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

Cong Wang*

March 18, 2024

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) leverage data from the control group to impute the missing counterfactual for the treated post treatment. Eventhough SCM relaxes some strict assumptions, it requires the treated unit to be inside the convex hull formulated by control units to avoid 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 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 interactive fixed effect (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 in managing high dimensional data but also adapts the versatility of IFE 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) proposed by Xu (2017). To model the entire data generating process, the CSC-IPCA estimator is designed to overcome the constraints of unconfoundedess assumption required by methodologies such as matching (Abadie and Imbens (2011, 2006)), 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)) implicitly imputes the missing counterfactuals for treated units with meticulously selected control units. The DID method (Card and Krueger (1993); Ashenfelter (1978)), on the other hand, explicitly imputes the missing counterfactuals by differencing the treated and control units, based on the assumption of parallel trends. 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 also highlighted by Athey et al. (2021) with their matrix completion (MC) method.

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. Firstly, it assumes a simple fixed effect model as with Bai and Perron (2003) with only the interactive component between factor loadings and time varying factors:

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

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:

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

This approach presents three major advantages over the CSC-IFE method. 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 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 at time t. The total number of units is $N = N_{treat} + N_{\text{ctrl}}$, where N_{treat} indicates the number of units in the treatment group, and N_{ctrl} represents those in the control group. Each unit is observed over T time periods, ranging from period 1 to period T. Let T_{pre} denote the number of pre-treatment periods, and T_{post} the number of post-treatment periods. The unit is treated at time $T_{\text{pre}}+1$, and the treatment effect is initially observed at time $T_{\text{pre}}+1$ and continues to be observed thereafter, a scenario commonly referred to as staggered adoption. For the data generating process of Y_{it} we assume the following functinoal form:

Assumption 1 *Functional form:*

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

where D_{it} is a binary treatment indicator and δ_{it} signifies the treatment effect, which exhibits variation across units and through times[1]. The expression $\Lambda_{it} = [\lambda_{it}^1, \ldots, \lambda_{it}^K]$ represents a vector of factor loadings (the number of common factors is K.), whereas $F_t = [f_t^1, \ldots, f_t^K]'$ corresponds to a vector of time-varying common factors, and μ_{it} is the idiosyncratic error term. A key distinction of the proposed model from that delineated in Xu (2017) is the incorporation of factor loadings B_{it}, which are instrumented by observed covariates X_{it}. This integration permits Λ_{it} to exhibit variability across time and units, thereby introducing an additional layer of heterogeneity into the model.

The vector $X_{it} = [x_{it}^1, \ldots, x_{it}^L]$ consists of observed covariates, where L denotes the number of covariates. The factor loadings Λ_{it} are assumed to be a linear function of these observed covariates X_{it}, with Γ acting as the $L \times K$ coefficient matrix, and $H_{it} = [\eta_{it}^1, \ldots, \eta_{it}^L]$ comprising the vector of error terms. Another key difference from the CSC-IFE approach is that we

[1]The symble "$\circ$" represents point-wise product.

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[2]. The loagic 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 allows us to handle high-dimensional datasets, especially when dealing with a large number of covariates, offering another advantage over the CSC-IFE method.

Following Neyman (1932) potential outcome framework (also discussed by Rubin (1974, 2005)), we observe the actual outcome for the treated $Y(1)$ and with the modelled data generating process we can impute the missing counterfactual $\hat{Y}(0)$. The difference between the actual outcome and imputed missing counterfactual will be the treatment effect for treated (ATT), which is defined as:

$$\widehat{ATT}_t = \frac{1}{N_{treat}} \sum_{i \in N_{treat}} \Big(Y_{it}(1) - \hat{Y}_{it}(0) \Big) = \frac{1}{N_{treat}} \sum_{i \in N_{treat}} \hat{\delta}_{it}. \tag{4}$$

2.1 Assumptions for identification

Assumption 2 *Unconfoundedness:*

$$\epsilon_{it} \perp D_{js}, X_{js}, F_s \quad \forall i, j, s, t. \tag{5}$$

Assumption 2 stipulates that the error term for any unit at any time period is independent of treatment assignment, observed covariates, and unobserved time-varying factors. This independence is a crucial condition that lends substance to model Equation **??** and is imperative for the consistent estimation of Γ.

Assumption 3 *Regularity conditions: (1)* Γ *is bounded and has a finite second moment, (2)* F_t *is bounded and has a finite second moment, (3)* X_{it} *is bounded and has a finite second moment.*

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

The regularity conditions outlined in Assumption 3 are essential for the consistent estimation of Γ and F_t. Specifically, these conditions ensure that the matrix $\Gamma' X_t' X_t \Gamma$, which is involved in inversion processes, remains nonsingular (where X_t denotes the $N \times L$ matrix consisting of the cross-section of $x_{i,t}$).

