DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 24-02

Robust Inference when Nuisance Parameters may be Partially Identified with Applications to Synthetic Controls

Joseph Fry
University of Colorado Boulder

October 13, 2024

Department of Economics

University of Colorado Boulder
Boulder, Colorado 80309

Robust Inference when Nuisance Parameters may be Partially Identified with Applications to Synthetic Controls

Joseph Fry[†]

[†]University of Colorado at Boulder, 256 UCB, 80309 Boulder, Colorado, United States

E-mail address: joseph.fry@colorado.edu

October 13, 2024

Abstract

When conducting inference for the average treatment effect on the treated with a Synthetic Control Estimator, the vector of control weights is a nuisance parameter which is often constrained, high dimensional, and may be only partially identified even when the average treatment effect on the treated is point-identified. All three of these features of a nuisance parameter can lead to failure of asymptotic normality for the estimate of the parameter of interest when using standard methods. I provide a new method yielding asymptotic normality for an estimate of the parameter of interest, even when all three of these complications are present. This is accomplished by first estimating the nuisance parameter using a regularization penalty to achieve a form of identification, and then estimating the parameter of interest using moment conditions that have been orthogonalized with respect to the nuisance parameter. I present high-level sufficient conditions for the estimator and verify these conditions in an example involving Synthetic Controls. In simulations, this Orthogonalized Synthetic Control inference method has desirable size and power properties relative to existing inference methods.

1 Introduction

In the context of method of moments estimation, if the moment conditions jointly identify a subvector of the parameters, then a standard estimation method such as General Method of Moments (GMM) is generally a consistent estimator for that subvector. However, if the remaining subvector is unidentified, then this remaining parameter cannot be consistently estimated and as a result the estimates for the identified subvector are not asymptotically normal (see, for example, Andrews and Cheng (2012)), which significantly complicates inference. Additionally, if a nuisance parameter is at the boundary of the parameter space or close to the boundary relative to the sample size, then this can also result in our estimates for the parameter of interest not being asymptotically normal (see, for example, Andrews (1999) and Geyer (1994)). Lastly, if the full vector is high dimensional, this can complicate standard asymptotic normality results even when the subvector we would like to perform inference on is low dimensional. I propose an estimation method that aims to simultaneously overcome these complications to obtain an estimate for an identified parameter of interest that is asymptotically normal even when a nuisance parameter is partially identified, on or near the boundary of the parameter space, and, in some cases, high dimensional.

The procedure I propose can be decomposed into three steps. In the first step, regularized estimates of all parameters are found by minimizing a penalty function subject to the constraint that the sample moment conditions are close to zero. The primary purpose of this penalty function is to make the estimated nuisance parameter converge to a unique element of the identified set. We must therefore choose which element of the identified set we would like our estimate to converge to. Since this choice affects the asymptotic variance of our subsequent estimate of the parameter of interest, I base the penalty function on an estimate of the asymptotic variance as a function of the nuisance parameter, provided that this asymptotic variance function can be estimated sufficiently accurately. In cases where the asymptotic variance cannot be accurately estimated, such as when using time-series or panel data and the degree of temporal dependence is high relative to the number of time periods, I discuss alternative ways of choosing the penalty function in Section 3. The second step in my procedure is to use Neyman orthogonalization to construct a set of moment conditions that are orthogonal with respect to the nuisance parameter. This involves introducing a new parameter that is chosen to make the derivative of the moment conditions with respect to the original nuisance parameter equal to zero. After the estimated nuisance parameters have been plugged into the orthogonal moment conditions, the third step is to use these moment conditions to re-estimate the parameter of interest. In Section 2, I give two suggestions of how this can be done and provide asymptotic normality results for both methods.

My primary application of this approach is as a Synthetic Control Estimator (SCE) which gives an asymptotically normal estimate of the average treatment effect on the treated (ATT). Several other inference methods have been proposed for SCEs. The placebo method of Abadie et al. (2010) is commonly used in practice and, as previously noted (see Abadie et al. (2010) and Abadie et al. (2015)), corresponds to a traditional Fisher Randomization Test when treatment is randomly assigned. While this would mean that this test have exact size from a design-based perspective, this condition is unrealistic in most current SCE applications. Chernozhukov et al. (2024) propose a t-test inference method based on a K-fold cross-fitting procedure and Li (2020) propose a subsampling method. They each show that their method has asymptotically correct size when both the number of pre-treatment time periods T_0 and post-treatment time periods T_1 are large, which I also show for my method in section 4. Chernozhukov et al. (2024) also prove that their estimator is asymptotically normal, although it relies on the bias of their SCE being the same in the pre-treatment and post-treatment time periods.

Several papers proposing methods closely related to the SCE have also provided asymptotic normality results for the estimated ATT. Arkhangelsky et al. (2021) introduce a Synthetic Difference-in-Differences estimator, which involves a weighted Difference-in-Differences regression. They establish consistency and asymptotic normality of the estimated ATT, although this relies on having the number of control units go to infinity and the Euclidean norm of the control weights converge to zero at a sufficiently fast rate. Carvalho et al. (2018) provide a Lasso-based estimator of the ATT that is asymptotically normal and that allows that number of units to be large, but their inference method relies on consistently estimating the long-run variance, unlike my method and the t-test of Chernozhukov et al. (2024). Two other inference methods are the conformal inference method of Chernozhukov et al. (2021) and the End-of-Sample Instability Test, originally introduced by Andrews (2003) and applied to SCE by Cao and Dowd (2019). Chernozhukov et al. (2021)'s and Cao and Dowd (2019)'s methods require stronger assumptions on the degree of temporal dependence, but they both have the potential advantage that the sizes of their tests are asymptotically correct when T_1 is fixed and only $T_0 \to \infty$. The asymptotic results of Li (2020) and Cao and Dowd (2019) rely on keeping the number of control units fixed as the number of time periods grows, which can provide a poor approximation in applications where the number of controls is not small relative to the number of time periods. I compare my method with many of these approaches in simulations and show that it controls size and has the highest power among the methods that control size.

A large body of work considers inference in cases where a set of moment equations only partially identifies a vector of parameters (e.g., Chernozhukov et al. (2007) and Romano and Shaikh (2010)). Hansen (1996) offers a method for inference when a nuisance parameter is unidentified under the null hypothesis, but it requires simulating the sampling distribution of the estimated nuisance parameter. Chaudhuri and Zivot (2011) provide an inference method for GMM estimators when a nuisance parameter may be weakly identified, and Andrews and Cheng (2012) propose an inference method for extremum estimators when a subvector may be weakly identified. Han and McCloskey (2019) generalize Andrews and Cheng (2012)'s results to the case where the entire vector is allowed to be weakly identified by introducing a method of reparameterization. Cox (2022) also builds on this work for the case of, possibly constrained, minimum distance estimators. However, these methods generally do not allow the vector of parameters to be high dimensional. Additionally, by only considering cases where the parameter of interest is identified, we can focus on conducting the estimation in a way that makes inference simpler for that parameter (by making its estimated values asymptotically normal) and possibly in a way that makes the estimates of that parameter more precise (by minimizing its asymptotic variance). On the other hand, methods that conduct inference on the whole identified set have the advantage that they can be used when both the parameter of interest and the nuisance parameter are partially identified.

This work also extends the literature on the Neyman Orthogonalized Score. While the technique dates back to Neyman (1959), several more recent papers have used it as a way to achieve asymptotic normality after obtaining an estimate of a high-dimensional nuisance parameter using a regularization penalty or machine learning technique (e.g., Belloni et al. (2018), Ning and Liu (2017), Chernozhukov et al. (2015), Belloni et al. (2014), and Chernozhukov et al. (2018)). Many of these methods estimate the nuisance parameter with LASSO. While this can be a powerful technique when the nuisance parameter is high dimensional but has a sparse, point-identified value, if the nuisance parameter is partially identified, the LASSO penalty may often be insufficient for the estimates to converge to a specific vector. I help extend this literature by showing how the Neyman Orthogonalized Score can be applied in cases where the nuisance parameter is partially identified. The literature on optimal instruments is also related, and in particular, Singh et al. (2020) take a similar approach as I do here, where they choose the function of the instruments that minimizes the estimated asymptotic variance for the parameter of interest. In section 5, I give an example of how my method can be applied to the optimal choice of instruments.

In Section 2, I discuss how a set of moment conditions can be combined with initial regularized estimates of the parameters to create the orthogonal moment conditions and then how these orthogonal moment conditions can be used to estimate the parameter of interest. I discuss this first to show what properties we want the regularized estimates of the parameters to satisfy, and how the limiting values of the estimated nuisance parameters influence the asymptotic variance for the estimated parameter of interest. In Section 3, I then show when the regularized estimates satisfy these conditions and discuss how to adjust the procedure to handle cases where the asymptotic variance cannot be consistently estimated. In Section 4, I show when the high level conditions in the previous sections are satisfied for a SCE under a linear factor model structure. I apply this SCE by replicating the work of Andersson (2019) estimating the effect of Sweden's carbon pricing policies on CO2 emissions. My method finds that their results are statistically significant at commonly used levels, whereas several other methods do not. I also conduct simulations to compare the performance of my inference method with existing approaches for SCEs. Lastly, I discuss other applications, possible extensions, and limitations of the method in Section 5.

2 Neyman Orthogonalization

I assume that the researcher has a set of moment conditions they wish to use that are a function of a vector of parameters $\theta = (\beta, \delta)$ where the subvector $\beta \in B \subseteq \mathbb{R}^p$ is the parameter of interest and $\delta \in D_n \subseteq \mathbb{R}^J$ is a nuisance parameter. I let $\Theta_n = B \times D_n$ equal the parameter space for the entire vector. I use the subscript n to highlight the fact that the parameter space for the nuisance parameter may change as the sample size grows. This is particularly relevant when δ is high dimensional, so J is allowed to grow with n. Then I let $g(\theta) = E[\sum_{i=1}^n g_i(\theta)/n]$ and $\hat{g}(\theta) = \sum_{i=1}^n g_i(\theta)/n$ denote Q-dimensional vectors for the population and sample moment conditions with a sample size of n. Note that since I wish to allow for cases where observations are not identically distributed, I allow g to change with n. I am interested in the case where these moment conditions jointly identify the true value of the parameter of interest β, but may only partially identify δ. The true value of the parameter of interest and the identified set of the nuisance parameter may also change with the sample size, so I denote the identified set as

$$\Theta_{0,n} \coloneqq \{\theta \in \Theta_n : g(\theta) = 0\} = \{\beta_{0,n}\} \times D_{0,n}, \tag{1}$$

where $\beta_{0,n}$ is the true value of β and $D_{0,n}$ is the identified set for δ. Allowing for drifting sequences of $\beta_{0,n}$ is useful for multiple reasons. First, it is useful for analyzing whether the method is robust to $\beta_{0,n}$ being close to the boundary of the parameter space and being close to values where δ is unidentified. Additionally, in the SCE application, $\beta_{0,n}$ corresponds to the average of a sequence of dynamic treatment effects, so we want to allow that average to change as our sample size changes.

Using this set of the moment conditions, the orthogonalized moment conditions are given by

$$M(\theta, \eta) = \eta g(\theta), \tag{2}$$

where $\eta \in H \subseteq \mathbb{R}^{m \times Q}$ is an additional nuisance parameter and m is the number of orthogonalized moment conditions. Generally, I recommend having η be unconstrained so that $H = \mathbb{R}^{m \times Q}$, however in specific applications such as the SCE case, there can be reason to constrain η for achieving identification as I discuss in Section 4. Because the orthogonalized moments are linear combinations of the original moment conditions, there is no reason to choose $m > Q$. Furthermore, if the q-th element of the original moment conditions $g_q(\theta)$ does not vary with β at all, then $g_q(\beta, \delta_{0,n}) = 0$ for any $\delta_{0,n} \in D_{0,n}$. Therefore, if m is greater than the number of elements of g which are nontrivial functions of β, the elements of $M(\beta, \delta_{0,n}, \eta) = \eta g(\beta, \delta_{0,n})$ are linearly dependent functions of the elements of $g(\beta, \delta_{0,n})$, for any fixed value of η and $\delta_{0,n} \in D_{0,n}$. This is relevant because the orthogonalized moment conditions are used to estimate β after plugging in values of the nuisance parameters. Therefore, I set m equal to the number of moment conditions in g that are non-trivial functions of β.

We want to choose η so that asymptotically $M(\theta, \eta)$ is insensitive to δ. The first step in the estimation procedure is to obtain initial estimates of the parameters by picking values that minimize a penalty function among all the values that make the sample moment conditions close to zero, which I discuss how to do in Section 3. Let $\hat{\theta} = (\hat{\beta}, \hat{\delta})$ and $\hat{\eta}$ be equal to an initial regularized estimates of the parameters, and suppose that the distance from our estimates to the values $\theta_{0,n} = (\beta_{0,n}, \delta_{0,n})$ and $\eta_{0,n}$ is converging to zero as $n \to \infty$.[1] Then we want the sequence of matrices $\eta_{0,n}$ to satisfy

$$\partial_\delta M(\theta_{0,n}, \eta_{0,n}) = 0, \tag{3}$$

so that these moment conditions are not sensitive to δ at $(\theta_{0,n}, \eta_{0,n})$. Generally, there may exist many choices of η that satisfy this condition, particularly if $rank(\partial_\delta g(\theta_{0,n})) < Q$ which may happen when δ is partially identified. Therefore, η itself can be thought of as a partially identified nuisance parameter. By construction, $\partial_{vec(\eta)} M(\theta_{0,n}, \eta) = 0$ for any $\theta_{0,n} \in \Theta_{0,n}$ because $g(\theta_{0,n}) = 0$. Therefore, $M(\theta_{0,n}, \eta)$ is also orthogonal with respect to η. This allows η to be estimated in the same manner as δ, where a regularization penalty is used to make $\hat{\eta}$ converge to a specific element of the set $H_{0,n} \coloneqq \{\eta \in H : \partial_\delta M(\theta_{0,n}, \eta) = 0\}$. Note that this set depends on which $\delta_{0,n} \in D_{0,n}$ is selected, so it is necessary to either estimate δ first or estimate δ and η jointly. This means that a penalty function will also be used to select $\hat{\eta}$ from among all the values of $\eta \in H$ which make the sample moment conditions approximately orthogonal with respect to δ. This is very similar to the Neyman Near-Orthogonal Score introduced by Chernozhukov et al. (2018), although they handle the estimation of δ differently since they impose that the original nuisance parameter is point-identified.

Another property we want $\eta_{0,n}$ to satisfy is for $M(\beta, \delta_{0,n}, \eta_{0,n}) = \eta_{0,n} g(\beta, \delta_{0,n})$ to identify β. However, this can be handled by choosing our penalty function so that it diverges to infinity at values of η that make β unidentified. This naturally arises when the penalty function is based on the asymptotic variance for our estimate of β, since the expression for the asymptotic variance generally diverges as η approaches a point where β is unidentified. After estimates of the nuisance parameters $\hat{\delta}$ and $\hat{\eta}$ have been obtained, they can be plugged into the sample orthogonalized moments where

[1] Since the dimension of δ may be growing, which metrics this holds under is key for the results below. Assumption 2.1 contains the details on under exactly which metrics this convergence must hold.

$$\hat{M}(\beta, \hat{\delta}, \hat{\eta}) = \hat{\eta}\hat{g}(\beta, \hat{\delta}). \tag{4}$$

Because of equation (3), under suitable conditions, the sample moment conditions are not sensitive to the values of the nuisance parameters when $\beta = \beta_{0,n}$ and $(\hat{\delta}, \hat{\eta})$ is close to $(\delta_{0,n}, \eta_{0,n})$. In my asymptotic results, I focus on the cases where the dimensions of the parameter of interest and the moment conditions are fixed but the dimension of δ may either be fixed or growing with the sample size. This is relevant to the SCE case where δ is the vector of control weights, which may be of a similar size to the number of time periods.

Assumption 2.1 As $n \to \infty$ while p and Q are fixed and either J fixed or $J \to \infty$, we have that

1. $||\hat{\delta} - \delta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 = o_p(1/(\log(J)\log(n)))$.[2]

2. $g(\theta)$ is twice continuously differentiable on Θ_n and for each $q \in \{1, ..., Q\}$, $||\partial_\delta g_q(\theta_{0,n})||_\infty = O(\log(J))$ and $||\partial_\delta \hat{g}_q(\theta_{0,n}) - \partial_\delta g_q(\theta_{0,n})||_\infty = O_p(\log(J)/\sqrt{n})$ where $\hat{g}_q$ is the q-th element of $\hat{g}$.

3. There exists $\epsilon > 0$ such that for each $q \in \{1, ..., Q\}$, $\sup_{\delta:||\delta-\delta_{0,n}||_1 \leq \epsilon} \max eig(\partial^2_\delta \hat{g}_q(\beta_{0,n}, \delta)) = O_p(\log(J))$ where $\hat{g}_q$ is the q-th element of $\hat{g}$ and $\max eig$ denotes the maximum eigenvalue of a matrix.

4. Either $||\hat{\delta} - \delta_{0,n}||_1 = o_p(n^{-1/4}/\sqrt{\log(J)})$ and $||\hat{\eta} - \eta_{0,n}||_1 = o_p(n^{-1/4}/\sqrt{\log(J)})$ or $\hat{g}$ is linear in θ and $||\hat{\eta}\partial_\delta \hat{g}(\theta)||_\infty = O_p(\log(J)\log(n)/\sqrt{n})$.

For Assumption 2.1.2, I show in Appendix B, that if a triangular array $\{X_i\}_{n\in\mathbb{N}}$ where $X_i = \{X_{i1}, ..., X_{iJ}\}$ is α-mixing with exponential speed, is mean-invariant, has uniformly bounded fourth moments, and has an exponential-type bound on the tails of their distributions, then $\max_{1\leq j\leq J} |E[\sum_{i=1}^n X_{ij}/n] - \sum_{i=1}^n X_{ij}/n| = O_p(\log(J)/\sqrt{n})$ when $n, J \to \infty$ with $J/n^\gamma \to 0$ for some $\gamma > 0$. This allows for cases where there is a significant degree of dependence across

[2] I use $||\cdot||_1$ to denote the L1 norm and $||\cdot||_\infty$ to denote the L^∞ norm for both vectors and matrices. Similarly, I use $||\cdot||_2$ to denote both the L2 norm for vectors and the Frobenius norm for matrices.

the observations and J is growing faster than n. This is important is some applications, such as the SCE case, where there may be a significant degree of temporal dependence and the length of the control weights (which corresponds to J) may be greater than n. Assumption 2.1.3 imposes that there is a bound on how locally convex or concave g is at $\theta_{0,n}$, which holds trivially when g is linear in θ.

Assumption 2.1.1 imposes that distances between $\hat{\delta}$ and $\delta_{0,n}$ as well as between $\hat{\eta}$ and $\eta_{0,n}$ using the L1 norm are converging to zero at the given rates. For $\hat{\delta}$, the faster that its dimension is growing, the faster its rate of convergence must be. Furthermore, an additional condition on its rate of convergence must be imposed when the moment conditions are a non-linear function of δ, as shown in Assumption 2.1.4. The reason why weaker conditions are needed when $\hat{g}$ is linear in δ is because in this case, making $\partial_{\delta}\hat{M}(\beta_{0,n}, \delta_{0,n}, \eta_{0,n})$ close to zero makes $\partial_{\delta}\hat{M}(\beta_{0,n}, \delta, \eta_{0,n})$ close to zero for any δ. In cases where it is hard to achieve a rate of convergence for the nuisance parameters that is faster than $n^{-1/4}$, Mackey et al. (2018) shows that making the moment conditions h-th order orthogonal can allow this condition to be weakened to $o_p(n^{-1/(2h+2)})$. I show how to obtain $\hat{\delta}$ and $\hat{\eta}$ so that Assumptions 2.1.1 and 2.1.4 are satisfied under plausible conditions in Section 3. Together with the orthogonality condition, this gives the following adaptivity condition.

Lemma 2.1 (Adaptivity Condition) Suppose $(\beta_{0,n}, \delta_{0,n}) \in \Theta_{0,n}$, $\eta_{0,n}$ satisfies equation (3), and Assumption 2.1 holds. Then as $n \to \infty$ with p and Q fixed with either J fixed or $J \to \infty$,

$$\sqrt{n}(\hat{M}(\beta_{0,n}, \hat{\delta}, \hat{\eta}) - \hat{M}(\beta_{0,n}, \delta_{0,n}, \eta_{0,n})) = o_p(1).$$

Because of this adaptivity condition, an estimator of β using $\hat{M}(\beta, \hat{\delta}, \hat{\eta})$ is asymptotically equivalent to an estimator using $\hat{M}(\beta, \delta_{0,n}, \eta_{0,n})$. One way to use these orthogonalized moment conditions to estimate β is via a GMM estimator:

$$\tilde{\beta}_{GMM} = \arg\min_{\beta \in B} \hat{M}(\beta, \hat{\delta}, \hat{\eta})' W_n \hat{M}(\beta, \hat{\delta}, \hat{\eta}), \tag{5}$$

where W_n is a $m \times m$ weighting matrix. As is common for GMM estimators, when a consistent estimator of the asymptotic variance of $\sqrt{n}\hat{M}(\theta_{0,n}, \eta_{0,n})$ is available, it is most efficient to let W_n be equal to the inverse of the estimated asymptotic variance. I discuss the choice of W_n further in Section 3. In many cases, it is possible to show that $\sqrt{n}\hat{M}(\theta_{0,n}, \eta_{0,n})$ is asymptotically normal since it is a linear function of a vector of averages. This allows for modified versions of standard arguments for the asymptotic normality of GMM estimators to be applied.

Assumption 2.2 As $n \to \infty$ while Q and p are fixed and either J fixed or $J \to \infty$, we have that

1. $\sup_{\theta \in \Theta_n} ||\hat{g}(\theta) - g(\theta)||_2 = o_p(1)$
2. For all $\epsilon > 0$, there exists γ_ϵ such that $P(\sup_{\theta \in \Theta_n : ||\theta - \theta_{0,n}||_1 < \gamma_\epsilon} ||\partial_\beta \hat{g}(\theta) - \partial_\beta \hat{g}(\theta_{0,n})||_2 > \epsilon) \to 0$.
3. $\sqrt{n}\hat{M}(\theta_{0,n}, \eta_{0,n}) \xrightarrow{d} N(0, V_M)$ for some sequence of positive definite matrix V_M.
4. $\partial_\beta M(\beta_{0,n}, \delta_{0,n}, \eta_{0,n}) \to M_\beta$ for some matrix M_β with $rank(M_\beta) = p$ and $||\beta_1 - \beta_2||_2 \leq C||M(\beta_1, \delta_{0,n}, \eta_{0,n}) - M(\beta_2, \delta_{0,n}, \eta_{0,n})||_2$ for all $\beta_1, \beta_2 \in B$ and all n for some $C > 0$.
5. $W_n - W \xrightarrow{p} 0$ for some positive definite matrix W and the sequence $\eta_{0,n}$ is bounded.
6. There exists $\epsilon > 0$ such that $\{\beta : ||\beta - \beta_{0,n}||_1 < \epsilon\} \subseteq B$ for all n.

For the uniform convergence condition in Assumption 2.1.1, I provide an example of when this will hold in Appendix B, provided $\hat{g}$ satisfies a stochastic Lipschitz continuity condition. Note that Assumption 2.2.6 imposes that $\beta_{0,n}$ are interior points of B and bounded away from the boundary of B since extremum estimators like the one defined by equation (5) are generally not be asymptotically normal when the parameter is on the boundary of the parameter space or close to the boundary. Even when $\hat{g}$ is linear in β, $\partial_\beta \hat{g}(\theta)$ may still vary with δ so Assumption 2.2.2 is imposed to ensure that $\partial_\beta \hat{g}(\beta_{0,n}, \hat{\delta})$ converges to $\partial_\beta \hat{g}(\beta_{0,n}, \delta_{0,n})$ as $\hat{\delta}$ converges to $\delta_{0,n}$. This condition trivially holds when the cross partial derivatives $\partial_{\beta\delta} \hat{g}(\theta)$

are equal to zero. Assumption 2.2.3 can be directly combined with the adaptivity condition to show the asymptotic normality of $\sqrt{n}\hat{M}(\beta_{0,n}, \hat{\delta}, \hat{\eta})$ and Assumption 2.2.4 guarantees the identification of β. The penalty function can be chosen to help ensure that Assumption 2.2.4 holds and $\eta_{0,n}$ is bounded.

Because $\hat{M}(\theta_{0,n}, \eta_{0,n})$ involves taking sample averages, in the case where J is fixed it can be satisfied under standard conditions that allow Central Limits Theorems to be applied. I provide examples of this in Appendix B, including conditions that allow for quite general forms of non-stationarity. In cases where $J \to \infty$, arguments justifying Assumption 2.2.3 are more complicated. However, as I show with the SCE case, when $\delta_{0,n}$ is high dimensional but sparse, Assumption 2.2.3 holds under very similar conditions to the low dimensional case. As for the convergence of W_n, I discuss this further in Section 3 when describing how to choose the weighting matrix.

For the case when the parameter of interest may be close to or at the boundary, we can use a "One-Step" estimator $\tilde{\beta}_{OS}$, equal to

$$\tilde{\beta}_{GMM} - (\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}))^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}). \tag{6}$$

This estimator can be thought of as minimizing the quadratic approximation of the GMM objective function at the initial estimate $\tilde{\beta}_{GMM}$. This means that when β is unconstrained, the estimators $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$ are identical. The conditions for the asymptotic normality of the One-Step estimator are the same as for the GMM estimator, except $\beta_{0,n}$ is allowed to be on the boundary or close to the boundary of the parameter space. This follows from the same reasoning as Theorem 1 in Ketz (2018) for his "quasi-unconstrained" estimator.

Proposition 2.1 (Asymptotic Normality) Suppose $(\beta_{0,n}, \delta_{0,n}) \in \Theta_{0,n}$, $\eta_{0,n}$ satisfies equation (3), and Assumptions 2.1 and 2.2.1-2.2.5 hold. Then as $n \to \infty$ with p and Q fixed with either J fixed or $J \to \infty$, $\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n}) = O_p(1)$ and

$$\sqrt{n}(\tilde{\beta}_{OS} - \beta_{0,n}) \xrightarrow{d} N(0, V),$$

where $V = (M_\beta' W M_\beta)^{-1} M_\beta' W V_M W M_\beta (M_\beta' W M_\beta)^{-1}$. If Assumption 2.6 additionally holds, then as $n \to \infty$ with p and Q fixed with either J fixed or $J \to \infty$,

$$\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n}) \xrightarrow{d} N(0, V).$$

Therefore, when the same weighting matrix W_n is used, the GMM and One-Step estimators achieve the same asymptotic variance. When $W = V_M^{-1}$, V simplifies to $(M_\beta' V_M^{-1} M_\beta)^{-1}$. In general, the choice of the nuisance parameters may influence the precision of $\tilde{\beta}_{GMM}$ and β_{OS} both through changing how much variability there is in the sample moment conditions (i.e., V_M) and through how sensitive those moment conditions are to β (i.e., M_β).

3 Regularized Estimation

I now discuss estimation of the nuisance parameters. The results in the previous section hold for many possible values of $(\delta_{0,n}, \eta_{0,n}) \in D_{0,n} \times H_{0,n}$. For some penalty function $f(\theta, \eta)$, I define the optimal nuisance parameters as

$$(\delta_{0,n}, \eta_{0,n}) = \underset{\delta \in D_{0,n}, \eta \in H_{0,n}}{\arg\min} f(\beta_{0,n}, \delta, \eta). \tag{7}$$

The primary purpose of the penalty function is to select a unique pair of elements from the identified sets. Therefore, we want there to be unique elements of the identified sets that minimize $f(\theta, \eta)$. However, the penalty not only influences whether the estimated nuisance parameters each converge to a particular element of the identified sets, but also which elements of the identified sets they converge to. In subsection 3.2 below, I show how in some cases a form of relative asymptotic efficiency can be achieved by making the penalty function depend on the asymptotic variance of $\tilde{\beta}_{GMM}$.

