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

Cong Wang

Sapienza University of Rome

September 5, 2024

Background & Motivation

All the methods for causal inference can be viewed as missing data imputation methods, where some are more explicit than others. – Imbens and Rubin (2015).

- Matching method explicity impute the missing counterfactual of trated X_i by control $X_{j(i)}$.
- Difference-in-difference (DID) method implicitly impute the missing counterfactual by differencing the treated and controls before and after the treatment.
- Novel synthetic control method (SCM) explicitly impute the missing counterfactual of Y_{it} with a weighted average of control units, without extrapolation.

$$\hat{Y}_{it} = \sum_{j=1}^{J} W_j Y_{jt}$$

Impute Counterfactual by Modeling the DGPs

- After transforming the causal inference problem into a missing data imputation problem, it is natural to think about modeling the data generating process (DGPs).
- Factor model is popular for its flexibility and sparsity in DGPs modeling. (Bai (2003), Stock and Watson (2002), etc.)
- Recent advance in asset pricing using IPCA for stock return prediction (Kelly et al. (2020))

A brief history of using the factor model for causal inference:

1. Pure factor models. Hsiao et al. (2012) appear to be the first to use factor models for causal inference.

$$y_{it} = \lambda_i F_t + \epsilon_{it}$$

2. Interactive fixed effects model. Xu (2016) use factor model plus regression terms.

$$y_{it} = \lambda_i F_t + X_{it}\beta + \epsilon_{it}$$

Set up

- Y_{it} is the observed outcome for unit $i = 1, 2, \cdots, N$ at time $t = 1, \cdots, T$.
- Total number of observed units is $N = N_{treat} + N_{ctrl}$, for N_{ctrl} number of units in the control group $\mathcal{C}$ and N_{treat} units in the treated group $\mathcal{T}$.
- Each unit is observed over $T = T_{pre} + T_{post}$ periods.
- Our target estimand is the average treatment effect for the treated units (ATT) in the post-treatment periods.

Assumption

Functional form:

$$Y_{it} = D_{it} \circ \delta_{it} + \Lambda_{it} F_t' + \mu_{it}$$
$$\Lambda_{it} = X_{it}\Gamma + H_{it}$$

Set up

- where $X_{it} = [x_{it}^1, \cdots, x_{it}^L]$ is a vector of observed covariates. $F_t = [f_t^1, \cdots, f_t^K]$ is a vector of unobserved time-varying factors. $\Lambda_{it} = [\lambda_{it}^1, \cdots, \lambda_{it}^K]$ is a vector of unobserved factor loading instrumented by covariates X_{it}.
- We use Neyman (1932) and Rubin (2003) potential outcome framework to specify the potential outcome for treated and control units:

$$\begin{cases} Y_{it}^1 = \delta_{it} + X_{it}\Gamma F_t' + \epsilon_{it} & \text{if } i \in \mathcal{T} \ \& \ t > T_{pre} \\ Y_{it}^0 = X_{it}\Gamma F_t' + \epsilon_{it} & \text{otherwise.} \end{cases}$$

Treatment Assignment

- To simplify the estimation, we focus on the block assignment scenario where all the treated units are treated at the same time (can be relaxed) and the treatment once turned on can not be turned off.

