\boldsymbol{Y}$, given by $\widehat{\boldsymbol{M}}_n \colon \mathbb{R}^n \to \mathcal{M}_{\text{all}}$, where $\mathcal{M}_{\text{all}} = \{\boldsymbol{M} | \boldsymbol{M} \subseteq \{1, ..., p\},\ \text{rank}(\boldsymbol{X}_{\boldsymbol{M}}) = |\boldsymbol{M}|\}$, is the space of all possible full-rank submodels. Let $\boldsymbol{M}_{\text{true}} \in \mathcal{M}_{\text{all}}$ be the index set of the true data generating mechanism upon assuming such linear model exists. From now on, we use $\widehat{\boldsymbol{M}}$ to represent the output of the map $\widehat{\boldsymbol{M}}_n(\boldsymbol{Y})$, i.e. $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}_n(\boldsymbol{Y})$.

If a submodel $\boldsymbol{M}$ is given *a priori*, then the ordinary least-squares (OLS) estimator of $\boldsymbol{\beta}_{\boldsymbol{M}}$ in (2) is given by

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}} = (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1}\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{Y}. \tag{4}$$

The same estimator is also used for $\boldsymbol{b}_{\boldsymbol{M}}$ in (3), and to avoid introducing more notation, we use $\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ as its estimator. Components of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ are denoted by $\widehat{\beta}_{j \cdot \boldsymbol{M}}$, for all $j \in \boldsymbol{M}$.

Moreover, related to our motivating example in Section 3 based on (1), we denote $\widehat{\boldsymbol{\beta}}^{\text{Method}}_{\widehat{\boldsymbol{M}}}$ as the OLS estimator corresponding to a submodel $\widehat{\boldsymbol{M}}$ selected

by a "Method" considered below. Define

$$T_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}} = \frac{(\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}} - \beta_j^0)}{\text{sd}(\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}})} \quad \text{and} \quad T_{j\cdot\widehat{\boldsymbol{M}},0}^{\text{Method}} = \frac{\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}}{\text{sd}(\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}})}, \tag{5}$$

where $\text{sd}(\cdot)$ stands for standard deviation. The quantity $T_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}$ will be used for constructing naïve confidence intervals for β_j^0, and $T_{j\cdot\widehat{\boldsymbol{M}},0}^{\text{Method}}$ will be used for testing the null hypothesis $\beta_j^0 = 0$. Also, $T_{j\cdot\boldsymbol{M}}$ and $T_{j\cdot\boldsymbol{M},0}$ are respectively defined when replacing $\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}$ in $T_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}$ and $T_{j\cdot\widehat{\boldsymbol{M}},0}^{\text{Method}}$ by $\widehat{\beta}_{j\cdot\boldsymbol{M}}$, for any fixed submodel $\boldsymbol{M} \in \mathcal{M}_{\text{all}}$.

In addition, a penalized least-squares estimator of $\boldsymbol{\beta}^0$ in (1) is given by

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \text{J}_n(\boldsymbol{\beta};\boldsymbol{\lambda}) \right\}, \tag{6}$$

where $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^n (Y_i - \boldsymbol{x_i}^\top\boldsymbol{\beta})^2$, $\boldsymbol{\lambda}$ is the tuning parameter (possibly multi-dimensional) and $\text{J}_n(\boldsymbol{\beta};\boldsymbol{\lambda})$ is a penalty function. In the sequel, we consider these penalties:

1. LASSO: $\text{J}_n(\boldsymbol{\beta};\boldsymbol{\lambda}) = \lambda||\boldsymbol{\beta}||_1$, where $\boldsymbol{\lambda} = \lambda$;
2. Elastic net (Zou & Hastie 2005): $\text{J}_n(\boldsymbol{\beta};\boldsymbol{\lambda}) = \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||^2$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$;
3. SCAD (Fan & Li 2001): $\text{J}_n(\boldsymbol{\beta};\boldsymbol{\lambda}) = n\sum_{j=1}^p P(|\beta_j|;\boldsymbol{\lambda})$, where $P(|\beta_j|;\boldsymbol{\lambda})$ is given by

$$P(|\beta_j|;\boldsymbol{\lambda}) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \le \lambda; \\ -\frac{(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta_j| \le a\lambda; \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda. \end{cases}$$

   Here, $\boldsymbol{\lambda} = (\lambda, a)$, where $a \ge 2$.
4. Adaptive LASSO (AdaLASSO) (Zou 2006): $\text{J}_n(\boldsymbol{\beta};\boldsymbol{\lambda}) = \lambda||\boldsymbol{w} * \boldsymbol{\beta}||_1$, $\boldsymbol{\lambda} = \lambda$ and the data-dependent weights $\boldsymbol{w} = (w_1, ..., w_p)^\top$ is recommended to be chosen as $w_j = 1/|\widehat{\beta}_j|$ for some consistent estimator $\widehat{\beta}_j$ of β_j^0. Moreover, where $*$ represents elementwise multiplication of two vectors.

The books by Fan et al. (2020), Wainwright (2019), and Hastie et al. (2015) provide a comprehensive review on the above modern regularization techniques and related theories in the context of different statistical models, including linear and generalized linear regression models.

3. Motivating example

We now give some tangible insights into the unfavorable impacts of ignoring the data-dependent model selection on statistical inference. Through a simulated example based on the linear regression model (1), we achieve this by:

(a) plotting the empirical densities and the boxplots of $\widehat{\beta}^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}}$ against $\widehat{\beta}_{j\cdot\boldsymbol{M}_{\text{true}}}$ in FIG 1, where $j \in \boldsymbol{M}_{\text{true}} = \{1, 3, 4, 5, 7\}$. Our choices of "Method" to obtain $\widehat{\boldsymbol{M}}$ are the LASSO, elastic net and SCAD, where the required tuning parameters are chosen based on a 10-fold cross validation.

(b) plotting the empirical densities (FIG 2) and the boxplots (FIG 3) of $T^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}}$ and $T^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}},0}$ against their respective counterparts, $T_{j\cdot\boldsymbol{M}_{\text{true}}}$ and $T_{j\cdot\boldsymbol{M}_{\text{true}},0}$, where $j \in \boldsymbol{M}_{\text{true}}$;

(c) tabulating the empirical coverage probabilities of the naïve confidence intervals for the true nonzero β^0_j, where $j \in \boldsymbol{M}_{\text{true}}$ in TABLE 1. The naïve confidence intervals are constructed by using $T^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}}$ in (5), and are given by

$$\Big(\widehat{\beta}^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}} - t(n-|\widehat{\boldsymbol{M}}|;1-\alpha/2)\text{sd}(\widehat{\beta}^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}}),\ \widehat{\beta}^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}} + t(n-|\widehat{\boldsymbol{M}}|;1-\alpha/2)\text{sd}(\widehat{\beta}^{\text{Method}}_{j\cdot\widehat{\boldsymbol{M}}})\Big), \tag{7}$$

where $t(n-|\widehat{\boldsymbol{M}}|;1-\alpha/2)$ is the $100(1-\alpha/2)^{\text{th}}$ percentile of a t-distribution with $n-|\widehat{\boldsymbol{M}}|$ degrees of freedom.

In this example, for parts (a) and (b) we only report the results for $j = 5$ as the plots for $j = 1, 3, 4, 7$ show similar behaviors. For part (c), we report empirical coverage probabilities of 95% ($\alpha = 0.05$) naïve confidence intervals

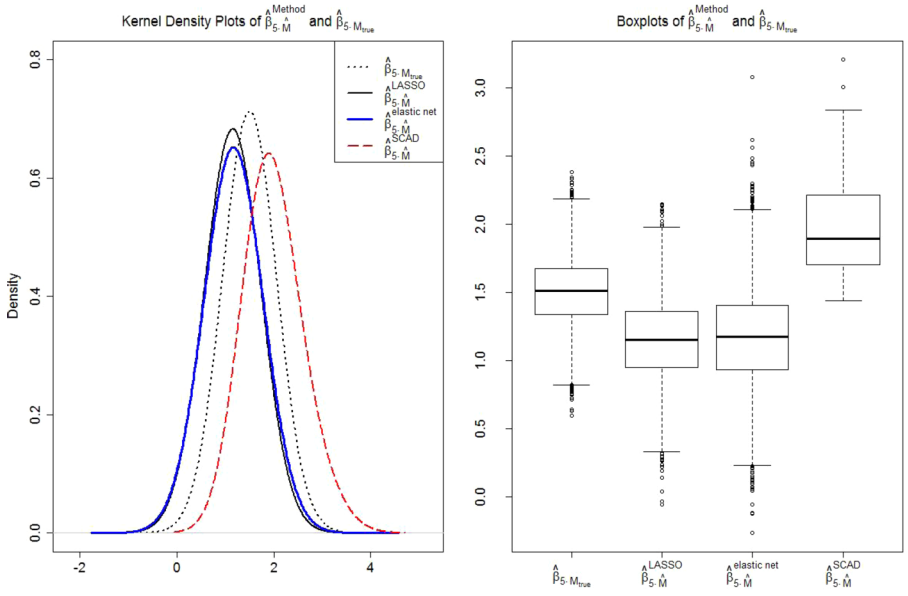

FIG 1. *Density plots (left) and box plots (right) of* $\widehat{\beta}^{Method}_{5\cdot\widehat{\boldsymbol{M}}}$ *against* $\widehat{\beta}_{5\cdot\boldsymbol{M}_{true}}$.

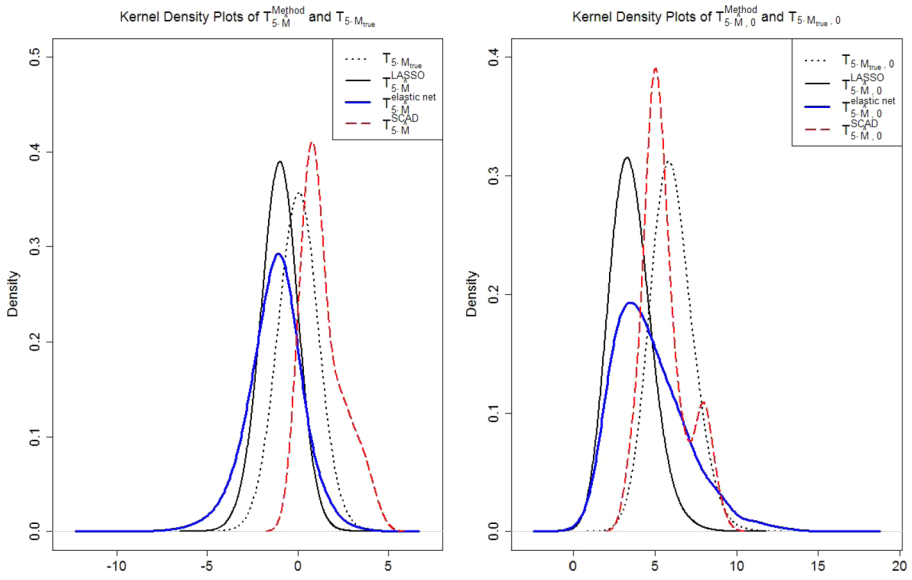

FIG 2. *Density plots of* $T^{Method}_{5\cdot\widehat{\boldsymbol{M}}}$ *against* $T_{5\cdot\boldsymbol{M}_{true}}$ *(left) and* $T^{Method}_{5\cdot\widehat{\boldsymbol{M}},0}$ *against* $T_{5\cdot\boldsymbol{M}_{true},0}$ *(right).*

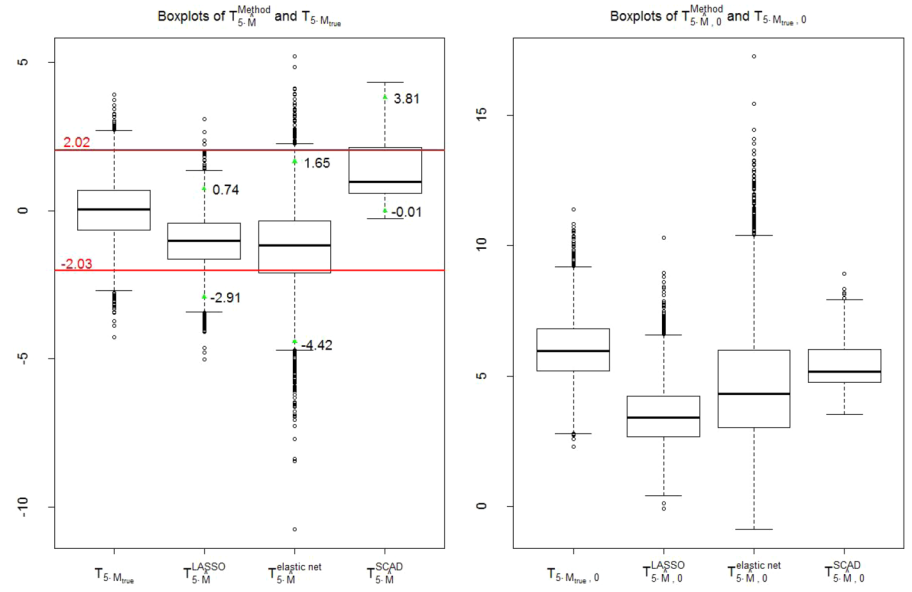

FIG 3. *Box-plots of* $T^{Method}_{5\cdot\widehat{\boldsymbol{M}}}$ *against* $T_{5\cdot\boldsymbol{M}_{true}}$ *(left) and* $T^{Method}_{5\cdot\widehat{\boldsymbol{M}},0}$ *against* $T_{5\cdot\boldsymbol{M}_{true},0}$ *(right). In the left panel, the* 2.5^{th} *and the* 97.5^{th} *percentiles of each boxplot are indicated.*

TABLE 1
Empirical coverage probabilities of 95% *naïve confidence intervals for the true non-zero* β_j^0*'s.*

	Empirical coverage probabilities				
Model Selector + Estimation	β_1^0	β_3^0	β_4^0	β_5^0	β_7^0
LASSO + OLS	.691	.486	.695	.736	.560
Elastic net + OLS	.749	.493	.815	.867	.619
SCAD + OLS	.362	.736	.682	.727	.723

in (7) for parameters $\beta_j^0, j = 1, 3, 4, 5, 7$, in TABLE 1 below. Details of model setting and `R` implementation, and additional plots for $j = 1, 3, 4, 7$ are given in Section S1 of the Supplementary Material (Zhang, Khalili & Asgharian 2022).

FIG 1, 2 and 3 show that regardless of model selectors, there exists substantial discrepancies between the sampling distribution of post-selection quantities, $\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}, T_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}, T_{j\cdot\widehat{\boldsymbol{M}},0}^{\text{Method}}$, and their counterparts, $\widehat{\beta}_{j\cdot\boldsymbol{M}_{\text{true}}}, T_{j\cdot\boldsymbol{M}_{\text{true}}}$ and $T_{j\cdot\boldsymbol{M}_{\text{true}},0}$. More specifically, bimodality, longer tail and positive or negative skewness appear in the sampling distribution of these post-selection quantities in contrast to their counterparts. Thus, we conclude that conducting model selection and inference on the same data can distort the sampling distribution of parameter estimators and the quantities in (5). In particular, the sampling distributions of these quantities after model selection no longer follow t-distribution as seen in FIG 2 and 3. Therefore, for example, in the construction of post-selection confidence intervals based on $T_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}$, simply using the tails of a t-distribution would be inaccurate to capture the tail behaviors of $T_{j\cdot\widehat{\boldsymbol{M}}}^{\text{Method}}$. This contributes to the under-coverage of naïve confidence intervals in (7), a phenomenon noted in TABLE 1. As it can be seen from the table, the under-coverage can be severe.

In addition, our simulations indicate that there is no tractable pattern of the post-selection sampling distributions of parameter estimators and quantities in (5), and the coverage probability of naïve confidence intervals in (7). We provide explanations for these phenomena in Sections 4.

4. Post-selection sampling distribution of parameter estimators and properties of coverage probabilities of naïve confidence intervals

We now review existing results about finite- and large-sample distributional properties of post-selection parameter estimators and properties of coverage probabilities of naïve confidence intervals in (7).

The research in entangling the distributional properties of post-selection parameter estimators is mainly pioneered by Benedikt M. Pötscher and Hannes Leeb. Through a series of works (Pötscher 1991, Pötscher & NovÁk 1998, Leeb & Pötscher 2003, 2005, 2006, 2008), the following points have been highlighted:

(i) *Deviation from Gaussianity.* The finite- (in both known and unknown variance) and large-sample conditional (on the order of selected models) and unconditional distributions of post-model-selection parameter estimators,

in general, deviate from Gaussian. Indeed, these distributions are mixtures of distributions, and each mixture component corresponds to a model in the model space. As such, these distributions have complicated forms and are difficult to work with when performing statistical inference.

(ii) *Non-uniform convergence*. The aforementioned mixture distributions depend on the unknown true parameter $\boldsymbol{\beta}^0$. More specifically, regardless of the use of consistent or inconsistent model selectors, the convergence of the mixture distributions to their limiting distributions depends on how frequently different models are selected. However, these frequencies depend on the true parameter $\boldsymbol{\beta}^0$, and consequently, the convergence of post-model-selection estimators to their limiting distributions is non-uniform with respect to $\boldsymbol{\beta}^0$. Thus, irrespective of the sample size, inferential validity of using the large-sample distribution to estimate the corresponding finite-sample counterpart is not guaranteed.

(iii) *Nonexistence of uniformly consistent estimator*. A uniformly consistent estimator of conditional (on the selected submodel) or unconditional distribution of post-model-selection parameter estimator does not exist. Therefore, it is impossible to estimate their exact distributions in a uniform and consistent fashion.

These results provide theoretical explanations to the discrepancies between the sampling distributions of the OLS estimator $\widehat{\beta}_{j\cdot \boldsymbol{M}_{\text{true}}}$ and post-selection OLS estimators $\widehat{\beta}^{\text{Method}}_{j\cdot \widehat{\boldsymbol{M}}}$ as seen in FIG 1. Since the computation of the quantities in (5) involves these parameter estimators, discrepancies between their true and post-selection sampling distributions are expected as observed in FIG 2 and 3.

Next, we review results concerning finite- and large-sample properties of naïve confidence intervals, serving as an explanation for the under-coverage observed in TABLE 1. The relevant work was initiated by fixing the true regression coefficient. In various settings, Hurvich & Tsai (1990), Regal & Hook (1991) and Zhang (1992) showed, both theoretically and via simulations, that finite- and large-sample conditional (on selected models) and unconditional coverage probabilities of naïve confidence intervals are lower than nominal coverage probabilities.

