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

Cong Wang†

September 9, 2024

Job Market Paper, latest version available here.

Abstract

We propose a novel method for causal inference within the frameworks of counterfactual and synthetic control methods. Building on the Generalized Synthetic Control method developed by Xu (2017), the Instrumented Principal Component Analysis method instruments factor loadings with predictive covariates rather than including them as direct regressors. These instrumented factor loadings exhibit time-varying dynamics, offering a better economic interpretation. Covariates are instrumented through a transformation matrix, Γ, when we have a large number of covariates it can be easily reduced in accordance with a small number of latent factors helping us to effectively handle high-dimensional datasets and making the model parsimonious. Most importantly, our simulations show that this method is less biased in the presence of unobserved covariates compared to other mainstream approaches. In the empirical application, we use the proposed method to evaluate the effect of Brexit on foreign direct investment to the UK.

Keywords: Synthetic Control, Principal Component Analysis, Factor Model, Causal Inference

JEL Codes: G11, G12, G30

*We thank Matteo Lacopini, Emanuele Bacchiocchi for helpful discussion on this paper. There is a Github repository for this paper, available at https://github.com/CongWang141/JMP.git, which contains the latest version of the paper, the code, and the data.

†Department of Economics and Law, Sapienza University of Rome.

1 Introduction

In this paper, we introduce a novel counterfactual imputation method for causal inference, called the Counterfactual and Synthetic Control method with Instrumented Principal Component Analysis (CSC-IPCA). This method combines the dimension reduction capabilities of Principal Component Analysis (PCA) described by Jollife and Cadima (2016) to handle high-dimensional datasets with the versatility of the factor models studied by Bai and Perron (2003), Bai (2009), among others, which accommodate a wide range of data-generating processes (DGPs). The CSC-IPCA method represents a significant advancement over the Generalized Synthetic Control (GSC) method proposed by Xu (2017), which utilizes the Interactive Fixed Effects (IFE) approach to model DGPs and impute missing counterfactuals for causal inference.

The main difference between our method and CSC-IFE[1] lies in how we handle covariates. CSC-IFE combines the structural component $\Lambda_i F_t$ with the regressors $X_{it}\beta$, as shown in the following equation:

$$y_{it} = \Lambda_i F_t + X_{it}\beta + \epsilon_{it} \tag{1}$$

Instead of including the covariates X_{it} linearly as regressors, the CSC-IPCA method instruments the factor loadings Λ_{it} with predictive covariates through a transformation matrix Γ. This method is constructed as fallowing: first, it assumes a simple factor model, as in Bai and Perron (2003), with only the structural component combined with factor loadings Λ_i and common factors F_t:

$$y_{it} = \Lambda_i F_t + \epsilon_{it} \tag{2}$$

[1]In this paper, we consider the Generalized Synthetic Control (GSC) method as part of the broader counterfactual and synthetic control framework. Therefore, throughout the paper, we refer to the GSC method as the Counterfactual and Synthetic Control method with Interactive Fixed Effects (CSC-IFE).

Next, it instruments the static factor loadings Λ_i with covariates X_{it} instead of including them as regressors, allowing the factor loadings to incorporate time-varying properties and become dynamic:

$$\Lambda_{it} = X_{it}\Gamma + H_{it} \tag{3}$$

The static factor loadings Λ_i in Equation 2 are assumed to be time-invariant by most studies in the related literature. However, in many economonic and social science context, the factor loadings are not constant but fluctuate over time in response to relevant covariates. By instrumenting the factor loadings Λ_i with covariates X_{it} through Equation 3, we can capture the time-varying properties of the factor loadings. The matrix Γ, serving as an $L \times K$ mapping function from covariates (with the number of L) to factor loadings (with the number of K), also acts as dimension reduction operation, which aggregates all the information from the covariates into a smaller number of factor loadings, making the model parsimonious.

The CSC-IPCA method offers several key benefits. First, it inherits the dimension reduction capabilities of conventional PCA, where the transformation matrix Γ serves as a dimensionality reduction operator. This enables efficient handling of high-dimensional datasets with a large number of predictive covariates while maintaining the sparsity of the factor model. This feature is particularly valuable when working with financial data (Feng et al. (2020)) and high-dimensional macroeconomic time series data (Brave (2009)).

Second, unlike conventional static factor models, the instrumented factor loadings in CSC-IPCA exhibit time-varying dynamics. This is particularly realistic in many economic and social science contexts. For example, consider a company that increases its investment in R&D, transitioning from a conservative stance to a more aggressive one. This change can also impact its profitability, potentially shifting it from a robust to a weaker position. As a result, the unit effect evolves along with its investment strategy. In such cases, static factor loadings fail to capture the time-varying dynamics of the company's changing fundamentals.