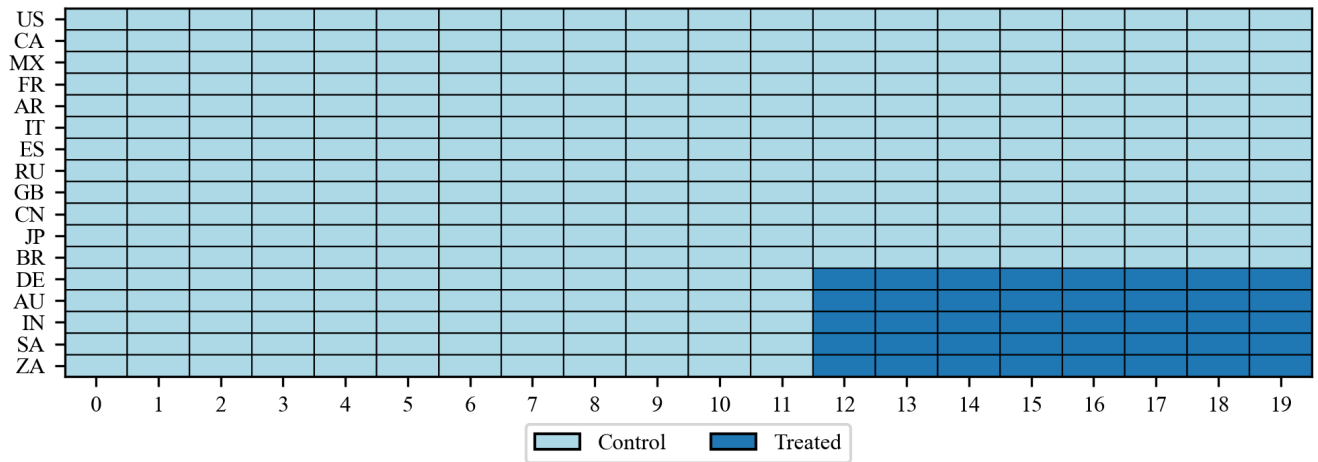
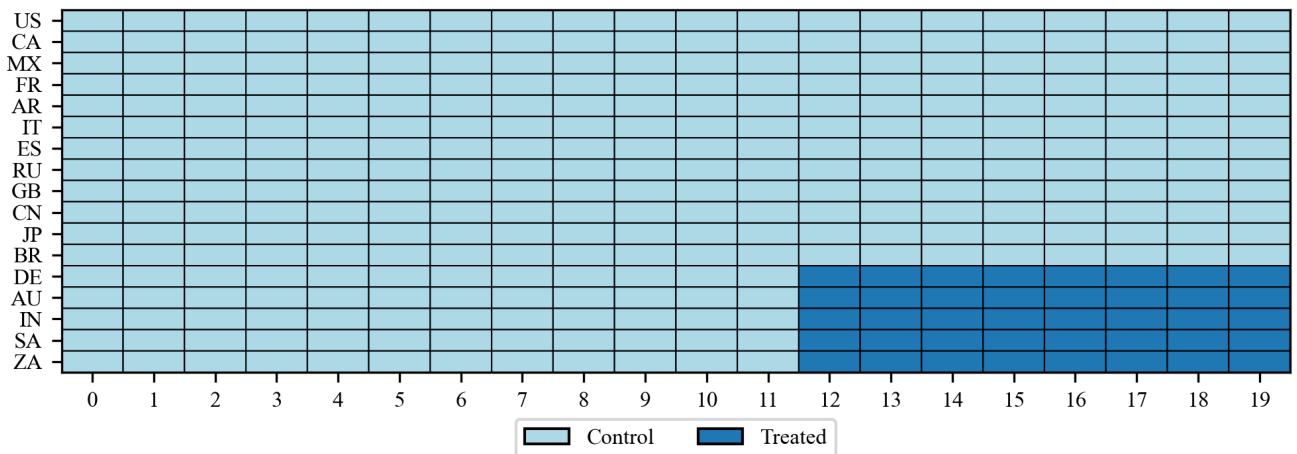


Figure: Block assignment scenario

Estimand

Under Nemany's potential outcome framework, our estimand ATT can be expressed as:

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



Advantages of the CSC-IPCA Method

- It incorporates the strengths of previous SCM approaches, such as the relaxation of the untestable parallel trends assumption (PTA).
- It complements the SCM method when targeted outcomes are outside the convex hull formulated by controls.
- It eliminates the need for correct model specification.
  1. Estimates data with different DGPs other than the functional form, our method still excels.
  2. Covariates are incorporated into the model through a mapping matrix Γ instead of a specific model specification.

Advantages of the CSC-IPCA Method

- It inherits the ability of PCA to effectively handle high-dimensional data and enhances the value extracted from numerous covariates.
  1. The mapping matrix Γ conducts a dimensional reduction process, making it easier to handle high-dimensional financial data.
  2. Prediction information from covariates is more effectively handled. When there are unobserved covariates, this method is less biased compared to its peers.
- It generates time-varying factor loadings which have better economic interpretation and better estimate for common factors.
  1. Instrumented factor loadings Λ_{it} inherit time-varying properties, making them more realistic in practice.
  2. Dynamic factor loadings help us better estimate common factors.