In large-sample theory, Kabaila (1995) explored the pitfall of assigning the criterion that asymptotic nominal coverage rate is achieved for a *fixed* regression coefficient as a benchmark to assess the asymptotic validity of a confidence region. In particular, using Hodge's example of a "superefficient" estimator, the author showed that even if the above criterion is met by some confidence interval, for any given n, there exists a regression coefficient such that this confidence interval fails to deliver the desired coverage probability. In other words, achieving proper asymptotic coverage probability for a fixed regression coefficient does not guarantee the validity of the associated confidence interval. As such, Kabaila (1995) argued that asymptotic coverage rate of a valid confidence region should be achieved by its *minimal* coverage probability over the space of all possible values of $\boldsymbol{\beta}^0$. Based on this criterion, a novel Monte Carlo simulation algorithm was proposed by Kabaila (2005) to compute the coverage probability of naïve

confidence intervals at any given parameter value, which is also shown to be lower than the nominal coverage probability based on simulated data. Further attempts were made by Kabaila & Leeb (2006) and Kabaila & Giri (2009), where upper bounds of the large- and finite-sample minimal coverage probability of naïve confidence interval are respectively derived. The authors showed that this upper bound could be far from the nominal coverage probability.

More detailed explanations of the above mentioned contents are respectively given in Sections S2 and S3 of our Supplementary Material.

Having recognized the challenges in performing statistical inference after data-driven model selection, researchers have since been devising post-selection inferential methodologies that take into account model selection uncertainty. In the next section, we discuss two types of statistical estimation targets.

5. Statistical estimation targets

One primary challenge of post-selection inference is the randomness in the estimation target. For instance, under the setting of the linear regression model (1), the OLS estimator corresponding to a submodel $\boldsymbol{M} \in \mathcal{M}_{\text{all}}$ is given by $\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}} = (\boldsymbol{X}_{\boldsymbol{M}}^{\top}\boldsymbol{X}_{\boldsymbol{M}})^{-1}\boldsymbol{X}_{\boldsymbol{M}}^{\top}\boldsymbol{Y}$, which is an unbiased estimator of $(\boldsymbol{X}_{\boldsymbol{M}}^{\top}\boldsymbol{X}_{\boldsymbol{M}})^{-1}\boldsymbol{X}_{\boldsymbol{M}}^{\top}\boldsymbol{X}\boldsymbol{\beta}^{0}$. However, upon introducing the uncertainty inherited by a data-drive model selector $\widehat{\boldsymbol{M}}_n$, it is not clear what the estimation target of the least-squares estimator $(\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^{\top}\boldsymbol{X}_{\widehat{\boldsymbol{M}}})^{-1}\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^{\top}\boldsymbol{Y}$ is. This makes the interpretation of an estimation target so formed ambiguous. Thus, it is imperative to first formulate a meaningful estimation target. With such motivation, we now discuss two classes of estimation targets that have been the focus of contemporary high-dimensional and post-selection inference literature, namely the population-based target in (1), and the projection-based target in (3).

5.1. Population-based target

According to the full model interpretation of parameters (Berk et al. 2013), the linear model (1) is regarded as the true data generating mechanism for the response $\boldsymbol{Y}$. As such, this model has the special status of embracing a complete set of predictors which are causal for the response vector $\boldsymbol{Y}$. In other words, the variation in $\boldsymbol{Y}$ can be explained by that set of predictors. Under this framework, although a submodel may be selected and the coefficient estimates of the de-selected predictors are given a numerical value of zero, these predictors do exist when conducting inference, and they have tangible and meaningful interpretations. Therefore, the regression coefficient β_j^0 represents the expected change in the response $\boldsymbol{Y}$ per unit change in x_j, holding all the remaining covariates fixed, regardless of being selected or de-selected. Thus, the coefficients β_j^0's are referred to as the population-based targets and they indeed contribute to the interpretation of linear model (1) according to the aforementioned full model interpretation of parameters.

It is apparent that the above view is legitimate only within the presumed framework (1). In fact, it hinges entirely on the first-order-correctness condition $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}^0$. On the other hand, another line of thought argues that some limitations may associate with the full model interpretation of parameters, which motivates the introduction of the projection-based target as discussed in the next section.

5.2. Projection-based target

There are several limitations of the interpretation of $\boldsymbol{\beta}^0$ according to the full model view of parameters. First, the concept of the so-called full model to entangle all the variability in $\boldsymbol{Y}$ is controversial. Given the complexity of the mother nature, it may neither be accurate nor feasible to measure or include all the pertinent predictors. Moreover, Berk et al. (2013) point out that due to the issue of predictor redundancy, which is common in social and biological sciences, the full model may not even contain the parameter of interest. Second, the full model view may give rise to complications when conducting statistical inference. In fact, as shown by Leeb & Pötscher (2005), the sampling distribution of coefficient estimator for the parameter of interest depends on all the other true regression coefficients. Since these coefficients are unknown, it is uncertain whether there exists discrepancy between the targeted sampling distribution and its counterpart when we have a fixed model. As a consequence, this may lead to unreliable statistical inference based on a wrong sampling distribution of the estimator for the parameter of interest. Indeed, its true sampling distribution cannot be estimated in a uniform and consistent manner; see Section 4, point (iii). Third, it is contentious to presume a linear structure of the true model. A plausible perspective toward statistical modeling of real life problems is that none of the arbitrarily constructed candidate models and observed data adequately represent the full reality (Bancroft & Han 1977, Box 1976). In fact, the true data generating mechanism is perhaps much more complex. Thus, simply assuming a specific structure for the true model may lack solid scientific verification.

Considering the above limitations with respect to the full model view of parameters, Berk et al. (2013) proposed three principles pertaining to the interpretation of regression parameters of a submodel: (i) "the full model has no special status other than being the repository of available predictors"; (ii) "the coefficients of excluded predictors are not zero; they are not defined and therefore do not exist"; (iii) "the meaning of a predictor's coefficient depends on which other predictors are included in the selected model".

Governed by the above three guidelines, Berk et al. (2013) subsequently proposed the projection-based target $\boldsymbol{b}_{\boldsymbol{M}}$ as defined in (3). It is seen that $\boldsymbol{b}_{\boldsymbol{M}}$ is the orthogonal projection of $\boldsymbol{\mu}$ onto the column space of $\boldsymbol{X}_{\boldsymbol{M}}$. In general, $\boldsymbol{b}_{\boldsymbol{M}}$ and $\boldsymbol{\beta}_{\boldsymbol{M}}$ in (2) are different. In fact, if the first-order-correctness condition holds so that $\boldsymbol{\mu} = \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\beta}_{\boldsymbol{M}}$ for the submodel $\boldsymbol{M}$, then the two targets are equal.

It has been discussed that the formulation of projection-based target $\boldsymbol{b}_{\boldsymbol{M}}$ is

advantageous over population-based target $\boldsymbol{\beta}^0$ in two aspects: (I) the first-order-correctness condition is not required; (II) simplicity in interpretation. Abandoning the assumption of the first-order-correctness is more appealing in many real life statistical modeling problems. Indeed, the central concept underpinning the definition of $\boldsymbol{b}_{\boldsymbol{M}}$ is that although it is unrealistic to assume a specific structure of the true model, it is always legitimate to *approximate* the true model by a linear structure containing a selected submodel $\boldsymbol{M}$. As such, with a selected submodel $\boldsymbol{M}$, each component of projected regression coefficient, denoted by $b_{j\cdot\boldsymbol{M}}$, $j \in \boldsymbol{M}$, is interpreted as the approximated expected change by submodel $\boldsymbol{M}$ in the response $\boldsymbol{Y}$ per unit change in x_j, holding all other covariates in $\boldsymbol{M}$ fixed. This interpretation is in fact the second advantage of $\boldsymbol{b}_{\boldsymbol{M}}$ as it is interpreted only within the framework of a submodel $\boldsymbol{M}$ thereby avoiding the issue caused by covariates that are not included in $\boldsymbol{M}$.

We next review the inferential methods for the two estimation targets discussed above respectively in Sections 6 and 7, by focusing on confidence interval constructions.

6. Post-selection inferential methodologies

Focusing on the projection-based targets, we now provide a review of three existing post-selection statistical inferential methods. Particular emphasis is on the construction of valid post-selection confidence intervals. A collection of recent work in post-model-selection that complements the foregoing discussions is presented in Section 6.4. Inferential methods designed for the population-based targets have been surveyed and compared through simulations by Dezeure et al. (2015) which we discuss and further elaborate on in Section 7.

6.1. Universally valid post-selection inference (PoSI)

Analogous to Scheffé's S-method (Scheffé 1959) on valid simultaneous inference by virtue of Scheffé's constant, Berk et al. (2013) proposed the so-called PoSI method, which is capable of producing universally valid post-selection confidence intervals for the projection-based regression coefficients $b_{j\cdot\widehat{\boldsymbol{M}}}$, $j \in \widehat{\boldsymbol{M}}$, regardless of model selection procedures and selected submodel $\widehat{\boldsymbol{M}}$. Here "valid" means the confidence intervals achieve at least the nominal coverage probability $1-\alpha$, for any $\alpha \in (0,1)$, by taking into account the model selection stage. The merit of the PoSI method is that even if a selected submodel deviates from the true model, it still guarantees valid inference for that selected submodel. We now describe the method.

Given a submodel $\widehat{\boldsymbol{M}} \in \mathcal{M}_{\text{all}}$ selected by a generic model selection procedure $\widehat{\boldsymbol{M}}_n$, we consider the following confidence interval for $b_{j\cdot\widehat{\boldsymbol{M}}}$,

$$\text{CI}_{j\cdot\widehat{\boldsymbol{M}}}(K) = \left(\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}} - K\widehat{\sigma}[(\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^\top \boldsymbol{X}_{\widehat{\boldsymbol{M}}})^{-1}]_{jj}^{1/2},\ \widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}} + K\widehat{\sigma}[(\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^\top \boldsymbol{X}_{\widehat{\boldsymbol{M}}})^{-1}]_{jj}^{1/2}\right), \tag{8}$$

where $\widehat{\beta}_{j\cdot\widehat{\boldsymbol{M}}}$ is the OLS estimator corresponding to a submodel $\widehat{\boldsymbol{M}}$, K is a constant to be specified below, and $[(\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^\top \boldsymbol{X}_{\widehat{\boldsymbol{M}}})^{-1}]_{jj}$ is the j^{th} diagonal element of the matrix $[(\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^\top \boldsymbol{X}_{\widehat{\boldsymbol{M}}})^{-1}]$. The error variance σ^2 is estimated by $\widehat{\sigma}^2 = \text{SSE}/(n-p)$, where SSE is the sum of squares of errors obtained from fitting a linear regression model including all the covariates (full model).

If in (8) K is replaced by $t(n-|\widehat{\boldsymbol{M}}|; 1-\alpha/2)$, we obtain the naïve confidence interval in (7). We have demonstrated in our motivating example, TABLE 1 in Section 3, that such interval fails to attain the nominal coverage probabilities for the population-based targets. We refer to Remark 6.1.2 for a recent result regarding the coverage probability of the naïve confidence intervals when the population-based target is replaced by the projection-based target.

The goal of the PoSI approach is to have a potentially larger value of the constant K to capture the extra randomness brought about by data-dependent model selector $\widehat{\boldsymbol{M}}_n$. In this way, the resulting confidence interval becomes wider and thus capable of achieving the nominal coverage probability, regardless of model selection procedures and selected submodels. This objective can be formulated as follows. The constant K in (8) is chosen such that

$$\mathbb{P}\left(b_{j\cdot\widehat{\boldsymbol{M}}} \in \text{CI}_{j\cdot\widehat{\boldsymbol{M}}}(K),\ \forall j \in \widehat{\boldsymbol{M}}\right) \geq 1-\alpha,\ \alpha \in (0,1). \tag{9}$$

More specifically, Berk et al. (2013) define the PoSI-constant, $K_{\text{PoSI}} \equiv K_{\text{PoSI}}(\boldsymbol{X}, \mathcal{M}_{\text{all}}, \alpha, r)$, by

$$K_{\text{PoSI}}(\boldsymbol{X}, \mathcal{M}_{\text{all}}, \alpha, r) = \min\left\{K \in \mathbb{R} \;\middle|\; \mathbb{P}\left(\max_{\boldsymbol{M}\in\mathcal{M}_{\text{all}}} \max_{j\in\boldsymbol{M}} |t_{j\cdot\boldsymbol{M}}| \leq K\right) \geq 1-\alpha\right\}, \tag{10}$$

where

$$\begin{aligned} t_{j\cdot\boldsymbol{M}} &= \frac{\widehat{\beta}_{j\cdot\boldsymbol{M}} - b_{j\cdot\boldsymbol{M}}}{\widehat{\sigma}[(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}}, \\ &= \frac{\boldsymbol{e}_j^\top (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1} \boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{Y} - \boldsymbol{e}_j^\top (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1} \boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{\mu}}{\widehat{\sigma}[(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}}, \\ &= \frac{\boldsymbol{e}_j^\top (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1} \boldsymbol{X}_{\boldsymbol{M}}^\top (\boldsymbol{Y} - \boldsymbol{\mu})}{\widehat{\sigma}[(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}}, \\ &= \frac{\boldsymbol{e}_j^\top (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1} \boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{\epsilon}}{\widehat{\sigma}[(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}}, \quad \text{and} \quad r = n-p. \end{aligned} \tag{11}$$

Recall that $\boldsymbol{e}_j \in \mathbb{R}^{|\boldsymbol{M}|}$ is the j^{th} standard basis. Now, since for any randomly selected submodel $\widehat{\boldsymbol{M}}$,

$$\max_{j\in\widehat{\boldsymbol{M}}} |t_{j\cdot\widehat{\boldsymbol{M}}}| \leq \max_{\boldsymbol{M}\in\mathcal{M}_{\text{all}}} \max_{j\in\boldsymbol{M}} |t_{j\cdot\boldsymbol{M}}|,$$

we then have

$$\mathbb{P}\left(\max_{j\in\widehat{\boldsymbol{M}}}|t_{j\cdot\widehat{\boldsymbol{M}}}|\leq K_{\mathrm{PoSI}}\right)\geq 1-\alpha.$$

Thus, by choosing $K = K_{\mathrm{PoSI}}$, (9) is satisfied.

We notice that the constant K_{PoSI} is the $100(1-\alpha/2)^{\mathrm{th}}$ percentile of the random variable $T = \max_{\boldsymbol{M}\in\mathcal{M}_{\mathrm{all}}} \max_{j\in\boldsymbol{M}} |t_{j\cdot\boldsymbol{M}}|$ in (10), whose distribution is needed for the computation of K_{PoSI}. On the other hand, for each $\boldsymbol{M}\in\mathcal{M}_{\mathrm{all}}$, the random variable $t_{j\cdot\boldsymbol{M}}$ as shown in (11) mainly depends on the design matrix $\boldsymbol{X}_{\boldsymbol{M}}$ and the error term $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$, where ϵ_i's are independent and $\epsilon_i \sim N(0,\sigma^2)$. Thus, the distribution of T can be approximated using Monte Carlo simulations. Specifically, we first obtain several copies of the random variable T based on independent draws of $\boldsymbol{\epsilon}$, where $\epsilon_i \sim N(0,\widehat{\sigma}^2)$ are generated independently. Then, given α, the constant K_{PoSI} satisfying (10) is approximated by the $100(1-\alpha/2)^{\mathrm{th}}$ percentile of the empirical distribution based on the obtained copies of T. Recall that $\widehat{\sigma}^2$ is computed based on the full model. In practice, this computation has been implemented in the `R` package `PoSI` (Buja & Zhang 2015).

Another choice for the constant K in (8) is the Scheffé's constant (Scheffé 1959), given by $K_{\mathrm{Sch}} = \sqrt{d\times F(d,r;1-\alpha)}$ under the linear regression model (1), where d is the rank of the full design matrix $\boldsymbol{X}$ and $F(d,r;1-\alpha)$ is the $100(1-\alpha)^{\mathrm{th}}$ percentile of an F-distribution with d and r degrees of freedom, i.e. $\mathbb{P}(F\geq F(d,r;1-\alpha)) = 1-\alpha$. In comparison, the Scheffé's constant is more conservative than the PoSI-constant, as Berk et al. (2013) showed that $K_{\mathrm{PoSI}}\leq K_{\mathrm{Sch}}$, for all design matrices $\boldsymbol{X}$ and any model universe $\mathcal{M}_{\mathrm{all}}$.

It is worthwhile to note that the calculation of the PoSI-constant, K_{PoSI}, is independent of any quantities derived from selected submodels. In other words, the formulation of the PoSI-constant is *pre-experimental* as it does not depend on any experiment outcomes. Although the construction of the PoSI confidence interval unavoidably involves some *post-experimental* quantities which are dependent on the experiment outcomes (such as the post-selection point estimators), the inclusion of the PoSI-constant indeed contributes to the universal validity in covering a random projected target with nominal coverage probability.

However, the strong universal protection resulted from the pre-experimental nature of the PoSI approach implies that this method is necessarily conservative, since its inferential validity is safeguarded against all possible selected submodels. Indeed, our simulations in Section 8 indicate that conditioning on a selected submodel, the PoSI confidence intervals are in general wider than the EPoSI confidence intervals (introduced in the next section). Yet, we demonstrate that this conservativeness of the PoSI approach is moderate, since when conditioning on both the selected submodel and corresponding sign vector of point estimator(s), the PoSI confidence intervals are not necessarily wider.

Apart from the above characteristics, one limitation of the PoSI approach is its relatively high computational cost for computing the PoSI constant K_{PoSI}. In fact, the authors recommended to use this approach for designs with $p\approx 20$

only. Therefore, this method is not feasible in high- or ultrahigh-dimensional settings unless some pre-screening procedure (Fan & Lv 2008) is performed.

Remark 6.1.1. The order of the magnitude of the PoSI-constant is studied by Berk et al. (2013) and Bachoc et al. (2018). Berk et al. (2013) prove that for a specific type of exchangeable design matrix, $K_{\text{PoSI}} = O(\sqrt{\log p})$ and this rate is also achieved for orthogonal design matrix. In full generality, however, it turns out that $K_{\text{PoSI}} = O(\sqrt{p})$, which is the rate for the Scheffé's constant. Bachoc et al. (2018) derive an upper bound of K_{PoSI} when the design matrix satisfies a Restricted Isometry Property condition (Foucart & Rauhut 2013). In practice, this upper bound becomes useful when it is computable while K_{PoSI} itself is not. In this case, K_{PoSI} can be replaced by the upper bound for constructing confidence intervals.

