Online Supplemental Appendix to "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls"

Victor Chernozhukov* Kaspar Wüthrich† Yinchu Zhu‡

April 5, 2021

A Extensions

A.1 Testing Hypotheses about Average Effects over Time

In addition to testing sharp null hypotheses, researchers are often also interested in testing hypotheses about average effects over time, $\bar{\theta} = T_*^{-1} \sum_{t=T_0+1}^{T} \theta_t$:

$$H_0 : \bar{\theta} = \bar{\theta}^0 \tag{A.1}$$

*Massachusetts Institute of Technology. Email: vchern@mit.edu

†University of California San Diego. Email: kwuthrich@ucsd.edu

‡Brandeis University. Email: yinchuzhu@brandeis.edu

Hypothesis (A.1) can be tested by collapsing the data into averages of non-overlapping blocks of T_* observations over the time dimension. To simplify the exposition, we assume that T/T_* is an integer. Note that Assumption 1 implies the following model for the average potential outcomes $\bar{Y}_{1r}^N = T_*^{-1}\sum_{t=r}^{r+T_*-1} Y_t^N$ and $\bar{Y}_{1r}^I = T_*^{-1}\sum_{t=r}^{r+T_*-1} Y_t^I$:

$$\left.\begin{array}{l} \bar{Y}_{1r}^N = \bar{P}_r^N + \bar{u}_r \\ \bar{Y}_{1r}^I = \bar{P}_r^N + \bar{\theta}_r + \bar{u}_r \end{array}\right| \quad E(\bar{u}_r) = 0, \quad r = 1, T_*+1, 2T_*+1, \dots, T_0 - T_* + 1, T_0 + 1,$$

where $\bar{P}_r^N = T_*^{-1}\sum_{t=r}^{r+T_*-1} P_t^N$ and $\bar{u}_r = T_*^{-1}\sum_{t=r}^{r+T_*-1} u_t$. Define the aggregated (collapsed) data under the null as $\bar{\mathbf{Z}} = (\bar{Z}_1, \dots, \bar{Z}_{T_0+1})'$, where

$$\bar{Z}_r = \begin{cases} \left(\bar{Y}_{1r}^N, \bar{Y}_{2r}^N, \dots, \bar{Y}_{J+1r}^N, \bar{X}_{1r}', \dots, \bar{X}_{J+1r}'\right)', & r < T_0 + 1 \\ \left(\bar{Y}_{1r}^I - \bar{\theta}^o, \bar{Y}_{2r}^N, \dots, \bar{Y}_{J+1r}^N, \bar{X}_{1r}', \dots, \bar{X}_{J+1r}'\right)', & r = T_0 + 1 \end{cases}$$

and $\bar{X}_{jr} = T_*^{-1}\sum_{t=r}^{r+T_*-1} X_{jt}$ for $j = 1, \dots, J+1$. Note that testing hypothesis (A.1) is equivalent to testing a hypothesis concerning a per-period effect based on the aggregated data $\bar{\mathbf{Z}}$. Specifically, we estimate the average proxy $\hat{\bar{P}}_r^N$ based on the aggregated data $\bar{\mathbf{Z}}$ and obtain the residuals $\hat{\bar{u}} = (\hat{\bar{u}}_1, \hat{\bar{u}}_{T_*+1}, \dots, \hat{\bar{u}}_{T_0+1})'$, where $\hat{\bar{u}}_r = \bar{Y}_{1r}^N - \hat{\bar{P}}_r^N$. The test statistic is $S\left(\hat{\bar{u}}\right)$, and p-values can be obtained based on permutations of $\hat{\bar{u}}$ as described in Section 2.2.

The key assumption underlying this procedure is that the average mean proxy $\bar{P}_r^N$ can be identified and consistently estimated based on the aggregated data $\bar{\mathbf{Z}}$. This is the case for SC and the other regression-based estimators discussed in Sections 2.3.2–2.3.3, provided that $E(\bar{u}_r \bar{Y}_{jr}^N) = 0$ for $2 \leq j \leq J+1$ and that the sufficient conditions for consistent estimation in Section 4.2 hold for the aggregate data. By contrast, identification and estimation of $\bar{P}_r^N$ may not be possible for nonlinear and dynamic models. Under consistent estimation of $\hat{\bar{P}}_r^N$, the formal properties of the test follow from the results in Section 3.1 since stationarity and weak dependence of $\{u_t\}$ imply stationarity and weak dependence of $\{\bar{u}_r\}$. Alternatively, if $\hat{\bar{P}}_r^N$ can be shown to be stable and the aggregate data are stationary and weakly dependent, the properties of the test follow from the results in Section 3.2. Finally, we emphasize that the effective sample size is T/T_* instead of T, such that T needs to be substantially larger than T_*.

A.2 Multiple Treated Units

Our method can be extended to accommodate multiple treated units by collapsing the data into averages across the treated units. Consider a setup with L treated units, indexed by $j = 1, \dots, L$, and J control units, indexed by $j = L+1, \dots, J+L$. Suppose that Assumption

1 holds for all treated units:

$$\left.\begin{array}{l} Y_{jt}^N = P_{jt}^N + u_{jt} \\ Y_{jt}^I = P_{jt}^N + \theta_{jt} + u_{jt} \end{array}\right| \quad E(u_{jt}) = 0, \quad t = 1, \dots, T, \quad j = 1, \dots, L.$$

Under this assumption, hypotheses about the unit-specific policy effects $\{\theta_{jt}\}$ can be tested by separately applying the proposed inference procedure to each treated unit. In addition, one is often also interested in conducting inferences about the average treatment effects on the treated units, $\{\bar{\theta}_t\}$, where $\bar{\theta}_t = L^{-1} \sum_{j=1}^{L} \theta_{jt}$.

Specifically, consider the following null hypothesis:

$$H_0 : \left(\bar{\theta}_{T_0+1}, \dots, \bar{\theta}_T\right) = \left(\bar{\theta}_{T_0+1}^0, \dots, \bar{\theta}_T^0\right). \tag{A.2}$$

To test hypothesis (A.2), note that if Assumption 1 holds for all treated units, we have the following model for the average potential outcomes $\bar{Y}_t^N = L^{-1} \sum_{j=1}^{L} Y_{jt}^N$ and $\bar{Y}_t^I = L^{-1} \sum_{j=1}^{L} Y_{jt}^I$:

$$\left.\begin{array}{l} \bar{Y}_t^N = \bar{P}_t^N + \bar{u}_t \\ \bar{Y}_t^I = \bar{P}_t^N + \bar{\theta}_t + \bar{u}_t \end{array}\right| \quad E(\bar{u}_t) = 0, \quad t = 1, \dots, T,$$

where $\bar{P}_t^N = L^{-1} \sum_{j=1}^{L} P_{jt}^N$ and $\bar{u}_t = L^{-1} \sum_{j=1}^{L} u_{jt}$. Define the data under the null as $\bar{\mathbf{Z}} = (\bar{Z}_1, \dots, \bar{Z}_T)'$, where

$$\bar{Z}_t = \begin{cases} \left(\bar{Y}_t^N, Y_{L+1t}^N, \dots, Y_{J+Lt}^N, \bar{X}_t', X_{L+1t}', \dots, X_{J+Lt}'\right)', & t \leq T_0, \\ \left(\bar{Y}_t^I - \bar{\theta}_t^o, Y_{L+1t}^N, \dots, Y_{J+Lt}^N, \bar{X}_t', X_{L+1t}', \dots, X_{J+Lt}'\right)', & t > T_0, \end{cases}$$

and $\bar{X}_t = L^{-1} \sum_{j=1}^{L} X_{jt}$. To test hypothesis (A.2), we compute the estimated average proxy $\hat{\bar{P}}_t^N$ based on the aggregated data $\bar{\mathbf{Z}}$ and obtain the residuals $\hat{\bar{u}} = (\hat{\bar{u}}_1, \dots, \hat{\bar{u}}_T)'$, where $\hat{\bar{u}}_t = \bar{Y}_t^N - \hat{\bar{P}}_t^N$ for $t = 1, \dots, T$. The test statistic is $S\left(\hat{\bar{u}}\right)$, and p-values can be obtained based on permutations of $\hat{\bar{u}}$ as described in Section 2.2. The formal properties of this test follow from the results in Section 3.

A.3 Placebo Tests

Here we propose easy-to-implement placebo tests for assessing the credibility of inferences based on our method. We recommend applying these placebo tests when using our inference procedures.

Following Abadie et al. (2015), the idea is to consider a placebo intervention before the actual intervention took place. For a given $\tau \geq 1$, we use our method to test the null hypothesis

$$H_0 : \theta_{T_0-\tau+1} = \cdots = \theta_{T_0} = 0 \tag{A.3}$$

based in the pre-treatment data $\mathbf{Z} = (Z_1, \ldots, Z_{T_0})'$. Using an appropriate CSC method, we compute the counterfactual mean proxies $\hat{P}_t^N$ based on $\mathbf{Z}$ and obtain the residuals

$$\hat{u} = (\hat{u}_1, \ldots, \hat{u}_{T_0})', \quad \hat{u}_t = Y_{1t}^N - \hat{P}_t^N, \quad t = 1, \ldots, T_0.$$

We then apply the proposed inference method, treating $\{1, \ldots, T_0 - \tau\}$ as the pre-treatment period and $\{T_0 - \tau + 1, \ldots, T_0\}$ as the post-treatment period. The theoretical properties of such placebo tests follow directly from the results in Section 3. As illustrated in our application, it may be useful to complement the formal testing results with plots of the pre-treatment residuals $(\hat{u}_1, \ldots, \hat{u}_{T_0})$.

A rejection of the null hypothesis (A.3) undermines the credibility of the assumptions underlying our procedure and the inferences on the policy effects in the post-treatment period. While non-rejections provide evidence in favor of our method, it is important to emphasize that such non-rejections do not "prove" that our method is valid. Moreover, by construction, the placebo tests cannot be used to assess some of the key assumptions underlying our approach, such as the invariance of the distribution of $\{u_t\}$ under the intervention.

Given our inference method's genericness, the placebo tests discussed here can be used to assess and compare the credibility of different CSC methods. For example, in our empirical application, the placebo tests provide evidence in favor of SC and constrained Lasso, but suggest that the difference-in-differences results need to be interpreted with caution.

B Interpretation as a Structural Breaks Test

A key assumption underlying our method is the invariance of the distribution of $\{u_t\}$ under the intervention (Assumption 1). In settings where this assumption fails, following the literature on end-of-sample structural breaks tests (e.g., Andrews, 2003), our procedure can be used as a test of the null hypothesis that the policy has no impact whatsoever against the alternative hypothesis that $\theta \neq 0$ and or the policy affects the distribution of $\{u_t\}$.

Consider the following testing problem:

$$H_0 : \begin{cases} Y_{1t}^N = P_t^N + u_t, \quad E(u_t) = 0, \quad t = 1, \ldots, T, \quad \theta = 0, \quad \text{and} \\ \{u_t\}_{t=0}^T \text{ is stationary and weakly dependent.} \end{cases} \tag{B.1}$$

against

$$H_1 : \begin{cases} Y_{1t}^N = P_t^N + u_t, \quad t = 1, \ldots, T, \quad \theta \neq 0 \ \text{ and or the distribution of} \\ \{u_t\}_{t=T_0+1}^T \text{ differs from that of } \{u_t\}_{t=s}^{s+T_*-1} \text{ for } s = 1, \ldots, T_0 - T_* + 1. \end{cases} \tag{B.2}$$

Hypothesis (B.1) can be tested by applying our procedure to the data under the null where $Y_{1t}^N = Y_{1t}^I = Y_{1t}$ for all $t = 1, \ldots, T$. The theoretical size properties of this test follow directly from the results in Section 3. This test has power against location shifts induced by $\theta \neq 0$ as well as changes in the distribution of $\{u_t\}$ that increase the quantiles of $S(u)$ (e.g., scale shifts); see Andrews (2003, Section 2.5) for a related discussion.

C Prediction Sets for Random Policy Effects

In the main text, we assume that the policy effect sequence $\{\theta_t\}$ is fixed (Assumption 1). Here we show that our procedure generates valid prediction sets for θ_t when θ_t is assumed to be random, as, for example, in Cattaneo et al. (2021).[1] Our analysis here is in the same spirit as the classical conformal prediction literature, which aims at constructing prediction intervals for future values of a (random) target quantity of interest. Specifically, we will show that when θ_t is random, Algorithm 1 provides unconditionally valid $(1-\alpha)$-prediction sets with non-asymptotic performance guarantees.

To state the result, define $\mathbf{Z}^* = (Z_1^*, \ldots, Z_T^*)'$ with $Z_t^* = \left(Y_{1t}^N, Y_{2t}^N, \ldots, Y_{J+1t}^N, X_{1t}', \ldots, X_{J+1t}'\right)'$ for $1 \leq t \leq T_0 + T_*$. Notice that $\mathbf{Z}^*$ contains the true counterfactuals.

Theorem C.1 (Prediction Sets). *Assume that T_* is fixed. Suppose that $Y_{1t}^N = P_t^N + u_t$, $1 \leq t \leq T$, where $\{u_t\}$ is a centered and stationary stochastic process, and that the policy effects $\theta_t := Y_{1t}^I - Y_{1t}^N$ are random. Suppose that Assumption 3 holds for $\hat{P}_N$ computed using $\mathbf{Z}^*$. Impose Assumption 2.1 if $\Pi = \Pi_{all}$. Impose Assumption 2.2, if $\Pi = \Pi_{\rightarrow}$. Assume the statistic $S(u)$ has a density function bounded by D. Then*

$$|P\left(\theta_t \in \mathcal{C}_{1-\alpha}(t)\right) - (1-\alpha)| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T),$$

where $\mathcal{C}_{1-\alpha}(t)$ is defined in Algorithm 1 and $\tilde{\delta}_T = (T_*/T_0)^{1/4}(\log T)$. The constant C depends on T_*, M and D, but not on T.

[1] Following the literature on conformal prediction (e.g., Lei and Wasserman, 2014; Lei et al., 2018), we use the terminology "prediction set" instead of "confidence set" because the resulting sets are not conventional confidence sets; see Cattaneo et al. (2021) for a discussion of the differences between confidence intervals and prediction intervals for SC.

Let us briefly discuss the interpretation of our model when θ_t is regarded as random. Suppose that the potential outcomes are $Y_{1t}^I = P_t^I + u_t^I$ and $Y_{1t}^N = P_t^N + u_t^N$. The counterfactual mean proxies and the prediction errors under the policy, $\{P_t^I\}$ and $\{u_t^I\}$, may differ from the counterfactual mean proxies and the errors in the absence of the policy, $\{P_t^N\}$ and $\{u_t^N\}$. Algorithm 1 provides prediction intervals for

$$\theta_t := Y_{1t}^I - Y_{1t}^N = \Delta_t^P + \Delta_t^u,$$

where $\Delta_t^P := P_t^I - P_t^N$ and $\Delta_t^u := u_t^I - u_t^N$. Here θ_t captures both the effects of the policy on the mean proxy, Δ_t^P, and the effect on distribution of the error, Δ_t^u (e.g., a scale shift). Therefore, we are only assuming that in the absence of the policy, $\{u_t^N\}$ is a stationary process. All the changes to the mean, the variance, or other features of the distribution due to the policy are captured by the policy effects $\{\theta_t\}$.

In sum, with random policy effects, our procedure simply provides a prediction set for the policy effects defined as the difference between the two potential outcomes.

D Model-free Exact Validity under Exchangeability

Every permutation procedure that is approximately valid in time series settings should have good properties in "ideal" settings where the data are iid or exchangeable. The following theorem, which is based on standard arguments, shows that under exchangeability of the data, our conformal inference approach achieves exact finite sample size control. This result is model-free in the sense that we do not need to use a correct or consistent estimator for the counterfactual mean proxy. As a result, our procedure controls size under arbitrary forms of misspecification and is fully robust against overfitting.

Theorem D.1 (Exact Validity)**.** *Let Π be $\Pi_{\rightarrow}$ or Π_{all}. Suppose that the data $\{Z_t\}_{t=1}^T$ are iid or exchangeable with respect to Π under the null hypothesis and that $\hat{u}_t = g(Z_t, \hat{\beta})$, where the estimator $\hat{\beta} = \hat{\beta}(\{Z_t\}_{t=1}^T)$ is invariant with respect to any permutation of the data. Then, under the null hypothesis, $\{\hat{u}_t\}_{t=1}^T$ is an exchangeable sequence and the permutation p-value is unbiased in level:*

$$P(\hat{p} \leq \alpha) \leq \alpha.$$

Moreover, if $\{S(\hat{u}_\pi)\}_{\pi \in \Pi}$ has a continuous distribution,

$$\alpha - \frac{1}{|\Pi|} \leq P(\hat{p} \leq \alpha).$$

Theorem D.1 requires that the estimators are invariant under permutations of the data under the null hypothesis. Invariance holds for regression-based estimators such as SC, constrained Lasso, or penalized regression, provided that the null hypothesis is imposed for estimation but may fail for dynamic models such as linear and non-linear autoregressive models.

An inspection of the proof of Theorem D.1 shows that our procedure achieves finite sample size control whenever the residuals $\{\hat{u}_t\}_{t=1}^T$ are exchangeable. We demonstrate that exchangeability of $\{\hat{u}_t\}_{t=1}^T$ is implied if the data $\{Z_t\}_{t=1}^T$ are iid or exchangeable. Exchangeability $\{\hat{u}_t\}_{t=1}^T$ may hold even if the data are not exchangeable. For example, in the difference-in-difference model, the outcome data can have an arbitrary common trend eliminated by differencing, making it possible for the residuals to be iid or exchangeable with non-iid data.

The finite sample validity under exchangeability is crucial for the robustness of our proposal. While high-dimensional CSC approaches may overfit in small samples, Theorem D.1 shows that the proposed method does not suffer from overfitting or misspecification. The fundamental reason is that we exploit symmetry rather than to completely rely on the consistency of the estimator. Imposing the null hypothesis when estimating the model leads to invariance of the estimator under permutations, exchangeable residuals, and finite sample validity. As a result, the proposed procedure is more robust than alternative approaches that estimate the proxies based on the pre-treatment data without imposing the null hypothesis.

Even in time series settings where exchangeability fails, imposing the null hypothesis for estimation is crucial for achieving a good performance in typical CSC applications where T_0 is rather small (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Doudchenko and Imbens, 2016; Cunningham and Shah, 2018). Figure I.1 augments Figure 1 with results for $T = 100$ ($T_0 = 99, T_* = 1$). In the empirically relevant case where $T_0 = 19$, estimating P_t^N under the null yields an excellent performance. Irrespective of the degree of persistence, size accuracy is substantially better than when P_t^N is estimated based on pre-treatment data only. Even when $T_0 = 99$, which is much larger than the T_0 in many CSC applications, imposing the null yields notable performance improvements. In fact, the size accuracy is better for $T_0 = 19$ when P_t^N is estimated under the null than for $T_0 = 99$ when P_t^N is estimated based on the pre-treatment data only.

E Sufficient Conditions for Estimator Stability

In this section, we provide sufficient conditions for the estimator stability Assumption 4. We first present generic sufficient conditions for low-dimensional models. For high-dimensional models, the theoretical analysis is more difficult and a case-by-case analysis is needed. We are not aware of any theoretical work that establishes Assumption 4 for any high-dimensional model. Here we verify the stability condition for constrained Lasso; stability of Ridge regression is verified in Appendix F.

E.1 Generic Sufficient Condition for Low-dimensional Models

Consider $\hat{\beta}(\mathbf{Z}) = \arg\min_{\beta\in\mathcal{B}} \hat{L}(\mathbf{Z};\beta)$, where $\hat{L}(\mathbf{Z};\beta)$ is a loss function and $\mathcal{B}\subset\mathbb{R}^p$ for a fixed p. Let $\mathcal{H}$ be a set of subsets of $\{1,\ldots,T\}$. Notice that Assumption 4 only requires $\mathcal{H}$ to be a singleton, but in this subsection and the next, we allow $\mathcal{H}$ to be a class of subsets.

Lemma E.1. *Suppose that the following conditions hold:*

1. $\sup_{\beta\in\mathcal{B}} |\hat{L}(\mathbf{Z};\beta) - L(\beta)| = o_P(1)$ *for some non-random* $L(\cdot)$.
2. $\max_{H\in\mathcal{H}} \sup_{\beta\in\mathcal{B}} |\hat{L}(\mathbf{Z}_H;\beta) - L(\beta)| = o_P(1)$.
3. $L(\cdot)$ *is continuous at* β_*, $\min_\beta L(\beta)$ *has a unique minimum at* β_* *and* $\mathcal{B}$ *is compact.*

Then $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}) - \hat{\beta}(\mathbf{Z}_H)\|_2 = o_P(1)$.

In the literature of misspecified models, β_* is usually referred to as the pseudo-true value (e.g., White, 1996). In M-estimation with $\hat{L}(\mathbf{Z};\beta) = T^{-1}\sum_{t=1}^T l(Z_t;\beta)$, one can often show $\sup_\beta |\hat{L}(\mathbf{Z};\beta) - L(\beta)| = o_P(1)$ with $L(\beta) = El(Z_1;\beta)$; in GMM models with $\hat{L}(\mathbf{Z};\beta) = \|T^{-1}\sum_{t=1}^T \psi(Z_t;\beta)\|_2$, one can often use $L(\beta) = \|E\psi(Z_1;\beta)\|_2$.

The proof of Lemma E.1 shows that $\|\hat{\beta}(\mathbf{Z}) - \beta_*\|_2 = o_P(1)$ and $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 = o_P(1)$. In other words, the stability of the estimator arises from the consistency to the pseudo-true value β_*. Such consistency holds under very weak conditions. We essentially only require a uniform law of large numbers. This can be verified for many low-dimensional models under weakly dependent data. The conclusion of Lemma E.1 translates to Assumption 4 once we derive a bound on $\max_{\pi\in\Pi}\sup_{\beta_1\neq\beta_2} |S(\mathbf{Z}^\pi;\beta_1) - S(\mathbf{Z}^\pi;\beta_2)|/\|\beta_1-\beta_2\|_2$; this requires knowledge of the model structure.

For example, suppose that $\max_{\pi\in\Pi}\sup_{\beta_1\neq\beta_2} |S(\mathbf{Z}^\pi;\beta_1) - S(\mathbf{Z}^\pi;\beta_2)|/\|\beta_1-\beta_2\|_2 = O_P(1)$.[2] Then $\max_{\pi\in\Pi} |S(\mathbf{Z}^\pi;\hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}^\pi;\hat{\beta}(\mathbf{Z}_H))| = o_P(1)$. Here is how we apply Theorem 2 to obtain the asymptotic size control. Fix an arbitrary $\delta > 0$. We can simply choose the constant function $\varrho_T(x) = \delta$ for Assumption 4. Since $\max_{\pi\in\Pi} |S(\mathbf{Z}^\pi;\hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}^\pi;\hat{\beta}(\mathbf{Z}_H))| = o_P(1)$, Assumption 4 holds for some $\gamma_{1,T} = o(1)$ (due to the definition of convergence in probability). Assume that $\Psi(x,\beta)$ has bounded derivative with respect to x. Then we can choose $\xi_T = K_1$ for a large constant K_1 and $\gamma_{2,T} = 0$. As a result, Theorem 2 has

$$|P(\hat{p}\leq\alpha) - \alpha|$$

[2] We only need $S(\mathbf{Z};\beta)$ to be Lipschitz with respect to β. In the simple example of $S(\mathbf{Z};\beta) = |Y_{T_0+1} - X'_{T_0+1}\beta|$, this only requires $\|X_{T_0+1}\|_2$ to be bounded.

$$
\begin{aligned}
&\leq C_1\sqrt{\xi_T\varrho_T(T_0/R+2k)} + C_1\left(T_0^{-1}R[\log(T_0/R)]^{1/D_3}\right)^{1/4} + C_1\exp\left(-(k-T_*+1)^{1/D_3}\right)\\
&\quad + C_1\sqrt{\gamma_{1,T}} + C_1\sqrt{\gamma_{2,T}}\\
&= C_1\sqrt{K_1\delta} + C_1\left(T_0^{-1}R[\log(T_0/R)]^{1/D_3}\right)^{1/4} + C_1\exp\left(-(k-T_*+1)^{1/D_3}\right) + C_1\sqrt{o(1)}.
\end{aligned}
$$

Since $R \asymp T_0/\log(T_0)$, we have $T_0^{-1}R[\log(T_0/R)]^{1/D_3} = o(1)$. Choosing $k \asymp \log(T_0)$ and assuming that T_* is fixed, we obtain $\exp\left(-(k-T_*+1)^{1/D_3}\right) = o(1)$. Therefore, the above display implies

$$
|P(\hat{p} \leq \alpha) - \alpha| = C_1\sqrt{K_1\delta} + o(1).
$$

Since $\delta > 0$ is arbitrary, we have $|P(\hat{p} \leq \alpha) - \alpha| = o_P(1)$.