Estimation

To combine the functional form we get the following structure component:

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

CSC-IPCA method is estimated by minimizing the sum of squared residuals of the following objective function:

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

- The optimization, as defined in the equation above, is quadratic with respect to either Γ or $\boldsymbol{f}_t$, when the other is held constant.
- We can use alternating least squares (ALS) method for the numerical solution of this optimization problem.

Estimation

Step 1: Estimate the common factors $\hat{F}_t$ and the mapping matrix $\hat{\Gamma}_{\text{ctrl}}$ with an ALS algorithm, based exclusively on data from the control group for the whole time period.

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

With a fixed Γ, the solutions for $\boldsymbol{f}_t$ are t-separable and can be obtained via cross-sectional OLS for each t:

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

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

$$\hat{\gamma} = \left(\sum_{i,t} (\boldsymbol{x}_{it}' \otimes \boldsymbol{f}_t)(\boldsymbol{x}_{it} \otimes \boldsymbol{f}_t')\right)^{-1} \left(\sum_{i,t} (\boldsymbol{x}_{it}' \otimes \boldsymbol{f}_t) y_{it}\right).$$

Estimation

Step 2: Estimate 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.

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

The Γ_{treat} is estimated through:

$$\hat{\gamma} = \left(\sum_{i,t} (\boldsymbol{x}_{it}' \otimes \boldsymbol{f}_t)(\boldsymbol{x}_{it} \otimes \boldsymbol{f}_t') \right)^{-1} \left(\sum_{i,t} (\boldsymbol{x}_{it}' \otimes \boldsymbol{f}_t) y_{it} \right).$$

$$for\ i \in \mathcal{T}, T <= T_{pre}.$$

Estimation

- The estimation of $\boldsymbol{f}_t$ and Γ is not deterministic.
- We can find any arbitrary rotation matrix R, such that $\boldsymbol{x}_{it}\Gamma R R^{-1}\boldsymbol{f}'_t$ yields the same structural component.
- We put specific constraints on the mapping matrix $\Gamma_{norm} = \Gamma_{treat} R$ and factor $\boldsymbol{f}_{norm} = R^{-1}\boldsymbol{f}_t$ for identification.

Estimation

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

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

1. Cholesky decomposition to get a upper triangular matrix $R_1 = cholesky(\Gamma'\Gamma)$,
2. Singular value decomposition on $R_1 \boldsymbol{f}_t \boldsymbol{f}'_t R'_1$ to get $R_2 = U$ where $U \Sigma V' = svd(R_1 \boldsymbol{f}_t \boldsymbol{f}'_t R'_1)$.
3. Finally, the rotation matrix R is given by: $R = R_1^{-1} R_2$.

Estimation

Step 4: The final step involves imputing the counterfactual outcome $\hat{Y}_{it}^0$ 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:

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

The estimated average treatment effect for treated is:

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

Hyperparameter tuning

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;

Inference

We use conformal inference (Chernozhukov et al. (2021)) to construct the confidence interval.

1. We 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 $\tilde{Y}_{it} = Y_{it} - \theta_{it}$.
2. Following the estimation procedure to estimate the time-varying factor F_t with only control data as before, and update the Γ for the newly adjusted treated units with the **entire set of treated units**.
3. 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:

$$S(\hat{\mu}) = \left(\frac{1}{\sqrt{T_{post}}} \sum_{t > T_{pre}} |\hat{\mu}|^q \right)$$

Inference

4. We employ $q = 1$ for the permanent intervention effect as designed in our study.
5. Block permute the residuals and calculate the test statistic in each permutation. The P-value is defined as:

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

6. Repeat the above procedures with different nulls to get different P-values and construct confidence intervals at different significance levels.