Remark 6.1.2. Zhao, Shojaie & Witten (2021) show that when we apply the LASSO for variable selection and then refit a linear regression model (1) corresponding to the selected variables, the resulting ordinary least-squares estimator is asymptotically normally distributed. As a consequence, the naïve confidence interval, $\text{CI}_{j\cdot\widehat{\boldsymbol{M}}}(z_{1-\alpha/2})$ in (8), where $z_{1-\alpha/2}$ denotes the $100(1-\alpha/2)^{\text{th}}$ percentile of a standard normal distribution, asymptotically attains the nominal coverage probability of $1-\alpha$ for the *projected* target. These results, however, depend on several assumptions regarding the design matrix $\boldsymbol{X}$, its dimensions n and p, the tuning parameter λ in running the LASSO for variable selection, and the true signal strength.

6.2. Exact post-selection inference (EPoSI)

In contrast to the pre-experimental proposal by Berk et al. (2013), Lee et al. (2016) and Lee & Taylor (2014) present a post-experimental approach, which depends on the model selection procedure and the selected submodel. Specifically, they construct confidence intervals with the exact $(1-\alpha)$ coverage probability via conditioning on a selected submodel. Therefore, their method is known as the conditionally exact post-selection inference and the resulting confidence interval for $b_{j\cdot\boldsymbol{M}}$, denoted by $\text{CI}_{j\cdot\boldsymbol{M}}$, satisfies

$$\mathbb{P}\Big(b_{j\cdot\boldsymbol{M}} \in \text{CI}_{j\cdot\boldsymbol{M}} \;\Big|\; \widehat{\boldsymbol{M}} = \boldsymbol{M}\Big) = 1-\alpha,\ j \in \boldsymbol{M}, \tag{12}$$

where the event $\{\widehat{\boldsymbol{M}} = \boldsymbol{M}\}$ means $\{\boldsymbol{Y} \in \mathbb{R}^n : \widehat{\boldsymbol{M}}_n(\boldsymbol{Y}) = \boldsymbol{M}\}$. Since this method involves a number of delicate steps, we first provide a compendium of this approach in Algorithm 1, before supplying details below.

The crux of the EPoSI approach involves two stages: first, determination of the conditional distribution of $\Big(\widehat{\beta}_{j\cdot\boldsymbol{M}} \;\Big|\; \widehat{\boldsymbol{M}} = \boldsymbol{M}\Big)$, which corresponds to **Steps 1-3** below, and it turns out to be a truncated normal distribution; second, construction of the confidence intervals as shown in (12), which corresponds **Steps 4-5**. Specifically, the construction of confidence intervals shown in (12)

Algorithm 1 EPoSI Algorithm

1: For model selection procedures possessing polyhedral selection property, the so-called selection event $E_n(\boldsymbol{M}, \boldsymbol{s})$, defined as $E_n(\boldsymbol{M}, \boldsymbol{s}) = \left\{\boldsymbol{Y} \in \mathbb{R}^n : \widehat{\boldsymbol{M}}_n(\boldsymbol{Y}) = \boldsymbol{M}, \widehat{\boldsymbol{s}}_n(\boldsymbol{Y}) = \boldsymbol{s}\right\} \equiv \left\{\widehat{\boldsymbol{M}} = \boldsymbol{M}, \widehat{\boldsymbol{s}} = \boldsymbol{s}\right\}$, can be shown to be equivalent to $\left\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\ \boldsymbol{Y} \leq \boldsymbol{B}(\boldsymbol{M}, \boldsymbol{s})\right\}$ for some affine matrix $\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})$ and vector $\boldsymbol{B}(\boldsymbol{M}, \boldsymbol{s})$.

2: The above formulation of selection event can be further decomposed as $\left\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\ \boldsymbol{Y} \leq \boldsymbol{B}(\boldsymbol{M}, \boldsymbol{s})\right\} = \left\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{\eta_j}^\top \boldsymbol{Y} \in [\mathcal{V}^-(\boldsymbol{Y}), \mathcal{V}^+(\boldsymbol{Y})], \mathcal{V}^0(\boldsymbol{Y}) \geq 0\right\}$, where $\boldsymbol{\eta_j}^\top \boldsymbol{Y}$ gives the ordinary least-squares estimator for $b_{j \cdot \boldsymbol{M}}$, which is $\widehat{\beta}_{j \cdot \boldsymbol{M}}$. Thus, the conditional distribution of $\left(\widehat{\beta}_{j \cdot \boldsymbol{M}} | E_n(\boldsymbol{M}, \boldsymbol{s})\right)$ is equivalent to $\left(\boldsymbol{\eta_j}^\top \boldsymbol{Y} \middle| \left\{\boldsymbol{\eta_j}^\top \boldsymbol{Y} \in [v^-, v^+]\right\}\right)$.

3: By the normality assumption of $\boldsymbol{Y}$, $\left(\boldsymbol{\eta_j}^\top \boldsymbol{Y} \middle| \left\{\boldsymbol{\eta_j}^\top \boldsymbol{Y} \in [v^-, v^+]\right\}\right)$ follows a truncated normal distribution with cumulative distribution function given by $F^{[v^-, v^+]}_{\xi_j, \tau_j}(\cdot)$, where ξ_j and τ_j are the mean and variance of this truncated normal, respectively.

4: By Probability Integral Transform, $\left(F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\xi_j, \tau_j}(\boldsymbol{\eta_j}^\top \boldsymbol{Y}) \;\middle|\; E_n(\boldsymbol{M}, \boldsymbol{s})\right) \sim \text{Unif}(0, 1)$.

5: The above result in Step 4 enables us to construct the confidence interval $\text{CI}_{j \cdot \boldsymbol{M}}$ in (12). In practice, the upper and lower bounds of $\text{CI}_{j \cdot \boldsymbol{M}}$ can be computed using the `R` package `selectiveInference` (Tibshirani, Tibshirani, Taylor, Loftus & Reid 2016).

can be accomplished via first considering the distribution of $\left(\widehat{\beta}_{j \cdot \boldsymbol{M}} \;\middle|\; \widehat{\boldsymbol{M}} = \boldsymbol{M}, \widehat{\boldsymbol{s}} = \boldsymbol{s}\right)$. Next, we provide details for the EPoSI framework.

Step 1 We first recall that the ordinary least-squares estimator for $b_{j \cdot \boldsymbol{M}}$ is $\widehat{\beta}_{j \cdot \boldsymbol{M}}$. To obtain $\text{CI}_{j \cdot \boldsymbol{M}}$ as in (12), we need to find the conditional distribution of $\widehat{\beta}_{j \cdot \boldsymbol{M}}$, given $\widehat{\boldsymbol{M}} = \boldsymbol{M}$. To achieve that, we first investigate the conditional distribution of $\widehat{\beta}_{j \cdot \boldsymbol{M}}$ given the so-called selection event

$$E_n(\boldsymbol{M}, \boldsymbol{s}) = \left\{\boldsymbol{Y} \in \mathbb{R}^n : \widehat{\boldsymbol{M}}_n(\boldsymbol{Y}) = \boldsymbol{M}, \widehat{\boldsymbol{s}}_n(\boldsymbol{Y}) = \boldsymbol{s}\right\} \equiv \left\{\widehat{\boldsymbol{M}} = \boldsymbol{M}, \widehat{\boldsymbol{s}} = \boldsymbol{s}\right\}. \tag{13}$$

Here, $\widehat{\beta}_{j \cdot \boldsymbol{M}}$ can be written as $\boldsymbol{\eta}_j^\top \boldsymbol{Y}$, where $\boldsymbol{\eta}_j^\top = \boldsymbol{e}_j^\top (\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}})^{-1} \boldsymbol{X}_{\boldsymbol{M}}^\top$ and $\boldsymbol{e}_j \in \mathbb{R}^{|\boldsymbol{M}|}$ is the j^{th} standard basis. Moreover, $\widehat{\boldsymbol{s}}_n$ is the selection of sign vector of the non-zero coefficient estimate(s) corresponding to the model selection procedure $\widehat{\boldsymbol{M}}_n$, $\widehat{\boldsymbol{s}}$ is the selected sign vector which is random and $\boldsymbol{s}$ is the selected sign vector corresponding to submodel $\boldsymbol{M}$.

Then, to ascertain the exact conditional distribution of $\boldsymbol{\eta}_j^\top \boldsymbol{Y} \mid E_n(\boldsymbol{M}, \boldsymbol{s})$, it is required that the selection event $E_n(\boldsymbol{M}, \boldsymbol{s})$ satisfies

$$E_n(\boldsymbol{M}, \boldsymbol{s}) = \left\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\ \boldsymbol{Y} \leq \boldsymbol{B}(\boldsymbol{M}, \boldsymbol{s})\right\}, \tag{14}$$

where $\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})$ and $\boldsymbol{B}(\boldsymbol{M}, \boldsymbol{s})$ respectively denote some affine matrix and vector depending on the model selection procedure, selected submodel, and sign vector of the corresponding non-zero coefficient estimate(s). Here, for two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, we say $\boldsymbol{u} \leq \boldsymbol{v}$ if and only if every component of $\boldsymbol{u}$ is less than or equal to every component of $\boldsymbol{v}$. Geometrically, (14) is equivalent to characterizing the

response vector falling into a single polytope. Thus, this requirement is known as the *polyhedral selection property*. Examples of model selection procedures having this property include, but not limited to, LASSO and elastic net with fixed tuning parameters, marginal screening, nonnegative least squares and orthogonal matching pursuit (Tropp & Gilbert 2007). The specific forms of the affine matrix and vector $\boldsymbol{A(M,s)}$ and $\boldsymbol{B(M,s)}$ corresponding to LASSO are given in Section S4 of our Supplementary Material.

Step 2 With the polyhedral selection property (14), Lee et al. (2016) and Lee & Taylor (2014) show that the single polytope $\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{A(M,s)}\ \boldsymbol{Y} \leq \boldsymbol{B(M,s)}\}$ can be further decomposed as

$$\Big\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{A(M,s)}\ \boldsymbol{Y} \leq \boldsymbol{B(M,s)}\Big\} = \Big\{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{\eta_j^\top Y} \in \big[\mathcal{V}^-(\boldsymbol{Y}), \mathcal{V}^+(\boldsymbol{Y})\big], \mathcal{V}^0(\boldsymbol{Y}) \geq 0\Big\}, \quad (15)$$

where the exact forms of quantities $\mathcal{V}^-(\boldsymbol{Y}), \mathcal{V}^+(\boldsymbol{Y})$ and $\mathcal{V}^0(\boldsymbol{Y})$ are given in Section S4 of our Supplementary Material. These quantities are shown to be independent of $\boldsymbol{\eta_j^\top Y}$. In summary, with the polyhedral selection property, the selection event $E_n(\boldsymbol{M}, \boldsymbol{s})$ is equivalent to the event of all $\boldsymbol{Y} \in \mathbb{R}^n$ such that $\boldsymbol{\eta_j^\top Y}$ is truncated to some interval $[\mathcal{V}^-(\boldsymbol{Y}), \mathcal{V}^+(\boldsymbol{Y})]$ with $\mathcal{V}^0(\boldsymbol{Y}) \geq 0$.

Step 3 Based on the characterizations (14) and (15), and the assumption that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, we have

$$\boldsymbol{\eta_j^\top Y} \Big| \Big\{\boldsymbol{\eta_j^\top Y} \in [v^-, v^+]\Big\} \sim \mathrm{TN}\Big(\xi_j, \tau_j; v^-, v^+\Big), \quad (16)$$

where $v^-, v^+ \in \mathbb{R}$, $\xi_j = \boldsymbol{\eta_j^\top \mu}$, $\tau_j = \sigma^2 ||\boldsymbol{\eta}_j||^2$, and $\mathrm{TN}(\mu, \sigma^2; a, b)$ denotes a Gaussian distribution with mean μ and variance σ^2, truncated to the interval $[a, b]$.

Step 4 Let $F^{[a,b]}_{\mu,\sigma^2}(x)$ denote the distribution function of such a truncated Gaussian random variable. By the Probability Integral Transformation, (14) and (16), we have

$$\Big(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\xi_j,\tau_j}(\boldsymbol{\eta_j^\top Y}) \ \Big|\ E_n(\boldsymbol{M}, \boldsymbol{s})\Big) \sim \mathrm{Unif}(0,1), \quad (17)$$

where $\mathcal{V}^- = \mathcal{V}^-(\boldsymbol{Y})$, $\mathcal{V}^+ = \mathcal{V}^+(\boldsymbol{Y})$ and $\mathrm{Unif}(0,1)$ denotes a continuous uniform random variable on the interval $(0,1)$.

Step 5 The finite-sample pivot (17) enables us to construct valid confidence intervals for the $b_{j\cdot\boldsymbol{M}}$ as follows. Let $\mathrm{L}_j^{\boldsymbol{s}}$ and $\mathrm{U}_j^{\boldsymbol{s}}$ be two quantities satisfying $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\mathrm{L}_j^{\boldsymbol{s}},\tau_j}(\boldsymbol{\eta_j^\top Y}) = 1 - \alpha/2$ and $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\mathrm{U}_j^{\boldsymbol{s}},\tau_j}(\boldsymbol{\eta_j^\top Y}) = \alpha/2$. Then due to the fact that $F^{[a,b]}_{\mu,\sigma^2}(x)$ is monotone decreasing in μ (Lee et al. 2016), we have

$$\mathbb{P}\Big(b_{j\cdot\boldsymbol{M}} \in [\mathrm{L}_j^{\boldsymbol{s}}, \mathrm{U}_j^{\boldsymbol{s}}] \ \Big|\ E_n(\boldsymbol{M}, \boldsymbol{s})\Big) = 1 - \alpha, \quad (18)$$

where $E_n(\boldsymbol{M}, \boldsymbol{s})$ satisfies the requirement in (14).

To construct the confidence intervals in (12), we then consider a union of polytopes:

$$\begin{aligned}\left\{\widehat{\boldsymbol{M}} = \boldsymbol{M}\right\} &= \bigcup_{\boldsymbol{s}\in\{-1,1\}^{|\boldsymbol{M}|}} \left\{\widehat{\boldsymbol{M}} = \boldsymbol{M}, \widehat{\boldsymbol{s}} = \boldsymbol{s}\right\} \\ &= \bigcup_{\boldsymbol{s}\in\{-1,1\}^{|\boldsymbol{M}|}} E_n(\boldsymbol{M}, \boldsymbol{s}) \\ &= \bigcup_{\boldsymbol{s}\in\{-1,1\}^{|\boldsymbol{M}|}} \left\{\boldsymbol{Y} : \boldsymbol{\eta_j}^\top \boldsymbol{Y} \in [\mathcal{V}^-(\boldsymbol{Y}), \mathcal{V}^+(\boldsymbol{Y})], \mathcal{V}^0(\boldsymbol{Y}) \geq 0\right\} \text{ by (15)}, \\ &= \left\{\boldsymbol{Y} : \boldsymbol{\eta_j}^\top \boldsymbol{Y} \in [\widetilde{\mathcal{V}}^-(\boldsymbol{Y}), \widetilde{\mathcal{V}}^+(\boldsymbol{Y})], \widetilde{\mathcal{V}}^0(\boldsymbol{Y}) \geq 0\right\}, \end{aligned} \tag{19}$$

where the union is taken over $2^{|M|}$ sign vectors. By the same argument as conditioning on a single polytope in (17), we obtain the uniformly distributed pivotal quantity as

$$\left(F_{\xi_j,\tau_j}^{[\widetilde{\mathcal{V}}^-,\widetilde{\mathcal{V}}^+]}(\boldsymbol{\eta_j}^\top \boldsymbol{Y}) \;\middle|\; \widehat{\boldsymbol{M}} = \boldsymbol{M}\right) \sim \text{Unif}(0,1), \tag{20}$$

from which the lower and upper bound L_j and U_j can be constructed in the same fashion as the single polytope-based interval in (18). As such, the EPoSI confidence interval $\text{CI}_{j\cdot\boldsymbol{M}}$ in (12) is given by $[\text{L}_j, \text{U}_j]$.

One advantage of the EPoSI is its capability of constructing valid post-selection confidence intervals in both low- and high-dimensional configurations. Taylor & Tibshirani (2018) explore generalization of the EPoSI approach to a large class of ℓ_1-penalized regression models, including generalized linear models, Cox's proportional hazards model, and the graphical LASSO (Friedman et al. 2008). On the other hand, one limitation of the EPoSI approach is with respect to the lengths of the resulting confidence intervals in both single and union of polytopes cases. More specifically, by conditioning on a single polytope (14) or a union of polytopes (19), the resulting confidence intervals are sometimes wider than the competing confidence intervals when the signal strengths are weak (Lee et al. 2016), or may even have infinite lengths (Kivaranovic & Leeb 2021). These phenomena have also been demonstrated by our simulations in TABLE 6-8, Section 8.

To gain insights and explore possible remedial procedures to this limitation, Fithian et al. (2017) explain that including redundant information in the selection event in (15), also known as the conditioning event, gives rise to reduced power of associated hypothesis tests, contributing to unacceptably wide confidence intervals. The authors then propose a generic framework of uniformly most powerful unbiased selective level-α test for exponential family models, and conclude that hypothesis tests based on data splitting procedures are always inadmissible. In other words, there always exists a testing procedure exhibiting higher power than data splitting. Tian & Taylor (2018) propose to incorporate a noise ω in the original response $\boldsymbol{Y}$ and perform subsequent inference procedures outlined in this section on the randomized response $\boldsymbol{Y}^* = \boldsymbol{Y} + \omega$. It is

theoretically shown that the resulting estimator derived from the randomized response $\boldsymbol{Y}^*$ is uniformly consistent. Moreover, via simulations, Tian & Taylor (2018) demonstrate that assigning ω to be either Gaussian or logistic noise leads to shortened EPoSI confidence intervals. This framework of randomized noise augmentation is then applied by Hyun et al. (2021) to post-selection inference for change point detection, which demonstrates enhanced power. As discussed by Fithian et al. (2017) and Tian & Taylor (2018), conditioning on unnecessary information implies less information for statistical inference, and hence resulting in wider confidence intervals. Motivated by this, Liu et al. (2018) construct inferential framework with so-called minimal conditioning, focusing on a different estimation target than the projection-based one in (3). Specifically, they divide the selected model $\widehat{\boldsymbol{M}}_n(\boldsymbol{Y}) = \boldsymbol{M}$ into two groups of variables, namely the high and low value targets. Only the high value targets, denoted by $\widehat{\boldsymbol{H}}_n(\boldsymbol{Y}) = \boldsymbol{H}$, are included in the new conditioning event and instead of (12), the authors aim to achieve

$$\mathbb{P}\Big(b_{j\cdot\boldsymbol{H}} \in \mathrm{CI}_{j\cdot\boldsymbol{H}} \;\Big|\; j \in \boldsymbol{M}, \widehat{\boldsymbol{H}}_n(\boldsymbol{Y}) = \boldsymbol{H}\Big) = 1 - \alpha,$$

where the new target for inference, $b_{j\cdot\boldsymbol{H}}$, is the j^{th} coordinate of the vector $(\boldsymbol{X}_{\boldsymbol{H}}^\top \boldsymbol{X}_{\boldsymbol{H}})^{-1}\boldsymbol{X}_{\boldsymbol{H}}^\top \mathbb{E}(\boldsymbol{Y})$, given the high value target $\boldsymbol{H}$, and $\mathrm{CI}_{j\cdot\boldsymbol{H}}$ is the corresponding EPoSI confidence interval. To select the high value targets, two solutions are proposed, namely stable-ℓ_1 and stable-t approach. Based on this set up, similar derivation of truncation region and truncated Gaussian result as those in (15) and (16) are constructed. Via simulation, Liu et al. (2018) verify that their framework has less chance of producing infinite confidence intervals. Besides Liu et al. (2018), Jewell et al. (2021), Mehrizi & Chenouri (2021) and Chen et al. (2021) also have proposed EPoSI-based frameworks that condition on less information, with specific application to inference for change point detection and graph fused LASSO, respectively.

