Generalized Synthetic Control Method: Causal Inference with Instrumented Principal Component Analysis

Cong Wang[*]

March 8, 2024

Abstract

The fundamental problem of causal inference is missing data. Based on identification assumptions such as unconfoundedness, traditional methodologies impute the missing counterfactuals either implicitly or explicitly. However, the required assumptions are either too strong or untestable. Synthetic control methods (SCM) leverage data from the control group to impute the missing counterfactual for the treated post-treatment. Eventhough SCM relaxes these strict assumptions, it requires the treated unit to be inside the convex hull formulated by control units, in 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 model by integrating covariates into dynamic factor loadings adding additional robustness. This methodology confers multiple benefits: firstly, it incorporates the strengths of previous SCM approaches, such as the relaxation of the Parallel Trends Assumption (PTA) and conditional randomization of treatment assignment. Secondly, it does not require the targeted outcomes inside the convex hull formulated by control units. Thirdly, it eliminates the need for correct functional form assumptions required by IFE. Finally, it efficiently manages 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 propose a new counterfactual imputation method that leverages the dimension reduction capability of PCA method (Jollife and Cadima (2016)) and the flexibility of the IFE model (Bai and Perron (2003)) to both enhance the value extracted from numerous covariates and improve the predictive accuracy. We name the newly proposed method the generalized synthetic control method with instrumented principal component analysis (GSC-IPCA), aligning with the previous generalized synthetic control method with interactive fixed effects (GSC-IFE) proposed by Xu (2017). To model the entire data generating process, the GSC-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 SCM (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 a counterfactual, 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 implicitly imputes the missing counterfactuals for treated units with meticulously selected control units. While 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).

The GSC-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 extends the linear IFE model by first performing a dimension reduction operation to distill information from multiple covariates, thus obtaining unit-specific factor loadings, which inherit time-varying properties[1]. Subsequently, these factor loadings are interacted with time-varying factors. A critical feature of our approach is that the unit-specific factor loadings are instrumented by covariates, which provide additional robustness. Our simulations reveal that the GSC-IPCA estimator maintains its unbiasedness even in scenarios where some covariates remain unobserved, in contrast to the GSC-IFE estimator, which tends to exhibit bias under comparable conditions.

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.

Assumption 1 *Functional form:*

$$\begin{aligned} Y_{it} &= D_{it} \circ \delta_{it} + B_{it} F_t + \mu_{it}, \\ B_{it} &= X_{it} \Gamma + H_{it} \end{aligned} \tag{1}$$

where D_{it} is a binary treatment indicator and δ_{it} signifies the treatment effect, which exhibits variation across units and through times[2]. The expression $B_{it} = [\beta_{it}^1, \ldots, \beta_{it}^K]$ represents a vector of factor loadings (the number of common factors is K.), whereas $F_t = [f_t^1, \ldots, f_t^K]'$

[1] The functional form of the data generating process in Xu (2017) only sepcifies unit-varying factor loadings. Here we make the unit specific factor loadings to accomandate time-variying property.

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

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 B_{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 B_{it} are theorized 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.

Upon examining the functional form presented in Equation 1, we can amalgamate the two segments to formulate the ensuing equation:

$$Y_{it} = D_{it} \circ \delta_{it} + (X_{it}\Gamma)F_t + \epsilon_{it}, \quad \epsilon_{it} = \mu_{it} + H_{it}F_t. \tag{2}$$

The factor component of the model, $B_{it}F_t = \beta_{it1}f_{1t} + \beta_{it2}f_{2t} + \cdots + \beta_{itk}f_{kt}$, where $B_{it} = X_{it}\Gamma$, is assumed to adopt a linear, additive form. Despite appearing to be restrictively structured, this approach is capable of capturing a vast array of unobserved heterogeneities. It is inclusive of all specifications present in the interactive fixed effects model within the GS-IFE, such as unit and time fixed effects, unit-specific linear or quadratic time trends, and autoregressive processes. Beyond the additive integration of the treatment effect as delineated in Equation 2, the model imposes no additional constraints on the functional form of the treatment effect. This level of flexibility enables the straightforward application of PCA for estimating the factor loadings and common factors, thereby facilitating the imputation of counterfactual outcomes for treated units.

The main quantity of interest of this paper is the average treatment effect (ATE) for the treated, which is defined as:

$$\widehat{ATT}_t = \frac{1}{N_{treat}} \sum_{it} \Big(Y_{it}(1) - \hat{Y}_{it}(0) \Big) = \frac{1}{N_{treat}} \sum_{it} \hat{\delta}_{it}. \quad for \quad \forall i > N_{co}, \, \forall t > T_{pre}. \tag{3}$$

2.1 Assumptions for identification

Assumption 2 *Unconfoundedness:*

$$\epsilon_{it} \perp D_{js}, X_{js}, F_s \quad \forall i, j, s, t. \tag{4}$$

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 2 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.*

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 GSC-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: $\delta_{it} = Y_{it}(1) - \hat{Y}_{it}(0)$, where $\hat{Y}_{it}(0)$ is derived through a three-step imputation process.

