Biometrika (2023), **??**, ??, *pp.* 1–18
Printed in Great Britain Advance Access publication on ??

Hybrid Confidence Intervals for Informative Uniform Asymptotic Inference After Model Selection

BY A. MCCLOSKEY

Department of Economics, University of Colorado at Boulder, 256 UCB, Boulder, Colorado 80309, U.S.A.

adam.mccloskey@colorado.edu

SUMMARY

I propose a new type of confidence interval for correct asymptotic inference after using data to select a model of interest without assuming any model is correctly specified. This hybrid confidence interval is constructed by combining techniques from the selective inference and post-selection inference literatures to yield a short confidence interval across a wide range of data realizations. I show that hybrid confidence intervals have correct asymptotic coverage, uniformly over a large class of probability distributions that do not bound scaled model parameters. I illustrate the use of these confidence intervals in the problem of inference after using the LASSO objective function to select a regression model of interest and provide evidence of their desirable length and coverage properties in small samples via a set of Monte Carlo experiments that entail a variety of different data distributions as well as an empirical application to the predictors of diabetes disease progression.

Some key words: Confidence Interval; LASSO; Misspecification; Post-Selection Inference; Selective Inference; Uniform Asymptotics

1. INTRODUCTION

A large portion of the statistics literature in recent years has been dedicated to advancing inference methods that are valid after using data to select a model of interest without assuming the correct specification of any model in the selection set. Many, if not most, of these methods can be roughly broken down into two strands: methods that are valid for a selected parameter conditional on the model that is selected using a particular model selection criterion and methods that are unconditionally valid irrespective of the particular model selection criterion used. The former is often referred to as "selective inference" (e.g., Lee et al., 2016) and the latter is often referred to as "post-selection inference" (e.g., Berk et al., 2013), terms I will use throughout this paper.

Apart from the differences in coverage guarantees from confidence intervals constructed from these two approaches, they also feature complementary strengths and weaknesses in terms of informativeness. While selective intervals tend to be short when the model selected by the data is selected with (unconditional) high probability, they can become exceedingly wide when this selection event occurs with low probability. Indeed, standard selective intervals approach naive intervals (with incorrect coverage for the selected parameter) based upon inverting t-tests while ignoring data-driven model selection when the selection event occurs with high probability (Andrews et al., 2020). On the other hand, their expected length may be infinite (Kivaranovic and Leeb, 2021). Post-selection intervals do not suffer this latter drawback but can be very conservative in scenarios where the model selection criterion is known and the model selected by the data is selected with high probability, leading to coverage probabilities well in excess of their nominal levels and unnecessarily wide intervals.

In this paper, I propose a new class of hybrid confidence intervals for inference after model selection that aims to draw on the complementary informativeness strengths of selective and post-selection intervals. Motivated by the fact that selective intervals can become extremely wide when the model selected by the data is selected with low (unconditional) probability, by relaxing the conditional coverage requirement it may be possible to attain intervals with guaranteed unconditional coverage and better length properties. Although post-selection intervals have guaranteed unconditional coverage, selective intervals are substantially shorter when the model selected by the data is selected with high probability. In order to use the relative strengths of these two confidence intervals in terms of their lengths, the hybrid approach to confidence interval construction uses the data to transition a post-selection interval toward a selective interval when the latter is short. The hybrid intervals introduced here make use of similar reasoning to the hybrid intervals of Andrews et al. (2020) and Andrews et al. (2021), but applied to a model selection framework that generalizes that of Andrews et al. (2020) to incorporate many popular model selection criteria used in linear regression.

Like selective intervals, but unlike post-selection intervals, hybrid intervals require knowledge of the model selection criterion used by the researcher and the model selected by the data. The hybrid approach relaxes the conditional (on the selected model) coverage requirement of standard selective intervals to produce confidence intervals that are unconditionally valid. This unconditional coverage guarantee is analogous to the unconditional coverage guarantee of post-selection intervals with one difference: since we know the model selection criterion being used when constructing hybrid intervals, we attain validity on average across the models potentially selected using this particular model selection criterion, rather than across all possible model selection criteria. (Selective intervals posess the same unconditional coverage properties as hybrid intervals. For brevity, we refer the interested reader to Tibshirani et al. (2016) and Tibshirani et al. (2018) for detailed discussions on the interpretation of unconditional coverage when the model selection criterion is known.) The relaxation of the conditional coverage requirement can be viewed as an alternative approach for improving the length properties of selective intervals to standard sample splitting, the “data-carving” approach of Fithian et al. (2017) or the randomized response approach of Tian and Taylor (2018). However, in contrast to these latter approaches, the hybrid approach does not discard or add noise to the data in the model selection stage. It therefore yields valid inference on an object of interest that is selected from the full set of data rather than a fraction of it.

The hybrid approach to confidence interval construction improves upon the typical length properties of both the selective and post-selection intervals by modifying the conditioning event of selective intervals to restrict the selected parameter of interest to lie within a post-selection interval with a higher coverage probability. Since this latter addition to the conditioning event is not necessarily satisfied by the data, the hybrid approach takes the selective interval based upon this event but modifies which truncated normal quantiles are used in its constructoin to maintain correct coverage. By construction, the maximum length of the hybrid interval is bounded above by the length of the corresponding higher coverage post-selection interval, yielding finite expected length and breaking the negative result of Kivaranovic and Leeb (2021). At the same time, hybrid intervals can be configured to *nearly* approach naive intervals when the model selected by the data is selected with high probability in the sense that they approach naive intervals of a slightly higher coverage level.

Under a strengthening of the general model selection framework of Markovic et al. (2018) and an assumption implying the existence of a uniformly asymptotically valid post-selection confidence interval, I establish the uniform asymptotic validity of the hybrid intervals I propose. Standard uniform laws of large numbers and central limit theorems and results in e.g., Kuchibhotla et al. (2018) and Bachoc et al. (2020) can be used to verify these assumptions in linear regression model selection contexts. In the absence of knowledge of a true model, these assumptions typically require the practically-relevant setting of random, rather than fixed, regressors. As an illustration, in the supplementary materials of this paper I show how to verify these assumptions in the context of performing inference on a population regression coefficient for a predictor of interest after using the LASSO objective function to select the control variables that enter the regression. Importantly, this framework does not impose distributional assumptions on the data or that any parameters are known a priori.

The uniform asymptotic validity results can be applied to many other examples such as inference in a linear regression model selected by LASSO with randomized cross-validation, a randomized information criterion, a fixed number of steps along the forward stepwise and least angle regression algorithms and along the solution path of LASSO. Unlike the point-wise asymptotic results of e.g., Tian and Taylor (2017), the *uniform* asymptotic results I establish in this paper provide better approximations to finite sample coverage across a broad range of data-generating processes. And unlike the results of e.g., Tibshirani et al. (2018) or Andrews et al. (2021) but in line with e.g., Bachoc et al. (2020), the uniform asymptotic validity results I establish in this paper do not require one to bound the magnitude of various parameters such as (scaled) population regression coefficients.

Finally, I investigate the finite-sample performance of the hybrid intervals relative to the selective intervals of Lee et al. (2016), standard intervals formed after sample splitting and the post-selection intervals of Bachoc et al. (2020) when using LASSO as the model selection criterion in a set of Monte Carlo experiments and an empirical data application. In the Monte Carlo experiments, I draw data from a variety of distributions in a small sample setting to examine how well the uniform asymptotic results translate to challenging finite-sample settings that significantly depart from Gaussianity. Under a wide range of values for the LASSO penalty parameter, I find that the length distribution of hybrid intervals compares very favorably to those of selective and post-selection intervals. The length gains relative to selective intervals are acutely pronounced at higher quantiles of the relative length distributions. As a stark example, the 95^{th} quantile of the length distribution of the selective intervals can be more than 34 times larger than that of the hybrid intervals under the Monte Carlo designs I examine. In addition, I find that the hybrid intervals exhibit approximately correct finite-sample coverage. In the empirical application, I analyze the diabetes dataset from Efron et al. (2004) that was used to study selective intervals in Lee et al. (2016) and again find favorable empirical performance for the hybrid intervals.

2. Basic Ideas Behind the Hybrid Approach

To impart intuition, I focus on a particular application for the construction of confidence intervals after model selection, noting that a general approach that covers a wide range of applications is given in the following section. In particular, consider the standard linear regression framework for which a response variable y_i is modeled as a linear function of a predictor variable of interest z_i and some subset of the control variables $X_{1i}, \ldots, X_{pi}$ for $i = 1, \ldots, n$. Without imposing any assumptions about the true underlying relationship between the response, predictor of interest and controls, the researcher chooses a model $M = (E, s_E)$ with $E \subset \{1, \ldots, p\}$ being the subset of indices corresponding to the controls in the model and $s_E \subset \{-1, 1\}^{|E|}$ being the set of signs of the coefficients on these controls. The researcher's target parameter of interest is equal to the population linear regression coefficient θ_E defined by

$$(\theta_E, \beta_E) = \underset{\theta\in\mathbb{R}, b_E\in\mathbb{R}^{|E|}}{\operatorname{argmin}} \mathbb{E}\|y - z\theta - X_E b_E\|^2,$$

where $y = (y_1, \ldots, y_n)^T$, $z = (z_1, \ldots, z_n)^T$ may be fixed or random and X_E is the submatrix of the fixed or random design matrix $X = (x_1, \ldots, x_p)$ corresponding to model $M = (E, s_E)$ with $x_k = (X_{k1}, \ldots, X_{kn})^T$.

To establish the basic arguments, let us temporarily assume we have confidence intervals with correct finite-sample coverage, regression coefficient estimators that are normally distributed in finite samples and that variance parameters are known. There is now a large literature enabling the construction of selective intervals with (asymptotically) correct coverage for a population regression coefficient conditional on the model $\widehat{M} = (\widehat{E}, \hat{s}_{\widehat{E}})$ selected by the user for a variety of model selection criteria (e.g., Lee et al., 2016; Tibshirani et al., 2016; Tibshirani et al., 2018). That is, we have at our disposal a level $1-\alpha$ selective interval $CI^{S,\alpha}_{\widehat{M}}$ such that

$$\mathbb{P}\left(\theta_{\widehat{E}} \in CI^{S,\alpha}_{\widehat{M}} | \widehat{M} = M\right) \geq 1 - \alpha \tag{1}$$

for all $M = (E, s_E)$ with $E \subset \{1, \ldots, p\}$ (or some other relevant subset of the universe of models). On the other hand, there is a growing literature enabling the construction of post-selection intervals with correct unconditional coverage for a regression coefficient chosen by *any* model selection technique (e.g., Berk et al., 2013; Kuchibhotla et al., 2020; Bachoc et al., 2020). That is, for any potentially data-dependent $\widehat{M} \subset \{1, \ldots, p\}$, we have at our disposal a level $1-\alpha$ post-selection interval $CI^{P,\alpha}_{\widehat{M}}$ such that

$$\mathbb{P}\left(\theta_{\widehat{E}} \in CI^{P,\alpha}_{\widehat{M}}\right) \geq 1 - \alpha \tag{2}$$

for all $\widehat{M} = (\widehat{E}, \hat{s}_{\widehat{E}})$ with $\widehat{E} \subset \{1, \ldots, p\}$ (or some other relevant subset of the universe of models).

Selective confidence intervals are typically constructed by expressing the model selection event $\{\widehat{M} = M\}$ in terms of a data-dependent truncation interval for the OLS estimator $\widehat{\theta}_{\widehat{E}}$ of $\theta_{\widehat{E}}$, where

$$(\widehat{\theta}_E, \widehat{\beta}_E) = \underset{\theta\in\mathbb{R}, b_E\in\mathbb{R}^{|E|}}{\operatorname{argmin}} \|y - z\theta - X_E b_E\|^2.$$

This truncation interval depends upon a sufficient statistic $Z_{\widehat{M}}$ for the unknown nuisance parameter $\beta_{\widehat{E}}$ that is independent of $\widehat{\theta}_{\widehat{E}}$ after conditioning on the realization of $\widehat{M} = (\widehat{E}, \hat{s}_{\widehat{E}})$, i.e., $\{\widehat{M} = M\} = \{\widehat{\theta}_E \in [\mathcal{V}^-_M(Z_M), \mathcal{V}^+_M(Z_M)]\}$. (There is an additional element of the conditioning set $\mathcal{V}^0_M(Z_M) \geq 0$ that is suppressed in this section for simplicity of exposition.) The typical construction of a selective interval then proceeds by invoking the fact that $\widehat{\theta}_{\widehat{E}} | \widehat{M} = M$ is distributed according to a normal distribution with mean θ_E truncated to the interval $[\mathcal{V}^-_M(Z_M), \mathcal{V}^+_M(Z_M)]$, and collecting all null hypothesized values of θ_E for which a test based upon this distribution evaluated at the realized value of Z_M would fail to reject at level α. On the other hand, post-selection intervals typically take the form $CI^{P,\alpha}_{\widehat{M}} = \widehat{\theta}_{\widehat{E}} \pm \sigma_{\widehat{M}} K_\alpha$, where σ_M is the standard deviation of $\widehat{\theta}_E$ and K_α is a constant that guarantees (2) holds.

