Andrews, Isaiah; Kitagawa, Toru; McCloskey, Adam

Working Paper

Inference on winners

cemmap working paper, No. CWP31/18

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Andrews, Isaiah; Kitagawa, Toru; McCloskey, Adam (2018) : Inference on winners, cemmap working paper, No. CWP31/18, Centre for Microdata Methods and Practice (cemmap), London,
https://doi.org/10.1920/wp.cem.2018.3118

This Version is available at:
https://hdl.handle.net/10419/189743

cemmap
centre for microdata methods and practice

Inference on winners

Isaiah Andrews
Toru Kitagawa
Adam McCloskey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP31/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Inference on Winners*

Isaiah Andrews† Toru Kitagawa‡ Adam McCloskey§

May 10, 2018

Abstract

Many questions in econometrics can be cast as inference on a parameter 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 sets and median-unbiased estimators that are valid conditional on the parameter selected and so overcome this winner's curse. If one requires validity only on average over target parameters that might have been selected, we develop hybrid procedures that combine conditional and projection confidence sets and offer further performance gains that are attractive relative to existing alternatives.

Keywords: Winner's Curse, Selective Inference

JEL Codes: C12, C13

*We would like to thank Tim Armstrong, Frank Schoerfheide and Jesse Shapiro and seminar participants at Brandeis, Brown, BU, and Yale for helpful comments. Andrews gratefully aknowledges financial support from the NSF under grant number 1654234. Kitagawa gratefully aknowledges 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).

†Department of Economics, MIT, iandrews@mit.edu

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

§Department of Economics, Brown University, adam_mccloskey@brown.edu

1 Introduction

Many problems in econometrics can be cast as inference on target parameters selected through optimization over a finite set. In a randomized trial considering multiple treatments, 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. Perhaps less obviously, in structural break and tipping point models, researchers first estimate the location of a break or tipping point by minimizing the sum of squared residuals and then seek to estimate the magnitude of the discontinuity taking the estimated break location as given.

Estimators that do not account for data-driven selection of the target parameters can be badly biased, and conventional confidence sets that add and subtract a standard normal quantile times the standard error may severely under-cover. To illustrate, consider inference on the true average effect of the treatment that performed best in a randomized trial. Due to data-driven selection of the treatment of interest, the conventional estimate for its average effect will be biased upwards, and the usual 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 sets.

This paper develops estimators and confidence sets that eliminate these biases. There are two distinct perspectives from which to consider bias and coverage. The first conditions on the target parameter selected, for example on the identity of the best-performing treatment, while the second is unconditional and averages over possible target parameters. Conditional validity is more demanding but, as we discuss in the next section, may be desirable in some settings, for example when one wants to ensure validity conditional on the recommendation made to a policymaker. Both perspectives differ from inference on the effectiveness of the "true" best treatment as in e.g. Chernozhukov et al. (2013), in that we consider inference on the effectiveness of the (estimated) best-performing treatment in the experiment rather than the (unknown) best-performing treatment in the population.

Considering first conditional inference, we derive optimal unbiased and equal-tailed confidence sets. Our results build on the rapidly growing literature on selective

inference (e.g. Fithian et al. (2017); Lee et al. (2016); Harris et al. (2016); Tian and Taylor (2016)), which derives optimal conditional confidence sets in a range of other settings. We further observe that the results of Pfanzagl (1994) imply optimal median-unbiased estimators for conditional settings, which does not appear to have been previously noted in the selective inference literature. Hence, for settings where conditional validity is desired, we propose optimal inference procedures that eliminate the winner's curse noted above. We further show that in cases where this 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 so our corrections do not sacrifice efficiency in such cases.

A common alternative remedy for the biases we consider is sample splitting. In settings with identically distributed data, choosing the target parameter using the first part of the data and constructing estimates and confidence sets using the second part ensures unbiasedness of estimates and validity of conventional confidence sets. Indeed, split-sample confidence sets are valid conditional on the target parameter. Such procedures have three undesirable properties, however. First, the target parameter is generally more variable than if constructed using the full data. Second, only the second part of the data is used for inference, which Fithian et al. (2017) show implies that split-sample procedures are dominated by optimal conditional procedures applied using the same sample split. Third, non-uniqueness of the sample split means that the results are random or non-unique even conditional on the data. In work in progress, we develop implementable procedures that dominate conventional sample splitting in our setting.

We next turn to unconditional inference. One approach to constructing unconditional confidence sets is projection, which was previously used by e.g. Romano and Wolf (2005) and Kitagawa and Tetenov (2018). To obtain a projection confidence set, we form a simultaneous confidence band for all potential target parameters and take the implied set of values for the target parameter of interest. The resulting confidence sets 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 simulations 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 perform well in all cases, we next

introduce hybrid intervals which combine conditioning and projection. These maintain most of the good performance of our conditional procedures in cases for which the winner's curse does not arise, but are subsets of (conservative) projection intervals by construction, and so limit the maximal under-performance relative to projection confidence sets. We also introduce hybrid estimators which allow a controlled degree of bias while limiting the deviation from the conventional estimate.

Since we are not aware of any other procedures with guaranteed validity conditional on the target parameter and since our conditional procedures are optimal in this class, our simulations focus on unconditional performance. The simulation designs are based on empirical welfare maximization applications from Kitagawa and Tetenov (2018) and tipping point applications from Card et al. (2008). In both settings, we find that while our conditional procedures exhibit good unconditional performance in cases where the objective function determining the target parameter has a clear, well-separated optimum, their unconditional performance can be quite poor in other cases, including in calibrations to the data. By contrast, our hybrid procedures perform quite well: our hybrid confidence sets are shorter than the previously available alternative (projection intervals) in all specifications, and are shorter than conditional intervals in all but the well-separated case (where they are nearly tied). Hybrid estimators eliminate nearly all the bias of conventional estimates, and are much less dispersed than our exactly median unbiased estimates. These results show that while optimal conditional inference is possible, conditional validity can come at the cost of unconditional performance. By combining conditional and projection approaches our hybrid procedures yield better performance than either, and offer a substantial improvement over existing alternatives.

This paper is related to the literature on tests of superior predictive performance (e.g. White (2000); Hansen (2005); Romano and Wolf (2005)). This 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.[1] 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.

As suggested above, our results are also closely related to the growing literature on selective inference. Fithian et al. (2017) describe a general conditioning approach

[1] As noted above, Romano and Wolf (2005) also propose a version of projection intervals.

applicable to a wide range of settings, while a rapidly growing literature including e.g. Lee et al. (2016); Harris et al. (2016); Tian and Taylor (2016) works out the details of this approach for a range of settings. Likewise, our conditional confidence sets examine the implications of the conditional approach in our setting. In a particularly related paper, Tian et al. (2016) consider inference conditional on the solution to a penalized convex optimization problem falling in a given set, though neither our setting nor theirs nests the other.

Beyond the new setting considered, we make three main contributions relative to the selective inference literature. First, we observe that the same structure used to develop optimal conditional confidence sets also allows construction of optimal quantile unbiased estimators using the results of Pfanzagl (1994). Second, we note that conditional validity, as generally imposed in this literature, may come at a substantial cost of unconditional performance, relative to unconditional alternatives. Finally, for settings where unconditional validity is sufficient we introduce novel hybrid procedures which outperform both conditional procedures and the available unconditional alternatives.

In the next section, we begin by introducing the problem we consider, and the techniques we propose, in the context of a toy example. Section 3 introduces the normal model in which we develop our main results, and shows how it arises as an asymptotic approximation to empirical welfare maximization and structural break examples. Section 4 develops our optimal conditional procedures, discusses their properties, and compares them to sample splitting. Section 5 introduces projection confidence intervals and our hybrid procedures. Finally, Sections 6 and 7 report results for simulations calibrated to empirical welfare maximization and tipping point applications, respectively. All proofs, along with other supporting material, are given in the supplement.

2 A Stylized Example

We begin by illustrating the problem we consider, along with the solutions we propose, in a stylized example based on Manski (2004). In the treatment choice problem of Manski (2004) a treatment rule assigns treatments to subjects based on observable characteristics. Given a social welfare criterion and (quasi-)experimental data, Kitagawa and Tetenov (2018) propose what they call empirical welfare maximization

(EWM), which selects the treatment rule that maximizes the sample analog of the social welfare criterion over a class of candidate rules.

For simplicity suppose there are only two candidate policies: θ_1 corresponding to "treat everyone" and θ_2 corresponding to "treat no one." Suppose further that our social welfare function is the average of an outcome variable Y. If we have a sample of independent observations $i \in \{1, ..., n\}$ from a randomized trial where a binary treatment $D_i \in \{0, 1\}$ is randomly assigned to subjects with $Pr\{D_i = 1\} = d$, then as in Kitagawa and Tetenov (2018) the scaled empirical welfare under (θ_1, θ_2) is

$$(X_n(\theta_1), X_n(\theta_2)) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i Y_i}{d}, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(1 - D_i) Y_i}{1 - d} \right).$$

EWM selects the rule $\hat{\theta} = \text{argmax}_{\theta \in \{\theta_1, \theta_2\}} X_n(\theta)$.[2]

Kitagawa and Tetenov (2018) show that the welfare from the policy selected by EWM converges to the optimal social welfare at the minimax optimal rate, providing a strong argument for this approach. As noted by Kitagawa and Tetenov (2018), however, even after choosing a policy we often want estimates and confidence intervals for its implied social welfare in order to learn about the size of policy impact and to communicate with stakeholders (rather than finding or revising a policy recommendation). For a fixed policy θ, the empirical welfare $X_n(\theta)$ is unbiased for the true (scaled) social welfare $\mu_n(\theta)$ from the corresponding policy.[3] By contrast, the empirical welfare of the estimated optimal policy, $X_n(\hat{\theta})$, is biased upwards relative to the true social welfare $\mu_n(\hat{\theta})$ since we are more likely to select a given policy when the empirical welfare overestimates the true welfare. Likewise confidence sets for $\mu_n(\hat{\theta})$ that ignore estimation of θ may cover $\mu_n(\hat{\theta})$ less often than we intend. This is a form of the 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. Under mild conditions the central limit theorem implies

[2] If the summands are weighted by the sample propensity scores instead, we obtain Manski's conditional empirical success rule and the asymptotically optimal rules of Hirano and Porter (2009) with a symmetric loss.

[3] $X_n(\theta)$ is exactly mean-unbiased, while the asymptotic approximation (1) below shows that it is asymptotically median-unbiased.

that our estimates of social welfare are asymptotically normal,

$$\begin{pmatrix} X_n(\theta_1) - \mu_n(\theta_1) \\ X_n(\theta_2) - \mu_n(\theta_2) \end{pmatrix} \Rightarrow N\left(0, \begin{pmatrix} \Sigma(\theta_1) & \Sigma(\theta_1,\theta_2) \\ \Sigma(\theta_1,\theta_2) & \Sigma(\theta_2) \end{pmatrix}\right) \tag{1}$$

where the asymptotic variance can be consistently estimated, while the scaled social welfare μ_n cannot be. To simplify the analysis, for this section we assume that $\Sigma(\theta_1,\theta_2) = 0$.[4] Motivated by (1), we abstract from approximation error and assume that we observe

$$\begin{pmatrix} X(\theta_1) \\ X(\theta_2) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(\theta_1) \\ \mu(\theta_2) \end{pmatrix}, \begin{pmatrix} \Sigma(\theta_1) & 0 \\ 0 & \Sigma(\theta_2) \end{pmatrix}\right)$$

for $\Sigma(\theta_1)$ and $\Sigma(\theta_2)$ known, and that $\hat{\theta} = \text{argmax}_{\theta\in\Theta} X(\theta)$, $\Theta = \{\theta_1, \theta_2\}$.

As suggested above, $X(\hat{\theta})$ is biased upwards as an estimator of $\mu(\hat{\theta})$. Indeed, this bias arises both conditional on $\hat{\theta}$ and unconditionally. To see this note that $\hat{\theta} = \theta_1$ if $X(\theta_1) > X(\theta_2)$, where we ignore ties (which occur with zero probability). Conditional on $\hat{\theta} = \theta_1$ and $X(\theta_2)$, $X(\theta_1)$ follows a normal distribution truncated below at $X(\theta_2)$. Since this holds for all $X(\theta_2)$, $X(\theta_1)$ has positive median bias conditional on $\hat{\theta} = \theta_1$,[5]

$$Pr_\mu\left\{X(\hat{\theta}) \geq \mu(\hat{\theta}) | \hat{\theta} = \theta_1\right\} > \frac{1}{2} \text{ for all } \mu. \tag{2}$$

Since the same argument holds for θ_2, $\hat{\theta}$ is likewise biased upwards unconditionally

$$Pr_\mu\left\{X(\hat{\theta}) \geq \mu(\hat{\theta})\right\} > \frac{1}{2} \text{ for all } \mu. \tag{3}$$

Note that (3) differs from (2) in that the target parameter is random, due to its dependence on $\hat{\theta}$. Unsurprisingly given this bias, the conventional confidence set which adds and subtracts a quantile of the standard normal distribution times the standard error need not have correct coverage.

To illustrate these issues, Figure 1 plots the coverage of conventional confidence sets as well as the median bias of conventional estimates, in an example with $\Sigma(\theta_1) = \Sigma(\theta_2) = 1$. For comparison we also consider cases with ten and fifty policies, $|\Theta| = 10$

[4] One can show that $\Sigma(\theta_1,\theta_2) = -\mu(\theta_1)\mu(\theta_2)$, so this restriction arises naturally if one models μ as shrinking with the sample size to keep it on the same order as sampling uncertainty, $\mu_n = \frac{1}{\sqrt{n}}\mu^*$.

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

and $|\Theta| = 50$, where we again set $\Sigma(\theta) = 1$ for all θ and for ease of reporting assume that all the policies other than the first are equally effective, $\mu(\theta_2) = \mu(\theta_3) = ... = \mu(\theta_{-1})$. The first panel of Figure 1 shows that while the conventional confidence set appears to have reasonable coverage when there are only two policies, its coverage can fall substantially when $|\Theta| = 10$ or $|\Theta| = 50$.[6] 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, and 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 $|\Theta| = 2$ for simplicity, let $F_{TN}(x(\theta_1); \mu(\theta_1), x(\theta_2))$ denote the truncated normal distribution function for $X(\theta_1)$ truncated below at $x(\theta_2)$ when the true social welfare for θ_1 is $\mu(\theta_1)$. For fixed $(x(\theta_1), x(\theta_2))$ this function is strictly decreasing in $\mu(\theta_1)$, and for $\hat{\mu}_\alpha$ the solution to $F_{TN}(X(\theta_1); \hat{\mu}_\alpha, x(\theta_2)) = 1 - \alpha$, Proposition 1 below shows that

$$Pr_\mu\left\{\hat{\mu}_\alpha \geq \mu(\hat{\theta}) | \hat{\theta} = \theta_1\right\} = \alpha \text{ for all } \mu.$$

Hence, $\hat{\mu}_\alpha$ is α-quantile unbiased for $\mu(\hat{\theta})$ conditional on $\hat{\theta} = \theta_1$, and the analogous statement likewise holds conditional on $\hat{\theta} = \theta_2$. Indeed, Proposition 1 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 $1 - \alpha$

$$Pr\left\{\mu(\hat{\theta}) \in CS | \hat{\theta} = \theta_j\right\} \geq 1 - \alpha \text{ for } j \in \{1, 2\} \text{ and all } \mu. \tag{4}$$

While the equal-tailed confidence interval is easy to compute, there are other confidence sets available in this setting. As in Lehmann and Scheffé (1955) and Fithian et al. (2017) it is possible to construct a uniformly most accurate unbiased (UMAU) confidence set, CS_U, conditional on $\hat{\theta}$, i.e., the average length of CS_U is shortest among any unbiased confidence sets conditional on $\hat{\theta}$. To construct CS_U, we collect the parameter values not rejected by a uniformly most powerful unbiased test

[6] For example, these could correspond to cases where we consider "treat no one" along with nine or forty-nine different treatments, respectively.

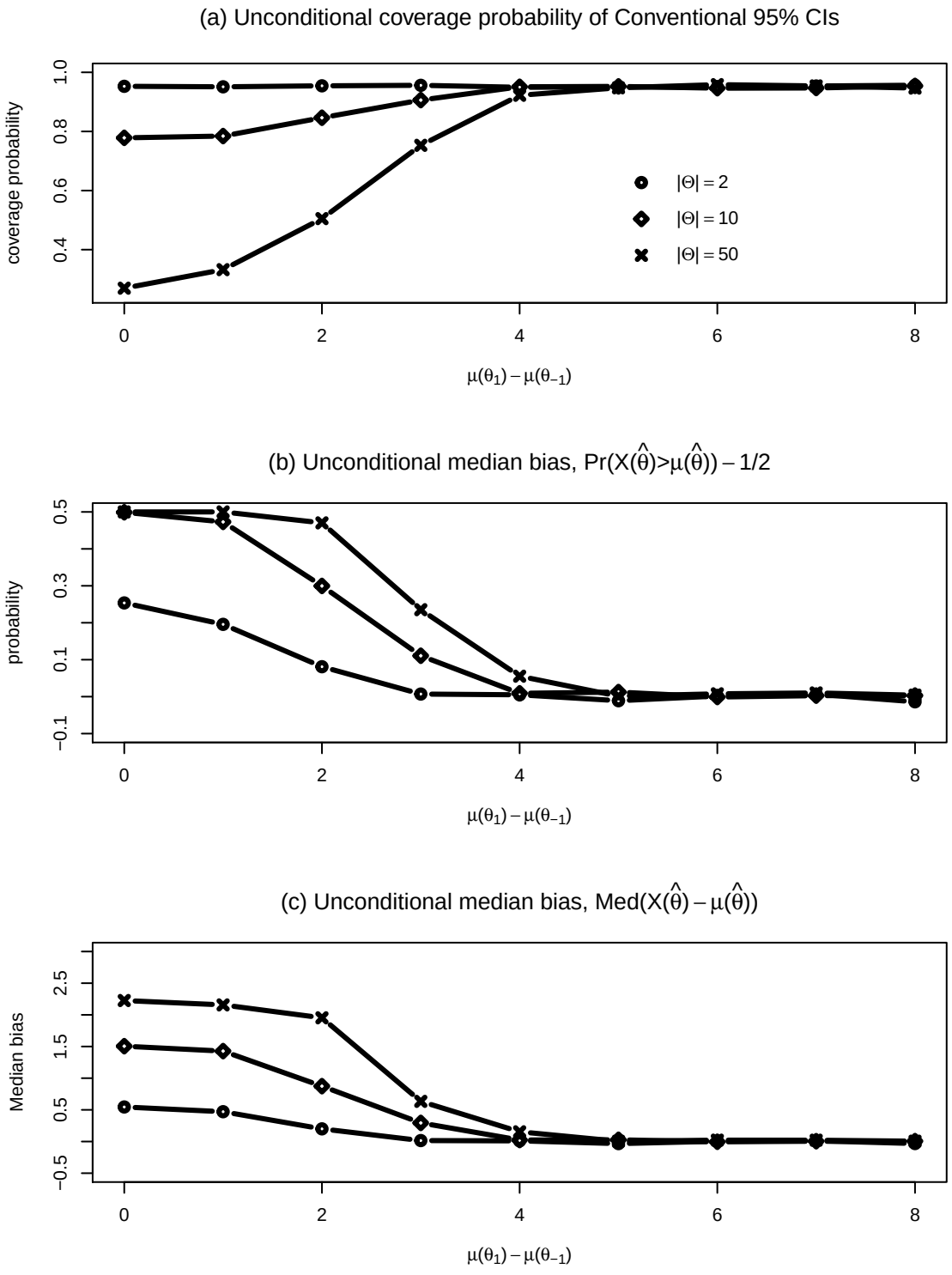

Figure 1: Performance of conventional procedures in examples with 2, 10, and 50 policies.

conditional on $\hat{\theta}$. While straightforward to implement, the exact form of this test is somewhat involved and so is deferred to Section 4 below. The equal tailed confidence set CS_{ET} is not unbiased, so there is not a clear ranking between CS_{ET} and CS_U.