Simulation

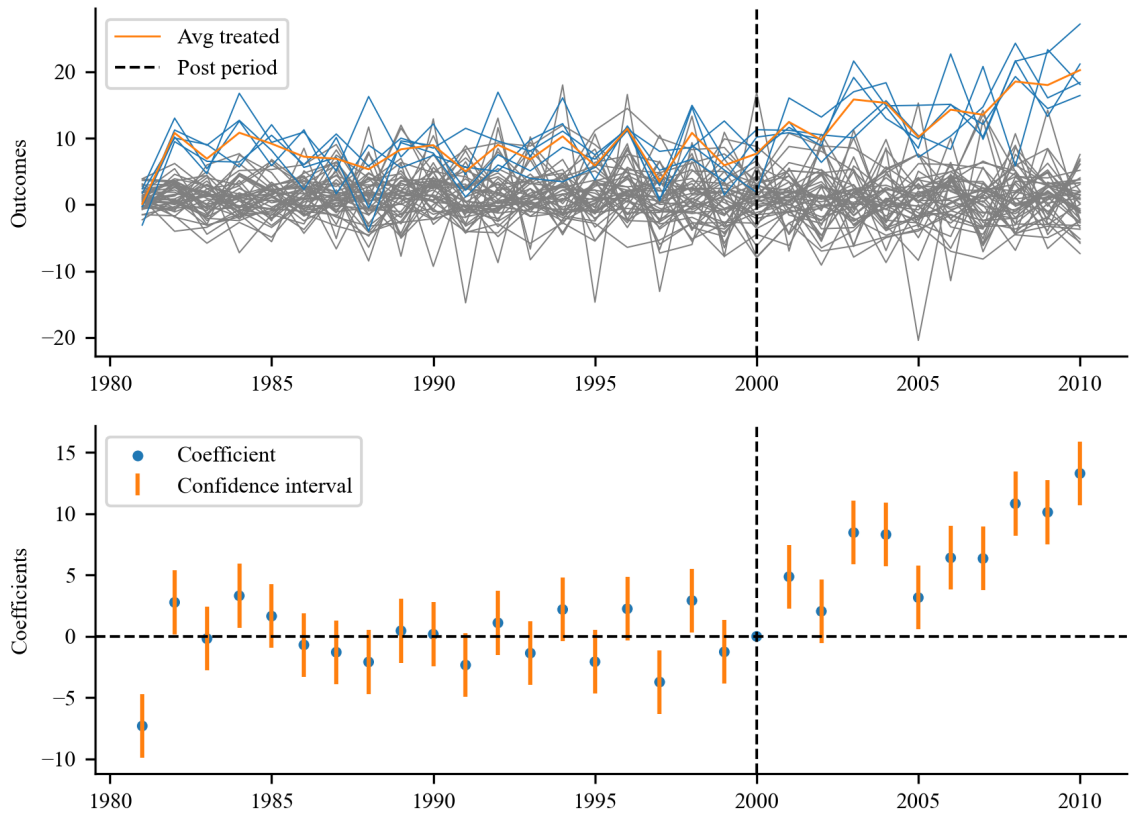
We use the following DGPs to simulate the data:

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

- $L = 10$ and $K = 3$.
- $X_{it} = [x_{it}^1, \ldots, x_{it}^L]$ 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.
- $F_t = [f_t^1, \ldots, f_t^3]$ denotes the vector of time-varying common factors, adhering to a similar VAR(1) process.
- The coefficient vector $\beta = [\beta^1, \ldots, \beta^L]$ associated with the covariates is drawn uniformly from $(0, 1)$.
- 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. The heterogeneous treatment effect is modeled as $\delta_{it} = \bar{\delta}_{it} + e_{it}$.
  $\bar{\delta}_t = [0, \cdots, 0, 1, 2, \ldots, T_{post}]$ represents a time-varying treatment effect.