E.2 Constrained Lasso

Here we propose sufficient conditions for estimator stability for constrained Lasso. In contrast to Sections 2.3.2 and 4.2, we do not impose correct specification but study the behavior of the constrained Lasso estimator under potential misspecification. To make this explicit, we use β instead of w to denote the coefficient vector in this subsection. Here, it is possible that $EX_t(Y_t - X_t'\beta) \neq 0$ for any $\beta \in \mathcal{W}$. In practice, this arises when the relationship between X_t and Y_t is non-linear or when the constraint set $\mathcal{W}$ is too small. For example, the true parameter could be non-sparse with exploding ℓ_1-norm (e.g., $\beta = (1,\ldots,1)'/\sqrt{J}$).

We first introduce some additional notation. Define $Y_t = Y_{1t}^N$ and $X_t = (Y_{2t}^N,\ldots,Y_{J+1t}^N)'$ and let $\{(\tilde{Y}_t,\tilde{X}_t)\}_{t=1}^T$ be iid from the distribution of (Y_1,X_1) and independent of the data $\{(Y_t,X_t)\}_{t=1}^T$. The constrained Lasso objective functions based on the data under the original data and after switching out observations with $t\in H$ are given by

$$
\hat{Q}(\beta) = \frac{1}{T}\sum_{t=1}^T (Y_t - X_t'\beta)^2 \quad \text{and} \quad \hat{Q}_H(\beta) = T^{-1}\sum_{t=1}^T (Y_{t,H} - X_{t,H}'\beta)^2,
$$

where $(Y_{t,H},X_{t,H}) = (Y_t,X_t)$ for $t \notin H$ and $(Y_{t,H},X_{t,H}) = (\tilde{Y}_t,\tilde{X}_t)$ for $t\in H$. The corresponding constrained Lasso estimators are

$$
\hat{\beta}(\mathbf{Z}) = \arg\min_{\beta\in\mathcal{W}} \hat{Q}(\beta) \quad \text{and} \quad \hat{\beta}(\mathbf{Z}_H) = \arg\min_{\beta\in\mathcal{W}} \hat{Q}_H(\beta),
$$

where $\mathcal{W} \subseteq \{v \in \mathbb{R}^J : \|v\|_1 \leq K\}$ and $K > 0$ is a constant. Furthermore, we define $\hat{\Sigma} = T^{-1}\sum_{t=1}^T X_tX_t'$ and $\hat{\mu} = T^{-1}\sum_{t=1}^T X_tY_t$. Similarly, for $H \subset \{1,\ldots,T\}$, let $\hat{\Sigma}_H = T^{-1}\sum_{t=1}^T X_{t,H}X_{t,H}'$ and $\hat{\mu}_H = T^{-1}\sum_{t=1}^T X_{t,H}Y_{t,H}$. Finally, let $\mathcal{H}$ be a set of subsets of $\{1,\ldots,T\}$.

Lemma E.2. *Suppose that the following conditions hold:*

1. *with probability at least* $1-\gamma_{1,T}$, $\|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \leq c_T$ *and* $\|\hat{\mu}_H - \hat{\mu}\|_\infty \leq c_T$ *for all* $H \in \mathcal{H}$.
2. *with probability at least* $1-\gamma_{2,T}$, $\min_{\|v\|_0 \leq s} v'\hat{\Sigma}v/\|v\|_2^2 \geq \kappa_1$.
3. *with probability at least* $1-\gamma_{3,T}$, $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H)\|_0 \leq s/2$ *and* $\|\hat{\beta}(\mathbf{Z})\|_0 \leq s/2$.
4. $P(\max_{1\leq t\leq T} \|X_t\|_\infty \leq \kappa_2) = 1$.

Let $\hat{\varepsilon}_t = Y_t - X_t'\hat{\beta}(\mathbf{Z})$ *and* $\hat{\varepsilon}_{t,H} = Y_t - X_t'\hat{\beta}(\mathbf{Z}_H)$. *Then we have that*

$$P\left(\max_{H\in\mathcal{H}} \max_{1\leq t\leq T} |\hat{\varepsilon}_t - \hat{\varepsilon}_{t,H}| \leq 2\kappa_2\sqrt{\kappa_1 s c_T K(2K+1)}\right) \geq 1-\gamma_{1,T}-\gamma_{2,T}-\gamma_{3,T}.$$

Lemma E.2 provides sufficient conditions for perturbation stability. Inspecting the proof, we notice that the argument does not require the estimator to converge to anything. To our knowledge, this is the first result of this kind. In the conformal prediction literature, one-observation perturbation stability has been considered in Assumption A3 of Lei et al. (2018), who only verify it assuming correct model specification and consistent variable selection. There is also a strand of literature in statistics that considers misspecified models in high dimensions and focuses on the pseudo-true value. For example, for linear models, the pseudo-true value represents the best linear projection and is often assumed to be sparse, making it possible to establish consistency of Lasso to this pseudo-true value (e.g., Bühlmann and van de Geer, 2015). We do not make these assumptions. Lemma E.2 allows the model to be misspecified and the pseudo-true value may or may not be consistently estimated by constrained Lasso.

Lemma E.2 says that when the solution of constrained Lasso is sparse, the stability of $\hat{\Sigma}$ and $\hat{\mu}$ guarantees the stability of the estimator. When $|H| \asymp \log T_0$ and the observed variables are bounded, we can choose $c_T \asymp T_0^{-1}\log(T_0)$. The sparse eigenvalue condition can typically be verified whenever $s \leq cT$, where $c > 0$ is a constant that depends on the eigenvalues of $E\hat{\Sigma}$. Thus, Lemma E.2 would guarantee that when $\sup_{H\in\mathcal{H}} |H| \lesssim \log T_0$, we have

$$\max_{H\in\mathcal{H}} \max_{1\leq t\leq T} |\hat{\varepsilon}_t - \hat{\varepsilon}_{t,H}| = O_P(\sqrt{sT_0^{-1}\log T_0}).$$

Therefore, whenever the solutions $\hat{\beta}(\mathbf{Z})$ and $\hat{\beta}(\mathbf{Z}_H)$ are sparse enough with $s = o(T_0/\log(T_0))$, we can expect stability of the estimated residuals. One implication is that since $\|\hat{\beta}(\mathbf{Z})\|_0$ and $\|\hat{\beta}(\mathbf{Z}_H)\|_0$ are clearly bounded above by J, the stability should easily hold for $J \ll T_0/\log(T_0)$.

We now give an explicit formula for $\varrho_T(\cdot)$ in Assumption 4. Suppose that $S(\mathbf{Z},\beta) = |T_*^{-1}\sum_{t=T_0+1}^{T_0+T_*}(Y_t - X_t'\beta)|$. Then the above display implies that $\max_H |S(\mathbf{Z};\hat{\beta}(\mathbf{Z})) - S(\mathbf{Z};\hat{\beta}(\mathbf{Z}_H))| = O_P(\sqrt{sT_0^{-1}\log T_0})$. As in Appendix E.1, we can use Theorem 2 to obtain the asymptotic size control. Assume $s = o(T_0/\log(T_0))$. Let $q_T = (sT_0^{-1}\log(T_0))^{-1/4}$. Notice that $q_T \to \infty$. Since $q_T \to \infty$ and $|H| \lesssim \log(T_0)$, the above result of $\max_H |S(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - S(\mathbf{Z}; \hat{\beta}(\mathbf{Z}_H))| = O_P(\sqrt{sT_0^{-1}\log T_0})$ implies that there exists $\gamma_{1,T} = o(1)$ such that Assumption 4 holds with $\varrho_T(x) = q_T\sqrt{sT_0^{-1}x}$. Then by the same argument as in Appendix E.1, Theorem 2 implies

$$|P(\hat{p} \leq \alpha) - \alpha| \leq C_1\sqrt{K_1\varrho_T(T_0/R + 2k)} + o(1),$$

where K_1 is the same constant as in Appendix E.1. Since $k \leq T_0/R$, $R \asymp T_0/\log(T_0)$ and $q_T \to \infty$, we have

$$\varrho_T(T_0/R + 2k) \leq \varrho_T(3T_0/R) = \sqrt{3}q_T\sqrt{sT_0^{-1}(T_0/R)} \lesssim q_T\sqrt{sT_0^{-1}\log(T_0)} = 1/q_T \to 0.$$

Remark 1. Stability does not imply that the constrained Lasso residuals, $\hat{\varepsilon}_t$, are close to $\varepsilon_{*,t} = Y_t - X_t'\beta_*$, where $\beta_* = \arg\min_{\beta \in \mathcal{W}} E(Y_t - X_t'\beta)^2$ is a pseudo-true value. In Chernozhukov et al. (2019), we show that $\|\hat{\beta} - \beta_*\|_2 = O_P((T_0^{-1}\log J)^{1/4})$. However, this is far from enough to conclude that $|X_t'(\hat{\beta} - \beta_*)| = o_P(1)$ due to the high-dimensionality of X_t. The usual Cauchy-Schwarz bound $\|X_t\|_2\|\hat{\beta} - \beta_*\|_2$ would not converge to zero; the Hölder bound $\|X_t\|_\infty\|\hat{\beta} - \beta_*\|_1$ does not suffices either since $\|\hat{\beta} - \beta_*\|_1$ does not converge to zero. [When $\mathcal{W}$ is a bounded ℓ_1-ball, it is in fact impossible to achieve $\|\hat{\beta} - \beta_*\|_1 = o_P(1)$ (Ye and Zhang, 2010).] Even if β_* is assumed to be sparse, one would still require $\|\beta_*\|_0 = o(\sqrt{T_0}/\log J)$. In contrast, our stability condition discussed above only requires the much weaker condition of $\|\beta_*\|_0 = o(T_0/\log T_0)$. Therefore, stability condition could be satisfied even if the estimated residual does not converge to a pseudo-true target. $\square$

F Consistency and Estimator Stability

In this section, we illustrate the difference between stability and consistency using a simple and analytically tractable example: Ridge regression. We shall show that under correct specification, Ridge may not be consistent, while still satisfying estimator stability.

To keep theoretical analysis tractable, we work under stylized conditions. Given data $Z = (Y, X)$, we define the ridge estimator

$$\hat{\beta}_\lambda(Z) = (X'X + \lambda I_T)^{-1}X'Y,$$

where λ is a tuning parameter. Then we can compute the residuals $\hat{u}(Z; \lambda) = Y - X\hat{\beta}_\lambda(Z)$.

Lemma F.1. *Suppose that $Y = X\beta + u$, where $X = (X_1, \ldots, X_T)' \in \mathbb{R}^{T \times J}$ and $E(u \mid X) = 0$. Assume that $\|\beta\|_2$ is bounded away from zero and infinity, $E(uu' \mid X) = \sigma^2 I_T$ for* *a constant $\sigma > 0$, $J \lesssim T^{\kappa_0}$ with $\kappa_0 \in (0, 2/3)$ and for some constants $\kappa_1, \kappa_2 > 0$,*

$$P\left(\kappa_1 T \leq \lambda_{\min}(X'X) \leq \lambda_{\max}(X'X) \leq \kappa_2 T\right) \geq 1 - o(1)$$

and

$$P\left(\max_{1\leq t\leq T} \|X_t\| \leq \kappa_3\sqrt{J} \quad \text{and} \quad \|X'u\|_2 \leq \kappa_4\sqrt{JT}\right) \geq 1 - o(1).$$

Let $Z_H = (\tilde{Y}, \tilde{X})$ be the perturbed data with $|H| \asymp \log T$. For $\tilde{u} = \tilde{Y} - \tilde{X}\beta$, assume that

$$P\left(\|\tilde{X}'\tilde{u} - X'u\|_2 \leq \kappa_5\sqrt{J|H|}\right) \geq 1 - o(1)$$

and

$$P\left(\|\tilde{X}'\tilde{X} - X'X\| \leq \kappa_6(|H| + J)\right) \geq 1 - o(1),$$

where $\kappa_5, \kappa_6 > 0$ are constants. If $\lambda \asymp T$, then $E\|X(\hat{\beta}_\lambda(Z) - \beta)\|_2^2 \gtrsim T$ and $\|\hat{u}(Z;\lambda) - \hat{u}(Z_H;\lambda)\|_\infty = o_P(1)$[3]*.*

Lemma F.1 provides a robustness guarantee for the validity of the procedure. Under the ideal choice of the tuning parameter λ, we would expect consistency of $\hat{\beta}_\lambda(Z)$ and hence validity of the procedure. However, Lemma F.1 states that even when the tuning parameter is badly chosen such that consistency fails, one might still expect the estimator to be stable under perturbations, which is sufficient for the validity of our inference procedure. This is important in practice since computing optimal tuning parameters is often difficult.

Now we give very simple sufficient (but probably far from necessary) conditions for the assumptions of Lemma F.1. Suppose that the data is independent across t and rows of X are sub-Gaussian. Then standard results in random matrix theory can be used to verify these assumptions. For example, $X'X/T - E(X'X/T)$ has eigenvalues tending to zero (e.g., Theorem 5.39 and Remark 5.40 in Vershynin (2010)). Hence, as long as $E(X'X/T)$ has bounded eigenvalues, there exist constants κ_1, κ_2 that satisfy the condition in Lemma F.1. Let us further assume that entries of X are also bounded and entries of u are sub-Gaussian and independent of X. Then κ_3 can be chosen to be a large enough constant.

The existence of κ_4, κ_5 is a consequence of the Hanson-Wright inequality. We consider $\|X'u\|_2$ conditional on X. By Theorem 6.3.2 in Vershynin (2018) (proved using the Hanson-Wright inequality) applied to the conditional probability given X and by the definition of

[3]Notice that consistency is enough to derive asymptotic size control from Theorem 2; see the discussion at the end of Appendix E.1 for an explicit formula for quantities in Assumption 4.

sub-Gaussian variables, we have that with probability one, for any $z > 0$,

$$P\left(|\|X'u\|_2 - \sigma\|X\|_F| > z\|X\| \mid X\right) \le C_1 \exp\left(-C_2 z^2\right),$$

where $C_1, C_2 > 0$ are constants depending only on the sub-Gaussian norm of u. Choosing $z = \sqrt{\log(T)}$, we obtain

$$P\left(\|X'u\|_2 \le \sigma\|X\|_F + \sqrt{\log(T)}\|X\|\right) \le o(1).$$

Notice that $E\|X\|_F^2 = \sum_{j,t} EX_{j,t}^2 = O(JT)$ and $\|X\|_F^2 - E\|X\|_F^2 = \sum_t(\sum_j(X_{j,t}^2 - EX_{j,t}^2))$. By the boundedness of $X_{j,t}$, $\sum_j(X_{j,t}^2 - EX_{j,t}^2) = O(J)$ has mean zero and variance bounded by $O(J^2)$. It follows by the central limit theorem that $\|X\|_F^2 - E\|X\|_F^2 = \sum_t(\sum_j(X_{j,t}^2 - EX_{j,t}^2)) = \sqrt{T}J$. Since $E\|X\|_F^2 = O(JT)$, there exists a constant $C_3 > 0$ such that $P(\|X\|_F^2 < C_3JT) \ge 1 - o(1)$. Again by the random matrix theory (e.g., Theorem 5.39 and Remark 5.40 in Vershynin (2010)), $\|X\| = O_P(\sqrt{J+T})$. We have that with probability approaching one, $\|X'u\|_2 \le \sigma\|X\|_F + \sqrt{\log(T)}\|X\| \le \sqrt{C_3JT}\sigma + O(\sqrt{(J+T)\log(T)})$. Assume $J \gg \log(T)$. Since $(J+T)\log(T) \ll JT$, there exists κ_4 satisfying $P(\|X'u\|_2 \le \kappa_4\sqrt{JT}) \ge 1 - o(1)$. The same argument holds for κ_5; we just repeat the same argument with T replaced by $|H|$ once we realize $\tilde{X}'\tilde{u} - X'u = \sum_{t\in H}\tilde{X}_t\tilde{u}_t - \sum_{t\in H} X_t u_t$, where X_t denotes the t-th row of X. (Analogous notations apply to $\tilde{X}_t, u_t$, etc.)

To find κ_6, we observe that $\tilde{X}'\tilde{X} - X'X = \sum_{t\in H}\tilde{X}_t\tilde{X}_t' - \sum_{t\in H} X_tX_t'$. By the random matrix theory (e.g., Theorem 5.39 and Remark 5.40 in Vershynin (2010)), $\|\sum_{t\in H}(X_tX_t' - EX_tX_t')\| \le C_4\sqrt{J|H|}$ with probability approaching one, where $C_4 > 0$ is a constant. Again assuming that EX_tX_t' has eigenvalues bounded by a constant $C_5 > 0$, we have that $\|\sum_{t\in H} X_tX_t'\| \le C_5|H| + C_4\sqrt{J|H|}$ with probability approaching one. Recall the elementary inequality $\sqrt{J|H|} \le (J + |H|)/2$. Therefore, we can simply set $\kappa_6 = 4(C_4 + C_5)$ and obtain that with probability approaching one, $\|\sum_{t\in H} X_tX_t'\| \le (|H| + J)\kappa_6/2$. Similarly, we can get a bound for $\|\sum_{t\in H}\tilde{X}_t\tilde{X}_t'\|$.

The above analysis is based on simple sub-Gaussian or boundedness assumptions on (X, u). However, the purpose is to show that even in the simple case, Ridge regression can exhibit stability without being consistent (due to Lemma F.1). We expect similar results to hold in more general or complicated data-generating processes. We leave this extension for future research.

G Simulation Study

This section presents simulation evidence on the finite sample properties of our inference procedures. We consider the three CSC methods used in the empirical application in Section 5: difference-in-differences, canonical SC, and constrained Lasso with $K = 1$.

We consider different data generating processes (DGPs) for the treated unit all of which specify the treated outcome as a weighted combination of the control outcomes:

$$Y_{1t} = \begin{cases} \sum_{j=2}^{J+1} w_j Y_{jt}^N + u_t & \text{if } t \leq T_0, \\ \theta_t + \sum_{j=2}^{J+1} w_j Y_{jt}^N + u_t & \text{if } t > T_0, \end{cases}$$

where $u_t = \rho_u u_{t-1} + v_t$, $v_t \overset{iid}{\sim} N(0, 1 - \rho_u^2)$. Similar to Hahn and Shi (2017), the control outcomes are generated using a factor model:

$$Y_{jt}^N = \lambda_{1j} + F_{1t} + \lambda_{2j} F_{2t} + \epsilon_{jt},$$

where $\lambda_{1j} = (j-1)/J$, $\lambda_{2j} = (j-1)/J$, $F_{1t} \overset{iid}{\sim} N(0,1)$, and $\epsilon_{jt} = \rho_\epsilon \epsilon_{jt-1} + \xi_{jt}$, $\xi_{jt} \overset{iid}{\sim} N(0, 1-\rho_\epsilon^2)$. In the simulations, we vary ρ_u, ρ_ϵ, T_0, J, and F_{2t}. The DGPs differ with respect to the specification of the weights w.

	Weight Specification	**Correctly Specified Model(s)**
DGP1	$w = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	Difference-in-differences, SC, constrained Lasso
DGP2	$w = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, \ldots, 0\right)'$	SC, constrained Lasso
DGP3	$w = -1 \cdot \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)'$	constrained Lasso
DGP4	$w = (1, -1, 0, \ldots, 0)'$	–

We set $T_* = 1$ and consider the problem of testing the null hypothesis of a zero effect:

$$H_0 : \theta_T = 0.$$

The p-values are computed using the set of moving block permutations $\Pi_\rightarrow$. The nominal level is $\alpha = 0.1$.

We first analyze the performance of our procedure with stationary data, letting $F_{2t} \overset{iid}{\sim} N(0, 1)$. Table I.1 presents simulation evidence on the size properties of our method when the data are iid ($\rho_u = \rho_\epsilon = 0$), which implies exchangeability of the residuals (cf. Theorem D.1). As expected, our procedure achieves exact size control, irrespective of whether or not the model for P_t^N is correctly specified. To study the finite sample performance with dependent data, we set $\rho_u = \rho_\epsilon = 0.6$. Table I.2 shows that our method exhibits close-to-correct size under correct specification and under misspecification, confirming the theoretical results on the robustness of our procedure under estimator stability and stationarity.

To investigate the performance with non-stationary data, we use trending factors: $F_{2t} \sim N(t, 1)$. Tables I.3 and I.4 show that under correct specification, our method exhibits excellent size properties. However, unlike in stationary settings, misspecification can cause size distortions. This finding is expected given the theoretical results and discussions in Section 3.2.

Figure I.2 displays power curves for a setting where $T_0 = 19$ and $J = 50$ as in our empirical application, $\rho_u = \rho_\epsilon = 0.6$, and $F_{2t} \overset{iid}{\sim} N(0, 1)$. Under correct specification, our method exhibits excellent small sample power properties and comes close to achieving the oracle power bound based on the true marginal distribution of u_t. Moreover, we find that imposing additional constraints when estimating P_t^N (e.g., using SC instead of the more general constrained Lasso) does not improve power when these additional restrictions are correct but can cause power losses when they are not.

H Proofs

Additional Notation

We introduce some additional notations that will be used in the proofs. For $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$. For two positive sequences a_n, b_n (indexed by n), we use $a_n \ll b_n$ to denote $a_n = o(b_n)$. We use $\Phi(\cdot)$ to denote the cumulative distribution function of the standard normal distribution. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors or the spectral norm for matrices. We use $\overset{d}{=}$ to denote equal in distribution.

H.1 Proof of Theorem 1

The proof proceeds by verifying the high-level conditions in the following lemma. Let $n = |\Pi|$ so that $n = T!$ if $\Pi = \Pi_{\text{all}}$ and $n = T$ if $\Pi = \Pi_{\rightarrow}$.

Lemma H.1 (Approximate Validity under High-Level Conditions[4]). *Let $\{\delta_{1n}, \delta_{2n}, \gamma_{1n}, \gamma_{2n}\}$ be sequences of numbers converging to zero. Assume the following conditions.*

[4]In Chernozhukov et al. (2018), we use a version of Lemma H.1, which relies on permuting the data instead of permuting the residuals, to derive performance guarantees for prediction intervals obtained using classical conformal prediction methods with weakly dependent data. The proof of Lemma H.1 (presented in Section H.1.1 to make the exposition self-contained) follows from the same arguments as the proof of Theorem 2 in Chernozhukov et al. (2018) modified to the problem of permuting residuals.