Both CS_{ET} and CS_U have conditional coverage $1-\alpha$, and so by the law of iterated expectations have unconditional coverage $1-\alpha$ as well

$$Pr_\mu\left\{\mu(\hat{\theta}) \in CS\right\} \geq 1-\alpha \text{ for all } \mu. \tag{5}$$

Unconditional coverage is easier to attain, so relaxing the coverage requirement from (4) to (5) may yield tighter confidence sets in some cases. Conditional and unconditional coverage requirements address different questions, however, and which is more appropriate depends on the problem at hand. In the EWM problem, for instance, a policy maker who is told the recommended policy $\hat{\theta}$ along with a confidence interval may want the confidence interval to be valid conditional on the recommendation, which is precisely the conditional coverage requirement (4). In particular, this ensures that if one considers repeated instances in which EWM recommends a particular course of action (e.g. departure from the status quo), reported confidence sets 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 sets cover the true value with probability at least $1-\alpha$ on average across a range of recommendations, it suffices to impose the unconditional requirement (5).

We are unaware of prior work which ensures conditional coverage (4).[7] For unconditional coverage (5), however, Kitagawa and Tetenov (2018) propose an unconditional confidence set based on projecting a simultaneous confidence band for μ to obtain a confidence set for $\mu(\hat{\theta})$. 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

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

this set has correct unconditional coverage (5). Figure 2 plots the average (unconditional) length of 95% confidence sets CS_{ET}, CS_U, CS_P, along with the conventional confidence set, again in cases with $|\Theta| \in \{2, 10, 50\}$. As this figure illustrates, CS_{ET}

[7] As noted in the introduction and further discussed in Section 4.3 below, split-sample confidence intervals also have conditional coverage, but change the definition of $\hat{\theta}$.

and CS_U are shorter than CS_P when $|\mu(\theta_1)-\mu(\theta_{-1})|$ exceeds five, and in fact converge to the conventional interval as $|\mu(\theta_1)-\mu(\theta_{-1})|$ tends to infinity. When $|\mu(\theta_1)-\mu(\theta_{-1})|$ is small, on the other hand, CS_{ET} and CS_U can be substantially wider than CS_P. Both features become more pronounced as we increase the number of policies considered. In Figure 3 we plot the mean absolute error $E_\mu\left[|\hat{\mu}-\mu(\hat{\theta})|\right]$ for different estimators in this design, and find that the median-unbiased estimator likewise exhibits substantially larger mean absolute error than the conventional estimator $X(\hat{\theta})$ when $|\mu(\theta_1)-\mu(\theta_{-1})|$ is small.

Recall that CS_U is the optimal unbiased confidence set, while the endpoints of CS_{ET} are optimal quantile unbiased estimators. So long as we impose correct conditional coverage (4) and unbiasedness, there is therefore no scope to improve unconditional performance. If, on the other hand, we require only correct unconditional coverage (5), as for CS_P, improved unconditional performance is possible.

To improve performance, we consider hybrid confidence sets CS^H_{ET} and CS^H_U. As detailed in Section 5.2 below, these confidence sets are constructed analogously to CS_{ET} and CS_U, but further condition on the event that the true social welfare falls in the level $1-\beta$ projection interval CS^β_P for $\beta<\alpha$. This ensures that the hybrid confidence sets are never longer than the level $1-\beta$ unconditional interval, and so limits the performance deterioration when $|\mu(\theta_1)-\mu(\theta_2)|$ is small. These hybrid confidence sets have correct unconditional coverage (5), but do not in general have correct conditional coverage (4). 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 cases with 10 and 50 policies, the hybrid confidence sets have shorter average length than the unconditional interval CS_P for all parameter values considered. In Figure 3 we report results for a hybrid estimation procedure based on a similar approach (detailed in Section 5.3 below), and again find substantial performance improvements.

The improved unconditional performance of the hybrid intervals is achieved by requiring only unconditional rather than conditional, coverage. In particular, the projection confidence set CS_P and the hybrid confidence sets CS^H_{ET} and CS^H_U do not have correct conditional coverage (4). To illustrate, Figure 4 plots the conditional coverage of the intervals CS_U and CS_{ET} given $\hat{\theta}=\theta_1$ in the case with two policies, along with that of the projection and hybrid intervals. As expected, the conditional intervals have correct conditional coverage, while the hybrid and projection intervals

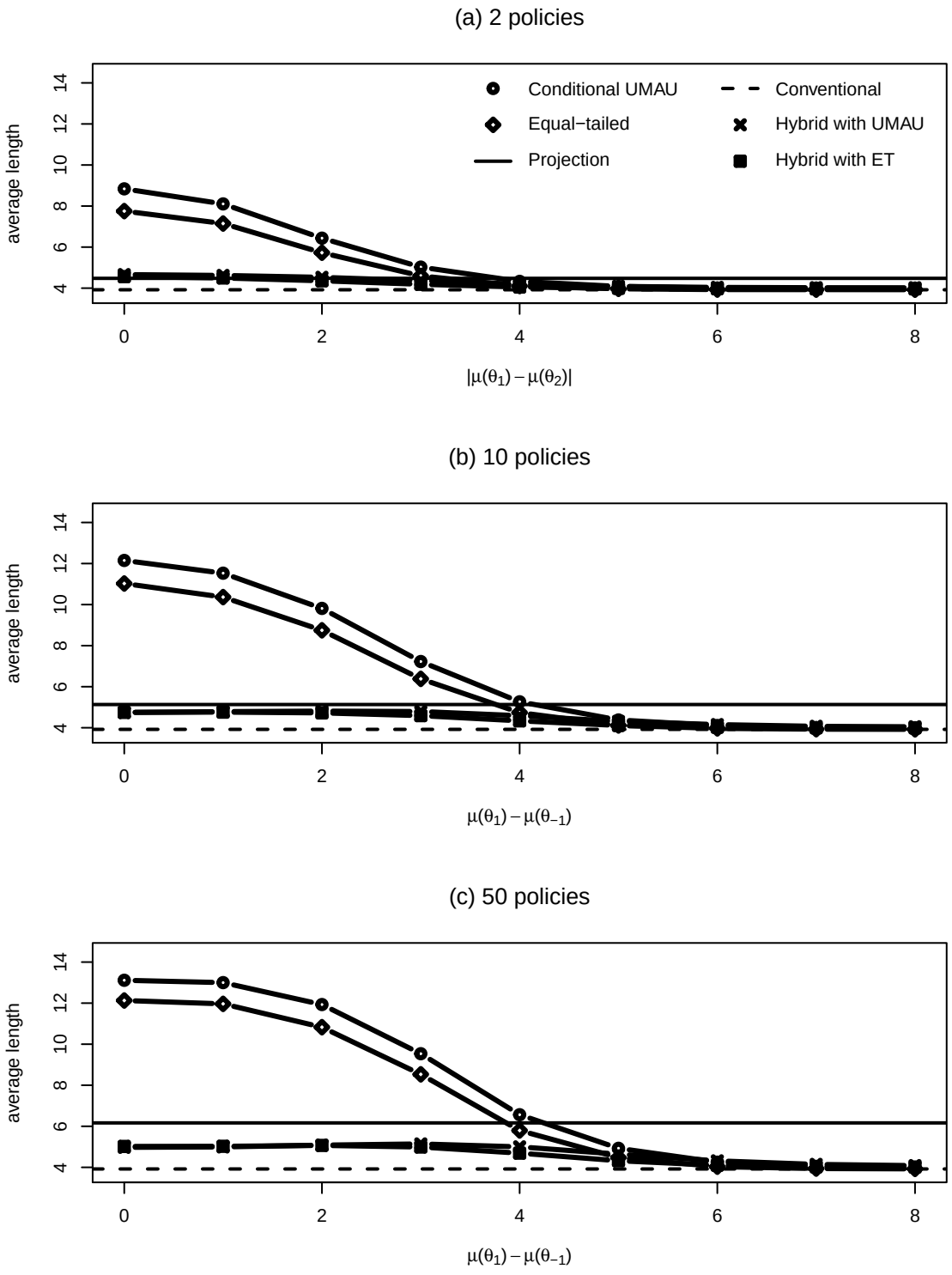

Figure 2: Average length of confidence sets for $\mu(\hat{\theta})$ in cases with 2, 10, and 50 policies.

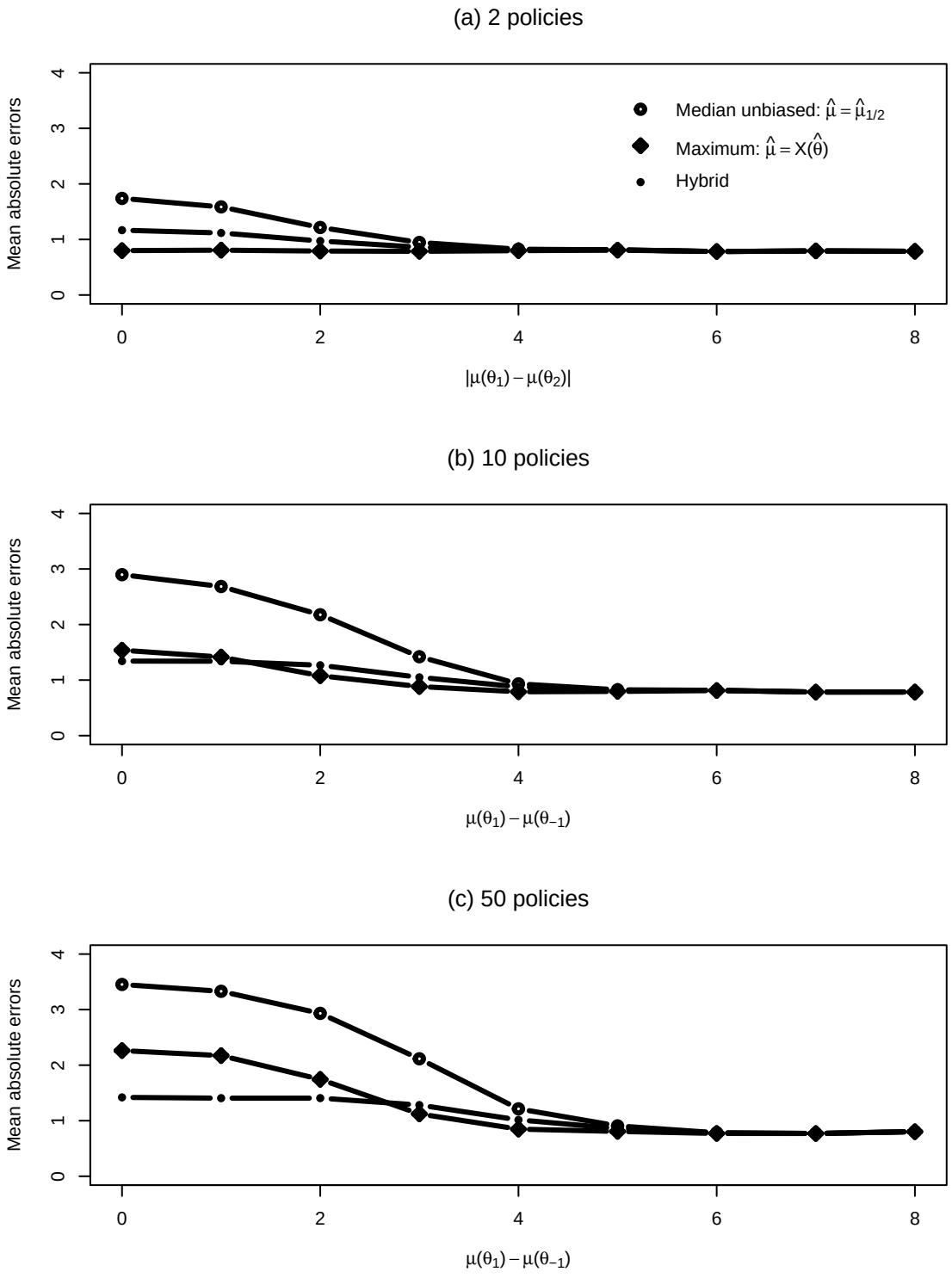

Figure 3: Mean absolute error of estimators of $\mu(\hat{\theta})$ in cases with 2, 10, and 50 policies.

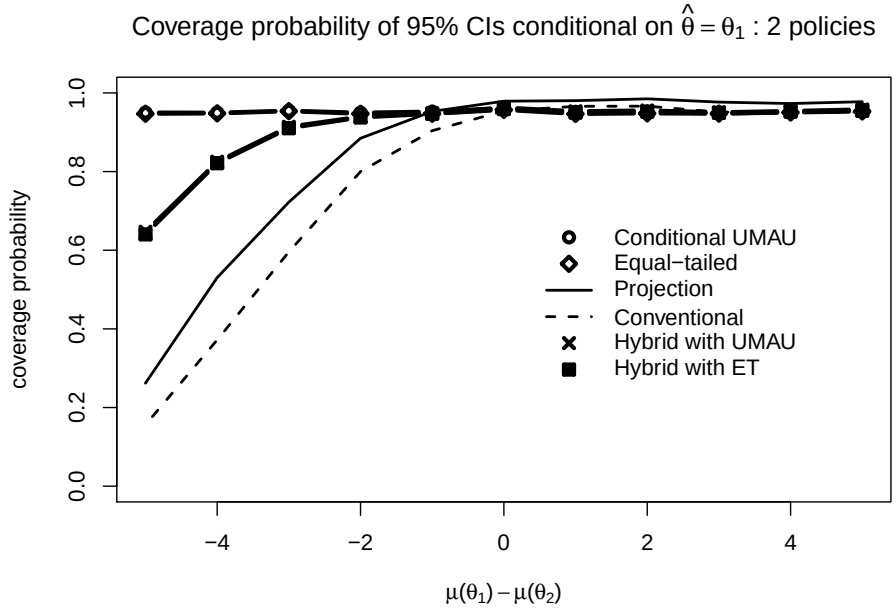

Figure 4: Conditional coverage in case with two policies.

do not. Coverage distortions appear when $\mu(\theta_1) \ll \mu(\theta_2)$. In this case $\hat{\theta} = \theta_2$ with high probability, but we will nonetheless sometimes have $\hat{\theta} = \theta_1$, and conditional on this event $X(\theta_1)$ will be far away from $\mu(\theta_1)$ with high probability. Hence, conditional on this event, projection and hybrid intervals under-cover.

3 Setting

This section introduces our general setting, which extends the stylized example of the previous section in several directions. We assume that we observe normal random vectors $\left(X\left(\theta\right)', Y\left(\theta\right)\right)'$ for $\theta \in \Theta$ with Θ a finite set, $X\left(\theta\right) \in \mathbb{R}^{d_X}$, and $Y\left(\theta\right) \in \mathbb{R}^1$. In particular, for $\Theta = \left\{\theta_1, ..., \theta_{|\Theta|}\right\}$, let $X = \left(X\left(\theta_1\right)', ..., X\left(\theta_{|\Theta|}\right)'\right)'$ and $Y = \left(Y\left(\theta_1\right), ..., Y\left(\theta_{|\Theta|}\right)\right)'$. Then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\mu, \Sigma\right) \tag{6}$$

for

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

and

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

We assume that the covariance function Σ is known, while the mean function μ is unknown and unrestricted unless noted otherwise. As above, we will show that this model arises naturally as an asymptotic approximation in a range of examples. For simplicity of exposition we assume throughout that $\Sigma_Y(\theta) = \Sigma_Y(\theta, \theta) > 0$ for all $\theta \in \Theta$, since otherwise there is no inference problem conditional on $\hat{\theta} = \theta$.

We are interested in inference on $\mu_Y(\hat{\theta})$, where $\hat{\theta}$ is determined based on X. We define $\hat{\theta}$ through either the *level maximization* problem where (for $d_X = 1$)

$$\hat{\theta} = \underset{\theta \in \Theta_F}{\operatorname{argmax}} \; X(\theta), \tag{7}$$

or the *norm maximization* problem where (for $d_X \geq 1$)

$$\hat{\theta} = \underset{\theta \in \Theta_F}{\operatorname{argmax}} \; \|X(\theta)\|, \tag{8}$$

where $\|\cdot\|$ denotes the Eulidean norm.[8] We will again be interested in constructing confidence sets for $\mu_Y(\hat{\theta})$ that are valid either conditional on the value of $\hat{\theta}$ or unconditionally, as well as median-unbiased estimates. We may also want to condition on some additional event $\hat{\gamma} = \tilde{\gamma}$, for $\hat{\gamma} = \gamma(X)$ a function of X which takes values in the finite set Γ. In such cases, we aim to construct confidence sets for $\mu_Y(\hat{\theta})$ which are valid conditional on the pair $(\hat{\theta}, \hat{\gamma})$. Examples of such additional conditioning events are discussed in examples below.