Step 1: The initial step entails estimating the time-varying factors $\hat{F}_t$ and the coefficient matrix $\hat{\Gamma}_{\text{ctrl}}$ utilizing an Alternating Least Squares (ALS) algorithm, based exclusively on data from the control group.

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

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:

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

4 Monte Carlo Simulation

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

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

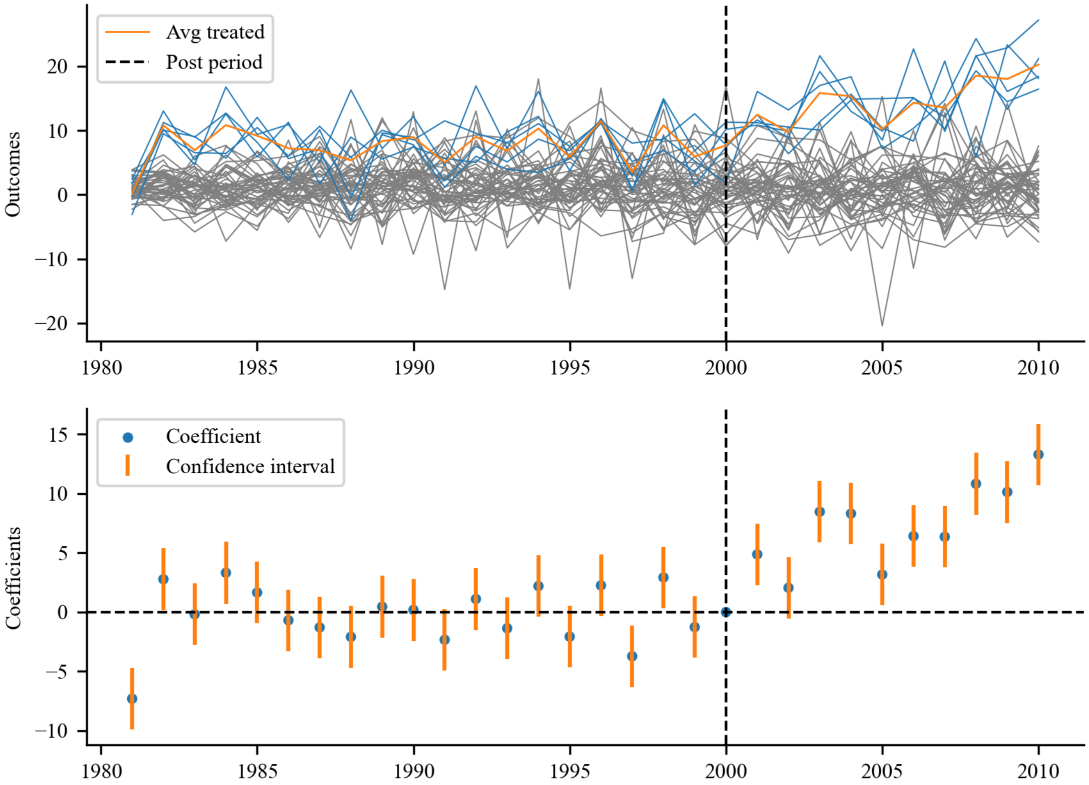
where $X_{it} = [x_{it}^1, \ldots, x_{it}^L]$ denotes a vector of $L \times 1$ time-varying covariates, which follows a VAR(1) process. $X_{it} = \mu_i + A_i X_{i,t-1} + \nu_{it}$, where A_i is a $L \times L$ variance-covariance matrix[3], The drift term μ_i equals 0 for control units and 2 for treated units,[4], 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 $\bar{\delta}_t = [0, \cdots, 0, 1, 2, \ldots, T_{post}]$

[3] 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.

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

represents a time-varying treatment effect[5]. 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: **GSC-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.

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.

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

In such cases, the synthetic control method will be biased, as it avoids extrapolation and typically fit poorly for treated units.