Assumption 4 *Asymptotic normality:*

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

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

Assumption 4 simply containes central limit theorems with respect to different variables, which are satisfied by various mixing processes.

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:

$$\arg\min_{\Gamma, F_t} \sum_i \sum_t \left(Y_{it} - (X_{it}\Gamma)F_t\right)^2. \tag{6}$$

Unlike the IFE method (Bai (2009)), our approach requires estimating only two parameters, Γ and F_t, simplifying the process. Different from principal component analysis (Jolliffe (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 coefficient matrix $\hat{\Gamma}_{\text{ctrl}}$ utilizing an ALS algorithm, based exclusively on data from the control group for the whole time period.

$$(\hat{\Gamma}_{ctrl}, \hat{F}_t) = \arg\min_{\Gamma, F_t} \sum_{i \in N_{ctrl}} \sum_{t \in T} (Y_{it} - (X_{it}\Gamma)F_t)' (Y_{it} - (X_{it}\Gamma)F_t) \,. \tag{3}$$

Step 2: The subsequent step involves estimating the coefficient 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.

$$\hat{\Gamma}_{treat} = \arg\min_{\Gamma} \sum_{i \in N_{treat}} \sum_{t \in T_{pre}} \left(Y_{it} - (X_{it}\Gamma)\hat{F}_t\right)' \left(Y_{it} - (X_{it}\Gamma)\hat{F}_t\right) . \tag{4}$$

Step 3: The final step involves imputing the counterfactual outcome $\hat{Y}_{it}(0)$ for treated unit i at time t by substituting the estimated coefficient matrix $\hat{\Gamma}_{treat}$ and the time-varying factors $\hat{F}_t$ into the following equation:

$$\hat{Y}_{it}(0) = (X_{it}\hat{\Gamma}_{treat})\hat{F}_t, \quad \forall i \in N_{treat}, \forall t \in T_{post}. \tag{5}$$

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 rotation 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 rotation 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 rotation matrices.

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 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.

3.2 Inference

In our approach to causal inference, we focus on modeling the whole data generating process to impute unobserved counterfactuals. Unlike approaches that rely on parametric bootstrap

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 Γ 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;

procedures for obtaining uncertainty estimates for inference, as used by Xu (2017), conformal inference (Chernozhukov et al. (2021)) has gained popularity in recent literature such as Ben-Michael et al. (2021), Roth et al. (2023), and Imbens (2024). This 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:

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

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

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 Γ 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;

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[3]. 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:

[3]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.

$$S(\hat{\mu}) = \left(\frac{1}{\sqrt{T_*}} \sum_{t \in T_1} |\hat{\mu}|^q \right) \tag{7}$$

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 permutation. The P-value is defined as:

$$\hat{p} = 1 - \hat{F}(S(\hat{u})), \text{ where } \hat{F}(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} 1\{S(\hat{u}_\pi) < x\}. \tag{8}$$

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 coverage probability.

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[4], The drift term μ_i equals 0 for control units and 2 for treated units,[5], 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-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 $L \times K$ coefficient matrix for the factor loadings, is drawn uniformly from $(-0.1, 0.1)$, with these covariates serving as instruments. 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[6]. Only the outcome Y_{it}, the covariates X_{it}, and the treatment indicator D_{it} are observed, while all other variables remain unobserved.