In the remainder of this section, we show how this class of problems arises in examples and discuss the choice between conditional and unconditional confidence sets in each case. We first revisit the empirical welfare maximization problem in a more general setting and show that it gives rise to the level maximization problem (7)

[8]For simplicity of notation we will assume $\hat{\theta}$ is unique unless noted otherwise. Our analysis does not rely on this assumption, however: see footnote 13 below.

asymptotically. We then discuss structural break models, and show that they reduce to the norm maximization problem (8) asymptotically. We also briefly discuss other examples giving rise to level and norm maximization problems.

Empirical Welfare Maximization In the empirical welfare maximization problem of Kitagawa and Tetenov (2018), as in the last section we aim to select a welfare-maximizing treatment rule from a set of policies Θ. Let us assume that we have a sample of independent observations $i \in \{1, ..., n\}$ from a randomized trial where treatment is randomly assigned conditional on observables C_i with $Pr\{D_i = 1|C_i\} = d(C_i)$. We consider policies that assign units to treatment based on the observables, where rule θ assigns i to treatment if and only if $C_i \in \mathcal{C}_\theta$. The scaled empirical welfare under policy θ is[9]

$$X_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{Y_i D_i}{d(C_i)} 1\{C_i \in \mathcal{C}_\theta\} + \frac{Y_i(1-D_i)}{1-d(C_i)} 1\{C_i \notin \mathcal{C}_\theta\} \right).$$

EWM again selects the policy that maximizes empirical welfare, $\hat{\theta} = \text{argmax}_{\theta \in \Theta_F} X_n(\theta)$.

The definition of Y_n in this setting depends on the object of interest. We may be interested in the overall social welfare, in which case we can define $Y_n = X_n$. Alternatively we could be interested in social welfare relative to the baseline of no treatment, in which case we can define $Y_n(\theta)$ as the difference in scaled empirical welfare between policy θ and the policy that treats no one, which we denote by $\theta = 0$,

$$Y_n(\theta) = X_n(\theta) - X_n(0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i D_i}{d(C_i)} - \frac{Y_i(1-D_i)}{1-d(C_i)} \right] 1\{C_i \in \mathcal{C}_\theta\}.$$

Likewise, we might be interested in the social welfare for a particular subgroup defined by the observables, say $\mathcal{S}$, in which case we can take

$$Y_n(\theta) = \frac{\sqrt{n} \sum_{i=1}^{n} \left(\frac{Y_i D_i}{d(C_i)} 1\{C_i \in \mathcal{S} \cap \mathcal{C}_\theta\} + \frac{Y_i(1-D_i)}{1-d(C_i)} 1\{C_i \in \mathcal{S} \setminus \mathcal{C}_\theta\} \right)}{\sum_{i=1}^{n} 1\{C_i \in \mathcal{S}\}}.$$

[9]Kitagawa and Tetenov (2018) primarily consider welfare relative to the baseline of no treatment, which yields the same optimal policy.

For $\mu_{X,n}$ and $\mu_{Y,n}$ the true scaled social welfare corresponding to X_n and Y_n,

$$\begin{pmatrix} X_n - \mu_{X,n} \\ Y_n - \mu_{Y,n} \end{pmatrix} \Rightarrow N(0, \Sigma) \tag{9}$$

under mild conditions, where the covariance Σ will depend on the data generating process and the definition of Y_n but is consistently estimable. By contrast, the scaling of X_n and Y_n means that $\mu_{X,n}$ and $\mu_{Y,n}$ are not consistently estimable. As in the last section, this suggests the asymptotic problem where we observe normal random vectors (X, Y) as in (6), Σ is known, and $\hat{\theta}$ is defined as in (7) so we consider the level maximization problem.[10]

As argued in the last section, if a policy maker is told the recommended policy $\hat{\theta}$ as well as a confidence set for $\mu_Y(\hat{\theta})$, it is natural to require that the confidence set be valid conditional on the recommendation. It may also be natural to condition on additional variables. For example, if a recommendation is made only when we reject the null hypothesis that no policy in Θ improves outcomes over the base case of no treatment, $H_0 : \max_{\theta \in \Theta_F} (\mu(\theta) - \mu(0)) \leq 0$, then it is also natural to condition inference on this rejection.[11] To cover this case we can define $\hat{\gamma} = \gamma(X)$ as a dummy for rejection of H_0. If on the other hand we care only about performance on average across a range of recommendations we need only impose unconditional coverage. $\triangle$

The level maximization problem arises in a number of other settings as well. For example, the literature on tests of superior predictive performance (c.f. White (2000); Hansen (2005); Romano and Wolf (2005)) considers the problem of testing whether some trading strategies or forecasting rules amongst a candiate set beat a benchmark. If we define $X_n = Y_n$ as the vector of performance measures for different strategies, X_n is asymptotically normal under mild conditions (see e.g. Romano and Wolf (2005)). If one wants to form a confidence set for the performance of the "best" strategy based on X_n (perhaps also conditioning on the result of a test for superior performance) this reduces to our level maximization problem asymptotically.

Another example comes from Bhattacharya (2009) and Graham et al. (2014), who

[10] Under mild regularity conditions, property (9) also holds in settings where the empirical welfare involves estimated propensity scores and/or estimated outcome regressions, e.g., the hybrid procedures of Kitagawa and Tetenov (2018) and the doubly robust welfare estimators used in Athey and Wager (2018).

[11] In case of $|\Theta| = 2$, conditioning on this rejection can be interpreted as conditioning on that the decision criterion of Tetenov (2012) supports the same policy.

considers the problem of optimally dividing individuals into groups to maximize peer effects. For X_n again a scaled objective function, the results of Bhattacharya (2009) show that his problem reduces to our level maximization problem asymptotically when one considers a finite set of assignments. More broadly, any time we consider an m-estimation problem with a finite parameter space and are interested in the value of the population objective or some other function at the estimated optimal value, this falls into our level maximization framework under mild conditions.

We next discuss an example of structural break estimation, showing that it gives rise to our level-maximization problem asymptotically.

Structural Break Estimation Suppose we observe time-series data on an outcome Y_t and a k-dimensional vector of regressors C_t for $t \in \{1, ..., T\}$. We assume there is a linear but potentially time-varying relationship between Y_t and C_t,

$$Y_t = C_t' \left(\beta + \varphi_T(t/T)\right) + U_t \tag{10}$$

where the residuals U_t are orthogonal to C_t. Similarly to Elliott and Müller (2014) the function $\varphi_T : [0, 1] \to \mathbb{R}^k$ determines the value of the time-varying coefficient $\beta + \varphi_T(t/T)$. This model nests the traditional structural break model (see e.g. Hansen (2001), Perron (2006), and references therein) by taking

$$\varphi_T(t/T) = 1(t/T > \theta)\delta, \tag{11}$$

where $\theta = \tau/T \in [0, 1]$ is the true "break fraction" and τ is the true break date. The model (10) is more general however, and allows the possibility that there are multiple breaks, that the parameters change continuously at times, or both.

The structural break model is widely used in practice, so we consider a researcher who fits the model (11). To allow the possibility of misspecification, however, we assume only that the data is generated by (10). To provide a good asymptotic approximation to finite sample behavior, we follow Elliott and Müller (2007) and Elliott and Müller (2014) and model parameter instability as on the same order as sampling uncertainty, with $\varphi_T(t/T) = \frac{1}{\sqrt{T}} g(t/T)$ for a fixed function g. We further assume that

$$\frac{1}{T}\sum_{t=1}^{[\theta T]} C_t C_t' \to_p \theta \Sigma_C, \quad \frac{1}{T}\sum_{t=1}^{[\theta T]} C_t C_t' g(t/T) \to_p \Sigma_{Cg}(\theta), \tag{12}$$

and

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[\theta T]} C_t U_t \Rightarrow \Omega^{1/2} W(\theta), \tag{13}$$

all uniformly in $0 \leq \theta \leq 1$. Here Σ_C is a full rank matrix, $\Sigma_{Cg} : [0,1] \to \mathbb{R}^k$ is a vector-valued function, Ω is a symmetric positive definite matrix which we assume is consistently estimable, and $W(\cdot)$ is a standard k-dimensional Wiener process. Condition (12) is a special case of Conditions 1(ii) and 1(iv) of Elliott and Müller (2007) and can be relaxed to their conditions at the expense of extra notation, while condition (13) is implied by a functional central limit theorem under mild assumptions.

The standard break-fraction estimator $\hat{\theta}$ chooses θ to minimize the sum of squared residuals in an OLS regression of Y_t on C_t and $\mathbf{1}(t/T > \theta)C_t$. If we define

$$X_T(\theta) = \begin{pmatrix} \left(\sum_{t=1}^{[\theta T]} C_t C_t'\right)^{-\frac{1}{2}} \left(\sum_{t=1}^{[\theta T]} C_t \eta_t\right) \\ \left(\sum_{t=[\theta T]+1}^{T} C_t C_t'\right)^{-\frac{1}{2}} \left(\sum_{t=[\theta T]+1}^{T} C_t \eta_t\right) \end{pmatrix},$$

for $\eta_t \equiv U_t + T^{-1/2} C_t' g(t/T)$, the proof of Proposition 1 in Elliott and Müller (2007) implies that

$$\hat{\theta} = \underset{\theta \in \Theta_T}{\operatorname{argmax}} \ \|X_T(\theta)\| + o_p(1) \tag{14}$$

for $\Theta_T = \left\{\theta \in \{\frac{1}{T}, \frac{2}{T}, ...1\} : \tilde{\lambda} \leq \theta \leq 1 - \tilde{\lambda}\right\}$, $\tilde{\lambda} \in (0, 1/2)$ a user-selected tuning parameter, and $o_p(1)$ an asymptotically negligible term. Hence, $\hat{\theta}$ is asymptotically equivalent to the solution to a norm-maximization problem analogous to (8).

Suppose we are interested in the break in the jth parameter, $\delta_j = e_j'\delta$ for e_j the jth standard basis vector.[12] In practice it is common to estimate δ by least squares imposing the estimated break date $\hat{\theta}$. When the structural break model (11) is misspecified, however, there is neither a "true" break fraction θ nor a "true" break coefficient δ. Instead, the population regression coefficient $\delta(\theta)$ imposing break fraction θ depends on θ. Thus, for break fraction θ, the coefficient of interest is $\delta_j(\theta)$. Denote the OLS estimate imposing break fraction θ by $\hat{\delta}(\theta)$, and define $Y_T(\theta) = \sqrt{T}\hat{\delta}_j(\theta)$. If we define $\mu_{Y,T}(\theta) = \sqrt{T}\delta_j(\theta)$ as the scaled coefficient of interest and

[12] By changing the definition of Y_T below, our results likewise apply to the pre-break parameters β_j and the post-break parameters $\beta_j + \delta_j$, amongst other possible objects of interest.

$\mu_{X,T}(\theta)$ as the asymptotic mean of $X_T(\theta)$, Section B.2 of the Appendix shows that

$$\begin{pmatrix} X_T(\theta) - \mu_{X,T}(\theta) \\ Y_T(\theta) - \mu_{Y,T}(\theta) \end{pmatrix} \Rightarrow N\left(0, \Sigma(\theta)\right) \tag{15}$$

uniformly over the sequence of parameter spaces Θ_T, where the covariance function Σ is consistently estimable but $\mu_{X,T}(\theta)$ and $\mu_{Y,T}(\theta)$ are not. As before, this suggests the asymptotic problem (6), where we now define $\hat{\theta}$ through norm maximization (8).

Since the estimated break fraction $\hat{\theta}$ is random, and the parameter of interest $\delta_j(\theta)$ depends on θ, it is important to account for this randomness in our inference procedures. In particular, it may be appealing to condition inference on the estimated break date $\hat{\theta}$, since we only seek to conduct inference on $\delta(\tilde{\theta})$ when $\hat{\theta} = \tilde{\theta}$. It may also be natural to condition inference on additional variables. For example, if we report a confidence set for the break magnitude $\delta(\hat{\theta})$ only when we reject the null hypothesis of parameter constancy, $H_0 : \varphi_T(\theta) = 0$ for all θ, it is natural to condition inference on this rejection. As above, this can be accomplished by defining $\hat{\gamma} = \gamma(X)$ as a dummy for rejection of H_0, and conditioning inference on $(\hat{\theta}, \hat{\gamma})$. Even if we only desire coverage of $\delta(\hat{\theta})$ on average over the distribution of $\hat{\theta}$, and so prefer to consider unconditional confidence sets, accounting for the randomness of $\hat{\theta}$ remains important. If on the other hand we are confident that the break model is correctly specified, so (11) holds, it will typically be more appealing to focus on inference for the "true" parameters as in Elliott and Müller (2014). $\triangle$

While our discussion of structural break estimation focuses on the linear model (10), Elliott and Müller (2014) show that structural break estimation in nonlinear GMM models with time-varying parameters gives rise to the same asymptotic problem. Hence, our results apply in that setting as well. Likewise, Wang (2017) shows that the same asymptotic problem arises in threshold models, including the tipping-point model of Card et al. (2008) that we study below. Further afield, one could generalize our approach to consider norm-minimzation rather than norm-maximization, and so derive results for general GMM-type problems with finite parameter spaces.

4 Conditional Inference

This section develops conditional inference procedures for our general setting. We seek confidence sets with correct coverage conditional on $\hat{\theta}$, potentially along with some other conditioning variable $\hat{\gamma}$,

$$Pr\left\{\mu_Y(\hat{\theta}) \in CS | \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\} \geq 1 - \alpha \text{ for all } \tilde{\theta} \in \Theta, \tilde{\gamma} \in \Gamma, \text{ and all } \mu. \tag{16}$$

As in the stylized example of Section 2, we derive both equal tailed and uniformly most powerful unbiased confidence sets.[13] We also derive optimal conditionally α-quantile unbiased estimators, which for $\alpha \in (0, 1)$ satisfy

$$Pr_\mu\left\{\hat{\mu}_\alpha \geq \mu_Y(\hat{\theta}) | \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\} = \alpha \text{ for all } \tilde{\theta} \in \Theta, \tilde{\gamma} \in \Gamma, \text{ and all } \mu. \tag{17}$$

To implement our conditional procedures we need to know the form of particular conditioning events. We derive these conditioning events for our level and norm maximization settings, and illustrate in our empirical welfare maximization and structural break examples. We then discuss sample splitting as an alternative conditional inference approach, and following Fithian et al. (2017) note that conventional sample splitting procedures are dominated.

4.1 Optimal Conditional Inference

Since $\hat{\theta}$ and $\hat{\gamma}$ are functions of X, we can re-write the conditioning event as

$$\left\{X : \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\} = \mathcal{X}(\tilde{\theta}, \tilde{\gamma}).$$

Thus, for conditional inference we are interested in the distribution of (X, Y) conditional on $X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})$. Our results below imply that under mild conditions, the elements of Y other than $Y(\tilde{\theta})$ do not help in constructing a quantile unbiased estimate or unbiased confidence set for $\mu_Y(\hat{\theta})$ once we condition on $X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})$. Hence, we limit attention to the conditional distribution of $(X, Y(\tilde{\theta}))$ given $X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})$.

Since $(X, Y(\tilde{\theta}))$ is jointly normal unconditionally, it is truncated normal conditional on $X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})$. Correlation between X and $Y(\tilde{\theta})$ implies that the conditional

[13] If $\hat{\theta}$ is not unique we change the conditioning event $\hat{\theta} = \tilde{\theta}$ to $\tilde{\theta} \in \text{argmax } X(\theta)$ or $\tilde{\theta} \in \text{argmax } \|X(\theta)\|$ for the level and norm maximization problems, respectively.

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

$$Z_{\tilde{\theta}} = X - \left(\Sigma_{XY}(\cdot, \tilde{\theta})/\Sigma_Y(\tilde{\theta})\right) Y(\tilde{\theta}), \tag{18}$$

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 $\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, Z_{\tilde{\theta}} = z\right\}$ is truncated normal:

$$Y(\hat{\theta})|\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, Z = z \sim \xi|\xi \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z), \tag{19}$$

for $\xi \sim N\left(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta})\right)$ normally distributed and

$$\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z) = \left\{y : z + \left(\Sigma_{XY}(\cdot, \tilde{\theta})/\Sigma_Y(\tilde{\theta})\right) y \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})\right\} \tag{20}$$

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

Using this result, it is straightforward to construct quantile unbiased estimators for $\mu_Y(\hat{\theta})$. Let $F_{TN}(y; \mu_Y(\tilde{\theta}), \tilde{\theta}, \tilde{\gamma}, z)$ denote the distribution function for the truncated normal distribution (19). This distribution function is strictly decreasing in $\mu_Y(\tilde{\theta})$. Define $\hat{\mu}_\alpha$ as the unique solution to

$$F_{TN}(Y(\hat{\theta}); \hat{\mu}_\alpha, \tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}}) = 1 - \alpha. \tag{21}$$

Proposition 1 below shows that $\hat{\mu}_\alpha$ is conditionally α-quantile unbiased in the sense of (17), 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 $1 - \alpha$.

While we are interested in inference conditional on $\hat{\theta} = \tilde{\theta}$ and $\hat{\gamma} = \tilde{\gamma}$, our derivation of the quantile unbiased estimator $\hat{\mu}_\alpha$ further conditions on $Z_{\tilde{\theta}}$. Since conditioning reduces the amount of information available for inference, we might be concerned that this estimator is inefficient. Our next result, based on Pfanzagl (1979) and

Pfanzagl (1994), shows that this is not the case, and that $\hat{\mu}_\alpha$ is optimal in the class of quantile unbiased estimators in a strong sense.

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 open and convex. Otherwise, there exists some μ^* such that the parameter space for μ is an open convex subset of $\left\{\mu^* + \Sigma^{\frac{1}{2}} v : v \in \mathbb{R}^{\dim(X,Y)}\right\}$ for $\Sigma^{\frac{1}{2}}$ the symmetric square root of Σ.

This assumption requires that the parameter space for μ be sufficiently rich, in the sense of containing an open set in the appropriate space.[14] When Σ is degenerate (for example when X and Y are perfectly correlated as in the EWM example with $X = Y$), this assumption further implies that (X, Y) have the same support for all values of μ. This rules out that there exists a pair μ_1, μ_2 of parameter values which can be perfectly distinguished based on the data. Under this assumption, $\hat{\mu}_\alpha$ is an optimal quantile unbiased estimator.