However, since the asymptotic variance is not observed, if the penalty function depends on this, then $f(\theta, \eta)$ is also unknown. Since $D_{0,n}$, and $H_{0,n}$ are unknown as well, the nuisance parameters are chosen to minimize an estimated penalty function $\hat{f}$ among all parameters that come close to setting the sample versions of the moment conditions equal to zero. We can define the feasible set for the regularized parameters as $\hat{\Theta}_0 = \{\theta \in \Theta_n : ||\hat{g}(\theta)||_\infty \leq \lambda_\delta\}$ and $\hat{H}_0 = \{\eta \in \mathbb{R}^{m \times Q} : ||\partial_\delta \eta \hat{g}(\theta)||_\infty \leq \lambda_\eta\}$, where λ_δ and λ_η are tuning parameters whose choice is discussed below. Then we can estimate the parameters using

$$(\hat{\beta}, \hat{\delta}, \hat{\eta}) = \underset{\theta \in \hat{\Theta}_0, \eta \in \hat{H}_0}{\arg\min} \; \hat{f}(\theta, \eta). \tag{8}$$

3.1 Rate of Convergence

For analyzing the asymptotic properties of this estimator, I consider two cases as before: one where the dimension of δ is fixed and one where the dimension is allowed to grow with the sample size. In the case for which J is fixed and $\hat{g}$ is linear in θ, Assumption 2.1 only requires that $||\hat{\delta} - \delta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 = o_p(1/\log(n))$ and $||\hat{\eta}\partial_\delta \hat{g}(\beta_{0,n}, \delta)||_\infty = O_p(n^{-1/2}/\log(n))$. This allows for very slow rates of convergence for $\hat{\delta}$ and $\hat{\eta}$. Since $||\hat{\eta}\partial_\delta \hat{g}(\hat{\theta})||_\infty \leq \lambda_\eta$, in the linear case, the second condition can be directly achieved by choosing λ_η to be $O_p(n^{-1/2}/\log(n))$. However, if $\hat{g}$ is non-linear in δ, then we also want $||\hat{\delta} - \delta_{0,n}||_2 = o_p(n^{-1/4})$ and $||\hat{\eta} - \eta_{0,n}||_2 = o_p(n^{-1/4})$ in order for Assumption 2.1.4 to hold. While this still allows for rates of convergence slower than the standard parametric $\sqrt{n}$ rate, stronger assumptions are imposed to satisfy these conditions for the non-linear case.

Assumption 3.1 As $n \to \infty$ while p and Q are fixed, and either J fixed or $J \to \infty$, we have that

1. g and $\hat{g}$ are continuously differentiable on Θ_n. For any $\zeta > 0$,

$$\sup_{(\theta,\eta)\in\Theta_n\times H: f(\theta,\eta)\leq f(\theta_{0,n},\eta_{0,n})+\zeta} ||\eta||_1$$

   is bounded and the set $\{(\theta,\eta) \in \Theta_n \times H : f(\theta,\eta) \leq f(\theta_{0,n},\eta_{0,n}) + \zeta\}$ is compact for all n.

2. For some sequences of positive constants $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $\sup_{\theta\in\Theta_n} ||\hat{g}(\theta)-g(\theta)||_\infty = O_p(a_n)$, $\sup_{\theta\in\Theta_{0,n}} ||\hat{g}(\theta) - g(\theta)||_\infty = O_p(b_n)$, $\sup_{\theta\in\Theta_n} ||\partial_\delta\hat{g}(\theta) - \partial_\delta g(\theta)||_\infty = O_p(a_n)$, and $\sup_{\theta\in\Theta_{0,n}} ||\partial_\delta\hat{g}(\theta) - \partial_\delta g(\theta)||_\infty = O_p(b_n)$.

3. For some sequences of non-negative constants $\{c_n\}_{n\in\mathbb{N}}$ $\sup_{\theta\in\hat{\Theta}_0,\eta\in\hat{H}_0} |\hat{f}(\theta,\eta) - f(\theta,\eta)| = O_p(c_n)$.

4. There exists constants $C_1, C_2 > 0$ such that $||g(\theta)||_\infty + ||\partial_\delta\eta g(\theta)||_\infty \geq C_2 \max\{||\theta - \Theta_{0,n}||_1 + ||\eta - H_{0,n}||_1, C_1\}$ for all $\theta \in \Theta, \eta \in H$.

5. There exists constants $C_3, C_4, C_5, C_6 > 0$ and $\gamma_1, \gamma_2 > 0$, such that for all $(\theta,\eta) \in \Theta_{0,n} \times H_{0,n}$ $|f(\theta,\eta) - f(\theta_{0,n},\eta_{0,n})| \geq C_4 \min\{((||\theta - \theta_{0,n}||_1 + ||\eta - \eta_{0,n}||_1)^{\gamma_1}, C_3\}$ and for all $(\theta_1,\eta_1),(\theta_2,\eta_2) \in \Theta_n \times H$, if $f(\theta_1,\eta_1), f(\theta_2,\eta_2) \leq f(\theta_{0,n},\eta_{0,n}) + C_5$ then $||\theta_1 - \theta_2||_1 + ||\eta_1 - \eta_2||_1 \geq C_6|f(\theta_1,\eta_1) - f(\theta_2,\eta_2)|^{\gamma_2}$.

For the estimator defined by equation (8), both the objective function and feasible set may be stochastic, so there can be uncertainty coming from both the estimated penalty function and the sample moment conditions. As a result, the rate of convergence for $\hat{\delta}$ and $\hat{\eta}$ depends both on the rate of convergence of the sample moments to the population moments and the rate of convergence of the estimated penalty function.

Assumption 3.1.1 allows for the parameter spaces to not be compact, as long as the feasible values of the parameters that make the penalty sufficiently small are compact. Assumption 3.1.4 can be viewed as an extension of Assumption 2.2.4. It provides a strong identification condition for the identified sets $\Theta_{0,n}$ and $H_{0,n}$. Strong partial identification conditions of this form are common in the literature on estimating identified sets and are imposed by others such as Chernozhukov et al. (2007). It holds in a variety of applications including the case of a SCE under a linear factor model data-generating process discussed in Section 4 but also in cases where the partial identification is caused by collinearity and an insufficient number of instruments. Assumption 3.1 weakens the assumptions imposed by Chernozhukov et al. (2007) by not requiring that the parameter space or the identified sets to be compact, and it allows for the dimension of θ to be growing. Part of the reason that these conditions can be weakened is that I do not need to show that $\hat{\Theta}_0$ and $\hat{H}_0$ are converging to $\Theta_{0,n}$ and $H_{0,n}$. Rather, I only need to show convergence of the feasible set on the subset of the parameter space where f is small, which is why the compactness condition in Assumption 3.1.1 is imposed.

Assumption 3.1.2 strengths Assumption 2.2.1 by imposing a specific rate for the uniform convergence of the sample moment conditions. Achieving this condition is difficult in some cases when B and D_n are not bounded, although these conditions could likely be weakened when $||\hat{g}(\theta)||_\infty$ and $||\eta\partial_\delta\hat{g}(\theta)||_\infty$ are convex functions. I impose a rate of convergence for the estimated penalty function in Assumption 3.1.3 and a condition relating $|f(\beta_{0,n}, \delta, \eta) - f(\theta_{0,n}, \eta_{0,n})|$ to $||\delta - \delta_{0,n}||_1$ and $||\eta - \eta_{0,n}||_1$ in Assumption 3.1.5. Note that for Assumption 3.1.5, it is only necessary that among elements of $D_{0,n}$ and $H_{0,n}$ that are close to $\delta_{0,n}$ and $\eta_{0,n}$, $(||\delta - \delta_{0,n}||_1 + ||\eta - \eta_{0,n}||_1)^{\gamma_1}$ can be bounded by $f(\beta_{0,n}, \delta, \eta) - f(\theta_{0,n}, \eta_{0,n})$. This allows us to guarantee that if elements of the identified sets achieve close to the minimum value of f, then they must be close to $\theta_{0,n}$ and $\eta_{0,n}$. It may often be the case that $(\delta_{0,n}, \eta_{0,n})$ is on the boundary of the identified sets $D_{0,n}$ and $H_{0,n}$ and is not a local minimum of $f(\beta_{0,n}, \delta, \eta)$ on $D_n \times H$. This is why the second part of Assumption 3.2.5 is needed, because it allows us to place a bound on the rate of change of $f(\theta, \eta)$ for values of the nuisance parameter that give a penalty value which is not significantly greater than the value at $(\theta_{0,n}, \eta_{0,n})$. As a result, values just outside $D_{0,n} \times H_{0,n}$ should not make f much lower than $f(\theta_{0,n}, \eta_{0,n})$. This still allows for the possibility that $f(\theta, \eta)$ may diverge to infinity at some values of (θ, η), such as values of η where β becomes unidentified when f depends on V.

When f depends on the V, Assumption 3.1.3 often holds when $\hat{f}$ is the same function of an estimate of the asymptotic variance and $\hat{V}$ is converging to V at a rate of c_n. In subsection 3.2 below, I discuss how the estimated variance is often converging at least as fast as a rate as the sample moment conditions. Alternatively, f can be chosen to be a known function, in which case Assumption 3.1.3 holds trivially with $\hat{f}(\theta, \eta) = f(\theta, \eta)$, so $c_n = 0$. Under similar regularity conditions on the sample moment conditions as before, when Θ_n is compact and J is fixed, Assumption 3.1.2 often holds with $a_n = 1/\sqrt{n}$. In the case where J is growing, a_n is generally growing with J but adding constraints on δ can help ensure that it is growing slowly in J (see Remark 3.1). For the SCE case, I use $f(\theta, \eta) = ||\delta||_2^2 + ||\eta||_2^2$ which, along with the properties of $D_{0,n}$ and $H_{0,n}$, allows for Assumption 3.1.5 to hold with $\gamma_1 = 2$ and $\gamma_2 = 1$.

Lemma 3.1 (Rate of Convergence of Regularized Estimates) Suppose that Assumption 3.1 holds, and $\lambda_\delta, \lambda_\eta \to 0$ such that $\min\{\lambda_\delta, \lambda_\eta\}/b_n \to \infty$ as $n \to \infty$ with p and Q fixed and either J fixed or $J \to \infty$. Then

$$||\hat{\delta} - \delta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 = O_p(\max\{\lambda_\delta, \lambda_\eta, a_n, c_n\}^{\gamma_2/\gamma_1}).$$

Lemma 3.1 requires that λ_δ and λ_η are shrinking more slowly than b_n. This guarantees that $\Theta_{0,n} \subseteq \hat{\Theta}$ and $H_{0,n} \subseteq \hat{H}$ with probability approaching one. Similarly as to with a_n and c_n, when the identified sets are compact and J is fixed, we often have that $b = 1/\sqrt{n}$. We can therefore satisfy this condition by having λ_η and λ_δ shrinking at a slightly slower rate than a_n and b_n (e.g., $\lambda_\delta, \lambda_\eta = O_p(n^{-1/2}\log(n))$). In such cases, we have $||\hat{\delta} - \delta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 = O_p((n^{-1/2}\log(n))^{\gamma_2/\gamma_1})$. For the linear case, we only need $||\hat{\delta} - \delta_{0,n}||_1 = o_p(1/\log(n))$ when J is fixed to satisfy Assumption 2.1.1 and 2.1.4, so Assumption 3.1.5 holding with any $\gamma_1, \gamma_2 > 0$ is sufficient. On the other hand, for the non-linear case, we want $||\hat{\delta} - \delta_{0,n}||_1 = o_p(n^{-1/4})$ so we need $\gamma_2/\gamma_1 > 1$. In cases where $\gamma_2/\gamma_1 \leq 1$, one potential solution is the approach of

Mackey et al. (2018), where further nuisance parameters are introduced to make the moment conditions h-th order orthogonal, in which case we only need that $||\hat{\delta}-\delta_{0,n}||_2 = o_p(n^{-1/(2h+2)})$. When $J \to \infty$, a_n as well as c_n if $c_n \neq 0$ are generally growing with J. However, as long as it is growing slowly in J and J is not growing too much faster than n, we can satisfy Assumption 2.1.1 and 2.1.4 under similar conditions as when J is fixed.[3] Having constraints on δ is beneficial for guaranteeing a_n and c_n grow slowly in J.

Remark 3.1 (Adding L1 norm Constraints) Suppose that we additionally add a constraint on the L1 norm of δ to the estimator defined by equation (8) uses $\tilde{\Theta}_n = \{\theta \in \Theta_n : ||\delta||_1 \leq \lambda_1\}$ as its parameter space, for some additional penalty term λ_1. Furthermore, suppose that g and $\hat{g}$ are linear in θ so $g(\theta) = E[\sum_{i=1}^n (X_0 - X_1\beta - X_2\delta)/n]$ and $\hat{g}(\theta) = \sum_{i=1}^n (X_{0,i} - X_{1,i}\beta - X_{2,i}\delta)/n$. Then for the uniform convergence conditions in Assumption 3.2.2, $\sup_{\theta\in\tilde{\Theta}_n} ||\hat{g}(\theta)-g(\theta)||_\infty \leq || \sum_{i=1}^n E[X_{0,i}]/n - \sum_{i=1}^n X_{0,i}/n||_\infty + \sup_{\beta\in B} ||(\sum_{i=1}^n E[X_{1,i}]/n - \sum_{i=1}^n X_{1,i}/n)\beta||_\infty + \lambda_1 \max_{1\leq j\leq J} |\sum_{i=1}^n E[X_{2,j,i}]/n - \sum_{i=1}^n X_{2,j,i}/n|$. Note that the last term is also a bound on $\sup_{\theta\in\tilde{\Theta}_n} ||\partial_\delta \hat{g}(\theta) - \partial_\delta g(\theta)||_\infty$. In many cases, the maximum in the last term can be growing slowly in J. Additionally, if it results in $\delta_{0,n}$ being sparse, this can make it easier to show that $\sqrt{n}\hat{M}(\beta_{0,n}, \delta_{0,n}, \eta_{0,n})$ is asymptotically normal. On the other hand, if λ_1 is kept too small so that $\{\delta : ||\delta||_1 \leq \lambda_1\} \cap D_{0,n} = \emptyset$, then it causes Lemma 3.1 to fail to hold. Also, further constraining δ may increase what the minimum asymptotic variance is obtainable is.

The conditions placed on λ_δ and λ_η in Lemma 3.1 do not tell us how to choose them in practice. Therefore, in order to reduce room for specification searching by researchers, it is important to have an algorithm for determining these tuning parameters. One plausible approach is based on estimating confidence sets for the identified sets $\Theta_{0,n}$ and $H_{0,n}$. When a parameter is partially identified by some population criterion function, one way to construct confidence sets for the identified set, used by Chernozhukov et al. (2007) and others, is to

[3] For example, if $a_n = b_n = c_n = n^{-1/2}\log(J)$ then we can choose λ_δ and λ_η to be $O_p(n^{-1/2}\log(J)\log(n))$ so that $||\hat{\delta} - \delta_{0,n}||_1 = O_p((n^{-1/2}\log(J)\log(n))^{\gamma_2/\gamma_1})$. Then as long as $\gamma_2, \gamma_1 > 0$ and $J/n^\gamma \to 0$ for some $\gamma > 0$, we have that $\hat{\delta}$ satisfies Assumptions 2.1.1 and 2.1.4 for the linear case.

use the set where the sample criterion function is below a particular value. This is similar to how $\hat{\Theta}_0$ and $\hat{H}_0$ are constructed. However, if λ_δ and λ_η are chosen so that $\hat{\Theta}_0$ and $\hat{H}_0$ are confidence sets for $\Theta_{0,n}$ and $H_{0,n}$, this still leaves the choice of the confidence level. In order for $\min\{\lambda_\delta, \lambda_\eta\}/b_n \to \infty$, we want this confidence level to be converging to 100% as $n \to \infty$ and there is still the question of how to choose the confidence level in practice. Furthermore, these methods often require resampling and therefore may be computationally intensive. For the simulation results in Section 4, I use a computationally simple method which guarantees that $\hat{\Theta}_0$ and $\hat{H}_0$ are non-empty but also that λ_δ and λ_η are shrinking at a rate of $b_n \log(n)$.

3.2 Variance Estimation and Inference

As mentioned earlier,

$$V = (M_\beta' W M_\beta)^{-1} M_\beta' W V_M W M_\beta (M_\beta' W M_\beta)^{-1}.$$

Therefore, we can estimate V with

$$\hat{V}(\hat{\theta}, \hat{\eta}) = (\hat{M}_\beta' W_n \hat{M}_\beta)^{-1} \hat{M}_\beta' W_n \hat{V}_M(\hat{\theta}, \hat{\eta}) W_n \hat{M}_\beta (\hat{M}_\beta' W_n \hat{M}_\beta)^{-1},$$

where $\hat{M}_\beta = \partial_\beta \hat{M}(\hat{\theta}, \hat{\eta})$ and $\hat{V}_M(\hat{\theta}, \hat{\eta})$ is an estimate of the asymptotic variance of the orthogonal moment conditions. In the case where the data are I.I.D., V_M is simply the variance matrix for the moment conditions, and if $\hat{V}_M(\theta, \eta)$ is the sample variance matrix then under simple regularity conditions $\hat{V}_M(\hat{\theta}, \hat{\eta})$ is $\sqrt{n}$-consistent when J is fixed. However, in contexts with time series or panel data, V_M often corresponds to the long-run variance of the moment conditions.

When V_M corresponds to the long-run variance and the structure of the dependence across observations is unknown, it is common to use estimators in the class of quadratic

Heteroscedastic Autocorrelation Consistent (HAC) variance estimators that take the form:

$$\hat{V}_M(\theta,\eta) = \sum_{i=1}^{n}\sum_{s=1}^{n} Q_K(\frac{i}{n},\frac{s}{n})\eta(g_i(\theta) - \hat{g}(\theta))(g_s(\theta) - \hat{g}(\theta))'\eta'/n, \tag{9}$$

where $Q_K(i,s)$ is a weighting function that depends on a smoothing parameter K. This includes kernel variance estimators such as those of Andrews (1991) and Newey and West (1987) as well as the orthonormal series variance estimators such as that of Phillips (2005). For conventional Kernel HAC estimators, $Q_K(i,s) = \mathcal{K}((i-s)/K)$ where $\mathcal{K}$ is a kernel and K is the bandwidth. For Series HAC estimators, $Q_K(i,s) = \sum_{k=1}^{K}\phi_k(i)\phi_k(s)/K$ where $\{\phi_k(s)\}_{k=1}^{K}$ are orthonormal basis functions taking values in $[0,1]$.

Asymptotic results for non-parametric long-run variance estimators that rely on $K \to \infty$ as $n \to \infty$ can often provide poor approximations in practice, particularly when the degree of temporal dependence is high relative to the sample size. Intuitively, this is because the uncertainty in our estimation of V_M significantly contributes to our uncertainty in our test statistic in these cases, but this is not captured by increasing-smoothing asymptotic results. SCEs are often used is settings with small to moderate sample sizes and with data that display a high degree of dependence over time, making the use of increasing-smoothing asymptotic results particularly questionable. This is illustrated by Chernozhukov et al. (2024) who show that their inference procedure for their SCE performs very poorly when they calculate their standard errors using a HAC estimator and rely on increasing-smoothing asymptotic results to obtain critical values.

The notion of fixed-smoothing asymptotics was first introduced by Vogelsang and Kiefer (2002), Kiefer and Vogelsang (2002), and Vogelsang and Kiefer (2005). As the name suggests, it involves keeping the degree of smoothing K fixed as the sample size grows. For both Kernel and Series HAC estimators, this results in $\hat{V}_M(\hat{\theta},\hat{\eta})$ converging to a stochastic matrix. The fact that $\hat{V}_M(\hat{\theta},\hat{\eta})$, and therefore $\hat{V}(\hat{\theta},\hat{\eta})$, are converging to something stochastic has several implications for this procedure, since $\hat{V}_M$ is potentially being used three different times. It may be used to estimate $\hat{\delta}$ and $\hat{\eta}$, it can be used to estimate $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$ by setting $W_n = \hat{V}_M(\hat{\theta}, \hat{\eta})^{-1}$, and lastly it may be used along with $\tilde{\beta}_{GMM}$ or $\tilde{\beta}_{OS}$ to construct a test statistic or confidence interval. If $\hat{V}$ is not converging to V, then $\hat{f}$ should not be a function of $\hat{V}$ in order to satisfy Assumption 3.1.3. Therefore, we want to define an alternative penalty function. One alternative is to have f depend on a known function which upper bounds $V(\theta, \eta)$. In this case, Assumption 3.1.3 can be trivially satisfied with $\hat{f} = f$. In Section 4, I give an example of how to do this for the SCE case.

For using $\hat{V}_M(\hat{\theta}, \hat{\eta})$ to weight the moment conditions, Hwang and Sun (2018) compare the performance of the One-Step and Two-Step GMM estimator under a fixed-smoothing asymptotic framework. They show that whether the Two-Step GMM procedure outperforms the One-Step GMM procedure depends on the values of long-run correlation coefficients. Because these long-run correlations can also not be consistently estimated under the fixed-smoothing asymptotic framework, it is generally not clear whether the One-Step or Two-Step GMM estimators performs better. As shown by Sun (2014a), under fixed-smoothing asymptotics, while the One-Step GMM estimator is still asymptotically normal, the Two-Step GMM estimator is asymptotically mixed normal. Because of this, it is easier to choose $\hat{\delta}$ and $\hat{\eta}$ to minimize an upper bound on the asymptotic variance and simpler to conduct inference when $W_n = I_m$. For these reasons, I focus on using $W_n = I_m$ and have the penalty function be based on the upper bound of the asymptotic variance when fixed-smoothing asymptotic results are relevant.

However, even when $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$ are asymptotically normal, common test statistics can still have nonstandard limiting distributions. For example, the Wald test statistic, rather than converging to a chi-squared distribution, converges to a distribution which depends on the kernel or basis function and the smoothing parameter. For Kernel HAC estimators, Sun (2014b) provides conditions under which an adjusted Wald statistic and an adjusted t-statistic have asymptotic distributions that can be approximated by F and t distributions, but do not converge exactly to F and t distributions. On the other hand, Sun (2013) gives versions of the Wald statistic and t-statistic that converge exactly to F and t distributions when $W_n = I_m$ and a Series HAC estimator is used. Furthermore, Lazarus et al. (2021) characterize the size-power frontier for Kernel and Series HAC estimators under a fixed-smoothing framework and find that there is little cost to restricting attention to tests which converge exactly to t and F distributions. I therefore focus on verifying that the conditions of Sun (2013) for the GMM estimator with $W_n = I_m$ and $\hat{V}_M$ being a Series HAC estimator, with the test statistic being the standard Wald statistic defined as

$$\mathbb{W}_n = n(\tilde{\beta}_{GMM} - \beta_{0,n})'\hat{V}(\hat{\theta}, \hat{\eta})^{-1}(\tilde{\beta}_{GMM} - \beta_{0,n}),$$

and the t-statistic is defined as

$$t_n = \sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n})/\sqrt{\hat{V}(\hat{\theta}, \hat{\eta})}$$

when $p = 1$. I impose the following conditions in order for this method to be able to conduct valid and standard inference in a fixed-smoothing asymptotic framework.

Assumption 3.2 Suppose that with p, K, and Q fixed, and either J fixed or $J \to \infty$, as $n \to \infty$, we have that

1. For all $\epsilon > 0$, there exists γ_ϵ, such that

   $P(\sup_{r\in[0,1],\theta\in\Theta_n:||\theta-\theta_{0,n}||_1<\gamma_\epsilon} ||\sum_{i=1}^{\lfloor rn\rfloor}(\partial_\beta g_i(\beta, \delta) - \partial_\beta g_i(\beta, \delta_{0,n}))/n||_2 > \epsilon) \to 0.$

2. $K \geq p$ and $\{\phi_k(x)\}_{k=0}^K$ with $\phi_0(x) = 1$ is a sequence of continuously differentiable and orthonormal basis functions in $\mathcal{L}^2[0, 1]$ satisfying $\int_0^1 \phi_k(x)dx = 0$.

3. Uniformly in $r, \lambda_n \in [0, 1]$, $\sum_{i=1}^{\lfloor rn\rfloor} \partial_\beta \hat{g}(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \delta_{0,n})/n - r\partial_\beta g(\theta_{0,n}) \xrightarrow{p} 0$.

4. $V_M^{-1/2} \sum_{i=1}^n \phi_k(\frac{i}{n})\eta_{0,n} g_i(\theta_{0,n})/\sqrt{n} \xrightarrow{d} \xi_k$ jointly for $k = 0, ..., K$ with $\xi_k \sim iidN(0, I_m)$.

Assumptions 3.2.2-3.2.4 contain the conditions imposed by Sun (2013) adjusted to this setting.[4] Assumption 3.1.2 is satisfied for commonly used based functions. For example, $\phi_k(x) = \sqrt{2}\sin(2\pi kx)$ and $\phi_k(x) = \sqrt{2}\cos(2\pi kx)$ satisfy the condition. However, there are basis functions such as $\phi_k(x) = \sqrt{2}\sin(\pi(0.5-k)x)$ that do not satisfy it because it does not satisfy the mean-zero condition. Assumption 3.1.3 is standard in the literature on fixed-smoothing asymptotics (see, for example, Vogelsang and Kiefer (2005)), and easily holds in cases where $\hat{g}$ is linear in θ. Assumption 3.2.4 holds when $\hat{B}(r) = \sum_{i=1}^{\lfloor rn \rfloor} \hat{g}_i(\theta_{0,n})/\sqrt{n}$ is converging weakly to a Gaussian process with almost surely continuous sample paths and independent increments. In Appendix B, I show that this can hold in cases where $g_i(\theta_{0,n})$ is not stationary.

One additional complication that is not present in Sun (2013) or other previous fixed-smoothing results is the plugged-in values of the nuisance parameters $\hat{\hat{\delta}}$ and $\hat{\eta}$. This is why Assumption 3.2.1 is imposed. Assumption 3.2.1 can be viewed as a slightly stronger version of Assumption 2.2.2 and it is serving the same role of bounding how sensitive $\partial_\beta g_i(\theta)$ is to δ. This has sufficient conditions similar to Assumption 2.2.2 and also holds trivially when the cross partial derivatives of the moment conditions are equal to zero. Additionally, similarly to with the adaptivity condition in Lemma 2.1, we want $\hat{V}(\beta_{0,n}, \hat{\delta}, \hat{\eta})$ to be asymptotically equivalent to $\hat{V}(\beta_{0,n}, \delta_{0,n}, \eta_{0,n})$. In order for this to be the case, I impose Assumption 2.1*, which slightly strengthens some of the conditions of Assumption 2.1 to hold with partial sums.

Assumption 2.1* As $n \to \infty$ while p and Q are fixed and either J fixed or $J \to \infty$, we have that

1. For each $q \in \{1, ..., m\}$, $||\partial_\delta E[\sum_{i=1}^t M_{qt}(\theta_{0,n}, \eta_{0,n})/n]||_\infty \to 0$ uniformly over $1 \leq t \leq n$.

2. $g(\theta)$ is twice continuously differentiable on Θ_n and for each $q \in \{1, ..., m\}$, uniformly over $1 \leq t \leq n$,

   $||\partial_\delta \sum_{i=1}^t M_{qi}(\theta_{0,n}, \eta_{0,n})/n - \partial_\delta E[\sum_{i=1}^t M_{qt}(\theta_{0,n}, \eta_{0,n})/n]||_\infty = O_p(\log(J)/\sqrt{n})$.

[4] Here, I imposed the conditions on the original moment conditions evaluated at $\theta_{0,n}$ rather than on the orthogonalized moment conditions used to estimate $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$. In the proof of Proposition 3.1, I verify that this implies that they hold for $\hat{M}(\beta, \hat{\delta}, \hat{\eta})$ under Assumption 2.1.

3. There exists $\epsilon > 0$ such that for each $q \in \{1, ..., Q\}$,

   $\sup_{1 \leq t \leq n, \delta: ||\delta - \delta_{0,n}||_1 \leq \epsilon} \max eig(\partial^2_\delta \sum_{i=1}^t g_{qt}(\beta_{0,n}, \delta)/n) = O_p(\log(J))$.