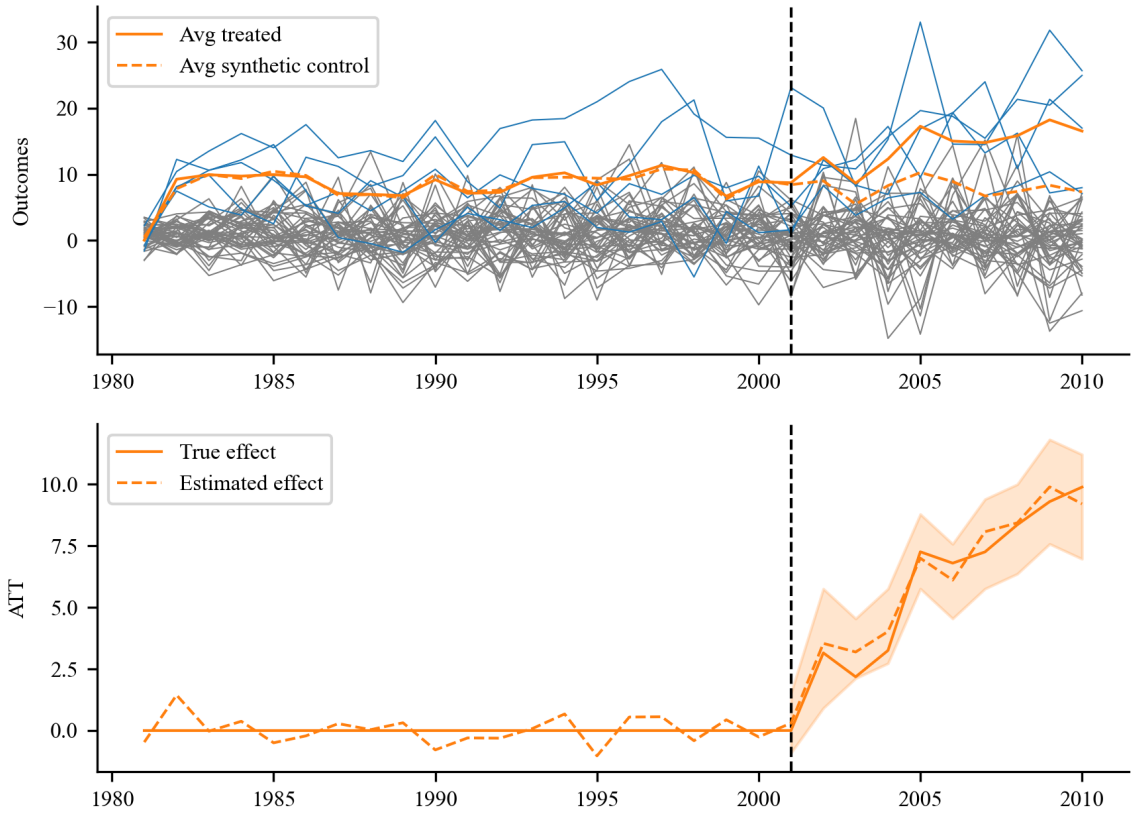
Simulation

- There is a possibility that treated units are not in the convex hull formulated by controls.
- From the simple event study plot, the parallel trend assumption is not satisfied.



