Low-Rank Approximations of Nonseparable Panel Models

Iván Fernández-Val
BU

Hugo Freeman
UCL

Martin Weidner
UCL

Nov 2020

(BOE conference)

Introduction

- Model:

$$Y_{it} = g(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}), \qquad i = 1, \ldots, N, \; t = 1, \ldots, T,$$

  where Y_{it} and $\boldsymbol{X}_{it}$ are observed, while $\boldsymbol{A}_i$, $\boldsymbol{B}_t$, $\boldsymbol{U}_{it}$ are unobserved, and $g(\cdot)$ is unknown.

- Panel data allows us to control for unobserved confounding variables $\boldsymbol{A}_i$ (constant over t) and $\boldsymbol{B}_t$ (constant across i). Those are allowed to be correlated to the observed covariates $\boldsymbol{X}_{it}$ ("fixed effect approach").

- Goal: estimate effect of $\boldsymbol{X}_{it}$ on Y_{it}, while controlling for $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$.

Example: empirical illustration

Effect of election day registration (EDR) laws on vote turnout in the US
(dataset from *Xu, 2017*)

- $N = 47$ states, $T = 24$ presidential elections (1920-2012).
- Y_{it} = voter turnout rate.
- $X_{it} \in \{0, 1\}$, indicator for EDR law that allows eligible voters to register on election day.
- 4 waves of EDR adoption: ME, MN and WI in 1976; WY, ID and NH in 1994; MT and IA in 2008; and CT in 2012

$\Rightarrow$ We want to estimate the average treatment effect on the treated, while controlling for state specific heterogeneity $\boldsymbol{A}_i$ and election specific heterogeneity $\boldsymbol{B}_t$.

Introduction

- We observe $Y_{it}(0) := Y_{it}$ for pairs (i, t) with $X_{it} = 0$.

$\Rightarrow$ Want to to impute the unobserved potential outcome $Y_{it}(0)$ for pairs (i, t) with $X_{it} = 1$.

- We are going to do this using matrix completion methods, which rely on the $N \times T$ matrix of expected outcomes $\mathrm{E}\left[Y_{it}(0) \,\middle|\, \boldsymbol{A}^N, \boldsymbol{B}^T\right]$ to have good low-rank approximations.

Econometric Applications of Matrix Completion Methods

- *Athey, Bayati, Doudchenko, Imbens & Khosravi (2017)* and *Bai and Ng (2019)* apply matrix completion methods to estimate ATE.
- *Chernozhukov, Hansen, Liao & Zhu (2018)* consider the case of "spiked low-rank matrices" whose rank is allowed to converge to infinity.
- *Archangelsky, Athey, Hirshberg, Imbens & Wager (2019)* derived consistency results for synthetic control estimators based on matrix completion methods.
- *Chen, Fan, Ma & Yang (2019)* provided non-asymptotic distributional guarantees for debiased convex and nonconvex matrix completion estimators under normality and missing at random.
- *Moon & Weidner (2018)*, *Beyhum & Gautier (2019)* consider nuclear norm regularized estimators of the linear model with factor structure.

etc.

Main contribution of our paper

- We do <u>not</u> assume that the true DGP has a low-rank structure, but allow for a general non-separable model $Y_{it} = g(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it})$.
- Our results highlight the potential of low-rank structures to approximate very general DGPs.
- We suggest a new estimation method for the treatment effects based on our DGP (where g is smooth, and $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are low-dimensional).
- (in practice, one might want to more parametric models like *Pesaran 2006* and *Bai 2009*, but it is useful to know that the general nonparametric model allows "identification" = consistent estimation as $N, T \to \infty$).

Principal component analysis (PCA)

- Notice that $\boldsymbol{Y} = (Y_{it})$ is an $N \times T$ matrix, and we are interested in applications where both N and T are large.

- Goal: Approximate the $N \times T$ matrix $\boldsymbol{Y}$ by a low-rank matrix:

$$Y_{it} \approx \sum_{r=1}^{R} \lambda_{ir} f_{tr}$$

$\Rightarrow$ calculate the singular value decomposition (SVD)

$$Y_{it} = \sum_{r=1}^{\max(N,T)} \underbrace{s_r\, u_{ir}}_{=\lambda_{ir}} \underbrace{v_{tr}}_{=f_{tr}}$$

(same as calculating the eigenvalue decomposition of $\boldsymbol{Y}\boldsymbol{Y}'$ or $\boldsymbol{Y}'\boldsymbol{Y}$)

