Matrix Completion Methods for Causal Panel Data Models

Susan Athey et al. (2021), JASA

Naoki Eguchi

2025.6.25 ミクロ計量経済学

Faculty of Medicine, Kyoto University

Introduction

What is ”matrix completion” ?

- Matrix completion (MC) is an imputation method for the missing in the matrix.
- In PCA (Principle Compoent Analysis), given M principle components, we approximate a true data matrix $\mathbf{Y} \in \mathbb{R}^{I \times J}$ by $\mathbf{A} \in \mathbb{R}^{I \times M}$ and $\mathbf{B} \in \mathbb{R}^{M \times J}$.

$$\min_{\mathbf{A}, \mathbf{B}} \left\{ \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \sum_{m=1}^{M} a_{im} b_{mj} \right)^2 \right\}$$

- If there is missing in matrix, using only observed data, we can approximate the original matrix, hence we can impute the missing by $\hat{\mathbf{A}}, \hat{\mathbf{B}}$.

$$\min_{\mathbf{A}, \mathbf{B}} \left\{ \sum_{(i,j) \in \mathcal{O}} \left(Y_{ij} - \sum_{m=1}^{M} a_{im} b_{mj} \right)^2 \right\}$$

- Applying this idea to panel data analysis, MC-based panel method is seen to a generalization of many model-based (factor-regression-based) panel methods.

Where to impute

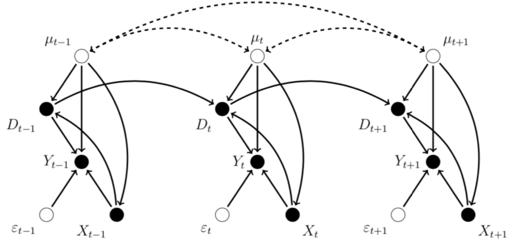
- As many panel data methods, we want to know ATT : $\mathbb{E}[Y_{it}(1) - Y_{it}(0)|W_i = 1]$.
- Thus, it boils down to estimate (impute) the counterfactual $Y_{it}(0)$.
    - Horizontal : Under unconfoundedness, we can impute counterfactual PO using observed outcomes for control units.
    - Vertical : By SCM, we can also impute it using weighted average outcomes for control units with most predictive weights trained with pre-treatment datas.

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark \\ \vdots & \vdots & \vdots \\ \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & ? \\ \vdots & \vdots & \vdots \\ \checkmark & \checkmark & ? \end{pmatrix}.$$

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \dots & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & ? & \dots & ? \end{pmatrix}$$

Xu (2024): Counterfactual estimation

- functional form: $Y_{it}(0) = f(\mathbf{X_{it}}) + h(\mathbf{U_{it}}) + \epsilon_{it}$
  $\rightarrow$No anticipation, carryover, feedback, LDV
- strict exogeneity: $\forall i, j \in \{1, \ldots, N\}, \forall s, t \in \{1, \ldots, T\}, \epsilon_{it} \perp\!\!\!\perp \{D_{js}, \mathbf{X_{js}}, \mathbf{U_{js}}\}$



- low-dimensional decomposition: $h(\mathbf{U_{it}}) = \{L_{it}\}, \mathrm{rank}(\mathbf{L}_{N\times T}) \ll \min\{N, T\}$
  $\rightarrow$ The rank (= number of factors) is FIXED !!
- MC panel method is a generalization of Xu's counterfactual estimation.

Set Up

Notation and Estimand

- Consider a setting with N units observed over T periods characterized by a binary treatment W_{it} and hence two POs $Y_{it}(1), Y_{it}(0)$.
    - $\mathbf{X} \in \mathbb{R}^{N \times P}$, $\mathbf{Z} \in \mathbb{R}^{T \times Q}$: observe (unit / time)-specific covariance matrix
- Estimand : $\mathbf{Y} = \{Y_{it}(0)^{1}\} = \begin{pmatrix} Y_{11}(0) & \cdots & Y_{1T}(0) \\ \vdots & \ddots & \vdots \\ Y_{N1}(0) & \cdots & Y_{NT}(0) \end{pmatrix}$ ($\leftarrow$ Matrix!!)
- $W_{it} = \begin{cases} 1 & \text{if } (i,t) \in \mathcal{M} : \text{Missing indice} \\ 0 & \text{if } (i,t) \in \mathcal{O} : \text{Observed indice as training data} \end{cases}$