6.3. EPoSI for sequential regression

Tibshirani, Taylor, Lockhart & Tibshirani (2016) show that the EPoSI method can also be applied at each step of certain sequential regression procedures. Their goal is to assess the significance of the latest selected covariate(s) through projecting $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ onto the space of current active covariates. The implication is that, at each step of a sequential regression procedure and given the set of current active predictors, conditionally exact post-selection confidence intervals for the projected regression coefficient of a newly-entered covariate can be constructed. We should be aware that the EPoSI approach is only applicable to those sequential regression procedures that at *each* step satisfy the polyhedral selection property described in Section 6.2. A non-exhaustive list of such sequential regression procedures includes a modified version of forward stepwise regression (Tibshirani, Taylor, Lockhart & Tibshirani 2016), the least angle regression (LARS) (Efron et al. 2004) and the LASSO solution path which is a modified LARS path.

To apply the EPoSI method on a sequential regression procedure, similar to (13), consider the selection event

$$E_{n,k}(\boldsymbol{M}_k, \boldsymbol{s}_k) = \Big\{ \boldsymbol{Y} \in \mathbb{R}^n : \widehat{\boldsymbol{M}}_{n,k}(\boldsymbol{Y}) = \boldsymbol{M}_k, \widehat{\boldsymbol{s}}_{n,k}(\boldsymbol{Y}) = \boldsymbol{s}_k \Big\},$$

where $\boldsymbol{M}_k$ and $\boldsymbol{s}_k$ are respectively the set of current active covariate(s) and sign vector of their corresponding estimated regression coefficient(s) after k steps. If the polyhedral selection property is satisfied, then by (14), we have

$$E_{n,k}(\boldsymbol{M}_k, \boldsymbol{s}_k) = \Big\{ \boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{A}_k(\boldsymbol{M}_k, \boldsymbol{s}_k)\ \boldsymbol{Y} \leq \boldsymbol{B}_k(\boldsymbol{M}_k, \boldsymbol{s}_k) \Big\},$$

for some step-specific affine matrix $\boldsymbol{A}_k(\boldsymbol{M}_k, \boldsymbol{s}_k) \equiv \boldsymbol{A}_k$ and vector $\boldsymbol{B}_k(\boldsymbol{M}_k, \boldsymbol{s}_k) \equiv \boldsymbol{B}_k$. Given $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$, valid conditional confidence intervals for the projected coefficient(s) of the newly-entered covariate(s) after k steps, conditioning on $E_{n,k}(\boldsymbol{M}_k, \boldsymbol{s}_k)$, can be constructed as prescribed in Section 6.2. Marginalizing over $2^{|\boldsymbol{M}_k|}$ possible sign patterns yields a confidence interval upon conditioning on $\{\widehat{\boldsymbol{M}}_{n,k}(\boldsymbol{Y}) = \boldsymbol{M}_k\}$.

In case of a modified forward stepwise regression, Tibshirani, Taylor, Lockhart & Tibshirani (2016) explicitly construct $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ through induction. They show that $\boldsymbol{B}_k = \boldsymbol{0}$, for all k, and both $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ have exactly $2pk - k(k+1)$ rows after k steps, where p is the total number of available covariates in the data. Moreover, $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are also formulated when applying this scheme to the LARS and LASSO solution path. In this case, the authors demonstrate that both the affine matrix $\boldsymbol{A}_k$ and vector $\boldsymbol{B}_k$ roughly reach $3pk$ rows after k steps, resulting in relatively heavier computation load. To remedy the computational burden in case of the LARS path, the authors propose a compact version of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$, denoted by $\widetilde{\boldsymbol{A}}_k$ and $\widetilde{\boldsymbol{B}}_k$. The compact versions only have $k+1$ rows after k steps, which immensely enhance the computational feasibility and efficiency of the EPoSI method. The authors show that applying $\widetilde{\boldsymbol{A}}_k$ and $\widetilde{\boldsymbol{B}}_k$ for uncertainty quantification about $\boldsymbol{w}_k^\top \boldsymbol{\mu}$ requires the contrast vector at k^{th} step, $\boldsymbol{w}_k$, to be in the column span of current active variables selected by LARS. As a consequence, they propose the so-called *spacing test* $H_{0,k} : \boldsymbol{w}_k^\top \boldsymbol{\mu} = 0$, to assess the significance of the latest selected variable at the k^{th} step when regressing $\boldsymbol{\mu}$ on the column space of $\boldsymbol{X}_{\boldsymbol{M}_k}$. The primary strength of the spacing test lies on the simplicity of its corresponding one- and two-sided spacing statistics (Tibshirani, Taylor, Lockhart & Tibshirani 2016). Recently, Azais et al. (2018) have examined the distribution of the spacing test statistic under the alternative hypothesis. The authors have studied the power of this test and proven that it is unbiased for LARS.

Under the framework of the linear regression model (1), Lockhart et al. (2014) present a closely related method to test the significance of all the $k-1$ predictors entered before the k^{th} knot λ_k along the LASSO solution path. Their goal is to test $H_0 : \text{supp}(\boldsymbol{\beta}^0) \subseteq \boldsymbol{M}_{k-1}$ against its alternative, where $\text{supp}(\boldsymbol{\beta}^0) = \{j : \beta_j^0 \neq 0, 1 \leq j \leq p\}$. Their proposed statistic for testing the null hypothesis H_0 is called covariance test statistic and is given by

$$C_k = \frac{1}{\sigma^2} \cdot \Big(\Big\langle \boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{M}_k}\, \widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_{k+1}) \Big\rangle - \Big\langle \boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{M}_{k-1}}\, \widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_{k+1}) \Big\rangle \Big), \quad (21)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_{k+1})$ is the estimated coefficient(s) along the LASSO solution path with tuning parameter $\lambda = \lambda_{k+1}$ using predictors in $\boldsymbol{M}_k$, and $\widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_{k+1})$ is the estimated coefficient(s) obtained by re-applying LASSO with tuning parameter $\lambda = \lambda_{k+1}$ to the active variables in $\boldsymbol{M}_{k-1}$. In other words, C_k in (21) measures how much change in the covariance between the response vector $\boldsymbol{Y}$ and the fitted values can be attributed to the newly-entered predictor(s) at the tuning parameter value $\lambda = \lambda_{k+1}$.

Lockhart et al. (2014) show that under $H_0 : \text{supp}(\boldsymbol{\beta}^0) \subseteq \boldsymbol{M}_{k-1}$, and certain conditions on the design matrix $\boldsymbol{X}$, as $(n, p) \to \infty$,

$$C_k \xrightarrow{d} \text{Exp}(1),$$

where Exp(1) is the exponential random variable with scale parameter 1. Thus, the approximated exponential distribution is used to test H_0 at a designated significance level. Tibshirani, Taylor, Lockhart & Tibshirani (2016) establish asymptotic equivalence between the covariance test statistic C_k and a modified version of their spacing test statistic, R_k, as $\exp(C_k) = R_k(1 + o_p(1))$.

6.4. Recent work

PoSI and related work: Following the idea of PoSI method (Berk et al. 2013) delineated in Section 6.1, Bachoc et al. (2020) develop a general framework to construct asymptotically (p fixed and $n \to \infty$) valid marginal confidence interval for the projection-based target $\boldsymbol{b}_{\widehat{\boldsymbol{M}}}$ defined in (3), irrespective of model selectors and selected submodel $\widehat{\boldsymbol{M}}$. Remarkably, the large-sample coverage rate attained by such interval for $\boldsymbol{b}_{\widehat{\boldsymbol{M}}}$ is also uniformly achieved over an arbitrary sequence of underlying data generating mechanisms. One salient feature of this method is that it does not require the homoscedasticity and normality assumptions, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, as in Berk et al. (2013). Moreover, this approach is applicable to a wide range of model settings, including homoscedastic or heteroscedastic linear regression models and binary regression models. Specifically, when fitting homoscedastic linear regression models to non-Gaussian data, the proposed confidence interval has the same form as (8) with $K = K_{\text{UPoSI}}$, known as the *u*niform PoSI-constant. In comparison with the PoSI method, K_{UPoSI} is computed as the $100(1-\alpha)^{\text{th}}$ percentile of the maximum of Gaussian random variables, whereas the PoSI-constant, K_{PoSI}, is that of the maximum of the t-type random variables as in (10). In addition, in constructing $\text{CI}_{j \cdot \widehat{\boldsymbol{M}}}(K_{\text{UPoSI}})$, the estimator of the error variance, $\widehat{\sigma}$, is obtained based on the selected submodel. In contrast, this estimator is calculated based on the full model in formulating the PoSI confidence interval $\text{CI}_{j \cdot \widehat{\boldsymbol{M}}}(K_{\text{PoSI}})$. In practice, the algorithm for computing K_{UPoSI} is given in the supplementary material of Bachoc et al. (2019).

So far, the aforementioned methodologies have been proposed upon conditioning on all the regressors contained in the design matrix $\boldsymbol{X}$. The implication is that the distribution of $\boldsymbol{X}$ is independent of any model parameter of interest,

say $\boldsymbol{\theta}$, to be estimated. Such ancillarity assumption of the distribution of $\boldsymbol{X}$ with respect to $\boldsymbol{\theta}$ can be mathematically expressed as

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) \cdot f_{\boldsymbol{X}}(\boldsymbol{x}), \tag{22}$$

where $f_{\boldsymbol{X},\boldsymbol{Y}}$, $f_{\boldsymbol{Y}|\boldsymbol{X}}$ and $f_{\boldsymbol{X}}$ respectively denote the joint distribution of $(\boldsymbol{X},\boldsymbol{Y})$, conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and marginal distribution of $\boldsymbol{X}$. Buja et al. (2015) (Section 4) give a nice account of the dependence of model coefficient on the distribution of $\boldsymbol{X}$, especially when the true mean response vector $\boldsymbol{\mu}$ is nonlinear in $\boldsymbol{X}$ and $\boldsymbol{X}$ is heteroscedastic. As such, the ancillarity assumption in (22) is violated. In view of this, Buja et al. (2015), Kuchibhotla et al. (2018*a*), Kuchibhotla et al. (2018*b*) and Rinaldo et al. (2019) adopt an assumption-lean framework. Specifically, the classical assumptions that $\boldsymbol{X}$ is fixed, the response vector $\boldsymbol{Y}$ is normally distributed and homoscedastic are abandoned.

For any submodel $\boldsymbol{M} \in \mathcal{M}_{\text{all}}$, Kuchibhotla et al. (2018*a*) and Kuchibhotla et al. (2018*b*) focus on the so-called population version of the least-squares estimator

$$\boldsymbol{\zeta}_{\boldsymbol{M}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{|\boldsymbol{M}|}} \mathbb{E}\|\boldsymbol{Y} - \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\beta}\|^2 = \left[\mathbb{E}\left(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}}\right)\right]^{-1} \mathbb{E}\left(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{Y}\right), \tag{23}$$

where the expectation is with respect to the joint distribution of $(\boldsymbol{X},\boldsymbol{Y})$. Kuchibhotla et al. (2018*a*) obtain the estimation error bounds, in terms of both ℓ_1- and ℓ_2-norms, for $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}} - \boldsymbol{\zeta}_{\boldsymbol{M}})$ and the linear representation

$$\left(\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}} - \boldsymbol{\zeta}_{\boldsymbol{M}}\right) - \left[\mathbb{E}\left(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}}\right)\right]^{-1}\left(\boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{Y} - \boldsymbol{X}_{\boldsymbol{M}}^\top \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\zeta}_{\boldsymbol{M}}\right)\right),$$

uniformly over the set of all submodels in $\mathcal{M}_{\text{all}}$ with sizes less than a prespecified value k. Moreover, the corresponding rates of convergence are also derived. Kuchibhotla et al. (2018*b*) construct both finite- and large-sample valid simultaneous confidence regions for $\boldsymbol{\zeta}_{\boldsymbol{M}}$. In practice, these confidence regions are computed based on high-dimensional central limit theorem (Chernozhukov et al. 2017) and multiplier bootstrap (Zhang & Cheng 2014).

On the other hand, for a selected submodel $\widehat{\boldsymbol{M}}$ (a random quantity), Rinaldo et al. (2019) study the inference for $\boldsymbol{\zeta}_{\widehat{\boldsymbol{M}}}$, referred to as the linear projection parameter, which is the best linear predictor of $\boldsymbol{Y}$ using $\boldsymbol{X}_{\widehat{\boldsymbol{M}}}$. The authors propose two constructions of simultaneous confidence regions for $\boldsymbol{\zeta}_{\widehat{\boldsymbol{M}}}$. Respectively, they are based on a combination of sample splitting and bootstrap, and sampling splitting and normal approximation. The asymptotic coverage rates of these two constructions are also investigated. Lei et al. (2018) extend the aforementioned assumption-lean framework to predictive inference and construct marginally valid prediction intervals for the predicted responses.

From the perspective of algorithmic stability, Zrnic & Jordan (2020) propose a construction of valid post-selection confidence intervals for $\boldsymbol{\zeta}_{\widehat{\boldsymbol{M}}}$ using the generic form specified in (8) with an appropriate choice of the quantity K (other than

the PoSI constant K_{PoSI} (10)). Heuristically, a model selection procedure $\widehat{\boldsymbol{M}}_n$ is stable if given two realizations of data, the probability that two submodels selected by $\widehat{\boldsymbol{M}}_n$ are different is small. Their method is based on the logic that for a stable model selection procedure, the submodel selected is almost independent of the data. As a consequence, it is reasonable to conduct both model selection and valid statistical inference on the same data. In other words, a more stable model selection procedure leads to confidence intervals that are shorter, less variable in terms of lengths, and closer to the confidence intervals obtained as if the model was fixed a priori. Guided by this, given a (η, τ, ν)-stable model selection procedure $\widehat{\boldsymbol{M}}_n$ (*ibid.*, Definition 2) and to achieve (9), a valid $100(\delta + \tau + \nu)\%$ post-selection confidence interval (8) is obtained as follows: (i) when σ is known, $K = z(1 - \delta(1 - \nu)/(2|\widehat{\boldsymbol{M}}|e^{\eta}))$, where $z(1 - \alpha)$ is the $100(1 - \alpha)^{\text{th}}$ percentile of the standard normal distribution, (ii) when σ is unknown, $K = t(n - |\widehat{\boldsymbol{M}}|, 1 - \delta(1 - \nu)/(2|\widehat{\boldsymbol{M}}|e^{\eta}))$. Moreover, they also provide stabilized modifications of the LASSO, marginal screening and forward stepwise selection procedures. In comparison with the PoSI approach discussed in Section 6.1, their approach bypasses the computation of the PoSI constant. As such, it is more computationally efficient.

EPoSI theories: Besides the aforementioned advance in post-model-selection inference, the exact post-selection inference (EPoSI) and EPoSI for sequential regression (sequential EPoSI), respectively discussed in Sections 6.2 and 6.3, have also been further studied. Two main assumptions of the existing frameworks are that (i) the variance of responses, σ^2, is known, and (ii) the response $\boldsymbol{Y}$ follows a Gaussian distribution. To address the first assumption, Tian et al. (2018) apply the EPoSI approach to square-root LASSO (Belloni et al. 2011) for inference on selected submodel after model selection, where an estimate for σ^2 is derived based on the square-root LASSO. Tibshirani et al. (2018) propose an alternative solution via constructing a computationally efficient bootstrapped version of the truncated normal statistics in (16) which does not depend on σ^2. To remove the second assumption on Gaussianity of $\boldsymbol{Y}$, Tian & Taylor (2017) and Tibshirani et al. (2018) respectively examine large-sample conditional framework of the EPoSI and sequential EPoSI. Specifically, both work show that under certain conditions on the distribution of $\boldsymbol{Y}$, selection procedures and unknown regression coefficients, the pivotal quantity in (17) and its counterpart for sequential EPoSI converge ($n \to \infty$, p constant) to the uniform distribution so that subsequent construction of post-selection confidence interval can be conducted in the same fashion. Taylor & Tibshirani (2018) further explore the asymptotic aspects of EPoSI (assuming p is fixed) and show that this framework can be generalized for statistical inference of a large class of ℓ_1-penalized regression models, including generalized linear models, Cox's proportional hazards model, and the graphical LASSO (Friedman et al. 2008). Zhao, Small & Ertefaie (2021) employ the EPoSI approach in a two-stage proposal for solving effect modification problem and demonstrate theoretically and via simulations that this method is asymptotically valid. Hyun et al. (2018) extend the framework of EPoSI to the

so-called generalized LASSO estimators with specific examples of fused LASSO, trend filtering and graph fused LASSO and useful application to changepoint detection problem.

EPoSI and its extension: Apart from the above theoretical development of the EPoSI framework, progress has been made to extend this idea to several important problems in statistics, including change point detection, graphs, clustering, and regression trees, some of which we review below.