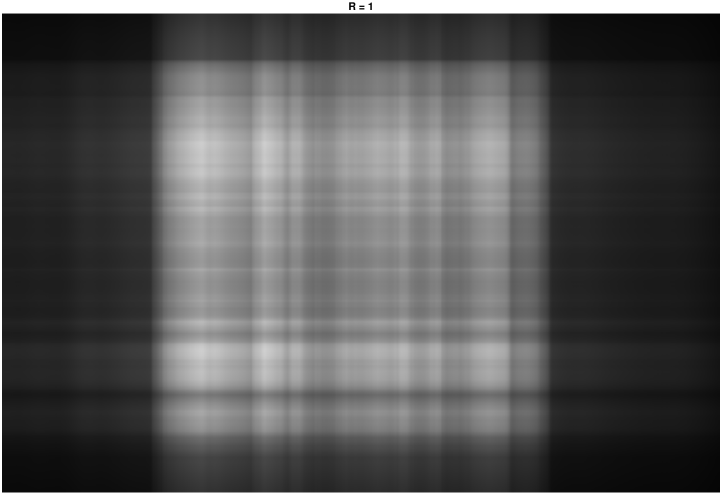
$\Rightarrow$ only keep the R largest singular values s_r for the approximation.

Grayscale Image Example

orginal image

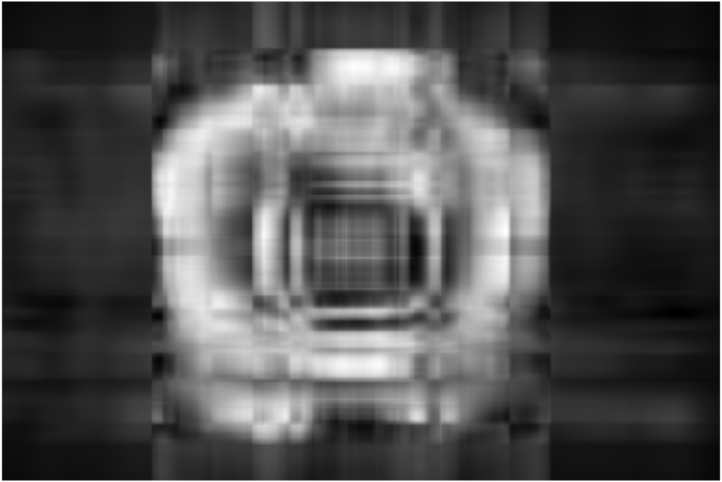
- This grayscale image can be interpreted as 750×1125 matrix.

Grayscale Image Example (cont.)



- Using 1 principal component to reconstruct the image.

Grayscale Image Example (cont.)



- Using 5 principal components to reconstruct the image.

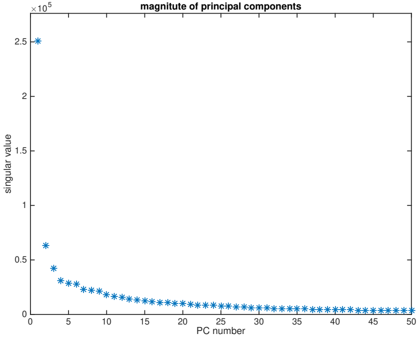
Grayscale Image Example (cont.)

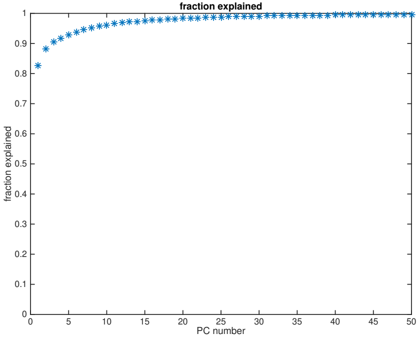
- Using 20 principal components to reconstruct the image.

Grayscale Image Example (cont.)

- Using 50 principal components to reconstruct the image.

Grayscale Image Example (cont.)





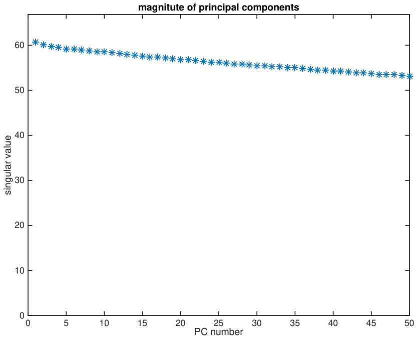
- The singular values are quickly decreasing with R.
- The fraction of total variation explained quickly approaches one as R increases.
- Analogous plots for actual economic variables.
  (e.g. Y_{it} = GDP of country i at time t)

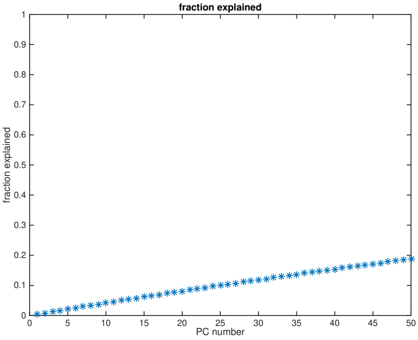
Is the same true for any large matrix?