Under Assumptions 2.1 and 2.1*, when $\hat{V}_M$ is a Series HAC estimator,

$$\hat{V}_M(\beta_{0,n}, \hat{\delta}, \hat{\eta}) - \hat{V}_M(\beta_{0,n}, \delta_{0,n}, \eta_{0,n}) \xrightarrow{p} 0.$$

Since the partial sums $\sum_{i=1}^t \eta g_i(\theta)$ include fewer terms than the full sum $\sum_{i=1}^n \eta g_i(\theta)$, they will generally be smaller asymptotically, so the bounds on partial sums in Assumptions 2.1* will have similar sufficient conditions to before. Together with Assumption 3.2, we can show that the Wald and t statistics will converge to F and t distributions.

Proposition 3.1 Suppose $\tilde{\beta}_{GMM}$ is estimated using equation (5) with $W_n = I_m$ and Assumptions 2.1*, 2.1, 2.2, and 3.2 hold with $K \geq p$. When $n \to \infty$ with K, p, and Q fixed and either J fixed or $J \to \infty$,

$$\frac{K-p+1}{pK} \mathbb{W}_n \xrightarrow{d} F_{p,K-p+1} \text{ and } t_n \xrightarrow{d} t_K \text{ when } p = 1,$$

where $F_{p,K-p+1}$ is an F distribution with $p, K-p+1$ degrees of freedom and t_K is a t-distribution with K degrees of freedom.

Here, I've focused on test statistics that use $\tilde{\beta}_{GMM}$, but because $\tilde{\beta}_{OS}$ has the same asymptotic distribution, Proposition 3.1 holds for the Wald and t statistics that use $\tilde{\beta}_{OS}$ estimated with $W_n = I_m$ if we simply use $\tilde{\beta}_{OS}$ in place of $\tilde{\beta}_{GMM}$ in Assumption 3.2.3. If we were instead to focus on asymptotics with $K \to \infty$ where $V(\theta_{0,n}, \eta_{0,n})$ is consistently estimated, then we have that $\mathbb{W}_n \xrightarrow{d} \chi^2_p/p$, where χ^2_p is chi-squared distribution with p degrees of freedom, and $t_n \xrightarrow{d} N(0,1)$. Note that $F_{p,K-p+1} \xrightarrow{d} \chi^2_p/p$ and $t_K \xrightarrow{d} N(0,1)$ as $K \to \infty$, so critical values obtained from the fixed-smoothing asymptotic results are approximately the same as the critical values from increasing smoothing asymptotic results when K is large. For applications in which increasing smoothing asymptotic results are likely to provide an accurate approximation, consistency for a variety of HAC estimators has been shown under fairly general conditions that include allowing for non-stationarity. For example, suppose that the sequence $\{g_i(\theta_{0,n})\}_{i\in\mathbb{N}}$ is mean-zero, α-mixing, and there exists $\nu > 1$ such that $\sup_i E[||g_i(\theta_{0,n})||_2^\nu] < \infty$ and $\sum_{s=1}^{\infty} s^2\alpha(s)^{\frac{\nu-1}{\nu}} < \infty$ where $\alpha(s)$ are the mixing coefficients. Andrews (1991) shows that a class of Kernel HAC estimators are consistent under these conditions if $n, K \to \infty$ such that $K^2/n \to 0$.[5]

As noted in the literature, non-parametric long-run variance estimators can often be converging to the true long-run variance at a rate faster than $\sqrt{n}$. For Series HAC estimators, Phillips (2005) provides a Mean Squared Error (MSE)-optimal choice of K under stationarity conditions such that $K = O(n^{4/5})$ and a convergence rate of $n^{4/5}$ is achieved. Whereas $K = O(n^{2/3})$ when using the Coverage Probability Error (CPE)-optimal choice of Sun (2013). Andrews (1991) provides rates of convergence for a class of Kernel HAC estimators when the data is weakly stationary. He shows that when using a Quadratic Spectral Kernel with the optimal choice of bandwidth, it is also possible to achieve a $n^{4/5}$ rate of convergence, while other choices still achieve a rate faster than $\sqrt{n}$. However, $\hat{V}(\theta,\eta)$ depends on $\hat{V}_M(\theta,\eta)$ and $\partial_\beta\hat{M}(\theta,\eta)$, and $\partial_\beta\hat{M}(\theta,\eta)$ is usually converging at the standard parametric rate $\sqrt{n}$ when J is fixed. This means that when the penalty function f depends on V and the fixed-smoothing asymptotic results are not relevant, Assumption 3.1 usually holds with $c_n = n^{-1/2}$ when J is fixed, so the estimated penalty function is often converging at the same rate as the sample moment conditions.

Intuitively, choosing the penalty function so that the nuisance parameters are minimizing some estimate of the asymptotic variance seems like it should provide a relatively more efficient estimator, at least when the variance estimator is sufficiently accurate. However, if β is not a scalar, then there may be a trade-off between more precisely estimating different subvectors of β. However, in most applications, either β is a scalar or, if $p > 1$, then the

[5] This follows from Lemma 1 and Theorem 1(a) of Andrews (1991). The class of Kernel estimators includes many commonly used ones such as the truncated, Bartlett, Parzen, Tukey-Hanning, and Quadratic Sprectral kernels.

object of interest is $h(\beta)$ for some known function $h : \mathbb{R}^p \to \mathbb{R}$. When β is a scalar, we can choose $\hat{f}(\theta, \eta) = \hat{V}(\theta, \eta)$. If the object of interest is $h(\beta)$ and h is continuously differentiable, then $\sqrt{n}(h(\tilde{\beta}_{GMM}) - h(\beta_{0,n}))$ is asymptotically normal by the Delta method and we can choose $\hat{f}(\theta, \eta) = \partial_\beta h(\beta)\hat{V}(\theta, \eta)\partial_\beta h(\beta)'$.

In summary, in applications where fixed-smoothing asymptotic results provide a better approximation due to the degree of dependence being high relative to the sample size, I recommend: not having $\hat{f}$ depend on $\hat{V}$, having $W_n = I_m$, and using a Series HAC estimator with K chosen according to Sun (2013). Otherwise, greater efficiency can be achieved by having $\hat{f}$ depend on $\hat{V}$, having $W_n = \hat{V}_M(\hat{\theta}, \hat{\eta})^{-1}$, and having $\hat{V}_M$ be either a Series or Kernel HAC estimator with K chosen to provide the fastest rate of convergence.

4 Synthetic Control Application

First introduced by Abadie and Gardeazabal (2003) and Abadie et al. (2010), SCEs have become a popular choice in contexts with panel data where a single unit becomes and stays treated and there is a large pool of never-treated units which can be used as control units. The method involves constructing a weighted average of the control units or a Synthetic Control (SC) unit by minimizing the difference between this SC and the treated unit on a set of pre-treatment predictor variables, so that this SC can be used as an estimate of the treated unit's counterfactual outcomes in the post-treatment time periods. I focus on analyzing my method in this original context, where there is a single unit that becomes and stays treated. Researchers are most commonly interested in conducting inference on the average treatment effect on the treated unit and the weights on the control units are a nuisance parameter, so I use β to denote the ATT and δ to denote the control weights. When applying my method as an SCE, I refer to my estimator as the Orthogonalized SCE.

4.1 Verifying the Conditions for the Formal Results

I first discuss how to implement the estimator as an SCE and verify the conditions imposed in the formal results above hold when the data follow a linear factor model. Linear factor models, also known as interactive fixed effect models, have been a common setting to explore the properties of SCEs, beginning with Abadie et al. (2010). I index the units $\{0, 1, ..., J\}$ where $i = 0$ is the treated unit and $\mathcal{J} = \{1, ..., J\}$ denotes the set of control units. The control units never receive treatment, whereas the treated unit becomes and stays treated after a known point in time. I denote the sets of indices for time periods prior to its treatment and after its treatment as $\mathcal{T}_0$ with $T_0 = |\mathcal{T}_0|$ and $\mathcal{T}_1$ with $T_1 = |\mathcal{T}_1|$ respectively. Here, the asymptotics are slightly different from before because there are two variables T_0 and T_1 that capture our sample size rather than just n. I focus on verifying the assumptions in the previous sections for $n = \min\{T_0, T_1\}$.

Assumption 4.1 (Linear Factor Model) For all units $i \in \{0, ..., J\}$ and time periods $t \in \mathcal{T}_0 \bigcup \mathcal{T}_1$, outcomes follow a linear factor model with R factors so that

$$Y_{it} = \beta_t d_{it} + f_t \mu_i + \epsilon_{it},$$

where d_{it} is an indicator function equal to 1 if and only if $i = 0$ and $t \in \mathcal{T}_1$ and equal to 0 otherwise. The factor loadings μ_i, dynamic treatment effects β_t, and treatment assignment d_{it} are fixed, but the latent factors f_t and idiosyncratic shocks ϵ_{it} are stochastic.[6]

I let μ denote the $R \times (J + 1)$ matrix of factor loadings with μ_i being its i-th column, f denote the $(T_0 + T_1) \times R$ matrix of realizations of the factors with f_t being is t-th row, and ϵ denote the $(J+1) \times (T_0 + T_1)$ matrix of idiosyncratic shocks. Additionally, I use the subscript $\mathcal{J}$ to denote the sub-matrix for only the units $j \in \mathcal{J}$ and the superscripts *pre* and *post* to denote the sub-matrices for only pre-treatment and post-treatment values respectively. I

[6] I have the treatment effects be fixed because the parameter of interest $\beta_{0,n}$ in the previous sections was assumed to be fixed. However, in the SCE case, generalizing the results to allow $\beta_{0,n}$ to be stochastic is straightforward, in which case confidence intervals can then be interpreted as prediction intervals.

define the average treatment effect on the treated to be $\beta_{0,n} = \sum_{t\in\mathcal{T}_1}\beta_t/T_1$. This means that $\beta_{0,n}$ may be changing as T_1 grows, just as $\beta_{0,n}$ is allowed to change with the sample size as in the previous sections. If the idiosyncratic shocks are mean-zero and the SC has the same factor loadings as the treated unit (i.e., $\mu_0 = \mu_{\mathcal{J}}\delta$), then we can identify $\beta_{0,n}$ using the moment condition $\sum_{t\in\mathcal{T}_1} E[Y_{0t} - \beta - Y'_{\mathcal{J},t}\delta]/T_1 = 0$. Therefore, we want to use this moment condition plus a set of moment conditions that identify the set $D_{0,n} = \{\delta \in \Delta^J : \mu_0 = \mu_{\mathcal{J}}\delta\}$, where $\Delta^J \coloneqq \{\delta \geq 0 : ||\delta||_1 = 1\}$ is the $J-1$-dimensional unit simplex.

Several of the difficulties mentioned earlier can arise in characterizing the asymptotic distribution of the estimated average treatment effect due to its dependence on $\hat{\delta}$. Not only may there be many $\delta \in D_{0,n}$ so δ is partially identified, but analysis of the asymptotic distribution of the control weights is also complicated by the fact that it is often high-dimensional (relative to T_0 and T_1). Lastly, even when J is fixed and δ is point-identified, $\sqrt{T_0}(\hat{\delta} - \delta_{0,n})$ generally has a non-standard asymptotic distribution due to the constraints $\hat{\delta} \in \Delta^J$.[7] Other inference methods, like Cao and Dowd (2019) and Li (2020), also have the limitations of assuming that the control weights are point-identified and treating the number of units as fixed in their asymptotic results. One exception is Zhang et al. (2023) who obtain a $\sqrt{n}$-consistent estimator while allowing δ to be partially identified, and another is Chernozhukov et al. (2024) whose method is discussed further below in subsection 4.3.

Several recent papers have discussed how to estimate the SC using a set of moment conditions, including Fry (2024), Powell (2021), and Shi et al. (2023). A linear instrumental variables approach along the lines of Fry (2024) or Shi et al. (2023) is the most straightforward to verify the conditions of sections 2 and 3. In this case, the moment conditions are

$$g(\beta,\delta) = \begin{pmatrix} \sum_{t\in\mathcal{T}_0} E[Z'_t(Y_{0t} - Y_{\mathcal{J},t}\delta)]/T_0 \\ \sum_{t\in\mathcal{T}_1} E[Y_{0t} - \beta - Y_{\mathcal{J},t}\delta]/T_1 \end{pmatrix} \text{ and } \hat{g}(\beta,\delta) = \begin{pmatrix} Z^{pre}(Y_0^{pre'} - Y_{\mathcal{J}}^{pre'}\delta)/T_0 \\ \sum_{t\in\mathcal{T}_1}(Y_{0t} - \beta - Y_{\mathcal{J},t}\delta)/T_1 \end{pmatrix},$$

[7] This is illustrated by Li (2020) and Fry (2024) who use the method of Andrews (1999) to characterize the asymptotic distribution of the estimated average treatment effect.

where Z^{pre} is a $(Q-1) \times T_0$ matrix of containing the values of $Q-1$ instruments in pre-treatment time periods. In Fry (2024), the vector of instruments is a constant and units which are not included in the set of controls but are also untreated in pre-treatment time periods. The exclusion restriction $E[Z_{qt}(Y_{0t} - \sum_{j=1}^{J} \delta_j Y_{jt})] = 0$ holds for these other units in pre-treatment time periods when the idiosyncratic shocks are uncorrelated across units and the factors are uncorrelated with the idiosyncratic shocks. Intuitively, when the factors are responsible for the covariance across units, we can guarantee that the SC has the same exposure to latent factors as the treated unit by estimating the SC to have the same covariance with other units. In the empirical application below, I provide a practical example for how the set of instruments can be chosen.[8] Other potentially valid choices of instruments exist, such as using lagged values of the outcome variable or using shift-share instruments. Additionally, it may be possible to reframe other estimators, such as the Debiased OLS estimator of Chernozhukov et al. (2024) discussed below, as method of moments estimators, in which case the same orthogonalization technique could be employed.

Following my recommendation from section 2, I set $m = 1$ so η is $1 \times Q$ and there is only a single orthogonal moment condition. This means that $\partial_\beta M(\theta, \eta) = -\eta_Q$ where η_Q is the Q-th element of η. Also, since there is only a single orthogonalized moment condition that is linear in β, both the One-Step and GMM estimators in Section 2 are equivalent to picking the value of β that sets $\hat{M}(\beta, \hat{\delta}, \hat{\eta})$ equal to zero for any positive definite weighting matrix W_n. Therefore, the estimator is simply given by:

$$\tilde{\beta}_{GMM} = \tilde{\beta}_{OS} = \sum_{t \in \mathcal{T}_1} (Y_{0t} - Y_{\mathcal{J},t}\hat{\delta})/T_1 + \sum_{q=1}^{Q-1} \hat{\eta}_q / \hat{\eta}_Q \sum_{t \in \mathcal{T}_0} Z_{qt}(Y_{0t} - Y_{\mathcal{J},t}\hat{\delta})/T_0.$$

Under Assumptions 4.2 and 4.3 below, the l-th row and q-th column of the asymptotic

[8] Fry (2024) also provides several model selection methods for splitting untreated units into a set of controls and set of instruments. However, because of the potential problems for inference that using a data-driven model selection procedure may introduce, I focus on cases where the units used as instruments are only the units which are known to not be valid controls but plausibly valid instruments. I discuss how this can be done when discussing the empirical application below.

variance of the original moment conditions V_g is given by

$$\lim_{T_0,T_1\to\infty} \min\{T_0,T_1\} \sum_{t\in\mathcal{T}_{I_{l=Q}}} \sum_{s\in\mathcal{T}_{I_{q=Q}}} E[g_{l,t}(\theta_{0,n})g_{q,s}(\theta_{0,n})']/(|\mathcal{T}_{I_{l=Q}}\mathcal{T}_{I_{q=Q}}|), \tag{10}$$

where $I_{q=Q}$ is an indicator function that is equal to 1 if and only if $q = Q$. I define $\hat{V}_M$ to be a Series HAC estimator and $\hat{V}(\hat{\theta},\hat{\eta}) = \hat{V}_M(\hat{\theta},\hat{\eta})/\hat{\eta}_Q^2$. As mentioned before, fixed-smoothing asymptotics provide a better approximation for SCE applications because of the small sample sizes and high degree of temporal dependence in the data. Furthermore, the values of the smoothing parameter K are often small in applications. For example, in the empirical application below, K is equal to 4 when the method of Sun (2013) is used. Therefore, I use the method suggested in Section 3 for such cases by not using $\hat{V}$ when constructing the penalty function $\hat{f}$. Also, for testing a null hypothesis $H_{0,n} : \beta_{0,n} = \bar{\beta}$, I use the test statistic $\sqrt{\min\{T_0,T_1\}}(\tilde{\beta}_{GMM} - \bar{\beta})/\sqrt{\hat{V}(\hat{\theta},\hat{\eta})}$ and critical values from a t distribution with K degrees of freedom.

Even if the penalty function does not depend on the long-run variance estimator, we can still choose the penalty function to be equal to an upper bound of the asymptotic variance as a function of the nuisance parameters, where the upper bound is tight in special cases. Because $\hat{g}$ is linear in δ, each of the elements of the asymptotic variance of the original moment conditions V_g involves a quadratic form $\delta'\Omega\delta$ for some positive definite matrix Ω. Also, since $V(\theta,\eta) = \eta V_g(\theta)\eta'/\eta_Q^2$ where $V_g(\theta)$ is positive definite for any fixed $\theta \in \Theta_n$, choosing $\hat{f}(\theta,\eta) = f(\theta,\eta) = ||\delta||_2^2 + ||\eta||_2^2/\eta_Q^2$ minimizes an upper bound on $V(\theta,\eta)$ and does not involve $\hat{V}(\theta,\eta)$. However, $||\eta||_2^2/\eta_Q^2 = \sum_{q=1}^{Q-1}(\eta_q/\eta_Q)^2 + 1$ may not have a unique minimum on

$$H_{0,n} = \{\eta \in \mathbb{R}^Q : \eta \begin{pmatrix} E[Z^{pre}\lambda^{pre}\mu_{\mathcal{J}}]/T_0 \\ \sum_{t\in\mathcal{T}_1} E[\lambda_t\mu_{\mathcal{J}}]/T_1 \end{pmatrix} = 0\}.$$

Therefore, I normalize $\eta_Q = 1$. I show in the proof of Proposition 4.1 that this allows $\eta_{0,n}$ to be identified. Then the penalty can simply be set to $\hat{f}(\theta,\eta) = f(\theta,\eta) = ||\delta||_2^2 + ||\eta||_2^2$ and have the parameter space for η be equal to $H = \{\eta \in \mathbb{R}^Q : \eta_Q = 1\}$. I mention when the upper bound being minimized is tight below when discussing Assumption 4.3. In order for Assumption 3.1 to be satisfied, I impose the following conditions:

Assumption 4.2 As $T_0, T_1, J \to \infty$ while Q is fixed,

1. $E[Z_t\epsilon_{it}] = 0$ for all i, t and $\max_{0\leq i\leq J}\{|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}||Z_t\epsilon_{it}||_1\} = O_p(\log(J)/\sqrt{T_0})$.

2. $Z^{pre'}f^{pre}/T_0 = E[Z^{pre'}f^{pre}/T_0] + O_p(1/\sqrt{T_0})$.

3. $E[Z^{pre'}f^{pre}/T_0]'E[Z^{pre'}f^{pre}/T_0] \to \Omega_0$ and $E[\sum_{t\in\mathcal{T}_1} f_t/T_1] \to \Omega_1$ where Ω_0 is full rank. The sequence of factor loadings μ_i is uniformly bounded.

4. $\max_{0\leq i\leq J}\{|\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\epsilon_{it}|\} = O_p(\log(J)/\sqrt{T_1})$ and $\sum_{t\in\mathcal{T}_1} f_t/T_1 = E[\sum_{t\in\mathcal{T}_1} f_t/T_1]+O_p(1/\sqrt{T_1})$.

5. $D_{0,n}$ is non-empty for all $J > C$ for some integer $C \geq 1$.

Assumption 4.2.1 ensures that the instruments satisfy the exclusion restriction, and Assumption 4.2.3 guarantees that the instruments are relevant and that there are enough of them to identify $D_{0,n}$. Furthermore, following the reasoning discussed in Remark 3.1, by setting $D_n = \Delta^J$, the rate of convergence conditions in Assumption 4.2 guarantee that the sample moment conditions converge to the population moment conditions at a rate of $\log(J)/\sqrt{\min\{T_0, T_1}$ uniformly in $\delta \in D_n$. Assumption 4.2.5, imposes that once enough control units have been added, the factor loadings of the treated unit μ_0 fall in the convex hull of the factor loadings of the control units. Note that the factors may not have the same average values before and after treatment, which allows for cases where there is a dependence between the values of the latent factors and the timing of treatment. In order for $\sqrt{n}\hat{M}(\beta_{0,n}, \delta_{0,n}, \eta_{0,n}) = \sqrt{\min\{T_0, T_1\}}\eta_{0,n}\hat{g}(\beta_{0,n}, \delta_{0,n})$ to be asymptotically normal and for the other conditions of Propositions 2.1 and 3.1 to hold, I impose the following additional assumption:

Assumption 4.3 As $T_0, T_1, J \to \infty$ while Q and K are fixed,

1. For $q, l \in \{1, ..., Q-1\}$, $E[(Z_q^{pre}\epsilon^{pre})'(Z_l^{pre}\epsilon^{pre})/T_0] \to \Sigma^{q,l}$ and $E[\sum_{t\in\mathcal{T}_1} \epsilon_{it}/\sqrt{T_1})^2] \to \Sigma_i$ for each $i \in \{0, ..., J\}$, where $\Sigma^{q,l}$ are $(J+1)\times(J+1)$ non-stochastic positive definite matrices and Σ_i are positive constants.

2. $\{\phi_k(x)\}_{k=1}^K$ are chosen to satisfy Assumption 3.2.2.

3. There exists a set $\mathcal{P} \subset \mathbb{N}$ with $|\mathcal{P}| < \infty$ and $\sum_{j\in\mathcal{P}} \delta_{0,n,j} = 1$ for all n, where $\delta_{0,n,j}$ is the j-th element of $\delta_{0,n}$.

4. The sequence $(Z_t, \epsilon_t)_{t\in\mathbb{N}}$ is α-mixing with mixing coefficients $\alpha(t)$.

5. There exists $\gamma > 2$ such that $\sup_{t\in\mathcal{T}_0} E[||Z_t(\epsilon_0 - \epsilon_{\mathcal{J},t})\delta_{0,n}]||^\gamma]$ and $\sup_{t\in\mathcal{T}_1} E[|\epsilon_{0t} - \epsilon_{\mathcal{J},t}\delta_{0,n}|^\gamma]$ are bounded and $\sum_{t\in\mathbb{N}} \alpha(t)^{1-2/\gamma} < \infty$.

Assumption 4.3.3 imposes that the optimal control weights are sparse. This sparsity condition is useful for verifying the identification condition in Assumption 3.1.5. It is also useful for obtaining the asymptotic normality of $\hat{g}(\beta_{0,n}, \delta_{0,n})$, since then it is sufficient to have asymptotic normality of sample averages involving the sparse set of control units that receive positive weight. In this context, sparsity of the optimal control weights is plausible for two reasons. First, because there are only R latent factors, it is plausible that there exists $\delta \in \Delta^J$ such that $\mu_0 = \mu_{\mathcal{J}}\delta$ and $||\delta||_0$ is equal to or only slightly larger than R. Second, the simplex constraints have a tendency to select such sparse solutions, which is why most empirical applications with many controls find sparse weights. Assumption 4.3.1 helps us specify what the asymptotic variance of the sample moment conditions is. Along with the mixing condition and requiring a bound on the moments of $\hat{g}(\beta_{0,n}, \delta_{0,n})$, this is sufficient to apply a Functional Central Limit Theorem to the partial sums of the sample moment conditions, which allows for the estimator to be asymptotically normal and our t-statistic to have a t-distribution asymptotically. Also note that the upper bound of the asymptotic variance being minimized is tight in the special case where $\Sigma^{q,l} = I_{J+1}$ for all $q, s \in \{1, ..., Q\}$ and Σ_i is constant across units. This means that we should generally expect the upper bound to be closer to being tight when idiosyncratic shocks are homogeneous across units and the instruments are not redundant.

Note that for $\delta_{0,n} \in D_{0,n}$,

$$\hat{g}(\beta_{0,n}, \delta_{0,n}) = \begin{pmatrix} \sum_{t \in \mathcal{T}_0} Z_t(\epsilon_{0t} - \epsilon'_{\mathcal{J},t}\delta_{0,n})/T_0 \\ \sum_{t \in \mathcal{T}_1}(\epsilon_{0t} - \epsilon'_{\mathcal{J},t}\delta_{0,n})/T_1 \end{pmatrix} = \begin{pmatrix} \sum_{t \in \mathcal{T}_0} Z_t(\epsilon_{0t} - \sum_{j \in \mathcal{P}} \epsilon_{jt}\delta_{0,n,j})/T_0 \\ \sum_{t \in \mathcal{T}_1}(\epsilon_{0t} - \sum_{j \in \mathcal{P}} \epsilon_{j,t}\delta_{0,n,j})/T_1 \end{pmatrix}.$$

Therefore, using the adaptivity condition of Lemma 2.1,

$$\tilde{\beta}_{GMM} = \tilde{\beta}_{OS} = \sum_{t \in \mathcal{T}_1}(\epsilon_{0t} - \sum_{j \in \mathcal{P}} \epsilon_{j,t}\delta_{0,n,j})/T_1 + \sum_{q=1}^{Q-1} \eta_{0,n,q} \sum_{t \in \mathcal{T}_0} Z_{qt}(\epsilon_{0t} - \sum_{j \in \mathcal{P}} \epsilon_{jt}\delta_{0,n,j})/T_0 + o_p(1).$$

As a result, the asymptotic distribution of the estimator depends only on a finite number of averages of the idiosyncratic shocks and averages of the product of the idiosyncratic shocks with the instruments. When T_0 is large relative to T_1, the uncertainty in $\tilde{\beta}_{GMM}$ comes from the idiosyncratic shocks in the post-treatment time periods, whereas when T_1 is large relative to T_0 the uncertainty in $\tilde{\beta}_{GMM}$ comes from the instruments and idiosyncratic shocks in the pre-treatment time periods.

Proposition 4.1 Under Assumptions 4.1 and 4.2 for the estimator described above, the conditions of Assumption 3.1 are satisfied with $\gamma_1 = 2$, $\gamma_2 = 1$, $a_n = \log(J)/\sqrt{\min\{T_0, T_1\}}$, $b_n = \log(J)/\sqrt{\min\{T_0, T_1\}}$, and $c_n = 0$. Furthermore, if Assumption 4.3 also holds, λ_δ and λ_η from equation (8) satisfy $\max\{\lambda_\delta, \lambda_\eta\}^{1/2}\log(J) \to 0$, $\min\{\lambda_\delta, \lambda_\eta\}\sqrt{\min\{T_0, T_1\}/\log(J)} \to \infty$, and $T_1/T_0 \to a$ for some $a > 0$ as $T_0, T_1 \to \infty$, then the conditions of Propositions 2.1 and 3.1 hold. More specifically, we have that

$$\sqrt{\min\{T_0, T_1\}}(\tilde{\beta}_{GMM} - \beta_{0,n})/\sqrt{V} \xrightarrow{d} N(0,1) \text{ and}$$

$$\sqrt{\min\{T_0, T_1\}}(\tilde{\beta}_{GMM} - \beta_{0,n})/\sqrt{\hat{V}(\hat{\theta}, \hat{\eta})} \xrightarrow{d} t_K.$$

For empirical applications, I use an algorithm for selecting λ_δ and λ_η so that

$$\lambda_\delta, \lambda_\eta = O_p(\log(J)\log(\min\{T_0, T_1\})/\sqrt{\min\{T_0, T_1\}}).$$

Therefore, the condition on λ_δ and λ_η in Proposition 4.1 is satisfied as long as

$$\log(J)^2 \log(\min\{T_0, T_1\})/\sqrt{\min\{T_0, T_1\}} \to 0.$$

This allows for J to be growing significantly faster than T_0 and T_1.

