Counterfactual and Synthetic Control Method: Causal Inference with Instrumented Principal Component Analysis

Cong Wang[*]

March 31, 2024

Job Market Paper

Abstract

The fundamental problem of causal inference lies in the absence of counterfactuals. Traditional methodologies address this by implicitly or explicitly imputing the missing counterfactuals, relying on identification assumptions that are often untestable and too strong. Synthetic control methods (SCM) utilizes weighted average of control units to impute the missing counterfactual for the treated. Eventhough SCM relaxes some strict assumptions, it requires the treated unit to be inside the convex hull formulated by control units avoiding extrapolation. In recent advances, researchers modelling the entire data generating process (DGP) to impute the missing counterfactual explicitly. This paper expands the interactive fixed effect (IFE) model by integrating covariates into dynamic factor loadings adding additional robustness. This methodology offers multiple benefits: firstly, it incorporates the strengths of previous SCM approaches, such as the relaxation of the untestable parallel trends assumption (PTA). Secondly, it does not require the targeted outcomes inside the convex hull formulated by control units. Thirdly, it eliminates the need for correct model specification required by IFE model. Finally, it inherits the ability of principal component anlaysis (PCA) to effectively handle high dimensional data and enhances the value extracted from numerous covariates.

Keywords: Synthetic Control, Principal Component Analysis, Causal Inference

JEL Codes: G11, G12, G30

[*] Department of Economics and Law, Sapienza University of Rome

1 Introduction

In this paper, we introduce a novel counterfactual imputation method that combines the dimension reduction capabilities of principal component analysis (Jollife and Cadima (2016)) with the flexibility of the interactive fixed effect model (Bai and Perron (2003); Bai (2009)). This approach not only harnesses the strengths of PCA method in managing high dimensional data but also adapts the versatility of IFE method to accommodate a wide range of data generating processes. We name the newly proposed method the counterfactual and synthetic control method with instrumented principal component analysis (CSC-IPCA), aligning with the previous counterfactual and synthetic control method with interactive fixed effects (CSC-IFE)[1] proposed by Xu (2017). To model the entire data generating process, the CSC-IPCA estimator is designed to overcome the constraints of untestable and stringent assumptions such as unconfoundedness and common support for matching (Abadie and Imbens (2011, 2006)), parallel trend assumption for diffference-in-differences (DID) (Card and Krueger (1993)), and others. Furthermore, it addresses the limitation observed in the original synthetic control method (Abadie et al. (2010)) and its variants (Ben-Michael et al. (2021), Arkhangelsky et al. (2021)), which necessitate the outcomes of treated units to lie inside or not far from the convex hull formed by control units.

Causal inference in economics and other social sciences is frequently complicated by the absence of counterfactuals, which is essential for evaluating the impact of a treatment or policy intervention. Imbens and Rubin (2015) state that at some level, all methods for causal inference can be viewed as imputation methods, although some are more explicitly than others. For instance, under certain assumptions, the matching method (Abadie and Imbens (2006, 2011)) explicitly imputes the missing counterfactuals for treated units with meticulously selected control units. The DID method (Card and Krueger (1993); Ashenfelter

[1] The concept of counterfactual and synthetic control (CSC) is a generalization for a group of methodolies for causal inference proposed by Chernozhukov et al. (2021), the original name of this methodology is actually "generalized synthetic control method with interactive fixed effect model" named by Xu (2017).

(1978)), on the other hand, implicitly imputes the missing counterfactuals by differencing the control units before and after treatment. While the SCM method explicitly imputes the missing counterfactual with weighted average of control units. Our method aligns with the recent trend in causal inference literature, aiming to explicitly impute the missing counterfactual by modeling the entire data generating process, a strategy highlighted by Athey et al. (2021) with their matrix completion (MC) method, and Xu (2017) with their interactive fixed effect method.

Fixed effects models have long been explored in the econometrics literature related to modeling panel data, with significant contributions by Bai and Perron (2003), Pesaran (2006), Stock and Watson (2002), Eberhardt and Bond (2009), among others. However, within the context of causal inference, Hsiao et al. (2012) stands out as the first work proposing the use of this method specifically for predicting missing counterfactuals in sythetic control settings, followed by Gobillon and Magnac (2016); Xu (2017); Chan et al. (2016) and Li (2018). The CSC-IPCA estimator builds upon the instrumented principal component analysis first introduced by Kelly et al. (2020, 2019) in the context of predicting stock returns. It belongs to the fixed effect method family, while through the instrumented factor loadings we can estimate the time varying latent factors more accuately. Firstly, it assumes a simple fixed effect model as with Bai and Perron (2003) with only the interactive component between factor loadings Λ_i and time varying factors F_t:

$$Y_{it} = \Lambda_i F_t' + \mu_{it} \tag{1}$$

Secondly, it instruments the factor loadings Λ_i with covariates X_{it}, which allows us to capture the time varying properties of the factor loadings:

$$\Lambda_{it} = X_{it}\Gamma + H_{it} \tag{2}$$

The static factor loadings Λ_i in Equation 1 are assumed to be constant by most studies in the related literature. However, in many economic and social science contexts, the factor loadings are not fixed but fluctuate over time in response to relevant covariates. This adaptation is particularly realistic in many economic and social science contexts[2]. By instrumenting the factor loadings Λ_i with covariates X_{it} through Equation 2, we can capture the time varying properties of the factor loadings. The matrix Γ, serving as an $L \times K$ mapping function from covariates (number of L) to factor loadings (number of K), also acts as a dimension reduction operation, which aggregates all the information from the covariates into a smaller number of factors, making the model parsimonious.

The instrumented factor loadings enhance our estimation of the time varying latent factors F_t. Unlike the prediction problem addressed in Kelly et al. (2020, 2019), here we utilize only the control units to estimate F_t and then update Γ with the treated units before treatment. This represents the crucial distinction between our method and the IPCA method for prediction, where the mapping matrix Γ is assumed to be constant for all units across all periods.

This approach presents three major advantages over the CSC-IFE model and other fixed effect methods for causal inference. First, it eliminates the necessity for correct model specification, a crucial demand of the CSC-IFE method. In CSC-IFE method, aside from the interactive component $\lambda_i F_t$, the model also mandates that covariates be included in the functional form linearly as regressors $X_{it}\beta$. Second, in the CSC-IPCA method, the unit specific factor loadings Λ_{it} are instrumented by covariates, introducing time varying characteristics to the factor loadings. This adaptation is particularly realistic in many economic and social science contexts, where unit effects are not fixed, but fluctuating over time in response to covariates. Third, the CSC-IPCA method incorporates a dimension reduction operation by matrix Γ. This step is particularly beneficial for high-dimensional datasets with a large number of covariates. This feature makes it particularly valuable for financial data (Feng

[2] Considering a company that increases its investment in R&D, it transitions from a conservative stance to a more aggressive one. This can also impact its profitability, potentially moving it from a robust position to a weaker one. The unit effect, hence, changes along with its investment strategy.

et al. (2020)) and high dimensional macroeconomic time series data (Brave (2009)).

2 Framework

Consider Y_{it} as the observed outcome for a specific unit i ($i = 1, \ldots, N$) and time t ($t = 1, \ldots, T$). The total number of observed units from the panel is $N = N_{treat} + N_{ctrl}$, where N_{treat} indicates the number of units in the treatment group $\mathcal{T}$ and N_{ctrl} proxies the number of units in the control group $\mathcal{C}$. Each unit is observed over $T = T_{pre} + T_{post}$ periods, where T_{pre} is the number of periods before treatment and T_{post} is the number of periods post treatment. We observe the treatment effect at $T_{pre} + 1$ right after the beginning of the treatment and continue to observe there after till the end the observing periods, a scenario commonly referred to as staggered adoption. Following Equation 1, 2, we assume that the outcome variable Y_{it} is given by a simple interactive fixed effect model.