- Can the first few principal components always explain a large fraction of the data?

Is the same true for any large matrix?

- Can the first few principal components always explain a large fraction of the data?
- No
  e.g., for a 750×1125 matrix with $e_{it} \sim \text{i.i.d.}\,\mathcal{N}(0,1)$ (pure noise!) we find:





When can low-rank approximation explain the mean of Y_{it}?

- Factor Model / Interactive Fixed Effects Model:

$$Y_{it} = \sum_{r=1}^{R} \lambda_{ir} f_{tr} + e_{it},$$

where λ_{ir} are unobserved "factor loading" (R individual specific effects), f_{tr} are unobserved "factors" (R time specific effects), and e_{it} are unobserved "idiosyncratic errors" (mean zero noise).

$\Rightarrow$ see e.g. *Stock and Watson (2002)*, *Bai and Ng (2002)*, *Bai (2003)*, ...

$\Rightarrow$ In that case the PCA estimators $\widehat{\lambda}_{ir}$ and $\widehat{f}_{tr}$ (after appropriate normalization choice) converge to λ_{ir} and f_{tr} as $N, T \to \infty$.

When can low-rank approximation explain the mean of Y_{it}?

- Nonseparable model: (no covariates, yet)

$$Y_{it} = g(\boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}),$$

where we assume that the noise term satisfies

$$\boldsymbol{U}_{it} \stackrel{d}{=} \boldsymbol{U}_{js} \mid \boldsymbol{A}^N, \boldsymbol{B}^T$$

- By defining $m(\boldsymbol{A}_i, \boldsymbol{B}_t) := \mathrm{E}\left[Y_{it} \mid \boldsymbol{A}_i, \boldsymbol{B}_t\right]$ and $E_{it} := Y_{it} - m(\boldsymbol{A}_i, \boldsymbol{B}_t)$ we can rewrite the model as

$$Y_{it} = m(\boldsymbol{A}_i, \boldsymbol{B}_t) + E_{it}$$

$\Rightarrow$ $m(\boldsymbol{A}_i, \boldsymbol{B}_t)$ can be well-approximated by a low rank matrix if

(1) $\dim(\boldsymbol{A}_i)$ and $\dim(\boldsymbol{B}_t)$ are relatively small.

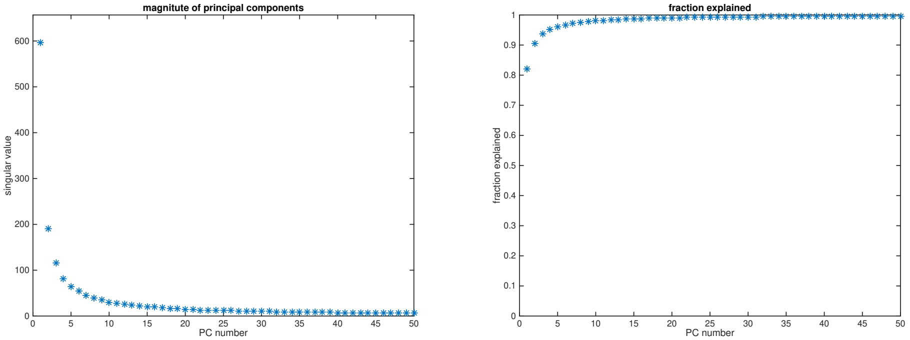
(2) $m(\cdot, \cdot)$ is well-behaved. (e.g. sufficiently smooth)

Simple example

Binary choice mean function:

$$m(A_i, B_t) = \mathbb{1}(A_i + B_t > 0), \qquad \text{with } A_i, B_t \sim \text{i.i.d.}\,\mathcal{N}(0,1)$$

$\Rightarrow$ again simulating a 750×1125 matrix from this DGP gives



Full model with covariates

- Model:

$$Y_{it} = g(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}), \quad i \in \mathbb{N} = \{1, \ldots, N\},\ t \in \mathbb{T} = \{1, \ldots, T\},$$

  where Y_{it}, $\boldsymbol{X}_{it}$ observed; $\boldsymbol{A}_i$, $\boldsymbol{B}_t$, $\boldsymbol{U}_{it}$ unobserved; g unknown.

- Assumptions:

$$\boldsymbol{U}_{it} \overset{d}{=} \boldsymbol{U}_{js} \mid \boldsymbol{X}^{NT}, \boldsymbol{A}^N, \boldsymbol{B}^T, \qquad \text{for all } i, j \in \mathbb{N},\ t, s \in \mathbb{T},$$

  and

$$\boldsymbol{U}_{it} \perp\!\!\!\perp \boldsymbol{X}_{js} \mid \boldsymbol{A}^N, \boldsymbol{B}^T, \qquad \text{for all } i, j \in \mathbb{N},\ t, s \in \mathbb{T},$$

Motivation for this model

- This model can be motivated from a purely statistical perspective as a latent variable model using the Aldous-Hoover representation for exchangeable arrays, e.g. Xu, Massouli and Lelarge (2014), Chatterjee (2015), Orbanz and Roy (2015), and Li and Bell (2017).