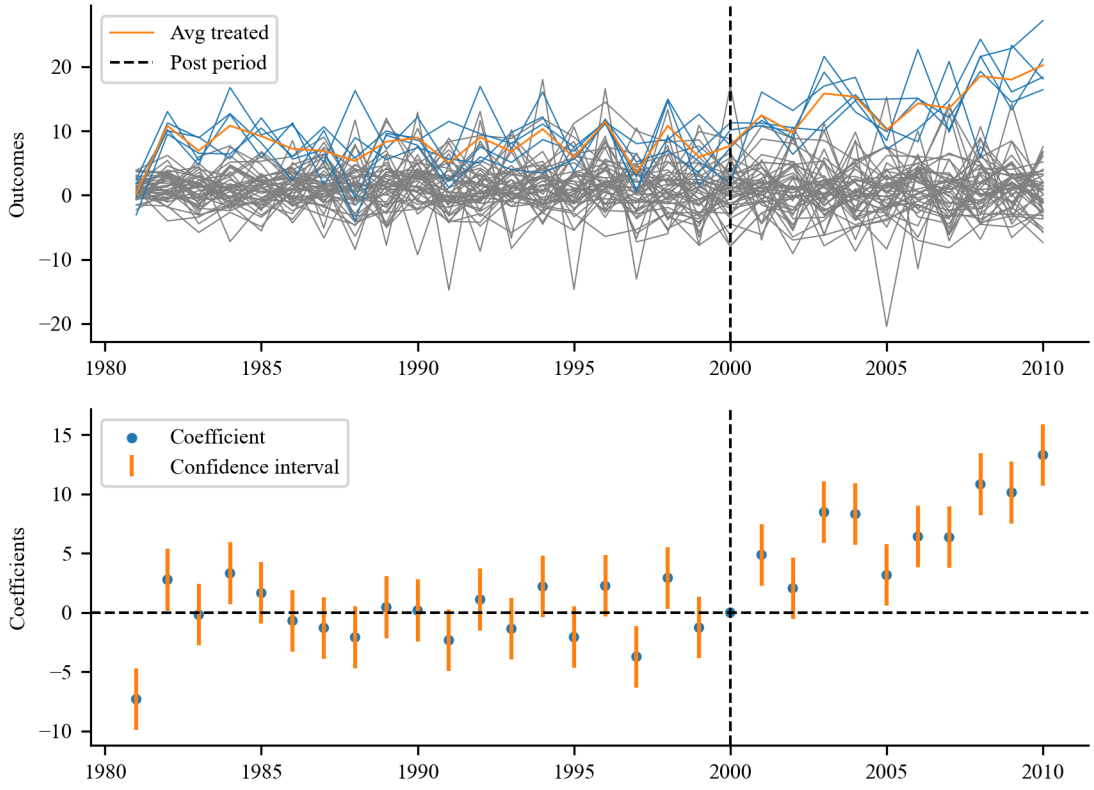
Figure 1 represents the simulated data following our data generating process. Observations from the upper panel indicate that the parallel trend assumption is not met. To verify this, we plot a simple event study, clearly revealing a failure in the parallel trend assumption. Furthermore, outcomes for 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]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.

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

[6]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**



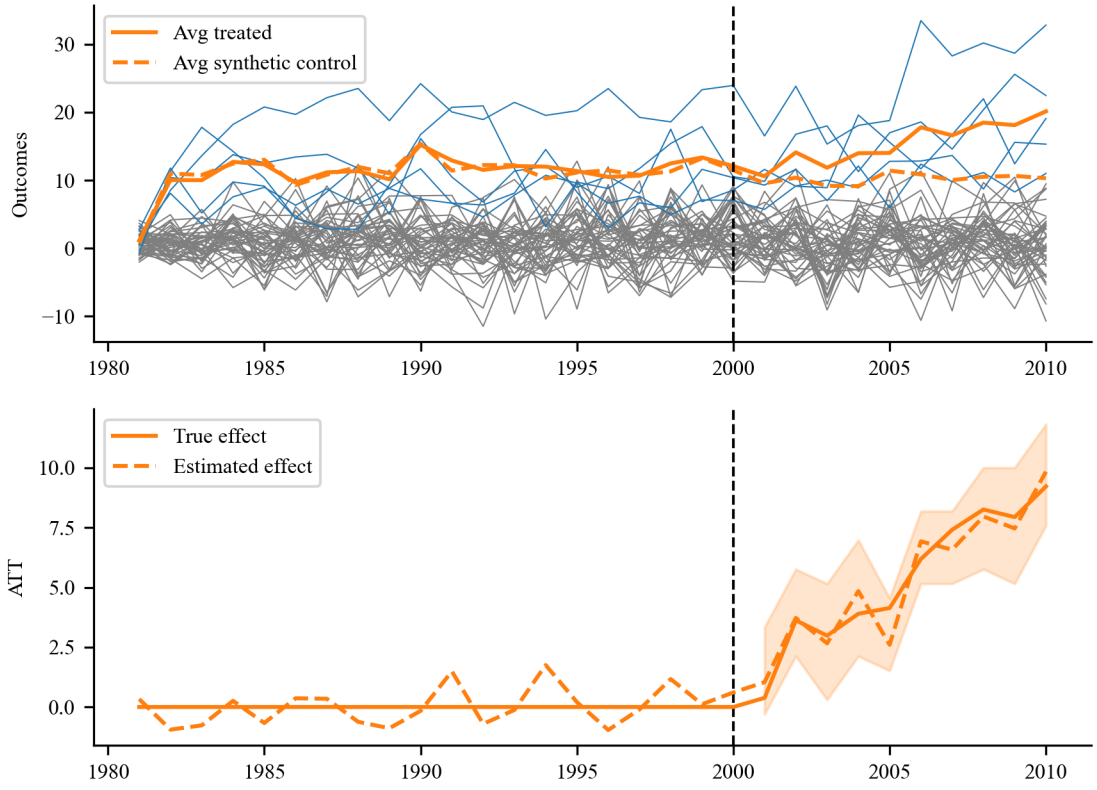
In this graphic, the upper panel plots simulated data following the above data generating process. The light blue lines represent treated units and the light gray lines represent controls. Key parameters are$N_{treat} = 5, N_{ctrl} = 45, T_0 = 20, T_1 = 10, L = 10$. The lower panel plots a simple event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

