Matrix Completion Methods for Causal Panel Data Models

Susan Athey, Mohsen Bayati,

Nikolay Doudchenko, Guido Imbens, & Khashayar Khosravi

(Stanford University)

NBER Summer Institute

Cambridge, July 27th, 2017

- We are interested in estimating the (average) effect of a binary treatment on a scalar outcome.

- We have data on N units, for T periods.

- We observe
  - the treatment, $W_{it} \in \{0, 1\}$,
  - the realized outcome Y_{it},
  - time invariant characteristics of the units X_i,
  - unit-invariant characteristics of time Z_t,
  - time and unit specific characteristics V_{it}

We observe (in addition to covariates):

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & \dots & Y_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \dots & Y_{NT} \end{pmatrix} \qquad \text{outcome.}$$

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \dots & 0 \end{pmatrix} \qquad \text{treatment.}$$

- **rows are units**, **columns are time periods**. (Important because some, but not all, methods treat units and time periods asymmetric)

In terms of potential outcomes:

$$\mathbf{Y}(0) = \begin{pmatrix} ? & ? & \checkmark & \dots & ? \\ \checkmark & \checkmark & ? & \dots & \checkmark \\ ? & \checkmark & ? & \dots & \checkmark \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ ? & \checkmark & ? & \dots & \checkmark \end{pmatrix} \quad \mathbf{Y}(1) = \begin{pmatrix} \checkmark & \checkmark & ? & \dots & \checkmark \\ ? & ? & \checkmark & \dots & ? \\ \checkmark & ? & \checkmark & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & ? & \checkmark & \dots & ? \end{pmatrix}.$$

In order to estimate the average treatment effect for the treated, (or other average, e.g., overall average effect)

$$\tau = \frac{\sum_{i,t} W_{it} \Bigg(Y_{it}(1) - Y_{it}(0) \Bigg)}{\sum_{it} W_{it}},$$

We need to **impute** the missing potential outcomes in at least one of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$.

Focus on problem of imputing missing in $\mathbf{Y}$ (either $\mathbf{Y}(0)$ or $\mathbf{Y}(1)$)

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

$\mathcal{O}$ and $\mathcal{M}$ are sets of indices (it) with $Y_{i,t}$ observed and missing, with cardinalities $|\mathcal{O}|$ and $|\mathcal{M}|$. Covariates, time-specific, unit-specific, time/unit-specific.

- This is a **Matrix Completion Problem**.

General set up:

$$\mathbf{Y}_{N\times T} = \mathbf{L}_{N\times T} + \varepsilon_{N\times T}$$

- Key assumption "Matrix Unconfoundedness":

$$\mathbf{W}_{N\times T} \perp\!\!\!\perp \varepsilon_{N\times T} \;\Big|\; \mathbf{L}_{N\times T}$$

(but $\mathbf{W}$ may depend on $\mathbf{L}$)

- In addition:

$$\mathbf{L}_{N\times T} \approx \mathbf{U}_{N\times R}\mathbf{V}_{T\times R}^{\top}$$

well approximated by matrix with rank R low relative to N and T.

- Classification of practical problems depending on
  - magnitude of T and N,
  - pattern of missing data, fraction of observed data $|\mathcal{O}|/(|\mathcal{O}|+|\mathcal{M}|)$ close to zero or one.

- Different structure on $\mathbf{L}$ in
  - average treatment effect under unconfoundedness lit.
  - synthetic control literature
  - panel data / DID / fixed effect literature
  - machine learning literature

Classification of Problem I: Magnitude of N and T

Thin Matrix (N large, T small), typical cross-section setting:

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} ? & \checkmark & ? \\ \checkmark & ? & \checkmark \\ ? & ? & \checkmark \\ \checkmark & ? & \checkmark \\ ? & ? & ? \\ \vdots & \vdots & \vdots \\ ? & ? & \checkmark \end{pmatrix} \qquad \text{(many units, few time periods)}$$

Fat Matrix (N small, T large), time series setting:

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} ? & ? & \checkmark & \checkmark & \checkmark & \dots & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & ? & \dots & \checkmark \\ ? & \checkmark & ? & \checkmark & ? & \dots & \checkmark \end{pmatrix} \quad \text{(few units, many periods)}$$

Or approx **square** matrix, N and T comparable magnitude.

Classification of Problem II: Pattern of Missing Data

Most of econometric causal literature focuses on case with block of Treated Units / Time Periods

$$\mathbf{Y}_{N\times T}=\left(\begin{array}{ccc|ccc}\checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark\\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark\\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark\\ \hline \checkmark & \checkmark & \checkmark & ? & \dots & ?\\ \checkmark & \checkmark & \checkmark & ? & \dots & ?\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ \checkmark & \checkmark & \checkmark & ? & \dots & ?\end{array}\right)=\left(\begin{array}{cc}\mathbf{Y}_{C,\text{pre}}(0) & \mathbf{Y}_{C,\text{post}}(0)\\ \mathbf{Y}_{T,\text{pre}}(0) & ?\end{array}\right)$$

Easier because it allows for complete-data modeling of

- cond. distr. of $\mathbf{Y}_{C,\text{post}}(0)$ given $\mathbf{Y}_{C,\text{pre}}(0)$ (matching) or
- cond. distr. of $\mathbf{Y}_{T,\text{pre}}(0)$ given $\mathbf{Y}_{C,\text{pre}}(0)$ (synt. control).