4.2 Empirical Application

I now explore how my method works in practice by replicating the work of Andersson (2019) and then examining its performance in simulations fitted to their data. Andersson (2019) evaluate the impact of Sweden's Carbon Tax on CO2 emissions from transport per capita in the country. The carbon tax was introduced at US \$30 per ton of CO2 in 1990 and increased slightly during the 1990s to US \$44 in 2000. Then, from 2001 to 2004, the rate was increased to US \$109, and as of 2023 it is around \$125. When the carbon tax was implemented, it complemented an existing energy tax and there was also an addition of a Value-Added-Tax (VAT) of 25 percent in 1990. The primary treatment effect they are interested in is the combined effect of the carbon tax and VAT starting in 1990 on CO2 per capita. While we could view this as a case with two continuous treatment variables (the Carbon tax rate and the VAT tax rate), if we are only interested in estimating the average difference between actual CO2 per capita and CO2 per capita without the Carbon tax and VAT, then we can view the Carbon and VAT together as a single binary treatment.

Andersson (2019) uses pre-treatment time periods of 1960 through 1989 and the post-treatment periods of 1990 through 2005. The set of units they use as controls are the 14 OECD countries: Australia, Belgium, Canada, Denmark, France, Greece, Iceland, Japan, New Zealand, Poland, Portugal, Spain, Switzerland, and United States. They arrived at this set of 14 control countries by starting with 24 other OECD countries for which data was available. They excluded 9 of these countries: Ireland, Finland, Norway, Netherlands, Germany, Italy, United Kingdom, Austria, Turkey, and Luxembourg. In the case of Finland, Norway, the Netherlands, Germany, Italy, and the United Kingdom, their justification was based on these countries implementing a Carbon tax or making significant changes to their fuel taxes. Because these units were experiencing similar interventions, using them as controls could lead us to underestimate the effect of the policies in Sweden. However, due in part to Sweden being one of the first countries to adopt a Carbon tax, all these other interventions happened in the post-treatment years of 1990 to 2005. Since the post-treatment data for the instruments are not used, these policy changes do not necessarily present a problem for using these countries' CO2 per capita from transport data as instruments. In the case of Austria and Luxembourg, their justification for excluding them was based on concerns of "fuel tourism". They exclude Turkey because its CO2 emissions data was significantly different from the other OECD countries throughout the entire sample, and they exclude Ireland based on economic shocks that happened in the post-treatment time periods which did not also occur in Sweden.[9] As discussed in Fry (2024), the key assumption for a country to be a valid instrument unit is that the factors playing a role in determining both its CO2 emissions from transport and Sweden's CO2 emissions from transport should be the same as the factors playing a role in determining both its CO2 emissions from transport and CO2 emissions from transport in the control countries, in the pre-treatment time periods. I estimate the Orthogonalized SCE using the same set of controls as Andersson (2019) and the set of instruments being Ireland, Finland, Norway, the Netherlands, Germany, Italy, the United Kingdom, and a constant, although I find similar results when also excluding Ireland from the set of instruments.

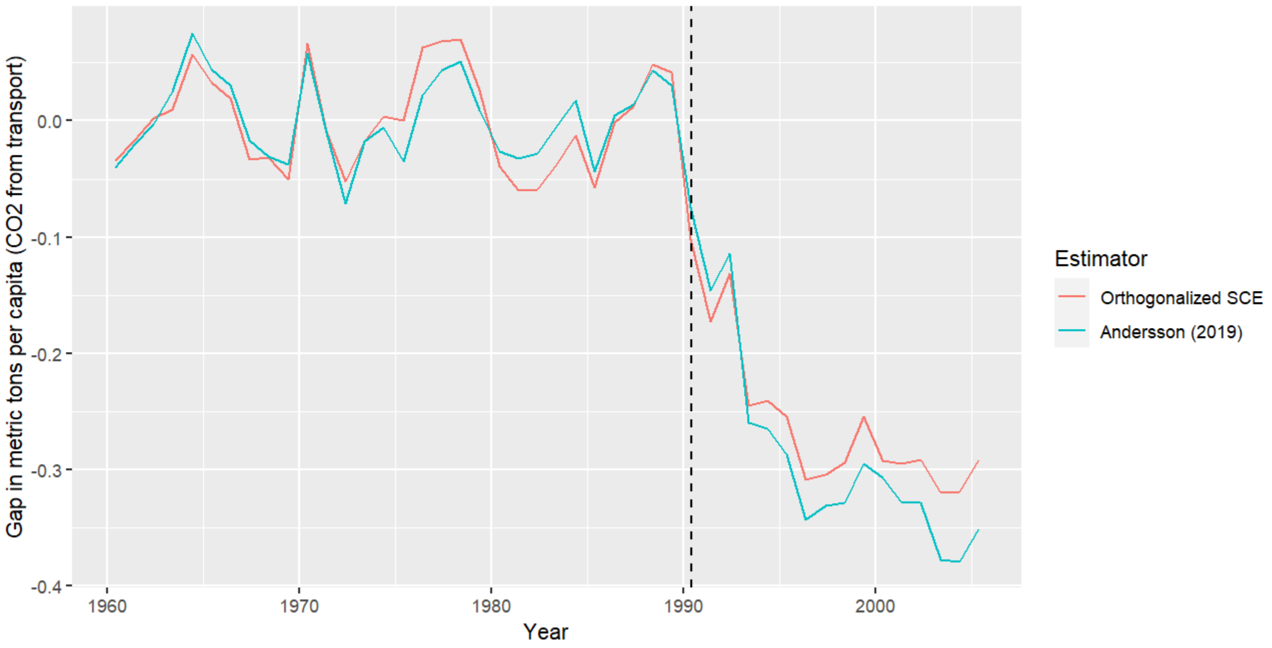
In the main specification of Andersson (2019), their predictor variables are CO2 from transport per capita in 1970, 1980, and 1989, as well as GDP per capita, motor vehicles

[9] More specifically, they cite the Celtic Tiger expansion period in Ireland.

(per 1,000 people), gasoline consumption per capita, and urban population averaged for the period 1980–1989. They weight these predictors using the approach of Abadie et al. (2010). As is common when the SC's weights are constrained to be a convex combination of the control units, the control weights they find end up being rather sparse with only 6 of the 14 control countries included receiving weight greater than 1%: Belgium (0.195), New Zealand (0.177), Denmark (0.384), Greece (0.090), Switzerland (0.061), and the United States (0.088). Using this SC as their counterfactual, they estimate an effect of -0.29 metric tons of CO2 emissions per capita in an average year, which is a 10.9 percent reduction, for the 1990–2005 period. Aggregating over the total population and the 1990-2005 period, the total cumulative reduction in emissions for the post-treatment period is 40.5 million tons of CO2. They perform several placebo tests and robustness checks, including the popular placebo test of Abadie and Gardeazabal (2003) and Abadie et al. (2010). This method involves additionally estimating a synthetic unit for every control unit using the other control units. A test statistic is then constructed for the treated unit and every control unit by calculating the MSPE (mean squared prediction error) of each synthetic unit in the post-treatment time periods, and then either dividing by the synthetic unit's pre-treatment MSPE or excluding certain synthetic units with especially large pre-treatment MSPE. A p-value can then be calculated by looking at what quantile the treated unit's test statistic falls in. In Andersson (2019), when they exclude the synthetic units with a pre-treatment MSPE at least 20 times larger than Synthetic Sweden's pre-treatment MSPE, it leaves 9 control countries. The gap in emissions for Sweden in the post-treatment period is the largest of all remaining countries, giving a p-value of $1/10 = .1$. When using the ratio of post-treatment MSPE to pre-treatment MSPE, the p-value is $1/15 = .067$. In both cases, the test statistic is the most extreme for the actually treated unit, but the p-value fails to fall below the most common thresholds for statistical significance because of the small number of control units.

When re-estimating the average treatment effect, I use the same set of pre-treatment time periods and post-treatment time periods. The weights of Synthetic Sweden are somewhat

Figure 1: Gap between Sweden and Synthetic Sweden



less sparse than before with the countries receiving positive weight being: Australia (0.087), Belgium (0.113), Denmark (0.322), Greece (0.089), Japan (0.0190), New Zealand (0.105), Switzerland (0.201), and the United States (0.064). It is unsurprising that the weights are still sparse but with slightly more countries receiving non-zero weight since the penalty function encourages the weights to be more spread out but the simplex constraints are still imposed. That said, the weights are quite similar, with all 6 countries that received positive weight before still receiving positive weight and Denmark still receiving the most weight. The weights on the moment conditions $\hat{\eta}$ are fairly spread out across the pre-treatment moment conditions with the largest weight being placed on the one using the United Kingdom as an instrument and the smallest weight being placed on the one using the Netherlands as an instrument.[10] The estimated average effect of the Carbon tax and VAT from 1990 to 2005 is a decrease of 0.26 metric tons of CO2 per capita per year, which is similar to the estimate of Andersson (2019). Where the method introduced here allows for a more substantial

[10] More specifically, the weights on the pre-treatment moment conditions are: Finland (-6.166), Germany (-10.590), Ireland (7.883), Italy (7.912), Netherlands (-0.662), Norway (11.617), United Kingdom (-22.402), and the constant (17.882).

difference is in terms of inference. Using the t-test described above, the p-value for the null hypothesis that these taxes had no average effect on CO2 emissions from transport per capita in the post-treatment time periods is 0.0067.[11] This supports the results of the original paper, by showing that under plausible assumptions, the results would be very surprising if these policies had no effect on average. However, in addition to the method of Abadie et al. (2010) not necessarily providing statistically significant results, if the inference methods of Chernozhukov et al. (2021) or Cao and Dowd (2019) discussed below are used, we calculate p-values greater than .1 causing us to fail to reject the null of no effect at common levels of statistical significance.[12] This may be due to these methods having lower power in this case.

4.3 Simulations Based on the Empirical Application

Several other methods of inference for SCEs have recently been proposed in addition to the method of Abadie et al. (2010). I discuss the differences between these methods and the t-test using the Orthogonalized SCE and then compare their performance in simulations. I also compare the bias and MSE of the Orthogonalized SCE with existing variations of the SCE. I conduct the simulations by fitting a linear factor model to the pre-treatment CO2 emissions from transport per capita data from Andersson (2019). I first estimate the number of factors using the Singular Value Thresholding method of Gavish and Donoho (2014).[13] Using this method, I estimate that there are five factors. I then estimate the factor loadings and factor realizations using Principal Components Analysis. In order to allow the number of time periods to vary, I fit the estimated values of the factors $\hat{f}$ and the residuals $\hat{\epsilon}_{it} = Y_{it} - \sum_{r=1}^{5} \hat{f}_{tr}\hat{\mu}_{ri}$ to models and use these models for a parametric bootstrap.

[11] When estimating the long-run variance of the moment conditions, the smoothing parameter K is estimated to be 4, illustrating the empirical relevance of the fixed-smoothing asymptotics. For the series $\phi_k(x)$, I choose $\phi_k(x) = \sqrt{2}\sin(2\pi xk)$ for even k and $\phi_k(x) = \sqrt{2}\cos(2\pi xk)$ for odd k. Therefore, Assumption 3.2.2 is satisfied.

[12] Using the subsampling method with 300 iterations and a subsample size of 10 gives a p-value of .14. Using the conformal inference method with moving block permutations gives a p-value of 0.39. Using the End-of-Sample Instability test gives a p-value of 0.43. Using the t-test cross-fitting method of Chernozhukov et al. (2024) with $K = 3$ gives a p-value of .009.

[13] Here, the number of factors is chosen to be equal to the number of singular values greater than the median singular value times 2.858.

For the factors, I fit each factor to an ARIMA model.[14] For the idiosyncratic shocks, I sample them from mean-zero normal distributions that are independent across both unit and time. I also set each of the variances of the idiosyncratic shocks to be the same over time but allow for heteroscedasticity across units by using the sample variance of $\{\hat{\epsilon}_{it}\}_{t=1960}^{1989}$ for each i, which is consistent with my formal results. The factor loadings are fixed across the simulation draws.

For the bias and MSE, I compare the Orthogonalized SCE to the estimators of Chernozhukov et al. (2024) as well as an OLS version of the SCE which chooses the control weights to minimize pre-treatment MSE. The OLS-SCE is the SCE analyzed by Ferman (2021), Ferman and Pinto (2021), Li (2020), and others.[15] I also compare the bias and MSE of my estimator to a "Naive IV-SCE", which is an estimator that uses the same control weights estimated using the instruments, but does not perform the orthogonalization step. In other words, it simply estimates the ATT by calculating the average difference between the treated unit and the SC in the post-treatment time periods. For inference, I compare my t-test procedure to several other methods that have recently been proposed in the literature.

The method that is most comparable to mine is the t-test cross-fitting procedure of Chernozhukov et al. (2024), where the pre-treatment time periods are split into K blocks and control weights $\hat{\delta}_k$ are estimated using OLS withholding the k-th block, H_k. This handles the bias in estimating the control weights with OLS by subtracting $\sum_{t \in H_k}(Y_{0t} - Y'_{\mathcal{J},t}\hat{\delta}_k)/|H_k|$ from $\sum_{t \in \mathcal{T}_1}(Y_{0t} - Y'_{\mathcal{J},t}\hat{\delta}_k)/T_1$, giving K different estimates $\hat{\beta}_k$ which can be averaged: $\hat{\beta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k$. Their test statistic then relies on standardizing $\hat{\beta}$ using the variation across the estimates $\hat{\beta}_k$. Following the suggestion of Chernozhukov et al. (2024), I use $K = 3$ in the simulations. While Ferman and Pinto (2021) show that using OLS to estimate δ can lead the SCE to be asymptotically biased, Chernozhukov et al. (2024)'s debiasing approach can fix this while avoiding the need for instruments. However, it relies on the bias being constant over

[14] This is done using the auto.arima() function in the forecast package in R, where AIC is used for model selection and QMLE is used to estimate the parameters.

[15] As noted by Kaul et al. (2021), for this choice of predictors, this is also an optimal solution to Abadie et al. (2010)'s algorithm for weighting predictors, one of the most commonly used methods in practice.

time. One similarity between our approaches is that their test statistic asymptotically follows a t-distribution but does not require consistently estimating the asymptotic variance of the estimated ATT. Chernozhukov et al. (2024) prove that their estimator is asymptotically normal, despite not orthogonalizing with respect to the weights. However, Chernozhukov et al. (2024)'s result relies on each of the sets of control weights $\hat{\delta}_k$ being approximately independent of the shocks that occur in $\mathcal{T}_1$ and H_k and the bias of the OLS estimator being the same over time. In cases where the timing of treatment is influenced by the values of the latent factors, we would usually expect the bias in the post-treatment time periods to be different from in the pre-treatment time periods. Using Neyman orthogonalization allows us to achieve asymptotic normality without having to impose these conditions.

Li (2020) propose a subsampling method that they show has asymptotically correct size when both T_0 and T_1 are large. However, they use an I.I.D. subsampling method rather than a block subsampling method which requires stronger independence conditions and Andrews and Guggenberger (2010) show that subsampling and m out of n bootstrap methods can have incorrect asymptotic size when the parameter is close to the boundary of the parameter space. Also, they use the OLS-SCE estimator, which, as mentioned, can be asymptotically biased in the factor model setting. Chernozhukov et al. (2021) provide a method for conformal inference that can be used with different SCEs provided that the estimator satisfies certain consistency conditions. For the simulations, I use the moving block version of the conformal inference method in order to better deal with the temporal dependence of the observations. The End-of-Sample Instability test was originally introduced by Andrews (2003), suggested for SCE by Hahn and Shi (2017), and formally analyzed and extended by Cao and Dowd (2019). It involves testing for a structural break in the sequence of differences between the treated unit and SC. For the placebo method of Abadie et al. (2010), I include the version that excludes the synthetic units with a pre-treatment MSPE at least 20 times larger than the treated unit's pre-treatment MSPE. Using the version that uses the ratio of post-treatment MSPE to pre-treatment MSPE provides similar results, except has higher power when $\alpha = .1$.[16]

Table 1: Main Estimation Accuracy Results

	Naive GMM-SCE	Orthogonalized GMM-SCE	OLS-SCE	Debiased OLS-SCE
Bias Magnitude				
$T_0 = 30, T_1 = 16$	0.037	0.010	0.099	0.099
$T_0 = 60, T_1 = 16$	0.014	0.002	0.093	0.047
$T_0 = 30, T_1 = 32$	0.032	0.013	0.131	0.136
$T_0 = 60, T_1 = 32$	0.022	0.002	0.107	0.089
$\hat{\beta}$ **MSE**				
$T_0 = 30, T_1 = 16$	0.043	0.023	0.018	0.017
$T_0 = 60, T_1 = 16$	0.038	0.011	0.015	0.007
$T_0 = 30, T_1 = 32$	0.039	0.033	0.028	0.027
$T_0 = 60, T_1 = 32$	0.041	0.010	0.019	0.013

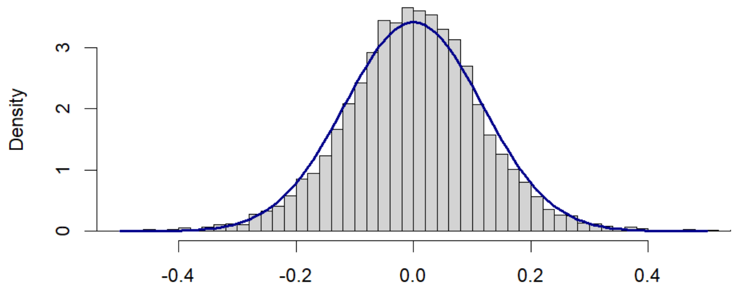
Notes: All simulations are conducted with a thousand replications.

Chernozhukov et al. (2021)'s and Cao and Dowd (2019)'s methods require stronger conditions on the idiosyncratic shocks, but they both have the potential advantage that the sizes of their tests are asymptotically correct when T_1 is fixed and only $T_0 \to \infty$. This suggests that they may be preferable when T_1 is very small. On the other hand, these methods and the placebo method are designed to test the sharp null of no effect in every post-treatment time period, rather than testing a null hypothesis about the ATT. While it depends on context, usually testing the sharp null hypothesis is of less interest. When conducting the simulations for the sizes of the tests, $\beta_t = 0$ for all $t \in \mathcal{T}_1$, so both the sharp null and the null hypothesis of $\beta_{0,n} = 0$ are true.

Table 1 includes the results for the bias and MSE of the OLS-SCE, the Debiased OLS method of Chernozhukov et al. (2024), the Orthogonalized SCE, and the naive IV-SCE which skips the orthogonalization step. Although the primary purpose of the orthogonalization step is to achieve asymptotic normality, we can see that in these simulations the orthogonalized version of the estimator also has lower bias and MSE than the naive version. While the Debiased OLS method does start to provide lower bias than the regular OLS estimator as T_0

[16] For this inference method in the simulations, I also estimate the weights using OLS.

Figure 2: Normality of the Orthogonalized SCE when $T_0 = 30$ and $T_1 = 16$



grows, the bias is still notable, indicting that the assumption of constant bias across the pre-treatment and post-treatment time periods fails to hold. Overall, the Orthogonalized SCE provides the lowest bias and the MSE results are more mixed, with the Debiased OLS-SCE usually providing slightly lower MSE than the Orthogonalized SCE.

Figure 2 shows the histogram for estimates of $\beta_{0,n}$ for the Orthogonalized SCE from 10000 replications when the sample size is the same as in Andersson (2019). We can see that even with this relatively small sample size, the normal approximation holds relatively well with only slightly greater concentration near zero and slightly more outliers than would be expected.

Table 2 contains the size results for the inference methods discussed above when the nominal size is .1, .05, and .01. It is worth noting that the End-of-Sample Instability test and subsampling method are more computationally intensive than the conformal inference method and two t-test methods.[17] Looking at the results, we can see that the rejection rates for the Orthogonalized SCE are all below the nominal levels, even when there are only four post-treatment time periods, indicting that it is succeeding in controlling size. As T_0 increases, it tends to become slightly more conservative. The conformal inference method is similarly only experiencing very small over-rejection when $T_0 = 30$ and $T_1 = 16$

[17] The average runtime of the subsampling method is around 10 times longer and the End-of-Sample Instability test is around 100 times longer than that the average runtime of the Orthogonalized SCE t-test.

Table 2: Main Size Results

	Orthogonalized SCE t-test	End-of-Sample Instability test	Conformal Inference	Cross-Fitting t-test	Subsampling Method	Placebo Method
Rejection Rates with $\alpha = .1$						
$T_0 = 30, T_1 = 4$	0.086	0.228	0.101	0.265	0.282	0.002
$T_0 = 30, T_1 = 16$	0.061	0.158	0.109	0.378	0.363	0.002
$T_0 = 60, T_1 = 16$	0.051	0.202	0.089	0.413	0.378	0.000
$T_0 = 30, T_1 = 32$	0.073	0.208	0.091	0.520	0.482	0.001
$T_0 = 60, T_1 = 32$	0.042	0.188	0.085	0.356	0.466	0.000
Rejection Rates with $\alpha = .05$						
$T_0 = 30, T_1 = 4$	0.045	0.190	0.033	0.164	0.212	0.000
$T_0 = 30, T_1 = 16$	0.030	0.111	0.065	0.237	0.259	0.000
$T_0 = 60, T_1 = 16$	0.023	0.162	0.053	0.267	0.287	0.000
$T_0 = 30, T_1 = 32$	0.038	0.175	0.044	0.346	0.384	0.000
$T_0 = 60, T_1 = 32$	0.027	0.159	0.035	0.227	0.357	0.000
Rejection Rates with $\alpha = .01$						
$T_0 = 30, T_1 = 4$	0.016	0.128	0.000	0.036	0.120	0.000
$T_0 = 30, T_1 = 16$	0.008	0.072	0.000	0.065	0.130	0.000
$T_0 = 60, T_1 = 16$	0.003	0.111	0.000	0.071	0.137	0.000
$T_0 = 30, T_1 = 32$	0.007	0.121	0.000	0.10	0.227	0.000
$T_0 = 60, T_1 = 32$	0.006	0.118	0.000	0.074	0.198	0.000

Notes: All simulations are conducted with a thousand replications.

or $T_1 = 4$ and then also becomes more conservative as the sample size grows. The placebo method never calculates a p-value below .05 by construction because J is too small, and is also very conservative when $\alpha = .1$. On the other hand, the cross-fitting t-test method and the subsampling method have significant over-rejection even as T_0 and T_1 grow. The over-rejection is smaller but still quite notable for the End-of-Sample Instability test.

Table 3 contains the power for the same set of inference methods. The simulations are done under the same conditions as before, but now under the alternative hypothesis of $\beta_t = -.5$ in each post-treatment time period so $\beta_{0,n} = -.5$. In Appendix C, I also include the size-adjusted power. The power is adjusted for size by finding the threshold for the p-value that makes the method's actual size equal to the nominal size under the null, and then seeing how often the p-value falls below this threshold under the alternative. While this adjustment is infeasible to do in application, it is useful for comparing the power of inference methods with very different sizes. The conformal inference method and the placebo method generally have the lowest power, although the power of the conformal inference method

Table 3: Main Power Results

	Orthogonalized SCE t-test	End-of-Sample Instability test	Conformal Inference	Cross-Fitting t-test	Subsampling Method	Placebo Method
Rejection Rates with $\alpha = .1$						
$T_0 = 30, T_1 = 4$	0.808	0.849	0.696	0.975	0.282	0.168
$T_0 = 30, T_1 = 16$	0.796	0.827	0.304	0.968	1.000	0.458
$T_0 = 60, T_1 = 16$	0.902	0.889	0.655	1.000	1.000	0.020
$T_0 = 30, T_1 = 32$	0.633	0.855	0.008	0.946	0.996	0.232
$T_0 = 60, T_1 = 32$	0.900	0.888	0.034	0.970	0.998	0.000
Rejection Rates with $\alpha = .05$						
$T_0 = 30, T_1 = 4$	0.644	0.816	0.367	0.897	0.212	0.000
$T_0 = 30, T_1 = 16$	0.634	0.796	0.160	0.862	0.999	0.000
$T_0 = 60, T_1 = 16$	0.788	0.860	0.366	0.996	1.000	0.000
$T_0 = 30, T_1 = 32$	0.466	0.829	0.002	0.825	0.990	0.000
$T_0 = 60, T_1 = 32$	0.775	0.863	0.006	0.880	0.997	0.000
Rejection Rates with $\alpha = .01$						
$T_0 = 30, T_1 = 4$	0.306	0.754	0.000	0.477	0.120	0.000
$T_0 = 30, T_1 = 16$	0.259	0.746	0.000	0.451	0.995	0.000
$T_0 = 60, T_1 = 16$	0.390	0.815	0.000	0.821	0.991	0.000
$T_0 = 30, T_1 = 32$	0.160	0.780	0.000	0.443	0.972	0.000
$T_0 = 60, T_1 = 32$	0.376	0.812	0.000	0.521	0.987	0.000

Notes: All simulations are conducted with a thousand replications.

starts to increase when T_1 is very small and the power of the placebo method would likely increase if J was larger. While the cross-fitting t-test and the End-of-Sample Instability test have higher power before adjusting for their over-rejection under the null, after adjusting for size, the cross-fitting t-test has less power and the power of the End-of-Sample Instability test is similar to that of the Orthogonalized SCE t-test. The subsampling method's power generally remains the highest even after adjusting for its size, although it drastically drops when T_1 is made very small. Overall, however, the Orthogonalized SCE t-test consistently has the highest power of the tests that control for size in these simulations.

One potential point of caution is that in finite samples, the p-values may be fairly sensitive to the choice of the smoothing parameter K for the Series HAC estimator. In the simulations, I use the method of Sun (2013) to choose K and it appears to perform quite well. In applications, it may be worth checking the robustness of statistical significance of results to moderate changes in K. Also, one obvious drawback of this method is that it requires a set of units to be used as instruments while the others do not. This motivates further investigating what are alternative sets of moment conditions that can identify the ATT in such cases.

5 Other Applications and Limitations

5.1 Other Applications

In section 4, I focused on the traditional SCE case where there is a treated single unit that receives some binary intervention. However, the method can be extended to other contexts, provided that there is a set of moment conditions that identify whatever function of treatment effects is of interest. For example, if there is a set of treated units with indices in $\mathcal{N}_1$ who become treated at the same time, then the same moment equations could be used with Y_{0t} replaced with $\sum_{i \in \mathcal{N}_1} Y_{it}/|\mathcal{N}_1|$. In a staggered adoption setting, this could then be extended by estimating separate control weights for each treatment block while using units in other treatment blocks as instruments.[18]

Another potential application of this method beyond the SCE case is estimation of a scalar regression coefficient β using many instruments. Consider the linear model

$$Y_i = X_i \beta_{0,n} + \epsilon_i,$$

where Y is our outcome variable, X is an endogenous explanatory variable, and ϵ is unobserved. We could also extend this to include covariates W if we use that Frisch–Waugh–Lovell Theorem to project Y and X onto the orthogonal complement of W. Suppose we have a set of instruments $Z_i = (Z_{1i}, ..., Z_{Ji})$ which each satisfy the exclusion restriction so $E[Z_{ji}\epsilon_i] = 0$ for each $j \in \{1, ..., J\}$, but many of them may be weak or irrelevant. We can use the moment

[18] Since units need to be untreated during the time periods for which they are being used as instruments, it would be important for each instrument's moment equation to only use time periods in which that instrument is untreated.

condition

$$g(\theta) = E[(Z_i\delta)(Y_i - X_i\beta)].$$

If we are in a context where Z_i is high-dimensional, imposing the restriction that $\delta \in D \coloneqq \{\delta : ||\delta||_1 = 1\}$ can help us to achieve the necessary rate of convergence for $\hat{\delta}$ similarly as in the SCE case and as discussed in Remark 3.1. Because the instruments satisfy the exclusion restriction, $g(\theta) = E[(Z_i\delta)X_i(\beta_{0,n} - \beta)]$. Thus, $g(\theta) = 0$ if and only if $E[(Z_i\delta)X_i] = 0$ or $\beta = \beta_{0,n}$. The first option can be ruled out by choosing $\hat{\delta}$ to minimize our estimate for the asymptotic variance of $\tilde{\beta}_{GMM}$ (i.e., $\hat{f}(\theta,\eta) = \hat{V}(\theta,\eta)$), since this expression diverges if the linear combination of instruments is chosen to be irrelevant. As a result, it is still true that $g(\beta, \delta_{0,n})$ identifies β.