Proposition 1
Let $\hat{\mu}_\alpha$ be the unique solution of (21). $\hat{\mu}_\alpha$ is conditionally α-quantile unbiased in the sense of (17). 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 increasing as d moves away from $\mu_Y(\tilde{\theta})$ for all fixed $\mu_Y(\tilde{\theta})$,

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

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

Proposition 1 shows that $\hat{\mu}_\alpha$ is optimal in the strong sense that it has lower risk (expected loss) for any loss function which is quasiconvex in the estimate for every true parameter value. Hence, the endpoints of CS_{ET} are optimal quantile unbiased estimators.

Rather than considering equal-tailed intervals, we can alternatively consider unbiased confidence sets. Following Lehmann and Romano (2005), we say that a level

[14] The assumption that the parameter space is open can be relaxed at the cost of complicating the statements below.

$1-\alpha$ two-sided confidence set 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 set 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 sets, a level $1-\alpha$ confidence set CS is unbiased if and only if $\phi = 1\left\{\mu_Y(\tilde{\theta}) \in CS\right\}$ is an unbiased test for the corresponding family of hypotheses.[15] 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 the optimal unbiased tests conditional on $\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, Z_{\tilde{\theta}} = z_{\tilde{\theta}}\right\}$. Thus, conditioning on $Z_{\tilde{\theta}}$ again does not come at the cost of power.

Optimal unbiased 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

$$\phi_{TS,\alpha}\left(\mu_{Y,0}\right) = 1\left\{Y(\tilde{\theta}) \notin \left[c_l\left(Z\right), c_u\left(Z\right)\right]\right\} \tag{22}$$

where $c_l\left(z\right)$, $c_u\left(z\right)$ solve

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

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

$$\phi_{OS-,\alpha}(\mu_{Y,0}) = 1\left\{F_{TN}(Y(\tilde{\theta}); \mu_{Y,0}, \tilde{\theta}, \tilde{\gamma}, z) \leq \alpha\right\} \tag{23}$$

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

$$\phi_{OS+,\alpha}(\mu_{Y,0}) = 1\left\{F_{TN}(Y(\tilde{\theta}); \mu_{Y,0}, \tilde{\theta}, \tilde{\gamma}, z) \geq 1-\alpha\right\}. \tag{24}$$

Proposition 2

If Assumption 1 holds, $\phi_{TS,\alpha}$, $\phi_{OS-,\alpha}$, and $\phi_{OS+,\alpha}$ are uniformly most powerful unbi-

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

ased 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 sets we collect the values not rejected by these tests. In particular, define the two-sided uniformly most accurate unbiased confidence set as

$$CS_U = \{\mu_{Y,0} : \phi_{TS,\alpha}(\mu_{Y,0}) = 0\},$$

and note that 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

$$CS_{U,-} = \{\mu_{Y,0} : \phi_{OS-,\alpha}(\mu_{Y,0}) = 0\} = (-\infty, \hat{\mu}_{1-\alpha}],$$

$$CS_{U,+} = \{\mu_{Y,0} : \phi_{OS+,\alpha}(\mu_{Y,0}) = 0\} = [\hat{\mu}_{\alpha}, \infty),$$

where we have used the definition of the one-sided tests to express these in terms of our quantile unbiased estimators. 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.

4.2 Conditioning Sets

Thus far we have left the conditioning events $\mathcal{X}(\tilde{\theta}, \tilde{\gamma})$ and $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z)$ unspecified. To implement our conditional procedures we need tractable representations of both. We first derive the form of these conditioning events for the level maximization problem (7) and the norm maximization problem (8) without additional conditioning variables $\hat{\gamma}$. We then discuss the effect of adding additional conditioning variables and illustrate in our EWM and structural break examples.

In level maximization problems without additional conditioning variables, we are interested in inference conditional on $X \in \mathcal{X}(\tilde{\theta})$ for

$$\mathcal{X}(\tilde{\theta}) = \left\{X : X(\tilde{\theta}) = \max_{\theta \in \Theta} X(\theta)\right\}.$$

The following result, based on Lemma 5.1 of Lee et al. (2016), derives the form of

$\mathcal{Y}(\tilde{\theta}, z)$ in this setting.

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

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

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

and

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

If $\mathcal{V}(\tilde{\theta}, z) \geq 0$, *then*

$$\mathcal{Y}(\tilde{\theta}, z) = \left[\mathcal{L}(\tilde{\theta}, z), \mathcal{U}(\tilde{\theta}, z) \right].$$

If $\mathcal{V}(\tilde{\theta}, z) < 0$, *then* $\mathcal{Y}(\tilde{\theta}, z) = \emptyset$.

Thus, the conditioning event $\mathcal{Y}(\tilde{\theta}, z)$ is an interval bounded above and below by easy-to-calculate functions of z. While we must have $\mathcal{V}(\tilde{\theta}, z) \geq 0$ for this interval to be non-empty, $Pr_\mu \left\{ \mathcal{V}(\hat{\theta}, Z_{\hat{\theta}}) < 0 \right\} = 0$ 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 $\mathcal{U}(\hat{\theta}, Z_{\hat{\theta}})$.

In the norm maximization problem the conditioning event is

$$\mathcal{X}(\tilde{\theta}) = \left\{ X : \|X(\tilde{\theta})\| = \max_{\theta \in \Theta} \|X(\theta)\| \right\}.$$

This conditioning event involves nonlinear constraints so the results of Lee et al. (2016) do not apply. The expression for $\mathcal{Y}(\tilde{\theta}, z)$ is more involved, but remains easy to calculate in applications.

Lemma 2
Define

$$A(\tilde{\theta}, \theta) = \Sigma_Y(\tilde{\theta})^{-2} \sum_{i=1}^{d_X} \left[\Sigma_{XY,i}(\tilde{\theta})^2 - \Sigma_{XY,i}(\theta, \tilde{\theta})^2 \right],$$

$$B_Z(\tilde{\theta},\theta) = 2\Sigma_Y(\tilde{\theta})^{-1}\sum_{i=1}^{d_X}\left[\Sigma_{XY,i}(\tilde{\theta})Z_{\tilde{\theta},i}(\tilde{\theta}) - \Sigma_{XY,i}(\theta,\tilde{\theta})Z_{\tilde{\theta},i}(\theta)\right],$$

and

$$C_Z(\tilde{\theta},\theta) = \sum_{i=1}^{d_X}\left[Z_{\tilde{\theta},i}(\tilde{\theta})^2 - Z_{\tilde{\theta},i}(\theta)^2\right].$$

For

$$D_Z(\tilde{\theta},\theta) = B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta),$$

$$G_Z(\tilde{\theta},\theta) = \frac{-B_Z(\tilde{\theta},\theta) + \sqrt{D_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)},\ K_Z(\tilde{\theta},\theta) = \frac{-B_Z(\tilde{\theta},\theta) - \sqrt{D_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)}$$

$$\textit{and } H_Z(\tilde{\theta},\theta) = \frac{-C_Z(\tilde{\theta},\theta)}{B_Z(\tilde{\theta},\theta)},$$

define

$$\ell_Z^1(\tilde{\theta}) = \max\left\{\max_{\theta\in\Theta:A(\tilde{\theta},\theta)<0,D_Z(\tilde{\theta},\theta)\geq 0} G_Z(\tilde{\theta},\theta), \max_{\theta\in\Theta:A(\tilde{\theta},\theta)=0,B_Z(\tilde{\theta},\theta)>0} H_Z(\tilde{\theta},\theta)\right\},$$

$$\ell_Z^2(\tilde{\theta},\theta) = \max\left\{\max_{\theta\in\Theta:A(\tilde{\theta},\theta)<0,D_Z(\tilde{\theta},\theta)\geq 0} G_Z(\tilde{\theta},\theta), \max_{\theta\in\Theta:A(\tilde{\theta},\theta)=0,B_Z(\tilde{\theta},\theta)>0} H_Z(\tilde{\theta},\theta), G_Z(\tilde{\theta},\theta)\right\},$$

$$u_Z^1(\tilde{\theta},\theta) = \min\left\{\min_{\theta\in\Theta:A(\tilde{\theta},\theta)<0,D_Z(\tilde{\theta},\theta)\geq 0} K_Z(\tilde{\theta},\theta), \min_{\theta\in\Theta:A(\tilde{\theta},\theta)=0,B_Z(\tilde{\theta},\theta)<0} H_Z(\tilde{\theta},\theta), K_Z(\tilde{\theta},\theta)\right\},$$

$$u_Z^2(\tilde{\theta}) = \min\left\{\min_{\theta\in\Theta:A(\tilde{\theta},\theta)<0,D_Z(\tilde{\theta},\theta)\geq 0} K_Z(\tilde{\theta},\theta), \min_{\theta\in\Theta:A(\tilde{\theta},\theta)=0,B_Z(\tilde{\theta},\theta)<0} H_Z(\tilde{\theta},\theta)\right\},$$

and

$$\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}}) = \min_{\theta\in\Theta:A(\tilde{\theta},\theta)=B_Z(\tilde{\theta},\theta)=0} C_Z(\tilde{\theta},\theta).$$

If $\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}}) \geq 0$ *then*

$$\mathcal{Y}(\tilde{\theta},Z_{\tilde{\theta}}) = \bigcap_{\theta\in\Theta:A(\tilde{\theta},\theta)>0,D_Z(\tilde{\theta},\theta)\geq 0}\left[\ell_Z^1(\tilde{\theta}),u_Z^1(\tilde{\theta},\theta)\right]\cup\left[\ell_Z^2(\tilde{\theta},\theta),u_Z^2(\tilde{\theta})\right].$$

If $\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}}) < 0$, *then* $\mathcal{Y}(\tilde{\theta},Z_{\tilde{\theta}}) = \emptyset$.

While the expression for $\mathcal{Y}(\tilde{\theta},z)$ in this setting is long, it is easy to calculate in practice, and can be expressed as a finite union of intervals using DeMorgan's laws.

Moreover, as in the level maximization case, $Pr_\mu\left\{\mathcal{V}(\hat{\theta}, Z_{\hat{\theta}}) < 0\right\} = 0$ for all μ, so we can ignore this constraint in applications.

Our derivations have so far assumed we have no additional conditioning variables $\hat{\gamma}$. If we also condition on $\hat{\gamma} = \tilde{\gamma}$, then for $\mathcal{X}_\gamma(\tilde{\gamma}) = \{X : \gamma(X) = \tilde{\gamma}\}$, 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 (20), $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z) = \mathcal{Y}(\tilde{\theta}, z) \cap \mathcal{Y}_\gamma(\tilde{\gamma}, z)$. The form of $\mathcal{X}_\gamma(\tilde{\gamma})$ and $\mathcal{Y}_\gamma(\tilde{\gamma}, z)$ depends on the conditioning variables $\hat{\gamma}$ considered. We next discuss the effect of conditioning on the outcomes of pretests in our empirical welfare maximization and structural break examples.

Empirical Welfare Maximization (continued) Suppose that we report estimates and confidence sets for welfare only if the improvement in empirical welfare from the estimated optimal policy over a baseline policy $\theta = 0$ exceeds a threshold c, $X(\hat{\theta}) - X(0) \geq c$. For instance, we might report results only when the test of White (2000) rejects the null that no policy has performance exceeding the baseline, $H_0 : \max_\Theta \mu_X(\theta) \leq \mu_X(0)$. This implies that we report results if and only if $X(\hat{\theta}) - X(0) \geq c$ for c a critical value depending on Σ. We can set $\gamma(X) = 1\left\{X(\hat{\theta}) - X(0) \geq c\right\}$, and it is natural to condition inference on $\hat{\gamma} = 1$.

The conditioning event in this setting is $\mathcal{X}_\gamma(1) = \left\{X : X(\hat{\theta}) - X(0) \geq c\right\}$, and one can show that, assuming $\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(0) > 0$ for simplicity,

$$\mathcal{Y}_\gamma(1, Z_{\tilde{\theta}}) = \left\{ y : y \geq \frac{\Sigma_Y(\tilde{\theta})\left(c - Z_{\tilde{\theta}}(\tilde{\theta}) + Z_{\tilde{\theta}}(0)\right)}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(0)} \right\}.$$

See Section B.1 of the Supplement for details, as well as expressions for other values of $\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(0)$. In the present case, provided $\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) \geq 0$, $\mathcal{Y}(\tilde{\theta}, 1, Z_{\tilde{\theta}}) = \left[\mathcal{L}^*(\tilde{\theta}, Z_{\tilde{\theta}}), \mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta}})\right]$ where $\mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta}})$ is the upper bound derived in Lemma 1, while

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

for $\mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta}})$ again as in Lemma 1. Hence, when $\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(0) > 0$, conditoning on $\hat{\gamma} = 1$ simply modifies the lower bound $\mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta}})$. Likewise, when $\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(0) < 0$

or $\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(0) = 0$, conditioning on $\hat{\gamma} = 1$ modifies $\mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta}})$ and $\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}})$, respectively. $\triangle$

Structural Break Estimation (continued) Suppose we report estimates and confidence sets for the break parameter $\delta(\hat{\theta})$ only if we reject the null hypothesis of no structural break, $H_0 : \delta(\theta) = 0$ for all $\theta \in \Theta$. Suppose, in particular, that we test this hypothesis with the sup-Wald test of Andrews (1993). As shown in Elliott and Müller (2014), in our setting such a test rejects if and only if $\|X(\hat{\theta})\| > c$ for a critical value c that depends on Σ. We can set $\gamma(X) = 1\left\{\|X(\hat{\theta})\| > c\right\}$, and it is again natural to condition inference on $\hat{\gamma} = 1$.

In this setting $\mathcal{X}_\gamma(1) = \left\{X : \|X(\hat{\theta})\| > c\right\}$. As before the expressions for the conditioning sets are involved but straightforward to compute. In particular, for $\bar{\mathcal{V}}(Z_{\tilde{\theta}})$, $\bar{\mathcal{L}}(Z_{\tilde{\theta}})$, and $\bar{\mathcal{U}}(Z_{\tilde{\theta}})$ defined in Section B.2 of the Supplement, if $\bar{\mathcal{V}}(Z_{\tilde{\theta}}) \geq 0$ then $\mathcal{Y}_\gamma(1, Z_{\tilde{\theta}}) = \left[\bar{\mathcal{L}}(Z_{\tilde{\theta}}), \bar{\mathcal{U}}(Z_{\tilde{\theta}})\right]$, while $\mathcal{Y}_\gamma(1, Z_{\tilde{\theta}}) = \emptyset$ otherwise. Thus,

$$\mathcal{Y}(\tilde{\theta}, 1, Z_{\tilde{\theta}}) = \bigcap_{\theta \in \Theta : A(\tilde{\theta}, \theta) > 0, D_Z(\tilde{\theta}, \theta) \geq 0} \left[\ell_Z^{1*}(\tilde{\theta}), u_Z^{1*}(\tilde{\theta}, \theta)\right] \cup \left[\ell_Z^{2*}(\tilde{\theta}, \theta), u_Z^{2*}(\tilde{\theta})\right]$$

when $\mathcal{V}^*(\tilde{\theta}, Z_{\tilde{\theta}}) \geq 0$, and $\mathcal{Y}(\tilde{\theta}, 1, Z_{\tilde{\theta}}) = \emptyset$ otherwise, where

$$\left(\ell_Z^{j*}, u_Z^{j*}\right) = \left(\max\left\{\ell_Z^j, \bar{\mathcal{L}}(Z_{\tilde{\theta}})\right\}, \min\left\{u_Z^j, \bar{\mathcal{U}}(Z_{\tilde{\theta}})\right\}\right) \text{ for j} \in \{1, 2\},$$

and $\mathcal{V}^*(\tilde{\theta}, Z_{\tilde{\theta}}) = \min\left\{\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}), \bar{\mathcal{V}}(Z_{\tilde{\theta}})\right\}$. $\triangle$

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

4.3 Comparison to Sample Splitting

A common remedy in practice for the problems we study is to split the sample. If we have iid observations and calculate $\hat{\theta}$ based on the first half of the data, conventional estimates and confidence intervals for $\mu_Y(\hat{\theta})$ that use only the second half of the data will be (conditionally) valid. Hence it is natural to ask how our conditioning approach

compares to sample splitting. Fithian et al. (2017) discuss this issue at length: here we briefly summarize the implications of their results in our setting.

For ease of exposition we focus on even sample splits. Asymptotically, such splits yield a pair of independent and identically distributed normal draws (X^1, Y^1) and (X^2, Y^2) both of which follow (6).[16] Sample splitting procedures calculate $\hat{\theta}$ as in (7) and (8) for level and norm maximization, respectively, but replace X by X^1. Inference on $\mu_Y(\hat{\theta})$ is then conducted using (X^2, Y^2). In particular, the conventional 95% sample-splitting confidence interval for $\mu_Y(\hat{\theta})$,

$$\left[Y^2(\hat{\theta}) - 1.96\sqrt{\Sigma_Y(\hat{\theta})}, Y^2(\hat{\theta}) + 1.96\sqrt{\Sigma_Y(\hat{\theta})}\right], \tag{25}$$

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

Empirical Welfare Maximization (continued) Suppose we split the sample in half at random. Define (X_n^1, Y_n^1) and (X_n^2, Y_n^2) analogously to (X_n, Y_n), now using only the first and second halves of the data, respectively. (X_n^1, Y_n^1) and (X_n^2, Y_n^2) still converge in distribution as in (9). Moreover, (X_n^1, Y_n^1) and (X_n^2, Y_n^2) are independent, both in finite samples and asymptotically. Let $\hat{\theta} = \text{argmax}_{\theta\in\Theta} X_n^1(\theta)$. If we use Y_n^1 for inference, the same issues as discussed above arise due to dependence between X_n^1 and Y_n^1. If we instead use only Y_n^2, then since $\hat{\theta}$ and Y_n^2 are independent we can rely on conventional estimates and confidence intervals. $\triangle$

While sample splitting resolves the inference problem this comes at a cost. First, $\hat{\theta}$ 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 larger expected welfare loss (see, e.g., Theorems 2.1 and 2.2 in Kitagawa and Tetenov (2018)). Moreover, even after conditioning on $\hat{\theta}$ the minimal sufficient statistic for μ is the the full-sample average $\frac{1}{2}(X^1, Y^1) + \frac{1}{2}(X^2, Y^2)$. Hence, using only (X^2, Y^2) for inference sacrifices information.