Two important special cases:

Single Treated Unit (Abadie et al Synthetic Control)

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \checkmark & \checkmark & ? & \dots & ? & \leftarrow \text{(treated unit)} \end{pmatrix}$$

Single Treated Period (Most of Treatment Effect Lit)

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & \checkmark & \checkmark & \dots & ? \\ & & & & \uparrow \\ & & & & \text{(treated period)} \end{pmatrix}$$

Other Important Assignment Patterns

Staggered Adoption (e.g., adoption of technology, Athey and Stern, 1998)

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

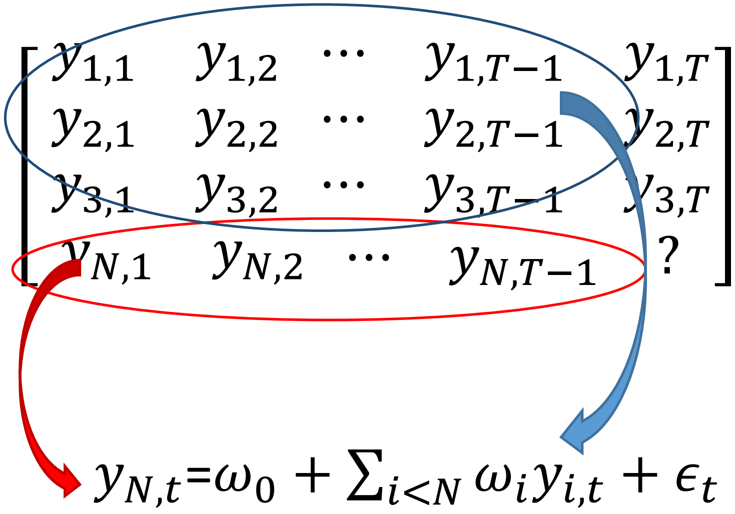
Netflix Problem

- Very very large N (number of individuals),
- Large T (number of movies),
- raises computational issues
- General missing data pattern,
- Fraction of observed data is close to zero, $|\mathcal{O}| << |\mathcal{M}|$

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

Fat Matrix: Vertical Regression

- **Outcome:**
  - Target unit outcome in period *t*
- **Covariates:**
  - Other unit’s outcomes in same period.
- **Observation** is a **time period**.
- What is **stable:**
  - Patterns **across units**
- **Identification**:
  - $Y_{N,t}(0) \perp W_{N,t} | Y_{1,t}, . . , Y_{N-1,t}$
- **Examples:**
  - Synthetic control: $\omega_i \geq 0, \sum_{i>1} \omega_i = 1$
  - Doudchenko-Imbens: estimate ω_i w/ elastic net



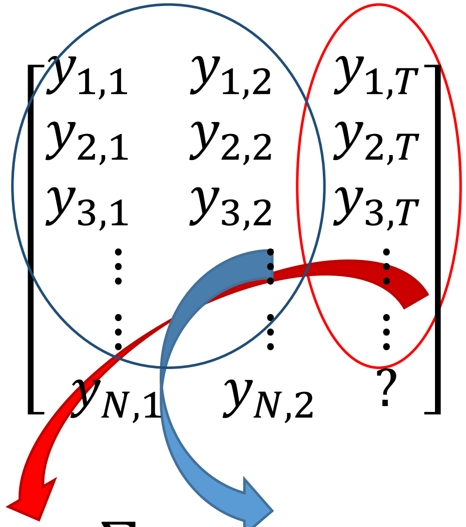
$$\begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,T-1} & y_{1,T} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,T-1} & y_{2,T} \\ y_{3,1} & y_{3,2} & \cdots & y_{3,T-1} & y_{3,T} \\ y_{N,1} & y_{N,2} & \cdots & y_{N,T-1} & ? \end{bmatrix}$$

$$y_{N,t} = \omega_0 + \sum_{i<N} \omega_i y_{i,t} + \epsilon_t$$

$$\hat{y}_{N,T} = \sum_{i>1} \widehat{\omega}_i y_{iT}$$

Thin Matrix: Horizontal Regression

- **Outcome:**
  - Target time period outcome
- **Covariates:**
  - Other time period outcome for same unit
- **Observation** is a **unit**.
- What is **stable:**
  - Time patterns within a unit
- **Identification:**
  - $Y_{i,T}(0) \perp W_{i,T} | Y_{i,1}, \ldots, Y_{i,T-1}$
- **Examples:**
  - Matching, ATE literature: avg. outcomes from units with most similar $\boldsymbol{y}_{i,-T}$
  - With regularization: Chernozhukov et al, Athey, Imbens and Wager (2017)
  - Closely related to transposed versions of Synthetic Controls, Elastic Net



$$\begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,T} \\ y_{2,1} & y_{2,2} & y_{2,T} \\ y_{3,1} & y_{3,2} & y_{3,T} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ y_{N,1} & y_{N,2} & ? \end{bmatrix}$$

$$y_{i,T} = \sum_{t<T} \omega_t y_{i,t} + \epsilon_i$$

$$\hat{y}_{N,T} = \sum_{t<T} \widehat{\omega}_t y_{N,t}$$

General Matrix: Matrix Regression (Panel)

$$\begin{bmatrix} y_{1,1} & \cdots & y_{1,T} \\ y_{2,1} & \cdots & y_{2,T} \\ y_{3,1} & \cdots & y_{3,T} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ y_{N,1} & \cdots & ? \end{bmatrix}$$

$$y_{i,t} = \gamma_i + \delta_t + \epsilon_{i,t}$$

$$\hat{y}_{NT} = \hat{\gamma}_N + \hat{\delta}_T$$

$$Y_{N\times T} = L_{N\times T} + \epsilon_{N\times T} = \begin{bmatrix} \gamma_1 & 1 \\ \vdots & \vdots \\ \gamma_N & 1 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ \delta_1 & \cdots & \delta_T \end{bmatrix} + \epsilon_{N\times T}$$

- Panel data regression. Exploit additive structure in unit and time effects.
- Identification: $Y_{i,t}(0) \perp W_{i,t} | \gamma_i, \delta_t$
- Matrix formulation of identification: $Y_{N\times T}(0) \perp W_{N\times T} | L_{N\times T}$

II. How/why do we regularize:

Potentially many parameters when (i) vertical regression on thin matrix, (ii) horizontal regression on fat matrix, iii) matrix is approx square:

> “Regularization theory was one of the first signs of the existence of intelligent inference.” (Vapnik, 1999, p. 9)

- Need regularization to avoid overfitting.