For this application, the moment condition is already orthogonal with respect to δ when $\beta = \beta_{0,n}$, so there is no need to perform the orthogonalization step. In other words, we can let $\hat{\eta} = \eta_{0,n} = 1$, so that the convergence condition for $\hat{\eta}$ is trivially satisfied. We can let $\hat{g}(\theta) = \sum_{i=1}^{n}(Z_i\delta)(Y_i - X_i\beta)/n$. Since this is again a case where solving for both $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$ is equivalent to setting a single moment condition equal to zero, our estimate has the standard two-stage least squares form $\frac{\sum_{i=1}^{n}(Z_i\hat{\delta})Y_i/n}{\sum_{i=1}^{n}(Z_i\hat{\delta})X_i/n}$. If the data are identically distributed, its asymptotic variance is $V_g/E[(Z_i\delta_{0,n})X_i]^2$ where V_g is the asymptotic variance of $\sum_{i=1}(Z_i\delta_{0,n})\epsilon_i/\sqrt{n}$. So then for some estimate of this asymptotic variance $\hat{V}_g(\theta)$, we can define

$$\hat{\theta} = \underset{\beta\in\mathbb{R},\delta\in\mathbb{R}^J}{\arg\min}\, \hat{V}_g(\theta)/(\sum_{i=1}^{n}(Z_i\delta)X_i/n)^2 \text{ such that } |\sum_{i=1}^{n}(Z_i\delta)(Y_i - X_i\beta)/n| \le \lambda_\delta,\ ||\delta||_1 = 1.$$

Because of the constraints, we generally choose a sparse subset of the instruments, which provide us with a more precise estimate of $\beta_{0,n}$. We should expect this estimator to perform well when there is a sparse subset of strong instruments. Other methods have been proposed for this case, such as Belloni et al. (2012) who use a two-stage estimator with LASSO used in the first stage. They show that their estimator can be semi-parametrically efficient under homoscedasticity conditions. Further work is needed to compare the performance of this estimator to existing options and explore variations (e.g., using a jackknife estimator for $\tilde{\beta}_{GMM}$ when the data are I.I.D.).

This method could also be applied in cases where θ is point-identified but constrained. Generally, constraints can present a complication for inference if $\theta_{0,n}$ is at or close to the boundary of the parameter space, but the Neyman orthogonalization allows us to handle constraints on the nuisance parameter and the one-step estimator $\tilde{\beta}_{OS}$ allows for the parameter of interest to be at or near the boundary of the parameter space. One example is models with control variables that have sign-restricted coefficients, and the number of controls can be large if regularization can be used to achieve the rate of convergence requirements in section 2. This can also be applied to random coefficient models, such as that of Berry et al. (1995), since the variance parameters are constrained to be non-negative.

5.2 Inference with Full Vector Partial Identification

So far, I have assumed that the parameter of interest is point-identified. However, ideas from this procedure may still be of use when β is also partially identified. In some cases, it may be possible to reparametrize the model in order to obtain a point-identified subvector we want to conduct inference on. Alternatively, if we wish to test a null hypothesis that some vector $\bar{\beta}$ is in the identified set for β, then we could use $n\hat{M}(\bar{\beta},\hat{\delta},\hat{\eta})'\hat{V}_M(\bar{\beta},\hat{\delta},\hat{\eta})^{-1}\hat{M}(\bar{\beta},\hat{\delta},\hat{\eta})$ to form a test statistic, where $\delta_{0,n}$ and $\eta_{0,n}$ can be chosen to minimize $\hat{V}_M(\bar{\beta},\delta,\eta)$ since this may increase the power of this test. Consider the following model:

$$Y_i = X_i\beta_{0,n} + h(V_i,\delta_{0,n}) + \epsilon_i,$$

where $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}^J$, and the function h is known. If both X and the elements of V are endogenous observable variables, then we can try to identify the whole vector with a set of instruments. However, if we have fewer instruments than the length of $\theta = (\beta,\delta)$, then the entire vector θ is partially identified. If we have a set of instruments $Z_i \in \mathbb{R}^Q$ and we assume for simplicity that (Y_i, X_i, V_i, Z_i) is I.I.D., then the identified set is $\theta \in \mathbb{R}^{J+1}$ such that

$$E[Z_i Y_i] - E[Z_i X_i]\beta - E[Z_i h(V_i, \delta)] = 0.$$

Since β appears in all the moment conditions, we can have $m = Q$ so $\eta \in \mathbb{R}^{Q \times Q}$. Our identified set for η is similar to the SCE example where $\eta_{0,n}$ should satisfy $\partial_\delta \eta_{0,n} E[Z_i h(V_i, \delta)] = 0$. Supposing that we want to test whether $\beta = 0$ is in the identified set, we can use the test statistic $n\hat{M}(0, \hat{\delta}, \hat{\eta})' \hat{V}_M(0, \hat{\delta}, \hat{\eta})^{-1} \hat{M}(0, \hat{\delta}, \hat{\eta})$. For any θ with $\beta = 0$, we have $V_g(\theta) = E[(Z_i(Y_i - h(V_i, \delta)))(Z_i(Y_i - h(V_i, \delta)))']$. Since the sampling is I.I.D., we can let $\hat{V}_M$ be the sample variance of the orthogonalized moment conditions. Since g is linear in β, verifying many of the conditions in section 2 is similar to the SCE case. However, a faster rate of convergence for $||\hat{\delta} - \delta_{0,n}||_2$ is required since the moment conditions are not linear in δ. On the other hand, the complication of δ being high-dimensional is not present. Because the data are I.I.D., when the sample is large we are able to accurately estimate V_g and V_M so it may be reasonable to make $\hat{f}$ a function of $\hat{V}_M$.

Partial identification of θ can also arise when the number of moment conditions exceeds the number of parameters to be estimated. For example, Chalak and Kim (2024) study measurement error models where the latent factors and the observable proxies for the latent factors can both directly affect the outcome variable. They find that under independence conditions on the errors, higher order moment conditions can be used to partially identify the entire vector. Here, the identified set is a finite set of points. If we are interested in testing a particular null hypothesis for a subvector we could take the same approach discussed above. However, in this case, identification for the whole vector or a subvector can be achieved by placing sign restrictions on either the entire vector of parameters or a subvector respectively. This can place us in the case where $\beta_{0,n}$ is point-identified and θ is constrained, in which case this method can be employed as previously discussed.

5.3 Weak Identification Cases

Assumption 3.1.4 can be considered a strong partial identification condition, as it bounds how far values of δ can be from its identified set in terms of how small the value of δ makes the population moment conditions at $\beta_{0,n}$. This means the results in the previous sections allow for δ to be strongly point-identified, strongly partially identified, or completely unidentified, but cases of weak identification have been ruled out. Weak identification of $D_{0,n}$ can present a problem for this method because it relies on being able to consistently estimate an element of the identified set $\delta_{0,n}$ which, roughly speaking, relies on being able to consistently estimate the identified set $D_{0,n}$. This cannot be done when $D_{0,n}$ is weakly identified, but whether $\hat{\delta}$ is converging to an element of $D_{0,n}$ be may still influence the asymptotic distribution of $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$.

To illustrate the problem, consider the example of the non-linear regression model,

$$Y_i = \beta_{0,n} h(X_i, \delta_{0,n}) + U_i,$$

where Y and X are observed and the function h is known. Suppose the moment conditions $g(\beta, \delta) = E[X_i U_i] = E[X_i(Y_i - \beta h(X_i, \delta))]$ and $\hat{g}(\beta, \delta) = \sum_{i=1}^{n}(X_i(Y_i - \beta h(X_i, \delta)))/n$ are used. Here, $D_{0,n} = \mathbb{R}^J$ when $\beta_{0,n} = 0$ and $D_{0,n}$ contains a single element $\delta_{0,n}$ otherwise under restrictions on h. When $\beta_{0,n}/\lambda_\delta \to b \in [0, \infty)$, the feasible values for $\hat{\delta}$ generally expand to include all of $\mathbb{R}^J$, whereas when $\beta_{0,n}/\lambda_\delta \to \infty$, $\hat{\Theta}_0$ shrinks to just include a single point. While standard estimators of β are asymptotically normal when $\sqrt{n}\beta_{0,n} \to \infty$, they are $\sqrt{n}$-consistent but not asymptotically normal in the weak identification case of $\sqrt{n}\beta_{0,n} \to b \in [0, \infty)$ (see, Andrews and Cheng (2012) and Han and McCloskey (2019)). When δ is weakly identified, the orthogonalized estimators $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$ are $\sqrt{n}$-consistent by a similar reasoning as with standard estimators. However, they are generally not asymptotically normal. For example, in the case with $\sqrt{n}\beta_{0,n} \to b$, $\beta_{0,n}/\lambda_\delta \to 0$ so $\hat{\delta}$ converges to the global minimizer of $f(\beta_{0,n}, \delta, \eta)$ on $\mathbb{R}^J$, say δ^*. This may in turn result in $\hat{\eta}$ converging to a different value η^*. Using the adaptivity condition, we would then expect that $\sqrt{n}\hat{M}(\beta_{0,n}, \hat{\delta}, \hat{\eta}) = \sqrt{n}\hat{M}(\beta_{0,n}, \delta^*, \eta^*) + o_p(1)$. While $\sqrt{n}\hat{M}(\beta_{0,n}, \delta^*, \eta^*)$ may still be asymptotically normal, it is not centered around zero unless $\sqrt{n}\beta_{0,n} \to 0$. As a result, $\tilde{\beta}_{GMM}$ and $\tilde{\beta}_{OS}$ generally have non-standard limiting distributions. On the other hand, when $\sqrt{n}\beta_{0,n} \to \infty$, it is possible to consistently estimate $\delta_{0,n}$, so the method works similarly to before. Hence, the problem arises when $\sqrt{n}\beta_{0,n} \to b \in (0, \infty)$. Intuitively, this is because we either want to be close enough to the unidentified case where the true value of δ is irrelevant (i.e., $\sqrt{n}\beta_{0,n} \to 0$ so $\sqrt{n}M(\beta_{0,n}, \delta^*, \eta^*) \to 0$) or close enough to the strongly identified case where we can accurately estimate $\delta_{0,n}$ (i.e., $\sqrt{n}\beta_{0,n} \to \infty$ so $||\hat{\delta} - \delta_{0,n}||_1 \xrightarrow{p} 0$).

Appendix A

Proof of Lemma 3.1: I first show that $||\hat{\theta} - \theta_{0,n}||_1 \xrightarrow{p} 0$ and $||\hat{\eta} - \eta_{0,n}||_1 \xrightarrow{p} 0$, both when J fixed and when $J \to \infty$. Let $\epsilon > 0$ and let $N_\epsilon(\theta_{0,n}, \eta_{0,n})$ be an open ϵ-ball using the $||\cdot||_1$ norm centered at $(\theta_{0,n}, \eta_{0,n})$. For some $\zeta > 0$ to be specified below, let $S_0 \coloneqq \{(\theta, \eta) \in \Theta_{0,n} \times H_{0,n} : f(\theta, \eta) \leq f(\theta_{0,n}, \eta_{0,n}) + \zeta\}$. Then let $\tau = d_H(S_0, \{(\theta_{0,n}, \eta_{0,n})\}, ||\cdot||_1)$, where $d_H(, , ||\cdot||_1)$ denotes the Hausdorff distance using the $||\cdot||_1$ norm. Note $N_\epsilon(\theta_{0,n}, \eta_{0,n}) \cap S_0 = \emptyset$ exactly when $\tau < \epsilon$, so $N_{\min\{\epsilon, \tau/2\}}(\theta_{0,n}, \eta_{0,n}) \cap S_0 \neq \emptyset$ for all n. Because S_0 is compact and f is continuous on S_0 for each n by Assumption 3.1.1, $S_0 \cap N^c_{\min\{\epsilon, \tau/2\}}(\theta_{0,n}, \eta_{0,n})$ is also compact for each n. Then we can define

$$\gamma := \min_{(\theta,\eta) \in S_0 \cap N^c_{\min\{\epsilon,\tau/2\}}(\theta_{0,n}, \eta_{0,n})} f(\theta, \eta) - f(\theta_{0,n}, \eta_{0,n}).$$

Note that by Assumption 3.1.5,

$$C_6 |f(\theta_1, \eta_1) - f(\theta_2, \eta_2)|^{\gamma_2} \leq ||\theta_1 - \theta_2||_1 + ||\eta_1 - \eta_2||_1$$

when $f(\theta_1, \eta_1), f(\theta_2, \eta_2) \leq f(\theta_{0,n}, \eta_{0,n}) + C_5$. So if ζ is chosen such that $\zeta \leq C_5$, then it holds that there exists κ such that for all $(\theta_1, \eta_1), (\theta_2, \eta_2) \in S := \{(\theta, \eta) \in \Theta_n \times H \ : \ f(\theta, \eta) \leq f(\theta_{0,n}, \eta_{0,n}) + \zeta\}$, if $||\theta_1 - \theta_2||_1 + ||\eta_1 - \eta_2||_1 < \kappa$, then $|f(\theta_1, \eta_1) - f(\theta_2, \eta_2)| < \gamma/4$. Letting S_0^κ denote the closed κ blowup of S_0 using the $|| \cdot ||_1$ norm, we obtain that:

$$\min_{(\theta,\eta) \in S_0^\kappa \cap N^c_{\min\{\epsilon, \tau/2\}}(\theta_{0,n}, \eta_{0,n})} f(\theta, \eta) - f(\theta_{0,n}, \eta_{0,n}) > 3/4\gamma.$$

Note that since $(\hat{\theta}, \hat{\eta}) \in \hat{\Theta}_0 \times \hat{H}_0$, then

$$(\hat{\theta}, \hat{\eta}) \in \hat{S}_0 := \{(\theta, \eta) \in \hat{\Theta}_0 \times \hat{H}_0 : f(\theta, \eta) \leq f(\theta_{0,n}, \eta_{0,n}) + \zeta\}$$

when $|f(\hat{\theta}, \hat{\eta}) - f(\theta_{0,n}, \eta_{0,n})| \leq 3/4\gamma$ as long as ζ is chosen so that $\zeta \geq 3/4\gamma$. Therefore, since $d_H(\hat{S}_0, S_0; || \cdot ||_1) < \kappa$ implies $\hat{S}_0 \subset S_0^\kappa$, it follows that

$$P(||\hat{\theta} - \theta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 < \epsilon) \geq P(||\hat{\theta} - \theta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 < \min\{\epsilon, \tau/2\})$$

$$\geq P(|f(\hat{\theta}, \hat{\eta}) - f(\theta_{0,n}, \eta_{0,n})| \leq 3/4\gamma; d_H(\hat{S}_0, S_0, || \cdot ||_1) < \kappa).$$

Let $(\theta_p, \eta_p) = \arg\min_{(\theta,\eta) \in \hat{S}_0} ||\theta - \theta_{0,n}||_1 + ||\eta - \eta_{0,n}||_1$. If $d_H(\hat{S}_0, S_0, || \cdot ||_1) < \kappa$, then $||\theta_p - \theta_{0,n}||_1 + ||\eta_p - \eta_{0,n}||_1 < \kappa$ and therefore $f(\theta_p, \eta_p) < f(\theta_{0,n}, \eta_{0,n}) + \gamma/4$. By definition of $\hat{\theta}$ and $\hat{\eta}$, $\hat{f}(\hat{\theta}, \hat{\eta}) \leq \hat{f}(\theta_p, \eta_p)$. This implies

$$f(\hat{\theta}, \hat{\eta}) - f(\theta_p, \eta_p) \leq |\hat{f}(\hat{\theta}, \hat{\eta}) - f(\hat{\theta}, \hat{\eta})| + |\hat{f}(\theta_p, \eta_p) - f(\theta_p, \eta_p)|.$$

Thus $f(\theta_p, \eta_p) < f(\theta_{0,n}, \eta_{0,n}) + \gamma/4$ and $|\hat{f}(\hat{\theta}, \hat{\eta}) - f(\hat{\theta}, \hat{\eta})| + |\hat{f}(\theta_p, \eta_p) - f(\theta_p, \eta_p)| < \gamma/4$ together imply $f(\hat{\theta}, \hat{\eta}) - f(\theta_{0,n}, \eta_{0,n}) < \gamma/2$, in which case $||\hat{\theta} - \theta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 < \epsilon$. Therefore, we have that

$$P(||\hat{\theta} - \theta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 < \epsilon)$$

$$\geq P(|\hat{f}(\hat{\theta}, \hat{\eta}) - f(\hat{\theta}, \hat{\eta})| + |\hat{f}(\theta_p, \eta_p) - f(\theta_p, \eta_p)| < \gamma/4; d_H(\hat{S}_0, S_0, || \cdot ||_1) < \kappa)$$

$$\geq P(2 \sup_{(\theta,\eta)\in\hat{\Theta}_0\times\hat{H}_0} |f(\theta,\eta) - \hat{f}(\theta,\eta)| < \gamma/4; d_H(\hat{S}_0, S_0, ||\cdot||_1) < \kappa).$$

Then because $\sup_{(\theta,\eta)\in\hat{\Theta}_0\times\hat{H}_0} |f(\theta,\eta) - \hat{f}(\theta,\eta)| = O_p(c_n)$ and $d_H(\hat{S}_0, S_0, ||\cdot||_1) = O_p(a_n)$ by Lemma B1, we have that $||\hat{\theta} - \theta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 < \epsilon$ with probability approaching one (wpa1). So $||\hat{\theta} - \theta_{0,n}||_1 = o_p(1)$ and $||\hat{\eta} - \eta_{0,n}||_1 = o_p(1)$.

Now I show the rate of convergence for $||\hat{\theta} - \theta_{0,n}||_1$ and $||\hat{\eta} - \eta_{0,n}||_1$. Let $\{\epsilon_n\}_{n\in\mathbb{N}}$ be a decreasing sequence where $\epsilon_n > 0$ and $\epsilon_n \to 0$. Let

$$S_0 := \{(\theta,\eta) \in \Theta_{0,n}\times H_{0,n} : f(\theta,\eta) \leq f(\theta_{0,n},\eta_{0,n})+\zeta; ||\theta-\theta_{0,n}||_1+||\eta-\eta_{0,n}||_1 < \min\{C_3, C_5\}/2\},$$

$$\hat{S}_0 := \{(\theta,\eta) \in \hat{\Theta}_0 \times \hat{H}_0 : f(\theta_{0,n},\eta_{0,n}) + \zeta; ||\theta - \theta_{0,n}||_1 + ||\eta - \eta_{0,n}||_1 < \min\{C_3, C_5\}/2\}, \text{ and}$$

$$S := \{(\theta,\eta) \in \Theta_n \times H : f(\theta,\eta) \leq f(\theta_{0,n},\eta_{0,n})+\zeta; ||\theta-\theta_{0,n}||_1+||\eta-\eta_{0,n}||_1 < \min\{C_3, C_5\}/2\}.$$

By the consistency of $\hat{\theta}$ and $\hat{\eta}$, $(\hat{\theta},\hat{\eta}) \in \hat{S}_0$ wpa1, so since $\hat{S}_0 \subseteq \hat{\Theta}_0 \times \hat{H}_0$,

$$(\hat{\theta},\hat{\eta}) = \arg\min_{(\theta,\eta)\in\hat{S}_0} \hat{f}(\theta,\eta) \text{ wpa1}.$$

Similarly to before, let $\tau_n = d_H(S_0, \{(\theta_{0,n},\eta_{0,n})\}, ||\cdot||_1)$ and let $N_\epsilon(\theta_{0,n},\eta_{0,n})$ be an open ϵ-ball centered at $(\theta_{0,n},\eta_{0,n})$ so that $S_0 \cap N^c_{\min\{\epsilon_n,\tau_n/2\}}(\theta_{0,n},\eta_{0,n}) \neq \emptyset$ for all n. Because S_0 is compact for each n, so $S_0 \cap N^c_{\min\{\epsilon_n,\tau_n/2\}}(\theta_{0,n},\eta_{0,n})$ is compact for each n.

By Assumption 3.2.5, for any $(\theta,\eta) \in \Theta_{0,n} \times H_{0,n}$, we have that

$$|f(\theta_{0,n},\eta_{0,n}) - f(\theta,\eta)| \geq C_4 \min\{||\theta - \theta_{0,n}||_1^{\gamma_1} + ||\eta - \eta_{0,n}||_1^{\gamma_1}, C_3\}.$$

Then for sufficiently large n, $2\epsilon_n^{\gamma_1} < C_3$ so

$$\nu_n := \min_{(\theta,\eta)\in S_0\cap N^c_{\epsilon_n(\theta_{0,n},\eta_{0,n})}} f(\theta,\eta) - f(\theta_{0,n},\eta_{0,n}) \geq 2C_4\epsilon_n^{\gamma_1}$$

for all (θ,η) with $||\theta-\theta_{0,n}||_1, ||\eta-\eta_{0,n}||_1 \leq \epsilon_n$. Also by Assumption 3.2.5, for all $(\theta_1,\eta_1), (\theta_2,\eta_2) \in$

S,

$$C_6|f(\theta_1,\eta_1)-f(\theta_2,\eta_2)|^{\gamma_2} \leq ||\theta_1-\theta_2||_1+||\eta_1-\eta_2||_1.$$

Let $\kappa_n = C_6/4\nu_n^{1/\gamma_2}$. So when $\kappa_n < C_5$, $||\theta-\theta_{0,n}||_1+||\eta-\eta_{0,n}||_1 < \kappa_n$ implies $|f(\theta_1,\eta_1) - f(\theta_2,\eta_2)| < \nu_n^{1/\gamma_2}/4$. Let $S_0^{\kappa_n}$ denote the closed κ_n blowup of S_0 using the $||\cdot||_1$ norm. We have that for all $(\theta,\eta)\in S_0^{\kappa_n}$, in which case we obtain that:

$$\min_{(\theta,\eta)\in S_0^{\kappa_n}\cap N^c_{\min\{\epsilon,\tau_n/2\}}(\theta_{0,n},\eta_{0,n})} f(\theta,\eta)-f(\theta_{0,n},\eta_{0,n}) > 3/4\nu_n^{1/\gamma_2}.$$

Therefore, since $d_H(\hat{S}_0,S_0;||\cdot||_1)<\kappa_n$ implies $\hat{S}_0\subset S_0^{\kappa_n}$, it follows that for n sufficiently large

$$P(||\hat{\theta}-\theta_{0,n}||_1+||\hat{\eta}-\eta_{0,n}||_1 < \epsilon_n) \geq P(|f(\hat{\theta},\hat{\eta})-f(\theta_{0,n},\eta_{0,n})| \leq 3/4\nu_n^{1/\gamma_2}; d_H(\hat{S}_0,S_0,||\cdot||_1)<\kappa_n).$$

Let

$$(\theta_p,\eta_p) = \arg\min_{(\theta,\eta)\in\hat{S}_0} ||\theta-\theta_{0,n}||_1+||\eta-\eta_{0,n}||_1.$$

If $d_H(\hat{S}_0,S_0,||\cdot||_1)<\kappa_n$, then $||\theta_p-\theta_{0,n}||_1+||\eta_p-\eta_{0,n}||_1<\kappa_n$ and therefore $f(\theta_p,\eta_p) < f(\theta_{0,n},\eta_{0,n})+\nu_n^{1/\gamma_2}/4$. By definition of $\hat{\theta}$ and $\hat{\eta}$, $\hat{f}(\hat{\theta},\hat{\eta})\leq\hat{f}(\theta_p,\eta_p)$. This implies

$$f(\hat{\theta},\hat{\eta})-f(\theta_p,\eta_p) \leq |\hat{f}(\hat{\theta},\hat{\eta})-f(\hat{\theta},\hat{\eta})|+|\hat{f}(\theta_p,\eta_p)-f(\theta_p,\eta_p)|.$$

Thus $f(\theta_p,\eta_p) < f(\theta_{0,n},\eta_{0,n})+\nu_n^{1/\gamma_2}/4$ and $|\hat{f}(\hat{\theta},\hat{\eta})-f(\hat{\theta},\hat{\eta})|+|\hat{f}(\theta_p,\eta_p)-f(\theta_p,\eta_p)| < \nu_n^{1/\gamma_2}/4$ together imply $f(\hat{\theta},\hat{\eta})-f(\theta_{0,n},\eta_{0,n}) < \nu_n^{1/\gamma_2}/2$, in which case $||\hat{\theta}-\theta_{0,n}||_1+||\hat{\eta}-\eta_{0,n}||_1<\epsilon_n$. Therefore, we have that

$$P(||\hat{\theta}-\theta_{0,n}||_1+||\hat{\eta}-\eta_{0,n}||_1 < \epsilon_n)$$

$$\geq P(|\hat{f}(\hat{\theta},\hat{\eta})-f(\hat{\theta},\hat{\eta})|+|\hat{f}(\theta_p,\eta_p)-f(\theta_p,\eta_p)| < \nu_n^{1/\gamma_2}/4; d_H(\hat{S}_0,S_0,||\cdot||_1)<\kappa_n)$$

$$\geq P(2\sup_{(\theta,\eta)\in\hat{\Theta}_0\times\hat{H}_0}|f(\theta,\eta)-\hat{f}(\theta,\eta)| < (2C_4)^{1/\gamma_2}\epsilon_n^{\gamma_1/\gamma_2}/4; d_H(\hat{S}_0,S_0,||\cdot||_1) < C_6/4(2C_4)^{1/\gamma_2}\epsilon_n^{\gamma_1/\gamma_2}.)$$

$$= P((\sup_{(\theta,\eta)\in\hat{\Theta}_0\times\hat{H}_0} |f(\theta,\eta) - \hat{f}(\theta,\eta)|)^{\gamma_2/\gamma_1} < (2C_4)^{1/\gamma_1}\epsilon_n/4^{\gamma_2/\gamma_1}; (d_H(\hat{S}_0, S_0, ||\cdot||_1))^{\gamma_2/\gamma_1}$$

$$< (C_6/4)^{\gamma_2/\gamma_1}(2C_4)^{1/\gamma_1}\epsilon_n),$$

where the second inequality follows from $\nu_n > 2C_4\epsilon_n^{\gamma_1}$ and $\kappa_n = C_6/4\nu_n^{1/\gamma_2} > C_6/4(2C_4)^{1/\gamma_2}\epsilon_n^{\gamma_1/\gamma_2}$. Then because $\sup_{(\theta,\eta)\in\hat{S}_0} |f(\theta,\eta) - \hat{f}(\theta,\eta)| = O_p(c_n)$ and $d_H(\hat{S}_0, S_0, ||\cdot||_1) = O_p(\max\{\lambda_\delta, \lambda_\eta, a_n\})$ by Lemma B1, for any sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ with $\epsilon_n > 0$ and $\epsilon_n/(\max\{\lambda_\delta, \lambda_\eta, a_n, c_n\}^{\gamma_2/\gamma_1}) \to \infty$, we have that $P(||\hat{\theta} - \theta_{0,n}||_1 < \epsilon_n) \to 1$ and $P(||\hat{\eta} - \eta_{0,n}||_1 < \epsilon_n) \to 1$. Therefore,

$$||\hat{\delta} - \delta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 \leq ||\hat{\theta} - \theta_{0,n}||_1 + ||\hat{\eta} - \eta_{0,n}||_1 = O_p(\max\{\lambda_\delta, \lambda_\eta, a_n, c_n\}^{\gamma_2/\gamma_1}).$$

Proof of Lemma 2.1: Let $\gamma = (\delta, \eta_q)$ and $\hat{\gamma} = (\hat{\delta}, \hat{\eta}_q)$ where η_q is the q-th row of η. Since the q-th element on the orthogonalized sample moment conditions $\hat{M}_q$ for $q \in \{1, ..., m\}$ are twice continuously differentiable in γ, for each q there exists $\bar{\gamma}$ with $||\bar{\gamma} - \gamma_{0,n}||_1 \leq ||\hat{\gamma} - \gamma_{0,n}||_1$ such that:

$$\sqrt{n}(\hat{M}_q(\beta_{0,n}, \hat{\gamma}) - \hat{M}_q(\beta_{0,n}, \gamma_{0,n})) = \sqrt{n}\partial_\gamma \hat{M}_q(\beta_{0,n}, \gamma_{0,n})(\hat{\gamma} - \gamma_{0,n}) + \sqrt{n}(\hat{\gamma} - \gamma_{0,n})'\partial_\gamma^2 \hat{M}_q(\beta_{0,n}, \bar{\gamma})(\hat{\gamma} - \gamma_{0,n}).$$