Fithian et al. (2017) formalize this point, and show that sample splitting tests are inadmissible. Corollary 1 of Fithian et al. (2017), applied in our setting, shows that for any sample splitting test there exists a test that uses the full data and has weakly

[16] Uneven sample splits still result in independent draws with the same mean but different variances.

higher power against all alternatives and strictly higher power at some alternatives. This result extends directly to quantile unbiased estimators, and shows that for any quantile unbiased split-sample estimator, there exists a full-sample quantile unbiased estimator which is more concentrated around the true parameter value in the sense of Proposition 1.

Hence, while split-sample methods allow valid inference, they are dominated. Splitting the sample changes the definition of $\hat{\theta}$ and the conditioning event $\left\{\hat{\theta} = \tilde{\theta}\right\}$, however, so sample splitting approaches are not directly comparable to our conditioning approach developed above.

While conventional sample splitting methods are dominated, calculating $\hat{\theta}$ based on only part of the data may increase the amount of information available for inference on $\mu_Y(\hat{\theta})$ and so allow tighter confidence intervals. Thus, depending on how we weight nosier values of $\hat{\theta}$ against more precise inference on $\mu_Y(\hat{\theta})$ it may be helpful to split the sample, though we should continue to apply conditional inference procedures. See Tian and Taylor (2016) and Tian et al. (2016) for related discussion.

4.4 Behavior When $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\}$ 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 set

$$CS_N = \left[Y(\hat{\theta}) - c_{\alpha/2,N}\sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_{\alpha/2,N}\sqrt{\Sigma_Y(\hat{\theta})}\right] \tag{26}$$

for $c_{\alpha,N}$ the $1-\alpha$-quantile of the standard normal distribution, $\hat{\mu}_N$ is biased and CS_N has incorrect coverage conditional on $\hat{\theta} = \tilde{\theta}$, $\hat{\gamma} = \tilde{\gamma}$. These biases are mild when $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\}$ is close to one, however, since in this case the conditional distribution is close to the unconditional one. Intuitively, $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}\right\}$ is close to one for some $\tilde{\theta}$ when $\mu_X(\theta)$ or $\|\mu_X(\theta)\|$ has a well-separated maximum in the level and norm maximization problems, respectively. This section shows that our procedures converge to conventional ones in this case.

In particular, suppose first that for some sequence of values $\mu_{Y,m}$ and $z_{\tilde{\theta},m}$ the conditional probability $\Pr_{\mu_{Y,m}}\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} | Z_{\tilde{\theta}} = z_{\tilde{\theta},m}\right\} \to 1$ as $m \to \infty$. Then our conditional confidence sets and estimates converge to the usual confidence sets and estimates.

Lemma 3

Consider any sequence of values $\mu_{Y,m}$ and $z_{\tilde{\theta},m}$ such that

$$Pr_{\mu_{Y,m}}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\right\}\rightarrow 1.$$

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 $CS_U \rightarrow_p CS_N$, $CS_{ET} \rightarrow_p CS_N$, and $\hat{\mu}_{\frac{1}{2}} \rightarrow_p Y(\tilde{\theta})$, where for confidence sets $\rightarrow_p$ denotes convergence in probability of the endpoints.

Lemma 3 discusses probabilities conditional $Z_{\tilde{\theta}}$. If we consider a sequence of values $\mu_{Y,m}$ such that $Pr_{\mu_{Y,m}}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\} \rightarrow_p 1$, however, the same result holds both conditioning only on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$ and unconditionally.

Proposition 3

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

These results provide an additional argument for using our procedures: they remain valid even 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.

5 Unconditional Inference

Rather than requiring validity conditional on $(\hat{\theta},\hat{\gamma})$ we can instead require coverage only on average, yielding the unconditional coverage requirement

$$Pr\left\{\mu(\hat{\theta})\in CS\right\}\geq 1-\alpha \text{ for all } \mu. \tag{27}$$

All confidence sets with correct conditional coverage in the sense of (16) also have correct unconditional 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 set CS with correct conditional coverage (16) also has correct unconditional coverage (27).

Uniqueness of $\hat{\theta}$ implies that the conditioning events $\mathcal{X}(\tilde{\theta}, \tilde{\gamma})$ partition the support of X up to sets of measure zero. 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 4
Suppose that for all $\theta, \tilde{\theta} \in \Theta$ *such that* $\theta \neq \tilde{\theta}$, *either* $Var\left(X(\theta)|X(\tilde{\theta})\right) \neq 0$ *or* $Var\left(X(\tilde{\theta})|X(\theta)\right) \neq 0$. *Then* $\hat{\theta}$ *is unique with probability one for all* μ.

Proposition 4 shows that the conditional confidence sets derived in the last section are valid unconditional confidence sets as well. As shown in Section 4.4, these conditional confidence sets also converge to the usual confidence sets when $(\hat{\theta}, \hat{\gamma})$ takes some value $(\tilde{\theta}, \tilde{\gamma})$ with high probability, and so perform well in this case. Unconditional coverage is less demanding than conditional coverage, however, so relaxing the coverage requirement from (16) to (27) may allow us to obtain shorter confidence sets in other cases.

In this section we explore the benefits of such a relaxation. We begin by introducing unconditional confidence sets based on projections of simultaneous confidence bands for μ_Y as in Kitagawa and Tetenov (2018). We then introduce hybrid confidence sets that combine projection confidence sets with conditioning arguments. We do not know of estimators for $\mu_Y(\hat{\theta})$ which are unconditionally α-quantile unbiased but not conditionally unbiased, but introduce hybrid estimators which allow a small unconditional bias.

5.1 Projection Confidence Sets

One approach to obtain an unconditional confidence set for $\mu_Y(\hat{\theta})$, used in Kitagawa and Tetenov (2018) and, in the one-sided case, Romano and Wolf (2005), is to start with a joint confidence set for μ and project on the dimension corresponding to $\hat{\theta}$. Formally, 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

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

then $Pr_\mu\{\mu \in CS_\mu\} = 1 - \alpha$ for all μ, so CS_μ is a level $1 - \alpha$ confidence set for μ.[17] If we then define

$$CS_P = \left\{\tilde{\mu}(\hat{\theta}) : \exists \mu \in CS_\mu \text{ such that } \mu(\hat{\theta}) = \tilde{\mu}(\hat{\theta})\right\}$$

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

as the projection of CS_μ on the parameter space for $\mu(\hat{\theta})$, then since $\{\mu \in CS_\mu\}$ implies $\mu(\hat{\theta}) \in CS_P$, CS_P satisfies the unconditional coverage requirement (27). As noted in Section 2, however, CS_P does not generally have correct conditional coverage.

The width of the confidence set CS_P depends on the variance $\Sigma_Y(\hat{\theta})$ but does not otherwise depend on the data. To account for the randomness of $\hat{\theta}$, the critical value c_α is larger than the conventional two-sided normal critical value. This means that CS_P will be conservative in cases where $\hat{\theta}$ takes a given value $\tilde{\theta}$ with high probability. We next consider hybrid confidence sets, which combine projection and conditioning arguments.

5.2 Hybrid Confidence Sets

Conditional confidence sets have coverage exactly $1 - \alpha$ and so are non-conservative. We showed in Section 4.4 that these confidence sets converge to conventional confidence sets when $(\hat{\theta}, \hat{\gamma})$ takes a given value with high probability. On the other hand, our simulation results in Section 2 also show that conditional confidence sets can perform poorly in cases where the maximum is not well-separated (see also the discussion in Section 2.5 of Fithian et al. 2017). The simulation results of Section 2 suggest that the projection intervals introduced in the last section can perform better in cases where the maximum is not well-separated, but these intervals are longer than the conditional intervals in the case where the maximum is well-separated. Hence, neither the conditional nor the projection intervals seem entirely satisfactory. To bridge the gap between these procedures we introduce hybrid confidence sets, which combine projection and conditioning arguments.

Hybrid confidence sets are constructed to be subsets of the level $1 - \beta$ projection

[17] Note that we consider a studentized confidence band that adjusts the width based on $\Sigma_Y(\hat{\theta})$, while Kitagawa and Tetenov (2018) consider an unstudentized band. Romano and Wolf (2005) argue for studentization in a closely related problem.

confidence set for $0 \leq \beta < \alpha$

$$CS_P^\beta = \left[Y(\hat{\theta}) - c_\beta \sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_\beta \sqrt{\Sigma_Y(\hat{\theta})} \right].$$

The hybrid confidence set collects the values $\mu_{Y,0} \in CS_P^\beta$ not rejected by a hybrid test. Like our conditional tests, hybrid tests condition $\left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\}$ but hybrid tests of $H_0 : \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ further condition on the event that the null value is contained in the projection confidence set, $\mu_{Y,0} \in CS_P^\beta$. This changes the conditioning event to

$$\mathcal{Y}^H(\tilde{\theta}, \tilde{\gamma}, \mu_{Y,0}, z) = \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z) \cap \left[\mu_{Y,0} - c_\beta \sqrt{\Sigma_Y(\tilde{\theta})}, \mu_{Y,0} + c_\beta \sqrt{\Sigma_Y(\tilde{\theta})} \right].$$

Similar to our conditional confidence sets we construct hybrid confidence sets by inverting both equal-tailed and uniformly most powerful unbiased hybrid tests. To construct the conditional equal tailed test, we define $\phi^H_{OS-,\alpha}$ and $\phi^H_{OS+,\alpha}$ analogously to $\phi_{OS-,\alpha}$ and $\phi_{OS+,\alpha}$ in (23) and (24), respectively, except that we use the conditioning event $\mathcal{Y}^H(\tilde{\theta}, \tilde{\gamma}, \mu_{Y,0}, Z_{\tilde{\theta}})$ rather than $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}})$. The equal-tailed hybrid test of $H_0 : \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ is

$$\phi^H_{ET,\alpha}(\mu_{Y,0}) = \max \left\{ \phi^H_{OS-,\alpha/2}(\mu_{Y,0}), \phi^H_{OS+,\alpha/2}(\mu_{Y,0}) \right\},$$

which rejects if either of the upper or lower size $\alpha/2$ one-sided tests rejects. The level $1-\alpha$ equal-tailed hybrid confidence set is

$$CS^H_{ET} = \left\{ \mu_{Y,0} \in CS_P^\beta : \phi^H_{ET, \frac{\alpha-\beta}{1-\beta}}(\mu_{Y,0}) = 0 \right\},$$

which collects the set of values in CS_P^β which are not rejected by $\phi^H_{ET,\alpha}$.

To form a hybrid confidence set based on inverting unbiased tests, we likewise define $\phi^H_{TS,\alpha}$ analogously to $\phi_{TS,\alpha}$ in (22), using the conditioning event $\mathcal{Y}^H(\tilde{\theta}, \tilde{\gamma}, \mu_{Y,0}, Z_{\tilde{\theta}})$ rather than $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}})$. By the results of Proposition 2, we know that $\phi^H_{TS,\alpha}(\mu_{Y,0})$ is the uniformly most powerful unbiased test of $H_0 : \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ conditional on $\left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, \mu_{Y,0} \in CS_P^\beta \right\}$. The corresponding level $1-\alpha$ confidence set is then

$$CS^H_U = \left\{ \mu_{Y,0} \in CS_P^\beta : \phi^H_{U, \frac{\alpha-\beta}{1-\beta}}(\mu_{Y,0}) = 0 \right\}.$$

For $\beta = 0$ the hybrid confidence sets coincide with the conditional confidence sets CS_{ET} and CS_U. For $\beta > 0$, on the other hand, the hybrid confidence sets are contained in CS_P^β, and the size of the hybrid tests are correspondingly adjusted downwards. This adjustment is necessary because the true value $\mu_Y(\hat{\theta})$ sometimes falls outside CS_P^β, so if we did not account for this our hybrid confidence sets would under-cover. With this adjustment, however, hybrid confidence sets have coverage at least $1-\alpha$ both conditionally and unconditionally.

Proposition 5

The hybrid confidence sets CS_{ET}^H and CS_U^H have conditional coverage $\frac{1-\alpha}{1-\beta}$

$$Pr_\mu\left\{\mu(\tilde{\theta}) \in CS_{ET}^H | \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\} = \frac{1-\alpha}{1-\beta},$$

$$Pr_\mu\left\{\mu(\tilde{\theta}) \in CS_U^H | \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, \mu_Y(\tilde{\theta}) \in CS_P^\beta\right\} = \frac{1-\alpha}{1-\beta},$$

for all $\tilde{\theta} \in \Theta$, $\tilde{\gamma} \in \Gamma$, and all μ. Moreover, provided $\hat{\theta}$ is unique with probability one for all μ, both confidence sets have unconditional coverage at least $1-\alpha$,

$$\inf_\mu Pr_\mu\left\{\mu(\hat{\theta}) \in CS_{ET}^H\right\} \geq 1-\alpha, \quad \inf_\mu Pr_\mu\left\{\mu(\hat{\theta}) \in CS_U^H\right\} \geq 1-\alpha.$$

Hybrid confidence sets strike a balance between the conditional and projection approaches. The maximal length of hybrid intervals is bounded above by the length of CS_P^β. For β small hybrid confidence sets will be close to conditional confidence sets and thus, to the conventional confidence set, when $\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\}$ with high probability, though for $\beta > 0$ hybrid confidence intervals do not fully converge to conventional confidence sets as $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}\right\} \to 1$.[18] In our simulations we find that the performance of the hybrid and conditional approaches is quite similar in these well-separated cases. Hence, while one could modify the definition of hybrid confidence sets to restore full equivalence in the well-separated case, for simplicity we do not pursue this extension here.

While hybrid intervals combine the conditional and projection approaches, they

[18] Indeed, one can directly choose β to yield a given maximal power loss for the hybrid tests relative to conditional tests in the well-separated case. Such a choice of β will depend on Σ, however, so for simplicity we instead use $\beta = \alpha/10$ in our simulations. For similar reasons, both Romano et al. (2014) and McCloskey (2017) find this choice to perform well in two different settings when using a Bonferroni correction.

can yield performance more appealing than either. Specifically, recall that in Section 2 we found that hybrid confidence intervals had a shorter average length for many parameter values than did either the conditional or projection approaches used in isolation. Our simulation results in Sections 6 and 7 below provide further evidence of out-performance in realistic settings.

It is worth contrasting our hybrid approach with more conventional Bonferroni corrections as in e.g. Romano et al. (2014); McCloskey (2017). A simple Bonferroni approach for our setting intersects a level $1-\beta$ projection confidence interval CS_P^β with a level $1-\alpha+\beta$ conditional interval that conditions only on $\left\{\hat{\theta}=\tilde{\theta}, \hat{\gamma}=\tilde{\gamma}\right\}$. Bonferroni intervals differ from our hybrid intervals in two respects. First, they use a level $1-\alpha+\beta$ conditional confidence interval, while that of hybrid approach uses a level $\frac{1-\alpha}{1-\beta}$ conditional interval, where $\frac{1-\alpha}{1-\beta} < 1-\alpha+\beta$. Second, the conditional interval used by the Bonferroni approach does not condition on $\mu_Y(\tilde{\theta}) \in CS_P^\beta$, while that used by the hybrid approach does. Consequently, the hybrid interval never contains the endpoints of CS_P^β, while the same is not true of Bonferroni intervals.

5.3 Hybrid Estimators

The simulation results of Section 2 showed that our median unbiased estimator can sometimes be much more dispersed than the conventional estimator $\hat{\mu} = Y(\hat{\theta})$. While we do not know of an alternative approach to construct exactly median unbiased estimators in our setting, a version of our hybrid approach yields estimators which control both median bias and mean absolute error relative to $\hat{\mu} = Y(\hat{\theta})$.

To construct hybrid estimators we again condition on both $\left\{\hat{\theta}=\tilde{\theta}, \hat{\gamma}=\tilde{\gamma}\right\}$ and $\mu_Y(\tilde{\theta}) \in CS_P^\beta$. Conditional on these events and $Z_{\tilde{\theta}}$ we know that $Y(\tilde{\theta})$ again lies in $\mathcal{Y}^H(\tilde{\theta}, \tilde{\gamma}, \mu_Y(\tilde{\theta}), Z_{\tilde{\theta}})$. Let $F_{TN}^H(y; \mu_Y(\tilde{\theta}), \tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}})$ denote the conditional distribution function of $Y(\tilde{\theta})$, and define $\hat{\mu}_\alpha^H$ to solve

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

Proposition 6
For $\alpha \in (0,1)$, $\hat{\mu}_\alpha^H$ is uniquely defined, 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 6 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,

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

Hence, the absolute median bias of $\hat{\mu}^H_{\frac{1}{2}}$ (measured as the deviation of the exceedance probability from 1/2) is bounded above by β, and goes to zero as $\beta \to 0$. On the other hand, since $\hat{\mu}^H_{\frac{1}{2}} \in CS_P^\beta$ we have $\left|\hat{\mu}^H_{\frac{1}{2}} - Y(\hat{\theta})\right| \leq c_\beta\sqrt{\Sigma_Y(\tilde{\theta})}$, so the difference between $\hat{\mu}^H_{\frac{1}{2}}$ and the conventional estimator $Y(\hat{\theta})$ is bounded above by half the width of CS_P^β. Hence, the mean absolute error of $\hat{\mu}^H_{\frac{1}{2}}$ can likewise differ from that of $Y(\hat{\theta})$ by no more than this amount, which tends to zero as $\beta \to 1$. Hence, as β varies the hybrid estimator interpolates between the median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ and the conventional estimator $Y(\hat{\theta})$.

6 Simulations: Empirical Welfare Maximization

Our first set of simulations considers the EWM setting introduced in Section 3. We calibrate our simulations to experimental data from the National Job Training Partnership Act (JTPA) Study, which was previously used by Kitagawa and Tetenov (2018) to study empirical welfare maximization. A detailed description of the study can be found in Bloom et al. (1997).

We have data on $n = 11,204$ individuals i and the treatment D_i is binary; $D_i = 1$ indicates assignment to a job training program and $D_i = 0$ indicates non-assignment. The probability of assignment is constant, $d(c) = \Pr(D_i = 1|C_i = c) = 2/3$. We consider rules that assign treatment based on years of education C_i. In the data, C takes integer values ranging from 6 to 18 years. As in Section 3, rule θ assigns i to treatment if and only if $C_i \in \mathcal{C}_\theta$.