- **How** you do the regularization is important for substantive and computational reasons: lasso/elastic-net/ridge are better than best subset in simple regression setting.

Literature:

Regularize →	No Regular.	Best Subset	ℓ_1/LASSO, ℓ_2 et al
Regression ↓			
Horizontal	earlier causal effect lit.	–	Chernozhukov et al Athey et al
Vertical	Abadie-Diam., Hainmueller	–	Doudch.-Imb. & Abadie-L'Hour
Matrix	two-way fixed effect literature	Bai (2003) Xu (2017)	**Current Paper**

Econometric Literature I: Treatment Effect / Matching-Regression

- Thin matrix (many units, few periods), single treated period (period T).

Strategy: Use controls to regress $Y_{i,T}$ on lagged outcomes $Y_{i,1},\ldots,Y_{i,T-1}$. N_C obs, $T-1$ regressors.

- Does not work well if $\mathbf{Y}$ is fat (few units, many periods).

- Key identifying assumption: $Y_{iT}(0) \perp\!\!\!\perp W_{iT}|Y_{i1},\ldots,Y_{iT-1}$

Econometric Literature II: Abadie-Diamond-Hainmueller Synthetic Control Literature

- Fat matrix, single treated unit (unit N), treatment starts in period T_0.

Strategy: Use pretreatment periods to regress $Y_{N,t}$ on contemporaneous outcomes $Y_{1,t},\ldots,Y_{N-1,t}$. T_0-1 obs, N regressors. Weights (regression coefficients) are nonnegative and sum to one, no intercept.

- Does not work well if matrix is thin (many units).

- Key identifying assumption: $Y_{Nt}(0) \perp\!\!\!\perp W_{Nt}|Y_{1t},\ldots,Y_{N-1t}$

Econometric Literature III: Doudchenko-Imbens

- Fat matrix or similar N, T, single treated unit (unit N), treatment starts in period T_0.

Strategy: Use pretreatment periods to regress $Y_{N,t}$ on contemporaneous outcomes $Y_{1,t},\ldots,Y_{N-1,t}$. using elastic net regularization. $T_0 - 1$ obs, N regressors.

- Allows for negative weights, weights summing to something other than one, non-zero intercept, typically requires **regularization**.

Econometric Literature IV: Transposed Abadie-Diamond-Hainmueller or Doudchenko-Imbens (Reverse role of time and units compared to ADH or DI)

- fat matrix, single treated unit (N), treatment in period T.

Strategy: Use control units to regress Y_{iT} on lagged outcomes $Y_{i1}, \ldots, Y_{iT-1}$. using elastic net regularization. N_C obs, $T-1$ regressors.

- Allows for negative weights, weights summing to something other than one, non-zero intercept.

Similar to regression estimator for matching setting, with regularization.

Econometric Literature V: Fixed Effect Panel Data Literature / Difference-In-Differences

- T and N similar, general pattern for treatment assignment.

Model:

$$Y_{it} = \alpha_i + \gamma_t + \varepsilon_{i,t}$$

- **Symmetric** in role of units and time periods.

- Suppose $T = 2$, $N = 2$, $W_{2,2} = 1$, $W_{i,t} = 0$ if $(i, t) \neq (2, 2)$, then we have a classic DID setting, leading to imputed value

$$\widehat{Y}_{2,2} = Y_{1,2} + \left(Y_{2,1} - Y_{1,1} \right)$$

Questions: What to do if we are unsure about thin/fat/square, with staggered adoption or general assignment mechanism?

- We generalize interactive fixed effects model (Bai, 2003, 2009; Xu 2017, Gobillon and Magnac, 2013; Kim and Oka, 2014), allowing for large rank $\mathbf{L}$.

- We propose a new estimator with novel regularization:
  - can deal with staggered/general missing data patterns
  - Computationally feasible bec. convex optimization probl.
  - Reduces to matching under assump. in thin case.
  - Reduces to synt. control under assump. in fat case.

X_i is P-vector, Z_t is Q vector.

Model (generalized version of Xu, 2017):

$$Y_{it} = L_{it} + \sum_{p=1}^{P} \sum_{q=1}^{Q} X_{ip} H_{pq} Z_{qt} + \gamma_i + \delta_t + V_{it}\beta + \varepsilon_{it}$$

Unobserved: L_{it}, γ_i, δ_t, H_{pq}, β, ε_{it}.

• We do not necessarily need the fixed effects γ_i and δ_t, these can be subsumed into $\mathbf{L}$.

If $L_{it} = \gamma_i + \delta_t$, then $\mathbf{L}$ is a rank 2 matrix:

$$\mathbf{L} = \begin{pmatrix} \gamma_{N\times 1} & \iota_{N\times 1} \end{pmatrix} \begin{pmatrix} \iota_{T\times 1} & \delta_{T\times 1} \end{pmatrix}^\top$$

$$= \begin{pmatrix} \gamma_1 & 1 \\ \gamma_2 & 1 \\ \vdots & \vdots \\ \gamma_N & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ \delta_1 & \delta & \dots & \delta_T \end{pmatrix}$$

• It may be convenient to include the fixed effects given that we regularize $\mathbf{L}$.