(E) With probability $1-\gamma_{1n}$: the randomization distribution

$$\tilde{F}(x) := \frac{1}{n} \sum_{\pi \in \Pi} \mathbf{1}\{S(u_\pi) < x\},$$

is approximately ergodic for $F(x) = P\left(S(u) < x\right)$, namely

$$\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F\left(x\right) \right| \leq \delta_{1n},$$

(A) With probability $1-\gamma_{2n}$, estimation errors are small:

1. *the mean squared error is small, $n^{-1} \sum_{\pi \in \Pi} \left[S(\hat{u}_\pi) - S(u_\pi)\right]^2 \leq \delta_{2n}^2$;*
2. *the pointwise error at $\pi = \text{Identity}$ is small, $|S(\hat{u}) - S(u)| \leq \delta_{2n}$;*
3. *The pdf of $S(u)$ is bounded above by a constant D.*

Suppose in addition that the null hypothesis is true. Then, the approximate p-value obeys for any $\alpha \in (0,1)$

$$\left|P\left(\hat{p} \leq \alpha\right) - \alpha\right| \leq 4\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n}.$$

With this result at hand, the proof of the theorem is a consequence of following four lemmas, which verify the approximate ergodicity conditions (E) and conditions on the estimation error (A) of Lemma H.1. Putting the bounds together and optimizing the error yields the result of the theorem.

The following lemma verifies approximate ergodicity (E) (which allows for large T_*) for the case of moving block permutations.

Lemma H.2 (Approximate Ergodicity under Moving Block Permutations)**.** *Let Π be the moving block permutations. Suppose that $\{u_t\}_{t=1}^T$ is stationary and strong mixing. Assume the following conditions: (1) $\sum_{k=1}^\infty \alpha_{mixing}(k)$ is bounded by a constant M, (2) $T_0 \geq T_* + 2$, and (3) $S(u)$ has bounded pdf. Then there exists a constant $M' > 0$ depending only on M such that for any $\delta_{1n} > 0$,*

$$P\left(\sup_{x \in \mathbb{R}} \left|\tilde{F}(x) - F(x)\right| \leq \delta_{1n}\right) \geq 1 - \gamma_T,$$

where $\gamma_T = \left(M' \sqrt{\frac{T_*}{T_0}} \log T_0 + \frac{T_*+1}{T_0+T_*}\right) / \delta_{1n}$.

The following lemma verifies approximate ergodicity (E) (which allows for large T_*) for the case of iid permutations.

Lemma H.3 (Approximate Ergodicity under iid Permutations)**.** *Let* Π *be the set of all permutations. Suppose that* $\{u_t\}_{t=1}^T$ *is iid. Assume that* $S(u)$ *only depends on the last* T_* *entries of* u*. If* $T_0 \geq T_* + 2$*, then*

$$P\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}(x) - F(x)\right| \leq \delta_{1n}\right) \geq 1 - \gamma_T,$$

where $\gamma_T = \sqrt{\pi/(2\lfloor T/T_*\rfloor)}/\delta_{1n}$*.*

The following lemma verifies the condition on the estimation error (A) for moving block permutations.

Lemma H.4 (Bounds on Estimation Errors under Moving Block Permutations)**.** *Consider moving block permutations* Π*. Let* T_* *be fixed. Suppose that for some constant* $Q > 0$*,* $|S(u) - S(v)| \leq Q\|D_{T_*}(u-v)\|_2$ *for any* $u, v \in \mathbb{R}^T$ *and* $D_{T_*} := \mathrm{Blockdiag}(0_{T_*}, I_{T_*})$*. Then Condition (A) (1)-(2) is satisfied if there exist sequences* $\gamma_T, \delta_{2n} = o(1)$ *such that with probability at least* $1 - \gamma_T$*,*

$$\|\hat{P}^N - P^N\|_2/\sqrt{T} \leq \delta_{2n} \text{ and } |\hat{P}_t^N - P_t| \leq \delta_{2n} \text{ for } T_0 + 1 \leq t \leq T.$$

The following lemma verifies the condition on the estimation error (A) for iid permutations.

Lemma H.5 (Bounds on Estimation Errors under iid Permutations)**.** *Consider the set of all permutations* Π*. Let* T_* *be fixed. Suppose that for some constant* $Q > 0$*,* $|S(u) - S(v)| \leq Q\|D_{T_*}(u-v)\|_2$ *for any* $u, v \in \mathbb{R}^T$ *and* $D_{T_*} := \mathrm{Blockdiag}(0, I_{T_*})$*. Then Condition (A) (1)-(2) is satisfied if there exist sequences* $\gamma_T, \delta_{2n} = o(1)$ *such that with probability at least* $1 - \gamma_T$*,*

$$\|\hat{P}^N - P^N\|_2/\sqrt{T} \leq \delta_{2n} \text{ and } |\hat{P}_t^N - P_t| \leq \delta_{2n} \text{ for } T_0 + 1 \leq t \leq T.$$

Now we conclude the proof of Theorem 1.

For the moving block permutations, let $\delta_{1n} = (T_*/T_0)^{1/4}$. Then we apply Lemma H.1 together with Lemmas H.2 and H.4, obtaining

$$\begin{aligned}
|P(\hat{p} \leq \alpha) - \alpha| &\leq 4\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \gamma_{1n} + \gamma_{2n} \\
&\leq 4\delta_{1n} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}}) + \left(M'\sqrt{\frac{T_*}{T_0}}\log T_0 + \frac{T_* + 1}{T_0 + T_*}\right)/\delta_{1n} + \gamma_{2n} \\
&\leq 4(T_*/T_0)^{1/4} + 4\delta_{2n} + 2D(\delta_{2n} + 2\sqrt{\delta_{2n}})
\end{aligned}$$

$$+\left(M'\sqrt{\frac{T_*}{T_0}}\log T_0+\frac{T_*+1}{T_0+T_*}\right)(T_*/T_0)^{-1/4}+\gamma_{2n}.$$

The final result for moving block permutations follows by straight-forward computations and the observations that $\delta_{2n}=O(\sqrt{\delta_{2n}})$ (due to $\delta_{2n}=o(1)$).

For iid permutations, we also use $\delta_{1n}=(T_*/T_0)^{1/4}$. Then we apply Lemma H.1 together with Lemmas H.3 and H.5, obtaining

$$\begin{aligned}
|P(\hat{p}\leq\alpha)-\alpha| &\leq 4\delta_{1n}+4\delta_{2n}+2D(\delta_{2n}+2\sqrt{\delta_{2n}})+\gamma_{1n}+\gamma_{2n}\\
&\leq 4\delta_{1n}+4\delta_{2n}+2D(\delta_{2n}+2\sqrt{\delta_{2n}})+\sqrt{2\pi/\lfloor T/T_*\rfloor}/\delta_{1n}+\gamma_{2n}\\
&\leq 4(T_*/T_0)^{1/4}+4\delta_{2n}+2D(\delta_{2n}+2\sqrt{\delta_{2n}})+\sqrt{2\pi/\lfloor T/T_*\rfloor}(T_*/T_0)^{-1/4}+\gamma_{2n}\\
&\lesssim (T_*/T_0)^{1/4}+\delta_{2n}+\sqrt{\delta_{2n}}+\gamma_{2n}.
\end{aligned}$$

This completes the proof for iid permutations.

H.1.1 Proof of Lemma H.1

The proof proceeds in two steps.[5]

Step 1: We bound the difference between the p-value and the oracle p-value, $\hat{F}(S(\hat{u}))-F(S(u))$.

Let $\mathcal{M}$ be the event that the conditions (A) and (E) hold. By assumption,

$$P(\mathcal{M})\geq 1-\gamma_{1n}-\gamma_{2n}. \tag{H.1}$$

Notice that on the event $\mathcal{M}$,

$$\begin{aligned}
\left|\hat{F}(S(\hat{u}))-F(S(u))\right| &\leq \left|\hat{F}(S(\hat{u}))-F(S(\hat{u}))\right|+|F(S(\hat{u}))-F(S(u))|\\
&\overset{\text{(i)}}{\leq} \sup_{x\in\mathbb{R}}\left|\hat{F}(x)-F(x)\right|+D\,|S(\hat{u})-S(u)|\\
&\leq \sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\tilde{F}(x)\right|+\sup_{x\in\mathbb{R}}\left|\tilde{F}(x)-F(x)\right|+D\,|S(\hat{u})-S(u)|\\
&\leq \sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\tilde{F}(x)\right|+\delta_{1n}+D\,|S(\hat{u})-S(u)|\\
&\leq \sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\tilde{F}(x)\right|+\delta_{1n}+D\delta_{2n},
\end{aligned}\tag{H.2}$$

where (i) holds by the fact that the bounded pdf of $S(u)$ implies Lipschitz property for F.

[5]The proof follows from the same arguments as the proof of Theorem 2 in Chernozhukov et al. (2018) modified to the problem of permuting residuals and is presented for completeness.

Let $A=\left\{\pi\in\Pi:\ |S(\hat{u}_\pi)-S(u_\pi)|\geq\sqrt{\delta_{2n}}\right\}$. Observe that on the event $\mathcal{M}$, by Chebyshev inequality

$$|A|\delta_{2n}\leq\sum_{\pi\in\Pi}\left(S(\hat{u}_\pi)-S(u_\pi)\right)^2\leq n\delta_{2n}^2$$

and thus $|A|/n\leq\delta_{2n}$. Also observe that on the event $\mathcal{M}$, for any $x\in\mathbb{R}$,

$$\begin{aligned}
&\left|\hat{F}(x)-\tilde{F}(x)\right|\\
&\leq\frac{1}{n}\sum_{\pi\in A}\left|\mathbf{1}\left\{S(\hat{u}_\pi)<x\right\}-\mathbf{1}\left\{S(u_\pi)<x\right\}\right|+\frac{1}{n}\sum_{\pi\in(\Pi\setminus A)}\left|\mathbf{1}\left\{S(\hat{u}_\pi)<x\right\}-\mathbf{1}\left\{S(u_\pi)<x\right\}\right|\\
&\overset{\text{(i)}}{\leq}\frac{|A|}{n}+\frac{1}{n}\sum_{\pi\in(\Pi\setminus A)}\mathbf{1}\left\{|S(u_\pi)-x|\leq\sqrt{\delta_{2n}}\right\}\leq\frac{|A|}{n}+\frac{1}{n}\sum_{\pi\in\Pi}\mathbf{1}\left\{|S(u_\pi)-x|\leq\sqrt{\delta_{2n}}\right\}\\
&\leq\frac{|A|}{n}+P\left(|S(u)-x|\leq\sqrt{\delta_{2n}}\right)\\
&\qquad+\sup_{z\in\mathbb{R}}\left|\frac{1}{n}\sum_{\pi\in\Pi}\mathbf{1}\left\{|S(u_\pi)-z|\leq\sqrt{\delta_{2n}}\right\}-P\left(|S(u)-z|\leq\sqrt{\delta_{2n}}\right)\right|\\
&=\frac{|A|}{n}+P\left(|S(u)-x|\leq\sqrt{\delta_{2n}}\right)\\
&\qquad+\sup_{x\in\mathbb{R}}\left|\left[\tilde{F}\left(z+\sqrt{\delta_{2n}}\right)-\tilde{F}\left(z-\sqrt{\delta_{2n}}\right)\right]-\left[F\left(z+\sqrt{\delta_{2n}}\right)-F\left(z-\sqrt{\delta_{2n}}\right)\right]\right|\\
&\leq\frac{|A|}{n}+P\left(|S(u)-x|\leq\sqrt{\delta_{2n}}\right)+2\sup_{z\in\mathbb{R}}\left|\tilde{F}(z)-F\left(z\right)\right|\\
&\overset{\text{(ii)}}{\leq}\frac{|A|}{n}+2D\sqrt{\delta_{2n}}+2\delta_{1n}\overset{\text{(iii)}}{\leq}\delta_{1n}+2\delta_{2n}+2D\sqrt{\delta_{2n}},
\end{aligned}\tag{H.3}$$

where (i) follows by the elementary inequality of $|\mathbf{1}\{S(\hat{u}_\pi)<x\}-\mathbf{1}\{S(u_\pi)<x\}|\leq\mathbf{1}\{|S(u_\pi)-x|\leq|S(\hat{u}_\pi)-S(u_\pi)|\}$, (ii) follows by the bounded pdf of $S(u)$ and (iii) follows by $|A|/n\leq\delta_{2n}$. Since the above display holds for each $x\in\mathbb{R}$, it follows that on the event $\mathcal{M}$,

$$\sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\tilde{F}(x)\right|\leq\delta_{1n}+2\delta_{2n}+2D\sqrt{\delta_{2n}}.\tag{H.4}$$

We combine (H.2) and (H.4) and obtain that on the event $\mathcal{M}$,

$$\left|\hat{F}(S(\hat{u}))-F(S(u))\right|\leq2\delta_{1n}+2\delta_{2n}+D(\delta_{2n}+2\sqrt{\delta_{2n}}).\tag{H.5}$$

Step 2: Here we derive the desired result. Notice that

$$\begin{aligned}
&\left|P\left(1-\hat{F}(S(\hat{u}))\leq\alpha\right)-\alpha\right|\\
&=\left|E\left(\mathbf{1}\left\{1-\hat{F}(S(\hat{u}))\leq\alpha\right\}-\mathbf{1}\left\{1-F(S(u))\leq\alpha\right\}\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq E\left|\mathbf{1}\left\{1-\hat{F}(S(\hat{u}))\leq\alpha\right\}-\mathbf{1}\left\{1-F(S(u))\leq\alpha\right\}\right| \\
&\overset{\text{(i)}}{\leq} P\left(|F(S(u))-1+\alpha|\leq\left|\hat{F}(S(\hat{u}))-F(S(u))\right|\right) \\
&\leq P\left(|F(S(u))-1+\alpha|\leq\left|\hat{F}(S(\hat{u}))-F(S(u))\right| \text{ and } \mathcal{M}\right)+P(\mathcal{M}^c) \\
&\overset{\text{(ii)}}{\leq} P\left(|F(S(u))-1+\alpha|\leq 2\delta_{1n}+2\delta_{2n}+D(\delta_{2n}+2\sqrt{\delta_{2n}})\right)+P\left(\mathcal{M}^c\right) \\
&\overset{\text{(iii)}}{\leq} 4\delta_{1n}+4\delta_{2n}+2D(\delta_{2n}+2\sqrt{\delta_{2n}})+\gamma_{1n}+\gamma_{2n},
\end{aligned}
$$

where (i) follows by the elementary inequality $|\mathbf{1}\{1-\hat{F}(S(\hat{u}))\leq\alpha\}-\mathbf{1}\{1-F(S(u))\leq\alpha\}|\leq \mathbf{1}\{|F(S(u))-1+\alpha|\leq|\hat{F}(S(\hat{u}))-F(S(u))|\}$, (ii) follows by (H.5), (iii) follows by the fact that $F(S(u))$ has the uniform distribution on $(0,1)$ and hence has pdf equal to 1, and by (H.1). The proof is complete.

H.1.2 Proof of Lemma H.2

We define

$$
s_t=\begin{cases}(\sum_{s=t}^{t+T_*-1}|u_s|^q)^{1/q} & \text{if } 1\leq t\leq T_0\\ (\sum_{s=t}^{T}|u_s|^q+\sum_{s=1}^{t-T_0-1}|u_s|^q)^{1/q} & \text{otherwise.}\end{cases}
$$

It is straight-forward to verify that

$$
\{S(u_\pi):\ \pi\in\Pi\}=\{s_t:\ 1\leq t\leq T\}.
$$

Let $\tilde{\alpha}_{\text{mixing}}$ be the strong-mixing coefficient for $\{s_t\}_{t=1}^{T_0}$. Notice that $\{s_t\}_{t=1}^{T_0}$ is stationary (although $\{s_t\}_{t=1}^{T}$ is clearly not). Let $\check{F}(x)=T_0^{-1}\sum_{t=1}^{T_0}\mathbf{1}\{s_t\leq x\}$. The bounded pdf of $S(u)$ implies the continuity of $F(\cdot)$. It follows, by Proposition 7.1 of Rio (2017), that

$$
E\left(\sup_{x\in\mathbb{R}}\left|\check{F}(x)-F(x)\right|^2\right)\leq\frac{1}{T_0}\left(1+4\sum_{k=0}^{T_0-1}\tilde{\alpha}_{\text{mixing}}(t)\right)\left(3+\frac{\log T_0}{2\log 2}\right)^2. \tag{H.6}
$$

Notice that $\tilde{\alpha}_{\text{mixing}}(t)\leq 2$ and that $\tilde{\alpha}_{\text{mixing}}(t)\leq\alpha_{\text{mixing}}\left(\max\{t-T_*,0\}\right)$ so that

$$
\begin{aligned}
\sum_{k=0}^{T_0-1}\tilde{\alpha}_{\text{mixing}}(t) &= \sum_{k=0}^{T_*}\tilde{\alpha}_{\text{mixing}}(t)+\sum_{k=T_*+1}^{T_0-1}\tilde{\alpha}_{\text{mixing}}(t)\leq 2(T_*+1)+\sum_{k=1}^{T_0-T_*-1}\alpha_{\text{mixing}}(k)\\
&\leq 2(T_*+1)+\sum_{k=1}^{\infty}\alpha_{\text{mixing}}(k).
\end{aligned}
$$

Since $\sum_{k=1}^{\infty} \alpha_{\text{mixing}}(k)$ is bounded by M, it follows by (H.6) that

$$E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|^2\right) \leq B_T := \frac{1 + 4(2(T_* + 1) + M)}{T_0} \left(3 + \frac{\log T_0}{2\log 2}\right)^2.$$

By Liapunov's inequality,

$$E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|\right) \leq \sqrt{E\left(\sup_{x\in\mathbb{R}} \left|\check{F}(x) - F(x)\right|^2\right)} \leq \sqrt{B_T}.$$

Since $(T_0 + T_*)\tilde{F}(x) - T_0\check{F}(x) = \sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \leq x\}$, it follows that

$$\begin{aligned}
\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - \check{F}(x)\right| &= \sup_{x\in\mathbb{R}} \left|\left(\frac{T_0}{T_0 + T_*}\check{F}(x) + \frac{1}{T_0 + T_*}\sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \leq x\}\right) - \check{F}(x)\right| \\
&= \sup_{x\in\mathbb{R}} \left|\frac{1}{T_0 + T_*}\check{F}(x) + \frac{1}{T_0 + T_*}\sum_{t=T_0+1}^{T_0+T_*} \mathbf{1}\{s_t \leq x\}\right| \leq \frac{T_* + 1}{T_0 + T_*},
\end{aligned}$$

where the last inequality follows by $\sup_{x\in\mathbb{R}} |\check{F}(x)| \leq 1$ and the boundedness of the indicator function. Combining the above two displays, we obtain that

$$E\left(\sup_{x\in\mathbb{R}} \left|\tilde{F}(x) - F(x)\right|\right) \leq \sqrt{B_T} + \frac{T_* + 1}{T_0 + T_*}.$$

The desired result follows by Markov's inequality.

H.1.3 Proof of Lemma H.3

The proof follows by an argument given by Romano and Shaikh (2012) for subsampling. We give a complete argument for our setting here for clarity and completeness.

Recall that Π is the set of all bijections π on $\{1, ..., T\}$. Let $k_T = \lfloor T/T_* \rfloor$. Define the blocks of indices

$$b_i = (T - iT_* + 1, T - iT_* + 2, ..., T - iT_* + T_*) \in \mathbb{R}^{T_*}, \qquad i = 1, ..., k_T$$

Since $S(u)$ only depends on u_{b_1}, the last T_* entries of u, we can define

$$Q(x; u_{b_1}) = \mathbf{1}\{S(u) \leq x\} - F(x).$$

Therefore,

$$\tilde{F}(x) - F(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} Q(u_{\pi(b_1)}; x).$$

Define $\pi(b_i) := \pi_{|b_i}(b_i)$ to mean the restriction of the permutation map $\pi : \{1, \dots, T\} \to \{1, \dots, T\}$ to the domain b_i.

Notice that for $1 \le i \le k_T$, the value of $\sum_{\pi \in \Pi} Q(u_{\pi(b_i)}; x)$ does not depend on i. It follows that

$$\begin{aligned}\tilde{F}(x) - F(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} Q(u_{\pi(b_1)}; x) &= \frac{1}{k_T} \sum_{i=1}^{k_T} \left(\frac{1}{|\Pi|} \sum_{\pi \in \Pi} Q(u_{\pi(b_i)}; x) \right) \\ &= \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \left[\frac{1}{k_T} \sum_{i=1}^{k_T} Q(u_{\pi(b_i)}; x) \right].\end{aligned}$$

Hence by Jensen's inequality

$$E\left(\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F(x) \right| \right) \le \frac{1}{|\Pi|} \sum_{\pi \in \Pi} E\left(\sup_{x \in \mathbb{R}} \left| \frac{1}{k_T} \sum_{i=1}^{k_T} Q(u_{\pi(b_i)}; x) \right| \right).$$

To compute the above expectation, we observe that for any $\pi \in \Pi$,

$$\begin{aligned}E\left(\sup_{x \in \mathbb{R}} \left| \frac{1}{k_T} \sum_{i=1}^{k_T} Q(u_{\pi(b_i)}; x) \right| \right) &= \int_0^1 P\left(\sup_{x \in \mathbb{R}} \left| \frac{1}{k_T} \sum_{i=1}^{k_T} Q(u_{\pi(b_i)}; x) \right| > z \right) dz \\ &\le \int_0^1 2\exp\left(-2k_T z^2\right) dz < \int_0^\infty 2\exp\left(-2k_T z^2\right) dz = \sqrt{\pi/(2k_T)},\end{aligned}$$

where the first inequality follows by the Dvoretsky-Kiefer-Wolfwitz inequality (e.g., Theorem 11.6 in Kosorok (2007)) and the fact that for any $\pi \in \Pi$, $\{Q(u_{\pi(b_i)}; x)\}_{i=1}^{k_T}$ is a sequence of iid random variables (since π is a bijection and $\{b_i\}_{i=1}^{k_T}$ are disjoint blocks of indices); the last equality follows from the properties of the normal density. Therefore, the above two display imply that

$$E\left(\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F(x) \right| \right) \le \sqrt{\pi/(2k_T)}.$$

The desired result follows by Markov's inequality.

H.1.4 Proof of Lemma H.4

Due to the Lipschitz property of $S(\cdot)$, we have

$$\begin{aligned}\sum_{\pi\in\Pi}[S(\hat{u}_\pi)-S(u_\pi)]^2 \le Q\sum_{\pi\in\Pi}\|D_{T_*}(\hat{u}_\pi-u_\pi)\|_2^2 = Q\sum_{\pi\in\Pi}\sum_{t=T_0+1}^{T_0+T_*}\left(\hat{u}_{\pi(t)}-u_{\pi(t)}\right)^2 \\ = Q\sum_{t=T_0+1}^{T_0+T_*}\sum_{\pi\in\Pi}\left(\hat{u}_{\pi(t)}-u_{\pi(t)}\right)^2 = QT_*\|\hat{u}-u\|_2^2 = QT_*\|\hat{P}^N-P^N\|^2\end{aligned}$$

where the penultimate equality follows by the observation that for moving block permutation Π,

$$\sum_{\pi\in\Pi}\left(\hat{u}_{\pi(t)}-u_{\pi(t)}\right)^2 = \|\hat{u}-u\|_2^2.$$

Hence condition (A) (1) follows with a rescaled value of δ_n. Condition (A) (2) holds by the Lipschitz property of $S(\cdot)$:

$$|S(\hat{u})-S(u)| \le Q\|D_{T_*}(\hat{u}-u)\|_2 \le Q\sqrt{\sum_{t=T_0+1}^{T_0+T_*}(\hat{u}_t-u_t)^2}$$

Hence, Condition (A) (2) follows since $\|\hat{P}_t^N-P_t^N\| = |\hat{u}_t-u_t| \le \delta_n$ for $T_0+1\le t\le T$ with high probability. The proof is complete.