Assumption 1 *Functional form:*

$$\begin{aligned} Y_{it} &= D_{it} \circ \delta_{it} + \Lambda_{it} F_t' + \mu_{it}, \\ \Lambda_{it} &= X_{it}\Gamma + H_{it} \end{aligned} \tag{3}$$

The primary distinction of this functional form from existing interactive fixed effect models (Gobillon and Magnac (2016); Chan et al. (2016)) is that the factor loading Λ_{it} is instrumented by observed covariates X_{it}, which makes the conventionally static factor loadings exhibit time varying features. Specifically, $F_t = [f_t^1, \ldots, f_t^K]$ is vector of unobserved common factors, $\Lambda_{it} = [\lambda_{it}^1, \ldots, \lambda_{it}^K]$ represents a vector of factor loadings associated with K common factors. Meanwhile, the vector $X_{it} = [x_{it}^1, \ldots, x_{it}^L]$ comprises observed covariates, with L indicating the total number of covariates. The transformation matrix Γ, which is of size $L \times K$, maps the information from observed covariates X_{it} to factor loadings Λ_{it}. This integration permits Λ_{it} to exhibit variability across time and units, thereby introducing an additional layer of heterogeneity into the model. Another key difference from the CSC-IFE approach by Xu (2017) is that we retain only the interacted component $\Lambda_{it}F_t$ between factors and factor loadings, the linear part of covariates $X_{it}\beta$ is not included in the function form[3]. The logic behind is that we believe that the unit specific factor loadings, instrumented by covariates, have included all the predictive information from these covariates. This functional form exhibit two more advantages over the CSC-IFE model. Firstly, it does not require the correct model specification, a crucial demand of the CSC-IFE method. Secondly, it incorporates a dimension reduction operation by matrix Γ, which allows us to handle high dimensional datasets, especially when dealing with a large number of covariates.

The remainder of the model adheres to conventional standards, where D_{it} denotes a binary treatment indicator, and δ_{it} represents the treatment effect, which varies across units and over time. For computational simplicity, we assume $D_{it} = 1$ for a unit i in the set $\mathcal{T}$ and for times $t > T_{pre}$, with all other D_{it} set to 0. This setup typifies a standard staggered adoption scenario. The model easily accommodates variations in treatment timing by removing the constraint that treatment must commence precisely at $t = T_{pre}+1$. The term μ_{it} signifies the idiosyncratic error associated with the outcome variable Y_{it}. Additionally, $H_{it} = [\eta^1_{it}, \ldots, \eta^L_{it}]$ constitutes the vector of error terms linked to L observed covariates.

Following Neyman (1932) potential outcome framework (also discussed by Rubin (1974, 2005)), we observe the actual outcome for the treated and control:

$$\begin{cases} Y^1_{it} = \delta_{it} + X_{it}\Gamma F'_t + \epsilon_{it} & if\ i \in \mathcal{T}\ \&\ t > T_{pre} \\ Y^0_{it} = X_{it}\Gamma F'_t + \epsilon_{it} & otherwise. \end{cases} \tag{4}$$

where Equation 4 represents the actual outcome for the treated and control units combined the two parts together in Equation 3. Our goal is to impute the missing counterfactual $\hat{Y}^0_{it} = X_{it}\hat{\Gamma}\hat{F}_t$ for the treated units $i \in \mathcal{T}$ when $t > T_{pre}$, where the $\hat{\Gamma}$ and $\hat{F}_t$ are estimated parameters. We then calculate the treatment effect for the treated (ATT) as the difference between the actual outcome and imputed missing counterfactuals, which is defined as:

[3] The functional form of the CSC-IFE is $Y_{it} = D_{it} \circ \delta_{it} + \Lambda_i F_t + X_{it}\beta + \mu_{it}$

$$\widehat{ATT}_t = \frac{1}{N_{treat}} \sum_{i \in \mathcal{T}} \left(Y_{it}^1 - \hat{Y}_{it}^0 \right) = \frac{1}{N_{treat}} \sum_{i \in \mathcal{T}} \hat{\delta}_{it}. \tag{5}$$

2.1 Assumptions for identification

To ensure the identification of the treatment effect, we introduce a series of assumptions about the data generating process, in addition to assumptions regarding the functional form. These assumptions are crucial for achieving consistent estimation of the treatment effect.

Assumption 2 *Assumptions for consistency:*

(1) Covariate orthogonality.

(2) Bounded second moment.

(3) Compact parameter space for Γ.

(4) Bounded covariates and their variance.

The assumptions outlined above serve as regularity conditions necessary for ensuring consistency. The first condition reflects the exclusion restriction commonly associated with instrumental variable regression, indicating that the instruments (in this case, covariates) are orthogonal to the error terms. This orthogonality ensures that the covariates do not correlate with any unobserved factors influencing the outcome. Similarly, the functional form described in Equation 3 aligns with the framework utilized in two-stage instrumental regression. The second condition imposes a limitation on the second moment of a series of random variables, ensuring that their variances remain finite and do not escalate indefinitely. This is a crucial stipulation for maintaining statistical control over the model's error terms.

The third condition mandates that the parameter space for the mapping matrix Γ be compact, thereby circumventing issues related to rank deficiency. This compactness is vital for the estimability of Γ and the existence of its inverse, mirroring standard factor analysis assumptions. The fourth condition, common in econometric analyses, requires that the covariates X_{it} are almost surely bounded, meaning their values do not tend toward infinity with probability one. Moreover, this condition stipulates that the variance of these covariates must also be bounded away from zero to ensure that the matrix $\Gamma' X_t' X_t \Gamma$ remains nonsingular. For detailed explanations of each condition, see Appendix Assumption A.2, which presents comprehensive details for the assumption conditions.

Assumption 3 *Assumptions for asymptotic normality.*

(1) Panel-wise and cross-sectional central limit theorem.

(2) Bounded dependence for covariates and errors.

(3) Cross-sectional homoskedasticity of covariates.

To derive the asymptotic properties of the CSC-IPCA estimator, we introduce additional assumptions. Assumption 3 encompasses panel-wise and cross-sectional central limit theorems for various variables, which are fulfilled by diverse mixing processes. These conditions are pivotal for determining the asymptotic distribution of factor and mapping matrix estimations. For an in-depth discussion, we refer to Kelly et al. (2020). The requirement of bounded dependence stipulates that both time series and cross-sectional dependencies between covariates and errors are bounded, a crucial step for establishing asymptotic normality. Meanwhile, the assumption of cross-sectional homoskedasticity simplifies the expressions for the asymptotic variances of the estimated mapping matrix Γ. It should be noted that relaxing this assumption would not alter the convergence rate. Detailed explanations of each assumption's conditions are provided in Appendix Assumption A.3.

The requirement of bounded dependence stipulates that both time series and cross-sectional dependencies of $X_{it}\epsilon_{it}$ are bounded, a crucial step for establishing asymptotic normality. Meanwhile, the assumption of cross-sectional homoskedasticity simplifies the expressions for the asymptotic variances of the estimated mapping matrix Γ. It should be noted that relaxing this assumption would not alter the convergence rate. Detailed explanations of each assumption's conditions are provided in Appendix Assumption A.3.

3 Estimation

The CSC-IPCA estimator of the treatment effect for a treated unit i at time t is defined as the difference between the observed outcome and its estimated counterfactual: $\delta_{it} = Y_{it}^1 - \hat{Y}_{it}^0$. To combine the functional form in Equation 3, we get the structural component of the potential outcome $Y_{it} = (X_{it}\Gamma)F_t'$. The CSC-IPCA method is estimated by minimizing the sum of squared residuals of the following objective function:

$$\arg\min_{\Gamma, F_t} \sum_i \sum_t \left(Y_{it} - (X_{it}\Gamma)F_t'\right)\left(Y_{it} - (X_{it}\Gamma)F_t'\right)'. \tag{6}$$

Unlike the IFE method (Bai (2009); Xu (2017)), our approach requires estimating only two parameters, Γ and F_t, simplifying the process. Different from principal component analysis (Jolliffe (2002); Stock and Watson (2002)), our method involves using covariates to instrument the factor loadings component. This necessitates the estimation of Γ rather than Λ_i, so we can not directly use eigenvalue decomposition. While the objective function in Equation 6 formulates the problem as minimizing a quadratic function with a single unknown variable (e.g., Γ) while holding the other variable (e.g., F_t) constant. This structure enables the application of the alternating least squares (ALS) method for a numerical solution. Generally, the imputation for the missing counterfactual $Y_{it}(0)$ is executed in three steps:

Step 1: The initial step entails estimating the time-varying factors $\hat{F}_t$ and the mapping matrix $\hat{\Gamma}_{\text{ctrl}}$ utilizing an ALS algorithm, based exclusively on data from the control group for the whole time period.

