Inference on Winners*

Isaiah Andrews† Toru Kitagawa‡ Adam McCloskey§

December 22, 2021

Abstract

Many empirical questions concern target parameters selected through optimization. For example, researchers may be interested in the effectiveness of the best policy found in a randomized trial, or the best-performing investment strategy based on historical data. Such settings give rise to a winner's curse, where conventional estimates are biased and conventional confidence intervals are unreliable. This paper develops optimal confidence intervals and median-unbiased estimators that are valid conditional on the target selected and so overcome this winner's curse. If one requires validity only on average over targets that might have been selected, we develop hybrid procedures that combine conditional and projection confidence intervals to offer further performance gains relative to existing alternatives.

Keywords: Winner's Curse, Selective Inference

JEL Codes: C12, C13

*We thank Tim Armstrong, Stéphane Bonhomme, Raj Chetty, Gregory Cox, Áureo de Paula, Nathaniel Hendren, Patrick Kline, Hannes Leeb, Anna Mikusheva, Magne Mogstad, José Luis Montiel Olea, Mikkel Plagborg-Møller, Jack Porter, Adam Rosen, Frank Schoerfheide, Jesse Shapiro, and participants at numerous seminars and conferences for helpful comments. We also thank Raj Chetty and Nathaniel Hendren for extremely generous assistance on the application using data from Chetty et al. (2018), and thank Jeff Rowley, Peter Ruhm, and Nicolaj Thor for outstanding research assistance. Andrews gratefully acknowledges financial support from the NSF under grant number 1654234. Kitagawa gratefully acknowledges financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) and the European Research Council (Starting grant No. 715940). Initial version posted May 10, 2018.

†Department of Economics, Harvard University, iandrews@fas.harvard.edu

‡CeMMAP and Department of Economics, University College London, t.kitagawa@ucl.ac.uk

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

1 Introduction

A wide range of empirical questions involve inference on target parameters selected through optimization over a finite collection of candidates. In a randomized trial considering multiple treatments, for instance, one might want to learn about the true average effect of the treatment that performed best in the experiment. In finance, one might want to learn about the expected return of the trading strategy that performed best in a backtest.

Estimators that do not account for data-driven selection of the target parameter can be badly biased, and conventional t-test-based confidence intervals may severely under-cover. To illustrate the problem, consider inference on the true average effect of the treatment that performed best in a randomized trial.[1] Since it ignores the data-driven selection of the treatment of interest, the conventional estimate for this average effect will be biased upwards. Similarly, the conventional confidence interval will under-cover, particularly when the number of treatments considered is large. This gives rise to a form of winner's curse, where follow-up trials will be systematically disappointing relative to what we would expect based on conventional estimates and confidence intervals. This form of winner's curse has previously been discussed in contexts including genome-wide association studies (e.g. Zhong and Prentice, 2009; Xu et al., 2011; Ferguson et al., 2013) and online A/B tests (Lee and Shen, 2018).

This paper develops estimators and confidence intervals that eliminate the winner's curse. There are two distinct perspectives from which to consider bias and coverage. The first requires validity conditional on the target selected, for example on the identity of the best-performing treatment, while the second is unconditional and requires validity on average over possible target parameters. Conditional validity is more demanding but may be desirable in some settings, for example when one wants to ensure validity conditional on the recommendation made to a policy maker. Both perspectives differ from inference on the effectiveness of the "true" best treatment, as in e.g. Chernozhukov et al. (2013) and Rai (2018), in that we consider inference on the effectiveness of the (observed) best-performing treatment

[1] Such a scenario seems to be empirically relevant, as a number of recently published randomized trials in economics either were designed with the intent of recommending a policy or represent a direct collaboration with a policy maker. For example, Khan et al. (2016) assess how incentives for property tax collectors affect tax revenues in Pakistan, Banerjee et al. (2018) evaluate the efficacy of providing information cards to potential recipients of Indonesia's *Raskin* programme, and Duflo et al. (2018) collaborate with the Gujarat Pollution Control Board (an Indian regulator tasked with monitoring industrial emissions in the state) to evaluate how more frequent but randomized inspection of plants performs relative to discretionary inspection. Baird et al. (2016) find that deworming Kenyan children has substantial beneficial effects on their health and labor market outcomes into adulthood, and Björkman Nyqvist and Jayachandran (2017) find that providing parenting classes to Ugandan mothers has a greater impact on child outcomes than targeting these classes at fathers.

in the sample rather than the (unobserved) best-performing treatment in the population.[2]

For conditional inference, we derive optimal median-unbiased estimators and equal-tailed confidence intervals. We further show that in cases where the winner's curse does not arise (for instance because one treatment considered is vastly better than the others) our conditional procedures coincide with conventional ones. Hence, our corrections do not sacrifice efficiency in such cases.

An alternative approach to conditional inference is sample splitting. In settings with independent observations, choosing the target parameter using one subset of the data and constructing estimates and confidence intervals using the remaining subset ensures unbiasedness of estimates and validity of conventional confidence intervals conditional on the target parameter. The split-sample target parameter is necessarily more variable than the full-data target, however. Moreover, since only a subset of the data is used for inference, split-sample procedures are inefficient within the class of procedures with the same target. In the supplement to this paper we build on our conditional inference results to develop computationally tractable confidence intervals and estimators that dominate conventional sample-splitting.

We next turn to unconditional inference. One approach to constructing valid unconditional confidence intervals is projection, applied in various settings by e.g. Romano and Wolf (2005), Berk et al. (2013), and Kitagawa and Tetenov (2018a). To obtain a projection confidence interval, we form a simultaneous confidence band for all potential targets and take the implied set of values for the target of interest. The resulting confidence intervals have correct unconditional coverage but, unlike our conditional intervals, are wider than conventional confidence intervals even when the latter are valid. On the other hand, we find in simulation that projection intervals outperform conditional intervals in cases where there is substantial randomness in the target parameter, e.g. when there is not a clear best treatment.

Since neither conditional nor projection intervals are uniformly best from an unconditional perspective, we introduce hybrid estimators and confidence intervals that combine conditioning and projection. These maintain most of the good performance of our conditional approach in cases for which the winner's curse does not arise, while improving on conditional procedures in cases where these underperform, e.g. by limiting the maximum length of hybrid intervals relative to projection intervals. In simulations calibrated to our applications we find that hybrid intervals are typically shorter than both conditional and projection intervals, often by a large margin.

[2]See Dawid (1994) for an early discussion of this distinction, and an argument in favor of inference on the best-performing treatment in the sample.

We derive our main results in the context of a finite-sample normal model with an unknown mean vector and a known covariance matrix. This model can be viewed as an asymptotic approximation to many different non-normal finite-sample settings. To formalize this connection, we note, and prove in the appendix, that feasible versions of our procedures, based on non-normal data and plugging in estimated variances, are uniformly asymptotically valid over a large class of data-generating processes.

We illustrate our results with two applications. The first uses data from Karlan and List (2007) to conduct inference on the effect of the best-performing treatment in an experiment studying the impact of matching incentives on charitable giving. Simulations calibrated to these data show that conventional estimates ignoring selection are substantially upward biased, while our corrections reduce bias and increase coverage. Applied to the original Karlan and List (2007) data, our corrections suggest substantially less optimism about the effect of the best-performing treatment than conventional approaches, with point estimates below the lower bound of the conventional confidence intervals.

For our second application, we consider the problem of targeting neighborhoods based on estimated economic mobility. In cooperation with the Seattle and King County public housing authorities, Bergman et al. (2020) conduct an experiment encouraging housing voucher recipients to move to high-opportunity neighborhoods, which are selected based on census-tract level estimates of economic mobility from Chetty et al. (2018). We consider an analogous exercise in the 50 largest commuting zones in the US, selecting top tracts based on estimated economic mobility and examining conventional and corrected inference on the average mobility in selected tracts. Calibrating simulations to the Chetty et al. (2018) data, we find that conventional approaches suffer from severe bias in many commuting zones. These biases are reduced, but not eliminated, by the empirical Bayes corrections used by Chetty et al. (2018) and many others in the applied literature. Intuitively, commonly-applied empirical Bayes approaches correspond to a normal prior on unit-level causal effects conditional on covariates. Bayesian arguments (discussed in Appendix E) imply that these methods correct the winner's curse when the normal prior matches the distribution of true effects, but not in general otherwise. Turning to the original Chetty et al. (2018) data, our corrected estimates imply lower mobility, and higher uncertainty, for selected tracts than conventional approaches, but nonetheless strongly indicate gains from moving to selected tracts. Our confidence intervals likewise suggest substantially higher uncertainty than empirical Bayes credible sets, though we do not find a clear ordering between our bias-corrected point estimates of economic mobility and the empirical Bayes point estimates of Chetty et al. (2018).

The choice between conditional and unconditional inference methods is necessarily context-specific, as it depends on the extent to which we care about validity conditional on selecting a given target. We report results for both conditional and unconditional approaches in each application, but view conditional inference as particularly natural for the first application. In this setting the "winning" treatment is easily interpretable, raising the question of what we can conclude conditional on identity of this treatment. In our second application, by contrast, our primary goal is to assess the efficacy of targeting the top third of census tracts in each commuting zone, with less focus on the precise collection of tracts selected. We therefore view unconditional inference as more natural in this context.

It is important to emphasize that our goal is to evaluate the effectiveness of a recommended policy or treatment, taking the rule for selecting a recommendation as given, rather than to improve the rule. Our procedures thus play a role similar to that of ex-post policy evaluations, with the difference that we can produce estimates without waiting for a policy to be implemented. Like ex-post evaluations, these estimates may be useful for a variety of purposes, including understanding the true effectiveness of a selected policy and forecasting the effects of future implementations. Our results are also useful in settings where ex-post evaluation is possible, since comparison of our estimates with ex-post results can shed light on whether differences between observed performance and conventional ex-ante estimates can be explained solely by the winner's curse.

Related Literature This paper is related to the literature on tests of superior predictive performance (e.g. White (2000); Hansen (2005); Romano and Wolf (2005)). That literature studies the problem of testing whether some strategy or policy beats a benchmark, while we consider the complementary question of inference on the effectiveness of the estimated "best" policy. Our conditional inference results combine naturally with the results of this literature, allowing one to condition inference on e.g. rejecting the null hypothesis that no policy outperforms a benchmark.

Our results build upon, and contribute to, the rapidly growing literature on selective inference. Fithian et al. (2017) describe a general approach to constructing optimal conditional confidence sets in a wide range of settings, while a rapidly growing literature including e.g. Harris et al. (2016), Lee et al. (2016), Tian and Taylor (2018), and our own follow-up work in Andrews et al. (2020b), works out the details of this approach in particular settings. More specifically, this literature primarily focuses on inference after regressor selection in the linear regression model using various model selection criteria, while Andrews et al. (2020b) focuses on inference after estimating a break location in a break or threshold

regression model. Like this literature, our analysis of conditional confidence intervals examines the implications of the conditional approach in our setting. Our results are also related to the growing literature on unconditional post-selection inference, including Berk et al. (2013), Bachoc et al. (2020), and Kuchibhotla et al. (2020). This literature considers analogs of our projection confidence intervals for inference following model selection (see also Laber and Murphy, 2011). Recent work by Guo and He (2021) proposes tightening projection confidence intervals in the context of a winner's curse for selected subgroups via a sequence of tuning parameters that drifts as the sample size grows.

Beyond the new setting considered, we make three main theoretical contributions relative to the selective and post-selection inference literatures. First, when one only requires unconditional validity, we introduce the class of hybrid inference and estimation procedures. We find that hybrid procedures offer large gains in unconditional performance relative both to conditional procedures and to existing unconditional alternatives.[3] Two of our own follow-up papers, Andrews et al. (2020b) and McCloskey (2020), adapt the hybrid procedures we introduce here to different settings, relying on the theoretical results established in this paper. Second, for settings where conditional inference is desired, we observe that the same structure used in the literature to develop optimal conditional confidence intervals also allows construction of optimal quantile unbiased estimators, using results from Pfanzagl (1994) on optimal estimation in exponential families.[4] Third, our uniform asymptotic results are the first of their kind in the conditional inference literature.[5]

Finally, there is a distinct but complementary literature that studies inference on ranks based on some measure of interest. For example, this literature allows one to form valid confidence intervals for the identification of best performing unit, rather than for the performance of the unit selected as best by the data. Conventional inference procedures for these problems fail for similar reasons that give rise to a winner's curse. Recent work by Mogstad et al. (2020) overcomes this inference failure and studies, among other settings, inference on ranks in neighborhood targeting, as in our second application.

In the next section, we begin by introducing the problem we consider and the techniques

[3] A related hybridization, combining conditional and unconditional inference, is used in Andrews et al. (2019) to improve power for tests of parameters identified by moment inequalities.

[4] Eliasz (2004) previously used results from Pfanzagl (1994) to study quantile-unbiased estimation in a different setting, targeting coefficients on highly persistent regressors.

[5] McCloskey (2020) uses our uniformity results to establish uniform asymptotic validity for hybrid confidence intervals for inference after model selection, while Tibshirani et al. (2018) and Andrews et al. (2020b) establish uniform asymptotic validity for conditional confidence intervals in different settings from ours, but only under particular local sequences.

we propose in the context of a stylized example. Section 3 introduces the normal model, develops our conditional procedures, and briefly discusses sample splitting. Section 4 introduces projection confidence intervals and our hybrid procedures. Section 5 discusses practical implementation in and translates our normal model results to uniform asymptotic results. Finally, Sections 6 and 7 discuss applications to data from Karlan and List (2007) and Bergman et al. (2020), respectively. The supplement to this paper contains proofs of our theoretical results and additional theoretical, numerical and empirical results.

2 A Stylized Example

We begin by illustrating the problem we consider, along with the solutions we propose, in a stylized example. Suppose we have data from a randomized trial of a binary treatment (e.g. participation in a job training program), where individuals $i \in \{1,...,n\}$ were randomly assigned to treatment ($D_i = 1$) or control ($D_i = 0$), with $\frac{n}{2}$ individuals in each group. We are interested in an outcome Y_i (e.g. a dummy for employment in the next year), and compute the treatment and control means,

$$(X_n^*(1), X_n^*(0)) = \left(\frac{2}{n} \sum_{i=1}^{n} D_i Y_i, \frac{2}{n} \sum_{i=1}^{n} (1 - D_i) Y_i \right).$$

If trial particpants are a random sample from some population, then for $Y_{i,1}$ and $Y_{i,0}$ equal to the potential outcomes for i under treatment and control, respectively, $(X_n^*(1), X_n^*(0))$ unbiasedly estimate the average potential outcomes $(\mu^*(1), \mu^*(0)) = (E[Y_{i,1}], E[Y_{i,0}])$ in the population.

For policymakers and researchers interested in maximizing the average outcome, it is natural to focus on the treatment that performed best in the experiment. Formally, let $\Theta = \{0,1\}$ denote the set of policies (just control and treatment in this example) and define $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} X_n^*(\theta)$ as the policy yielding the highest average outcome in the experiment. While $X_n^*(\theta)$ unbiasedly estimates $\mu^*(\theta)$ for fixed policies θ, $X_n^*(\hat{\theta}_n)$ systematically over-estimates $\mu^*(\hat{\theta}_n)$ since we are more likely to select a given policy when the experiment over-estimates its effectiveness. Likewise, confidence intervals for $\mu^*(\hat{\theta}_n)$ that ignore estimation of θ may cover $\mu^*(\hat{\theta}_n)$ less often than we intend. Hence, if a policymaker deploys the treatment $\hat{\theta}_n$, or a researcher examines it in a follow-up experiment, the results will be systematically disappointing relative to the original trial. This is a form of winner's curse: estimation error leads us to over-predict the benefits of our chosen policy and to misstate our uncertainty about its effectiveness.

To simplify the analysis and develop corrected inference procedures, we turn to asymptotic approximations. For $X_n = \sqrt{n}X_n^* = \sqrt{n}(X_n^*(0),X_n^*(1))$ and $(\mu_n(1),\mu_n(0)) = \sqrt{n}(\mu^*(1),\mu^*(0))$, provided the potential outcomes $(Y_{i,0},Y_{i,1})$ have finite variance,

$$\begin{pmatrix} X_n(0)-\mu_n(0) \\ X_n(1)-\mu_n(1) \end{pmatrix} \Rightarrow N\left(0,\begin{pmatrix} \Sigma(0) & 0 \\ 0 & \Sigma(1) \end{pmatrix}\right), \tag{1}$$

where $\Rightarrow$ denotes convergence in distribution and the asymptotic variance Σ can be consistently estimated while the scaled average outcomes μ_n cannot be. Motivated by (1), let us abstract from approximation error and assume that we observe

$$\begin{pmatrix} X(0) \\ X(1) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(0) \\ \mu(1) \end{pmatrix},\begin{pmatrix} \Sigma(0) & 0 \\ 0 & \Sigma(1) \end{pmatrix}\right)$$

for $\Sigma(0)$ and $\Sigma(1)$ known, and that $\hat{\theta} = \text{argmax}_{\theta\in\Theta} X(\theta)$ with $\Theta = \{0,1\}$.[6]

As discussed above, $X(\hat{\theta})$ is biased upwards as an estimator of $\mu(\hat{\theta})$. This bias arises both conditional on $\hat{\theta}$ and unconditionally. To see this note that $\hat{\theta} = 1$ if $X(1) > X(0)$, where ties occur with probability zero. Conditional on $\hat{\theta} = 1$ and $X(0) = x(0)$, however, $X(1)$ follows a normal distribution truncated below at $x(0)$. Since this holds for all $x(0)$, $X(1)$ has positive median bias conditional on $\hat{\theta} = 1$:[7]

$$Pr_\mu\Big\{X(\hat{\theta}) \geq \mu(\hat{\theta}) | \hat{\theta} = 1\Big\} > \frac{1}{2} \text{ for all } \mu.$$

Since the same argument holds for $\hat{\theta} = 0$, $\hat{\theta}$ is also biased upwards unconditionally:

$$Pr_\mu\Big\{X(\hat{\theta}) \geq \mu(\hat{\theta})\Big\} > \frac{1}{2} \text{ for all } \mu.$$

Similarly, conventional t-statistic-based confidence intervals need not have correct coverage.

To illustrate these issues, Figure 1 plots the coverage of conventional confidence intervals, as well as the median bias of conventional estimates, in an example with $\Sigma(1) = \Sigma(0) = 1$. For comparison we also consider cases with ten and fifty policies (e.g. additional treatments) $|\Theta| = 10$ and $|\Theta| = 50$, where we again set Σ to be diagonal with $\Sigma(\theta) = 1$ for all θ and, for

[6]Finite-sample results in this normal model correspond to asymptotic results for cases where the difference in outcomes $E[Y_{i,1}] - E[Y_{i,0}]$ is of order $\frac{1}{\sqrt{n}}$, so the optimal policy $\theta^* = \text{argmax}_{\theta\in\Theta}\mu^*(\theta)$ is weakly identified. We defer an in-depth discussion of asymptotics to Section 5 and Appendix D.

[7]It also has positive mean bias, but we focus on median bias for consistency with our later results.

ease of reporting, assume that all the policies other than the first (policy θ_1) are equally effective, with average outcome $\mu(\theta_{-1})$. The first panel of Figure 1 shows that while the conventional confidence interval has reasonable coverage when there are only two policies, its coverage can fall substantially when $|\Theta|=10$ or $|\Theta|=50$. The second panel shows that the median bias of the conventional estimator $\hat{\mu}=X(\hat{\theta})$, measured as the deviation of the exceedance probability $Pr_\mu\{X(\hat{\theta})\geq\mu(\hat{\theta})\}$ from $\frac{1}{2}$, can be quite large. The third panel shows that the same is true when we measure bias as the median of $X(\hat{\theta})-\mu(\hat{\theta})$. In all cases we find that performance is worse when we consider a larger number of policies, as is natural since a larger number of policies allows more scope for selection.

Our results correct these biases. Returning to the case with $\Theta=\{0,1\}$ for simplicity, let $F_{TN}(x(1);\mu(1),x(0))$ denote the truncated normal distribution function for $X(1)$, truncated below at $x(0)$, when the true mean is $\mu(1)$. This function is strictly decreasing in $\mu(1)$, and for $\hat{\mu}_\alpha$ the solution to $F_{TN}(X(1);\hat{\mu}_\alpha,X(0))=1-\alpha$, Proposition 2 below shows that

$$Pr_\mu\Big\{\hat{\mu}_\alpha\geq\mu(\hat{\theta})|\hat{\theta}=1\Big\}=\alpha \quad \text{for all } \mu.$$

Hence, $\hat{\mu}_\alpha$ is α-quantile unbiased for $\mu(\hat{\theta})$ conditional on $\hat{\theta}=1$, and the analogous statement holds conditional on $\hat{\theta}=0$. Indeed, Proposition 2 shows that $\hat{\mu}_\alpha$ is the optimal α-quantile unbiased estimator conditional on $\hat{\theta}$.

Using this result, we can eliminate the biases discussed above. The estimator $\hat{\mu}_{1/2}$ is median unbiased and the equal-tailed confidence interval $CS_{ET}=\left[\hat{\mu}_{\alpha/2},\hat{\mu}_{1-\alpha/2}\right]$ has conditional coverage $1-\alpha$, where we say that a confidence interval CS has conditional coverage $1-\alpha$ if

$$Pr\Big\{\mu(\hat{\theta})\in CS|\hat{\theta}=\tilde{\theta}\Big\}\geq 1-\alpha \text{ for } \tilde{\theta}\in\Theta \text{ and all } \mu. \tag{2}$$

By the law of iterated expectations, CS_{ET} also has unconditional coverage $1-\alpha$:

$$Pr_\mu\Big\{\mu(\hat{\theta})\in CS\Big\}\geq 1-\alpha \text{ for all } \mu. \tag{3}$$

Unconditional coverage is easier to attain, so relaxing the coverage requirement from (2) to (3) may allow shorter confidence intervals in some cases. Conditional and unconditional coverage requirements address different questions, however, and which is more appropriate depends on the problem at hand. For instance, if a researcher recommends the policy $\hat{\theta}$ to a policymaker, it may also be natural to report a confidence interval that is valid conditional on the recommendation, which is precisely the conditional coverage requirement (2).

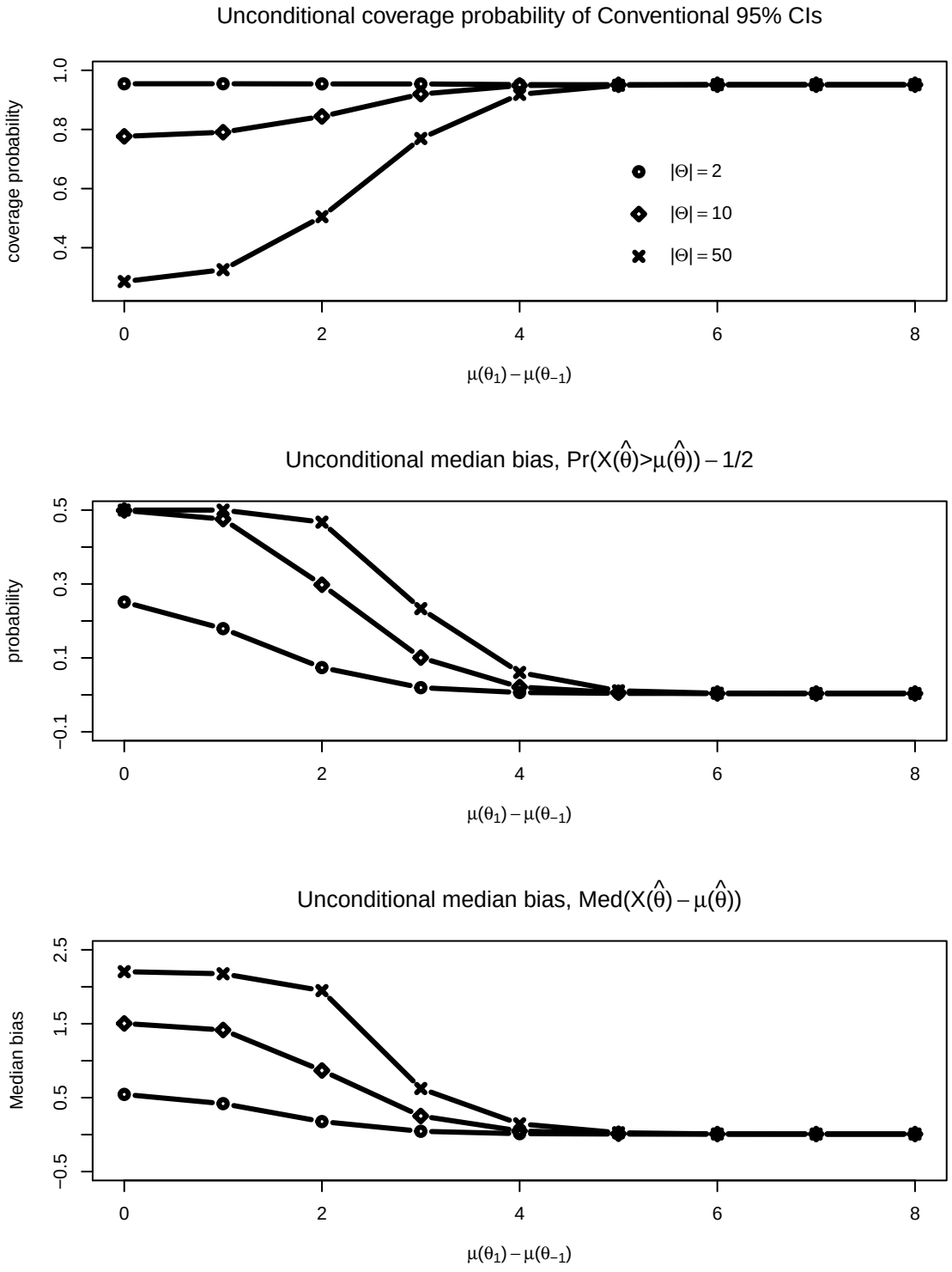

Figure 1: Performance of conventional procedures in examples with 2, 10, and 50 policies.

Conditional coverage ensures that if one considers repeated instances in which researchers recommend a particular course of action (e.g. departure from the status quo), reported confidence intervals will in fact cover the true effects a fraction $1-\alpha$ of the time. On the other hand, if we only want to ensure that our confidence intervals cover the true value with probability at least $1-\alpha$ on average across the distribution of recommendations, it suffices to impose the unconditional coverage requirement (3).

We are unaware of alternatives in the literature that ensure conditional coverage (2). For unconditional coverage (3), however, one can form an unconditional confidence interval by projecting a simultaneous confidence set for μ. In particular, let c_α denote the $1-\alpha$ quantile of $\max_j|\xi_j|$ for $\xi=(\xi_1,\xi_2)'\sim N(0,I_2)$ a two-dimensional standard normal random vector. If we define CS_P as

$$CS_P=\left[X(\hat{\theta})-c_\alpha\sqrt{\Sigma(\hat{\theta})},X(\hat{\theta})+c_\alpha\sqrt{\Sigma(\hat{\theta})}\right],$$

this set has correct unconditional coverage (3).

Figure 2 plots the (unconditional) median length of 95% confidence intervals CS_{ET} and CS_P, along with the conventional confidence interval, again in cases with $|\Theta|\in\{2,10,50\}$. We focus on median length, rather than mean length, because the results for Kivaranovic and Leeb (2020) imply that CS_{ET} has infinite expected length. As Figure 2 illustrates, the median length of CS_{ET} is shorter than the (nonrandom) length of CS_P in all cases when $|\mu(\theta_1)-\mu(\theta_{-1})|$ exceeds four, and converges to the length of the conventional interval as $|\mu(\theta_1)-\mu(\theta_{-1})|$ grows larger. When $|\mu(\theta_1)-\mu(\theta_{-1})|$ is small, on the other hand, CS_{ET} can be substantially wider than CS_P. This reflects that in these cases, $X(\hat{\theta})$ is frequently close to the next-best treatment. For a truncated normal distribution, an observation close to the lower endpoint provides evidence of a small mean, but with little precision about the exact value, leading to long confidence intervals.

These features become still more pronounced as we increase the number of policies considered, and are still more pronounced for higher quantiles of the length distribution. To illustrate, Figure 3 plots the 95th percentile of the distribution of length in the case with $|\Theta|=50$ policies, while results for other quantiles and specifications are reported in Appendix F.

Figure 4 plots the median absolute error $Med_\mu\left(|\hat{\mu}-\mu(\hat{\theta})|\right)$ for different estimators $\hat{\mu}$, and shows that the median-unbiased estimator likewise exhibits larger median absolute error than the conventional estimator $X(\hat{\theta})$ when $|\mu(\theta_1)-\mu(\theta_{-1})|$ is small. This feature is again more pronounced as we increase the number of policies considered, or if we consider

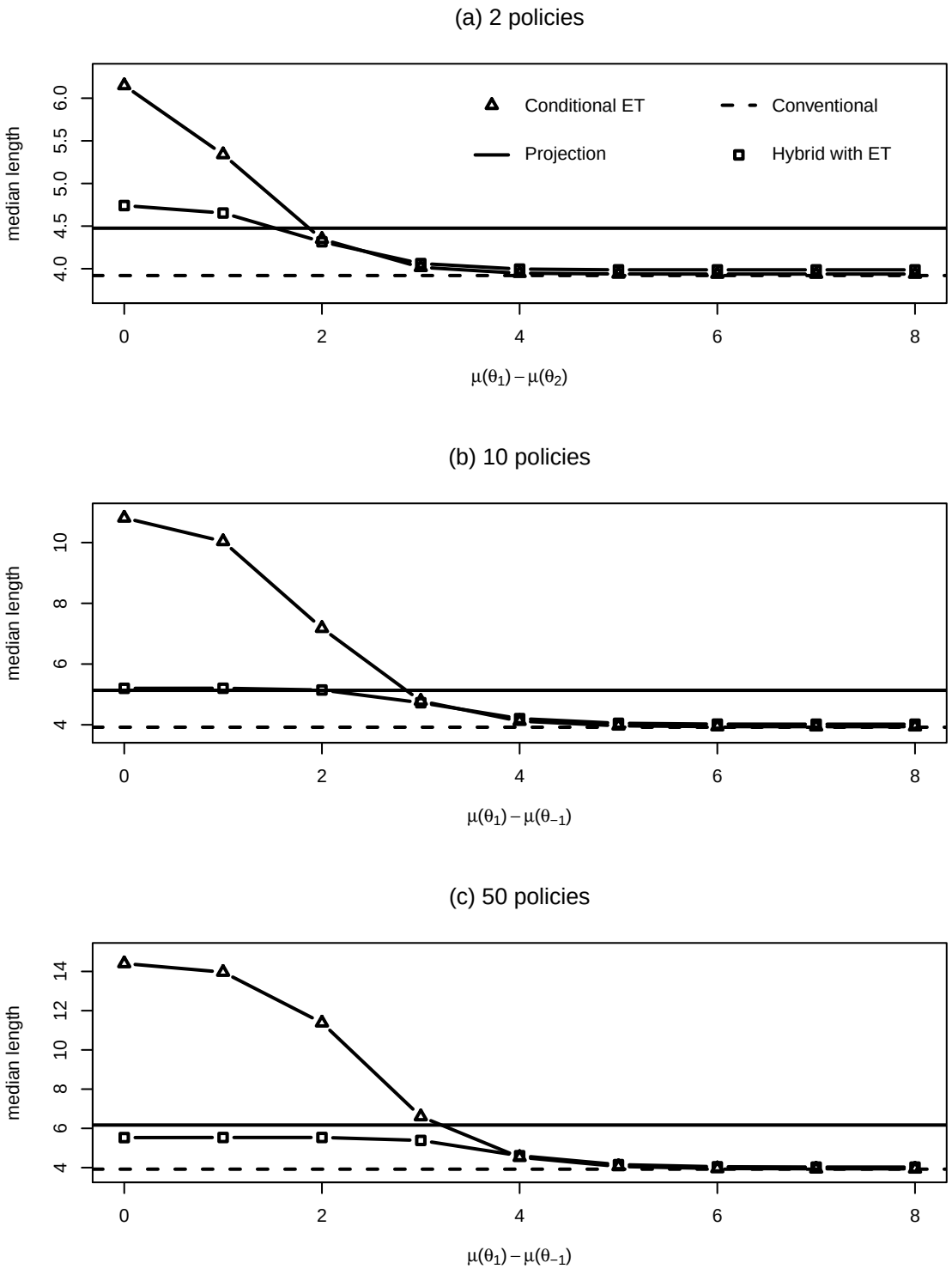

Figure 2: Median length of confidence intervals for $\mu(\hat{\theta})$ in cases with 2, 10, and 50 policies.

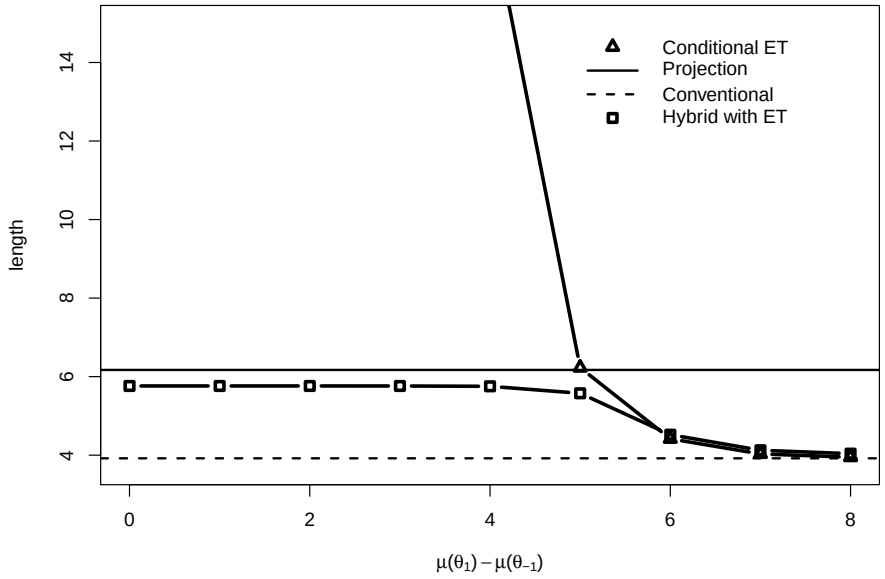

Figure 3: 95th percentile of length of confidence intervals for $\mu(\hat{\theta})$ in case with 50 policies.

higher quantiles as in Appendix F.

Recall that $\hat{\mu}_{\frac{1}{2}}$ and the endpoints of CS_{ET} are optimal quantile unbiased estimators. So long as we impose median unbiasedness and correct conditional coverage, there is hence little scope to improve conditional performance. If we instead focus on unconditional bias and coverage, by contrast, improved performance is possible.