The magnitude of the first term on the right-hand side is less than or equal to

$$\sqrt{n}||\partial_\gamma \hat{M}_q(\beta_{0,n}, \gamma_{0,n})||_\infty ||\hat{\gamma} - \gamma_{0,n}||_1 \leq \sqrt{n}O_p(\log(J)/\sqrt{n})o_p(1/\log(J)) = o_p(1).$$

The second term on the right-hand side is equal to

$$\sqrt{n}(\hat{\delta} - \delta_{0,n})'\partial_\delta^2 \hat{M}_q(\beta_{0,n}, \bar{\gamma})(\hat{\delta} - \delta_{0,n}) + 2\sqrt{n}(\hat{\eta}_q - \eta_{0,n,q})\partial_\delta \hat{g}(\beta_{0,n}, \bar{\delta})(\hat{\delta} - \delta_{0,n}).$$

In the linear case, $\sqrt{n}(\hat{\delta} - \delta_{0,n})'\partial_\delta^2 \hat{M}_q(\beta_{0,n}, \bar{\gamma})(\hat{\delta} - \delta_{0,n}) = 0$ and $\partial_\delta \hat{g}(\beta_{0,n}, \bar{\delta}) = \partial_\delta \hat{g}(\theta_{0,n}) = \partial_\delta \hat{g}(\hat{\theta})$. Therefore,

$$||(\hat{\eta}_q - \eta_{0,n,q})\partial_\delta \hat{g}(\beta_{0,n}, \bar{\delta})(\hat{\delta} - \delta_{0,n})||_\infty =$$

$$||\hat{\eta}_q \partial_\delta \hat{g}(\hat{\theta})(\hat{\delta} - \delta_{0,n}) - \eta_{0,n,q} \partial_\delta \hat{g}(\theta_{0,n})(\hat{\delta} - \delta_{0,n})||_\infty \leq$$

$$(||\hat{\eta}_q \partial_\delta \hat{g}(\hat{\theta})||_\infty + ||\eta_{0,n,q} \partial_\delta \hat{g}(\theta_{0,n})||_\infty)||\hat{\delta} - \delta_{0,n}||_1 = (||\hat{\eta}_q \partial_\delta \hat{g}(\hat{\theta})||_\infty + ||\partial_\delta \hat{M}_q(\theta_{0,n}, \eta_{0,n})||_\infty)||\hat{\delta} - \delta_{0,n}||_1$$

$$= (O_p(\log(J)\log(n)/\sqrt{n}) + O_p(\log(J)/\sqrt{n}))o_p(1/(\log(J)\log(n))) = o_p(1/\sqrt{n}).$$

Otherwise, for the non-linear case, there exists $\bar{\delta}^*$ such that $||\bar{\delta}^* - \delta_{0,n}||_1 \leq ||\hat{\delta} - \delta_{0,n}||_1$ and

$$\partial_\delta \hat{g}(\beta_{0,n}, \bar{\delta}) = \partial_\delta \hat{g}(\theta_{0,n}) + (\bar{\delta} - \delta_{0,n})' \partial^2_\delta \hat{g}(\beta_{0,n}, \bar{\delta}^*).$$

Therefore, using Assumptions 2.1.2 and 2.1.4,

$$\sqrt{n}||(\hat{\eta}_q - \eta_{0,n,q}) \partial_\delta \hat{g}(\beta_{0,n}, \bar{\delta})(\hat{\delta} - \delta_{0,n})||_\infty \leq$$

$$\sqrt{n}||\hat{\eta}_q - \eta_{0,n,q}||_1(||\partial_\delta \hat{g}(\theta_{0,n})||_\infty ||\hat{\delta} - \delta_{0,n}||_1$$

$$+ \max_{s \in \{1,\ldots,Q\}} |(\bar{\delta} - \delta_{0,n})' \partial^2_\delta \hat{g}_s(\beta_{0,n}, \bar{\delta}^*)(\hat{\delta} - \delta_{0,n})|)$$

$$\leq \sqrt{n} o_p(n^{-1/4}/\sqrt{\log(J)}) O_p(\log(J)) o_p(n^{-1/4}/\sqrt{\log(J)})$$

$$+ \sqrt{n} o_p(n^{-1/4}/\sqrt{\log(J)})||\hat{\delta} - \delta_{0,n}||_2^2 \max_{s \in \{1,\ldots,Q\}} \max eig(\partial^2_\delta \hat{g}_s(\beta_{0,n}, \bar{\delta}^*)).$$

Then, using $\epsilon > 0$ defined by Assumption 2.1.3, because $||\bar{\delta}^* - \delta_{0,n}||_1 \leq ||\hat{\delta} - \delta_{0,n}||_1 < \epsilon$ wpa1, then wpa1,

$$\max_{s \in \{1,\ldots,Q\}} \max eig(\partial^2_\delta \hat{g}_s(\beta_{0,n}, \bar{\delta}^*)) \leq \max_{s \in \{1,\ldots,Q\}} \sup_{\delta : ||\delta - \delta_{0,n}||_1 < \epsilon} \max eig(\partial^2_\delta \hat{g}_s(\beta_{0,n}, \delta)) = O_p(\log(J)) \text{ and}$$

$$\max eig(\partial^2_\delta \hat{M}_q(\beta_{0,n}, \bar{\gamma})) \leq \sup_{\gamma : ||\gamma - \gamma_{0,n}||_1 < \epsilon} \max eig(\partial^2_\delta \hat{M}_q(\beta_{0,n}, \gamma)) = O_p(\log(J)).$$

Hence,

$$\sqrt{n}||(\hat{\eta}_q - \eta_{0,n,q}) \partial_\delta \hat{g}(\beta_{0,n}, \bar{\delta})(\hat{\delta} - \delta_{0,n})||_\infty \leq$$

$$\leq \sqrt{n} o_p(n^{-1/4}/\sqrt{\log(J)}) O_p(\log(J)) o_p(n^{-1/4}/\sqrt{\log(J)})$$

$$+\sqrt{n}o_p(n^{-1/4}/\log(J))o_p(n^{-1/2}/\log(J))O_p(\log(J)) = o_p(1).$$

Also,

$$|\sqrt{n}(\hat{\delta} - \delta_{0,n})'\partial_{\delta}^2 \hat{M}_q(\beta_{0,n}, \bar{\gamma})(\hat{\delta} - \delta_{0,n})| \leq \sqrt{n}||\hat{\delta} - \delta_{0,n}||_2^2 \max eig(\partial_{\delta}^2 \hat{M}_q(\beta_{0,n}, \bar{\gamma}))$$

$$\leq \sqrt{n}o_p(n^{-1/2}/\log(J))O_p(\log(J)) = o_p(1).$$

Therefore,

$$\sqrt{n}(\hat{M}_q(\beta_{0,n}, \hat{\gamma}) - \hat{M}_q(\beta_{0,n}, \gamma_{0,n})) = o_p(1),$$

for each $q \in \{1, ..., m\}$.

Proof of Proposition 2.1: I first show the consistency of $\tilde{\beta}_{GMM}$. Note that

$$||\hat{M}(\beta, \hat{\delta}, \hat{\eta}) - \hat{M}(\beta, \delta_{0,n}, \eta_{0,n})||_2 = ||\hat{\eta}\hat{g}(\beta, \hat{\delta}) - \eta_{0,n}\hat{g}(\beta, \delta_{0,n})||_2 \leq ||\hat{\eta}||_2||\hat{g}(\beta, \hat{\delta}) - g(\beta, \hat{\delta})||_2$$

$$+||\hat{\eta}||_2||g(\beta, \hat{\delta}) - g(\beta, \delta_{0,n})||_2 + ||\hat{\eta} - \eta_{0,n}||_2||g(\beta, \delta_{0,n}||_2$$

$$+||\eta_{0,n}||_2||g(\beta, \delta_{0,n}) - \hat{g}(\beta, \delta_{0,n})||_2.$$

Note that since $||\eta_{0,n}||_2 = O(1)$ and $||\hat{\eta} - \eta_{0,n}||_2 = o_p(1)$, $||\hat{\eta}||_2 = O_p(1)$. Then $\sup_{\beta \in B} ||g(\beta, \hat{\delta}) - g(\beta, \delta_{0,n})||_2 = o_p(1)$ by the continuous mapping theorem, and $\sup_{\beta \in B} ||\hat{g}(\beta, \hat{\delta}) - g(\beta, \hat{\delta})||_2 = o_p(1)$ and $\sup_{\beta \in B} ||g(\beta, \delta_{0,n}) - \hat{g}(\beta, \delta_{0,n})||_2 = o_p(1)$ by Assumption 2.1.1. Hence,

$$\sup_{\beta \in B} ||\hat{M}(\beta, \hat{\delta}, \hat{\eta}) - \hat{M}(\beta, \delta_{0,n}, \eta_{0,n})||_2 = o_p(1).$$

Therefore, since $W_n \xrightarrow{p} W$ where W is positive definite, we have that

$$\sup_{\beta \in B} |\hat{M}(\beta, \hat{\delta}, \hat{\eta})'W_n\hat{M}(\beta, \hat{\delta}, \hat{\eta}) - \hat{M}(\beta, \delta_{0,n}, \eta_{0,n})'W_n\hat{M}(\beta, \delta_{0,n}, \eta_{0,n})| = o_p(1).$$

Let $\hat{Q}(\beta)$ be equal to the objective function from equation (5). Combining this with Assumption 2.1.1 and $W_n \xrightarrow{p} W$ gives $\sup_{\beta \in B} |\hat{Q}(\beta) - Q(\beta)| = o_p(1)$, where

$$Q(\beta) = M(\beta, \delta_{0,n}, \eta_{0,n})' W M(\beta, \delta_{0,n}, \eta_{0,n}).$$

Let $\epsilon > 0$. By the definition of $\tilde{\beta}_{GMM}$, $\hat{Q}(\tilde{\beta}_{GMM}) < \hat{Q}(\beta_{0,n}) + \epsilon/3$. Then by the uniform convergence we have that $Q(\tilde{\beta}_{GMM}) < \hat{Q}(\tilde{\beta}_{GMM}) + \epsilon/3$ and $\hat{Q}(\beta_{0,n}) < Q(\beta_{0,n}) + \epsilon/3$ wpa1. Combining these inequalities gives $Q(\tilde{\beta}_{GMM}) < Q(\beta_{0,n}) + \epsilon$ wpa1. By the strong identification condition of Assumption 2.2.4 and W being positive definite, for some $\bar{C} > 0$, $||\tilde{\beta}_{GMM} - \beta_{0,n}||_2^2/\tilde{C} \leq C^2 ||M(\tilde{\beta}_{GMM}, \delta_{0,n}, \eta_{0,n})||_2^2/\tilde{C} \leq Q(\tilde{\beta}_{GMM}) < \epsilon$ wpa1. So we have that $\tilde{\beta}_{GMM} - \beta_{0,n} \xrightarrow{p} 0$.

By Assumption 2.2.4, $\partial_\beta^2 Q(\beta_{0,n}) = \partial_\beta M(\theta_{0,n}, \eta_{0,n})' W \partial_\beta M(\theta_{0,n}, \eta_{0,n}) \rightarrow M_\beta' W M_\beta$ and $M_\beta' W M_\beta$ is positive definite. Using Assumption 2.2.2 and the consistency of $\hat{\delta}$ and $\hat{\eta}$, $\partial_\beta^2 \hat{Q}(\beta_{0,n}) = \partial_\beta \hat{g}(\beta_{0,n}, \hat{\delta})' \hat{\eta}' W_n \hat{\eta} \partial_\beta \hat{g}(\beta_{0,n}, \hat{\delta}) - \partial_\beta M(\theta_{0,n}, \eta_{0,n})' W \partial_\beta M(\theta_{0,n}, \eta_{0,n}) \xrightarrow{p} 0$. So $\partial_\beta^2 \hat{Q}(\beta_{0,n}) \xrightarrow{p} M_\beta' W M_\beta$. Since $\hat{Q}(\beta)$ is twice continuously differentiable,

$$\hat{Q}(\tilde{\beta}_{GMM}) = \hat{Q}(\beta_{0,n}) + \partial_\beta \hat{Q}(\beta_{0,n})(\tilde{\beta}_{GMM} - \beta_{0,n}) + (\tilde{\beta}_{GMM} - \beta_{0,n})' \partial_\beta^2 \hat{Q}(\bar{\beta})(\tilde{\beta}_{GMM} - \beta_{0,n})/2,$$

for some $\tilde{\beta}$ with $||\tilde{\beta} - \beta_{0,n}||_2 \leq ||\tilde{\beta}_{GMM} - \beta_{0,n}||_2$. Then since $\hat{Q}(\tilde{\beta}_{GMM}) \leq \hat{Q}(\beta_{0,n})$, we have that

$$0 \geq \partial_\beta \hat{Q}(\beta_{0,n})(\tilde{\beta}_{GMM} - \beta_{0,n}) + (\tilde{\beta}_{GMM} - \beta_{0,n})' \partial_\beta^2 \hat{Q}(\bar{\beta})(\tilde{\beta}_{GMM} - \beta_{0,n})/2.$$

Using Assumption 2.2.2 and the consistency of $\tilde{\beta}_{GMM}$,

$$||\partial_\beta^2 \hat{Q}(\bar{\beta}) - \partial_\beta^2 \hat{Q}(\beta_{0,n})||_2 \leq \sup_{\beta : ||\beta - \beta_{0,n}||_2 < ||\tilde{\beta}_{GMM} - \beta_{0,n}||_2} ||\partial_\beta^2 \hat{Q}(\beta) - \partial_\beta^2 \hat{Q}(\beta_{0,n})||_2 = o_p(1).$$

Then

$$0 \geq \partial_\beta \hat{Q}(\beta_{0,n})(\tilde{\beta}_{GMM} - \beta_{0,n}) + (\tilde{\beta}_{GMM} - \beta_{0,n})' M_\beta' W M_\beta (\tilde{\beta}_{GMM} - \beta_{0,n})/2 + o_p(||\tilde{\beta}_{GMM} - \beta_{0,n}||_2^2).$$

Multiplying both sides by $\frac{n}{(1+\sqrt{n}||\tilde{\beta}_{GMM}-\beta_{0,n}||_2)^2}$ gives

$$\frac{\sqrt{n}||\tilde{\beta}_{GMM}-\beta_{0,n}||_2}{(1+\sqrt{n}||\tilde{\beta}_{GMM}-\beta_{0,n}||_2)^2}\sqrt{n}\partial_\beta\hat{Q}(\beta_{0,n})$$

$$+\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})'M_\beta'WM_\beta\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})/2+o_p(1)\leq 0.$$

Then if $\sqrt{n}||\tilde{\beta}_{GMM}-\beta_{0,n}||_2\to\infty$, $\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})'M_\beta'WM_\beta\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})/2\leq o_p(1)$. But since $'M_\beta'WM_\beta$ is positive definite, this implies that $\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})=o_p(1)$ which is a contradiction. Therefore, $\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})=O_p(1)$.

I now show that asymptotic normality of $\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})$ when Assumption 2.2.6 additionally holds. Since $\beta_{0,n}$ is bounded away from the boundary of B, wpa1 the first order condition is satisfied,

$$\hat{M}(\tilde{\beta}_{GMM},\hat{\delta},\hat{\eta})'W_n\partial_\beta\hat{M}(\tilde{\beta}_{GMM},\hat{\delta},\hat{\eta})=0.$$

Using the Mean Value Theorem, for some $\bar{\beta}$ such that $||\bar{\beta}-\beta_{0,n}||_2\leq||\tilde{\beta}_{GMM}-\beta_{0,n}||_2$, we have that

$$(\hat{M}(\beta_{0,n},\hat{\delta},\hat{\eta})+\partial_\beta\hat{M}(\bar{\beta},\hat{\delta},\hat{\eta})(\tilde{\beta}_{GMM}-\beta_{0,n}))'W_n\partial_\beta\hat{M}(\tilde{\beta}_{GMM},\hat{\delta},\hat{\eta})=0.$$

Therefore,

$$\sqrt{n}(\tilde{\beta}_{GMM}-\beta_{0,n})=$$

$$(\partial_\beta\hat{M}(\bar{\beta},\hat{\delta},\hat{\eta})'W_n\partial_\beta\hat{M}(\tilde{\beta}_{GMM},\hat{\delta},\hat{\eta}))^{-1}\partial_\beta\hat{M}(\tilde{\beta}_{GMM},\hat{\delta},\hat{\eta})'W_n\sqrt{n}\hat{M}(\beta_{0,n},\hat{\delta},\hat{\eta}).$$

Using the adaptivity condition of Lemma 2.1,

$$\sqrt{n}\hat{M}(\beta_{0,n},\hat{\delta},\hat{\eta})=\sqrt{n}\hat{M}(\theta_{0,n},\eta_{0,n})+o_p(1)\xrightarrow{d}N(0,V_M).$$

Again using Assumption 2.2.2, $\partial_\beta\hat{M}(\bar{\beta},\hat{\delta},\hat{\eta})-\partial_\beta\hat{M}(\theta_{0,n},\eta_{0,n})=o_p(1)$ and $\partial_\beta\hat{M}(\tilde{\beta}_{GMM},\hat{\delta},\hat{\eta})-$

$\partial_\beta \hat{M}(\theta_{0,n}, \eta_{0,n}) = o_p(1)$. Then since $\partial_\beta \hat{M}(\theta_{0,n}, \eta_{0,n}) = \partial_\beta M(\theta_{0,n}, \eta_{0,n}) + o_p(1) = M_\beta + o_p(1)$ and $W_n \xrightarrow{p} W$, we have that

$$\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n}) \xrightarrow{d} N(0, V),$$

where $(M_\beta' W M_\beta)^{-1} M_\beta' W V_M W M_\beta (M_\beta' W M_\beta)^{-1}$.

I now show the asymptotic normality of $\tilde{\beta}_{OS}$. Because $\hat{M}$ is twice continuously differentiable, using the Mean Value Theorem,

$$\sqrt{n}(\tilde{\beta}_{OS} - \beta_{0,n}) =$$

$$\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n} - (\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}))^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}))$$

$$= \sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n})$$

$$+ \sqrt{n}(\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}))^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \hat{M}(\beta_{0,n}, \hat{\delta}, \hat{\eta})$$

$$- \sqrt{n}(\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}))^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \partial_\beta \hat{M}(\bar{\beta}, \hat{\delta}, \hat{\eta})(\tilde{\beta}_{GMM} - \beta_{0,n}).$$

As shown above, $\partial_\beta \hat{M}(\beta_{0,n}, \delta_{0,n}, \eta_{0,n}) \xrightarrow{p} M_\beta$, $\partial_\beta \hat{M}(\bar{\beta}, \hat{\delta}, \hat{\eta}) \xrightarrow{p} M_\beta$, and $\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}) \xrightarrow{p} M_\beta$. Using this along with $W_n \xrightarrow{p} W$ and the adaptivity condition of Lemma 2.1, the second term is equal to

$$(\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}))^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})' W_n \sqrt{n} \hat{M}(\theta_{0,n}, \eta_{0,n}) + o_p(1)$$

$$= (M_\beta' W M_\beta)^{-1} M_\beta' W \sqrt{n} \hat{M}(\theta_{0,n}, \eta_{0,n}) + o_p(1)$$

and $(\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}) W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})')^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}) W_n \partial_\beta \hat{M}(\bar{\beta}, \hat{\delta}, \hat{\eta})' \xrightarrow{p} I_p$. Then because $\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n}) = O_p(1)$, $\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n})$ multiplied by

$$(I_p - (\partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}) W_n \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta})')^{-1} \partial_\beta \hat{M}(\tilde{\beta}_{GMM}, \hat{\delta}, \hat{\eta}) W_n \partial_\beta \hat{M}(\bar{\beta}, \hat{\delta}, \hat{\eta})'),$$

is converging in probability to zero. Therefore, we have that

$$\sqrt{n}(\tilde{\beta}_{OS} - \beta_{0,n}) =$$

$$(M_\beta' W M_\beta)^{-1} M_\beta' W \sqrt{n} \hat{M}(\theta_{0,n}, \eta_{0,n}) \xrightarrow{d} N(0, V).$$

Proof of Proposition 3.1: The proof proceeds by verifying Assumption 3.1 of Sun (2013) and then replicating the argument in the proof of Theorem 3.1 of Sun (2013) for the case with the sample moment conditions being $\hat{M}(\beta, \hat{\delta}, \hat{\eta})$ and with drifting sequences of the true parameter $\beta_{0,n}$. Proposition 2.1 holds so $||\tilde{\beta}_{GMM} - \beta_{0,n}||_1 \xrightarrow{p} 0$. Assumption 2.2 imposes that $\beta_{0,n}$ is an interior point of B bounded away from the boundary and Assumption 2.1 imposes that $\hat{g}$ is twice continuously differentiable in θ which implies that $\hat{M}(\beta, \hat{\delta}, \hat{\eta})$ is twice continuously differentiable in β. Then note that for $\lambda_n, r \in [0, 1]$,

$$\sum_{i=1}^{\lfloor rn \rfloor} \partial_\beta \hat{M}_i(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \hat{\delta}, \hat{\eta}) = \hat{\eta}(\sum_{i=1}^{\lfloor rn \rfloor} \partial_\beta g_i(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \hat{\delta})/n -$$

$$\hat{\eta} \sum_{i=1}^{\lfloor rn \rfloor} \partial_\beta g_i(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \delta_{0,n}))/n + \hat{\eta} \sum_{i=1}^{\lfloor rn \rfloor} \partial_\beta g_i(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \delta_{0,n})/n.$$

The difference between the first and second terms is converging in probability to zero uniformly in $\lambda_n, r \in [0, 1]$ by Assumption 3.2.1. Therefore, by Assumption 3.2.3 and $||\hat{\eta} - \eta_{0,n}||_1 = o_p(1)$, we have that

$$\sum_{i=1}^{\lfloor rn \rfloor} \partial_\beta \hat{M}_i(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \hat{\delta}, \hat{\eta}) \xrightarrow{p} r M_\beta,$$

uniformly in $\lambda_n, r \in [0, 1]$. Also, M_β is full rank by Assumption 2.2. By Lemma 2.1* in Appendix B and Assumption 3.2.4,

$$V_M^{-1/2} \sum_{t=1}^{n} \phi_k(\frac{t}{n}) \hat{\eta} g_t(\beta_{0,n}, \hat{\delta})/\sqrt{n}$$

$$= V_M^{-1/2}\sum_{t=1}^{n}(\phi_k(\frac{t}{n}) - \phi_k(\frac{t+1}{n}))(\sum_{i=1}^{t}\hat{\eta}g_i(\beta_{0,n},\hat{\delta})/\sqrt{n})$$

$$V_M^{-1/2}\sum_{t=1}^{n}(\phi_k(\frac{t}{n}) - \phi_k(\frac{t+1}{n}))(\sum_{i=1}^{t}\eta_{0,n}g_i(\theta_{0,n})/\sqrt{n}) + o_p(1)$$

$$= V_M^{-1/2}\sum_{t=1}^{n}\phi_k(\frac{t}{n})\eta_{0,n}g_t(\theta_{0,n})/\sqrt{n} + o_p(1) \xrightarrow{d} \xi_k,$$

for each $k \in \{0, ..., K\}$, where I have normalized $\phi_k(\frac{n+1}{n}) = 0$.

I now extend the proof of Theorem 3.1 of Sun (2013). Let $S_t(\beta) = \sum_{i=1}^{t}\hat{\eta}g_i(\beta,\hat{\delta})$. Then since the moment conditions are twice continuously differentiable in β,

$$S_t(\tilde{\beta}_{GMM})/\sqrt{n} = S_t(\beta_{0,n})/\sqrt{n} + (\sum_{i=1}^{t}\partial_\beta g_i(\tilde{\beta}_t)/n)\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n}),$$

where $\tilde{\beta}_t = \beta_{0,n} + \lambda_n \odot (\tilde{\beta}_{GMM} - \beta_{0,n})$ for some $\lambda_n \in [0,1]^p$ where $\odot$ denotes the element-wise product. From Proposition 2.1 and $W = I_m$,

$$\sqrt{n}(\tilde{\beta}_{GMM} - \beta_{0,n}) = \sqrt{n}(M_\beta' M_\beta)^{-1}\hat{M}(\beta_{0,n},\hat{\delta},\hat{\eta}) + o_p(1) = (M_\beta' M_\beta)^{-1}S_n(\beta_{0,n})/\sqrt{n} + o_p(1).$$

Therefore,

$$S_t(\tilde{\beta}_{GMM})/\sqrt{n} = S_t(\beta_{0,n})/\sqrt{n} + (\sum_{i=1}^{t}\partial_\beta g_i(\tilde{\beta}_t)/n)((M_\beta' M_\beta)^{-1}S_n(\beta_{0,n}) + o_p(1))$$

$$= S_t(\beta_{0,n})/\sqrt{n} - \frac{i}{n}S_n(\beta_{0,n})/\sqrt{n} + o_p(1),$$

uniformly over t. Then,

$$\sum_{i=1}^{n}\phi_k(\frac{i}{n})\hat{\eta}g_i(\tilde{\beta}_{GMM},\hat{\delta})/\sqrt{n} = \sum_{i=1}^{n}(\phi_k(\frac{i}{n}) - \phi_k(\frac{i+1}{n}))\frac{1}{\sqrt{n}}(S_i(\beta_{0,n}) - \frac{t}{n}S_n(\beta_{0,n})) + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_k(\frac{i}{n})(\hat{\eta}g_i(\beta_{0,n},\hat{\delta}) - \frac{1}{n}S_n(\beta_{0,n})) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_k(\frac{i}{n})\hat{\eta}g_i(\beta_{0,n},\hat{\delta}) + o_p(1).$$

Again, for convince defining $\phi_k(\frac{n+1}{n}) = 0$. Then

$$V_M^{-1/2}\sum_{i=1}^{n}\hat{\eta}\phi_k(\frac{t}{n})g_i(\tilde{\beta}_{GMM},\hat{\delta})/\sqrt{n} \xrightarrow{d} \xi_k,$$

jointly for $k = 0, 1, ..., K$. Therefore,

$$\frac{K-p+1}{K}\mathbb{W}_n \xrightarrow{d} \frac{K-p+1}{K}\xi_0'(\sum_{k=1}^{K}\xi_k\xi_k'/K)^{-1}\xi_0 = F_{p,K-p+1}$$

and

$$t_n \xrightarrow{d} \xi_0^2/\sqrt{\sum_{k=1}^{K}\xi_k^2/K} = t_K$$

when $p = 1$.

Proof of Proposition 4.1: I first verify that Assumption 3.1 holds when Assumptions 4.1 and 4.2 hold. Assumption 3.1.1 holds because $D_n = \Delta^J$ is compact for all J, $f(\theta,\eta) = ||\delta||_2^2 + ||\eta||_2^2$, and $\hat{g}$ and g are linear in δ. For Assumption 3.1.2, note that for each $q \in 1, ..., Q-1$,

$$\sup_{\delta\in\Delta^J}|Z_q^{pre}(Y_0^{pre'} - Y_{\mathcal{J}}^{pre'}\delta)/T_0 - E[Z_q^{pre}(Y_0^{pre'} - Y_{\mathcal{J}}^{pre'}\delta)/T_0]|$$

$$= \sup_{\delta\in\Delta^J}|Z_q^{pre}(Y_0^{pre'} - Y_{\mathcal{J}}^{pre'}\delta)/T_0 - E[Z_q^{pre}f^{pre}/T_0](\mu_0 - \mu_{\mathcal{J}}\delta)|$$

$$\leq \sup_{\delta\in\Delta^J}\{|(Z_q^{pre}f^{pre}/T_0 - E[Z_q^{pre}f^{pre}/T_0])(\mu_0 - \mu_{\mathcal{J}}\delta)'|$$

$$+|Z_q^{pre}(\epsilon_0^{pre'} - \epsilon_{\mathcal{J}}^{pre'}\delta)/T_0|\} \leq$$

$$\max_{0\leq j\leq J}||\mu_0 - \mu_j||_2||Z_q^{pre}f^{pre}/T_0 - E[Z_q^{pre}f^{pre}/T_0]||_2 + 2\max_{0\leq j\leq J}|Z_q^{pre}\epsilon_j^{pre'}/T_0|$$

$$= O(1)O_p(1/\sqrt{n}) + O_p(\log(J)/\sqrt{T_0}).$$

Similarly,

$$\sup_{\beta,\delta\in\Delta^J}|\sum_{t\in\mathcal{T}_1}(Y_{0t}-Y_{\mathcal{J},t}\delta)/T_1-\beta-(E[\sum_{t\in\mathcal{T}_1}f_t/T_1](\mu_0-\mu_{\mathcal{J}}\delta)+\beta_{0,n}-\beta)|$$

$$\leq\max_{0\leq j\leq J}||\sum_{t\in\mathcal{T}_1}f_t/T_1-E[\sum_{t\in\mathcal{T}}f_t/T_1]||_2||\mu_0-\mu_j||_2+2\max_{0\leq j\leq J}|\sum_{t\in\mathcal{T}_1}\epsilon_{jt}/T_1|$$

$$=O_p(1/\sqrt{T_1})O(1)+O_p(\log(J)/\sqrt{T_1}).$$

Therefore, $\sup_{\theta\in\Theta_n}||\hat{g}(\theta)-g(\theta)||_\infty=O_p(\log(J)/\sqrt{\min\{T_0,T_1\}})$. This also shows that

$$\sup_{\theta\in\Theta_n}||\partial_\delta\hat{g}(\theta)-\partial_\delta g(\theta)||_\infty=O_p(\log(J)/\sqrt{\min\{T_0,T_1\}}),$$

so Assumption 3.1.2 holds with $a_n,b_n=\log(J)/\sqrt{\min\{T_0,T_1\}}$. Because $\hat{f}(\theta,\eta)=f(\theta,\eta)$, Assumption 3.1.3 holds trivially with $c_n=0$.

