Hybrid Confidence Intervals for Informative Uniform Asymptotic Inference After Model Selection*

Adam McCloskey†

November 11, 2020

Abstract

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 confidence interval that draws on the relative strengths of these two inference techniques with desirable length properties 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. 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 properties in finite samples via a set of Monte Carlo exercises that is calibrated to real-world data.

Keywords: Confidence Interval, Selective Inference, Post-Selection Inference, LASSO, Uniform Asymptotics, Misspecification

*I thank Arun Kuchibhotla for helpful comments and discussions.

†Department of Economics, University of Colorado, adam.mccloskey@colorado.edu

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 (PoSI) inference (e.g., Berk et al., 2013), terms I will use throughout this paper.

Apart from the differences in coverage guarantees from confidence intervals (CIs) constructed from these two approaches, they also feature complementary strengths and weaknesses in terms of informativeness. While selective CIs 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 CIs approach "naive" CIs (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., 2020b). On the other hand, their expected length may be infinite (Kivaranovic and Leeb, 2020). PoSI CIs 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 CIs.

In this paper, I propose a new class of hybrid CIs for inference after model selection that aims to draw on the complementary informativeness strengths of selective and PoSI CIs. Like selective CIs, but unlike PoSI CIs, these hybrid CIs 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 CIs to produce CIs that are valid "on average" across the models potentially selected by the data, a feature shared by PoSI CIs. This approach focuses on the case for which the predictor variable of interest enters all models under consideration, so that this unconditional coverage requirement is well-defined (see Berk et al., 2013 and Lee et al., 2016 for a discussion of this issue). The relaxation of the conditional coverage requirement can be viewed as an alternative approach for improving the length properties

of selective CIs to the "data-carving" approach of Fithian et al. (2017) or the randomized response approach of Tian and Taylor (2018), for which noise is added to the data in the model selection stage. The hybrid CIs introduced in this paper make use of similar reasoning to the hybrid CIs of Andrews et al. (2020b) and Andrews et al. (2020a), but applied to a model selection framework that generalizes that of Andrews et al. (2020b) to incorporate many popular model selection criteria used in linear regression.

The hybrid approach to CI construction improves upon the typical length properties of both the selective and PoSI CIs by modifying the conditioning event of selective CIs to restrict the selected parameter of interest to lie within a PoSI CI with higher nominal coverage probability. Since this latter addition to the conditioning event is not necessarily satisfied by the data, the hybrid approach takes the selective CI based upon this event but modifies the truncated normal quantiles used in its constructoin to maintain correct coverage. It also limits the maximum length of the hybrid CI to be bounded above by the length of the corresponding higher coverage PoSI CI, yielding finite expected length and breaking the negative result of Kivaranovic and Leeb (2020). At the same time, hybrid CIs can be configured to *nearly* approach "naive" CIs when the model selected by the data is selected with high probability in the sense that they approach "naive" CIs 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 PoSI CI, I establish the uniform asymptotic validity of the hybrid CIs 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 particular linear regression model selection contexts. Indeed, I discuss 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.

Using insights of Markovic et al. (2018), 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 or selected by a randomized information criterion. Unlike the point-wise asymptotic results of e.g., Tian and Taylor (2017), the *uniform* asymptotic results I establish in this paper ought to 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. (2020a) 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 CIs relative to the standard selective CIs of Lee et al. (2016) and PoSI CIs of Bachoc et al. (2020) when using LASSO as the model selection criterion in a set of Monte Carlo experiments calibrated to real-world data. Using the diabetes dataset from Efron et al. (2004) (also examined by Lee et al., 2016), I draw Monte Carlo samples via nonparametric bootstrap on the original dataset in order to produce empirically-relevant simulated datasets that significantly depart from Gaussianity. Under a wide range of values for the LASSO penalty parameter, I find that the length distribution of hybrid CIs compares very favorably to those of selective and PoSI CIs. The length gains relative to selective CIs 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 CIs can be more than 16 times larger than that of the hybrid CIs under the Monte Carlo designs I examine. In addition, I find that the hybrid CIs exhibit correct finite-sample coverage.

The remainder of this paper is organized as follows. Section 2 introduces the basic intuition for and construction of hybrid CIs. Section 3 lays out the general model selection framework under study involving an affine constraint on a vector of statistics. Section 4 details the general construction of hybrid CIs and includes the main theoretical results of this paper on the correct uniform asymptotic coverage of hybrid CIs. In Section 5, I show how the general framework of Sections 3 and 4 specialize to constructing a hybrid CI for a regression coefficient after using the LASSO objective function to select the regression model of interest. Section 6 presents Monte Carlo simulation results for the post-LASSO model selection exercise calibrated to a real-world dataset. Finally, Section 7 concludes while proofs of the theoretical results in the paper are contained in the Technical Appendix. All tables and figures are collected at the end of the document.

2 Basic Ideas Behind the Hybrid Approach

Motivated by the fact that selective CIs can become extremely wide under certain realizations of the data, by relaxing the conditional coverage requirement it may be possible to attain CIs with guaranteed unconditional coverage and better length properties. 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 $X_{1i},...,X_{pi}$ for $i=1,...,n$. Without imposing any assumptions about the true underlying relationship between the response, predictor of interest and controls, the researcher chooses a model $M \subset \{1,...,p\}$ as the subset of indices corresponding to the controls of interest. The researcher's target parameter of interest is equal to the population linear regression coefficient θ_M defined by

$$(\theta_M, \beta_M) = \text{argmin}_{\theta \in \mathbb{R}, b_M \in \mathbb{R}^{|M|}} E\|y - z\theta - X(M)b_M\|^2,$$

where $y = (y_1,...,y_n)'$, $z = (z_1,...,z_n)'$ and $X(M)$ is the submatrix of the design matrix $X = (x_1,...,x_p)$ corresponding to model M with $x_k = (X_{k1},...,X_{kn})'$.