- We think of it as a structural model where the unobserved effects $\boldsymbol{A}_i$ and $\boldsymbol{B}_t$ are associated with individual heterogeneity and aggregate shocks, respectively.

- Our model similar to the nonseparable panel model in Chernozhukov, Fernández-Val, Hahn and Newey (2013), but we incorporate time effects $\boldsymbol{B}_t$, which allow the relationship between Y_{it} and $\boldsymbol{X}_{it}$ to vary over time in an unrestricted fashion.

Parameters of interest

- The structural function itself g is generally not identified.
- Let $Y_{it}(\boldsymbol{x}) := g(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}(\boldsymbol{x}))$ be the potential outcome obtained by setting exogenously $\boldsymbol{X}_{it} = \boldsymbol{x}$ and drawing $\boldsymbol{U}_{it}(\boldsymbol{x}) \overset{d}{=} \boldsymbol{U}_{it} \mid \boldsymbol{A}^N, \boldsymbol{B}^T$. Average structural functions (ASFs):

$$\mu(\boldsymbol{x}) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{E}\left[Y_{it}(\boldsymbol{x}) \,\middle|\, \boldsymbol{A}^N, \boldsymbol{B}^T \right]$$

- In the paper we also consider $\mu_t(\boldsymbol{x})$ specific on $t = 1, \ldots, T$, conditional ASF (e.g. for ATT estimation), and and also discuss quantile treatment effect.
- In the following we will focus on the case $\boldsymbol{X}_{it} \in \{0, 1\}$, implying that

$$\mu(1) - \mu(0)$$

is the average treatment effect.

$\Rightarrow$ How to estimate those effects?

PCA = Least Squares Estimator

- Without covariates the PCA estimator reads

$$\left\{\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{f}}\right\} \in \underset{\{\boldsymbol{\lambda}\in\mathbb{R}^{N\times R}, \boldsymbol{f}\in\mathbb{R}^{T\times R}\}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \sum_{r=1}^{R}\lambda_{ir}\, f_{tr}\right)^2$$

$\Rightarrow$ $\widehat{\mathrm{E}}\left[Y_{it} \,\middle|\, \boldsymbol{A}^N, \boldsymbol{B}^T\right] = \sum_{r=1}^{R} \widehat{\lambda}_{ir}\, \widehat{f}_{tr}$.

$\Rightarrow$ easy and fast to compute via SVD.

- With covariates we need $\mathrm{E}\left[Y_{it}(x) \,\middle|\, \boldsymbol{A}^N, \boldsymbol{B}^T\right]$ for $x \in \{0,1\}$.

$\Rightarrow$ for each $x \in \{0,1\}$ the outcome $Y_{it}(x)$ is only observed for a subset $\mathbb{D}(x)$ of pairs (i,t).

$\Rightarrow$ PCA for unbalanced panels?

Matrix Completing (nuclear norm minimization)

- The problem

$$\underset{\{\boldsymbol{\lambda}\in\mathbb{R}^{N\times R},\boldsymbol{f}\in\mathbb{R}^{T\times R}\}}{\text{argmin}} \sum_{(i,t)\in\mathbb{D}(x)} \left(Y_{it} - \boldsymbol{\lambda}_i'\boldsymbol{f}_t\right)^2$$

can equivalently also be expressed as

$$\min_{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}} \sum_{(i,t)\in\mathbb{D}(x)} \left(Y_{it} - \Gamma_{it}\right)^2 \quad \text{s.t.} \quad \text{rank}(\boldsymbol{\Gamma}) \leq R,$$

where $\boldsymbol{\Gamma}$ is an $N \times T$ matrix.

- Used here:

$$\boldsymbol{\Gamma} = \boldsymbol{\lambda}\,\boldsymbol{f}' \quad \Leftrightarrow \quad \text{rank}(\boldsymbol{\Gamma}) \leq R \quad \Leftrightarrow \quad \sum_{r=1}^{\min(N,T)} \mathbb{1}\left(s_r(\boldsymbol{\Gamma}) > 0\right) \leq R,$$

where $s_1(\boldsymbol{\Gamma}) \geq s_2(\boldsymbol{\Gamma}) \geq \ldots \geq s_{\min(N,T)}(\boldsymbol{\Gamma}) \geq 0$ are the singular values of $\boldsymbol{\Gamma}$.