Hyun et al. (2021) study post-detection inference of change point with the following setup,

$$Y_i \sim N(\theta_i, \sigma^2),\ i = 1, ..., n, \tag{24}$$

where the mean vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top$ follows a piecewise constant structure. Specifically, for location indices $\boldsymbol{b} = (b_1, ..., b_n)^\top$, $1 \leq b_1 \ \cdots < b_t \leq n-1$, the mean parameters satisfy $0 = b_0$, $b_{t+1} = n$, $\theta_{b_j+1} = ... = \theta_{b_{j+1}}$, for $j = 0, ..., t$, where $t \in \{0\} \cup \mathbb{Z}^+$ is the number of change points. The goal is to use a data-dependent method to test whether there is a change in the mean at these locations. Upon characterizing the changepoint detection (using methods such as binary segmentation, wild binary segmentation, circular segmentation and fused LASSO) as polyhedral selection events in (14), Hyun et al. (2021) apply the EPoSI framework and obtain p-values for the hypothesis testing of interest. Based on (24) and using changepoint detection methods such as binary segmentation, ℓ_0 segmentation and fused LASSO, Jewell et al. (2021) propose an EPoSI-based construction that conditions on less information, resulting in tests with improved powers. On the other hand, Mehrizi & Chenouri (2021) study a variant of the change point detection problem where the underlying mean vector is a piecewise polynomial function. Under this setting, change points identified by a filtering algorithm forms a polyhedral set. Thus, the EPoSI scheme can be used to obtain p-values and confidence intervals. In addition, two extensions with less conditioning requirements, respectively conditioning on only the target change point, and only the target change point and its adjacent neighbors being in the selected change points set, are proposed to produce shorter confidence intervals.

Chen et al. (2021) present an EPoSI-based framework for testing the significance of a difference in the means of two connected components obtained from the graph fused LASSO. Their method conditions on less information and is shown to be more powerful than the method of Hyun et al. (2018). In clustering, Gao et al. (2021) consider testing for a difference in means of a pair of clusters identified via a data-dependent procedure. Upon showing that sample splitting is not a valid procedure in this case, the authors propose a method that conditions on the selected clusters and as opposed to (16), a truncated chi-squared counterpart is obtained in the computation of p-values. Neufeld et al. (2021) apply EPoSI on inference associated with the Classification and Regression Tree algorithm. Specifically, the authors obtain EPoSI-based p-values for testing a difference in the mean response between a pair of terminal nodes and confidence interval for the mean response within a single terminal.

7. Statistical inference for population-based target: bias correction and sampling techniques

In this section, our goal is to discuss another line of work on inference about the population-based target $\boldsymbol{\beta}^0$, even though these are not considered as post-model-selection methods. We review three methods, bias-correction, sampling techniques, and an optimization perspective for constructing p-values and confidence regions for $\boldsymbol{\beta}^0$ under the framework of linear regression model (1) (unless otherwise specified).

7.1. Bias-correction

The general idea of the methods in this category is to (asymptotically) remove or quantify the bias of a regularized estimator, such as the LASSO or ridge estimator, in estimating $\boldsymbol{\beta}^0$. Consequently, the distribution of the bias-corrected estimator can be approximated by a tractable distribution, which then enables us to construct p-values for testing the hypotheses $H_j : \beta_j^0 = 0$, for $j = 1, ..., p$, and confidence intervals for β_j^0's.

De-biasing LASSO: A well-known bias-correction technique is the de-biasing LASSO approach pioneered by Javanmard & Montanari (2014*b*) and Zhang & Zhang (2014), where the initial scaled LASSO estimator undergoes a bias-correction stage, leading to a so-called low-dimensional projection estimator (LDPE). The resulting corrected estimator thus becomes de-sparsified but asymptotically unbiased with a limiting Gaussian distribution. This idea is further studied by van der Geer et al. (2014) and Javanmard & Montanari (2014*a*).

The general form of a de-biasing LASSO estimator is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{DB}}(\lambda) = \widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}(\lambda) + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}(\lambda))/n,$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}(\lambda)$ is the LASSO estimator with tuning parameter λ in (6), and $\widehat{\boldsymbol{\Theta}}$ is a $p \times p$ matrix which has different constructions according to the aforementioned works. Assuming that the inverse of population variance-covariance matrix is sparse, one example of such construction is $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}^{-1}$, where $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top\boldsymbol{X}/n$ is the sample variance-covariance matrix, assuming the existence of its inverse. In high-dimensional settings, $\widehat{\boldsymbol{\Theta}}$ is obtained using the graphical LASSO (Friedman et al. 2008). By relaxing the sparsity assumption on inverse of the covariance, another example of $\widehat{\boldsymbol{\Theta}}$ is provided by Javanmard & Montanari (2014*a*), which is obtained by minimizing the error term $\boldsymbol{\Delta_n}$ and the variance of the Gaussian term $\boldsymbol{Z_n}$ in (25) below. Under different assumptions and constructions of $\widehat{\boldsymbol{\Theta}}$, and appropriate choice of a sequence of tuning parameters $\lambda = \lambda_n$, a uniform asymptotic normality result is established as follows,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{DB}}(\lambda_n) - \boldsymbol{\beta}^0) = \boldsymbol{Z_n} + \boldsymbol{\Delta_n}, \tag{25}$$

where

$$\boldsymbol{Z_n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Theta}}^\top), \quad \text{and} \quad \sup_{\boldsymbol{\beta}^0 \in \mathbb{R}^p : ||\boldsymbol{\beta}^0||_0 \leq s_0} ||\boldsymbol{\Delta_n}||_\infty = o_{\mathbb{P}}(1),$$

with $s_0 = |\{1 \leq j \leq p : \beta_j^0 \neq 0\}|$ being the size of the active set, which is commonly referred to the sparsity constant. One possible estimator for σ^2 is the scaled LASSO estimator (Zhang & Zhang 2014). Therefore, similar to the standard maximum likelihood theory, (25) is used to construct asymptotically valid confidence intervals for $\boldsymbol{\beta}^0$ and to obtain p-values for testing the hypotheses $H_j : \beta_j^0 = 0$, for $j = 1, ..., p$.

To compare these aforementioned four proposals (Zhang & Zhang 2014, Javanmard & Montanari 2014*a*,*b*, van der Geer et al. 2014), in addition to different construction of $\widehat{\boldsymbol{\Theta}}$, we note that both Javanmard & Montanari (2014*a*) and Javanmard & Montanari (2014*b*) also consider a random design matrix case, whereas the other two works focus on deterministic designs. Moreover, Javanmard & Montanari (2014*b*) require $n \geq s_0 \log(p/s_0)$. In contrast, Javanmard & Montanari (2014*a*) and van der Geer et al. (2014) require $n \gg (s_0 \log(p))^2$.

Bias-corrected ridge estimator: Bühlmann (2013) proposes a bias-correction method by focusing on correcting the bias of the ridge estimator. Define $\boldsymbol{\theta}^0$ by $\boldsymbol{\theta}^0 = \boldsymbol{P_X}\boldsymbol{\beta}^0$, where $\boldsymbol{P_X} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^-\boldsymbol{X}$, and "$-$" denotes the pseudo-inverse of a matrix. Consider the ridge estimator

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda ||\boldsymbol{\beta}||^2 \right\} = \frac{1}{n} \left(\widehat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I_p} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \quad (26)$$

where λ is the tuning parameter, $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top \boldsymbol{X}/n$ and $\boldsymbol{I_p}$ is the $p \times p$ identity matrix. Here, $p = p(n) \to \infty$ as $n \to \infty$. Bühlmann argues that the ridge estimator (26) is a reasonable estimator for $\boldsymbol{\theta}^0$ since the bias of $\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)$ in estimating $\boldsymbol{\theta}^0$ can be controlled in a fashion as discussed below in (28). Therefore, the bias of $\widehat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)$ in estimating the population-based regression coefficient $\boldsymbol{\beta}^0$ now can be decomposed componentwise as

$$\begin{aligned} \mathbb{E}(\widehat{\beta}_{\text{ridge},j}(\lambda) - \beta_j^0) &= \left(\mathbb{E}(\widehat{\beta}_{\text{ridge},j}(\lambda)) - \theta_j^0\right) + \left(\theta_j^0 - \beta_j^0\right) \\ &\leq |\mathbb{E}(\widehat{\beta}_{\text{ridge},j}(\lambda)) - \theta_j^0| + |\theta_j^0 - \beta_j^0|. \end{aligned} \quad (27)$$

Based on (27), Bühlmann (2013) then shows that $|\mathbb{E}(\widehat{\beta}_{\text{ridge},j}(\lambda)) - \theta_j^0|$ is a minor term which can be controlled, while $|\theta_j^0 - \beta_j^0|$, is a major term, which can be estimated by a newly-proposed post-corrected estimator given in (29). Specifically, for the minor bias term, we have

$$\max_j |\mathbb{E}(\widehat{\beta}_{\text{ridge},j}(\lambda)) - \theta_j^0| \leq \lambda ||\boldsymbol{\theta}^0|| \left(\sigma_{\min \neq 0}(\widehat{\boldsymbol{\Sigma}})\right)^{-1}, \ j = 1, ..., p, \quad (28)$$

where $\sigma_{\min \neq 0}(\widehat{\boldsymbol{\Sigma}})$ is the minimal non-zero eigenvalue of the matrix $\widehat{\boldsymbol{\Sigma}}$. In light of the bias-and-variance trade-off, λ is chosen to achieve a relatively small upper

bound for $|\mathbb{E}(\widehat{\beta}_{\text{ridge},j}(\lambda)) - \theta_j^0|$ while attaining a moderate variance of $\widehat{\beta}_{\text{ridge},j}(\lambda)$. Now, to address the major bias term $|\theta_j^0 - \beta_j^0|$, it is worthwhile to note that

$$\frac{\theta_j^0 - \beta_j^0}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} = \beta_j^0 - \frac{\beta_j^0}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} + \sum_{k:k=1,k\neq j}^{p} \frac{(\boldsymbol{P}_{\boldsymbol{X}})_{j,k}}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} \beta_k^0, \ j = 1, ..., p,$$

which implies

$$\beta_j^0 = \frac{\theta_j^0}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} - \sum_{k:k=1,k\neq j}^{p} \frac{(\boldsymbol{P}_{\boldsymbol{X}})_{j,k}}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} \beta_k^0.$$

Next, θ_j^0 is estimated by $\widehat{\beta}_{\text{ridge},j}(\lambda)$ and one possible initial estimator for β_k^0 in the second summand is the LASSO estimator, $\widehat{\beta}_{\text{LASSO},k}(\lambda)$. This leads to the post-correction ridge estimator

$$\widehat{\beta}_{\text{corr},j}(\lambda) = \frac{\widehat{\beta}_{\text{ridge},j}(\lambda)}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} - \sum_{k:k=1,k\neq j}^{p} \frac{(\boldsymbol{P}_{\boldsymbol{X}})_{j,k}}{(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}} \widehat{\beta}_{\text{LASSO},k}(\lambda), \ j = 1, ..., p. \tag{29}$$

Under a sparsity condition that $s_0 = |\{1 \leq j \leq p : \beta_j^0 \neq 0\}| = o((n/\log p)^{\xi})$ for some $0 < \xi < 1/2$, and compatibility condition for $\widehat{\boldsymbol{\Sigma}}$ (van der Geer 2007), Bühlmann (2013) shows

$$\sigma^{-1}\boldsymbol{V}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\text{corr}}(\lambda) - \boldsymbol{\beta}^0) \approx \boldsymbol{Z} + \sigma^{-1}\boldsymbol{V}^{-1/2}\boldsymbol{U}, \tag{30}$$

where

$$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$
$$\boldsymbol{U} = (U_1, ..., U_p) \text{ and } |U_j| \leq \widetilde{U}_j := \max_{i:i\neq j} |(\boldsymbol{P}_{\boldsymbol{X}})_{j,i}/(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}| \, (\log(p)/n)^{1/2-\kappa}.$$

Here, $\boldsymbol{V}$ is the variance-covariance matrix of the ridge estimator and κ is the parameter controlling sparsity. It is crucial to note that differing from the de-biasing LASSO estimator approach, the bias term $\boldsymbol{U}$ in (30) does not vanish asymptotically. In fact, the upper bound of $\boldsymbol{U}$ is carried over to the construction of confidence intervals for $\boldsymbol{\beta}^0$ and p-values for testing the hypotheses $H_j : \beta_j^0 = 0, j = 1, ..., p$. To be more specific, for example, the $100(1-\alpha)\%$ confidence interval for β_j^0 is

$$\big(\widehat{\beta}_{\text{corr},j}(\lambda) - \Delta_j, \widehat{\beta}_{\text{corr},j}(\lambda) + \Delta_j\big),$$

where

$$\Delta_j = \widetilde{U}_j + \frac{\sigma V_{j,j}^{-1/2}}{|(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}|} \Phi^{-1}(1 - \alpha/2),$$

and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal. On the other hand, the p-values for testing the hypotheses $H_j : \beta_j^0 = 0$ are given by

$$2\bigg(1 - \Phi\big(\sigma^{-1}V_{j,j}^{-1/2}|(\boldsymbol{P}_{\boldsymbol{X}})_{j,j}|(|\widehat{\beta}_{\text{corr},j}(\lambda)| - \widetilde{U}_j)_+\big)\bigg).$$

Other bias correction techniques: Apart from the aforementioned research in the linear regression context, there has been further developments on this topic under different settings, some of which are presented below.

Neykov et al. (2018) have proposed post-corrected parameter estimators and asymptotically valid confidence regions for parameters of interest in the context of estimating equations. The bias-correction step of this method is based on projecting an estimating equation under consideration onto a certain sparse direction, which is obtained via solving a tractable linear optimization problem. A Z-estimator (van der Vaart & Wellner 1996) for the parameter of interest is then obtained as the root of the projected estimating equation. Neykov et al. (2018) show that the resulting parameter estimator is consistent and asymptotically normal, hence facilitating the construction of a valid confidence region. This method is applicable to instrumental variable regression, graphical models, linear discriminant analysis and autoregressive models.

In the context of program evaluation and various causal inference applications, Belloni et al. (2015) have developed bias-correction methods targeting inference about causal effects. Specifically, using the absolute deviations (LAD) method for regression model with homoscedastic error, they construct uniformly valid confidence regions for the parameters of interest in the presence of a high-dimensional nuisance parameter. In contrast with the foregoing discussions, the bias-correction step of their method is based on an orthogonal moment equation for the parameters of interest. It has been shown that the resulting LAD regression parameter estimator is uniformly asymptotically normal, hence can be used to construct an asymptotically valid confidence region for the parameters of interest (Belloni et al. 2014, Belloni, Chernozhukov & Wei 2013).

7.2. Re-sampling

The re-sampling techniques such as data-splitting and bootstrapping are also used to perform valid inference for the population-based regression coefficient $\boldsymbol{\beta}^0$, which we review in this section.

Data-splitting: The main idea behind this method is aligned with Fisher's view discussed in the Introduction that the same data cannot be used for exploration and validation for statistical modeling. As such, a portion of the data is only used for model selection while the remaining portion is used to perform inference based on that selected submodel. Wasserman & Roeder (2009) propose a data splitting procedure, also referred to as screen and clean, in which the data are randomly divided into three parts, $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_3$, of approximately equal sizes. For a given tuning parameter λ, the LASSO is used to obtain an active set based on $\mathcal{D}_1$. Then cross validation based on $\mathcal{D}_2$ is used to select an optimal tuning parameter, yielding an estimated active set. Hypothesis testing for the regression coefficients corresponding to the estimated active set is then performed using the ordinary least-squares based on $\mathcal{D}_3$. This procedure produces one p-value for testing the effect of every selected covariate.

The resulting p-values in Wasserman & Roeder (2009) are sensitive to the randomness due to the data splitting, which can be a cause of concern when the sample size is small. To address the issue, Meinshausen et al. (2009) propose a so-called multi sample-splitting method, in which the data are randomly split into two parts B times, obtaining $(\mathcal{D}_1^b, \mathcal{D}_2^b), b = 1, ..., B$. This procedure results in B p-values for testing $H_0 : \beta_j^0 = 0, j = 1, ..., p$. They then adopt a quantile approach to aggregate these p-values and obtain a single p-value corresponding to each covariate, as described in Algorithm 2. For a selected submodel $\widehat{\boldsymbol{M}}$, having assumed the *sure screening property*, that is $\lim_{n\to\infty} \mathbb{P}(\boldsymbol{M}_{\text{true}} \subseteq \widehat{\boldsymbol{M}}) = 1$, and the *sparsity property*, that is $|\widehat{\boldsymbol{M}}| < n/2$, it is shown that the aggregated p-values are capable of (asymptotically) controlling the family-wise error rate and false discovery rate. Model selection procedures satisfying these two requirements include LASSO, L_2 boosting (Friedman 2001, Bühlmann 2006), orthogonal matching pursuit (Tropp & Gilbert 2007) and sure independent screening (Fan & Lv 2008). One key difference between the single and multi sample-splitting described above is that the former assigns p-values to only the selected covariates while the later assigns p-values to all the covariates.

Algorithm 2 Multi Sample-splitting Algorithm

1: Randomly split the data into two disjoint parts, $\mathcal{D}_1^b$ and $\mathcal{D}_2^b$, of approximately equal sizes $n/2$.

2: Estimate the active set $\widehat{\boldsymbol{M}}^b$ based on $\mathcal{D}_1^b$.

3: Based on $\mathcal{D}_2^b$, estimate the regression coefficients corresponding to the active set $\widehat{\boldsymbol{M}}^b$ using the ordinary least-squares and obtain p-values p_j^b corresponding to each $j \in \widehat{\boldsymbol{M}}^b$, and $p_j^b = 1$ for $j \notin \widehat{\boldsymbol{M}}^b$.

4: Define the adjusted p-values as $p_{\text{adj},j}^b = \min\left(p_j^b \times |\widehat{\boldsymbol{M}}^b|, 1\right)$ for $j = 1, ..., p$.

5: Repeat Steps 1 to 4 B times for $b = 1, ..., B$.

6: For $j = 1, ..., p$, obtain the aggregated p-values P_j over the B p-values, $\{p_{\text{adj},j}^b, b = 1, ..., B\}$, as

$$P_j = \min\left\{1, (1 - \log\gamma_{\min}) \inf_{\gamma\in(\gamma_{\min},1)} Q_j(\gamma)\right\},\ j = 1, ..., p,$$

where $\gamma \in (0, 1)$ and $\gamma_{\min}$ is a lower bound for γ, usually 0.05. Here, $Q_j(\gamma)$ is defined by

$$Q_j(\gamma) = \min\left\{1, q_\gamma\left(\left\{p_{\text{adj},j}^b/\gamma, b = 1, ..., B\right\}\right)\right\},\ j = 1, ..., p,$$

where $q_\gamma(\cdot)$ denotes the empirical γ-quantile function.