To establish the basic arguments, let us temporarily assume we have CIs with correct finite-sample coverage, regression coefficients that are normally distributed in finite samples and that variance parameters are known. There is now a large literature enabling the construction of selective CIs with (asymptotically) correct coverage for a population regression coefficient conditional on the model $\widehat{M} \subset \{1,...,p\}$ 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 CI $CI^{S,\alpha}_{\widehat{M}}$ such that

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

for all $M \subset \{1,...,p\}$ (or some other relevant subset of the universe of models). On the other hand, there is a growing literature enabling the construction of PoSI CIs with correct unconditional coverage for a regression coefficient chosen by *any* model selection technique (e.g., Berk et al., 2013; Kuchibhotla et al., 2019; Bachoc et al., 2020). That is, for any $\widehat{M} \subset \{1,...,p\}$, we have at our disposal a level $1-\alpha$ PoSI CI $CI^{P,\alpha}_{\widehat{M}}$ such that

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

for all $\widehat{M} \subset \{1,...,p\}$ (or some other relevant subset of the universe of models).

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

$$(\widehat{\theta}_M, \widehat{\beta}_M) = \text{argmin}_{\theta \in \mathbb{R}, b_M \in \mathbb{R}^{|M|}} \|y - z\theta - X(M)b_M\|^2.$$

This truncation interval depends upon a sufficient statistic $Z_{\widehat{M}}$ for the unknown nuisance

parameter $\beta_{\widehat{M}}$ that is independent of $\widehat{\theta}_{\widehat{M}}$ after conditioning on the realization of $\widehat{M}$, i.e., $\{\widehat{M}=M\}=\{\widehat{\theta}_M\in[\mathcal{V}_M^-(Z_M),\mathcal{V}_M^+(Z_M)]\}$.[1] The typical construction of a selective CI then proceeds by invoking the fact that $\widehat{\theta}_{\widehat{M}}|\widehat{M}=M$ is distributed according to a normal distribution with mean θ_M truncated to the interval $[\mathcal{V}_M^-(Z_M),\mathcal{V}_M^+(Z_M)]$, and collecting all null hypothesized values of θ_M 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, PoSI CIs typically take the form

$$CI_{\widehat{M}}^{P,\alpha}=\widehat{\theta}_{\widehat{M}}\pm\sigma_{\widehat{M}}K_\alpha,$$

where σ_M is the standard deviation of $\widehat{\theta}_M$ and K_α is a constant that guarantees (2) holds.

My proposal in this context is to form a level $1-\alpha$ hybrid CI $CI_{\widehat{M}}^{H,\alpha}$ that is constructed in analogy with the selective CI 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 $\theta_{\widehat{M}}$ lies inside of a level $1-\gamma>1-\alpha$ PoSI CI:

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

where

$$\begin{aligned}\mathcal{V}_M^{-,H}(Z_M,\theta_M)&=\max\left\{\mathcal{V}_M^-(Z_M),\theta_M-\sigma_M K_\gamma\right\},\\\mathcal{V}_M^{+,H}(Z_M,\theta_M)&=\min\left\{\mathcal{V}_M^+(Z_M),\theta_M+\sigma_M K_\gamma\right\}\end{aligned}$$

(using the convention that $[a,b]=\emptyset$ if $b<a$). A hybrid CI is then constructed by invoking the fact that $\widehat{\theta}_{\widehat{M}}|\{\widehat{M}=M\}\cap\{\theta_{\widehat{M}}\in CI_{\widehat{M}}^{P,\gamma}\}$ is distributed according to a normal distribution with mean θ_M truncated to the interval $[\mathcal{V}_M^{-,H}(Z_M,\theta_M),\mathcal{V}_M^{+,H}(Z_M,\theta_M)]$. It is defined as all null hypothesized values of θ_M 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 $(\alpha-\gamma)/(1-\gamma)$.

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 $\mathbb{P}(\theta_{\widehat{M}}\in CI_{\widehat{M}}^{P,\gamma})<1$. To see why this adjusted level yields correct unconditional coverage of the hybrid CI, note that analogous arguments to

[1] There is an additional element of the conditioning set $\mathcal{V}_M^0(Z_M)\geq 0$ that is suppressed in this section for simplicity of exposition.

those used to guarantee correct conditional coverage (1) can be used to show

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

for all $M \subset \{1,...,p\}$ so that

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

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

Note further that, by construction $CI^{H,\alpha}_{\widehat{M}} \subset CI^{P,\gamma}_{\widehat{M}}$, limiting its length by the length of $CI^{P,\gamma}_{\widehat{M}}$. This feature serves to substantially reduce the length of $CI^{H,\alpha}_{\widehat{M}}$ relative to $CI^{S,\alpha}_{\widehat{M}}$ under data realizations that lead to $\widehat{\theta}_{\widehat{M}}$ being close to one of the bounds of the truncation interval $[\mathcal{V}^-_M(Z_M), \mathcal{V}^+_M(Z_M)]$. Conversely, if γ is relatively small and $\widehat{\theta}_{\widehat{M}}$ is not close to one of the bounds of the truncation interval, the bounds of $CI^{H,\alpha}_{\widehat{M}}$ will be close to those of $CI^{S,\alpha}_{\widehat{M}}$, retaining its short length under this more favorable data realization.