To improve performance, we consider hybrid inference, which combines the conditional and unconditional approaches. Hybrid inference first computes a level $\beta<\alpha$ projection interval CS_P^β, and then considers conditional inference given $\hat{\theta}$ and $\mu(\hat{\theta})\in CS_P^\beta$. In the case with $\Theta=\{0,1\}$, for instance, if $\hat{\theta}=1$ and the true mean is $\mu(1)$ then the conditional distribution of $X(1)$ given $\hat{\theta}=1$, $X(0)=x(0)$, and $\mu(1)\in CS_P^\beta$ is a $N(\mu(1),\Sigma(1))$ distribution truncated to the interval

$$\left[\max\left\{x(0),\mu(1)-c_\beta\sqrt{\Sigma(1)}\right\},\mu(1)+c_\beta\sqrt{\Sigma(1)}\right].$$

For the corresponding distribution function $F_{TN}^H(x(1);\mu(1),x(0))$, the hybrid estimator $\hat{\mu}_\alpha^H$ solves $F_{TN}^H(X(1);\hat{\mu}_\alpha^H,X(0))=1-\alpha$. Arguments analogous to those in the conditional case imply that $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on the (potentially false) event

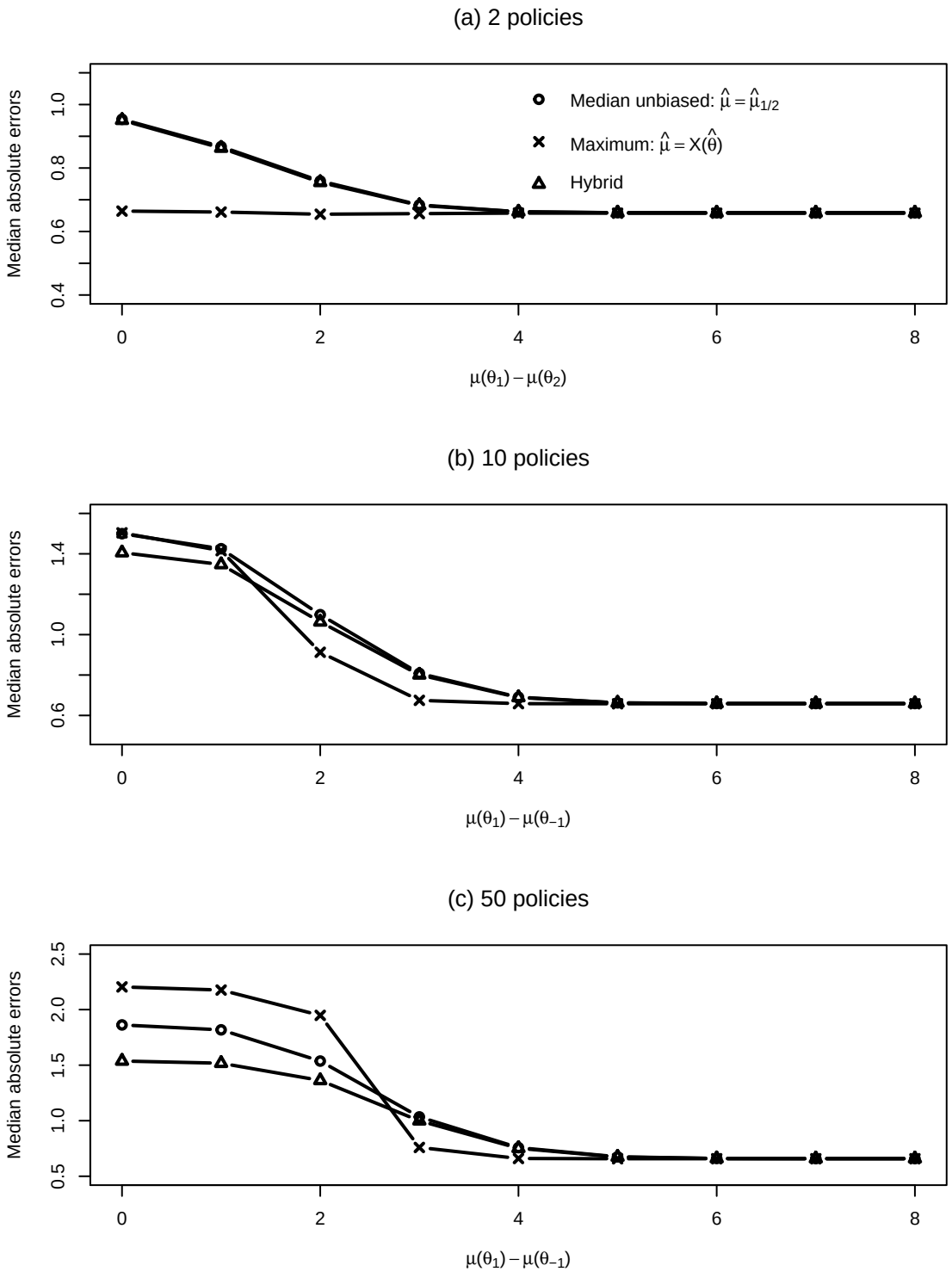

Figure 4: Median absolute error of estimators of $\mu(\hat{\theta})$ in cases with 2, 10, and 50 policies.

$\left\{\hat{\theta}=1,\mu(\hat{\theta})\in CS_P^\beta\right\}$. Since $Pr_\mu\left\{\mu(\hat{\theta})\in CS_P^\beta\right\}\geq 1-\beta$ one can further show that the unconditional quantile bias of $\hat{\mu}_\alpha^H$ is bounded, in the sense that

$$\left|Pr_\mu\left\{\hat{\mu}_\alpha^H\geq\mu(\hat{\theta})\right\}-\alpha\right|\leq\beta\cdot\max\{\alpha,1-\alpha\}.$$

We again form level $1-\alpha$ equal-tailed confidence intervals based on these estimates, where to account for the dependence on the projection interval we adjust the quantile considered and take $CS_{ET}^H=\left[\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)}},\hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)}}\right]$. See Section 4.2 for details on this adjustment. By construction, hybrid intervals are never longer than the level $1-\beta$ projection interval CS_P^β.

Due to their dependence on the projection interval, hybrid intervals do not in general have correct conditional coverage (2). By relaxing the conditional coverage requirement, however, we obtain major improvements in unconditional performance, as illustrated in Figure 2. In particular, we see that in the case with 50 policies, the hybrid confidence intervals have shorter median length than the unconditional interval CS_P for all parameter values considered. The gains relative to conditional confidence intervals are large for many parameter values, and are still more pronounced for higher quantiles of the length distribution, as in Figure 3 and Appendix F. In Figure 4 we report results for the hybrid estimator $\hat{\mu}^H_{\frac{1}{2}}$, and again find substantial performance improvements.

The improved unconditional performance of the hybrid confidence intervals is achieved by requiring only unconditional, rather than conditional, coverage. To illustrate, Figure 5 plots the conditional coverage given $\hat{\theta}=\theta_1$ in the case with two policies. As expected, the conditional interval has correct conditional coverage, while coverage distortions appear for the hybrid and projection intervals when $\mu(\theta_1)\ll\mu(\theta_2)$. In this case $\hat{\theta}=\theta_2$ with high probability but the data will nonetheless sometimes realize $\hat{\theta}=\theta_1$. Conditional on this event, $X(\theta_1)$ will be far away from $\mu(\theta_1)$ with high probability, so projection and hybrid confidence intervals under-cover. As $\mu(\theta_1)-\mu(\theta_2)$ diverges to $-\infty$, their conditional coverage probabilities given $\hat{\theta}=\theta_1$ approach 0.

3 Conditional Inference

This section introduces our general setting, which extends the stylized example of the previous section in several directions, and develops conditional inference procedures. We then discuss sample splitting as an inefficient conditional inference method and briefly discuss the construction of dominating procedures. Finally, we show that our conditional procedures converge to conventional ones when the latter are valid.

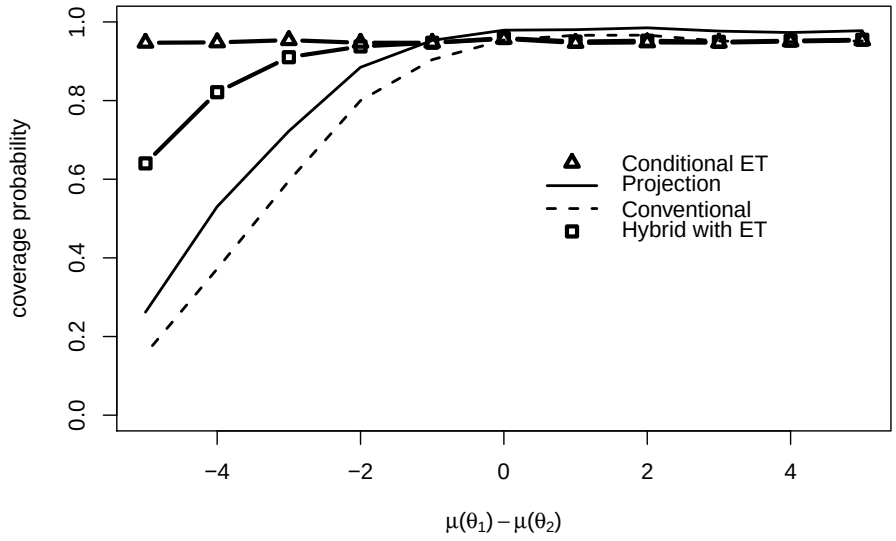

Figure 5: Coverage conditional on $\hat{\theta}=\theta_1$ in case with two policies.

3.1 Setting

Suppose we observe a collection of normal random vectors $(X(\theta),Y(\theta))'\in\mathbb{R}^2$ for $\theta\in\Theta$ where Θ is a finite set. For $\Theta=\{\theta_1,...,\theta_{|\Theta|}\}$, let $X=\big(X(\theta_1),...,X\big(\theta_{|\Theta|}\big)\big)'$ and $Y=\big(Y(\theta_1),...,Y\big(\theta_{|\Theta|}\big)\big)'$. Then

$$\begin{pmatrix} X \\ Y \end{pmatrix}\sim N(\mu,\Sigma) \tag{4}$$

for

$$E\left[\begin{pmatrix} X(\theta) \\ Y(\theta) \end{pmatrix}\right]=\mu(\theta)=\begin{pmatrix} \mu_X(\theta) \\ \mu_Y(\theta) \end{pmatrix},$$

$$\Sigma(\theta,\tilde{\theta})=\begin{pmatrix} \Sigma_X(\theta,\tilde{\theta}) & \Sigma_{XY}(\theta,\tilde{\theta}) \\ \Sigma_{YX}(\theta,\tilde{\theta}) & \Sigma_Y(\theta,\tilde{\theta}) \end{pmatrix}=Cov\left(\begin{pmatrix} X(\theta) \\ Y(\theta) \end{pmatrix},\begin{pmatrix} X(\tilde{\theta}) \\ Y(\tilde{\theta}) \end{pmatrix}\right).$$

We assume that Σ is known, while μ is unknown and unrestricted unless noted otherwise. For brevity of notation, we abbreviate $\Sigma(\theta,\theta)$ to $\Sigma(\theta)$. We assume throughout that $\Sigma_Y(\theta)>0$ for all $\theta\in\Theta$, since the inference problem we study is trivial when $\Sigma_Y(\theta)=0$. As discussed in Section 5 below, this model arises naturally as an asymptotic approximation.

We are interested in inference on $\mu_Y(\hat{\theta})$, where $\hat{\theta}$ is determined based on X. We define

$\hat{\theta}$ through the *level maximization*,[8]

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} X(\theta). \tag{5}$$

In a follow-up paper, Andrews et al. (2020b), we develop results on inference when $\hat{\theta}$ instead maximizes $\|X(\theta)\|$ and $X(\theta)$ may be vector-valued.

We are interested in constructing estimates and confidence intervals for $\mu_Y(\hat{\theta})$ that are valid either conditional on the value of $\hat{\theta}$ or unconditionally. In many cases, as in Section 2 above, we are interested in the mean of the same variable that drives selection, so $X = Y$ and $\mu_X = \mu_Y$. In other settings, however, we may select on one variable but want to do inference on the mean of another. Continuing with the example discussed in Section 2, for instance, we might select $\hat{\theta}$ based on outcomes for all individuals, but want to conduct inference on average outcomes for some subgroup defined using covariates. In this case, $Y(\theta)$ corresponds to the estimated average outcome for the group of interest under treatment θ.

3.2 Conditional Inference

We first consider conditional inference, seeking estimates of $\mu_Y(\hat{\theta})$ which are quantile unbiased conditional on $\hat{\theta}$:

$$Pr_\mu\left\{\hat{\mu}_\alpha \geq \mu_Y(\hat{\theta}) | \hat{\theta} = \tilde{\theta}\right\} = \alpha \text{ for all } \tilde{\theta} \in \Theta \text{ and all } \mu. \tag{6}$$

Since $\hat{\theta}$ is a function of X, we can re-write the conditioning event in terms of the sample space of X as $\left\{X : \hat{\theta} = \tilde{\theta}\right\} = \mathcal{X}(\tilde{\theta})$.[9] Thus, for conditional inference we are interested in the distribution of (X,Y) conditional on $X \in \mathcal{X}(\tilde{\theta})$. Our results below imply that the elements of Y other than $Y(\tilde{\theta})$ do not help in constructing a quantile-unbiased estimate or confidence interval for $\mu_Y(\hat{\theta})$ conditional on $X \in \mathcal{X}(\tilde{\theta})$. Hence, we limit attention to the conditional distribution of $(X,Y(\tilde{\theta}))$ given $X \in \mathcal{X}(\tilde{\theta})$.

Since $(X,Y(\tilde{\theta}))$ is jointly normal unconditionally, it has a multivariate truncated normal distribution conditional on $X \in \mathcal{X}(\tilde{\theta})$. Correlation between X and $Y(\tilde{\theta})$ implies that the conditional distribution of $Y(\tilde{\theta})$ depends on both the parameter of interest $\mu_Y(\hat{\theta})$ and μ_X. To eliminate dependence on the nuisance parameter μ_X, we condition on a sufficient

[8] For simplicity of notation we assume $\hat{\theta}$ is unique almost surely unless noted otherwise.

[9] If $\hat{\theta}$ is not unique we change the conditioning event from $\hat{\theta} = \tilde{\theta}$ to $\tilde{\theta} \in \operatorname{argmax} X(\theta)$.

statistic. Without truncation and for any fixed $\mu_Y(\tilde{\theta})$, a minimal sufficient statistic for μ_X is

$$Z_{\tilde{\theta}} = X - \Big(\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta})\Big)Y(\tilde{\theta}), \tag{7}$$

where we use $\Sigma_{XY}(\cdot,\tilde{\theta})$ to denote $Cov(X,Y(\tilde{\theta}))$. $Z_{\tilde{\theta}}$ corresponds to the part of X that is (unconditionally) orthogonal to $Y(\tilde{\theta})$ which, since $(X,Y(\tilde{\theta}))$ are jointly normal, means that $Z_{\tilde{\theta}}$ and $Y(\tilde{\theta})$ are independent. Truncation breaks this independence, but $Z_{\tilde{\theta}}$ remains minimal sufficient for μ_X. The conditional distribution of $Y(\hat{\theta})$ given $\Big\{\hat{\theta}=\tilde{\theta}, Z_{\tilde{\theta}}=z\Big\}$ is truncated normal:

$$Y(\hat{\theta})|\hat{\theta}=\tilde{\theta}, Z_{\tilde{\theta}}=z \sim \xi|\xi\in\mathcal{Y}(\tilde{\theta},z), \tag{8}$$

where $\xi\sim N\Big(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta})\Big)$ is normally distributed and

$$\mathcal{Y}(\tilde{\theta},z) = \Big\{y : z + \Big(\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta})\Big)y \in \mathcal{X}(\tilde{\theta})\Big\} \tag{9}$$

is the set of values for $Y(\tilde{\theta})$ such that the implied X falls in $\mathcal{X}(\tilde{\theta})$ given $Z_{\tilde{\theta}}=z$. Thus, conditional on $\hat{\theta}=\tilde{\theta}$, and $Z_{\tilde{\theta}}=z$, $Y(\hat{\theta})$ follows a one-dimensional truncated normal distribution with truncation set $\mathcal{Y}(\tilde{\theta},z)$.

The following result, based on Lemma 5.1 of Lee et al. (2016), characterizes $\mathcal{Y}(\tilde{\theta},z)$.

Proposition 1
Let $\Sigma_{XY}(\tilde{\theta}) = Cov(X(\tilde{\theta}),Y(\tilde{\theta}))$. *Define*

$$\mathcal{L}(\tilde{\theta},Z_{\tilde{\theta}}) = \max_{\theta\in\Theta:\Sigma_{XY}(\tilde{\theta})>\Sigma_{XY}(\tilde{\theta},\theta)} \frac{\Sigma_Y(\tilde{\theta})\Big(Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})\Big)}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta},\theta)},$$

$$\mathcal{U}(\tilde{\theta},Z_{\tilde{\theta}}) = \min_{\theta\in\Theta:\Sigma_{XY}(\tilde{\theta})<\Sigma_{XY}(\tilde{\theta},\theta)} \frac{\Sigma_Y(\tilde{\theta})\Big(Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})\Big)}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta},\theta)},$$

and

$$\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}}) = \min_{\theta\in\Theta:\Sigma_{XY}(\tilde{\theta})=\Sigma_{XY}(\tilde{\theta},\theta)} -\Big(Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})\Big).$$

If $\mathcal{V}(\tilde{\theta},z)\geq 0$, *then* $\mathcal{Y}(\tilde{\theta},z) = \Big[\mathcal{L}(\tilde{\theta},z),\mathcal{U}(\tilde{\theta},z)\Big]$. *If* $\mathcal{V}(\tilde{\theta},z)<0$, *then* $\mathcal{Y}(\tilde{\theta},z)=\emptyset$.

Thus, $\mathcal{Y}(\tilde{\theta},z)$ is an interval bounded above and below by functions of z. While we must have $\mathcal{V}(\tilde{\theta},z)\geq 0$ for this interval to be non-empty, $Pr_\mu\Big\{\mathcal{V}(\hat{\theta},Z_{\hat{\theta}})<0\Big\}=0$ for all μ so this

constraint holds almost surely when we consider the value $\hat{\theta}$ observed in the data. Hence, in applications we can safely ignore this constraint and calculate only $\mathcal{L}(\hat{\theta},Z_{\hat{\theta}})$ and $\mathcal{U}(\hat{\theta},Z_{\hat{\theta}})$.

Using this result, it is straightforward to construct quantile-unbiased estimators for $\mu_Y(\hat{\theta})$. Let $F_{TN}(y;\mu_Y(\tilde{\theta}),\tilde{\theta},z)$ denote the distribution function for the truncated normal distribution (8). This function is strictly decreasing in $\mu_Y(\tilde{\theta})$. Define $\hat{\mu}_\alpha$ as the unique solution to $F_{TN}(Y(\hat{\theta});\hat{\mu}_\alpha,\tilde{\theta},Z_{\tilde{\theta}})=1-\alpha$. Proposition 2 below shows that $\hat{\mu}_\alpha$ is conditionally α-quantile-unbiased in the sense of (6), so $\hat{\mu}_{\frac{1}{2}}$ is median-unbiased while the equal-tailed interval $CS_{ET}=\left[\hat{\mu}_{\alpha/2},\hat{\mu}_{1-\alpha/2}\right]$ has conditional coverage $1-\alpha$

$$Pr\left\{\mu_Y(\hat{\theta})\in CS_{ET}|\hat{\theta}=\tilde{\theta}\right\}\geq 1-\alpha \text{ for } \tilde{\theta}\in\Theta \text{ and all } \mu. \tag{10}$$

Moreover results in Pfanzagl (1979) and Pfanzagl (1994) on optimal estimation for exponential families imply that $\hat{\mu}_\alpha$ is optimal in the class of quantile-unbiased estimators.

To establish optimality, we add the following assumption:

Assumption 1
If $\Sigma=Cov((X',Y')')$ has full rank, then the parameter space for μ is $\mathbb{R}^{2|\Theta|}$. Otherwise, there exists some μ^* such that the parameter space for μ is $\left\{\mu^*+\Sigma^{\frac{1}{2}}v : v\in\mathbb{R}^{2|\Theta|}\right\}$, where $\Sigma^{\frac{1}{2}}$ is the symmetric square root of Σ.

This assumption requires that the parameter space for μ be sufficiently rich. When Σ is degenerate (for example when $X=Y$, as in Section 2), this assumption further implies that (X,Y) have the same support for all values of μ. This rules out cases in which a pair of parameter values μ_1, μ_2 can be perfectly distinguished based on the data. Under this assumption, $\hat{\mu}_\alpha$ is an optimal quantile-unbiased estimator.

Proposition 2
For $\alpha\in(0,1)$, $\hat{\mu}_\alpha$ is conditionally α-quantile-unbiased in the sense of (6). If Assumption 1 holds, then $\hat{\mu}_\alpha$ is the uniformly most concentrated α-quantile-unbiased estimator, in that for any other conditionally α-quantile-unbiased estimator $\hat{\mu}^*_\alpha$ and any loss function $L\left(d,\mu_Y(\tilde{\theta})\right)$ that attains its minimum at $d=\mu_Y(\tilde{\theta})$ and is quasiconvex in d for all $\mu_Y(\tilde{\theta})$,

$$E_\mu\left[L\left(\hat{\mu}_\alpha,\mu_Y(\tilde{\theta})\right)|\hat{\theta}=\tilde{\theta}\right]\leq E_\mu\left[L\left(\hat{\mu}^*_\alpha,\mu_Y(\tilde{\theta})\right)|\hat{\theta}=\tilde{\theta}\right]$$

for all μ and all $\tilde{\theta}\in\Theta$.

Proposition 2 shows that $\hat{\mu}_\alpha$ is optimal in the strong sense that it has lower expected loss than any other quantile-unbiased estimator for a large class of loss functions.

Other Selection Events We have discussed inference conditional on $\hat{\theta}=\tilde{\theta}$, but the same approach applies, and is optimal, for more general conditioning events. For instance, in the context of Section 2 a researcher might deliver a recommendation to a policymaker only when a statistical test indicates that the best-performing treatment outperforms some benchmark (see Tetenov, 2012). In this case, it is natural to also condition inference on the result of this test. Analogously, one may wish to conduct inference on the performance of an estimated best trading strategy or forecasting rule after finding a rejection when testing for superior predictive ability according to methods of e.g. White (2000), Hansen (2005) or Romano and Wolf (2005). Appendix A discusses the conditional approach in this more general case and derives the additional conditioning event in the context of the example just described.

Uniformly Most Accurate Unbiased Confidence Intervals In addition to equal-tailed confidence intervals, classical results on testing in exponential families discussed in Fithian et al. (2017) also permit the construction of uniformly most accurate unbiased confidence intervals. A level $1-\alpha$ confidence set is unbiased if its probability of covering a false parameter value is bounded above by $1-\alpha$, and uniformly most accurate unbiased confidence intervals minimize the probability of covering all incorrect parameter values over the class of unbiased confidence sets. Details of how to construct these confidence intervals are deferred to Appendix A for brevity.

3.3 Comparison to Sample Splitting

An alternative remedy for winner's curse bias is to split the sample. If we have iid observations and select $\hat{\theta}^1$ based on the first half of the data, conventional estimates and confidence intervals for $\mu_Y(\hat{\theta}^1)$ that use only the second half of the data will be conditionally valid given $\hat{\theta}_1$. Hence, it is natural to ask how the analog of our conditioning approach applied to inference on $\mu_Y(\hat{\theta}^1)$, conditional on the realization of $\hat{\theta}^1$, compares to this conventional sample splitting approach.

Asymptotically, even sample splits yield a pair of independent and identically distributed normal draws (X^1,Y^1) and (X^2,Y^2), both of which follow (4), albeit with a different scaling for (μ,Σ) than in the full-sample case.[10] Sample splitting procedures calculate $\hat{\theta}^1$ as

[10] Appendix C considers cases with general sample splits and describes the scaling for (μ,Σ). Intuitively, the scope for improvement over conventional split-sample inference is increasing in the fraction of the data used to construct X_1.

in (5), replacing X by X^1. Inference on $\mu_Y(\hat{\theta}^1)$ is then conducted using Y^2. In particular, the conventional 95% sample-splitting confidence interval for $\mu_Y(\hat{\theta}^1)$,

$$\left[Y^2(\hat{\theta}^1)-1.96\sqrt{\Sigma_Y(\hat{\theta}^1)},Y^2(\hat{\theta}^1)+1.96\sqrt{\Sigma_Y(\hat{\theta}^1)}\right],$$

has correct (conditional) coverage, and $Y^2(\hat{\theta}^1)$ is median-unbiased for $\mu_Y(\hat{\theta}^1)$.

Conventional sample splitting resolves the winner's curse, but comes at a cost. First, $\hat{\theta}^1$ is based on less data than in the full-sample case, which is unappealing since a policy recommendation estimated with a smaller sample size leads to a lower expected welfare (see, e.g., Theorems 2.1 and 2.2 in Kitagawa and Tetenov (2018b)). Moreover, even after conditioning on $\hat{\theta}^1$, the full-sample average $\frac{1}{2}(X^1,Y^1)+\frac{1}{2}(X^2,Y^2)$ remains minimal sufficient for μ. Hence, using only Y^2 for inference sacrifices information.

Fithian et al. (2017) formalize this point and show that conventional sample splitting tests (and thus confidence intervals) are inefficient.[11] Motivated by this result, in Appendix C we derive optimal estimators and confidence intervals for $\mu_Y(\hat{\theta}^1)$ that are valid conditional on $\hat{\theta}^1$. These optimal split-sample procedures involve distributions that are difficult to compute, however, so we also propose computationally straightforward alternatives. These alternatives dominate conventional split-sample methods for inference on $\mu_Y(\hat{\theta}^1)$, but are in turn dominated by the (intractable) optimal split-sample procedures. The split sample methods we introduce in Appendix C are related to conditionally valid methods for inference in adaptive clinical trial designs proposed in the biostatistics literature (e.g., Cohen and Sackrowitz, 1989; Sampson and Sill, 2005). See Appendix C for details.

3.4 Behavior When $Pr_\mu\left\{\hat{\theta}=\tilde{\theta}\right\}$ is Large

As discussed in Section 2, if we ignore selection and compute the conventional (or "naive") estimator $\hat{\mu}_N=Y(\hat{\theta})$ and the conventional confidence interval

$$CS_N=\left[Y(\hat{\theta})-c_{\alpha/2,N}\sqrt{\Sigma_Y(\hat{\theta})},Y(\hat{\theta})+c_{\alpha/2,N}\sqrt{\Sigma_Y(\hat{\theta})}\right],$$

where $c_{\alpha,N}$ is the $1-\alpha$-quantile of the standard normal distribution, $\hat{\mu}_N$ is biased and CS_N has incorrect coverage conditional on $\hat{\theta}=\tilde{\theta}$. These biases are mild when $Pr_\mu\left\{\hat{\theta}=\tilde{\theta}\right\}$ is close to one, however, since in this case the conditional distribution is close to the unconditional

[11] Corollary 1 of Fithian et al. (2017) applied in our setting shows that for any sample splitting test based on Y^2, there exists a test that uses the full data and has weakly higher power against all alternatives and strictly higher power against some alternatives.

one. Intuitively, $Pr_\mu\left\{\hat{\theta}=\tilde{\theta}\right\}$ is close to one for some $\tilde{\theta}$ when $\mu_X(\theta)$ has a well-separated maximum. Our procedures converge to conventional ones in this case.

Proposition 3
Consider any sequence of values μ_m such that $Pr_{\mu_m}\left\{\hat{\theta}=\tilde{\theta}\right\}\to 1$. Then under μ_m we have $CS_{ET}\to_p CS_N$ and $\hat{\mu}_{\frac{1}{2}}\to_p Y(\tilde{\theta})$ both conditional on $\hat{\theta}=\tilde{\theta}$ and unconditionally, where for confidence intervals $\to_p$ denotes convergence in probability of the endpoints.

This result provides an additional argument for using our procedures: they remain valid when conventional procedures fail, but coincide with conventional procedures when the latter are valid. On the other hand, as we saw in Section 2, there are cases where our conditional procedures have poor unconditional performance.

4 Unconditional Inference

Rather than requiring validity conditional on $\hat{\theta}$, one might instead require coverage only on average, yielding the unconditional coverage requirement

$$Pr\left\{\mu_Y(\hat{\theta})\in CS\right\}\geq 1-\alpha \text{ for all } \mu. \tag{11}$$

All confidence intervals with correct conditional coverage in the sense of (10) also have correct unconditional coverage provided $\hat{\theta}$ is unique with probability one.

Proposition 4
Suppose that $\hat{\theta}$ is unique with probability one for all μ. Then any confidence interval CS with correct conditional coverage (10) also has correct unconditional coverage (11).

Uniqueness of $\hat{\theta}$ implies that the conditioning events $\mathcal{X}(\tilde{\theta})$ partition the support of X with measure zero overlap. The result then follows from the law of iterated expectations.

A sufficient condition for almost sure uniqueness of $\hat{\theta}$ is that Σ_X has full rank. A weaker sufficient condition is given in the next lemma. Cox (2018) gives sufficient conditions for uniqueness of a global optimum in a much wider class of problems.

Lemma 1
Suppose that for all θ, $\tilde{\theta}\in\Theta$ such that $\theta\neq\tilde{\theta}$, $X(\theta)$ and $X(\tilde{\theta})$ are not perfectly (positively) correlated. Then $\hat{\theta}$ is unique with probability one for all μ.

While the conditional confidence intervals derived in the last section are unconditionally valid, unconditional coverage is less demanding than conditional coverage. Hence, if we

are only concerned with unconditional coverage, relaxing the coverage requirement may allow us to obtain shorter confidence intervals in some settings.

This section explores the benefits of such a relaxation. We begin by introducing unconditional confidence intervals based on projections of simultaneous confidence bands for μ. We then introduce hybrid estimators and confidence intervals that combine projection intervals with conditioning arguments.

4.1 Projection Confidence Intervals

One approach to obtain an unconditional confidence interval for $\mu_Y(\hat{\theta})$ is to start with a joint confidence interval for μ and project on the dimension corresponding to $\hat{\theta}$. This approach was used by Romano and Wolf (2005) in the context of multiple testing, and by Kitagawa and Tetenov (2018a) for inference on an estimated optimal policy. This approach has also been used in a large and growing statistics literature on post-selection inference including e.g. Berk et al. (2013), Bachoc et al. (2020) and Kuchibhotla et al. (2020).

To formally describe the projection approach, let c_α denote the $1-\alpha$ quantile of $\max_\theta |\xi(\theta)|/\sqrt{\Sigma_Y(\theta)}$ for $\xi \sim N(0, \Sigma_Y)$. If we define

$$CS_\mu = \left\{ \mu_Y : |Y(\theta) - \mu_Y(\theta)| \leq c_\alpha \sqrt{\Sigma_Y(\theta)} \text{ for all } \theta \in \Theta \right\},$$

then CS_μ is a level $1-\alpha$ confidence set for μ_Y. If we then define

$$CS_P = \left\{ \tilde{\mu}_Y(\hat{\theta}) : \exists \mu \in CS_\mu \text{ such that } \mu_Y(\hat{\theta}) = \tilde{\mu}_Y(\hat{\theta}) \right\} = \left[Y(\hat{\theta}) - c_\alpha \sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_\alpha \sqrt{\Sigma_Y(\hat{\theta})} \right]$$

as the projection of CS_μ on the parameter space for $\mu_Y(\hat{\theta})$, then since $\mu_Y \in CS_\mu$ implies $\mu_Y(\hat{\theta}) \in CS_P$, CS_P satisfies the unconditional coverage requirement (11). As noted in Section 2, however, CS_P does not generally have correct conditional coverage.

The width of the confidence interval CS_P depends on the variance $\Sigma_Y(\hat{\theta})$ but does not otherwise depend on the data.[12] To account for the randomness of $\hat{\theta}$, the critical value c_α is typically larger than the conventional two-sided normal critical value. Hence, CS_P will be conservative in cases where $\hat{\theta}$ takes a given value $\tilde{\theta}$ with high probability. To improve performance in such cases, we propose a hybrid inference approach.

[12] One could consider alternative projection intervals, for instance optimized to have shorter length at some $\hat{\theta}$ values in exchange for greater length at others. See Freyberger and Rai (2018) and Frandsen (2020).

4.2 Hybrid Inference

As shown in Section 2, the conditional and projection approaches each have good unconditional performance in some cases, but neither is fully satisfactory. Hybrid inference combines the approaches to obtain good performance over a wide range of parameter values.

To construct hybrid estimators, we condition both on $\hat{\theta}=\tilde{\theta}$ and on the event that $\mu_Y(\hat{\theta})$ lies in the level $1-\beta$ projection confidence interval CS_P^β for $0\leq\beta<\alpha$. Hence, the conditioning event becomes

$$\mathcal{Y}^H(\tilde{\theta},\mu_Y(\tilde{\theta}),z)=\mathcal{Y}(\tilde{\theta},z)\cap\left[\mu_Y(\tilde{\theta})-c_\beta\sqrt{\Sigma_Y(\tilde{\theta})},\mu_Y(\tilde{\theta})+c_\beta\sqrt{\Sigma_Y(\tilde{\theta})}\right].$$

Let $F_{TN}^H(y;\mu_Y(\tilde{\theta}),\tilde{\theta},z)$ denote the conditional distribution function of $Y(\tilde{\theta})$, and define $\hat{\mu}_\alpha^H$ to solve $F_{TN}^H(Y(\hat{\theta});\hat{\mu}_\alpha^H,\hat{\theta},Z_{\tilde{\theta}})=1-\alpha$. The hybrid estimator $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on $\mu(\hat{\theta})\in CS_P^\beta$.

Proposition 5

For $\alpha\in(0,1)$, $\hat{\mu}_\alpha^H$ is unique and $\hat{\mu}_\alpha^H\in CS_P^\beta$. If $\hat{\theta}$ is unique almost surely for all μ, $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on $\mu_Y(\hat{\theta})\in CS_P^\beta$:

$$Pr_\mu\left\{\hat{\mu}_\alpha^H\geq\mu_Y(\hat{\theta})|\mu_Y(\hat{\theta})\in CS_P^\beta\right\}=\alpha \text{ for all } \mu.$$

Proposition 5 implies several notable properties for the hybrid estimator. First, since $Pr_\mu\left\{\mu_Y(\hat{\theta})\in CS_P^\beta\right\}\geq 1-\beta$ by construction, one can show that

$$\left|Pr_\mu\left\{\hat{\mu}_\alpha^H\geq\mu_Y(\hat{\theta})\right\}-\alpha\right|\leq\beta\cdot\max\{\alpha,1-\alpha\} \text{ for all } \mu.$$

This implies that the absolute median bias of $\hat{\mu}_{\frac{1}{2}}^H$ (measured as the deviation of the exceedance probability from 1/2) is bounded above by $\beta/2$. On the other hand, since $\hat{\mu}_{\frac{1}{2}}^H\in CS_P^\beta$ we have $\left|\hat{\mu}_{\frac{1}{2}}^H-Y(\hat{\theta})\right|\leq c_\beta\sqrt{\Sigma_Y(\tilde{\theta})}$, so the difference between $\hat{\mu}_{\frac{1}{2}}^H$ and the conventional estimator $Y(\hat{\theta})$ is bounded above by half the width of CS_P^β. As β varies, the hybrid estimator interpolates between the median-unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ and the conventional estimator $Y(\hat{\theta})$.

