Inference After Estimation of Breaks*

Isaiah Andrews† Toru Kitagawa‡ Adam McCloskey§

May 28, 2020

Abstract

In an important class of econometric problems, researchers select a target parameter by maximizing the Euclidean norm of a data-dependent vector. Examples that can be cast into this frame include threshold regression models with estimated thresholds and structural break models with estimated breakdates. Estimation and inference procedures that ignore the randomness of the target parameter can be severely biased and misleading when this randomness is non-negligible. This paper studies conditional and unconditional inference in such settings, accounting for the data-dependent choice of target parameters. We detail the construction of quantile-unbiased estimators and confidence sets with correct coverage, and prove their asymptotic validity under data generating process such that the target parameter remains random in the limit. We also provide a novel sample splitting approach that improves on conventional split-sample inference.

Keywords: Selective Inference, Sample Splitting, Structural Breaks, Threshold Regression, Misspecification

JEL Codes: C12, C13

*We thank Frank Schoerfheide, Jesse Shapiro, Tim Vogelsang, two anonymous referees and participants at multiple seminars for helpful comments. Some of the results in this paper previously appeared in the 2018 working paper version of I. Andrews et al. (2019). Andrews gratefully acknowledges financial support from the NSF under grant number 1654234. Kitagawa gratefully acknowledges financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) and the European Research Council (Starting grant No. 715940).

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

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

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

1 Introduction

In a variety of economic settings, researchers select a target parameter by maximizing the Euclidean norm of a data-dependent vector. For example, in threshold regressions and structural break models researchers commonly estimate the location of a break or threshold by minimizing a residual sum of squares or, equivalently, maximizing an explained sum of squares. Researchers then estimate and form confidence sets for the magnitude of the discontinuity, taking the estimated threshold or break as given (see e.g. Hansen (2000) and Perron (2006)). Estimation and inference procedures that do not account for the data-driven selection of the target parameter in such settings can perform very poorly when the maximizer is variable. In the structural break and threshold regression settings, this corresponds to the empirically-relevant case where the location of the threshold/break is unknown, and is not obvious from the data. In such cases conventional estimators may be badly biased, and conventional confidence intervals may under-cover.

This paper builds on the results of our companion paper, I. Andrews et al. (2019) (henceforth abbreviated AKM), to derive quantile-unbiased estimators and valid confidence sets. In AKM we develop results on estimation and inference following abstract data-dependent selection of a target parameter in a normal model, proving a variety of validity and optimality results. The form of the resulting procedures depends on the nature of selection, and AKM works out details for the case where the target parameter is chosen by maximizing the *level* of an asymptotically normal random variable. The translation of results from the normal model to results on asymptotic validity also depends on the form of selection considered, and AKM further shows uniform asymptotic validity for the level-maximization case. In the present paper, we derive the form of the AKM estimators and confidence sets in the settings where the target parameter is chosen by maximizing the *norm*, rather than the level, of asymptotically normal random variables and prove asymptotic validity of these procedures for the first time in this class of problems. For asymptotic validity, we focus on sequences of parameter spaces such that the norm-maximizing value is random, even asymptotically.

Our results also build on the rapidly expanding statistics literature on selective inference, which has primarily considered inference on regression parameters after using popular model-selection tools. See e.g. Harris et al. (2016), Lee et al. (2016), Tian and Taylor (2018), Fithian et al. (2017), Tibshirani et al. (2018), and references therein. To implement selective inference procedures one needs a tractable representation of the selection event of interest as a function of sufficient statistics for nuisance parameters. We derive such

a representation for norm-maximization problems, while existing papers in the literature have developed analogous representations for other questions, for example inference after LASSO model selection in Lee et al. (2016). Tibshirani et al. (2018) also develop results on asymptotic validity for parameter spaces similar to those we consider, but their results cover neither the norm-maximization setting we study nor some of the estimators and confidence intervals we discuss.[1] In contrast to the level-maximization setting studied by AKM, we show that norm-maximization induces a non-convex conditioning event. Deriving a tractable form for this conditioning event and establishing validity of the resulting procedures constitute the main theoretical contributions of this paper.

Split-sample inference provides an alternative solution to the failures of conventional estimators and confidence sets in this setting. In a split-sample approach, one selects the target parameter based on one part of the data (or, alternatively, a noised-up version of the full data), and then conducts inference using the remainder of the data. In AKM we derive an improved split-sample approach for abstract selection events, which dominates conventional split-sample inference. In the present paper, we work out the details of improved split-sample inference in norm-maximization settings.

We illustrate how our estimation and inference procedures can be applied to structural break and threshold regression settings. Building upon Elliott and Müller (2007, 2014) and Lee and Wang (2020), we show how to cast estimation and inference in these models as norm-maximization problems. In a notable departure from these papers, our estimators and confidence intervals do not require the threshold or structural break model to be correctly specified. This added generality is important, since researchers sometimes fit a threshold model as a parsimonious approximation in settings where the parameters may in fact change in a more continuous manner. Hence, if we perform inference after estimating the breaks, the pseudo-true parameter defined based on an estimated change point becomes a natural object of interest.

Finally, we examine the performance of our proposed methods in threshold regression simulations calibrated to data from Card et al. (2008). These authors studied the dynamics of neighborhood segregation by comparing the change in white share between 1980 and 1990 to the minority share in 1980. They fit a model which allows for "tipping point" dynamics, where an increase in the minority share beyond some threshold leads to a discontinuous

[1] In particular, the inference procedures studied in Tibshirani et al. (2018) are all what we term *conditional* below, in that they condition on a selection event, while we consider both conditional and unconditional inference procedures.

decrease in the white share ("white flight"). Their theoretical model, by contrast, predicts large but potentially continuous changes as a function of the minority share, and so suggests the discontinuous tipping point model may be misspecified. Our simulations calibrated to this application highlight that conventional estimation and inference procedures can perform very poorly in terms of bias and coverage when the target parameter is selected through norm-maximization. By contrast, our new procedures perform well in terms of both bias and coverage, and outperform existing alternatives. In particular, Card et al. (2008) originally conducted inference based on a split-sample approach, and we find substantial performance gains from our improved split-sample methods

The next section illustrates the pitfalls of conventional inference, and outlines the goals of our corrections, in a stylized norm-maximization problem. Section 3 discusses the norm-maximization problem in the context of a normal model, and shows that both the structural break and threshold regression examples are asymptotically normal under a small-break asymptotic approximation. We then derive the expressions needed to implement the AKM inference procedures in a norm-maximization setting. Section 4 establishes the asymptotic validity of our estimators and confidence intervals in norm-maximization settings. Finally, Section 5 discusses implementation of our improved split-sample procedures, while Section 6 wraps up with our simulation study based on Card et al. (2008). Proofs of the theoretical results in Section 3 can be found in the Appendix. Proofs of some of the theoretical results in Section 4, results on alternative confidence interval constructions and further simulation results can be found in the Supplemental Appendix.

2 Norm Maximization in a Stylized Example

We begin by considering a stylized example inspired by Romano and Wolf (2005). In particular, suppose we compare two investment strategies in a backtest, and seek to estimate and form a confidence interval for the expected return of the strategy with the largest absolute historical average return.[2] To further simplify the analysis for this section, let us suppose the returns of the two strategies are statistically independent. Such independence is neither required nor imposed in the rest of the paper.

Let $R_{i,1}$ and $R_{i,2}$ denote the observed returns of the two investment strategies in period i for a sample of observations $i = 1,...,n$ and $\bar{R} = (\bar{R}_1, \bar{R}_2)' = n^{-1}\sum_{i=1}^{n} R_i$. We assume returns are stationary, and when $|\bar{R}_1| \geq |\bar{R}_2|$ we are interested in inference on $E[R_{i,1}]$, while

[2]Treating negative and positive historical returns symmetrically can be justified by the ability to take short positions.

when $|\bar{R}_2| > |\bar{R}_1|$ we are interested in inference on $E[R_{i,2}]$. Standard weak dependence and moment conditions imply a central limit theorem:

$$\sqrt{n}(\bar{R} - E[R_i]) \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

where Σ is a diagonal, consistently estimable variance matrix.

To capture the feature that average returns are small relative to sampling uncertainty, let us further model $E[R_i] = \mu/\sqrt{n}$ for a fixed vector μ, so

$$X_n = \sqrt{n}\bar{R} \xrightarrow{d} X \sim \mathcal{N}(\mu,\Sigma). \tag{1}$$

This "small return" approximation ensures that both investment strategies are chosen with positive probability even in large samples, and hence that our asymptotic analysis captures the finite-sample uncertainty about which strategy has the highest absolute return.[3] If we were to instead fix $E[R_i]$, then so long as the elements of $|E[R_i]|$ are not equal, the strategy with the largest absolute expected return would be chosen with probability one asymptotically.

For $\hat{j} = \text{argmax}_j |X_j|$, our inference problem thus asymptotically resembles that of estimation and inference on $\mu_{\hat{j}}$ based on observing $X \sim \mathcal{N}(\mu,\Sigma)$ for known Σ. We thus study this asymptotic problem, in the hope (borne out by the results of Section 4) that finite-sample results for procedures based on (X,Σ) will translate to asymptotic results for procedures based on X_n and a variance estimator $\widehat{\Sigma}_n$.

Since X_j is an unbiased estimator for μ_j, one may be tempted to use $X_{\hat{j}}$ to estimate $\mu_{\hat{j}}$ and to form a level-$(1-\alpha)$ confidence interval in the standard way as

$$[X_{\hat{j}} - \sqrt{\Sigma_{\hat{j}\hat{j}}} z_{1-\alpha/2}, X_{\hat{j}} + \sqrt{\Sigma_{\hat{j}\hat{j}}} z_{1-\alpha/2}], \tag{2}$$

for $z_{1-\alpha/2}$ the $1-\alpha/2$ quantile of a standard normal distribution. Recall, however, that $\hat{j}$ is random, so in general $X_{\hat{j}}$ is biased and (2) does not have correct coverage for $\mu_{\hat{j}}$. To understand why, suppose that $\hat{j} = 1$, or equivalently that $|X_1| \geq |X_2|$. Conditional on $\hat{j} = 1$, $X_{\hat{j}}$ is distributed as a normal variable with mean μ_1 and variance Σ_{11}, truncated to lie outside the random interval $(-|X_2|, |X_2|)$. We thus see that the distribution of $X_{\hat{j}}$ conditional on the realized value of $\hat{j} = 1$ is neither normal nor symmetric about μ_1 for $\mu_1 \neq 0$.

[3] These drifting sequence asymptotics can be considered a form of weak-identification asymptotics, where the best performing investment strategy is weakly identified.

To better understand the behavior of $X_{\hat{j}}$ in this setting, let us further condition on the realized value of X_2. The conditional mean of $X_{\hat{j}}$ given $\hat{j}=1$ and $X_2=x_2$ is

$$E[X_{\hat{j}}|\hat{j}=1,X_2=x_2]=E[X_1||X_1|\geq|x_2|]=\mu_1+\sigma_1\frac{\phi\left(\frac{|x_2|-\mu_1}{\sqrt{\Sigma_{11}}}\right)-\phi\left(\frac{-|x_2|-\mu_1}{\sqrt{\Sigma_{11}}}\right)}{1-\Phi\left(\frac{|x_2|-\mu_1}{\sqrt{\Sigma_{11}}}\right)+\Phi\left(\frac{-|x_2|-\mu_1}{\sqrt{\Sigma_{11}}}\right)}. \quad (3)$$

The conditional bias thus has the same sign as $\phi\left(\frac{|x_2|-\mu_1}{\sqrt{\Sigma_{11}}}\right)-\phi\left(\frac{-|x_2|-\mu_1}{\sqrt{\Sigma_{11}}}\right)$, which for $x_2\neq 0$ has the same sign as μ_1. Since the sign of the bias is the same for almost every x_2, the bias of $X_{\hat{j}}$ conditional on $\hat{j}=1$ likewise has the same sign as μ_1. Since the analogous argument holds for $\hat{j}=2$, we thus see that $X_{\hat{j}}$ is biased away from zero conditional on $\hat{j}=j$, in the sense that $|E[X_{\hat{j}}|\hat{j}=j]|>|\mu_j|$ whenever $\mu_j\neq 0$. As this bias suggests, the conventional confidence set (2) will undercover.