$$(\hat{\Gamma}_{ctrl}, \hat{F}_t) = \arg\min_{\Gamma, F_t} \sum_{i \in \mathcal{C}} \sum_{t \leq T} \left(Y_{it} - (X_{it}\Gamma)F_t'\right)\left(Y_{it} - (X_{it}\Gamma)F_t'\right)'. \tag{7}$$

Step 2: The subsequent step involves estimating the mapping matrix $\hat{\Gamma}_{treat}$ for treated unit i at time t, employing the previously estimated time-varying factors $\hat{F}_t$ and the observed covariates X_{it}, using only pretreatment data from the treated units.

$$\hat{\Gamma}_{treat} = \arg\min_{\Gamma} \sum_{i \in \mathcal{T}} \sum_{t \le T_{pre}} \left(Y_{it} - (X_{it}\Gamma)\hat{F}'_t \right) \left(Y_{it} - (X_{it}\Gamma)\hat{F}'_t \right)'. \tag{8}$$

Step 3: The third step includes normalizing the estimated mapping matrix $\hat{\Gamma}_{treat}$ and $\hat{F}_t$ by a set of constraints:

$$\begin{aligned} \Gamma_{norm} &= \hat{\Gamma}_{treat} R, \\ F_{norm} &= R^{-1}\hat{F}_t, \\ s.t. \Gamma'_{norm}\Gamma_{norm} &= \mathcal{I}, \quad F_{norm}F'_{norm}/T = \text{Diagonal}. \end{aligned} \tag{9}$$

Similar to most factor analysis methods, the estimated Γ and F_t are not deterministic. There exist infinite numbers of "rotated" parameters ΓR and $R^{-1}F_t$ that yeild the same objective function value (i.e. $X_{it}\Gamma F_t = X_{it}\Gamma R R^{-1} F_t$). To make the estimation identifiable, it's necessary to impose some constraints on the estimated Γ and F_t. Following Connor and Korajczyk (1993); Stock and Watson (2002); Bai and Ng (2002), the aforementioned restrictions reduce the model's complexity and make it easier to understand and interpret the relationships between the factors F_t and factor loadings Λ_{it}. Unlike the estimation methods in Bai (2009); Xu (2017) where normalization constraints are set before the estimation, the structural component $X_{it}\Gamma F_t$ allows us to normlize the estimated Γ and F_t after the estimation. This enables us to easily find the rotation matrix R that satisfies the above constraints.

Step 4: The final step involves imputing the counterfactual outcome $\hat{Y}^0_{it}$ for treated unit i at time t by substituting the estimated mapping matrix $\hat{\Gamma}_{norm}$ and the time varying factors $\hat{F}_{norm}$ into the following equation:

$$\hat{Y}_{it}(0) = (X_{it}\hat{\Gamma}_{norm})\hat{F}'_{norm}, \quad \forall i \in \mathcal{T}, \quad \& \quad T_{pre} < t \leq T. \tag{10}$$

The main difference between CSC-IPCA and the instrumented principal component analysis (IPCA) as proposed by Kelly et al. (2020) lies in the purpose of prediction. In the IPCA method, the authors predict the next period stock returns using all covariates from the preceding period, under the assumption that the mapping matrix Γ remains constant across all observations. In contrast, CSC-IPCA introduces a pivotal distinction: it operates under the assumption that treated and control groups are characterized by unique mapping matrices, Γ_{treat} and Γ_{ctrl}. This assumption is vital for the unbiased estimation of the ATT, setting the CSC-IPCA method apart by directly addressing heterogeneity in the treatment effect through the specification of group-specific mapping matrices. The detailed estimation procedures are presented in the Appendix A.2.

3.1 Hyper parameter tuning

Similar to CSC-IFE methods, researchers often encounter the challenge of selecting the appropriate number of factors, K, without prior knowledge of the true data generating process. To facilitate this selection, we introduce data driven approaches for determining the hyperparameter K. Utilizing both control and treated units as training and validation data respectively offers a practical solution. To enhance the robustness of this process, we propose two validation methods for hyperparameter tuning. Algorithm 1 describes a bootstrap method to ascertain K. This approach involves repeatedly sampling N_{ctrl} control units for training data and N_{treat} treated units for validation data, both with replacement. The optimal K is then determined by minimizing the average of sum squared errors across these validations. We also propose a leave-one-out cross validation method, as detailed in Appendix A.3.

Algorithm 1: Bootstrap Hyperparameter Tuning

```
Data: $Y, X$
Result: Optimal hyperparameter $k$
Determine the maximum possible hyperparameter $K$ and the number of repetitions $N$;
Initialize an array $MSE$ to store the average of sum squared error for each $k$;
for $k \leftarrow 1$ to $K$ do
    Initialize sum of squared errors: $SSE_k \leftarrow 0$;
    for $n \leftarrow 1$ to $N$ do
        Construct a bootstrap training dataset $(Y^b_{ctrl}, X^b_{ctrl})$ by sampling $N_{ctrl}$ control observations with replacement;
        Construct a bootstrap validation dataset $(Y^b_{treat}, X^b_{treat})$ by sampling $N_{treat}$ treated observations with replacement;
        Estimate parameters $\Gamma$ and $F_t$ using the training data via the ALS method;
        Use the estimated $\hat{\Gamma}$ and $\hat{F}_t$ to predict $\hat{Y}^b_{treat}$ with the validation data;
        Compute the sum of squared error for the validation data:
          $SE_n \leftarrow \sum \left( Y^b_{treat} - \hat{Y}^b_{treat} \right)^2$;
        Accumulate the sum of squared errors: $SSE_k \leftarrow SSE_k + SE_n$;
    end
    Calculate the average sum squared error for $k$: $MSE[k] \leftarrow \frac{SSE_k}{N}$;
end
Select $k$ corresponding to the minimum value in $MSE$;
```

3.2 Inference

In the context of causal inference, the pioneering application of the interactive fixed effect model is attributed to Hsiao et al. (2012). Nonetheless, a formal framework for inference using this method was not established until the works of Chan et al. (2016) and Li (2018). These methods of inference rely on a large number of control units and pre treatment periods to develop asymptotic properties. In recent years, conformal inference (Chernozhukov et al. (2021)) has gained popularity in causal inference literature, such as Ben-Michael et al. (2021), Roth et al. (2023), and Imbens (2024). Conformal inference is a nonparametric method that provides and exact and robust inference without requiring the specification of a model. Our causal inference framework, designed for predicting missing counterfactuals, allows us to construct inference procedures based on conformal prediction introduced by Shafer and

Vovk (2008) for the robustness against misspecification. The causal effect is identified as the difference between the observed outcomes and these estimated counterfactuals, expressed mathematically as:

$$\theta_{it} = Y_{it} - \hat{Y}_{it}, \quad \forall i \in N_{tr} \ \& \ \forall t \in T_{post},$$

where θ_{it} denotes the treatment effect for unit i at time t, and $\hat{Y}_{it}$ represents the imputed counterfactual outcome.

To conduct conformal inference, we first postulate a sharp null hypothesis, $H_0 : \theta_{it} = \theta_{it}^0$. Under this null hypothesis, we adjust the outcome for treated units post treatment as $\tilde{Y}_{it} = Y_{it} - \theta_{it}$. We then replace the original dataset with this adjusted part, $\tilde{Y}_{it}$. Secondly, following the estimation procedures described in Section 3 to estimate the time varying factor F with only control data as before, and update the Γ for the newly adjusted treated units with the entire set of treated units[4]. The concept revolves around updating the Γ using all the treated units, under the assumed null hypothesis, to minimize the occurrence of large residuals after intervention.

Thirdly, we estimate the treatment effect and compute the residuals for the treated units in the post treatment period. The test statistic showing how large the residual is under the null:

$$S(\hat{\mu}) = \left(\frac{1}{\sqrt{T_*}} \sum_{t \in T_1} |\hat{\mu}|^q \right) \tag{11}$$

Where $\hat{\mu}$ represents the residual for the treated units in the post-treatment periods, we employ $q = 1$ for the permanent intervention effect as designed in our study. A high value of the test statistic indicates a poor post treatment fit, suggesting that the treatment effect postulated by the null is unlikely to be observed, hence leading to the null's rejection.