As with the quantile-unbiased estimator $\hat{\mu}_\alpha$, we can form confidence intervals based on hybrid estimators. In particular, the set $[\hat{\mu}_{\alpha/2}^H,\hat{\mu}_{1-\alpha/2}^H]$ has coverage $1-\alpha$ conditional on $\mu_Y(\hat{\theta})\in CS_P^\beta$. This is not fully satisfactory, however, as $Pr_\mu\{\mu_Y(\hat{\theta})\in CS_P^\beta\}<1$. Hence, to ensure correct coverage, we define the level $1-\alpha$ hybrid confidence interval

as $CS^H_{ET} = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)}}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)}}\right]$. With this adjustment, hybrid confidence intervals have coverage at least $1-\alpha$ both conditional on $\mu_Y(\hat{\theta}) \in CS^\beta_P$ and unconditionally.

Proposition 6

Provided $\hat{\theta}$ is unique with probability one for all μ, the hybrid confidence interval CS^H_{ET} has coverage $\frac{1-\alpha}{1-\beta}$ conditional on $\mu_Y(\hat{\theta}) \in CS^\beta_P$:

$$Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS^H_{ET} | \mu_Y(\hat{\theta}) \in CS^\beta_P\right\} = \frac{1-\alpha}{1-\beta} \text{ for all } \mu.$$

Moreover, the unconditional coverage is between $1-\alpha$ and $\frac{1-\alpha}{1-\beta} \leq 1-\alpha+\beta$:

$$\inf_\mu Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS^H_{ET}\right\} \geq 1-\alpha, \quad \sup_\mu Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS^H_{ET}\right\} \leq \frac{1-\alpha}{1-\beta}.$$

Hybrid confidence intervals strike a balance between the conditional and projection approaches. The maximal length of hybrid confidence intervals is bounded above by the length of CS^β_P. For small β, hybrid confidence intervals will be close to conditional confidence intervals, and thus to conventional confidence intervals, when $\hat{\theta} = \tilde{\theta}$ with high probability. For $\beta > 0$, however, hybrid confidence intervals do not fully converge to conventional confidence intervals as $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}\right\} \to 1$. Nevertheless, our simulations in Section 2 find similar performance for the hybrid and conditional approaches in well-separated cases.

While hybrid confidence intervals combine the conditional and projection approaches, they can yield overall performance more appealing than either. In Section 2 we found that hybrid confidence intervals had a shorter median length for many parameter values than did either the conditional or projection approaches used in isolation. Our simulation results below provide further evidence of outperformance in realistic settings.

Choice of β To use the hybrid approach we must select the coverage β of the initial projection interval CS^β_P. Intuitively this choice trades off the length of CS^β_P, which bounds the worst-case length of CS^H_{ET} in the poorly-separated case, against the length of CS^H_{ET} in the well-separated case. For a given Σ we can precisely quantify this tradeoff, calculating the length of CS^β_P and the length of CS^H_{ET} in the well-separated case for each β and selecting a point on the resulting frontier. This frontier is Σ-specific, however, so this analysis does not deliver a general recommendation.

As an alternative, we note that the length of CS^H_{ET} in the well-separated case is bounded above by that of the level $\frac{1-\alpha}{1-\beta}$ conventional confidence interval. Specifically, for the standard

choice of $\alpha=5\%$, choosing $\beta=\frac{\alpha}{10}=0.5\%$ implies that the CS^H_{ET} has half-length no more than 2.0025 standard errors in the well-separated case. We view this as a small increase relative to the half-length of the conventional 5% interval, 1.96 standard errors, and so suggest this as a default choice, focusing on $\beta=\frac{\alpha}{10}$ in our simulations and applications.[13]

Comparison to Bonferroni Adjustment It is worth contrasting our hybrid approach with Bonferroni corrections as in e.g. Romano et al. (2014) and McCloskey (2017). A simple Bonferroni approach for our setting intersects a level $1-\beta$ projection confidence interval CS^β_P with a level $1-\alpha+\beta$ conditional interval that conditions only on $\hat{\theta}=\tilde{\theta}$. Bonferroni intervals differ from our hybrid approach in two respects. First, they use a level $1-\alpha+\beta$ conditional confidence interval, while the hybrid approach uses a level $\frac{1-\alpha}{1-\beta}$ conditional interval, where $\frac{1-\alpha}{1-\beta}\leq 1-\alpha+\beta$. Second, the conditional interval used by the Bonferroni approach does not condition on $\mu_Y(\tilde{\theta})\in CS^\beta_P$, while that used by the hybrid approach does. Consequently, one can show that hybrid confidence intervals exclude the endpoints of CS^β_P almost surely, while the same is not true of Bonferroni intervals.

5 Feasible Inference and Large-Sample Results

Our results have so far assumed that (X,Y) are jointly normal with known variance Σ. While exact normality is rare in practice, researchers commonly use estimators that are asymptotically normal with consistently estimable asymptotic variance. Our results for the finite-sample normal model translate to asymptotic results in this case.

Specifically, suppose that for sample size n we construct a vector of statistics X_n, that $\hat{\theta}_n=\text{argmax}_{\theta\in\Theta}X_n(\theta)$, and that we are interested in the mean of $Y_n(\hat{\theta}_n)$. In the treatment choice example discussed in Section 2, for instance, θ indexes treatments, $X_n(\theta)$ is $\sqrt{n}$ times the sample average outcome under treatment θ, and $Y_n(\theta)=X_n(\theta)$. We suppose that (X_n,Y_n) are jointly asymptotically normal once recentered by vectors $(\mu_{X,n},\mu_{Y,n})$,

$$\begin{pmatrix} X_n-\mu_{X,n} \\ Y_n-\mu_{Y,n} \end{pmatrix} \Rightarrow N(0,\Sigma).$$

In the treatment choice example $\mu_{X,n}(\theta)=\mu_{Y,n}(\theta)=\sqrt{n}E[Y_{i,\theta}]$ is the average potential outcome under treatment θ, scaled by $\sqrt{n}$. We further assume that we have a consistent estimator $\widehat{\Sigma}$ for the asymptotic variance Σ. In the treatment choice example, for instance, we can take $\widehat{\Sigma}$ to be the matrix with the sample variance of the outcome for each the

[13]Romano et al. (2014) and McCloskey (2017) likewise find this choice to perform well in two different settings when using a Bonferroni correction

treatment group along the diagonal and zeros elsewhere.

More broadly, (X_n,Y_n) can be any vectors of asymptotically normal estimators, and we can calculate $\widehat{\Sigma}$ as we would for inference on $(\mu_{X,n},\mu_{Y,n})$, including corrections for clustering, serial correlation, and the like in the usual way.[14] Feasible inference based on our approach simply substitutes (X_n,Y_n) and $\widehat{\Sigma}$ in place of (X,Y) and Σ in all expressions. Appendix D shows that this plug-in approach yields asymptotically valid inference on $\mu_{Y,n}(\hat{\theta}_n)$. This result is trivial when the sequence of vectors $\mu_{X,n}$ has a well-separated maximizer $\theta^* = \text{argmax}_{\theta\in\Theta}\mu_{X,n}(\theta)$ with $\mu_{X,n}(\theta^*) \gg \max_{\theta\in\Theta\setminus\theta^*}\mu_{X,n}(\theta)$ for large n, since in this case $\hat{\theta}_n = \theta^*$ with high probability, and the selection problem vanishes. In Section 2, for instance, if we fix a data-generating process with $E[Y_{i,1}] > E[Y_{i,0}]$ and take $n\to\infty$, then $Pr\{\hat{\theta}_n = 1\}\to 1$.

Based on this argument, it could be tempting to conclude that inference ignoring the winner's curse will be approximately valid so long as there is not an exact tie for the treatment yielding the highest average outcome. In finite samples, however, near-ties yield very similar behavior to exact ties. Moreover, no matter how large the sample size, we can have near-ties sufficiently close that inference ignoring selection remains unreliable. Hence, what matters for inference is neither whether there are exact ties, nor the sample size as such (beyond the minimum needed to justify the normal approximation), but instead how close the best-performing treatments are to each other *relative* to the degree of sampling uncertainty. Depending on the data generating process, selection issues can thus remain important no matter how large the sample. To obtain reliable large-sample approximations, we thus seek *uniform* asymptotic results, which for sufficiently large samples guarantee performance over a wide class of data generating processes. Appendix D establishes that plug-in versions of our proposed procedures are uniformly asymptotically valid in this sense.

6 Application: Charitable Giving

Karlan and List (2007) partner with a political charity to conduct a field experiment examining the effectiveness of matching incentives at increasing charitable giving. In matched donations, a lead donor pledges to 'match' any donations made by other donors up to some threshold, effectively lowering the price of political activism for other donors.

Karlan and List (2007) use a factorial design. Potential donors, who were previous donors to the charity, were mailed a four page letter asking for a donation. The contents of the letter were randomized, with one third of the sample assigned to a control group

[14] We have scaled (X_n,Y_n) by $\sqrt{n}$ for expositional purposes, but dropping this scaling yields inference on the correspondingly scaled version of $\mu_{Y,n}(\hat{\theta}_n)$. Hence, one can use estimators and estimated variances with the natural scale in a given setting.

Index	Treatment	Description
0	0	Control group with no matched donations
Match ratio		
1	1:1	An additional dollar up to the match limit
2	2:1	Two additional dollars up to the match limit
3	3:1	Three additional dollars up to the match limit
Match size		
1	\$25,000	Up to \$25,000 is pledged
2	\$50,000	Up to \$50,000 is pledged
3	\$100,000	Up to \$100,000 is pledged
4	Unstated	The pledged amount is not stated
Ask amount		
1	Same	The individual is asked to give as much as their largest past donation
2	25% more	The individual is asked to give 25% more than their largest past donation
3	50% more	The individual is asked to give 50% more than their largest past donation

Table 1: Treatment arms for Karlan and List (2007). Individuals were assigned to the control group or to the treatment group, in the ratio 1:2. Treated individuals were randomly assigned a match ratio, a match size and an ask amount with equal probability. There are 36 possible combinations, plus the control group. The leftmost column specifies a reference index used throughout this section for convenience.

that received a standard letter with no match. The remaining two thirds received a letter with the line "now is the time to give!" and details for a match. Treated individuals were randomly assigned with equal probability to one of 36 separate treatment arms. Treatment arms are characterized by a match ratio, a match size, and an ask amount, for which further details are given in Table 1. The outcome of interest is the average dollar amount that individuals donated to the charity in the month following the solicitation.

In total, 50,083 individuals were contacted, of which 16,687 were randomly assigned to the control group, while 33,396 were randomly assigned to one of the 36 treatment arms. The (unconditional) average donation was \$0.81 in the control group and \$0.92 in the treatment group. Conditional on giving, these figures were \$45.54 and \$44.35, respectively. The discrepancy reflects the low response rate; only 1,034 of 50,083 individuals donated.

Treatment	Average donation	Standard error	95% CI
(1,3,2)	1.52	0.35	[0.83,2.20]
(2,1,3)	1.51	0.46	[0.61,2.41]
(2,1,1)	1.42	0.39	[0.66,2.19]
(3,1,3)	1.40	0.36	[0.70,2.11]

Table 2: The average donations for the four best treatment arms according to the data, $n=50{,}083$. Treatments are indexed by the indicators for (Match ratio, Match size, Ask amount) defined in Table 1. The reported 95% confidence intervals are the conventional ones that do not take selection into account.

Table 2 reports average revenue from the four best-performing treatment arms, along with standard errors and conventional confidence intervals. Taken at face value, the results for the best-performing arm suggest that a similarly-situated nonprofit considering a campaign that promises a dollar-for-dollar match up to \$100,000 in donations and asks individuals to donate 25% more than their largest past donation could expect to raise \$1.52 per potential donor, on average, with a confidence interval of \$0.83 to \$2.20. This estimate and confidence interval are clearly subject to winner's curse bias, however: we are picking the best-performing arm out of 37 in the experiment, which will bias our estimates and confidence intervals upward.

Simulation Results To investigate the extent of winner's curse bias and the finite-sample performance of our corrections, we calibrate simulations to this application. We simulate datasets by resampling observations with replacement from the Karlan and List (2007) data (i.e. by drawing nonparametric bootstrap samples). In each simulated sample we re-estimate the effectiveness of each treatment arm, pick the best-performing arm, and study the performance of estimates and confidence intervals, treating the estimates for the original Karlan and List (2007) data as the true values. The underlying data here are non-normal and we re-estimate the variance in each simulation draw. Hence, these results also speak to the finite-sample performance of the normal approximation. We report results based on 10,000 simulation draws.

Since revenue does not account for the cost of the fund-raising campaign, it is impossible for the solicitation to raise a negative amount. We therefore set the parameter space for $\mu(\hat{\theta})$ to $\mathbb{R}_+$, and trim the point estimators and the confidence intervals at zero, $\hat{\mu}^{trim} \equiv \max\{0,\hat{\mu}\}$ and $CS^{trim} = [0,\infty) \cap CS$. This trimming does not affect the coverage of the confidence intervals, and also preserves the α-quantile unbiasedness of the estimators

Winner											
(1,3,2)	(1,4,2)	(1,4,3)	(2,1,1)	(2,1,3)	(2,2,2)	(2,3,3)	(2,4,1)	(2,4,2)	(3,1,1)	(3,1,3)	(3,3,1)
16.0%	11.4%	1.3%	13.0%	18.9%	10.8%	1.3%	1.5%	2.8%	5.1%	10.0%	3.6%

Table 3: Frequency of simulation replications where each treatment is estimated to perform best in simulations calibrated to Karlan and List (2007). Treatments are indexed by the indicators for (Match ratio, Match size, Ask amount) defined in Table 1. 31 of the 37 treatments are best in at least one replication; those that won in at least 1% of simulated samples are reported.

	Estimate		
	Conventional	Median unbiased	Hybrid
Median bias	0.61	-0.18	-0.18
Probability bias	0.50	-0.07	-0.07
Median absolute error	0.61	0.65	0.64

Table 4: Performance measures for alternative estimators in simulations calibrated to Karlan and List (2007). Probability bias is $Pr\{\hat{\mu}^{trim} > \mu(\hat{\theta})\} - \frac{1}{2}$.

so long as the true value $\mu(\hat{\theta})$ is greater than zero.

There is substantial variability in the "winning" arm: 31 of the 37 treatments won in at least one simulation draw and 12 treatment arms won in at least 1% of simulated samples. Table 3 lists these 12 treatments. The variability of the winning arm suggests that there is scope for a winner's curse in this setting.

Table 4 examines the performance of naive, median unbiased, and hybrid estimates, reporting (unconditional) median bias, probability bias ($Pr\{\hat{\mu}^{trim} > \mu(\hat{\theta})\} - \frac{1}{2}$), and median absolute error. Trimming the estimators at zero does not affect the reported performance measures. Naive estimates suffer from substantial bias in this setting: they have a median bias of \$0.61, and over-estimate the revenue generated by the selected arm 100% of the time, up to rounding. The median unbiased and hybrid estimators substantially improve both measures of bias, though given the finite-sample setting they do not eliminate it completely and are both somewhat downward biased, though to a lesser degree.[15] All three estimators perform similarly in terms of median absolute error.

Tables 5 and 6 report results for confidence intervals. Specifically, we consider the naive,

[15]This is a particularly challenging setting for the normal approximation, as the outcomes distribution is highly skewed due to the large number of zeros. In particular, there are on average only 20 nonzero outcomes per non-control treatment (out of approximately 930 observations in each treatment group).

projection, conditional, and hybrid confidence intervals with nominal coverage 95%. Table 5 reports unconditional coverage and median length, while Table 6 reports conditional coverage probabilities given $\hat{\theta}$ values among the 12 treatments listed in Table 3. Naive confidence intervals slightly undercover unconditionally, with coverage 92%. Their conditional coverage varies depending on which treatment is the winner. If the winning treatment is one of the six best-performing treatments, the conditional coverage is at least 95%, while otherwise the naive confidence intervals under-cover with coverage probability as low as 65%.

	Unconditional coverage	Median length	
		Trimmed	Untrimmed
Naive CS	0.92	1.88	1.88
CS_P	1.00	3.08	3.08
CS_{ET}	0.97	2.69	5.91
CS^H_{ET}	0.97	2.52	2.56

Table 5: Unconditional coverage probabilities of the confidence intervals in simulations calibrated to Karlan and List (2007). Unconditional median lengths are reported for the trimmed and untrimmed confidence intervals.

Treatment	Average donation	Conditional coverage			
θ	$\mu(\theta)$	Naive CS	CS_P	CS_{ET}	CS^H_{ET}
(1,3,2)	1.52	0.95	1	0.98	0.98
(2,1,3)	1.51	0.97	1	0.97	0.97
(2,1,1)	1.42	0.94	1	0.97	0.97
(3,1,3)	1.40	0.95	1	0.97	0.97
(2,2,2)	1.34	0.96	1	0.97	0.98
(1,4,2)	1.27	0.99	1	0.97	0.97
(3,3,1)	1.26	0.84	1	0.96	0.97
(3,1,1)	1.24	0.89	1	0.97	0.97
(2,4,2)	1.22	0.79	1	0.99	0.99
(2,3,3)	1.12	0.65	1	0.98	0.98
(2,4,1)	1.10	0.81	1	0.97	0.97
(1,4,3)	1.03	0.78	1	0.96	0.97

Table 6: Conditional coverage probabilities, $Pr\{\mu(\hat{\theta}) \in CS^{trim} | \hat{\theta} = \theta\}$, of the confidence intervals for each of the 12 treatments in Table 3. The treatments are sorted according to the average donation.

Treatment (1,3,2)		Estimates	Equal-tailed CI
Naive		1.52	[0.83,2.20]
Projection		–	[0.40,2.63]
Conditional	– trimmed	0	[0, 1.42]
	– untrimmed	-7.49	[-47.66,1.42]
Hybrid		0.20	[0.19,1.47]

Table 7: Naive and bias-corrected estimates and confidence intervals for best-performing treatment in Karlan and List (2007) data.

Projection confidence intervals over-cover unconditionally and conditionally for these treatments, with coverage 100%. Conditional and hybrid confidence intervals slightly over-cover, with unconditional and conditional coverage about 97%, and have unconditional median (trimmed) length around 35% larger than naive intervals and around 20% shorter than projection intervals. It is important to emphasize, however, that the conditional coverage for projection and hybrid intervals is particular to the data generating process considered here: as illustrated in Figure 5, these intervals do not ensure conditional coverage in general.

The median length of conditional intervals more than doubles if we leave their lower bound untrimmed. In contrast, the median length of the hybrid confidence intervals is basically unaffected by trimming. This is because despite the similarity of their upper bounds, the lower bound of the conditional confidence intervals tends to be negative and substantially lower than the lower bound of the hybrid confidence intervals. In other words, if the parameter space is unconstrained, the hybrid confidence intervals are substantially shorter than conditional confidence intervals. The good performance of the hybrid approach in applications with unconstrained parameter space is encouraging, and in line with the results in Section 2.

Empirical results Returning to the Karlan and List (2007) data, Table 7 reports corrected estimates and confidence intervals for the best-performing treatment in the experiment. We repeat the naive estimate and confidence interval for comparison. The median unbiased estimate makes an aggressive downwards correction to the naive estimate, suggesting negative revenue (-$7.49) from the winning arm if not trimmed. The conditional confidence interval is tight, ranging from 0 to $1.42, if trimmed at zero, and otherwise extremely wide, ranging from -$47.66 to $1.42. The hybrid estimate also shifts the conventional estimate downwards, but much less so. Moreover, the hybrid confidence interval is no wider than the naive interval, and excludes both zero and the naive estimate.

These results suggest that future fundraising campaigns deploying the winning strategy in the experiment are likely to raise some revenue, but substantially less than would be expected based on the naive estimates.

Conditional inference seems potentially natural in this application. The data highlight an interpretable combination of treatment parameters (1:1 match, $100,000 pledged, with an ask 25% above an individual's highest past donation) as best-performing, raising the question of what we can conclude about this particular treatment, given that it was the best in the experiment. This is precisely the question answered by the conditional approach. By contrast, while the hybrid approach ensures correct coverage on average across different "winning" treatments which could arise, it offers no guarantees given the particular winner observed in the Karlan and List (2007) data.

7 Application: Neighborhood Effects

We next discuss simulation and empirical results based on Chetty et al. (2018) and Bergman et al. (2020). Earlier work, including Chetty and Hendren (2016a) and Chetty and Hendren (2016b) argues that the neighborhood in which a child grows up has a long-term causal impact on income in adulthood, and, moreover, that these impacts are closely related to the adult income of children who spend their entire childhood in a given neighborhood.

Motivated by these findings, Bergman et al. (2020) partnered with the public housing authorities in Seattle and King County in Washington State in an experiment helping housing voucher recipients move to a set of higher-opportunity target neighborhoods. Bergman et al. (2020) choose target neighborhoods based on the Chetty et al. (2018) "Opportunity Atlas." This atlas compiles census-tract level estimates of economic mobility for communities across the United States. Bergman et al. (2020) define target neighborhoods by selecting approximately the top third of tracts in Seattle and King County based on estimated economic mobility.[16] They then make "relatively minor" adjustments to the set of target tracts based on other criteria (Bergman et al., 2020, Appendix A).

A central question in this setting is whether families moving to the target tracts will in fact experience the positive outcomes predicted based on the Opportunity Atlas estimates and the hypothesis of neighborhood effects. Once long-term outcomes for the experimental sample are available, one can begin address this question by comparing outcomes for children in treated families to the Opportunity Atlas estimates used to select the target tracts in

[16]They measure economic mobility in terms of the average household income rank in adulthood for children growing up at the 25th percentile of the income distribution. See Chetty et al. (2018) for details.

the first place. Such a comparison is complicated by the winner's curse, however: the Atlas estimates were already used to select the target tracts, so the naive estimate for the causal effect of the selected tracts will be systematically biased upwards. It is therefore useful to examine the extent of the winner's curse, and the impact of our corrections, in this setting.

Motivated by related issues, Chetty et al. (2018) and Bergman et al. (2020) do not focus on naive estimates, but instead adopt a shrinkage or empirical Bayes approach. Their estimates correspond to Bayesian posterior means under a prior that takes tract-level economic mobility to be normally distributed conditional on a vector of observable tract characteristics, and then estimates mean and variance hyperparameters from the data. If one takes this prior seriously and abstracts from estimation of the hyperparameters (for instance because the number of tracts is large and we plug in consistent estimates), the posterior median for average economic mobility over selected tracts will be median-unbiased under the prior, and Bayesian credible sets will have correct coverage, again under the prior. See Section E in the supplement for further discussion. The efficacy of Bayesian approaches for correcting selection issues hinges crucially on correct specification of the prior, however, whereas our results ensure correct coverage and controlled median bias for all possible distributions of economic mobility across tracts. Throughout this section, we therefore include empirical Bayes procedures in our analysis as a point of comparison.[17]

Simulation Results To examine the extent of winner's curse bias and the performance of different corrections, we calibrate simulations to the Opportunity Atlas data. For each of the 50 largest commuting zones (CZs) in the United States we treat the (un-shrunk) tract-level Opportunity Atlas estimates as the true values. We then simulate estimates by adding normal noise with standard deviation equal to the Opportunity Atlas standard error.[18]

We select the top third of tracts in each commuting zone based on these simulated estimates.[19] To cast this into our setting, let $\mathcal{T}$ be the set of tracts in a given CZ and

[17] Armstrong et al. (2020) propose an approach to robustify empirical Bayes confidence intervals to the choice of priors. Applied in the present setting, this approach would ensure correct coverage for tract-level economic mobility on average across all tracts in a given commuting zone. This approach does not focus on high-ranking tracts, however, and so does not, and is not intended to, address the winner's curse. Hence, we do not report results for this approach.

[18] We base our estimates in this setting on the public Opportunity Atlas estimates and standard errors since we do not have access to the underlying microdata. We also do not have access to the correlation structure of the estimate across tracts. Such correlations arise from individuals who move across tracts, and there are few movers between most pairs of tracts, so we expect that these omitted correlations are small.

[19] We select the target tracts based on the un-shrunk estimates, rather than shrunk estimates as in Bergman et al. (2020). We do this because we find that selecting based on un-shrunk estimates yields slightly higher average quality for selected tracts than selecting on shrunk estimates, and because selection

Θ the set of selections from $\mathcal{T}$ containing one third of tracts, $\Theta = \{\theta \subset \mathcal{T} : |\theta| = \lfloor |\mathcal{T}|/3 \rfloor\}$. For $\hat{\mu}_t$ the estimated effect of tract t, define $X(\theta)$ as the average estimate over tracts in θ, $X(\theta) = \frac{1}{|\theta|}\sum_{t\in\theta}\hat{\mu}_t$. Target tracts are selected as $\hat{\theta} = \text{argmax}_{\theta\in\Theta} X(\theta)$, and the naive estimate for the average effect over selected tracts is $X(\hat{\theta})$. Correspondingly, for μ_t the neighborhood effect for tract t, let $\mu_X(\theta)$ be the average neighborhood effect over tracts in θ, $\mu_X(\theta) = \frac{1}{|\theta|}\sum_{t\in\theta}\mu_t$. We are interested in the difference between the average quality of selected tracts and the average over all tracts in the same commuting zone, $\mu_Y(\hat{\theta}) = \frac{1}{|\hat{\theta}|}\sum_{t\in\hat{\theta}}\mu_t - \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\mu_t$, and so define $Y(\theta) = \frac{1}{|\theta|}\sum_{t\in\theta}\hat{\mu}_t - \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\hat{\mu}_t$.[20] We study the performance of naive estimates and confidence intervals, empirical Bayes estimates and credible sets, and our corrected estimates and confidence intervals.

Figure 6 reports results based on ten thousand simulation draws. Panel (a) plots the average true upward mobility for selected tracts $E\left[\frac{1}{|\hat{\theta}|}\sum_{t\in\hat{\theta}}\mu_t\right]$ less the average over all tracts in the same CZ $\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\mu_t$, $E\left[\mu_Y(\hat{\theta})\right]$, across the 50 CZs considered. Selected tracts are better than average across all 50 CZs, though the precise degree of improvement varies. Panel (b) shows median bias for the estimators we consider, where the quantity of interest is again the difference between average upward mobility for selected tracts, less the average over all tracts in the same CZ. As expected the naive estimator is biased upwards, while the sign of the bias for empirical Bayes differs across CZs. The conditional estimator is median unbiased up to simulation error, while the hybrid estimator is very close to median unbiased. Panel (c) plots the median absolute estimation error across the four estimators. The naive estimator has the largest median absolute estimation error in most CZs, while the empirical Bayes typically has the smallest. The conditional and hybrid estimators are in the middle, with quite similar median absolute estimation errors for this application. Finally, panels (d) and (e) plot the coverage and median length of confidence intervals. We see that the naive confidence interval severely under-covers, with coverage close to zero in all 50 CZs. The coverage of empirical Bayes intervals differs widely across CZs,

based on shrunk estimates introduces nonlinearity (due to estimation of the hyperparameters) which complicates conditional and hybrid inference.

[20] Under the neighborhood effects model, $\mu_Y(\hat{\theta})$ corresponds to the average effect of moving a household from a randomly selected tract in the CZ to a randomly selected target tract. This need not correspond to the average treatment effect from the experiment in Bergman et al. (2020), since treatment and control households are not in general uniformly distributed across these sets of tracts. Indeed, some treated households settle in non-target tracts, and some control households settle in target tracts. Given realized location choices for treatment and control households, one could re-define $\mu_Y(\hat{\theta})$ accordingly. We do not pursue this extension, however, as data on location choice under treatment only exists for the Seattle CZ, where Bergman et al. (2020) conducted their experiment. A previous version of the paper applied an incorrect scaling when calculating $\mu_Y(\hat{\theta})$. We thank Magne Mogstad and Larry Katz flagging scaling errors.

ranging from less than 1% to over 90%.[21] Conditional confidence intervals have coverage equal to 95% up to simulation error in all CZs, while the hybrid intervals have coverage very close to 95%, and below 96%, in all cases. Finally, projection intervals have coverage equal to 100%, up to simulation error, in all CZs. Turning to median length, we see that hybrid intervals are longer than empirical Bayes and naive confidence intervals, but are considerably shorter than conditional and projection intervals in many cases.

Empirical Results Figure 7 plots results for the Opportunity Atlas data. As in the simulations we select the top third of census tracts in each CZ based on the naive estimates and then report naive, empirical Bayes, and hybrid estimates and intervals, as well as projection intervals, for the average upward mobility across selected tracts, less the average over the commuting zone. For visibility, we defer results for conditional intervals to Figure 8 of Appendix E. From these results, we see that both the empirical Bayes and hybrid adjustments shift the naive estimates and intervals downward. There is not a clear pattern to the shifts in the estimates: in some cases the empirical Bayes estimate is below the hybrid, while in other cases the order is reversed. As expected given our simulation results, the (coverage-maintaining) hybrid intervals are wider than the (under-covering) empirical Bayes intervals, but considerably shorter than projection intervals. Hybrid and projection intervals exclude zero in all CZs, suggesting that, under the hypothesis of neighborhood effects, there is real scope for selecting better neighborhoods based on the Opportunity Atlas, albeit less than the naive estimates suggest.[22]

The results for conditional procedures in Figure 8 of Appendix E are qualitatively similar, but the width of the conditional intervals is extremely variable across CZs. Specifically, while conditional intervals are quite similar to hybrid intervals in some CZs, they are much

[21] If one selects target tracts based on the empirical Bayes, rather than naive, estimates, this reduces the bias of EB estimates, with the average median bias across the 50 CZs falling from approximately -0.0023 to approximately -0.0001. Empirical Bayes credible sets continue to under-cover, however, with average coverage rising from approximately 50% to approximately 59%.

[22] It is useful to compare our results with those of Mogstad et al. (2020), who study the problem of inference on ranks and consider the Opportunity Atlas data for Seattle as an example. They show that if one forms simultaneous confidence sets for individual tracts, one can say very little about which tracts are best. Hence, we can say little about the effect of moving an individual from an arbitrary non-target tract to arbitrary target tract, and can likewise say little about the average treatment effect of shifting households from one group of tracts to the other if we allow arbitrary location choices within each group of tracts. We consider a complementary exercise, inference on the average quality of selected sets of tracts, corresponding to an average treatment effect under uniformly distributed location choices. For this problem, we find strong evidence that selected tracts are, as a group, better than average. These exercises answer different questions, and the more positive results obtained in our case reflect that it is statistically easier to distinguish average mobility across groups of selected tracts than it is to rank individual tracts.

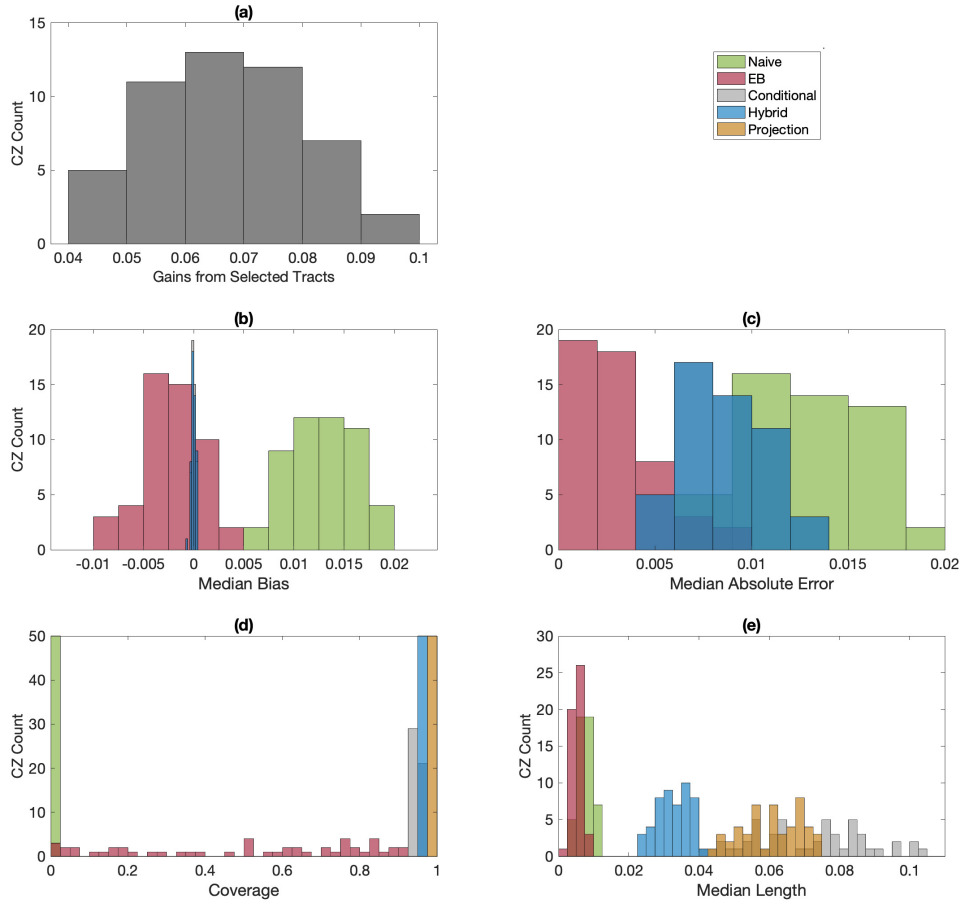

Figure 6: Simulation results from calibration to Chetty et al. (2018) Opportunity Atlas. Panel (a) shows the distribution of average improvement in economic mobility in selected tracts, relative to within-CZ average, across 50 largest CZs. An effect size of 0.01 corresponds to predicted household income rank in adulthood increasing by 1 percentile (for a child spending their entire childhood in a given tract). Panel (b) shows the median bias of different estimators across the 50 CZs. Panel (c) plots the median absolute error across the same CZs. Panel (d) shows coverage of confidence intervals across the 50 largest CZs, while panel (e) plots their median length.

longer in others.[23] Conditional intervals lie above zero in 20 of the 50 commuting zones, but include zero in the other 30. Hence, if we are satisfied with unconditional coverage we find

[23]Interestingly, the hybrid interval for Seattle, the site of Bergman et al. (2020)'s experiment, is very short, and substantially shifted downwards relative to the naive and empirical Bayes intervals. This reflects that fact that the "next best" tracts in this case are extremely close to some of the included tracts. This leads to very strong downward correction by the conditional approach, and a hybrid interval concentrated near the lower bound of the projection interval CS_P^β.

strong evidence that selected tracts are better than average, while if we demand conditional coverage results are more mixed, and depend on which commuting zone we consider.

Precision aside, unconditional inference seems potentially more natural than conditional inference in this application. Within a given CZ, our primary interest is in the efficacy of targeting the top third of tracts based on estimated economic mobility, rather than in the precise combination of tracts selected. Hence, it is natural to focus on unconditional approaches, which give guarantees on average across potential selections. Moreover, in this application we consider multiple CZs, putting the focus even more squarely on average results from targeting tracts in this way, rather than the particular selection in a given CZ.

8 Conclusion

This paper considers a form of the winner's curse that arises when we select a target parameter for inference based on optimization. We propose confidence intervals and quantile unbiased estimators for the target parameter that are optimal conditional on its selection. We hence recommend our conditional inference procedures when it is appropriate to remove uncertainty about the choice of target parameters from inferential statements. These conditionally valid procedures are also unconditionally valid, but we find that they sometimes have unappealing (unconditional) performance relative to existing alternatives. If one is satisfied with correct unconditional coverage and (in the case of estimation) a small, controlled degree of bias, we propose hybrid procedures which combine conditioning with projection confidence intervals.