My proposal in this context is to form a level $1-\alpha$ hybrid confidence interval $CI^{H,\alpha}_{\widehat{M}}$ that is constructed in analogy with the selective interval after modifying the conditioning event and appropriately adjusting the corresponding coverage level. More specifically, this modified conditioning event is equal to the intersection of the model selection event expressed in terms of the sufficient statistic Z_M and the (potentially false) event that $\theta_{\widehat{E}}$ lies inside of a level $1-\gamma > 1-\alpha$ post-selection interval:

$$\begin{aligned} &\left\{\widehat{M}=M\right\} \cap \left\{\theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}\right\} \\ &= \left\{\widehat{\theta}_E \in [\mathcal{V}^-_M(Z_M), \mathcal{V}^+_M(Z_M)]\right\} \cap \left\{\widehat{\theta}_E \in [\theta_E - \sigma_M K_\gamma, \theta_E + \sigma_M K_\gamma]\right\} \\ &= \left\{\widehat{\theta}_E \in [\mathcal{V}^{-,H}_M(Z_M,\theta_E), \mathcal{V}^{+,H}_M(Z_M,\theta_E)]\right\}, \end{aligned}$$

where $\mathcal{V}^{-,H}_M(Z_M,\theta_E) = \max\left\{\mathcal{V}^-_M(Z_M),\ \theta_E - \sigma_M K_\gamma\right\}$ and $\mathcal{V}^{+,H}_M(Z_M,\theta_E) = \min\left\{\mathcal{V}^+_M(Z_M),\ \theta_E + \sigma_M K_\gamma\right\}$ (using the convention that $[a,b]=\emptyset$ if $b<a$). A hybrid interval is then constructed by invoking the fact that $\widehat{\theta}_{\widehat{E}} | \{\widehat{M}=M\} \cap \{\theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}\}$ is distributed according to a normal distribution with mean θ_E truncated to the interval $[\mathcal{V}^{-,H}_M(Z_M,\theta_E), \mathcal{V}^{+,H}_M(Z_M,\theta_E)]$. It is defined as all null hypothesized values of θ_E for which a test based upon this distribution evaluated at the realized value of Z_M would fail to reject at the adjusted level of $(\alpha-\gamma)/(1-\gamma)$.

When $\mathbb{P}(\widehat{M}=M)$ is small, one (or both) of the bounds of the truncation interval used to form a selective interval $[\mathcal{V}^-_M(Z_M), \mathcal{V}^+_M(Z_M)]$ tends to be very close to $\widehat{\theta}_E$ so that a test based upon the distribution of $\widehat{\theta}_E | \{\widehat{M}=M\}$ fails to reject for many hypothesized values of the mean θ_E of $\widehat{\theta}_E$, including very large ones. Since it is based upon inverting such a test, this induces the selective interval to become very long. The additional condition $\theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}$ used to form the hybrid interval bounds the value θ_E can take from above and below so that by conditioning on $\{\widehat{\theta}_E \in [\mathcal{V}^{-,H}_M(Z_M,\theta_E), \mathcal{V}^{+,H}_M(Z_M,\theta_E)]\}$ rather than $\{\widehat{\theta}_E \in [\mathcal{V}^-_M(Z_M), \mathcal{V}^+_M(Z_M)]\}$, the values of θ_E under consideration in the formation of the hybrid interval are bounded above and below by $\widehat{\theta}_E \pm \sigma_M K_\gamma$. In contrast to the selective interval, this fact implies that the level $1-\alpha$ hybrid interval lies inside of the level $1-\gamma$ post-selection interval $CI^{P,\gamma}_{\widehat{M}}$ so that the former is never longer than the latter. In addition, the latter confidence interval is not much wider than the level $1-\alpha$ post-selection interval $CI^{P,\alpha}_{\widehat{M}}$ by virtue of the fact that the length of $CI^{P,\alpha}_{\widehat{M}}$ depends upon α as a constant times $\sqrt{\log(\alpha^{-1})}$. This is shown by Bachoc et al. (2018) under Gaussian errors with a fixed design matrix but the argument can be extended via the central limit theorem (see Bachoc et al., 2020 and Kuchibhotla et al., 2021). On the other hand, when $\mathbb{P}(\widehat{M}=M)$ is large, the truncation interval used to form the selective interval tends to be very wide so that the distribution of $\widehat{\theta}_E | \{\widehat{M}=M\}$ is close to the unconditional distribution of $\widehat{\theta}_E$ and the selective interval is very close to its short naive counterpart. Since $\mathcal{V}^-_M(Z_M)$ is large and negative and $\mathcal{V}^+_M(Z_M)$ is large and positive in this case, the hybrid truncation interval $[\mathcal{V}^{-,H}_M(Z_M,\theta_E), \mathcal{V}^{+,H}_M(Z_M,\theta_E)]$ becomes very close to the selective truncation interval yielding a hybrid interval that is very close to the $(1-\alpha)/(1-\gamma)$ selective and naive confidence intervals.

The reason behind inverting tests at the adjusted level $(\alpha-\gamma)/(1-\gamma)$ (rather than α) is to account for the fact that the modified conditioning event is not necessarily satisfied by a given realization of the data since $\mathbb{P}(\theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}) < 1$. To see why this adjusted level yields correct unconditional coverage of the hybrid confidence interval, note that analogous arguments to those used to guarantee correct conditional coverage (1) can be used to guarantee

$$\mathbb{P}\left(\theta_{\widehat{E}} \in CI^{H,\alpha}_{\widehat{M}} \,\middle|\, \widehat{M} = M, \theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}\right) \geq \frac{1-\alpha}{1-\gamma} \tag{3}$$

for all $M = (E, s_E)$ with $E \subset \{1, \ldots, p\}$ so that

$$\mathbb{P}\left(\theta_{\widehat{E}} \in CI^{H,\alpha}_{\widehat{M}}\right) \geq \mathbb{P}\left(\theta_{\widehat{E}} \in CI^{H,\alpha}_{\widehat{M}} \,\middle|\, \theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}\right) \mathbb{P}\left(\theta_{\widehat{E}} \in CI^{P,\gamma}_{\widehat{M}}\right) \geq 1-\alpha$$

for all $\widehat{M} = (\widehat{E}, \hat{s}_{\widehat{E}})$ with $\widehat{E} \subset \{1, \ldots, p\}$, where the final inequality follows from (2), (3) and the law of iterated expectations.

3. General Asymptotic Model Selection Framework

In this section I introduce a set of very general assumptions to incorporate several forms of model selection and post-selection targets of inferential interest. I discuss how these assumptions apply in many model selection settings. In the supplementary material, I provide more detail about how these assumptions hold in the context of inference on a regression coefficient of interest after using LASSO to select controls variables.

Suppose we use a data set of n observations that is realized from an unknown probability measure $\mathbb{P} \in \mathcal{P}_n$ to select a model M from a finite set of models $\mathcal{M} = \{1, \ldots, |\mathcal{M}|\}$. I require the set of probability measures $\mathcal{P}_n$ to satisfy a uniform version of the model selection condition of Markovic et al (2018). Letting $\widehat{M_n}$ denote the (random) model selected by the data, suppose that the event that a given model $M \in \mathcal{M}$ is selected is equivalent to a random vector $D_n(M)$ satisfying an affine constraint according to the following assumption.

Assumption 1. For all $M \in \mathcal{M}$, $\widehat{M}_n = M$ if and only if $A_M D_n(M) \leq \widehat{a}_{M,n}$, where A_M is a fixed matrix and $\widehat{a}_{M,n}$ is a random vector such that for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\|\widehat{a}_{M,n} - a_{M,n}(\mathbb{P})\| > \varepsilon\right) = 0$$

for some vector-valued sequence of functions $a_{M,n}(\mathbb{P})$ such that for some finite $\bar{\lambda}$, $\|a_{M,n}(\mathbb{P})\| \leq \bar{\lambda}$ for all $\mathbb{P} \in \mathcal{P}_n$ and $n \geq 1$.

Several papers in the selective inference literature have shown that, under an appropriate definition of a model, the model selection event $\{\widehat{M}_n = M\}$ is equivalent to the event that an affine constraint holds on a lower-dimensional statistic $D_n(M)$ that is a function of the underlying data in accordance with Assumption 1. The appropriate definition of a model for Assumption 1 to be satisfied in the context of linear regression is usually the set of non-zero regression coefficients and their signs. The affine constraints typically arise from Karush-Khun-Tucker necessary and sufficient conditions from optimizing the objective function determining which model is selected, by comparing the values of different statistics across the steps of an iterative selection procedure and/or by comparing the values of statistics evaluated at different tuning parameters when model selection tuning parameters are data-dependent. For example, Lee et al. (2016) show that for a fixed LASSO penalty parameter λ, the LASSO model selection event characterizing the set of non-zero regression coefficients and their signs is equivalent to $\{A_M D_n(M) \leq \widehat{a}_{M,n}\}$ for which A_M is a matrix of zeros, ones and negative ones, $D_n(M)$ is a vector of scaled least squares regression estimates and inner products of regressors and residuals and $\widehat{a}_{M,n}$ is a vector that is a function of λ and the design matrix. This vector $\widehat{a}_{M,n}$ naturally satisfies the uniform convergence in probability in the assumption by a uniform law of large numbers and standard moments conditions on the underlying design matrix.

As another example, Tibshirani et al. (2016) show that for a fixed number of steps along the forward stepwise and least angle regression algorithms, the selection event characterizing the active regressors and their signs is equivalent to $\{A_M D_n(M) \leq \widehat{a}_{M,n}\}$ for which A_M is a matrix of ones and negative ones, $D_n(M)$ is a vector with elements that are functions of inner products of orthognally projected regressors and dependent variables and $\widehat{a}_{M,n} = 0$. Using the equivalence between the solution path of LASSO and a modified version of least angle regression (Efron et al., 2004), Tibshirani et al. (2016) provide a similar affine characterization for models chosen along the solution path of LASSO (rather than for a fixed LASSO penalty parameter). Finally, Markovic et al. (2018) and Tian and Taylor (2018) show the affine characterization of the selection event holds for various selection procedures when random noise is added to the selection criterion and/or the selection criterion uses a data-dependent tuning parameter (such as LASSO with λ chosen via cross-validation).

I am interested in constructing a confidence interval for a scalar parameter that is chosen based upon the selected model $\widehat{M}_n$. I denote this target parameter as $\mu_{T,n}(\widehat{M}_n)$, where by a slight abuse of notation, $\mu_{T,n}(M)$ denotes $\mu_{T,n}(M;\mathbb{P})$, the M^{th} element of $\mu_{T,n}(\mathbb{P})$ defined in Assumption 2. Recall that M indexes a model, which in the context of linear regression is not equal the number of regressors. Rather, in a typical application of model selection for p regressors subject to selection, M indexes both the set of non-zero regression coefficients amongst the p coefficients along with their signs so that there is a total of $|\mathcal{M}| = 3^p$ models being selected amongst.

For any M, assume that in the absence of data-dependent selection, there is an asymptotically Gaussian statistic $T_n(M)$ centered around $\mu_{T,n}(M)$. Further assume that the full vectors of statistics $T_n(M)$ and the statistics determining selection $D_n(M)$ are uniformly jointly asymptotically normal under $\mathbb{P} \in \mathcal{P}_n$ with centering vectors $\mu_{T,n}$ and $\mu_{D,n}$ and limiting covariance matrix Σ that may depend upon $\mathbb{P}$. I use the following strengthening of Markovic et al. (2018) to full joint convergence of all statistics because I focus on unconditional inferential statements that do not condition on the selected model. For any matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its minimum and maximum eigenvalues.

