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

Today's Agenda ; Keyword : Imputation

-

Imputation

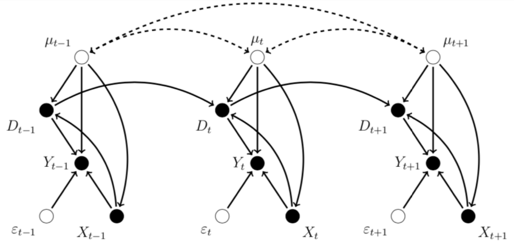
- As many panel data methods, we want to know ATT : $\mathbb{E}[Y_{it}(1) - Y_{it}(0)|W_i = 1]$.
- Thus, it boils down to estimate (impute) the counterfactual $Y_{it}(0)$.
  - Horizontal : Under unconfoundedness, we can impute counterfactual PO using observed outcomes for control units.
  - Vertical : 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: $h(\mathbf{U_{it}}) = \{L_{it}\}, \text{rank}(\mathbf{L}_{N\times T}) \ll \min\{N, T\}$
  $\rightarrow$ The rank (= number of factors) is FIXED !!

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!!)
- $W_{it} = \begin{cases} 1 & \text{if } (i,t) \in \mathcal{M} \\ 0 & \text{if } (i,t) \in \mathcal{O} \end{cases}$

[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

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

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 $\mathbf{L}^*$
- 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)
  → 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 $\hat{\mathbf{L}} + \hat{\Gamma}\mathbf{1}_T' + \mathbf{1}_N\hat{\Delta}'$

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

  - $\Gamma \in \mathbb{R}^N$: 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 $\mathbf{L}$)
-

The relationship with horizontal and vertical regressions

- Let the estimand $\mathbf{Y}$ partition as $\begin{pmatrix} \mathbf{Y_0} & \mathbf{y_1} \\ \mathbf{y_2}' & ? \end{pmatrix}$, 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

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

Theorem 1. In the case with only the (N, T) entry missing, we have,

(i) (nuclear norm matrix completion)

$$(R^{\text{mc-nnm}}, \mathbf{A}_\lambda^{\text{mc-nnm}}, \mathbf{B}_\lambda^{\text{mc-nnm}}, \gamma_\lambda^{\text{mc-nnm}}, \delta_\lambda^{\text{mc-nnm}}) =$$

$$\underset{R,\mathbf{A},\mathbf{B},\gamma,\delta}{\operatorname{argmin}} \left\{ Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2 \right\},$$

(ii) (horizontal regression, defined if $N > T$)

$$(R^{\mathrm{hr}}, \mathbf{A}^{\mathrm{hr}}, \mathbf{B}^{\mathrm{hr}}, \gamma^{\mathrm{hr}}, \delta^{\mathrm{hr}}) = \underset{R,\mathbf{A},\gamma,\delta}{\operatorname{argmin}}\, Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = T - 1, \qquad \mathbf{A} = \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{y}_2^\top \end{pmatrix}, \qquad \gamma = 0,$$
$$\delta_1 = \delta_2 = \cdots = \delta_{T-1} = 0,$$

(iii) (vertical regression, defined if $T > N$),

$$(R^{\mathrm{vt}}, \mathbf{A}^{\mathrm{vt}}, \mathbf{B}^{\mathrm{vt}}, \gamma^{\mathrm{vt}}, \delta^{\mathrm{vt}}) = \underset{R,\mathbf{A},\mathbf{B},\gamma,\delta}{\operatorname{argmin}}\, Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = N - 1, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix},$$
$$\gamma_1 = \gamma_2 = \cdots = \gamma_{N-1} = 0, \qquad \delta = 0,$$

(iv) (synthetic control),

$$
\begin{aligned}
&(R^{\text{sc-adh}}, \mathbf{A}^{\text{sc-adh}}, \mathbf{B}^{\text{sc-adh}}, \gamma^{\text{sc-adh}}, \delta^{\text{sc-adh}}) \\
&= \operatorname*{argmin}_{R,\mathbf{A},\mathbf{B},\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta)\,,
\end{aligned}
$$

subject to

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

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

Theoritical Bounds
for the Estimation Error

Two illustrations

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, Yiqing Xu (2024), *A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data*, American Journal of Political Science.
- 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.