We consider two classes of policies. The first, which we call threshold policies, treat all individuals with fewer than θ years of education, $\mathcal{C}_\theta = \{C : C \leq \theta\}$. The second, which we call interval policies, treat all individuals with between θ_l and θ_u years of education, $\mathcal{C}_\theta = \{C : \theta_l \leq C \leq \theta_u\}$, where a policy θ consists of a (θ_l, θ_u) pair. The total number of policies $|\Theta|$ is equal to 13 and 91 for the threshold and interval cases, respectively. We define $X_n(\theta)$ as a scaled estimate for the increase in income from policy θ relative to the baseline of no treatment. For Y_i individual

income measured in hundreds of thousands of dollars,

$$X_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{Y_i D_i}{d(C_i)} 1\{C_i \in \mathcal{C}_\theta\} - \frac{Y_i(1-D_i)}{1-d(C_i)} 1\{C_i \notin \mathcal{C}_\theta\} \right),$$

and we consider inference on the average increase in income, so $Y_n = X_n$.

For our simulations, we draw normal vectors X with known variance Σ_X equal to a (consistent) estimate for the asymptotic variance of X_n based on the JTPA data, and take $\hat{\theta} = \text{argmax}_\theta X(\theta)$. The mean vector $\mu_{X,n}$ of X_n is not consistently estimable due to the $\sqrt{n}$ scaling, so we consider three specifications for the mean μ_X of X. Specification (i) sets $\mu_X = 0$, so all policies yield the same welfare as the baseline of no treatment. Specification (ii) sets $\mu_X = (0, -10^5, ..., -10^5)$, so one policy is much more effective than the others. Finally, specification (iii) sets $\mu_X = X_n$ for X_n calculated in the JTPA data. Intuitively, we expect that specification (i) will be unfavorable to conditional confidence sets, since in Section 2 these performed poorly when all policies were equally effective. Specification (ii) should be favorable to conditional confidence sets, since in this case one policy is much more effective than the others and $\hat{\theta}$ selects this policy with high probability. Hence, we are in the "well-separated" case and the results of Section 4.4 apply. Finally, specification (iii) is calibrated to the data, and it is not obvious which approaches will perform well in this setting.

To the best of our knowledge our conditional confidence sets are the only procedures available with correct conditional coverage given $\hat{\theta}$.[19] Hence, we focus on unconditional performance, and compare the conditional confidence sets CS_{ET} and CS_U, the hybrid confidence sets CS^H_{ET} and CS^H_U, and the projection confidence set CS_P. The conditional and hybrid confidence sets are novel to this paper, but (unstudentized) projection confidence sets were previously considered for this problem by Kitagawa and Tetenov (2018). We take $\alpha = 0.05$ in all cases and so consider 95% confidence sets. For hybrid confidence sets we set $\beta = \alpha/10 = .005$. All reported results are based on 10^4 simulation draws.

Table 1 reports the unconditional coverage $Pr_\mu\{\mu(\hat{\theta}) \in CS\}$ of all five confidence sets, along with the conventional confidence set CS_N as in (26). As expected, all confidence sets other than CS_N have correct coverage in all settings considered. The

[19] As noted in Section 4.3, if we instead calculated $\hat{\theta}$ based on only part of the data one could use sample-splitting to obtain confidence sets with conditional coverage.

conditional confidence sets are exact, with coverage equal to 95% up to simulation error. By contrast, hybrid confidence sets tend to be slightly conservative though generally by no more than $\beta = .005$, and projection confidence sets are often quite conservative, with coverage approaching one when we consider interval policies.

Table 1: Unconditional Coverage Probability

DGP	CS_{ET}	CS_U	CS^H_{ET}	CS^H_U	CS_P	CS_N
Class of Threshold Policies						
(i)	0.952	0.951	0.955	0.953	0.987	0.921
(ii)	0.95	0.951	0.952	0.952	0.993	0.949
(iii)	0.952	0.951	0.951	0.953	0.992	0.951
Class of Interval Policies						
(i)	0.948	0.948	0.955	0.955	0.993	0.827
(ii)	0.947	0.947	0.956	0.955	0.998	0.947
(iii)	0.946	0.945	0.953	0.952	0.998	0.947

We next compare the length of confidence sets. Projection confidence sets were proposed in the previous literature, and their length is proportional to the standard error $\sqrt{\Sigma_X(\hat{\theta})}$ for the welfare of the estimated optimal policy. Hence, CS_P provides a natural benchmark against which to compare the length of our new confidence sets. In Table 2 we compare our new confidence sets to this benchmark in two ways, first reporting the average length of CS_{ET}, CS_U, CS^H_{ET}, and CS^H_U relative to CS_P (that is, the ratio of the average of their lengths), and then reporting in what fraction of simulation draws our new confidence sets are longer than CS_P.

Focusing first on specification (i), where $\mu_X = 0$, we see that conditional confidence sets are longer than CS_P on average and in most simulation draws for both the threshold and interval policy specifications. Hence, as expected this case is unfavorable to these confidence sets. By contrast, our hybrid confidence sets are shorter than the projection sets both on average and in the substantial majority of simulation draws. Turning next to specification (ii), where μ_X has a well-separated maximum, we see that as expected conditional confidence sets are much shorter than projection confidence sets and are shorter on average and in all simulation draws. Hybrid confidence sets perform nearly as well. Finally in (iii), where μ_X is calibrated to the data, we see that the performance of the conditional confidence sets is between cases (i) and (ii) for the threshold policy specification, but even worse than case (i) (in terms of average length) for the interval policy specification. By contrast, hybrid confidence

sets again perform well.

Overall, these simulation results strongly favor the hybrid confidence sets relative to both the conditional and projection sets. We do not find a strong advantage for either CS^H_{ET} or CS^H_U, though when the two differ CS^H_{ET} generally performs better. Since CS^H_{ET} is also typically easier to calculated, these simulation results thus suggest using CS^H_{ET} in this setting.

Table 2: Length of Confidence Sets Relative to CS_P in EWM Simulations

DGP	Average Length Relative to CS_P				Probability Longer than CS_P			
	CS_{ET}	CS_U	CS^H_{ET}	CS^H_U	CS_{ET}	CS_U	CS^H_{ET}	CS^H_U
Class of Threshold Policies								
(i)	6.49	5.55	0.91	0.91	0.699	0.779	0	0.119
(ii)	0.75	0.75	0.75	0.75	0	0	0	0
(iii)	2.78	2.51	0.87	0.9	0.324	0.433	0.045	0.275
Class of Interval Policies								
(i)	10.77	8.97	0.83	0.83	0.782	0.876	0	0
(ii)	0.63	0.63	0.65	0.65	0	0	0	0
(iii)	25.75	21.42	0.78	0.81	0.33	0.427	0	0

We next consider the properties of our point estimators. The initial columns of Table 3 report the simulated median bias of our median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$, our hybrid estimator $\hat{\mu}^H_{\frac{1}{2}}$, and the conventional estimator $X(\hat{\theta})$, measured both as the difference in the exceedance probability from $\frac{1}{2}$ and as the studentized median estimation error. The hybrid estimator is quite close to being median unbiased.

The final three columns of Table 3 report the mean absolute studentized error for the estimators considered. These results show that the median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ has a much larger mean absolute error than the conventional estimator $X(\hat{\theta})$ in all designs except the well-separated case (ii), where all three estimators perform similarly. The hybrid estimator $\hat{\mu}^H_{\frac{1}{2}}$ likewise has a larger mean absolute error than the conventional estimator, but the difference is much smaller. Hence, we see that our hybrid estimator greatly reduces MAE relative to the median unbiased estimator, at the cost of only a very small increase in median bias. The choice between the hybrid and conventional estimator in this setting is less clear, however, and depends on one's relative dislike of bias and mean absolute error.

Overall, the results of this section confirm our theoretical results. Conditional confidence sets and estimators perform well when the optimal policy is well-separated

Table 3: Bias and Mean Absolute Error of Point Estimators

DGP	$Pr_\mu\left\{\hat{\mu} > \mu_X(\hat{\theta})\right\} - \frac{1}{2}$			$Med_\mu\left(\frac{\hat{\mu}-\mu_X(\hat{\theta})}{\sqrt{\Sigma_X(\hat{\theta})}}\right)$			$E_\mu\left[\frac{\lvert\hat{\mu}-\mu_X(\hat{\theta})\rvert}{\sqrt{\Sigma_X(\hat{\theta})}}\right]$		
	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$X(\hat{\theta})$	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$X(\hat{\theta})$	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$X\left(\hat{\theta}\right)$
Class of Threshold Policies									
(i)	-0.007	-0.007	0.391	-0.02	-0.02	0.82	5.02	1.28	0.96
(ii)	-0.001	0.001	0.001	0	0	0	0.79	0.79	0.79
(iii)	-0.001	-0.001	0.104	0	0	0.25	2.02	1.03	0.78
Class of Interval Policies									
(i)	0	0.003	0.5	0	0.02	1.3	10.36	1.46	1.38
(ii)	-0.002	0.001	0.001	0	0	0	0.81	0.81	0.82
(iii)	0	0.001	0.148	0	0	0.35	4.1	1.14	0.8

but can otherwise underperform existing alternatives. Hybrid confidence sets outperform existing alternatives in all cases, nearly matching conditional confidence sets in the well-separated case and maintaining much better performance in other settings. Hybrid estimators eliminate almost all median bias while obtaining a substantially smaller mean absolute error than the exact median-unbiased estimator. Hence, we find strong evidence favoring our hybrid confidence sets relative to the available alternatives, and evidence favoring our hybrid estimators if bias reduction is desired.

7 Simulations: Tipping Point Estimation

Our second set of simulation results is based on the tipping point model of Card et al. (2008). Card et al. study the evolution of neighborhood composition as a function of minority population share. In particular, for Y_i the normalized change in the white population of census tract i between 1980 and 1990, $C_{i,1}$ a vector of controls, and $C_{i,2}$ the minority share in 1980, Card et al. consider the specification

$$Y_i = \beta + C_{i,1}'\alpha + \delta 1\{C_{i,2} > \theta\} + U_i,$$

which allows the white population share to change discontinuously when the minority share exceeds some threshold θ. They then fit this model, including the break point θ, by least squares. See Card et al. (2008) for details on the data and motivation.

Wang (2017) shows that if we model the coefficient δ as on the same order as sampling uncertainty, the model of Card et al. (2008) is asymptotically equivalent to a version of the structural breaks model we introduced in Section 3. Hence, we can immediately apply our results for that model to the present setting. For details on this equivalence result, see Wang (2017). Following Wang (2017), we consider data from Chicago and Los Angeles, estimating the model separately in each city.

We define X_n as discussed in Section 3, and $\hat{\theta}$ is again asymptotically equivalent to the solution to the norm-maximization problem $\text{argmax}_{\theta\in\Theta} \|X_n(\theta)\|$. We define $Y_n(\theta) = \sqrt{n}\hat{\delta}(\theta)$ to be proportional to the estimated break coefficient imposing tipping point θ, so we again consider the problem of inference on the break coefficient while acknowledging randomness in the estimated breakdate.

Our simulations draw normal random vectors (X, Y), now from the limiting normal model derived in Section 3. This model depends on matrices analagous to Σ_C and Ω in Section 3 which we (consistently) estimate from the Card et al. (2008) data. It also depends on the analog of the function $\Sigma_{cg}(\cdot)$. Since this is not consistently estimable, we consider four specifications. Specification (i) assumes there is no break, corresponding to $\delta = 0$. Specification (ii) assumes that there is a single large break, setting $\delta = -100\%$ (the largest possible break in the context of this model) and taking the true break point $\hat{\theta}$ to equal the estimate in the Card et al. (2008) data. Finally, specification (iii) calibrates (the analog of) $\Sigma_{cg}(\cdot)$ to the data, corresponding to analog of model (10) where the intercept term in the regression may depend arbitrarily upon a neighborhood's minority share. This specification implies that the break model is misspecified, but as discussed above our approach remain applicable in this case, unlike the results of Wang (2017). Indeed, Card et al. (2008) acknowledge that the tipping point model only approximates their underlying theoretical model of neighborhood ethnic composition, so misspecification seems likely in this setting.

We again focus on the unconditional performance of our proposed procedures, along with existing alternatives. All reported results are based on 10^4 simulation draws. Table 4 reports coverage for the confidence sets CS_{ET}, CS_U, CS^H_{ET}, CS^H_U, and CS_P, along with the conventional confidence set CS_N. As for the simulations calibrated to the EWM application, we see that all confidence sets other than CS_N have correct coverage, CS_P often over-covers, the conditional confidence sets have exact coverage and the hybrid confidence sets exhibit minimal over-coverage. In this application, the conventional confidence set CS_N can be seen to exhibit severe under-

coverage for some simulation designs.

Table 4: Unconditional Coverage Probability

DGP	CS_{ET}	CS_U	CS^H_{ET}	CS^H_U	CS_P	CS_N
Chicago Data Calibration						
(i)	0.949	0.947	0.949	0.95	0.95	0.704
(ii)	0.949	0.952	0.954	0.955	0.995	0.95
(iii)	0.952	0.949	0.957	0.958	0.992	0.932
Los Angeles Data Calibration						
(i)	0.954	0.95	0.954	0.953	0.95	0.49
(ii)	0.948	0.948	0.953	0.953	0.998	0.948
(iii)	0.949	0.947	0.955	0.954	0.998	0.944

Table 5 compares the lengths of our confidence sets to that of CS_P. For each confidence set we again report both average length relative to CS_P and the frequency with which the confidence set is longer than CS_P. Here we see that the conditional confidence sets can be relatively long on average. We also see that the use of hybrid confidence sets provides marked performance improvements across the specifications considered. Remarkably, neither of the hybrid confidence sets is longer than CS_P in any simulation draw across all specifications examined. The overall message is similar to that of the previous section: hybrid confidence sets possess clear advantages for unconditional inference and CS^H_{ET} seems to be the most compelling option, especially given its computational simplicity.

Table 5: Length of Confidence Sets Relative to CS_P in Tipping Point Simulations

	Average Length Relative to CS_P				Probability Longer than CS_P			
	CS_{ET}	CS_U	CS^H_{ET}	CS^H_U	CS_{ET}	CS_U	CS^H_{ET}	CS^H_U
Chicago Data Calibration								
(i)	2.54	2.65	0.89	0.88	0.896	0.978	0	0
(ii)	0.73	0.75	0.72	0.72	0.012	0.019	0	0
(iii)	1.45	1.73	0.83	0.85	0.377	0.541	0	0
Los Angeles Data Calibration								
(i)	1.8	1.97	0.84	0.82	0.863	0.965	0	0
(ii)	0.67	0.7	0.66	0.66	0.015	0.027	0	0
(iii)	1.06	1.22	0.73	0.75	0.201	0.323	0	0

Finally, we consider the properties of our point estimators. The initial columns of Table 6 report median bias measured both with the deviation of the exceedance

probability from $\frac{1}{2}$ and with the studentized median estimation error. We again see that $\hat{\mu}_{\frac{1}{2}}$ is median-unbiased (up to simulation error) and that $\hat{\mu}^H_{\frac{1}{2}}$ exhibits minimal median bias. By contrast, the conventional estimator $Y(\hat{\theta})$ has substantial median bias as measured by the studentized median estimation error, though very little as measured by the exceedance probability in specification (i). This latter feature reflects the fact that the density of $Y(\hat{\theta})$ in this specification has very little mass at zero.

Turning to mean absolute studentized error, we see that all estimators perform similarly when the series has a single large break. By contrast, in specifications (i) (no break) and (iii) (fully data-calibrated model that does not impose a break), the median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ has a substantially larger mean absolute error than the conventional estimator $Y(\hat{\theta})$. The hybrid estimator has a smaller mean absolute studentized error than the median unbiased estimator across these specifications. Its mean absolute error is the smallest of the three estimators in specification (i) and lies between those of the other estimators in specification (iii).

Table 6: Bias and Mean Absolute Error in Tipping Point Simulations

	$Pr\left\{\hat{\mu} > \mu_Y(\hat{\theta})\right\} - \frac{1}{2}$			$Med\left(\frac{\hat{\mu}-\mu_Y(\hat{\theta})}{\sqrt{\Sigma_Y(\hat{\theta})}}\right)$			$E\left[\left\lvert\frac{\hat{\mu}-\mu_Y(\hat{\theta})}{\sqrt{\Sigma_Y(\hat{\theta})}}\right\rvert\right]$		
	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$Y(\hat{\theta})$	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$Y(\hat{\theta})$	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$Y(\hat{\theta})$
Chicago Data Calibration									
(i)	0.002	0.002	-0.003	0.01	0.01	-0.69	3.59	1.4	1.69
(ii)	-0.003	-0.003	-0.007	-0.01	-0.01	-0.01	0.84	0.82	0.81
(iii)	0.002	0.002	-0.16	0.01	0	-0.41	1.93	1.11	0.85
Los Angeles Data Calibration									
(i)	0.003	0.004	-0.005	0.02	0.02	-1.05	3.44	1.44	2.04
(ii)	-0.003	-0.003	-0.001	-0.01	-0.01	-0.02	0.83	0.81	0.8
(iii)	-0.005	-0.005	-0.098	-0.01	-0.01	-0.23	1.46	1	0.81

Overall, the results of this section again suggest excellent performance for our hybrid confidence sets and estimators relative to existing alternatives.

8 Conclusion

This paper considers a form of the winners' curse that arises when we select a target parameter for inference based on optimization. We propose confidence sets and quantile unbiased estimators for the target parameter that are optimal conditional

on its selection. We hence recommend our conditional inference procedure when it is appropriate to remove uncertainty about the choice of target parameters from inferential statements. The conditionally valid procedures are indeed unconditionally valid, but we find that these conditionally valid procedures can have unappealing (unconditional) performance relative to existing alternatives. If one is satisfied with unconditional coverage and (in the case of estimation) a small, controlled degree of bias, we propose hybrid inference procedures which combine conditioning with projection confidence sets. Examining performance in simulations calibrated to empirical welfare maximization and tipping point applications, we find that our hybrid approach performs well in both cases.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.

Athey, S. and Wager, S. (2018). Efficient policy learning. *arXiv preprint*. arXiv 1702.0289.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association*, 104:486–500.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997). The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of Human Resources*, 32(3):549–576.

Card, D., Mas, A., and Rothstein, J. (2008). Tipping and the dynamics of segregation. *Quarterly Journal of Economics*, 123:177–216.

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

Cox, G. (2018). Almost sure uniqueness of a global minimum without convexity. Unpublished Manuscript.