Assumption 2. For $T_n = (T_n(1), \ldots, T_n(|\mathcal{M}|))^T \in \mathbb{R}^{|\mathcal{M}|}$ and $D_n = (D_n(1)^T, \ldots, D_n(|\mathcal{M}|)^T)^T$ and the class of Lipschitz functions that are bounded in absolute value by one and have Lipschitz constant bounded by one, BL_1, there exist sequences of functions $\mu_{T,n}(\mathbb{P})$ and $\mu_{D,n}(\mathbb{P})$ and a function $\Sigma(\mathbb{P})$ such that for $(T_{\mathbb{P}}^{*T}, D_{\mathbb{P}}^{*T})^T \sim \mathcal{N}(0, \Sigma(\mathbb{P}))$ with

$$\Sigma = \begin{pmatrix} \Sigma_T & \Sigma_{TD} \\ \Sigma_{DT} & \Sigma_D \end{pmatrix},$$

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_n} \sup_{f\in BL_1} \left| \mathbb{E}_{\mathbb{P}}\left[f\begin{pmatrix} T_n - \mu_{T,n}(\mathbb{P}) \\ D_n - \mu_{D,n}(\mathbb{P}) \end{pmatrix} \right] - \mathbb{E}_{\mathbb{P}}\left[f\begin{pmatrix} T_{\mathbb{P}}^* \\ D_{\mathbb{P}}^* \end{pmatrix} \right] \right| = 0.$$

Furthermore, for some finite $\bar{\lambda} > 0$, $1/\bar{\lambda} \leq \Sigma_T(M, M; \mathbb{P}) \leq \bar{\lambda}$ and $1/\bar{\lambda} \leq \lambda_{\min}(\Sigma_D^{(M)}(\mathbb{P})) \leq \lambda_{\max}(\Sigma_D^{(M)}(\mathbb{P})) \leq \bar{\lambda}$ for all $M \in \mathcal{M}$ and $\mathbb{P} \in \mathcal{P}_n$, where $\Sigma_D^{(M)}$ is the covariance matrix of $D^*(M)$.

The statistics $T_n(M)$ typically take the form of scaled (linear functionals of) sample regression coefficient estimates under the imposed model M with $\mu_{T,n}(M)$ being equal to the scaled (linear functionals of) population regression counterparts. As discussed following Assumption 1, the statistics $D_n(M)$ typically take the form of scaled sample regression parameter estimates and functions of inner products of dependent variables, regressors and residuals (with the additon of noise terms for applications involving randomization). Thus, T_n and D_n naturally satisfy joint uniform central limit theorems under standard assumptions on the data. Indeed, Markovic et al. (2018) verify marginal central limit theorems on $(T_n(M), D_N(M)^T)^T$ for a given M for inference on regression coefficients after LASSO model selection with a fixed penalty parameter and a penalty parameter chosen via randomized cross-validation and after randomized information criteria-based selection. It is straightforward to extend these results to show that the joint uniform central limit theorem in Assumption 2 holds for these problems. Similarly, Tibshirani et al. (2018) show that both linear functionals of sample regression coefficient estimates and the selection events involved in models selected for a fixed number of steps along the forward stepwise and least angle regression algorithms and along the solution path of LASSO are functions of a "master statistic", enabling a joint uniform central limit theorem of the form given in Assumption 2. Although these latter results impose a fixed design matrix, they easily extend to random designs under suitable assumptions. It is important to note that Assumption 2 does not impose eigenvalue bounds on the full covariance matrix Σ, allowing this matrix to be singular.

In order to form asymptotically valid hybrid intervals, I require the use of uniformly consistent estimators $\widehat{\Sigma}_{T,n}$ and $\widehat{\Sigma}_{DT,n}$ for the covariance matrices in Assumption 2. Let $\|\cdot\|$ denote the Frobenius norm.

Assumption 3. There exist estimators $\widehat{\Sigma}_{T,n}$ and $\widehat{\Sigma}_{DT,n}$ such that for all $\varepsilon > 0$,

$$\lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\|\widehat{\Sigma}_{T,n}-\Sigma_T(\mathbb{P})\|>\varepsilon\right)=0 \text{ and } \lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\|\widehat{\Sigma}_{DT,n}-\Sigma_{DT}(\mathbb{P})\|>\varepsilon\right)=0.$$

Unlike the post-selection intervals of Bachoc et al. (2020), using an inconsistent conservative estimator of Σ_T that consistently overestimates its diagonal values will not lead to hybrid intervals with correct asymptotic coverage. In the construction of the hybrid confidence interval, consistent estimation of Σ_T and Σ_{DT} is crucial to forming a random vector $Z_{M,n}$ that is asymptotically independent of the statistic $T_n(M)$. The necessity of Assumption 3 therefore limits the scope of application of hybrid intervals relative to post-selection ones. Nevertheless, in the context of selection in the linear regression model, the (co)variances Σ_T and Σ_{DT} are possible to consistently estimate when (i) the regressors are either random or constant (corresponding to an intercept term) or (ii) the true model is known a priori. The former has been shown by, e.g., Kuchibhotla et al. (2018) and references therein and the latter is well known. Although much of the selective inference and post-selection literature has focused on the case of a fixed design matrix, case (i) is more relevant for applications involving data that are sampled from an underlying population distribution, a commonly encountered scenario in practice. For typical applications, Kuchibhotla et al. (2018) show how the elements of Σ_T and Σ_{DT} can be estimated consistently using standard heteroskedasticity-robust methods in linear regression contexts with iid data under standard moment conditions, applying to both homoskedastic and heteroskedastic data. Freedman (1981) and Buja et al. (2019) show how the elements of Σ_T and Σ_{DT} can be estimated via pairs bootstrap (see also Markovic et al., 2018). The consistency arguments can be strengthened to the uniform consistency requirement of Assumption 3 straightforwardly.

I make one final high-level assumption on the existence of a post-selection interval with correct unconditional uniform asymptotic coverage of the parameter of interest $\mu_{T,n}(\widehat{M}_n)$.

Assumption 4. For any $\alpha \in (0,1)$, we have a confidence interval of the form

$$CI^{P,\alpha}_{n,\widehat{M}_n} = T_n(\widehat{M}_n) \pm \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n)}K_{n,\alpha}$$

that satisfies $\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\alpha}_{n,\widehat{M}_n}\right)\geq 1-\alpha$ and for any $\varepsilon>0$, $\lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(|K_{n,\alpha}-K_\alpha(\mathbb{P})|>\varepsilon\right)=0$ for some function $K_\alpha(\mathbb{P})$ such that for some finite $\bar{\lambda}$, $0\leq K_\alpha(\mathbb{P})\leq\bar{\lambda}$ for all $\mathbb{P}\in\mathcal{P}_n$.

For inference on (linear functionals of) population regression coefficients after model selection in the linear regression framework, $K_{n,\alpha}$ typically takes the form of (an upper bound on) the $(1-\alpha)$-quantile of the maximum of a sequence of correlated standard normal random variables, for which the correlation matrix is derived from $\widehat{\Sigma}_{T,n}$, or an asymptotically equivalent bootstrap version. The uniformly consistent estimation of $\widehat{\Sigma}_{T,n}$ implied by Assumption 3 then immediately implies the uniform consistency of $K_{n,\alpha}$ for K_α required by Assumption 4 while the results of Bachoc et al. (2020) imply the uniform coverage requirement of the assumption for several examples of post-selection intervals.

4. Hybrid Confidence Intervals and Uniform Asymptotic Validity

We are now equipped with the ingredients needed to define the $(1-\alpha)$-level hybrid confidence interval, $CI^{H,\alpha}_{n,\widehat{M}_n}$, for $\mu_{T,n}(\widehat{M}_n)$. To begin describing the hybrid interval construction, it is useful to express the conditioning event $\{\widehat{M}_n=M\}$ in terms of a data-dependent interval for the target statistic $T_n(\widehat{M}_n)$: $[\mathcal{V}^-_{M,n}(Z_{M,n}),\mathcal{V}^+_{M,n}(Z_{M,n})]$. The bounds of this interval are functions of a random vector $Z_{M,n}$. Exact expressions for this random vector and the bounds are provided in the supplementary material.

The hybrid interval is constructed from the distribution function of $T_n(\widehat{M}_n)$ after conditioning on the events $\{\widehat{M}_n=M\}$ and $\{\mu_{T,n}(\widehat{M}_n)\in CI^{P,\gamma}_{n,\widehat{M}_n}\}$ for some $\gamma\in(0,\alpha)$. More specifically, let $F_{TN}(\cdot;\mu,\sigma^2,\mathcal{L},\mathcal{U})$ denote the truncated normal distribution function of $\xi|\{\mathcal{L}\leq\xi\leq\mathcal{U}\}$ for $\xi\sim\mathcal{N}(\mu,\sigma^2)$. For $\alpha\in(0,1)$, define $\widehat{\mu}^{H,\alpha}_{T,n}(\widehat{M}_n)$ to solve

$$F_{TN}\left(T_n(\widehat{M}_n);\mu,\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n),\mathcal{V}^{-,H}_{\widehat{M}_n,n}(Z_{\widehat{M}_n,n},\mu),\mathcal{V}^{+,H}_{\widehat{M}_n,n}(Z_{\widehat{M}_n,n},\mu)\right)=1-\alpha$$

in μ, where expressions for $\mathcal{V}^{-,H}_{M,n}(z,\mu)$ and $\mathcal{V}^{+,H}_{M,n}(z,\mu)$ are provided in the supplementary material. In turn, $CI^{H,\alpha}_{n,\widehat{M}_n}$ is defined as

$$CI^{H,\alpha}_{n,\widehat{M}_n}=\left[\widehat{\mu}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n),\widehat{\mu}^{H,1-\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n)\right],\qquad(4)$$

where $\widehat{\mu}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n)$ and $\widehat{\mu}^{H,1-\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n)$ are used instead of $\widehat{\mu}^{H,\alpha/2}_{T,n}(\widehat{M}_n)$ and $\widehat{\mu}^{H,1-\alpha/2}_{T,n}(\widehat{M}_n)$ to account for the fact that the probability of the conditioning event $\{\mu_{T,n}(\widehat{M}_n)\in CI^{P,\gamma}_{n,\widehat{M}_n}\}$ is only bounded below by $1-\gamma$ under all sequences of probability measures $\{\mathbb{P}_n\}$ (by Assumption 4). For simplicity, I focus on the two-sided equal-tailed version of the hybrid interval as defined in (4) but note that the uniform asymptotic validity results presented here also apply to one-sided and non-equal-tailed versions for which $\widehat{\mu}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n)$ and $\widehat{\mu}^{H,1-\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n)$ are replaced by any $\widehat{\mu}^{H,q_1}_{T,n}(\widehat{M}_n)$ and $\widehat{\mu}^{H,1-q_2}_{T,n}(\widehat{M}_n)$ such that $q_1+q_2=(\alpha-\gamma)/(1-\gamma)$.

Once the post-selection constant $K_{n,\gamma}$ is found, construction of the hybrid interval is computationally straightforward since it just involves finding the zeros of two continuous functions. Since $K_{n,\gamma}$ must be computed to form the hybrid interval, this implies that the hybrid and post-selection intervals share approximately the same degree of computational complexity. Berk et al.

(2013) provide code for efficiently computing non-conservative $K_{n,\gamma}$ values (in the sense that the coverage requirement in Assumption 4 holds with equality) in linear regression contexts when 20 or less covariates are subjected to model selection. Berk et al. (2013) and Bachoc et al. (2020) also discuss computationally straightforward methods for computing conservative values of $K_{n,\gamma}$ that satisfy Assumption 4 even when the number of models under consideration is very large by appealing to bounds on the quantiles of the maximum of correlated Gaussian random variables.

I now state a result establishing the uniform asymptotic coverage of $CI^{H,\alpha}_{n,\widehat{M}_n}$ conditional on the realization of the selected model $\widehat{M}_n$ and the possibly false event $\{\mu_{T,n}(\widehat{M}_n) \in CI^{P,\gamma}_{n,\widehat{M}_n}\}$. The proof of both this result and the one following it are contained in the supplementary material.

PROPOSITION 1. *Under Assumptions 1–4,*

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_n} \left| \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n) \in CI^{H,\alpha}_{n,\widehat{M}_n} \,\middle|\, \widehat{M}_n = M, \mu_{T,n}(\widehat{M}_n) \in CI^{P,\gamma}_{n,\widehat{M}_n}\right) - \frac{1-\alpha}{1-\gamma} \right| \times \mathbb{P}\left(\widehat{M}_n = M, \mu_{T,n}(\widehat{M}_n) \in CI^{P,\gamma}_{n,\widehat{M}_n}\right) = 0$$

for all $M \in \mathcal{M}$.