Our results suggest a range of opportunities for future work. First, rather than considering inference on $\mu_Y(\hat{\theta})$, under suitable assumptions one could build on our results to forecast $Y(\hat{\theta})$. Alternatively, while conditional and projection confidence intervals have antecedents in the literature on inference after model selection, including in Berk et al. (2013) and Fithian et al. (2017), there is no analog of our hybrid approach in this literature. Our positive simulation results for the hybrid method suggest that this approach might yield appealing performance in a range of post-selection-inference settings. Even if a fully conditional approach is desired in the post-selection problem, as in Fithian et al. (2017), one could consider the analog of our optimal median-unbiased estimates that condition on the selected model. Finally, the problem of estimating the value of a dynamic treatment rule (c.f. Chakraborty and Murphy, 2014; Han, 2020) is closely related to our setting, so it seems likely that our results could prove useful there as well.

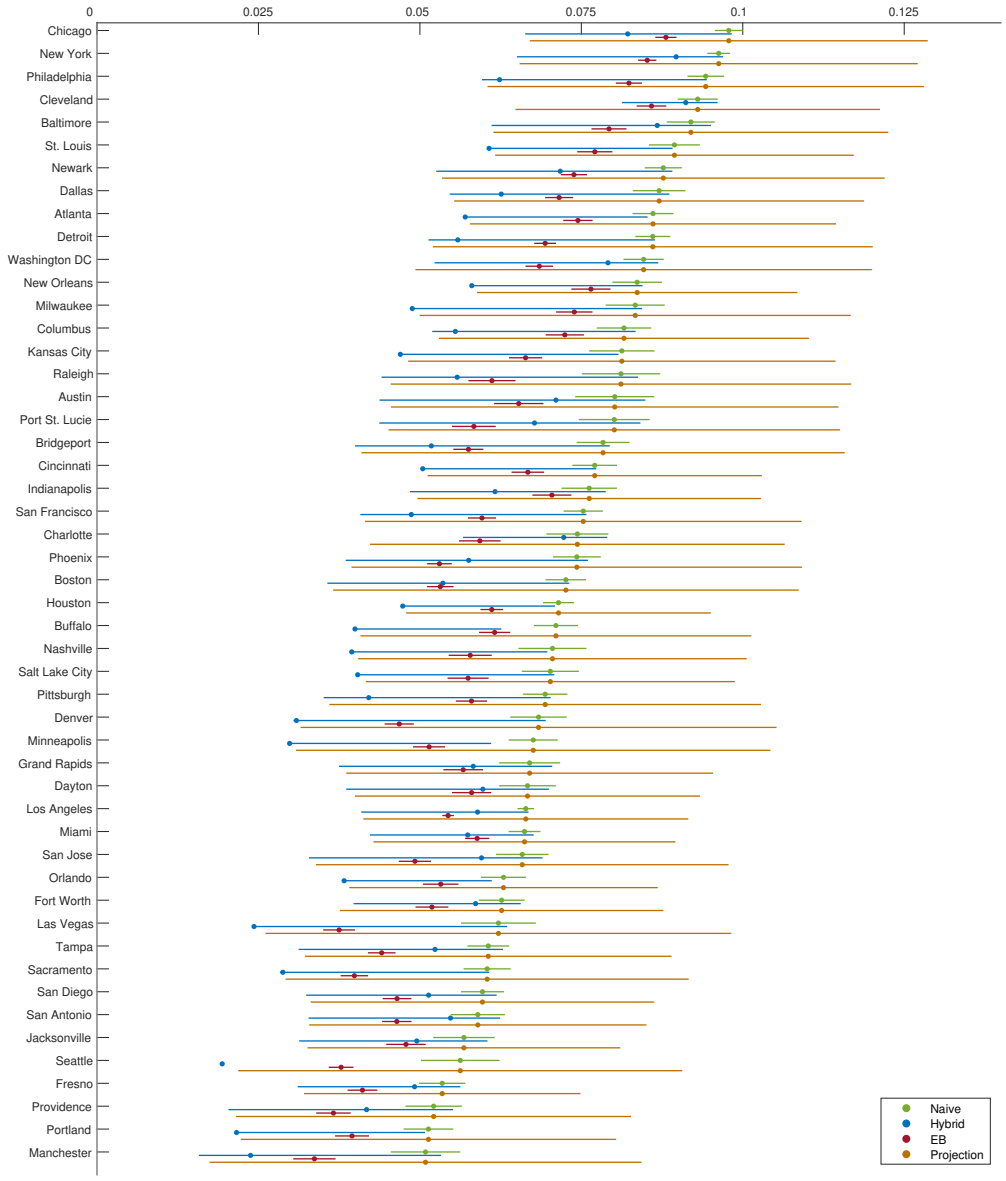

Figure 7: Estimates and confidence intervals for average economic mobility for selected census tracts based on Chetty et al. (2018) Opportunity Atlas, relative to the within-CZ average. CZs are ordered by the magnitude of the naive estimate.

References

Andrews, D., Cheng, X., and Guggenberger, P. (2020a). Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics*, Forthcoming.

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

Andrews, I., Roth, J., and Pakes, A. (2019). Inference for linear conditional moment inequalities. Unpublished Manuscript.

Armstrong, T., Kolesar, M., and Plagborg-Moller, M. (2020). Robust empirical bayes confidence intervals. Unpublished Manuscript.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *Annals of Statistics*, 48(1):440–463.

Baird, S., Hicks, J. H., Kremer, M., and Miguel, E. (2016). Worms at work: Long-run impacts of a child health investment. *The Quarterly Journal of Economics*, 131(4):1637–1680.

Banerjee, A., Hanna, R., Kyle, J., Olken, B. A., and Sumarto, S. (2018). Tangible information and citizen empowerment: Identification cards and food subsidy programs in Indonesia. *Journal of Political Economy*, 126(2):451–491.

Bergman, P., Chetty, R., DeLuca, S., Hendren, N., Katz, L. F., and Palmer, C. (2020). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. Unpublished Manuscript.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41(2):802–831.

Björkman Nyqvist, M. and Jayachandran, S. (2017). Mothers care more, but fathers decide: Educating parents about child health in Uganda. *American Economic Review*, 107(5):496–500.

Chakraborty, B. and Murphy, S. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Applications*, 1:447–64.

Chernozhukov, V., Lee, S., and Rosen, A. (2013). Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737.

Chetty, R., Friedman, J., Hendren, N., Jones, M. R., and Porter, S. R. (2018). The opportunity atlas: Mapping the childhood roots of social mobility. Unpublished Manuscript.

Chetty, R. and Hendren, N. (2016a). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *Quarterly Journal of Economics*, 133(3):1107–1162.

Chetty, R. and Hendren, N. (2016b). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *Quarterly Journal of Economics*, 133(3):1163–1228.

Cohen, A. and Sackrowitz, H. B. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters*, 8:273–278.

Cox, G. (2018). Almost sure uniqueness of a global minimum without convexity. *Annals of Statistics*, 48(1):584–606.

Dawid, A. P. (1994). Selection paradoxes in bayesian inference. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 24:211–220.

Duflo, E., Greenstone, M., Pande, R., and Ryan, N. (2018). The value of regulatory discretion: Estimates from environmental inspections in India. *Econometrica*, 86(6):2123–2160.

Eliasz, P. (2004). Optimal median unbiased estimation of coefficients on highly persistent regressors. Unpublished Manuscript.

Ferguson, J. P., Cho, J. H., Yang, C., and Zhao, H. (2013). Empirical bayes correction for the winner's curse in genetic association studies. *Genetic Epidemiology*, 37(1):60–68.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection. *arXiv*.

Frandsen, B. (2020). A rational approach to inference on multiple parameters. Unpublished Manuscript.

Freyberger, J. and Rai, Y. (2018). Uniform confidence bands: Characterization and optimality. *Journal of Econometrics*, 1(204):119–130.

Guo, X. and He, X. (2021). Inference on selected subgroups in clinical trials. Forthcoming in the Journal of the American Statistical Association.

Han, S. (2020). Identification in nonparametric models for dynamic treatment effects. *Journal of Econometrics*, Forthcoming.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380.

Harris, X. T., Panigrahi, S., Markovic, J., Bi, N., and Taylor, J. (2016). Selective sampling after solving a convex problem. *arXiv*.

Karlan, D. and List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793.

Khan, A. Q., Khwaja, A. I., and Olken, B. A. (2016). Tax farming redux: Experimental evidence on performance pay for tax collectors. *The Quarterly Journal of Economics*, 131(1):219–271.

Kitagawa, T. and Tetenov, A. (2018a). Supplement to "who should be treated? empirical welfare maximization methods for treatment choice". *Econometrica*.

Kitagawa, T. and Tetenov, A. (2018b). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

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

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

Laber, E. and Murphy, S. (2011). Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 106(495):904–913.

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(3):907–927.

Lee, M. R. and Shen, M. (2018). Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *KDD*.

Lehmann, E. and Scheffé, H. (1955). Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics*, 15(3):219–236.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, third edition.

McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics*, 200(1):17–35.

McCloskey, A. (2020). Hybrid confidence intervals for informative uniform asymptotic inference after model selection. Working Paper.

Mogstad, M., Romano, J. P., Shaikh, A., and Wilhelm, D. (2020). Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. Unpublished Manuscript.

Pfanzagl, J. (1979). On optimal median unbiased estimators in the presence of nuisance parameters. *Annals of Statistics*, 7(1):187–193.

Pfanzagl, J. (1994). *Parametric Statistical Theory*. De Gruyter.

Rai, Y. (2018). Statistical inference for treatment assignment policies. Unpublished Manuscript.

Romano, J. P., Shaikh, A., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Sampson, A. R. and Sill, M. W. (2005). Drop-the-losers design: normal case. *Biometrical Journal*, 47:257–268.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Annals of Statistics*, 46(2):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(3):1255–1287.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Xu, L., Craiu, R. V., and Sun, L. (2011). Bayesian methods to overcome the winner's curse in genetic studies. *Annals of Applied Statistics*, 5(201-231).

Zhong, H. and Prentice, R. (2009). Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology*, 34(1):78–91.

Supplement to the paper

Inference on Winners

Isaiah Andrews Toru Kitagawa Adam McCloskey

December 22, 2021

This supplement contains proofs and additional results for the paper "Inference on Winners." Section A generalizes the conditional inference results discussed in the main text, introducing additional conditioning variables and unbiased confidence intervals. Section B proves our results for the finite-sample normal model. Section C constructs procedures that dominate conventional sample splitting as discussed in Section 3.3 of the paper. Section D proves the uniform asymptotic results discussed in the main text. Section E provides additional results and discussion to complement the application in Section 7 of the main text. Finally, Section F reports additional simulation results for the stylized example of Section 2 of the paper.

A Conditional Inference

This section extends the conditional inference results developed in Section 3 of the main text in two directions, first allowing dependence on additional conditioning variables, and then introducing uniformly most accurate unbiased confidence intervals.

A.1 Additional Conditioning Events

Suppose that in addition to conditioning on $\{\hat{\theta}=\tilde{\theta}\}$, we also want to condition on an additional event $\{\hat{\gamma}=\tilde{\gamma}\}$, for $\hat{\gamma}=\gamma(X)$ some function of X. We thus seek estimators that are quantile-unbiased conditional on $(\hat{\theta},\hat{\gamma})$,

$$Pr_{\mu}\Big\{\hat{\mu}_{\alpha}\geq\mu_{Y}(\hat{\theta})|\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\Big\}=\alpha \text{ for all } \tilde{\theta}\in\Theta,\ \tilde{\gamma}\in\Gamma, \text{ and all } \mu, \tag{12}$$

and confidence sets with correct conditional coverage

$$Pr_{\mu}\Big\{\mu_{Y}(\hat{\theta})\in CS|\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\Big\}\geq 1-\alpha \text{ for all } \tilde{\theta}\in\Theta,\ \tilde{\gamma}\in\Gamma, \text{ and all } \mu.$$

As in the main text, we re-write the conditioning event in terms of the sample space of X as $\left\{X:\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}=\mathcal{X}(\tilde{\theta},\tilde{\gamma})$, and study the conditional distribution of $(X,Y(\tilde{\theta}))$ given $X\in\mathcal{X}(\tilde{\theta},\tilde{\gamma})$. For $Z_{\tilde{\theta}}$ as defined in (7), let

$$\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)=\left\{y:z+\left(\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta})\right)y\in\mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}.$$

Conditional on $\hat{\theta}=\tilde{\theta}$, $\hat{\gamma}=\tilde{\gamma}$, and $Z_{\tilde{\theta}}=z$, $Y(\hat{\theta})$ again follows a one-dimensional normal distribution $N(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta}))$ truncated to $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$.

To characterize $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$, note that for for $\mathcal{X}(\tilde{\theta})$ as derived in the main text, we can write $\mathcal{X}(\tilde{\theta},\tilde{\gamma})=\mathcal{X}(\tilde{\theta})\cap\mathcal{X}_\gamma(\tilde{\gamma})$. Likewise, for $\mathcal{Y}_\gamma(\tilde{\gamma},z)$ defined analogously to (9), $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)=\mathcal{Y}(\tilde{\theta},z)\cap\mathcal{Y}_\gamma(\tilde{\gamma},z)$. The form of $\mathcal{X}_\gamma(\tilde{\gamma})$ and $\mathcal{Y}_\gamma(\tilde{\gamma},z)$ depends on the conditioning variables $\hat{\gamma}$ considered.

To construct quantile-unbiased estimators, let $F_{TN}(y;\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},z)$ denote the distribution function for a $N(\mu_Y(\tilde{\theta},\Sigma_Y(\tilde{\theta})))$ distribution truncated to $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$. This function is strictly decreasing in $\mu_Y(\tilde{\theta})$, so define $\hat{\mu}_\alpha$ as the unique solution to

$$F_{TN}(Y(\hat{\theta});\hat{\mu}_\alpha,\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})=1-\alpha. \tag{13}$$

Proposition 7

Let $\hat{\mu}_\alpha$ solve (13). $\hat{\mu}_\alpha$ is conditionally α-quantile-unbiased in the sense of (12). If Assumption 1 holds, then $\hat{\mu}_\alpha$ is the uniformly most concentrated α-quantile-unbiased estimator in that for any other conditionally α-quantile-unbiased estimator $\hat{\mu}^*_\alpha$ and any loss function $L\left(d,\mu_Y(\tilde{\theta})\right)$ that attains its minimum at $d=\mu_Y(\tilde{\theta})$ and is quasiconvex in d for all $\mu_Y(\tilde{\theta})$,

$$E_\mu\left[L\left(\hat{\mu}_\alpha,\mu_Y(\tilde{\theta})\right)|\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right]\leq E_\mu\left[L\left(\hat{\mu}^*_\alpha,\mu_Y(\tilde{\theta})\right)|\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right]$$

for all μ and all $\tilde{\theta}\in\Theta$, $\tilde{\gamma}\in\Gamma$.

Proposition 7 shows that optimality of $\hat{\mu}_\alpha$ extends to general conditioning variables $\hat{\gamma}$. Hence, $\hat{\mu}_{\frac{1}{2}}$ is an optimal median-unbiased estimator, while $CS_{ET}=[\hat{\mu}_{\frac{\alpha}{2}},\hat{\mu}_{1-\frac{1}{2}}]$ is an optimal equal-tailed confidence interval.

A.1.1 Additional Conditioining Example: Outperforming a Benchmark

Here, we derive the truncation set $\mathcal{Y}_\gamma(\tilde{\gamma},z)$ corresponding to the additional selection event discussed as an example in Section 3 of the main text. We would like to condition on rejection of the null hypothesis that no treatment outperforms a benchmark, $H_0:\max_{\theta\in\Theta}\mu_X(\theta)\leq\mu_X(0)$. For $X(0)$ the observed outcome for the under the benchmark

treatment, a non-studentized test of the null hypothesis rejects when the difference between the empirical outcome at the best treatment and the benchmark exceeds a critical value threshold: $X(\hat{\theta})-X(0)\geq c$ for some $c>0$. In this case, $\hat{\gamma}=\gamma(X)=1\Big\{X(\hat{\theta})-X(0)\geq c\Big\}$ and we wish to condition inference on rejection of H_0 for which $\hat{\gamma}=1$.[24] In order to implement our procedures in this context, we need a tractable expression for $\mathcal{Y}_\gamma(1,z)$.

Note that we can write

$$\Big\{X(\tilde{\theta})-X(0)\geq c\Big\}=\left\{Z_{\tilde{\theta}}(\tilde{\theta})-Z_{\tilde{\theta}}(0)+\frac{\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)}{\Sigma_Y(\tilde{\theta})}Y(\tilde{\theta})\geq c\right\}.$$

Rearranging, we see that

$$\mathcal{Y}_\gamma(1,Z_{\tilde{\theta}})=\begin{cases}\left\{y:y\geq\frac{\Sigma_Y(\tilde{\theta})\left(c-Z_{\tilde{\theta}}(\tilde{\theta})+Z_{\tilde{\theta}}(0)\right)}{\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)}\right\} & \text{if } \Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)>0\\ \left\{y:y\leq\frac{\Sigma_Y(\tilde{\theta})\left(c-Z_{\tilde{\theta}}(\tilde{\theta})+Z_{\tilde{\theta}}(0)\right)}{\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)}\right\} & \text{if } \Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)<0\\ \mathbb{R} & \text{if } \Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)=0 \\ & \quad \text{and } Z_{\tilde{\theta}}(\tilde{\theta})-Z_{\tilde{\theta}}(0)\geq c\\ \emptyset & \text{if } \Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)=0\\ & \quad \text{and } Z_{\tilde{\theta}}(\tilde{\theta})-Z_{\tilde{\theta}}(0)<c.\end{cases}.$$

Thus, if for example $\Sigma_{XY}(\tilde{\theta})>\Sigma_{XY}(\tilde{\theta},0)$, then $\mathcal{Y}(\tilde{\theta},1,z)=\mathcal{Y}(\tilde{\theta},z)\cap\mathcal{Y}_\gamma(1,z)=\Big[\mathcal{L}^*(\tilde{\theta},Z_{\tilde{\theta}}),\mathcal{U}(\tilde{\theta},Z_{\tilde{\theta}})\Big]$, where $\mathcal{U}(\tilde{\theta},Z_{\tilde{\theta}})$ is the upper bound defined in Proposition 1 while

$$\mathcal{L}^*(\tilde{\theta},Z_{\tilde{\theta}})=\max\left\{\mathcal{L}(\tilde{\theta},Z_{\tilde{\theta}}),\frac{\Sigma_Y(\tilde{\theta})\Big(c-Z_{\tilde{\theta}}(\tilde{\theta})+Z_{\tilde{\theta}}(0)\Big)}{\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)}\right\},$$

for $\mathcal{L}(\tilde{\theta},Z_{\tilde{\theta}})$ defined as in Proposition 1. Hence, when $\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)>0$, conditoning on $\hat{\gamma}=1$ simply modifies the lower bound $\mathcal{L}(\tilde{\theta},Z_{\tilde{\theta}})$. Likewise, when $\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)<0$ or $\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},0)=0$, conditioning on $\hat{\gamma}=1$ modifies $\mathcal{U}(\tilde{\theta},Z_{\tilde{\theta}})$ and $\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}})$, respectively.

As this example illustrates, it is straightforward to incorporate additional conditioning variables $\hat{\gamma}$ in the level maximization problems we study here provided one can characterize the set $\mathcal{Y}_\gamma(\tilde{\gamma},z)$. While such characterizations are easy to obtain in many cases, they depend on the conditioning variable considered and must be derived on a case-by-case basis.

[24] We could equally well consider studentized tests.

A.2 Unbiased Confidence Intervals

Rather than considering equal-tailed intervals, we can alternatively consider unbiased confidence intervals. Following Lehmann and Romano (2005), we say that a level $1-\alpha$ two-sided confidence interval CS is unbiased if its probability of covering any given false parameter value is bounded above by $1-\alpha$. Likewise, a one sided lower (upper) confidence interval is unbiased if its probability of covering a false parameter value above (below) the true value is bounded above by $1-\alpha$. Using the duality between tests and confidence intervals, a level $1-\alpha$ confidence interval CS is unbiased if and only if $\phi(\mu_{Y,0})=1\{\mu_{Y,0}\notin CS\}$ is an unbiased test for the corresponding family of hypotheses.[25] The results of Lehmann and Scheffé (1955) applied in our setting imply that optimal unbiased tests conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$ are the same as optimal unbiased tests conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma},Z_{\tilde{\theta}}=z_{\tilde{\theta}}\right\}$. These optimal tests take a simple form.

Define a size α test of the two-sided hypothesis $H_0:\mu_Y(\tilde{\theta})=\mu_{Y,0}$ as

$$\phi_{TS,\alpha}(\mu_{Y,0})=1\left\{Y(\tilde{\theta})\notin[c_l(Z_{\tilde{\theta}}),c_u(Z_{\tilde{\theta}})]\right\}$$

where $c_l(z)$, $c_u(z)$ solve

$$Pr\{\zeta\in[c_l(z),c_u(z)]\}=1-\alpha,\ E[\zeta 1\{\zeta\in[c_l(z),c_u(z)]\}]=(1-\alpha)E[\zeta]$$

for ζ that follows a truncated normal distribution

$$\zeta\sim\xi|\xi\in\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z),\quad \xi\sim N\left(\mu_{Y,0},\Sigma_Y(\tilde{\theta})\right).$$

Likewise, define a size α test of the one-sided hypothesis $H_0:\mu_Y(\tilde{\theta})\geq\mu_{Y,0}$ as

$$\phi_{OS-,\alpha}(\mu_{Y,0})=1\left\{F_{TN}(Y(\tilde{\theta});\mu_{Y,0},\tilde{\theta},\tilde{\gamma},z)\leq\alpha\right\}$$

and a test of $H_0:\mu_Y(\tilde{\theta})\leq\mu_{Y,0}$ as

$$\phi_{OS+,\alpha}(\mu_{Y,0})=1\left\{F_{TN}(Y(\tilde{\theta});\mu_{Y,0},\tilde{\theta},\tilde{\gamma},z)\geq1-\alpha\right\}.$$

Proposition 8

[25] That is, $H_0:\mu_Y(\tilde{\theta})=\mu_{Y,0}$ for a two-sided confidence interval, $H_0:\mu_Y(\tilde{\theta})\geq\mu_{Y,0}$ for a lower confidence interval and $H_0:\mu_Y(\tilde{\theta})\leq\mu_{Y,0}$ for an upper confidence interval.

If Assumption 1 holds, $\phi_{TS,\alpha}$, $\phi_{OS-,\alpha}$, and $\phi_{OS+,\alpha}$ are uniformly most powerful unbiased size α tests of their respective null hypotheses conditional on $\hat{\theta}=\tilde{\theta}$ and $\hat{\gamma}=\tilde{\gamma}$.

To form uniformly most accurate unbiased confidence intervals we collect the values not rejected by these tests. The two-sided uniformly most accurate unbiased confidence interval is $CS_U=\{\mu_{Y,0}:\phi_{TS,\alpha}(\mu_{Y,0})=0\}$. CS_U is unbiased and has conditional coverage $1-\alpha$ by construction. Likewise, we can form lower and upper one-sided uniformly most accurate unbiased confidence intervals as $CS_{U,-}=\{\mu_{Y,0}:\phi_{OS-,\alpha}(\mu_{Y,0})=0\}=(-\infty,\hat{\mu}_{1-\alpha}]$, and $CS_{U,+}=\{\mu_{Y,0}:\phi_{OS+,\alpha}(\mu_{Y,0})=0\}=[\hat{\mu}_{\alpha},\infty)$, respectively. Hence, we can view CS_{ET} as the intersection of level $1-\frac{\alpha}{2}$ uniformly most accurate unbiased upper and lower confidence intervals. Unfortunately, no such simplification is generally available for CS_U, though Lemma 5.5.1 of Lehmann and Romano (2005) guarantees that this set is an interval.

A.3 Behavior When $Pr_{\mu}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$ is Large

In Proposition 3 of the main text, we showed that our median-unbiased estimators and equal-tailed confidence intervals converge to conventional ones when $Pr_{\mu}\left\{\hat{\theta}=\tilde{\theta}\right\}\to 1$. The same result holds for general conditioning events and unbiased confidence intervals.

Lemma 2
Consider any sequence of values $\mu_{Y,m}$ and $z_{\tilde{\theta},m}$ such that $Pr_{\mu_{Y,m}}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\right\}\to 1$. Then under $\mu_{Y,m}$, conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma},Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\right\}$ we have $CS_U\to_p CS_N$, $CS_{ET}\to_p CS_N$, and $\hat{\mu}_{\frac{1}{2}}\to_p Y(\tilde{\theta})$.

Proposition 9
Consider any sequence of values μ_m such that $Pr_{\mu_m}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}\to 1$. Then under μ_m, we have $CS_U\to_p CS_N$, $CS_{ET}\to_p CS_N$, and $\hat{\mu}_{\frac{1}{2}}\to_p Y(\tilde{\theta})$ both conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$ and unconditionally.

B Proofs

We first prove the results stated in Section A, and then build on these to prove the results for the finite-sample normal model discussed in the main text.

B.1 Proofs for Results in Section A

Proof of Proposition 7 For ease of reference, let us abbreviate $(Y(\tilde{\theta}),\mu_Y(\tilde{\theta}),Z_{\tilde{\theta}})$ by $(\tilde{Y},\tilde{\mu}_Y,\tilde{Z})$. Let $Y(-\tilde{\theta})$ collect the elements of Y other than $Y(\tilde{\theta})$ and define $\mu_Y(-\theta)$

analagously. Let

$$Y^* = Y(-\tilde{\theta}) - Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix},$$

$$\mu_Y^* = \mu_Y(-\tilde{\theta}) - Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ \begin{pmatrix} \tilde{\mu}_Y \\ \mu_X \end{pmatrix},$$

and $\tilde{\mu}_Z = \mu_X - \left(\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta})\right)\mu_Y$. Here we use A^+ to denote the Moore-Penrose pseudoinverse of a matrix A. Note that $(\tilde{Z},\tilde{Y},Y^*)$ is a one-to-one transformation of (X,Y), and thus that observing $(\tilde{Z},\tilde{Y},Y^*)$ is equivalent to observing (X,Y). Likewise, $(\tilde{\mu}_Z,\tilde{\mu}_Y,\mu_Y^*)$ is a one-to-one linear transformation of (μ_X,μ_Y), and if the set of possible values for the latter contains an open set, that for the former does as well (relative to the appropriate linear subspace).

Note, next, that since $(\tilde{Z},\tilde{Y},Y^*)$ is a linear transformation of (X,Y), $(\tilde{Z},\tilde{Y},Y^*)$ is jointly normal (with a potentially degenerate distribution). Note next that $(\tilde{Z},\tilde{Y},Y^*)$ are mutually uncorrelated, and thus independent. That $\tilde{Z}$ and $\tilde{Y}$ are uncorrelated is straightforward to verify. To show that Y^* is likewise uncorrelated with the other elements, note that we can write $Cov\left(Y^*,(\tilde{Y},X')'\right)$ as

$$Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) - Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right).$$

For $V\Lambda V'$ an eigendecomposition of $Var\left((\tilde{Y},X')'\right)$ (so $VV'=I$), note that we can write

$$Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) = VDV'$$

for D a diagonal matrix with ones in the entries corresponding to the nonzero entries of Λ and zeros everywhere else. For any column v of V corresponding to a zero entry of D, $v'Var\left(\left(\tilde{Y},X'\right)'\right)v=0$, so the Cauchy-Schwarz inequality implies that

$$Cov\left(Y\left(-\tilde{\theta}\right), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)v = 0.$$

Thus,

$$Cov\left(Y\left(-\tilde{\theta}\right),\begin{pmatrix}\tilde{Y}\\ X\end{pmatrix}\right)VDV' = Cov\left(Y\left(-\tilde{\theta}\right),\begin{pmatrix}\tilde{Y}\\ X\end{pmatrix}\right)VV' = Cov\left(Y\left(-\tilde{\theta}\right),\begin{pmatrix}\tilde{Y}\\ X\end{pmatrix}\right),$$

so Y^* is uncorrelated with $\left(\tilde{Y},X'\right)'$.

Using independence, the joint density of $(\tilde{Z},\tilde{Y},Y^*)$ absent truncation is given by

$$f_{N,\tilde{Z}}(\tilde{z};\tilde{\mu}_Z)f_{N,\tilde{Y}}(\tilde{y};\tilde{\mu}_Y)f_{N,Y^*}(\tilde{y}^*;\mu_Y^*)$$

for f_N normal densities with respect to potentially degenerate base measures:

$$f_{N,\tilde{Z}}(\tilde{z};\tilde{\mu}_Z) = \tilde{\det}(2\pi\Sigma_{\tilde{Z}})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\tilde{z}-\tilde{\mu}_Z)'\Sigma^+_{\tilde{Z}}(\tilde{z}-\tilde{\mu}_Z)\right)$$

$$f_{N,\tilde{Y}}(\tilde{y};\tilde{\mu}_Y) = (2\pi\Sigma_{\tilde{Y}})^{-\frac{1}{2}}\exp\left(-\frac{(\tilde{y}-\tilde{\mu}_Y)^2}{2\Sigma_{\tilde{Y}}}\right)$$

$$f_{N,Y^*}(y^*;\mu_Y^*) = \tilde{\det}(2\pi\Sigma_{Y^*})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y^*-\tilde{\mu}_Y^*)'\Sigma^+_{Y^*}(y^*-\mu_Y^*)\right),$$

where $\tilde{\det}(A)$ denotes the pseudodeterminant of a matrix A, $\Sigma_{\tilde{Z}} = Var(\tilde{Z})$, $\Sigma_{\tilde{Y}} = \Sigma_Y(\tilde{\theta})$, and $\Sigma_{Y^*} = Var(Y^*)$.

The event $\left\{X \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}$ depends only on $(\tilde{Z},\tilde{Y})$ since it can be expressed as

$$\left\{\left(\tilde{Z}+\frac{\Sigma_{XY}(\cdot,\tilde{\theta})}{\Sigma_Y(\tilde{\theta})}\tilde{Y}\right) \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\},$$

so conditional on this event Y^* remains independent of $(\tilde{Z},\tilde{Y})$. In particular, we can write the joint density conditional on $\left\{X \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}$ as

$$\frac{1\left\{\left(\tilde{z}+\Sigma_{XY}(\cdot,\tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1}\tilde{y}\right) \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}}{Pr_{\tilde{\mu}_Z,\tilde{\mu}_Y}\left\{X \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}} f_{N,\tilde{Z}}(\tilde{z};\tilde{\mu}_Z)f_{N,\tilde{Y}}(\tilde{y};\tilde{\mu}_Y)f_{N,Y^*}(\tilde{y}^*;\mu_Y^*). \tag{14}$$

The density (14) has the same structure as (5.5.14) of Pfanzagl (1994), and satisfies properties (5.5.1)-(5.5.3) of Pfanzagl (1994) as well. Part 1 of the proposition then follows immedi-

ately from Theorem 5.5.9 of Pfanzagl (1994). Part 2 of the proposition follows by using Theorem 5.5.9 of Pfanzagl (1994) to verify the conditions of Theorem 5.5.15 of Pfanzagl (1994). □

Proof of Proposition 8 In the proof of Proposition 7, we showed that the joint density of $(\tilde{Z},\tilde{Y},Y^*)$ (defined in that proof) has the exponential family structure assumed in equation 4.10 of Lehmann and Romano (2005). Moreover, Assumption 1 implies that the parameter space for (μ_X,μ_Y) is convex and is not contained in any proper linear subspace. Thus, the parameter space for $(\tilde{\mu}_Z,\tilde{\mu}_Y,\mu_Y^*)$ inherits the same property, and satisfies the conditions of Theorem 4.4.1 of Lehmann and Romano (2005). The result follows immediately. □

Proof of Lemma 2 Recall that conditional on $Z_{\tilde{\theta}}=z_{\tilde{\theta}}$, $\hat{\theta}=\tilde{\theta}$ and $\hat{\gamma}=\tilde{\gamma}$ if and only if $Y(\tilde{\theta})\in\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z_{\tilde{\theta}})$. Hence, the assumption of the lemma implies that

$$Pr_{\mu_{Y,m}}\left\{Y(\tilde{\theta})\in\mathcal{Y}(\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})|Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\right\}\to 1.$$

Note, next, that both the conventional and conditional confidence intervals are equivariant under shifts, in the sense that the conditional confidence interval for $\mu_Y(\tilde{\theta})$ based on observing $Y(\tilde{\theta})$ conditional on $Y(\tilde{\theta})\in\mathcal{Y}(\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})$ is equal to the conditional confidence interval for $\mu_Y(\tilde{\theta})$ based on observing $Y(\tilde{\theta})-\mu_Y^*(\tilde{\theta})$ conditional on $Y(\tilde{\theta})-\mu_Y^*(\tilde{\theta})\in\mathcal{Y}(\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})-\mu_Y^*(\tilde{\theta})$ for any constant $\mu_Y^*(\tilde{\theta})$. Hence, rather than considering a sequence of values $\mu_{Y,m}$, we can fix some μ_Y^* and note that $Pr_{\mu_Y^*}\left\{Y(\tilde{\theta})\in\mathcal{Y}_m^*|Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\right\}\to 1$, where $\mathcal{Y}_m^*=\mathcal{Y}(\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})-\mu_{Y,m}(\tilde{\theta})+\mu_Y^*(\tilde{\theta})$. Confidence intervals for $\mu_{Y,m}(\tilde{\theta})$ in the original problem are equal to those for $\mu_Y^*(\tilde{\theta})$ in the new problem, shifted by $\mu_{Y,m}(\tilde{\theta})-\mu_Y^*(\tilde{\theta})$. Hence, to prove the result it suffices to prove the equivalence of conditional and conventional confidence intervals in the problem with μ_Y fixed (and likewise for estimators).

To prove the result, we make use of the following lemma, which is proved below. First, we must introduce the following notation. Let $(c_{l,ET}(\mu_{Y,0},\mathcal{Y}),c_{u,ET}(\mu_{Y,0},\mathcal{Y}))$ denote the critical values for an equal-tailed test of $H_0:\mu_Y(\tilde{\theta})=\mu_{Y,0}$ for $Y(\tilde{\theta})\sim N\left(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta})\right)$ conditional on $Y(\tilde{\theta})\in\mathcal{Y}$. That is, $(c_{l,ET}(\mu_{Y,0},\mathcal{Y}),c_{u,ET}(\mu_{Y,0},\mathcal{Y}))$ solve $F_{TN}(c_{l,ET}(\mu_{Y,0},\mathcal{Y});\mu_{Y,0},\mathcal{Y})=\frac{\alpha}{2}$ and $F_{TN}(c_{u,ET}(\mu_{Y,0},\mathcal{Y});\mu_{Y,0},\mathcal{Y})=1-\frac{\alpha}{2}$, where $F_{TN}(\cdot;\mu_{Y,0},\mathcal{Y})$ is the distribution function for the normal distribution $N\left(\mu_{Y,0},\Sigma_Y(\tilde{\theta})\right)$ truncated to $\mathcal{Y}$. Similarly, let $(c_{l,U}(\mu_{Y,0},\mathcal{Y}),c_{u,U}(\mu_{Y,0},\mathcal{Y}))$ denote the critical values for the corresponding unbiased test. That is, $(c_{l,U}(\mu_{Y,0},\mathcal{Y}),c_{u,U}(\mu_{Y,0},\mathcal{Y}))$ solve $Pr\{\zeta\in[c_{l,U}(\mu_{Y,0},\mathcal{Y}),c_{u,U}(\mu_{Y,0},\mathcal{Y})]\}=1-\alpha$ and $E\left[\zeta 1\{\zeta\in[c_{l,U}(\mu_{Y,0},\mathcal{Y}),c_{u,U}(\mu_{Y,0},\mathcal{Y})]\}\right]=(1-\alpha)E\left[\zeta\right]$ for $\zeta\sim\xi|\xi\in\mathcal{Y}$ where $\xi\sim N\left(\mu_{Y,0},\Sigma_Y(\tilde{\theta})\right)$.