By Assumption 4.2.3, there exists a subset of $\mathcal{I}$ instruments with $|\mathcal{I}|=R$ and the minimum singular value of $E[Z_{\mathcal{I}}^{pre}f^{pre}/T_0]$ is bounded below by some constant $C>0$. For Assumption 3.2.4, with loss of generality, assume that the row of Z^{pre} are order so that the indices in $\mathcal{I}$ correspond to the first R rows of Z^{pre}. Note that for any δ, there exists $\tilde{\delta}$ where $\tilde{\delta}_j=\delta_j$ for $j>R+1$ and $\mu_0=\mu_{\mathcal{J}}\tilde{\delta}$ because μ_0 has R elements. Then,

$$E[Z^{pre}f^{pre}/T_0](\mu_0-\mu_{\mathcal{J}}\delta)=E[Z^{pre}f^{pre}/T_0](\mu_0-\mu_{\mathcal{J}}\tilde{\delta})-E[Z^{pre}f^{pre}/T_0]\mu_{\mathcal{J}}(\delta-\tilde{\delta})$$

$$=E[Z^{pre}f^{pre}/T_0]\mu_{\mathcal{I}}(\tilde{\delta}_{\mathcal{I}}-\delta_{\mathcal{I}}).$$

Therefore, following page 1262 of Chernozhukov et al. (2007),

$$||g(\theta)||_\infty\geq||g(\theta)||_2/\sqrt{Q}\geq||E[Z^{pre}f^{pre}/T_0](\mu_0-\mu_J\delta)||_2/\sqrt{Q}\geq C||\tilde{\delta}_{\mathcal{I}}-\delta_{\mathcal{I}}||_2/\sqrt{Q}$$

$$\geq C/\sqrt{QR}||\tilde{\delta}_{\mathcal{I}}-\delta_{\mathcal{I}}||_1=C/\sqrt{QR}||\tilde{\delta}-\delta||_1\geq C/\sqrt{RQ}||\delta-D_{0,n}||_1$$

where the last inequality follows from $\tilde{\delta} \in D_{0,n}$. Similarly, note that for any $\eta \in H$, there exists $\tilde{\eta}$ such that $\tilde{\eta} \in H_{0,n}$ for all n and $\eta_q = \tilde{\eta}_q$ for all $q > R$. Then

$$||\eta \partial_\delta g(\delta)||_\infty \geq ||(\eta - \tilde{\eta})\partial_\delta g(\delta)||_\infty = ||(\eta - \tilde{\eta})E[Z^{pre} f^{pre}/T_0]||_\infty$$

$$\geq C||(\eta_R - \tilde{\eta}_R)E[Z_{\mathcal{I}}^{pre} f^{pre}/T_0]||_2/\sqrt{R}$$

$$\geq C||\eta_{\mathcal{I}} - \tilde{\eta}_{\mathcal{I}}||_2/\sqrt{R} \geq C||\eta_{\mathcal{I}} - \tilde{\eta}_{\mathcal{I}}||_1/R = C/R||\eta - \tilde{\eta}||_1,$$

where C is equal to the smallest singular value $\partial_{\delta_{\mathcal{I}}} g(\delta)$ and the equality follows from the fact that $\eta_Q = \tilde{\eta}_Q = 1$. Therefore, Assumption 3.1.4 holds.

For the first part of Assumption 3.1.5, note that the for each n the identified set $D_{0,n} \times H_{0,n}$ and the set

$$S = \{(\delta, \eta) : ||\delta||_2^2 + ||\eta||_2^2 \leq ||\delta_{0,n}||_2^2 + ||\eta_{0,n}||_2^2\}$$

are convex. By the separating hyperplane theorem, there exists a hyperplane described by the linear equations $Ax = b$ such that for all $\gamma = (\delta, \eta) \in S$ we have that $A\gamma \geq b$ and for all $\gamma \in D_{0,n} \times H_{0,n}$, $A\gamma \leq b$. Then since $(\delta_{0,n}, \eta_{0,n}) \in S$ and $(\delta_{0,n}, \eta_{0,n}) \in D_{0,n} \times H_{0,n}$, $A(\delta_{0,n}, \eta_{0,n}) = b$ so $(\delta_{0,n}, \eta_{0,n})'\gamma \geq 0$ for all γ in the half space with $A\gamma \leq b$. Then it follows from the Law of Cosines that

$$||\gamma||_2^2 - ||(\delta_{0,n}, \eta_{0,n})||_2^2 \geq ||\gamma - (\delta_{0,n}, \eta_{0,n})||_2^2$$

for all $\gamma \in D_{0,n} \times H_{0,n}$. Because $\delta_{0,n}$ only takes non-zero values for the elements in $\mathcal{P}$ and $||\delta_{0,n}||_1 = ||\delta||_1 = 1$, so

$$||\delta - \delta_{0,n}||_1 = ||\delta_{\mathcal{P}} - \delta_{0,n,\mathcal{P}}||_1 + ||\delta_{\mathcal{P}^c}||_1 = ||\delta_{\mathcal{P}} - \delta_{0,n,\mathcal{P}}||_1 + (1 - ||\delta_{\mathcal{P}}||_1)$$

$$= ||\delta_{\mathcal{P}} - \delta_{0,n,\mathcal{P}}||_1 + (||\delta_{0,n,\mathcal{P}}||_1 - ||\delta_{\mathcal{P}}||_1) \leq 2||\delta_{\mathcal{P}} - \delta_{0,n,\mathcal{P}}||_1 \leq 2\sqrt{|\mathcal{P}|}||\delta_{\mathcal{P}} - \delta_{0,n,\mathcal{P}}||_2,$$

where the second to last inequality follows from the reverse triangle inequality. Since the dimension of η is always equal to Q, $||\eta - \eta_{0,n}||_1 \leq \sqrt{Q}||\eta - \eta_{0,n}||_2$. Therefore,

$$||\gamma||_2^2 - ||(\delta_{0,n}, \eta_{0,n})||_2^2 \geq ||\gamma - (\delta_{0,n}, \eta_{0,n})||_2^2 \geq (2\sqrt{|\mathcal{P}|}||\delta - \delta_{0,n}||_1 + \sqrt{Q}||\eta - \eta_{0,n}||_1)^2.$$

Therefore, we have that for all n, for all $\theta \in \Theta_{0,n}$ and $\eta \in H_{0,n}$,

$$|f(\theta, \eta) - f(\theta_{0,n}, \eta_{0,n})| \geq C_4(||\delta - \delta_{0,n}||_1 + ||\eta - \eta_{0,n}||_1)^2 = C_4(||\theta - \theta_{0,n}||_1 + ||\eta - \eta_{0,n}||_1)^2$$

for some constant $C_4 > 0$ which does not depend on θ, η, or n.

For the second part of Assumption 3.1.5, note that for any $\delta_1, \delta_2 \in D_n$,

$$2||\delta_1 - \delta_2||_1 \geq 2||\delta_1 - \delta_2||_2 \geq (||\delta_1||_2 + ||\delta_2||_2)|\ ||\delta_1||_2 - ||\delta_2||_2| = 2|\ ||\delta_1||_2^2 - ||\delta_2||_2^2|.$$

Similarly, for any $\delta_1, \delta_2 \in D_n$ and $\eta_1, \eta_2 \in H$, if $C_5 \geq ||\delta_1||_2^2 + ||\eta_1||_2^2 \geq ||\eta_1||_2^2$ and $C_5 \geq ||\delta_2||_2^2 + ||\eta_2||_2^2 \geq ||\eta_2||_2^2$ for some $C_5 > 0$, then,

$$2C_5||\eta_1 - \eta_2||_1 \geq 2C_5||\eta_1 - \eta_2||_2 \geq (||\eta_1||_2 + ||\eta_2||_2)|\ ||\eta_1||_2 - ||\eta_2||_2\ | = |\ ||\eta_1||_2^2 - ||\eta_2||_2^2\ |.$$

Then for any n and any $\theta_1, \theta_2 \in \Theta_n$ and $\eta_1, \eta_2 \in H$ with $f(\theta_1, \eta_1), f(\theta_2, \eta_2) \leq C_5$,

$$||\theta_1 - \theta_2||_1 + ||\eta_1 - \eta_2||_1 \geq ||\delta_1 - \delta_2||_1 + ||\eta_1 - \eta_2||_1 \geq |f(\theta_1, \eta_1) - f(\theta_2, \eta_2)|/\max\{2C_5, 2\}.$$

Therefore, Assumption 3.1.5 holds with $\gamma_1 = 2$ and $\gamma_2 = 1$. I now verify the conditions of Assumption 2.1. Since Assumption 3.1 holds, by Lemma 3.1 we have that

$$||\hat{\delta} - \delta_{0,n}||_1 = O_p(\max\{\log(J)/\sqrt{\min\{T_0, T_1\}}, \lambda_\delta, \lambda_\eta\}^{\frac{1}{2}}) = o_p(1/(\log(J)\log(\min\{T_0, T_1\}))).$$

For Assumption 2.1.4, $\hat{g}$ is linear in θ and

$$||\hat{\eta}\partial_\delta \hat{g}(\theta)||_\infty \leq \lambda_\eta = O_p(\log(J)\log(\min\{T_0,T_1\})/\sqrt{\min\{T_0,T_1\}}).$$

Assumption 2.1.3 holds trivially since $\hat{g}$ is linear in θ and Assumption 2.1.2 is shown above. To verify Assumption 2.1*, first note that $\eta_{0,n}$ satisfies,

$$\sum_{t\in\mathcal{T}_0}\sum_{q=1}^{Q-1} E[\eta_{0,n,q}Z_{qt}Y_{\mathcal{J},t}]/T_0 + \sum_{t\in\mathcal{T}_1} E[Y_{\mathcal{J},t}]/T_1 =$$

$$(\sum_{t\in\mathcal{T}_0}\sum_{q=1}^{Q-1} E[\eta_{0,n,q}Z_{qt}f_t]/T_0 + \sum_{t\in\mathcal{T}_1} E[f_t]/T_1)\mu_{\mathcal{J}} = 0.$$

Therefore,

$$\sum_{t\in\mathcal{T}_0}\sum_{q=1}^{Q-1} E[\eta_{0,n,q}Z_{qt}f_t]/T_0 + \sum_{t\in\mathcal{T}_1} E[f_t]/T_1 = 0,$$

for each $\eta_{0,n}$. As a result, Assumption 4.2.3 implies that $\eta_{0,n} \to \eta^*$ for some η^*. Then for partial sums of the population moment conditions with pre-treatment time periods $\{-t_0, ..., -1\}$ and post-treatment time periods $\{0, ..., t_1 - 1\}$,

$$\sum_{i=-t_0}^{-1}\sum_{q=1}^{Q-1} E[\eta_{0,n,q}Z_{qi}Y_{\mathcal{J},i}]/T_0 + \sum_{i=0}^{t_1-1} E[Y_{\mathcal{J},i}]/T_1 \to 0$$

as $T_0, T_1 \to \infty$, uniformly over $t_0 \leq T_0$ and $t_1 \leq T_1$, so Assumption 2.1.1* holds. For Assumption 2.1.2*,

$$||\sum_{i=-t_0}^{-1}\sum_{q=1}^{Q-1} \eta_{0,n,q}Z_{qi}Y_{\mathcal{J},i}/T_0 + \sum_{i=0}^{t_1-1} Y_{\mathcal{J},i}/T_1 - \sum_{i=-t_0}^{-1}\sum_{q=1}^{Q-1} E[\eta_{0,n,q}Z_{qi}Y_{\mathcal{J},i}]/T_0 - \sum_{i=0}^{t_1-1} E[Y_{\mathcal{J},i}]/T_1||_\infty \leq$$

$$||\eta_{0,n}||_1(||\sum_{i=-t_0}^{-1}\sum_{q=1}^{Q-1} Z_{qi}\epsilon_{\mathcal{J},i}/T_0||_\infty + ||\sum_{i=-t_0}^{-1}\sum_{q=1}^{Q-1}(E[Z_{qi}f_i] - Z_{qi}f_i)\mu_{\mathcal{J}}/T_0||_\infty$$

$$+||\sum_{i=0}^{t_1-1}(E[f_i]-f_i)\mu_{\mathcal{J}}/T_1||_\infty + ||\sum_{i=0}^{t_1-1}\epsilon_{\mathcal{J},i}/T_1||_\infty)$$

$$= O(1)O_p(\log(J)/\sqrt{T_0}) + O(1)O_p(\log(J)/\sqrt{T_1}) = O_p(\log(J)/\sqrt{\min\{T_0,T_1\}})$$

uniformly over $1 \leq t_0 \leq T_0$ and $1 \leq t_1 \leq T_1$ as $T_0, T_1 \to \infty$. Assumption 2.1.3* holds trivially because $\hat{g}$ is linear in θ.

I now verify the conditions of Assumption 2.2. Since

$$\sup_{\theta\in\Theta_n} ||\hat{g}(\theta) - g(\theta)||_2 = O_p(\log(J)/\sqrt{\min\{T_0,T_1\}})$$

and $\log(J)/\sqrt{\min\{T_0,T_1\}} \to 0$, Assumption 2.2.1 holds. Because $g(\theta)$ and $\hat{g}(\theta)$ are linear in θ, Assumption 2.2.2 holds. Also $\partial_\beta M(\beta_{0,n}, \delta_{0,n}, \eta_{0,n}) = -1 \neq 0$ and $||\beta_1 - \beta_2||_2 = ||M(\beta_1, \delta_{0,n}, \eta_{0,n}) - M(\beta_2, \delta_{0,n}, \eta_{0,n})||_2$ for all $\beta_1, \beta_2 \in B$ so Assumption 2.2.4 holds. Then for Assumption 2.2.5, since there is single moment condition we can set $W_n = W = 1$ and

$$\hat{M}(\beta, \delta_{0,n}, \eta_{0,n}) - M(\beta, \delta_{0,n}, \eta_{0,n}) = \eta_{0,n}(\hat{g}(0, \delta_{0,n}) - g(0, \delta_{0,n})) \xrightarrow{p} 0$$

as shown above. I prove Assumption 2.2.3 along with Assumption 3.2.

Assumption 4.3 directly guarantees that Assumption 3.2.2 is satisfied. For Assumption 3.2.1, it holds trivially because $g_i(\theta)$ is linear in θ. Similarly for Assumption 3.2.3, $\partial_\beta \hat{M}(\beta, \delta, \eta) = \eta_Q$, so

$$\sum_{i=1}^{\lfloor rn \rfloor} \partial_\beta \hat{M}(\beta_{0,n} + \lambda_n(\tilde{\beta}_{GMM} - \beta_{0,n}), \hat{\delta}, \hat{\eta})/n - r\partial_\beta M(\theta_{0,n}, \eta_{0,n}) = -r\hat{\eta}_Q + r\eta_{0,n,Q} \xrightarrow{p} 0$$

where $\partial_\beta M(\theta_{0,n}, \eta_{0,n}) = -\eta_{0,n,Q}$ is full rank since $\eta_{0,n,Q} = 1 \neq 0$.

Note that we can equivalently define the optimal control weights as

$$\delta_{0,n} = \arg\min_{\delta\in\Delta^J:\mu_0=\mu_{\mathcal{J}}\delta} ||\delta||_2^2.$$

However, the elements of $\delta_{0,n}$ are only taking non-zero values for indices in $\mathcal{P}$ by Assumption 4.3.3, $\delta_{0,n}$ must also be equal to zero for all indices not in $\mathcal{P}$ and if $\mathcal{P} \subseteq \mathcal{J}$, then

$$\delta_{0,n,\mathcal{P}} = \underset{\delta \in \Delta^{|\mathcal{P}|}: \mu_0 = \mu_{\mathcal{P}}\delta}{\arg\min} ||\delta||_2^2,$$

where $\delta_{0,n,\mathcal{P}}$ denotes the subvector of $\delta_{0,n}$ with indices in $\mathcal{P}$. As a result, for sufficiently large J so that $\mathcal{P} \subseteq \mathcal{J}$, $\delta_{0,n,\mathcal{P}}$ does not vary with J or n.

Then for Assumption 3.2.4 and 2.2.3, first note that for $q \in \{1, .., Q-1\}$ and $k \in \{0, 1, ..., K\}$,

$$\sum_{t \in \mathcal{T}_0} \phi_k(\frac{t}{T_0}) g_{q,t}(\theta_{0,n})/\sqrt{T_0} = \sum_{t \in \mathcal{T}_0} \phi_k(\frac{t}{T_0}) Z_{qt}(\epsilon_{0t} - \sum_{j \in \mathcal{P}} \delta_{0,n,j}\epsilon_{jt})/\sqrt{T_0},$$

$$= \sum_{t \in \mathcal{T}_0} \phi_k(\frac{t}{T_0}) Z_{qt}\epsilon_{0t}/\sqrt{T_0} + \sum_{j \in \mathcal{P}} \delta_{0,n,j} \sum_{t \in \mathcal{T}_0} \phi_k(\frac{t}{T_0}) Z_{qt}\epsilon_{jt}/\sqrt{T_0}$$

and similarly for $q = Q$,

$$\sum_{t \in \mathcal{T}_1} \phi_k(\frac{t}{T_1}) g_{q,t}(\theta_{0,n})/\sqrt{T_1} = \sum_{t \in \mathcal{T}_1} \phi_k(\frac{t}{T_1})(\beta_t - \beta_{0,n}) + \sum_{t \in \mathcal{T}_1} \phi_k(\frac{t}{T_1})\epsilon_{0t} + \sum_{j \in \mathcal{P}} \delta_{0,n,j} \sum_{t \in \mathcal{T}_1} \phi_k(\frac{t}{T_1})\epsilon_{jt}/\sqrt{T_1}$$

because $\mu_0 = \mu_{\mathcal{J}}\delta_{0,n}$ and the sparsity of $\delta_{0,n}$. As noted earlier, for T_0 and T_1 sufficiently large, $\delta_{0,n}$ does not vary with n. Therefore, the limit in equation (10) does exist and is equal to V_g, for some fixed positive definite matrix V_g. Then because of Assumption 4.3.1 and $T_1/T_0 \to a > 0$, the conditions of Lemma B3 are satisfied for the sequence of random vectors

$$\begin{pmatrix} \sqrt{\frac{\min\{T_0, T_1\}}{T_0}} g_{1t}(\theta_{0,n}) \\ ... \\ \sqrt{\frac{\min\{T_0, T_1\}}{T_0}} g_{q-1.t}(\theta_{0,n}) \end{pmatrix}$$

with $t \in \mathcal{T}_0$ and the sequence of random variables $\sqrt{\frac{\min\{T_0, T_1\}}{T_1}} g_{Qt}(\theta_{0,n})$ with $t \in \mathcal{T}_1$. Then, by

Lemma B3, we have that

$$V_g^{-1/2}\sqrt{\min\{T_0,T_1\}}\begin{pmatrix}\sum_{t\in\mathcal{T}_0}\phi_k(\frac{t}{T_0})g_{1,t}(\theta_{0,n})/T_0\\ \cdots\\ \sum_{t\in\mathcal{T}_0}\phi_k(\frac{t}{T_0})g_{Q-1,t}(\theta_{0,n})/T_0\\ \sum_{t\in\mathcal{T}_1}\phi_k(\frac{t}{T_1})g_{Q,t}(\theta_{0,n})/T_1\end{pmatrix}$$

$$= V_g^{-1/2}\begin{pmatrix}\sum_{t\in\mathcal{T}_0}\phi_k(\frac{t}{T_0})\sqrt{\frac{\min\{T_0,T_1\}}{T_0}}g_{1,t}(\theta_{0,n})/\sqrt{T_0}\\ \cdots\\ \sum_{t\in\mathcal{T}_0}\phi_k(\frac{t}{T_0})\sqrt{\frac{\min\{T_0,T_1\}}{T_0}}g_{Q-1,t}(\theta_{0,n})/\sqrt{T_0}\\ \sum_{t\in\mathcal{T}_1}\phi_k(\frac{t}{T_1})\sqrt{\frac{\min\{T_0,T_1\}}{T_1}}g_{Q,t}(\theta_{0,n})/\sqrt{T_1}\end{pmatrix}\xrightarrow{d}\zeta_k,$$

jointly for $k\in\{0,1,...,K\}$ with $\zeta_k\sim iidN(0,I_Q)$. Then since $V_M(\theta_{0,n},\eta_{0,n})=\eta_{0,n}V_g(\theta_{0,n})\eta_{0,n}$ and $\hat{M}(\theta_{0,n},\eta_{0,n})=\eta_{0,n}\hat{g}(\theta_{0,n})$ we also have that

$$V_M^{-1/2}\sqrt{\min\{T_0,T_1\}}\eta_{0,n}\begin{pmatrix}\sum_{t\in\mathcal{T}_0}\phi_k(\frac{t}{T_0})g_{1,t}(\theta_{0,n})/T_0\\ \cdots\\ \sum_{t\in\mathcal{T}_0}\phi_k(\frac{t}{T_0})g_{Q-1,t}(\theta_{0,n})/T_0\\ \sum_{t\in\mathcal{T}_1}\phi_k(\frac{t}{T_1})g_{Q,t}(\theta_{0,n})/T_1\end{pmatrix}\xrightarrow{d}\xi_k,$$

jointly for $k\in\{0,1,...,K\}$ with $\xi_k\sim iidN(0,1)$ so Assumptions 2.2.3 and 3.2.4 hold.

Appendix B

Lemma B1 (Rate of Convergence of the Estimated Identified Set) Suppose that the conditions of Lemma 3.1 hold. Let $S_0:=\{(\theta,\eta)\in\Theta_{0,n}\times H_{0,n}:f(\theta,\eta)\leq f(\theta_{0,n},\eta_{0,n})+\zeta\}$ and $\hat{S}_0:=\{(\theta,\eta)\in\hat{\Theta}_0\times\hat{H}_0:f(\theta,\eta)\leq f(\theta_{0,n},\eta_{0,n})+\zeta\}$ for some $\zeta>0$. Then $d_H(\hat{S}_0,S_0,||\cdot||_1)=O_p(\max\{a_n\lambda_\delta,\lambda_\eta\})$.

Proof:

Note that by the identification condition on g in Assumption 3.1.4, for any $(\theta,\eta)\in\hat{S}_0$,

$$C_2 \min\{||\theta - \Theta_{0,n}||_1 + ||\eta - H_{0,n}||_1, C_1\} \leq ||g(\theta)||_\infty + ||\eta \partial_\delta g(\theta)||_\infty \leq$$

$$||\hat{g}(\theta)||_\infty + ||g(\theta) - \hat{g}(\theta)||_\infty + ||\eta \partial_\delta \hat{g}(\theta)||_\infty + ||\eta(\partial_\delta g(\theta) - \partial_\delta \hat{g}(\theta))||_\infty$$

$$\leq \lambda_\delta + ||g(\theta) - \hat{g}(\theta)||_\infty + \lambda_\eta + ||\eta||_1 ||\partial_\delta g(\theta) - \partial_\delta \hat{g}(\theta)||_\infty.$$

By Assumption 3.1.1, $\sup_{(\theta,\eta)\in\hat{S}_0} ||\eta||_1 \leq \sup_{(\theta,\eta)\in\Theta_n\times H: f(\theta,\eta)\leq f(\theta_{0,n},\eta_{0,n})+\zeta} ||\eta||_1 = O(1)$. Then because $\lambda_\delta, \lambda_\eta, \sup_{\theta\in\Theta_n} ||g(\theta) - \hat{g}(\theta)||_\infty, \sup_{\theta\in\Theta_n} ||\partial_\delta g(\theta) - \partial_\delta \hat{g}(\theta)||_\infty = o_p(1)$, this implies that $\sup_{(\theta,\eta)\in\hat{S}_0} ||\theta - \Theta_{0,n}||_1 + ||\eta - H_{0,n}||_1 < C_1$ wpa1. Then wpa1,

$$\sup_{(\theta,\eta)\in\hat{S}_0} ||\theta - \Theta_{0,n}||_1 + ||\eta - H_{0,n}||_1 \leq \lambda_\delta + \lambda_\eta + \sup_{\theta\in\Theta_n} ||g(\theta) - \hat{g}(\theta)||_\infty + \sup_{(\theta,\eta)\in\hat{S}_0} ||\eta\|_1 ||\partial_\delta g(\theta) - \partial_\delta \hat{g}(\theta)||_\infty$$

$$= \lambda_\delta + \lambda_\eta + O_p(a_n) + O(1)O_p(a_n) = O_p(\max\{\lambda_\delta, \lambda_\eta, a_n\}).$$

Also note that since $\sup_{(\theta,\eta)\in S_0} ||\hat{g}(\theta)||_\infty \leq \sup_{\theta\in\Theta_{0,n}} ||\hat{g}(\theta)||_\infty = O_p(b_n)$,

$$\sup_{(\theta,\eta)\in S_0} ||\partial_\delta \eta g(\theta)||_\infty \leq \sup_{(\theta,\eta)\in S_0} ||\eta||_1 ||\partial_\delta \hat{g}(\theta) - \partial_\delta g(\theta)||_\infty = O(1)O_p(b_n) = O_p(b_n),$$

and $\max\{\lambda_\delta, \lambda_\eta\} b_n \to 0$, $\sup_{\theta\in\Theta_{0,n}} ||\hat{g}(\theta)||_\infty < \lambda_\delta$ and $\sup_{\theta\in\Theta_{0,n},\eta\in H_{0,n}} ||\partial_\delta \eta \hat{g}(\theta)||_\infty < \lambda_\eta$ wpa1. Hence, $S_0 \subset \hat{S}_0$ wpa1. Therefore, $d_H(S_0, \hat{S}_0, ||\cdot||_1) = O_p(\max\{a_n, \lambda_\delta, \lambda_\eta\})$.

Lemma B2 (Convergence of Maximum of Averages) For random vectors $\{X_i\}_{i\in\mathbb{N}}$ with X_i is taking values in $\mathbb{R}^J$ where J is growing with n, suppose that $\{X_i\}_{i\in\mathbb{N}}$ is α-mixing with exponentially decaying mixing coefficients, mean-zero, bounded four moments, and there exists constants $C_1, C_2, q > 0$ such that $\sup_i P(|Z_{ji}| > a) \leq c_1 exp(-c_2 a^q))$ for all $a > 0$. Then as $n, J \to \infty$, $\max_{1\leq j\leq J} |\sum_{i=1}^n X_{jt}/n| = O_p(\log(J)/\sqrt{n})$.