Matrix Completing (nuclear norm minimization)

- $\mathrm{rank}(\boldsymbol{\Gamma}) \leq R$ is a non-convex constraint.

- Convex relaxation of this constraint:

$$\underbrace{\sum_{r=1}^{\min(N,T)} s_r(\boldsymbol{\Gamma})}_{=:\|\boldsymbol{\Gamma}\|_1} \leq \mathrm{const.}$$

  where $\|\boldsymbol{\Gamma}\|_1$ is the nuclear norm (or trace norm).

- An estimate for $\boldsymbol{\Gamma} = \boldsymbol{\lambda f}'$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}(x) &= \underset{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}}{\mathrm{argmin}} \sum_{(i,t)\in\mathbb{D}(x)} \left(Y_{it} - \boldsymbol{\Gamma}_{it}\right)^2 \quad \text{s.t.} \quad \|\boldsymbol{\Gamma}\|_1 \leq \mathrm{const.} \\ &= \underset{\boldsymbol{\Gamma}\in\mathbb{R}^{N\times T}}{\mathrm{argmin}} \sum_{(i,t)\in\mathbb{D}(x)} \left(Y_{it} - \Gamma_{it}\right)^2 + \rho\,\|\Gamma\|_1,\end{aligned}$$

  where $\rho > 0$ is a penalty parameter. This is a convex problem.

- See Recht, Fazel and Parrilo (2010) and Hastie, Tibshirani and Wainwright (2015) for surveys on "matrix completion".

Estimation of ASF and ATE

- Matrix completion via nuclear norm minimization:

$$\widehat{\boldsymbol{\Gamma}}(x) = \underset{\boldsymbol{\Gamma} \in \mathbb{R}^{N \times T}}{\operatorname{argmin}} \sum_{(i,t) \in \mathbb{D}(x)} (Y_{it} - \Gamma_{it})^2 + \rho \, \|\boldsymbol{\Gamma}\|_1,$$

- Average across i, t to estimate ASF

$$\widehat{\mu}(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[D_{it}(x) Y_{it} + \{1 - D_{it}(x)\} \widehat{\Gamma}_{it}(x) \right],$$

where $D_{it}(x) := \mathbb{1}\{X_{it} = x\}$.

- Finally,

$$\widehat{\mathrm{ATE}} = \widehat{\mu}(1) - \widehat{\mu}(0).$$

- Analogously for $\widehat{\mu}(0|1)$ to get ATT, and for time specific effects.

Sampling assumptions

- Remember:

$$
\begin{aligned}
Y_{it} &= g(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t, \boldsymbol{U}_{it}) \\
&= m(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t) + E_{it},
\end{aligned}
$$

where

$$
\begin{aligned}
m(\boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t) &:= \mathrm{E}\left[Y_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}, \boldsymbol{A}_i, \boldsymbol{B}_t\right], \\
E_{it} &:= Y_{it} - m(\boldsymbol{X}_{it}, \boldsymbol{A}_i, \boldsymbol{B}_t)
\end{aligned}
$$

We assume that:

- $\boldsymbol{A}_i$ is independent and identically distributed across i.
- $\boldsymbol{B}_t$ is independent and identically distributed over t.
- E_{it} is in independent across i and over t, conditional on $\boldsymbol{X}^{NT}$, $\boldsymbol{A}^N$, $\boldsymbol{B}^T$, with uniformly bounded fourth moments.

Smoothness assumption

- Let
$$m(x, \boldsymbol{a}, \boldsymbol{b}) = \sum_{j=1}^{\infty} s_j(x)\, u_j(x, \boldsymbol{a})\, v_j(x, \boldsymbol{b})$$
be the functional singular value decomposition of $m(x, \boldsymbol{a}, \boldsymbol{b})$. We assume that
$$\sum_{j=1}^{\infty} s_j(x) < \infty.$$
- For example, if $(\boldsymbol{a}, \boldsymbol{b}) \mapsto m(x, \boldsymbol{a}, \boldsymbol{b})$ is continuously differentiable up to order s, then
$$s_j(x) \lesssim j^{-\frac{s}{d_a \wedge d_b}},$$
by Theorem 3.3 of Griebel and Harbrecht (2013), where $d_a \wedge d_b$ is the minimum of d_a and d_b. This implies that $\sum_{j=1}^{\infty} s_j(x) < \infty$ if $s > d_a \wedge d_b$.

Consistency of Matrix Completion Estimator

Let $n(x) = |\mathbb{D}(x)|$.

