Tuning Parameter Selection in the Synthetic Control Method

Ryan Lee*

July 30, 2017

Abstract

In this paper I show an asymptotically optimal choice of a weighting matrix used in the synthetic control method. The synthetic control method takes a weighted average of outcomes for untreated units to estimate the outcome under no treatment for a treated unit. This can then be used to estimate a treatment effect for the treated unit. The weights are chosen such that the weighted average of the outcomes in the pretreatment time periods and of covariates approximates that of the treated unit. In practice, these weights are chosen to minimize a distance which depends on a weighting matrix. I show asymptotic optimality of a leave-one-out cross-validation procedure to choose this weighting matrix. This amounts to performing the synthetic control method, in turn, as if each of the untreated units were instead treated and assessing the prediction on the untreated units for a given weighting matrix. This is not straightforward because there is dependence across these synthetic control estimates.

*Northwestern University, Department of Economics

1 Introduction

The synthetic control method (SCM) introduced in Abadie and Gardeazabal (2003) has rapidly gained traction in applied work. As often happens with new methods, there are issues that need to be addressed. In the SCM a tuning parameter choice needs attention for two reasons. First, the standard method for choosing this tuning parameter has not been shown to be optimal in any sense. Second, and perhaps more importantly, the current method restricts the set of specifications available to researchers. This restriction eliminates specifications that previous research has motivated. These two points are expanded upon as the focus of this paper.

In the first use of the synthetic control method, Abadie and Gardeazabal (2003) examine the effects of terrorist conflict in the Basque Country, using per capita GDP as the outcome variable. Panel data on real per capita GDP are available before and after times of peak ETA terrorist activity, as well as other characteristics that determine growth for 17 regions of Spain. One might expect the researchers to use the difference in differences (DiD) approach here. This would involve choosing one of the 16 non-Basque regions of Spain and taking the change in per capita GDP before and after the terrorist activity and subtracting that quantity from the change in per capita GDP for the Basque Country. To see why they may not want to do this, let us compare this setting to that of a well-known DiD application.

Card and Krueger (1994) examine a 1992 minimum wage increase in New Jersey, surveying employment at individual fast food restaurants in New Jersey and Pennsylvania before and after the minimum wage increase. The first distinction here is that Card and Krueger (1994) have data at the individual restaurant level, whereas Abadie and Gardeazabal (2003) have aggregate per capita GDP for entire regions of Spain. The assumptions for identification in DiD, as in Angrist and Pischke (2009), is that

$$\mathbb{E}\left[Y_{ist}^N | s, t\right] = \gamma_s + \phi_t \tag{1}$$

where Y_{ist}^N is the potential outcome under no treatment. In Card and Krueger (1994), the outcome is employment for restaurant i, in state s (New Jersey or Pennsylvania), at time t (pre- or post- minimum wage increase). This potential outcome is observed in the pre- and post- time periods for Pennsylvania, and in the pre- time period for New Jersey. Therefore,

$$\mathbb{E}\left[Y_{ist}^N | s = \text{Penn.}, t = \text{Post}\right] - \mathbb{E}\left[Y_{ist}^N | s = \text{Penn.}, t = \text{Pre}\right] = \phi_{\text{Post}} - \phi_{\text{Pre}}$$

is identified. Also,

$$\mathbb{E}\left[Y_{ist}^N | s = \text{NJ}, t = \text{Post}\right] - \mathbb{E}\left[Y_{ist}^N | s = \text{NJ}, t = \text{Pre}\right] = \phi_{\text{Post}} - \phi_{\text{Pre}}$$

holds. The right-hand-side of the previous two equations are the same, giving this assumption the name "common trends." Therefore, $\mathbb{E}\left[Y_{ist}^T - Y_{ist}^N | s = \text{NJ}, t = \text{Post}\right]$ can be identified from (1) in this setting, where Y_{ist}^T is the potential outcome under treatment.

Returning to the topic of Basque Country terrorism, the researchers make the case that the determinants of GDP growth in the Basque Country were different from those of other regions of Spain before the terrorist activity. This makes the common trends assumption difficult to believe because, in the absence of the terrorist activity, the per capita GDPs likely would have grown at different rates anyway. Additionally, the researchers have data on all 17 Spanish regions for many years before peak ETA terrorist activity—in comparison to the two states and two time periods in Card and Krueger (1994). The SCM allows them to leverage the length and width of this panel to replace the common trends assumption.

The model that motivates the SCM is as follows:

$$\begin{aligned} Y_{it}^N &= \theta_t' Z_i + \lambda_t' \mu_i + \epsilon_{it} \\ Y_{it}^T &= \alpha_{it} + Y_{it}^N. \end{aligned} \tag{2}$$

where $\mathbb{E}\left[\epsilon_{it} | \mu_1, Z_1, ..., \mu_{J+1}, Z_{J+1}\right] = 0$, Y_{it}^N and Y_{it}^T represent the potential outcomes for unit i at time t under no treatment and under treatment, respectively, and α_{it} is the treatment effect for unit i in time t. Z_i and μ_i represent an $r \times 1$ vector of unit-specific covariates and an $F \times 1$ vector of (unobserved) unit-specific factor loadings, and θ_t and λ_t are time-specific vectors of coefficients and factors. Our panel extends for T periods where T_0 is the last pretreatment period. $T = T_0 + 1$ is assumed in this paper for simplicity of the discussion. Unit 1 is the only treated unit and units $2, ..., J + 1$ are the untreated units, sometimes referred to as the "pool." The observed outcome is

$$Y_{it} = D_{it} Y_{it}^T + (1 - D_{it}) Y_{it}^N \tag{3}$$

where D_{it} is an indicator for receiving treatment. $D_{it} = 1$ only for $i = 1$ and $t = T$ and is 0 otherwise. α_{1T} is the parameter we are interested in estimating.

Part of the appeal of this model is the general factor structure, which can be shown to be more general than in DiD. If we consider (2) without covariates, then

$$\mathbb{E}\left[Y_{it}^N\right] = \mathbb{E}\left[\lambda_t' \mu_i\right].$$

If

$$\lambda_t = \begin{pmatrix} 1 \\ \phi_t \end{pmatrix}$$

then $\mathbb{E}\left[\lambda_t' \mu_i\right]$ could be decomposed into the structure used in DiD. However, λ_t is not limited to this particular vector, meaning that the common trends assumption in DiD does not need to hold.