In this paper we build on the results of AKM to overcome the problem of biased estimation and undercoverage for norm maximization settings. In particular, we develop a conditionally α-quantile-unbiased estimator $\hat{\mu}_\alpha$, which has the property that

$$P_\mu\left\{\hat{\mu}_\alpha\geq\mu_{\hat{j}}|\hat{j}=j\right\}=\alpha \text{ for } j\in\{1,2\} \text{ and all } \mu.$$

These estimators can be used to form equal-tailed $1-\alpha$ level confidence intervals $[\hat{\mu}_{\frac{\alpha}{2}},\hat{\mu}_{1-\frac{\alpha}{2}}]$, which have correct conditional coverage given $\hat{j}$,

$$P_\mu\left\{\mu_{\hat{j}}\in CS|\hat{j}=j\right\}=1-\alpha \text{ for } j\in\{1,2\} \text{ and all } \mu.$$

If the level-maximization criterion considered in AKM were used instead, i.e., $\hat{j}=\text{argmax}_j X_j$, the conditioning event in (3) would be given by a half-open interval $\{X_1\geq x_2\}$ rather than the union of the disconnected intervals $\{X_1\geq|x_2|\}\cup\{X_1\leq-|x_2|\}$. Below we show that in general, norm-maximization problems lead to conditioning events that can be represented as finite unions of disconnected intervals, unlike level-maximization problems for which the conditioning events are intervals.

Conditional quantile unbiasedness and conditional coverage are demanding requirements, and simulations in AKM suggest that they can come at the cost of unconditional performance.[4] Hence, for cases where we are satisfied with controlled unconditional bias,

[4]Whether a conditional or unconditional coverage constraint is more appropriate is context-specific. We refer the interested reader to AKM for further discussion of this point.

following AKM we also introduce estimators $\hat{\mu}_\alpha^H$ such that

$$|P_\mu\{\hat{\mu}_\alpha^H \geq \mu_{\hat{j}}\} - \alpha| \leq \beta \cdot \max\{\alpha, 1-\alpha\} \text{ for all } \mu,$$

where β is a user-selected constant, and confidence sets with correct unconditonal coverage,

$$Pr_\mu\{\mu_{\hat{j}} \in CS\} \geq 1-\alpha \text{ for all } \mu.$$

3 Norm Maximization in the Normal Model

This section introduces a finite-dimensional normal model with norm maximization which generalizes the stylized example in the last section, and shows that this model arises as an asymptotic approximation to threshold regression and structural break models. We then briefly introduce the inference procedures of AKM and derive the expressions needed to use these procedures in the norm-maximization setting.

As in the general setting of AKM, assume we observe normal random vectors $\left(X(\theta)', Y(\theta)\right)'$ for $\theta \in \Theta$ where Θ is a finite set, $X(\theta) \in \mathbb{R}^{d_X}$, and $Y(\theta) \in \mathbb{R}$. In particular, for $\Theta = \{\theta_1, ..., \theta_{|\Theta|}\}$, let $X = \left(X(\theta_1)', ..., X(\theta_{|\Theta|})'\right)'$ and $Y = \left(Y(\theta_1), ..., Y(\theta_{|\Theta|})\right)'$. Then

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

for

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

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

We assume that Σ is known, while μ is unknown and unrestricted unless noted otherwise. We abbreviate $\Sigma(\theta, \theta)$ by $\Sigma(\theta)$. We assume throughout that $\Sigma_Y(\theta) > 0$ for all $\theta \in \Theta$, since the inference problem is trivial when $\Sigma_Y(\theta) = 0$. We distinguish between the blocks of $\Sigma(\theta, \tilde{\theta})$, i.e., $\Sigma_X(\theta, \tilde{\theta})$, $\Sigma_{XY}(\theta, \tilde{\theta})$, $\Sigma_{YX}(\theta, \tilde{\theta})$ and $\Sigma_Y(\theta, \tilde{\theta})$, since two of these blocks are used explicitly in the formulation of our estimation and inference approaches later in the paper (see Section 3.2). Here, and throughout the text θ and $\tilde{\theta}$ denote fixed elements of Θ.

We are interested in inference on $\mu_Y(\hat{\theta})$ where $\hat{\theta}$ is chosen by norm maximization

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

for $\|\cdot\|$ the Euclidean norm. The stylized example of the previous section corresponds to the special case for which $Y = X$, $d_X = 1$ and $|\Theta| = 2$.

3.1 Threshold Regression and Structural Break Estimation

Suppose we observe data on an outcome Y_i, a scalar threshold regressor Q_i and a k-dimensional vector of regressors C_i for $i \in \{1, ..., n\}$. We assume there is a linear but potentially regressor-dependent relationship between Y_i and C_i:

$$Y_i = C_i'(\beta + \varphi_n(Q_i)) + U_i, \tag{6}$$

where $Q_i \in \mathbb{R}$ and the residuals U_i are orthogonal to Q_i and C_i. Similarly to Elliott and Müller (2014), the function $\varphi_n : \mathbb{R} \to \mathbb{R}^k$ determines the value of the regressor-dependent coefficient $\beta + \varphi_n(Q_i)$. This model nests the traditional threshold regression model (see e.g. Hansen (2000) and references therein) by taking

$$\varphi_n(Q_i) = 1\{Q_i > \theta\}\delta, \tag{7}$$

where $\theta \in \mathbb{R}$ is the "true" threshold. This model also nests a time-varying parameters regression model where the observations $i = 1, ..., n$ are ordered and $Q_i = i/n$ denotes the position of observation i in the sample. The traditional structural change model is a special case of this time-varying parameters model (see Bai (1997) and Perron (2006)) for which (7) holds and $\theta \in (0,1)$ is the "true" break fraction. For the remainder of this paper we focus terminology on the threshold regression, with the understanding that the analysis also applies to the special case of a regression model with a structural break.

The threshold model (7) is often used as a parsimonious approximation to the more general linear regression model (6) with regressor-dependent coefficents. For instance, as noted above Card et al. (2008) use the threshold model to approximate a theoretical model where the regressor-dependent coefficients may change smoothly. Hansen (1997, 2000) also notes that the threshold regression model is often used as a misspecified but parsimonious approximation to a more general class of nonlinear regression models. Since the threshold regression model is widely used in practice as an approximating model, we consider a

researcher who fits the model (7), but to allow for the possibility of misspecification we assume only that the data are generated by (6). Note that this modeling framework also covers the case of a standard linear regression model with no change in its coefficients, for which $\varphi_n(\cdot)=0$ in formulation (6) or $\delta=0$ in formulation (7).

To provide a good asymptotic approximation to finite sample behavior, we follow Elliott and Müller (2007, 2014) and Lee and Wang (2020) and model parameter instability as being on the same order of magnitude as sampling uncertainty, with $\varphi_n(Q_i)=\frac{1}{\sqrt{n}}g(Q_i)$ for a fixed function g. As in the stylized example of the previous section, this DGP allows the asymptotic problem to reflect an important feature of the finite sample problem in many applications, namely that the data provide limited information about the regressor-dependent coefficient function $\varphi_n(\cdot)$. See Elliott and Müller (2007, 2014) for further justification of this drifting sequence DGP.[5]

We further assume that

$$\frac{1}{n}\sum_{i=1}^{n} C_i C_i' 1\{Q_i \le \theta\} \to_p \Sigma_C(\theta), \quad \frac{1}{n}\sum_{i=1}^{n} C_i C_i' g(Q_i) 1\{Q_i \le \theta\} \to_p \Sigma_{Cg}(\theta), \tag{8}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} C_i U_i 1\{Q_i \le \theta\} \Rightarrow G(\theta), \tag{9}$$

all uniformly in $\theta \in \mathbb{R}$. Here $\Sigma_C : \mathbb{R} \to \mathbb{R}^{k\times k}$ is a consistently-estimable matrix-valued function and $\Sigma_C(\theta)$ is full rank for all θ in the interior of the support of Q_i, $\Sigma_{Cg}:\mathbb{R}\to\mathbb{R}^k$ is a vector-valued function, and $G(\cdot)$ is a k-dimensional mean zero Gaussian process with a consistently estimable covariance function that is positive definite when evaluated at points in the interior of the support of Q_i. Conditions (8) and (9) are analogous to Conditions 1(ii) and 1(iv) of Elliott and Müller (2007) for structural break models in a time-series setting. See Hansen (2000) and Lee and Wang (2020) for sufficient conditions that give rise to (8) and (9). Although these high-level conditions cover both threshold and structural break

[5]While we model the degree of parameter instability as local to zero, if one instead models the degree of parameter instability as fixed but takes $\varphi_n=\varphi$ to be continuous, rather than a step function, estimates based on (7) also exhibit nonstandard behavior, with a $n^{-\frac{1}{3}}$ rate of convergence for the estimated threshold rather than the n^{-1} rate obtained for the correctly-specified case. See Bühlmann and Yu (2002) and Banerjee and McKeague (2007). Song et al. (2016) shows that nonstandard asymptotic behavior arises even when the threshold model is only locally misspecified, while Hansen (2017) obtains nonstandard asymptotic results in the regression kink case, where there is a discontinuity in the derivative rather than the level. Inference results that, like those in the present paper, account for estimation of the threshold are important in all of these settings, and are discussed in the references above.

applications, appropriate low-level sufficient conditions, and the form of $\Sigma_C(\cdot)$, $\Sigma_{Cg}(\cdot)$, and $G(\cdot)$, will differ across applications.

The standard threshold estimator $\hat{\theta}_n$ chooses θ to minimize the sum of squared residuals in an OLS regression of Y_i on C_i and $\mathbf{1}\{Q_i>\theta\}C_i$ across a finite grid of thresholds Θ. For

$$X_n(\theta)=\begin{pmatrix} (\sum_{i=1}^n C_iC_i'1\{Q_i\leq\theta\})^{-\frac{1}{2}}(\sum_{i=1}^n C_i\eta_i1\{Q_i\leq\theta\}) \\ (\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\})^{-\frac{1}{2}}(\sum_{i=1}^n C_i\eta_i1\{Q_i>\theta\}) \end{pmatrix},$$

with $\eta_i\equiv U_i+n^{-1/2}C_i'g(Q_i)$, arguments analogous to those in the proof of Proposition 1 in Elliott and Müller (2007) imply that $\hat{\theta}_n=\text{argmax}_{\theta\in\Theta}\|X_n(\theta)\|+o_p(1)$, where $o_p(1)$ is an asymptotically negligible term.

Suppose we are interested in the approximate change in the jth parameter $\delta_j=e_j'\delta$, where e_j is the j^{th} standard basis vector.[6] In practice it is common to estimate δ by least squares imposing the estimated threshold $\hat{\theta}_n$. When the threshold regression model (7) is misspecified, however, there is neither a "true" threshold θ nor a "true" change coefficient δ. Instead, the population regression coefficient $\delta(\theta)$ imposing threshold θ depends on θ. Thus, we are interested in $\delta_j(\hat{\theta}_n)$, the population regression coefficient at the estimated threshold. Denote the OLS estimate imposing threshold θ by $\hat{\delta}_j(\theta)$ and define $Y_n(\theta)=\sqrt{n}\hat{\delta}_j(\theta)$. If we define $\mu_Y(\theta)=\lim_{n\to\infty}\sqrt{n}\delta_j(\theta)$ as the scaled coefficient of interest and $\mu_X(\theta)$ as the population analog of $X_n(\theta)$,[7] Section A.1 of the Appendix shows that

$$\begin{pmatrix} X_n(\theta) \\ Y_n(\theta) \end{pmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix} \mu_X(\theta) \\ \mu_Y(\theta) \end{pmatrix},\Sigma(\theta)\right) \tag{10}$$

uniformly over a parameter space Θ contained in the interior of the support of Q_i, where the covariance matrix $\Sigma(\theta)$ is consistently estimable but $\mu_X(\theta)$ and $\mu_Y(\theta)$ are not. This corresponds to the asymptotic problem (4) where $\hat{\theta}$ is defined through norm-maximization (5).[8]

Inference in Threshold and Break Models Since the estimated threshold $\hat{\theta}_n$ is random and the parameter of interest $\delta_j(\theta)$ (or equivalently $\mu_{Y,n}(\theta)$) 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 threshold $\hat{\theta}_n$, since we only seek

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

[7]See Section A.1 of the Appendix for precise definitions of these quantities.

[8]Estimators and confidence intervals for the object of interest $\delta_j(\hat{\theta})$ can be obtained by a simple $\sqrt{n}$-rescaling of the corresponding estimators and confidence intervals for $\mu_Y(\hat{\theta})$.

to conduct inference on $\delta_j(\tilde{\theta})$ when $\hat{\theta}_n = \tilde{\theta}$. Even if we only desire coverage of $\delta_j(\hat{\theta}_n)$ on average over the distribution of $\hat{\theta}_n$, and so prefer unconditional estimators and confidence intervals, accounting for the randomness of $\hat{\theta}_n$ remains important.

It may also be natural to condition inference on additional variables. For example, if we report a confidence interval for the change coefficient $\delta_j(\hat{\theta}_n)$ only when we reject the null hypothesis of parameter constancy, $H_0 : \varphi_n(\theta) = 0$ for all θ, it is natural to condition inference on this rejection. This can be accomplished by defining $\hat{\gamma}_n = \gamma(X_n)$ to be a dummy for rejection of H_0, and conditioning inference on $(\hat{\theta}_n, \hat{\gamma}_n)$. See Propositions 1 and 2 below for further details.