Finally, we block permute the residuals and calculate the test statistic in each permuta-

[4] As a clarification, in the estimation section, we update Γ using only the treated units before treatment. However, for inference, we use the entire set of treated units to update Γ.

tion. The P-value is defined as:

$$\hat{p} = 1 - \hat{F}(S(\hat{u})), \text{ where } \hat{F}(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} 1\{S(\hat{u}_\pi) < x\}. \tag{12}$$

where Π represents the set of all block permutations, the test statistic for each permutation is denoted by $S(\hat{\mu}_\pi)$, , with x being the test statistic calculated from the unpermuted residuals. By employing different sets of nulls, we can compute a confidence interval at a specified confidence level.

4 Monte Carlo Simulation

In this section, we employ Monte Carlo simulations to assess the performance of the CSC-IPCA estimator in finite sample settings. We juxtapose the CSC-IPCA estimator against the CSC-IFE and the original SC estimators. Our comparative analysis focuses on key metrics including bias, mean squared errors, and converage properties.

We initiate our analysis with a data generating process that incorporates $L = 10$ and $K = 3$ time-varying covariates and common factors, along with unit and time fixed effects:

$$Y_{it} = D_i\delta_t' + X_{it}\beta' + (X_{it}\Gamma)F_t' + \alpha_i + \xi_t + \epsilon_{it}. \tag{6}$$

where $X_{it} = [x_{it}^1, \ldots, x_{it}^L]$ denotes a vector of $L \times 1$ time-varying covariates, which follows a VAR(1) process. $X_{it} = \mu_i + A_i X_{i,t-1} + \nu_{it}$, where A_i is a $L \times L$ variance-covariance matrix[5], The drift term μ_i equals 0 for control units and 2 for treated units,[6], and ν_{it} is a $L \times 1$ vector of i.i.d. standard normal errors. While $F_t = [f_t^1, \ldots, f_t^3]$ denotes the vector of time-

[5] In our methodology, the variance-covariance matrix is not constrained to be diagonal, thus allowing covariates within each unit to be correlated, reflecting the typical scenario in most economic time series data. To emphasize the independence among different units, we generate N unique variance-covariance matrices, each corresponding to a unit, ensuring cross-sectional independence and preserving time-series correlation. Moreover, we impose a condition on these matrices by requiring the eigenvalues of A_i to have characteristic roots that reside inside the unit circle, thereby assuring the stationarity of the VAR(1) process.

[6] This configuration underscores that the treatment assignment is not random; rather, it depends on the covariates X_{it}.

varying common factors, adhering to a similar VAR(1) process, the variable ϵ_{it} represents the idiosyncratic error term. Unit and time fixed effects, α_i and ξ_t respectively, are uniformly drawn from the interval $(0, 1)$. The coefficient vector $\beta = [\beta^1, \ldots, \beta^L]$ associated with the covariates is drawn uniformly from $(0, 1)$, and Γ, the $L \times K$ mapping matrix for the factor loadings, is drawn uniformly from $(-0.1, 0.1)$. The treatment indicator D_{it} is binary, defined as $D_{it} = 1$ for treated units during post-treatment periods, and $D_{it} = 0$ otherwise. The heterogeneous treatment effect is modeled as $\delta_{it} = \bar{\delta}_{it} + e_{it}$, where e_{it} is i.i.d as standard normal, and $\bar{\delta}_t = [0, \cdots, 0, 1, 2, \ldots, T_{post}]$ represents a time-varying treatment effect[7]. Only the outcome Y_{it}, the covariates X_{it}, and the treatment indicator D_{it} are observed, while all other variables remain unobserved.

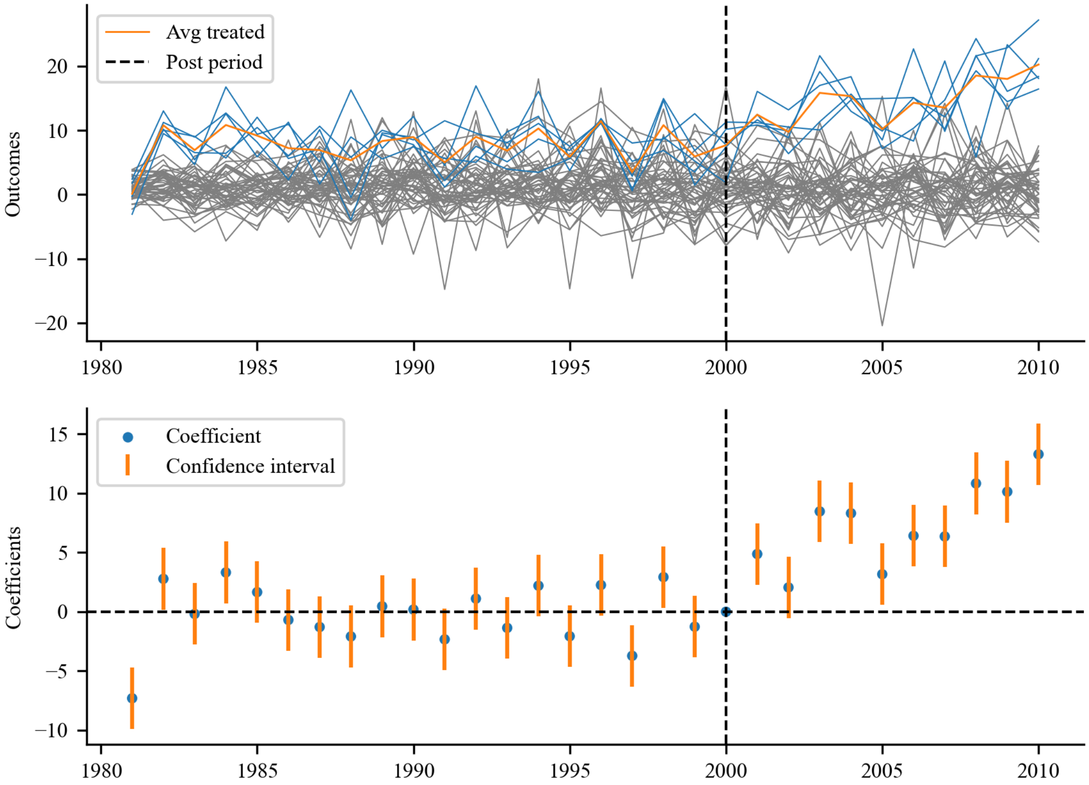
Figure 1 represents the simulated data following our data generating process. Observations from the upper panel indicate that the parallel trend assumption is not met. To verify this, we plot a simple event study, clearly revealing a failure in the parallel trend assumption. Furthermore, outcomes for treated units are marginally higher than for control units. In such cases, the synthetic control method will be biased, as it avoids extrapolation and typically fit poorly for treated units.

4.1 A simulated example

Following this data generating process, figure 2 illustrates both the raw data and the imputed counterfactual outcomes as estimated by the CSC-IPCA method. In the upper panel, control units are represented in gray and treated units in light blue, with the average outcome for treated units highlighted in orange. The imputed synthetic average for treated outcomes is also shown, delineated by an orange dashed line. The CSC-IPCA method is capable of capturing the trajectory of the average outcome for treated units before treatement. The lower panel of Figure 2 shows the estimated ATT (dashed line) with the true ATT (solid line). The CSC-IPCA method is able to capture the true ATT, as evidenced by the close

[7] Here we simplify the treatment effect to be constant across units, however the heterogeneous treatment effect across units can also be easily employed.