[1] 以降は簡単のため，$Y_{it}(0) = Y_{it}$ とし，“(0)” を省略して表記する．

Patterns of data matrix

- Ordinary case (rich data wrt. units and times)

$$
\mathbf{Y}_{N\times T} = \begin{pmatrix}
\checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\
\checkmark & \checkmark & \checkmark & ? & \dots & ? \\
\checkmark & \checkmark & \checkmark & ? & \dots & ? \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\checkmark & \checkmark & \checkmark & ? & \dots & ?
\end{pmatrix}
$$

- Staggered adoption

$$
\mathbf{Y}_{N\times T} = \begin{array}{cl}
\begin{pmatrix}
\checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark & \dots & ? \\
\checkmark & \checkmark & ? & ? & \dots & ? \\
\checkmark & \checkmark & ? & ? & \dots & ? \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\checkmark & ? & ? & ? & \dots & ?
\end{pmatrix} &
\begin{array}{l}
\text{(never adopter)} \\
\text{(late adopter)} \\
\\
\text{(medium adopter)} \\
\\
\text{(early adopter)}
\end{array}
\end{array}
$$

Horizontal regression and unconfoundedness : thin matrix ($N \gg T$)

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark \\ \vdots & \vdots & \vdots \\ \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & ? \\ \vdots & \vdots & \vdots \\ \checkmark & \checkmark & ? \end{pmatrix}.$$

1. Regress the last period outcome on the lagged outcomes. (among untreated)
2. Predict the missing POs using the estimated regression.

$$\forall (i,T) \in \mathcal{M},\ \hat{Y}_{iT} = \hat{\beta}_0 + \sum_{t=1}^{T-1} \hat{\beta}_t Y_{it}, \text{ where } \hat{\beta} = \arg\min_{\beta} \sum_{i:(i,T)\in\mathcal{O}} (Y_{iT} - \beta_0 - \sum_{t=1}^{T-1} \beta_t Y_{it})^2.$$

$\rightarrow$ Nonparametrically,

Vertical regression and synthesis control : fat matrix ($T \gg N$)

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \dots & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & ? & \dots & ? \end{pmatrix}$$

1. Regress the outcomes for treated unit prior to the treatment on the outcomes for the control units in the same periods.
2. Predct the missing POs using the estimated regression.

$$\forall (N,t) \in \mathcal{M},\ \hat{Y}_{Nt} = \hat{\gamma}_0 + \sum_{i=1}^{N-1} \hat{\gamma}_i Y_{it}, \text{ where } \hat{\gamma} = \arg\min_{\gamma} \sum_{t:(N,t)\in\mathcal{O}} (Y_{Nt} - \gamma_0 - \sum_{i=1}^{N-1} \gamma_i Y_{it})^2.$$

$\rightarrow$ Vertical regression is generalization of ADH(2010) in that it relaxes two restrictions :

- the coefficients $\hat{\gamma}$ are nonnegative. (Interpretability ; What is a negative weight?)
- the intercept in this regression is 0. (This is seen to be plausible in recent literatures.)

Matrix Completion

Model

- Under no covariates, we model the $N \times T$ matrix of complete matrix $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{L}^* + \epsilon, \text{ where } \mathbb{E}[\epsilon|\mathbf{L}^*] = 0.$$

Assumption 1

- ϵ is independent of $\mathbf{L}^*$ (strict exogeneity)
- The element of ϵ are $\sigma - sub - Gaussian$ and independent each other.
  $\Leftrightarrow \forall t,\ \mathbb{E}[\exp(t\epsilon)] \leq \exp(\frac{\sigma^2 t^2}{2})$.

- The goal is to estimate the matrix $\mathbf{L}^*$. (low-rank assumption)
  $\rightarrow$ Note that two types[2] of fixed effects are included.

[2] これら以外にも Interactive fixed effect といったあらゆる factor を"少数まで"許容する

MC-NNM (Matrix Completion with Nuclear Norm Minimization) estimator

- MC-NNM estimator for $\mathbf{L}^*$ is given by $\hat{\mathbf{L}} + \hat{\Gamma}\mathbf{1}_T^\top + \mathbf{1}_N\hat{\Delta}^\top$

$$(\hat{\mathbf{L}}, \hat{\Gamma}, \hat{\Delta}) = \arg\min_{\mathbf{L},\Gamma,\Delta}\{\frac{1}{|\mathcal{O}|}||\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L} - \Gamma\mathbf{1}_T^\top - \mathbf{1}_N\Delta^\top)||_F^2 + \lambda||\mathbf{L}||_*\}$$