Lemma 3

Suppose that we observe $Y(\tilde{\theta}) \sim N\left(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta})\right)$ *conditional on* $Y(\tilde{\theta})$ *falling in a set* $\mathcal{Y}$. *If we hold* $\left(\Sigma_Y(\tilde{\theta}),\mu_{Y,0}\right)$ *fixed and consider a sequence of sets* $\mathcal{Y}_m$ *such that* $Pr\left\{Y(\tilde{\theta}) \in \mathcal{Y}_m\right\} \to 1$, *we have that for*

$$\phi_{ET}(\mu_{Y,0}) = 1\left\{Y(\tilde{\theta}) \notin [c_{l,ET}(\mu_{Y,0},\mathcal{Y}_m),c_{u,ET}(\mu_{Y,0},\mathcal{Y}_m)]\right\} \tag{15}$$

and

$$\phi_U(\mu_{Y,0}) = 1\left\{Y(\tilde{\theta}) \notin [c_{l,U}(\mu_{Y,0},\mathcal{Y}_m),c_{u,U}(\mu_{Y,0},\mathcal{Y}_m)]\right\}, \tag{16}$$

$$(c_{l,ET}(\mu_{Y,0},\mathcal{Y}_m),c_{u,ET}(\mu_{Y,0},\mathcal{Y}_m)) \to \left(\mu_{Y,0} - c_{\frac{\alpha}{2},N}\sqrt{\Sigma_Y(\tilde{\theta})},\mu_{Y,0} + c_{\frac{\alpha}{2},N}\sqrt{\Sigma_Y(\tilde{\theta})}\right)$$

and

$$(c_{l,U}(\mu_{Y,0},\mathcal{Y}_m),c_{u,U}(\mu_{Y,0},\mathcal{Y}_m)) \to \left(\mu_{Y,0} - c_{\frac{\alpha}{2},N}\sqrt{\Sigma_Y(\tilde{\theta})},\mu_{Y,0} + c_{\frac{\alpha}{2},N}\sqrt{\Sigma_Y(\tilde{\theta})}\right).$$

To complete the proof, first note that CS_{ET} and CS_U are formed by inverting (families of) equal-tailed and unbiased tests, respectively. Let CS_m denote a generic conditional confidence interval formed by inverting a family of tests

$$\phi_m(\mu_{Y,0}) = 1\left\{Y(\tilde{\theta}) \notin [c_l(\mu_{Y,0},\mathcal{Y}_m^*),c_u(\mu_{Y,0},\mathcal{Y}_m^*)]\right\}.$$

Hence, we want to show that

$$CS_m \to_p \left[Y(\tilde{\theta}) - c_{\frac{\alpha}{2},N}, Y(\tilde{\theta}) + c_{\frac{\alpha}{2},N}\right], \tag{17}$$

as $m \to \infty$, for CS_m formed by inverting either (15) or (16).

We note that CS_m is a finite interval for all m, which holds trivially for the equal-tailed confidence interval CS_{ET}, and holds for C_U by Lemma 5.5.1 of Lehmann and Romano (2005). For each value $\mu_{Y,0}$ our Lemma 3 implies that

$$\phi_m(\mu_{Y,0}) \to_p 1\left\{Y\left(\tilde{\theta}\right) \notin \left[\mu_{Y,0} - c_{\frac{\alpha}{2},N}, \mu_{Y,0} + c_{\frac{\alpha}{2},N}\right]\right\}$$

for ϕ_m equal to either (15) or (16). This convergence in probability holds jointly for all finite collections of values $\mu_{Y,0}$, however, which implies (17). The same argument works

for the median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$, which can also be viewed as the upper endpoint of a one-sided 50% confidence interval. □

Proof of Proposition 9 We prove this result for the unconditional case, noting that since $Pr_{\mu_m}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}\to 1$, the result conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$ follows immediately.

Note that $Pr_{\mu_m}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}\to 1$ implies $Pr_{\mu_{Y,m}}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}=z\right\}\to_p 1$ for all z. Hence, for $g(\mu_Y,z)=Pr_{\mu_Y}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}=z\right\}$, we see that $g(\mu_{Y,m},z)\to_p 1$ for all z. Note, next, that for d the Euclidian distance between the endpoints, if we define $h_\varepsilon(\mu_Y,z)=Pr_{\mu_Y}\{d(CS_U,CS_N)>\varepsilon|Z_{\tilde{\theta}}=z\}$, Lemma 2 implies that for any sequence $(\mu_{Y,m},z_m)$ such that $g(\mu_{Y,m},z_m)\to 1$, $h_\varepsilon(\mu_{Y,m},z_m)\to 0$. Hence, if we define $\mathcal{G}(\delta)=\{(\mu_Y,z):g(\mu_Y,z)>1-\delta\}$ and $\mathcal{H}(\varepsilon)=\{(\mu_Y,z):h_\varepsilon(\mu_Y,z)<\varepsilon\}$, for all $\varepsilon>0$ there exists $\delta(\varepsilon)>0$ such that $\mathcal{G}(\delta(\varepsilon))\subseteq\mathcal{H}(\varepsilon)$.

Hence, since our argument above implies that for all $\delta>0$, $Pr_{\mu_m}\{(\mu_{Y,m},z)\in\mathcal{G}(\delta)\}\to 1$ for all z, we see that for all $\varepsilon>0$, $Pr_{\mu_m}\{(\mu_{Y,m},z)\in\mathcal{H}(\varepsilon)\}\to 1$ for all z as well, which suffices to prove the desired claim for confidence intervals. The same argument likewise implies the result for our median unbiased estimator. □

Proof of Lemma 3 Note that we can assume without loss of generality that $\mu_{Y,0}=0$ and $\Sigma_Y(\tilde{\theta})=1$ since we can define $Y^*(\tilde{\theta})=\left(Y(\tilde{\theta})-\mu_{Y,0}\right)/\sqrt{\Sigma_Y(\tilde{\theta})}$ and consider the problem of testing that the mean of $Y^*(\tilde{\theta})$ is zero (transforming the set $\mathcal{Y}_m$ accordingly). After deriving critical values (c_l^*,c_u^*) in this transformed problem, we can recover critical values for our original problem as $(c_l,c_u)=\sqrt{\Sigma_Y(\tilde{\theta})}(c_l^*,c_u^*)+\mu_{Y,0}$. Hence, for the remainder of the proof we assume that $\mu_{Y,0}=0$ and $\Sigma_Y(\tilde{\theta})=1$.

Equal-Tailed Test We consider first the equal-tailed test. Note that this test rejects if and only if $Y(\tilde{\theta})\notin[c_{l,ET}(\mathcal{Y}),c_{u,ET}(\mathcal{Y})]$, where we suppress the dependence of the critical values on $\mu_{Y,0}=0$ for simplicity, and $(c_{l,ET}(\mathcal{Y}),c_{u,ET}(\mathcal{Y}))$ solve $F_{TN}(c_{l,ET}(\mathcal{Y}),\mathcal{Y})=\frac{\alpha}{2}$ and $F_{TN}(c_{u,ET}(\mathcal{Y}),\mathcal{Y})=1-\frac{\alpha}{2}$, for $F_{TN}(\cdot,\mathcal{Y})$ the distribution function of a standard normal random variable truncated to $\mathcal{Y}$. Recall that we can write the density corresponding to $F_{TN}(y,\mathcal{Y})$ as $\frac{1\{y\in\mathcal{Y}\}}{Pr\{\xi\in\mathcal{Y}\}}f_N(y)$ where f_N is the standard normal density and $Pr\{\xi\in\mathcal{Y}\}$ is the probability that $\xi\in\mathcal{Y}$ for $\xi\sim N(0,1)$. Hence, we can write $F_{TN}(y,\mathcal{Y})=\frac{\int_{-\infty}^{y}1\{\tilde{y}\in\mathcal{Y}\}f_N(\tilde{y})d\tilde{y}}{Pr\{\xi\in\mathcal{Y}\}}$.

Note next that for all y we can write $F_{TN}(y,\mathcal{Y}_m)=a_m(y)+F_N(y)$, where F_N is the standard normal distribution function and $a_m(y)=\frac{\int_{-\infty}^{y}1\{\tilde{y}\in\mathcal{Y}_m\}f_N(\tilde{y})d\tilde{y}}{Pr\{\xi\in\mathcal{Y}_m\}}-F_N(y)$. Recall, however, that $Pr\{\xi\in\mathcal{Y}_m\}\to 1$ and

$$\left|\int_{-\infty}^{y}1\{\tilde{y}\in\mathcal{Y}_m\}f_N(\tilde{y})d\tilde{y}-F_N(y)\right|=\left|\int_{-\infty}^{y}[1\{\tilde{y}\in\mathcal{Y}_m\}-1]f_N(\tilde{y})d\tilde{y}\right|$$

$$= \int_{-\infty}^{y} 1\{\tilde{y} \notin \mathcal{Y}_m\} f_N(\tilde{y}) d\tilde{y} \leq Pr\{\xi \notin \mathcal{Y}_m\} \to 0$$

for all y, so $a_m(y) \to 0$ for all y. Theorem 2.11 in van der Vaart (1998) then implies that $a_m(y) \to 0$ uniformly in y as well.

Note next that $F_{TN}\left(c_{l,ET}(\mathcal{Y}_m), \mathcal{Y}_m\right) = a_m\left(c_{l,ET}(\mathcal{Y}_m)\right) + F_N\left(c_{l,ET}(\mathcal{Y}_m)\right) = \frac{\alpha}{2}$ implies $c_{l,ET}(\mathcal{Y}_m) = F_N^{-1}\left(\frac{\alpha}{2} - a_m(c_{l,ET}(\mathcal{Y}_m))\right)$, and thus that $c_{l,ET}(\mathcal{Y}_m) \to F_N^{-1}\left(\frac{\alpha}{2}\right)$. Using the same argument, we can show that $c_{u,ET}(\mathcal{Y}_m) \to F_N^{-1}\left(1 - \frac{\alpha}{2}\right)$, as desired.

Unbiased Test We next consider the unbiased test. Recall that critical values $c_{l,U}(\mathcal{Y})$, $c_{u,U}(\mathcal{Y})$ for the unbiased test solve $Pr\{\zeta \in [c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})]\} = 1 - \alpha$ and $E[\zeta 1\{\zeta \in [c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})]\}] = (1-\alpha)E[\zeta]$ for $\zeta \sim \xi | \xi \in \mathcal{Y}$ where $\xi \sim N(0,1)$.

Note that for ζ_m the truncated normal random variable corresponding to $\mathcal{Y}_m$, we can write $Pr\{\zeta_m \in [c_l, c_u]\} = a_m(c_l, c_u) + (F_N(c_u) - F_N(c_l))$ with

$$a_m(c_l, c_u) = (F_N(c_l) - Pr\{\zeta_m \leq c_l\}) - (F_N(c_u) - Pr\{\zeta_m \leq c_u\}).$$

As in the argument for equal-tailed tests above, we see that both $F_N(c_u) - Pr\{\zeta_m \leq c_u\}$ and $F_N(c_l) - Pr\{\zeta_m \leq c_l\}$ converge to zero pointwise, and thus uniformly in c_u and c_l by Theorem 2.11 in van der Vaart (1998). Hence, $a_m(c_l, c_u) \to 0$ uniformly in (c_l, c_u).

Note, next, that we can write $E[\zeta_m 1\{\zeta_m \in [c_l, c_u]\}] = [\xi 1\{\xi \in [c_l, c_u]\}] + b_m(c_l, c_u)$ for

$$b_m(c_l, c_u) = E[\zeta_m 1\{\zeta_m \in [c_l, c_u]\}] - [\xi 1\{\xi \in [c_l, c_u]\}] = \int_{c_l}^{c_u} \left(\frac{1\{y \in \mathcal{Y}_m\}}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) y f_N(y) dy.$$

Note, however, that $\int_{c_l}^{c_u} (1\{y \in \mathcal{Y}_m\} - 1) y f_N(y) dy \leq E[|\xi| 1\{\xi \notin \mathcal{Y}_m\}]$. Hence, since

$$\begin{aligned}
&\left| \int_{c_l}^{c_u} \left(\frac{1\{y \in \mathcal{Y}_m\}}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) y f_N(y) dy \right| \\
&\leq \left| \int_{c_l}^{c_u} (1\{y \in \mathcal{Y}_m\} - 1) y f_N(y) dy \right| + \left| \int_{c_l}^{c_u} \left(\frac{1\{y \in \mathcal{Y}_m\}}{Pr\{\xi \in \mathcal{Y}_m\}} - 1\{y \in \mathcal{Y}_m\} \right) y f_N(y) dy \right| \\
&\leq E[|\xi| 1\{\xi \notin \mathcal{Y}_m\}] + \left| \left(\frac{1}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) \right| \int_{c_l}^{c_u} 1\{y \in \mathcal{Y}_m\} |y| f_N(y) dy \\
&\leq \sqrt{P(\xi \notin \mathcal{Y}_m)} + \left| \left(\frac{1}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) \right| E[|\xi|]
\end{aligned}$$

by the Cauchy-Schwartz Inequality, where the right hand side tends to zero and doesn't depend on (c_l, c_u), $b_m(c_l, c_u)$ converges to zero uniformly in (c_l, c_u).

Next, let us define $(c_{l,m},c_{u,m})$ as the solutions to $Pr\{\zeta_m \in [c_l,c_u]\} = 1-\alpha$ and $E[\zeta_m 1\{\zeta_m \in [c_l,c_u]\}] = (1-\alpha)E[\zeta_m]$. From our results above, we can re-write the problem solved by $(c_{l,m},c_{u,m})$ as $F_N(c_u) - F_N(c_l) = 1-\alpha-a_m(c_l,c_u)$, $E[\xi 1\{\xi \in [c_l,c_u]\}] = (1-\alpha)E[\zeta_m] - b_m(c_l,c_u)$. Letting $\bar{a}_m = \sup_{c_l,c_u}|a_m(c_l,c_u)|$, and $\bar{b}_m = \sup_{c_l,c_u}|b_m(c_l,c_u)|$ we thus see that $(c_{l,m},c_{u,m})$ solves $F_N(c_u) - F_N(c_l) = 1-\alpha-a_m^*$ and $E[\xi 1\{\xi \in [c_l,c_u]\}] = (1-\alpha)E[\zeta_m] - b_m^*$ for some $a_m^* \in [-\bar{a}_m,\bar{a}_m]$, $b_m^* \in \left[-\bar{b}_m,\bar{b}_m\right]$. We will next show that for any sequence of values (a_m^*,b_m^*) such that $a_m^* \in [-\bar{a}_m,\bar{a}_m]$ and $b_m^* \in \left[-\bar{b}_m,\bar{b}_m\right]$ for all m, the implied solutions $c_{l,m}(a_m^*,b_m^*)$, $c_{u,m}(a_m^*,b_m^*)$ converge to $F_N^{-1}\left(\frac{\alpha}{2}\right)$ and $F_N^{-1}\left(1-\frac{\alpha}{2}\right)$. This follows from the next lemma, which is proved below.

Lemma 4

Suppose that $c_{l,m}$ and $c_{u,m}$ solve $Pr\{\xi \in [c_l,c_u]\} = 1-\alpha+a_m$ and $E[\xi 1\{\xi \in [c_l,c_u]\}] = d_m$ for a_m, $d_m \to 0$. Then $(c_{l,m},c_{u,m}) \to \left(-c_{\frac{\alpha}{2},N},c_{\frac{\alpha}{2},N}\right)$.

Using this lemma, since $E[\zeta_m] \to 0$ as $m \to \infty$ we see that for any sequence of values $(a_m^*,b_m^*) \to 0$, $(c_{l,m}(a_m^*,b_m^*),c_{u,m}(a_m^*,b_m^*)) \to \left(-c_{\frac{\alpha}{2},N},c_{\frac{\alpha}{2},N}\right)$. However, since $\bar{a}_m,\bar{b}_m \to 0$ we know that the values a_m^* and b_m^* corresponding to the true $c_{l,m}$, $c_{u,m}$ must converge to zero. Hence $(c_{l,m},c_{u,m}) \to \left(-c_{\frac{\alpha}{2},N},c_{\frac{\alpha}{2},N}\right)$ as we wanted to show. □

Proof of Lemma 4 Note that the critical values solve

$$f(a_m,d_m,c) = \begin{pmatrix} F_N(c_u) - F_N(c_l) - (1-\alpha) - a_m \\ \int_{c_l}^{c_u} y f_N(y) dy - d_m \end{pmatrix} = 0.$$

We can simplify this expression, since $\frac{\partial}{\partial y} f_N(y) = -y f_N(y)$, so $\int_{c_l}^{c_u} y f_N(y) dy = f_N(c_l) - f_N(c_u)$.

We thus must solve the system of equations $g(c) - v_m = 0$, for

$$g(c) = \begin{pmatrix} F_N(c_u) - F_N(c_l) \\ f_N(c_l) - f_N(c_u) \end{pmatrix}, \quad v_m = \begin{pmatrix} a_m + (1-\alpha) \\ d_m \end{pmatrix}.$$

Note that for $v_m = (1-\alpha,0)'$ this system is solved by $c = \left(-c_{\frac{\alpha}{2},N},c_{\frac{\alpha}{2},N}\right)$. Further,

$$\frac{\partial}{\partial c} g(c) = \begin{pmatrix} -f_N(c_l) & f_N(c_u) \\ -c_l f_N(c_l) & c_u f_N(c_u) \end{pmatrix},$$

which evaluated at $c=\left(-c_{\frac{\alpha}{2},N},c_{\frac{\alpha}{2},N}\right)$ is equal to

$$\begin{pmatrix} -f_N\left(c_{\frac{\alpha}{2},N}\right) & f_N\left(c_{\frac{\alpha}{2},N}\right) \\ c_{\frac{\alpha}{2},N}f_N\left(c_{\frac{\alpha}{2},N}\right) & c_{\frac{\alpha}{2},N}f_N\left(c_{\frac{\alpha}{2},N}\right) \end{pmatrix}$$

and has full rank for all $\alpha\in(0,1)$. Thus, by the implicit function theorem there exists an open neighborhood V of $v_\infty=(1-\alpha,0)$ such that $g(c)-v=0$ has a unique solution $c(v)$ for $v\in V$ and $c(v)$ is continuously differentiable. Hence, if we consider any sequence of values $v_m\to(1-\alpha,0)$, we see that $c(v_m)\to\begin{pmatrix} -c_{\frac{\alpha}{2},N} \\ c_{\frac{\alpha}{2},N} \end{pmatrix}$, again as we wanted to show. □

B.2 Proofs for Results in Main Text

Proof of Proposition 1 Let us assume without loss of generality that $\tilde{\theta}=\theta_1$. Note that the conditioning event $\{\max_{\theta\in\Theta}X(\theta)=X(\theta_1)\}$ is equivalent to $\{MX\geq 0\}$, where

$$M\equiv\begin{pmatrix} 1 & -1 & 0 & 0 & ... & 0 \\ 1 & 0 & -1 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & ... & -1 \end{pmatrix}$$

is a $(|\Theta|-1)\times|\Theta|$ matrix and the inequality is taken element-wise. Let $A=\begin{bmatrix} -M & 0_{(|\Theta|-1)\times|\Theta|} \end{bmatrix}$, where $0_{(|\Theta|-1)\times|\Theta|}$ denotes the $(|\Theta|-1)\times|\Theta|$ matrix of zeros. Let $W=(X',Y')'$ and note that we can re-write the event of interest as $\{W:AW\leq 0\}$ and that we are interested in inference on $\eta'\mu$ for η the $2|\Theta|\times 1$ vector with one in the $(|\Theta|+1)$st entry and zeros everywhere else. Define

$$\tilde{Z}^*_{\tilde{\theta}}=W-cY(\tilde{\theta}),$$

for $c=Cov(W,Y(\tilde{\theta}))/\Sigma_Y(\tilde{\theta})$, noting that the definition of $Z_{\tilde{\theta}}$ in (7) corresponds to extracting the elements of $\tilde{Z}^*_{\tilde{\theta}}$ corresponding to X. By Lemma 5.1 of Lee et al. (2016),

$$\{W:AW\leq 0\}=\left\{W:\mathcal{L}(\tilde{\theta},\tilde{Z}^*_{\tilde{\theta}})\leq Y(\tilde{\theta})\leq\mathcal{U}(\tilde{\theta},\tilde{Z}^*_{\tilde{\theta}}),\mathcal{V}(\tilde{\theta},\tilde{Z}^*_{\tilde{\theta}})\geq 0\right\},$$

where for $(v)_j$ the jth element of a vector v,

$$\mathcal{L}(\tilde{\theta},z)=\max_{j:(Ac)_j<0}\frac{-(Az)_j}{(Ac)_j}$$

$$\mathcal{U}(\tilde{\theta},z)=\min_{j:(Ac)_j>0}\frac{-(Az)_j}{(Ac)_j}$$

$$\mathcal{V}(\tilde{\theta},z)=\min_{j:(Ac)_j=0}-(Az)_j.$$

Note, however, that

$$\left(A\tilde{Z}^*_{\tilde{\theta}}\right)_j=Z_{\tilde{\theta}}(\theta_j)-Z_{\tilde{\theta}}(\theta_1)$$

and

$$(Ac)_j=-\frac{\Sigma_{XY}(\theta_1,\theta_1)-\Sigma_{XY}(\theta_1,\theta_j)}{\Sigma_Y(\theta_1)}.$$

Hence, we can re-write

$$\frac{-(A\tilde{Z}^*_{\tilde{\theta}})_j}{(Ac)_j}=\frac{\Sigma_Y(\theta_1)(Z_{\tilde{\theta}}(\theta_j)-Z_{\tilde{\theta}}(\theta_1))}{\Sigma_{XY}(\theta_1,\theta_1)-\Sigma_{XY}(\theta_1,\theta_j)},$$

$$\mathcal{L}(\tilde{\theta},\tilde{Z}^*_{\tilde{\theta}})=\max_{j:\Sigma_{XY}(\theta_1,\theta_1)>\Sigma_{XY}(\theta_1,\theta_j)}\frac{\Sigma_Y(\theta_1)(Z_{\tilde{\theta}}(\theta_j)-Z_{\tilde{\theta}}(\theta_1))}{\Sigma_{XY}(\theta_1,\theta_1)-\Sigma_{XY}(\theta_1,\theta_j)},$$

$$\mathcal{U}(\tilde{\theta},\tilde{Z}^*_{\tilde{\theta}})=\min_{j:\Sigma_{XY}(\theta_1,\theta_1)<\Sigma_{XY}(\theta_1,\theta_j)}\frac{\Sigma_Y(\theta_1)(Z_{\tilde{\theta}}(\theta_j)-Z_{\tilde{\theta}}(\theta_1))}{\Sigma_{XY}(\theta_1,\theta_1)-\Sigma_{XY}(\theta_1,\theta_j)},$$

and

$$\mathcal{V}(\tilde{\theta},\tilde{Z}^*_{\tilde{\theta}})=\min_{j:\Sigma_{XY}(\theta_1,\theta_1)=\Sigma_{XY}(\theta_1,\theta_j)}-(Z_{\tilde{\theta}}(\theta_j)-Z_{\tilde{\theta}}(\theta_1)).$$

Note, however, that these are functions of $Z_{\tilde{\theta}}$, as expected. The result follows. □

Proof of Proposition 2 Follows as a special case of Proposition 7. □

Proof of Proposition 3 Follows as a special case of Proposition 9. □

Proof of Proposition 4 Provided $\hat{\theta}$ is unique with probability one, we can write

$$Pr_\mu\left\{\mu(\hat{\theta})\in CS\right\}=\sum_{\tilde{\theta}\in\Theta}Pr_\mu\left\{\hat{\theta}=\tilde{\theta}\right\}Pr_\mu\left\{\mu(\tilde{\theta})\in CS|\hat{\theta}=\tilde{\theta}\right\}.$$

Since $\sum_{\tilde{\theta}\in\Theta}Pr_\mu\left\{\hat{\theta}=\tilde{\theta}\right\}=1$, the result of the proposition follows immediately. □

Proof of Lemma 1 The assumption of the lemma implies that $X(\tilde{\theta})-X(\theta)$ has a non-degenerate normal distribution for all μ. Since Θ is finite, almost-sure uniqueness of $\hat{\theta}$ follows immediately. □

Proof of Proposition 5 We first establish uniqueness of $\hat{\mu}_\alpha^H$. To do so, it suffices to show that $F_{TN}^H(Y(\tilde{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},Z_{\tilde{\theta}})$ is strictly decreasing in $\mu_Y(\tilde{\theta})$. Note first that this holds for the truncated normal assuming truncation that does not depend on $\mu_Y(\tilde{\theta})$ by Lemma A.1 of Lee et al. (2016). When we instead consider $F_{TN}^H(Y(\tilde{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},Z_{\tilde{\theta}})$, we impose truncation to

$$Y(\tilde{\theta}) \in \left[\mu_Y(\tilde{\theta}) - c_\beta\sqrt{\Sigma_Y(\tilde{\theta})}, \mu_Y(\tilde{\theta}) + c_\beta\sqrt{\Sigma_Y(\tilde{\theta})}\right].$$

Since this interval shifts upwards as we increase $\mu_Y(\tilde{\theta})$, $F_{TN}^H(Y(\hat{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},Z_{\tilde{\theta}})$ is a-fortiori decreasing in $\mu_Y(\tilde{\theta})$. Uniqueness of $\hat{\mu}_\alpha^H$ for $\alpha \in (0,1)$ follows. Note, next, that $F_{TN}^H(Y(\tilde{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},Z_{\tilde{\theta}}) \in \{0,1\}$ for $\mu_Y(\tilde{\theta}) \notin CS_P^\beta$ from which we immediately see that $\hat{\mu}_\alpha^H \in CS_P^\beta$.

Finally, note that for $\mu_Y(\tilde{\theta})$ the true value, $F_{TN}^H(Y(\hat{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},Z_{\tilde{\theta}}) \sim U[0,1]$ conditional on $\left\{\hat{\theta}=\tilde{\theta}, Z_{\hat{\theta}}=z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\}$. Since $F_{TN}^H(Y(\hat{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},Z_{\tilde{\theta}})$ is decreasing in $\mu_Y(\tilde{\theta})$,

$$Pr_\mu\left\{\hat{\mu}_\alpha^H \geq \mu_Y(\tilde{\theta}) | \hat{\theta}=\tilde{\theta}, Z_{\hat{\theta}}=z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\}$$

$$= Pr_\mu\left\{F_{TN}^H(Y(\hat{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}}) \geq 1-\alpha | \hat{\theta}=\tilde{\theta}, Z_{\hat{\theta}}=z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\} = \alpha,$$

and thus $\hat{\mu}_\alpha^H$ is α-quantile-unbiased conditional on $\left\{\hat{\theta}=\tilde{\theta}, Z_{\hat{\theta}}=z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\}$. We can drop the conditioning on $Z_{\tilde{\theta}}$ by the law of iterated expectations, and α-quantile unbiasedness conditional on $\mu_Y(\tilde{\theta}) \in CS_P^\beta$ follows by the same argument as in the proof of Proposition 4.

Proof of Proposition 6 The first part of the proposition follows immediately from Proposition 5. For the second part of the proposition, note that

$$Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS_{ET}^H\right\} = Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS_P^\beta\right\} \times$$

$$\sum_{\tilde{\theta}\in\Theta} Pr_\mu\left\{\hat{\theta}=\tilde{\theta} | \mu_Y(\hat{\theta}) \in CS_P^\beta\right\} Pr_\mu\left\{\mu_Y(\tilde{\theta}) \in CS_{ET}^H | \hat{\theta}=\tilde{\theta}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\}$$

$$= Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS_P^\beta\right\} \frac{1-\alpha}{1-\beta} \geq (1-\beta)\frac{1-\alpha}{1-\beta} = 1-\alpha,$$

where the second equality follows from the first part of the proposition. The upper bound follows by the same argument and the fact that $Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CS_P^\beta\right\} \leq 1$. □

C Alternatives to Conventional Sample Splitting

In Section 3.3 of the main text, we discuss the relationship of our conditional approach to conventional sample splitting methods and note that the results of Fithian et al. (2017) imply that traditional sample splitting methods are dominated in our setting. For a given sample split, we derive optimal split-sample confidence intervals and estimators as well as easy-to-implement confidence intervals and estimators that dominate their conventional split-sample counterparts in the asymptotic version of the split-sample problem.

The Split-Sample Limit Experiment Let τ denote the given fraction of the full sample used to compute the estimated maximum and let (X_n^1,Y_n^1) and (X_n^2,Y_n^2) correspond to the first and second portions of the data, with

$$(X_n^1,Y_n^1)=\tau^{-1/2}(X_{[\tau\cdot n]},Y_{[\tau\cdot n]}),$$

$$(X_n^2,Y_n^2)=(1-\tau)^{-1}\big((X_n,Y_n)-\sqrt{\tau}(X_{[\tau\cdot n]+1},Y_{[\tau\cdot n]+1})\big)$$

and $[a]$ denoting the nearest integer to $a\in\mathbb{R}$. Finally, let $\hat{\theta}_n^1=\text{argmax}_{\theta\in\Theta}X_n^1(\theta)$ or $\hat{\theta}_n^1=\text{argmax}_{\theta\in\Theta}\|X_n^1(\theta)\|$, as in Andrews et al. (2020b), denote the estimated maximum from the first part of the sample. In large samples, (X_n^1,Y_n^1), (X_n^2,Y_n^2) and $\hat{\theta}_n^1$ behave according to[26]

$$\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix}\sim N(\mu,\Sigma),$$

$$\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix}\sim N\big(\mu,c^{-1}\Sigma\big)$$

and

$$\hat{\theta}^1=\text{argmax}_{\theta\in\Theta}X^1(\theta)$$

or

$$\hat{\theta}^1=\text{argmax}_{\theta\in\Theta}\big\|X^1(\theta)\big\|,$$

where $c=(1-\tau)/\tau$ and (X^1,Y^1) is independent of (X^2,Y^2). This is the generalization of the asymptotic problem discussed in Section 3.3 of the main text to arbitrary given sample

[26]The quantity Σ in the exposition of this section corresponds to the quantity Σ in the main text, multiplied by τ^{-1}.

splits.[27]

Traditional sample splitting methods for inference on $\mu_Y(\hat{\theta}^1)$ are based entirely on $Y^2(\hat{\theta}^1)$. Since Y^2 is independent of X^1, and thus of $\hat{\theta}^1$, this ensures the (conditional) median-unbiasedness of conventional split-sample estimates $Y^2(\hat{\theta}^1)$ and the (conditional) validity of conventional split-sample confidence intervals

$$CS_{SS} = \left[Y^2(\hat{\theta}^1) - \sqrt{c^{-1}\Sigma_Y(\hat{\theta}^1)}c_{\alpha/2,N}, Y^2(\hat{\theta}^1) + \sqrt{c^{-1}\Sigma_Y(\hat{\theta}^1)}c_{\alpha/2,N} \right]$$

for $\mu_Y(\hat{\theta}^1)$ but does not make full use of the information in the data. To derive optimal procedures in the sample splitting framework, we first derive a sufficient statistic for the unknown parameter μ conditional on $\left\{\hat{\theta}^1 = \tilde{\theta}\right\}$ and then apply classical exponential family results as in Section 3 of the main text.

Optimal Estimators and Confidence Intervals The joint (unconditional) density of (X^1, Y^1, X^2, Y^2) is proportional to

$$\exp\left(-\frac{1}{2}\left(\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} - \mu\right)' \Sigma^{-1}\left(\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} - \mu\right)\right) \exp\left(-\frac{c}{2}\left(\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix} - \mu\right)' \Sigma^{-1}\left(\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix} - \mu\right)\right).$$

The conditional density given $\left\{\hat{\theta}^1 = \tilde{\theta}\right\}$ is thus propotional to

$$\frac{1\left\{X^1 \in \mathcal{X}^1\left(\tilde{\theta}\right)\right\}}{Pr_\mu\left\{X^1 \in \mathcal{X}^1\left(\tilde{\theta}\right)\right\}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} - \mu\right)' \Sigma^{-1}\left(\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} - \mu\right)\right) \times$$

$$\exp\left(-\frac{c}{2}\left(\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix} - \mu\right)' \Sigma^{-1}\left(\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix} - \mu\right)\right)$$

with $\mathcal{X}^1(\tilde{\theta}) = \{X^1 : \hat{\theta} = \tilde{\theta}\}$, which we can re-write as

$$g_1\left(X^1, Y^1\right) g_2\left(X^2, Y^2\right) h(\mu) \exp\left(\left(\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} + c\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix}\right)' \Sigma^{-1}\mu\right)$$

[27] For simplicity of exposition, in this section we suppress the possibility of using additional conditioning variables $\hat{\gamma}_n = \gamma\left(X_n^1\right)$ with asymptotic counterpart $\hat{\gamma} = \gamma\left(X^1\right)$.

for

$$g_1\left(X^1,Y^1\right)=1\left\{X^1\in\mathcal{X}^1\left(\tilde{\theta}\right)\right\}\exp\left(-\frac{1}{2}\begin{pmatrix}X^1\\Y^1\end{pmatrix}'\Sigma^{-1}\begin{pmatrix}X^1\\Y^1\end{pmatrix}\right),$$

$$g_2\left(X^2,Y^2\right)=\exp\left(-\frac{c}{2}\begin{pmatrix}X^2\\Y^2\end{pmatrix}'\Sigma^{-1}\begin{pmatrix}X^2\\Y^2\end{pmatrix}\right),$$

and

$$h(\mu)=\frac{1}{Pr_\mu\left\{X^1\in\mathcal{X}^1\left(\tilde{\theta}\right)\right\}}\exp\left(-\frac{1+c}{2}\mu'\Sigma^{-1}\mu\right).$$

This exponential family structure shows that $\begin{pmatrix}X^*\\Y^*\end{pmatrix}=\left(\begin{pmatrix}X^1\\Y^1\end{pmatrix}+c\begin{pmatrix}X^2\\Y^2\end{pmatrix}\right)$ is sufficient for μ. Hence, for any function of (X^1,Y^1,X^2,Y^2), there exists a (potentially randomized) function of (X^*,Y^*) with the same distribution for all μ. Thus, to study questions of optimality it is without loss to limit attention to confidence intervals and estimators that depend only on (X^*,Y^*).