3 General Asymptotic Model Selection Framework

As in Markovic et al. (2018), suppose we use a dataset 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 $\mathcal{M} = \{1,...,|\mathcal{M}|\}$. 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, further 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 $A_M D_n(M) \leq a_{M,n}$, where A_M is a fixed matrix and $a_{M,n}$ is a random vector such that for all $\varepsilon > 0$,

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

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 $n \geq 1$.

I am interested in constructing a CI for a scalar parameter that is chosen based upon

the selected model $\widehat{M}_n$. I denote this target parameter as $\mu_{T,n}(\widehat{M}_n)$ and assume that in the absence of data-dependent selection, there is an asymptotically Gaussian statistic $T_n(M)$ centered around $\mu_{T,n}(M)$. I 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 $\mathbb{P}$. The strengthening of Markovic et al. (2018) to full joint convergence of all statistics is used here because I focus on unconditional inferential statements that do not condition on the selected model.

Assumption 2
For $T_n=(T_n(1),...,T_n(|\mathcal{M}|))'$ and $D_n=(D_n(1)',...,D_n(|\mathcal{M}|)')'$ 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 $(T^{*\prime}_{\mathbb{P}},D^{*\prime}_{\mathbb{P}})'\sim\mathcal{N}(0,\Sigma(\mathbb{P}))$ with

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

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

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

$$\Sigma_D^{(M)}=\Sigma_D\left(\sum_{m=1}^{M-1}\dim(D^*(m))+1:\sum_{m=1}^{M}\dim(D^*(m)),\sum_{m=1}^{M-1}\dim(D^*(m))+1:\sum_{m=1}^{M}\dim(D^*(m))\right)$$

is the covariance matrix of $D(M)$ and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a matrix A.

Markovic et al. (2018) detail several examples for which point-wise marginal versions of Assumptions 1 and 2 hold, including inference after LASSO with a fixed or randomized cross-validated penalty parameter and inference after randomized selection procedures. Some of their arguments need to be extended slightly to show the full joint and uniform convergence required by Assumption 2 (rather than point-wise convergence of subvectors of T_n and D_n). Andrews et al. (2020b) also provide examples under which Assumptions 1 and 2 hold.

In order to form asymptotically valid hybrid CIs, I require the use of a uniformly consistent estimator $\widehat{\Sigma}_n$ for the covariance matrix in Assumption 2.

Assumption 3

For all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\Big(\|\widehat{\Sigma}_n - \Sigma(\mathbb{P})\| > \varepsilon \Big) = 0.$$

To begin describing the hybrid CI construciton, it is useful to express the conditioning event in Assumption 1 in terms of a data-dependent interval for the target statistic $T_n(\widehat{M}_n)$. The bounds of this interval are expressed in terms of a directly-computable random vector $Z_{M,n}$ that is asymptotically independent of $T_n(M)$:

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

where $\widehat{\Sigma}^{(M)}_{DT,n} = \widehat{\Sigma}_{DT,n}\Big(\sum_{m=1}^{M-1} \dim(D_n(m)) + 1 : \sum_{m=1}^{M} \dim(D_n(m)), M \Big)$ 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:*

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

where

$$\begin{aligned}
\mathcal{V}^-_{M,n}(z) &= \max_{j:(A_M \widehat{\Sigma}^{(M)}_{DT,n} / \widehat{\Sigma}_{T,n}(M,M))_j < 0} \frac{a_{M,n,j} - (A_M z)_j}{(A_M \widehat{\Sigma}^{(M)}_{DT,n} / \widehat{\Sigma}_{T,n}(M,M))_j} \\
\mathcal{V}^+_{M,n}(z) &= \min_{j:(A_M \widehat{\Sigma}^{(M)}_{DT,n} / \widehat{\Sigma}_{T,n}(M,M))_j > 0} \frac{a_{M,n,j} - (A_M z)_j}{(A_M \widehat{\Sigma}^{(M)}_{DT,n} / \widehat{\Sigma}_{T,n}(M,M))_j} \\
\mathcal{V}^0_{M,n}(z) &= \min_{j:(A_M \widehat{\Sigma}^{(M)}_{DT,n} / \widehat{\Sigma}_{T,n}(M,M))_j = 0} a_{M,n,j} - (A_M z)_j.
\end{aligned}$$

I make one final high-level assumption on the existence of a PoSI CI with correct unconditional uniform asymptotic coverage of the parameter of interest $\mu_{T,n}(\widehat{M}_n)$. Bachoc et al. (2020) provide several examples of confidence intervals that satisfy this assumption in various settings.