Using the results from Proposition 1, we can show that $CI^{H,\alpha}_{n,\widehat{M}_n}$ has uniformly correct unconditional coverage at level $1-\alpha$ and a controlled degree of nonsimilarity. This is the main theoretical result of the paper.

PROPOSITION 2. *Under Assumptions 1–4,*

$$\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI^{H,\alpha}_{n,\widehat{M}_n}\right) \geq 1-\alpha$$

and

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI^{H,\alpha}_{n,\widehat{M}_n}\right) \leq \frac{1-\alpha}{1-\gamma}.$$

The user is free to choose any value of γ with $0 \leq \gamma \leq \alpha$ in the construction of the hybrid interval. There is no value of γ that is optimal uniformly across the parameter space in terms of interval length measures. Rather, as we can see from the lower and upper bounds given in the above proposition, γ controls the degree of (asymptotic) non-similarity of the hybrid confidence interval with the interval being closer to (asymptotically) similar when γ is small. Similar confidence intervals are not necessarily desirable in this context. In fact when $\gamma = 0$, the hybrid interval is identical to the (similar) level $1-\alpha$ selective interval. On the other hand, for $\gamma = \alpha$, the hybrid interval is equal to the (non-similar) level $1-\gamma$ post-selection interval.

The choice of γ trades off the length properties of the hybrid interval over different realizations of the data. I recommend a small but non-negligible value of γ such as $\gamma = \alpha/10$ to attain an interval that is not much longer than the selective interval when the model selected by the data is selected with high probability (so that the selective interval is short) without compromising a lot of length when this does not occur. Proposition 3 of Andrews et al. (2020) implies that when a given model is selected with probability approaching one, the $(1-\alpha)$–level hybrid interval converges to an interval contained in a $(1-\alpha)/(1-\gamma)$–level naive interval, where $\gamma \in (0,\alpha)$ is chosen by the user. When using the recommended value of $\gamma = \alpha/10$ to construct the hybrid interval, this means that 99%, 95% and 90% hybrid intervals converge to intervals contained in 99.1%, 95.5% and 90.9% naive intervals. See Section 5 and Andrews et al. (2021) and Andrews et al. (2020) for further evidence that this choice works well in practice in related contexts.

5. FINITE-SAMPLE PROPERTIES OF CONFIDENCE INTERVALS

In order to investigate the finite-sample properties of hybrid confidence intervals and compare them to existing confidence intervals in a variety of settings, I examine Monte Carlo experiments for the application of inference on a regression coefficient of interest after using LASSO to select controls variables, detailed in the supplementary material. The data-generating processes I study in these Monte Carlo experiments are designed to closely match those studied in the simulations of Tibshirani et al. (2018) and Bachoc et al. (2020) who focus on the different application of inference on the variable selected across the steps of the least angle regression algorithm. More specifically, I consider data generated from the standard linear regression model

$$y = \theta z + X\beta + u \tag{5}$$

where y is an $n \times 1$ vector of observations of the outcome of interest, $\theta \in \mathbb{R}$, z is an $n \times 1$ vector of observations of the predictor of interest, $\beta \in \mathbb{R}^p$, X is an $n \times p$ matrix of observations of control variables that are selected by the LASSO objective function and u is an $n \times 1$ vector of independent and identically distributed error terms that is independent of X. With this knowledge of the data-generating process, the target parameter of inferential inference after model selection can be written as $\mu_{T,n}(\widehat{M}_n) = \sqrt{n} e_1^T (\mathbb{E}[W_{E,i} W_{E,i}^T])^{-1} \mathbb{E}[W_{E,i} W_i^T]\delta$ for $E = \widehat{E}_n$, where e_1 denotes the first standard basis vector, $W = (z, X)$, $W_E = (z, X_E)$, X_E equals the submatrix of X composed of the columns of X corresponding to E and $\delta = (\theta, \beta^T)^T$. (The rows of an arbitrary matrix or vector B are denoted as B_i.)

For this simulation study, I generate data that entail significant departures from Gaussianity in a relatively small sample of $n = 50$ in order to assess the relevance of the asymptotic guarantees provided by Proposition 2 and the relative performance of the hybrid interval in a small sample setting. Across all simulation designs, I set $\alpha = 0.05$ for nominal confidence interval coverage of 95%, $p = 10$ potential control variables, $\theta = 0$ and $\beta = (-4, 4, 0 \ldots, 0)^T$. All quantities are computed across $1,000$ independent simulation replications. The full matrix of regressors W is generated in two ways. In the independent design case, the columns of W are drawn from independent distributions, where each column is drawn from an independent $\mathcal{N}(0, 1)$, Bernoulli or skew normal $(0, 1, 5)$ distribution with equal probability. In the dependent design case, each row of W is generated independently from a multivariate normal distribution with mean vector zero and covariance matrix $(e^{-0.1|i-j|})_{1 \leq i,j \leq p}$. In both cases, each column is subsequently normalized to have unit Euclidean norm and Y is then generated according to (5) after sampling the entries of u independently in four ways: from a normal, skew normal (with shape parameter five), Laplace or uniform distribution, all with mean zero and unit variance.

In each simulated data set, I perform the LASSO model selection exercise described in the supplementary material for several different values of the LASSO penalty parameter $\lambda \in \{1, 2, 4, 8, 16, 100\}$ and construct confidence intervals for the target parameter $\mu_{T,n}(\widehat{M}_n)$ ($\mu_{T,n}(\widehat{M}_{n/2})$ for the split-sample interval, see below) selected by the LASSO objective function. Specifically, I calculate the naive interval that ignores model selection, a split-sample interval, the selective interval, the hybrid interval using the recommended value of $\gamma = \alpha/10$ and the post-selection interval. The naive interval is simply based on inverting the standard asymptotic t-test at the nominal level. Both the hybrid and post-selection intervals computed in these simulations use the less conservative post-selection construction that incorporates the fact that the regressor z is the predictor of interest and therefore not subject to selection. The split-sample interval is constructed as follows: the first $n/2$ observations are used to select the model, yielding $\widehat{M}_{n/2}$, while the final $n/2$ observations are used to construct a standard interval based on inverting the standard asymptotic t-test at the nominal level. It is important to note that although the split-sample interval is known to have correct asymptotic coverage, it is not for the same object of interest, $\mu_{T,n}(\widehat{M}_n)$, for which the other intervals are designed. Instead, the spit-sample interval has correct asymptotic coverage for the scaled population regression coefficient evaluated at the model selected by the first half of the data only, i.e., $\mu_{T,n}(\widehat{M}_{n/2})$. This is especially important to keep in mind when evaluating the tradeoffs of the various confidence intervals because selecting the model from only a portion of the data will yield a selected model with less desirable statistical properties (e.g., larger prediction errors). Nevertheless, we include this comparison since it is commonly used as a valid method for inference after model selection.

The results of the Monte Carlo experiments are very similar across some of the error distributions and LASSO penalty parameters. To save on space, I report a subset of results that illustrate the main features and tradeoffs of the full set of experiments. In particular, I report results for the normal and skew normal error distributions and small, medium and large values of λ. The values of λ that fall into these categories are determined by the corresponding largest probability that a particular model is selected across simulation draws. A larger penalty parameter is needed in the dependent design cases to produce similar model selection probabilities to the independent design cases because dependent design matrices effectively reduce the signal-to-noise ratio in the model selection problem. The value $\lambda = 1$ is considered small for all designs and error distributions since the largest model selection probabilities when $\lambda = 1$ are 0.057, 0.003, 0.001 and 0.002 for independent designs with normal errors, independent designs with skew normal errors, dependent designs with normal errors and dependent designs with skew normal errors, respectively. The value $\lambda = 4$ generates moderate model selection probabilities for independent designs, equal to 0.323 and 0.257 for normal and skew normal errors, while $\lambda = 16$ generates moderate model selection probabilities for dependent designs, equal to 0.285 and 0.214 for normal and skew normal errors. Finally, $\lambda = 16$ generates large model selection probabilities equal to 1 for independent designs while $\lambda = 100$ does so for dependent designs.

To begin, Table 1 displays the simulated (unconditional) coverage probabilities of the five confidence intervals for the different penalty parameter values. The selective, hybrid and post-selection intervals all have finite-sample coverage close to the nominal level of 95%, where the selective and hybrid intervals tend to slightly under-cover and the post-selection intervals tend to over-cover. The small coverage distortions of the selective and hybrid intervals are to be expected from such small non-Gaussian data sets and they diminish for larger samples. On the other hand, both the naive and split-sample intervals exhibit more sizable under-coverage. This under-coverage is to be expected from the naive interval since it is known to incorrectly cover after using the data to select the model, even in large samples. The split-sample intervals only exhibit notable coverage distortions in the dependent design cases. Since split-sample intervals are known to have correct coverage in large samples, this is likely due to the fact that these intervals are constructed using only $n/2 = 25$ data points and the strong positive correlation between the regressors effectively reduces this sample size further relative to the independent design cases. It is also interesting to note that even for the relatively small sample size of $n = 50$, the hybrid intervals roughly obey the asymptotic upper bound on coverage given in Proposition 2: $(1 - \alpha)/(1 - \gamma) = 0.955$.

Next, Figs. 1–3 plot the ratios of the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles across simulation draws of the lengths of the five confidence intervals relative to those corresponding to the post-selection interval for small λ (Fig. 1), medium λ (Fig. 2) and large λ (Fig. 3). There are four panels within each figure corresponding to how the design matrix and errors are generated. The ratios of the length quantiles of the post-selection interval relative to itself, which is always equal to one, is also included in the figures in order to detect when the other intervals' length

Table 1. *Unconditional Coverage Probabilities*

	Confidence Interval				
λ	Naive	SS	Sel	HySI	PoSI
	Indep Design, Normal Errors				
1	0.876	0.929	0.915	0.919	0.975
4	0.909	0.941	0.906	0.910	0.984
16	0.925	0.950	0.925	0.931	0.990
	Indep Design, Skew Normal Errors				
1	0.898	0.927	0.942	0.940	0.969
4	0.919	0.952	0.931	0.932	0.987
16	0.920	0.943	0.919	0.927	0.989
	Dep Design, Normal Errors				
1	0.867	0.951	0.955	0.957	0.985
16	0.900	0.847	0.919	0.922	0.994
100	0.927	0.928	0.928	0.932	0.989
	Dep Design, Skew Normal Errors				
1	0.892	0.933	0.951	0.948	0.982
16	0.915	0.887	0.929	0.934	0.989
100	0.946	0.937	0.947	0.953	0.993

This table reports unconditional coverage probabilities for the selected population coefficient on the predictor of interest after using LASSO to choose the control variables in the regression across Monte Carlo replications, all evaluated at the nominal coverage level of 95%. Coverage probabilities are reported for naive, split-sample (SS), selective (Sel), hybrid (HySI) and post-selection (PoSI) confidence intervals for a sample size of $n = 50$. The coverage probabilities are reported for values of the LASSO penalization parameter λ corresponding to small, moderate and large model selection probabilities. The design matrix is generated with independent (upper half of table) or correlated (lower half) columns and the error terms have normal or skew normal distributions.

quantiles are shorter than those of the post-selection interval. Even though the naive intervals do not have correct coverage in large samples, I also include their ratios of length quantiles as a lower bound to show how close the other intervals come to attaining it.

From Fig. 1 we can see that for a small LASSO penalty parameter, the length quantiles of the selective interval are almost uniformly dominated by all other intervals. For this low level of penalization the probability that LASSO chooses any given model is low, with maximum model selection probabilities between 0.001 and 0.057, leading to excessively long selective intervals. On the other hand, the hybrid intervals tend to have similar length properties to those of the post-selection intervals, with some small increases in the dependent design cases (lower two panels). The split-sample intervals also have similar length properties although, unlike the hybrid intervals, their length quantiles always exceed those of the post-selection intervals with the additional drawback that their target model of interest is much less precisely selected.

Fig. 2 displays somewhat similar features to Fig. 1 although for the moderate-sized penalty parameter corresponding to this figure, we can start to see that the hybrid intervals become notably shorter than the post-selection intervals in most cases, especially when the selective intervals are shorter. The selective intervals tend to be shorter than in Fig. 1 because the higher penalty parameter increases the probability that LASSO chooses a given model, with corresponding maximum model selection probabilities between 0.214 and 0.323. In combination with the results in Fig. 1, we can start to see the benefits of hybridization: the hybrid interval borrows the strengths of both the selective and post-selection intervals for different realizations of the data.