Now that we have derived a sufficient statistic (X^*,Y^*) for μ, we turn to the question of how to construct optimal estimators and confidence intervals for $\mu_Y(\hat{\theta}_1)$ conditional on $\left\{\hat{\theta}_1=\tilde{\theta}\right\}$. Note that the unconditional density of (X^*,Y^*) is proportional to

$$\exp\left(-\frac{1}{2+2c}\left(\begin{pmatrix}X^*\\Y^*\end{pmatrix}-(1+c)\mu\right)'\Sigma^{-1}\left(\begin{pmatrix}X^*\\Y^*\end{pmatrix}-(1+c)\mu\right)\right).$$

The density of (X^*,Y^*) given $\left\{\hat{\theta}^1=\tilde{\theta}\right\}$ is thus proportional to

$$\frac{Pr\left\{X^1\in\mathcal{X}^1\left(\tilde{\theta}\right)|X^*,Y^*\right\}}{Pr_\mu\left\{X^1\in\mathcal{X}^1\left(\tilde{\theta}\right)\right\}}\exp\left(-\frac{1}{2+2c}\left(\begin{pmatrix}X^*\\Y^*\end{pmatrix}-(1+c)\mu\right)'\Sigma^{-1}\left(\begin{pmatrix}X^*\\Y^*\end{pmatrix}-(1+c)\mu\right)\right),$$

where we have used sufficiency to drop dependence of the numerator on μ.

This joint distribution has the same exponential family structure used to derive the optimal estimators and confidence intervals in the main text (see the proofs of Propositions 7 and 8). Hence, the same arguments deliver optimal procedures for the split-sample

setting. Specifically, for

$$Z^*_{\tilde{\theta}} = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} - \left(Cov\left(\begin{pmatrix} X^* \\ Y^* \end{pmatrix}, Y^*\left(\tilde{\theta}\right) \right) / \Sigma_{Y^*}\left(\tilde{\theta}\right) \right) Y^*\left(\tilde{\theta}\right),$$

where Σ_{Y^*} denotes the variance of Y^*, we can re-write

$$\exp\left(\left(\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} + c \begin{pmatrix} X^2 \\ Y^2 \end{pmatrix} \right) \Sigma^{-1}\mu \right) = \exp\left(Y^*\left(\tilde{\theta}\right) \mu_{Y^*}\left(\tilde{\theta}\right) / \Sigma_{Y^*}\left(\tilde{\theta}\right) + Z^*_{\tilde{\theta}} \Sigma^+_{Z^*} \mu_{Z^*} \right)$$

for Σ_{Z^*} the variance of Z^*, A^+ the Moore-Penrose pseudoinverse of a matrix A, and

$$\mu_{Z^*} = (1+c)\mu - \left(Cov\left(\begin{pmatrix} X^* \\ Y^* \end{pmatrix}, Y^*\left(\tilde{\theta}\right) \right) / Var\left(Y^*\left(\tilde{\theta}\right) \right) \right) \mu_{Y^*}\left(\tilde{\theta}\right).$$

This expression shows that when we are interested in inference on $\mu_Y(\tilde{\theta})$ conditional on $\left\{\hat{\theta}^1 = \tilde{\theta}\right\}$, μ_{Z^*} is the nuisance parameter, and $Z^*_{\tilde{\theta}}$ is minimal sufficient for this parameter relative to observing (X^1, Y^1, X^2, Y^2).

If we let $F^*_{SS}(Y^*(\tilde{\theta}); \mu_{Y^*}(\tilde{\theta}), \tilde{\theta}, z^*)$ denote the conditional distribution function of $Y^* | Z^* = z^*, \hat{\theta}^1 = \tilde{\theta}$, then the same arguments used to prove Proposition 7 show that the optimal α quantile-unbiased estimator $\hat{\mu}^*_{SS,\alpha}$ for $\mu(\hat{\theta}^1)$ in the sample splitting problem solves

$$F^*_{SS}(Y^*(\hat{\theta}^1); (1+c)\hat{\mu}^*_{SS,\alpha}, \tilde{\theta}, Z^*_{\tilde{\theta}}) = 1 - \alpha.$$

Likewise, the same arguments used to prove Proposition 8 show that the optimal two-sided unbiased test conditional on $\left\{\hat{\theta}^1 = \tilde{\theta}\right\}$ rejects $H_0 : \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ when

$$Y^*(\tilde{\theta}) \notin \left[c_l\left(Z^*_{\tilde{\theta}}\right), c_u\left(Z^*_{\tilde{\theta}}\right) \right],$$

where $c_l(z)$, $c_u(z)$ solve

$$Pr\{\zeta \in [c_l(z), c_u(z)]\} = 1 - \alpha,\ E[\zeta 1\{\zeta \in [c_l(z), c_u(z)]\}] = (1-\alpha)E[\zeta]$$

with ζ distributed according to $F^*_{SS}(\cdot; (1+c)\mu_{Y,0}, \tilde{\theta}, z)$. These optimal procedures condition on $Z^*_{\tilde{\theta}}$ rather than (X^1, Y^1) and so, unlike conventional sample splitting, continue to treat (X^1, Y^1) as random for inference.

Feasible Dominating Estimators and Confidence Intervals To implement the optimal split-sample proecdures for $\mu_Y(\hat{\theta}^1)$, we need to evaluate (or at least be able to draw from) the conditional distribution $F^*_{SS}(\cdot;(1+c)\mu_{Y,0},\tilde{\theta},z)$. Unfortunately, however, it is not computationally straightforward to do so since $Y^*|Z^*=z^*,\hat{\theta}^1=\tilde{\theta}$ is distributed as a normal random variable truncated to a dependent random set. We thus introduce side constraints to derive procedures that, although they are not fully optimal in the unconstrained problem, are computationally straightforward to implement and dominate conventional sample splitting procedures for inference on $\mu_Y(\hat{\theta}^1)$ conditional on the realized value of $\hat{\theta}^1$. These computationally feasible procedures are optimal within the class of split-sample procedures that condition on $\{\hat{\theta}^1=\tilde{\theta}\}$ and the realizations of

$$Z^i_{\tilde{\theta}}=X^i-\left(\Sigma_{XY}\left(\cdot,\tilde{\theta}\right)/\Sigma_Y\left(\tilde{\theta}\right)\right)Y^i\left(\tilde{\theta}\right)$$

for $i=1,2$, where $(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})$ is a sufficient statistic for the nuisance parameter μ_X. Since $Y^2(\hat{\theta}^1)|\{\hat{\theta}^1=\tilde{\theta},(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})=(z^1,z^1)\}\sim Y^2(\tilde{\theta})$, the conventional split-sample estimator $Y^2(\hat{\theta}^1)$ and confidence interval CS_{SS} fall within the class of split-sample conditional procedures that condition on $\{\hat{\theta}^1=\tilde{\theta}\}$ and $(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})$. These conventional procedures are therefore dominated by the optimal procedures within this class, which we now describe.

Standard exponential family arguments show that $(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})$ is sufficient for the nuisance parameter μ_X and, conditional on $\{\hat{\theta}^1=\tilde{\theta}\}$ and $(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})$, optimal estimation and inference is based upon the conditional distribution of $Y^*(\tilde{\theta})$. Note that since $Y^2(\tilde{\theta})$ is independent of $(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})$ and both $\hat{\theta}^1$ and $Y^2(\tilde{\theta})$ are independent of $Z^2_{\tilde{\theta}}$,

$$Y^*(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},(Z^1_{\tilde{\theta}},Z^2_{\tilde{\theta}})=(z^1,z^2)\}\sim Y^1(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},Z^1_{\tilde{\theta}}=z^1\}+cY^2(\tilde{\theta}).$$

Thus, the feasible dominating split-sample procedures rely upon the computation of the distribution function of $Y^1(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},Z^1_{\tilde{\theta}}=z^1\}+cY^2(\tilde{\theta})$. We now describe a fast method for computing this object.

In analogy with full sample inference, let

$$\mathcal{Y}^1(\tilde{\theta},z^1)=\left\{y^1:z^1+\left(\Sigma_{XY}\left(\cdot,\tilde{\theta}\right)/\Sigma_Y\left(\tilde{\theta}\right)\right)y^1\in\mathcal{X}^1(\tilde{\theta})\right\}$$

so that conditional on $\{\hat{\theta}^1=\tilde{\theta}\}$ and $Z^1_{\tilde{\theta}}=z^1$, $Y^1(\tilde{\theta})$ follows a one-dimensional truncated normal distribution with truncation set $\mathcal{Y}^1(\tilde{\theta},z^1)$. Note that in both the level and norm maximization contexts, $\mathcal{Y}^1(\tilde{\theta},z^1)$ can be expressed as a finite union of disjoint

intervals: $\mathcal{Y}^1(\tilde{\theta},z^1)=\bigcup_{k=1}^K[\ell_k(z^1),u_k(z^1)]$, where the dependence of $\ell_k(z^1)$ and $u_k(z^1)$ for $k=1,...,K$ on $\tilde{\theta}$ is suppressed for notational simplicity. Note that $Y^1(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},Z^1_{\tilde{\theta}}=z^1\}$ is distributed as $\xi^1|\xi^1\in\mathcal{Y}^1(\tilde{\theta},z^1)$, where $\xi^1\sim N(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta}))$. The density function of $Y^1(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},Z^1_{\tilde{\theta}}=z^1\}$ is thus

$$f^1(y^1)=\frac{\sum_{k=1}^K f_N\left((y^1-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)1(\ell_k(z^1)\leq y^1\leq u_k(z^1))}{\sqrt{\Sigma_Y(\tilde{\theta})}\sum_{k=1}^K\left(F_N\left((u_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)-F_N\left((\ell_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)\right)}$$

and $cY^2(\tilde{\theta})$ has density function $f^2(y^2)=c^{-1/2}\Sigma_Y(\tilde{\theta})^{-1/2}f_N\left((y^2-c\mu)/\sqrt{c\Sigma_Y(\tilde{\theta})}\right)$. Therefore, since $Y^1(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},Z^1_{\tilde{\theta}}=z^1\}$ and $cY^2(\tilde{\theta})$ are independent, the density function of $Y^*(\tilde{\theta})|\{\hat{\theta}^1=\tilde{\theta},Z^1_{\tilde{\theta}}=z^1\}$ is equal to

$$\frac{\sum_{k=1}^K\int_{\ell_k(z^1)}^{u_k(z^1)}f_N\left((t-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)f_N\left((y^*-t-c\mu_Y(\tilde{\theta}))/\sqrt{c\Sigma_Y(\tilde{\theta})}\right)dt}{\sqrt{c}\Sigma_Y(\tilde{\theta})\sum_{k=1}^K\left(F_N\left((u_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)-F_N\left((\ell_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)\right)}$$

with corresponding distribution function

$$F^A_{SS}(y^*;\mu_Y(\tilde{\theta}),\tilde{\theta},z^1)$$

$$=\frac{\sum_{k=1}^K\int_{\ell_k(z^1)}^{u_k(z^1)}f_N\left((t-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)F_N\left((y^*-t-c\mu_Y(\tilde{\theta}))/\sqrt{c\Sigma_Y(\tilde{\theta})}\right)dt}{\sqrt{\Sigma_Y(\tilde{\theta})}\sum_{k=1}^K\left(F_N\left((u_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)-F_N\left((\ell_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)\right)}$$

$$=\frac{E\left[F_N\left((y^*-\xi^1-c\mu_Y(\tilde{\theta}))/\sqrt{c\Sigma_Y(\tilde{\theta})}\right)1\left(\xi^1\in\bigcup_{k=1}^K[\ell_k(z^1),u_k(z^1)]\right)\right]}{\sum_{k=1}^K\left(F_N\left((u_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)-F_N\left((\ell_k(z^1)-\mu_Y(\tilde{\theta}))/\sqrt{\Sigma_Y(\tilde{\theta})}\right)\right)},$$

where the expectation is taken with respect to $\xi^1\sim N(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta}))$. This latter expression for $F^A_{SS}(y^*;\mu_Y(\tilde{\theta}),\tilde{\theta},z^1)$ is very easy to compute by generating normal random variables in standard software packages. This makes the computation of optimal estimators, tests and confidence intervals within the class discussed here computationally straightforward.

Similarly to the optimal case above, the same arguments used to prove Proposition

2 show that the optimal α quantile-unbiased estimator $\hat{\mu}^A_{SS,\alpha}$ for $\mu_Y(\hat{\theta}^1)$ in the sample splitting problem that conditions on $\{\hat{\theta}^1=\tilde{\theta}\}$ and the realizations of $Z^1_{\tilde{\theta}}$ and $Z^2_{\tilde{\theta}}$ solves

$$F^A_{SS}(Y^*(\hat{\theta}^1);\hat{\mu}^A_{SS,\alpha},\tilde{\theta},Z^1_{\tilde{\theta}})=1-\alpha.$$

Therefore, our (equal-tailed) alternative split-sample confidence interval for $\mu_Y(\hat{\theta}^1)$ is $C^A_{SS}=[\hat{\mu}^A_{SS,\alpha/2},\hat{\mu}^A_{SS,1-\alpha/2}]$. Likewise, the same arguments used to prove Proposition 8 show that the optimal two-sided unbiased test rejects $H_0:\mu_Y(\tilde{\theta})=\mu_{Y,0}$ when

$$Y^*(\tilde{\theta})\notin\left[c_l\left(Z^1_{\tilde{\theta}}\right),c_u\left(Z^1_{\tilde{\theta}}\right)\right],$$

where $c_l(z)$, $c_u(z)$ solve

$$Pr\{\zeta\in[c_l(z),c_u(z)]\}=1-\alpha,\ E[\zeta 1\{\zeta\in[c_l(z),c_u(z)]\}]=(1-\alpha)E[\zeta]$$

with ζ distributed according to $F^A_{SS}(\cdot;\mu_{Y,0},\tilde{\theta},z)$. These dominating procedures condition on the realization of $Z^1_{\tilde{\theta}}$ rather than (X^1,Y^1), and so unlike conventional sample splitting continue to treat (X^1,Y^1) as random for inference.

As mentioned in Section 3, the split sample methods we introduce here are related to conditionally valid methods in the biostatistics literature on inference in adaptive clinical trial designs (e.g., Cohen and Sackrowitz, 1989 and Sampson and Sill, 2005). However, these latter methods apply to a specific form for μ_Y, assume independence across the entries of X^1 and condition on the entire ordering of the entries of X^1, entailing a loss of power relative to our split sample methods which condition only on $\hat{\theta}^1$.

D Uniform Asymptotic Validity

This section establishes uniform asymptotic validity for plug-in versions of the procedures discussed in the main text. One could use arguments along the same lines as those below to derive results for additional conditioning variables $\hat{\gamma}_n$, but since such arguments would be case-specific, we do not pursue such an extension here.

Feasible finite-sample estimators and confidence intervals are denoted as their counterparts in Sections 3–4, with the addition of an n subscript. We suppose that the sample of size n is drawn from some (unknown) distribution $P\in\mathcal{P}_n$. We first impose that (X_n,Y_n) are uniformly asymptotically normal under $P\in\mathcal{P}_n$, where the centering vectors $(\mu_{X,n},\mu_{Y,n})$ and the limiting variance Σ may depend on P.

Assumption 2

For 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 $\mu_{X,n}(P)$ and $\mu_{Y,n}(P)$ and a function $\Sigma(P)$ such that for $\xi_P \sim N(0,\Sigma(P))$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \sup_{f\in BL_1} \left| E_P\left[f\left(\begin{array}{c} X_n-\mu_{X,n}(P) \\ Y_n-\mu_{Y,n}(P) \end{array} \right) \right] - E[f(\xi_P)] \right| = 0.$$

Uniform convergence in bounded Lipschitz metric is one formalization for uniform convergence in distribution. When X_n and Y_n are scaled sample averages based on independent data, as in Section 2, Assumption 2 will follow from moment bounds, while for dependent data it will follow from moment and dependence bounds.

We next assume that the asymptotic variance is uniformly consistently estimable.

Assumption 3

The estimator $\widehat{\Sigma}_n$ is uniformly consistent in the sense that for all $\varepsilon>0$

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

Provided we use a variance estimator appropriate to the setting (e.g. the sample variance for iid data, long-run variance estimators for time series, and so on) Assumption 3 will follow from the same sorts of sufficient conditions as Assumption 2.

Finally, we restrict the asymptotic variance.

Assumption 4

There exists a finite $\bar{\lambda}>0$ such that

$$1/\bar{\lambda} \leq \Sigma_X(\theta;P), \Sigma_Y(\theta;P) \leq \bar{\lambda}, \text{ for all } \theta\in\Theta \text{ and all } P\in\mathcal{P}_n,$$

$$1/\bar{\lambda} \leq \sqrt{\Sigma_X(\theta;P)\Sigma_X(\tilde{\theta};P)} - \Sigma_X(\theta,\tilde{\theta};P) \text{ for all } \theta,\tilde{\theta}\in\Theta \text{ with } \theta\neq\tilde{\theta} \text{ and all } P\in\mathcal{P}_n.$$

The upper bounds on $\Sigma_X(\theta;P)$ and $\Sigma_Y(\theta;P)$ ensure that the random variables ξ_P in Assumption 2 are stochastically bounded, while the lower bounds ensure that each entry (X_n,Y_n) has a nonzero asymptotic variance. The second condition ensures that no two elements of X_n are perfectly (positively) correlated asymptotically, and hence, by Lemma 1, guarantees that $\hat{\theta}_n$ is unique with probability tending to one. Note that this condition is weaker than a standard assumption bounding the eigenvalues of $\Sigma_X(P)$ away from zero.

High-Dimensional Settings Our asymptotic analysis considers settings where $|\Theta|$, and hence the dimension of X_n and Y_n, are fixed as $n \to \infty$. One might also be interested in settings where $|\Theta|$ grows with n, but this will raise complications for both the normal approximation and estimation of the asymptotic variance. Such an extension is interesting, but beyond the scope of this paper.

Variance Estimation Practically, even for fixed $|\Theta|$ one might still worry about the difficulty of estimating Σ in finite samples, since this matrix has $|\Theta|(|\Theta|+1)/2$ entries. Fortunately, in many cases Σ has additional structure which renders variance estimation more tractable than in the fully general case. Suppose, for instance, that we want to conduct inference on the best-performing treatment from a randomized trial, as in Section 2 above and Section 6 below. In this case, provided trial participants are drawn independently, elements of $X_n(\theta)$ corresponding to distinct treatments are uncorrelated and Σ is diagonal. In other cases, such as Section 7 below, $|\Theta|$ may be large, but the elements of X_n are formed by taking combinations of a much lower-dimensional set of random variables. In this case, Σ_X can be written as a known linear transformation of a much lower-dimensional variance matrix.

D.1 Uniform Asymptotic Validity

In the finite-sample normal model, we study both conditional and unconditional properties of our methods. We would like to do the same in our asymptotic analysis, but may have $Pr\left\{\hat{\theta}_n = \tilde{\theta}\right\} \to 0$ for some $\tilde{\theta}$, in which case conditioning on $\hat{\theta}_n = \tilde{\theta}$ is problematic. To address this, we multiply conditional statements by the probability of the conditioning event.

Asymptotic uniformity results for conditional inference procedures were established by Tibshirani et al. (2018) and Andrews et al. (2020b) for settings where the target parameter is chosen in other ways. Their results, however, limit attention to classes of data generating processes with asymptotically bounded means $(\mu_{X,n}, \mu_{Y,n})$. This rules out e.g. the conventional pointwise asymptotic case that fixes P and takes $n \to \infty$. We do not require such boundedness. Moreover, the results of Tibshirani et al. (2018) do not cover quantile-unbiased estimation, and also do not cover hybrid procedures, which are new to the literature.[28]

Our proofs are based on subsequencing arguments as in D. Andrews et al. (2020a), though due to the differences in our setting (our interest in conditional inference, and the fact that our target is random from an unconditional perspective) we cannot directly apply their results. We first establish the asymptotic validity of our quantile-unbiased estimators.

[28] In a follow-up paper, Andrews et al. (2020b), we apply the conditional and hybrid approaches developed here to settings where $\hat{\theta} = \text{argmax} \|X(\theta)\|$.

Proposition 10

Under Assumptions 2-4, for $\hat{\mu}_{\alpha,n}$ the α-quantile unbiased estimator,

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\hat{\mu}_{\alpha,n}\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)|\hat{\theta}_n=\tilde{\theta}\right\}-\alpha\right|Pr_P\left\{\hat{\theta}_n=\tilde{\theta}\right\}=0, \tag{18}$$

for all $\tilde{\theta}\in\Theta$, and

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\hat{\mu}_{\alpha,n}\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)\right\}-\alpha\right|=0. \tag{19}$$

This immediately implies asymptotic validity of equal-tailed confidence intervals.

Corollary 1

Under Assumptions 2-4, for $CS_{ET,n}$ the level $1-\alpha$ equal-tailed confidence interval

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_{ET,n}|\hat{\theta}_n=\tilde{\theta}\right\}-(1-\alpha)\right|Pr_P\left\{\hat{\theta}_n=\tilde{\theta}\right\}=0,$$

for all $\tilde{\theta}\in\Theta$, and

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_{ET,n}\right\}-(1-\alpha)\right|=0.$$

We can likewise establish uniform asymptotic validity of projection confidence intervals.

Proposition 11

Under Assumptions 2-4, for $CS_{P,n}$ the level $1-\alpha$ projection confidence interval,

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_{P,n}\right\}\geq 1-\alpha. \tag{20}$$

To state results for hybrid estimators and confidence intervals, let $C_n^H\left(\tilde{\theta};P\right)=1\left\{\hat{\theta}_n=\tilde{\theta},\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_{P,n}^{\beta}\right\}$ be an indicator for the hybrid conditioning event that $\hat{\theta}_n$ is equal to $\tilde{\theta}$ and the parameter of interest $\mu_Y(\tilde{\theta})$ falls in the level β projection confidence interval $CS_{P,n}^{\beta}$. We can establish quantile unbiasedness of hybrid estimators given this event, along with bounded unconditional bias.

Proposition 12

Under Assumptions 2-4, for $\hat{\mu}_{\alpha,n}^H$ the α-quantile unbiased hybrid estimator based on $CS_{P,n}^{\beta}$,

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\hat{\mu}_{\alpha,n}^H\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)|C_n^H\left(\tilde{\theta};P\right)=1\right\}-\alpha\right|E_P\left\{C_n^H\left(\tilde{\theta};P\right)\right\}=0, \tag{21}$$

for all $\tilde{\theta} \in \Theta$, *and*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \left| Pr_P\left\{ \hat{\mu}^H_{\alpha,n} \geq \mu_{Y,n}\left(\hat{\theta}_n; P\right)\right\} - \alpha \right| \leq \max\{\alpha, 1-\alpha\}\beta. \tag{22}$$

Validity of hybrid estimators again implies validity of hybrid confidence intervals.

Corollary 2

Under Assumptions 2-4, for $CS^H_{ET,n}$ *the level* $1-\alpha$ *equal-tailed hybrid confidence interval based on* $CS^{\beta}_{P,n}$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \left| Pr_P\left\{ \mu_{Y,n}\left(\hat{\theta}_n; P\right) \in CS^H_{ET,n} | C^H_n\left(\tilde{\theta}; P\right) = 1\right\} - \frac{1-\alpha}{1-\beta} \right| E_P\left\{ C^H_n\left(\tilde{\theta}; P\right)\right\} = 0, \tag{23}$$

for all $\tilde{\theta} \in \Theta$,

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_n} Pr_P\left\{ \mu_{Y,n}\left(\hat{\theta}_n; P\right) \in CS^H_{ET,n}\right\} \geq 1-\alpha, \tag{24}$$

and

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} Pr_P\left\{ \mu_{Y,n}\left(\hat{\theta}_n; P\right) \in CS^H_{ET,n}\right\} \leq \frac{1-\alpha}{1-\beta} \leq 1-\alpha+\beta. \tag{25}$$

Hence, our procedures are uniformly asymptotically valid, unlike conventional inference.[29]

D.2 Auxiliary Lemmas

This section collects lemmas that we will use to prove our uniformity results.

Lemma 5

Under Assumption 4, for any sequence of confidence intervals CS_n, *any sequence of sets* $\mathcal{C}_n(P)$ *indexed by* P, $C_n(P) = 1\left\{ \left(X_n, Y_n, \widehat{\Sigma}_n \right) \in \mathcal{C}_n(P)\right\}$, *and any constant* α, *to show that*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \left| Pr_P\left\{ \mu_{Y,n}\left(\hat{\theta}_n; P\right) \in CS_n | C_n(P) = 1\right\} - \alpha \right| Pr_P\{C_n(P) = 1\} = 0$$

it suffices to show that for all subsequences $\{n_s\} \subseteq \{n\}$, $\{P_{n_s}\} \in \mathcal{P}^{\infty} = \times_{n=1}^{\infty} \mathcal{P}_n$ *with:*

1. $\Sigma(P_{n_s}) \to \Sigma^* \in \mathcal{S}$ *for*

$$\mathcal{S} = \left\{ \Sigma : 1/\bar{\lambda} \leq (\Sigma_X(\theta), \Sigma_Y(\theta)) \leq \bar{\lambda}, 1/\bar{\lambda} \leq \sqrt{\Sigma_X(\theta; P)\Sigma_X(\tilde{\theta}; P)} - \Sigma_X(\theta, \tilde{\theta}; P)\right\}, \tag{26}$$

[29] The bootstrap also fails to deliver uniform validity, as it implicitly tries to estimate the difference between the "winning" policy and the others, which cannot be done with sufficient precision. We are unaware of results for subsampling, m-out-of-n bootstrap, or other resampling-based approaches for this setting.

2. $Pr_{P_{n_s}}\{C_{n_s}(P_{n_s})=1\}\to p^*\in(0,1]$, *and*

3. $\mu_{X,n_s}(P_{n_s})-\max_\theta\mu_{X,n_s}(\theta;P_{n_s})\to\mu_X^*\in\mathcal{M}_X^*$ *for*

$$\mathcal{M}_X^*=\left\{\mu_X\in[-\infty,0]^{|\Theta|}:\max_\theta\mu_X(\theta)=0\right\},$$

we have

$$\lim_{s\to\infty}Pr_{P_{n_s}}\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{n_s}|C_{n_s}(P_{n_s})=1\right\}=\alpha. \tag{27}$$

Lemma 6

For collections of sets $\mathcal{C}_{n,1}(P),...,\mathcal{C}_{n,J}(P)$, *and* $C_{n,j}(P)=1\left\{\left(X_n,Y_n,\widehat{\Sigma}_n\right)\in\mathcal{C}_{n,j}(P)\right\}$, *if* $\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}Pr_P\{C_{n,j}(P)=1,C_{n,j'}(P)=1\}=0$ *for all* $j\neq j'$ *and*

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,j}(P)=1\right\}-(1-\alpha)\right|Pr_P\{C_{n,j}(P)=1\}=0$$

for all j, *then*

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}\geq(1-\alpha)\cdot\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_j Pr_P\{C_{n,j}(P)=1\},$$

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}\leq1-\alpha\cdot\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_j Pr_P\{C_{n,j}(P)=1\}.$$

To state the next lemma, define

$$\mathcal{L}\left(\tilde{\theta},Z,\Sigma\right)=\max_{\theta\in\Theta:\Sigma_{XY}\left(\tilde{\theta}\right)>\Sigma_{XY}\left(\tilde{\theta},\theta\right)}\frac{\Sigma_Y\left(\tilde{\theta}\right)\left(Z(\theta)-Z\left(\tilde{\theta}\right)\right)}{\Sigma_{XY}\left(\tilde{\theta}\right)-\Sigma_{XY}\left(\tilde{\theta},\theta\right)} \tag{28}$$

$$\mathcal{U}\left(\tilde{\theta},Z,\Sigma\right)=\min_{\theta\in\Theta:\Sigma_{XY}\left(\tilde{\theta}\right)<\Sigma_{XY}\left(\tilde{\theta},\theta\right)}\frac{\Sigma_Y\left(\tilde{\theta}\right)\left(Z(\theta)-Z\left(\tilde{\theta}\right)\right)}{\Sigma_{XY}\left(\tilde{\theta}\right)-\Sigma_{XY}\left(\tilde{\theta},\theta\right)}, \tag{29}$$

where we define a maximum over the empty set as $-\infty$ and a minimum over the empty set as $+\infty$. For

$$\begin{pmatrix}X_n^*\\Y_n^*\end{pmatrix}=\begin{pmatrix}X_n-\max_\theta\mu_{X,n}(\theta;P)\\Y_n-\mu_{Y,n}(P)\end{pmatrix},$$

we next show that using $\left(X_n^*,Y_n^*,\widehat{\Sigma}_n\right)$ in our calculations yields the same bounds $\mathcal{L}$ and $\mathcal{U}$ as using $\left(X_n,Y_n,\widehat{\Sigma}_n\right)$, up to additive shifts

Lemma 7

For $\mathcal{L}\left(\tilde{\theta},Z,\Sigma\right)$ *and* $\mathcal{U}\left(\tilde{\theta},Z,\Sigma\right)$ *as defined in (28) and (29), and*

$$Z_{\tilde{\theta},n}=X_n-\frac{\widehat{\Sigma}_{XY,n}\left(\cdot,\tilde{\theta}\right)}{\widehat{\Sigma}_{Y,n}\left(\tilde{\theta}\right)}Y_n\left(\tilde{\theta}\right),\quad Z^*_{\tilde{\theta},n}=X^*_n-\frac{\widehat{\Sigma}_{XY,n}\left(\cdot,\tilde{\theta}\right)}{\widehat{\Sigma}_{Y,n}\left(\tilde{\theta}\right)}Y^*_n\left(\tilde{\theta}\right),$$

we have

$$\mathcal{L}\left(\tilde{\theta},Z^*_{\tilde{\theta},n},\widehat{\Sigma}_n\right)=\mathcal{L}\left(\tilde{\theta},Z_{\tilde{\theta},n},\widehat{\Sigma}_n\right)-\mu_{Y,n}\left(\tilde{\theta};P\right),\ \ \mathcal{U}\left(\tilde{\theta},Z^*_{\tilde{\theta},n},\widehat{\Sigma}_n\right)=\mathcal{U}\left(\tilde{\theta},Z_{\tilde{\theta},n},\widehat{\Sigma}_n\right)-\mu_{Y,n}\left(\tilde{\theta};P\right).$$

For brevity, going forward we use the shorthand notation

$$\left(\mathcal{L}\left(\tilde{\theta},Z_{\tilde{\theta},n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\tilde{\theta},Z_{\tilde{\theta},n},\widehat{\Sigma}_n\right),\mathcal{L}\left(\tilde{\theta},Z^*_{\tilde{\theta},n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\tilde{\theta},Z^*_{\tilde{\theta},n},\widehat{\Sigma}_n\right)\right)=(\mathcal{L}_n,\mathcal{U}_n,\mathcal{L}^*_n,\mathcal{U}^*_n).$$

Lemma 8

Under Assumptions 2 and 3, for any $\{n_s\}$ *and* $\{P_{n_s}\}$ *satisfying conditions (1)-(3) of Lemma 5 and any* $\tilde{\theta}$ *with* $\mu^*_X\left(\tilde{\theta}\right)>-\infty$, $\left(Y^*_{n_s},\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s},\widehat{\Sigma}_{n_s},\hat{\theta}_{n_s}\right)\to_d\left(Y^*,\mathcal{L}^*,\mathcal{U}^*,\Sigma^*,\hat{\theta}\right)$, *where the objects on the right hand side are calculated based on* (Y^*,X^*,Σ^*) *for* $\left(X^{*\prime},Y^{*\prime}\right)'\sim N(\mu^*,\Sigma^*)$ *with* $\mu^*=(\mu^{*\prime}_X,0')'$.

Lemma 9

For F_N *again the standard normal distribution function, the function*

$$F_{TN}(Y(\theta);\mu,\Sigma_Y(\theta),\mathcal{L},\mathcal{U})=\frac{F_N\left(\frac{Y(\theta)\wedge\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)-F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)-F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}1(Y(\theta)\geq\mathcal{L})\tag{30}$$

is continuous in $(Y(\theta),\mu,\Sigma_Y(\theta),\mathcal{L},\mathcal{U})$ *on the set*

$$\left\{(Y(\theta),\mu,\Sigma_Y(\theta))\in\mathbb{R}^3,\mathcal{L}\in\mathbb{R}\cup\{-\infty\},\mathcal{U}\in\mathbb{R}\cup\{\infty\}:\Sigma_Y(\theta)>0,\mathcal{L}<Y(\theta)<\mathcal{U}\right\}.$$

D.3 Proofs for Auxiliary Lemmas

Proof of Lemma 5 To prove that

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_n(P)=1\right\}-\alpha\right|Pr_P\{C_n(P)=1\}=0$$

it suffices to show that

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\Big(Pr_P\Big\{\mu_{Y,n}\Big(\hat{\theta}_n;P\Big)\in CS_n|C_n(P)=1\Big\}-\alpha\Big)Pr_P\{C_n(P)=1\}\geq 0 \tag{31}$$

and

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\Big(Pr_P\Big\{\mu_{Y,n}\Big(\hat{\theta}_n;P\Big)\in CS_n|C_n(P)=1\Big\}-\alpha\Big)Pr_P\{C_n(P)=1\}\leq 0. \tag{32}$$

We prove that to show (31), it suffices to show that for all $\{n_s\}$, $\{P_{n_s}\}$ satisfying conditions (1)-(3) of the lemma,

$$\liminf_{s\to\infty}Pr_{P_{n_s}}\Big\{\mu_{Y,n_s}\Big(\hat{\theta}_{n_s};P_{n_s}\Big)\in CS_{n_s}|C_{n_s}(P_{n_s})=1\Big\}\geq\alpha. \tag{33}$$

An argument along the same lines implies that to prove (32) it suffices to show that

$$\limsup_{s\to\infty}Pr_{P_{n_s}}\Big\{\mu_{Y,n_s}\Big(\hat{\theta}_{n_s};P_{n_s}\Big)\in CS_{n_s}|C_{n_s}(P_{n_s})=1\Big\}\leq\alpha. \tag{34}$$

Note, however, that (33) and (34) together are equivalent to (27).

Towards contradiction, suppose that (31) fails, so

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\Big(Pr_P\Big\{\mu_{Y,n}\Big(\hat{\theta}_n;P\Big)\in CS_n|C_n(P)=1\Big\}-\alpha\Big)Pr_P\{C_n(P)=1\}<-\varepsilon,$$

for some $\varepsilon>0$ but that (33) holds for all sequences satisfying conditions (1)-(3) of the lemma. Then there exists an increasing sequence of sample sizes n_q and some sequence $\{P_{n_q}\}$ with $P_{n_q}\in\mathcal{P}_{n_q}$ for all q such that

$$\limsup_{q\to\infty}\Big(Pr_{P_{n_q}}\Big\{\mu_{Y,n_q}\Big(\hat{\theta}_{n_q};P_{n_q}\Big)\in CS_{n_q}|C_{n_q}(P_{n_q})=1\Big\}-\alpha\Big)Pr_{P_{n_q}}\{C_{n_q}(P_{n_q})=1\}<-\varepsilon. \tag{35}$$

We want to show that there exists a further subsequence $\{n_s\}\subseteq\{n_q\}$ satisfying (1)-(3) in the statement of the lemma, and so establish a contradiction.

Note that since the set $\mathcal{S}$ defined in (26) is compact (e.g. in the Frobenius norm), and Assumption 4 implies that $\Sigma(P_{n_q})\in\mathcal{S}$ for all q, there exists a further subsequence $\{n_r\}\subseteq\{n_q\}$ such that

$$\lim_{r\to\infty}\Sigma(P_{n_r})\to\Sigma^*$$

for some $\Sigma^*\in\mathcal{S}$.