Lemma

Let above assumptions hold, and let $\rho/\sqrt{N+T} \to \infty$ and $\rho\sqrt{NT}/n(x) \to 0$ as $N, T \to \infty$. Then,

$$\frac{1}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} \left[\widehat{\Gamma}_{it}(x) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)\right]^2 = o_P(1).$$

This is just a technical lemma, because the consistency result we would like to obtain is

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\widehat{\Gamma}_{it}(x) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)\right]^2 = o_P(1),$$

Restricted strong convexity

- The existing literature on matrix completion relies on the concept of restricted strong convexity to derive the desired result on the last slide. Under certain conditions on a matrix $\boldsymbol{M}$ with entries M_{it}, and on $\boldsymbol{X}^{NT}$ (which determines the set $\mathbb{D}(x)$), there exists a constant $c > 0$ such that with high probability

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} M_{it}^2 \leq \frac{c}{n(x)} \sum_{(i,t)\in\mathbb{D}(x)} M_{it}^2.$$

- See e.g. Theorem 1 in Negahban and Wainwright (2012), Lemma 12 in Klopp et al. (2014), and Lemma 3 in Athey, Bayati, Doudchenko, Imbens and Khosravi (2017).

- Thus, if the matrix $\boldsymbol{M}$ with entries $M_{it} = \widehat{\Gamma}_{it}(x) - m(x, \boldsymbol{A}_i, \boldsymbol{B}_t)$ satisfy restricted strong convexity, then the desired result follows from Lemma 1.

Main consistency theorem

We do not show

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\widehat{\Gamma}_{it}(x) - m(x,\boldsymbol{A}_i,\boldsymbol{B}_t)\right]^2 = o_P(1),$$

in our paper, but instead directly establish consistency of

$$\mu(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m(x,\boldsymbol{A}_i,\boldsymbol{B}_t).$$

Theorem

Under appropriate assumptions (see paper) we have

$$\widehat{\mu}(x) = \mu(x) + o_P(1).$$

For this result we require X_{it} to be weakly correlated across i and over t, and also $\Pr\left(X_{it} = x \mid \boldsymbol{A}^N, \boldsymbol{B}^T\right) > 0$ for all i and t.

Debiasing Using Matching Methods

- The matrix completion (MC) estimator has two sources of bias:
  - low-rank approximation bias
  - shrinkage bias

$\Rightarrow$ Those biases make inference based on the MC estimator very difficult. We therefore consider alternative debiased estimators.

Debiasing Using Matching Methods

- Let $\widehat{\boldsymbol{\lambda}}_i(x)$ and $\widehat{\boldsymbol{f}}_t(x)$ be the $R \times 1$ vectors that satisfy

$$\widehat{\Gamma}_{it}(x) = \widehat{\boldsymbol{\lambda}}_i(x)' \widehat{\boldsymbol{f}}_t(x),$$

- Simple matching estimator: for values $x \neq X_{it}$ we construct counterfactuals by

$$\breve{\Gamma}_{it}(x) = Y_{i^*(i,t,x),t^*(i,t,x)},$$

where $i^*(i,t,x) \in \mathbb{N}$ and $t^*(i,t,x) \in \mathbb{T}$ are a solutions to

$$\begin{array}{ll} \min_{j \in \mathbb{N}, s \in \mathbb{T}} & \left\| \widehat{\boldsymbol{\lambda}}_i(x) - \widehat{\boldsymbol{\lambda}}_j(x) \right\|^2 + \left\| \widehat{\boldsymbol{f}}_t(x) - \widehat{\boldsymbol{f}}_s(x) \right\|^2 \\ \text{s.t.} & X_{js} = x. \end{array}$$

- Estimate $\mu(x)$ by

$$\breve{\mu}(x) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[D_{it}(x) Y_{it} + \{1 - D_{it}(x)\} \breve{\Gamma}_{it}(x) \right],$$

Debiasing Using Matching Methods

- Two-way matching estimator: for values $x \neq X_{it}$ we construct counterfactuals by

$$\widetilde{\Gamma}_{it}(x) = Y_{i,t^*(i,t,x)} + Y_{i^*(i,t,x),t} - Y_{i^*(i,t,x),t^*(i,t,x)},$$

  where $i^*(i,t,x) \in \mathbb{N}$ and $t^*(i,t,x) \in \mathbb{T}$ are a solutions to

$$\begin{array}{cc} \min_{j\in\mathbb{N},s\in\mathbb{T}} & \left\|\widehat{\boldsymbol{\lambda}}_i(x) - \widehat{\boldsymbol{\lambda}}_j(x)\right\|^2 + \left\|\widehat{\boldsymbol{f}}_t(x) - \widehat{\boldsymbol{f}}_s(x)\right\|^2 \\ \text{s.t.} & X_{is} = X_{jt} = X_{js} = x. \end{array}$$

- Estimate $\mu(x)$ by

$$\widetilde{\mu}(x) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[D_{it}(x)Y_{it} + \{1 - D_{it}(x)\}\widetilde{\Gamma}_{it}(x)\right],$$

- Here, we find the match (j,s) with $X_{js} = x$ that not only is the closest to (i,t) in terms of the estimated factor structure, but also corresponds to a unit j with $X_{jt} = x$ and a time period s with $X_{is} = x$. Then, we estimate the counterfactual $\Gamma_{it}(x)$ as a linear combination of Y_{jt}, Y_{is} and Y_{js}.