Too many parameters (especially $N \times T$ matrix $\mathbf{L}$), so we **need** regularization:

We shrink $\mathbf{L}$ and $\mathbf{H}$ towards zero.

For $\mathbf{H}$ we use Lasso-type element-wise ℓ_1 norm: defined as $\|\mathbf{H}\|_{1,e} = \sum_{p=1}^{P} \sum_{q=1}^{Q} |H_{pq}|$.

How do we regularize $\mathbf{L}_{N\times T}$?

In linear regression with many regressors,

$$Y_i = \sum_{l=1}^{K} \beta_k X_{ik} + \varepsilon_i,$$

we often regularize by adding a penalty term $\lambda\|\beta\|$ where

$$\|\beta\| = \|\beta\|_0 = \sum_{k=1}^{K} \mathbf{1}_{|\beta_k|\neq 0} \quad \text{best subset selection}$$

$$\|\beta\| = \|\beta\|_1 = \sum_{k=1}^{K} |\beta_k| \quad \text{LASSO}$$

$$\|\beta\| = \|\beta\|_2^2 = \sum_{k=1}^{K} |\beta_k|^2 \quad \text{ridge}$$

Matrix norms for $N \times T$ Matrix $\mathbf{L}_{N\times T}$

$$\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\mathbf{\Sigma}_{N\times T}\mathbf{R}_{T\times T} \quad \text{(singular value decomposition)}$$

$\mathbf{S}$, $\mathbf{R}$ unitary, Σ is rectang. diagonal with entries $\sigma_i(\mathbf{L})$ that are the **singular values**. Rank($\mathbf{L}$) is # of non-zero $\sigma_i(\mathbf{L})$.

$$\|\mathbf{L}\|_F^2 = \sum_{i,t} |L_{it}|^2 = \sum_{j=1}^{\min(N,T)} \sigma_i^2(\mathbf{L}) \quad \text{(Frobenius, like ridge)}$$

$$\|\mathbf{L}\|_* = \sum_{j=1}^{\min(N,T)} \sigma_i(\mathbf{L}) \quad \textbf{(nuclear norm, like LASSO)}$$

$$\|\mathbf{L}\|_R = \text{rank}(\mathbf{L}) = \sum_{j=1}^{\min(N,T)} \mathbf{1}_{\sigma_i(\mathbf{L})>0} \quad \text{(Rank, like subset)}$$

Xu (2017) focuses on case with block assignment,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{C,\text{pre}} & \mathbf{Y}_{C,\text{post}} \\ \mathbf{Y}_{T,\text{pre}} & ? \end{pmatrix}$$

Following Bai (2009), Xu fixes the rank $R(\mathbf{L})$ so we can write $\mathbf{L}$ as a matrix with an R-factor structure:

$$\mathbf{L} = \mathbf{U}\mathbf{V}^\top = \begin{pmatrix} \mathbf{U}_C \\ \mathbf{U}_T \end{pmatrix} \begin{pmatrix} \mathbf{V}_{\text{pre}} \\ \mathbf{V}_{\text{post}} \end{pmatrix}^\top$$

where

$$\mathbf{U} \text{ is } N \times R, \qquad \mathbf{V} \text{ is } T \times R$$

Xu (2017) two-step method:

First, use all controls to estimate $\mathbf{U}_C$, $\mathbf{V}_{\text{pre}}$, $\mathbf{V}_{\text{post}}$:

$$\min_{\mathbf{U}_C,\mathbf{V}_{\text{pre}},\mathbf{V}_{\text{post}}} \left\| \mathbf{Y}_C - \mathbf{U}_C \begin{pmatrix} \mathbf{V}_{\text{pre}} \\ \mathbf{V}_{\text{post}} \end{pmatrix}^\top \right\|$$

Second, use the treated units in pre period to estimate $\mathbf{U}_T$ given $\hat{\mathbf{V}}_{\text{pre}}$:

$$\min_{\mathbf{U}_T} \left\| \mathbf{Y}_{T,\text{pre}} - \mathbf{U}_T \hat{\mathbf{V}}_{\text{pre}}^\top \right\|$$

Choose rank of $\mathbf{L}$ through crossvalidation (equivalent to regularization through rank).

Two Issues

- Xu's approach does not work with staggered adoption (there may be only few units who never adopt), or general assignment pattern.

- Xu's method is not efficient because it does not use the $\mathbf{Y}_{T,\text{pre}}$ data to estimate $\mathbf{V}$.

Modified Xu (2017) method:

$$\min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \sum_{(i,t)\in\mathcal{O}} (Y_{it} - L_{it})^2 + \lambda_L \|\mathbf{L}\|_R$$

- More efficient, uses all data.

- Works with staggered adoption and general missing data pattern.

- Computationally **intractable** with large N and T because of non-convexity of objective function (like best subset selection in regression).

Our proposed method: regularize using using nuclear norm:

$$\hat{\mathbf{L}} = \min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \sum_{(i,t)\in\mathcal{O}} (Y_{it} - L_{it})^2 + \lambda_L \|\mathbf{L}\|_*$$

• The nuclear norm $\|\cdot\|_*$ generally leads to a low-rank solution for $\mathbf{L}$, the way LASSO leads to selection of regressors.

• Problem is convex, so fast solutions available.