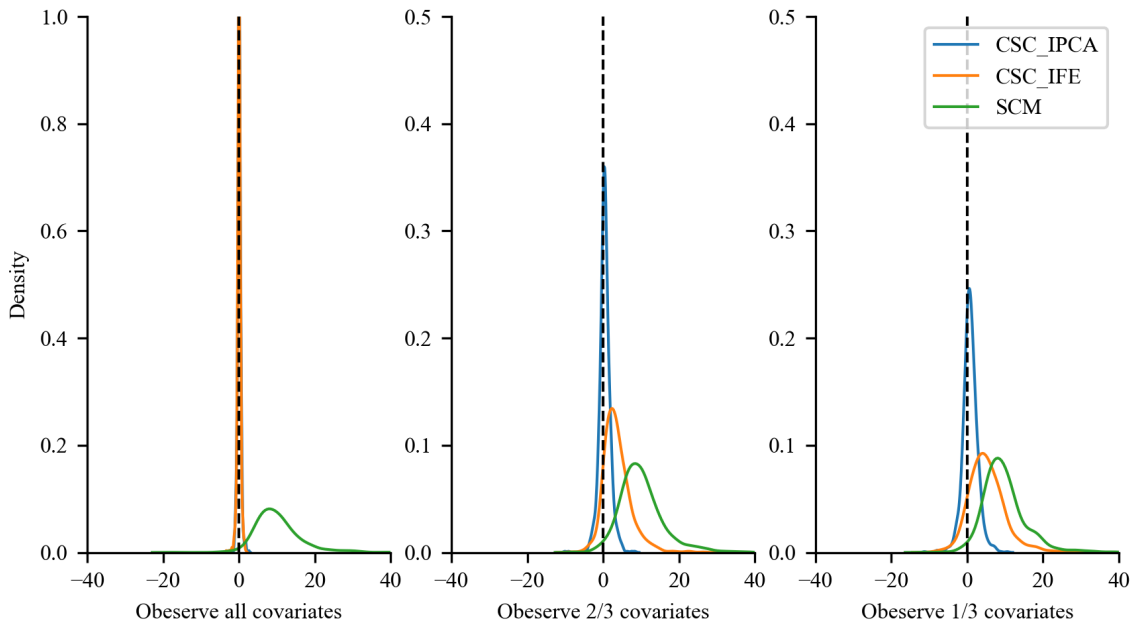
An example

- The upper panel shows the average synthetic control's outcome perfectly overlaps with the actual average treated outcome before the treatment.
- The lower panel shows the average treatment effect before and after treatment with a 90% confidence interval.



Bias comparison

- When all covariates are observed, both CSC-IPCA and CSC-IFE demonstrate unbiasedness and effectively estimate the true ATT.
- SCM exhibits an upward bias for the poor pre-treatment fit.
- As the number of unobserved covariates increases, both CSC-IPCA and CSC-IFE lose efficiency, but the CSC-IPCA estimator remains less unbiased than CSC-IPCA estimator.



Finite sample property

- The bias, RMSE, and STD are estimated based on 1000 simulations.
- $N_{treat} = 5$, $T_{post} = 5$, $L = 10$. We vary the number of control units N_{ctrl}, pre-treatment period T_{pre}, and the proportion of observed covariates α.
- The convergence rate of the CSC-IPCA estimator is the smaller one of $\mathcal{O}_p\left(\sqrt{N_{ctrl}}\right)$ and $\mathcal{O}_p\left(\sqrt{N_{treat}\,T_{pre}}\right)$.

Table 1: **Finite Sample Properties**

α		1/3	2/3	1	1/3	2/3	1	1/3	2/3	1
T_0	N_{ctrl}	Bias			RMSE			STD		
10	10	2.328	0.703	0.130	4.770	3.068	1.642	4.175	3.032	1.684
10	20	1.367	0.312	0.053	3.484	2.209	0.914	3.260	2.219	1.008
10	40	1.026	0.196	0.051	2.776	1.752	0.714	2.616	1.781	0.821
20	10	2.957	1.029	0.217	4.817	2.696	1.135	3.814	2.544	1.179
20	20	1.435	0.438	0.055	3.280	1.754	0.745	2.982	1.773	0.860
20	40	1.093	0.167	0.042	2.613	1.348	0.602	2.430	1.409	0.757
40	10	2.905	1.232	0.145	4.911	3.035	0.969	3.972	2.797	1.065
40	20	1.670	0.399	0.019	3.592	1.718	0.724	3.221	1.737	0.861
40	40	0.876	0.295	0.006	2.675	1.418	0.574	2.556	1.441	0.697

Identification assumptions

Assumption

Assumption for consistency:

1. *Covariate orthogonality:* $E\left[\boldsymbol{x}_{it}'\epsilon_{it}\right] = \mathbf{0}_{L\times 1}$,
2. *The following moments exist:* $\mathrm{E}\|\boldsymbol{f}_t\boldsymbol{f}_t'\|^2$, $\mathrm{E}\|\boldsymbol{x}_{it}'\epsilon_{it}\|^2$, $\mathrm{E}\|\boldsymbol{x}_{it}'\boldsymbol{x}_{it}\|^2$, $\mathrm{E}\left[\|\boldsymbol{x}_{it}'\boldsymbol{x}_{it}\|^2\|\boldsymbol{f}_t\boldsymbol{f}_t'\|^2\right]$,
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,* $\boldsymbol{x}_{it}$ *is bounded, and define* $\Omega_t^{xx} := \mathrm{E}\left[\boldsymbol{x}_{it}'\boldsymbol{x}_{it}\right]$, *then almost surely,* $\Omega_t^{xx} > \epsilon$ *for some* $\epsilon > 0$.

