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

Cong Wang†

September 8, 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_{tr} 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).

In the formal result, we derive the asympotic properties based on the unbaisedness and efficient estimation of both Γ and F_t. We show that the convergence rate of our estimand is the smaller one between $\mathcal{O}_p(\sqrt{N_{ctrl}})$ and $\mathcal{O}_p(\sqrt{N_{treat}T_{pre}})$.

References

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.

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

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.