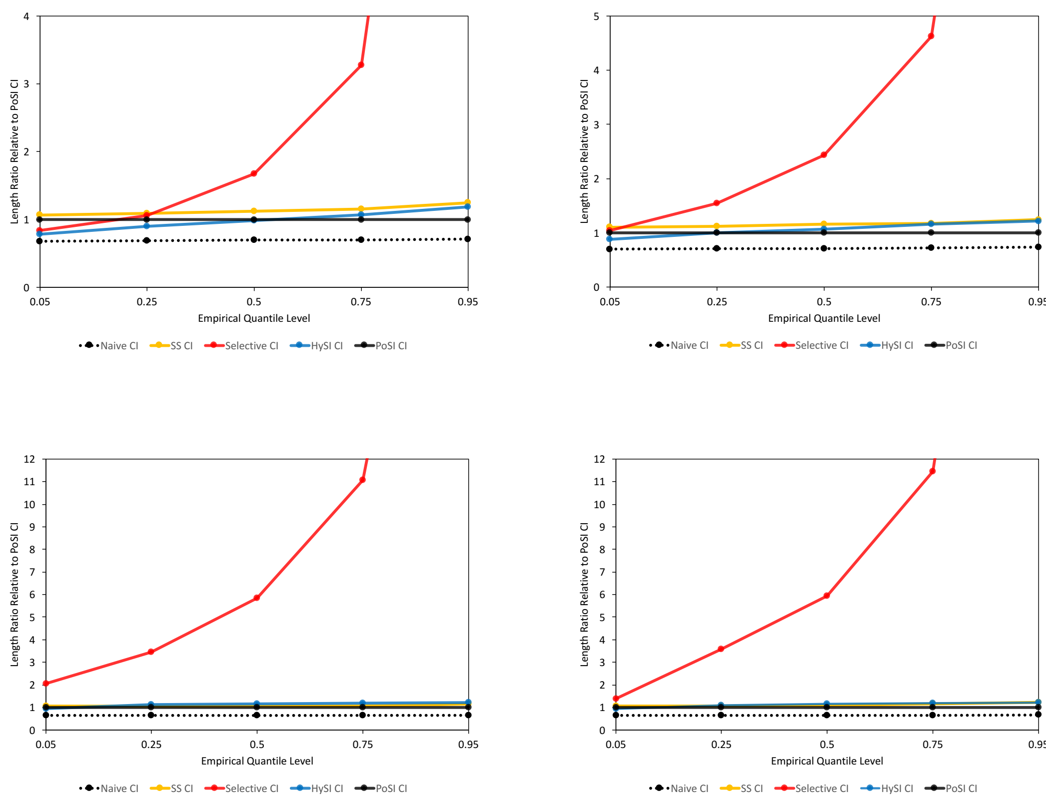

Fig. 1. This figure plots the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles of the lengths of the 95% naive (dotted black), split-sample (orange), selective (red), hybrid (blue) and post-selection (black) confidence intervals divided by the corresponding length quantiles of the 95% post-selection interval across Monte Carlo replications for inference after using LASSO to choose the control variables in the regression with a sample size of $n = 50$ and a small LASSO penalization parameter $\lambda = 1$. The upper-left plot corresponds to a design matrix with independent columns and error terms with a normal distribution. The upper-right plot corresponds to a design matrix with independent columns and error terms with a skew normal distribution. The lower-left plot corresponds to a design matrix with correlated columns and error terms with a normal distribution. The lower-right plot corresponds to a design matrix with correlated columns and error terms with a skew normal distribution.

Fig. 3 clearly show the benefits of using hybrid instead of post-selection when model selection probabilities are high (due to the large penalty parameter). Here we can see that the length quantiles of both the selective and hybrid intervals are nearly identical to those of the naive interval and substantially smaller than those of the post-selection interval. This is a clear illustration that the hybrid intervals attain nearly the same short lengths as the selective interval when they are short while guarding against the excessive lengths of the latter for unfavorable realizations of the data. Large penalty parameters, with corresponding large model selection probabilities, yield length quantiles of the selective and hybrid intervals that are nearly indistinguishable from those of the naive intervals, entailing length reductions of 34–35% relative to the post-selection interval across all quantile levels.

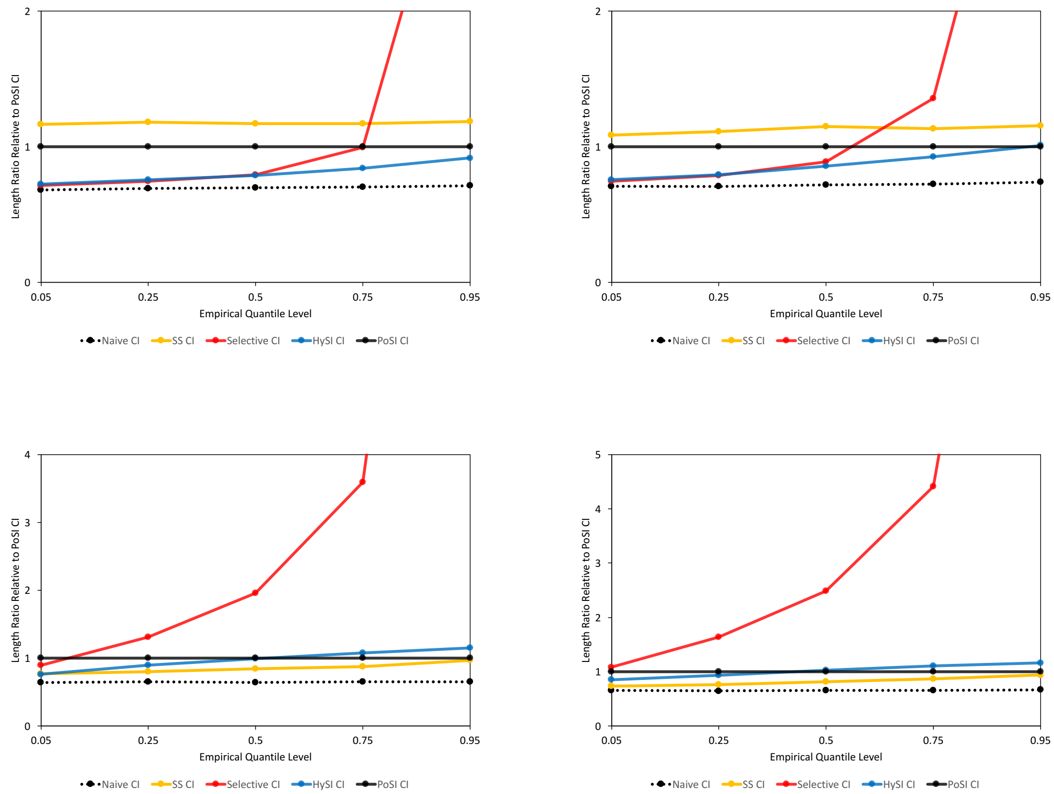

Fig. 2. This figure plots the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles of the lengths of the 95% naive (dotted black), split-sample (orange), selective (red), hybrid (blue) and post-selection (black) confidence intervals divided by the corresponding length quantiles of the 95% post-selection interval across Monte Carlo replications for inference after using LASSO to choose the control variables in the regression with a sample size of $n = 50$ and a medium LASSO penalization parameter. The upper-left plot corresponds to a design matrix with independent columns and error terms with a normal distribution for $\lambda = 4$. The upper-right plot corresponds to a design matrix with independent columns and error terms with a skew normal distribution for $\lambda = 4$. The lower-left plot corresponds to a design matrix with correlated columns and error terms with a normal distribution for $\lambda = 16$. The lower-right plot corresponds to a design matrix with correlated columns and error terms with a skew normal distribution for $\lambda = 16$.

6. Empirical Application to Diabetes Data

I further investigate the properties of hybrid confidence intervals in an empirical application to the diabetes data set from Efron et al. (2004). This data set was also examined by Lee et al. (2016) in their empirical application of selective inference after using LASSO as a model selection device, thus serving as a benchmark application for inference after using LASSO. Departing from the exact exercise performed by Lee et al. (2016), I perform the LASSO model selection exercise described in the supplementary material multiple times for the response of interest y being equal to a quantitative measure of disease progression one year after baseline. In each empirical exercise, I set one of the 10 regressors in the data set as a predictor of interest z while allowing the remaining nine regressors to be potential control variables X selected by LASSO. I perform these exercises for two values of the LASSO penalty parameter $\lambda \in \{50, 190\}$ to illustrate the merits of the hybrid intervals relative to other confidence intervals under low and

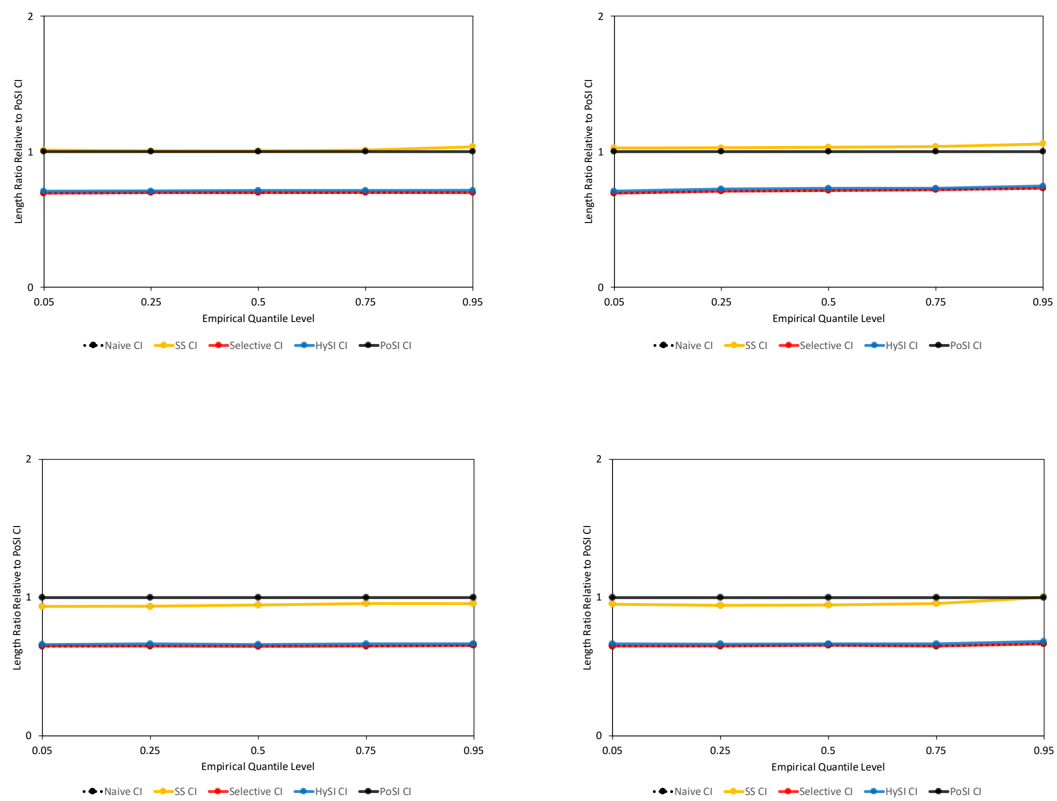

Fig. 3. This figure plots the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles of the lengths of the 95% naive (dotted black), split-sample (orange), selective (red), hybrid (blue) and post-selection (black) confidence intervals divided by the corresponding length quantiles of the 95% post-selection interval across Monte Carlo replications for inference after using LASSO to choose the control variables in the regression with a sample size of $n = 50$ and a large LASSO penalization parameter. The upper-left plot corresponds to a design matrix with independent columns and error terms with a normal distribution for $\lambda = 16$. The upper-right plot corresponds to a design matrix with independent columns and error terms with a skew normal distribution for $\lambda = 16$. The lower-left plot corresponds to a design matrix with correlated columns and error terms with a normal distribution for $\lambda = 100$. The lower-right plot corresponds to a design matrix with correlated columns and error terms with a skew normal distribution for $\lambda = 100$.

high levels of penalization. The penalty parameter of $\lambda = 190$ was examined by Lee et al. (2016) for this data set and corresponds to LASSO selecting three to four control variables across the different predictors of interest. On the other hand, $\lambda = 50$ corresponds to LASSO selecting six to seven controls.