The following thought experiments, first expressed in Abadie et al. (2010), give idealized conditions where an unbiased or asymptotically unbiased estimate of α_{1T} could be constructed. First, suppose that there exist weights $\{w_j^*\}_{j=2}^{J+1}$ such that

$$\sum_{j=2}^{J+1} w_j^* Z_j = Z_1 \quad \& \quad \sum_{j=2}^{J+1} w_j^* \mu_j = \mu_1. \tag{4}$$

This condition, in addition to (2), implies unbiasedness of $Y_{1T} - \sum_{j=2}^{J+1} w_j^* Y_{jT}$ as an estimate for α_{1T}. However, verifying (4) is not possible as the μ_j's are unobserved. So, the weights in (4) may exist, but we cannot find them. Therefore, suppose that we can find weights that satisfy the following instead of (4):

$$\sum_{j=2}^{J+1} w_j^* Z_j = Z_1 \quad \& \quad \sum_{j=2}^{J+1} w_j^* Y_{jt} = Y_{1t} \quad \text{for} \quad t = 1, ..., T_0. \tag{5}$$

Assuming (5), as well as the ϵ_{it}'s being independent across i and t with an even moment existing, $\mathbb{E}\left[\epsilon_{it} | Z_1, \mu_1, ..., Z_{J+1}, \mu_{J+1}\right] = 0$, and conditions regulating how the λ_t's change over pretreatment periods, then $Y_{1T} - \sum_{j=2}^{J+1} w_j^* Y_{jT}$ is an asymptotically unbiased estimate for α_{1T} as T_0, the number of pretreatment periods, goes to ∞.

The weights from (5) in the previous thought experiment are not guaranteed to exist, and often do not exist in applications. Instead, weights such that (5) hold *approximately* are found by minimizing a pseudometric between X_1, a vector of pretreatment variables, and the weighted average of this vector for the untreated units. Given the argument for an asymptotically unbiased estimate relied on (5), a natural choice for X_1 would be to include all covariates in Z and all pretreatment outcomes Y_{1t}, i.e.

$$X_1 = \begin{pmatrix} Z_1 \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1T_0} \end{pmatrix} \tag{6}$$

The pseudometric from Abadie et al. (2010) is

$$\sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)'} \tag{7}$$

where X_0 is a J column matrix whose columns contain the same variables as X_1 except for units $j = 2, ..., J+1$, i.e.

$$X_0 = \begin{pmatrix} Z_2 & Z_3 & \dots & Z_{J+1} \\ Y_{2,1} & Y_{3,1} & \dots & Y_{J+1,1} \\ Y_{2,2} & Y_{3,2} & \dots & Y_{J+1,2} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{2,T_0} & Y_{3,T_0} & \dots & Y_{J+1,T_0} \end{pmatrix}. \tag{8}$$

W is a length J vector of positive weights summing to one, and V is a diagonal positive semidefinite matrix. V is the focus of this paper. It would be possible to fix some V and (7) would still be a pseudometric. Informally, the idea why tuning V could improve the estimate of α_{1T} is that some dimensions of X_1 may be more important than others in determining Y_{1T}^N. In Section 2 it is shown that the standard way of choosing V does not allow for X_1 as in (6); the standard method assigns zeros to the diagonals in V that correspond to the covariates Z. This results in discretionary choices for X_1, or worse, relevant data being dismissed by the synthetic control method. Moreover, the standard choice of V has not been shown to be optimal in any sense. The choice of V that is discussed in this paper involves a leave-one-out cross validation (CV) across units j, does not suffer from the issues involving X_1 just discussed, and is optimal in terms of mean squared error under some additional conditions for large J.

The rest of this paper is organized as follows. Section 2 provides the details of the SCM, including the current standard choice of V, demonstrates how this choice of V does not allow specification in (6), and explains the cross validation choice of V. Section 3 gives the asymptotic argument for the optimality of cross validation. Section 4 supports Section 3 with simulation evidence. Section 5 concludes.

2 Synthetic Control Method and Cross Validation

The argument for the synthetic control method, as laid out in Abadie et al. (2010) and in Section 1, is the asymptotic unbiasedness of the estimate

$$\hat{\alpha}_{1T} = Y_{1T} - \sum_{j=2}^{J+1} w_j^* Y_{jT} \tag{9}$$

for α_{1T} when $\{w_j^*\}_{j=2}^{J+1}$ satisfy Equation (5). Again, weights that satisfy Equation (5) are not guaranteed to exist, so the following minimization is an attempt to make Equation (5) hold approximately.

$$W(V) = \arg\min_{W \in \mathcal{W}} (X_1 - X_0 W)' V (X_1 - X_0 W) \tag{10}$$

where $\mathcal{W} = \{W \in \mathbb{R}^J_{\geq 0} : \|W\|_1 = 1\}$, V is a positive semidefinite $J \times J$ diagonal matrix, and W is written explicitly a function of V. Previous research has made informal arguments for these restrictions on $\mathcal{W}$ instead of allowing for some other subset of $\mathbb{R}^J$. Comparing results under alternative choices of $\mathcal{W}$ is not the focus of this paper, and it may be difficult to develop theory to compare potential choices of $\mathcal{W}$. If V is positive semidefinite, the objective in (10) is a pseudometric between X_1 and $X_0 W$ when the square root is taken, which has no effect on the arg min. Additionally, it may be difficult to develop theory comparing this loss to other possible loss functions, so only losses of the form in (10) are considered. Only allowing V to be diagonal, reducing the dimension of the parameter, is more a matter of practicality; a brief note is made in Section 3 of how this relates to our cross validation optimality result.