Assumption 4

For any $\alpha\in(0,1)$*, we have a CI 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 $\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\Big(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\alpha}_{n,\widehat{M}_n}\Big)\geq 1-\alpha$ *and for any* $\varepsilon>0$

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

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* $\mathbb{P}\in\mathcal{P}_n$*.*

4 Hybrid CIs and Uniform Asymptotic Validity

We are now equipped with the ingredients needed to define the $(1-\alpha)$-level hybrid CI, $CI^{H,\alpha}_{n,\widehat{M}_n}$, for $\mu_{T,n}(\widehat{M}_n)$. This CI is constructed from the distribution function of $T_n(\widehat{M}_n)$ after conditioning on the events $\{\widehat{M}_n=M\}$ and

$$\left\{\mu_{T,n}(\widehat{M}_n)\in CI^{P,\gamma}_{n,\widehat{M}_n}\right\}$$

$$=\left\{\mu_{T,n}(\widehat{M}_n)-\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n)}K_{n,\gamma}\leq T_n(\widehat{M}_n)\leq\mu_{T,n}(\widehat{M}_n)+\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n)}K_{n,\gamma}\right\}$$

(by Assumption 4) 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

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

where

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

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

In turn, $CI^{H,\alpha}_{n,\widehat{M}_n}$ is defined as

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

where $\widehat{\mu}_{T,n}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}(\widehat{M}_n)$ and $\widehat{\mu}_{T,n}^{H,1-\frac{\alpha-\gamma}{2(1-\gamma)}}(\widehat{M}_n)$ are used instead of $\widehat{\mu}_{T,n}^{H,\alpha/2}(\widehat{M}_n)$ and $\widehat{\mu}_{T,n}^{H,1-\alpha/2}(\widehat{M}_n)$ to account for the fact that the conditioning event $\{\mu_{T,n}(\widehat{M}_n) \in CI_{n,\widehat{M}_n}^{P,\gamma}\}$ does not occur with probability one under all sequences of probability measures $\{\mathbb{P}_n\}$. For simplicity, I focus on the two-sided equal-tailed version of the hybrid CI 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 $\widehat{\mu}_{T,n}^{H,\frac{\alpha-\gamma}{2(1-\gamma)}}(\widehat{M}_n)$ and $\widehat{\mu}_{T,n}^{H,1-\frac{\alpha-\gamma}{2(1-\gamma)}}(\widehat{M}_n)$ are replaced by any $\widehat{\mu}_{T,n}^{H,q_1}(\widehat{M}_n)$ and $\widehat{\mu}_{T,n}^{H,1-q_2}(\widehat{M}_n)$ such that $q_1+q_2=(\alpha-\gamma)/(1-\gamma)$.

I now state a result establishing the asymptotic coverage of $CI_{n,\widehat{M}_n}^{H,\alpha}$ conditional on the realization of the selected model $\widehat{M}_n$ and the possibly false event $\{\mu_{T,n}(\widehat{M}_n) \in CI_{n,\widehat{M}_n}^{P,\gamma}\}$.

Proposition 1

Under Assumptions 1–4,

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

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

Using the results from Proposition 1, we can show that $CI_{n,\widehat{M}_n}^{H,\alpha}$ has 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,

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

and

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

5 Application to Inference After LASSO Model Selection

I now specialize the general framework to the problem of constructing a hybrid CI for a regression coefficient of interest after using LASSO to determine which covariates enter the regression model. Formally, suppose we have data $(z,X,y) \in \mathbb{R}^n \times \mathbb{R}^{n\times p} \times \mathbb{R}^n$. In order to satisfy Assumptions 1–4, I assume that the rows of y and z are random variables and

the rows of X are either random vectors or have entries equal to one (corresponding to an intercept term). This is usually the case in practice and is sufficient for invoking laws of large numbers and central limit theorems for identically distributed data without specifying a "true" model (see e.g., Kuchibhotla et al., 2018 and references therein for details). 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

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

with $y^* = (I - P_z)y$, $X^* = (I - P_z)X$ 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 corresonding to the selected controls s_E. In other words, a given model M is defined as a tuple (E, s_E) (see Lee et al., 2016 and Markovic et al., 2018).

The analysis in Lee et al. (2016) and Markovic et al. (2018) shows that $\widehat{M}_n = (\widehat{E}_n, sign(\widehat{\beta}_E)) = (E, s_E) = M$ if and only if $A_M D_n(M) \leq a_{M,n}$, where

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

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

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

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 $\tilde{X}'_{E,i}$ denote the i^{th} row of the submatrix of $\tilde{X}$ composed of the columns of $\tilde{X}$ corresponding to E and $\tilde{X}'_{-E,i}$ denote the i^{th} row of the

submatrix of $\tilde{X}$ composed of the remaining columns. Assumption 1 thus holds with

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

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

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

$$\mu_{T,n}(\mathbb{P}, M) = \sqrt{n} e_1' (\mathbb{E}_{\mathbb{P}}[W_E' W_E])^{-1} \mathbb{E}_{\mathbb{P}}[W_E' y]$$

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. 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 $(T_n', D_n')'$. Assumption 3 holds under standard moment, stationarity and dependence conditions on $\mathcal{P}_n$ when using a standard uniformly consistent covariance matrix estimator. See Kuchibhotla et al. (2018) and references therein for details about consistent covariance matrix estimation under minimal assumptions when regressors are random without assuming correct specification of the linear regression model.

Finally, Assumption 4 holds by the results of Bachoc et al. (2020) when using one of the PoSI CIs discussed in that paper. In particular, we may use the following inputs when constructing $CI^{P,\gamma}_{n,\widehat{M}_n}$: $\widehat{\Sigma}_{T,N}$ equal to the corresponding submatrix of the consistent

covariance matrix estimator $\widehat{\Sigma}_n$ and $K_{n,\alpha}$ equal to the $(1-\alpha)$-quantile of

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

with $\Omega = corr(\widehat{\Sigma}_{T,n}) \equiv diag(\widehat{\Sigma}_{T,n})^{\dagger/2} \widehat{\Sigma}_{T,n} diag(\widehat{\Sigma}_{T,n})^{\dagger/2}$, 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,

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

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 $\Omega = corr(\Sigma_T(\mathbb{P}))$. For $\alpha \neq 1$, $0 \leq K_\alpha(\mathbb{P}) \leq \bar{\lambda}$ for some finite $\bar{\lambda}$ and any probability measure $\mathbb{P}$. Finally, Theorem 2.3 of Bachoc et al. (2020) provides sufficient conditions on $\mathcal{P}_n$ that imply $\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\mu_{T,n}(\widehat{M}_n;\mathbb{P}) \in CI^{P,\alpha}_{n,\widehat{M}_n}\right) \geq 1-\alpha$.