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 alignment between the dashed and solid lines.

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



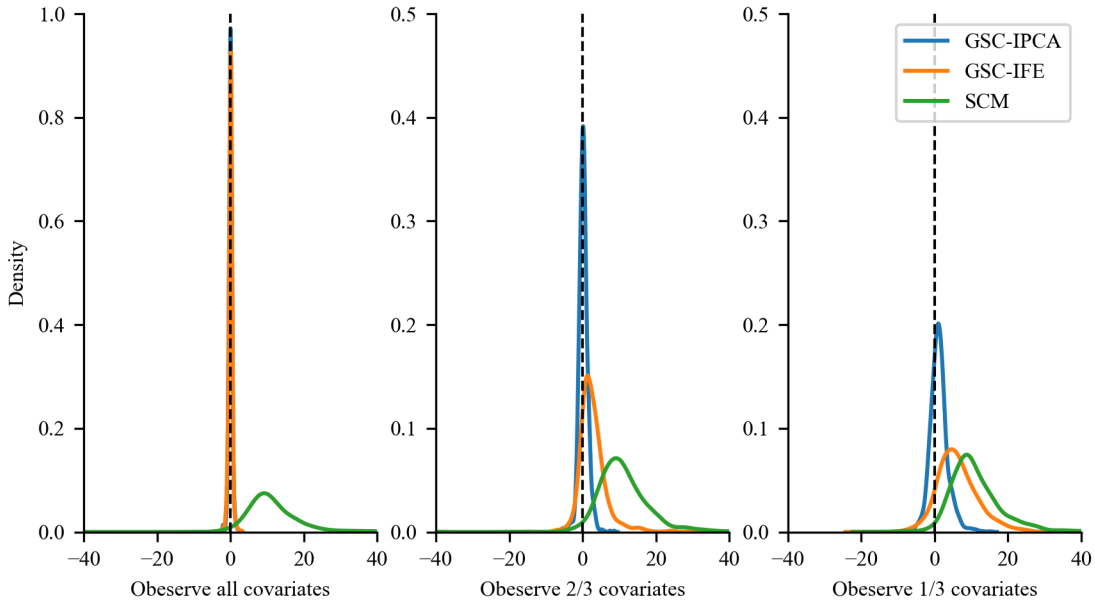
This graphic plots the CSC-IPCA method estimated ATT for simulated data $N_{treat} = 5, N_{ctrl} = 45, T_0 = 20, T_1 = 10, L = 10$.

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, 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 $N_{treat} = 5, T_1 = 5$. We vary the number of control units N_{co}, pre treatment

period T_0 , 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[7]. 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 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.

[7]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 $STD = \frac{1}{T_1}\sum_{t\in T_1}\left(\widehat{ATT}_t - \frac{1}{T_1}\sum_{t\in T_1}\widehat{ATT}_t\right)^2$

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 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.

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.

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.

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.

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

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.

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

A.1 Estimation of the CSC-IPCA Estimator Using the ALS Algorithm

As outlined in Equation 3, the data generating process can be described by:

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

Equation 3 details the derivation of the CSC-IPCA estimator from the minimization problem:

$$(\hat{\Gamma}, \hat{F}_t) = \arg\min_{\Gamma, F_t} \sum_{i \in N} \sum_{t \in T} (Y_{it} - (X_{it}\Gamma)F_t)' (Y_{it} - (X_{it}\Gamma)F_t).$$

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 3, 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:

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

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

$$\hat{\gamma} = \left(\sum_{i,t} (x_{i,t}' \otimes f_t)(x_{i,t} \otimes f_t') \right)^{-1} \left(\sum_{i,t} (x_{i,t}' \otimes f_t) y_{i,t} \right).$$

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 relative change in the optimization problem from Equation 3, ensures termination when

this change falls below a predefined threshold, set at $10e^{-6}$ in our implementation.