The standard method for choosing V is to minimize the mean squared prediction error in the pre treatment periods for the treated unit. Defining

$$Y_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,T_0} \end{pmatrix} \quad \& \quad Y_0 = \begin{pmatrix} Y_{2,1} & Y_{3,1} & \dots & Y_{J+1,1} \\ Y_{2,2} & Y_{3,2} & \dots & Y_{J+1,2} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{2,T_0} & Y_{3,T_0} & \dots & Y_{J+1,T_0} \end{pmatrix}$$

then this choice of V, V^{MSPE}, is

$$V^{MSPE} = \arg\min_{V \in \mathcal{V}} (Y_1 - Y_0 W(V))' (Y_1 - Y_0 W(V)) \tag{11}$$

where $\mathcal{V}$ is the set of positive semidefinite diagonal $J \times J$ matrices. Abadie et al. (2010) only provide informal arguments to support this specification. Furthermore, (11) rules out

the use of X_1 in Equation (6). To see why this is the case, suppose the specification in (6) is used while (11) is used to choose V. This results in choice of V as follows:

$$\tilde{V} = \begin{pmatrix} 0_{r\times r} & 0_{r\times T_0} \\ 0_{T_0\times r} & I_{T_0\times T_0} \end{pmatrix}.$$

This is important because $\tilde{V}$ has zeros along the diagonal that correspond to the covariates. Therefore, the covariates do not play a role in calculating the synthetic control weights. If a researcher has covariates that are a strong predictors of Y_{jT}^N, the potential outcome under no treatment in the treatment period, then this could be a hindrance to finding good weights. To understand why $\tilde{V}$ is the result of using (6) in combination with (11), consider the analogy of running OLS of a dependent variable Y on Y and X. Clearly, you would get a coefficient of 1 on Y and a coefficient of 0 on X. Mechanically, this is similar to getting $\tilde{V}$ as above. For a more formal explanation, notice that

$$\begin{aligned}
&(X_1 - X_0W)' \begin{pmatrix} 0_{r\times r} & 0_{r\times T_0} \\ 0_{T_0\times r} & I_{T_0\times T_0} \end{pmatrix} (X_1 - X_0W) \\
&\quad = \left(\begin{pmatrix} Z_1 \\ Y_{1,1} \\ \vdots \\ Y_{1,T_0} \end{pmatrix} - \begin{pmatrix} Z_2 & \dots & Z_{J+1} \\ Y_{2,1} & \dots & Y_{J+1,1} \\ \vdots & \ddots & \vdots \\ Y_{2,T_0} & \dots & Y_{J+1,T_0} \end{pmatrix} W \right)' \\
&\qquad\qquad * \begin{pmatrix} 0_{r\times r} & 0_{r\times T_0} \\ 0_{T_0\times r} & I_{T_0\times T_0} \end{pmatrix} \left(\begin{pmatrix} Z_1 \\ Y_{1,1} \\ \vdots \\ Y_{1,T_0} \end{pmatrix} - \begin{pmatrix} Z_2 & \dots & Z_{J+1} \\ Y_{2,1} & \dots & Y_{J+1,1} \\ \vdots & \ddots & \vdots \\ Y_{2,T_0} & \dots & Y_{J+1,T_0} \end{pmatrix} W \right) \\
&\quad = (Y_1 - Y_0W)' (Y_1 - Y_0W)
\end{aligned}$$

The above equalities tell us that $W(\tilde{V}) = \arg\min_{W\in\mathcal{W}} (Y_1 - Y_0W)' (Y_1 - Y_0W)$. But, we have that

$$\begin{aligned}
\min_{W\in\mathcal{W}} (Y_1 - Y_0W)' (Y_1 - Y_0W) &\leq \min_{V\in\mathcal{V}} (Y_1 - Y_0W(V))' (Y_1 - Y_0W(V)) \\
&\leq \left(Y_1 - Y_0W\left(\tilde{V}\right)\right)' \left(Y_1 - Y_0W\left(\tilde{V}\right)\right)
\end{aligned}$$

The first inequality is because $W(V) \in \mathcal{W}$ for all V by (10) and the second inequality is because $\tilde{V} \in \mathcal{V}$. However, the first and last expression are equal, so all of the above inequalities are equalities. This idea has previously appeared in Kaul et al. (2015)

As a result of this issue, many papers take one of two approaches. One is to include covariates in X_1 as well as linear combinations of the pretreatment outcomes, such as the average, not making use of all pretreatment observations. Such papers include Pinotti (2015), Ando (2015), and Kleven et al. (2013). The other approach avoids covariates altogether and includes only pretreatment outcomes in X_1. Papers taking this approach include Acemoglu et al. (2016) and Jardim et al. (2017). Although the argument previously laid out shows that (11) and (6) cannot be used in conjunction with each other, papers go through with this estimation nonetheless. Hinrichs (2012) examines the effect affirmative action bans had on the ethnic composition of college enrollments, using a ban in California. The predictors used—variables in X_1—include percent underrepresented minority (the outcome of interest) in the pretreatment years as well as the percent underrepresented minority in the state and per capita state income. As a result, the synthetic control method calculates weights as if the researcher did not provide the percent underrepresented minority in the state and per capita state income. Bohn et al. (2014) examine a 2007 Arizona law aimed at reducing unauthorized employment from undocumented immigrants and are interested in the effect on the composition of the state's population. In addition to approximating the pretreatment outcomes, they also use covariates including the state's workforce in various industrial categories, the proportion of the state's population in educational categories, and the state's unemployment rate. These covariates are given a zero diagonal element in V and do not have any effect on the synthetic control weights. Whenever (11) and (6) are used together, this produces an estimate as if the covariates were not included, despite the researchers' implicitly expressed desires to use the covariates. The cross validation replacement for (11), which is introduced next, does not suffer from this problem.

Cross Validation The cross validation estimate of V is defined here, which is shown to be optimal asymptotically in terms of mean squared error of the estimate of $\hat{\alpha}_{1T}$. But first, notice that

$$\hat{\alpha}_{1T} - \alpha_{1T} = Y_{1T} - \sum_{j=2}^{J+1} w_j Y_{jT} - \alpha_{1T} = Y_{1T}^T - \sum_{j=2}^{J+1} w_j Y_{jT}^N - \alpha_{1T} = Y_{1T}^N - \sum_{j=2}^{J+1} w_j Y_{jT}^N$$

Since unit 1 is treated in period T, we observe Y_{1T}^T, not Y_{1T}^N. However, we do observe Y_{jT}^N for $j = 2, ..., J+1$. This suggests that, with some assumptions on the joint distributions of the variables, performing the SCM for the untreated units could provide information about

the mean squared error of $\hat{\alpha}_{1T}$. To continue, some additional notation must be introduced. Let X_{1i} be the vector analogous to X_1, except for any unit $i = 1, ..., J+1$, and let X_{0i} be a matrix whose columns consist of vectors X_{1j} for $j \neq i$.[1] The cross validation choice of V is

$$\hat{V}_J = \arg\min_{V \in \mathcal{V}} \sum_{i=2}^{J+1} (Y_{iT} - \tilde{Y}_{iT}(V))^2 \tag{12}$$

where $\tilde{Y}_{iT}(V)$ is the synthetic control estimate of Y_{iT}^N where the calculations are done as if unit i is treated instead of unit 1. The weights for unit i are calculated analogously to (10) as follows

$$W^*(X_{1i}, X_{0i}, V) = \arg\min_{W \in \mathcal{W}} (X_{1i} - X_{0i}W)'V(X_{1i} - X_{0i}W) \tag{13}$$

and the weighted average using the vector of weights $W^*(X_{1i}, X_{0i}, V)$ to calculate $\tilde{Y}_{iT}(V)$ is taken using Y_{jT} over units $j \neq i$. In the data setting of the SCM, Doudchenko and Imbens (2016) also consider this type of cross validation except to choose a penalty parameter in a modified version of the SCM. Abadie et al. (2010) and Abadie et al. (2015) also suggest use of a validation period to choose V, though no result as in Section 3 has been shown.