If we are confident that the threshold model is correctly specified, so that (7) holds in the data, it is conceptually more appealing to focus on inference for the "true" parameters as in Elliott and Müller (2014) and Lee and Wang (2020). However, we note that these latter inference procedures can become computationally intractable when C_i has more than a few elements, so that even in the correctly-specified setting inference on $\delta_j(\hat{\theta}_n)$ may be the only feasible option. In addition, Proposition 5 of AKM implies that when the break magnitude $\|\delta\|$ is large enough that $\hat{\theta}_n$ takes a single value with very high probability, the conditional inference procedures here collapse to standard efficient inference on the "true" δ_j.[9] On the other hand, the computationally feasible standard threshold and structural break confidence intervals are based upon normal approximations to $\hat{\delta}_j(\hat{\theta}_n)$ (see e.g., Bai (1997), Bai and Perron (1998) and Hansen (2000)). Therefore, when the true break magnitude is not large enough to overwhelm sampling variability (in accord with the asymptotic approximations of this section), they can have poor coverage for either $\hat{\delta}_j(\hat{\theta}_n)$ or the "true" δ_j.

One solution to this problem that has been used in the literature is sample splitting, where the first part of the sample is used to form $\hat{\theta}_n$ and the second part is used to form $\hat{\delta}(\cdot)$, so that a conventional normal approximation can be applied to $\hat{\delta}_j(\hat{\theta}_n)$.[10] See e.g. Card et al. (2008). Like the methods proposed in this paper, this form of split-sample inference is valid for $\delta_j(\hat{\theta}_n)$, not the "true" δ_j. However, when $\hat{\theta}_n$ is formed using only the first fraction of the data, it is a less precise estimator of the (pseudo-)true break fraction than when using the full data as we do here. Even in the case for which one wishes to only use a fraction of the data to form $\hat{\theta}_n$, Section 5 provides an improved split-sample inference approach that dominates the standard method.

[9] Elliott and Müller (2014) employ a switching scheme such that their approach nearly reduces to standard inference, and increase their critical values to account for the switch.

[10] While direct application of this approach fails in structural break settings since the data may be non-stationary, an analogous effect can be achieved by adding normal noise: see Section 5 below for details.

Other Norm-Maximization Examples While our discussion of threshold regression estimation focuses on the linear model (6), Elliott and Müller (2014) show that structural break estimation in nonlinear models with time-varying parameters gives rise to the same asymptotic problem. Hence, our results apply in that setting as well. Likewise, the same asymptotic problem arises in nonlinear threshold models.[11] Further afield, one could generalize our approach to consider norm-minimization rather than norm-maximization, and so derive results for GMM-type problems with finite parameter spaces.

3.2 Inference in the Normal Model

AKM studies estimators for $\mu_Y(\hat{\theta})$ that are quantile-unbiased conditional on the realization of $\hat{\theta}$, perhaps along with the value of an additional conditioning variable $\hat{\gamma}=\gamma(X)$. The mean vector μ_X is a nuisance parameter in this problem, and AKM notes that

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

is a sufficient statistic for the unknown mean vector μ_X. Hence, the inference procedures derived in AKM condition on $(\hat{\theta},\hat{\gamma})=(\tilde{\theta},\tilde{\gamma})$ and $Z_{\tilde{\theta}}=z$. Conditional on $Z_{\tilde{\theta}}=z$, the event that $(\hat{\theta},\hat{\gamma})=(\tilde{\theta},\tilde{\gamma})$ is equivalent to the event that $Y(\tilde{\theta})\in\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$ for a set $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$ which depends on the nature of $\hat{\theta}$ and $\hat{\gamma}$. AKM derives the form of $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$ for level-maximization problems, while we derive the form for norm-maximization problems here.

AKM introduces an efficient conditionally α-quantile-unbiased estimator of $\mu_Y(\hat{\theta})$, $\hat{\mu}_\alpha$, with the property

$$Pr_\mu\Big\{\hat{\mu}_\alpha\geq\mu_Y(\tilde{\theta})|\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\Big\}=\alpha \text{ for all } \mu,\ \tilde{\theta},\ \tilde{\gamma}.$$

This estimator enables the construction of an equal-tailed confidence interval $CS=[\hat{\mu}_{\frac{\alpha}{2}},\hat{\mu}_{1-\frac{\alpha}{2}}]$ with conditional coverage $1-\alpha$:

$$Pr_\mu\Big\{\mu_Y(\tilde{\theta})\in CS|\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\Big\}=1-\alpha \text{ for all } \mu,\ \tilde{\theta},\ \tilde{\gamma}.$$

AKM shows that the estimator $\hat{\mu}_\alpha$ and confidence set CS posesses the additional appealing properties that for any sequence of means μ_m such that $Pr_{\mu_m}\Big\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\Big\}\to 1$, $\hat{\mu}_\alpha\to_p Y(\tilde{\theta})$ while CS converges to the conventional confidence interval that ignores selection. Hence,

[11] In a manuscript circulated after the initial public version of this paper, Hyun et al. (2018) consider the related problem of conditional inference for changepoint detection, but the changepoint estimation methods they consider cannot be cast as norm-maximization, so their results do not overlap with ours.

in cases where selection-corrected inference is unnecessary, there is no "price" for using the conditional procedures discussed above.

Imposing conditional unbiasedness and coverage can be costly from an unconditional perspective. Results in Kivaranovic and Leeb (2020) imply that conditional confidence sets can have infinite expected length. If one cares only about unconditional coverage, an alternative is to start with a joint confidence interval for μ_Y and project on the dimension corresponding to $\hat{\theta}$. This general approach has been applied in various contexts in the literature – see AKM for examples. To formally discuss this approach, let c_α denote the $1-\alpha$ quantile of $\max_{\theta\in\Theta}|\xi(\theta)|/\sqrt{\Sigma_Y(\theta)}$ for $\xi\sim N(0,\Sigma_Y)$. Define the level $1-\alpha$ projection confidence interval as

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

This interval has correct unconditional coverage $Pr_\mu\left\{\mu_Y(\hat{\theta})\in CS_P^\alpha\right\}\geq 1-\alpha$ for all μ, but does not in general have correct conditional coverage in the sense that we may have $Pr_\mu\left\{\mu_Y(\hat{\theta})\in CS_P^\alpha|\hat{\theta}=\tilde{\theta}\right\}<1-\alpha$ for some $\tilde{\theta}$ and μ. Moreover, the length of CS_P^α does not depend on (X,Y), so in cases where $Pr_\mu\left\{\hat{\theta}=\tilde{\theta}\right\}\approx 1$ for some $\tilde{\theta}$, CS_P^α will tend to be much longer than the conditional confidence sets discussed above.

In order to overcome some of the weaknesses of both the conditional and projection procedures, AKM introduces a hybrid estimator, $\hat{\mu}_\alpha^H$, and confidence interval, CS^H, that condition both on $\hat{\theta}$ and on the (possibly incorrect) event that $\mu_Y(\hat{\theta})\in CS_P^\beta$ for some $0\leq\beta\leq\alpha$. This latter conditioning limits the worst-case dispersion of the hybrid estimator and the worst-case length of the hybrid confidence interval. AKM shows that the unconditional level-α quantile bias of $\hat{\mu}_\alpha^H$ is controlled,

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

Hence, for example, the absolute median bias of $\hat{\mu}_{\frac{1}{2}}^H$ (measured as the deviation of the exceedance probability from 1/2) is bounded above by $\beta/2$. Using these estimators, we form the level $1-\alpha$ equal-tailed hybrid confidence intervals as

$$CS^H = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{2-2\beta}}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{2-2\beta}}\right].$$

Note that we have adjusted the quantiles considered to account for the possibility that

CS_P^β may not cover $\mu_Y(\hat{\theta})$. Finally, AKM also introduces confidence intervals that are uniformly most accurate unbiased in the conditional problem as well as analogous hybrid confidence intervals. For brevity, we defer discussion of these procedures to Section B of the Supplemental Appendix.

3.3 Conditioning Sets for Norm-Maximization Problems

To implement AKM procedures in a given setting, we need tractable representations for the sets $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$. AKM derives such representations in cases where $\hat{\theta}$ is selected by maximizing the *level* of $X(\theta)$ ($d_X=1$), but do not consider the norm-maximization setting studied here. Since the set of X values such that $\hat{\theta}=\tilde{\theta}$, $X\in\mathcal{X}(\tilde{\theta})=\left\{X:\|X(\tilde{\theta})\|=\max_{\theta\in\Theta}\|X(\theta)\|\right\}$, involves nonlinear constraints, other results in the existing literature (e.g., Lee et al. (2016)) likewise do not apply. Hence, in this section we derive $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$ for norm-maximization settings.

In norm-maximization settings without additional conditioning variables, the general expression for $\mathcal{Y}(\tilde{\theta},z)$ is long, but easy to calculate in applications.

Proposition 1

Define

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

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

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

For

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

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

define

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

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

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

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

and

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

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* $\mathcal{Y}(\tilde{\theta},Z_{\tilde{\theta}})=\emptyset$.

Note that $Pr_\mu\left\{\mathcal{V}(\hat{\theta},Z_{\hat{\theta}})<0\right\}=0$ for all μ so we can ignore this constraint in applications. While the expression for $\mathcal{Y}(\tilde{\theta},z)$ is long, it simplifies substantially in some special cases.

Corollary 1

Suppose $d_X=1$, $X=Y$, *and* Σ_X *is full rank.*

(i) Define $H_Z(\tilde{\theta},\theta)=-\Sigma_X(\tilde{\theta})Z_{\tilde{\theta}}(\theta)/(2\Sigma_X(\theta,\tilde{\theta}))$,

$$G_Z(\tilde{\theta},\theta)=\frac{\Sigma_X(\tilde{\theta})\Sigma_X(\theta,\tilde{\theta})Z_{\tilde{\theta}}(\theta)-\Sigma_X(\tilde{\theta})^2|Z_{\tilde{\theta}}(\theta)|}{\Sigma_X(\tilde{\theta})^2-\Sigma_X(\theta,\tilde{\theta})^2},$$
$$K_Z(\tilde{\theta},\theta)=\frac{\Sigma_X(\tilde{\theta})\Sigma_X(\theta,\tilde{\theta})Z_{\tilde{\theta}}(\theta)+\Sigma_X(\tilde{\theta})^2|Z_{\tilde{\theta}}(\theta)|}{\Sigma_X(\tilde{\theta})^2-\Sigma_X(\theta,\tilde{\theta})^2}.$$

Then, for

$$\ell_Z^1(\tilde{\theta})=\max\left\{\max_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|>\Sigma_X(\tilde{\theta})}G_Z(\tilde{\theta},\theta),\max_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|=\Sigma_X(\tilde{\theta}),\Sigma_X(\theta,\tilde{\theta})Z_{\tilde{\theta}}(\theta)<0}H_Z(\tilde{\theta},\theta)\right\},$$

$$\ell_Z^2(\tilde{\theta},\theta)=\max\left\{\max_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|>\Sigma_X(\tilde{\theta})}G_Z(\tilde{\theta},\theta),\max_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|=\Sigma_X(\tilde{\theta}),\Sigma_X(\theta,\tilde{\theta})Z_{\tilde{\theta}}(\theta)<0}H_Z(\tilde{\theta},\theta),K_Z(\tilde{\theta},\theta)\right\},$$

$$u_Z^1(\tilde{\theta},\theta)=\min\left\{\min_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|>\Sigma_X(\tilde{\theta})}K_Z(\tilde{\theta},\theta),\min_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|=\Sigma_X(\tilde{\theta}),\Sigma_X(\theta,\tilde{\theta})Z_{\tilde{\theta}}(\theta)>0}H_Z(\tilde{\theta},\theta),G_Z(\tilde{\theta},\theta)\right\},$$

and

$$u_Z^2(\tilde{\theta}) = \min\left\{ \min_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|>\Sigma_X(\tilde{\theta})} K_Z(\tilde{\theta},\theta), \min_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|=\Sigma_X(\tilde{\theta}),\Sigma_X(\theta,\tilde{\theta})Z_{\tilde{\theta}}(\theta)>0} H_Z(\tilde{\theta},\theta) \right\},$$

$$\mathcal{Y}(\tilde{\theta},Z_{\tilde{\theta}}) = \bigcap_{\theta\in\Theta:|\Sigma_X(\theta,\tilde{\theta})|<\Sigma_X(\tilde{\theta})} \left[\ell_Z^1(\tilde{\theta}),u_Z^1(\tilde{\theta},\theta)\right] \cup \left[\ell_Z^2(\tilde{\theta},\theta),u_Z^2(\tilde{\theta})\right].$$

(ii) If, moreover, $|\Sigma_X(\theta,\tilde{\theta})| < \Sigma_X(\tilde{\theta})$ *for all* θ, $\tilde{\theta}\in\Theta$ *such that* $\theta\neq\tilde{\theta}$,

$$\mathcal{Y}(\tilde{\theta},Z_{\tilde{\theta}}) = \left(-\infty, \min_{\theta\in\Theta} G_Z(\tilde{\theta},\theta)\right] \cup \left[\max_{\theta\in\Theta} K_Z(\tilde{\theta},\theta), \infty\right).$$

Part (i) of the corollary covers the stylized example discussed in Section 2, and its generalization to cases with dependence and more than two strategies. The condition $|\Sigma_X(\theta,\tilde{\theta})| < \Sigma_X(\tilde{\theta})$ in part (ii) holds automatically when X is comprised of studentized statistics and the elements of X are not perfectly dependent.