Elliott, G. and Müller, U. K. (2007). Confidence sets for the date of a single break in linear time series regressions. *Journal of Econometrics*, 141(2):1196–1218.

Elliott, G. and Müller, U. K. (2014). Pre and post break parameter inference. *Journal of Econometrics*, 180:141–157.

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

Graham, B. S., Imbens, G. W., and Ridder, G. (2014). Complementarity and aggregate implications of assortative matching: A nonparametric analysis. *Quantitative Economics*, 5:29–66.

Hansen, B. E. (2001). The new econometrics of structural change: Dating breaks in u.s. labor productivity. *Journal of Economic Perspectives*, 15(4):117–128.

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*.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77:1683–1701.

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

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

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 ed. edition.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

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

Perron, P. (2006). Dealing with structural breaks. In *Palgrave Handbook of Econometrics*, volume 1: Econometric Theory, pages 278–352. Palgrave.

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.

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.

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

Tian, X., Bi, N., and Taylor, T. (2016). MAGIC: a general, powerful and tractable method for selective inference. *arXiv*.

Tian, X. and Taylor, J. (2016). Selective inference with a randomized response. *Annals of Statistics (forthcoming)*.

Wang, Y. (2017). Inference in the threshold model. Working Paper.

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

Supplement to the paper

Inference on Winners

Isaiah Andrews Toru Kitagawa Adam McCloskey

May 10, 2018

This appendix contains proofs and supplementary results for the paper "Inference on Winners." Section A collects proofs for the results stated in the main text. Section B contains additional details and derivations for the Empirical Welfare Maximization and structural break examples.

A Proofs

Proof of Proposition 1 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), so if the set of possible values for the latter contains an open set, that for the former does as well.

Note, next, that since $(\tilde{Z}, \tilde{Y}, Y^*)$ is a linear transformation of (X, Y), $(\tilde{Z}, \tilde{Y}, Y^*)$ is jointly normal (albeit with a 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 $Var\left((\tilde{Y}, X')'\right)$, however, 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. However, 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

$$\frac{1\left\{\left(\tilde{z} + \Sigma_{XY}(\cdot,\tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1}\tilde{y}\right) \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}}{Pr_{\tilde{\mu}_Z,\tilde{\mu}_Y}\left\{X \in \mathcal{X}(\tilde{\theta},\tilde{\gamma})\right\}} f_{N,\tilde{Z}}(\tilde{z};\tilde{\mu}_Z) f_{N,\tilde{Y}}(\tilde{y};\tilde{\mu}_Y) f_{N,Y^*}(\tilde{y}^*;\mu_Y^*). \quad (28)$$

The density (28) 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 theorem then follows immediately Theorem 5.5.9 of Pfanzagl (1994). Part 2 of the theorem 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 2 In the proof of Proposition 1, 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, if we restrict attention to the linear space $\left\{\mu^* + \Sigma^{\frac{1}{2}}v : v \in \mathbb{R}^{\dim(X,Y)}\right\}$, we see 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 then follows immediately. □

Proof of Lemma 1 Let us number the elements of Θ as $\left\{\theta_1,\theta_2,...,\theta_{|\Theta|}\right\}$, where $X(\theta_1)$ is the first element of X, $X(\theta_2)$ is the second element, and so on. Let us further assume without loss of generality that $\tilde{\theta} = \theta_1$. Note that the conditioning

event $\{\max_{\theta\in\Theta} X(\theta) = X_1\}$ is equivalent to $\{MX \geq 0\}$, where

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

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

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

for $c = Cov(W, Y(\tilde{\theta}))/\Sigma_{YY}(\tilde{\theta})$, noting that the definition of $Z_{\tilde{\theta}}$ in (18) corresponds to extracting the elements of $Z^*_{\tilde{\theta}}$ corresponding to X. By Lemma 5.1 of Lee et al. (2016),

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

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

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

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

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

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

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

and

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

Note, however, that these depend only on the first $|\Theta|$ terms of $Z^*_{\tilde{\theta}}$, and thus are functions of $Z_{\tilde{\theta}}$, as expected. The result follows immediately. □

Proof of Lemma 2 Note the following equivalence of events:

$$\begin{aligned}
\{\hat{\theta} = \tilde{\theta}\} &= \left\{ \sum_{i=1}^{d_X} X_i(\tilde{\theta})^2 \geq \sum_{i=1}^{d_X} X_i(\theta)^2 \ \forall \theta \in \Theta \right\} \\
&= \left\{ \sum_{i=1}^{d_X} \left[Z_{\tilde{\theta},i}(\tilde{\theta}) + \Sigma_{XY,i}(\tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1} Y(\tilde{\theta}) \right]^2 \right. \\
&\qquad \left. \geq \sum_{i=1}^{d_X} \left[Z_{\tilde{\theta},i}(\theta) + \Sigma_{XY,i}(\theta,\tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1} Y(\tilde{\theta}) \right]^2 \ \forall \theta \in \Theta \right\} \\
&= \left\{ A(\tilde{\theta},\theta) Y(\tilde{\theta})^2 + B_Z(\tilde{\theta},\theta) Y(\tilde{\theta}) + C_Z(\tilde{\theta},\theta) \geq 0 \ \forall \theta \in \Theta \right\}, \qquad (29)
\end{aligned}$$

for $A(\tilde{\theta},\theta)$, $B_Z(\tilde{\theta},\theta)$, and $C_Z(\tilde{\theta},\theta)$ as defined in the statement of the lemma.

By the quadratic formula, (29) is equivalent to the event

$$\begin{aligned}
&\left\{ \frac{-B_Z(\tilde{\theta},\theta) + \sqrt{B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)} \leq Y(\tilde{\theta}) \right. \\
&\left. \leq \frac{-B_Z(\tilde{\theta},\theta) - \sqrt{B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)} \ \forall \theta \in \Theta \text{ s.th. } A(\tilde{\theta},\theta) < 0 \right.
\end{aligned}$$

$$
\begin{aligned}
&\text{and } B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta) \geq 0, \\
&Y(\tilde{\theta}) \leq \frac{-B_Z(\tilde{\theta},\theta) - \sqrt{B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)} \text{ or} \\
&Y(\tilde{\theta}) \geq \frac{-B_Z(\tilde{\theta},\theta) + \sqrt{B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)} \ \forall \theta \in \Theta \text{ s.th. } A(\tilde{\theta},\theta) > 0 \\
&\text{and } B_Z(\tilde{\theta},\theta)^2 - 4A(\tilde{\theta},\theta)C_Z(\tilde{\theta},\theta) \geq 0, \\
&Y(\tilde{\theta}) \geq \frac{-C_Z(\tilde{\theta},\theta)}{B_Z(\tilde{\theta},\theta)} \ \forall \theta \in \Theta \text{ s.th. } A(\tilde{\theta},\theta) = 0 \text{ and } B_Z(\tilde{\theta},\theta) > 0, \\
&Y(\tilde{\theta}) \leq \frac{-C_Z(\tilde{\theta},\theta)}{B_Z(\tilde{\theta},\theta)} \ \forall \theta \in \Theta \text{ s.th. } A(\tilde{\theta},\theta) = 0 \text{ and } B_Z(\tilde{\theta},\theta) < 0, \\
&C_Z(\tilde{\theta},\theta) \geq 0 \ \forall \theta \in \Theta \text{ s.th. } A(\tilde{\theta},\theta) = 0 \text{ and } B_Z(\tilde{\theta},\theta) = 0 \Big\} \\
&= \Bigg\{ Y(\tilde{\theta}) \in \bigcap_{\theta\in\Theta: A(\tilde{\theta},\theta)<0, D_Z(\tilde{\theta},\theta)\geq 0} \left[\frac{-B_Z(\tilde{\theta},\theta) + \sqrt{D_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)}, \frac{-B_Z(\tilde{\theta},\theta) - \sqrt{D_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)} \right] \\
&\quad \cap \bigcap_{\theta\in\Theta: A(\tilde{\theta},\theta)>0, D_Z(\tilde{\theta},\theta)\geq 0} \left(-\infty, \frac{-B_Z(\tilde{\theta},\theta) - \sqrt{D_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)} \right] \cup \left[\frac{-B_Z(\tilde{\theta},\theta) + \sqrt{D_Z(\tilde{\theta},\theta)}}{2A(\tilde{\theta},\theta)}, \infty \right) \\
&\quad \cap \bigcap_{\theta\in\Theta: A(\tilde{\theta},\theta)=0, B_Z(\tilde{\theta},\theta)>0} \Big[H_Z(\tilde{\theta},\theta), \infty \Big) \cap \bigcap_{\theta\in\Theta: A(\tilde{\theta},\theta)=0, B_Z(\tilde{\theta},\theta)<0} \Big(-\infty, H_Z(\tilde{\theta},\theta) \Big] \Bigg\} \\
&\quad \cap \left\{ \min_{\theta\in\Theta: A(\tilde{\theta},\theta)=B_Z(\theta\tilde{,}\theta)=0} C_Z(\tilde{\theta},\theta) \geq 0 \right\} \\
&= \Bigg\{ Y(\tilde{\theta}) \in \left[\max_{\theta\in\Theta: A(\tilde{\theta},\theta)<0, D_Z(\tilde{\theta},\theta)\geq 0} G_Z(\tilde{\theta},\theta), \min_{\theta\in\Theta: A(\tilde{\theta},\theta)<0, D_Z(\tilde{\theta},\theta)\geq 0} K_Z(\tilde{\theta},\theta) \right] \\
&\quad \cap \left[\max_{\theta\in\Theta: A(\tilde{\theta},\theta)=0, B_Z(\tilde{\theta},\theta)>0} H_Z(\tilde{\theta},\theta), \infty \right) \cap \left(-\infty, \min_{\theta\in\Theta: A(\tilde{\theta},\theta)=0, B_Z(\tilde{\theta},\theta)<0} H_Z(\tilde{\theta},\theta) \right] \\
&\quad \cap \bigcap_{\theta\in\Theta: A(\tilde{\theta},\theta)>0, D_Z(\tilde{\theta},\theta)\geq 0} \Big(-\infty, K_Z(\tilde{\theta},\theta) \Big] \cup \Big[G_Z(\tilde{\theta},\theta), \infty \Big) \Bigg\} \cap \Big\{ \mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) \geq 0 \Big\} \\
&= \left\{ Y(\tilde{\theta}) \in \bigcap_{\theta\in\Theta: A(\tilde{\theta},\theta)>0, D_Z(\tilde{\theta},\theta)\geq 0} \Big[\ell_Z^1(\tilde{\theta},\theta), u_Z^1(\tilde{\theta},\theta) \Big] \cup \Big[\ell_Z^2(\tilde{\theta},\theta), u_Z^2(\tilde{\theta},\theta) \Big] \right\} \cap \Big\{ \mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) \geq 0 \Big\}
\end{aligned}
$$

for $\ell_Z^1(\tilde{\theta})$, $\ell_Z^2(\tilde{\theta},\theta)$, $u_Z^1(\tilde{\theta},\theta)$, $u_Z^2(\tilde{\theta})$, and $\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}})$ again defined in the statement of the

lemma. The result follows immediately. □

Proof of Lemma 3 Recall that conditional on $Z_{\tilde{\theta}} = z_{\tilde{\theta}}$, $\hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}$ only if $Y(\tilde{\theta}) \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z_{\tilde{\theta}})$, and the reverse holds almost surely. Hence, the assumptions of the proposition imply 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 sets are equivariant under shifts, in the sense that the conditional confidence set 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 set 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 sets 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 sets 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.

Lemma 5
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*

$$\phi_{ET}(\mu_{Y,0}) = 1\left\{Y(\tilde{\theta}) \notin \left[c_{l,ET}\left(\mu_{Y,0}, \mathcal{Y}_m\right), c_{u,ET}\left(\mu_{Y,0}, \mathcal{Y}_m\right)\right]\right\}, \tag{30}$$

and

$$\phi_U(\mu_{Y,0}) = 1\left\{Y(\tilde{\theta}) \notin \left[c_{l,U}\left(\mu_{Y,0}, \mathcal{Y}_m\right), c_{u,U}\left(\mu_{Y,0}, \mathcal{Y}_m\right)\right]\right\}, \tag{31}$$

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

and

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

for $c_{N,\frac{\alpha}{2}}$ *the* $1-\frac{\alpha}{2}$ *quantile of a standard normal distribution.*

To complete the proof, let CS_m denote a generic conditional confidence set formed by inverting a family of tests

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

We want to show that

$$CS_m\rightarrow_p\left[Y(\tilde{\theta})-c_{N,\frac{\alpha}{2}},Y(\tilde{\theta})+c_{N,\frac{\alpha}{2}}\right], \tag{32}$$

as $m\rightarrow\infty$, for CS_m formed by inverting either (30) or (31).

We assume that CS_m is a finite interval for all m, which holds trivially for the equal-tailed confidence set CS_{ET}, and holds for C_U by Lemma 5.5.1 of Lehmann and Romano. Since we consider intervals, by convergence in probability we will mean convergence in probability of the endpoints. For each value $\mu_{Y,0}$ our Lemma 5 implies that

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

for ϕ_m equal to either (30) or (31). This convergence in probability holds jointly for all finite collections of values $\mu_{Y,0}$, however, which implies (32). 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 3 We prove this result for the unconditional case, noting that since $Pr_{\mu_m}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}\rightarrow 1$, the result conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$ follows immediately.

Note that by the law of iterated expectations, $Pr_{\mu_{Y,m}}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}\rightarrow 1$ implies that $Pr_{\mu_{Y,m}}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}\right\}\rightarrow_p 1$. Hence, if we define

$$g(\mu_Y,z)=Pr_{\mu_Y}\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}=z\right\}$$

we see that $g(\mu_{Y,m}, Z_{\tilde{\theta}}) \to_p 1$.

Note, next, that if for any metric d for the difference between confidence intervals (e.g. the sum of the distances between the endpoints) if we define

$$h_\varepsilon(\mu_Y, z) = Pr_{\mu_Y}\left\{d(CS_U, CS_N) > \varepsilon | Z_{\tilde{\theta}} = z\right\},$$

Lemma 3 states 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 $\mathcal{H}(\varepsilon) = \{(\mu_Y, z) : h(\mu_Y, z) < \varepsilon\}$, we see that 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 $\delta > 0$,

$$Pr_{\mu_{Y,m}}\left\{(\mu_{y,m}, Z_{\tilde{\theta}}) \in \mathcal{G}(\delta)\right\} \to 1,$$

we see that for all $\varepsilon > 0$,

$$Pr_{\mu_{Y,m}}\left\{(\mu_{y,m}, Z_{\tilde{\theta}}) \in \mathcal{H}(\varepsilon)\right\} \to 1,$$

as well, which suffices to prove the desired claim for confidence sets. The same argument likewise implies the result for our median unbiased estimator. □

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

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

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

Proof of Lemma 4 Consider first the level-maximization case. Note that 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.

For norm-maximization, assume without loss of generality that $Var\left(X(\theta) | X(\tilde{\theta})\right) \neq 0$. Note that $\|X(\theta)\|$ is continuously distributed conditional on $X(\tilde{\theta}) = x(\tilde{\theta})$ for all $x(\tilde{\theta})$ and all μ, so $Pr_\mu\left\{\|X(\theta)\| = \|X(\tilde{\theta})\|\right\} = 0$. Almost-sure uniqueness of $\hat{\theta}$ again follows immediately from finiteness of Θ. □

Proof of Proposition 5 The first part of the proposition follows immediately from Proposition 2. For the second part of the proposition, note that for CS^H either of the hybrid confidence sets,

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

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

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

where the second equality follows from the first part of the proposition. □

Proof of Proposition 6 We first establish uniqueness of $\hat{\mu}_\alpha^H$. To do so, it suffices to show that $F_{TN}^H(Y(\tilde{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})$ 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 $F_{TN}^H(Y(\tilde{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})$, 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 $\mu_Y(\tilde{\theta})$, $F_{TN}^H(Y(\hat{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}})$ is a fortiori decreasing in $\mu_Y(\tilde{\theta})$. Uniqueness of $\hat{\mu}_\alpha^H$ for $\alpha \in (0,1)$ follows. Moreover, we see that $\hat{\mu}_Y^H$ is strictly increasing in $Y(\tilde{\theta})$ conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma},Z_{\hat{\theta}}=z_{\tilde{\theta}},\mu_Y(\tilde{\theta}) \in CS_P^\beta\right\}$. Note, next, that $F_{TN}^H(Y(\tilde{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}}) \in \{0,1\}$ 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,

$$F_{TN}^H(Y(\hat{\theta});\mu_Y(\tilde{\theta}),\tilde{\theta},\tilde{\gamma},Z_{\tilde{\theta}}) \sim U[0,1]$$

conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma},Z_{\hat{\theta}}=z_{\tilde{\theta}},\mu_Y(\tilde{\theta}) \in CS_P^\beta\right\}$. Monotonicity of $\hat{\mu}_Y^H$ in $Y(\tilde{\theta})$ implies that

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

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

and thus that $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on $\left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma}, Z_{\hat{\theta}} = z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CS_P^\beta \right\}$. 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 Lemma 5 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. Recall that the equal tailed test rejects if and only if

$$F_{TN}\left(Y(\tilde{\theta}), \mathcal{Y} \right) \notin \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2} \right], \tag{33}$$

for $F_{TN}(y, \mathcal{Y})$ the standard normal density truncated to $\mathcal{Y}$. Note, however, that (33) is equivalent to

$$Y(\tilde{\theta}) \notin \left[c_{l,ET}\left(\mathcal{Y}\right), c_{u,ET}\left(\mathcal{Y}\right) \right]$$

where we suppress the dependence of the critical values on $\mu_{Y,0} = 0$ for simplicity, and $\left(c_{l,ET}\left(\mathcal{Y}\right), c_{u,ET}\left(\mathcal{Y}\right) \right)$ solve

$$F_{TN}\left(c_{l,ET}\left(\mathcal{Y}\right), \mathcal{Y} \right) = \frac{\alpha}{2}$$

$$F_{TN}\left(c_{u,ET}\left(\mathcal{Y}\right), \mathcal{Y} \right) = 1 - \frac{\alpha}{2}.$$

Recall that we can write the density of $F_{TN}\left(y, \mathcal{Y}\right)$ as $\frac{1\{y \in \mathcal{Y}\}}{Pr\{\xi \in \mathcal{Y}\}} f_N\left(y\right)$ where f_N is the standard normal density and $Pr\left\{\xi \in \mathcal{Y}\right\}$ is the probability that $\xi \in \mathcal{Y}$ for $\xi \sim N\left(0,1\right)$. Hence, we can write

$$F_{TN}\left(y, \mathcal{Y}\right) = \frac{\int_{-\infty}^{y} 1\left\{\tilde{y} \in \mathcal{Y}\right\} f_N\left(\tilde{y}\right) d\tilde{y}}{Pr\left\{\xi \in \mathcal{Y}\right\}}.$$

Note that 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

$$F_{TN}(c_{l,ET}(\mathcal{Y}_m), \mathcal{Y}_m) = a_m(c_{l,ET}(\mathcal{Y}_m)) + F_N(c_{l,ET}(\mathcal{Y}_m)) = \frac{\alpha}{2}$$

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

$$Pr\{\zeta \in [c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})]\} = 1 - \alpha$$

$$E[\zeta 1\{\zeta \in [c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})]\}] = (1 - \alpha) E[\zeta]$$

for $\zeta \sim \xi | \xi \in \mathcal{Y}$ where $\xi \sim N(0, 1)$.