Fig. 4 plots the naive, split-sample (using half of the data for model selection), selective, hybrid (with $\gamma = \alpha/10$) and post-selection nominal 95% confidence intervals for $\lambda = 50$ and $\lambda = 190$ and each of the 10 predictors of interest: Age, Sex, Body-Mass Index (BMI), Blood Pressure (BP) and six different blood serum measurements (S1–S6). Before comparing the intervals, I reiterate that the naive interval does not have correct 95% coverage and that the split-sample intervals cover a different, arguably inferior, target of interest based upon a model selected with half as much data. The left panel of Fig. 4 provides a striking illustration of how much shorter the hybrid intervals can become relative to selective intervals at this lower level of LASSO penalization: the

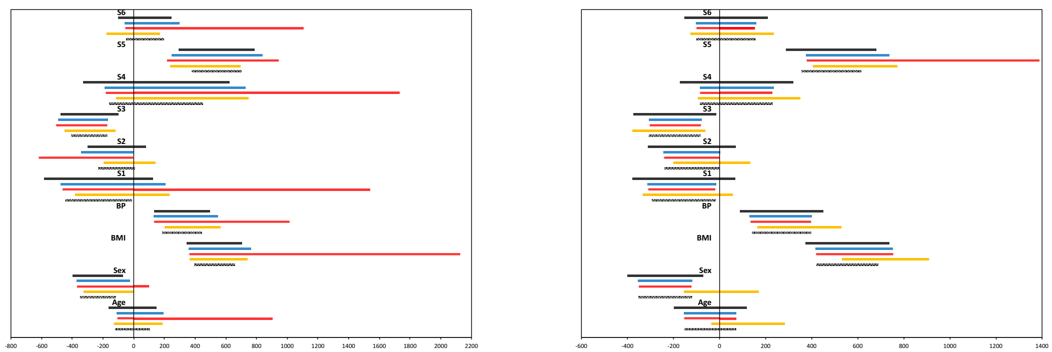

Fig. 4. This figure plots the naive (dotted black), split-sample (orange), selective (red), hybrid (blue) and post-selection (black) confidence intervals for the population coefficient for each of the 10 regressors in the diabetes data set after using LASSO to choose the control variables in the regression with the LASSO penalization parameter set to $\lambda = 50$ (left) and $\lambda = 190$ (right).

hybrid interval is shorter than the selective interval across all predictors of interest, with a length averaging 52% of the latter across the predictors and several length reductions in excess of 65%. In comparison to the post-selection intervals, the hybrid intervals tend to be very similar across predictors of interest with an average length increase of 2%.

In contrast to the left panel, the right panel of Fig. 4 illustrates a more favorable case for the selective intervals at this higher level of LASSO penalization. For all but one predictor of interest (S5), the hybrid intervals are very similar to the selective intervals in these cases where it performs well, entailing slight length increases over the latter of 0–3%. However, the selective interval for S5 is excessively long while the hybrid interval for this same predictor is not, providing a length reduction of nearly 65%. In comparison to the post-selection intervals, the hybrid intervals are shorter for all predictors of interest with an average length reduction in excess of 25% across predictors and reaching more than 35% for several of them. Finally, it is interesting to note the large difference between the split-sample confidence interval and the selective, hybrid and naive intervals when Sex is the predictor of interest: the split-sample interval is nearly centered at zero while the other three nearly coincide and indicate strong evidence that Sex is a strong predictor of diabetes disease progression.

In summary, Fig. 4 provides real world evidence that hybrid intervals perform very similarly to selective intervals (and also naive intervals) in scenarios that are favorable to the latter while transitioning more closely to post-selection intervals in scenarios for which selective intervals become very long.

7. Discussion

Two questions that I did not address in this paper but may be worth investigating in follow-up research are whether post-selection confidence intervals that do not satisfy the structure imposed by Assumption 4 can be used as an ingredient in the construction of hybrid intervals and whether hybrid intervals can be constructed to have correct asymptotic coverage for high-dimensional models with a diverging number of parameters. The first question is interesting in light of recent work dedicated to producing post-selection intervals that are either shorter and/or easier to compute in the presence of many models under consideration (see e.g., Kuchibhotla et al., 2020). For the second question, results in Tibshirani et al. (2018) suggest that uniform asymptotic coverage of hybrid intervals may not be possible in high-dimensional models. On the other hand, results in Tian and Taylor (2017) suggest that point-wise asymptotic coverage may be attainable.

Acknowledgement

I thank Arun Kuchibhotla (Carnegie Mellon University) for very helpful comments and discussions. I am also grateful to anonymous referees and an associate editor for providing feedback that helped improve the paper.

Supplementary material

Supplementary material available at *Biometrika* online includes a theoretical result establishing an expression for the truncation bounds used in the hybrid confidence interval construction with corresponding analytical expressions for these bounds, details on how the general framework of hybrid confidence interval construction specializes to the problem of forming a hybrid confidence interval for a regression coefficient of interest after using LASSO to determine which covariates enter the regression model and the proofs of the theoretical results in this paper.

References

Andrews, I., Kitagawa, T., and McCloskey, A. (2020). Inference on winners. Working Paper.

Andrews, I., Kitagawa, T., and McCloskey, A. (2021). Inference after estimation of breaks. *Journal of Econometrics*, 224:39–59.

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

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

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

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2019). Models as approximations, part I: Consequences illustrated with line regression. *Statistical Science*, 34:523–544.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Fithian, W., Sun, D. L., and Taylor, J. (2017). Optimal inference after model selection. arXiv preprint arXiv:1410.2597v4.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics*, 9:1218–1228.

Kivaranovic, D. and 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.

Kuchibhotla, A. K., Brown, L. D., and Buja, A. (2018). Model-free study of ordinary least squares linear regression. arXiv preprint arXiv:1809.10538v1.

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., and Zhao, L. H. (2020). Valid post-selection inference in model-free linear regression. *Annals of Statistics*, 48:2953–2981.

Kuchibhotla, A. K., Mukherjee, S., and Banerjee, D. (2021). High-dimensional CLT: Improvements, non-uniform extensions and large deviations. *Bernoulli*, 27:192–217.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the LASSO. *Annals of Statistics*, 44:907–927.

Markovic, J., Xia, L., and Taylor, J. (2018). Unifying approach to selective inference with applications to cross-validation. arXiv preprint arXiv:1703.06559v3.

Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scandanavian Journal of Statistics*, 44:480–499.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Annals of Statistics*, 46:679–710.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics*, 46:1255–1287.

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

[*Received on 23 November 2021. Editorial decision on ? ?? ??*]

Biometrika (2023), **??**, ??, *pp*. 1–8
Printed in Great Britain Advance Access publication on ??

Supplementary material for "Hybrid Confidence Intervals for Informative Uniform Asymptotic Inference After Model Selection"

BY A. MCCLOSKEY
Department of Economics, University of Colorado at Boulder, 256 UCB, Boulder, Colorado 80309, U.S.A.
adam.mccloskey@colorado.edu

SUMMARY

The first section of this supplementary material contains a theoretical result establishing an expression for the truncation bounds used in the hybrid confidence interval construction with corresponding analytical expressions for these bounds. The second section specializes the general framework of hybrid confidence interval construction to the general problem of constructing a hybrid confidence interval for a regression coefficient of interest after using LASSO to determine which covariates enter the regression model. The third section of this supplementary material provides the proofs of the theoretical results in "Hybrid Confidence Intervals for Informative Uniform Asymptotic Inference After Model Selection."

1. TRUNCATION BOUNDS FOR HYBRID CONFIDENCE INTERVALS

The bounds of the truncation interval for the target statistic $T_n(\widehat{M}_n)$ can be expressed in terms of a directly-computable random vector $Z_{M,n}$ that is asymptotically independent of $T_n(M)$:

$$Z_{M,n} = D_n(M) - \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right) T_n(M),$$

where $\widehat{\Sigma}^{(M)}_{DT,n}$ is the estimated covariance vector between $T_n(M)$ and $D_n(M)$. The following lemma follows from a slight extension of the arguments used to prove Lemma 5.1 in Lee et al. (2016).

LEMMA 1. *Under Assumption 1, the conditioning set for any model $M \in \mathcal{M}$ being selected can be expressed as follows:*

$$\left\{\widehat{M}_n = M\right\} = \left\{\mathcal{V}^-_{M,n}(Z_{M,n}) \le T_n(M) \le \mathcal{V}^+_{M,n}(Z_{M,n}), \mathcal{V}^0_{M,n}(Z_{M,n}) \ge 0\right\},$$

where

$$\mathcal{V}^-_{M,n}(z) = \max_{j:(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j<0} \frac{\widehat{a}_{M,n,j} - (A_M z)_j}{(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j}$$

$$\mathcal{V}^+_{M,n}(z) = \min_{j:(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j>0} \frac{\widehat{a}_{M,n,j} - (A_M z)_j}{(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j}$$

$$\mathcal{V}^0_{M,n}(z) = \min_{j:(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j=0} \widehat{a}_{M,n,j} - (A_M z)_j.$$

Proof. By Assumption 1,

$$
\begin{aligned}
\left\{\widehat{M}_n = M\right\} &= \{A_M D_n(M) \leq \widehat{a}_{M,n}\} \\
&= \left\{A_M \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right) T_n(M) \leq \widehat{a}_{M,n} - A_M Z_{M,n}\right\} \\
&= \left\{\left(A_M \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right)\right)_j T_n(M) \leq \widehat{a}_{M,n,j} - (A_M Z_{M,n})_j\right\} \\
&= \left\{\begin{matrix} T_n(M) \leq \frac{\widehat{a}_{M,n,j} - (A_M Z_{M,n})_j}{(A_M \widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j}, \text{ for } j : (A_M \widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j > 0 \\ T_n(M) \geq \frac{\widehat{a}_{M,n,j} - (A_M Z_{M,n})_j}{(A_M \widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j}, \text{ for } j : (A_M \widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j < 0 \\ 0 \leq \widehat{a}_{M,n,j} - (A_M Z_{M,n})_j \qquad \text{for } j : (A_M \widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j = 0 \end{matrix}\right\}.
\end{aligned}
$$

The statement of the lemma immediately follows. □

Upon intersecting the event

$$\left\{\widehat{M}_n = M\right\} = \left\{\mathcal{V}^-_{M,n}(Z_{M,n}) \leq T_n(M) \leq \mathcal{V}^+_{M,n}(Z_{M,n}), \mathcal{V}^0_{M,n}(Z_{M,n}) \geq 0\right\}$$

with the event

$$
\begin{aligned}
&\left\{\mu_{T,n}(\widehat{M}_n) \in CI^{P,\gamma}_{n,\widehat{M}_n}\right\} \\
&= \left\{\mu_{T,n}(\widehat{M}_n) - \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)} K_{n,\gamma} \leq T_n(\widehat{M}_n) \leq \mu_{T,n}(\widehat{M}_n) + \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)} K_{n,\gamma}\right\},
\end{aligned}
$$

(by Assumption 4) we obtain the truncation interval for $T_n(\widehat{M}_n)$ used to construct the hybrid confidence interval:

$$
\begin{aligned}
&\left\{\widehat{M}_n = M\right\} \cap \left\{\mu_{T,n}(\widehat{M}_n) \in CI^{P,\gamma}_{n,\widehat{M}_n}\right\} \\
&= \left\{\mathcal{V}^{-,H}_{M,n}(Z_{M,n}, \mu_{T,n}(M)) \leq T_n(M) \leq \mathcal{V}^{+,H}_{M,n}(Z_{M,n}, \mu_{T,n}(M)), \mathcal{V}^0_{M,n}(Z_{M,n}) \geq 0\right\},
\end{aligned}
$$

where

$$\mathcal{V}^{-,H}_{M,n}(z,\mu) = \max\left\{\mathcal{V}^-_{M,n}(z),\ \mu - \sqrt{\widehat{\Sigma}_{T,n}(M,M)} K_{n,\gamma}\right\},$$

$$\mathcal{V}^{+,H}_{M,n}(z,\mu) = \min\left\{\mathcal{V}^+_{M,n}(z),\ \mu + \sqrt{\widehat{\Sigma}_{T,n}(M,M)} K_{n,\gamma}\right\}.$$

2. Application to Inference After LASSO Model Selection

I now specialize the general framework to the problem of constructing a hybrid confidence interval for a regression coefficient of interest after using LASSO to determine which covariates enter the regression model, dropping the simplifying assumptions of Section 2 of the main text. Formally, suppose we have data $(z, X, y) \in \mathbb{R}^n \times \mathbb{R}^{n\times p} \times \mathbb{R}^n$ for which the rows of y and z are identically distributed random variables and the rows of X are either identically distributed random vectors or have entries equal to one (corresponding to an intercept term). (The rows of an arbitrary matrix or vector B are denoted as B_i.) We are interested in the population regression coefficient corresponding to the predictor of interest z after selecting which of the control variables in X should enter the regression model according to the non-zero subset of the vector $\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\|y^* - X^*\beta\|_2^2 + \lambda\|\beta\|_1$$

with $y^* = (I - P_z)y$ and $X^* = (I - P_z)X$ for $P_z = zz^T/z^Tz$ and λ being the LASSO penalty parameter. Letting $\widehat{E}_n$ denote the set of non-zero coefficients of $\widehat{\beta}$, we can characterize a model M as a set of LASSO-selected controls E and the sign of the LASSO regression coefficients corresponding to the selected controls s_E. In other words, a given model M is defined as a tuple (E, s_E).