3 Cross Validation Optimality

The goal of the following lemmas and theorems is to show optimality of our choice of V in terms of the mean squared error of our estimate $\hat{\alpha}_{1T}$.[2] $L_J(V)$ is defined as $\mathbb{E}\left[(\hat{\alpha}_{1T} - \alpha_{1T})^2\right]$. $\hat{V}_J$ is the cross-validation choice of V, and V_J^* is a minimizer of L_J in $\mathcal{V}$. The aim of this section is to show

$$|L_J(\hat{V}_J) - L_J(V_J^*)| \xrightarrow[J\to\infty]{p} 0 \tag{14}$$

where

$$\hat{V}_J = \arg\min_{V \in \mathcal{V}} \sum_{i=2}^{J+1} (Y_{iT} - \tilde{Y}_{iT}(V))^2 \tag{15}$$

[1] In practice this does matrix's columns are not *all* X_{1j} for $j \neq i$, i.e. may have fewer than J columns. In Section 3 it is shown that if there are too many columns in X_{0i}, then it is difficult to guarantee that the synthetic control weights are unique. This would mean that the synthetic control estimate would not necessarily be uniquely defined by (13). It is not new to the SCM to not use all observations in creating the synthetic control–see, for example, Abadie et al. (2015), which created a synthetic control for West Germany using 23 OECD countries in 1990, but Turkey, Luxembourg, and Iceland were prevented from having positive weights. This is addressed further in Section (3).

[2] If we have multiple post-treatment time periods, we could instead consider $\mathbb{E}\left[\sum_{t=T_0+1}^{T} (\hat{\alpha}_{1t} - \alpha_{1t})^2\right]$.

$\tilde{Y}_{iT}(V)$ is the synthetic control estimate of Y_{iT}^N for unit $i \neq 1$ using the other untreated units as control units written explicitly as a function of V. Equation (14) says that the mean squared error of the synthetic control estimate of α_{1T} using the cross validation choice of V, $\hat{V}_J$, converges in probability to the mean squared error of the synthetic control estimate of α_{1T} that uses the best possible V, in terms of mean squared error. Trivially, $L_J(V_J^*)$ would be smaller if $\mathcal{V}$ is allowed to include any positive semidefinite matrix, instead of just those that are diagonal. The convention of using diagonal V is retained for this paper, although this section could be extended to more general V.

In order to show optimality of our estimator, we first need to show that the synthetic control weights are almost surely uniquely defined for all positive definite diagonal V under some conditions. All proofs for this section can be found in Appendix A.

Lemma 1. *$W^*(X_{1i}, X_{0i}, V) = \arg\min_{W \in \mathcal{W}}(X_{1i} - X_{0i}W)'V(X_{1i} - X_{0i}W)$ is unique almost surely if*

1. *V is positive definite, and*
2. *letting Θ denote the space of vectors whose elements sum to zero and has dimension equal to the number of columns in X_{0i}, Null$(X_{0i}) \cap \Theta$ is trivial (i.e. $\{\vec{0}\}$) a.s.*

The latter condition could be reasonably satisfied if X_{1i} has more linearly independent continuously distributed variables than the number of columns of X_{0i} and the X_{1i} are i.i.d.

In light of this, it can be seen that X_{0i} cannot be allowed to have J columns, since as J grows it cannot be guaranteed that the arg min in Lemma 1 is unique. To address this problem, I still allow the number of untreated units, J, to grow, but I assume there is a J_0 such that, if we restrict the "pool" to a set of J_0 untreated units, then the weights are almost surely unique. In other words, there are J_0 units that are allowed to have positive weights in (13). These J_0 units are chosen, independent of everything else, as a random sample from $\{2, ..., i-1, i+1, ..., J+1\}$. To formalize this, there are J_0 columns of X_{0i}, which correspond to the X_{1j} for a simple random sample from $\{2, ..., i-1, i+1, ..., J+1\}$. J_0 is fixed and the subscript on L_J can be dropped. Let J_0^i denote the subset of $\{2, ..., i-1, i+1, ..., J+1\}$ sampled union with $\{i\}$.

There are other possibilities to modify (13) to guarantee the argmin is unique. This approach is similar to approaches already taken in that a subset of untreated units are allowed to have positive weights in the synthetic control estimate, e.g. Footnote 1. The reader may be concerned that the weights are still random due to the draws of J_0^i. Taking expectations over J_0^i removes this issue, which can be done computationally. A simple inequality shows that this further reduces the means squared error.

Lemma 2. *Let $\tilde{Y}_{iT}(V)$ denote the SC estimate for unit $i \neq 1$ using units $\{2, ..., i-1, i+1, ..., J+1\}$ as control units (written explicitly as a function of V), $W^*(X_{1i}, X_{0i}, V)$ as in Lemma 1. Under the assumptions of Lemma 1, $(Y_{iT} - \tilde{Y}_{iT}(V))^2$ is almost surely continuous in V for $V \in \mathcal{V}$ where $\mathcal{V}$ includes positive definite diagonal matrices.*

Notice that Lemmas 1 and 2 refers to positive definite V. In order to extend this to positive semidefinite V, define $W^*(X_{1i}, X_{0i}, V)$ for positive semidefinite as the limit of $W^*(X_{1i}, X_{0i}, V)$ for positive definite V, since Lemma 2 shows $\tilde{Y}_{it}(V)$ is continuous in V for positive definite V.[3]

Recapping Lemmas 1 and 2, $W^*(X_{1i}, X_{0i}, V)$ is a.s. unique and continuous in V. $W^*(X_{1i}, X_{0i}, V)$ can be defined for positive semidefinite V as the limit using positive definite V.