Khalili & Vidyashankar (2018) propose a multi sample-splitting approach which assigns p-values only to the selected covariates. Their framework is applicable to linear, generalized linear and finite mixture models, and is shown to attain asymptotic control of family-wise error rate at a given significance level.

To perform statistical inference on a low-dimensional parameter of interest in the presence of a high-dimensional nuisance parameter, Chernozhukov et al. (2018) propose an asymptotically and uniformly valid inference method that is based on cross-fitting. Specifically, in this method, the original data is first randomly partitioned into K folds with approximately equal sizes, indexed by I_j, $j = 1, ..., K$, and let $I_j^c = \{1, ..., n\} \setminus I_j$. For each $j = 1, ..., K$, an estimator for

the nuisance parameter is obtained using an existing machine learning method, such as the LASSO or boosting, based on the subset of data indexed by I_j^c. Then, for each $j = 1, ..., K$, an estimator for the parameter of interest is obtained via solving a Neyman orthogonal score equation upon plugging in the estimator for the nuisance parameter. A final estimator for the parameter of interest, referred to as the debiased machine learning (DML) estimator, is obtained by averaging over all the estimators across all $j = 1, ..., K$. Chernozhukov et al. (2018) theoretically prove that the DML estimator is asymptotically normal, which can then be used to construct asymptotically and uniformly valid confidence intervals for a parameter of interest. Though not being a paradigm of re-sampling procedures, as a closely related work, Belloni, Chernozhukov & Hansen (2013) consider a similar problem by focusing on statistical inference for a treatment effect in the presence of high-dimensional nuisance covariates x_j, $j \in \{1, ..., p\}$, in a partially linear regression model. Their development involves two variable selection steps using a version of the LASSO (called feasible LASSO) defined in Belloni et al. (2012): (i) selecting nuisance variables indexed by $M_1 \subseteq \{1, ..., p\}$ via regressing the treatment effect on all the x_j, and (ii) selecting nuisance variables indexed by $M_2 \subseteq \{1, ..., p\}$ via regressing the response on all the x_j. Then, an estimator for the treatment effect is obtained by regressing the response on the treatment effect and nuisance covariates indexed by $M_1 \cup M_2$. As such, this estimator is referred to as the post-double-selection (PDS) estimator. Similar to Chernozhukov et al. (2018), they show that the PDS estimator is asymptotically normal, leading to the construction of asymptotically and uniformly valid confidence interval for the parameter of interest, that is the treatment effect.

Bootstrapping: The idea of this approach is that the distribution of a penalized regression estimator, such as the LASSO, AdaLASSO estimator in (6) or de-biasing LASSO estimator (Zhang & Zhang 2014), can be well approximated by their bootstrap counterparts (see (33) below). Confidence intervals can then be constructed based on the distribution of the bootstrap estimators. More specifically, let

$$T_n = \sqrt{n}\big(\widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n) - \boldsymbol{\beta}^0\big),$$

where $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)$ is the LASSO estimator in (6). The goal is to approximate the distribution of T_n. To accomplish this, Chatterjee & Lahiri (2011) propose a residual-based bootstrap LASSO estimator, described in Algorithm 3 below. The bootstrap-based confidence region for $\boldsymbol{\beta}^0$ is given by

$$\text{R}_{\text{Boot}}(\alpha) = \big\{\boldsymbol{v} \in \mathbb{R}^p : \sqrt{n}||\boldsymbol{v} - \widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)|| \leq \widehat{t}_n(\alpha)\big\},\ \alpha \in (0, 1), \tag{31}$$

where $\widehat{t}_n(\alpha)$ is the α^{th} quantile of the empirical distribution of $\big\{||T^b_{\text{Boot}}||, b = 1, ..., B\big\}$. Here, T^b_{Boot} is defined by

$$T^b_{\text{boot}} = \sqrt{n}\big(\widetilde{\boldsymbol{\beta}}^b_{\text{Boot}}(\lambda_n) - \widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)\big), b = 1, ..., B, \tag{32}$$

where $\widetilde{\boldsymbol{\beta}}^b_{\text{Boot}}(\lambda_n)$ is the modified residual-based bootstrap LASSO estimator and $\widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)$ is the thresholded LASSO estimator, defined in Algorithm 3 [1].

Under certain conditions on the design matrix $\boldsymbol{X}$, the tuning parameter λ_n and error term $\boldsymbol{\epsilon}$ in (1), it is shown that as $n \to \infty$,

$$\rho\Big(\mathbb{P}\big(T_{\text{Boot}} \in E\big),\ \mathbb{P}\big(T_n \in E\big)\Big) \to 0,\ \text{for any } E \in \mathcal{B}(\mathbb{R}^p), \tag{33}$$

where T_{Boot} is the generic random variable representing $\big\{||T^b_{\text{Boot}}||, b = 1, ..., B\big\}$, $\mathcal{B}(\cdot)$ is the Borel σ-filed and $\rho(\cdot, \cdot)$ is the Lévy-Prohorov metric over the set of all probability measures on the space $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$. Using (33), it is then shown that the bootstrap-based confidence region $\text{R}_{\text{Boot}}(\alpha)$ in (31) is a uniformly asymptotically valid confidence region for $\boldsymbol{\beta}^0$, that is

$$\lim_{n\to\infty} \mathbb{P}\big(\boldsymbol{\beta}^0 \in \text{R}_{\text{Boot}}(\alpha)\big) = \alpha,\ \text{for all } \boldsymbol{\beta}^0 \in \mathbb{R}^p.$$

Algorithm 3 Residual-based Bootstrap LASSO Estimator

1: Obtain the thresholded LASSO estimator, $\widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)$, which is defined, component-wise, as
$$\widetilde{\beta}_{\text{LASSO},j}(\lambda_n) = \widehat{\beta}_{\text{LASSO},j}(\lambda_n)\mathbb{1}\big(|\widehat{\beta}_{\text{LASSO},j}(\lambda_n)| \geq a_n\big),\ j = 1, ..., p,$$
where $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)$ is the LASSO estimator in (6) and $\{a_n\}_{n=1}^{\infty}$ is a sequence of real numbers such that $a_n + (n^{-1/2}\log n)a_n^{-1} \to 0$ as $n \to \infty$.

2: Compute the modified residuals $\{\widetilde{r}_i, i = 1, ..., n\}$ by
$$\widetilde{r}_i = Y_i - \boldsymbol{x_i}^\top \widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n),\ i = 1, ..., n.$$

3: For $b = 1, ..., B$: draw a random sample $\{\widetilde{e}^b_1, ..., \widetilde{e}^b_n\}$ from the centered residuals $\{\widetilde{r}_i - 1/n\sum_{i=1}^n \widetilde{r}_i, i = 1, ..., n\}$ with replacement.

4: Set
$$\widetilde{Y}^b_i = \boldsymbol{x_i}^\top \widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n) + \widetilde{e}^b_i,\ i = 1, ..., n.$$

5: Then the modified residual-based bootstrap LASSO estimator, $\widetilde{\boldsymbol{\beta}}^b_{\text{Boot}}(\lambda_n)$, is given by
$$\widetilde{\boldsymbol{\beta}}^b_{\text{Boot}}(\lambda_n) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \sum_{i=1}^n (\widetilde{Y}^b_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 + \lambda_n ||\boldsymbol{\beta}||_1 \Big\}.$$

Steps 3 to 6 lead to the set $\big\{\widetilde{\boldsymbol{\beta}}^b_{\text{Boot}}(\lambda_n), b = 1, ..., B\big\}$, which is then used to obtain $\big\{||T^b_{\text{Boot}}||, b = 1, ..., B\big\}$ in (32).

Chatterjee & Lahiri (2013) apply the bootstrap technique based on the AdaLASSO estimator to approximate the distribution of a studentized version of T_n, denoted by $R_n = T_n/\widehat{\sigma}_n$, where $\widehat{\sigma}^2_n$ is the sample variance of the AdaLASSO-based residuals. They show analytically and via simulations that the bootstrap-based confidence intervals so obtained has better coverage compared to those obtained from the oracle-based normal approximation to the distribution of T_n, even when the regression parameter dimension is unbounded.

[1] For an appropriate choice of tuning parameter $\lambda = \lambda_n$ and threshold parameter as defined in Algorithm 3, see Section 5 of Chatterjee & Lahiri (2011).

Liu & Yu (2013) propose a residual-based bootstrap approach, where residuals are computed in a similar fashion as in Step 2 of Algorithm 3, where $\widetilde{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda_n)$ is replaced by a modified least-squares or ridge coefficient estimator, based on a submodel $\widehat{\boldsymbol{M}}$ selected by LASSO. The modified least-squares estimator is based on thresholding the singular values of the matrix $\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^{\top}\boldsymbol{X}_{\widehat{\boldsymbol{M}}}/n$. The reason for this modification is that the matrix inversion required in the OLS estimator (4) may not be stable when the smallest nonzero eigenvalue of $\boldsymbol{X}_{\widehat{\boldsymbol{M}}}^{\top}\boldsymbol{X}_{\widehat{\boldsymbol{M}}}/n$ is close to 0. Similar to Chatterjee & Lahiri (2011), the asymptotically valid confidence region $\text{R}_{\text{Boot}}(\alpha)$ in (31) can then be constructed.

On the other hand, focusing on valid statistical inference for an individual or a group of entries of $\boldsymbol{\beta}^0$, Dezeure et al. (2017*b*) study the application of bootstrap procedures to the de-biasing LASSO estimator. They demonstrate that bootstrapping the de-biasing LASSO estimator is capable of producing asymptotically valid: p-values for testing $H_0 : \beta_j^0 = 0$, $j \in G \subseteq \{1, ..., p\}$; and simultaneous confidence region for $\beta_j^0, j \in G \subseteq \{1, ..., p\}$. Specifically, three versions of bootstrap methods are presented, namely the residual bootstrap, multiplier wild bootstrap, and a so-called paired bootstrap method. For each of these procedures, they show the consistency of the resulting bootstrapped estimator in the sense of (33).

7.3. Optimization Perspective

Apart from the bias-correction and re-sampling techniques, methods from optimization theory, in particular techniques associated with variational inequalities (Hartman & Stampacchia 1966) and normal maps (Robinson 1992, 1993), have been used to provide valid asymptotic confidence intervals for $\boldsymbol{\beta}^0$. Generally speaking, a local minimizer of an objective function over a compact and convex set S in $\mathbb{R}^n$ is also a solution to a so-called variational inequality involving the gradient of the objective function. Moreover, the variational inequality can be equivalently written as a function, called a normal map, which is induced by the objective function and the set S. These two mathematical objects can then be connected by the fact that a solution in S to the variational inequality is a root to the associated normal map.

Assuming that the design matrix $\boldsymbol{X}$ is random and the dimension p is fixed, Lu et al. (2017) consider solving the so-called random-design population version of the LASSO problem

$$\min_{\beta_0, \boldsymbol{\beta}} \left[\mathbb{E}\big(Y - \beta_0 - \sum_{j=1}^{p} \beta_j x_j\big)^2 + \lambda ||\boldsymbol{\beta}||_1 \right], \tag{34}$$

where the solution to (34), denoted by $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$, is called the population penalized (LASSO) parameter. By transforming (34) to its corresponding variational inequality and normal map formulations, the authors propose a four-step method to construct asymptotically valid confidence interval for $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$ based on the LASSO estimator $\widehat{\boldsymbol{\beta}}_{\text{LASSO}}(\lambda)$:

Step 1: Transform the minimization problem in (34) to its normal maps;
Step 2: Obtain the asymptotic distributions for the solutions to the normal maps;
Step 3: Construct individual and simultaneous confidence intervals for the solutions obtained in Step 2;
Step 4: Assuming the linear regression model in (1) is the true model, convert the confidence intervals obtained in Step 3 to the ones for the population-based target $\boldsymbol{\beta}^0$.

This framework is extended by Yu et al. (2020) to penalized regression methods based a general penalty term $\mathrm{J}_n(\boldsymbol{\beta}; \boldsymbol{\lambda})$, including the adaptive LASSO, log penalty (Friedman 2012), transformed ℓ_1 penalty (Nikolova 2000), SCAD and the MCP (Zhang 2010).

8. Simulation study

In this section, we compare various post-selection confidence intervals through simulated data. In light of the simulation results by Dezeure et al. (2015) for the population-based targets, here our focal point is a comparison between the confidence intervals for the projection-based targets using the PoSI and EPoSI approaches, though we also include the naïve and Scheffé's confidence intervals. Details of data generation are presented in Section 8.1. Assessing metrics are defined in Section 8.2. Specifics pertinent to method, code and implementation are described in Section 8.3, followed by an analysis of simulation results in Section 8.4. Since the PoSI approach is computationally expensive for $p > 20$, we design simulations for $p \leq 20$ so that the PoSI and EPoSI confidence intervals can be compared.

8.1. Data generation

Design matrices $\boldsymbol{X}$ are simulated with $n = 40$ and 60, and $p = p(n) = \lceil n^{2/3} \rceil + 2$, where $\lceil \cdot \rceil$ denotes the ceiling function. Thus, the number of covariates are respectively 14 and 18 for $n = 40$ and 60. With these dimensions, *five* types of design matrix with the following variance-covariance structures are considered:

1. **Equicorrelated covariance**: $\boldsymbol{X}_{\mathrm{Eq}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$, where the (i, j)-entry of $\boldsymbol{\Sigma}_1$ is 0.8 if $i \neq j$, and 1 if $i = j$.
2. **Toeplitz covariance**: $\boldsymbol{X}_{\mathrm{Tp}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_2)$, where the (i, j)-entry of $\boldsymbol{\Sigma}_2$ is $0.9^{|i-j|}$.
3. **Exponential decay covariance**: $\boldsymbol{X}_{\mathrm{Ep}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_3})$, where the (i, j)-entry of $\boldsymbol{\Sigma}_3^{-1}$ is $0.5^{|i-j|/5}$.
4. **Exchangeable design matrix**: Define $\boldsymbol{X}^{(p)}(a)$ as in Example 1 of Berk et al. (2013), which is formulated as $\boldsymbol{X}^{(p)}(a) = \boldsymbol{I}_p + a\boldsymbol{E}_p$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and $\boldsymbol{E}_p$ is a $p \times p$ matrix with all entries equal to 1. Then we adopt the same approach as proposed by Leeb et al. (2015), who set $a = 10$ and $\boldsymbol{X}_{\mathrm{Ec}} = \boldsymbol{U}\boldsymbol{X}^{(p)}(a)$, where $\boldsymbol{U}$ has dimension $n \times p$ under

low-dimensional setting and is formed by putting together p orthonormal n-tuples obtained by first draw a set of $i.i.d.$ standard Gaussian n-tuples and then applying the Gram-Schmidt procedure.

5. **Design matrix from a real data set**: we include a design matrix (denoted by $\boldsymbol{X}_{\mathrm{Rl}}$) from the standardized diabetes data from Efron et al. (2004). $\boldsymbol{X}_{\mathrm{Rl}}$ has $n = 442$ observations and $p = 10$ covariates. The 10 covariates are the first 10 columns of the diabetes data, namely age (`age`), sex (`sex`), body mass index (`BMI`), average blood pressure (`map`) and six blood serum measurements (`tc`, `ldl`, `hdl`, `tch`, `ltg`, `glu`).

Given a design matrix $\boldsymbol{X}$ according to any of the scenarios 1-5, it remains fixed throughout the simulations. For the response vector $\boldsymbol{Y}$, we noticed in Section 5.2 that the construction of the PoSI and EPoSI confidence intervals do not require the first order correctness condition $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}^0$. In our simulations, yet, we assume this condition for simplicity. In other words, the $n \times 1$ dimensional response vectors $\boldsymbol{Y}$ are generated from the linear regression model (1). Given a design matrix $\boldsymbol{X}$, we generate 10000 copies of $\boldsymbol{Y}$, randomly from *four* linear regression models with error variance $\sigma^2 = 2$ and the regression coefficients specified in TABLE 2.

TABLE 2
Settings for the true linear data generating models.

True Models	$\boldsymbol{M}_{\mathrm{true}}$	Signal Strengths for the Non-zero β_j^0's
Model I	$\{1, 2, 3\}$	$\beta_j^0 \sim \mathrm{Unif}(2, 3)$ for $j = 1, 2$ and 3
Model II	$\{1, 2, 3, 4, 5\}$	$\beta_j^0 \sim \mathrm{Unif}(2, 3)$ for $j = 1, 2, 3, 4$ and 5
Model III	$\{1, 2, 3\}$	$\beta_j^0 \sim \mathrm{Unif}(4, 5)$ for $j = 1, 2$ and 3
Model IV	$\{1, 2, 3, 4, 5\}$	$\beta_j^0 \sim \mathrm{Unif}(4, 5)$ for $j = 1, 2, 3, 4$ and 5

Models I and II have weaker signal strengths compared to Models III and IV.

8.2. Assessing metrics

We set the nominal coverage probability for the confidence intervals to be $1 - \alpha = 0.95$. To compare various post-selection confidence intervals, we consider two assessing metrics: the empirical average coverage probability ($\mathrm{CP}_{j\cdot\boldsymbol{M}}$), and the empirical average length of confidence intervals ($\mathrm{L}_{j\cdot\boldsymbol{M}}$), conditioning on a submodel $\boldsymbol{M}$ selected by $\widehat{\boldsymbol{M}}_n$. In particular,

$$\mathrm{CP}_{j\cdot\boldsymbol{M}} = \frac{\sum_{i=1}^{10000} \mathbb{1}\{b_{j\cdot\boldsymbol{M}} \in \mathrm{CI}_{j\cdot\boldsymbol{M}}^i\}}{\sum_{i=1}^{10000} \mathbb{1}\{\widehat{\boldsymbol{M}} = \boldsymbol{M}\}},$$

and

$$\mathrm{L}_{j\cdot\boldsymbol{M}} = \frac{\sum_{i=1}^{10000} \mathrm{Length}\{\mathrm{CI}_{j\cdot\boldsymbol{M}}^i\}}{\sum_{i=1}^{10000} \mathbb{1}\{\widehat{\boldsymbol{M}} = \boldsymbol{M}\}},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, $\mathrm{CI}_{j\cdot\boldsymbol{M}}^i$ is the confidence interval for the j^{th} projection-based regression coefficient $b_{j\cdot\boldsymbol{M}}$ within the submodel $\boldsymbol{M}$ corresponding to the i^{th} random sample, and $\mathrm{Length}\{\cdot\}$ is the length of the confidence

interval. In case of the EPoSI confidence intervals, the above quantities are also computed by conditioning on both a selected submodel $\boldsymbol{M}$ and the sign vector $\boldsymbol{s}$ of its associated regression parameter estimate(s).