Identification assumption

Assumption

Assumptions for asymptotic normality:

1. *As* $N, T \to \infty$, $\frac{1}{\sqrt{NT}} \sum_{i,t} vect\left(\boldsymbol{x}'_{it} \epsilon_{it} \boldsymbol{f}'_t\right) \xrightarrow{d} Normal\left(0, \Omega^{x\epsilon f}\right)$,
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* $N, T \to \infty$, $\frac{1}{\sqrt{T}} \sum_t vect\left(\boldsymbol{f}_t \boldsymbol{f}'_t - \mathrm{E}[\boldsymbol{f}_t \boldsymbol{f}'_t]\right) \xrightarrow{d}$ $Normal\left(0, \Omega^f\right)$.
4. *Bounded dependence:* $\frac{1}{NT} \sum_{i,j,t,s} \|\tau_{ij,ts}\| < \infty$, *where* $\tau_{ij,ts} := \mathrm{E}\left[\boldsymbol{x}'_{it} \epsilon_{it} \epsilon'_{js} \boldsymbol{x}_{js}\right]$
5. *Constant second moments of the covariates:* $\Omega^{xx}_t = \mathrm{E}\left[X_t X'_t\right]$ *is constant across time periods.*

Formal result

we can formulate a target function for Γ as follows:

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

The Hessian matrix $H(\Gamma)$ is defined as the second derivative of the target function $G(\Gamma)$ with respect to Γ: $H(\Gamma) = \frac{\partial^2 G(\Gamma)}{\partial \Gamma \partial \Gamma'}$.
To satisfy the normalization criteria, we define the following identification function:

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

where $V^{ff} = E\left[\boldsymbol{f}_t \boldsymbol{f}'_t \right]$, meanwhile, veca$(\cdot)$ and vecb$(\cdot)$ vectorize the upper triangular entries of a square matrix.
The Jacobian matrix $J(\Gamma)$ as the derivative of the identification function $I(\Gamma)$ with respect to Γ: $J(\Gamma) = \frac{\partial I(\Gamma)}{\partial \Gamma}$.

Formal result

Proposition

Under the above assumptions, 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* $T/N \to \infty$,

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

where $H^0 := \frac{\partial S(\Gamma)}{\partial \gamma}|_{\gamma=\gamma^0}$ and $J^0 := \frac{\partial I(\Gamma)}{\partial \gamma}|_{\gamma=\gamma^0}$,
$\mathbb{V}^{[1]} = \left(Q^0 \otimes \mathcal{I}_K\right) \Omega^{x\epsilon f} \left(Q^{0\prime} \otimes \mathcal{I}_K\right)$, 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 the normality assumption.
Proof: refer to Kelly et al. (2020)

Formal result

Proposition

Under the Assumptions, 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 $\forall t$,

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

Proof: Decompose the left-hand side equation:

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

where $\tilde{\epsilon}_t$ is the estimated error term with estimated Γ and true $\boldsymbol{f}_t$. 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:

$$\begin{aligned}\sqrt{N}\left(\hat{\Gamma}' X_t' X_t \hat{\Gamma}\right)^{-1} \hat{\Gamma}' X_t' \boldsymbol{\epsilon}_t &= \sqrt{N}\left(\Gamma' X_t' X_t \Gamma\right)^{-1} \Gamma' X_t' \boldsymbol{\epsilon}_t + \mathcal{O}_p(1) \\ &\xrightarrow{d} \mathit{Normal}(0, \mathbb{V}_t^{[2]})\end{aligned}$$

Formal result

Theorem

Under Assumptions, the CSC-IPCA estimator $\mathrm{E}\left(\widehat{ATT}_t | D, X, \Gamma, F\right) \xrightarrow{P} ATT_t$, *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:

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

The third step converts the vector-matrix multiplication into vector multiplications with the Kronecker product, $\boldsymbol{x}_{it}\Gamma \boldsymbol{f}_t' = \boldsymbol{x}_{it}(\mathcal{I}_L \otimes \boldsymbol{f}_t)\boldsymbol{\gamma}$.