Lemma 3. *Define $\tilde{Y}_{iT}(V)$ to be the SC estimate of Y_{iT}^N with X_{0i} created as defined above (excluding unit 1). Under the assumptions of Lemma 1 and Lemma 2, and*

1. *i.i.d. (X_{1i}, Y_{iT}^N) across i, and*
2. *finite second moments of Y_{iT}^N*
3. *X_{0i} is as described previously*

then $L(V)$ is continuous in $\mathcal{V}$.

The proof of the following theorem is similar to that of Huber et al. (1967).

Theorem 1 (Main Result)**.** *The assumptions of Lemma 1-3 and finite fourth moment of Y_{iT}^N imply:*

1. $$\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V))^2 \xrightarrow{p} L(V)$$

   uniformly over $\mathcal{V}$, the space of positive semidefinite diagonal matrices whose diagonal elements sum to one, and
2. *Letting V^* denote an element of $\arg\min_{V\in\mathcal{V}} L(V)$ and $\hat{V}_J$ denoting the V found through cross validation $L(\hat{V}_J)$ converges in probability to $L(V^*)$.*

Part 2 of Theorem 1 has been the goal in this section and follows quickly from Part 1. This gives the sense in which cross validation is optimal. This says that the mean squared

[3] In practice, the diagonal elements of V can be bounded below by an arbitrarily small positive constant.

error of the SCM estimate of α_{1T} using cross validation converges in probability to the mean squared error of the SCM estimate of α_{1T} *as if* we plugged in the best V (in terms of MSE). This has shown the optimality of the cross validation choice of V, $\hat{V}_J$, in terms of the mean squared error of $\hat{\alpha}_1$.

4 Simulations

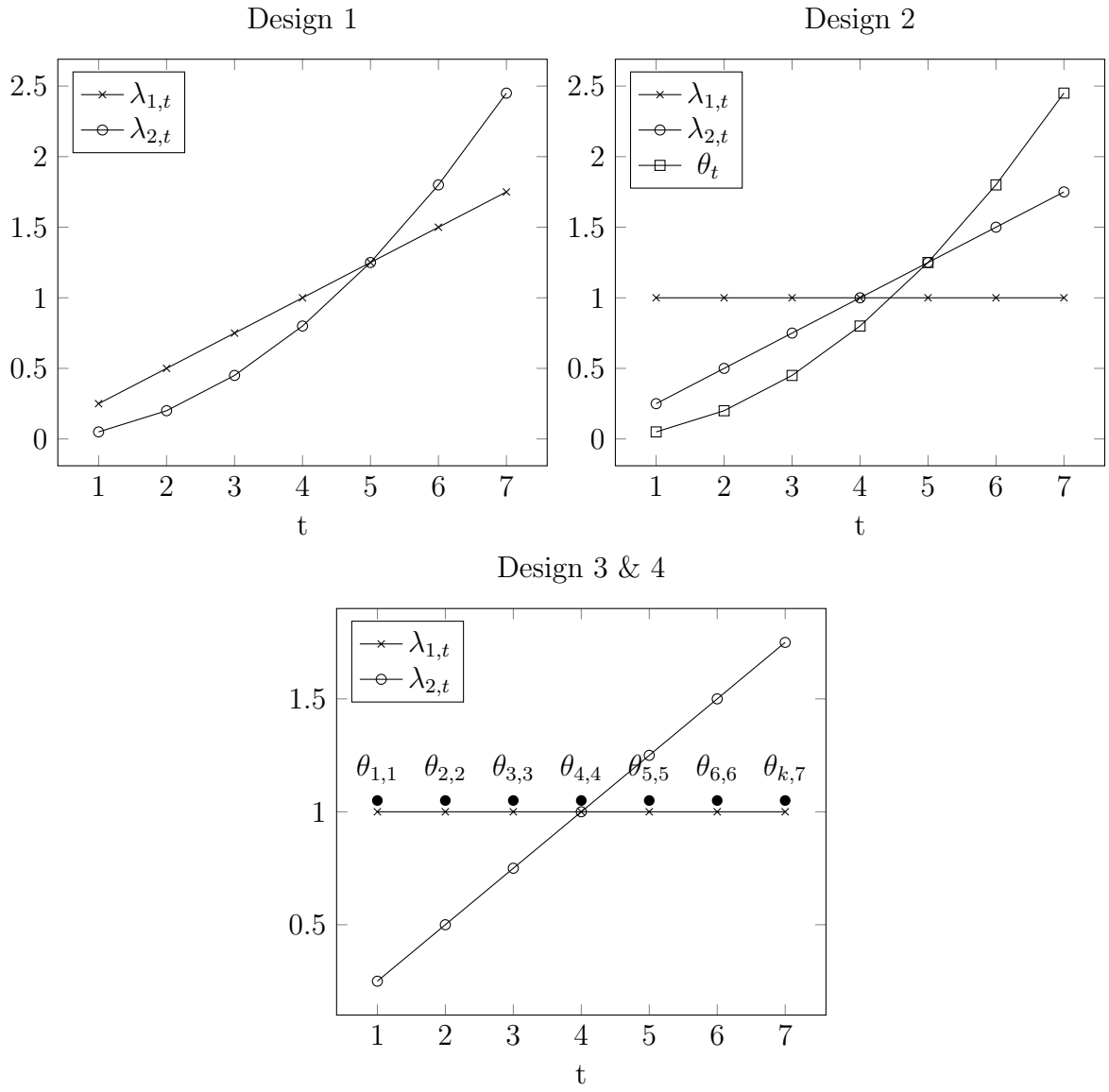
The simulations performed for this section support the asymptotic results demonstrated in Section 3 as the mean squared error of the estimate generally decreases for the simulations using the cross validated choice of V. Additionally, the argument specifying X_1 to include covariates and pretreatment outcomes is supported.

The simulations are designed as follows. The outcomes under no treatment Y_{iT}^N are generated according to Equation (2) where the covariates, factor loadings, and error terms are independent of each other and across i and, for all simulations, are drawn according to the distributions

$$Z_i \sim U\left([0,1]^r\right) \qquad \mu_i \sim U\left([0,1]^F\right) \qquad \epsilon_i \sim 0.2 * N\left(0, I_{T\times T}\right) \tag{16}$$

The dimensions r and F can be deduced from the design choices of θ_t and λ_t. All of the designs have 6 pretreatment periods, T_0, and 1 treatment period. Figure 1 graphs how the coefficients θ_t and factors λ_t evolve across time periods for the four different designs. Table 1 shows the specifications of X_1 across the different designs for both the MSPE and cross validation. In evaluating the cross validation choice of V, X_1 is always as in (6), whereas for the MSPE choice of V varies between two common choices of X_1. Remember that the mean squared error of $\hat{\alpha}_{1T}$ does not depend on α_{1T} because $\hat{\alpha}_{1T} - \alpha_{1T}$ is the same for all values of α_{1T}, so α_{1T} does not need to be stated in the design.