Note, next, that $Pr_{P_{n_r}}\{C_{n_r}(P_{n_r})=1\}\in[0,1]$ for all r, and so converges along a subsequence $\{n_t\}\subseteq\{n_r\}$. However, (35) implies that $Pr_{P_{n_r}}\{C_{n_r}(P_{n_r})=1\}\geq\frac{\varepsilon}{\alpha}$ for all r, and

thus that $Pr_{P_{n_t}}\{C_{n_t}(P_{n_t})=1\}\to p^*\in\left[\frac{\varepsilon}{\alpha},1\right]$.

Finally, let us define $\mu^*_{X,n}(P)=\mu_{X,n}(P)-\max_\theta\mu_{X,n}(\theta;P)$, and note that $\mu^*_{X,n}(P)\leq 0$ by construction. Since $\mu^*_{X,n}(P)$ is finite-dimensional and $\max_\theta\mu^*_{X,n}(P;\theta)=0$, there exists some $\theta\in\Theta$ such that $\mu^*_{X,n}(P;\theta)$ is equal to zero infinitely often. Let $\{n_u\}\subseteq\{n_t\}$ extract the corresponding sequence of sample sizes. The set $[-\infty,0]^{|\Theta|}$ is compact under the metric $d(\mu_X,\tilde{\mu}_X)=\|F_N(\mu_X)-F_N(\tilde{\mu}_X)\|$ for $F_N(\cdot)$ the standard normal cdf applied elementwise, and $\|\cdot\|$ the Euclidean norm. Hence, there exists a further subsequence $\{n_s\}\subseteq\{n_u\}$ along which $\mu^*_{X,n_s}(P_{n_s})$ converges to a limit in this metric. Note, however, that this means that $\mu^*_{X,n_s}(P_{n_s})$ converges to a limit $\mu^*\in\mathcal{M}^*$ in the usual metric.

Hence, we have shown that there exists a subsequence $\{n_s\}\subseteq\{n_q\}$ that satisfies (1)-(3). By supposition, (33) must hold along this subsequence. Thus,

$$\liminf_{n\to\infty}\left(Pr_{P_{n_s}}\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{n_s}|C_{n_s}(P_{n_s})=1\right\}-\alpha\right)Pr_P\{C_{n_s}(P_{n_s})=1\}\geq 0,$$

which contradicts (35). Hence, we have established a contradiction and so proved that (33) for all subsequences satisfying conditions (1)-(3) of the lemma implies (31). An argument along the same lines shows that (34) along all subsequences satisfying conditions (1)-(3) of the lemma implies (32). □

Proof of Lemma 6 Define $C_{n,J+1}(P)=1\{C_{n,j}(P)=0 \text{ for all } j\in\{1,...,J\}\}$. Note that

$$\begin{aligned}&Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}\\=&\textstyle\sum_{j=1}^{J+1}Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,j}(P)=1\right\}Pr_P\{C_{n,j}(P)=1\}+o(1)\end{aligned}$$

where the $o(1)$ term is negligible uniformly over $P\in\mathcal{P}_n$ as $n\to\infty$. Hence,

$$\begin{aligned}&Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}-(1-\alpha)\\=&\textstyle\sum_{j=1}^{J+1}\left(Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,j}(P)=1\right\}-(1-\alpha)\right)Pr_P\{C_{n,j}(P)=1\}+o(1)\end{aligned}$$

and

$$\begin{aligned}&\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}-(1-\alpha)\\=&\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_{j=1}^{J+1}\left(Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,j}(P)=1\right\}-(1-\alpha)\right)Pr_P\{C_{n,j}(P)=1\}\\=&\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\left(Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,J+1}(P)=1\right\}-(1-\alpha)\right)Pr_P\{C_{n,J+1}(P)=1\}\end{aligned}$$

$$\geq -(1-\alpha)\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n} Pr_P\{C_{n,J+1}(P)=1\}$$

$$=-(1-\alpha)\left(1-\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_{j=1}^{J} Pr_P\{C_{n,j}(P)=1\}\right)$$

which immediately implies that

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n} Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}\geq(1-\alpha)\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_{j=1}^{J} Pr_P\{C_{n,j}(P)=1\}.$$

Likewise,

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n} Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}-(1-\alpha)$$

$$=\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\sum_{j=1}^{J+1}\left(Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,j}(P)=1\right\}-(1-\alpha)\right)Pr_P\{C_{n,j}(P)=1\}$$

$$=\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left(Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n|C_{n,J+1}(P)=1\right\}-(1-\alpha)\right)Pr_P\{C_{n,J+1}(P)=1\}$$

$$\leq\alpha\cdot\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n} Pr_P\{C_{n,J+1}(P)=1\}=\alpha\left(1-\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_{j=1}^{J} Pr_P\{C_{n,j}(P)=1\}\right).$$

This immediately implies that

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n} Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_n\right\}\leq 1-\alpha\cdot\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}\sum_{j=1}^{J} Pr_P\{C_{n,j}(P)=1\},$$

as we wanted to show. □

Proof of Lemma 7 Note that

$$Z^*_{\tilde{\theta},n}=Z_{\tilde{\theta},n}-\max_{\theta}\mu_{X,n}(\theta;P)+\widehat{\Sigma}_{XY,n}\left(\cdot,\tilde{\theta}\right)\frac{\mu_{Y,n}\left(\tilde{\theta};P\right)}{\widehat{\Sigma}_{Y,n}\left(\tilde{\theta}\right)},$$

so

$$Z^*_{\tilde{\theta},n}(\theta)-Z^*_{\tilde{\theta},n}\left(\tilde{\theta}\right)=Z_{\tilde{\theta},n}(\theta)-Z_{\tilde{\theta},n}\left(\tilde{\theta}\right)+\left(\widehat{\Sigma}_{XY,n}\left(\theta,\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n}\left(\tilde{\theta}\right)\right)\frac{\mu_{Y,n}\left(\tilde{\theta};P\right)}{\widehat{\Sigma}_{Y,n}\left(\tilde{\theta}\right)}.$$

The result follows immediately. □

Proof of Lemma 8 By Assumption 2

$$\begin{pmatrix} X_{n_s}-\mu_{X,n_s}(P_{n_s}) \\ Y_{n_s}-\mu_{Y,n_s}(P_{n_s}) \end{pmatrix} \to_d N(0,\Sigma^*).$$

Hence, by Slutsky's lemma

$$\begin{pmatrix} X^*_{n_s} \\ Y^*_{n_s} \end{pmatrix} = \begin{pmatrix} X_{n_s}-\max_\theta \mu_{X,n_s}(\theta;P_{n_s}) \\ Y_{n_s}-\mu_{Y,n_s}(P_{n_s}) \end{pmatrix} \to_d \begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N(\mu^*,\Sigma^*).$$

We begin by considering one $\theta \in \Theta \setminus \left\{\tilde{\theta}\right\}$ at a time. Since $\widehat{\Sigma}_{n_s} \to_p \Sigma^*$ by Assumption 3, if $\Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right) \neq 0$ then

$$\frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta},n_s}(\theta) - Z^*_{\tilde{\theta},n_s}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right) - \widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)} \to_d \frac{\Sigma^*_Y\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta}}(\theta) - Z^*_{\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right)},$$

where the terms on the right hand side are based on (X^*,Y^*,Σ^*). The limit is finite if $\mu^*_X(\theta) > -\infty$, while otherwise $\mu^*_X(\theta) = -\infty$ and

$$\frac{\Sigma^*_Y\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta}}(\theta) - Z^*_{\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right)} = \begin{cases} -\infty & \text{if } \Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right) > 0 \\ +\infty & \text{if } \Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right) < 0 \end{cases}.$$

If instead $\Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right) = 0$, then since $\Sigma^*_X(\tilde{\theta},\theta) < \sqrt{\Sigma^*_X(\tilde{\theta})\Sigma^*_X(\theta)}$,

$$Z^*_{\tilde{\theta}}(\theta) - Z^*_{\tilde{\theta}}\left(\tilde{\theta}\right) = X^*(\theta) - X^*\left(\tilde{\theta}\right)$$

is normally distributed with non-zero variance. Hence, in this case

$$\left| \frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{n_s,\tilde{\theta}}(\theta) - Z^*_{n_s,\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right) - \widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)} \right| \to \infty. \tag{36}$$

Let us define

$$\Theta^*\left(\tilde{\theta}\right) = \left\{\theta \in \Theta \setminus \tilde{\theta} : \Sigma^*_{XY}\left(\tilde{\theta}\right) - \Sigma^*_{XY}\left(\tilde{\theta},\theta\right) \neq 0\right\}.$$

The argument above implies that

$$\max_{\theta\in\Theta^*(\tilde{\theta}):\widehat{\Sigma}_{XY,n_s}(\tilde{\theta})>\widehat{\Sigma}_{XY,n_s}(\tilde{\theta},\theta)} \frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta},n_s}(\theta)-Z^*_{\tilde{\theta},n_s}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)}$$

$$\rightarrow_d \mathcal{L}^* = \max_{\theta\in\Theta:\Sigma^*_{XY}(\tilde{\theta})>\Sigma^*_{XY}(\tilde{\theta},\theta)} \frac{\Sigma^*_Y\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta}}(\theta)-Z^*_{\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\Sigma^*_{XY}\left(\tilde{\theta}\right)-\Sigma^*_{XY}\left(\tilde{\theta},\theta\right)}, \tag{37}$$

and

$$\min_{\theta\in\Theta^*(\tilde{\theta}):\widehat{\Sigma}_{XY,n_s}(\tilde{\theta})<\widehat{\Sigma}_{XY,n_s}(\tilde{\theta},\theta)} \frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta},n_s}(\theta)-Z^*_{\tilde{\theta},n_s}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)}$$

$$\rightarrow_d \mathcal{U}^* = \min_{\theta\in\Theta:\Sigma^*_{XY}(\tilde{\theta})<\Sigma^*_{XY}(\tilde{\theta},\theta)} \frac{\Sigma^*_Y\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta}}(\theta)-Z^*_{\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\Sigma^*_{XY}\left(\tilde{\theta}\right)-\Sigma^*_{XY}\left(\tilde{\theta},\theta\right)}. \tag{38}$$

Since

$$\max_{\theta\in\Theta:\widehat{\Sigma}_{XY,n_s}(\tilde{\theta})>\widehat{\Sigma}_{XY,n_s}(\tilde{\theta},\theta)} \frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta},n_s}(\theta)-Z^*_{\tilde{\theta},n_s}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)} \leq Y^*_{n_s}(\tilde{\theta})$$

$$\leq \min_{\theta\in\Theta:\widehat{\Sigma}_{XY,n_s}(\tilde{\theta})<\widehat{\Sigma}_{XY,n_s}(\tilde{\theta},\theta)} \frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{\tilde{\theta},n_s}(\theta)-Z^*_{\tilde{\theta},n_s}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)}$$

with probability one for all n_s and $Y_{n_s} \xrightarrow{d} Y^*$, (36) implies

$$\frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{n_s,\tilde{\theta}}(\theta)-Z^*_{n_s,\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)} \rightarrow -\infty$$

when $\Sigma^*_{XY}\left(\tilde{\theta}\right)=\Sigma^*_{XY}\left(\tilde{\theta},\theta\right)$ for all $\theta,\tilde{\theta}\in\Theta$ such that $\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)>\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)$. Similarly,

$$\frac{\widehat{\Sigma}_{Y,n_s}\left(\tilde{\theta}\right)\left(Z^*_{n_s,\tilde{\theta}}(\theta)-Z^*_{n_s,\tilde{\theta}}\left(\tilde{\theta}\right)\right)}{\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)-\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)} \rightarrow \infty$$

when $\Sigma^*_{XY}\left(\tilde{\theta}\right)=\Sigma^*_{XY}\left(\tilde{\theta},\theta\right)$ for all $\theta,\tilde{\theta}\in\Theta$ such that $\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta}\right)<\widehat{\Sigma}_{XY,n_s}\left(\tilde{\theta},\theta\right)$. Thus,

the same convergence results as (37)–(38) continue to hold when we minimize and maximize over Θ rather than $\Theta^*(\tilde{\theta})$. Hence, $\left(\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s}\right) \to_d (\mathcal{L}^*,\mathcal{U}^*)$. Moreover, $\hat{\theta}_{n_s}$ is almost everywhere continuous in $X^*_{n_s}$, so $\left(Y^*_{n_s},\widehat{\Sigma}_{n_s},\hat{\theta}_{n_s}\right) \to_d \left(Y^*,\Sigma^*,\hat{\theta}\right)$ by the continuous mapping theorem, and this convergence holds jointly with that for $\left(\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s}\right)$. Hence, we have established the desired convergence. □

Proof of Lemma 9 Continuity for $\Sigma_Y(\theta) > 0, \mathcal{L} < Y(\theta) < \mathcal{U}$ with all elements finite is immediate from the functional form. Moreover, for fixed $(Y(\theta),\mu,\Sigma_Y(\theta)) \in \mathbb{R}^3$ with $\Sigma_Y(\theta) > 0$ and $\mathcal{L} < Y(\theta) < \mathcal{U}$,

$$\lim_{\mathcal{U}\to\infty} \frac{F_N\left(\frac{Y(\theta)\wedge\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)} 1(Y(\theta) \geq \mathcal{L}) = \frac{F_N\left(\frac{Y(\theta)-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\infty}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}$$

$$\lim_{\mathcal{L}\to-\infty} \frac{F_N\left(\frac{Y(\theta)\wedge\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)} 1(Y(\theta) \geq \mathcal{L}) = \frac{F_N\left(\frac{Y(\theta)-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}$$

and

$$\lim_{(\mathcal{L},\mathcal{U})\to(-\infty,\infty)} \frac{F_N\left(\frac{Y(\theta)\wedge\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L}-\mu}{\sqrt{\Sigma_Y(\theta)}}\right)} 1(Y(\theta) \geq \mathcal{L}) = \frac{F_N\left(\frac{Y(\theta)-\mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\infty}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}.$$

Hence, we obtain the desired result. □

D.4 Proofs for Uniformity Results

Proof of Proposition 10 Note that

$$\hat{\mu}_{\alpha,n} \geq \mu_{Y,n}\left(\hat{\theta}_n;P\right) \Longleftrightarrow \mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS_{U,-,n}$$

for $CS_{U,-,n} = (-\infty,\hat{\mu}_{\alpha,n}]$. Hence, by Lemma 5, to prove that (18) holds it suffices to show that for all $\{n_s\}$ and $\{P_{n_s}\}$ such that conditions (1)-(3) of the lemma hold with $C_n(P) = 1\left\{\hat{\theta}_n = \tilde{\theta}\right\}$, we have

$$\lim_{s\to\infty} Pr_{P_{n_s}}\left\{\hat{\mu}_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right) \in CS_{U,-,n_s} | \hat{\theta}_{n_s} = \tilde{\theta}\right\} = \alpha. \tag{39}$$

To this end, recall that for $F_{TN}(Y(\theta);\mu,\Sigma_Y(\theta),\mathcal{L},\mathcal{U})$ as defined in (30), the estimator $\hat{\mu}_{\alpha,n}$ solves $F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right)\right)=1-\alpha$. This cdf is strictly decreasing in μ as argued in the proof of Proposition 5, and is increasing in $Y_n\left(\hat{\theta}\right)$. Hence, $\hat{\mu}_{\alpha,n}\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)$ if and only if

$$F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu_{Y,n}\left(\hat{\theta}_n;P\right),\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right)\right)\geq 1-\alpha.$$

Note, next, that by Lemma 7 and the form of the function F_{TN},

$$\begin{aligned}&F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu_{Y,n}\left(\hat{\theta}_n;P\right),\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right)\right)\\&=F_{TN}\left(Y_n^*\left(\hat{\theta}_n\right);0,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}\left(\hat{\theta}_n,Z^*_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\hat{\theta}_n,Z^*_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right)\right),\end{aligned}$$

so $\hat{\mu}_{\alpha,n}\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)$ if and only if

$$F_{TN}\left(Y_n^*\left(\hat{\theta}_n\right);0,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}\left(\hat{\theta}_n,Z^*_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mathcal{U}\left(\hat{\theta}_n,Z^*_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right)\right)\geq 1-\alpha.$$

Lemma 8 shows that $\left(Y_n^*\left(\hat{\theta}_{n_s}\right),\widehat{\Sigma}_{Y,n_s}\left(\hat{\theta}_{n_s}\right),\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s},\hat{\theta}_{n_s}\right)$ converges in distribution as $s\to\infty$, so since F_{TN} is continuous by Lemma 9 while $\text{argmax}_\theta X^*(\theta)$ is almost surely unique and continuous for X^* as in Lemma 8, the continuous mapping theorem implies that

$$\begin{aligned}&\left(F_{TN}\left(Y^*_{n_s}\left(\hat{\theta}_{n_s}\right);0,\widehat{\Sigma}_{Y,n_s}\left(\hat{\theta}_{n_s}\right),\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s}\right),1\left\{\hat{\theta}_{n_s}=\tilde{\theta}\right\}\right)\\&\to_d\left(F_{TN}\left(Y^*\left(\hat{\theta}\right);0,\Sigma^*_Y\left(\hat{\theta}\right),\mathcal{L}^*,\mathcal{U}^*\right),1\left\{\hat{\theta}=\tilde{\theta}\right\}\right).\end{aligned}$$

Since we can write

$$Pr_{P_{n_s}}\left\{F_{TN}\left(Y^*_{n_s}\left(\hat{\theta}_{n_s}\right);0,\widehat{\Sigma}_{Y,n_s}\left(\hat{\theta}_{n_s}\right),\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s}\right)\geq 1-\alpha|\hat{\theta}_{n_s}=\tilde{\theta}\right\}$$

$$=\frac{E_{P_{n_s}}\left[1\left\{F_{TN}\left(Y^*_{n_s}\left(\hat{\theta}_{n_s}\right);0,\widehat{\Sigma}_{Y,n_s}\left(\hat{\theta}_{n_s}\right),\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s}\right)\geq 1-\alpha\right\}1\left\{\hat{\theta}_{n_s}=\tilde{\theta}\right\}\right]}{E_{P_{n_s}}\left[1\left\{\hat{\theta}_{n_s}=\tilde{\theta}\right\}\right]},$$

and by construction (see also Proposition 7 in the main text),

$$F_{TN}\left(Y^*\left(\hat{\theta}\right);0,\Sigma^*_Y\left(\hat{\theta}\right),\mathcal{L}^*,\mathcal{U}^*,\hat{\theta}\right)|\hat{\theta}=\tilde{\theta}\sim U[0,1],$$

and $Pr\left\{\hat{\theta}=\tilde{\theta}\right\}=p^*>0$, we thus have that

$$Pr_{P_{n_s}}\left\{F_{TN}\left(Y^*_{n_s}\left(\hat{\theta}_{n_s}\right);0,\widehat{\Sigma}_{Y,n_s}\left(\hat{\theta}_{n_s}\right),\mathcal{L}^*_{n_s},\mathcal{U}^*_{n_s}\right)\geq 1-\alpha|\hat{\theta}_{n_s}=\tilde{\theta}\right\}$$

$$\rightarrow Pr\left\{F_{TN}\left(Y^*\left(\hat{\theta}\right);0,\Sigma^*_Y\left(\hat{\theta}\right),\mathcal{L}^*,\mathcal{U}^*\right)\geq 1-\alpha|\hat{\theta}=\tilde{\theta}\right\}=\alpha,$$

which verifies (39).

Since this argument holds for all $\tilde{\theta}\in\Theta$, and Assumptions 2 and 4 imply that for all $\theta,\tilde{\theta}\in\Theta$ with $\theta\neq\tilde{\theta}$, $\lim_{n\rightarrow\infty}\sup_{P\in\mathcal{P}_n}Pr_P\left\{X_n(\theta)=X_n\left(\tilde{\theta}\right)\right\}=0$, Lemma 6 implies (19). □

Proof of Corollary 1 By construction, $CS_{ET,n}=\left[\hat{\mu}_{\alpha/2,n},\hat{\mu}_{1-\alpha/2,n}\right]$, and $\hat{\mu}_{1-\alpha/2,n}>\hat{\mu}_{\alpha/2,n}$ for all $\alpha<1$. Hence,

$$Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in CS_{ET,n}|\hat{\theta}_n=\tilde{\theta}\right\}$$

$$=Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\leq\hat{\mu}_{1-\alpha/2,n}|\hat{\theta}_n=\tilde{\theta}\right\}-Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\leq\hat{\mu}_{\alpha/2,n}|\hat{\theta}_n=\tilde{\theta}\right\},$$

so the result is immediate from Proposition 10 and Lemma 6. □

Proof of Proposition 11 By the same argument as in the proof of Lemma 5, to show that (20) holds it suffices to show that for all $\{n_s\}$, $\{P_{n_s}\}$ satisfying conditions (1)-(3) of Lemma 5, $\liminf_{n\rightarrow\infty}Pr_{P_{n_s}}\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}\right\}\geq 1-\alpha$.

To this end, note that

$$\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}\text{ if and only if }Y^*_{n_s}\left(\hat{\theta}_{n_s}\right)\in\left[-c_\alpha\left(\widehat{\Sigma}_{Y,n_s}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_{n_s}\right)},c_\alpha\left(\widehat{\Sigma}_{Y,n_s}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_{n_s}\right)}\right]$$

for $c_\alpha(\Sigma_Y)$ the $1-\alpha$ quantile of $\max_\theta|\xi(\theta)|/\sqrt{\Sigma_Y(\theta)}$ where $\xi\sim N(0,\Sigma_Y)$. Next, note that $c_\alpha(\Sigma_Y)$ is continuous in Σ on $\mathcal{S}$ as defined in (26). Hence, for all θ, $c_\alpha(\Sigma_Y)\sqrt{\Sigma_Y(\theta)}$ is continuous as well. Assumptions 2 and 3 imply that $\left(Y^*_{n_s},\widehat{\Sigma}_{n_s},\hat{\theta}_{n_s}\right)\rightarrow_d\left(Y^*,\Sigma^*,\hat{\theta}\right)$, which by the continuous mapping theorem implies

$$\left(Y^*_{n_s}\left(\hat{\theta}_{n_s}\right),c_\alpha\left(\widehat{\Sigma}_{Y,n_s}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_{n_s}\right)}\right)\rightarrow_d\left(Y^*\left(\hat{\theta}\right),c_\alpha(\Sigma^*_Y)\sqrt{\Sigma^*_Y\left(\hat{\theta}\right)}\right).$$

Hence, since $Pr\left\{\left|Y^*\left(\hat{\theta}\right)\right|-c_\alpha(\Sigma^*_Y)\sqrt{\Sigma^*_Y\left(\hat{\theta}\right)}=0\right\}=0$,

$$Pr_{P_{n_s}}\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}\right\}\rightarrow Pr\left\{Y^*\left(\hat{\theta}\right)\in\left[-c_\alpha(\Sigma^*_Y)\sqrt{\Sigma^*_Y\left(\hat{\theta}\right)},c_\alpha(\Sigma^*_Y)\sqrt{\Sigma^*_Y\left(\hat{\theta}\right)}\right]\right\}\qquad(40)$$

where the right hand side is at least $1-\alpha$ by construction. $\square$

Proof of Proposition 12 Note that $\hat{\mu}^H_{\alpha,n} \geq \mu_{Y,n}\left(\hat{\theta}_n;P\right)$ if and only if $\mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS^H_{U,-,n}$ for $CS^H_{U,-,n}=(-\infty,\hat{\mu}^H_{\alpha,n}]$. Hence, by Lemma 5, to prove that (21) holds it suffices to show that for all $\{n_s\}$ and $\{P_{n_s}\}$ such that conditions (1)-(3) of the lemma hold with $C_n(P)=1\left\{\hat{\theta}_n=\tilde{\theta},\mu_{Y,n}\left(\hat{\theta}_n;P_n\right)\in CS^{\beta}_{P,n}\right\}$, we have

$$\lim_{s\to\infty} Pr_{P_{n_s}}\left\{\hat{\mu}_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS^H_{U,-,n}|\hat{\theta}_{n_s}=\tilde{\theta},\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS^{\beta}_{P,n_s}\right\}=\alpha.$$

Recall that for $F_{TN}(Y(\theta);\mu,\Sigma_Y(\theta),\mathcal{L},\mathcal{U})$ defined as in (30), $\hat{\mu}^H_{\alpha,n}$ solves

$$F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}^H_n\left(\mu,\hat{\theta}_n\right),\mathcal{U}^H_n\left(\mu,\hat{\theta}_n\right)\right)=1-\alpha,$$

for

$$\mathcal{L}^H_n\left(\mu,\hat{\theta}_n\right)=\max\left\{\mathcal{L}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mu-c_\alpha\left(\widehat{\Sigma}_{Y,n}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_n\right)}\right\},$$

$$\mathcal{U}^H_n\left(\mu,\hat{\theta}_n\right)=\min\left\{\mathcal{U}\left(\hat{\theta}_n,Z_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),\mu+c_\alpha\left(\widehat{\Sigma}_{Y,n}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_n\right)}\right\}.$$

The proof of Proposition 5 shows that $F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}^H_n\left(\mu,\hat{\theta}_n\right),\mathcal{U}^H_n\left(\mu,\hat{\theta}_n\right)\right)$ is strictly decreasing in μ, so for a given value $\mu_{Y,0}$,

$$\hat{\mu}^H_{\alpha,n}\geq\mu_{Y,0}\Longleftrightarrow F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu_{Y,0},\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}^H_n\left(\mu_{Y,0},\hat{\theta}_n\right),\mathcal{U}^H_n\left(\mu_{Y,0},\hat{\theta}_n\right)\right)\geq 1-\alpha.$$

As in the proof of Proposition 10

$$\begin{aligned}&F_{TN}\left(Y_n\left(\hat{\theta}_n\right);\mu_{Y,n}\left(\hat{\theta}_n;P_n\right),\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}^H_n\left(\mu_{Y,n}\left(\hat{\theta}_n;P_n\right),\hat{\theta}_n\right),\mathcal{U}^H_n\left(\mu_{Y,n}\left(\hat{\theta}_n;P_n\right),\hat{\theta}_n\right)\right)\\&=F_{TN}\left(Y^*_n\left(\hat{\theta}_n\right);0,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}^{H*}_n\left(\hat{\theta}_n\right),\mathcal{U}^{H*}_n\left(\hat{\theta}_n\right)\right),\end{aligned}$$

where

$$\mathcal{L}^{H*}_n\left(\hat{\theta}_n\right)=\max\left\{\mathcal{L}\left(\hat{\theta}_n,Z^*_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),-c_\alpha\left(\widehat{\Sigma}_{Y,n}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_n\right)}\right\},$$

$$\mathcal{U}^{H*}_n\left(\hat{\theta}_n\right)=\min\left\{\mathcal{U}\left(\hat{\theta}_n,Z^*_{\hat{\theta}_n,n},\widehat{\Sigma}_n\right),c_\alpha\left(\widehat{\Sigma}_{Y,n}\right)\sqrt{\widehat{\Sigma}_Y\left(\hat{\theta}_n\right)}\right\}$$

so $\hat{\mu}^H_{\alpha,n}\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)$ if and only if $F_{TN}\left(Y^*_n\left(\hat{\theta}_n\right);0,\widehat{\Sigma}_{Y,n}\left(\hat{\theta}_n\right),\mathcal{L}^{H*}_n\left(\hat{\theta}_n\right),\mathcal{U}^{H*}_n\left(\hat{\theta}_n\right)\right)\geq 1-\alpha$.

Lemma 8 implies that $\left(Y^*_{n_s},\widehat{\Sigma}_{Y,n_s},\mathcal{L}^{H*}_{n_s}\left(\tilde{\theta}\right),\mathcal{U}^{H*}_{n_s}\left(\tilde{\theta}\right),\hat{\theta}_{n_s}\right)\rightarrow_d\left(Y^*,\Sigma^*_Y,\mathcal{L}^{H*}\left(\tilde{\theta}\right),\mathcal{U}^{H*}\left(\tilde{\theta}\right),\hat{\theta}\right)$,

where $\mathcal{L}^{H*}\left(\tilde{\theta}\right)$ and $\mathcal{U}^{H*}\left(\tilde{\theta}\right)$ are equal to $\mathcal{L}_n^{H*}\left(\tilde{\theta}\right)$ and $\mathcal{U}_n^{H*}\left(\tilde{\theta}\right)$ after replacing $(X_n,Y_n,\widehat{\Sigma}_n)$ with (X,Y,Σ^*). Then by the continuous mapping theorem and (40),

$$\begin{aligned}&\left(F_{TN}\left(Y_{n_s}^*\left(\hat{\theta}_{n_s}\right);0,\widehat{\Sigma}_{Y,n_s}\left(\hat{\theta}_{n_S}\right),\mathcal{L}_{n_s}^{H*}\left(\tilde{\theta}\right),\mathcal{U}_{n_s}^{H*}\left(\tilde{\theta}\right)\right),1\left\{\hat{\theta}_{n_s}=\tilde{\theta},\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}^{\beta}\right\}\right)\\ \to_d&\left(F_{TN}\left(Y^*\left(\hat{\theta}\right);0,\Sigma_Y^*\left(\hat{\theta}\right),\mathcal{L}^{H*}\left(\tilde{\theta}\right),\mathcal{U}^{H*}\left(\tilde{\theta}\right)\right),1\left\{\hat{\theta}=\tilde{\theta},Y^*\left(\hat{\theta}\right)\in\left[-c_\alpha(\Sigma_Y^*)\sqrt{\Sigma_Y^*\left(\hat{\theta}\right)},c_\alpha(\Sigma_Y^*)\sqrt{\Sigma_Y^*\left(\hat{\theta}\right)}\right]\right\}\right).\end{aligned}$$

Hence, by the same argument as in the proof of Proposition 10,

$$\lim_{s\to\infty}Pr_{P_{n_s}}\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{U,-,n_s}^{H}|\hat{\theta}_{n_s}=\tilde{\theta},\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}^{\beta}\right\}=\alpha,$$

as we aimed to show.

To prove (22), note that for $\widetilde{CS}_{U,+,n}^{H}=(\hat{\mu}_{\alpha,n}^{H},\infty)$,

$$\hat{\mu}_{\alpha,n}^{H}\geq\mu_{Y,n}\left(\hat{\theta}_n;P\right)\Longleftrightarrow\mu_{Y,n}\left(\hat{\theta}_n;P\right)\notin\widetilde{CS}_{U,+,n}^{H}$$

and thus that the argument above proves that

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\left|Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right)\in\widetilde{CS}_{U,+,n}^{H}|C_n^H\left(\tilde{\theta};P\right)\right\}-(1-\alpha)\right|Pr_P\left\{C_n^H\left(\tilde{\theta};P\right)\right\}=0$$

for $C_n^H\left(\tilde{\theta};P\right)$ as in the statement of the proposition. Since

$$\sum_{\tilde{\theta}\in\Theta}Pr_P\left\{\hat{\theta}_{n_s}=\tilde{\theta},\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}^{\beta}\right\}=Pr_P\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}^{\beta}\right\}+o(1),\qquad(41)$$

and Proposition 11 shows that

$$\liminf_{s\to\infty}\inf_{P\in\mathcal{P}_{n_s}}Pr_P\left\{\mu_{Y,n_s}\left(\hat{\theta}_{n_s};P_{n_s}\right)\in CS_{P,n_s}^{\beta}\right\}\geq1-\beta,$$

Lemma 6 together with (21) implies that

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n}Pr_P\left\{\hat{\mu}_{\alpha,n}^{H}<\mu_{Y,n}\left(\hat{\theta}_n;P\right)\right\}\geq(1-\alpha)(1-\beta)=(1-\alpha)-\beta(1-\alpha)$$

and

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}Pr_P\left\{\hat{\mu}_{\alpha,n}^{H}<\mu_{Y,n}\left(\hat{\theta}_n;P\right)\right\}\leq1-\alpha(1-\beta)=(1-\alpha)+\beta\alpha$$

from which the second result of the proposition follows immediately. □

Proof of Corollary 2 Note that by construction $CS^H_{ET,n} = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)},n}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)},n}\right]$, where $\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)},n} < \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)},n}$ provided $\frac{\alpha-\beta}{1-\beta} < 1$. Hence,

$$Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS^H_{ET,n} | C^H_n\left(\tilde{\theta},P\right)\right\}$$

$$= Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) \leq \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)},n} | C^H_n\left(\tilde{\theta},P\right)\right\} - Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) < \hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)},n} | C^H_n\left(\tilde{\theta},P\right)\right\},$$

so Proposition 12 immediately implies (23).

Equation (41) in the proof of Proposition 12 together with Lemma 6 implies that

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_n} Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS^H_{ET,n}\right\} \geq \frac{1-\alpha}{1-\beta}(1-\beta) = 1-\alpha$$

so (24) holds. We could likewise get an upper bound on coverage using Lemma 6, but obtain a sharper bound by proving the result directly. Specifically, note that

$$\mu_{Y,n}\left(\hat{\theta}_n;P_n\right) \in CS^H_{ET,n} \Rightarrow \mu_{Y,n}\left(\hat{\theta}_n;P_n\right) \in CS^\beta_{P,n}.$$

Hence,

$$Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS^H_{ET,n}\right\}$$
$$= Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS^H_{ET,n} | \hat{\mu}_{Y,n}\left(\hat{\theta}_n;P_n\right) \in CS^\beta_{P,n}\right\} Pr\left\{\mu_{Y,n}\left(\hat{\theta}_n;P_n\right) \in CS^\beta_{P,n}\right\}.$$

By the first part of the proposition, this implies that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} Pr_P\left\{\mu_{Y,n}\left(\hat{\theta}_n;P\right) \in CS^H_{ET,n}\right\} \leq \frac{1-\alpha}{1-\beta} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} Pr\left\{\mu_{Y,n}\left(\hat{\theta}_n;P_n\right) \in CS^\beta_{P,n}\right\}$$

$$\leq \frac{1-\alpha}{1-\beta},$$

so (25) holds as well. □

E Additional Results for Neighborhoods Application

E.1 Empirical Bayes and Winner's Curse

Chetty et al. (2018) focus on what they term forecast-unbiased estimates, which correspond to posterior means from a correlated random effects model which treats tract quality as normally and homoskedastically distributed conditional on a set of observable tract characteristics, with a mean that changes linearly in the tract characteristics. Specifically, for W_t the characteristics of

tract t, and μ_t tract quality, these estimates correspond to posterior means under the prior π that takes μ_t independent across tracts, with

$$\mu_t|W_t \sim N\big(W_t'\beta,\omega^2\big). \tag{42}$$

They then plug in estimates of ω and β.