Using the Karush-Khun-Tucker conditions for optimizing the LASSO objective function, Lee et al. (2016) show that $\widehat{M}_n = (\widehat{E}_n, \operatorname{sign}(\widehat{\beta}_{\widehat{E}_n})) = (E, s_E) = M$ if and only if $A_M D_n(M) \leq \widehat{a}_{M,n}$, where

$$A_M = \begin{pmatrix} -\operatorname{diag}(s_E) & 0 \\ 0 & I_{p-|E|} \\ 0 & -I_{p-|E|} \end{pmatrix},$$

$$D_n(M) = \begin{pmatrix} \sqrt{n}(X_E^{*T}X_E^*)^{-1}X_E^{*T}y^* \\ n^{-1/2}X_{-E}^{*T}(y^* - X_E^*(X_E^{*T}X_E^*)^{-1}X_E^{*T}y^*) \end{pmatrix},$$

$$\widehat{a}_{M,n} = \begin{pmatrix} -\lambda\sqrt{n}\operatorname{diag}(s_E)(X_E^{*T}X_E^*)^{-1}s_E \\ \lambda n^{-1/2}1_{p-|E|} - \lambda n^{-1/2}X_{-E}^{*T}X_E^*(X_E^{*T}X_E^*)^{-1}s_E \\ \lambda n^{-1/2}1_{p-|E|} + \lambda n^{-1/2}X_{-E}^{*T}X_E^*(X_E^{*T}X_E^*)^{-1}s_E \end{pmatrix},$$

with X_E^* equal to the submatrix of X^* composed of the columns of X^* corresponding to E and X_{-E}^* equal to the submatrix of X^* composed of the remaining columns. Let $\tilde{X} = X - z(\mathbb{E}_{\mathbb{P}}[z'z])^{-1}\mathbb{E}_{\mathbb{P}}[z'X]$ and let $\tilde{X}_E^T$ denote the submatrix of $\tilde{X}$ composed of the columns of $\tilde{X}$ corresponding to E and $\tilde{X}_{-E}^T$ denote the submatrix of $\tilde{X}$ composed of the remaining columns. Assumption 1 thus holds with

$$a_{M,n}(\mathbb{P}) = \begin{pmatrix} -\lambda n^{-1/2}\operatorname{diag}(s_E)(\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{X}_{E,i}^T])^{-1}s_E \\ \lambda n^{-1/2}1_{p-|E|} - \lambda n^{-1/2}\mathbb{E}_{\mathbb{P}}[\tilde{X}_{-E,i}\tilde{X}_{E,i}^T](\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{X}_{E,i}^T])^{-1}s_E \\ \lambda n^{-1/2}1_{p-|E|} + \lambda n^{-1/2}\mathbb{E}_{\mathbb{P}}[\tilde{X}_{-E,i}\tilde{X}_{E,i}^T](\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{X}_{E,i}^T])^{-1}s_E \end{pmatrix}$$

under standard moment, stationarity and dependence conditions on $\mathcal{P}_n$ that imply a uniform law of large numbers for $\widehat{a}_{M,n}$ and uniform moment bounds on $\mathbb{E}_{\mathbb{P}}[\tilde{X}_{-E,i}\tilde{X}_{E,i}^T]$ and $\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{X}_{E,i}^T]$.

Letting $W_E = (z, X_E)$ and e_1 denote the first standard basis vector, we are interested in forming a confidence interval that covers the (scaled) population regression coefficient on z in the selected model as the target parameter

$$\mu_{T,n}(\mathbb{P}, M) = \sqrt{n}e_1^T(\mathbb{E}_{\mathbb{P}}[W_{E,i}W_{E,i}^T])^{-1}\mathbb{E}_{\mathbb{P}}[W_{E,i}y_i]$$

for $E = \widehat{E}_n$ using the corresponding sample regression coefficient

$$T_n(M) = \sqrt{n}e_1'(W_E'W_E)^{-1}W_E'y$$

as a statistic. (Confidence intervals for unscaled population regression coefficients are formed by simply dividing the confidence intervals for $\mu_{T,n}(\mathbb{P}, M)$ by $\sqrt{n}$.) With these definitions in mind, as well as

$$\mu_{D,n}(\mathbb{P},M)=\begin{pmatrix}\sqrt{n}(\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{X}_{E,i}'])^{-1}\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{y}_i]\\ \sqrt{n}(\mathbb{E}_{\mathbb{P}}[\tilde{X}_{-E,i}\tilde{y}_i]-\mathbb{E}_{\mathbb{P}}[\tilde{X}_{-E,i}\tilde{X}_{E,i}'](\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{X}_{E,i}'])^{-1}\mathbb{E}_{\mathbb{P}}[\tilde{X}_{E,i}\tilde{y}_i])\end{pmatrix},$$

Assumption 2 holds under standard moment, stationarity and dependence conditions on $\mathcal{P}_n$ that imply a multivariate uniform central limit theorem for the vector $(T_n', D_n')'$. For Assumption 3, consider the heteroskedasticity-robust estimators $\widehat{\Sigma}_{T,n}$ and $\widehat{\Sigma}_{DT,n}$ for which

$$\widehat{\Sigma}_{T,n}(M,M^T)$$
$$=e_1^T\left(\frac{1}{n}\sum_{i=1}^n W_{E,i}W_{E,i}^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n W_{E,i}W_{E^T,i}^T\hat{u}_{E,i}\hat{u}_{E^T,i}\right)\left(\frac{1}{n}\sum_{i=1}^n W_{E^T,i}W_{E^T,i}^T\right)^{-1}e_1,$$

$$\widehat{\Sigma}_{DT,n}((M-1)p+1:Mp,M^T)$$
$$=\begin{pmatrix}\left(\frac{1}{n}\sum_{i=1}^n X_{E,i}^*X_{E,i}^{*T}\right)^{-1} & 0\\ 0 & I\end{pmatrix}\left(\frac{1}{n}\sum_{i=1}^n\begin{pmatrix}X_{E,i}^*\\ X_{-E,i}^*\end{pmatrix}W_{E^T,i}^Tu_{E,i}^*\hat{u}_{E^T,i}\right)$$
$$\times\left(\frac{1}{n}\sum_{i=1}^n W_{E^T,i}W_{E^T,i}^T\right)^{-1}e_1$$

where $\hat{u}_{E,i}=y_i-W_{E,i}^T\widehat{\beta}_{M,n}$ and $u_{E,i}^*=y_i^*-X_{E,i}^*\widehat{\beta}_{M,n}^*$ with $\widehat{\beta}_{M,n}=(W_E^TW_E)^{-1}W_E^Ty$ and $\widehat{\beta}_{M,n}^*=(X_E^{*T}X_E^*)^{-1}X_E^{*T}y^*$. Slight extensions of the arguments in Kuchibhotla et al. (2018) from pointwise to uniform consistency provide that Assumption 3 holds for $\widehat{\Sigma}_{T,n}$ and $\widehat{\Sigma}_{DT,n}$ when the data are independent under standard moment conditions on $\mathcal{P}_n$.

Finally, Assumption 4 holds by the results of Bachoc et al. (2020) when using one of the post-selection confidence intervals discussed in that paper. In particular, let $K_{n,\alpha}$ equal the $(1-\alpha)$-quantile of

$$\max_i|Z_i| \text{ for } Z\sim\mathcal{N}(0,\Omega)$$

with $\Omega=\operatorname{corr}(\widehat{\Sigma}_{T,n})\equiv\operatorname{diag}(\widehat{\Sigma}_{T,n})^{\dagger/2}\widehat{\Sigma}_{T,n}\operatorname{diag}(\widehat{\Sigma}_{T,n})^{\dagger/2}$, where $A^\dagger$ denotes the Moore-Penrose inverse of matrix A and $A^{1/2}$ denotes the symmetric nonnegative definite square root of a symmetric nonnegative definite matrix A. By Assumption 3,

$$\lim_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(|K_{n,\alpha}-K_\alpha(\mathbb{P})|>\varepsilon\right)=0$$

for any $\varepsilon>0$, where $K_\alpha(\mathbb{P})$ is equal to the $(1-\alpha)$-quantile of $\max_i|Z_i|$ for $Z\sim\mathcal{N}(0,\Omega)$ with $\Omega=\operatorname{corr}(\Sigma_T(\mathbb{P}))$. For $\alpha\neq 1$, $0\leq K_\alpha(\mathbb{P})\leq\bar{\lambda}$ for some finite $\bar{\lambda}$ and any probability measure $\mathbb{P}$. Theorem 2.3 of Bachoc et al. (2020) provides sufficient conditions on $\mathcal{P}_n$ that imply $\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI_{n,\widehat{M}_n}^{P,\alpha}\right)\geq 1-\alpha$ for $CI_{n,\widehat{M}_n}^{P,\alpha}$ formed according to Assumption 4. The form of post-selection interval introduced here is a less conservaitve version that incorporates the fact that the predictor of interest z is protected from variable selection, referred to as "PoSI1" by Berk et al. (2013).

3. PROOFS OF THEORETICAL RESULTS

The following lemma is useful for proving the correct uniform asymptotic coverage of the hybrid confidence interval $CI^{H,\alpha}_{n,\widehat{M}_n}$.

LEMMA 2. *For* $Z^*_{M,n} = D_n(M) - \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right) T^*_n(M)$ *with* $T^*_n(M) = T_n(M) - \mu_{T,n}(M)$, $\mathcal{V}^-_{M,n}(Z^*_{M,n}) = \mathcal{V}^-_{M,n}(Z_{M,n}) - \mu_{T,n}(M)$ *and* $\mathcal{V}^+_{M,n}(Z^*_{M,n}) = \mathcal{V}^+_{M,n}(Z_{M,n}) - \mu_{T,n}(M)$.

Proof. Noting that

$$\begin{aligned} Z^*_{M,n} &= D_n(M) - \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right) T_n(M) + \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right) \mu_{T,n}(M) \\ &= Z_{M,n} + \left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right) \mu_{T,n}(M), \end{aligned}$$

we have

$$\begin{aligned} \mathcal{V}^-_{M,n}(Z^*_{M,n}) &= \max_{j:(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j<0} \frac{\widehat{a}_{M,n,j} - (A_M Z^*_{M,n})_j}{(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j} \\ = \max_{j:(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j<0} & \frac{\widehat{a}_{M,n,j} - (A_M Z_{M,n})_j - \left(A_M\left(\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M)\right)\mu_{T,n}(M)\right)_j}{(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j} \\ &= \max_{j:(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j<0} \frac{\widehat{a}_{M,n,j} - (A_M Z_{M,n})_j}{(A_M\widehat{\Sigma}^{(M)}_{DT,n}/\widehat{\Sigma}_{T,n}(M,M))_j} - \mu_{T,n}(M) \\ &= \mathcal{V}^-_{M,n}(Z_{M,n}) - \mu_{T,n}(M). \end{aligned}$$

The proof for $\mathcal{V}^+_{M,n}(Z^*_{M,n})$ is entirely analogous and therefore omitted. □

Proof of Proposition 1. By the same argument used in the proof of Proposition 5 in Andrews et al. (2020),

$$F_{TN}(t;\mu,\Sigma_T(M,M),\mathcal{V}^{-,H}_{M,n}(z,\mu),\mathcal{V}^{+,H}_{M,n}(z,\mu))$$

is decreasing in μ so that $\widehat{\mu}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n) \geq \mu_{T,n}(\widehat{M}_n)$ is equivalent to

$$\begin{aligned} F_{TN}\left(T_n(\widehat{M}_n);\mu_{T,n}(\widehat{M}_n),\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n),\mathcal{V}^{-,H}_{\widehat{M}_n,n}(Z_{\widehat{M}_n,n},\mu_{T,n}(\widehat{M}_n)),\mathcal{V}^{+,H}_{\widehat{M}_n,n}(Z_{\widehat{M}_n,n},\mu_{T,n}(\widehat{M}_n))\right) \\ \geq 1 - \frac{\alpha-\gamma}{2(1-\gamma)}. \end{aligned}$$