Figure 1: **CSC-IPCA Data Generating Process**



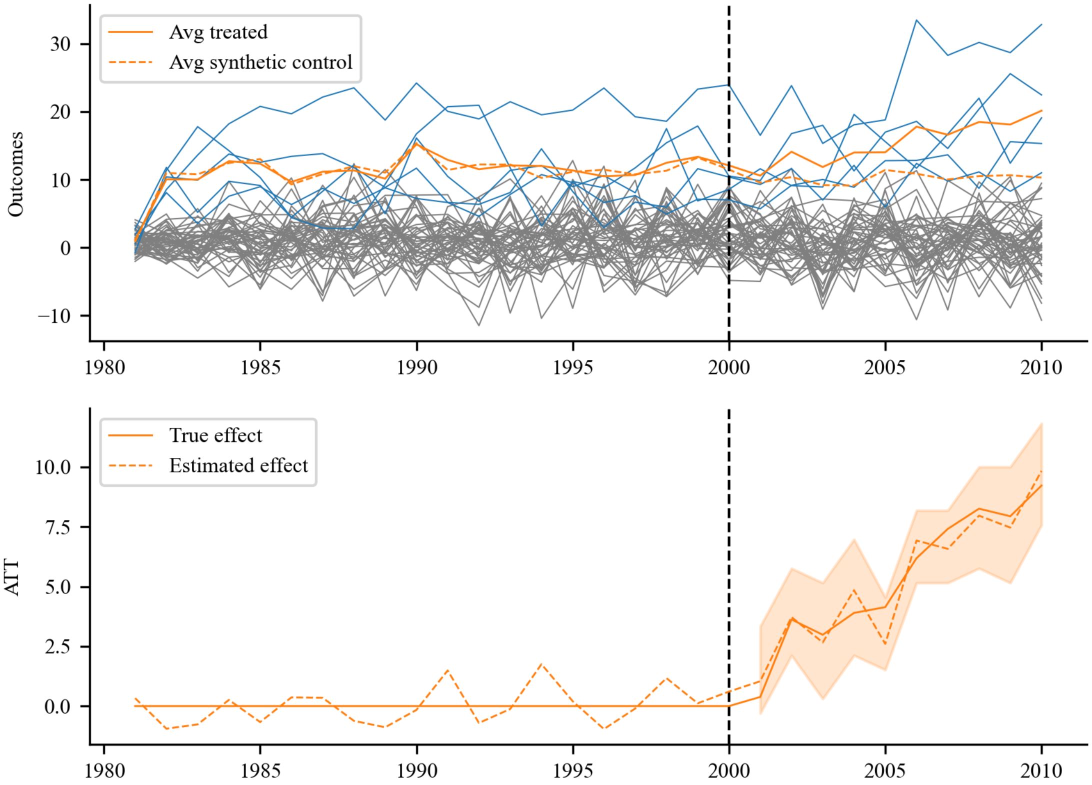
In this graphic, the upper panel plots simulated data following the above data generating process. The light blue lines represent treated units and the light gray lines represent controls. Key parameters are$N_{treat} = 5, N_{ctrl} = 45, T_0 = 20, T_1 = 10, L = 10$. The lower panel plots a simple event study.

alignment between the dashed and solid lines.

4.2 Bias comparision

Based on the same data generating process and parameters, we compare the CSC-IPCA, CSC-IFE, and SCM estimators with 1000 simulations. Figure 3 illustrates the bias among these estimation methods. In panel 1, where all covariates are observed, both CSC-IPCA and CSC-IFE demonstrate unbiasedness and effectively estimate the true ATT. However,

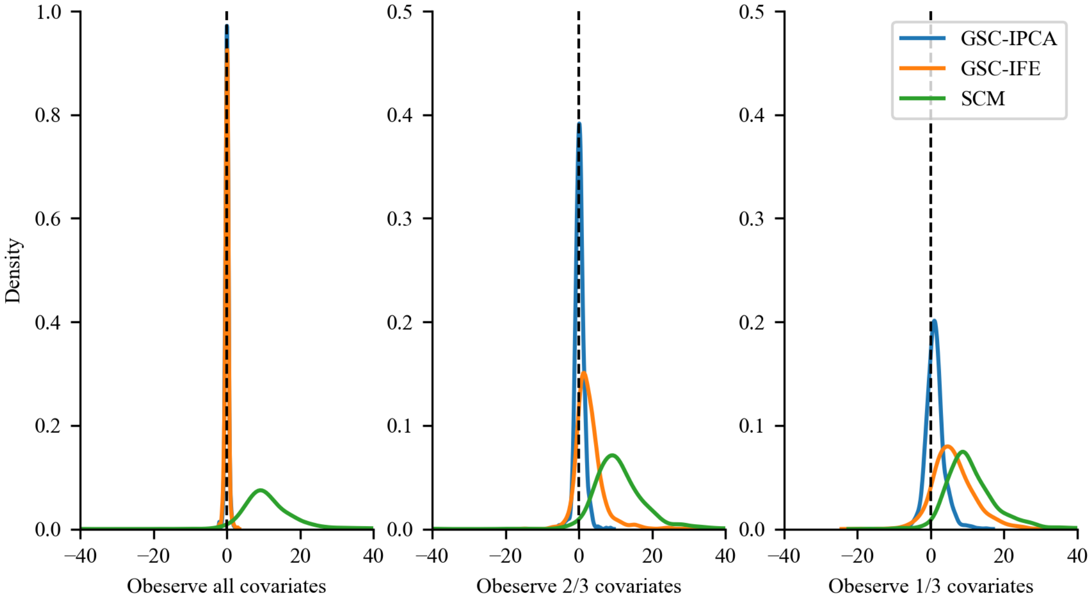
Figure 2: **CSC-IPCA Estimated ATT for Simulated Sample**



This graphic plots the CSC-IPCA method estimated ATT for simulated data $N_{treat} = 5, N_{ctrl} = 45, T_0 = 20, T_1 = 10, L = 10$.

due to the outcomes of treated units falling outside the convex hull of control units, the SCM exhibits an upward bias. This scenario is common in empirical studies where only a limited number of covariates are observed. As depicted in Figure 3, from left to right, we progressively observe all 10, then 6 (2/3), and finally 3 (1/3) covariates. With an increase in the number of unobserved covariates, both CSC-IPCA and CSC-IFE lose efficiency; however, the CSC-IPCA estimator remains unbiased.

Figure 3: **Bias Comparing with Other Methods**



This graphic plots the CSC-IPCA method estimated ATT for simulated data $N_{treat} = 5, N_{ctrl} = 45, T_0 = 20, T_1 = 10, L = 10$.

4.3 Finate sample properties

We present the Monte Carlo simulation results in Table 1 to investigate the finate sample properties of the CSC-IPCA estimator. The number of treated units and post treatment period are fixed to $N_{treat} = 5, T_{post} = 5$. We vary the number of control units N_{ctrl}, pre treatment period T_{pre} , and the proportion of observed covariates α with the total number of covariates $L = 10$ to investigate the finate sample properties. As showing in the table 1, the bias, RMSE, and STD are estimated based on 1000 simulations[8]. The results indicate that the bias of the CSC-IPCA estimator decreases as the number of control units and pre treatment period increases. The bias decreases the most when the proportion of observed covariates increases from $\frac{1}{3}$ to 1 (all covariates are observed). We observe similar pattern in

[8] The root mean squared error (RMSE) is defined as $RMSE = \sqrt{\frac{1}{T_1}\sum_{t\in T_1}\left(ATT_t - \widehat{ATT}_t\right)^2}$. The standard deviation (STD) is defined as $STD = \frac{1}{T_1}\sum_{t\in T_1}\left(\widehat{ATT}_t - \frac{1}{T_1}\sum_{t\in T_1}\widehat{ATT}_t\right)^2$

RMSE and STD. It is worth noting that if we observe all the covariates (i.e., $\alpha = 1$), the bias, RMSE, and STD of the CSC-IPCA estimator are all reduce to the lowest level even with a small number of control units and pre treatment period.

Table 1: **Finate Sample Properties**

α		$\frac{1}{3}$	$\frac{2}{3}$	1	$\frac{1}{3}$	$\frac{2}{3}$	1	$\frac{1}{3}$	$\frac{2}{3}$	1
T_0	N_{co}	Bias			RMSE			STD		
10	10	2.382	0.747	0.189	4.619	3.011	1.712	3.975	2.943	1.732
10	20	1.452	0.420	0.063	3.538	2.180	0.984	3.273	2.186	1.076
10	40	0.920	0.222	0.008	2.747	1.745	0.789	2.650	1.786	0.917
20	10	2.534	1.121	0.237	4.441	3.015	1.192	3.688	2.829	1.271
20	20	1.520	0.421	0.048	3.276	1.840	0.872	2.946	1.849	0.977
20	40	1.008	0.258	0.036	2.632	1.451	0.539	2.498	1.505	0.705
40	10	2.746	1.148	0.227	4.982	2.863	1.167	4.166	2.665	1.201
40	20	1.733	0.540	0.089	3.964	1.783	0.732	3.607	1.757	0.874
40	40	0.807	0.281	0.044	2.530	1.632	0.531	2.457	1.654	0.677

This table presents the finate sample properties of the CSC-IPCA method estimated ATT for simulated data. The number of treated units and post treatment period are fixed to $N_{treat} = 5, T_1 = 5$. We vary the number of control units N_{co}, pre treatment period T_0, and proportion of observed covariates α to investigate the finate sample properties, the total number of covariates is $L = 10$. The bias, RMSE, and STD are estimated based on 1000 simulations.