If we take the model (42) seriously and abstract from estimation of ω and β (for instance because the number of tracts is large and we plug in consistent estimates), Bayesian posterior means solve the winner's curse problem under the prior. Specifically, note that the posterior mean for μ_t given the vector of estimates $\hat{\mu}$ is simply the mean given $\hat{\mu}_t$, $E_\pi[\mu_t|\hat{\mu}] = E_\pi[\mu_t|\hat{\mu}_t]$. The law of iterated expectations implies, however, that $E_\pi[\mu_t|\hat{\mu}]$ is unbiased for μ_t conditional on $\hat{\mu}$, so for any event $\mathcal{E}$ such that $Pr_\pi\{\hat{\mu}\in\mathcal{E}\}>0$,

$$E_\pi[\mu_t - E_\pi[\mu_t|\hat{\mu}_t]|\hat{\mu}\in\mathcal{E}] = 0.$$

Likewise, since we model $\hat{\mu}_t$ as normally distributed conditional on μ_t, the posterior mean is also the posterior median, so

$$Pr_\pi\{E_\pi[\mu_t|\hat{\mu}_t] > \mu_t|\hat{\mu}\in\mathcal{E}\} = \frac{1}{2},$$

and $E_\pi[\mu_t|\hat{\mu}_t]$ is median-unbiased under the prior conditional on the event $\mathcal{E}$. Note, however, that selection of a particular set of target tracts can be written as an event $\mathcal{E}$, so this argument implies that Bayesian posterior means are immune to the winner's curse under the prior. This depends crucially on the prior, however, since if we calculate the outer probability with respect to some other distribution of effect sizes $\tilde{\pi}\neq\pi$, we typically have

$$Pr_{\tilde{\pi}}\{E_\pi[\mu_t|\hat{\mu}_t] > \mu_t|\hat{\mu}\in\mathcal{E}\} \neq \frac{1}{2}.$$

E.2 Additional Figure for Movers Application

Figure 8 plots naive, conditional, and projection intervals for the Opportunity Atlas application described in the main text.

F Additional Simulation Results for Stylized Example

In the stylized example discussed in Section 2 of the main text, we focus on the median length of confidence intervals and the median absolute error of estimators. In this section, we report results for other quantiles, in particular that τ-th quantiles for $\tau\in\{0.05,0.25,0.5,0.75,0.95\}$.

Figures 9 and 10 show the unconditional quantiles of the length of the 95% confidence intervals CS_U and CS_{ET}, for cases with $|\Theta|$=2, 10, and 50 policies. In each case and for each

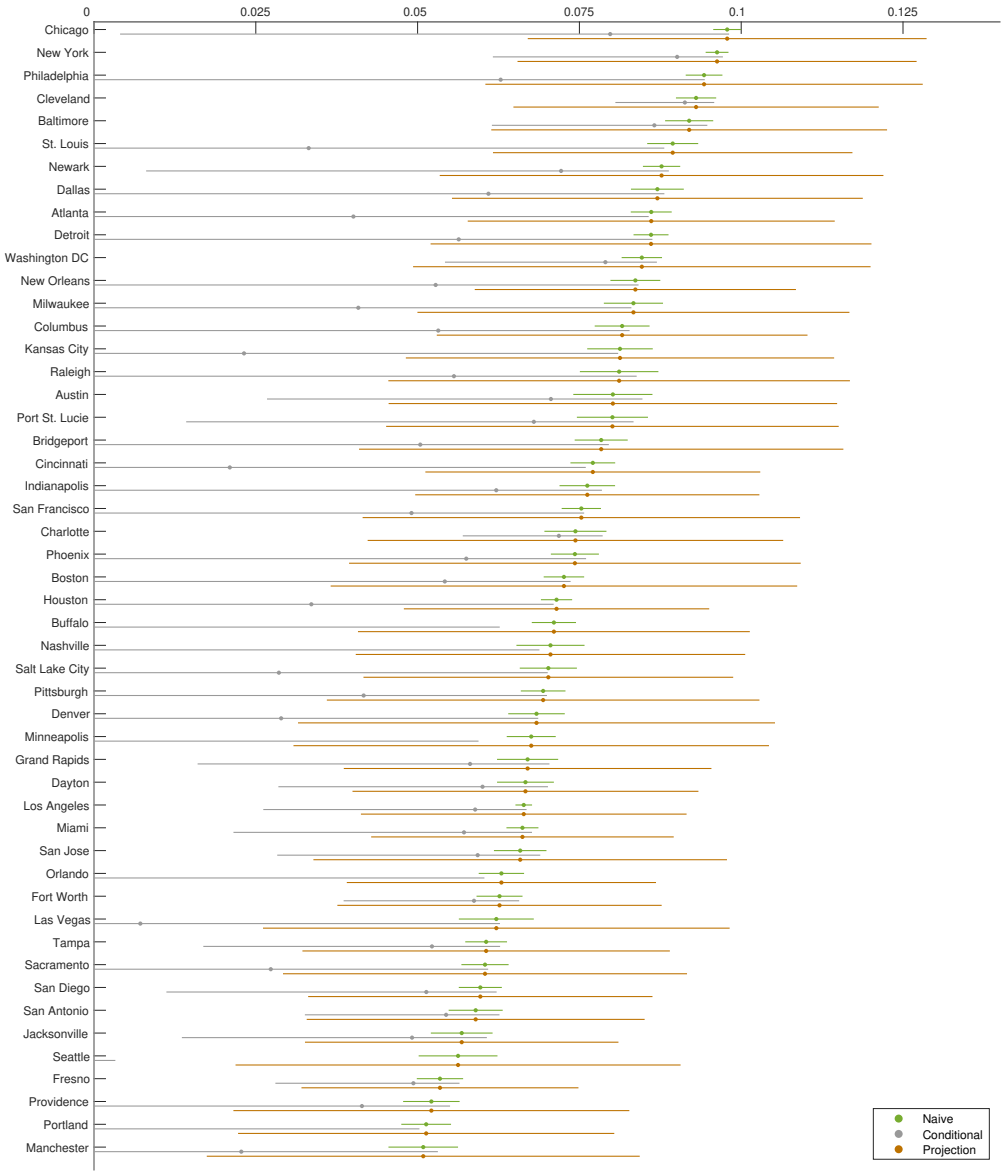

Figure 8: Estimates and confidence intervals for average economic mobility for selected census tracts based on Chetty et al. (2018) Opportunity Atlas, relative to the within-CZ average.

$\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$, the τ-th quantile is monotonically decreasing in $\mu(\theta_1) - \mu(\theta_{-1})$. Noting the different scales of the y-axes, we see that the upper quantiles grow as the number of policies increase, particularly for small $\mu(\theta_1) - \mu(\theta_{-1})$.

Figures 11 and 12 show the unconditional quantiles of the length of 95% hybrid confidence intervals CS_U^H and CS_{ET}^H with $\beta = 0.005$. Compared with Figures 9 and 10, the upper quantiles are much smaller, especially for small $\mu(\theta_1) - \mu(\theta_{-1})$. This substantial reduction in length directly comes from the construction of the hybrid confidence intervals, which ensures that CS_U^H and CS_{ET}^H are contained in CS_P^β. For the case of $|\Theta| = 50$, even the 95% quantiles of the length of CS_U^H and CS_{ET}^H are shorter than the length of CS_P uniformly over the range of $\mu(\theta_1) - \mu(\theta_{-1})$ values we consider.

Figures 13, 14, and 15 examine the performance of point estimators for $\mu(\hat{\theta})$. They plot the unconditional quantiles of the absolute error of the conventional estimator, the median unbiased estimator, and the hybrid estimator, respectively. In spite of the severe median bias shown in Figure 1 in the main text, the distribution of the conventional estimator is relatively concentrated compared to that of the median unbiased estimator. In particular, the upper quantiles of the absolute errors of $\hat{\mu}_{1/2}$ are very large for small $\mu(\theta_1) - \mu(\theta_{-1})$ (similar to the quantile plots of the length of CS_U and CS_{ET} shown in Figures 9 and 10).

At the cost of a small median bias, the hybrid estimator substantially reduces the absolute errors (Figure 15). The 95% quantile of the absolute errors of the hybrid estimator is overall similar to the 95% quantile of the absolute errors of the conventional estimator with a notable exception of the case of 2 policies. In contrast, for $|\Theta| = 10$ and 50, and for quantiles other than 95%, the hybrid estimator outperforms the conventional estimator over a wide range of values for $\mu(\theta_1) - \mu(\theta_{-1})$. These numerical results show that the hybrid estimator successfully reduces bias without greatly inflating the variability of the estimator.

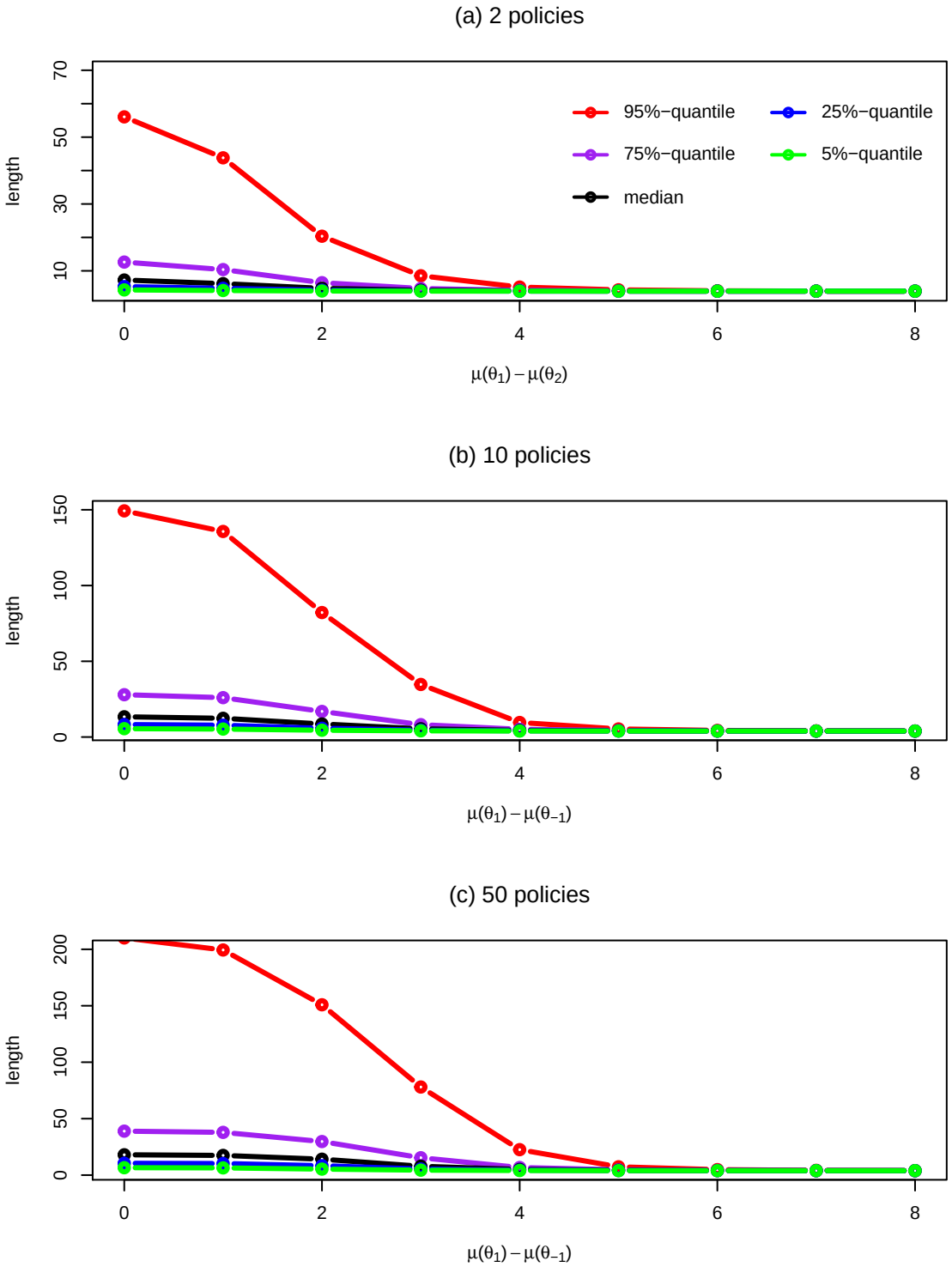

Figure 9: Quantiles of the length of 95% conditionally UMAU confidences sets CS_U.

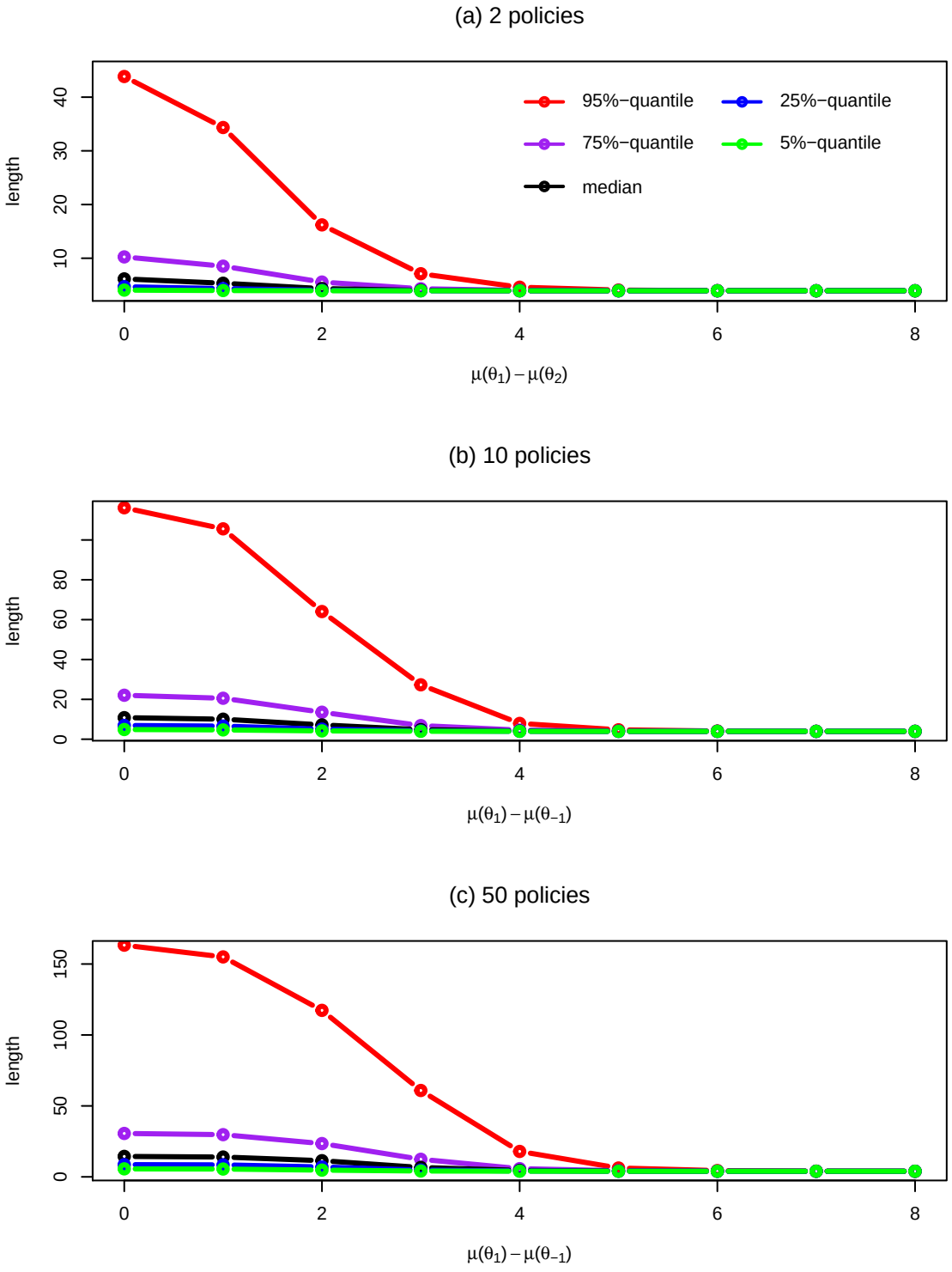

Figure 10: Quantiles of the length of 95% conditionally equal-tailed confidences sets CS_{ET}.

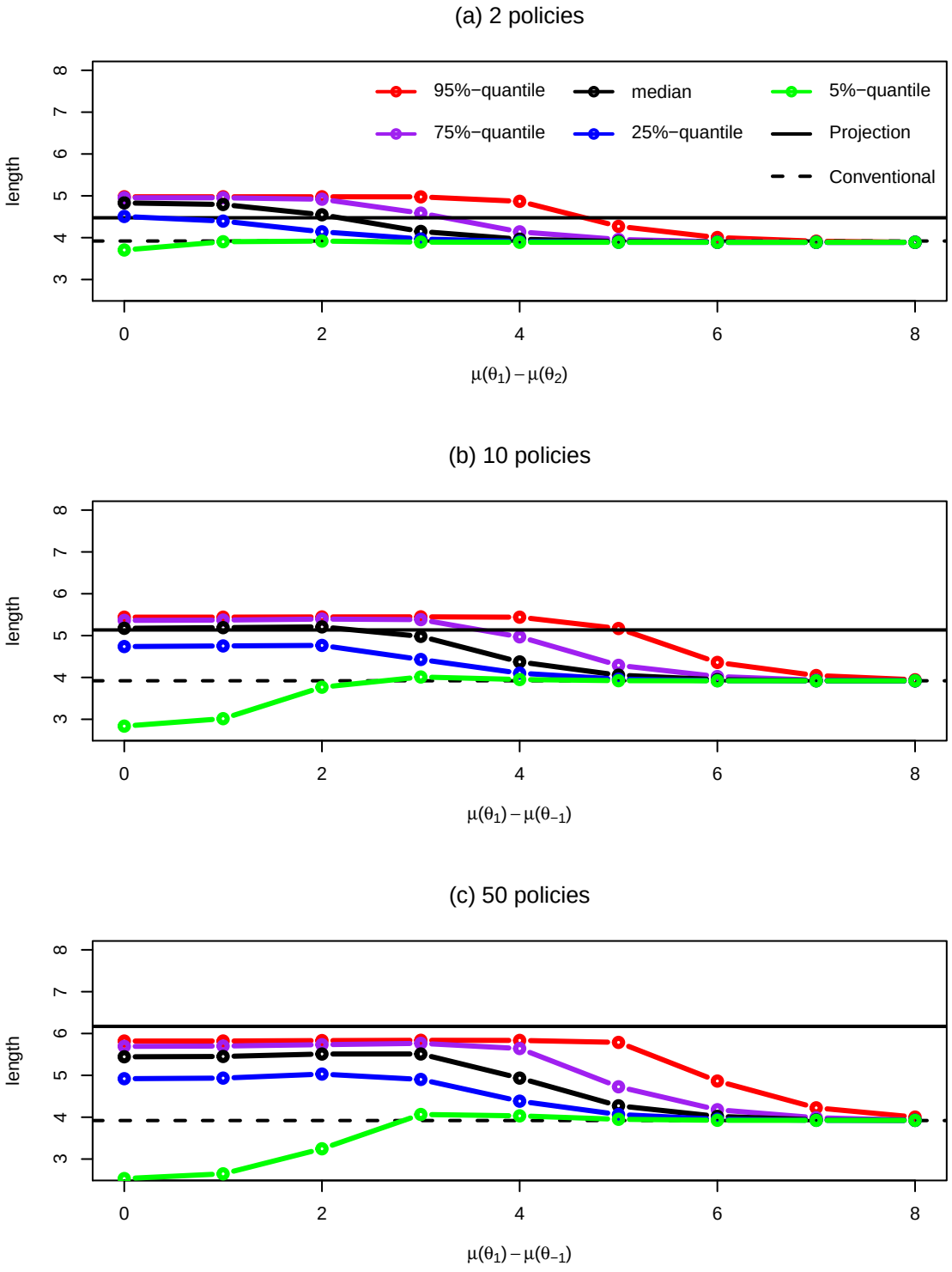

Figure 11: Quantiles of the length of 95% hybrid confidence intervals CS_U^H, with $\beta = 0.005$.

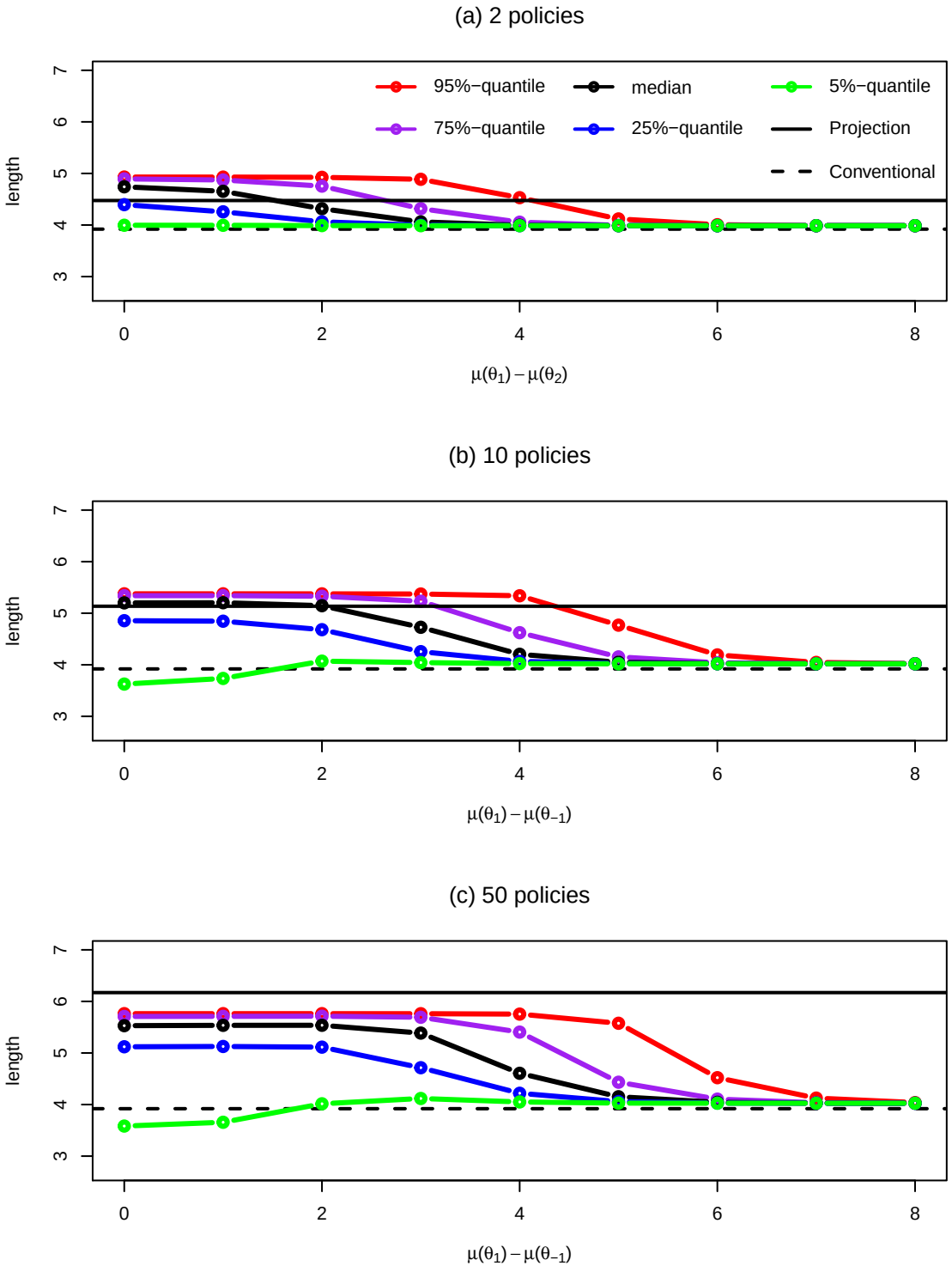

Figure 12: Quantiles of the length of 95% hybrid confidence intervals CS^H_{ET}, with $\beta = 0.005$.

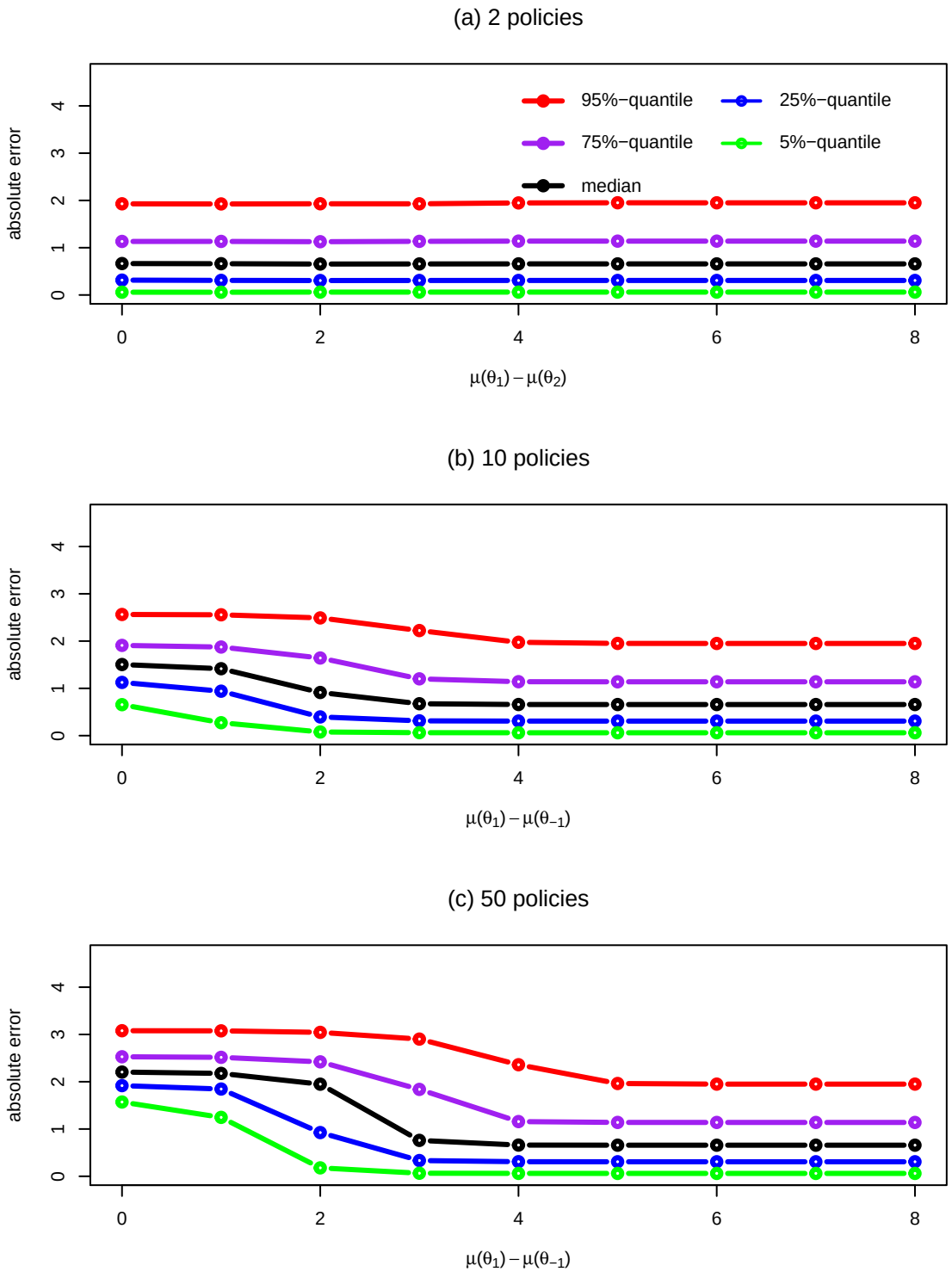

Figure 13: Quantiles of the absolute error of the conventional estimator (i.e. of $|X(\hat{\theta}) - \mu(\hat{\theta})|$).

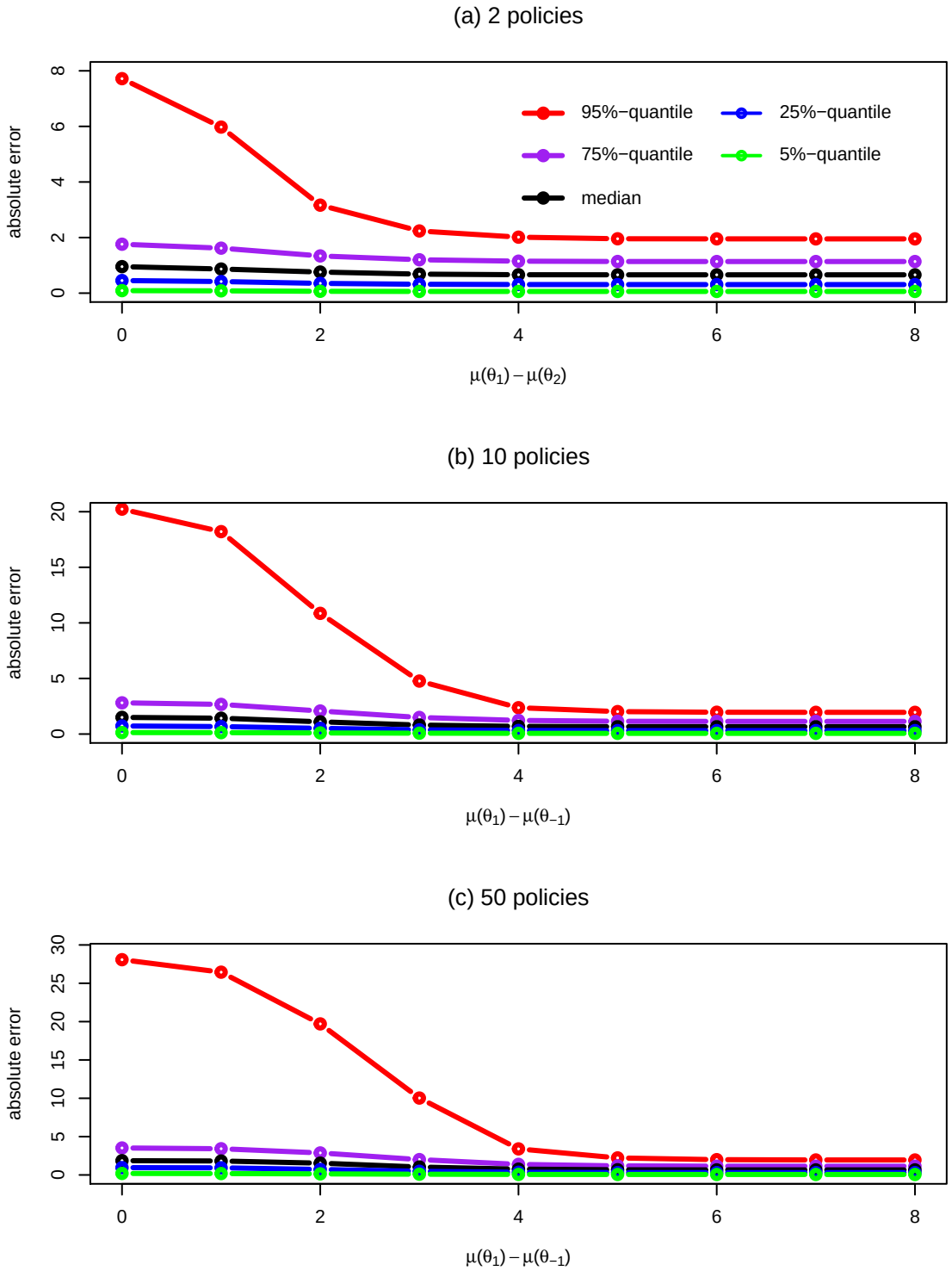

Figure 14: Quantiles of the absolute error of the conditionally optimal median unbiased estimator (i.e. of $|\hat{\mu}_{1/2} - \mu(\hat{\theta})|$).

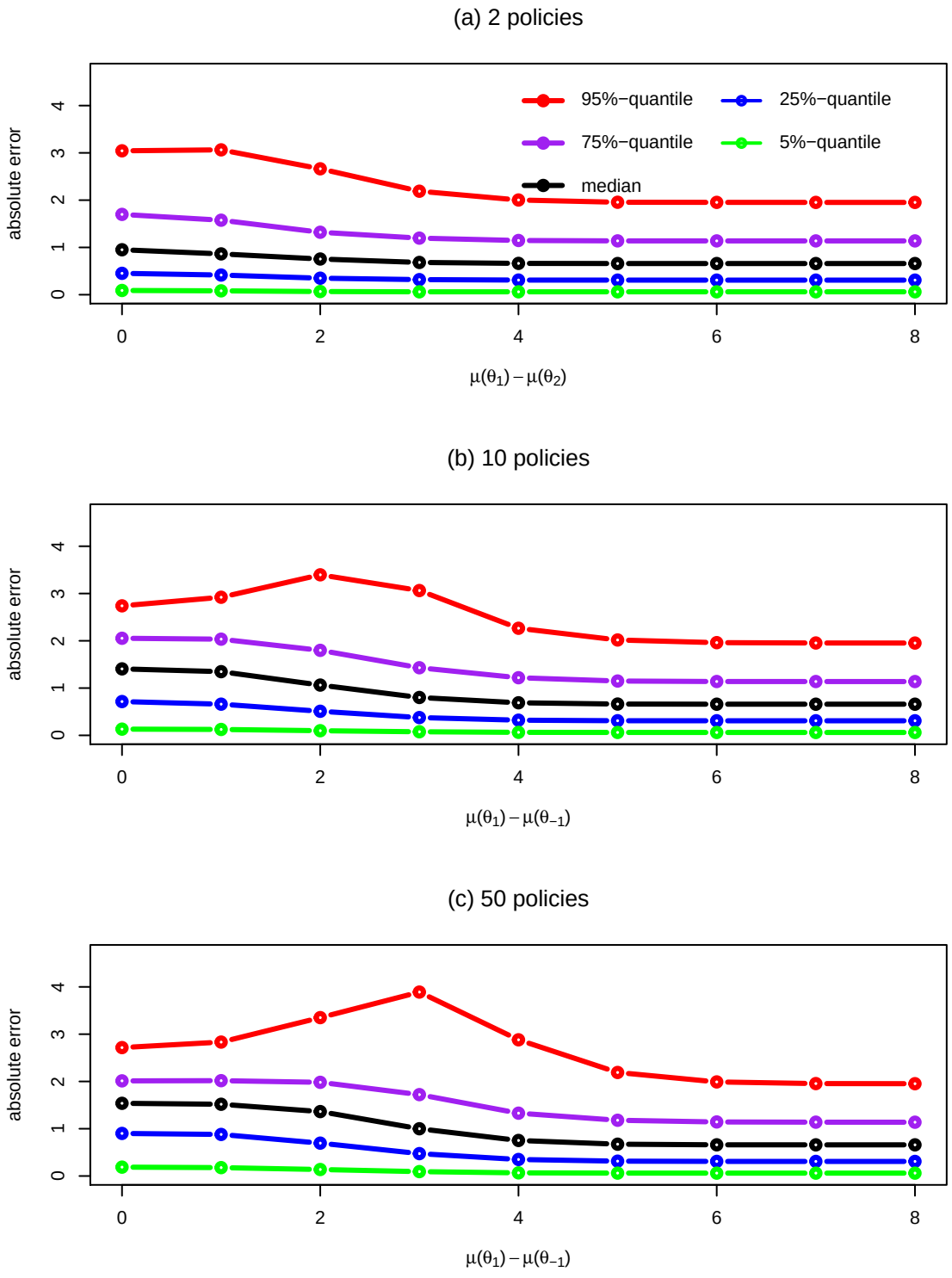

Figure 15: Quantiles of the absolute error of the hybrid estimator (i.e. of $|\hat{\mu}^H_{1/2} - \mu(\hat{\theta})|$) with $\beta = 0.005$.