So far this section has considered conditioning on $\hat{\theta}=\tilde{\theta}$, AKM also allows conditioning on another random variable $\hat{\gamma}=\gamma(X)$. Such conditioning can be used to incorporate the outcome of a pre-test or other data-driven selection in order to address pretest bias and coverage distortions. The form of $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$ will depend on the nature of $\hat{\gamma}$. Here we derive $\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z)$ in the case where $\hat{\gamma}$ encodes the outcome of the sup-Wald pretest, which is a natural pretest for tipping point and structural break applications.

Threshold Regression and Structural Break Estimation (continued) Suppose that we report estimates and confidence intervals for the change parameter $\delta_j(\hat{\theta})$ only if we reject the null hypothesis of no threshold, $H_0 : \delta(\theta) = 0$ for all $\theta \in \Theta$. Suppose, in particular, that we test H_0 with the sup-Wald test of D. Andrews (1993). Analogous results to those in Elliott and Müller (2014) show that the asymptotic version of such a test rejects asymptotically if and only if $\|X(\hat{\theta})\| > c$ for a critical value c that depends on Σ. $\triangle$

Let $\hat{\gamma} \in \{0,1\}$ be a dummy variable for rejection by the sup-Wald pretest, $\hat{\gamma} = 1\left\{\|X(\hat{\theta})\| > c\right\}$. We study inference conditional on $\hat{\theta}=\tilde{\theta}$ and $\hat{\gamma}=1$. We can express

$$\mathcal{Y}(\tilde{\theta},\tilde{\gamma},z) = \mathcal{Y}(\tilde{\theta},z) \cap \mathcal{Y}_\gamma(\tilde{\gamma},z),$$

with $\mathcal{Y}(\tilde{\theta},z)$ the conditioning set based on $\hat{\theta}$ alone (derived in Proposition 1), and $\mathcal{Y}_\gamma(\tilde{\gamma},z)$ the conditioning set based on $\hat{\gamma}$. The next result derives the form of $\mathcal{Y}_\gamma(1,z)$ for the

sup-Wald test, while $\mathcal{Y}_\gamma(0,z)=\mathcal{Y}_\gamma(1,z)^c$.

Proposition 2
Suppose $\hat{\gamma}=1\Big\{\|X(\hat{\theta})\|>c\Big\}$. *Define*

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

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

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

For

$$\bar{\mathcal{L}}(Z_{\tilde{\theta}})\equiv\frac{-\bar{B}_Z(\tilde{\theta})-\sqrt{D_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})},$$
$$\bar{\mathcal{U}}(Z_{\tilde{\theta}})\equiv\frac{-\bar{B}_Z(\tilde{\theta})+\sqrt{D_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})},$$
$$\bar{\mathcal{V}}(Z_{\tilde{\theta}})\equiv[1\{\bar{A}(\tilde{\theta})=0\}+1\{\bar{A}(\tilde{\theta})>0,D_Z(\tilde{\theta})<0\}]\bar{C}_Z(\tilde{\theta}),$$

if $\bar{\mathcal{V}}(Z_{\tilde{\theta}})\geq 0$ *then* $\mathcal{Y}_\gamma(1,Z_{\tilde{\theta}})=\big(\bar{\mathcal{L}}(Z_{\tilde{\theta}}),\bar{\mathcal{U}}(Z_{\tilde{\theta}})\big)^c$, *while* $\mathcal{Y}_\gamma(1,Z_{\tilde{\theta}})=\emptyset$ *otherwise.*

4 Practical Implementation and Uniform Asymptotic Validity

This section shows that the desirable finite-sample properties of the AKM estimators and confidence intervals in the normal model translate to asymptotic results in norm-maximization problems satisfying regularity conditions. In particular, we show that feasible implementations of the AKM procedures are uniformly asymptotically valid over classes of norm-maximization problems such that the mean vectors μ_X and μ_Y are asymptotically bounded. We begin by discussing our asymptotic setting and assumptions and relate our assumptions to the threshold regression and structural break examples. We then turn to asymptotic results for feasible versions of the AKM procedures.

4.1 Asymptotic Setting and Assumptions

In analogy with the normal model of Section 3, but dropping the assumption of finite-sample normality with known variance, assume we observe random vectors $\left(X_n(\theta)',Y_n(\theta)\right)'$ for $\theta\in\Theta$, where Θ is a finite set, $X_n(\theta)\in\mathbb{R}^{d_X}$, and $Y_n(\theta)\in\mathbb{R}$. In particular, for $\Theta=\left\{\theta_1,...,\theta_{|\Theta|}\right\}$, let $X_n=\left(X_n(\theta_1)',...,X_n\left(\theta_{|\Theta|}\right)'\right)'$ and $Y_n=\left(Y_n(\theta_1),...,Y_n\left(\theta_{|\Theta|}\right)\right)'$. We suppose that the data in the sample of size n are distributed according to $P\in\mathcal{P}_n$ for $\mathcal{P}_n$ a sample-size dependent class of distributions, and assume that with appropriate recentering, (X_n,Y_n) are jointly asymptotically normal uniformly over $P\in\mathcal{P}_n$.

Assumption 1
For BL_1 the class of Lipschitz functions that are bounded in absolute value by one and have Lipschitz constant bounded by one, and $\xi_P\sim N(0,\Sigma(P))$,

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

for some sequence of functions $\mu_{X,n}(\cdot)$ and $\mu_{Y,n}(\cdot)$.

Bounded Lipschitz distance metrizes convergence in distribution, so uniform convergence in bounded Lipschitz, as we assume here, is one formalization for uniform convergence in distribution. Intuitively speaking, this assumption requires that

$$\left(X_n'-\mu_{X,n}(P)',Y_n'-\mu_{Y,n}(P)'\right)'$$

be asymptotically $N(0,\Sigma(P))$ distributed, uniformly over $P\in\mathcal{P}_n$.

In many cases we can take $(\mu_{X,n}(P),\mu_{Y,n}(P))$ to be the mean of (X_n,Y_n) under P,

$$E_P\left[\begin{array}{c}X_n(\theta)\\Y_n(\theta)\end{array}\right]=\mu_n(\theta;P)=\left(\begin{array}{c}\mu_{X,n}(\theta;P)\\\mu_{Y,n}(\theta;P)\end{array}\right).$$

We do not impose this as an assumption, however, since the finite-sample mean may be poorly-behaved in some settings, including in structural break and tipping point applications, so it may be preferable to define $(\mu_{X,n}(P),\mu_{Y,n}(P))$ in some other application-specific way. For instance, see the next section for definitions in the tipping-point example.

We are interested in estimation and inference on $\mu_{Y,n}(\hat{\theta}_n;P)$ for the true but unknown

DGP P, in the norm-maximization problem where

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} \|X_n(\theta)\| + o_p(1),$$

and the $o_p(1)$ term is uniformly asymptotically negligible over $P \in \mathcal{P}_n$.

We next restrict the asymptotic variance $\Sigma(P)$.

Assumption 2

There exists a finite $\bar{\lambda} > 0$ such that for $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the minimum and maximum eigenvalues of a matrix A,

$$1/\bar{\lambda} \leq \lambda_{\min}(\Sigma_X(P)) \leq \lambda_{\max}(\Sigma_X(P)) \leq \bar{\lambda} \text{ for all } P \in \mathcal{P}_n$$

and

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

This assumption bounds the variance matrix $\Sigma_X(P)$ above and away from singularity, and likewise bounds the diagonal elements of $\Sigma_Y(P)$ above and away from zero. This ensures that the set of covariance matrices consistent with $P \in \mathcal{P}_n$ is a subset of a compact set, and that $\|X_n(\theta)\|$ has a unique maximum with probability tending to one.

Our estimators and confidence intervals depend not only on (X_n, Y_n), but also on an estimator $\widehat{\Sigma}_n$ of Σ. We assume that this estimator is uniformly consistent.

Assumption 3

$\widehat{\Sigma}_n$ is uniformly consistent for $\Sigma(P)$,

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

for all $\varepsilon > 0$.

Finally, we assume that $(\mu_{X,n}(P), \mu_{Y,n}(P))$ are asymptotically bounded.

Assumption 4

There exists a finite constant $C > 0$ such that

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_n} (\|\mu_{X,n}(P)\| + \|\mu_{Y,n}(P)\|) \leq C.$$

This assumption requires that $\|\mu_{X,n}(P)\|$ and $\|\mu_{Y,n}(P)\|$ be uniformly bounded over $\mathcal{P}_n$ by a constant that does not depend on the sample size. Given the scaling of (X_n,Y_n) in our threshold regression and structural break examples, this corresponds to the case with local parameter instability. More broadly, this condition implies that the norm-maximization problem remains non-trivial even asymptotically, in the sense that we do not have $Pr_P\{\hat{\theta}=\tilde{\theta}\}\to 1$ for any $\tilde{\theta}$. While it may be possible to relax this assumption, it holds in all settings we have encountered that give rise to the norm-maximization problem asymptotically.[12]

Threshold Regression and Structural Break Estimation (continued) As shown in Section 3, the asymptotically normal norm-maximization problem arises when we follow Elliott and Müller (2007, 2014) and Lee and Wang (2020) and model the degree of parameter instability as shrinking with the sample size at the $\sqrt{n}$ rate. The quantities (X_n,Y_n) in this example are detailed in Section 3 above, while we can take $(\mu_{X,n},\mu_{Y,n})$ to be their population analogs. In particular, let

$$\mu_{X,n}(\theta;P)=\sqrt{n}\begin{pmatrix} E_P[C_iC_i'1\{Q_i\le\theta\}]^{-\frac{1}{2}}E_P[C_i\eta_i1\{Q_i\le\theta\}] \\ E_P[C_iC_i'1\{Q_i>\theta\}]^{-\frac{1}{2}}E_P[C_i\eta_i1\{Q_i>\theta\}] \end{pmatrix}.$$

Calculations in Section A.1 of the Appendix show that we can write the population regression coefficient $\delta(\theta,P)$ imposing break point θ as $\delta(\theta,P)\equiv\mathcal{A}(\theta;P)^{-1}\mathcal{B}(\theta;P)$ for

$$\begin{aligned}\mathcal{A}(\theta;P)&=E_P[C_iC_i'1\{Q_i>\theta\}]-E_P[C_iC_i'1(\{Q_i>\theta\}]E_P[C_iC_i']^{-1}E_P[C_iC_i'1\{Q_i>\theta\}],\\ \mathcal{B}(\theta;P)&=E_P[C_iC_i'1\{Q_i>\theta\}g(Q_i)]-E_P[C_iC_i'1\{Q_i>\theta\}]E_P[C_iC_i']^{-1}E_P[C_iC_i'g(Q_i)],\end{aligned}$$

so we can define $\mu_{Y,n}(\theta;P)=\sqrt{n}\delta(\theta,P)$. Note that while $\mu_{X,n}$ and $\mu_{Y,n}$ correspond naturally to X_n and Y_n, respectively, in general $E_P[X_n]\neq\mu_{X,n}$ and $E_P[Y_n]\neq\mu_{Y,n}$.

In Section A.1 of the Appendix, we show that the elements of $\Sigma(\theta,\tilde{\theta};P)$ are functions of $\Sigma_C(\theta;P)$, $\Sigma_C(\tilde{\theta};P)$ and $E_P[G(\theta)G(\tilde{\theta})']$ so that we can construct an estimator $\widehat{\Sigma}_n$ by plugging in consistent estimators of these quantities. In particular, we can estimate $\Sigma_C(\theta;P)=E_P[C_iC_i'1\{Q_i\le\theta\}]$ by $\widehat{\Sigma}_C(\theta)=\frac{1}{n}\sum_{i=1}^n C_iC_i'1\{Q_i\le\theta\}$ and $E_P[G(\theta)G(\tilde{\theta})']=E_P[C_iC_i'1\{Q_i\le\theta\}]$ for $\theta\le\tilde{\theta}$ and iid data by $\widehat{\Sigma}_C(\theta)=\frac{1}{n}\sum_{i=1}^n C_iC_i'\hat{U}_i^21\{Q_i\le\theta\}$ with $\hat{U}_i=Y_i-C_i'(\hat{\beta}-\hat{\varphi}_n(Q_i))$ for consistent estimators $\hat{\beta}$ of β and $\hat{\varphi}_n(\cdot)$ of $\varphi_n(\cdot)$.[13]

[12]Note, moreover, that the proofs of uniform asymptotic validity for a related class of selective inference procedures in Tibshirani et al. (2018) rely on a similar condition, though the proofs of AKM for the level-maximization setting do not.