  - $\Gamma \in \mathbb{R}^N$: unit-varying (and time-fixed) effect (individual effect)
  - $\Delta \in \mathbb{R}^T$: time-varying (and unit-fixed) effect (time effect)
  - matrix indicator function : $\mathbf{P}_{\mathcal{O}}(\mathbf{A}) = \begin{cases} A_{it} & \text{if } (i,t) \in \mathcal{O} \\ 0 & \text{if } (i,t) \notin \mathcal{O} \end{cases}$ (NA is regarded as 0)
  - Frobenius norm : $||\mathbf{A}||_F^2 = \sum_{i=1}^N \sum_{t=1}^T A_{it}^2$ (行列版の mean squared error を計算している)
- Regularization term $\lambda||L||_*$ leads to the low rank of $\mathbf{L}$.

$\rightarrow$ minimize $\lambda||L||_* \Leftrightarrow$ the sparsity of Singular value $\sigma_i(\mathbf{L})(>0) \Leftrightarrow$ low rank of $\mathbf{L}$

- **Fact 1.** (Singular value decomposition) *Every real matrix* $L \in \mathbb{R}^N \times \mathbb{R}^T$ *can be decomposed using a onthogonal matrix* $\mathbf{S} \in \mathbb{R}^N \times \mathbb{R}^{\min(N,T)}$, $\mathbf{R} \in \mathbb{R}^T \times \mathbb{R}^{\min(N,T)}$ *by*

$$\mathbf{L} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{R}', \text{where } \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots \sigma_{\min(N,T)}), \mathbf{S}'\mathbf{S} = \mathbf{I}_{\min(\mathbf{N},\mathbf{T})} = \mathbf{R}'\mathbf{R}$$

- **Fact 2.** *The number of non-zero singular value = rank* $\mathbf{L}$
  - Nuclear norm : $||L||_* = \sum_{i=1}^{\min(N,T)} \sigma_i(\mathbf{L})$

    $\rightarrow$ minimize $\lambda ||L||_* \Leftrightarrow$ the sparsity of Singular value $\sigma_i(\mathbf{L})(> 0) \Leftrightarrow$ low rank of $\mathbf{L}$
- Since the rank of $\mathbf{L}$ corresponds to the number of factor, this assumption of low rank is quite plausible.
- Although the law rank matrix CAN include two fixed effects, these ”strong” factors are separately estimated for improving the quality of the practical imputations.

Algorithm for calculating $\hat{\mathrm{L}}$

- For simplicity, assume that there are no fixed effects. (only estimate $\mathbf{L}$)
- **Fact 3.** *For* $\mathbf{A} = \mathbf{S\Sigma R}$*, the minimizer is obtained analytically.*

  $\mathbf{S\tilde{\Sigma}R}^\top = \arg\min_{\mathbf{A}}\{\frac{1}{2}||\mathbf{L} - \mathbf{A}||_F^2 + \lambda||\mathbf{A}||_*\}$, where $\tilde{\Sigma} = \mathrm{diag}(\{\max(\sigma_i(\mathbf{A}) - \lambda, 0)\}_i)$

  $\rightarrow$ You can see elements with a small singular value (=weak factor) will be vanished.
- We perform this minimization over and over until the matrix converges.
  - Define $\mathrm{shrink}_\lambda(\mathbf{A}) = \mathbf{S\tilde{\Sigma}R}^\top$ and start with the initial choice $\mathbf{L_1}(\lambda, \mathcal{O}) = \mathbf{P}_\mathcal{O}(\mathbf{Y})$ (The missing starts with 0.)

$$\mathbf{L}_{k+1}(\lambda, \mathcal{O}) = \mathrm{shrink}_{\frac{\lambda|\mathcal{O}|}{2}}\{\mathbf{P}_\mathcal{O}(\mathbf{Y}) + \mathbf{P}_\mathcal{O}^\top(\mathbf{L}_k(\lambda, \mathcal{O}))\}$$

  - $\mathbf{P}_\mathcal{O}(\mathbf{A}) = \begin{cases} A_{it} & \text{if } (i,t) \in \mathcal{O} \\ 0 & \text{if } (i,t) \notin \mathcal{O} \end{cases}, \quad \mathbf{P}_\mathcal{O}^\top(\mathbf{A}) = \begin{cases} 0 & \text{if } (i,t) \in \mathcal{O} \\ A_{it} & \text{if } (i,t) \notin \mathcal{O} \end{cases}$
  - The limiting matrix $\hat{\mathbf{L}}(\lambda, \mathcal{O}) = \lim_{k\to\infty} \mathbf{L}_k(\lambda, \mathcal{O})$ is MC-NNM estimator given λ.