H.1.5 Proof of Lemma H.5

For $t,s\in\{1,...,T\}$, we define $A_{t,s}=\{\pi\in\Pi:\ \pi(t)=s\}$. Recall that Π is the set of all bijections on $\{1,...,T\}$. Thus, $|A_{t,s}|=(T-1)!$. It follows that for any $t\in\{1,...,T\}$,

$$\begin{aligned}\sum_{\pi\in\Pi}\left(\hat{u}_{\pi(t)}-u_{\pi(t)}\right)^2 &= \sum_{s=1}^{T}\sum_{\pi\in A_{t,s}}\left(\hat{u}_{\pi(t)}-u_{\pi(t)}\right)^2 \\ &= \sum_{s=1}^{T}\sum_{\pi\in A_{t,s}}\left(\hat{u}_s-u_s\right)^2 = \sum_{s=1}^{T}|A_{t,s}|\left(\hat{u}_s-u_s\right)^2 = (T-1)!\times\|\hat{u}-u\|_2^2.\end{aligned} \tag{H.7}$$

Due to the Lipschitz property of $S(\cdot)$, we have that

$$\frac{1}{|\Pi|}\sum_{\pi\in\Pi}[S(\hat{u}_\pi)-S(u_\pi)]^2 \le \frac{Q}{|\Pi|}\sum_{\pi\in\Pi}\|D_{T_*}(\hat{u}_\pi-u_\pi)\|_2^2 = \frac{Q}{|\Pi|}\sum_{\pi\in\Pi}\sum_{t=T_0+1}^{T_0+T_*}\left(\hat{u}_{\pi(t)}-u_{\pi(t)}\right)^2$$

$$= \frac{Q}{|\Pi|} \sum_{t=T_0+1}^{T_0+T_*} \sum_{\pi \in \Pi} \left(\hat{u}_{\pi(t)} - u_{\pi(t)} \right)^2 = \frac{Q}{|\Pi|} T_*(T-1)! \times \|\hat{u} - u\|_2^2 = QT^{-1}T_* \|\hat{u} - u\|_2^2,$$

where the penultimate equality follows by (H.7) and the last equality follows by $|\Pi| = T!$. Thus, part 1 of Condition (A) follows since T_* is fixed.

To see part 2 of Condition (A), notice that the Lipschitz property of $S(\cdot)$ implies

$$|S(\hat{u}) - S(u)| \leq Q\|D_{T_*}(\hat{u} - u)\|_2 \leq Q \sqrt{\sum_{t=T_0+1}^{T_0+T_*} (\hat{u}_t - u_t)^2}.$$

Hence, part 2 of Condition (A) follows since $|\hat{u}_t - u_t| \leq \delta_n$ for $T_0 + 1 \leq t \leq T$ with high probability. The proof is complete.

H.2 Proof of Theorem 2

We first state auxiliary results.

Lemma H.6. *Let $\{W_t\}_{t=1}^T$ be a stationary and β-mixing sequence with coefficient $\beta_{\text{mixing}}(\cdot)$. Let $G(x) = P(W_t \leq x)$. Then for any positive integer $1 \leq m \leq T/2$, we have*

$$E\left(\sup_{x \in \mathbb{R}} \left| T^{-1} \sum_{t=1}^{T} [\mathbf{1}\{W_t \leq x\} - G(x)] \right| \right) \leq 2\sqrt{T}\beta_{\text{mixing}}(m) + \sqrt{\pi m/(2T)} + (m-1)/T.$$

The next lemma was derived by Berbee (1987, Lemma 2.1); see Theorem 16.2.1 in Athreya and Lahiri (2006) and Lemma 7.1 of Chen et al. (2016) for popular versions. We state the result to make the exposition more self-contained. The proof is omitted.

Lemma H.7. *Let (W, R) be random vectors defined in the same probability space. Let $P_{W,R}$ denote the probability distribution of (W, R). Let P_R and P_W denote the probability distributions of R and W, respectively. Let $\|\cdot\|_{TV}$ denote the total variation metric. Define the β-mixing coefficient $\beta(W, R) = \|P_{W,R} - P_W \otimes P_R\|_{TV}/2$. Then the probability space can be extended to construct random element $\tilde{W}$ such that (1) $\tilde{W}$ and R are independent, (2) $\tilde{W}$ and W have the same distribution and (3) $P(\tilde{W} \neq W) \leq \beta(W, R)$.*

Now we prove Theorem 2. In this proof, universal constants refer to constants that depend only on $D_1, D_2, D_3 > 0$. Define $\tilde{F}(x) = R^{-1} \sum_{j=1}^R \tilde{F}_j(x)$, where

$$\tilde{F}_j(x) = m^{-1} \sum_{t \in H_j} \mathbf{1}\left\{ \phi\left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}) \right) \leq x \right\}.$$

We note that under $\Pi = \Pi_{\rightarrow}$, $\hat{F}(x)$ can be written as

$$\hat{F}(x) = T^{-1} \left(\sum_{t=1}^{T_0} \mathbf{1} \left\{ \phi \left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}) \right) \leq x \right\} + \sum_{t=T_0+1}^{T_0+T_*} \mathbf{1} \left\{ \phi \left(Z_{q(t)}, ..., Z_{q(t+T_*-1)}; \hat{\beta}(\mathbf{Z}) \right) \leq x \right\} \right), \tag{H.8}$$

where $q(t) = t\mathbf{1}\{t \leq T\} + (t-T)\mathbf{1}\{t > T\}$.

The rest of the proof proceeds in 4 steps. The first three steps bound $\sup_{x\in\mathbb{R}} |\hat{F}(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)|$, where we recall that $\Psi(x;\beta) = P(\phi(Z_t, ..., Z_{t+T_*-1}; \beta) \leq x)$. The fourth step derives the desired result.

Step 1: bound $\sup_{x\in\mathbb{R}} \left| \tilde{F}_j(x) - \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right|$.

Let $A_j = \bigcup_{t\in H_j} \{t, ..., t+T_*-1\}$. Since $k > T_*$, we have that $A_j \subset \widetilde{H}_j$ and $\min_{t\in A_j,\ s\in \widetilde{H}_j^c} |t-s| \geq k - T_* + 1$. This means that $\{Z_t\}_{t\in\widetilde{H}_j^c}$ and $\{Z_t\}_{t\in A_j}$ have a gap of at least $k-T_*+1$ time periods. By Lemma H.7 (applied with $W = \{Z_t\}_{t\in\widetilde{H}_j^c}$ and $R = \{Z_t\}_{t\in A_j}$), there exist random elements $\{\bar{Z}_t\}_{t\in A_j}$ (on an enlarged probability space) such that (1) $\{\bar{Z}_t\}_{t\in A_j}$ is independent of $\{Z_t\}_{t\in\widetilde{H}_j^c}$, (2) $\{\bar{Z}_t\}_{t\in A_j} \overset{d}{=} \{Z_t\}_{t\in A_j}$ and (3) $P(\{\bar{Z}_t\}_{t\in A_j} \neq \{Z_t\}_{t\in A_j}) \leq \beta_{\text{mixing}}(k - T_* + 1)$. Since $\{\tilde{Z}_t\}_{t\in\widetilde{H}_j}$ is independent of the data, we can construct $\{\bar{Z}_t\}_{t\in A_j}$ such that it is also independent of $\mathbf{Z}_{\widetilde{H}_j}$.

Define the event

$$\mathcal{M}_j = \left\{ \{\bar{Z}_t\}_{t\in A_j} = \{Z_t\}_{t\in A_j} \right\} \bigcap \left\{ \sup_{x\in\mathbb{R}} \left| \partial\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)/\partial x \right| \leq \xi_T \right\}$$
$$\bigcap \left\{ \max_{\pi\in\Pi} \left| S\left(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z})\right) - S\left(\mathbf{Z}^\pi, \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \right| \leq \varrho_T(|\widetilde{H}_j|) \right\}$$

as well as the functions

$$\begin{cases} \check{F}_j(x) = m^{-1} \sum_{t\in H_j} \mathbf{1} \left\{ \phi \left(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \leq x \right\} \\ \dot{F}_j(x) = m^{-1} \sum_{t\in H_j} \mathbf{1} \left\{ \phi \left(Z_t, ..., Z_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right) \leq x \right\} \end{cases}.$$

By the construction of $\{\bar{Z}_t\}_{t\in A_j}$ and Assumptions 4 and 5, $P(\mathcal{M}_j^c) \leq \beta_{\text{mixing}}(k - T_* + 1) + \gamma_{1,T} + \gamma_{2,T}$.

Notice that conditional on $\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})$, $\phi\left(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)$ is a stationary β-mixing across $t \in H_j$ with mixing coefficient $\tilde{\beta}_{\text{mixing}}(i) \leq \beta_{\text{mixing}}(i - T_* + 1)$ for $i \geq T_*$. Moreover, since $(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1})$ is independent of $\mathbf{Z}_{\widetilde{H}_j}$, we have

$$\Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) = P\left(\phi\left(\bar{Z}_t, ..., \bar{Z}_{t+T_*-1}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \leq x \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_j}) \right).$$

Hence, by Lemma H.6, we have that for any $m_1 \leq m/2$,

$$E\left(\sup_{x\in\mathbb{R}}\left|\check{F}_j(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|\right) \leq 2m^{1/2}\beta_{\mathrm{mixing}}(m_1-T_*+1)+\sqrt{\pi m_1/(2m)}+(m_1-1)/m.$$

We shall choose m_1 later. Observe that on the event $\mathcal{M}_j$, $\check{F}_j(\cdot)=\dot{F}_j(\cdot)$. Therefore,

$$E(a_j) \leq 2m^{1/2}\beta_{\mathrm{mixing}}(m_1-T_*+1)+\sqrt{\pi m_1/(2m)}+(m_1-1)/m+2P(\mathcal{M}_j^c), \tag{H.9}$$

where $a_j=\sup_{x\in\mathbb{R}}\left|\dot{F}_j(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|$.

Now we bound $\sup_{x\in\mathbb{R}}|\dot{F}_j(x)-\tilde{F}_j(x)|$. Fix an arbitrary $x\in\mathbb{R}$. Observe that on the event $\mathcal{M}_j$,

$$\begin{aligned}
&\left|\tilde{F}_j(x)-\dot{F}_j(x)\right| \\
&=\left|m^{-1}\sum_{t\in H_j}\left(\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z})\right)\leq x\right\}-\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x\right\}\right)\right| \\
&\leq m^{-1}\sum_{t\in H_j}\left|\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z})\right)\leq x\right\}-\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x\right\}\right| \\
&\overset{\text{(i)}}{\leq} m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\left|\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\right|\leq\varrho_T(|\widetilde{H}_j|)\right\} \\
&=m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x+\varrho_T(|\widetilde{H}_j|)\right\} \\
&\qquad -m^{-1}\sum_{t\in H_j}\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)< x-\varrho_T(|\widetilde{H}_j|)\right\} \\
&<\dot{F}_j\left(x+\varrho_T(|\widetilde{H}_j|)\right)-\dot{F}_j\left(x-2\varrho_T(|\widetilde{H}_j|)\right) \\
&\leq\Psi\left(x+\varrho_T(|\widetilde{H}_j|);\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x-2\varrho_T(|\widetilde{H}_j|);\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)+2a_j\overset{\text{(ii)}}{\leq}3\xi_T\varrho_T(|\widetilde{H}_j|)+2a_j,
\end{aligned}$$

where (i) follows by the elementary inequality $|\mathbf{1}\{x\leq z\}-\mathbf{1}\{y\leq z\}|\leq\mathbf{1}\{|y-z|\leq|x-y|\}$ for any $x,y,z\in\mathbb{R}$ and (ii) follows by the definition of $\mathcal{M}_j$. Since the above bound holds for any $x\in\mathbb{R}$ and $|\widetilde{H}_j|\leq m+2k$, we have that on the event $\mathcal{M}_j$,

$$\sup_{x\in\mathbb{R}}\left|\tilde{F}_j(x)-\dot{F}_j(x)\right|\leq 3\xi_T\varrho_T(m+2k)+2a_j.$$

By the definition of a_j, this means that on the event $\mathcal{M}_j$,

$$\sup_{x\in\mathbb{R}}\left|\tilde{F}_j(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|\leq 3\xi_T\varrho_T(m+2k)+3a_j.$$

By (H.9) and the fact that $\tilde{F}_j(\cdot)$ and $\Psi(\cdot,\cdot)$ take values in $[0,1]$, we have that for a universal constant $C_1 > 0$,

$$
\begin{aligned}
& E\left(\sup_{x\in\mathbb{R}}\left|\tilde{F}_j(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|\right) \qquad \text{(H.10)}\\
& \leq 3\xi_T\varrho_T(m+2k)+3E(a_j)+2P(\mathcal{M}_j^c)\\
& \leq 3\xi_T\varrho_T(m+2k)+6m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+3\sqrt{\pi m_1/(2m)}+3(m_1-1)/m+8P(\mathcal{M}_j^c)\\
& \overset{\text{(i)}}{\leq} C_1\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right),
\end{aligned}
$$

where (i) follows by $P(\mathcal{M}_j^c)\leq\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}$ and $m_1/m\leq\sqrt{m_1/m}$.

Step 2: bound $R^{-1}\sum_{j=1}^R\sup_{x\in\mathbb{R}}\left|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|$.

Let $\dot{\mathbf{Z}}=\{\dot{Z}_t\}_{t=1}^T$ satisfy that $\dot{\mathbf{Z}}\overset{d}{=}\mathbf{Z}$ and $\dot{\mathbf{Z}}$ is independent of $(\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T)$. Therefore, for any $1\leq j\leq R$,

$$
\begin{aligned}
\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) &= P\left(\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x\mid\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\\
&\overset{\text{(i)}}{=} P\left(\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x\mid\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right),
\end{aligned}
$$

where (i) follows by the fact that $\dot{\mathbf{Z}}$ and $(\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T)$ are independent. (This is the identity that $E(f(X;g(Y))\mid g(Y))=E(f(X;g(Y))\mid Y)$ for any measurable functions f and g if X and Y are independent. To see this, simply notice that the distribution of X given $g(Y)$ and the distribution of X given Y are both equal to the unconditional distribution of X.)

Define the event

$$
\mathcal{Q}_j=\left\{\sup_{x\in\mathbb{R}}\left|\partial\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)/\partial x\right|\leq\xi_T\right\}.
$$

Clearly, $P(\mathcal{Q}_j)\geq 1-\gamma_{2,T}$ by Assumption 5. Therefore, we have that on the event $\mathcal{Q}_j$,

$$
\begin{aligned}
&\left|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\\
&=\left|E\left(\mathbf{1}\left\{\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x\right\}-\mathbf{1}\left\{\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\leq x\right\}\mid\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right)\right|\\
&\leq E\left(\left|\mathbf{1}\left\{\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\leq x\right\}-\mathbf{1}\left\{\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\leq x\right\}\right|\mid\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\right)\\
&\overset{\text{(i)}}{\leq} E\Bigg[\mathbf{1}\Bigg\{\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\right|\\
&\qquad\qquad\leq\left|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\right|\Bigg\}\mid\mathbf{Z},\{\tilde{Z}_t\}_{t=1}^T\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= P\Bigg[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\Big| \\
&\qquad\qquad \leq \Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\Big| \mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^{T}\Bigg] \\
&\leq P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-x\Big| \leq 2\varrho_T(|\widetilde{H}_j|)\mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^{T}\right] \\
&\quad + P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\Big| > 2\varrho_T(|\widetilde{H}_j|)\mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^{T}\right] \\
&= \Psi\left(x+2\varrho_T(|\widetilde{H}_j|);\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x-2\varrho_T(|\widetilde{H}_j|);\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right) \\
&\quad + P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\Big| > 2\varrho_T(|\widetilde{H}_j|)\mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^{T}\right] \\
&\overset{\text{(ii)}}{\leq} 4\xi_T\varrho_T(m+2k) \\
&\quad + P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\Big| > 2\varrho_T(|\widetilde{H}_j|)\mid \mathbf{Z},\{\tilde{Z}_t\}_{t=1}^{T}\right],
\end{aligned}
$$

where (i) follows by the elementary inequality $|\mathbf{1}\{x\leq z\}-\mathbf{1}\{y\leq z\}|\leq \mathbf{1}\{|y-z|\leq |x-y|\}$ for any $x,y,z\in\mathbb{R}$ and (ii) follows by $|\widetilde{H}_j|\leq m+2k$ and the definition of $\mathcal{Q}_j$. Since the above bound does not depend on x, it holds uniformly in $x\in\mathbb{R}$ on the event $\mathcal{Q}_j$. Since $\Psi(\cdot,\cdot)$ is also bounded by one, we have that

$$
\begin{aligned}
&E\left(\sup_{x\in\mathbb{R}}\Big|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\Big|\right) \\
&\leq 4\xi_T\varrho_T(m+2k) \\
&\qquad + 2P(\mathcal{Q}_j^c)+P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)\Big| > 2\varrho_T(|\widetilde{H}_j|)\right] \\
&\leq 4\xi_T\varrho_T(m+2k)+2P(\mathcal{Q}_j^c) \\
&\qquad + P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z})\right)\Big| > \varrho_T(|\widetilde{H}_j|)\right] \\
&\qquad + P\left[\Big|\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\phi\left(\dot{Z}_{T_0+1},...,\dot{Z}_{T_0+T_*};\hat{\beta}(\mathbf{Z})\right)\Big| > \varrho_T(|\widetilde{H}_j|)\right] \\
&\overset{\text{(i)}}{\leq} 4\xi_T\varrho_T(m+2k)+2P(\mathcal{Q}_j^c)+2\gamma_{1,T} \overset{\text{(ii)}}{\leq} 4\xi_T\varrho_T(m+2k)+2\gamma_{1,T}+2\gamma_{2,T},
\end{aligned}
$$

where (i) follows by Assumption 4 and the fact that $|\widetilde{H}_j|=|\widetilde{H}_R|$ and (ii) follows by $P(\mathcal{Q}_j)\geq 1-\gamma_{2,T}$. Since the above bound holds for all $1\leq j\leq R$, we have

$$
E\left(R^{-1}\sum_{j=1}^{R}\sup_{x\in\mathbb{R}}\Big|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_j})\right)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\Big|\right)\leq 4\xi_T\varrho_T(m+2k)+2\gamma_{1,T}+2\gamma_{2,T}. \quad \text{(H.11)}
$$

Step 3: bound $\sup_{x\in\mathbb{R}}\Big|\hat{F}(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\Big|$.

By (H.8), we notice that

$$\sup_{x\in\mathbb{R}}\left|T\hat{F}(x)-m\sum_{j=1}^{R}\tilde{F}_j(x)\right|=\sup_{x\in\mathbb{R}}\Bigg|\sum_{t=T_0-mR+1}^{T_0}\mathbf{1}\left\{\phi\left(Z_t,...,Z_{t+T_*-1};\hat{\beta}(\mathbf{Z})\right)\leq x\right\}$$
$$+\sum_{t=T_0+1}^{T_0+T_*}\mathbf{1}\left\{\phi\left(Z_{q(t)},...,Z_{q(t+T_*-1)};\hat{\beta}(\mathbf{Z})\right)\leq x\right\}\Bigg|\leq T_*+(T_0-mR)\leq T_*+R-1.$$

Moreover, by (H.10) and (H.11), we have that

$$E\left(\sup_{x\in\mathbb{R}}\left|R^{-1}\sum_{j=1}^{R}\tilde{F}_j(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right)\right|$$
$$\leq C_2\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\mathrm{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\mathrm{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right)$$

for some universal constant $C_2>0$.

The above two displays imply that

$$\sup_{x\in\mathbb{R}}\left|\frac{T}{mR}\hat{F}(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\leq\frac{T_*+R-1}{mR}$$
$$+C_2\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\mathrm{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\mathrm{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right).$$

Since $\hat{F}(x)\in[0,1]$, we have that

$$\sup_{x\in\mathbb{R}}|(1-T/(mR))\hat{F}(x)|\leq\frac{T}{mR}-1\leq\frac{T-mR}{mR}\leq\frac{T_*+R-1}{mR}.$$

Since $mR\geq T_0/2$ (due to $R<T_0/2$), we have $(T_*+R-1)/(mR)\leq 2T_*T_0^{-1}+m^{-1}\lesssim\sqrt{m_1/m}$. Hence, the above two displays imply that for some universal constant $C_3>0$,

$$E\left(\sup_{x\in\mathbb{R}}\left|\hat{F}(x)-\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\right)$$
$$\leq C_3\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\mathrm{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\mathrm{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right).\tag{H.12}$$

Step 4: derive the desired result.

Let A_R be defined as in Step 1 with $j=R$. Following Step 1, we can construct random elements $\{\bar{Z}_t\}_{t\in A_R}$ (on an enlarged probability space) such that (1) $\{\bar{Z}_t\}_{t\in A_R}$ is independent of $\mathbf{Z}_{\widetilde{H}_R}$, (2) $\{\bar{Z}_t\}_{t\in A_R}\overset{d}{=}\{Z_t\}_{t\in A_R}$ and (3) $P(\{\bar{Z}_t\}_{t\in A_R}\neq\{Z_t\}_{t\in A_R})\leq\beta_{\mathrm{mixing}}(k-T_*+1)$.

Define $\bar{G}(\mathbf{Z}_{\widetilde{H}_R}) = \phi\left(\bar{Z}_{T_0+1}, ..., \bar{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)$. Since $\{T_0+1, ..., T_0+T_*\} \subset A_R$, we have that $(\bar{Z}_{T_0+1}, ..., \bar{Z}_{T_0+T_*})$ is independent of $\mathbf{Z}_{\widetilde{H}_R}$, which means that

$$P\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}) \leq x \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) = \Psi\left(x; \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \qquad \forall x \in \mathbb{R}.$$

Therefore,

$$\text{conditional on } \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R}),\ \Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \text{ has uniform distribution on } (0,1). \tag{H.13}$$

We also introduce the following notations to simplify the argument:
$\bar{G}(\mathbf{Z}) = \phi\left(\bar{Z}_{T_0+1}, ..., \bar{Z}_{T_0+T_*}; \hat{\beta}(\mathbf{Z})\right)$ and $G(\mathbf{Z}) = \phi\left(Z_{T_0+1}, ..., Z_{T_0+T_*}; \hat{\beta}(\mathbf{Z})\right)$.