[13]For dependent data, $E_P[G(\theta)G(\tilde{\theta})']=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n E_P[C_iC_j'U_iU_j1(Q_i\le\theta)1(Q_j\le\tilde{\theta})]$ can be consistently estimated using standard long-run variance estimation techniques.

In this setting, Assumptions 1-3 follow from standard conditions. In particular, Assumption 1 requires that (X_n,Y_n) be uniformly asymptotically normal over $\mathcal{P}_n$, and will follow from uniform versions of (8) and (9), along with bounds on $\Sigma_C(\theta;P)$. Assumption 2 bounds the behavior of Σ, and will follow from suitable uniform moment bounds. Finally, Assumption 3 will again follow from uniform moment bounds and, in the structural break setting, limits on the degree of dependence in the data. We note that these assumptions are fully compatible with the absence of a changing coefficient, in which case $g(\cdot)=0$.

Assumption 4 warrants additional discussion. This assumption holds if we take $\mathcal{P}_n$ to correspond to any finite collection of local sequences of the sort studied by Elliott and Müller (2007, 2014) and Lee and Wang (2020). If we instead take the degree of parameter instability to be fixed, one can show that the threshold regression and structural break models reduce to level maximization, as studied by AKM, asymptotically. Intuitively, for μ_X large,

$$\|X(\theta)\|^2 \approx \|\mu_X(\theta)\|^2 + 2\mu_X(\theta)'(X(\theta)-\mu_X(\theta)),$$

so the squared norm $\|X(\theta)\|^2$ behaves like a normal random variable.

The issue here is similar to the difference in the asymptotic distribution of the Vuong (1989) test between the nested and non-nested cases. As this analogy suggests, it may be possible to develop asymptotic results for threshold regression and structural break models that, analogous to the results of Shi (2015) and Schennach and Wilhelm (2017) for the Vuong test, cover cases with both fixed and local parameter instability. We are unaware of such results for existing procedures in threshold regression and structural break literatures, however, and this point is far afield from our primary focus here. Hence, in this paper we follow Elliott and Müller (2007, 2014) and Lee and Wang (2020) by limiting attention to cases with local parameter instability and refer readers interested in fixed parameter instability to the level-maximization results discussed in AKM. $\triangle$

Note that the stylized example discussed in Section 2 can likewise be recast as level maximization when μ_X grows large. In particular, for trading strategies with absolute average returns well-separated from zero, we can consistently estimate the sign of the average return, and so convert the problem to level maximization by choosing strategy j to maximize $\text{sign}\{\mu_j\}X_j$, where $\text{sign}\{x\}$ takes value 1 if $x>0$ and value -1 if $x<0$.

4.2 Uniformity Results for Estimators and Confidence Intervals

We next prove uniform asymptotic validity for feasible versions of the AKM procedures. These feasible versions are defined as in the normal model in Section 3.2, save that we

replace $\hat{\theta}$ by $\hat{\theta}_n$, Y by Y_n, Σ by $\hat{\Sigma}_n$, and $Z_{\tilde{\theta}}$ by

$$Z_{\tilde{\theta},n} = X_n - \left(\hat{\Sigma}_{XY,n}(\cdot,\tilde{\theta})/\hat{\Sigma}_{Y,n}(\tilde{\theta})\right) Y_n(\tilde{\theta})$$

in all expressions.

Asymptotic uniformity results for some conditional inference procedures that, like our corrections, rely on truncated normal distributions have been previously established by Tibshirani et al. (2018). However, their results do not cover the norm-maximization problems studied in this paper. Moreover, they do not cover the hybrid inference procedures of AKM, which are new to the literature, nor do they provide results for quantile-unbiased estimation. Our proofs are based on subsequencing arguments as in D. Andrews et al. (2018), though due to the differences in our setting (our interest in conditional inference, and the fact that our target is random from an unconditional perspective), we cannot directly apply their results.

4.2.1 Asymptotic Validity of Conditional Procedures

We begin our analysis of uniform asymptotic validity by establishing results for the feasible asymptotically α-quantile-unbiased estimator $\hat{\mu}_{\alpha,n}$. Just as $\hat{\mu}_\alpha$ is quantile-unbiased in the normal model, $\hat{\mu}_{\alpha,n}$ is asymptotically quantile-unbiased both conditional on the event $\left\{\hat{\theta}_n = \tilde{\theta}\right\}$ and unconditonally.

Proposition 3
Under Assumptions 1-4,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \left| Pr_P\left\{\hat{\mu}_{\alpha,n} \geq \mu_{Y,n}\left(\hat{\theta}_n;P\right) | \hat{\theta}_n = \tilde{\theta}\right\} - \alpha \right| = 0, \tag{11}$$

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

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \left| Pr_P\left\{\hat{\mu}_{\alpha,n} \geq \mu_{Y,n}\left(\hat{\theta}_n;P\right)\right\} - \alpha \right| = 0. \tag{12}$$

Arguments as in the proof of Proposition 3 imply analogous results for additional conditioning variables $\hat{\gamma}_n$, such as $\hat{\gamma}_n = 1\left\{\|X_n(\hat{\theta}_n)\| > c\right\}$. This is also true for the other conditional results in this subsection and the next. For the sake of brevity, we do not pursue such extensions here.

Proposition 3 immediately implies that the one-sided confidence intervals $(-\infty,\hat{\mu}_{1-\alpha,n}]$ and $[\hat{\mu}_{\alpha,n},\infty)$ have uniformly correct asymptotic coverage. We also consider equal-tailed

intervals $CS_n = [\hat{\mu}_{\alpha/2,n}, \hat{\mu}_{1-\alpha/2,n}]$. The following corollary shows that CS_n has correct asymptotic coverage for $\mu_{Y,n}(\hat{\theta}_n;P)$, both conditionally and unconditionally.

Corollary 2

Under Assumptions 1-4,

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

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

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

Arguments along the same lines as in the proof of Corollary 2 also imply uniform asymptotic validity of intervals which weight the two tails differently, viz, $[\hat{\mu}_{\delta,n}, \hat{\mu}_{1-\alpha-\delta,n}]$ for $0<\delta<\alpha$.

4.2.2 Unconditional Validity Results

In this section, we turn to asymptotic validity for the unconditional procedures discussed in Section 3.2, namely the projection and Hybrid approaches. Let us denote the feasible level $1-\alpha$ projection interval by $CS^{\alpha}_{P,n}$. This interval has asymptotically correct unconditional coverage for $\mu_{Y,n}(\hat{\theta}_n;P)$ uniformly over the class of DGPs $P\in\mathcal{P}_n$.

Proposition 4

Under Assumptions 1-4,

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

Next, consider feasible hybrid estimators $\hat{\mu}^H_{\alpha,n}$. While these estimators are not asymptoptically quantile-unbiased, their asymptotic quantile bias (as measured by the exceedence probability) is controlled.

Proposition 5

Under Assumptions 1-4,

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

for all $\tilde{\theta} \in \Theta$. *Moreover*

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

We can again use the estimators $\hat{\mu}^H_{\alpha,n}$ to form equal-tailed confidence intervals. As in Section 3.2, however, we need to adjust the quantiles we consider to account for the fact that CS^{β}_P may not cover $\mu_{Y,n}(\hat{\theta}_n;P)$. Hence, we define the feasible level $1-\alpha$ equal-tailed hybrid interval as

$$CS^H_n = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{1-\beta},n}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{1-\beta},n} \right].$$

Corollary 3

Under Assumptions 1-4,

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

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

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

and

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

5 Split-Sample Inference

We next briefly discuss feasible split-sample estimators and confidence intervals which dominate conventional split-sample inference as used in e.g. Card et al. (2008). These dominating procedures were introduced in a general asymptotic setting by AKM, and we refer the interested reader to AKM for theoretical details on dominance in the normal model. Asymptotic validity of these procedures under extensions of Assumptions 1-4 follows from arguments along the same lines as the proofs of the results in the last section, so we omit formal statements and proofs for brevity.

The problem we consider here is quite similar to that studied in Section 4.2, with the key difference being that only part of the sample is used to select the norm-maximizing value. In settings with iid data, this can be formalized as follows: for $\tau \in (0,1)$, assume we observe random vectors

$$(X^{1\prime}_n, Y^{1\prime}_n)' \equiv \tau^{-1/2}(X'_{[\tau n]}, Y'_{[\tau n]})'$$

and

$$(X_n^{2\prime},Y_n^{2\prime})' \equiv (1-\tau)^{-1}[(X_n',Y_n')' - \sqrt{\tau}(X_{[\tau n]+1}',Y_{[\tau n]+1}')'],$$

where $(X_n',Y_n')'$ is as defined in Section 4.2, and $[\tau n]$ is the closest whole number to τn. Intuitively, $(X_n^{1\prime},Y_n^{1\prime})'$ is the analog of $(X_n',Y_n')'$ formed from the first $[\tau n]$ observations, while $(X_n^{2\prime},Y_n^{2\prime})'$ is the analog of $(X_n',Y_n')'$ formed from the rest of the sample. Split-sample approaches then take

$$\hat{\theta}_n^1 = \underset{\theta\in\Theta}{\text{argmax}} \|X_n^1(\theta)\| + o_p(1)$$

and consider inference on $\mu_{Y,n}(\hat{\theta}_n^1;P)$. The conventional split-sample estimator and confidence interval for $\mu_{Y,n}(\hat{\theta}_n^1;P)$ are $Y_n^2(\hat{\theta}_n^1)$ and

$$CS_{SS,n} = \left[Y_n^2(\hat{\theta}_n^1) - \sqrt{\frac{1}{1-\tau}\hat{\Sigma}_{Y,n}(\hat{\theta}_n^1)}z_{1-\alpha/2}, Y_n^2(\hat{\theta}_n^1) + \sqrt{\frac{1}{1-\tau}\hat{\Sigma}_{Y,n}(\hat{\theta}_n^1)}z_{1-\alpha/2}\right],$$

where $\hat{\Sigma}_n$ is as defined in Section 4.2. Since $(X_n^{1\prime},Y_n^{1\prime})'$ and $(X_n^{2\prime},Y_n^{2\prime})'$ are based on different observations, they are independent by construction, and it is straightforward to show that $Y_n^2(\hat{\theta}_n^1)$ is asymptotically unbiased and $CS_{SS,n}$ has correct asymptotic coverage for $\mu_{Y,n}(\hat{\theta}_n^1;P)$ both conditional on the realization of $\hat{\theta}_n^1$ and unconditionally.

Direct sample splitting can also be applied in some stationary time-series applications, where asymptotic independence of $(X_n^{1\prime},Y_n^{1\prime})'$ and $(X_n^{2\prime},Y_n^{2\prime})'$ will follow from weak dependence assumptions. In structural break applications, however, we are fundamentally interested in non-stationarity, and splitting the sample is not a viable approach. In such cases, one can still employ an asymptotic analog of sample splitting. Specifically, for ξ a standard normal random vector independent of the data we can take

$$(X_n^{1\prime},Y_n^{1\prime})' = (X_n',Y_n')' - \sqrt{\frac{1-\tau}{\tau}}\widehat{\Sigma}_n^{\frac{1}{2}}\xi$$

$$(X_n^{2\prime},Y_n^{2\prime})' = (X_n',Y_n')' + \sqrt{\frac{\tau}{1-\tau}}\widehat{\Sigma}_n^{\frac{1}{2}}\xi,$$

and define $\hat{\theta}_n^1$ as above. Under Assumptions 1-4, $(X_n^{1\prime},Y_n^{1\prime})'$ and $(X_n^{2\prime},Y_n^{2\prime})'$ will be asymptotically independent, and asymptotic validity of conventional split-sample inference again follows.