Note that for ζ_m the truncated random variable corresponding to $\mathcal{Y}_m$ we can write

$$Pr\left\{\zeta_m \in [c_l, c_u]\right\} = a_m\left(c_l, c_u\right) + \left(F_N\left(c_u\right) - F_N\left(c_l\right)\right)$$

for

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

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

Note, next, that we can write

$$E\left[\zeta_m 1\left\{\zeta_m \in [c_l, c_u]\right\}\right] = \left[\xi 1\left\{\xi \in [c_l, c_u]\right\}\right] + b_m\left(c_l, c_u\right)$$

for

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

Note, however, that

$$\int_{c_l}^{c_u} \left(1\left\{y \in \mathcal{Y}_m\right\} - 1\right) y f_N\left(y\right) dy \leq E\left[\left|\xi\right| 1\left\{\xi \notin \mathcal{Y}_m\right\}\right].$$

Hence, since

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

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

Next, let us define $\left(c_{l,m}, c_{u,m}\right)$ as the solutions to

$$Pr\left\{\zeta_m \in [c_l, c_u]\right\} = 1 - \alpha$$

$$E\left[\zeta_m 1\left\{\zeta_m \in [c_l, c_u]\right\}\right] = (1-\alpha) E\left[\zeta_m\right].$$

From our results above, we can re-write the problem solved by $(c_{l,m}, c_{u,m})$ as

$$F_N(c_u) - F_N(c_l) = 1 - \alpha - a_m(c_l, c_u)$$

$$E\left[\xi 1\left\{\xi \in [c_l, c_u]\right\}\right] = (1-\alpha) E\left[\zeta_m\right] - (1-\alpha) b_m(c_l, c_u).$$

Letting

$$\bar{a}_m = \sup_{c_l, c_u} \left|a_m(c_l, c_u)\right|,$$

$$\bar{b}_m = (1-\alpha) \sup_{c_l, c_u} \left|b_m(c_l, c_u)\right|$$

we thus see that $(c_{l,m}, c_{u,m})$ solves

$$F_N(c_u) - F_N(c_l) = 1 - \alpha - a_m^*$$

$$E\left[\xi 1\left\{\xi \in [c_l, c_u]\right\}\right] = (1-\alpha) E\left[\zeta_m\right] - b_m^*$$

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 6

Suppose that $c_{l,m}$ and $c_{u,m}$ solve

$$Pr\left\{\xi \in [c_l, c_u]\right\} = 1 - \alpha + a_m$$

$$E\left[\xi 1\left\{\xi \in [c_l, c_u]\right\}\right] = d_m$$

for a_m, $d_m \to 0$. Then $(c_{l,m}, c_{u,m}) \to \left(-c_{N,\frac{\alpha}{2}}, c_{N,\frac{\alpha}{2}}\right)$.

Using this lemma, since $E[\zeta_m] \to 0$ as $m \to \infty$ we see that for any sequence of values $(a_m^*, b_m^*) \to 0$,

$$\left(c_{l,m}(a_m^*, b_m^*), c_{u,m}(a_m^*, b_m^*)\right) \to \left(-c_{N,\frac{\alpha}{2}}, c_{N,\frac{\alpha}{2}}\right).$$

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 $(c_{l,m}, c_{u,m}) \to \left(-c_{N,\frac{\alpha}{2}}, c_{N,\frac{\alpha}{2}}\right)$ as we

wanted to show. □

Proof of Lemma 6 Note that the critical values solve

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

for f_N and F_N the standard normal density and distribution function, respectively. 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

$$F_N(c_u) - F_N(c_l) = (1-\alpha) - a_m$$

$$f_N(c_l) - f_N(c_u) = d_m$$

or more compactly $g(c) - v_m = 0$, for

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

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

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

which 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

$$c(v_m) \to \begin{pmatrix} -c_{N,\frac{\alpha}{2}} \\ c_{N,\frac{\alpha}{2}} \end{pmatrix},$$

again as we wanted to show. □

B Additional Results

B.1 Details for Empirical Welfare Maximization Example

We derive the form of the conditioning event $\mathcal{Y}_\gamma\left(1, Z_{\tilde{\theta}}\right)$ discussed in Section 4.2, and including for cases when $\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta}, 0) \leq 0$. Note that we can write

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

Rearranging, we see that

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

B.2 Details for Structural Break Estimation Example

This section provides additional results to supplement our discussion of the structural break example in the text.

We begin by establishing the weak convergence (15). To do so, we show uniform convergence over all of $[0, 1]$, which implies uniform convergence over Θ_T. Note, in

particular, that under (12) and (13), the continuous mapping theorem implies that

$$X_T(\theta) \Rightarrow X(\theta)$$

$$= \begin{pmatrix} \theta^{-\frac{1}{2}}\Sigma_C^{-\frac{1}{2}}\Sigma_{Cg}(\theta) \\ (1-\theta)^{-\frac{1}{2}}\Sigma_C^{-\frac{1}{2}}(\Sigma_{Cg}(1)-\Sigma_{Cg}(\theta)) \end{pmatrix} + \begin{pmatrix} \theta^{-\frac{1}{2}}\Sigma_C^{-\frac{1}{2}}\Omega^{\frac{1}{2}}W(\theta) \\ (1-\theta)^{-\frac{1}{2}}\Sigma_C^{-\frac{1}{2}}\Omega^{\frac{1}{2}}(W(1)-W(\theta)) \end{pmatrix}$$

uniformly on $[0,1]$. Hence, if we define $\mu_{X,T}(\theta) = \mu_X(\theta)$ to equal the first term, we obtain the convergence (15) for X_T.

Likewise, standard regression algebra (e.g. the FWL theorem) shows that

$$\sqrt{T}\hat{\delta}(\theta) \equiv \mathcal{A}_T(\theta)^{-1}[\mathcal{B}_T(\theta) + \mathcal{C}_T(\theta)],$$

for

$$\mathcal{A}_T(\theta) \equiv T^{-1}\sum_{t=[\theta T]+1}^{T} C_t C_t' - \left(T^{-1}\sum_{t=[\theta T]+1}^{T} C_t C_t'\right)\left(T^{-1}\sum_{t=1}^{T} C_t C_t'\right)^{-1}\left(T^{-1}\sum_{t=[\theta T]+1}^{T} C_t C_t'\right)$$

$$\mathcal{B}_T(\theta) \equiv T^{-1}\sum_{t=[\theta T]+1}^{T} C_t C_t' g(t/T) - \left(T^{-1}\sum_{t=[\theta T]+1}^{T} C_t C_t'\right)\left(T^{-1}\sum_{t=1}^{T} C_t C_t'\right)^{-1}\left(T^{-1}\sum_{t=1}^{T} C_t C_t' g(t/T)\right)$$

$$\mathcal{C}_T(\theta) \equiv T^{-1/2}\sum_{t=[\theta T]+1}^{T} C_t U_t - \left(T^{-1}\sum_{t=[\theta T]+1}^{T} C_t C_t'\right)\left(T^{-1}\sum_{t=1}^{T} C_t C_t'\right)^{-1}\left(T^{-1/2}\sum_{t=1}^{T} C_t U_t\right).$$

Under (12) and (13), however, the continuous mapping theorem implies that

$$\mathcal{A}_{\mathcal{T}}(\theta) \to_p (1-\theta)\Sigma_C - (1-\theta)^2\Sigma_C\Sigma_C^{-1}\Sigma_C = \theta(1-\theta)\Sigma_C,$$
$$\mathcal{B}_{\mathcal{T}}(\theta) \to_p [\Sigma_{Cg}(1) - \Sigma_{Cg}(\theta)] - (1-\theta)\Sigma_C\Sigma_C^{-1}\Sigma_{Cg}(1) = \theta\Sigma_{Cg}(1) - \Sigma_{Cg}(\theta)$$
$$\mathcal{C}_{\mathcal{T}}(\theta) \Rightarrow \Omega^{1/2}(W(1) - W(\theta)) - (1-\theta)\Sigma_C\Sigma_C^{-1}\Omega^{1/2}W(1) = \Omega^{1/2}(\theta W(1) - W(\theta))$$

all uniformly over [0,1], where this convergence holds jointly with that for X_T. Hence, by another application of the continuous mapping theorem,

$$Y_T(\theta) = e_j'\sqrt{T}\hat{\delta}(\theta) \Rightarrow Y(\theta) = \frac{e_j'\Sigma_C^{-1}[\theta\Sigma_{Cg}(1) - \Sigma_{Cg}(\theta) + \Omega^{1/2}(\theta W(1) - W(\theta))]}{\theta(1-\theta)}.$$

Hence, if we define

$$\mu_Y(\theta) = \frac{e_j' \Sigma_C^{-1} [\theta \Sigma_{Cg}(1) - \Sigma_{Cg}(\theta)]}{\theta(1-\theta)}$$

then $\mu_{Y,T}(\theta) \to \mu_Y(\theta)$ uniformly in θ and we obtain the convergence (15), as desired.

Note that if the structural break model is correctly specified, so $g(t/T) = \mathbf{1}(t/T > \theta_0)d$ and $\Sigma_{Cg}(\theta) = \mathbf{1}(\theta > \theta_0)(\theta - \theta_0)\Sigma_C d$, then

$$\mu_Y(\theta) = \frac{d_j[\theta(1-\theta_0) - \mathbf{1}(\theta > \theta_0)(\theta - \theta_0)]}{\theta(1-\theta)} = \begin{cases} d_j \frac{\theta_0}{\theta} & \text{if } \theta > \theta_0, \\ d_j \frac{1-\theta_0}{1-\theta} & \text{if } \theta \leq \theta_0. \end{cases}$$

In particular, $\mu_Y(\theta_0) = d_j$, as desired. Given this structure, one can use our confidence interval constructions for $\mu_Y(\hat{\theta})$ to test for the presence of a structural break in parameter j under the maintained hypothesis that the structural break model is correctly specified. In particular, our confidence sets satisfy

$$\inf_d P_d(\mu_Y(\hat{\theta}) \in CS) \geq 1 - \alpha$$

so that a test defined as

$$\phi(CS) = \begin{cases} 1 & \text{if } 0 \notin CI \\ 0 & \text{if } 0 \in CI \end{cases}$$

has correct size under the (unconditional) null hypothesis $H_0 : d_j = 0$:

$$\begin{aligned} &P_{(d_1,\ldots,d_{j-1},0,d_{j+1},\ldots,d_k)}(\phi(CI) = 1) = P_{(d_1,\ldots,d_{j-1},0,d_{j+1},\ldots,d_k)}(0 \notin CS) \\ &= 1 - P_{(d_1,\ldots,d_{j-1},0,d_{j+1},\ldots,d_k)}(0 \in CS) \\ &= 1 - P_{(d_1,\ldots,d_{j-1},0,d_{j+1},\ldots,d_k)}(\mu_Y(\hat{\theta}) \in CS) \leq \alpha \end{aligned}$$

for all $d_1, \ldots, d_{j-1}, d_{j+1}, \ldots, d_k$.

Additional Conditioning Events Arguments as in the proof of Lemma 2 show that if we define

$$\bar{A}(\tilde{\theta}) \equiv \Sigma_Y(\tilde{\theta})^{-2} \sum_{i=1}^{d_X} \Sigma_{XY,i}(\tilde{\theta})^2,$$

$$\bar{B}_Z(\tilde{\theta}) \equiv 2\Sigma_Y(\tilde{\theta})^{-1} \sum_{i=1}^{d_X} \Sigma_{XY,i}(\tilde{\theta}) Z_{\tilde{\theta},i}(\tilde{\theta}),$$

$$\bar{C}_Z(\tilde{\theta}) \equiv \sum_{i=1}^{d_X} Z_{\tilde{\theta},i}(\tilde{\theta})^2 - c, \quad \bar{D}_Z(\tilde{\theta}) \equiv \bar{B}_Z(\tilde{\theta})^2 - 4\bar{A}(\tilde{\theta})\bar{C}_Z(\tilde{\theta}),$$

then for

$$\bar{\mathcal{L}}_1(Z_{\tilde{\theta}}) \equiv \begin{cases} \frac{-\bar{B}_Z(\tilde{\theta})+\sqrt{\bar{D}_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) \neq 0 \text{ and } \bar{D}_Z(\tilde{\theta}) \geq 0 \\ \frac{-\bar{C}_Z(\tilde{\theta})}{\bar{B}_Z(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) = 0 \text{ and } \bar{B}_Z(\tilde{\theta}) > 0 \\ -\infty & \text{otherwise,} \end{cases}$$

$$\bar{\mathcal{L}}_2(Z_{\tilde{\theta}}) \equiv \begin{cases} \frac{-\bar{B}_Z(\tilde{\theta})+\sqrt{\bar{D}_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) < 0 \text{ and } \bar{D}_Z(\tilde{\theta}) \geq 0 \\ \frac{-\bar{C}_Z(\tilde{\theta})}{\bar{B}_Z(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) = 0 \text{ and } \bar{B}_Z(\tilde{\theta}) > 0 \\ -\infty & \text{otherwise,} \end{cases}$$

$$\bar{\mathcal{U}}_1(Z_{\tilde{\theta}}) \equiv \begin{cases} \frac{-\bar{B}_Z(\tilde{\theta})-\sqrt{\bar{D}_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) < 0 \text{ and } \bar{D}_Z(\tilde{\theta}) \geq 0 \\ \frac{-\bar{C}_Z(\tilde{\theta})}{\bar{B}_Z(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) = 0 \text{ and } \bar{B}_Z(\tilde{\theta}) < 0 \\ \infty & \text{otherwise,} \end{cases}$$

$$\bar{\mathcal{U}}_2(Z_{\tilde{\theta}}) \equiv \begin{cases} \frac{-\bar{B}_Z(\tilde{\theta})-\sqrt{\bar{D}_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) \neq 0 \text{ and } \bar{D}_Z(\tilde{\theta}) \geq 0 \\ \frac{-\tilde{C}_Z(\tilde{\theta})}{\bar{B}_Z(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) = 0 \text{ and } \bar{B}_Z(\tilde{\theta}) < 0 \\ \infty & \text{otherwise.} \end{cases}$$

and

$$\bar{\mathcal{V}}(Z_{\tilde{\theta}}) \equiv 1\left\{(\bar{A}(\tilde{\theta}) = \bar{B}_Z(\tilde{\theta}) = 0)\right\}\tilde{C}_Z(\tilde{\theta})$$

we can write

$$\begin{aligned} \{\|X(\tilde{\theta})\|^2 \geq c\} = \{Y(\tilde{\theta}) \in [\bar{\mathcal{L}}_1(Z_{\tilde{\theta}}), \bar{\mathcal{U}}_1(Z_{\tilde{\theta}})], \bar{\mathcal{V}}(Z_{\tilde{\theta}}) \geq 0\} \\ \cup \{Y(\tilde{\theta}) \in [\bar{\mathcal{L}}_2(Z_{\tilde{\theta}}), \bar{\mathcal{U}}_2(Z_{\tilde{\theta}})], \bar{\mathcal{V}}(Z_{\tilde{\theta}}) \geq 0\}, \end{aligned}$$

However, $\tilde{A}(\tilde{\theta}) \geq 0$ by definition, and $\tilde{A}(\tilde{\theta}) = 0$ implies $\tilde{B}_Z(\tilde{\theta}) = 0$, so we can take

$$\bar{\mathcal{L}}_1(Z_{\tilde{\theta}}) = \begin{cases} \frac{-\bar{B}_Z(\tilde{\theta})+\sqrt{D_Z\left(\tilde{\theta}\right)}}{2\bar{A}(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) \neq 0 \text{ and } D_Z\left(\tilde{\theta}\right) \geq 0 \\ -\infty & \text{otherwise,} \end{cases}$$

$$\bar{\mathcal{L}}_2(Z_{\tilde{\theta}}) = -\infty$$

$$\bar{\mathcal{U}}_1(Z_{\tilde{\theta}}) = \infty$$

$$\bar{\mathcal{U}}_2(Z_{\tilde{\theta}}) = \begin{cases} \frac{-\bar{B}_Z(\tilde{\theta}) - \sqrt{D_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})} & \text{if } \bar{A}(\tilde{\theta}) \neq 0 \text{ and } D_Z(\tilde{\theta}) \geq 0 \\ \infty & \text{otherwise.} \end{cases}$$

Therefore we can drop $\bar{\mathcal{L}}_2(Z_{\tilde{\theta}})$ and $\bar{\mathcal{U}}_1(Z_{\tilde{\theta}})$ and define $\bar{\mathcal{L}}(Z_{\tilde{\theta}}) \equiv \bar{\mathcal{L}}_1(Z_{\tilde{\theta}})$, $\bar{\mathcal{U}}(Z_{\tilde{\theta}}) \equiv \bar{\mathcal{U}}_2(Z_{\tilde{\theta}})$. In this case, we see that if $\bar{\mathcal{V}}(Z_{\tilde{\theta}}) \geq 0$ then $\mathcal{Y}_\gamma(1, Z_{\tilde{\theta}}) = \left[\bar{\mathcal{L}}(Z_{\tilde{\theta}}), \bar{\mathcal{U}}(Z_{\tilde{\theta}})\right]$, while $\mathcal{Y}_\gamma(1, Z_{\tilde{\theta}}) = \emptyset$ otherwise. The result stated in the text follows immediately.