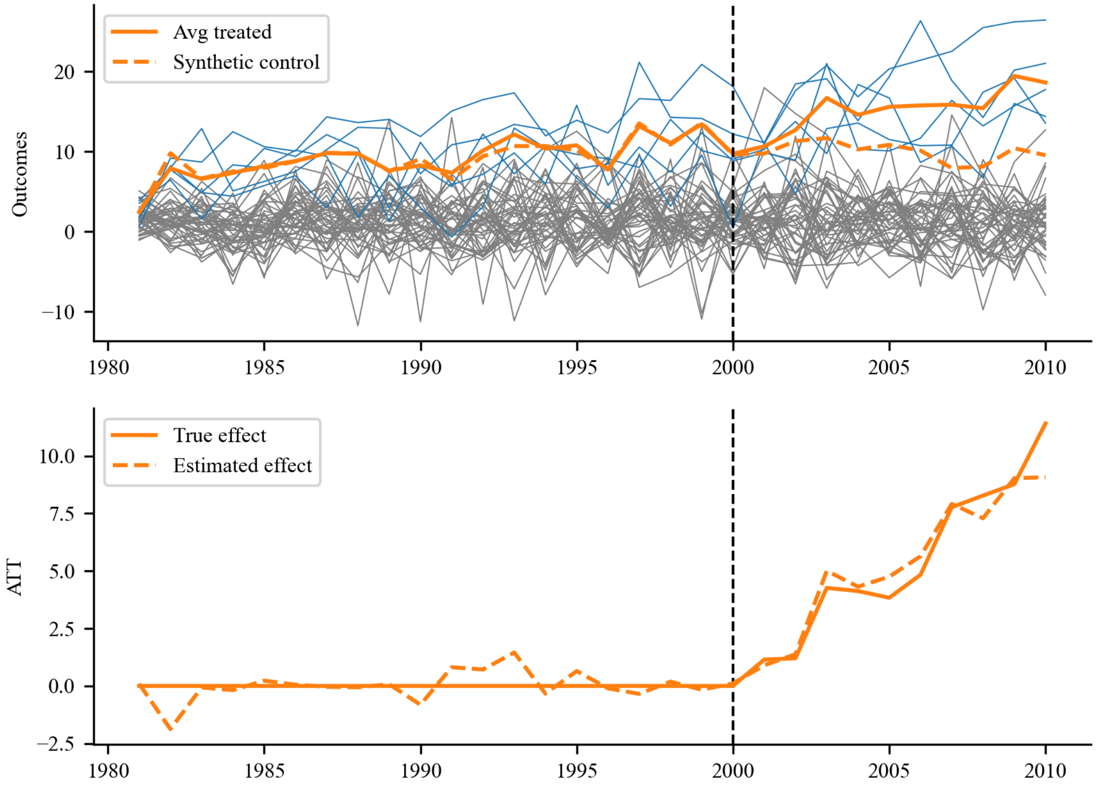
4.1 A simulated example

Following this data generating process, figure 2 illustrates both the raw data and the imputed counterfactual outcomes as estimated by the GSC-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 GSC-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 GSC-IPCA method is able to capture the true ATT, as evidenced by the close alignment between the dashed and solid lines.

4.2 Bias comparision

Based on the same data generating process and parameters, we compare the GSC-IPCA, GSC-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 GSC-IPCA and GSC-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 GSC-IPCA and GSC-IFE lose efficiency; however, the GSC-IPCA estimator remains unbiased.

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



This graphic plots the GSC-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

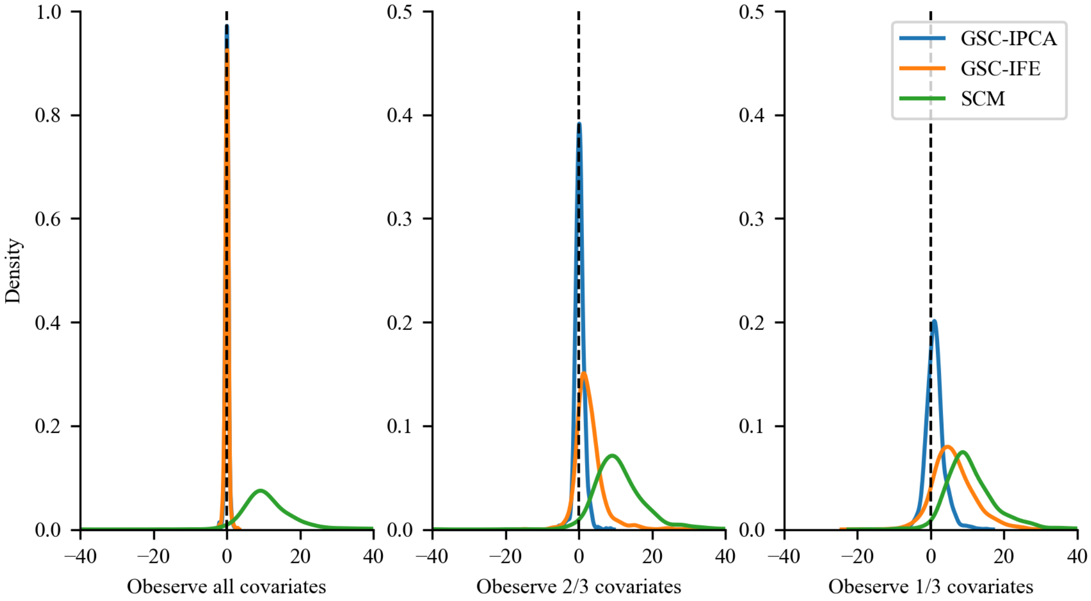
5 Empirical Application

In this section, we

6 Conclusion

Firms'

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



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

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.

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 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.

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

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.

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.

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 GSC-IPCA Estimator Using the ALS Algorithm

As outlined in Equation 2, 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 GSC-IPCA estimator from the minimization problem:

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

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.