Taking τ as given, we next describe how to construct estimators and confidence intervals for $\mu_{Y,n}(\hat{\theta}_n^1;P)$ that are conditionally and unconditionally valid, and dominate $Y_n^2(\hat{\theta}_n^1)$ and

$CS_{SS,n}$ in terms of concentration around $\mu_{Y,n}(\hat{\theta}_n^1;P)$ and confidence interval length. Let $F_{SS}^A(\cdot;\mu_Y(\tilde{\theta}^1),\tilde{\theta}^1,z^1)$ denote the distribution function of the random variable

$$\left(\xi^1+\frac{1-\tau}{\tau}\xi^2\right)\Big|\xi^1\in\mathcal{Y}(\tilde{\theta}^1,z^1)$$

where

$$\xi^1\sim\mathcal{N}\left(\mu_Y(\tilde{\theta}^1),\frac{1}{\tau}\hat{\Sigma}_{Y,n}(\tilde{\theta}^1)\right) \text{ and } \xi^2\sim\mathcal{N}\left(\mu_Y(\tilde{\theta}^1),\frac{1}{1-\tau}\hat{\Sigma}_{Y,n}(\tilde{\theta}^1)\right)$$

are independent, and $\mathcal{Y}(\tilde{\theta},z)$ is defined as in Proposition 1. Expressing $\mathcal{Y}(\tilde{\theta}^1,z^1)$ as a finite union of disjoint intervals using De Morgan's laws, $\mathcal{Y}(\tilde{\theta}^1,z^1)=\cup_{k=1}^K[\ell_k(z^1),u_k(z^1)]$, we obtain the following expression for $F_{SS}^A(y;\mu_Y(\tilde{\theta}^1),\tilde{\theta}^1,z^1)$:[14]

$$\frac{E\left[\Phi\left((y-\xi^1-\frac{1-\tau}{\tau}\mu_Y(\tilde{\theta}^1))/\sqrt{\frac{1-\tau}{\tau^2}\hat{\Sigma}_{Y,n}(\tilde{\theta}^1)}\right)1\left(\xi^1\in\bigcup_{k=1}^K[\ell_k(z^1),u_k(z^1)]\right)\right]}{\tau\sum_{k=1}^K\left(\Phi\left((u_k(z^1)-\mu_Y(\tilde{\theta}^1))/\sqrt{\tau^{-1}\hat{\Sigma}_{Y,n}(\tilde{\theta}^1)}\right)-\Phi\left((\ell_k(z^1)-\mu_Y(\tilde{\theta}^1))/\sqrt{\tau^{-1}\hat{\Sigma}_{Y,n}(\tilde{\theta}^1)}\right)\right)},$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable and the expectation is taken with respect to ξ^1.

The α-quantile asymptotically unbiased split-sample estimator $\hat{\mu}_{SS,\alpha,n}^A$ is the unique solution to

$$F_{SS}^A\left(Y_n^1(\hat{\theta}_n^1)+\frac{1-\tau}{\tau}Y_n^2(\hat{\theta}_n^1);\hat{\mu}_{SS,\alpha,n}^A,\tilde{\theta}^1,Z_{\tilde{\theta}^1,n}^1\right)=1-\alpha,$$

where $\tilde{\theta}^1=\hat{\theta}_n^1$ and

$$Z_{\tilde{\theta}^1,n}^1=X_n^1-\left(\hat{\Sigma}_{XY,n}(\cdot,\tilde{\theta}^1)/\hat{\Sigma}_{Y,n}(\tilde{\theta}^1)\right)Y_n^1(\tilde{\theta}^1).$$

The new dominating equal-tailed split-sample confidence interval is

$$CS_{SS,n}^A=[\hat{\mu}_{SS,\alpha/2,n}^A,\hat{\mu}_{SS,1-\alpha/2,n}^A].$$

The expression for $F_{SS}^A(y;\mu_Y(\tilde{\theta}^1),\tilde{\theta}^1,z^1)$ above makes the computation of $\hat{\mu}_{SS,\alpha,n}^A$ and $CS_{SS,n}^A$ very straightforward in practice.

6 Monte Carlo Simulations for the Threshold Regression Model

In this section, we conduct a simulation study based on the tipping point model of Card et al. (2008), a leading application of the threshold regression model discussed throughout

[14] See AKM for the full derivation.

this paper as a running example. Card et al. (2008) study the evolution of neighborhood composition as a function of minority population share. For Y_i the normalized change in the white population of census tract i between 1980 and 1990, C_i a vector of controls, and Q_i the minority share in 1980, Card et al. (2008) consider the specification

$$Y_i = \beta + C_i'\alpha + \delta 1\{Q_i > \theta\} + U_i.$$

This specification 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. We consider data from Chicago and Los Angeles with $n = 1{,}820$ and $n = 2{,}035$ observations, respectively, estimating the model separately in each city.[15]

Results in Hansen (2000) and Lee and Wang (2020) show that if we model the degree of parameter instability as on the same order as sampling uncertainty, this threshold regression model satisfies the high-level conditions (8)–(9) we introduced in Section 3.1. Hence, we can apply our results for the norm-maximization problem to the present setting. Specifically, we define X_n as discussed in Section 3.1 and $\hat{\theta}_n$ is again asymptotically equivalent to the solution to a norm-maximization problem $\text{argmax}_{\theta\in\Theta}\|X(\theta)\|$.[16] We define $Y_n(\theta) = \sqrt{n}\hat{\delta}(\theta)$ to be proportional to the estimated change coefficient imposing tipping point θ, so we again consider the problem of inference on the (scaled) change coefficient while acknowledging randomness in the estimated threshold.

Our simulations draw random vectors (X,Y) from the limiting normal model (10). This model depends on the function Σ_C and the covariance function of G in (9) which we (consistently) estimate from the Card et al. (2008) data. It also depends on the function $\Sigma_{Cg}(\cdot)$. Since this is not consistently estimable, we consider three specifications. Specification (i) assumes there is no coefficient change, corresponding to $\delta = 0$. Specification (ii) assumes that there is a single large change, setting $\delta = -100\%$ and taking the true threshold to equal the estimate in the Card et al. (2008) data. Finally, specification (iii) calibrates $\Sigma_{Cg}(\cdot)$ to the data, corresponding to the analog of model (6) where the intercept term in the regression may depend arbitrarily upon a neighborhood's minority share. This specification implies

[15] We focus on these cities since Card et al. (2008) note that their tipping point estimation method appears more appropriate for larger cities.

[16] While Card et al. (2008) optimize over all possible tipping points between 5% and 60%, consistent with our theoretical results we limit attention to a finite set of thresholds. In particular, we consider 100 evenly-spaced quantiles of the minority share, and then further restrict attention to thresholds between 5% and 60%. We also tried several other discretization schemes and found very similar results in all cases.

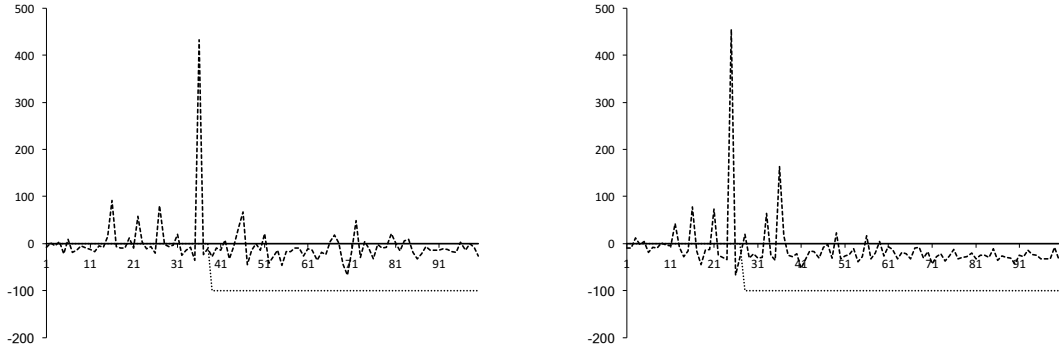

Figure 1: Specifications for $\varphi_n(\cdot)$ used in the simulations for Chicago (left) and LA (right) data. The horizontal axis corresponds to the empirical percentile of Q_i. The solid line corresponds to DGP (i), the dotted line corresponds to DGP (ii) and the dashed line corresponds to DGP (iii).

that the break model is misspecified but as discussed above, our approach remains applicable in this case, unlike the method of Lee and Wang (2020). 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.

Figure 1 above plots the function $\varphi_n(\cdot)$ for specifications (i)–(iii) corresponding to how the data are generated in the simulations, for both the Chicago and Los Angeles data.

We begin by considering the problem of inference on $\mu_Y(\hat{\theta})$. We focus on unconditional performance, as we are unaware of alternative procedures with conditional performance guarantees in this setting. All reported results are based on 10^4 simulation draws. Table 1 reports the unconditional coverage for the confidence intervals CS, CS^H, and CS^α_P, along with the conventional confidence interval $CS_N = [Y(\hat{\theta}) - \sqrt{\Sigma(\hat{\theta})} z_{1-\alpha/2}, Y(\hat{\theta}) + \sqrt{\Sigma(\hat{\theta})} z_{1-\alpha/2}]$. See the Supplemental Appendix for the corresponding results for the unbiased confidence intervals. In all cases, we consider confidence intervals with nominal coverage 95%, $\alpha = 0.05$. For hybrid confidence intervals, we set $\beta = \alpha/10$. From Table 1 we see that all confidence intervals other than CS_N have correct coverage, the projection confidence interval CS^α_P often over-covers, the conditional confidence interval CS has exact coverage and the hybrid confidence interval CS^H exhibits minimal over-coverage. In this application, the conventional confidence interval CS_N severely under-covers for some simulation designs.

Table 2 compares the lengths of our confidence intervals to that of CS^α_P. Since projection confidence intervals have been previously proposed in the literature and their length is proportional to the asymptotic standard error $\sqrt{\Sigma_Y(\hat{\theta})}$ of the estimated change coefficient,

Table 1: Unconditional Coverage Probability

DGP	CS	CS^H	CS_P^α	CS_N
Chicago Data Calibration				
(i)	0.948	0.949	0.95	0.750
(ii)	0.951	0.956	0.994	0.951
(iii)	0.947	0.951	0.990	0.934
Los Angeles Data Calibration				
(i)	0.949	0.949	0.95	0.615
(ii)	0.952	0.956	0.996	0.952
(iii)	0.951	0.955	0.996	0.95

This table reports the unconditional coverage probability of $\mu_Y(\hat{\theta})$ for the conditionally valid confidence interval (CS), the hybrid confidence interval (CS^H), the projection confidence interval (CS_P^α) and the conventional confidence interval (CS_N), all evaluated at the nominal coverage level of 95%. In the Chicago (Los Angeles) data calibrations, the covariance matrix Σ is set equal to a consistent estimate from the Chicago (Los Angeles) Card et al. (2008) data. The column "DGP" refers to the specification of the nuisance function $\Sigma_{Cg}(\cdot)$, which along with other parameters, determines the value of the mean vector μ (see Appendix A.1 for details). The function $\Sigma_{Cg}(\cdot)$ is set equal to the value it takes when there is no coefficient change in DGP (i), the value it takes when there is a single large coefficient change in DGP (ii) and its data-calibrated value in DGP (iii). For DGP (ii) the true threshold location is set to equal the estimate from the Card et al. (2008) data. All other parameters that determine μ are set equal to consistent estimates from the Card et al. (2008) data.

it provides a natural benchmark for comparison of our new confidence intervals. For each confidence interval we report both median length relative to CS_P^α and the frequency with which the confidence interval is longer than CS_P^α. Here we see that the conditional confidence interval can be relatively long, while the hybrid confidence interval provides marked performance improvements across the specifications considered. The benefits of the hybrid confidence interval can become even more pronounced at higher length quantiles. See Section C of the Supplemental Appendix. Remarkably, the hybrid confidence interval is not longer than CS_P^α in any simulation draw across all specifications examined. The overall message is that hybrid confidence interval possesses a clear advantage for unconditional inference and we recommend this approach for settings where unconditional coverage is desired.

Finally, we compare the conventional point estimator $Y(\hat{\theta})$ with $\hat{\mu}_{\frac{1}{2}}$ and $\hat{\mu}^H_{\frac{1}{2}}$. The initial columns of Table 3 report median bias measured both as the deviation of the exceedance probability from $\frac{1}{2}$ and as the median studentized estimation error. We 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, in specification (i) the conventional estimator $Y(\hat{\theta})$ has substantial median bias as measured by the median studentized estimation error, though very little as measured

Table 2: Length of Confidence Sets Relative to CS_P^α in Tipping Point Simulations

	Median Length Relative to CS_P^α		Probability Longer than CS_P^α	
DGP	CS	CS^H	CS	CS^H
		Chicago Data Calibration		
(i)	1.33	0.94	0.83	0
(ii)	0.72	0.74	0	0
(iii)	0.82	0.82	0.35	0
		Los Angeles Data Calibration		
(i)	1.26	0.86	0.58	0
(ii)	0.68	0.69	0	0
(iii)	0.68	0.70	0.15	0

This table reports the median length of the conditionally valid confidence interval (CS) and the hybrid confidence interval (CS^H), divided by the median length of the projection confidence interval (CS_P^α), as well as the frequency with which CS and CS^H is longer than CS_P^α. In the Chicago (Los Angeles) data calibrations, the covariance matrix Σ is set equal to a consistent estimate from the Chicago (Los Angeles) Card et al. (2008) data. The column "DGP" refers to the specification of the nuisance function $\Sigma_{Cg}(\cdot)$, which along with other parameters, determines the value of the mean vector μ (see Appendix A.1 for details). The function $\Sigma_{Cg}(\cdot)$ is set equal to the value it takes when there is no coefficient change in DGP (i), the value it takes when there is a single large coefficient change in DGP (ii) and its data-calibrated value in DGP (iii). For DGP (ii) the true threshold location is set to equal the estimate from the Card et al. (2008) data. All other parameters that determine μ are set equal to consistent estimates from the Card et al. (2008) data.

by the exceedance probability. This latter feature reflects the fact that the density of $Y(\hat{\theta})-\mu_Y(\hat{\theta})$ is bimodal with very little mass near zero in this specification.