8.3. Method, code and implementation

To satisfy the polyhedral selection property required by the EPoSI method, in our simulations we use the LASSO with a *fixed* tuning parameter λ as the model selector. The λ is chosen according to the strategy outlined in Section 4 of Negahban et al. (2012), who propose $\lambda = 2\mathbb{E}(||\boldsymbol{X}^\top \boldsymbol{\epsilon}||_\infty)$ so that convergence rate for errors of the LASSO estimate(s) can be controlled. Given a design matrix $\boldsymbol{X}$, we compute an empirical version of λ by randomly generating 1000 samples of $\boldsymbol{\epsilon}$ from the multivariate normal $\mathcal{N}(\boldsymbol{0}, \widehat{\sigma}^2 \boldsymbol{I})$, where $\widehat{\sigma}^2$ is the residual mean square obtained from the full model.

We compare the following five post-selection confidence intervals:

Naïve, Scheffé, PoSI, EPoSI1 and EPoSI2,

where EPOSI1 is the EPoSI confidence interval constructed by conditioning on a selected submodel $\boldsymbol{M}$, and EPoSI2 is the one derived by conditioning on both a selected submodel $\boldsymbol{M}$ and the sign vector $\boldsymbol{s}$ of its associated regression parameter estimate(s).

We use the functions `PoSI` and `summary.PoSI` from the `R` package `PoSI` (Buja & Zhang 2015) to compute the PoSI-constant and corresponding confidence intervals. For the EPoSI confidence intervals, we use the functions `selection.int`, `grid.search` and `myptruncnorm` [2].

8.4. Analysis of simulation results

The simulation results for Model I and the design matrices $\boldsymbol{X}_{\mathrm{Eq}}$, $\boldsymbol{X}_{\mathrm{Ec}}$ and $\boldsymbol{X}_{\mathrm{Rl}}$, with dimension $n = 40$ and $p = 14$, are respectively given in TABLE 3, 4 and 5. Each table contains the empirical average coverage probability ($\mathrm{CP}_{j\cdot\widehat{\boldsymbol{M}}}$), and the empirical average length of confidence intervals ($\mathrm{L}_{j\cdot\widehat{\boldsymbol{M}}}$), corresponding to the projection-based regression coefficients $b_{j\cdot\widehat{\boldsymbol{M}}}$, $j = 1$, 2 and 3, for the *three* most frequently submodels selected by the LASSO that contain $\boldsymbol{x}_j$ (sorted by the empirical model selection frequency in descending order). Simulation results corresponding to (1) the two other design matrices, $\boldsymbol{X}_{\mathrm{Tp}}$ and $\boldsymbol{X}_{\mathrm{Ep}}$, under Model I with dimension $n = 40$ and $p = 14$; (2) all the five design matrices under Models II, III and IV with dimension $n = 40$ and $p = 14$; (3) all the five design matrices under Models I, II, III and IV with dimension $n = 60$ and $p = 18$, show similar behaviors and are provided in Section S5 of the Supplementary Material.

From TABLE 3 to 5, we have observed a mix of under- and satisfactory coverage of naïve confidence intervals for projected regression coefficients. Such

[2]The `R` scripts of these three functions are kindly provided by Professor Jonathan Taylor from the Department of Statistics, Stanford University.

TABLE 3
$CP_{j\cdot\widehat{M}}$ and $L_{j\cdot\widehat{M}}$, $j = 1$, 2 and 3, of post-selection confidence intervals based on Model I and $\boldsymbol{X}_{Eq}$.

	Post-selection Confidence Intervals									
	Naïve		Scheffé		PoSI		EPoSI1		EPoSI2	
Top Selected Submodels	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$
$\{1,2,3\}$	.961	1.56	1.00	4.16	1.00	3.19	.949	1.51	.949	1.51
$\{1,2,3,11\}$	.992	1.61	1.00	4.30	1.00	3.29	.975	1.56	.966	2.63
$\{1,2,3,5\}$	.913	1.69	1.00	4.52	1.00	3.46	.896	1.93	.946	3.04
Top Selected Submodels	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$
$\{1,2,3\}$	.953	1.81	1.00	4.84	1.00	3.70	.945	1.75	.945	1.76
$\{1,2,3,11\}$	.934	1.94	1.00	5.18	1.00	3.96	.926	1.87	.922	3.13
$\{1,2,3,5\}$	.922	1.92	1.00	5.13	1.00	3.92	.913	1.86	.928	3.35
Top Selected Submodels	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$
$\{1,2,3\}$	.941	1.80	1.00	4.82	1.00	3.69	.941	1.74	.941	1.75
$\{1,2,3,11\}$	.975	1.88	1.00	5.02	1.00	3.84	.967	1.82	.983	3.00
$\{1,2,3,5\}$	.939	1.82	1.00	4.85	1.00	3.71	.939	1.76	.947	2.58

TABLE 4
$CP_{j\cdot\widehat{M}}$ and $L_{j\cdot\widehat{M}}$, $j = 1$, 2 and 3, of post-selection confidence intervals based on Model I and $\boldsymbol{X}_{Ec}$.

	Post-selection Confidence Intervals									
	Naïve		Scheffé		PoSI		EPoSI1		EPoSI2	
Top Selected Submodels	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$
$\{1,2,3\}$	.947	4.64	1.00	12.4	.999	9.47	.937	4.82	.958	5.75
$\{1,2\}$	.965	4.01	1.00	10.7	1.00	8.18	.960	3.90	1.00	4.47
$\{1,2,3,5\}$	.916	4.91	1.00	13.1	1.00	9.99	.907	5.13	.750	9.16
Top Selected Submodels	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$
$\{1,2,3\}$	.946	4.64	1.00	12.4	1.00	9.47	.947	4.62	.949	5.51
$\{1,2\}$	.965	4.01	1.00	10.7	1.00	8.18	.957	3.90	1.00	4.47
$\{1,2,3,5\}$	.903	4.91	1.00	13.1	1.00	9.99	.911	4.86	.833	9.58
Top Selected Submodels	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$
$\{1,2,3\}$	.962	4.64	1.00	12.5	.999	9.47	.966	4.79	.966	5.66
$\{1,2,3,5\}$	.987	4.91	1.00	13.1	1.00	9.99	.985	5.09	1.00	7.75
$\{1,2,3,4\}$	1.00	4.95	1.00	13.2	1.00	10.1	1.00	5.11	1.00	6.50

phenomena depend on the types of design matrices under consideration. For example, in the case of $\boldsymbol{X}_{\mathrm{Rl}}$, the coverage probabilities for all the three projected regression coefficients $b_{j\cdot\widehat{M}}$, $j = 1$, 2 and 3, fall below the desired nominal cov-

TABLE 5

$CP_{j\cdot\widehat{M}}$ and $L_{j\cdot\widehat{M}}$, $j = 1$, 2 and 3, of post-selection confidence intervals based on Model I and $\boldsymbol{X}_{Rl}$.

	Post-selection Confidence Intervals									
	Naïve		Scheffé		PoSI		EPoSI1		EPoSI2	
Top Selected Submodels	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$	$CP_{1\cdot\widehat{M}}$	$L_{1\cdot\widehat{M}}$
{1}	.946	5.56	1.00	12.2	.999	9.65	.945	5.84	.962	15.5
{1, 2}	.952	5.63	1.00	12.3	.999	9.78	.953	5.74	.963	19.2
{1, 3}	.973	5.66	1.00	12.4	1.00	9.81	.972	5.67	.972	20.3
Top Selected Submodels	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$	$CP_{2\cdot\widehat{M}}$	$L_{2\cdot\widehat{M}}$
{2}	.929	5.57	1.00	12.2	1.00	9.67	.929	5.65	.979	17.5
{1, 2}	.939	5.63	1.00	12.3	.999	9.78	.938	5.73	.974	20.2
{2, 3}	.936	5.59	1.00	12.2	1.00	9.70	.936	5.69	.971	19.2
Top Selected Submodels	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$	$CP_{3\cdot\widehat{M}}$	$L_{3\cdot\widehat{M}}$
{1, 3}	.944	5.66	1.00	12.4	1.00	9.81	.964	5.82	.991	20.1
{3}	.940	5.54	1.00	12.1	1.00	9.61	.940	6.17	.973	16.7
{2, 3}	.948	5.59	1.00	12.2	1.00	9.70	.948	5.85	.979	18.9

erage rate (between .929 and .948), except for $CP_{1\cdot\{1,2\}}$ and $CP_{1\cdot\{1,3\}}$. On the other hand, satisfactory coverage in general is observed in the case of $\boldsymbol{X}_{\mathrm{Ec}}$, where under-coverage only exists for $CP_{1\cdot\{1,2,3,5\}}$ and $CP_{2\cdot\{1,2,3,5\}}$. Moreover, in the case of $\boldsymbol{X}_{\mathrm{Eq}}$, both under- and satisfactory coverage are seen, which have approximately equal proportion. This empirical observation agrees with Leeb et al. (2015), indicating that although naïve confidence intervals in general fail to achieve desired nominal coverage probabilities for the population-based regression coefficients (also reported in TABLE 1), both under- and satisfactory coverage can be achieved for the projected regression coefficients. In addition, the extent of the under-coverage is moderate, depending on the specific structures of the design matrices.

In contrast, the Scheffé and PoSI confidence intervals consistently achieve over-coverage for all the projected regression coefficients, under all the design matrices considered here. Focusing on the EPoSI confidence intervals, empirical results demonstrate that similar to the naïve confidence intervals, a mix of under- and satisfactory coverage exists for both EPoSI1 and EPoSI2. For example, although over-coverage is attained by EPoSI2 for all the projected coefficients in the case of $\boldsymbol{X}_{\mathrm{Rl}}$, under-coverage is seen under the other two design matrices ($CP_{2\cdot\{1,2,3,11\}}$ in the case of $\boldsymbol{X}_{\mathrm{Eq}}$ for instance). This phenomenon also holds for EPoSI1. Furthermore, we can see that the coverage probabilities of EPoSI1 are aligned with those attained by the naïve confidence intervals.

In terms of the empirical average length ($L_{j\cdot\widehat{M}}$), from TABLE 3 to 5 we observe that the Scheffé confidence intervals are always wider than the PoSI confidence intervals. In fact, the ratio of the empirical average interval length

of these two confidence intervals is between 1.25 and 1.32, under the three design matrices $\boldsymbol{X}_{\mathrm{Eq}}$, $\boldsymbol{X}_{\mathrm{Ec}}$ and $\boldsymbol{X}_{\mathrm{Rl}}$. This observation agrees with the fact that the Scheffé's confidence intervals are more conservative (wider) than the PoSI confidence intervals, as discussed in Section 6.1.

Moreover, it is worthwhile to notice that conditioning on a selected submodel only, the EPoSI confidence intervals (EPoSI1) are always shorter than those obtained by conditioning on both a selected model and sign vector of corresponding point estimator(s) (EPoSI2). Furthermore, EPoSI1 are close to (or even shorter than) the naïve confidence intervals. For example, in the case of $\boldsymbol{X}_{\mathrm{Eq}}$ as reported in TABLE 3, the average length $\mathrm{L}_{1\cdot\{1,2,3\}}$ of EPoSI1, being 1.51, is shorter than that of naïve confidence interval, being 1.56. But in the case of $\boldsymbol{X}_{\mathrm{Ec}}$ as reported in TABLE 4, $\mathrm{L}_{1\cdot\{1,2,3,5\}}$ of EPoSI1 is wider. On the other hand, EPoSI2 are always wider than the naïve confidence intervals.

Comparing the EPoSI confidence intervals with those based on the PoSI and Scheffé, we have observed that under all the models and design matrices considered here, EPoSI1 are always shorter. In contrast, EPoSI2 can sometimes be surprisingly wider. As an example, for the projected regression coefficient $b_{1\cdot\boldsymbol{M}}$, in the case of $\boldsymbol{X}_{\mathrm{Rl}}$ (TABLE 5), when conditioning on the submodel $\{1,3\}$ and sign vector of its associated parameter point estimator(s), the ratio of empirical average interval length of EPoSI2 over the PoSI and Scheffé confidence intervals are approximately 2.07 and 1.63, respectively. This phenomenon is consistent with the explanations in Lee et al. (2016): when the signal is relatively week (as in Model I), the truncated Gaussian variable, $\boldsymbol{\eta}_j^{\top}\boldsymbol{Y}$ in (16), is near the boundary points of truncated intervals, leading to wider confidence intervals.

Yet, we wish to emphasize that the competition between the PoSI confidence intervals and EPoSI2 does depend on other factors as well, such as the design matrix and a particular selected submodel. For instance, in the case of $\boldsymbol{X}_{\mathrm{Eq}}$ and as seen in TABLE 3, the average length $\mathrm{L}_{2\cdot\{1,2,3\}}$ of the PoSI confidence interval, being 3.70, is larger than that of the EPoSI2, being 1.76.

In our simulation study, we have also observed that the EPoSI confidence intervals, EPoSI1 and EPoSI2, sometimes have *infinite* lengths. Therefore, the results corresponding to these intervals reported in TABLE 3 to 5 are only associated with the *finite* EPoSI confidence intervals. To be more specific, in TABLE 6 to 8, under Model I and the design matrices $\boldsymbol{X}_{\mathrm{Eq}}$, $\boldsymbol{X}_{\mathrm{Ec}}$ and $\boldsymbol{X}_{\mathrm{Rl}}$, we report the percentages of the time that the EPoSI confidence intervals have *finite* lengths, out of the total number of such intervals constructed for each selected submodel. As we can see from these three tables, the percentages of finite EPoSI1 and EPoSI2 differ dramatically according to the types of design matrix. In our simulations, both EPoSI1 and EPoSI2 in the case of $\boldsymbol{X}_{\mathrm{Ec}}$ have the lowest such percentages. For example, for the projected regression coefficient $b_{1\cdot\{1,2\}}$, only 75.3% of the EPoSI1 are finite, while almost all the EPoSI2 are infinite. Moreover, we also see that the percentages of finite EPoSI1 are always higher than those of EPoSI2. In a recent paper, Kivaranovic & Leeb (2021) show that the expected length of EPoSI2 interval is infinite, and they have also derived necessary and sufficient conditions under which the expected length of EPoSI1 interval is infinite.

TABLE 6
Percentages of finite EPoSI confidence intervals based on Model I and $\boldsymbol{X}_{Eq}$.

Projected Targets	Top Selected Submodels	EPoSI Confidence Intervals	
		EPOSI1	EPoSI2
$b_{1\cdot\widehat{M}}$	{1,2,3}	100%	99.6%
	{1,2,3,11}	99.2%	95.9%
	{1,2,3,5}	100%	96.5%
$b_{2\cdot\widehat{M}}$	{1,2,3}	100%	99.6%
	{1,2,3,11}	100%	95.1%
	{1,2,3,5}	100%	96.5%
$b_{3\cdot\widehat{M}}$	{1,2,3}	100%	99.6%
	{1,2,3,11}	100%	95.9%
	{1,2,3,5}	100%	98.3%

TABLE 7
Percentages of finite EPoSI confidence intervals based on Model I and $\boldsymbol{X}_{Ec}$.

Projected Targets	Top Selected Submodels	EPoSI Confidence Intervals	
		EPOSI1	EPoSI2
$b_{1\cdot\widehat{M}}$	{1,2,3}	93.5%	3.51%
	{1,2}	75.3%	.372%
	{1,2,3,5}	93.3%	4.01%
$b_{2\cdot\widehat{M}}$	{1,2,3}	56.8%	3.48%
	{1,2}	75.3%	.372%
	{1,2,3,5}	45.2%	4.01%
$b_{3\cdot\widehat{M}}$	{1,2,3}	56.4%	3.51%
	{1,2,3,5}	45.2%	4.01%
	{1,2,3,4}	50.7%	6.16%

TABLE 8
Percentages of finite EPoSI confidence intervals based on Model I and $\boldsymbol{X}_{Rl}$.

Projected Targets	Top Selected Submodels	EPoSI Confidence Intervals	
		EPOSI1	EPoSI2
$b_{1\cdot\widehat{M}}$	{1}	100%	95.7%
	{1,2}	99.7%	78.0%
	{1,3}	99.7%	75.1%
$b_{2\cdot\widehat{M}}$	{2}	99.9%	96.1%
	{1,2}	99.7%	77.8%
	{2,3}	100%	72.7%
$b_{3\cdot\widehat{M}}$	{1,3}	99.5%	73.7%
	{3}	99.8%	95.7%
	{2,3}	99.4%	73.3%

9. Conclusion

In this section, we summarize our observations and conclusions on various constructions of the confidence intervals for the projection- and the population-based regression coefficients.

Projection-based regression coefficients

1. Both EPoSI1 and EPoSI2 can have infinite lengths depending on the type of the design matrix. Moreover, EPoSI1 is more likely to be of finite length

than EPoSI2. This observation has also been discussed analytically by Kivaranovic & Leeb (2021). In view of this observation, we divide our discussions below into two groups, which are (i) either EPoSI1 or EPoSI2 is of finite length; (ii) neither is of finite length. The point 2 below pertains to the former case, while point 3 is about the latter.

2. p is relatively small (recommended cut-off point by Berk et al. (2013) is 20): In this scenario, it is practical to apply both the PoSI and the EPoSI methods. Considering coverage probabilities, under-coverage for the projected regression coefficients exists for both EPoSI1 and EPoSI2. In comparison, the PoSI confidence intervals consistently achieve the desired nominal coverage probability. In terms of the length of confidence intervals, the EPoSI1 are the shortest among all the competing confidence intervals and they can be even shorter than the naïve confidence intervals. On the other hand, the competition between the EPoSI2 and PoSI depends on several factors, one of which is the type of the design matrix. In some cases, the EPoSI2 can be wider than the PoSI confidence intervals.
   For p relatively large: In this case, the computational cost of the PoSI approach can be high. Similarly, since $2^{|\widehat{\boldsymbol{M}}|}$ calculations are required to obtain the EPoSI1, this interval is not recommended either due to heavy computational loads if p is too large. Therefore, in high-dimensional data where p is large, the EPoSI2 is more computationally feasible among the competing post-selection confidence intervals.
3. When neither EPoSI1 nor EPoSI2 is of finite length: In view of the computational cost associated with the PoSI confidence intervals in this case and having EPoSI1 and EPoSI2 of infinite lengths, we conclude that more research and new ideas are required in this case.