Proof: By Lemma 1 of Dendramis et al. (2021),

$$P(|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ji}| > a) \leq c_3[\exp(-c_4 a^2) + \exp(-c_5(\frac{a\sqrt{T_0}}{\log^2 T_0})^{q/(q+1)})]$$

for all $a > 0$ where c_3, c_4, and c_5 do not depend on i and j. Then let $a = \kappa \log J/2$ for some κ so that,

$$P(\max_{i\in\{1,...,J\}} |\frac{1}{n}\sum_{i=1}^{n} X_{ji}| > \kappa \log J/2) \leq$$

$$\sum_{j=1}^{J} P(|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ji}| > \kappa \log J/2\sqrt{n}) \leq$$

$$Jc_3 \exp(-c_4(\kappa/2)^2 \log J) + Jc_3 \exp(-c_5(\frac{\kappa \log J/2\sqrt{n}}{\log^2 n})^{q/(q+1)}) := r_J + r_{J,n}.$$

First consider the case where $J \to \infty$. Let $\gamma > 0$. Then we can choose κ such that $c_4(\kappa/2)^2 > 1 + \gamma$. Then $r_J \leq J \exp(-(1+\gamma)\log J) = c_3 J^{-\gamma} \to 0$ as $J \to \infty$.

Furthermore, if $J = o(n^{\zeta})$ for some $\delta > 0$. Then we have $n^{\frac{1}{4}} \geq J^{\frac{1}{4\zeta}}$ and $n^{\frac{1}{4}} > 2\log^2(n)$ as $n \to \infty$. Then

$$c_5(\frac{\kappa \log J/2\sqrt{n}}{\log^2 n})^{q/(q+1)} \geq c_5(\kappa(J^{\frac{1}{4\zeta}}\sqrt{\log J})^{q/(q+1)} > (1+\gamma)\log J$$

as $J \to \infty$. Therefore, $0 \leq r_{J,n} \leq r_J \to 0$ as $n, J \to \infty$. Hence $\max_{j\geq 1} |\frac{1}{n}\sum_{i=1}^{n} X_{ji}| = o_p(1)$.

Applying Lemma B2: Suppose $(Z_t, \epsilon_t)_{t\in\mathbb{Z}}$ is α-mixing with exponentially decaying mixing coefficients, $E[\epsilon_{jt}|Z_{qt}] = 0$ for all $q \in \{1, ..., Q-1\}$, $j \in \{1, ..., J\}$, $\sup_{i,t} E[\epsilon_{it}^4] < \infty$ and $\sup_{q,t} E[Z_{qt}^4] < \infty$, and

$$\sup_{i,t} P(|\epsilon_{it}| > a) \leq c_1 \exp(-c_2 a^{q_1}) \text{ and } \sup_{f,t} P(|Z_{qt}| > a) \leq c_1 \exp(-c_2 a^{q_2})$$

for all $a > 0$ for some $q_1, q_2 > 0$ and $c_1, c_2 > 0$ which do not depend on i, t and q. By Lemma A4 in Dendramis et al. (2021), an exponential tail bound also holds for products of the idiosyncratic shocks and factors so, for example, $\sup_{i,t,k} P(|Z_{qt}\epsilon_{it}| > a) \leq c_1 exp(-c_2 a^q)$

for all $a > 0$ where $q = q_1 q_2/(q_1 + q_2)$. Also $\sup_{q,j,t} E[(Z_{qt}\epsilon_{jt})^4] = \sup_{k,j,t} E[Z_{kt}^4]^{\frac{1}{2}} E[\epsilon_{jt}^4]^{\frac{1}{2}} < \infty$, $E[\epsilon_{jt} Z_{qt}] = E[\epsilon_{jt}|Z_{qt}]E[Z_{qt}] = 0$. As a result, the conditions of Lemma B2 is satisfied so $\max_{1\leq j\leq J} |\sum_{t\in\mathcal{T}_0} Z_{qt}\epsilon_{jt}/T_0| = O_p(\log(J)/\sqrt{T_0})$ for each $q \in \{1, ..., Q-1\}$. The same reasoning can be applied to show that $\max_{1\leq j\leq J} |\sum_{t\in\mathcal{T}_1} \epsilon_{jt}/T_1| = O_p(\log(J)/\sqrt{T_1})$.

Lemma B3 (Applying a Functional Central Limit Theorem) Suppose $\{\phi_k(x)\}_{k=0}^K$ satisfies Assumption 3.2.2 and there is a stochastic process $\{X_i\}_{i\in\mathcal{N}}$ with $E[X_i] = 0$, $E[X_i^2] < \infty$ for all $i \in \mathbb{N}$, and $E[(\sum_{i=1}^n X_i)^2/n] \to \sigma^2$ for some $\sigma^2 > 0$. Further suppose that the sequence is α-mixing with mixing coefficients $\alpha(k)$ and there exists $\gamma > 2$ such that $\sup_{i\in\mathbb{N}} E[|X_i|^\gamma] < \infty$ and $\sum_{k=1}^\infty \alpha(k)^{1-2/\gamma} < \infty$. Then

$$\sigma^{-1/2}\sum_{i=1}^n \phi_k(\frac{i}{n})X_i/\sqrt{n} \xrightarrow{d} \xi_k,$$

jointly for $k \in \{0, 1, ..., K\}$ with $\xi_k \sim iidN(0, 1)$.

Proof: The conditions of Theorem 0 of Herrndorf (1985) are satisfied, which provides a Functional Central Limit Theorem for partial sums for α-mixing processes. Therefore, the partial sums function $\hat{B}(r) = \sum_{i=1}^{\lfloor rn\rfloor} X_i/(\sigma\sqrt{n})$ converges weakly to the standard Wiener measure.

As pointed out by Phillips (2005) (see page 119), when Assumption 3.2.2 holds and $\hat{B}(r)$ is converging weakly to the standard Wiener measure, then standard functional limit arguments and Wiener integration show that

$$\sigma^{-1/2}\sum_{i=1}^n \phi_k(\frac{i}{n})X_i/\sqrt{n} \xrightarrow{d} \int_0^1 \phi_k(r)dB(r) = \xi_k,$$

where $\xi_k \sim N(0, 1)$, jointly for $k \in \{0, 1, ..., K\}$. Then, due to the orthogonality property of the basis functions, these ξ_k are uncorrelated and since they are jointly normal this also means that they are independent.

Lemma 2.1* (Partial Sums Adaptivity Condition) Suppose $(\beta_{0,n}, \delta_{0,n}) \in \Theta_{0,n}$, $\eta_{0,n}$ satisfies equation (3), and Assumptions 2.1 and 2.1* hold. Then, uniformly over t with

$1 \leq t \leq n$,

$$\sqrt{n}(\sum_{i=1}^{t} \hat{\eta} g_i(\beta_{0,n}, \hat{\delta})/n - \sum_{i=1}^{t} \eta_{0,n} g_i(\beta_{0,n}, \delta_{0,n})/n) = o_p(1).$$

Proof: Let $\gamma = (\delta, \eta_q)$ and $\hat{\gamma} = (\hat{\delta}, \hat{\eta}_q)$ where η_q is the q-th row of η. Since the q-th element on the orthogonalized sample moment conditions $\hat{M}_q$ for $q \in \{1, ..., m\}$ are twice continuously differentiable in γ, for each q there exists $\bar{\gamma}_t$ for each t with $1 \leq t \leq n$ and $||\bar{\gamma}_t - \gamma_0||_1 \leq ||\hat{\gamma} - \gamma_0||_1$ such that:

$$\sqrt{n}(\sum_{i=1}^{t} M_{qi}(\beta_{0,n}, \hat{\gamma})/n - \sum_{i=1}^{t} M_{qi}(\beta_{0,n}, \gamma_0)/n) = \sqrt{n}\partial_\gamma \sum_{i=1}^{t} M_{qi}(\beta_{0,n}, \gamma_0)/n(\hat{\gamma} - \gamma_0)$$

$$+\sqrt{n}(\hat{\gamma} - \gamma_0)'\partial_\gamma^2 \sum_{i=1}^{t} M_{qi}(\beta_{0,n}, \bar{\gamma}_t)/n(\hat{\gamma} - \gamma_0).$$

The magnitude of the first term on the right-hand side is less than or equal to

$$\sqrt{n}||\partial_\gamma \sum_{i=1}^{t} M_{qi}(\beta_{0,n}, \gamma_0)/n||_\infty ||\hat{\gamma} - \gamma_0||_1 \leq \sqrt{n} O_p(\log(J)/\sqrt{n}) o_p(1/\log(J)) = o_p(1).$$

The second term on the right-hand side is equal to

$$\sqrt{n}(\hat{\delta} - \delta_{0,n})'\partial_\delta^2 \hat{M}_q(\beta_{0,n}, \bar{\gamma}_t)(\hat{\delta} - \delta_{0,n}) + 2\sqrt{n}(\hat{\eta}_q - \eta_{0,n,q})\partial_\delta \sum_{i=1}^{t} g_i(\beta_{0,n}, \bar{\delta})/n(\hat{\delta} - \delta_{0,n}).$$

In the linear case, $\sqrt{n}(\hat{\delta} - \delta_{0,n})'\partial_\delta^2 \sum_{i=1}^{t} M_{qi}(\beta_{0,n}, \bar{\gamma}_t)/n(\hat{\delta} - \delta_{0,n}) = 0$ and

$$\partial_\delta \sum_{i=1}^{t} g_i(\beta_{0,n}, \bar{\delta})/n = \partial_\delta \sum_{i=1}^{t} g_i(\theta_{0,n})/n = \partial_\delta \sum_{i=1}^{t} g_i(\hat{\theta})/n.$$

Therefore,

$$||(\hat{\eta}_q - \eta_{0,n,q})\partial_\delta \sum_{i=1}^{t} g_i(\beta_{0,n}, \bar{\delta})/n(\hat{\delta} - \delta_{0,n})||_\infty =$$

$$||\hat{\eta}_q \partial_\delta \sum_{i=1}^{t} g_i(\hat{\theta})(\hat{\delta} - \delta_{0,n})/n - \eta_{0,n,q}\partial_\delta \sum_{i=1}^{t} g_i(\theta_{0,n})(\hat{\delta} - \delta_{0,n})/n||_\infty \leq$$

$$(||\hat{\eta}_q \partial_\delta \sum_{i=1}^{t} g_i(\hat{\theta})/n||_\infty + ||\eta_{0,n,q} \partial_\delta \sum_{i=1}^{t} g_i(\theta_{0,n})/n||_\infty)||\hat{\delta} - \delta_{0,n}||_1$$

$$= (||\hat{\eta}_q \partial_\delta \sum_{i=1}^{t} g_i(\hat{\theta})/n||_\infty + ||\partial_\delta \sum_{i=1}^{t} M_{qi}(\theta_{0,n}, \eta_{0,n})/n||_\infty)||\hat{\delta} - \delta_{0,n}||_1$$

$$= (O_p(\log(J)\log(n)/\sqrt{n}) + O_p(\log(J)/\sqrt{n}))o_p(1/(\log(J)\log(n))) = o_p(1/\sqrt{n}).$$

Otherwise, for the non-linear case, there exists $\bar{\delta}_t^*$ for each t with $1 \leq t \leq n$, such that $||\bar{\delta}_t^* - \delta_{0,n}||_1 \leq ||\hat{\delta} - \delta_{0,n}||_1$ and

$$\partial_\delta \sum_{i=1}^{t} g_i(\beta_{0,n}, \bar{\delta})/n = \partial_\delta \sum_{i=1}^{t} g_i(\theta_{0,n})/n + (\bar{\delta} - \delta_{0,n})' \partial_\delta^2 \sum_{i=1}^{t} g_i(\beta_{0,n}, \bar{\delta}_t^*)/n.$$

Therefore, using Assumptions 2.1.2 and 2.1.4,

$$\sqrt{n}||(\hat{\eta}_q - \eta_{0,n,q})\partial_\delta \sum_{i=1}^{t} g_i(\beta_{0,n}, \bar{\delta})/n(\hat{\delta} - \delta_{0,n})||_\infty \leq$$

$$\sqrt{n}||\hat{\eta}_q - \eta_{0,n,q}||_1(||\partial_\delta \sum_{i=1}^{t} g_i(\theta_{0,n})/n||_\infty ||\hat{\delta} - \delta_{0,n}||_1$$

$$+ \max_{s \in \{1,\ldots,Q\}} |(\bar{\delta} - \delta_{0,n})' \partial_\delta^2 \sum_{i=1}^{t} g_{si}(\beta_{0,n}, \bar{\delta}_t^*)/n(\hat{\delta} - \delta_{0,n})|)$$

$$\leq \sqrt{n} o_p(n^{-1/4}/\sqrt{\log(J)}) O_p(\log(J)) o_p(n^{-1/4}/\sqrt{\log(J)})$$

$$+ \sqrt{n} o_p(n^{-1/4}/\sqrt{\log(J)})||\hat{\delta} - \delta_{0,n}||_2^2 \max_{s \in \{1,\ldots,Q\}} \max eig(\partial_\delta^2 \sum_{i=1}^{t} g_{si}(\beta_{0,n}, \bar{\delta}_t^*)/n).$$

Then, using $\epsilon > 0$ defined by Assumption 2.1.4, because $||\bar{\delta}_t^* - \delta_{0,n}||_1 \leq ||\hat{\delta} - \delta_{0,n}||_1 + ||\hat{\eta}_q - \eta_{0,n,q}||_1 < \epsilon$ wpa1, then wpa1

$$\max_{s \in \{1,\ldots,Q\}} \max eig(\partial_\delta^2 \sum_{i=1}^{t} g_{si}(\beta_{0,n}, \bar{\delta}_t^*)/n)$$

$$\leq \max_{s\in\{1,...,Q\}} \sup_{\delta:||\delta-\delta_{0,n}||_1<\epsilon} \max eig(\partial_\delta^2 \sum_{i=1}^t g_{si}(\beta_{0,n},\delta)/n) = O_p(\log(J)) \text{ and}$$

$$\max eig(\partial_\gamma^2 \sum_{i=1}^t M_{qi}(\beta_{0,n},\bar{\gamma}_t)/n) \leq \sup_{\gamma:||\gamma-\gamma_0||_1<\epsilon} \max eig(\partial_\gamma^2 \sum_{i=1}^t M_{qi}(\beta_{0,n},\gamma)/n) = O_p(\log(J)).$$

Hence,

$$\sqrt{n}||(\hat{\eta}_q - \eta_{0,n,q})\partial_\delta \sum_{i=1}^t g_i(\beta_{0,n},\bar{\delta})/n(\hat{\delta}-\delta_{0,n})||_\infty \leq$$

$$\leq \sqrt{n}o_p(n^{-1/4}/\sqrt{\log(J)})O_p(\log(J))o_p(n^{-1/4}/\sqrt{\log(J)})$$

$$+\sqrt{n}o_p(n^{-1/4}/\log(J))o_p(n^{-1/2}/\log(J))O_p(\log(J)) = o_p(1).$$

Also,

$$|\sqrt{n}(\hat{\delta}-\delta_{0,n})'\partial_\delta^2 \sum_{i=1}^t M_{qi}(\beta_{0,n},\bar{\gamma}_t)/n(\hat{\delta}-\delta_{0,n})/n|$$

$$\leq \sqrt{n}||\hat{\delta}-\delta_{0,n}||_2^2 \max eig(\partial_\delta^2 \sum_{i=1}^t M_{qi}(\beta_{0,n},\bar{\gamma}_t)/n)$$

$$\leq \sqrt{n}O_p(\log(J))o_p(n^{-1/2}/\log(J)) = o_p(1).$$

Therefore, uniformly over t with $1 \leq t \leq n$,

$$(\sum_{i=1}^t M_{qi}(\beta_{0,n},\hat{\gamma}) - \hat{M}_{qi}(\beta_{0,n},\gamma_0))/\sqrt{n} = o_p(1),$$

for each $q \in \{1,...,m\}$.

Sufficient Condition for Assumption 2.2.2: Suppose that Θ_n is compact, $\hat{g}(\theta) - g(\theta) \xrightarrow{p} 0$ point-wise in θ, and there exists $\alpha > 0$ and $B_n = O_p(1)$ such that for all $\theta_1, \theta_2 \in B$, $\sup_{\theta\in\Theta_n} ||\hat{g}(\theta_1) - \hat{g}(\theta_2)||_2 \leq B_n||\theta_1-\theta_2||_2^\alpha$. Then since g is continuous, the conditions of Lemma 2.9 of Newey and McFadden (1994) are satisfied when J is fixed, so $\sup_{\theta\in\Theta_n} |\hat{g}(\theta) - g(\theta)| \xrightarrow{p} 0$ as $n \to \infty$ with J fixed.

Appendix C

I estimate the power with the true value of β_t being -.5 in each post-treatment time period. Table 4 contains the size-adjusted power results.

Table 4: Size-Adjusted Power Results

	Orthogonalized	End-of-Sample	Conformal	Cross-Fitting	Subsampling	Placebo
	SCE t-test	Instabilty test	Inference	t-test	Method	Method
Rejection Rates with $\alpha = .1$						
$T_0 = 30, T_1 = 4$	0.826	0.705	0.696	0.780	0.120	0.887
$T_0 = 30, T_1 = 16$	0.867	0.779	0.364	0.547	0.994	0.985
$T_0 = 60, T_1 = 16$	0.968	0.799	0.779	0.886	0.991	0.987
$T_0 = 30, T_1 = 32$	0.734	0.753	0.009	0.444	0.938	0.936
$T_0 = 60, T_1 = 32$	0.952	0.786	0.058	0.610	0.978	0.976
Rejection Rates with $\alpha = .05$						
$T_0 = 30, T_1 = 4$	0.660	0.527	0.574	0.552	0.120	0.803
$T_0 = 30, T_1 = 16$	0.765	0.000	0.235	0.399	0.989	0.975
$T_0 = 60, T_1 = 16$	0.902	0.668	0.366	0.710	0.977	0.972
$T_0 = 30, T_1 = 32$	0.556	0.640	0.005	0.278	0.938	0.820
$T_0 = 60, T_1 = 32$	0.843	0.679	0.015	0.380	0.955	0.939
Rejection Rates with $\alpha = .01$						
$T_0 = 30, T_1 = 4$	0.226	0.348	0.367	0.213	0.120	0.621
$T_0 = 30, T_1 = 16$	0.324	0.000	0.084	0.110	0.974	0.846
$T_0 = 60, T_1 = 16$	0.607	0.458	0.142	0.191	0.977	0.938
$T_0 = 30, T_1 = 32$	0.202	0.379	0.000	0.092	0.938	0.530
$T_0 = 60, T_1 = 32$	0.469	0.421	0.001	0.137	0.955	0.685

Notes: All simulations are conducted with a thousand replications.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science,* 59 (2): 495–510. https://doi.org/10.1111/ajps.12116.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *The American Economic Review* 93 (1): 113–132. 10.1257/000282803321455188.

Andersson, Julius J. 2019. "Carbon Taxes and CO2 Emissions: Sweden as a Case Study." *American Economic Journal: Economic Policy* 11 (4): 1–30. 10.1257/pol.20170144.

Andrews, D. W. K. 2003. "End-of-Sample Instability Tests." *Econometrica* 71 (6): 1661–1694. https://doi.org/10.1111/1468-0262.00466.

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59 (3): 817–858. https://doi.org/10.2307/2938229.

Andrews, Donald W. K. 1999. "Estimation When a Parameter is on a Boundary." *Econometrica* 67 (6): 1341–1383. https://doi.org/10.1111/1468-0262.00082.

Andrews, Donald W. K., and Xu Cheng. 2012. "Estimation and Inference With Weak, Semi-Strong, and Strong Identification." *Econometrica* 80 (5): 2153–2211. https://doi.org/10.3982/ECTA9456.

Andrews, Donald W. K., and Patrik Guggenberger. 2010. "Asymptotic Size and a Problem with Subsampling and with the m out of n Bootstrap." *Econometric Theory* 26 (2): 426–468. https://doi.org/10.1017/S0266466609100051.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111 (12): 4088–4118. 10.1257/aer.20190159.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica* 80 (6): 2369–2429. https://doi.org/10.3982/ECTA9626.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. 2018. "High-Dimensional Econometrics and Regularized GMM."

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50. 10.1257/jep.28.2.29.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–890. https://doi.org/10.2307/2171802.

Cao, Jianfei, and Connor Dowd. 2019. "Estimation and Inference for Synthetic Control Methods with Spillover Effects." Papers 1902.07343, arXiv.org, `https://ideas.repec.org/p/arx/papers/1902.07343.html`.

Carvalho, Carlos, Ricardo Masini, and Marcelo C. Medeiros. 2018. "ArCo: An artificial counterfactual approach for high-dimensional panel time-series data." *Journal of Econometrics* 207 (2): 352–380. https://doi.org/10.1016/j.jeconom.2018.07.005.

Chalak, Karim, and Daniel Kim. 2024. "Higher Order Moments for Differential Measurement Error, with Application to Tobin's q and Corporate Investment." *SSRN Electronic Journal*, `https://api.semanticscholar.org/CorpusID:269563508`.

Chaudhuri, Saraswata, and Eric Zivot. 2011. "A new method of projection-based inference in GMM with weakly identified nuisance parameters." *Journal of Econometrics* 164 (2): 239–251. https://doi.org/10.1016/j.jeconom.2011.05.012.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21 (1): C1–C68. 10.1111/ectj.12097.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments." *American Economic Review* 105 (5): 486–90. 10.1257/aer.p20151022.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models1." *Econometrica* 75 (5): 1243–1284. https://doi.org/10.1111/j.1468-0262.2007.00794.x.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu. 2021. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." *Journal of the American Statistical Association* 0 (0): 1–16. 10.1080/01621459.2021.1920957.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu. 2024. "A t-test for synthetic controls." *Working Paper*.

Cox, Gregory. 2022. "Weak Identification with Bounds in a Class of Minimum Distance Models."

Dendramis, Yiannis, Liudas Giraitis, and George Kapetanios. 2021. "Estimation of Time-Varying Covariance Matrices For Large Datasets." *Econometric Theory* 37 (6): 1100–1134. 10.1017/S0266466620000535.

Ferman, Bruno. 2021. "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls." *Journal of the American Statistical Association* 116 (536): 1764–1772. https://doi.org/10.1080/01621459.2021.1965613.

Ferman, Bruno, and Cristine Pinto. 2021. "Synthetic controls with imperfect pretreatment fit." *Quantitative Economics* 12 (4): 1197–1221. https://doi.org/10.3982/QE1596.

Fry, Joseph. 2024. "A method of moments approach to asymptotically unbiased Synthetic Controls." *Journal of Econometrics* 244 (1): 105846. https://doi.org/10.1016/j.jeconom.2024.105846.

Gavish, Matan, and David L. Donoho. 2014. "The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$." *IEEE Transactions on Information Theory* 60 (8): 5040–5053. 10.1109/TIT.2014.2323359.

Geyer, Charles J. 1994. "On the Asymptotics of Constrained M-Estimation." *The Annals of Statistics* 22 (4): 1993–2010. 10.1214/aos/1176325768.

Hahn, Jinyong, and Ruoyao Shi. 2017. "Synthetic Control and Inference." *Econometrics* 5 (4): 52. https://doi.org/10.3390/econometrics5040052.

Han, Sukjin, and Adam McCloskey. 2019. "Estimation and inference with a (nearly) singular Jacobian." *Quantitative Economics* 10 (3): 1019–1068. https://doi.org/10.3982/QE989.

Hansen, Bruce E. 1996. "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis." *Econometrica* 64 (2): 413–430. https://doi.org/10.2307/2171789.

Herrndorf, Norbert. 1985. "A functional central limit theorem for strongly mixing sequences of random variables." *Probability Theory and Related Fields* 69 541–550. https://doi.org/10.1007/BF00532665.

Hwang, Jungbin, and Yixiao Sun. 2018. "Should we go one step further? An accurate comparison of one-step and two-step procedures in a generalized method of moments framework." *Journal of Econometrics* 207 (2): 381–405. https://doi.org/10.1016/j.jeconom.2018.07.006.

Kaul, Ashok, Stefan Klöbner, Gregor Pfeifer, and Manuel Schieler. 2021. "Standard Synthetic Control Methods: The Case Of Using All Pre-Intervention Outcomes Together With Covariates." *Journal of Economic Literature* 59 (2): 391–425. https://doi.org/10.1080/07350015.2021.1930012.

Ketz, Philipp. 2018. "Subvector inference when the true parameter vector may be near or at the boundary." *Journal of Econometrics* 207 (2): 285–306. https://doi.org/10.1016/j.jeconom.2018.08.003.

Kiefer, Nicholas M., and Timothy J. Vogelsang. 2002. "Heteroskedasticity–Autocorrelation Robust Standard Errors Using The Bartlett Kernel Without Truncation." *Econometrica* 70 (5): 2093–2095. https://doi.org/10.1111/1468-0262.00366.

Lazarus, Eben, Daniel J. Lewis, and James H. Stock. 2021. "The Size-Power Tradeoff in HAR Inference." *Econometrica* 89 (5): 2497–2516. https://doi.org/10.3982/ECTA15404.

Li, Kathleen T. 2020. "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods." *Journal of the American Statistical Association* 115 (532): 2068–2083. https://doi.org/10.1080/01621459.2019.1686986.

Mackey, Lester, Vasilis Syrgkanis, and Ilias Zadik. 2018. "Orthogonal Machine Learning: Power and Limitations." In *Proceedings of the 35th International Conference on Machine Learning*, edited by Dy, Jennifer, and Andreas Krause Volume 80. of Proceedings of Machine Learning Research 3375–3383, PMLR, , 10–15 Jul, `https://proceedings.mlr.press/v80/mackey18a.html`.

Newey, Whitney K., and Daniel McFadden. 1994. *Chapter 36 Large sample estimation and hypothesis testing.* Volume 4. of Handbook of Econometrics, Elsevier, 2111-2245. https://doi.org/10.1016/S1573-4412(05)80005-4.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3): 703–708. https://doi.org/10.2307/1913610.

Neyman, Jerzy. 1959. "Optimal asymptotic tests of composite statistical hypotheses." *U. Grenander., ed., 'Probability and Statistics, the Harald Cramer Volume'* 2068–2083.

Ning, Yang, and Han Liu. 2017. "A GENERAL THEORY OF HYPOTHESIS TESTS AND CONFIDENCE REGIONS FOR SPARSE HIGH DIMENSIONAL MODELS." *The Annals of Statistics* 45 (1): 158–195, http://www.jstor.org/stable/44245775.

Phillips, Peter C.B. 2005. "HAC ESTIMATION BY AUTOMATED REGRESSION." *Econometric Theory* 21 (1): 116–142. 10.1017/S0266466605050085.

Powell, David. 2021. *Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?*. Santa Monica, CA: RAND Corporation, . 10.7249/WR1246.

Romano, Joseph P., and Azeem M. Shaikh. 2010. "Inference for the Identified Set in Partially Identified Econometric Models." *Econometrica* 78 (1): 169–211. https://doi.org/10.3982/ECTA6706.

Shi, Xu, Kendrick Li, Wang Miao, Mengtong Hu, and Eric Tchetgen Tchetgen. 2023. "Theory for identification and Inference with Synthetic Controls: A Proximal Causal Inference Framework."

Singh, Amandeep, Kartik Hosanagar, and Amit Gandhi. 2020. "Machine Learning Instrument Variables for Causal Inference." In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20 835–836, New York, NY, USA: Association for Computing Machinery, . 10.1145/3391403.3399466.

Sun, Yixiao. 2013. "A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator." *The Econometrics Journal* 16 (1): 1–26. https://doi.org/10.1111/j.1368-423X.2012.00390.x.

Sun, Yixiao. 2014a. "Fixed-Smoothing Asymptotics in a Two-Step Generalized Method of Moments Framework." *Econometrica* 82 (6): 2327–2370. https://doi.org/10.3982/ECTA11684.

Sun, Yixiao. 2014b. "Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference." *Journal of Econometrics* 178 (P3): 659–677. 10.1016/j.jeconom.2013.10.

Vogelsang, Timothy, and Nicholas Kiefer. 2002. "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size." *Econometric Theory* 18 1350–1366. 10.1017/S026646660218604X.

Vogelsang, Timothy, and Nicholas Kiefer. 2005. “A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests.” *Econometric Theory* 21 1130–1164. 10.1017/S0266466605050565.

Zhang, Jeffrey, Wei Li, Wang Miao, and Eric Tchetgen Tchetgen. 2023. “Proximal causal inference without uniqueness assumptions.” *Statistics and Probability Letters* 198 109836. 10.1016/j.spl.2023.109836.