Figure 1 – Coefficients and Factors for Four Simulations Designs



These three graphs display the factors and coefficients over the different time periods. In Design 3 & 4 the covariates belong to $\mathbb{R}^6$. The coefficients are defined as follows: $\theta_{k,t} = 1$ if $k = t$ or $t = 7(= T)$ and $\theta_{k,t} = 0$ otherwise. The first subscript in $\lambda_{1,t}$ and $\lambda_{2,t}$ refer to the first and second dimension of λ_t.

Table 1 – X_1 Specification for Both MSPE and CV for Four Simulation Designs

	MSPE				Cross Validation
	[1]	[2]	[3]	[4]	[1]-[4]
$X_1 =$	$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1T_0} \end{pmatrix}$	$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1T_0} \end{pmatrix}$	$\begin{pmatrix} Z_1 \\ \bar{Y}_1 \end{pmatrix}$	$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1T_0} \end{pmatrix}$	$\begin{pmatrix} Z_1 \\ Y_{11} \\ \vdots \\ Y_{1T_0} \end{pmatrix}$

This table displays the various specifications for X_1 across the four different simulation designs for estimating the treatment effect α_{1T} using both MSPE and cross validation. $\bar{Y}_1 = \frac{1}{T_0}\sum_{t=1}^{T_0} Y_{1t}$

.

Table 2 presents the simulation results for the four different simulation designs outlined in Equation (16), Figure 1, and Table 1 for both the MSPE and cross validation choice of V. These simulation were performed for sample sizes of 10, 20, 50, 100, and 200. Some of the samples previously mentioned used U.S. states, regions in Spain, OECD countries, or banks on the NYSE, which are similar in size to the sample sizes simulated. There are only a few simulations where the estimate using MSPE outperforms cross validation. Given the asymptotic nature of the result in Section 3, it makes sense that the improvements from using cross validation are larger when the sample size is larger.

Design 1 is somewhat of a best case scenario for using the MSPE choice of V. The data are not generated with covariates, so the estimation using cross validation cannot make use of the additional information that covariates would provide. Even then, the differences in the mean squared error of the estimates are modest for all sample sizes. Design 2 shows that these differences are slightly more pronounced in favor of cross validation when there is one covariate available.

Simulation Designs 3 & 4 demonstrate the usefulness of specifying X_1 to include all covariates and pretreatment outcomes. Much of the variation in Y_{iT}^N is due to the covariates. In Design 3, where the estimate using MSPE specifies X_1 using the covariates and only the average of the outcome in pretreatment periods, the improvement from using cross validation is up to about 12 percent. In Design 4, where the estimate using MSPE specifies X_1 using only pretreatment outcomes, the improvement from cross validation is even greater at up to about 16 percent.

Table 2 – Mean Squared Error of $\hat{\alpha}_{1T}$

	MSPE				Cross Validation			
J+1	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
10	0.2982	0.3646	0.6132	0.7228	0.3182	0.3334	0.6194	0.6685
20	0.2811	0.3544	0.6309	0.7120	0.2900	0.3099	0.6122	0.6635
50	0.2946	0.3527	0.6227	0.7323	0.2948	0.2980	0.5785	0.6511
100	0.2890	0.3605	0.6171	0.7339	0.2852	0.2980	0.5466	0.6316
200	0.2976	0.3556	0.6386	0.7343	0.2907	0.2923	0.5602	0.6164

This table presents the mean squared error of the estimates using MSPE and cross validation to choose V, across the four different simulation designs, and with $J + 1$, the sample size, being 10, 20, 50, 100, or 200. Each simulation uses 10,000 repetitions. Simulation results for cross validation between 3 and 4 should not be the same as J_0 varied across the two simulations. $J_0 = 5$ for [1], [2], and [4], and $J_0 = 6$ for [3].

5 Conclusion

A cross validation method for choosing a tuning parameter in the synthetic control method has been shown to be useful. First, it allows researchers to use a broader range of specifications, which no longer precludes specifications that have been motivated in previous research. Cross validation was also shown to be asymptotically optimal under some conditions. There results were supported by simulations. Finally, it was shown that the procedure for inference is valid if the i.i.d. assumption in Section 3 replaces the assumption that treatment was randomized.

References

Abadie, A., A. Diamond, and J. 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, 493–505.

——— (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510.

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

Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton (2016): "The value of connections in turbulent times: Evidence from the United States," *Journal of Financial Economics*, 121, 368 – 391.

Ando, M. (2015): "Dreams of urbanization: Quantitative case studies on the local impacts of nuclear power facilities using the synthetic control method," *Journal of Urban Economics*, 85, 68 – 85.

Angrist, J. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 1 ed.

Bohn, S., M. Lofstrom, and S. Raphael (2014): "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?" *The Review of Economics and Statistics*, 96, 258–269.

Card, D. and A. B. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–793.

Doudchenko, N. and G. W. Imbens (2016): "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis," Tech. rep., National Bureau of Economic Research.

Hinrichs, P. (2012): "The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities," *The Review of Economics and Statistics*, 94, 712–722.

Huber, P. J. et al. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical*

Statistics and Probability, Volume 1: Statistics, The Regents of the University of California.

Jardim, E., M. C. Long, R. Plotnick, E. van Inwegen, J. Vigdor, and H. Wething (2017): "Minimum Wage Increases, Wages, and Low-Wage Employment: Evidence from Seattle," Working Paper 23532, National Bureau of Economic Research.

Kaul, A., S. Klössner, G. Pfeifer, and M. Schieler (2015): "Synthetic control methods: Never use all pre-intervention outcomes as economic predictors," *Unpublished. URl: http://www. oekonometrie. uni-saarland. de/papers/SCM_ Predictors. pdf.*

Kleven, H. J., C. Landais, and E. Saez (2013): "Taxation and International Migration of Superstars: Evidence from the European Football Market," *American Economic Review*, 103, 1892–1924.

Lehmann, E. L. and J. P. Romano (2006): *Testing statistical hypotheses*, Springer Science & Business Media.