Population-based regression coefficients

Here, we summarize the conclusions made by Dezeure et al. (2015), who did a thorough study on the post-selection inference for the population-based targets. The authors compare the empirical coverage probabilities and lengths of confidence intervals constructed using the methods proposed by Zhang & Zhang (2014), Javanmard & Montanari (2014*a*), Bühlmann (2013), Meinshausen et al. (2009), Liu & Yu (2013) and Chatterjee & Lahiri (2013) through an empirical study, and make the following observations: (1) the empirical coverage probability of confidence intervals for the true zero regression coefficients is satisfactory for all the methods. Moreover, the empirical coverage probability for the true non-zero coefficients is in accord with the p-values for the corresponding hypothesis tests; (2) the empirical coverage probability for the smallest true non-zero coefficient (i.e. the weakest signal) is very low for all the methods; (3) confidence intervals constructed using the multi sample-splitting (Meinshausen et al. 2009) and ridge-based bias correction (Bühlmann 2013) approaches are in general wider than those derived from the other methods.

Acknowledgments

The authors would like to thank the co-editor Professor Richard Lockhart, an associate editor, and three referees for their thoughtful and constructive comments. This work is based on the master thesis of Dongliang Zhang written in the department of Mathematics and Statistics at McGill University. Dongliang Zhang also thanks Professors Martin Lindquist and Mei-Cheng Wang, his PhD advisors at Johns Hopkins University, for their support during the completion of this work.

Supplementary Material

Post-Model-Selection Inference in Linear Regression Models: An Integrated Review Supplementary Material
(doi: 10.1214/22-SS135SUPP; .pdf). Further details and discussions about properties of the methods presented in the paper are provided in the supplementary materials. The file also includes more simulation studies under different scenarios.

References

Azais, J., Castro, Y. D. & Mourareau, S. (2018), 'Power of the spacing test for least-angle regression', *Bernoulli* **24**, 465–492. MR3706766

Bachoc, F., Blanchard, G. & Neuvial, P. (2018), 'On the post selection inference constant under restricted isometry properties', *Electronic Journal of Statistics* **12**, 3736–3757. MR3878579

Bachoc, F., Leeb, H. & Pötscher, B. M. (2019), 'Valid confidence intervals for post-model-selection predictors', *The Annals of Statistics* **47**, 1475–1504. MR3911119

Bachoc, F., Preinerstorfer, D. & Steinberger, L. (2020), 'Uniformly valid confidence intervals post-model-selection', *The Annals of Statistics* **48**, 440–463. MR4065169

Bancroft, T. A. & Han, C.-P. (1977), 'Inference based on conditional specification: a note and a bibliography', *International Statistical Review* **45**, 117–127. MR0494634

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* **80**, 2369–2429. MR3001131

Belloni, A., Chernozhukov, V. & Hansen, C. (2013), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**, 608–650. MR3207983

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), 'Inference on treatment effects after selection amongst high-dimensional controls', *Review of Economic Studies* **81**, 608–650. MR3207983

Belloni, A., Chernozhukov, V. & Kato, K. (2015), 'Uniform post-selection inference for least absolute deviation regression and other z-estimation problems', *Biometrika* **102**, 77–94. MR3335097

Belloni, A., Chernozhukov, V. & Wang, L. (2011), 'Square-root lasso: pivotal recovery of sparse signals via conic programming', *Biometrika* **98**, 791–806. MR2860324

Belloni, A., Chernozhukov, V. & Wei, Y. (2013), 'Honest confidence regions for logistic regression with a large number of controls', *arXiv preprint arXiv preprint arXiv:1304.3969*.

Benjamini, Y. (2020), 'Selective inference: The silent killer of replicability', *Harvard Data Science Review* **2**.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013), 'Valid post-selection inference', *The Annals of Statistics* **41**, 802–837. MR3099122

Berk, R., Brown, L. D. & Zhao, L. (2009), 'Statistical inference after model selection', *Journal of Quantitative Criminology* **26**, 217–236.

Box, G. E. P. (1976), 'Science and statistics', *Journal of the American Statistical Association* **71**, 791–799. MR0431440

Breiman, L. (1992), 'The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error', *Journal of the American Statistical Association* **87**, 738–754. MR1185196

Bühlmann, P. (2006), 'Boosting for high-dimensional linear models', *The Annals of Statistics* pp. 559–583. MR2281878

Bühlmann, P. (2013), 'Statistical significance in high-dimensional linear models', *Bernoulli* **19**, 1212–1242. MR3102549

Buja, A., Berk, R. A., Brown, L. D., George, E. I., Pitkin, E., Traskin, M., Zhao, L. & Zhang, K. (2015), 'Models as approximations – a conspiracy of random regressors and model deviations against classical inference in regression', *Statistical Science* pp. 1–44. MR4048582

Buja, A. & Zhang, K. (2015), *PoSI: Valid Post-Selection Inference for Linear LS Regression*. R package version 1.0. MR3047169

Chatterjee, A. & Lahiri, S. N. (2011), 'Bootstrapping lasso estimators', *Journal of the American Statistical Association* **106**, 608–625. MR2847974

Chatterjee, A. & Lahiri, S. N. (2013), 'Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap', *The Annals of Statistics* **41**, 1232–1259. MR3113809

Chen, Y., Jewell, S. & Witten, D. (2021), 'More powerful selective inference for the graph fused lasso', *arXiv preprint arXiv:2109.1045*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**, C1–C68. MR3769544

Chernozhukov, V., Chetverikov, D. & Kato, K. (2017), 'Central limit theorems and bootstrap in high dimensions', *The Annals of Probability* **45**, 2309–2352. MR3693963

Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. (2015), 'High-dimensional inference: Confidence intervals, p-values and R-software hdi', *Statistical Science* **30**, 533–558. MR3432840

Dezeure, R., Bühlmann, P. & Zhang, C.-H. (2017*a*), 'High-dimensional simultaneous inference with the bootstrap', *TEST* **26**, 685–719. MR3713586

Dezeure, R., Bühlmann, P. & Zhang, C. H. (2017*b*), 'High-dimensional simultaneous inference with the bootstrap', *TEST* **26**, 685–719. MR3713586

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Annals of Statistics* **32**, 407–499. MR2060166

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**, 1348–1360. MR1946581

Fan, J., Li, R., Zhang, C. & Zou, H. (2020), *Statistical Foundations of Data Science*, CRC Data Science Series, 1 edn, Chapman and Hall.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultra-high dimensional feature space', *Journal of the Royal Statistical Society Series B* **70**, 849–911. MR2530322

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society A* **222**, 309–368.

Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd. MR1523616

Fithian, W., Sun, D. & Taylor, J. (2017), 'Optimal inference after model selection', *arXiv preprint arXiv:1410.2597*.

Foucart, S. & Rauhut, H. (2013), *A Mathematical Introduction to Compressive Sensing*, Basel:Birkhäuser. MR3100033

Friedman, J. (2012), 'Fast sparse regression and classification', *International Journal of Forecasting* **28**, 722–738.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *The Annals of Statistics* **29**, 1189–1232. MR1873328

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**, 432–441.

Gao, L., Bien, J. & Witten, D. (2021), 'Selective inference for hierarchical clustering', *arXiv preprint arXiv:2012.02936*.

Hartman, P. & Stampacchia, G. (1966), 'On some nonlinear elliptic differential functional equations', *Acta Math* **115**, 271–310. MR0206537

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, 1 edn, Chapman and Hall/CRC. MR3616141

Hurvich, C. M. & Tsai, C.-L. (1990), 'The impact of model selection on inference in linear regression', *The American Statistician* **44**, 214–271.

Hyun, S., G'Sell, M. & Tibshirani, R. (2018), 'Exact post-selection inference for the generalized lasso path', *Electronic Journal of Statistics* **12**, 1053–1097. MR3777139

Hyun, S., Lin, K., G'Sell, M. & Tibshirani, R. (2021), 'Post-selection inference for changepoint detection algorithms with application to copy number variation data', *Biometrics* **77**, 1037–1049. MR4320676

Ioannidis, J. (2005), 'Why most published research findings are false', *PLoS Medicine* **2**. MR2216666

Javanmard, A. & Montanari, A. (2014*a*), 'Confidence intervals and hypothesis testing for high-dimensional regression', *The Journal of Machine Learning Research* **15**, 2869–2909. MR3277152

Javanmard, A. & Montanari, A. (2014*b*), 'Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory', *IEEE Transactions on Information Theory* **60**, 6522–6554. MR3265038

Jewell, S., Fearnhead, P. & Witten, D. (2021), 'Testing for a change in mean after changepoint detection', *arXiv preprint arXiv:1910.04291*. MR4157975

Kabaila, P. (1995), 'The effect of model selection on confidence regions and prediction regions', *Econometric Theory* **11**, 537–549. MR1349934

Kabaila, P. (1998), 'Valid confidence intervals in regression after variable selection', *Econometric Theory* **14**, 463–482. MR1650037

Kabaila, P. (2005), 'On the coverage probability of confidence intervals in regression after variable selection', *Australian and New Zealand Journal of Statistics* **47**, 549–562. MR2235423

Kabaila, P. (2009), 'The coverage properties of confidence regions after model selection', *International Statistical Review* **77**, 405–414.

Kabaila, P. & Giri, K. (2009), 'Upper bounds on the minimum coverage probability of confidence intervals in regression after model selection', *Australian and New Zealand Journal of Statistics* **51**, 271–287. MR2569799

Kabaila, P. & Leeb, H. (2006), 'On the larger-sample minimal coverage probability of confidence intervals after model selection', *Journal of the American Statistical Association* **101**, 619–629. MR2256178

Khalili, A. & Vidyashankar, N. (2018), 'Hypothesis testing in finite mixture of regressions: Sparsity and model selection uncertainty', *Canadian Journal of Statistics*. MR3884430

Kivaranovic, D. & Leeb, H. (2021), 'On the length of post-model-selection confidence intervals conditional on polyhedral constraints', *Journal of the American Statistical Association* **116**, 845–857. MR4270029

Koopmans, T. C. (1949), 'Identification problems in economic model construction', *Econometrica* **17**, 125–144. MR0031703

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I. & Zhao, L. (2018*a*), 'A model free perspective for linear regression: Uniform-in-model bounds for post selection inference', *arXiv preprint arXiv:1802.05801v2*. MR4152630

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I. & Zhao, L. (2018*b*), 'Valid post-selection inference in assumption-lean linear regression', *arXiv preprint arXiv:1806.04119v1*. MR4152630

Lander, E. & Kruglyak, L. (1995), 'Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results', *Nature Genetics* **11**, 241–247.

Lee, J., Sun, D., Sun, Y. & Taylor, J. (2016), 'Exact post-selection inference, with application to the lasso', *The Annals of Statistics* **44**, 907–927. MR3485948

Lee, J. & Taylor, J. (2014), 'Exact post model selection inference for marginal screening', *Advances in Neural Information Processing Systems* **27**, 136–144.

Leeb, H. & Pötscher, B. M. (2003), 'The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations', *Econometric Theory* **19**, 100–142. MR1965844

Leeb, H. & Pötscher, B. M. (2005), 'Model selection and inference: Facts and fiction', *Econometric Theory* **21**, 21–59. MR2153856

Leeb, H. & Pötscher, B. M. (2006), 'Can one estimate the conditional distribution of post-model-selection estimators?', *The Annals of Statistics* **34**, 2554–2591. MR2291510

Leeb, H. & Pötscher, B. M. (2008), 'Can one estimate the unconditional distribution of post-model-selection estimators?', *Econometric Theory* **24**, 338–376. MR2422862

Leeb, H., Pötscher, B. M. & Ewald, K. (2015), 'On various confidence intervals post-model-selection', *Statistical Science* **30**, 216–227. MR3353104

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. & Wasserman, L. (2018), 'Distribution-free predictive inference for regression', *Journal of the American Statistical Association* **113**, 1094–1111. MR3862342

Liu, H. & Yu, B. (2013), 'Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression', *Electronic Journal of Statistics* **7**, 3124–3169. MR3151764

Liu, K., Markovic, J. & Tibshirani, R. (2018), 'More powerful post-selection inference, with application to the lasso', *arXiv preprint arXiv:1801.09037*.

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014), 'A significance test for the lasso', *The Annals of Statistics* **42**, 413–468. MR3210970

Lu, S., Liu, Y., Yin, L. & Zhang, K. (2017), 'Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization', *Journal of the Royal Statistical Society Series B* **79**, 589–611. MR3611761

Mann, C. (1994), 'Behavioral genetics in transition', *Science* **264**, 894–942.

Mehrizi, R. & Chenouri, S. (2021), 'Valid post-detection inference for change points identified using trend filtering', *arXiv preprint arXiv:2104.12022*.

Meinshausen, N. & Bühlmann, P. (2010), 'Stability selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473. MR2758523

Meinshausen, N., Meier, L. & Bühlmann, P. (2009), 'p-values for high dimensional regression', *Journal of the American Statistical Association* **104**, 1671–1681. MR2750584

Minnier, J., Tian, L. & Cai, T. (2011), 'A perturbation method for inference on regularized regression estimates', *Journal of the American Statistical Association* **106**, 1371–1382. MR2896842

Negahban, S. N., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012), 'A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers', *Statistical Science* **27**, 538–557. MR3025133

Neufeld, A., Gao, L. & Witten, D. (2021), 'Tree-values: selective inference for regression trees', *arXiv preprint arXiv:2106.07816*.

Neykov, M., Ning, Y., Liu, J. S. & Liu, H. (2018), 'A unified theory of confidence regions and testing for high-dimensional estimating equations', *Statistical Science* **33**, 427–443. MR3843384

Nikolova, M. (2000), 'Local strong homogeneity of a regularized estimator', *SIAM Journal on Applied Mathematics* **61**, 633–658. MR1780806

Ning, Y. & Liu, H. (2014), 'Sparc: Optimal estimation and asymptotic inference under semiparametric sparsity', *arXiv preprint arXiv:1412.2295*.

Ning, Y. & Liu, H. (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', *The Annals of Statistics* **45**, 158–195. MR3611489

Pearson, K. (1900), 'X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**, 157–175.

Pötscher, B. M. (1991), 'Effects of model selection on inference', *Econometric Theory* **7**, 163–185. MR1128410

Pötscher, B. M. (1995), 'Comment on "the effect of model selection on confidence regions and prediction regions" by P. Kabaila', *Econometric Theory* **11**, 550–559. MR1349934

Pötscher, B. M. & NovÁk, A. J. (1998), 'The distribution of estimators after model selection: large and small sample results', *Journal of Statistical Computation and Simulation* **60**, 19–56. MR1622696

Regal, R. R. & Hook, E. B. (1991), 'The effects of model selection on confidence interval for the size of a closed population', *Statistics in Medicine* **10**, 717–721.

Rinaldo, A., Wasserman, L. & G'Sell, M. (2019), 'Bootstrapping and sample splitting for high-dimensional, assumption-lean inference', *The Annals of Statistics* **47**, 3438–3469. MR4025748

Robinson, S. M. (1992), 'Normal maps induced by linear transformations', *Mathematics of Operations Research* **17**, 691–714. MR1177731

Robinson, S. M. (1993), 'Math. programming set. b', *Nonsingularity and symmetry for linear normal maps* **62**, 415–426. MR1247623

Scheffé, H. (1959), *The Analysis of Variance*, 1 edn, New York: Wiley. MR0116429

Shah, R. & Samworth, R. (2013), 'Variable selection with error control: another look at stability selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 55–80. MR3008271

Shapin, S. & Schaffer, S. (1985), *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life.*, Princeton University Press.

Stigler, S. (2005), 'Fisher in 1921', *Statistical Science* **20**, 32–49. MR2182986

Taylor, J. & Tibshirani, R. (2018), 'Post-selection inference for ℓ_1-penalized likelihood models', *The Canadian Journal of Statistics* **46**, 41–61. MR3767165

Tian, X., Loftus, J. R. & Taylor, J. (2018), 'Selective inference with unknown variance via the square-root lasso', *Biometrika* **105**, 755–768. MR3877864

Tian, X. & Taylor, J. (2018), 'Selective inference with a randomized response', *The Annals of Statistics* **46**, 679–710. MR3782381

Tian, X. & Taylor, T. (2017), 'Asymptotics of selective inference', *Scandinavian Journal of Statistics* **44**, 480–499. MR3658523

Tibshirani, R. (1996), 'Regression shrinkage and selection with the lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267–288.

Tibshirani, R., Rinaldo, A. & Wasserman, R. T. L. (2018), 'Uniform asymptotic inference and the bootstrap after model selection', *The Annals of Statistics* **46**, 1255–1287.

Tibshirani, R., Taylor, J., Lockhart, R. & Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', *Journal of the American Statistical Association* **111**, 600–620.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J. & Reid, S. (2016), 'selectiveinference: Tools for post-selection inference'. R package version 1.2.0. https://CRAN.R-project.org/package=selectiveInference

Tropp, J. & Gilbert, A. (2007), 'Signal recovery from random measurements via orthogonal matching pursuit', *IEEE Transactions on Information Theory* **53**, 4655–4666.

van der Geer, S. (2007), The deterministic lasso, *in* 'JSM Proceedings', Vol. 140, American Statistical Association.

van der Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* **42**, 1166–1202.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer.

Wainwright, M. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wasserman, L. & Roeder, K. (2009), 'High-dimensional variable selection', *The Annals of Statistics* **37**, 2178–2201.

Yu, G., Yin, L., Lu, S. & Liu, Y. (2020), 'Confidence intervals for sparse penalized regression with random designs', *Journal of the American Statistical Association* **115**, 794–809.

Zhang, C. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**, 894–942.

Zhang, C.-H. & Zhang, S. S. (2014), 'Confidence intervals for low-dimensional parameters in high-dimensional linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 217–242.

Zhang, P. (1992), 'Inference after variable selection in linear regression models', *Biometrika* **79**, 741–746.

Zhang, X. & Cheng, G. (2014), 'Bootstrapping high dimensional time series', *arXiv preprint arXiv:1406.1037v2*.

Zhang, D., Khalili, A. & Asgharian, M. (2022), 'Supplement to "Post-Model-selection inference in linear regression models: An integrated review"'. DOI: 10.1214/22-SS135SUPP

Zhao, Q., Small, D. & Ertefaie, A. (2021), 'Selective inference for effect modification via the lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Zhao, S., Shojaie, A. & Witten, D. (2021), 'In defense of the indefensible: A very naïve approach to high-dimensional inference', *Statistical Science* **36**, 562–577.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Zrnic, T. & Jordan, M. (2020), 'Post-selection inference via algorithmic stability', *arXiv preprint arXiv:2011.09462*.