6 Finite Sample Properties for Calibrated Experiments

In order to investigate the finite-sample properties of hybrid CIs and compare them to existing selective and PoSI CIs in an empirically-relevant setting, I examine Monte Carlo experiments calibrated to the diabetes dataset from Efron et al. (2004). This dataset was also examined by Lee et al. (2016) in their empirical application of selective inference after using LASSO as a model selection device. I simulate datasets by resampling the original observations of the diabetes dataset with replacement (i.e., by drawing datasets via nonparametric bootstrap), allowing for significant departures from Gaussianity and allowing the simulation results to reflect the estimation error in $\widehat{\Sigma}_n$.

In each simulated dataset, I perform the LASSO model selection exercise described in Section 5 for the response of interest y being equal to a quantitative measure of disease progression one year after baseline, several different predictors of interest z and several different values of the LASSO penalty parameter λ. Here, I report results for four different predictors of interest z: body mass index (BMI), blood pressure (BP), the third blood serum measurement in the dataset (S3) and the fifth blood serum measurement in the dataset (S5). These are the same four variables selected by the LASSO empirical exercise in Lee et al. (2016). I also report results for a penalty parameter λ roughly corresponding to a "high" degree of penalization $\lambda = 200$ and a "low" degree of penalization $\lambda = 50$. To save on space, I do not report results for other predictors of interest and penalty parameters since

the results are qualitatively quite similar. All results are based upon 1,000 simulation draws.

To begin, Table 1 displays the simulated (unconditional) coverage probabilities of the hybrid, selective and PoSI 95% CIs for the four predictors of interest and two penalty parameters. Throughout the Monte Carlo experiments in the construction of the hybrid CI, γ is set equal to the recommended level of 0.005, or $\alpha/10$. It is clear that all three CIs have essentially correct finite-sample coverage across all predictor and penalty parameter values, in line with asymptotic theory. The hybrid CIs tend to slightly over-cover while the PoSI CIs tend to significantly over-cover.

Next, Figures 1 and 2 plot the ratios of the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles of the lengths of the 95% hybrid and selective CIs relative to the corresponding length quantiles of the 95% PoSI CI across simulation draws for $\lambda = 200$ (Figure 1) and $\lambda = 50$ (Figure 2) and the four predictors of interest. The ratios of the length quantiles of the PoSI CI relative to itself, which is always equal to one, is also included in the figures in order to detect when the other CIs' length quantiles are shorter than those of the PoSI CI. From Figure 1 we can see that for a "high" degree of penalization, the length quantiles of the hybrid CI are shorter than those of the PoSI CI for all but one quantile-level/predictor of interest combination. Indeed, the length reduction relative to PoSI can be quite significant at the median level, where we can see length reductions larger than one third. From Figure 1, we can also see that although the selective CIs have the potential to be quite a bit shorter than the PoSI CIs, they can also become nearly five times as long at higher quantile levels. On the other hand, the hybrid CIs retain almost all of the length gains of the selective CIs at lower quantile levels without the length losses at higher quantile levels. This is the direct result of hybridization, borrwoing the strengths of both the selective and PoSI CIs for different realizations of the data.

From Figure 2, we can see that under "low" levels of penalization, the hybrid and PoSI CIs perform somewhat similarly in terms of length. However, the hybrid CI is still capable of fairly substantial reductions in length relative to the PoSI CI while not showing very large increases in length at any quantile-level/predictor of interest combination. Conversely, the selective CIs can perform quite poorly relative to both the hybrid and PoSI CIs under this level of penalizaiton, reaching length increases upwards of sixteen-fold. And we can again see that the hybrid CIs tend to perform nearly as well as the selective CIs when the latter are performing well, without the cost of very large lengths at higher quantile levels.

Finally, as an alternative measure of length performance, Table 2 reports the empirical probabilities that the 95% hybrid and selective CIs are longer than the 95% PoSI CI across

simulation draws for $\lambda = 200$ and $\lambda = 50$ and the four predictors of interest. For the larger penalty parameter, the hybrid CI is very likely and the selective CI is somewhat likely to be shorter than the PoSI CI. For the smaller penalty parameter, both the hybrid and selective CIs are likely to be longer than the PoSI CI except in the case for which the predictor of interest is S3. In all cases, the hybrid CI is substantially less likely to be longer than the PoSI CI than is the selective CI.

7 Conclusions and Extensions

In this paper, I introduce an alternative CI for inference after model selection to those that currently dominate the literature. By relaxing the coverage requirement of selective CIs, these hybrid CIs are able to borrow upon the relative strengths of selective and PoSI CIs to yield CIs with desirable length properties across a wide variety of data realizations.