Pinotti, P. (2015): "The Economic Costs of Organised Crime: Evidence from Southern Italy," *The Economic Journal*, 125, F203–F232.

A Proofs

In this appendix I provide proofs for the Lemmas and Theorem in Section 3.

Proof. **(Lemma 1)** $W^*(X_{1i}, X_{0i}, V)$ is unique if the objective for which it solves is strictly convex in $\mathcal{W}$. Using the definition of strict convexity, for every $W_1, W_2 \in \mathcal{W}$ s.t. $W_1 \neq W_2$ and $\alpha \in (0,1)$

$$((1-\alpha)W_1 + \alpha W_2)' X_{0i}' V X_{0i}((1-\alpha)W_1 + \alpha W_2) < (1-\alpha)W_1' X_{0i}' V X_{0i} W_1 + \alpha W_2' X_{0i}' V X_{0i} W_2$$

which is equivalent to

$$(W_1 - W_2)' X_{0i}' V X_{0i}(W_1 - W_2) > 0$$

Because V is positive definite, it is sufficient for $X_{0i}(W_1 - W_2) \neq \vec{0}$, which is almost surely true by the second condition. □

Proof. **(Lemma 2)** It suffices to show $W^*(X_{1i}, X_{0i}, V)$ is continuous a.s. for $\mathcal{V}$ positive definite diagonal matrices. Assume $\text{Null}(X_{0i}) \cap \Theta$ is trivial, which is a.s. true. Suppose for the sake of contradiction that $W^*(X_{1i}, X_{0i}, V)$ is not continuous at some $V_0 \in \mathcal{V}$
Therefore,

$$\exists\ \epsilon > 0 \ \text{ s.t. } \ \forall\ \delta > 0 \ \ \exists\ V \in N_\delta(V_0) \ \text{ s.t. } \ \|W^*(X_{1i}, X_{0i}, V_0) - W^*(X_{1i}, X_{0i}, V)\| \geq \epsilon$$

Let $\{\delta_n\}$ be a positive sequence such that $\delta_n \to 0$ and let V_n be the corresponding matrix such that the above statement holds. The norm considered for V is the Euclidean norm, since V can be though of as a vector of its diagonal elements. Define $W_n \equiv W^*(X_{1i}, X_{0i}, V_n)$ $(n \geq 1)$ and $W_0 \equiv W^*(X_{1i}, X_{0i}, V_0)$. Note that:

$$\begin{aligned}(X_{1i} - X_{0i}W_n)' V_n (X_{1i} - X_{0i}W_n) &\leq (X_{1i} - X_{0i}W_0)' V_n (X_{1i} - X_{0i}W_0) \\ &\underset{n\to\infty}{\longrightarrow} (X_{1i} - X_{0i}W_0)' V_0 (X_{1i} - X_{0i}W_0)\end{aligned} \tag{17}$$

where the inequality holds by optimality of the W_n and convergence is by continuity of the objective in V. By the Bolzano-Weierstrass theorem $\exists\{n_k\}$ such that $W_{n_k} \to \bar{W} \in \mathcal{W}$. Since the objective determining W^* is continuous in V and W

$$\begin{aligned}(X_{1i} - X_{0i}W_{n_k})' V_{n_k} (X_{1i} - X_{0i}W_{n_k}) &\underset{k\to\infty}{\longrightarrow} (X_{1i} - X_{0i}\bar{W})' V_0 (X_{1i} - X_{0i}\bar{W}) \\ &\leq (X_{1i} - X_{0i}W_0)' V_0 (X_{1i} - X_{0i}W_0)\end{aligned}$$

where the inequality follows from (17). The following equality must hold by the optimality of W_0. i.e.:

$$(X_{1i} - X_{0i}\bar{W})'V_0(X_{1i} - X_{0i}\bar{W}) = (X_{1i} - X_{0i}W_0)'V_0(X_{1i} - X_{0i}W_0)$$

However, the objective (13) is almost surely strictly convex and $\|\bar{W} - W_0\| \geq \epsilon$. This contradicts optimality of W_0. □

Proof. **(Lemma 3)** This follows from the Dominated Convergence Theorem.
Let $V \in \mathcal{V}$ be given and let V_n be a sequence in $\mathcal{V}$ such that $V_n \to V$.

$$|(Y_{iT}^N - \tilde{Y}_{iT})^2| \leq 4 \max_{l \in J_0^i} \{Y_{lT}^{N\,2}\}$$

The right of the above inequality is the dominating function. Because the second moment of Y_{jT}^N is assumed to exist, the expectation of the maximum of $J_0 + 1$ i.i.d. realizations of this variable must also exist. These observations are i.i.d. Therefore, the dominating function is integrable. $(Y_{iT}^N - \tilde{Y}_{iT}(V_n))^2 \underset{n\to\infty}{\longrightarrow} (Y_{iT}^N - \tilde{Y}_{iT}(V))^2$ a.e. pointwise; therefore, by the dominated convergence theorem

$$\lim_{n\to\infty} \mathbb{E}(Y_{iT}^N - \tilde{Y}_{iT}(V_n))^2 = \mathbb{E}(Y_{iT}^N - \tilde{Y}_{iT}(V))^2$$

This shows continuity of $\mathbb{E}(Y_{iT}^N - \tilde{Y}_{iT}(V))^2$ in V. □

Proof. **Part 1 of Theorem 1** Let $\epsilon > 0$ be given. Consider the term[4]

$$\sup_{V' \in N_\delta(V)} \left| (Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V))^2 \right| \tag{18}$$

Notice that by the continuity of $(Y_{iT}^N - \tilde{Y}_{iT}(V))^2$, (18) converges to zero as $\delta \to 0$ fixing V. Notice that (18) is bounded by $8 \max_{l \in J_0^i} \left\{ Y_{lT}^{N\,2} \right\}$, and the expectation of this exists. Therefore, by the dominated convergence theorem

$$\mathbb{E}\left[\sup_{V' \in N_\delta(V)} \left| (Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V))^2 \right| \right] \underset{\delta\to 0}{\longrightarrow} 0$$

[4] Assume it is measurable.