Debiasing Using Matching Methods

- We also consider matching estimates that use multiple matches for each pair (i, t), which is variance reducing.
- In the paper we show consistency of these matching estimators $\widetilde{\mu}(x)$ under appropriate assumptions, but full inference results are still missing.

Monte Carlo simulations

- Generate data for $N = T = 30$ from the model

$$Y_{it}(x) = x + g(A_i, B_t) + U_{it}(x), \qquad \text{for } \ x \in \{0, 1\},$$

  where $U_{it}(x) \sim \text{i.i.d.}\, \mathcal{N}(0, 1/4)$, $A_i, B_t \sim \text{i.i.d.} U(0, 1)$, and for g we use the Gaussian kernel similar to that used in Bordenave, Coste and Nadakuditi (2020) and Griebel and Harbrecht (2010).

- DGP for $X_{it} \in \{0, 1\}$ that resembles the empirical application.

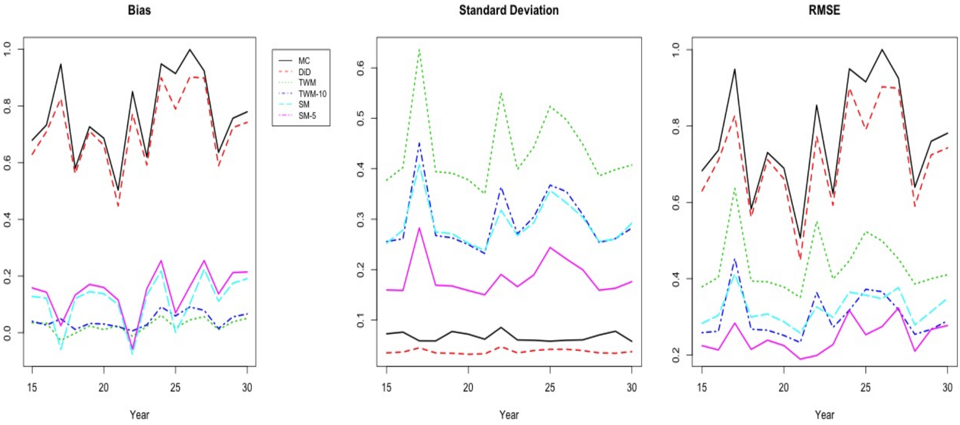
- Estimators:
  - naive difference in means (Dmeans)
  - difference-in-difference (DiD).
  - matrix completion (MC)
  - two-way matching method with k matches (TWM-k)
  - simple matching method with k matches (SM-k)

Monte Carlo simulations: results for $\mu(0 \mid 1)$

	Bias	St. Dev.	RMSE
Dmeans	0.59	0.02	0.59
DiD	0.70	0.03	0.70
MC	0.74	0.02	0.74
TWM-1	0.03	0.14	0.14
TWM-5	0.03	0.11	0.12
TWM-10	0.04	0.10	0.11
TWM-30	0.07	0.09	0.12
SM-1	0.12	0.10	0.16
SM-5	0.15	0.07	0.17
SM-10	0.19	0.06	0.20
SM-30	0.31	0.05	0.31

based on $1,000$ simulations

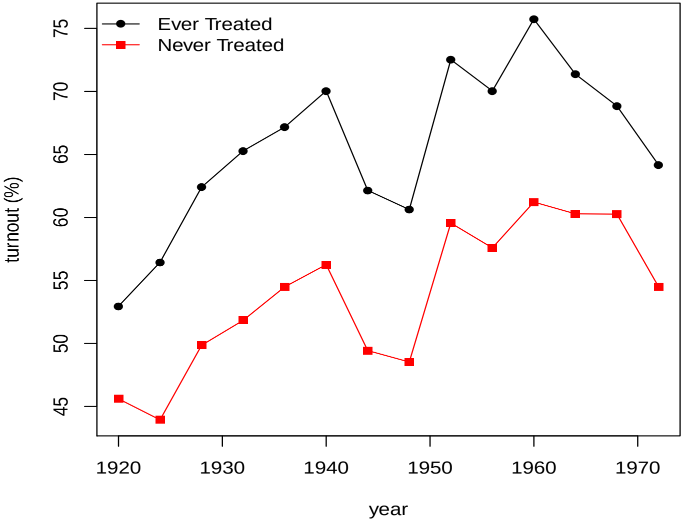
Monte Carlo simulations: results for $\mu_t(0 \mid 1)$