Formal result

The bias of the estimated ATT is the sum of four terms $A_{1,it}$, $A_{2,it}$, $A_{3,it}$, and ϵ_{it}. By proposition 1 and 2, we have the following results:

$$A_{1,it} = \boldsymbol{x}_{it} E_\Gamma \boldsymbol{f}_t' = \mathcal{O}_p\left(1/\sqrt{N_{treat}\, T_{pre}}\right).$$

$$A_{2,it} = -\boldsymbol{x}_{it} \Gamma \boldsymbol{e}_{f_t}' = \mathcal{O}_p\left(1/\sqrt{N_{ctrl}}\right).$$

$$A_{3,it} = -\boldsymbol{x}_{it} E_\Gamma \boldsymbol{e}_{f_t}' = \mathcal{O}_p\left(1/\sqrt{N_{treat}\, T_{pre} N_{ctrl}}\right).$$

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

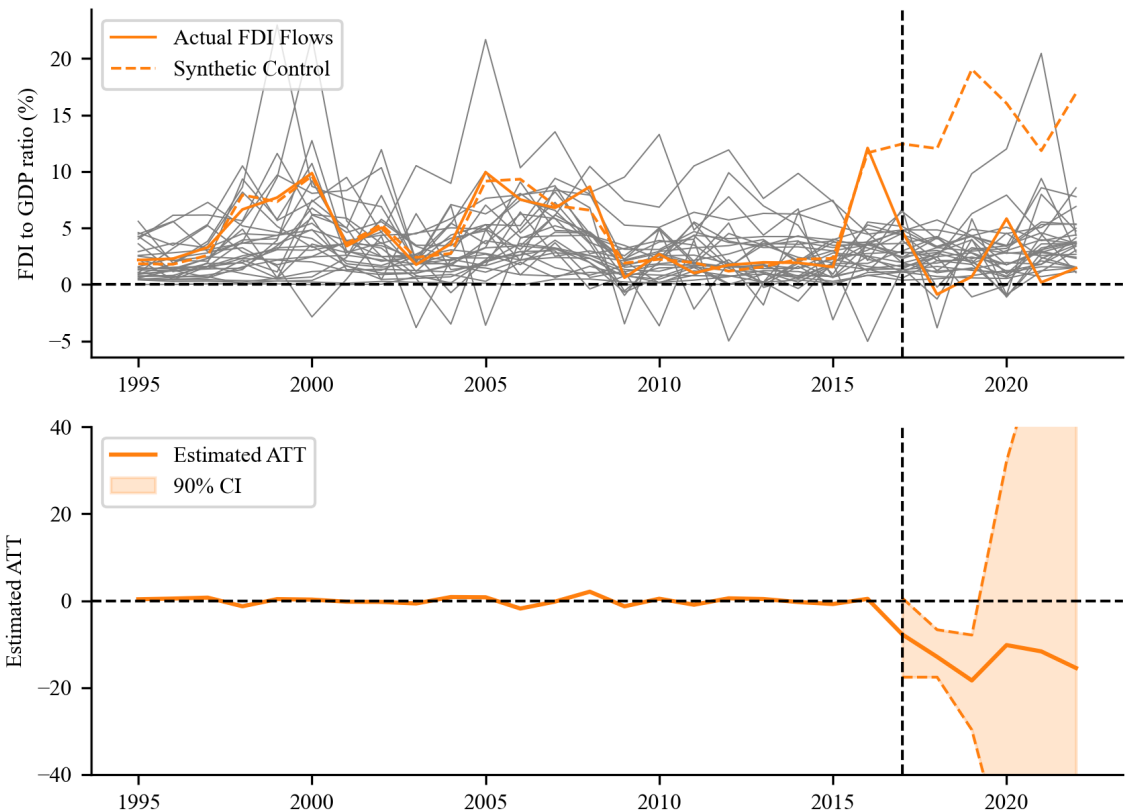
$$\begin{aligned}\hat{\delta}_{it} - \delta_{it} &= \mathcal{O}_p\left(\frac{1}{\sqrt{N_{treat}\, T_{pre}}}\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N_{ctrl}}}\right) + \mathcal{O}_p\left(\frac{1}{\sqrt{N_{treat}\, T_{pre} N_{ctrl}}}\right) + \mathcal{O}_p(1) \\ &= \mathcal{O}_p\left(\frac{1}{\sqrt{N_{ctrl}}}\right)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

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

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

Case study – Brexit on FDI in the UK

- We use OECD countries as control units and the UK as the treated unit.
- The treatment period is from 2017, and the pre-treatment period is from 2016 to 1995.
- The outcome variable is the foreign direct investment (FDI) inflow.
- The covariates include GDP, imports and exports, inflation, investment, employment, and demographic indicators.



Future Direction

- Solve overfitting.
- Better handle bad controls.