5 Empirical Application

In this section, we study the CSC-IPCA method with an empirical example. We apply the CSC-IPCA method to estimate the treatment effect of the Job Corps program on the earnings of participants. The Job Corps program is a federally funded education and vocational training program for disadvantaged youth in the United States. The program provides free education and vocational training to young people aged 16 to 24, with the aim of improving their employment prospects. The program has been evaluated in several studies. The data used in this study is from the National Job Corps Study (NJCS), which is a large-scale randomized controlled trial conducted in the 1990s. The NJCS data includes a sample of 16,000 young people who were randomly assigned to either the Job Corps program or a control group. The data includes information on the participants' earnings, education, and employment history. The data also includes a rich set of covari

6 Conclusion

Firms'

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. econometrica, 74(1):235–267, 2006.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics, 29(1):1–11, 2011.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. Journal of the American statistical Association, 105(490):493–505, 2010.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. American Economic Review, 111(12):4088–4118, 2021.

Orley Ashenfelter. Estimating the effect of training programs on earnings. The Review of Economics and Statistics, pages 47–57, 1978.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Journal of the American Statistical Association, 116(536):1716–1730, 2021.

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

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. Econometrica, 70(1):191–221, 2002.

Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. Journal of applied econometrics, 18(1):1–22, 2003.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. Journal of the American Statistical Association, 116(536):1789–1803, 2021.

Scott Brave. The chicago fed national activity index and business cycles. Chicago Fed Letter, (Nov), 2009.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, 1993.

Marc Chan, Simon Kwok, et al. Policy evaluation with interactive fixed effects. Preprint. Available at https://ideas. repec. org/p/syd/wpaper/2016-11. html, 2016.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. Journal of the American Statistical Association, 116(536):1849–1864, 2021.

Gregory Connor and Robert A Korajczyk. A test for the number of factors in an approximate factor model. the Journal of Finance, 48(4):1263–1291, 1993.

Markus Eberhardt and Stephen Bond. Cross-section dependence in nonstationary panel models: a novel estimator. 2009.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3):1327–1370, 2020.

Laurent Gobillon and Thierry Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. Review of Economics and Statistics, 98(3):535–551, 2016.

Nicholas J Higham. Cholesky factorization. Wiley interdisciplinary reviews: computational statistics, 1(2):251–254, 2009.

Cheng Hsiao, H Steve Ching, and Shui Ki Wan. A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. Journal of Applied Econometrics, 27(5):705–740, 2012.

Guido W Imbens. Causal inference in the social sciences. Annual Review of Statistics and Its Application, 11, 2024.

Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

Ian T Jollife and Jorge Cadima. Principal component analysis: A review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci, 374(2065):20150202, 2016.

Ian T Jolliffe. Principal component analysis for special types of data. Springer, 2002.

Bryan T Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics, 134(3):501–524, 2019.

Bryan T Kelly, Seth Pruitt, and Yinan Su. Instrumented principal component analysis. Available at SSRN 2983919, 2020.

Kathleen Li. Inference for factor model based average treatment effects. Available at SSRN 3112775, 2018.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science, pages 465–472, 1932.

M Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica, 74(4):967–1012, 2006.

Jonathan Roth, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. Journal of Econometrics, 235(2):2218–2244, 2023.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100(469):322–331, 2005.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. Journal of Machine Learning Research, 9(3), 2008.

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. Journal of the American statistical association, 97(460):1167–1179, 2002.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. Political Analysis, 25(1):57–76, 2017.

Appendix A

Within this section, we use lowercase letters to represent scalars, e.g., y_{it} would be a scalar of the outcome variable for unit i at time t. We use bold lowercase letters to represent vectors, e.g., $\mathbf{x}_{it}$ would be a vector of covariates for unit i at time t. We use uppercase letters to represent matrices, e.g., Γ represents the mapping matrix. We denote $\mathrm{E}\|\mathbf{f}_t\mathbf{f}_t'\|^2$ the Frobenius norm squared of the matrix $\mathbf{f}_t\mathbf{f}_t'$.

A.1 Identification assumptions

Assumption A.2 *Assumption for consistency:*

(1) Instrument orthogonality: $E\left[\boldsymbol{x}_{it}\epsilon_{it}\right] = \boldsymbol{0}_{L\times 1}$,

(2) The following moments exsit: $\mathrm{E}\|\mathbf{f}_t\mathbf{f}_t\|^2$, $\mathrm{E}\|\mathbf{x}_{it}\epsilon_{it}\|^2$, $\mathrm{E}\|\mathbf{x}_{it}'\mathbf{x}_{it}\|^2$, $\mathrm{E}\left[\|\mathbf{x}_{it}'\mathbf{x}_{it}\|^2\|\mathbf{f}_t\mathbf{f}_t\|^2\right]$,

(3) The parameter space Ψ *of* Γ *is compact and away from rank deficient:* $\det \Gamma'\Gamma > \epsilon$ *for some* $\epsilon > 0$,

(4) Almost surely, $\mathbf{x}_{it}$ *is bounded, and define* $\Omega_t^{xx} := \mathrm{E}\left[\mathbf{x}_{it}'\mathbf{x}_{it}\right]$, *then almost surely,* $\Omega_t^{xx} > \epsilon$ *for some* $\epsilon > 0$.

Assumption A.3 *Assumptions for asympototic normality:*

(1) As $N, T \to \infty$, $\frac{1}{\sqrt{NT}}\sum_{i,t} vect\left(\mathbf{x}_{i,t}'\epsilon_{i,t}\mathbf{f}_t'\right) \xrightarrow{d} Normal\left(0, \Omega^{x\epsilon f}\right)$,

(2) As $N \to \infty$, $\frac{1}{\sqrt{N}}\sum_i vect\left(X_i'\epsilon_i\right) \xrightarrow{d} Normal\left(0, \Omega^{x\epsilon}\right)$ *for* $\forall t$,

(3) As $N, T \to \infty$, $\frac{1}{\sqrt{T}}\sum_t vect\left(\mathbf{f}_t\mathbf{f}_t' - \mathrm{E}[\mathbf{f}_t\mathbf{f}_t']\right) \xrightarrow{d} Normal\left(0, \Omega^f\right)$.

(4) Bounded dependence: $\frac{1}{NT}\sum_{i,j,t,s}\|\tau_{ij,ts}\| < \infty$, *where* $\tau_{ij,ts} := \mathrm{E}\left[\mathbf{x}_{it}\mathbf{e}_{it}\mathbf{e}_{js}\mathbf{x}_{js}\right]$

(5) Constant second moments of the covariates: $\Omega_t^{xx} = \mathrm{E}\left[X_tX_t'\right]$ *is constant across time periods.*

A.2 Estimation of the CSC-IPCA estimator

As outlined in Equation 3, the common components of the data generating process are constructed by the interactive effect between time-varying factors F_t and dynamic factor loadings Λ_{it}, which is instrumented by the covariates X_{it} through the mapping matrix Γ. The data generating process can be formulated as follows:

$$Y_{it} = (X_{it}\Gamma)F_t + \epsilon_{it}, \quad \epsilon_{it} = \mu_{it} + H_{it}F_t. \tag{A.1}$$

Where the error term ϵ_{it} is combined with the time varying factors F_t and the idiosyncratic error μ_{it}. The error term μ_{it} is assumed to be independent of the covariates X_{it}, and the factor loadings Γ are assumed to be constant across all units. The objective function in Equation 6 is minimized to estimate the factor F_t and mapping matrix Γ. Equation A.2 details the first step to estimate the factor F_t and mapping matrix Γ with only the control units:

$$(\hat{\Gamma}, \hat{F}_t) = \underset{\Gamma, F_t}{\arg\min} \sum_{i \in \mathcal{T}} \sum_{t \leq T} \left(Y_{it} - (X_{it}\Gamma)F_t\right)' \left(Y_{it} - (X_{it}\Gamma)F_t\right). \tag{A.2}$$