For arbitrary $\alpha \in (0,1)$ and $c > 0$, we observe that

$$\begin{aligned}
&\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \\
&\overset{\text{(i)}}{=} \left|E\left(\mathbf{1}\left\{\hat{F}(G(\mathbf{Z})) < \alpha\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - E\left(\mathbf{1}\left\{\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) < \alpha\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right| \\
&\leq E\left(\left|\mathbf{1}\left\{\hat{F}(G(\mathbf{Z})) < \alpha\right\} - \mathbf{1}\left\{\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) < \alpha\right\}\right| \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\overset{\text{(ii)}}{\leq} E\left(\mathbf{1}\left\{\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq \left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right|\right\} \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&= P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq \left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\leq P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \leq c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\qquad + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) \\
&\overset{\text{(iii)}}{\leq} 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)
\end{aligned}$$

where (i) follows by (H.13), (ii) follows by the elementary inequality $|\mathbf{1}\{x < z\} - \mathbf{1}\{y < z\}| \leq \mathbf{1}\{|y - z| \leq |x - y|\}$ for any $x, y, z \in \mathbb{R}$ and (iii) follows by (H.13). Now we take expectation on both sides, obtaining

$$\begin{aligned}
&E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right) < \alpha \mid \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \alpha\right| \\
&\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(G(\mathbf{Z}))\right| > c\right) \\
&\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| > c\right) + P\left(\{\bar{Z}_t\}_{t \in A_R} \neq \{Z_t\}_{t \in A_R}\right) \\
&\leq 2c + P\left(\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| > c\right) + \beta_{\text{mixing}}(k - T_* + 1) \\
&\leq 2c + c^{-1} E\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R}); \hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right) - \hat{F}(\bar{G}(\mathbf{Z}))\right| + \beta_{\text{mixing}}(k - T_* + 1)
\end{aligned} \tag{H.14}$$

Define the event

$$\mathcal{A}=\left\{\sup_{x\in\mathbb{R}}\left|\partial\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)/\partial x\right|\leq\xi_T\right\}\bigcap\left\{\left|\bar{G}(\mathbf{Z})-\bar{G}(\mathbf{Z}_{\widetilde{H}_R})\right|\leq\varrho_T(|\widetilde{H}_R|)\right\}.$$

By Assumptions 4 and 5, $P(\mathcal{A}^c)\leq\gamma_{1,T}+\gamma_{2,T}$. Therefore,

$$\begin{aligned}
&E\left|\Psi\left(\bar{G}(\mathbf{Z});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\\
&=E\left(\left|\Psi\left(\bar{G}(\mathbf{Z});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\times\mathbf{1}_{\mathcal{A}}\right)\\
&\qquad\qquad+E\left(\left|\Psi\left(\bar{G}(\mathbf{Z});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\times\mathbf{1}_{\mathcal{A}^c}\right)\\
&\overset{\text{(i)}}{\leq}E\left(\left|\Psi\left(\bar{G}(\mathbf{Z});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\times\mathbf{1}_{\mathcal{A}}\right)+2P\left(\mathcal{A}^c\right)\\
&\leq\xi_T\varrho_T(|\widetilde{H}_R|)+2P\left(\mathcal{A}^c\right)\leq\xi_T\varrho_T(m+2k)+2\gamma_{1,T}+2\gamma_{2,T},
\end{aligned}$$

where (i) follows by the fact that $\Psi(\cdot,\cdot)\in[0,1]$. Hence, we have that

$$\begin{aligned}
&E\left|\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\hat{F}(\bar{G}(\mathbf{Z}))\right|\\
&\leq E\left|\Psi\left(\bar{G}(\mathbf{Z});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\hat{F}(\bar{G}(\mathbf{Z}))\right|+E\left|\Psi\left(\bar{G}(\mathbf{Z});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\Psi\left(\bar{G}(\mathbf{Z}_{\widetilde{H}_R});\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)\right|\\
&\leq E\sup_{x\in\mathbb{R}}\left|\Psi\left(x;\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\hat{F}(x)\right|+\xi_T\varrho_T(m+2k)+2\gamma_{1,T}+2\gamma_{2,T}\\
&\overset{\text{(i)}}{\leq}C_4\left(\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}\right)
\end{aligned}$$

for a universal constant $C_4>0$, where (i) follows by (H.12).

Now we combine (H.14) and the above display. We also choose

$$c\asymp\sqrt{\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}}.$$

Then we can find a universal constant $C_5>0$ such that

$$\begin{aligned}
&E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right)<\alpha\mid\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\alpha\right|\\
&\leq C_5\sqrt{\xi_T\varrho_T(m+2k)+m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)+\sqrt{m_1/m}+\beta_{\text{mixing}}(k-T_*+1)+\gamma_{1,T}+\gamma_{2,T}}.
\end{aligned}$$

Now we choose m_1 satisfying $m_1\asymp(\log m)^{1/D_3}$ and $m^{1/2}\beta_{\text{mixing}}(m_1-T_*+1)\lesssim m^{-1}$. Hence, for some universal constant $C_6>0$,

$$E\left|P\left(\hat{F}\left(G(\mathbf{Z})\right)<\alpha\mid\hat{\beta}(\mathbf{Z}_{\widetilde{H}_R})\right)-\alpha\right|$$

$$\leq C_6\sqrt{\xi_T\varrho_T(m+2k)}+C_6\left(m^{-1}(\log m)^{1/D_3}\right)^{1/4}+C_6\sqrt{\beta_{\text{mixing}}(k-T_*+1)}+C_6\sqrt{\gamma_{1,T}}+C_6\sqrt{\gamma_{2,T}}.$$

Since $m \asymp T_0/R$, the desired result follows once we notice that $\hat{p} \geq 1-\alpha$ and $\hat{F}\left(G(\mathbf{Z})\right) < \alpha$ are the same event.

H.2.1 Proof of Lemma H.6

Define $K = \lfloor T/m \rfloor$ and $\hat{F}(x) = m^{-1/2}\sum_{r=1}^{m}\hat{F}_r(x)$, where $\hat{F}_r(x) = K^{-1/2}\sum_{j=1}^{K}[\mathbf{1}\{W_{(j-1)m+r} \leq x\} - G(x)]$ for $1 \leq r \leq m$. Let $\Delta(x) = \sum_{t=mK+1}^{T}[\mathbf{1}\{W_t \leq x\} - G(x)]$. Let $L_T(x) = T^{-1/2}\sum_{t=1}^{T}[\mathbf{1}\{W_t \leq x\} - G(x)]$. Notice that

$$\sqrt{T}L_T(x) = \sqrt{mK}\hat{F}(x) + \Delta(x).$$

Since $|\mathbf{1}\{W_t \leq x\} - G(x)| \leq 1$, it follows that $\sup_{x\in\mathbb{R}}|\Delta(x)| \leq T - mK \leq m-1$ and thus

$$\sup_{x\in\mathbb{R}}\left|\sqrt{T}L_T(x) - \sqrt{mK}\hat{F}(x)\right| \leq m-1. \tag{H.15}$$

By Berbee's coupling (Lemma H.7), we can enlarge the probability space and define random variables $\{\bar{W}_t\}_{t=1}^{mK}$ such that (1) $\bar{W}_t \overset{d}{=} W_t$ for all $1 \leq t \leq mT$, (2) $\bar{W}_{(j-1)m+r}$ is independent across $1 \leq j \leq K$ for all r and (3) $P(\bigcup_{t=1}^{mK}\{\bar{W}_t \neq W_t\}) \leq mK\beta_{\text{mixing}}(m) \leq T\beta_{\text{mixing}}(m)$.

We now define $\bar{F}(x) = m^{-1/2}\sum_{r=1}^{m}\bar{F}_r(x)$, where $\bar{F}_r(x) = K^{-1/2}\sum_{j=1}^{K}[\mathbf{1}\{\bar{W}_{(j-1)m+r} \leq x\} - G(x)]$.

Since $\{\bar{W}_{(j-1)m+r}\}_{j=1}^{K}$ is independent, it follows by Dvoretzky-Kiefer-Wolfowitz inequality that for any $z > 0$,

$$P\left(\sup_{x\in\mathbb{R}}|\bar{F}_r(x)| > z\right) \leq 2\exp(-2z^2).$$

Therefore, we have that

$$E\left(\sup_{x\in\mathbb{R}}|\bar{F}_r(x)|\right) = \int_0^\infty P\left(\sup_{x\in\mathbb{R}}|\bar{F}_r(x)| > z\right)dz \leq 2\int_0^\infty \exp(-2z^2)dz = \sqrt{\pi/2}.$$

Hence, we have that

$$E\left(\sup_{x\in\mathbb{R}}|\bar{F}(x)|\right) \leq m^{-1/2}\sum_{r=1}^{m}E\left(\sup_{x\in\mathbb{R}}|\bar{F}_r(x)|\right) \leq \sqrt{\pi m/2}.$$

Since $\bar{F}(\cdot) = \hat{F}(\cdot)$ with probability at least $1 - T\beta_{\text{mixing}}(m)$, we have that

$$E\left(\sup_{x\in\mathbb{R}} |\bar{F}(x) - \hat{F}(x)|\right) \leq 2T\beta_{\text{mixing}}(m).$$

Therefore, $E\left(\sup_{x\in\mathbb{R}} |\hat{F}(x)|\right) \leq 2T\beta_{\text{mixing}}(m) + \sqrt{\pi m/2}$.

By (H.15) and $mK/T \leq 1$, we have that

$$E\left(\sup_{x\in\mathbb{R}} |L_T(x)|\right) \leq 2T\beta_{\text{mixing}}(m) + \sqrt{\pi m/2} + (m-1)/\sqrt{T}.$$

The proof is complete.

H.3 Proof of Theorem C.1

Recall $\mathbf{Z}^* = (Z_1^*, \ldots, Z_T^*)'$ with $Z_t^* = \left(Y_{1t}^N, Y_{2t}^N, \ldots, Y_{J+1t}^N, X_{1t}', \ldots, X_{J+1t}'\right)'$ for $1 \leq t \leq T_0 + T_*$. Let

$$\hat{p}_{\mathbf{Z}^*} = 1 - \hat{F}\left(S(\hat{u}(\mathbf{Z}^*))\right),$$

where $\hat{F}(x; \mathbf{Z}^*) = \frac{1}{|\Pi|}\sum_{\pi\in\Pi} \mathbf{1}\left\{S\left(\hat{u}_\pi(\mathbf{Z}^*)\right) < x\right\}$ and $\hat{u}(\mathbf{Z}^*) = Y^N - \hat{P}^N$ with $\hat{P}^N$ computed using $\mathbf{Z}^*$. By the proof of Theorem 1, we have

$$\left|P\left(\hat{p}_{\mathbf{Z}^*} \leq \alpha\right) - \alpha\right| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T),$$

where $\tilde{\delta}_T = (T_*/T_0)^{1/4}(\log T)$ and the constant C depends on T_*, M and D, but not on T. It follows that

$$\left|P\left(\hat{p}_{\mathbf{Z}^*} > \alpha\right) - (1-\alpha)\right| \leq C(\tilde{\delta}_T + \delta_T + \sqrt{\delta_T} + \gamma_T).$$

The desired result follows by observing that $\theta_t \in \mathcal{C}_{1-\alpha}(t)$ is the same event as $\hat{p}_{\mathbf{Z}^*} > \alpha$; this is simply because $\mathbf{Z}(\theta^0) = \mathbf{Z}^*$, where $\mathbf{Z}(\theta^0)$ is defined in Section 2.2.

H.4 Proof of Theorem D.1

Let $\{S^{(j)}(\hat{u})\}_{j=1}^n$ denoted the non-decreasing rearrangement of $\{S(\hat{u}_\pi) : \pi \in \Pi\}$, where $n = |\Pi|$, which we refer to as randomization quantiles. The p-value is

$$\hat{p} = \frac{1}{n}\sum_{\pi\in\Pi} \mathbf{1}(S(\hat{u}_\pi) \geq S(\hat{u})).$$

Note that

$$\mathbf{1}(\hat{p} \leq \alpha) = \mathbf{1}(S(\hat{u}) > S^{(k)}(\hat{u})),$$

where $k = k(\alpha) = n - \lfloor n\alpha \rfloor = \lceil n(1-\alpha) \rceil$.

The proof proceeds in three steps. First, we show that exchangeability of the data implies exchangeability of the residuals. Second, we show that exchangeability of the residuals implies that $P(\hat{p} \leq \alpha) \leq \alpha$. Third, we show that if there are no ties, $\alpha - 1/n \leq P(\hat{p} \leq \alpha)$. The proof follows from standard arguments (e.g., Hoeffding, 1952; Romano, 1990; Chernozhukov et al., 2018; Lei et al., 2018).

Step 1: By the iid or exchangeability property of data, we have that

$$\underbrace{\{g(Z_t, \hat{\beta}(\{Z_t\}_{t=1}^T))\}_{t=1}^T}_{\{\hat{u}_t\}_{t=1}^T} \overset{d}{=} \{g(Z_{\pi(t)}, \hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T)\}_{t=1}^T.$$

Since $\hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T)$ does not depend on π, we have

$$\{g(Z_{\pi(t)}, \hat{\beta}(\{Z_{\pi(t)}\}_{t=1}^T)\}_{t=1}^T = \underbrace{\{g(Z_{\pi(t)}, \hat{\beta}(\{Z_t\}_{t=1}^T)\}_{t=1}^T}_{\{\hat{u}_{\pi(t)}\}_{t=1}^T}.$$

Therefore, $\{\hat{u}_{\pi(t)}\}_{t=1}^T \overset{d}{=} \{\hat{u}_t\}_{t=1}^T$.

Step 2: Note that Π_{all} and $\Pi_{\rightarrow}$ form groups in the sense that $\Pi\pi = \Pi$ for all $\pi \in \Pi$. Therefore, the randomization quantiles are invariant,

$$S^{(k(\alpha))}(\hat{u}_\pi) = S^{(k(\alpha))}(\hat{u}), \text{ for all } \pi \in \Pi.$$

Therefore,

$$\sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u}_\pi)) = \sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u})) \leq \alpha n.$$

Since $\mathbf{1}(S(\hat{u}) > S^{(k(\alpha))}(\hat{u}))$ is equal in distribution to $\mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u}_\pi))$ for any $\pi \in \Pi$ by exchangeability (Step 1), we have that

$$\alpha \geq E \sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) > S^{(k(\alpha))}(\hat{u}_\pi))/n = E\mathbf{1}(S(\hat{u}) > S^{(k(\alpha))}(\hat{u})) = E\mathbf{1}(\hat{p} \leq \alpha).$$

Step 3: By continuity of the distribution of $\{S(\hat{u}_\pi)\}_{\pi \in \Pi}$, there are no ties with probability one. Therefore,

$$\sum_{\pi \in \Pi} \mathbf{1}(S(\hat{u}_\pi) \leq S^{(k(\alpha))}(\hat{u})) = k(\alpha) \leq n(1-\alpha) + 1$$

Because

$$\sum_{\pi\in\Pi}\mathbf{1}(S(\hat{u}_\pi)\le S^{(k(\alpha))}(\hat{u}))+\sum_{\pi\in\Pi}\mathbf{1}(S(\hat{u}_\pi)> S^{(k(\alpha))}(\hat{u}))=n,$$

we have

$$\sum_{\pi\in\Pi}\mathbf{1}(S(\hat{u}_\pi)> S^{(k(\alpha))}(\hat{u}))\ge n\alpha-1.$$

The result now follows by similar arguments as in Step 2.

H.5 Proof of Lemma 1

Let X_{jt} denote the (j,t) entry of the matrix $X\in\mathbb{R}^{T\times J}$. We assume the following conditions hold: (1) $E(u_tX_{jt})=0$ for $1\le j\le J$. (2) there exist constants $c_1,c_2>0$ such that $E|X_{jt}u_t|^2\ge c_1$ and $E|X_{jt}u_t|^3\le c_2$ for any $1\le j\le J$ and $1\le t\le T$; (3) for each $1\le j\le J$, the sequence $\{X_{jt}u_t\}_{t=1}^T$ is β-mixing and the β-mixing coefficient satisfies that $\beta(t)\le a_1\exp(-a_2t^\tau)$, where $a_1,a_2,\tau>0$ are constants. (4) there exists a constant $c_3>0$ such that $\max_{1\le j\le J}\sum_{t=1}^T X_{jt}^2u_t^2\le c_3^2T$ with probability $1-o(1)$. (5) $\log J=o(T^{4\tau/(3\tau+4)})$ and $w\in\mathcal{W}$. (6) There exists a sequence $\ell_T>0$ such that $(X_t'\delta)^2\le \ell_T\|X\delta\|_2^2/T,$ for all $w+\delta\in\mathcal{W}$ with probability $1-o(1)$ for $T_0+1\le t\le T$ and (7) $\ell_TB_T\to 0$ for $B_T=M[\log(T\vee J)]^{(1+\tau)/(2\tau)}T^{-1/2}$.

Then we claim that under conditions (1)-(5) listed above:

(1) There exist a constant $M>0$ depending only on K and the constants listed above such that with probability $1-o(1)$

$$\|X(\hat{w}-w)\|_2^2/T\le B_T=M[\log(T\vee J)]^{(1+\tau)/(2\tau)}T^{-1/2}$$

(2) Moreover, if (6) and (7) also hold, then

$$\frac{1}{T}\sum_{t=1}^T\left(\hat{P}_t^N-P_t^N\right)^2=o_P(1)\text{ and }\hat{P}_t^N-P_t^N=o_P(1),\text{ for any }T_0+1\le t\le T.$$

The following result is useful in deriving the properties of the ℓ_1-constrained estimator.

Lemma H.8. *Suppose that (1) $E(u_tX_{jt})=0$ for $1\le j\le J$. (2) $\max_{1\le j\le J,1\le t\le T}E|X_{jt}u_t|^3\le K_1$ for a constant $K_1>0$. (3) $\min_{1\le j\le J,1\le t\le T}E|X_{jt}u_t|^2\ge K_2$ for a constant $K_2>0$. (4) For each $1\le j\le J$, $\{X_{jt}u_t\}_{t=1}^T$ is β-mixing and the β-mixing coefficients satisfy $\beta(s)\le$*

$D_1 \exp(-D_2 s^\tau)$ *for some constants* $D_1, D_2, \tau > 0$. *Assume* $\log J = o(T^{4\tau/(3\tau+4)})$. *Then there exists a constant* $\kappa > 0$ *depending only on* K_1, K_2, D_1, D_2, τ *such that with probability* $1 - o(1)$

$$\max_{1\leq j\leq J} \left| \sum_{t=1}^{T} X_{jt} u_t \right| < \kappa [\log(T \vee J)]^{(1+\tau)/(2\tau)} \max_{1\leq j\leq J} \sqrt{\sum_{t=1}^{T} X_{jt}^2 u_t^2}$$

Proof. Define $W_{j,t} = X_{jt} u_t$. Let $m = \left\lfloor [4D_2^{-1} \log(JT)]^{1/\tau} \right\rfloor$ and $k = \lfloor T/m \rfloor$. For simplicity, we assume for now that T/m is an integer. Define

$$H_i = \{i, m+i, 2m+i, ..., (k-1)m+i\} \qquad \forall 1 \leq i \leq m.$$

By Berbee's coupling (Lemma H.7), there exist a sequence of random variables $\{\tilde{W}_{j,t}\}_{t\in H_i}$ such that (1) $\{\tilde{W}_{j,t}\}_{t\in H_i}$ is independent across t, (2) $\tilde{W}_{j,t}$ has the same distribution as $W_{j,t}$ for $t \in H_i$ and (3) $P\left(\bigcup_{t\in H_i} \{\tilde{W}_{j,t} \neq W_{j,t}\}\right) \leq k\beta(m)$.

By assumption, $\max_{j,t} E|X_{jt}u_t|^3$ is uniformly bounded above and $\min_{j,t} E|X_{jt}u_t|^2$ is uniformly bounded away from zero. It follows, by Theorem 7.4 of Peña et al. (2008), that there exist constants $C_0, C_1 > 0$ depending on K_1 and K_2 such that for any $0 \leq x \leq C_0 k^{1/6}$,

$$P\left(\left| \frac{\sum_{t\in H_i} \tilde{W}_{j,t}}{\sqrt{\sum_{t\in H_i} \tilde{W}_{j,t}^2}} \right| > x \right) \leq C_1 \left(1 - \Phi(x)\right),$$

where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Therefore, for any $0 \leq x \leq C_0 k^{1/6}$,

$$\begin{aligned} P\left(\left| \frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}} \right| > x \right) &\leq P\left(\left| \frac{\sum_{t\in H_i} \tilde{W}_{j,t}}{\sqrt{\sum_{t\in H_i} \tilde{W}_{j,t}^2}} \right| > x \right) + P\left(\bigcup_{t\in H_i} \{\tilde{W}_{j,t} \neq W_{j,t}\} \right) \\ &\leq C_1 \left(1 - \Phi(x)\right) + k\beta(m). \end{aligned} \tag{H.16}$$

The Cauchy-Schwarz inequality implies

$$\begin{aligned} \left| \sum_{t=1}^{T} W_{j,t} \right| \leq \sum_{i=1}^{m} \left| \frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}} \right| \sqrt{\sum_{t\in H_i} W_{j,t}^2} &\leq \sqrt{\sum_{i=1}^{m} \left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}} \right)^2} \times \sqrt{\sum_{i=1}^{m} \sum_{t\in H_i} W_{j,t}^2} \\ &= \sqrt{\sum_{i=1}^{m} \left(\frac{\sum_{t\in H_i} W_{j,t}}{\sqrt{\sum_{t\in H_i} W_{j,t}^2}} \right)^2} \times \sqrt{\sum_{t=1}^{T} W_{j,t}^2}. \end{aligned}$$

Hence,

$$\left|\frac{\sum_{t=1}^{T}W_{j,t}}{\sqrt{\sum_{t=1}^{T}W_{j,t}^2}}\right| \leq \sqrt{\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right)^2}.$$

It follows that for any $0 \leq x \leq C_0 k^{1/6}\sqrt{m}$,

$$\begin{aligned}
&P\left(\left|\frac{\sum_{t=1}^{T}W_{j,t}}{\sqrt{\sum_{t=1}^{T}W_{j,t}^2}}\right| > x\right) \leq P\left(\sqrt{\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right)^2} > x\right) \\
&= P\left(\sum_{i=1}^{m}\left(\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right)^2 > x^2\right) \leq \sum_{i=1}^{m} P\left(\left|\frac{\sum_{t\in H_i}W_{j,t}}{\sqrt{\sum_{t\in H_i}W_{j,t}^2}}\right| > \frac{x}{\sqrt{m}}\right) \\
&\overset{\text{(i)}}{\leq} m\left[C_1\left(1-\Phi(x/\sqrt{m})\right)+k\beta(m)\right] \overset{\text{(ii)}}{\leq} C_1 m\sqrt{\frac{m}{2\pi}}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 km\exp\left(-D_2 m^{\tau}\right) \\
&< C_1 m^{3/2}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 T\exp\left(-D_2 m^{\tau}\right)
\end{aligned}$$

where (i) follows by (H.16) and (ii) follows by the inequality $1-\Phi(a) \leq a^{-1}\phi(a)$ (with ϕ being the pdf of $N(0,1)$) and $\beta(m) \leq D_1\exp(-D_2 m^{\tau})$.

By the union bound, it follows that for any $0 \leq x \leq C_0 k^{1/6}\sqrt{m}$,

$$P\left(\max_{1\leq j\leq J}\left|\frac{\sum_{t=1}^{T}W_{j,t}}{\sqrt{\sum_{t=1}^{T}W_{j,t}^2}}\right| > x\right) \leq C_1 J m^{3/2}x^{-1}\exp\left(-\frac{x^2}{2m}\right) + D_1 JT\exp\left(-D_2 m^{\tau}\right).$$

Now we choose $x = 2\sqrt{m\log(Jm^{3/2})}$. Since $\log J = o(T^{4\tau/(3\tau+4)})$ and $k \asymp T/m$, it can be very easily verified that $x \ll C_0 k^{1/6}\sqrt{m}$ and the two terms on the right-hand side of the above display tend to zero. The desired result follows.

If T/k is not an integer, then we simply add one observation from $\{W_{j,t}\}_{t=km+1}^{T}$ to each of H_i for $1 \leq i \leq m$. The bound in (H.16) holds with C_1 large enough. The proof is complete. $\square$

Now we are ready to prove Lemma 1.

Proof of Lemma 1. Let $\Delta = \hat{w} - w$. Since $\|w\|_1 \leq K$, we have $\|Y - X\hat{w}\|_2^2 \leq \|Y - Xw\|_2^2$. Notice that $Y - Xw = u$ and $Y - X\hat{w} = u - X\Delta$. Therefore, $\|u - X\Delta\|_2^2 \leq \|u\|_2^2$, which means $\|X\Delta\|_2^2 \leq 2u'X\Delta$. Now we observe that

$$\|X\Delta\|_2^2 \leq 2u'X\Delta \overset{\text{(i)}}{\leq} 2\|Xu\|_\infty\|\Delta\|_1 \overset{\text{(ii)}}{\leq} 4K\|Xu\|_\infty, \tag{H.17}$$

where (i) follows by Hölder's inequality and (ii) follows by $\|\Delta\|_1 \leq 2K$ (since $\|\hat{w}\|_1 \leq K$ and $\|w\|_1 \leq K$). By Lemma H.8, there exists a constant $\kappa > 0$ such that

$$P\left(\max_{1\leq j\leq J}\left|\sum_{t=1}^{T} X_{jt}u_t\right| > \kappa[\log(T\vee J)]^{(1+\tau)/(2\tau)} \max_{1\leq j\leq J}\sqrt{\sum_{t=1}^{T} X_{jt}^2 u_t^2}\right) = o(1).$$

Since $P\left(\max_{1\leq j\leq J}\sum_{t=1}^{T} X_{jt}^2 u_t^2 \leq c_3^2 T\right) \to 1$, it follows that

$$P\left(\max_{1\leq j\leq J}\left|\sum_{t=1}^{T} X_{jt}u_t\right| > \kappa c_3[\log(T\vee J)]^{(1+\tau)/(2\tau)}\sqrt{T}\right) = o(1). \tag{H.18}$$

Part (1) follows by combining (H.17) and (H.18). Part (2) follows by part (1) and $\ell_T B_T = o(1)$. □

H.6 Proof of Lemma 2

We borrow results and notations from Bai (2003). Following standard notation, we use i instead of j to denote units. Here are the regularity conditions from Bai (2003).