- For the case with fixed effects, we replace $\mathbf{P}_{\mathcal{O}}(\mathbf{Y})$ with $\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \Gamma_k \mathbf{1}_T^\top - \mathbf{1}_N \Delta_k^\top)$
- After each iteration to obtain $\hat{\mathbf{L}}_{k+1}$, we can estimate Γ_{k+1} and Δ_{k+1} by using the first-order conditions.
    - More specifically, after estimating $\hat{\mathbf{L}}_{k+1}$, the objective function is as the quadratic form wrt. Γ_k, Δ_k, so we additionally minimize it and renew $\Gamma_{k+1}, \Delta_{k+1}$.
    - Finally, replace the $\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \Gamma_k \mathbf{1}_T^\top - \mathbf{1}_N \Delta_k^\top)$ with $\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \Gamma_{k+1} \mathbf{1}_T^\top - \mathbf{1}_N \Delta_{k+1}^\top)$, then proceed the algorithm to obtain $\hat{\mathbf{L}}_{k+2}$.
- we can interpert the term $\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \Gamma_k \mathbf{1}_T^\top - \mathbf{1}_N \Delta_k^\top)$ as an invariant (fixed) factor term.
- Actually, this separation of two FEs makes practical imputation greatly improved.

Inference : CV and CI

- The optimal value of λ is selected through cross-validation.
- The theory of asymptotic distribution of $\mathbf{L}^* - \hat{\mathbf{L}}$ to construct CI has not yet developed.
- Instead, using resampling method, we see the fluctulation of the imputed matrix and construct CI like premutation methods in ADH-synthetic control method.
  - Randomly select the subset $\mathcal{O}_k$ of $\mathcal{O}$ for $k = 1, ..., K$.
  - Then, using the sample in $\mathcal{O}_k$, calculate MC-NNM estimator $\hat{\mathbf{L}}^{(k)}$.
  - Finally, we construct a pointwise CI for $\{\hat{L}_{it}^{(k)}\}$.
- However, according to Choi and Yuan(2024) which is the extended literature of Athey et al.(2021), their debiased estimator for $\mathbf{L}^*$ is asymptotic normal (but pointwisely).

The relationship with horizontal and vertical regressions

Interpretation as generalization

- For simplicity, we assume when only one missing $Y_{NT}(0)$ exists i.e. $\mathcal{M} = (N, T)$
- Let the estimand $\mathbf{Y}$ partition as $\begin{pmatrix} \mathbf{Y_0} & \mathbf{y_1} \\ \mathbf{y_2}^\top & ? \end{pmatrix}$,
  where $\mathbf{Y_0} \in \mathbb{R}^{N-1} \times \mathbb{R}^{T-1}, \mathbf{y_1} \in \mathbb{R}^{N-1}, \mathbf{y_2} \in \mathbb{R}^{T-1}$.
- For a given positive integer R, define an $N \times R$ matrix $\mathbf{A}$, an $T \times R$ matrix $\mathbf{B}$, a N-dim. vector γ and a R-dim. vector δ, then, the objective function w.r.t. MSE is

$$Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) = \frac{1}{|\mathcal{O}|} ||\mathbf{P}_{\mathcal{O}}(\mathbb{Y} - \mathbf{A}\mathbf{B}^\top - \gamma \mathbf{1}_T^\top - \mathbf{1}_N \delta^\top)||_F^2$$

- However, there is no unique solution for $\mathbf{A}, \mathbf{B}$ unless restrict them.
  $\rightarrow$ Thus, we compare the restriction of each methods.

MC-NNM estimator

- Nuclear norm matrix completion

$$
\begin{aligned}
(R_\lambda^{\text{mc-nnm}},\, \mathbf{A}_\lambda^{\text{mc-nnm}},\, \mathbf{B}_\lambda^{\text{mc-nnm}},\, \gamma_\lambda^{\text{mc-nnm}},\, \delta_\lambda^{\text{mc-nnm}}) \\
= \arg \min_{R,A,B,\gamma,\delta} \Big\{ Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2 \Big\}.
\end{aligned}
$$

- **Fact 3.** $\|\mathbf{L}\|_* = \min_{\mathbf{A},\mathbf{B}\colon \mathbf{L}=\mathbf{A}\mathbf{B}'} \frac{1}{2}(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$.
- In second and third terms, we regularize $\mathbf{A}, \mathbf{B}$ so that the minimization problem has a unique solution.
  $\rightarrow$ Compared with other methods, there is a big differnece in that MC-NNM does not restrict the form of the matrix but just regularize it in a data-driven manner.