Turning to median absolute studentized error, we see that all estimators perform similarly when the series has a single large break. By contrast, the median-unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ performs better than the conventional estimator $Y(\hat{\theta})$ in specification (i) (no break) but performs worse in specification (iii). The hybrid estimator $\hat{\mu}^H_{\frac{1}{2}}$ is weakly better than the unbiased estimator in all cases, with perfomance gains in case (i) and equal performance in the other two cases. The performance gains are again more pronounced if one considers higher quantiles of the absolute error distribution, as reported in Section C of the Supplemental Appendix.

6.1 Split-Sample Procedures

We have so far focused on inference on $\mu_Y(\hat{\theta})$ and compared the performance of our conditional and hybrid procedures to the projection confidence interval CS_P^α and conventional estimator $Y(\hat{\theta})$. However Card et al. (2008) instead adopt a sample splitting approach, using two thirds of the data to select the break date and one third of the data for inference. In this section we compare the performance of this conventional split-sample procedure

Table 3: Bias and Median Absolute Error in Tipping Point Simulations

	$Pr_\mu\left\{\hat{\mu} > \mu_Y(\hat{\theta})\right\} - \frac{1}{2}$			$Med_\mu\left(\frac{\hat{\mu}-\mu_Y(\hat{\theta})}{\sqrt{\Sigma_Y(\hat{\theta})}}\right)$			$Med_\mu\left(\left\lvert\frac{\hat{\mu}-\mu_Y(\hat{\theta})}{\sqrt{\Sigma_Y(\hat{\theta})}}\right\rvert\right)$		
DGP	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$Y(\hat{\theta})$	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$Y(\hat{\theta})$	$\hat{\mu}_{\frac{1}{2}}$	$\hat{\mu}^H_{\frac{1}{2}}$	$Y(\hat{\theta})$
Chicago Data Calibration									
(i)	0	0	0.01	-0.01	0.01	0.64	1.51	1.38	1.52
(ii)	-0.01	-0.01	-0.01	-0.03	-0.03	-0.03	0.66	0.66	0.66
(iii)	-0.01	-0.01	-0.15	-0.03	-0.03	-0.37	0.83	0.83	0.71
Los Angeles Data Calibration									
(i)	0	0	0	0	0	-0.8	1.38	1.29	1.80
(ii)	0	0	0	0.01	0.01	0.01	0.67	0.67	0.67
(iii)	0	0	0.006	0	-0.01	-.016	0.74	0.74	0.68

This table reports the deviation of the probability that an estimator exceeds $\mu_Y(\hat{\theta})$ from 1/2, the median studentized estimation error, and the median studentized absolute estimation error for the conditionally median-unbiased estimator ($\hat{\mu}_{\frac{1}{2}}$), the hybrid estimator ($\hat{\mu}^H_{\frac{1}{2}}$) and the conventional estimator ($Y(\hat{\theta})$). In the Chicago (Los Angeles) data calibrations, the covariance matrix Σ is set equal to a consistent estimate from the Chicago (Los Angeles) Card et al. (2008) data. The column "DGP" refers to the specification of the nuisance function $\Sigma_{Cg}(\cdot)$, which along with other parameters, determines the value of the mean vector μ (see Appendix A.1 for details). The function $\Sigma_{Cg}(\cdot)$ is set equal to the value it takes when there is no coefficient change in DGP (i), the value it takes when there is a single large coefficient change in DGP (ii) and its data-calibrated value in DGP (iii). For DGP (ii) the true threshold location is set to equal the estimate from the Card et al. (2008) data. All other parameters that determine μ are set equal to consistent estimates from the Card et al. (2008) data.

to that of (asymptotic versions of) the dominating split-sample alternative discussed in Section 5. We consider the same calibrations to the Card et al. (2008) data as above and choose the sample split as in Card et al. (2008).

Table 4 compares asymptotic versions of the conventional split-sample confidence interval CS_{SS} and estimator $Y^2(\hat{\theta}^1)$ used by Card et al. (2008) to the asymptotic versions of our (equal-tailed) alternative split-sample confidence interval C^A_{SS} and median-unbiased estimator $\hat{\mu}^A_{\frac{1}{2},SS}$, where we drop the n subscript in the table to emphasize that we consider the asymptotic problem. These results clearly reflect the dominance of the alternative split-sample procedures, with substantial performance improvements for both confidence intervals and estimators across all calibrations. These improvements are largest in the true break case (ii), but are nearly as large in the data-calibrated case (iii). Section C of the Supplemental Appendix provides ratios of the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} quantiles of the lengths of CS^A_{SS} relative to the those of CS_{SS} as well as the quantiles of $\left|\hat{\mu}-\mu_Y(\hat{\theta}^1)\right|/\sqrt{\Sigma_Y(\hat{\theta}^1)}$ for $\hat{\mu}=\hat{\mu}^A_{\frac{1}{2},SS}$ and $\hat{\mu}=Y^2(\hat{\theta}^1)$. There, our new split-sample proce-

dures can be seen to dominate the conventional ones across all quantiles and simulation designs considered, often by very wide margins.

Table 4: Performance Measures of Split-Sample Procedures

DGP	Median Length Relative to CS_{SS} CS^A_{SS}	$Med_\mu\left(\frac{\lvert\hat{\mu}-\mu_Y(\hat{\theta}^1)\rvert}{\sqrt{\Sigma_Y(\hat{\theta}^1)}}\right)$ $\hat{\mu}^A_{\frac{1}{2},SS}$	$Y^2(\hat{\theta}^1)$
Chicago Data Calibration			
(i)	0.83	0.57	0.67
(ii)	0.58	0.38	0.66
(iii)	0.64	0.44	0.67
Los Angeles Data Calibration			
(i)	0.78	0.55	0.69
(ii)	0.58	0.39	0.67
(iii)	0.59	0.42	0.68

This table reports the median length of the alternative split-sample confidence interval (CS^A_{SS}), divided by the median length of the conventional split-sample confidence interval (CS_{SS}), and the median studentized absolute estimation error of the median-unbiased alternative split-sample estimator ($\hat{\mu}^A_{\frac{1}{2},SS}$) and of the conventional split-sample estimator ($Y^2(\hat{\theta}^1)$). In the Chicago (Los Angeles) data calibrations, the covariance matrix Σ is set equal to a consistent estimate from the Chicago (Los Angeles) Card et al. (2008) data. The column "DGP" refers to the specification of the nuisance function $\Sigma_{Cg}(\cdot)$, which along with other parameters, determines the value of the mean vector μ (see Appendix A.1 for details). The function $\Sigma_{Cg}(\cdot)$ is set equal to the value it takes when there is no coefficient change in DGP (i), the value it takes when there is a single large coefficient change in DGP (ii) and its data-calibrated value in DGP (iii). For DGP (ii) the true threshold location is set to equal the estimate from the Card et al. (2008) data. All other parameters that determine μ are set equal to consistent estimates from the Card et al. (2008) data.

References

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

Andrews, D. W. K., Cheng, X., and Guggenberger, P. (2018). Generic results for establishing the asymptotic size of confidence sets and tests. Forthcoming in *Journal of Econometrics*.

Andrews, I., Kitagawa, T., and McCloskey, A. (2019). Inference on winners. unpublished manuscript.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79:551–563.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78.

Banerjee, M. and McKeague, I. W. (2007). Confidence sets for split points in decision trees. *Ann. Statist.*, 35(2):543–574.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.*, 30(4):927–961.

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

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*.

Hansen, B. E. (1997). Inference in TAR models. *Studies in Nonlinear Dynamics and Econometrics*, 2:1–14.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68:575–603.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business & Economic Statistics*, 35(2):228–240.

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

Hyun, S., Lin, K., G'Sell, M., and Tibshirani, R. J. (2018). Post-selection inference for changepoint detection algorithms with application to copy number variation data. Unpublished Manuscript.

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

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.

Lee, Y. and Wang, Y. (2020). Inference in threshold models. Unpublished Manuscript.

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

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

Schennach, S. M. and Wilhelm, D. (2017). A simple parametric model selection test. *Journal of the American Statistical Association*, 112(520):1663–1674.

Shi, X. (2015). A nondegenerate vuong test. *Quantitative Economics*, 6:85–121.

Song, R., Banerjee, M., and Kosorok, M. R. (2016). Asymptotics for change-point models under varying degrees of mis-specification. *Ann. Statist.*, 44(1):153–182.

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

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

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

A Appendix: Proofs for Results in Section 3

Proof of Proposition 1 Note the following equivalence of events:

$$\begin{aligned}\{\hat{\theta}=\tilde{\theta}\} &= \left\{\sum_{j=1}^{d_X} X_j(\tilde{\theta})^2 \geq \sum_{j=1}^{d_X} X_j(\theta)^2 \ \forall \theta\in\Theta\right\} \\ &= \left\{\sum_{j=1}^{d_X}\left[Z_{\tilde{\theta},j}(\tilde{\theta})+\Sigma_{XY,j}(\tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1}Y(\tilde{\theta})\right]^2 \right. \\ &\qquad \left. \geq \sum_{j=1}^{d_X}\left[Z_{\tilde{\theta},j}(\theta)+\Sigma_{XY,j}(\theta,\tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1}Y(\tilde{\theta})\right]^2 \ \forall\theta\in\Theta\right\} \\ &= \left\{A(\tilde{\theta},\theta)Y(\tilde{\theta})^2+B_Z(\tilde{\theta},\theta)Y(\tilde{\theta})+C_Z(\tilde{\theta},\theta)\geq 0 \ \forall\theta\in\Theta\right\}, \end{aligned} \tag{13}$$

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

By the quadratic formula, (13) is equivalent to the event

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

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

for $D_Z(\tilde{\theta},\theta)$, $G_Z(\tilde{\theta},\theta)$, $H_Z(\tilde{\theta},\theta)$, $K_Z(\tilde{\theta},\theta)$, $\ell_Z^1(\tilde{\theta})$, $\ell_Z^2(\tilde{\theta},\theta)$, $u_Z^1(\tilde{\theta},\theta)$, $u_Z^2(\tilde{\theta})$, and $\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}})$ again defined in the statement of the proposition. The result follows immediately. □

Proof of Corollary 1 (i) The result follows directly from Proposition 1 after specializing the problem of AKM to the case for which $X = Y$ and $d_X = 1$. More specifically, in the notation of Proposition 1, $A(\tilde{\theta},\theta) = 1 - \Sigma_X(\theta,\tilde{\theta})^2 / \Sigma_X(\tilde{\theta})^2$ so that $A(\tilde{\theta},\theta) < (=) 0$ if and only if $|\Sigma_X(\theta,\tilde{\theta})| > (=) \Sigma_X(\tilde{\theta})$, $B_Z(\tilde{\theta},\theta) = -2 Z_{\tilde{\theta}}(\theta) \Sigma_X(\theta,\tilde{\theta}) / \Sigma_X(\tilde{\theta})$ so that $B_Z(\tilde{\theta},\theta) > 0$ if and only if $\Sigma_X(\theta,\tilde{\theta}) Z_{\tilde{\theta}}(\theta) < 0$ and $D_Z(\tilde{\theta},\theta) = 4 Z_{\tilde{\theta}}(\theta)^2$ so that $D_Z(\tilde{\theta},\theta) \geq 0$ holds everywhere. Moreover, $\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) \geq 0$ vacuously holds everywhere since $C_Z(\tilde{\theta},\theta) = -Z_{\tilde{\theta}}(\theta)^2$ so that $\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) < 0$ would imply that $Z_{\tilde{\theta}}(\theta) \neq 0$ for some $\theta \in \Theta$ for which both $|\Sigma_X(\theta,\tilde{\theta})| = \Sigma_X(\tilde{\theta})$ and $\Sigma_X(\theta,\tilde{\theta}) Z_{\tilde{\theta}}(\theta) = 0$. This is impossible given the full rank assumption on Σ_X, which implies $\Sigma_X(\tilde{\theta}) > 0$.