Two questions that I did not address in this paper but may be worth investigation in follow-up research are whether PoSI CIs that do not satisfy the structure imposed by Assumption 4 can be used as an ingredient in the construction of hybrid CIs and whether hybrid CIs 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 PoSI CIs that are either shorter and/or easier to compute in the presence of many models under consideration (see e.g., Kuchibhotla et al., 2019). For the second question, results in Tibshirani et al. (2018) suggest that uniform asymptotic coverage of hybrid CIs 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.

8 Technical Appendix

Proof of Lemma 1: By Assumption 1,

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

The statement of the lemma immediately follows. ■

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

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

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. (2020b),

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

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

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

Further, Lemma 2 implies

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

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

$$
\begin{aligned}
&F_{TN}\Big(T^*_n(\widehat{M}_n);0,\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n),\max\Big\{\mathcal{V}^{-}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}),-\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n)}K_{n,\gamma}\Big\},\\
&\qquad\min\Big\{\mathcal{V}^{+}_{\widehat{M}_n,n}(Z^*_{\widehat{M}_n,n}),\sqrt{\widehat{\Sigma}_{T,n}(\widehat{M}_n,\widehat{M}_n)}K_{n,\gamma}\Big\}\Big)\geq 1-\frac{\alpha-\gamma}{2(1-\gamma)}.
\end{aligned}
\tag{5}
$$

By a slight extension of Lemma 5 of Andrews et al. (2020b), 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. $\Sigma(\mathbb{P}_{n_s})\to\Sigma^*\in\mathcal{S}$

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

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. $\mathbb{P}_{n_s}\Big(\widehat{M}_{n_s}=M,\mu_{T,n_s}(\widehat{M}_{n_s})\in CI^{P,\gamma}_{n_s,\widehat{M}_{n_s}}\Big)\to p^*\in(0,1]$, 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

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

Let $\{\mathbb{P}_{n_s}\}$ be a sequence satisfying conditions 1.–5. Now under $\{\mathbb{P}_{n_s}\}$, $(T^*_{n_s},\widehat{\Sigma}_{n_s},K_{n_s,\gamma}) \xrightarrow{d} (T^*,\Sigma^*,K^*_\gamma)$ by Assumptions 2–4, where $T^*_n=(T^*_n(1),...,T^*_n(|\mathcal{M}|)'$ 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 $D^*(M)\sim\mathcal{N}(0,\Sigma^{(M)*}_D)$. This convergence occurs jointly with that for $(T^*_{n_s},\widehat{\Sigma}_{n_s},K_{n_s,\gamma})$. 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, nearly identical arguments to those used in the proof of Lemma 8 in Andrews et al. (2020b) 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 $\widehat{\Sigma}_n$ and $a_{M,n}$ with Σ^* and a^*_M and Z^*_M is defined identically to $Z^*_{M,n}$ after replacing $\widehat{\Sigma}_n$, $D_n(M)$ and $T^*_n(M)$ with Σ^*, $D^*(M)+\mu^*_D(M)$ 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

$$(T^*_{n_s},\widehat{\Sigma}_{n_s},K_{n_s,\gamma},\widehat{M}_{n_s},\mathcal{V}^-_{M,n_s}(Z^*_{M,n_s}),\mathcal{V}^+_{M,n_s}(Z^*_{M,n_s}))\xrightarrow{d}(T^*,\Sigma^*,K^*_\gamma,\widehat{M},\mathcal{V}^-_M(Z^*_M),\mathcal{V}^+_M(Z^*_M)) \tag{6}$$

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

By Lemma 9 of Andrews et al. (2020b), $F_{TN}(t;\mu,\Sigma_T(M,M),\mathcal{L},\mathcal{U})$ 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, (6) implies

$$\begin{aligned}
&\Bigg(F_{TN}(T^*_{n_s}(\widehat{M}_{n_s});0,\widehat{\Sigma}_{T,n_s}(\widehat{M}_{n_s},\widehat{M}_{n_s}),\max\left\{\mathcal{V}^-_{\widehat{M}_{n_s},n_s}(Z^*_{\widehat{M}_{n_s},n_s}),-\sqrt{\widehat{\Sigma}_{T,n_s}(\widehat{M}_{n_s},\widehat{M}_{n_s})}K_{n_s,\gamma}\right\},\\
&\quad\min\left\{\mathcal{V}^+_{\widehat{M}_{n_s},n_s}(Z^*_{\widehat{M}_{n_s},n_s}),\sqrt{\widehat{\Sigma}_{T,n_s}(\widehat{M}_{n_s},\widehat{M}_{n_s})}K_{n_s,\gamma}\right\},\mathbf{1}(\widehat{M}_{n_s}=M,\mu_{T,n_s}(\widehat{M}_{n_s})\in CI^{P,\gamma}_{n_s,\widehat{M}_{n_s}})\Bigg)\\
&\xrightarrow{d}\Bigg(F_{TN}(T^*(\widehat{M});0,\Sigma^*_T(\widehat{M},\widehat{M}),\max\left\{\mathcal{V}^-_{\widehat{M}}(Z^*_{\widehat{M}}),-\sqrt{\Sigma^*_T(\widehat{M},\widehat{M})}K^*_\gamma\right\},\\
&\quad\min\left\{\mathcal{V}^+_{\widehat{M}}(Z^*_{\widehat{M}}),\sqrt{\Sigma^*_T(\widehat{M},\widehat{M})}K^*_\gamma\right\},\mathbf{1}\left(\widehat{M}=M,-\sqrt{\Sigma^*_T(\widehat{M},\widehat{M})}K^*_\gamma\leq T^*(\widehat{M})\leq\sqrt{\Sigma^*_T(\widehat{M},\widehat{M})}K^*_\gamma\right)\Bigg),
\end{aligned} \tag{7}$$