Therefore, $\forall V$, $\exists \delta(V)$ such that $\delta < \delta(V)$ implies

$$\mathbb{E}\left[\sup_{V' \in N_\delta(V)} \left|(Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V))^2\right|\right] < \epsilon$$

The $\{N_{\delta(V)}(V)\}_{V \in \mathcal{V}}$ cover $\mathcal{V}$, and $\mathcal{V}$ is compact. Therefore, $\exists K$ such that K neighborhoods centered at $\{V_k\}_{k=1}^K$ of radius $\delta_k = \delta(V_k)$ cover $\mathcal{V}$.
Let $V \in \mathcal{V}$ be given. Let k be such that $V \in N_{\delta_k}(V_k)$, which exists since the sets cover $\mathcal{V}$. From here, we bound by terms which only depend on k. Since there are only finitely many possible k, the convergence is uniform.

$$\begin{aligned}\left|\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V))^2 - L(V)\right| \leq &\left|\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2\right| \\ &+ \left|\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2 - L(V_k)\right| + |L(V_k) - L(V)|\end{aligned}$$

Let $\mu_k \equiv \mathbb{E}\left(\sup_{V' \in N_\delta(V_k)} \left|(Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2\right|\right) < \epsilon$ and notice $|L(V_k) - L(V)| < \epsilon$.

$$\begin{aligned}\left|\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V))^2 - L(V)\right| &\leq \left|\frac{1}{J}\sum_{i=2}^{J+1}\sup_{V' \in N_\delta(V_k)}\left|(Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2\right| - \mu_k\right| + |\mu_k| \\ &\quad + \left|\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2 - L(V_k)\right| + |L(V_k) - L(V)| \\ &\leq \underbrace{\left|\frac{1}{J}\sum_{i=2}^{J+1}\sup_{V' \in N_\delta(V_k)}\left|(Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2\right| - \mu_k\right|}_{(*)} \\ &\quad + \underbrace{\left|\frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2 - L(V_k)\right|}_{(**)} + 2\epsilon\end{aligned}$$

Notice that the last expression only depends on V through V_k. There are finitely many V_k. If this expression converges in probability to zero, then we have uniform convergence. It remains to show that $(*)$ and $(**)$ converge in probability to zero. First, consider $(*)$.

Define $\zeta_i = \sup_{V' \in N_\delta(V_k)} \left|(Y_{iT}^N - \tilde{Y}_{iT}(V'))^2 - (Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2\right| - \mu_k$.

$$\mathbb{P}\left((*) > \epsilon\right) = \mathbb{P}\left(\left|\sum_{i=2}^{J+1} \zeta_i\right| > J\epsilon\right) \leq \frac{\mathbb{E}\left(\sum_{i=2}^{J+1} \zeta_i\right)^2}{(J\epsilon)^2} = \frac{\mathbb{E}\left[(\zeta_i)^2\right]}{J\epsilon^2} + \frac{(J-1)\mathbb{E}_{i\neq j}\left[\zeta_i\zeta_j\right]}{J\epsilon^2}$$

where the inequality is Chebyshev's inequality. Consider the first term. $\mathbb{E}\left[(\zeta_i)^2\right]$ exists because $\zeta_i^2 \leq \max_{l \in J_0^i}(Y_{lT}^{N\,2} + \epsilon)^2$. The fourth moment of Y_{iT}^N exists, so the fourth moment of the max of a finite number, $|J_0^i|$, of i.i.d. realizations of Y_{iT}^N must exist. Therefore, the first term converges to zero. Now consider the second term.

$$\mathbb{E}_{i\neq j}\left[\zeta_i\zeta_j\right] = \mathbb{E}_{i\neq j}\left[\zeta_i\zeta_j | J_0^i \cap J_0^j = \emptyset\right]\mathbb{P}\left(J_0^i \cap J_0^j = \emptyset\right) + \mathbb{E}_{i\neq j}\left[\zeta_i\zeta_j | J_0^i \cap J_0^j \neq \emptyset\right]\mathbb{P}\left(J_0^i \cap J_0^j \neq \emptyset\right)$$

$\mathbb{E}_{i\neq j}\left[\zeta_i\zeta_j | J_0^i \cap J_0^j = \emptyset\right] = 0$ because J_0^i's where chosen independently of the data. ζ_i and ζ_j are functions of independent observations and are mean zero. $\zeta_i\zeta_j$ can be bounded in a similar way that ζ_i^2 was bounded. $\mathbb{P}\left(J_0^i \cap J_0^j \neq \emptyset\right) \longrightarrow 0$, so $(*) \xrightarrow{p} 0$. Now consider $(**)$. Again for notational simplicity, define $\xi_i = (Y_{iT}^N - \tilde{Y}_{iT}(V_k))^2 - L(V_k)$. Showing $(**) \xrightarrow{p} 0$ is similar to that of $(*)$.

$$\mathbb{P}\left((**) > \epsilon\right) = \mathbb{P}\left(\left|\sum_{i=1}^{J+1} \xi_i\right| > J\epsilon\right) \leq \frac{\mathbb{E}\left(\sum_{i=2}^{J+1} \xi_i\right)^2}{(J\epsilon)^2} = \frac{\mathbb{E}\left[(\xi_i)^2\right]}{J\epsilon^2} + \frac{(J-1)\mathbb{E}_{i\neq j}\left[\xi_i\xi_j\right]}{J\epsilon^2}$$

$\mathbb{E}\left[(\xi_i)^2\right]$ exists because $\mathbb{E}\max_{l \in J_0^i} Y_{lT}^{N\,4}$ exists. The argument for $\mathbb{E}_{i\neq j}\left[\xi_i\xi_j\right]$ is analogous to $\mathbb{E}_{i\neq j}\left[\zeta_i\zeta_j\right]$. Therefore, the desired result has been shown. □

Proof. **Part 2 of Theorem 1**

$$\left|L(\hat{V}_J) - L(V^*)\right|$$

$$
\begin{aligned}
&\leq \left| L(\hat{V}_J) - \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 \right| + \left| \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 - L(V^*) \right| \\
&\leq \left| L(\hat{V}_J) - \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 \right| \\
&\qquad + \max\left\{ \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 - L(V^*), L(V^*) - \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 \right\} \\
&\leq \left| L(\hat{V}_J) - \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 \right| \\
&\qquad + \max\left\{ \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V^*))^2 - L(V^*), L(\hat{V}_J) - \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 \right\} \\
&\leq 2\left| L(\hat{V}_J) - \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(\hat{V}_J))^2 \right| + \left| \frac{1}{J}\sum_{i=2}^{J+1}(Y_{iT}^N - \tilde{Y}_{iT}(V^*))^2 - L(V^*) \right|
\end{aligned}
$$

Both terms converge in probability to zero by the uniform convergence in probability presented in Part 1 of Theorem 1. □