Last but not least, the most valuable benefit of the CSC-IPCA method is its reduced bias when unobserved covariates are present, compared to other similar methods. Instead of including covariates linearly as regressors which is a practice often criticized for model misspecification. The CSC-IPCA method incorporates covariates into the factor loadings through a mapping matrix. This approach provides a more efficient way of handling covariates, allowing for better extraction of predictive information and reducing exposure to model misspecification. Our simulation studies demonstrate that, in the presence of unobserved covariates, the CSC-IPCA method is the least biased among the methods considered.

The IPCA method was developed by Kelly et al. (2020), and applied by Kelly et al. (2019) for predicting stock returns in the asset pricing literature. The main difference between using the IPCA method for prediction and for causal inference lies in the assumption that the transformation matrix Γ differs between treated and control units. In the estimation process, we first use the control units to estimate the common factors F_t over the entire time period. Next, we update the transformation matrix Γ_{treat} for the treated units using data from the pre-treatment period. The subsequent step involves normalizing the common factors and the transformation matrix based on prespecified normalization restrictions. Finally, the estimated parameters are used to impute the missing counterfactuals for the treated units after the treatment, allowing us to evaluate the average treatment effect on the treated (ATT).

We provide bootstrap and leave-one-out cross-validation procedures for hyperparameter tuning to select the optimal number of latent factors, K. Additionally, we construct confidence intervals using the novel and increasingly popular conformal inference method developed by Chernozhukov et al. (2021). In our formal results, we derive the asymptotic properties based on the unbiased and efficient estimation of both Γ and F_t. We show that the convergence rate of our estimand, i.e. the ATT, is the smaller of $\mathcal{O}p(\sqrt{N_{ctrl}})$ and $\mathcal{O}p(\sqrt{N_{treat}T_{pre}})$, so large T_{pre} and N_{ctrl} would be necessary for us to get the accurate estimation.

In the empirical application, we use this newly developed method to assess the impact of Brexit on foreign direct investment (FDI) to the U.K. We use 9 covariates that we consider have predictive power over FDI.

2 Literature Review

Causal inference in economics and other social sciences is frequently complicated by the absence of counterfactuals, which are 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 missing data imputation methods, although some are more explicit than others. For instance, under certain assumptions, the matching method (Abadie and Imbens (2006, 2011)) explicitly imputes the missing counterfactual for treated units with meticulously selected controls. The DID method (Card and Krueger (1993); Ashenfelter (1978)), on the other hand, implicitly imputes the missing counterfactual by differencing the control units before and after treatment. Meanwhile, the SCM method explicitly imputes the missing counterfactual with a weighted average of control units. Our method aligns with the recent trend in the causal inference literature, aiming to explicitly impute the missing counterfactual by modeling the entire DGPs, a strategy highlighted by Athey et al. (2021) with their matrix completion (MC) method, and Xu (2017) with their CSC-IFE method.

As another branch of causal inference, modeling entire DGPs offers distinct advantages. This approach helps to overcome the constraints imposed by untestable and stringent assumptions, such as unconfoundedness and common support in matching methods (Rosenbaum and Rubin (1983); Rubin (1997)), as well as the parallel trends assumption in difference-in-differences (DID) models (Card and Krueger (1993)). Additionally, it addresses the limitations of the original Synthetic Control Method (SCM) (Abadie et al. (2010)) and its variants (Ben-Michael et al. (2021), Arkhangelsky et al. (2021)), which require the outcomes of treated units to lie within or near the convex hull formed by the control units.

Factor 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 these methods specifically for predicting missing counterfactuals in synthetic control settings, followed by Gobillon and Magnac (2016), Xu (2017), Chan et al. (2016), and Li (2018). Conventional factor models with static factor loadings fail to capture time-varying factor loadings that arise due to changes in a unit's fundamentals. Kelly et al. (2020) was the first to incorporate time-varying factor loadings by instrumenting them with covariates. The IPCA method has been successfully applied to stock return prediction by Kelly et al. (2019), demonstrating significant accuracy in out-of-sample predictions. Our paper is the first to apply this method to causal inference within the relevant literature.

This paper is structured as follows. Section 3 introduces the framework of the CSC-IPCA method, detailing the functional form and assumptions for identification. Section 4 outlines the estimation procedures, including hyperparameter tuning and inference. Section 5 presents the results of Monte Carlo simulations, comparing different estimation methods and providing finite sample properties. Section 6 demonstrates the application of the CSC-IPCA method in a real-world setting, evaluating the impact of Brexit on foreign direct investment (FDI) in the U.K. Section 7 concludes the paper with a summary of the main findings and potential future research directions. More detailed proofs and derivations are provided in Appendix B.