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

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

Given the equivalence in (5), Lemma 1 and (7), the result of the proposition follows from the same arguments used to prove the first part of Corollary 2 of Andrews et al. (2020b). ■

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

$$
\begin{aligned}
&\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\Big(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{H,\alpha}_{n,\widehat{M}_n}\Big)\\
&\geq\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\Big(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{H,\alpha}_{n,\widehat{M}_n}\Big|\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\gamma}_{n,\widehat{M}_n}\Big)\mathbb{P}\Big(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\gamma}_{n,\widehat{M}_n}\Big)\\
&=\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\sum_{M\in\mathcal{M}}\Big\{\mathbb{P}\Big(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{H,\alpha}_{n,\widehat{M}_n}\Big|\widehat{M}_n=M,\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\gamma}_{n,\widehat{M}_n}\Big)\\
&\qquad\qquad\times\mathbb{P}\Big(\widehat{M}_n=M,\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\gamma}_{n,\widehat{M}_n}\Big)\Big\}\\
&\geq\frac{1-\alpha}{1-\gamma}\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\sum_{M\in\mathcal{M}}\mathbb{P}\Big(\widehat{M}_n=M,\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\gamma}_{n,\widehat{M}_n}\Big)\\
&=\frac{1-\alpha}{1-\gamma}\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\Big(\mu_{T,n}(\widehat{M}_n;\mathbb{P})\in CI^{P,\gamma}_{n,\widehat{M}_n}\Big)\geq\frac{1-\alpha}{1-\gamma}(1-\gamma)=1-\alpha,
\end{aligned}
$$

where the second inequality follows from Lemma 6 of Andrews et al. (2020b) 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. (2020b). ■

References

Andrews, I., Kitagawa, T., and McCloskey, A. (2020a). Inference after estimation of breaks. Forthcoming in *Journal of Econometrics*.

Andrews, I., Kitagawa, T., and McCloskey, A. (2020b). 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.

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. Working Paper.

Kivaranovic, D. and Leeb, H. (2020). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. Forthcoming in *Journal of the American Statistical Association*.

Kuchibhotla, A. K., Brown, L. D., and Buja, A. (2018). Model-free study of ordinary least squares linear regression. Working Paper.

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., and Zhao, L. H. (2019). Valid post-selection inference in model-free linear regression. Forthcoming in *Annals of Statistics*.

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. Working Paper.

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.

Table 1: Unconditional Coverage Probabilities

	Predictor of Interest			
CI	BMI	BP	S3	S5
	$\lambda = 200$			
Hybrid	0.963	0.952	0.954	0.959
Selective	0.956	0.951	0.948	0.959
PoSI	0.989	0.989	1.000	1.000
	$\lambda = 50$			
Hybrid	0.953	0.957	0.968	0.979
Selective	0.949	0.954	0.966	0.978
PoSI	0.992	0.995	0.997	0.997

This table reports the unconditional coverage probability of hybrid, selective and PoSI CIs across nonparametric bootstrap replications, all evaluated at the nominal coverage level of 95%. The coverage probabilities are reported for the four predictors of interest z being equal to BMI, BP, S3 and S5 after selecting control variables via LASSO with a "high" degree of penalization $\lambda = 200$ and a "low" degree of penalization $\lambda = 50$.

Table 2: Probabilities of Being Longer than PoSI CI

	Predictor of Interest			
CI	BMI	BP	S3	S5
	$\lambda = 200$			
Hybrid	0.248	0.031	0.002	0.069
Selective	0.472	0.261	0.166	0.322
	$\lambda = 50$			
Hybrid	0.760	0.584	0.148	0.809
Selective	0.943	0.864	0.450	0.955

This table reports the frequency with which the 95% hybrid and selective CIs are longer than the 95% PoSI CI across nonparametric bootstrap replications. The probabilities are reported for the four predictors of interest z being equal to BMI, BP, S3 and S5 after selecting control variables via LASSO with a "high" degree of penalization $\lambda = 200$ and a "low" degree of penalization $\lambda = 50$.