Suppose that there exists a constant $D_0 > 0$ the following conditions hold:
(1) $\max_{1\leq t\leq T} E\|F_t\|_2^4 \leq D_0$, $\max_{1\leq j\leq N}\|\lambda_j\|_2^4 \leq D_0$, $\max_{jt} E|u_{jt}|^8 \leq D_0$ and $E(u_{jt}) = 0$.
(2) $\max_s N^{-1}\sum_{t=1}^{T}|\sum_{i=1}^{N} E(u_{is}u_{it})| \leq D_0$ and $\max_i \sum_{j=1}^{N}\max_{1\leq t\leq T}|E(u_{it}u_{jt})| \leq D_0$.
(3) $(NT)^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(u_{it}u_{js})| \leq D_0$ and $\max_{s,t} E|N^{-1/2}\sum_{i=1}^{N}[u_{is}u_{it} - E(u_{is}u_{it})]|^4 \leq D_0$.
(4) $N^{-1}\sum_{i=1}^{N} E\|T^{-1/2}\sum_{t=1}^{T} F_t u_{it}\|_2^2 \leq D_0$.
(5) $\max_t E\|(NT)^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{N} F_s[u_{is}u_{it} - E(u_{is}u_{it})]\|_2^2 \leq D_0$.
(6) $E\|(NT)^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{N} F_t\lambda_i' u_{it}\|_2^2 \leq D_0$.

Moreover, we assume the following conditions: (7) for each t, $N^{-1/2}\sum_{i=1}^{N}\lambda_i u_{it} \to^d N(0,\Gamma_t)$ as $N\to\infty$, where $\Gamma_t = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j' E(u_{it}u_{jt})$. (8) for each i, $T^{-1/2}\sum_{t=1}^{T} F_t u_{it} \to^d N(0,\Phi_i)$ as $T\to\infty$, where $\Phi_i = \lim_{T\to\infty} T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T} E(F_t F_s' u_{is}u_{it})$. (9) $N^{-1}\sum_{i=1}^{N}\lambda_i\lambda_i' \to \Sigma_\Lambda$ and $T^{-1}\sum_{t=1}^{T} F_t F_t' = \Sigma_F + o_P(1)$ for some $k\times k$ positive definite matrices Σ_Λ and Σ_F satisfying that $\Sigma_\Lambda\Sigma_F$ has distinct eigenvalues.

What follows below is the proof of the lemma. We recall some notations used by Bai (2003). Let $F = (F_1,\ldots,F_T)' \in \mathbb{R}^{T\times k}$ and $\Lambda = (\lambda_1,\ldots,\lambda_N)' \in \mathbb{R}^{N\times k}$. Define $H = (\Lambda'\Lambda/N)(F'\tilde{F}/T)V_{NT}^{-1}$, where $V_{NT} \in \mathbb{R}^{k\times k}$ is the diagonal matrix with the largest k eigenvalues of $Y^N(Y^N)'/(NT)$ on the diagonal and $\tilde{F}$ is the normalized F, namely $\tilde{F}'\tilde{F}/T = I_k$.

We start with the first equation in the proof of Theorem 3 in Bai (2003) (on page 166):

$$\hat{\lambda}_1'\hat{F}_t - \lambda_1'F_t = \left(\hat{F}_t - H'F_t\right)' H^{-1}\lambda_1 + \hat{F}_t'(\hat{\lambda}_1 - H^{-1}\lambda_1). \tag{H.19}$$

The rest of the proof proceeds in two steps. We first recall some results from Bai (2003) and then derive the desired result.

Step 1: recall useful results from Bai (2003). By Lemma A.1 of Bai (2003),

$$\sum_{t=1}^{T} \|\hat{F}_t - H'F_t\|_2^2 = O_P(T/\delta_{NT}^2), \tag{H.20}$$

where $\delta_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. By definition, $\hat{F}'\hat{F}/T = I_k$, which means

$$\sum_{t=1}^{T} \|\hat{F}_t\|_2^2 = \sum_{t=1}^{T} \text{trace}(\hat{F}_t\hat{F}_t') = \text{trace}(\hat{F}'\hat{F}) = kT. \tag{H.21}$$

By Theorem 2 of Bai (2003),

$$\hat{\lambda}_1 = H^{-1}\lambda_1 + O_P(\max\{T^{-1/2}, N^{-1}\}). \tag{H.22}$$

By the proof of part (i) in Theorem 2 of Bai (2003), H converges in probability to a nonsingular matrix; see page 166 of Bai (2003). Hence, $\|H^{-1}\| = O_P(1)$. By assumption, $\|\lambda_1\|_2 = O(1)$. Hence,

$$\|H^{-1}\lambda_1\|_2 = O_P(1). \tag{H.23}$$

Step 2: prove the desired result.

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \left(\hat{\lambda}_1'\hat{F}_t - \lambda_1'F_t\right)^2 &\overset{\text{(i)}}{\leq} 2\sum_{t=1}^{T} \left[\left(\hat{F}_t - H'F_t\right)' H^{-1}\lambda_1\right]^2 + 2\sum_{t=1}^{T} \left[\hat{F}_t'(\hat{\lambda}_1 - H^{-1}\lambda_1)\right]^2 \\
&\leq 2\sum_{t=1}^{T} \|\hat{F}_t - H'F_t\|_2^2 \times \|H^{-1}\lambda_1\|_2^2 + 2\sum_{t=1}^{T} \|\hat{F}_t\|_2^2 \times \|\hat{\lambda}_1 - H^{-1}\lambda_1\|_2^2 \\
&\overset{\text{(ii)}}{=} O_P(T/\delta_{NT}^2) \times O_P(1) + 2kT \times O_P(\max\{T^{-1}, N^{-2}\}) \\
&= O_P(T/\delta_{NT}^2),
\end{aligned}$$

where (i) follows by (H.19) and the elementary inequality of $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ and (ii) follows by (H.20), (H.21), (H.22) and (H.23). Since $n = |\Pi| = T$ for moving block permutation, we have

$$\frac{1}{n}\sum_{t=1}^{T}\left(\hat{\lambda}_1'\hat{F}_t-\lambda_1'F_t\right)^2=O_P\left(\frac{1}{\min\{N,T\}}\right).$$

Finally, notice that Theorem 3 of Bai (2003) implies $\hat{\lambda}_1'\hat{F}_t-\lambda_1'F_t=O_P(1/\delta_{NT})$. The proof is complete.

H.7 Proof of Lemma 3

We recite conditions from Bai (2009). Following standard notation, we use i instead of j to denote units.

Suppose that there exists a constant $D_1>0$ the following conditions hold:
(1) $\max_{i,t}E\|X_{it}\|_2^4\le D_1$, $\max_t E\|F_t\|_2^4\le D_1$, $\max_i E\|\lambda_i\|_2^4\le D_1$ and $\max_{i,t}E|u_{it}|^8\le D_1$.
(2) $N^{-1}\sum_{i=1}^N\sum_{j=1}^N\max_{t,s}|E(u_{it}u_{js})|\le D_1$and $T^{-1}\sum_{s=1}^T\sum_{t=1}^T\max_{i,j}|E(u_{it}u_{js})|\le D_1$.
(3) $(NT)^{-1}\sum_{i=1}^N\sum_{j=1}^N\sum_{s=1}^T\sum_{t=1}^T|E(u_{it}u_{js})|\le D_1$.
(4) $\max_{t,s}E\left|N^{-1/2}\sum_{i=1}^N[u_{is}u_{it}-E(u_{is}u_{it})]\right|^4\le D_1$.
(5) $T^{-2}N^{-1}\sum_{t,s,q,v}\sum_{i,j}|cov(u_{it}u_{ts},u_{jq}u_{jv})|\le D_1$
(6) $T^{-1}N^{-2}\sum_{t,s}\sum_{i,j,k,q}|cov(u_{it}u_{jt},u_{ks}u_{qs})|\le D_1$.
(7) the largest eigenvalue of $E(u_iu_i')$ is bounded by D_1, where $u_i=(u_{i1},...,u_{iT})'\in\mathbb{R}^T$.

Moreover, the following conditions also hold: (8) $u=(u_1,\dots,u_N)$ is independent of (X,F,Λ). (9) $F'F/T=\Sigma_F+o_P(1)$ and $\Lambda'\Lambda/N=\Sigma_\Lambda+o_P(1)$ for some matrices Σ_F and Σ_Λ. (10) N/T is bounded away from zero and infinity. (11) For $X_i=(X_{i1},...,X_{iT})'\in\mathbb{R}^{T\times k_x}$ and $M_F=I_T-F(F'F)^{-1}F'$, we have

$$\inf_{F:\ F'F/T=I_k}\frac{1}{NT}\sum_{i=1}^N X_i'M_FX_i-\frac{1}{T}\left[\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N X_i'M_FX_j\lambda_i'(\Lambda'\Lambda/N)^{-1}\lambda_j\right]>0.$$

What follows below is the proof of the lemma. We introduce some notations used in Bai (2009). Let $H=(\Lambda'\Lambda/N)(F'\hat{F}/T)V_{NT}^{-1}$, where V_{NT} is the diagonal matrix that contains the k largest eigenvalues of $(NT)^{-1}\sum_{i=1}^N(Y_i^N-X_i\hat{\beta})(Y_i^N-X_i\hat{\beta})'$ with $Y_i^N=(Y_{i1}^N,Y_{i2}^N,...,Y_{iT}^N)'\in\mathbb{R}^T$. Let $\delta_{NT}=\min\{\sqrt{N},\sqrt{T}\}$. The rest of the proof proceeds in two steps. We first derive bounds for $\sum_{t=1}^T(\hat{u}_{1t}-u_{1t})^2$ and then prove the pointwise result.

Step 1: derive bounds for $\sum_{t=1}^T(\hat{u}_{1t}-u_{1t})^2$.

Define $\Delta_\beta=\hat{\beta}-\beta$ and $\Delta_{F,t}=\hat{F}_t-H'F_t$. Denote $\Delta_F=(\Delta_{F,1},...,\Delta_{F,T})'\in\mathbb{R}^{T\times k}$. Notice that $\hat{F} - FH = \Delta_F$. As pointed out on page 1237 of Bai (2009),

$$\hat{\lambda}_1 = T^{-1}\hat{F}'(Y_1^N - X_1\hat{\beta}) = T^{-1}\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta). \tag{H.24}$$

Notice that

$$\begin{aligned}
&|\hat{u}_{1t} - u_{1t}|^2 = \left|F_t'\lambda_1 - \hat{F}_t'\hat{\lambda}_1 - X_{1t}'\Delta_\beta\right|^2 \\
&\overset{\text{(i)}}{=} \left|F_t'\lambda_1 - T^{-1}(H'F_t + \Delta_{F,t})'\hat{F}'(u_1 + F\lambda_1 - X_1\Delta_\beta) - X_{1,t}'\Delta_\beta\right|^2 \\
&\leq \left[\left|F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right| + \left|T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right| + \left|T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right| + |X_{1t}'\Delta_\beta|\right]^2 \\
&\lesssim \left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right]^2 + \left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right]^2 + \left[T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right]^2 + [X_{1t}'\Delta_\beta]^2,
\end{aligned} \tag{H.25}$$

where (i) follows by (H.24) and $\hat{F}_t = H'F_t + \Delta_{F,t}$. Therefore,

$$\begin{aligned}
&\sum_{t=1}^T (\hat{u}_{1t} - u_{1t})^2 \lesssim \sum_{t=1}^T \left[F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right]^2 + \sum_{t=1}^T \left[T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right]^2 \\
&\qquad + \sum_{t=1}^T \left[T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right]^2 + \sum_{t=1}^T [X_{1t}'\Delta_\beta]^2 \\
&\overset{\text{(i)}}{=} \lambda_1'\left(I_k - H\hat{F}'F/T\right)'(F'F)\left(I_k - H\hat{F}'F/T\right)\lambda_1 \\
&\qquad + T^{-2}\left(\hat{F}'F\lambda_1\right)'(\Delta_F'\Delta_F)\left(\hat{F}'F\lambda_1\right) + T^{-1}\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\|_2^2 + \|X_1\Delta_\beta\|_2^2 \\
&\overset{\text{(ii)}}{=} O_P\left(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-4}\right) + O_P\left(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2}\right) + O_P\left(1 + T\delta_{NT}^{-4} + T\|\Delta_\beta\|_2^2\right) + O_P(T\|\Delta_\beta\|_2^2) \\
&= O_P\left(1 + T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2}\right),
\end{aligned}$$

where (i) follows by $\sum_{t=1}^T \hat{F}_t\hat{F}_t' = \hat{F}'\hat{F} = TI_k$ and (ii) follows by Lemma H.9, together with $\|F\| = O_P(\sqrt{T})$, $\lambda_1 = O(1)$ and $\|\hat{F}\| = O_P(\sqrt{T})$. Since $N \asymp T$, Theorem 1 of Bai (2009) implies $\|\Delta_\beta\|_2 = O_P(1/\sqrt{NT}) = O_P(T^{-1})$. Therefore, the above display implies

$$\sum_{t=1}^T (\hat{u}_{1t} - u_{1t})^2 = O_P(1).$$

Step 2: show the pointwise result.

By (H.25), we have

$$|\hat{u}_{1t} - u_{1t}| \leq \left|F_t'\left(I_k - H\hat{F}'F/T\right)\lambda_1\right| + \left|T^{-1}\Delta_{F,t}'\hat{F}'F\lambda_1\right| + \left|T^{-1}\hat{F}_t'\hat{F}'(u_1 - X_1\Delta_\beta)\right| + |X_{1t}'\Delta_\beta|$$

$$\overset{\text{(i)}}{\leq} \|F_t\|_2 \cdot \|\lambda_1\|_2 \cdot O_P\left(\|\Delta_\beta\|_2 + \delta_{NT}^{-2}\right) + O_P\left(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2}\right) \cdot T^{-1}\|F\lambda_1\|_2$$
$$+ T^{-1}\|\hat{F}_t\|_2 \cdot O_P\left(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2\right) + \|X_{1t}\|_2 \cdot \|\Delta_\beta\|_2 \overset{\text{(ii)}}{\leq} O_P(T^{-1/2}),$$

where (i) follows by $I_k - H\hat{F}'F/T = O_P(\|\Delta_\beta\|_2 + \delta_{NT}^{-2})$, $\|\hat{F}\Delta_{F,t}\| = O_P(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2})$, and $\|\hat{F}'(u_1 - X_1\Delta_\beta)\|_2 = O_P(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2)$ (due to Lemma H.9), whereas (ii) follows by $\|\hat{F}_t\|_2 = O_P(1)$ (Lemma H.9), $\|X_{1t}\|_2 = O_P(1)$, $\|F_t\|_2 = O_P(1)$, $\lambda_1 = O(1)$, $\|\Delta_\beta\|_2 = O_P(T^{-1})$ and $\|F\lambda_1\|_2 = O_P(\sqrt{T})$.

Lemma H.9. *Suppose that the assumption of Lemma 3 holds. Let δ_{NT}, H, Δ_F and u_1 be defined as in the proof of Lemma 3. Then (1) $I_k - H\hat{F}'F/T = O_P(\|\Delta_\beta\|_2 + \delta_{NT}^{-2})$; (2) $\Delta_F'\Delta_F = O_P(T\|\Delta_\beta\|_2^2 + T\delta_{NT}^{-2})$; (3) $\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\| = O_P\left(\sqrt{T} + T\delta_{NT}^{-2} + T\|\Delta_\beta\|_2\right)$; (4) $\|X_1\Delta_\beta\|_2 = O_P(\sqrt{T}\|\Delta_\beta\|_2)$; (5) $\|\hat{F}\Delta_{F,t}\|_2 = O_P(T\|\Delta_\beta\|_2 + T\delta_{NT}^{-2})$; (6) $\|\hat{F}_t\|_2 = O_P(1)$ for $1 \leq t \leq T$.*

Proof. **Proof of part (1).** Lemma A.7(i) in Bai (2009) implies HH' converges in probability to a nonsingular matrix. Hence,

$$H = O_P(1) \quad \text{and} \quad H^{-1} = O_P(1). \tag{H.26}$$

Notice that

$$\begin{aligned} I_k - H\hat{F}'F/T &\overset{\text{(i)}}{=} I_k - H(FH + \Delta_F)'F/T = I_k - (HH')(F'F/T) - H\Delta_F'F/T \\ &\overset{\text{(ii)}}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}) - H\Delta_F'F/T \\ &\overset{\text{(iii)}}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}), \end{aligned} \tag{H.27}$$

where (i) holds by $\hat{F} = FH + \Delta_F$, (ii) holds by $I_k - (HH')(F'F/T) = O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2})$ (due to Lemma A.7(i) in Bai (2009)) and (iii) holds by (H.26) and $\Delta_F'F/T = O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2})$ (due to Lemma A.3(i) in Bai (2009)). This proves part (1).

Proof of part (2). Part (2) follows by Proposition A.1 of Bai (2009):

$$T^{-1}\Delta_F'\Delta_F = O_P(\|\Delta_\beta\|_2^2) + O_P(\delta_{NT}^{-2}). \tag{H.28}$$

Proof of part (3). To see part (3), first observe that the independence between u_1 and F implies that

$$E(\|F'u_1\|^2 \mid F) \leq \sum_{t=1}^{T} E(F_t'F_t u_{1t}^2 \mid F) = \sum_{t=1}^{T} F_t'F_t E(u_{1t}^2).$$

It follows that

$$E\left(\|F'u_1\|^2\right) \leq \sum_{t=1}^{T} E(F_t'F_t)E(u_{1t}^2) \overset{\text{(i)}}{\lesssim} T\sum_{t=1}^{T} E(u_{1t}^2) = O(T),$$

where (i) holds by the uniform boundedness of $E(F_t'F_t)$. This means that

$$\|F'u_1\|_2 = O_P(\sqrt{T}). \tag{H.29}$$

Notice that

$$\begin{aligned}\left\|\hat{F}'(u_1 - X_1\Delta_\beta)\right\|_2 &\leq \|H'F'u_1\|_2 + \left\|\left(\hat{F} - FH\right)' u_1\right\| + \|\hat{F}\| \cdot \|X_1\| \cdot \|\Delta_\beta\|_2 \\ &\overset{\text{(i)}}{=} \|H'F'u_1\|_2 + \left(O_P(T^{1/2}\|\Delta_\beta\|_2) + O_P(T\delta_{NT}^{-2})\right) + O_P(T\|\Delta_\beta\|_2) \\ &\overset{\text{(ii)}}{=} O_P(\sqrt{T}) + \left(O_P(T^{1/2}\|\Delta_\beta\|_2) + O_P(T\delta_{NT}^{-2})\right) + O_P(T\|\Delta_\beta\|_2),\end{aligned}$$

where (i) follows by $\left(\hat{F} - FH\right)' u_1/T = O_P(T^{-1/2}\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2})$ (due to Lemma A.4 in Bai (2009)) and the fact that $\|\hat{F}\| = O(\sqrt{T})$ and $\|X_1\| = O_P(\sqrt{T})$ (see the beginning of Appendix A in Bai (2009)), whereas (ii) follows by (H.26) and (H.29). We have proved part (3).

Proof of part (4). We notice that $\|X_1\| = O_P(\sqrt{T})$; see the beginning of Appendix A in Bai (2009). Part (4) follows by $\|X_1\Delta_\beta\| \leq \|X_1\| \cdot \|\Delta_\beta\|_2$.

Proof of part (5). Notice that

$$\|\hat{F}\Delta_{F,t}\|_2/T \leq \|\hat{F}\Delta_F\|/T \overset{\text{(i)}}{\leq} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}),$$

where (i) follows by Lemma A.3(ii) in Bai (2009). We have proved part (5).

Proof of part (6). Notice that

$$T^{-1}\|\Delta_{F,t}\|_2^2 \leq T^{-1}\Delta_F'\Delta_F = T^{-1}\hat{F}'\Delta_F - T^{-1}H'F'\Delta_F \overset{\text{(i)}}{=} O_P(\|\Delta_\beta\|_2) + O_P(\delta_{NT}^{-2}),$$

where (i) follows by Lemma A.3(i)-(ii) of Bai (2009). By Theorem 1 of Bai (2009) and by the assumption of $N \asymp T$, we have that $\|\Delta_{F,t}\|_2 = O_P(1)$. By $\|\hat{F}_t\|_2 \leq \|H'F_t\|_2 + \|\Delta_{F,t}\|_2$, $F_t = O_P(1)$ and $H = O_P(1)$, we can see that $\|\hat{F}_t\|_2 = O_P(1)$. The proof is complete. $\square$

H.8 Proof of Lemma 4

We start with the assumptions. Recall $N = J + 1$. Assume that (1) $\{u_j\}_{j=1}^N$ is independent across j conditional on M, (2) $\max_{1\leq j\leq N} T^{-1}\sum_{t=1}^T E(|u_{jt}|^{2\kappa_1} \mid M) = O_P(1)$ for some constant $\kappa_1 > 1$, (3) $\|N^{-1}\sum_{j=1}^N E(u_j u_j' \mid M)\| = O_P(1)$ and (4) there exists a sequence $\ell_T > 0$ such that
$\ell_T (NT)^{-1} K\sqrt{N \vee (N^{1/\kappa_1} T \log N)} = o(1)$ and with probability $1 - o(1)$, $T^{-1}\sum_{t=1}^T(\hat{M}_{1t} - M_{1t})^2 \leq \ell_T (NT)^{-1}\sum_{t=1}^T\sum_{j=1}^N(\hat{M}_{jt} - M_{jt})^2$ and $(\hat{M}_{1t} - M_{1t})^2 \leq \ell_T(NT)^{-1}\sum_{t=1}^T\sum_{j=1}^N(\hat{M}_{jt} - M_{jt})^2$ for $T_0 + 1 \leq t \leq T$.

We now prove Lemma 4. Define $\Delta = \hat{M} - M$. Let $Y \in \mathbb{R}^{T\times N}$ be the matrix whose (t, j) entry is Y_{jt}^N. For (j, t), define the matrix $Q_{jt} \in \mathbb{R}^{N\times T}$ by $Q_{is} = \mathbf{1}\{(i, s) = (j, t)\}$, i.e, a matrix of zeros except that the (j, t) entry is one. Then we can write the model as

$$Y_{jt}^N = \mathrm{trace}(Q_{jt}'M) + u_{jt} \qquad \text{for} \quad 1 \leq j \leq N,\ 1 \leq t \leq T. \tag{H.30}$$

Notice that the estimator $\hat{M}$ satisfies

$$\|\hat{M}\|_* \leq K$$

and

$$\sum_{t=1}^T\sum_{j=1}^N\left(Y_{jt}^N - \mathrm{trace}(Q_{jt}'\hat{M})\right)^2 \leq \sum_{t=1}^T\sum_{j=1}^N\left(Y_{jt}^N - \mathrm{trace}(Q_{jt}'M)\right)^2.$$

Plugging (H.30) into the above inequality and rearranging terms, we obtain

$$\begin{aligned}\sum_{t=1}^T\sum_{j=1}^N\left(\mathrm{trace}(Q_{jt}'\Delta)\right)^2 \leq 2\sum_{t=1}^T\sum_{j=1}^N u_{jt}\mathrm{trace}(Q_{jt}'\Delta) &= 2\mathrm{trace}\left(\left[\sum_{t=1}^T\sum_{j=1}^N u_{jt}Q_{jt}\right]'\Delta\right)\\ &\overset{\text{(i)}}{=} 2\mathrm{trace}(u'\Delta)\\ &\overset{\text{(ii)}}{\leq} 2\|u\|\cdot\|\Delta\|_*\\ &\overset{\text{(iii)}}{\leq} 4K\|u\|,\end{aligned} \tag{H.31}$$

where (i) follows by $\sum_{t=1}^T\sum_{j=1}^N u_{jt}Q_{jt} = u$, (ii) follows by the trace duality property (see e.g., McCarthy (1967), Rotfeld (1969) and Rohde and Tsybakov (2011)) and (iii) follows by the fact that $\|\hat{M}\|_* \leq K$ and $\|M\|_* \leq K$.