3 Framework

Consider Y_{it} as the observed outcome for a specific unit i ($i = 1, \dots, N$) at time t ($t = 1, \dots, T$). The total number of observed units in the panel is $N = N_{treat} + N_{ctrl}$, where N_{treat} represents the number of units in the treatment group $\mathcal{T}$ and N_{ctrl} represents 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 after treatment. We observe the treatment effect at $T_{pre} + 1$ right after the beginning of the treatment and continue to observe thereafter until the end of the observation periods, a scenario commonly referred to as block assignment[2]. Following Equations 2 and 3, we assume that the outcome variable Y_{it} is given by a simple factor model with factor loadings instrumented by covariates. The functional form is given by:

Assumption 1 *Functional form:*

$$\begin{aligned} Y_{it} &= D_{it} \circ \delta_{it} + \Lambda_{it} F_t' + \mu_{it}, \\ \Lambda_{it} &= X_{it}\Gamma + H_{it} \end{aligned} \tag{4}$$

The primary distinction of this functional form from existing fixed effects 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 a vector of K unobserved common factors, and $\Lambda_{it} = [\lambda_{it}^1, \ldots, \lambda_{it}^K]$ represents a vector of factor loadings. Meanwhile, the vector $X_{it} = [x_{it}^1, \ldots, x_{it}^L]$ comprises L observed 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 structural component $\Lambda_{it} F_t$ between common factors and factor loadings; the linear part of covariates $X_{it}\beta$ (as specified in Equation 1) is excluded from the functional form. The logic behind this is that we believe the unit-specific factor loadings, instrumented by covariates, have included all the predictive information from these predictive covariates. This functional form exhibits two major advantages

[2] We can also adopt this method for the more commonly observed staggered adoption scenario. We demonstrate different treatment assignment mechanisms in Appendix **??**

over the CSC-IFE model. Firstly, it suffers less from the risk of model miss-specification, as the model is sufficiently simpler. Secondly, it incorporates a dimension reduction operation via the matrix Γ, which allows us to handle high-dimensional datasets, especially when dealing with a large number of covariates. Thirdly, instead of accommodating predictive covariates as regressors we believe instrumenting factor loadings with covariates can better abstract information for outcome prediction.

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 unit i in the group of treated $\mathcal{T}$ and for period $t > T_{pre}$, with all other D_{it} set to 0. The model easily accommodates variations in treatment timing by removing the constraint that treatment must commence simultaneously for all treated units. The term μ_{it} signifies the idiosyncratic error associated with the outcome variable Y_{it}. Additionally, $H_{it} = [\eta_{it}^1, \ldots, \eta_{it}^K]$ constitutes the vector of error terms linked to K unobserved factor loadings.

Following Neyman (1932) potential outcome framework (also discussed by Rubin (1974, 2005)), we observe the actual outcome for the treated and untreated units for the entire period. If we combine the two components in Equation 4, we get the actual outcomes for treated and controls distinguishied by different Γ and treatment assignments, as presented in the following:

$$\begin{cases} Y_{it}^1 = \delta_{it} + X_{it}\Gamma_{treat}F_t' + \epsilon_{it} & if\ i \in \mathcal{T}\ \&\ t > T_{pre} \\ Y_{it}^0 = X_{it}\Gamma_{ctrl}F_t' + \epsilon_{it} & if\ i \in \mathcal{C}. \end{cases} \tag{5}$$

where Equation 5 represents the actual outcome for the treated and control units combined the two parts together in Equation 4. Our goal is to impute the missing counterfactual $\hat{Y}_{it}^0 = X_{it}\hat{\Gamma}_{treat}\hat{F}_t$ for the treated units $i \in \mathcal{T}$ when $t > T_{pre}$, where the $\hat{\Gamma}_{treat}$ and $\hat{F}_t$ are estimated parameters. We then calculate the ATT as the difference between the actual outcome and the imputed missing counterfactuals, which is defined as:

$$\widehat{ATT}_t = \frac{1}{N_{treat}} \sum_{i \in \mathcal{T}} \left(Y_{it}^1 - \hat{Y}_{it}^0 \right) = \frac{1}{N_{treat}} \sum_{i \in \mathcal{T}} \hat{\delta}_{it}. \tag{6}$$

4 Estimation

5 Monte Carlo Simulation

6 Empirical application

7 Conclusion

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.

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.

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.

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 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.

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.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

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

Donald B Rubin. Estimating causal effects from large data sets using propensity scores. Annals of internal medicine, 127(8_Part_2):757–763, 1997.

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

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 Technical Details

Appendix B Formal Result

Appendix C Simulation Study

Appendix D Empirical Applicatoin