Estimation: $\widehat{\mathbf{L}}$ is obtained via the following procedure*:

(1) Initialize $\widehat{\mathbf{L}}_1$ by $\mathbf{0}_{N \times T}$.

(2) For $k = 1, 2, \ldots$ repeat till convergence ($\mathcal{O}$ is where we observe $\mathbf{Y}$):

$$\widehat{\mathbf{L}}_{k+1} = \text{Shrink}_{\lambda}\left(P_{\mathcal{O}}(\mathbf{Y}) + P_{\mathcal{O}}^{\perp}(\widehat{\mathbf{L}}_k)\right)$$

Here $P_{\mathcal{O}}$, $P_{\mathcal{O}}^{\perp}$, and Shrink$_{\lambda}$ are matrix operators on $\mathbb{R}^{N \times T}$. For any $\mathbf{A}_{N \times T}$, $P_{\mathcal{O}}(\mathbf{A})$ is equal to $\mathbf{A}$ on $\mathcal{O}$ and is equal to 0 outside of $\mathcal{O}$. $P_{\mathcal{O}}^{\perp}(\mathbf{A})$ is the opposite; it is equal to 0 on $\mathcal{O}$ and is equal to $\mathbf{A}$ outside of $\mathcal{O}$.

For SVD $\mathbf{A} = \mathbf{S}\mathbf{\Sigma}\mathbf{R}'$ with $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots \sigma_{\min(N,T)})$,

$$\text{Shrink}_{\lambda}(\mathbf{A}) = \mathbf{S}\,\text{diag}(\sigma_1 - \lambda, \ldots, \sigma_{\ell} - \lambda, \underbrace{0, \ldots, 0}_{\min(N,T)-\ell})\,\mathbf{R}'.$$

where σ_{ℓ} is the smallest singular value of $\mathbf{A}$ that is larger than λ.

*More details in Mazumder, Hastie, and Tibshirani (2010)

General Case: We estimate $\mathbf{H}$, $\mathbf{L}$, δ, γ, and β as

$$\min_{\mathbf{H},\mathbf{L},\delta,\gamma}\left\{\frac{1}{|\mathcal{O}|}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}-L_{it}-\sum_{\substack{1\leq p\leq P\\1\leq q\leq Q}}X_{ip}H_{pq}Z_{qt}-\gamma_i-\delta_t-V_{it}\beta\right)^2+\lambda_L\|\mathbf{L}\|_*+\lambda_H\|\mathbf{H}\|_{1,e}\right\}$$

- The same estimation procedure as before applies here with an additional Shrink operator for $\mathbf{H}$.

- We choose λ_L and λ_H through crossvalidation.

Additional Generalizations I:

- Allow for propensity score weighting to focus on fit where it matters:

Model propensity score $E_{it} = \text{pr}(W_{it} = 1|X_i, Z_t, V_{it})$, $\mathbf{E}$ is $N \times T$ matrix with typical element E_{it}

Possibly using matrix completion:

$$\min_{\mathbf{E}} \frac{1}{NT} \sum_{i,t} (W_{it} - E_{it})^2 + \lambda_L \|\mathbf{E}\|_*$$

and then

$$\min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \sum_{(i,t)\in\mathcal{O}} \frac{\widehat{E}_{it}}{1 - \hat{E}_{it}} (Y_{it} - L_{it})^2 + \lambda_L \|\mathbf{L}\|_*$$

Additional Generalizations II:

- Take account of of time series correlation in $\varepsilon_{it} = Y_{it} - L_{it}$

Modify objective function from logarithm of Gaussian likelihood based on independence to have autoregressive structure.

Adaptive Properties of Matrix Regression I

Suppose N is large, T is small, $W_{it} = 0$ if $t < T$ (ATE under unconf setting), and the data-generating-process is

$$Y_{iT} = \mu + \sum_{t=1}^{T-1} \alpha_t Y_{it} + \varepsilon_{iT}, \qquad \varepsilon_{iT} \perp\!\!\!\perp (Y_{i1}, \ldots, Y_{i,T-1})$$

Then matrix regression ≈ horizontal regression, and $\gamma_i = 0$, $\delta = (0, 0, \ldots, \mu)$, and rank $T-1$ matrix

$$\mathbf{L} = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1,T-1} & \mu + \sum_{t=1}^{T-1} \alpha_t Y_{1t} \\ Y_{21} & Y_{22} & \ldots & Y_{2,T-1} & \mu + \sum_{t=1}^{T-1} \alpha_t Y_{2t} \\ Y_{31} & Y_{32} & \ldots & Y_{3,T-1} & \mu + \sum_{t=1}^{T-1} \alpha_t Y_{3t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & \ldots & Y_{N,T-1} & \mu + \sum_{t=1}^{T-1} \alpha_t Y_{Nt} \end{pmatrix} \qquad \text{(rank } T{-}1\text{)}$$

Adaptive Properties of Matrix Regression II

Suppose N is small, T is large, single treated unit, (synthetic control setting) and the data-generating-process is

$$Y_{Nt} = \mu + \sum_{i=1}^{N-1} \alpha_i Y_{it} + \varepsilon_{Nt}, \qquad \varepsilon_{Nt} \perp\!\!\!\perp (Y_{1t}, \ldots, Y_{N-1,t})$$

Then matrix regression $\approx$ vertical regression, and $\gamma_i = (0, 0, \ldots, \mu)$, $\delta = 0$, and rank $N-1$ matrix

$$\mathbf{L} = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1T} \\ Y_{21} & Y_{22} & \ldots & Y_{2T} \\ Y_{31} & Y_{32} & \ldots & Y_{3T} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N-1,1} & Y_{N-1,2} & \ldots & Y_{N-1,T} \\ \mu + \sum_{i=1}^{N-1} \omega_i Y_{i1} & \mu + \sum_{i=1}^{N-1} \omega_i Y_{i2} & \ldots & \mu + \sum_{i=1}^{N-1} \omega_i Y_{iT} \end{pmatrix}$$

Results I: If there are no covariates (just $\mathbf{L}$), $\mathcal{O}$ is sufficiently random, and $\varepsilon_{it} = Y_{it} - L_{it}$ are iid with variance σ^2.