Further, Lemma 2 implies

$$
\begin{aligned}
&F_{TN}\left(T_n(\widehat{M}_n); \mu_{T,n}(\widehat{M}_n), \widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n), \mathcal{V}^{-,H}_{\widehat{M}_n,n}(Z_{\widehat{M}_n,n}, \mu_{T,n}(\widehat{M}_n)), \mathcal{V}^{+,H}_{\widehat{M}_n,n}(Z_{\widehat{M}_n,n}, \mu_{T,n}(\widehat{M}_n))\right)\\
&= F_{TN}\left(T_n^*(\widehat{M}_n) + \mu_{T,n}(\widehat{M}_n); \mu_{T,n}(\widehat{M}_n), \widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n),\right.\\
&\qquad \max\left\{\mathcal{V}^{-}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}) + \mu_{T,n}(\widehat{M}_n), \mu_{T,n}(\widehat{M}_n) - \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}\right\},\\
&\qquad \left.\min\left\{\mathcal{V}^{+}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}) + \mu_{T,n}(\widehat{M}_n), \mu_{T,n}(\widehat{M}_n) + \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}\right\}\right)\\
&= F_{TN}\left(T_n^*(\widehat{M}_n); 0, \widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n), \max\left\{\mathcal{V}^{-}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}), -\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}\right\},\right.\\
&\qquad \left.\min\left\{\mathcal{V}^{+}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}), \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}\right\}\right)
\end{aligned}
$$

so that $\widehat{\mu}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}_{T,n}(\widehat{M}_n) \geq \mu_{T,n}(\widehat{M}_n)$ is equivalent to

$$
\begin{aligned}
&F_{TN}\left(T_n^*(\widehat{M}_n); 0, \widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n), \max\left\{\mathcal{V}^{-}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}), -\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}\right\},\right.\\
&\qquad \left.\min\left\{\mathcal{V}^{+}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}), \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}\right\}\right) \geq 1 - \frac{\alpha-\gamma}{2(1-\gamma)}. \qquad (1)
\end{aligned}
$$

By an extension of Lemma 5 of Andrews et al. (2020), to prove the statment of the proposition, it suffices to show that for all subsequences $\{n_s\} \subset \{n\}$, $\{\mathbb{P}_{n_s}\} \in \times_{n=1}^{\infty}\mathcal{P}_n$ with

1. $\Sigma(\mathbb{P}_{n_s}) \to \Sigma^* \in \mathcal{S}$

$$\mathcal{S} = \{\Sigma : 1/\bar{\lambda} \leq \Sigma_T(M, M) \leq \bar{\lambda}, 1/\bar{\lambda} \leq \lambda_{\min}(\Sigma_D^{(M)}) \leq \lambda_{\max}(\Sigma_D^{(M)}) \leq \bar{\lambda}\},$$

2. $K_\gamma(\mathbb{P}_{n_s}) \to K_\gamma^* \in [0, \bar{\lambda}]$,
3. $a_{M,n_s}(\mathbb{P}_{n_s}) \to a_M^* \in [-\bar{\lambda}, \bar{\lambda}]^{\dim(a_M)}$,
4. $\mathbb{P}_{n_s}\left(\widehat{M}_{n_s} = M, \mu_{T,n_s}(\widehat{M}_{n_s}) \in CI^{P,\gamma}_{n_s,\widehat{M}_{n_s}}\right) \to p^* \in (0, 1]$, and
5. $\mu_{D,n_s}(M; \mathbb{P}_{n_s}) \to \mu_D^*(M) \in [-\infty, \infty]^{\dim(D^*(M))}$

for some finite $\bar{\lambda}$, we have

$$\lim_{n\to\infty} \mathbb{P}_{n_s}\left(\mu_{T,n_s}(\widehat{M}_{n_s}) \in CI^{\alpha,H}_{n_s,\widehat{M}_{n_s}} \middle| \widehat{M}_{n_s} = M, \mu_{T,n_s}(\widehat{M}_{n_s}) \in CI^{P,\gamma}_{n_s,\widehat{M}_{n_s}}\right) = \frac{1-\alpha}{1-\gamma}.$$

Let $\{\mathbb{P}_{n_s}\}$ be a sequence satisfying conditions 1.–5. Now under $\{\mathbb{P}_{n_s}\}$, $(T^*_{n_s}, \widehat{\Sigma}_{T,n_s}, \widehat{\Sigma}_{DT,n_s}, K_{n_s,\gamma}) \xrightarrow{d} (T^*, \Sigma_T^*, \Sigma_{DT}^*, K_\gamma^*)$ by Assumptions 2–4, where $T_n^* = (T_n^*(1), \ldots, T_n^*(|\mathcal{M}|)^T$ and $T^* \sim \mathcal{N}(0, \Sigma_T^*)$. In addition, conditions 3.–4. along with Assumptions 1–2 imply that under $\{\mathbb{P}_{n_s}\}$, $\widehat{M}_{n_s} \xrightarrow{d} \widehat{M}$, where $\widehat{M} = M \in \mathcal{M}$ if and only if $A_M(D^*(M) + \mu_D^*(M)) \leq a_M^*$ with $D^*(M) \sim \mathcal{N}(0, \Sigma_D^{(M)*})$. This convergence occurs jointly with that for $(T^*_{n_s}, \widehat{\Sigma}_{T,n_s}, \widehat{\Sigma}_{DT,n_s}, K_{n_s,\gamma})$. Note that it is not possible for $(A_M\mu_D^*(M))_j = \infty$ for any j under conditions 3.–4. and Assumptions 1–2. Thus, under Assumptions 1–3, similar arguments to those used in the proof of Lemma 8 in Andrews et al. (2020) show that for any $M \in \mathcal{M}$, $(\mathcal{V}^-_{M,n_s}(Z^*_{M,n_s}), \mathcal{V}^+_{M,n_s}(Z^*_{M,n_s})) \xrightarrow{d} (\mathcal{V}^-_M(Z^*_M), \mathcal{V}^+_M(Z^*_M))$ under $\{\mathbb{P}_{n_s}\}$, where

$\mathcal{V}_M^-(z)$ and $\mathcal{V}_M^+(z)$ are defined identically to $\mathcal{V}_{M,n}^-(z)$ and $\mathcal{V}_{M,n}^+(z)$ after replacing $\widehat{\Sigma}_{T,n}$, $\widehat{\Sigma}_{DT,n}$ and $\widehat{a}_{M,n}$ with Σ_T^*, Σ_{DT}^* and a_M^* and Z_M^* is defined identically to $Z_{M,n}^*$ after replacing $\widehat{\Sigma}_{T,n}$, $\widehat{\Sigma}_{DT,n}$, $D_n(M)$ and $T_n^*(M)$ with Σ_T^*, Σ_{DT}^*, $D^*(M)+\mu_D^*(M)$ and $T^*(M)$. This convergence is joint with that of $(T_{n_s}^*, \widehat{\Sigma}_{n_s}, K_{n_s,\gamma}, \widehat{M}_{n_s})$ so that we may write

$$\begin{aligned}(T_{n_s}^*, \widehat{\Sigma}_{n_s}, K_{n_s,\gamma}, \widehat{M}_{n_s}, \mathcal{V}_{M,n_s}^-(Z_{M,n_s}^*), \mathcal{V}_{M,n_s}^+(Z_{M,n_s}^*)) \\ \xrightarrow{d} (T^*, \Sigma^*, K_\gamma^*, \widehat{M}, \mathcal{V}_M^-(Z_M^*), \mathcal{V}_M^+(Z_M^*))\end{aligned} \tag{2}$$

under $\{\mathbb{P}_{n_s}\}$ for any $M \in \mathcal{M}$.

By Lemma 9 of Andrews et al. (2020), $F_{TN}(t;\mu,\Sigma_T(M,M),\mathcal{L},\mathcal{U})$ is continuous over the set

$$\left\{(t,\mu,\Sigma_T(M,M)) \in \mathbb{R}^3, \mathcal{L} \in \mathbb{R}\cup\{-\infty\}, \mathcal{U} \in \mathbb{R}\cup\{\infty\} : \Sigma_T(M,M) > 0, \mathcal{L} < t < \mathcal{U}\right\}$$

so that with Assumption 4, (2) implies

$$\begin{aligned}&\Bigg(F_{TN}(T_{n_s}^*(\widehat{M}_{n_s}); 0, \widehat{\Sigma}_{T,n_s}(\widehat{M}_{n_s}, \widehat{M}_{n_s}), \max\left\{\mathcal{V}_{\widehat{M}_{n_s},n_s}^-(Z_{\widehat{M}_{n_s},n_s}^*), -\sqrt{\widehat{\Sigma}_{T,n_s}(\widehat{M}_{n_s}, \widehat{M}_{n_s})}K_{n_s,\gamma}\right\}, \\ &\min\left\{\mathcal{V}_{\widehat{M}_{n_s},n_s}^+(Z_{\widehat{M}_{n_s},n_s}^*), \sqrt{\widehat{\Sigma}_{T,n_s}(\widehat{M}_{n_s}, \widehat{M}_{n_s})}K_{n_s,\gamma}\right\}, \mathbf{1}(\widehat{M}_{n_s} = M, \mu_{T,n_s}(\widehat{M}_{n_s}) \in CI_{n_s,\widehat{M}_{n_s}}^{P,\gamma})\Bigg) \\ &\xrightarrow{d} \Bigg(F_{TN}(T^*(\widehat{M}); 0, \Sigma_T^*(\widehat{M}, \widehat{M}), \max\left\{\mathcal{V}_{\widehat{M}}^-(Z_{\widehat{M}}^*), -\sqrt{\Sigma_T^*(\widehat{M}, \widehat{M})}K_\gamma^*\right\}, \\ &\qquad \min\left\{\mathcal{V}_{\widehat{M}}^+(Z_{\widehat{M}}^*), \sqrt{\Sigma_T^*(\widehat{M}, \widehat{M})}K_\gamma^*\right\}, \\ &\qquad \mathbf{1}\left(\widehat{M} = M, -\sqrt{\Sigma_T^*(\widehat{M}, \widehat{M})}K_\gamma^* \le T^*(\widehat{M}) \le \sqrt{\Sigma_T^*(\widehat{M}, \widehat{M})}K_\gamma^*\right)\Bigg),\end{aligned} \tag{3}$$

since $\mu_{T,n}(\widehat{M}_n) \in CI_{n,\widehat{M}_n}^{P,\gamma}$ is equivalent to

$$-\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma} \le T_n^*(\widehat{M}_n) \le \sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n, \widehat{M}_n)}K_{n,\gamma}.$$

Given the equivalence in (1), Lemma 1 and (3), the result of the proposition follows from the same arguments used to prove the first part of Corollary 2 of Andrews et al. (2020). □

Proof of Proposition 2. To see why the first inequality holds, note the following:

$$\begin{aligned}&\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{H,\alpha}\right) \\ &\ge \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{H,\alpha} \,\middle|\, \mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{P,\gamma}\right) \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{P,\gamma}\right) \\ &= \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \sum_{M\in\mathcal{M}} \Bigg\{\mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{H,\alpha} \,\middle|\, \widehat{M}_n = M, \mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{P,\gamma}\right) \\ &\qquad\qquad \times \mathbb{P}\left(\widehat{M}_n = M, \mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{P,\gamma}\right)\Bigg\} \\ &\ge \frac{1-\alpha}{1-\gamma} \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \sum_{M\in\mathcal{M}} \mathbb{P}\left(\widehat{M}_n = M, \mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{P,\gamma}\right) \\ &= \frac{1-\alpha}{1-\gamma} \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI_{n,\widehat{M}_n}^{P,\gamma}\right) \ge \frac{1-\alpha}{1-\gamma}(1-\gamma) = 1-\alpha,\end{aligned}$$

where the second inequality follows from Lemma 6 of Andrews et al. (2020) and Proposition 1 and the final inequality holds by Assumption 4. The second inequality in the proposition follows from essentially the same argument used to prove the final part of Corollary 2 of Andrews et al. (2020). □

References

Andrews, I., Kitagawa, T., and McCloskey, A. (2020). Inference on winners. Working Paper.

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

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

Kuchibhotla, A. K., Brown, L. D., and Buja, A. (2018). Model-free study of ordinary least squares linear regression. arXiv prepreint arXiv:1809.10538v1.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the LASSO. *Annals of Statistics*, 44:907–927.

[*Received on 23 November 2021. Editorial decision on ? ?? ??*]