Horizontal regression (thin matrix, $N > T$)

- Horizontal regression estimator, defined if $N > T$: thin matrix

$$(R^{hr},\, \mathbf{A}^{hr},\, \mathbf{B}^{hr},\, \gamma^{hr},\, \delta^{hr}) = \arg\min_{R,A,B,\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

$$\text{subject to} \quad R = T-1, \quad \mathbf{A} = \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{y}_2^\top \end{pmatrix}, \quad \gamma = 0, \quad \delta_1 = \cdots = \delta_{T-1} = 0.$$

- The solution for $\mathbf{B}$ is

$$\mathbf{B}^{hr\top} = \begin{pmatrix} E_{T-1} & \hat{\beta} \end{pmatrix}, \quad (\hat{\beta}, \hat{\delta}_T) = \arg\min_{\beta,\delta_T} \sum_{i=1}^{N-1} \Big(Y_{iT} - \delta_T - \sum_{t=1}^{T-1} \beta_t\, Y_{it}\Big)^2.$$

- Restrict the form of $\mathbf{A}$ and (of course) assume no individual effect ($\gamma = 0$).

Vertical regression (fat matrix, $T > N$)

- Vertical regression estimator, defined if $T > N$: fat matrix

$$(R^{vt}, \mathbf{A}^{vt}, \mathbf{B}^{vt}, \gamma^{vt}, \delta^{vt}) = \arg\min_{R,A,B,\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

$$\text{subject to} \quad R = N-1, \quad \mathbf{B}^{\mathsf{T}} = \begin{pmatrix} \mathbf{Y}_0 & \mathbf{y}_1 \end{pmatrix}, \quad \gamma_1 = \cdots = \gamma_{N-1} = 0, \quad \delta = 0.$$

- The solution for $\mathbf{A}$ is

$$\mathbf{A}^{\text{vt}} = \begin{pmatrix} E_{N-1} \\ \hat{\alpha} \end{pmatrix}, \quad (\hat{\alpha}, \hat{\gamma}_N) = \arg\min_{\alpha,\gamma_N} \sum_{t=1}^{T-1} \left(Y_{Nt} - \gamma_N - \sum_{i=1}^{N-1} \alpha_i Y_{it}\right)^2.$$

- Restrict the form of $\mathbf{B}$ and (of course) assume no time effect ($\delta = 0$).

Another methods

(iv) (synthetic control),

$$(R^{\text{sc-adh}}, \mathbf{A}^{\text{sc-adh}}, \mathbf{B}^{\text{sc}-\text{adh}}, \gamma^{\text{sc-adh}}, \delta^{\text{sc-adh}}) = \underset{R, \mathbf{A}, \mathbf{B}, \gamma, \delta}{\operatorname{argmin}} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = N-1, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \qquad \delta = 0, \qquad \gamma = 0,$$

$$\forall\, i, A_{iT} \geq 0, \sum_{i=1}^{N-1} A_{iT} = 1,$$

(v) (vertical regression, elastic net),

$$(R^{\text{vt}-\text{en}}, \mathbf{A}^{\text{vt}-\text{en}}, \mathbf{B}^{\text{vt}-\text{en}}, \gamma^{\text{vt}-\text{en}}, \delta^{\text{vt}-\text{en}}) = \underset{R, \mathbf{A}, \mathbf{B}, \gamma, \delta}{\operatorname{argmin}} \left\{ Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) + \lambda \left[\frac{1-\alpha}{2} \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_F^2 + \alpha \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_1 \right] \right\},$$

subject to

$$R = N-1, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix},$$

$$\gamma_1 = \gamma_2 = \cdots = \gamma_{N-1} = 0, \qquad \delta = 0,$$

where $\mathbf{A}$ is partitioned as

$$\mathbf{A} = \begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{a}_1 \\ \mathbf{a}_2^\top & \mathbf{a}_3 \end{pmatrix},$$

Theoritical Bounds for the Estimation Error

Estimation error

- Define p_c to be a minimiun expected proportion of control (never-treated) unit.

$$p_c \equiv \min_{1 \leq i \leq N} \pi_T^{(i)}, \text{ where } \pi_T^{(i)} = \mathbb{P}(t_i = T)$$

Assumption 2

- p_c is sufficiently large even as $N, T \to \infty$ with $N \geq T$. (remained control units)

$$p_c \gtrsim \frac{\log^{3/2}(N+T)}{\sqrt{T}} \vee \frac{\sqrt{T}\log^{3/2}(N+T)}{N}$$

- True matrix $\mathbf{L}^*$ is a low rank.