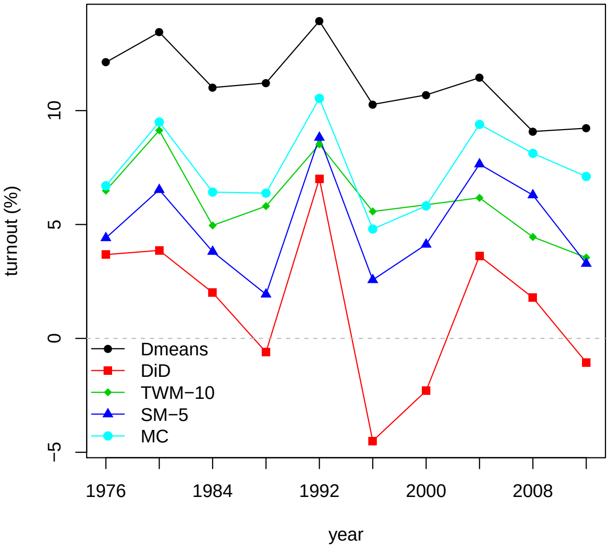
Election Day Registration (EDR) and Vote Turnout

- Effect of allowing vote registration in election day on vote turnout in the U.S. (*Xu, 2017*)
- Data: 24 presidential elections from 1920 to 2012, 47 states excluding Alaska, Hawaii and North Dakota (early adopter)
- Turnout rate, Y_{it}, is total ballots counted divided by voting-age population
- 4 waves of EDR adoption: ME, MN and WI in 1976; WY, ID and NH in 1994; MT and IA in 2008; and CT in 2012
- Focus on average treatment effect on the treated; staggered adoption (*Athey & Imbens, 2018*)
- Treated states have higher turnouts in pretreatment periods

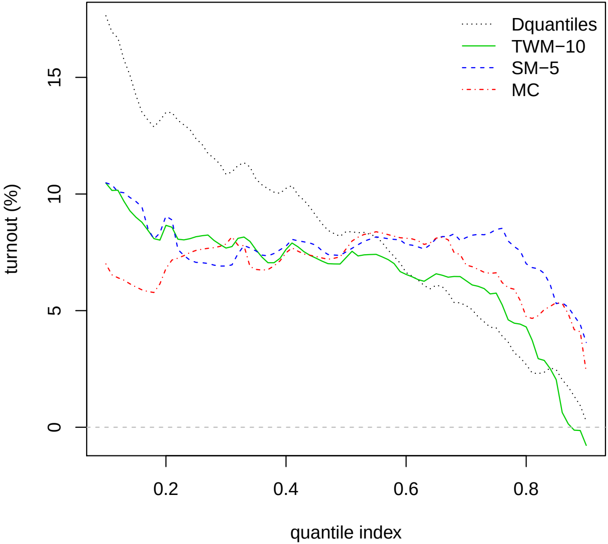
Assessing Pretreatment Parallel Trends



Average Treatment Effect on the Treated by Year



Quantile Treatment Effects on the Treated



Concluding Remarks

Main message:

- Low-rank approximations are useful for two-way fixed effects models even if the underlying DGP is not of low-rank.
- For unbalanced panels one can replace PCA with matrix completion estimators, e.g. *Athey, Bayati, Doudchenko, Imbens & Khosravi (2017)*.
- We can identify (via large N, T) interesting average effects in fully non-parametric panel data models with two-way effects.

Interesting future work:

- Choice of tuning parameters (penalty parameter ρ, or number of factors).
- Inference.
- How general can the DGP for X_{it} be?

Thank you

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2017). Matrix completion methods for causal panel data models.

Bordenave, C., S. Coste, and R. R. Nadakuditi (2020). Detection thresholds in very sparse matrix completion. *arXiv preprint arXiv:2005.06062*.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist. 43*(1), 177–214.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Griebel, M. and H. Harbrecht (2010). *Approximation of two-variate functions: Singular value decomposition versus regular sparse grids*. SFB 611.

Griebel, M. and H. Harbrecht (2013, 05). Approximation of bi-variate functions: singular value decomposition versus sparse grids. *IMA Journal of Numerical Analysis 34*(1), 28–54.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Klopp, O. et al. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli 20*(1), 282–303.

Li, K. T. and D. R. Bell (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics 197*(1), 65 – 75.

Negahban, S. and M. J. Wainwright (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research 13*(1), 1665–1697.

Orbanz, P. and D. M. Roy (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 437–461.

Recht, B., M. Fazel, and P. A. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review 52*(3), 471–501.

Xu, J., L. Massouli, and M. Lelarge (2014). Edge label inference in generalized stochastic block models: from spectral theory to impossibility results.