The alternating least squares (ALS) method is employed for the numerical solution of this optimization problem. Unlike PCA, the IPCA optimization challenge cannot be resolved through eigen decomposition. The optimization, as defined in equation above, is quadratic with respect to either Γ or F_t, when the other is held constant. This characteristic permits the analytical optimization of Γ and F_t sequentially. With a fixed Γ, the solutions for F_t are t-separable and can be obtained via cross-sectional OLS for each t:

$$\hat{F}_t(\Gamma) = (\Gamma' X_t' X_t \Gamma)^{-1} \Gamma' X_t' Y_t. \tag{A.3}$$

Conversely, with known F_t, the optimal Γ (vectorized as $\gamma = vect(\Gamma)$) is derived through pooled panel OLS of y_{it} against LK regressors, $x_{it} \otimes f_t$:

$$\hat{\gamma} = \left(\sum_{i,t} (x'_{i,t} \otimes f_t)(x_{i,t} \otimes f'_t) \right)^{-1} \left(\sum_{i,t} (x'_{i,t} \otimes f_t) y_{i,t} \right). \tag{A.4}$$

Inspired by PCA, the initial guess for F_t is the first K principal components of the outcome matrix Y_{it}. The ALS algorithm alternates between these two steps until convergence is achieved, typically reaching a local minimum rapidly. The convergence criterion, based on the minimization of relative change in the parameters F_t and Γ in each iteration, ensures termination when this change falls below a predefined threshold, set at 10^{-6} in our implementation.

As we have mentioned before the estimation of F_t and Γ is not diterministic. Bai (2009) and Xu (2017) set the constraints on the factor loadings and factors before the estimation to ensure the identifiability of the model. However, in our case, since the structural component is identified by the product between factors and factor loadings $X_{it}\Gamma F_t$, we can find any arbitrary rotation matrix R, such that $X_{it}\Gamma R R^{-1} F_t$ yeilds the same structural component. For a specific constraint on the mapping matrix $\Gamma_{norm} = \Gamma_{treat} R$ and factor $F_{norm} = R^{-1} F_t$, such that:

$$\begin{aligned} &\Gamma'_{norm}\Gamma_{norm} = \mathcal{I}, \\ &F_{norm}F'_{norm}/T = \text{Diagonal}. \end{aligned} \tag{A.5}$$

Where $\mathcal{I}$ is the identity matrix, and T is the number of time periods. The rotation matrix R can be easily found by the following steps: first, we use Cholesky decomposition (referred to Higham (2009) for a guidence) to decompose the product $\Gamma'\Gamma$ into a upper triangular matrix $R_1 = cholesky(\Gamma'\Gamma)$, then we perform singular value decomposition on $R_1 F_t F'_t R'_1$ to get $R_2 = U$ where $U\Sigma V' = svd(R_1 F_t F'_t R'_1)$. Finally, the rotation matrix R is given by:

$$R = R_1^{-1} R_2. \tag{A.6}$$

A.3 Hyperparameter tuning

We can also utilize leave-one-out cross validation to select the hyperparameter K, as detailed in Algorithm 2. This method involves excluding the t^{th} period data from the control group to serve as the training data, while similarly excluding the corresponding period data from the treated group to act as validation data. This process is repeated for each time period in the pretreatment phase, applying a predetermined number of factor loadings. The optimal number of factors, K, is identified as the one that yields the minimum average of sum squared errors across all iterations.

Algorithm 2: Leave-One-Out Cross-Validation for Hyperparameter k

```
Data: $Y, X$
Result: Optimal hyperparameter $k$
Determine the maximum possible hyperparameter $K$;
Initialize an array $MSE$ to store the average of sum squared error for each $k$;
for $k = 1$ to $K$ do
    Set sum of squared errors $SSE_k = 0$;
    for $t \leftarrow 1$ to $T_{pre}$ do
        Remove the $t^{th}$ period observation from control data, using the rest as
          training data $(Y_{ctrl}^{-t}, X_{ctrl}^{-t})$;
        Similarly, exclude the $t^{th}$ period observation from treated data, using the rest
          as validation data $(Y_{treat}^{-t}, X_{treat}^{-t})$;
        Estimate parameters $\Gamma$ and $F_t$ using the training data via the ALS method;
        Use the estimated $\hat{\Gamma}$ and $\hat{F}_t$ to predict $\hat{Y}_{treat}^{-t}$ with the validation data;
        Calculate the sum squared error $SE_t = \sum(Y_{treat}^{-t} - \hat{Y}_{treat}^{-t})^2$;
        Accumulate the sum of squared errors: $SSE_k \leftarrow SSE_k + SE_t$;
    end
    Calculate the average sum squared error for $k$: $MSE[k] = \frac{SSE_k}{T_{pre}}$;
end
Select $k$ corresponding to the minimum value in $MSE$;
```

Appendix B Formal Result

In this section, we derive the formal result for the CSC- IPCA estimator. We first establish the consistency and asymptotic properties of the mapping matrix Γ and the factor F_t. We then derive the formal result for the CSC-IPCA estimated ATT.

B.1 Mapping matrix estimation asympototic properties

In this section, we delve into the asymptotic properties of the estimation error associated with the mapping matrix. Kelly et al. (2020) have proven it in their paper, referring to Theorem 3. The following proposition 1, is a special case of their result. Based on our estimation methods, we estimate the mapping matrix Γ first by concentrating out the factor $\mathbf{f}_t$, as shown in Equation 6, we can formulate a target function for Γ as follows:

$$G(\Gamma) = \frac{1}{2NT} \sum_{i,t} \left(y_{it} - \mathbf{x}_{it} \Gamma \hat{\mathbf{f}}_t \right)^2 . \tag{B.7}$$

we define the score function $S(\Gamma)$ as the derivative of the target function $G(\Gamma)$ with respect to Γ: $S(\Gamma) = \frac{\partial G(\Gamma)}{\partial \Gamma}$. The Hessian matrix $H(\Gamma)$ is defined as the second derivative of the target function $G(\Gamma)$ with respect to Γ: $H(\Gamma) = \frac{\partial^2 G(\Gamma)}{\partial \Gamma \partial \Gamma'}$.

It is crucial to highlight that our normalization criterion, delineated in Equation 9, mandates that the mapping matrix Γ adheres to orthonormality and the factor $\mathbf{f}_t \mathbf{f}_t' / T$ is required to exhibit orthogonality. To satisfy these requirement we define the following identification function:

$$I(\Gamma) := \begin{bmatrix} \text{veca}(\Gamma^T \Gamma - \mathcal{I}_K) \\ \text{vecb}\left(\frac{1}{T} \sum_t \hat{\mathbf{f}}_t \hat{\mathbf{f}}_t' - V^{ff} \right) \end{bmatrix} \tag{B.8}$$

where veca$(\cdot)$ and vecb$(\cdot)$ vectorize the upper triangular entries of a square matrix. The difference is veca$(\cdot)$ includes the diagonal elements, while vecb$(\cdot)$ excludes them. We define the Jacobian matrix $J(\Gamma)$ as the derivative of the identification function $I(\Gamma)$ with respect to Γ: $J(\Gamma) = \frac{\partial I(\Gamma)}{\partial \Gamma}$.

Proposition 1 *Under Assumption A.2 and A.3, mapping matrix estimation error centered against the normalized true mapping matrix converges to a normal distribution at the rate of $\sqrt{NT}$: as $N, T \to \infty$ such that $T/N \to \infty$,*

$$\sqrt{NT}\left(\hat{\gamma}-\gamma\right)\xrightarrow{d}-\left(H^{0\prime}H^{0}+J^{0\prime}J^{0}\right)^{-1}H^{0\prime}Normal(0,\mathbb{V}_{Y}^{[1]})$$

where $H^0 := \frac{\partial S(\Gamma)}{\partial \gamma}|_{\gamma=\gamma^0}$ and $J^0 := \frac{\partial I(\Gamma)}{\partial \gamma}|_{\gamma=\gamma^0}$, $\mathbb{V}_Y^{[1]} = \left(Q^0 \otimes \mathcal{I}_K\right)\Omega^{x\epsilon f}\left(Q^{0\prime} \otimes \mathcal{I}_K\right)$, and $Q^0 := Q_t(\Gamma^0)$ given that $Q_t(\Gamma) := \mathcal{I}_L - \Omega_t^{xx}\left(\Gamma'\Omega_t^{xx}\Gamma\right)^{-1}\Gamma'$ is constant over t under Assumption A.3.