- Under this assumption, $\dfrac{||\mathbf{L}^* - \hat{\mathbf{L}}||_F}{\sqrt{NT}}$ has an upper bound with a high probability ($\mathbb{P} = 1 - \dfrac{2}{(N+T)^2}$), then, MC-NNM estimator has a consisitency.

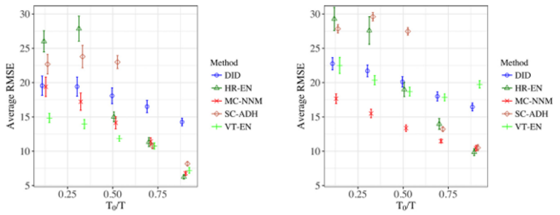
Empirical application

Comparing each method

- In a real data matrix $\mathbf{Y}$ where no units is treated, we choose units as ”hypothetical treated” units (=regard as missing) and aim to predict (impute) their value.
  $\rightarrow$ Technically, compare the average root-mean-squared-error(RMSE) to assess which of the algorithm generally perform well.
- The compared estimators (algorithms) are as follows.
  - DiD : 2 回差分により counterfactual を補完する
  - VT-EN : elastics net で vertical regression function を作り、その推定値で補完
  - HR-EN : elastics net で horizontal regression function を作り、その推定値で補完
  - SC-ADH : classical approach to construct ”synthetic control”.
  - MC-NNM : 本日の主役

ADH(2010): California smoking data

- $N = 38, T = 31$
  - Simultaneous adoption : Let randomly selected $N_t = 8$ units ”treated” in period $T_0 + 1$.
  - Staggered adoption : Let randomly selected $N_t = 35$ units ”treated” in some periods after period T.



(a) Simultaneous adoption, $N_t = 8$ (b) Staggered adoption, $N_t = 35$

- (a) : VT-EN performs well on the whole, DiD is poor.
- (b) : MC-NNM is superior (large sample is favorable!), and VT-EN is generally poor.

Final marks

Summary and Conclusion

- Matrix completion is a imputing method for the missing $Y_{it}(0)$ where $(i, t) \in \mathcal{M}$.
- MC-NNM estimator generalizes and nests many estimators such as DiD, ADH-SC, vertical or horizontal (penalized) regression, and interactive fixed effect model.
- The critical differnce with previous DiD estimators is that MC-NNM holistically regularizes latent factors through nuclear-norm minimization which induces a low-rank matrix (sparsity).
    - Intrinsic factor vs. Explicitly imposed factor
- Under appropriate conditions (low-rank structure, sufficient number of control units, sufficiently large N, T), the MC-NNM estimator achieves consistency.
- Practically, this estimator performs well with large N and T, and allows for a relatively large number of factors.

Further extensions

- To consider the model with covariates, we can separately include them in $\frac{1}{|\mathcal{O}|}||\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top - \gamma\mathbf{1}_T^\top - \mathbf{1}_N\delta^\top)||_F^2$.
- For further advanced developments, a recent study by Choi and Yuan (2024, JASA) proposes a debiased estimator for $\mathbf{L}^*$ with rigorous asymptotic inference (pointwise asymptotic normality).
- They introduce a cross-fitting (sample-splitting) strategy and further propose group-based ATT estimators, extending the matrix completion literature toward causal inference with stronger theoretical guarantees.

References

References

- Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi (2021), *Matrix Completion Methods for Causal Panel Data Models*, Journal of the American Statistical Association.
- Licheng Liu, Ye Wang, and Yiqing Xu (2024), *A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data*, American Journal of Political Science.
- Jungjun Choi and Ming Yuan(2024), *Matrix Completion When Missing Is Not at Random and Its Applications in Causal Panel Data Models*, Journal of the American Statistical Association.
- Martin J. Wainwright (2019), *High-Dimensional Statistics*, Cambridge University Press. (mainly Chapter 10)
- 冨岡亮太 (2015), スパース性に基づく機械学習, 講談社.
- 川口康平, 澤田真行 (2024), 因果推論の計量経済学, 日本評論社.