To bound $\|u\|$, we apply Lemma H.10. Note that the conditions of Lemma H.10 are satisfied by our assumption. Therefore, $E(\|u\| \mid M) = O_P\left(\sqrt{N \vee (N^{1/\kappa_1}T\log N)}\right)$. This and (H.31) imply that

$$(NT)^{-1}\sum_{t=1}^{T}\sum_{j=1}^{N}\left(\mathrm{trace}(Q'_{jt}\Delta)\right)^2 = O_P\left((NT)^{-1}K\sqrt{N\vee(N^{1/\kappa_1}T\log N)}\right).$$

The desired result follows by Assumption (4) listed at the beginning of the proof.

Lemma H.10. *Suppose that the following conditions hold:*
(i) $\{u_j\}_{j=1}^N$ *is independent across* j *conditional on* M.
(ii) $\max_{1\le j\le N} T^{-1}\sum_{t=1}^T E(|u_{jt}|^{2\kappa_1} \mid M) = O_P(1)$ *for some constant* $\kappa_1 > 1$.
(iii) $\|N^{-1}\sum_{j=1}^N E(u_j u_j' \mid M)\| = O_P(1)$.

Then $E(\|u\| \mid M) = O_P\left(\sqrt{N\vee(N^{1/\kappa_1}T\log N)}\right)$.

Proof. Recall the elementary inequality $|T^{-1}\sum_{t=1}^T a_t| \le [T^{-1}\sum_{t=1}^T |a_t|^\kappa]^{1/\kappa}$ for any $\kappa > 1$ (due to Liapunov's inequality). It follows that $T^{-1}\sum_{t=1}^T u_{jt}^2 \le [T^{-1}\sum_{t=1}^T |u_{jt}|^{2\kappa_1}]^{1/\kappa_1}$, which means

$$\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1} \le T^{\kappa_1-1}\sum_{t=1}^T |u_{jt}|^{2\kappa_1}. \tag{H.32}$$

Hence,

$$\begin{aligned}
E\left(\left[\max_{1\le j\le N}\sum_{t=1}^T u_{jt}^2\right] \mid M\right) &\overset{\text{(i)}}{\le} \left\{E\left[\max_{1\le j\le N}\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1} \mid M\right]\right\}^{1/\kappa_1}\\
&\le \left\{E\left[\sum_{i=1}^N\left(\sum_{t=1}^T u_{jt}^2\right)^{\kappa_1} \mid M\right]\right\}^{1/\kappa_1}\\
&\overset{\text{(ii)}}{\le} \left\{E\left[T^{\kappa_1-1}\sum_{j=1}^N\sum_{t=1}^T |u_{jt}|^{2\kappa_1} \mid M\right]\right\}^{1/\kappa_1}\\
&\le \left\{\left[NT^{\kappa_1}\max_{1\le i\le N} T^{-1}\sum_{t=1}^T E\left(|u_{jt}|^{2\kappa_1} \mid M\right)\right]\right\}^{1/\kappa_1}\\
&\overset{\text{(iii)}}{=} \{[NT^{\kappa_1}O_P(1)]\}^{1/\kappa_1} = O_P(N^{1/\kappa_1}T),
\end{aligned}$$

where (i) follows by Liapunov's inequality, (ii) follows by (H.32) and (iii) follows by the assumption that $\max_{1\le j\le N} T^{-1}\sum_{t=1}^T E(|u_{jt}|^{2\kappa_1} \mid M) = O_P(1)$. Therefore, it follows, by Theorem 5.48 and Remark 5.49 of Vershynin (2010), that

$$\begin{aligned}
E\left(\|u\| \mid M\right) \le \sqrt{E\left(\|u\|^2 \mid M\right)} \le \|E(u'u \mid M)/N\|^{1/2}\sqrt{N}\\
+ O\left(\sqrt{O(N^{1/\kappa_1}T)\log\min\left(O(N^{1/\kappa_1}T), N\right)}\right)
\end{aligned}$$

$$\overset{\text{(i)}}{\leq} O_P\left(\sqrt{N}\right) + O\left(\sqrt{N^{1/\kappa_1} T \log N}\right),$$

where (i) holds by the assumption of $\|E(u'u \mid M)/N\| = \|N^{-1}\sum_{j=1}^{N} E(u_j u_j' \mid M)\| = O_P(1)$. The proof is complete. $\square$

H.9 Proof of Lemma 5

By the analysis on page 215-216 of Hamilton (1994) (leading to Equation (8.2.29) therein), we have that $\hat{\rho} - \rho = o_P(1)$. Hence,

$$\sum_{t=K+1}^{T} (\hat{u}_t - u_t)^2 = \sum_{t=K+1}^{T} (y_t'(\rho - \hat{\rho}))^2 = (\hat{\rho} - \rho)'\left(\sum_{t=K+1}^{T} y_t y_t'\right)(\hat{\rho} - \rho) \leq \|\hat{\rho} - \rho\|_2^2 \times \left\|\sum_{t=K+1}^{T} y_t y_t'\right\|.$$

The analysis on page 215 of Hamilton (1994) (leading to Equation (8.2.26) therein) implies that

$$T^{-1}\sum_{t=K+1}^{T} y_t y_t' = E(y_t y_t') + o_P(1),$$

which means $\left\|\sum_{t=K+1}^{T} y_t y_t'\right\| = O_P(T)$. Since $\hat{\rho} - \rho = o_P(1)$, the above display implies that

$$\sum_{t=K+1}^{T} (\hat{u}_t - u_t)^2 = o_P(T).$$

Since $\hat{u}_t - u_t = y_t'(\rho - \hat{\rho})$, the pointwise consistency follows by $\hat{\rho} - \rho = o_P(1)$ and the fact that $y_t = O_P(1)$ for $T_0 + 1 \leq t \leq T$ (due to the stationarity property of u_t).

H.10 Proof of Lemma 6

By assumption, $\max_{K+1 \leq t \leq T} |\hat{P}_t^N - P_t^N| \leq \ell_T \|\hat{\rho} - \rho\|$. Therefore,

$$\frac{1}{T}\sum_{t=K+1}^{T} (\hat{P}_t^N - P_t^N)^2 \leq \ell_T^2 \|\hat{\rho} - \rho\|$$

and

$$\max_{T_0+1 \leq t \leq T} |\hat{P}_t^N - P_t^N| \leq \ell_T \|\hat{\rho} - \rho\|.$$

Since $\ell_T \|\hat{\rho} - \rho\| = o_P(1)$, the desired result follows.

H.11 Proof of Lemma 7

We first derive the following result that is useful in proving Lemma 7.

Lemma H.11. *Recall $\varepsilon_t = x_t'\rho + u_t$, where $\rho = (\rho_1, \rho_2, ..., \rho_K)' \in \mathbb{R}^K$ and $x_t = (\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-K})' \in \mathbb{R}^K$. Suppose that the following hold: (1) $\{u_t\}_{t=1}^T$ is an iid sequence with $E(u_1^4)$ uniformly bounded. (2) the roots of $1 - \sum_{j=1}^K \rho_j L^j = 0$ are uniformly bounded away from the unit circle.*

Then we have (i) $(T-K)^{-1}\sum_{t=K+1}^T u_t^2 = O_P(1)$; (ii) $(T-K)^{-1}\sum_{t=K+1}^T x_t u_t = o_P(1)$; (iii) $(T-K)^{-1}\sum_{t=K+1}^T \|x_t\|^2 = O_P(1)$. (iv) There exists a constant $\lambda_0 > 0$ such that the smallest eigenvalue of $(T-K)^{-1}\sum_{t=K+1}^T x_t x_t'$ is bounded below by λ_0 with probability approaching one.

Proof. **Proof of part (i)**. Part (i) follows by the law of large numbers; see e.g., Theorem 3.1 of White (2014).

Proof of part (ii). Let $\mathcal{F}_t$ be the σ-algebra generated by $\{u_s : s \leq t\}$. First notice that $\{x_t u_t\}_{t=K+1}^T$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}$. Since ε_t is a stationary process, we have that $E\|x_t u_t\|^2 = \sum_{j=1}^K E(\varepsilon_{t-j}^2 u_t^2) = \sum_{j=1}^K E(\varepsilon_{t-j}^2)E(u_t^2)$ is uniformly bounded. Hence, part (ii) follows by Exercise 3.77 of White (2014).

Proof of part (iii). To see part (iii), notice that $\|x_t\|^2 = x_t'x_t = \sum_{j=1}^K \varepsilon_{t-j}^2$. By the analysis on page 215 of Hamilton (1994), for each $1 \leq j \leq K$, $(T-K)^{-1}\sum_{t=K+1}^T \varepsilon_{t-j}^2 = E(\varepsilon_{t-j}^2) + o_P(1)$. Thus, part (iii) follows by

$$(T-K)^{-1}\sum_{t=K+1}^T \|x_t\|^2 = (T-K)^{-1}\sum_{j=1}^K \sum_{t=K+1}^T \varepsilon_{t-j}^2 = K\left(E(\varepsilon_t^2) + o_P(1)\right).$$

Proof of part (iv). Similarly, the analysis on page 215 of Hamilton (1994) implies that

$$(T-K)^{-1}\sum_{t=K+1}^T x_t x_t' = o_P(1) + Ex_t x_t'.$$

By Proposition 5.1.1 of Brockwell and Davis (2013), $E(x_t x_t')$ has eigenvalues bounded away from zero. Part (iv) follows. □

Now we are ready to prove Lemma 7.

Proof of Lemma 7. Define $\delta_t = \hat{\varepsilon}_t - \varepsilon_t$, $\Delta_t = \hat{x}_t - x_t$, $\tilde{u}_t = u_t + \delta_t - \Delta_t'\rho$ and $a_t = \tilde{u}_t - u_t$. Notice that

$$\hat{\varepsilon}_t = \delta_t + \varepsilon_t = \delta_t + x_t'\rho + u_t = \delta_t + (\hat{x}_t - \Delta_t)'\rho + u_t = \hat{x}_t'\rho + \tilde{u}_t. \tag{H.33}$$

Therefore,

$$\hat{\rho} = \left(\sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t'\right)^{-1} \left(\sum_{t=K+1}^{T} \hat{x}_t \hat{\varepsilon}_t\right) = \left(\sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t'\right)^{-1} \left(\sum_{t=K+1}^{T} \hat{x}_t (\hat{x}_t' \rho + \tilde{u}_t)\right)$$
$$= \rho + \left(\sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t'\right)^{-1} \left(\sum_{t=K+1}^{T} \hat{x}_t \tilde{u}_t\right). \tag{H.34}$$

The rest of the proof proceeds in three steps. First two steps show that $(T-K)^{-1}\sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t'$ is well-behaved and $(T-K)^{-1}\sum_{t=K+1}^{T} \hat{x}_t \tilde{u}_t = o_P(1)$. This would imply $\hat{\rho} = \rho + o_P(1)$. In the third step, we derive the final result.

Step 1: show that $\left[(T-K)^{-1}\sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t'\right]^{-1} = O_P(1)$.

It is not hard to see that $\|\Delta_t\|^2 = \sum_{s=t-1}^{t-K} \delta_s^2$. Therefore,

$$\sum_{t=K+1}^{T} \|\Delta_t\|^2 = \sum_{t=K+1}^{T} \sum_{s=t-1}^{t-K} \delta_s^2 \leq K \sum_{t=1}^{T} \delta_t^2 \overset{\text{(i)}}{=} o_P(T), \tag{H.35}$$

where (i) follows by the assumption of $T^{-1}\sum_{t=1}^{T} \delta_t^2 = o_P(1)$. Notice that

$$\left\|\sum_{t=K+1}^{T} (\hat{x}_t \hat{x}_t' - x_t x_t')\right\| = \left\|\sum_{t=K+1}^{T} (x_t \Delta_t' + \Delta_t x_t' + \Delta_t \Delta_t')\right\|$$
$$\leq 2 \sum_{t=K+1}^{T} \|x_t\| \cdot \|\Delta_t\| + \sum_{t=K+1}^{T} \|\Delta_t\|^2$$
$$\leq 2\sqrt{\left(\sum_{t=K+1}^{T} \|x_t\|^2\right)\left(\sum_{t=K+1}^{T} \|\Delta_t\|^2\right)} + \sum_{t=K+1}^{T} \|\Delta_t\|^2 \overset{\text{(i)}}{=} o_P(T), \tag{H.36}$$

where (i) follows by (H.35) and Lemma H.11. Thus,

$$\left\|\frac{1}{T-K}\sum_{t=K+1}^{T} (\hat{x}_t \hat{x}_t' - x_t x_t')\right\| = o_P(1).$$

By Lemma H.11, the smallest eigenvalue of $(T-K)^{-1}\sum_{t=K+1}^{T} x_t x_t'$ is bounded below by a positive constant with probability approaching one. It follows that

$$\left[(T-K)^{-1}\sum_{t=K+1}^{T} \hat{x}_t \hat{x}_t'\right]^{-1} = O_P(1). \tag{H.37}$$

Step 2: show that $(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t = o_P(1)$.

By Lemma H.11, we have

$$(T-K)^{-1}\sum_{t=K+1}^{T} x_t u_t = o_P(1). \tag{H.38}$$

Notice that

$$\begin{aligned}
\left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\hat{x}_t\tilde{u}_t - x_t u_t)\right\| &= \left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\Delta_t u_t + x_t a_t + \Delta_t a_t)\right\| \\
&\le \frac{1}{T-K}\sum_{t=K+1}^{T}(\|\Delta_t u_t\| + \|x_t a_t\| + \|\Delta_t a_t\|) \\
&\le \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}u_t^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|x_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}a_t^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|\Delta_t\|^2\right)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}a_t^2\right)}.
\end{aligned} \tag{H.39}$$

We observe that

$$\begin{aligned}
\sum_{t=K+1}^{T}a_t^2 = \sum_{t=K+1}^{T}(\delta_t - \Delta_t'\rho)^2 &\le 2\sum_{t=K+1}^{T}\delta_t^2 + 2\sum_{t=K+1}^{T}(\Delta_t'\rho)^2 \\
&\le 2\sum_{t=1}^{T}\delta_t^2 + 2\|\rho\|^2\sum_{t=K+1}^{T}\|\Delta_t\|^2 \overset{\text{(i)}}{=} O_P(T),
\end{aligned} \tag{H.40}$$

where (i) follows by (H.35) and the assumption of $T^{-1}\sum_{t=1}^{T}\delta_t^2 = o_P(1)$. Combining (H.39) with (H.35) and (H.40), we obtain

$$\begin{aligned}
&\left\|\frac{1}{T-K}\sum_{t=K+1}^{T}(\hat{x}_t\tilde{u}_t - x_t u_t)\right\| \\
&\le \sqrt{o_P(1)\left(\frac{1}{T-K}\sum_{t=K+1}^{T}u_t^2\right)} + \sqrt{\left(\frac{1}{T-K}\sum_{t=K+1}^{T}\|x_t\|^2\right)o_P(1)} + \sqrt{o_P(1)\times o_P(1)} \overset{\text{(i)}}{=} o_P(1),
\end{aligned} \tag{H.41}$$

where (i) follows by Lemma H.11. Now we combine (H.38) and (H.41), obtaining

$$(T-K)^{-1}\sum_{t=K+1}^{T}\hat{x}_t\tilde{u}_t = o_P(1). \tag{H.42}$$

By (H.34) together with (H.37) and (H.42), it follows that

$$\hat{\rho}-\rho = o_P(1). \tag{H.43}$$

Step 3: show the desired result.

Recall that $\hat{u}_t = \hat{\varepsilon}_t - \hat{x}_t'\hat{\rho}$. Hence,

$$\hat{u}_t - u_t = (\hat{\varepsilon}_t - \hat{x}_t'\hat{\rho}) - u_t \overset{\text{(i)}}{=} (\hat{x}_t'(\rho-\hat{\rho})+\tilde{u}_t) - u_t = \hat{x}_t'(\rho-\hat{\rho}) + a_t, \tag{H.44}$$

where (i) follows by (H.33). Therefore, we have

$$\begin{aligned}
\sum_{t=K+1}^{T}(\hat{u}_t-u_t)^2 &= \sum_{t=K+1}^{T}\left(\hat{x}_t'(\rho-\hat{\rho})+a_t\right)^2 \\
&\leq 2\sum_{t=K+1}^{T}\left(\hat{x}_t'(\hat{\rho}-\rho)\right)^2 + 2\sum_{t=K+1}^{T}a_t^2 \\
&\leq 2\|\hat{\rho}-\rho\|^2\sum_{t=K+1}^{T}\|\hat{x}_t\|^2 + 2\sum_{t=K+1}^{T}a_t^2 \\
&= 2\|\hat{\rho}-\rho\|^2\left(\sum_{t=K+1}^{T}\mathrm{trace}(x_tx_t') + \sum_{t=K+1}^{T}\mathrm{trace}(\hat{x}_t\hat{x}_t' - x_tx_t')\right) + 2\sum_{t=K+1}^{T}a_t^2 \\
&\overset{\text{(i)}}{\leq} o_P(1)\times(O_P(T)+o_P(T)) + o_P(T) = o_P(T),
\end{aligned}$$

where (i) follows by (H.36), (H.43), (H.40) and Lemma H.11.

To see the pointwise result, we notice that by (H.44) and (H.43), it suffices to verify that $a_t = o_P(1)$ and $\hat{x}_t = O_P(1)$ for $T_0+1\leq t\leq T$.

Since $\hat{x}_t - x_t = (\delta_{t-1},\delta_{t-2},...,\delta_{t-K})'$, the assumption of pointwise convergence of $\hat{\varepsilon}_t$ (i.e., $\delta_t = o_P(1)$ for $T_0+1-K\leq t\leq T$) implies that $\hat{x}_t - x_t = o_P(1)$ for $T_0+1\leq t\leq T$. Since $x_t = O_P(1)$ due to the stationarity condition, we have $\hat{x}_t = O_P(1)$ for $T_0+1\leq t\leq T$.

Since both δ_t and Δ_t are both $o_P(1)$ for $T_0+1\leq t\leq T$, so is $a_t = \delta_t - \Delta_t'\rho$. Hence, we have proved the pointwise result. The proof is complete. □

H.12 Proof of Lemma E.1

Fix an arbitrary $\eta > 0$. Define $a_\eta = \inf_{\|\beta-\beta_*\|_2 \geq \eta}(L(\beta) - L(\beta_*))/3$. By the compactness of $\mathcal{B}$ and the uniqueness of the minimum of $L(\cdot)$, we have $a_\eta > 0$.

(Otherwise, one can find a sequence $\{\beta_k\}_{k=1}^{\infty}$ with $\|\beta_k - \beta_*\|_2 \geq \eta$ for all $k \geq 1$ with $L(\beta_k) \to L(\beta_*)$. By compactness of $\mathcal{B}$ implies that some subsequence of β_k converges to a point $\beta_{**} \in \mathcal{B}$. Clearly, $\|\beta_{**} - \beta_*\|_2 \geq \eta$. The continuity of $L(\cdot)$ implies $L(\beta_{**}) = L(\beta_*)$. This contradicts the uniqueness of the minimum of $L(\cdot)$.)

Define the event

$$\mathcal{M} = \left\{ \sup_{\beta} |\hat{L}(\mathbf{Z}; \beta) - L(\beta)| \leq a_\eta \right\} \bigcap \left\{ \max_{H \in \mathcal{H}} \sup_{\beta} |\hat{L}(\mathbf{Z}_H; \beta) - L(\beta)| \leq a_\eta \right\}.$$

By the assumption, we know $P(\mathcal{M}) = 1 - o(1)$.

Notice that on the event $\mathcal{M}$,

$$L(\hat{\beta}(\mathbf{Z})) - L(\beta_*) \leq a_\eta + \hat{L}(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - L(\beta_*) \leq 2a_\eta + \hat{L}(\mathbf{Z}; \hat{\beta}(\mathbf{Z})) - \hat{L}(\mathbf{Z}; \beta_*) \leq 2a_\eta.$$

It follows by the definition of a_η that $\|\hat{\beta}(\mathbf{Z}) - \beta_*\|_2 \leq \eta$ on the event $\mathcal{M}$. Thus, $P(\|\hat{\beta} - \beta_*\|_2 \leq \eta) \geq P(\mathcal{M}) = 1 - o(1)$. Since $\eta > 0$ is arbitrary, we have $\|\hat{\beta}(\mathbf{Z}) - \beta_*\|_2 = o_P(1)$.

By the same analysis, we have that on the event $\mathcal{M}$, $\|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 \leq \eta$ for all $H \in \mathcal{H}$. Thus, on the event $\mathcal{M}$, $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 \leq \eta$. We have that $\max_{H\in\mathcal{H}} \|\hat{\beta}(\mathbf{Z}_H) - \beta_*\|_2 = o_P(1)$. The desired result follows.

H.13 Proof of Lemma E.2

For notational simplicity, we write $\hat{\beta} = \hat{\beta}(\mathbf{Z})$ and $\hat{\beta}_H = \hat{\beta}(\mathbf{Z}_H)$. Define the event $\mathcal{M} = \mathcal{M}_1 \bigcap \mathcal{M}_2$, where $\mathcal{M}_1 = \{\min_{\|v\|_0 \leq m} v'\hat{\Sigma}v/\|v\|_2^2 \geq \kappa_0\} \bigcap \{\|\hat{\beta}\|_0 \leq s/2\}$ and

$$\mathcal{M}_2 = \bigcap_{H \in \mathcal{H}} \left(\left\{ \|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \leq c_T,\ \|\hat{\mu}_H - \hat{\mu}\|_\infty \leq c_T \right\} \bigcap \left\{ \max_{H \in \mathcal{H}} \|\hat{\beta}_H\|_0 \leq s/2 \right\} \right).$$

By assumption, $P(\mathcal{M}) \geq 1 - \gamma_{1,T} - \gamma_{2,T} - \gamma_{3,T}$. The rest of the argument are statements on the event $\mathcal{M}$.

Fix $H \in \mathcal{H}$, let $\Delta = \hat{\beta}_H - \hat{\beta}$. Define $\xi = \hat{\mu} - \hat{\Sigma}\hat{\beta}$ and $\xi_H = \hat{\mu}_H - \hat{\Sigma}_H\hat{\beta}$. Since $\|\hat{\beta}\|_1 \leq K$, we have

$$\|\xi_H - \xi\|_\infty \leq \|\hat{\mu}_H - \hat{\mu}\|_\infty + \|\hat{\Sigma}_H - \hat{\Sigma}\|_\infty \|\hat{\beta}\|_1 \leq c_T(1 + K). \tag{H.45}$$

When $\Delta = 0$, the result clearly holds. Now we consider the case with $\Delta \neq 0$.

Step 1: show that on the event $\mathcal{M}$, $0 \leq \lambda^2\Delta'\hat{\Sigma}\Delta - \lambda\Delta'\hat{\mu} \leq c_T K^2 + 2c_T K$ for any $\lambda \in [0,1]$.