Proof: referring to Kelly et al. (2020).

B.2 Factor estimation asympototic properties

Proposition 2 *Under Assumption A.2 and A.3, factor estimation error centered against the normalized true factor converges to a normal distribution at the rate of* $\sqrt{N}$*: as* $N, T \to \infty$ *for* $\forall t$*,*

$$\sqrt{N}\left(\hat{\mathbf{f}}_t - \mathbf{f}_t\right) \xrightarrow{d} N\left(0, \mathbb{V}_t^{[2]}\right),$$

Proof: Decompose the left hand side euqation:

$$\begin{aligned}\sqrt{N}\left(\hat{\mathbf{f}}_t - \mathbf{f}_t\right) &= \sqrt{N}\left(\left(\hat{\Gamma}'X_t'X_t\hat{\Gamma}\right)^{-1}\hat{\Gamma}'X_t'\left(X_t\hat{\Gamma}\mathbf{f}_t + \tilde{\epsilon}_t\right) - \mathbf{f}_t\right)\\ &= \sqrt{N}\left(\left(\hat{\Gamma}'X_t'X_t\hat{\Gamma}\right)^{-1}\hat{\Gamma}'X_t'X_t\hat{\Gamma}\mathbf{f}_t - \mathbf{f}_t\right) + \sqrt{N}\left(\hat{\Gamma}'X_t'X_t\hat{\Gamma}\right)^{-1}\hat{\Gamma}'X_t'\tilde{\epsilon}_t\end{aligned}$$

where $\tilde{\epsilon}_t$ is the estimated error term. Given Proposition 1, $\hat{\Gamma} - \hat{\Gamma}^0 = \mathcal{O}_p\left(1/\sqrt{NT}\right)$. The first term is simply $\mathcal{O}_p\left(1/\sqrt{NT}\right)$. For the second term:

$$\begin{aligned}\sqrt{N}\left(\hat{\Gamma}'X_t'X_t\hat{\Gamma}\right)^{-1}\hat{\Gamma}'X_t'\epsilon_t =& \sqrt{N}\left(\Gamma'X_t'X_t\Gamma\right)^{-1}\Gamma'X_t'\epsilon_{\mathbf{t}} + \mathcal{O}_p(1)\\ &\xrightarrow{d} Normal(0, \mathbb{V}_t^{[2]})\end{aligned}$$

where the variance term $\mathbb{V}_t^{[2]}$, which is given by $\mathbb{V}_t^{[2]} = \left(\Gamma^\top\Omega_t^{xx}\Gamma\right)^{-1}\Gamma^\top\Omega_t^{x\epsilon}\Gamma\left(\Gamma^\top\Omega_t^{xx}\Gamma\right)^{-1}$. Assumption A.3 delineates the properties of Ω_t^{xx} and $\Omega_t^{x\epsilon}$. Notably, despite the presence of multiple matrix multiplications, the matrices Ω_t^{xx} and $\Omega_t^{x\epsilon}$ remain invariant under these operations. Consequently, the term $\mathbb{V}_t^{[2]}$ is constant across all observational units.

B.3 Consistency and asymptotic property of the ATT estimation

Theorem 1 *Under Assumptions 1, A.2, and A.3, the CSC-IPCA estimator* $\mathrm{E}\left(\widehat{ATT}_t|D,X,\Lambda,F\right)\xrightarrow{P} ATT_t$*, where* $ATT_t=\frac{1}{N_{treat}}\sum_{i\in\mathcal{T}}\delta_{it}$ *is the true treatment effect. for all* $t>T_{pre}$ *as both* $N_{ctrl},\ T_{pre}\to\infty$.

Proof: Denote i as the treated unit on which the treatment effect is of interest, the bias of estimated ATT is given by:

$$
\begin{aligned}
\hat{\delta}_{it}-\delta_{it} &= y^1_{it}-\hat{y}^0_{it}-\delta_{it},\\
&= \mathbf{x}_{it}\Gamma\mathbf{f}'_t-\mathbf{x}_{it}\hat{\Gamma}\hat{\mathbf{f}}'_t+\epsilon_{it},\\
&= \mathbf{x}_{it}\left(\left(\mathcal{I}\otimes\mathbf{f}_t\right)\boldsymbol{\gamma}-(\mathcal{I}\otimes\hat{\mathbf{f}}_t)\hat{\boldsymbol{\gamma}}\right)+\epsilon_{it},\\
&= \mathbf{x}_{it}\left(\left(\mathcal{I}\otimes\mathbf{f}_t\right)\boldsymbol{\gamma}-\mathcal{I}\otimes(\mathbf{f}_t+\mathbf{e}_{f_t})(\boldsymbol{\gamma}+\mathbf{e}_\gamma)\right)+\epsilon_{it},\\
&= \mathbf{x}_{it}\left((\mathcal{I}\otimes\mathbf{f}_t)\mathbf{e}_\gamma-(\mathcal{I}\otimes\mathbf{e}_{f_t}\boldsymbol{\gamma})-(\mathcal{I}\otimes\mathbf{e}_{f_t})\mathbf{e}_\gamma\right)+\epsilon_{it}\\
&= \mathbf{x}_{it}E_\Gamma\mathbf{f}'_t-\mathbf{x}_{it}\Gamma\mathbf{e}'_{f_t}-\mathbf{x}_{it}E_\Gamma\mathbf{e}'_{f_t}+\epsilon_{it},\\
&= A_{1,it}+A_{2,it}+A_{3,it}+\epsilon_{it}.
\end{aligned}
$$

where $\mathbf{e}_{f_t}=\mathbf{f}_t-\hat{\mathbf{f}}_t$ is a vector of estimation error of the factor $\mathbf{f}_t$, $\mathbf{e}_\gamma$ is vectorized estimation error of the mapping matrix $E_\Gamma=\Gamma-\hat{\Gamma}$. The third step converts the vector-matrix multiplication into vector multiplications with the Kronecker product, $\mathbf{x}_{it}\Gamma\mathbf{f}'_t=\mathbf{x}_{it}(\mathcal{I}\otimes\mathbf{f}_t)\boldsymbol{\gamma}$, where $\mathcal{I}$ is an $L\times L$ identity matrix. The bias of the estimated ATT is the sum of three terms $A_{1,it}$, $A_{2,it}$, $A_{3,it}$, and ϵ_{it}. By propersition 1 and 2, we have the following results:

$$
\begin{aligned}
A_{1,it} &= \mathbf{x}_{it}E_\Gamma\mathbf{f}'_t=\mathcal{O}_p\left(1/\sqrt{N_{treat}T_{pre}}\right).\\
A_{2,it} &= -\mathbf{x}_{it}\Gamma\mathbf{e}'_{f_t}=\mathcal{O}_p\left(1/\sqrt{N_{ctrl}}\right).\\
A_{3,it} &= -\mathbf{x}_{it}E_\Gamma\mathbf{e}'_{f_t}=\mathcal{O}_p\left(1/\sqrt{N_{treat}T_{pre}}\right).
\end{aligned}
$$

Since we estimate the factor $\mathbf{f}_t$ using only control units and update the mapping matrix Γ with treated units in the pre-treatment period, both $\mathbf{f}_t$ and Γ converge over different dimensions of T and N. Consequently, the error term ϵ_{it} is assumed to converge to zero, leading to the bias of the estimated ATT also converging to zero:

$$\begin{aligned}\hat{\delta}_{it} - \delta_{it} &= \mathcal{O}_p\left(\frac{1}{\sqrt{N_{\text{treat}} T_{\text{pre}}}}\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N_{\text{ctrl}}}}\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N_{\text{treat}} T_{\text{pre}}}}\right) + \epsilon_{it} \\ &= \mathcal{O}_p\left(\frac{1}{\sqrt{N_{\text{ctrl}}}}\right).\end{aligned}$$

Therefore, as $N_{\text{ctrl}},\ T_{\text{pre}} \to \infty$, the estimated ATT converges to the true ATT:

$$\mathrm{E}\left(\widehat{ATT}_t | D, X, \Lambda, F\right) \xrightarrow{P} ATT_t.$$