Recall $\|\mathbf{Y}\|_F = \sqrt{\sum_{i,t} Y_{it}^2}$ and $\|\mathbf{Y}\|_\infty = \max_{i,t} |Y_{it}|$.

Let $\mathbf{Y}^*$ be the matrix including all the missing values; e.g., $\mathbf{Y}(0)$. Our estimate $\hat{\mathbf{Y}}$ for $\mathbf{Y}^*$ is $\hat{\mathbf{L}}$.

The estimated matrix $\hat{\mathbf{Y}}$ is close to $\mathbf{Y}^*$ in the following sense*:

$$\frac{\left\|\hat{\mathbf{Y}} - \mathbf{Y}^*\right\|_F}{\|\mathbf{Y}^*\|_F} \leq C \max\left(\sigma, \frac{\|\mathbf{Y}^*\|_\infty}{\|\mathbf{Y}^*\|_F}\right) \frac{\text{rank}(\mathbf{L})(N+T)\ln(N+T)}{|\mathcal{O}|}.$$

Often the number of observed entries $|\mathcal{O}|$ is of order $N \times T$ so if $\text{rank}(\mathbf{L}) \ll \min(N, T)$ and $\|\mathbf{Y}^*\|_\infty / \|\mathbf{Y}^*\|_F < \infty$, as $N + T$ grows, the error goes to 0.

*Adapting the analysis of Negahban and Wainwright (2012)

Results II

To get confidence interval for $Y_{it}(1 - Y_{it}(0)$ (for treated unit with $W_{it} = 1$), we need confidence interval for L_{it} **and** distributional assumption on $\varepsilon_{it} = Y_{it}(0) - L_{it}$ (e.g., normal, $\mathcal{N}(0, \sigma^2)$).

- To estimate L_{it} consistently, and have distributional results, we need N and T to be large (even when rank($\mathbf{L}$) $= 1$).

- We assume $\mathbf{L}_{N \times T}$ is a rank R matrix, R fixed as N, T increase. (Can probably be relaxed to let R increase slowly.)

Large sample properties of $\widehat{L}_{it}$, following Bai (2003). Decompose $\mathbf{L}$ as a rank R matrix:

$$\mathbf{L}_{N\times T} = \mathbf{U}_{N\times R}\mathbf{V}'_{T\times R}$$

Define

$$\Sigma_U = \frac{1}{N}\mathbf{U}^\top\mathbf{U} \qquad \Omega_i = \mathbf{U}_i^\top\Sigma_U^{-1}\sigma^2\Sigma_U^{-1}\mathbf{U}_i$$

$$\Sigma_V = \frac{1}{T}\mathbf{V}^\top\mathbf{V} \qquad \Psi_t = \mathbf{V}_t^\top\Sigma_V^{-1}\sigma^2\Sigma_V^{-1}\mathbf{V}_t$$

Then

$$\left(\sqrt{\frac{\Omega_i}{N}+\frac{\Psi_t}{T}}\right)^{-1}\left(\widehat{L}_{it}-L_{it}\right)\xrightarrow{d}\mathcal{N}(0,1)$$

Illustrations

- To assess root-mean-squared-error, not to get point estimate. We take a complete matrix $\mathbf{Y}$, drop some entries and compare imputed to actual values. We compare five estimators

  - DID

  - SC-ADH (Abadie-Diamond-Hainmueller)

  - EN (Elastic Net, Doudchenko-Imbens)

  - EN-T (Elastic Net Transposed, Doudchenko-Imbens)

  - MC-NNM (Matrix Completion, Nuclear-Norm Min)

Illustration I California Smoking Example

Take Abadie-Diamond-Hainmueller California smoking data. Consider two settings:

- Case 1: Simultaneous adoption

$$W = \begin{array}{cl}
\left(\begin{array}{ccc|ccc}
0 & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & 0 & \dots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \dots & 0 \\
\hline
0 & 0 & 0 & 1 & \dots & 1 \\
0 & 0 & 0 & 1 & \dots & 1 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 1 & \dots & 1
\end{array}\right)
&
\begin{array}{l}
\\ \\ \\ \longleftarrow N_c \\ \\ \\ \\ \longleftarrow N_c + N_t = 38
\end{array}
\\
\begin{array}{ccc|ccc}
 & & \uparrow & & & \uparrow \\
 & & T_0 & & & T = 31
\end{array}
&
\end{array}$$

Illustration I California Smoking Example

Take Abadie-Diamond-Hainmueller California smoking data. Consider two settings:

- Case 2: Staggered adoption

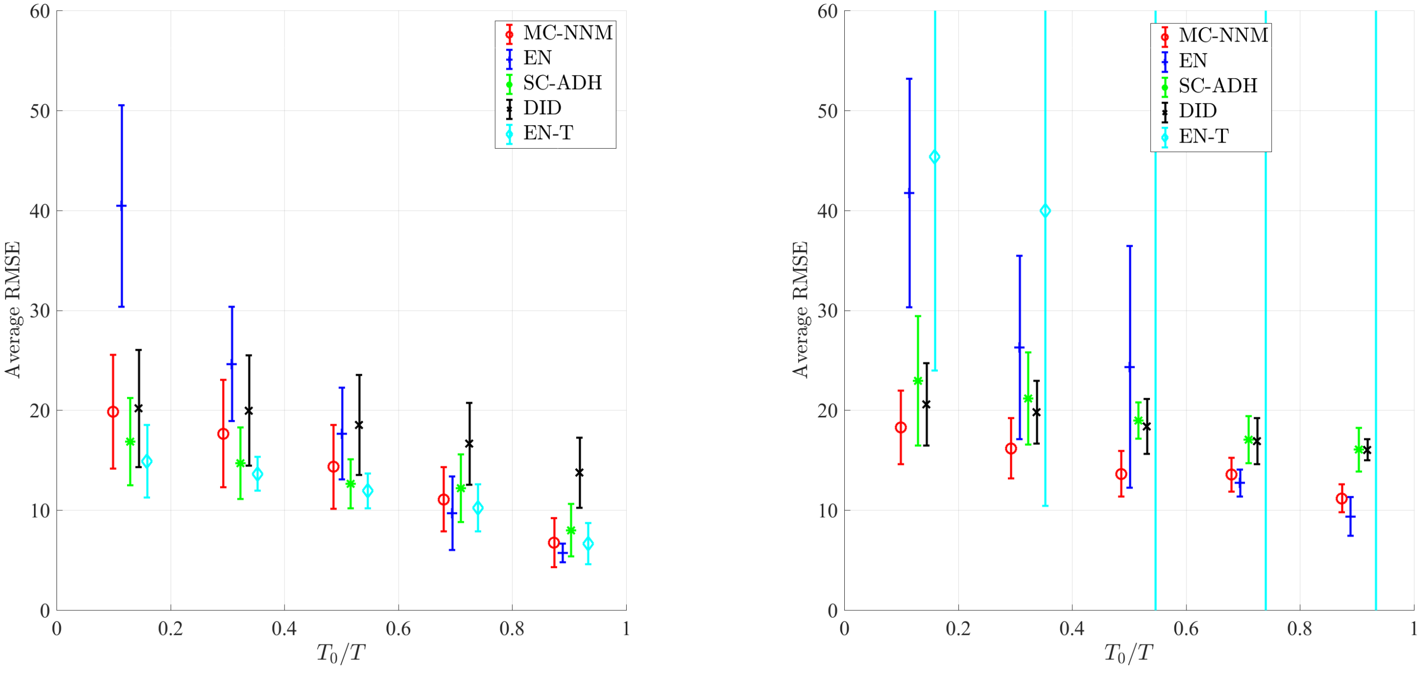
$$W = \left(\begin{array}{ccc|ccccccccc|l}
0 & 0 & 0 & 0 & \dots & & & & & & & 0 & \\
0 & 0 & 0 & 0 & \dots & & & & & & & 0 & \\
\vdots & \ddots & \vdots & \vdots & \ddots & & & & & & & \vdots & \\
0 & 0 & 0 & 0 & \dots & & & & & & & 0 & \leftarrow N_c \\
\hline
0 & 0 & 0 & 0 & \dots & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 & \\
0 & 0 & 0 & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots & \\
0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \\
0 & 0 & 0 & 0 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \\
0 & 0 & 0 & 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 & \leftarrow N_c + N_t = 38 \\
 & & \uparrow & & & & & & & & & \uparrow & \\
 & & T_0 & & & & & & & & & T = 31 &
\end{array}\right)$$

We report average RMSE for different ratios T_0/T.

Illustration I California Smoking Example $(N = 38, T = 31)$

Simultaneous adoption, $N_t = 8$

Staggered adoption, $N_t = 35$



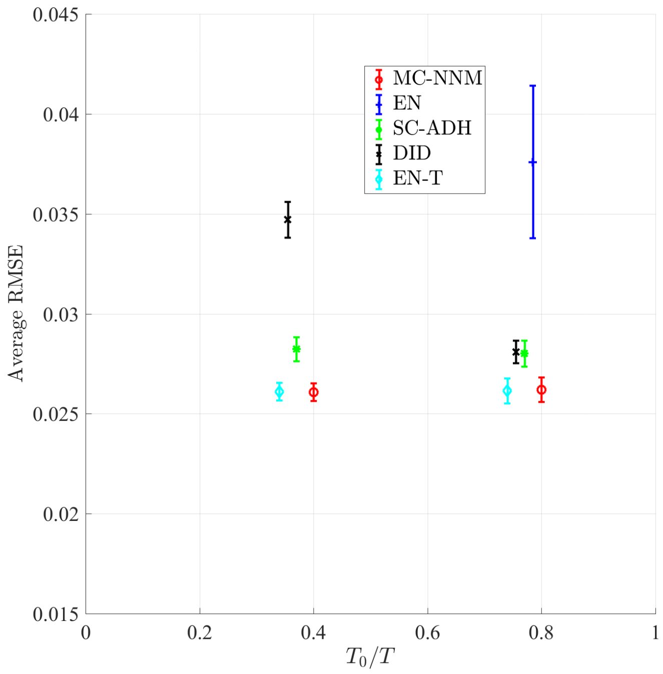
Illustrations II Stock Market Data

Daily returns on $\approx$ 2400 stocks, for $\approx$ 3000 days. We pick N stocks at random, for first T periods. This is our sample.