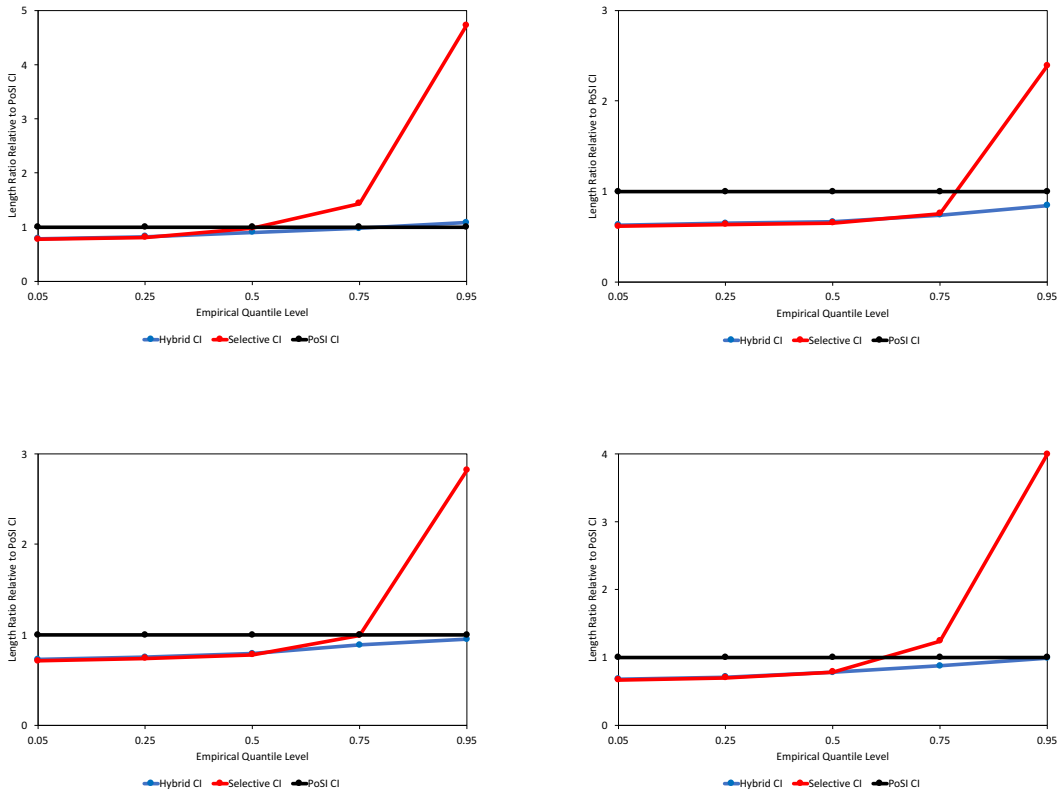

Figure 1: This figure plots the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles of the lengths of the 95% hybrid (blue), selective (red) and PoSI (black) CIs divided by the corresponding length quantiles of the 95% PoSI CI across nonparametric bootstrap replications for inference after controls variables X are selected via LASSO with a "high" degree of penalization $\lambda = 200$. The **upper-left** plot corresponds to the predictor of interest z being equal to BMI. The **upper-right** plot corresponds to the predictor of interest z being equal to S3. The **lower-left** plot corresponds to the predictor of interest z being equal to BP. The **lower-right** plot corresponds to the predictor of interest z being equal to S5.

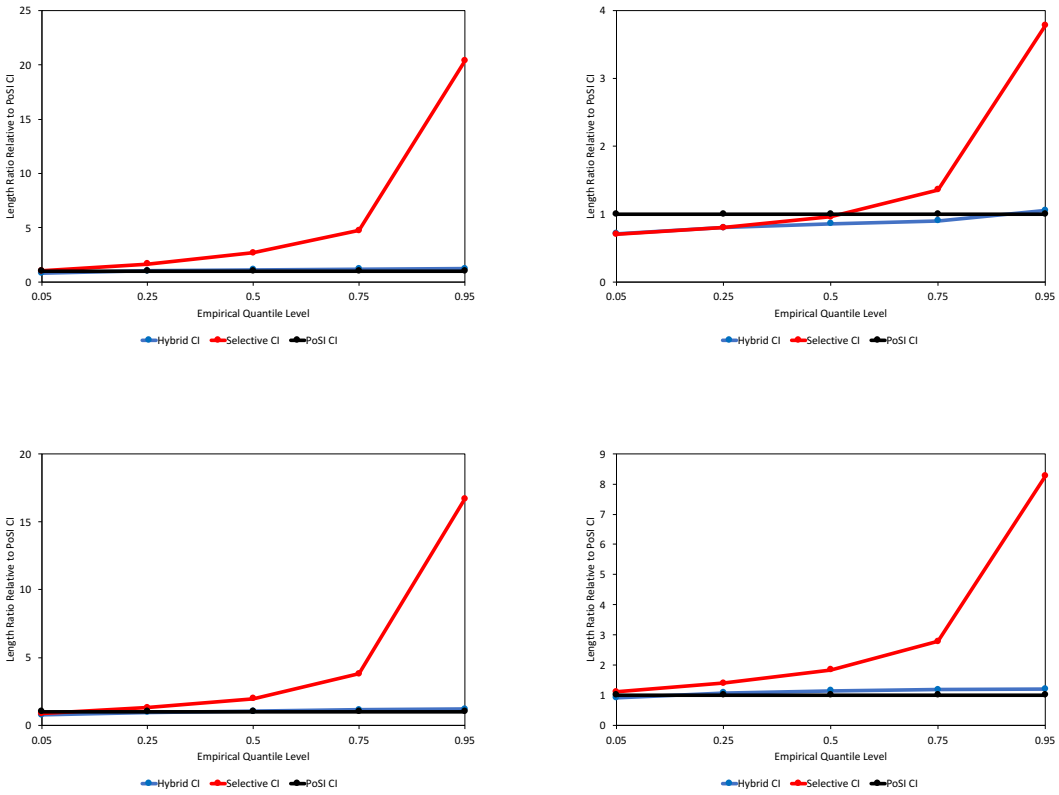

Figure 2: This figure plots the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} empirical quantiles of the lengths of the 95% hybrid (blue), selective (red) and PoSI (black) CIs divided by the corresponding length quantiles of the 95% PoSI CI across nonparametric bootstrap replications for inference after controls variables X are selected via LASSO with a "low" degree of penalization $\lambda = 50$. The **upper-left** plot corresponds to the predictor of interest z being equal to BMI. The **upper-right** plot corresponds to the predictor of interest z being equal to S3. The **lower-left** plot corresponds to the predictor of interest z being equal to BP. The **lower-right** plot corresponds to the predictor of interest z being equal to S5.