Recall that $\hat{Q}(\beta) = \beta'\hat{\Sigma}\beta - 2\hat{\mu}'\beta + T^{-1}\sum_{t=1}^{T} Y_t^2$. Since the term $T^{-1}\sum_{t=1}^{T} Y_t^2$ does not affect the minimizer, we modify $\hat{Q}$ by dropping this term. With a slight abuse of notation, we still use the symbol $\hat{Q}(\beta) = \beta'\hat{\Sigma}\beta - 2\hat{\mu}'\beta$ and $\hat{Q}_H(\beta) = \beta'\hat{\Sigma}_H\beta - 2\hat{\mu}_H'\beta$. Therefore,

$$\hat{Q}(\beta) - \hat{Q}_H(\beta) = \beta'(\hat{\Sigma} - \hat{\Sigma}_H)\beta - 2(\hat{\mu} - \hat{\mu}_H)'\beta.$$

Since $\sup_{\beta\in\mathcal{W}} \|\beta\|_1 \leq K$, we have that on the event $\mathcal{M}$,

$$\sup_{\beta\in\mathcal{W}} \left|\hat{Q}(\beta) - \hat{Q}_H(\beta)\right| \leq c_T K^2 + 2c_T K.$$

Let $\bar{\beta} = \hat{\beta} + \lambda\Delta$, where $\lambda \in [0,1]$. Then clearly, $\bar{\beta} = \lambda\hat{\beta}_H + (1-\lambda)\hat{\beta}$. By definition of $\hat{\beta}$, we have that $\hat{Q}(\hat{\beta}) \leq \hat{Q}(\bar{\beta})$, which means that

$$\lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta \geq 0. \tag{H.46}$$

Clearly, $\hat{Q}_H(\hat{\beta}_H) \leq \hat{Q}_H(\bar{\beta})$ for any $\lambda \in [0,1]$. Hence, $\hat{Q}_H(\hat{\beta} + \Delta) \leq \hat{Q}_H(\hat{\beta} + \lambda\Delta)$ for any $\lambda \in (0,1)$. By $\hat{Q}_H(\beta) = \beta'\hat{\Sigma}_H\beta - 2\hat{\mu}_H'\beta$, this simplifies to $(1+\lambda)\Delta'\hat{\Sigma}_H\Delta \leq 2\xi_H'\Delta$. Since λ can be arbitrarily close to one, this means $\Delta'\hat{\Sigma}_H\Delta \leq \xi_H'\Delta$. It follows that for any $\lambda \in [0,1]$, we have

$$\lambda^2\Delta'\hat{\Sigma}_H\Delta \leq 2\lambda\xi_H'\Delta.$$

Notice that

$$\begin{aligned} 0 \leq 2\lambda\xi_H'\Delta - \lambda^2\Delta'\hat{\Sigma}_H\Delta &= 2\lambda\xi'\Delta - \lambda^2\Delta'\hat{\Sigma}\Delta + 2\lambda(\xi_H - \xi)'\Delta + \lambda^2\Delta'(\hat{\Sigma} - \hat{\Sigma}_H)\Delta \\ &\leq 2\lambda\xi'\Delta - \lambda^2\Delta'\hat{\Sigma}\Delta + 2\lambda\|\xi_H - \xi\|_\infty\|\Delta\|_1 + \lambda^2\|\hat{\Sigma} - \hat{\Sigma}_H\|_\infty\|\Delta\|_1^2. \end{aligned}$$

It follows, by (H.45) and $\|\Delta\|_1 \leq \|\hat{\beta}_H\|_1 + \|\hat{\beta}\|_1 \leq 2K$, that

$$\lambda^2\Delta'\Sigma\Delta - 2\lambda\xi'\Delta \leq 2\lambda\|\xi_H - \xi\|_\infty\|\Delta\|_1 + \lambda^2\|\hat{\Sigma} - \hat{\Sigma}_H\|_\infty\|\Delta\|_1^2 \leq 4c_T(1+K)K + 4c_T K^2. \tag{H.47}$$

Since (H.46) and (H.47) hold for any $\lambda \in [0,1]$, we have that

$$0 \leq \lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta \leq 4c_T K(2K+1) \qquad \forall\lambda \in [0,1]. \tag{H.48}$$

Step 2: show the desired result.

Since $0 \leq \lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta$ for any $\lambda \in (0,1)$, we have that $\xi'\Delta \leq \lambda\Delta'\hat{\Sigma}\Delta/2$ for any

$\lambda \in (0,1)$. Thus,

$$\xi'\Delta \leq 0.$$

Hence, by the second inequality in (H.48), for any $\lambda \in [0,1]$,

$$\lambda^2\Delta'\hat{\Sigma}\Delta \leq \lambda^2\Delta'\hat{\Sigma}\Delta - 2\lambda\xi'\Delta \leq 4c_T K(2K+1).$$

Now we take $\lambda = 1$, which implies

$$\Delta'\hat{\Sigma}\Delta \leq 4c_T K(2K+1).$$

Since $\|\Delta\|_0 \leq \|\hat{\beta}_H\|_0 + \|\hat{\beta}\|_0 \leq s$ and $\|\Delta\|_1 \leq \sqrt{\|\Delta\|_0}\|\Delta\|_2$, it follows that

$$4c_T K(2K+1) \geq \Delta'\hat{\Sigma}\Delta \geq \kappa_1\|\Delta\|_2^2 \geq \kappa_1 s^{-1}\|\Delta\|_1^2.$$

Hence, $\|\Delta\|_1 \leq 2\sqrt{\kappa_1 s c_T K(2K+1)}$ and

$$\left|(Y_t - X_t'\hat{\beta}) - (Y_t - X_t'\hat{\beta}_H)\right| = |X_t'\Delta| \leq \|X_t\|_\infty\|\Delta\|_1 \leq 2\kappa_2\sqrt{\kappa_1 s c_T K(2K+1)}.$$

On the event $\mathcal{M}$, the above bound holds for all $H \in \mathcal{H}$. The desired result follows by $P(\mathcal{M}) \geq 1 - \gamma_{1,T} - \gamma_{2,T} - \gamma_{3,T}$.

H.14 Proof of Lemma F.1

For notational simplicity, we write $\hat{\beta}$ instead of $\hat{\beta}_\lambda$ and $c_T = |H|$. Let $\hat{\Omega} = X'X$, $\tilde{\Omega} = \tilde{X}'\tilde{X}$, $\hat{\mu} = X'u$ and $\tilde{\mu} = \tilde{X}'\tilde{u}$. We work on the event on which $\kappa_1 T \leq \lambda_{\min}(\hat{\Omega}) \leq \lambda_{\max}(\hat{\Omega}) \leq \kappa_2 T$, $\|\hat{\mu} - \tilde{\mu}\|_2 \leq \kappa_3\sqrt{Jc_T}$ and $\|\hat{\Omega} - \tilde{\Omega}\| \leq \kappa_4(c_T + J)$. Let $q_1 \geq q_2 \geq \cdots \geq q_J$ denote the eigenvalues of $\hat{\Omega}$.

Step 1: show inconsistency.

Notice that

$$\begin{aligned}
\hat{\beta} &= (\hat{\Omega} + \lambda I)^{-1}X'(X\beta + u) \\
&= (\hat{\Omega} + \lambda I)^{-1}\hat{\Omega}\beta + (\hat{\Omega} + \lambda I)^{-1}X'u \\
&= \beta + \left[(\hat{\Omega} + \lambda I)^{-1}\hat{\Omega} - I\right]\beta + (\hat{\Omega} + \lambda I)^{-1}X'u \\
&= \beta + (\hat{\Omega} + \lambda I)^{-1}\left[\hat{\Omega} - (\hat{\Omega} + \lambda I)\right]\beta + (\hat{\Omega} + \lambda I)^{-1}X'u \\
&= \beta - \lambda(\hat{\Omega} + \lambda I)^{-1}\beta + (\hat{\Omega} + \lambda I)^{-1}X'u.
\end{aligned} \tag{H.49}$$

Therefore,

$$
\begin{aligned}
& E\left(\|X(\hat{\beta}-\beta)\|_2^2 \mid X\right) \\
& = \|\lambda X(\hat{\Omega}+\lambda I)^{-1}\beta\|_2^2 + E\left(\|X(\hat{\Omega}+\lambda I)^{-1}X'u\|_2^2 \mid X\right) \\
& = \lambda^2\beta'(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}\beta + E\left(u'X(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}X'u \mid X\right) \\
& = \lambda^2\beta'(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}\beta + E\left(\operatorname{trace}\left[(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}X'uu'X\right] \mid X\right) \\
& = \lambda^2\beta'(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}\beta + \operatorname{trace}\left[(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}\right].
\end{aligned}
$$

Notice that

$$
\operatorname{trace}\left[(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}\right] = \sum_{i=1}^{J} q_i^2(q_i+\lambda)^{-2} \gtrsim \frac{T^2 J}{(T+\lambda)^2}
$$

and

$$
\lambda^2\beta'(\hat{\Omega}+\lambda I)^{-1}\hat{\Omega}(\hat{\Omega}+\lambda I)^{-1}\beta \geq \lambda^2\|\beta\|_2^2 \min_{1\leq i\leq J} q_i(q_i+\lambda)^{-2} \gtrsim \frac{\lambda^2 T}{(T+\lambda)^2}.
$$

Then we have

$$
E\left[T^{-1}\|X(\hat{\beta}-\beta)\|_2^2 \mid X\right] \gtrsim \frac{TJ+\lambda^2}{(T+\lambda)^2} \gtrsim \frac{TJ}{(T+\lambda)^2} + \frac{\lambda^2}{(T+\lambda)^2}.
$$

Since $\lambda \asymp T$, we have $E(\|X(\hat{\beta}-\beta)\|_2^2) \gtrsim T$.

Step 2: show stability.

Similar to (H.49), we notice that the perturbed estimator would be

$$
\tilde{\beta} = \beta - \lambda(\tilde{\Omega}+\lambda I)^{-1}\beta + (\tilde{\Omega}+\lambda I)^{-1}\tilde{\mu}.
$$

Then we bound

$$
\begin{aligned}
X_t'(\hat{\beta}-\tilde{\beta}) &= \lambda X_t'\left[(\tilde{\Omega}+\lambda I)^{-1} - (\hat{\Omega}+\lambda I)^{-1}\right]\beta + X_t'\left[(\hat{\Omega}+\lambda I)^{-1}\hat{\mu} - (\tilde{\Omega}+\lambda I)^{-1}\tilde{\mu}\right] \\
&= \lambda X_t'\left[(\tilde{\Omega}+\lambda I)^{-1} - (\hat{\Omega}+\lambda I)^{-1}\right]\beta + X_t'\left[(\hat{\Omega}+\lambda I)^{-1}(\hat{\mu}-\tilde{\mu}) + \left((\hat{\Omega}+\lambda I)^{-1} - (\tilde{\Omega}+\lambda I)^{-1}\right)\tilde{\mu}\right] \\
&= X_t'\left[(\tilde{\Omega}+\lambda I)^{-1} - (\hat{\Omega}+\lambda I)^{-1}\right](\lambda\beta-\tilde{\mu}) + X_t'(\hat{\Omega}+\lambda I)^{-1}(\hat{\mu}-\tilde{\mu}).
\end{aligned}
$$

Now we notice that

$$
\|(\tilde{\Omega}+\lambda I)^{-1} - (\hat{\Omega}+\lambda I)^{-1}\| = \left\|(\tilde{\Omega}+\lambda I)^{-1}\left[(\hat{\Omega}+\lambda I) - (\tilde{\Omega}+\lambda I)\right](\hat{\Omega}+\lambda I)^{-1}\right\|
$$

$$= \|(\tilde{\Omega} + \lambda I)^{-1}(\hat{\Omega} - \tilde{\Omega})(\hat{\Omega} + \lambda I)^{-1}\| = O_P\left(\frac{(J + c_T)}{(T + \lambda)^2}\right)$$

and $\|\lambda\beta - \tilde{\mu}\|_2 \leq \lambda + \|\tilde{\mu}\|_2 = O_P(\lambda + \sqrt{JT})$. This means that

$$\left|X_t'\left[(\tilde{\Omega} + \lambda I)^{-1} - (\hat{\Omega} + \lambda I)^{-1}\right](\lambda\beta - \tilde{\mu})\right| = O_P\left(\frac{(J + c_T)}{(T + \lambda)^2}\left(\lambda + \sqrt{JT}\right)\sqrt{J}\right).$$

Moreover, we also have

$$\left|X_t'(\hat{\Omega} + \lambda I)^{-1}(\hat{\mu} - \tilde{\mu})\right| \leq O_P\left(\sqrt{J}\frac{1}{\mu + \lambda}\sqrt{Jc_T}\right) = O_P\left(\frac{J\sqrt{c_T}}{T + \lambda}\right).$$

It follows that

$$\max_{1 \leq t \leq T} |X_t'(\hat{\beta} - \tilde{\beta})| = O_P\left(\frac{(J + c_T)}{(T + \lambda)^2}\left(\lambda + \sqrt{JT}\right)\sqrt{J} + \frac{J\sqrt{c_T}}{T + \lambda}\right).$$

Since $J \ll T$ and $T \gtrsim \lambda \gg J^{3/2}$, we have $\max_{1 \leq t \leq T} |X_t'(\hat{\beta} - \tilde{\beta})| = o_P(1)$.

I Tables and Figures Appendix

Table I.1: Size Properties: $F_{2t} \stackrel{iid}{\sim} N(0,1)$, $\rho_\epsilon = \rho_u = 0$

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.10	0.10	0.09	0.10	0.10	0.10	0.10	0.09	0.10
50	0.10	0.09	0.10	0.10	0.10	0.10	0.09	0.10	0.10
100	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.10	0.09	0.09	0.10	0.09	0.09	0.10	0.09	0.10
50	0.10	0.10	0.10	0.09	0.10	0.10	0.09	0.10	0.10
100	0.11	0.10	0.10	0.11	0.10	0.10	0.11	0.10	0.10
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.10	0.09	0.09	0.09	0.10	0.09	0.11	0.09	0.09
50	0.10	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10
100	0.10	0.09	0.10	0.10	0.09	0.10	0.10	0.10	0.09
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.10	0.10	0.10	0.10	0.09	0.10	0.10	0.10	0.10
50	0.11	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
100	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.11

Notes: Simulation design as described in the main text. Nominal level $\alpha = 0.1$. Based on simulations with 5000 repetitions.

Table I.2: Size Properties: $F_{2t} \overset{iid}{\sim} N(0,1)$, $\rho_\epsilon = \rho_u = 0.6$

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.13	0.13	0.12	0.12	0.11	0.10	0.12	0.11	0.11
50	0.11	0.11	0.10	0.12	0.12	0.12	0.12	0.12	0.12
100	0.11	0.10	0.10	0.11	0.11	0.12	0.12	0.12	0.11
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.12	0.12	0.11	0.12	0.11	0.11	0.12	0.11	0.11
50	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.11	0.12
100	0.10	0.10	0.11	0.11	0.11	0.12	0.11	0.11	0.12
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.10	0.10	0.10	0.10	0.09	0.09	0.11	0.10	0.10
50	0.10	0.10	0.10	0.10	0.10	0.10	0.12	0.12	0.12
100	0.10	0.10	0.10	0.10	0.10	0.10	0.12	0.13	0.12
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$	$J=20$	$J=50$	$J=100$
20	0.11	0.11	0.10	0.11	0.10	0.10	0.12	0.11	0.12
50	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12
100	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11

Notes: Simulation design as described in the main text. Nominal level $\alpha = 0.1$. Based on simulations with 5000 repetitions.

Table I.3: Size Properties: $F_{2t} \sim N(t, 1)$, $\rho_\epsilon = \rho_u = 0$

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.10	0.10	0.10	0.08	0.09	0.09	0.07	0.08	0.09
50	0.11	0.09	0.10	0.09	0.09	0.09	0.09	0.08	0.09
100	0.10	0.10	0.10	0.10	0.09	0.09	0.10	0.10	0.09
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.42	0.43	0.43	0.09	0.10	0.09	0.08	0.08	0.08
50	0.71	0.75	0.76	0.10	0.10	0.10	0.10	0.09	0.09
100	0.94	0.96	0.96	0.10	0.10	0.09	0.10	0.09	0.09
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.45	0.43	0.44	0.51	0.48	0.49	0.08	0.09	0.08
50	0.78	0.75	0.78	0.82	0.80	0.80	0.09	0.09	0.09
100	0.97	0.97	0.97	0.98	0.97	0.98	0.09	0.10	0.10
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.35	0.34	0.33	0.20	0.13	0.12	0.08	0.07	0.08
50	0.65	0.61	0.60	0.39	0.19	0.14	0.09	0.10	0.09
100	0.89	0.86	0.86	0.61	0.33	0.21	0.09	0.10	0.10

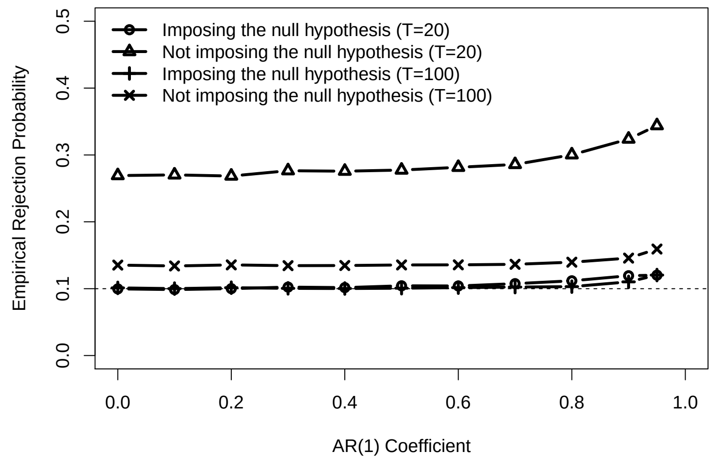
Notes: Simulation design as described in the main text. Nominal level $\alpha = 0.1$. Based on simulations with 5000 repetitions.

Table I.4: Size Properties: $F_{2t} \sim N(t, 1)$, $\rho_\epsilon = \rho_u = 0.6$

	DGP1								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.12	0.12	0.12	0.10	0.09	0.09	0.10	0.09	0.08
50	0.10	0.12	0.11	0.11	0.11	0.11	0.12	0.12	0.11
100	0.10	0.11	0.10	0.11	0.12	0.11	0.11	0.12	0.12
	DGP2								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.44	0.47	0.49	0.12	0.12	0.13	0.10	0.10	0.10
50	0.74	0.77	0.77	0.11	0.11	0.11	0.12	0.12	0.12
100	0.95	0.96	0.96	0.11	0.11	0.11	0.11	0.11	0.11
	DGP3								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.45	0.45	0.45	0.53	0.51	0.53	0.10	0.09	0.09
50	0.78	0.78	0.78	0.83	0.82	0.81	0.12	0.12	0.12
100	0.97	0.97	0.97	0.98	0.97	0.97	0.12	0.11	0.11
	DGP4								
	Diff-in-Diffs			Synthetic Control			Constrained Lasso		
T_0	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$	$J = 20$	$J = 50$	$J = 100$
20	0.38	0.37	0.36	0.21	0.14	0.12	0.10	0.10	0.10
50	0.66	0.63	0.63	0.39	0.20	0.14	0.12	0.12	0.11
100	0.89	0.86	0.86	0.62	0.34	0.20	0.10	0.11	0.11

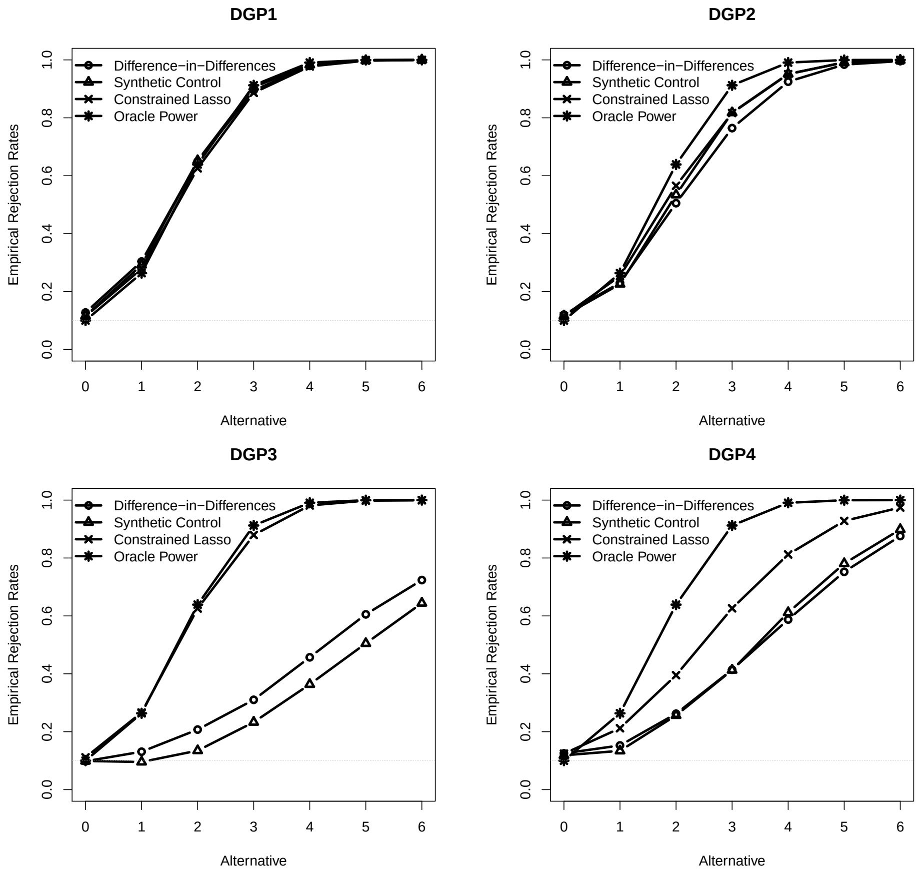
Notes: Simulation design as described in the main text. Nominal level $\alpha = 0.1$. Based on simulations with 5000 repetitions.

Figure I.1: Finite Sample Size Properties ($\alpha = 0.1$)



Notes: Empirical rejection probability from testing $H_0 : \theta_T = 0$. The data are generated as $Y_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N + u_t$, where $Y_{jt}^N \sim N(0,1)$ is iid across (j,t), $\{u_t\}$ is a Gaussian AR(1) process, $(w_2, \ldots, w_{J+1})' = (1/3, 1/3, 1/3, 0, \ldots, 0)'$, and $J = 50$. The weights are estimated using the canonical SC method (cf. Section 2.3.2).

Figure I.2: Power Curves



Notes: Simulation design as described in the main text. Nominal level $\alpha = 0.1$. Based on simulations with 5000 repetitions.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *The American Economic Review*, 93(1):113–132.

Andrews, D. W. (2003). End-of-sample instability tests. *Econometrica*, 71(6):1661–1694.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. Springer Science & Business Media.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Berbee, H. (1987). Convergence rates in the strong law for bounded mixing sequences. *Probability Theory and Related Fields*, 74(2):255–270.

Brockwell, P. J. and Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.

Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473.

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. arXiv:1912.07120.

Chen, X., Shao, Q.-M., Wu, W. B., and Xu, L. (2016). Self-normalized cramér-type moderate deviations under dependence. *The Annals of Statistics*, 44(4):1593–1617.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732–749. PMLR.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2019). Practical and robust t-test based inference for synthetic control and related methods. arXiv:1812.10820.

Cunningham, S. and Shah, M. (2018). Decriminalizing indoor prostitution: Implications for sexual violence and public health. *The Review of Economic Studies*, 85(3):1683–1715.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research.

Hahn, J. and Shi, R. (2017). Synthetic control and inference. *Econometrics*, 5(4):1–12.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23(2):169–192.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

McCarthy, C. A. (1967). C_p. *Israel Journal of Mathematics*, 5(4):249–271.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*. Springer.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798–2822.

Rotfeld, S. Y. (1969). The singular numbers of the sum of completely continuous operators. In *Spectral Theory*, pages 73–78. Springer.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge University Press.

White, H. (2014). *Asymptotic theory for econometricians*. Academic press.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540.