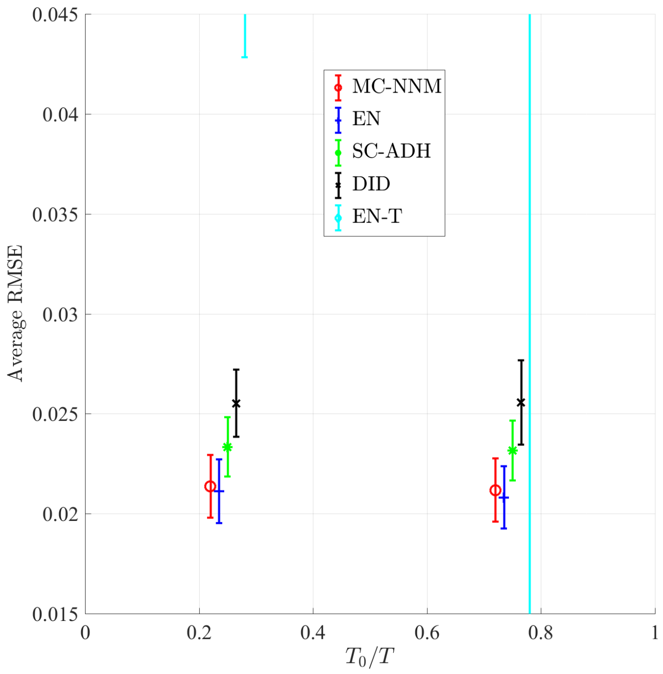
We then pick $\lfloor N/2 \rfloor$ stocks at random from the sample, consider the simultaneous adoption case with T_0 in $\{\lfloor 0.25T \rfloor, \lfloor 0.75T \rfloor\}$, impute the missing data and compare to actual data.

We repeat this 5 times for two pairs of (N, T): $(N, T) = (1000, 5)$ (thin) and $(N, T) = (5, 1000)$ (fat).

Illustrations II Stock Market Data



Thin: $(N, T) = (1000, 5)$



Fat: $(N, T) = (5, 1000)$

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. ”Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program.” *Journal of the American statistical Association* 105.490 (2010): 493-505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. ”Comparative politics and the synthetic control method.” *American Journal of Political Science* 59.2 (2015): 495-510.

Abadie, Alberto, and Jeremy L'Hour “A Penalized Synthetic Control Estimator for Disaggregated Data”

Athey, Susan, Guido W. Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. arXiv preprint arXiv:1604.07125v3 (2016).

Bai, Jushan. ”Inferential theory for factor models of large dimensions.” *Econometrica* 71.1 (2003): 135-171.

Bai, Jushan. ”Panel data models with interactive fixed effects.” *Econometrica* 77.4 (2009): 1229-1279.

Bai, Jushan, and Serena Ng. ”Determining the number of factors in approximate factor models.” *Econometrica* 70.1 (2002): 191-221.

Candés, Emmanuel J., and Yaniv Plan. ”Matrix completion with noise.” *Proceedings of the IEEE* 98.6 (2010): 925-936.

Candés, Emmanuel J., and Benjamin Recht. "Exact matrix completion via convex optimization." *Foundations of Computational mathematics* 9.6 (2009): 717.

Chamberlain, G., and M. Rothschild. "Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* 51 12811304, 1983.

Chernozhukov, Victor, et al. "Double machine learning for treatment and causal parameters." arXiv preprint arXiv:1608.00060 (2016).

Doudchenko, Nikolay, and Guido W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. No. w22791. National Bureau of Economic Research, 2016.

Gross, David. “Recovering Low-Rank Matrices From Few Coefficients in Any Basis”, *IEEE Transactions on Information Theory*, Volume: 57, Issue: 3, March 2011

Gobillon, Laurent, and Thierry Magnac. ”Regional policy evaluation: Interactive fixed effects and synthetic controls.” *Review of Economics and Statistics* 98.3 (2016): 535-551.

Keshavan, Raghunandan H., Andrea Montanari, and Sewoong Oh. “Matrix Completion from a Few Entries.” *IEEE Transactions on Information Theory*, vol. 56,no. 6, pp.2980-2998, June 2010

Keshavan, Raghunandan H., Andrea Montanari. and Sewoong Oh. “Matrix completion from noisy entries.” *Journal of Machine Learning Research* 11.Jul (2010): 2057-2078.

Kim,D., AND T.Oka (2014):“Divorce Law Reforms and Divorce Rates in the USA: An Interactive Fixed-Effects Approach, *Journal of Applied Econometrics*, 29, 231245 (2014).

Mazumder, Rahul and Hastie, Trevor and Tibshirani, Rob. “Spectral Regularization Algorithms for Learning Large Incomplete Matrices”, *Journal of Machine Learning*, 11 2287-2322 (2010).

Moon, Hyungsik Roger, and Martin Weidner. ”Linear regression for panel with unknown number of factors as interactive fixed effects.” *Econometrica* 83.4 (2015): 1543-1579.

Liang, Dawen, et al. ”Modeling user exposure in recommendation.” *Proceedings of the 25th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 2016.

Negahban, Sahand and Wainwright, Martin. “Estimation of (near) low-rank matrices with noise and high-dimensional scaling”, *Annals of Statistics*, Vol 39, Number 2, pp. 1069–1097.

Negahban, Sahand and Wainwright, Martin. “Restricted strong convexity and weighted matrix completion: Optimal bounds with noise” *Journal of Machine Learning Research*, 13: 1665-1697, May 2012

Recht, Benjamin. “A Simpler Approach to Matrix Completion”, *Journal of Machine Learning Research* 12:3413-3430, 2011

Xu, Yiqing. ”Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models.” *Political Analysis* 25.1 (2017): 57-76.