(ii) Specializing the result in part (i), we have

$$\{ X(\tilde{\theta}) : \hat{\theta} = \tilde{\theta} \} = \bigcap_{\theta \in \Theta} \left(-\infty, G_Z(\tilde{\theta},\theta) \right] \cup \left[K_Z(\tilde{\theta},\theta), \infty \right).$$

But note that

$$G_Z(\tilde{\theta},\theta) = \begin{cases} \frac{-\Sigma_X(\tilde{\theta}) |Z_{\tilde{\theta}}(\theta)|}{\Sigma_X(\tilde{\theta}) + \Sigma_X(\theta,\tilde{\theta})} & \text{if } Z_{\tilde{\theta}}(\theta) \geq 0 \\ \frac{-\Sigma_X(\tilde{\theta}) |Z_{\tilde{\theta}}(\theta)|}{\Sigma_X(\tilde{\theta}) - \Sigma_X(\theta,\tilde{\theta})} & \text{if } Z_{\tilde{\theta}}(\theta) < 0 \end{cases}$$

and

$$K_Z(\tilde{\theta},\theta) = \begin{cases} \frac{\Sigma_X(\tilde{\theta}) |Z_{\tilde{\theta}}(\theta)|}{\Sigma_X(\tilde{\theta}) - \Sigma_X(\theta,\tilde{\theta})} & \text{if } Z_{\tilde{\theta}}(\theta) \geq 0 \\ \frac{\Sigma_X(\tilde{\theta}) |Z_{\tilde{\theta}}(\theta)|}{\Sigma_X(\tilde{\theta}) + \Sigma_X(\theta,\tilde{\theta})} & \text{if } Z_{\tilde{\theta}}(\theta) < 0, \end{cases}$$

which implies $G_Z(\tilde{\theta},\theta) \leq 0$ and $K_Z(\tilde{\theta},\theta) \geq 0$ for all $\tilde{\theta},\theta \in \Theta$ and thus the result in part (ii). □

Proof of Proposition 2 Arguments as in the proof of Proposition 1 show that

$$\{ \|X(\tilde{\theta})\|^2 \geq c \} = \left\{ Y(\tilde{\theta}) \leq \frac{-\bar{B}_Z(\tilde{\theta}) - \sqrt{D_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})} \text{ or } Y(\tilde{\theta}) \geq \frac{-\bar{B}_Z(\tilde{\theta}) + \sqrt{D_Z(\tilde{\theta})}}{2\bar{A}(\tilde{\theta})}, D_Z(\tilde{\theta}) \geq 0 \right\}$$
$$\cap \{ \bar{C}_Z(\tilde{\theta}) \geq 0, D_Z(\tilde{\theta}) < 0 \}$$

if $\bar{A}(\tilde{\theta}) > 0$ and $\{ \|X(\tilde{\theta})\|^2 \geq c \} = \{ \bar{C}_Z(\tilde{\theta}) \geq 0 \}$ if $\bar{A}(\tilde{\theta}) = 0$, since $\bar{A}(\tilde{\theta}) \geq 0$ by definition. Then we can immediately see that if $\bar{\mathcal{V}}(Z_{\tilde{\theta}}) \geq 0$ then $\mathcal{Y}_\gamma(1, Z_{\tilde{\theta}}) = \left(\bar{\mathcal{L}}(Z_{\tilde{\theta}}), \bar{\mathcal{U}}(Z_{\tilde{\theta}}) \right)^c$, while

$\mathcal{Y}_\gamma(1,Z_{\tilde{\theta}})=\emptyset$ otherwise. □

A.1 Threshold Regression Limit Experiment Details

This section provides additional results to supplement our discussion of the threshold regression example in the text.

We begin by establishing the weak convergence (10). To do so, we show uniform convergence over any compact set $\tilde{\Theta}$ in the interior of the support of Q_i, which implies uniform convergence over Θ. Note, in particular, that under (8) and (9) the continuous mapping theorem implies that

$$X_n(\theta) \Rightarrow X(\theta)$$
$$= \begin{pmatrix} \Sigma_C(\theta)^{-1/2}\Sigma_{Cg}(\theta) \\ (\Sigma_C(\infty)-\Sigma_C(\theta))^{-1/2}(\Sigma_{Cg}(\infty)-\Sigma_{Cg}(\theta)) \end{pmatrix} + \begin{pmatrix} \Sigma_C(\theta)^{-1/2}G(\theta) \\ (\Sigma_C(\infty)-\Sigma_C(\theta))^{-1/2}(G(\infty)-G(\theta)) \end{pmatrix} \tag{14}$$

uniformly on $\tilde{\Theta}$, where we use the following slight abuse of notation:

$$\frac{1}{n}\sum_{i=1}^n C_iC_i' \to_p \Sigma_C(\infty), \quad \frac{1}{n}\sum_{i=1}^n C_iC_i'g(Q_i) \to_p \Sigma_{Cg}(\infty), \quad \text{and} \quad \frac{1}{\sqrt{n}}\sum_{i=1}^n C_iU_i \Rightarrow G(\infty).$$

Hence, if we define $\mu_X(\theta)$ to equal the first term, we obtain the convergence (10) for X_n.

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

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

for

$$\mathcal{A}_n(\theta) \equiv n^{-1}\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\} - \left(n^{-1}\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\}\right)\left(n^{-1}\sum_{i=1}^n C_iC_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\}\right),$$
$$\mathcal{B}_n(\theta) \equiv n^{-1}\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\}g(Q_i) - \left(n^{-1}\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\}\right)\left(n^{-1}\sum_{i=1}^n C_iC_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^n C_iC_i'g(Q_i)\right),$$
$$\mathcal{C}_n(\theta) \equiv n^{-1/2}\sum_{i=1}^n C_iU_i1\{Q_i>\theta\} - \left(n^{-1}\sum_{i=1}^n C_iC_i'1\{Q_i>\theta\}\right)\left(n^{-1}\sum_{i=1}^n C_iC_i'\right)^{-1}\left(n^{-1/2}\sum_{i=1}^n C_iU_i\right).$$

Under (8) and (9), however, the continuous mapping theorem implies that

$$\mathcal{A}_n(\theta) \to_p \Sigma_C(\infty)-\Sigma_C(\theta)-(\Sigma_C(\infty)-\Sigma_C(\theta))\Sigma_C(\infty)^{-1}(\Sigma_C(\infty)-\Sigma_C(\theta))$$

$$
\begin{aligned}
&= \Sigma_C(\theta) - \Sigma_C(\theta)\Sigma_C(\infty)^{-1}\Sigma_C(\theta) \equiv \mathcal{A}^*(\theta), \\
\mathcal{B}_n(\theta) &\to_p \Sigma_{Cg}(\infty) - \Sigma_{Cg}(\theta) - (\Sigma_C(\infty) - \Sigma_C(\theta))\Sigma_C(\infty)^{-1}\Sigma_{Cg}(\infty) \\
&= \Sigma_C(\theta)\Sigma_C(\infty)^{-1}\Sigma_{Cg}(\infty) - \Sigma_{Cg}(\theta) \equiv \mathcal{B}^*(\theta), \\
\mathcal{C}_n(\theta) &\Rightarrow G(\infty) - G(\theta) - (\Sigma_C(\infty) - \Sigma_C(\theta))\Sigma_C(\infty)^{-1}G(\infty) \\
&= \Sigma_C(\theta)\Sigma_C(\infty)^{-1}G(\infty) - G(\theta) \equiv \mathcal{C}^*(\theta)
\end{aligned}
$$

all uniformly on $\tilde{\Theta}$, where this convergence holds jointly with that for X_n. By another application of the continuous mapping theorem,

$$
Y_n(\theta) = e_j'\sqrt{n}\hat{\delta}(\theta) \Rightarrow Y(\theta) = e_j'\mathcal{A}^*(\theta)^{-1}[\mathcal{B}^*(\theta) + \mathcal{C}^*(\theta)]. \tag{15}
$$

Hence, if we define $\mu_Y(\theta) = e_j'\mathcal{A}(\theta)^{-1}\mathcal{B}(\theta)$, we obtain the convergence (10), as desired.

We now obtain explicit expressions for the elements of the variance matrix $\Sigma(P)$ in terms of the consistently estimable quantities $\Sigma_C(\theta)$ and $E[G(\theta)G(\tilde{\theta})']$, dropping explicit dependence of expectation operators on P to ease notation. From (14) we obtain

$$
\Sigma_X(\theta,\tilde{\theta}) = \begin{pmatrix} \Sigma_X^{11}(\theta,\tilde{\theta}) & \Sigma_X^{12}(\theta,\tilde{\theta}) \\ \Sigma_X^{21}(\theta,\tilde{\theta}) & \Sigma_X^{22}(\theta,\tilde{\theta}) \end{pmatrix},
$$

where

$$
\begin{aligned}
\Sigma_X^{11}(\theta,\tilde{\theta}) &= \Sigma_C(\theta)^{-1/2}E[G(\theta)G(\tilde{\theta})']\Sigma_C(\tilde{\theta})^{-1/2}, \\
\Sigma_X^{12}(\theta,\tilde{\theta}) &= \Sigma_C(\theta)^{-1/2}(E[G(\theta)G(\infty)'] - E[G(\theta)G(\tilde{\theta})'])(\Sigma_C(\infty) - \Sigma_C(\tilde{\theta}))^{-1/2}, \\
\Sigma_X^{21}(\theta,\tilde{\theta}) &= (\Sigma_C(\infty) - \Sigma_C(\theta))^{-1/2}(E[G(\infty)G(\tilde{\theta})'] - E[G(\theta)G(\tilde{\theta})'])\Sigma_C(\tilde{\theta})^{-1/2}, \\
\Sigma_X^{22}(\theta,\tilde{\theta}) &= (\Sigma_C(\infty) - \Sigma_C(\theta))^{-1/2}(E[G(\infty)G(\infty)'] - E[G(\infty)G(\tilde{\theta})'] - E[G(\theta)G(\infty)'] + E[G(\theta)G(\tilde{\theta})']) \\
&\quad \times (\Sigma_C(\infty) - \Sigma_C(\tilde{\theta}))^{-1/2}.
\end{aligned}
$$

From (14) and (15) we obtain

$$
\begin{aligned}
\Sigma_{XY}(\theta,\tilde{\theta}) &= \begin{pmatrix} \Sigma_C(\theta)^{-1/2}E[G(\theta)\mathcal{C}^*(\tilde{\theta})']\mathcal{A}^*(\tilde{\theta})^{-1}e_j \\ (\Sigma_C(\infty) - \Sigma_C(\theta))^{-1/2}(E[G(\infty)\mathcal{C}^*(\tilde{\theta})'] - E[G(\theta)\mathcal{C}^*(\tilde{\theta})'])\mathcal{A}^*(\tilde{\theta})^{-1}e_j \end{pmatrix}, \\
\Sigma_{YX}(\theta,\tilde{\theta}) &= \begin{pmatrix} \Sigma_C(\tilde{\theta})^{-1/2}E[G(\tilde{\theta})\mathcal{C}^*(\theta)']\mathcal{A}^*(\theta)^{-1}e_j \\ (\Sigma_C(\infty) - \Sigma_C(\tilde{\theta}))^{-1/2}(E[G(\infty)\mathcal{C}^*(\theta)'] - E[G(\tilde{\theta})\mathcal{C}^*(\theta)'])\mathcal{A}^*(\theta)^{-1}e_j \end{pmatrix}',
\end{aligned}
$$

where

$$E[G(\theta)\mathcal{C}^*(\tilde{\theta})'] = E[G(\theta)G(\infty)']\Sigma_C(\infty)^{-1}\Sigma_C(\tilde{\theta}) - E[G(\theta)G(\tilde{\theta})'].$$

Finally, from (15) we obtain

$$\Sigma_Y(\theta,\tilde{\theta}) = e_j'\mathcal{A}^*(\theta)^{-1}E[\mathcal{C}^*(\theta)\mathcal{C}^*(\tilde{\theta})']\mathcal{A}^*(\tilde{\theta})^{-1}e_j,$$

where

$$\begin{aligned}E[\mathcal{C}^*(\theta)\mathcal{C}^*(\tilde{\theta})'] = {} & \Sigma_C(\theta)\Sigma_C(\infty)^{-1}E[G(\infty)G(\infty)']\Sigma_C(\infty)^{-1}\Sigma_C(\tilde{\theta}) - \Sigma_C(\theta)\Sigma_C(\infty)^{-1}E[G(\infty)G(\tilde{\theta})'] \\ & - E[G(\theta)G(\infty)']\Sigma_C(\tilde{\theta})\Sigma_C(\infty)^{-1} + E[G(\theta)G(\tilde{\theta})'].\end{aligned}$$