Matrix Completion Methods for Causal Panel Data Models

Susan Athey, Mohsen Bayati,

Nikolay Doudchenko, Guido Imbens, & Khashayar Khosravi

(Stanford University)

Motivating Example I

• California's anti-smoking legislation (Proposition 99) took effect in 1989.

• **What is the causal effect of the legislation on smoking rates in California in 1989?**

• We **observe** smoking rates in California in 1989 given the legislation. We need to **impute** the **counterfactual** smoking rates in California in 1989 had the legislation not been enacted.

• We have data in the absence of smoking legislation in California prior to 1989, and for other states both before and in 1989. (and other variables, but not of essence)

Motivating Example II

In US, on any day, 1 in 25 patients suffers at least one Hospital Acquired Infection (HAI).

• Hospital acquired infections cause 75,000 deaths per year, cost 35 billion dollars per year

• 13 states have adopted a reporting policy (at different times during 2000-2010) that requires hospitals to report HAIs.

What is (average) causal effect of reporting policy on HAIs?

Set Up: we observe (in addition to covariates):

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

$$\mathbf{W} = \begin{pmatrix} 1 & 1 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \dots & 0 \end{pmatrix} \quad \text{(binary treatment).}$$

• rows of $\mathbf{Y}$ and $\mathbf{W}$ correspond to physical units, columns correspond to time periods.

In terms of potential outcome matrices $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$:

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

$Y_{it}(0)$ observed iff $W_{it} = 0$, $Y_{it}(1)$ observed iff $W_{it} = 1$.

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

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

We need to **impute** the missing potential outcomes in $\mathbf{Y}(0)$ (and in $\mathbf{Y}(1)$ for other estimands).

Often structure on $\mathbf{W}$: treated unit/time-period block

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

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

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & \text{(never adopter)} \\ 0 & 0 & 0 & 0 & \dots & 1 & \text{(late adopter)} \\ 0 & 0 & 0 & 0 & \dots & 1 & \\ 0 & 0 & 1 & 1 & \dots & 1 & \\ 0 & 0 & 1 & 1 & \dots & 1 & \text{(medium adopter)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 1 & 1 & 1 & \dots & 1 & \text{(early adopter)} \end{pmatrix}$$

Let us focus for the moment on case with a single treated unit/time-period

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

Two related literatures on causal inference:

1. causal literature with unconfoundedness

2. synthetic control literature

1. Program Evaluation Literature under unconfoundedness (Rosenbaum-Rubin, 1984, Imbens-Rubin 2015)

$$W_{iT} \perp\!\!\!\perp \Big(Y_{iT}(0), Y_{iT}(1)\Big) \,\Big|\, Y_{i1}, \ldots, Y_{iT-1}, X_i$$

• In biomedical applications often no lagged outcomes, only features/characteristics. In social science applications often lagged outcomes, e.g., earnings in labor market experiments.

• lagged outcomes $Y_{i1}, \ldots, Y_{iT-1}$ are analyzed just like other features/covariates/characteristics X_i, by adjusting for them.

• Huge literature in causal inference, many estimators (matching, propensity score, weighting, including settings with high-dimensional X_i), widely used in practice.

Parametric version: **Horizontal** Regression (unit of observation is row in $\mathbf{Y}_{N\times T}$)

Specification of regression function (ignoring other covariates):

$$Y_{iT} = \beta_0 + \sum_{t=1}^{T-1} \beta_t Y_{it} + \varepsilon_i \quad i = 1, \ldots, N-1$$

estimated on $N-1$ control units, with T regressors.

Prediction for treated unit N:

$$\widehat{Y}_{NT} = \widehat{\beta}_0 + \sum_{t=1}^{T-1} \widehat{\beta}_t Y_{Nt}$$

• Nonparametric (matching) version: find a control unit j with $Y_{Nt} \approx Y_{jt}$, for all $t < T$. (match on rows of $\mathbf{Y}$)

• Also propensity score estimators, and doubly robust methods.

• If T (number of regressors in regression) large relative to $N - 1$ (number of observations) use regularized regression (lasso, ridge, elastic net)

But: most popular in cases with

– **Thin** Matrix $\mathbf{Y}$, N large, T small, multiple treated units, only treated in last period.

2. Synthetic Control Literature (Abadie-Diamond-Hainmueller, JASA 2010, Imbens-Doudchenko, 2016)

Abadie-Diamond-Hainmueller: Find a set of nonnegative weights α_i

$$Y_{N,t} \approx \sum_{i=1}^{N-1} \alpha_i Y_{i,t} \ \ \forall t, \quad \text{with restriction} \ \sum_{i=1}^{N-1} \alpha_i = 1$$

Construct a "synthetic" version of California by taking a convex combination of other states, e.g.,

$$\text{CA} = 0.3 \times \text{UT} + 0.7 \times \text{NY}$$

Imbens & Doudchenko: relax adding up and non-negativity restrictions.

This leads to estimating **Vertical** Regression (unit of observation is column in $\mathbf{Y}_{N\times T}$):

$$Y_{Nt} = \alpha_0 + \sum_{i=1}^{N-1} \alpha_i Y_{it} + \eta_t \quad t = 1, \ldots, T-1$$

estimated on $T-1$ pre-treatment periods, with N regressors. Prediction for treated period:

$$\widehat{Y}_{NT} = \widehat{\alpha}_0 + \sum_{i=1}^{N-1} \widehat{\alpha}_i Y_{iT}$$

• Nonparametric version: find a pre-treatment period $s < T$ with $Y_{is} \approx Y_{iT}$, for all $i = 1, \ldots, N-1$. (match on columns of $\mathbf{Y}$)

• If N large relative to $T-1$, use regularized regression (lasso, ridge, elastic net).

But: most popular in cases with

– **Fat** Matrix $\mathbf{Y}$, N small, T large, multiple treated periods, only one treated unit.

Compare: **unconfoundedness / horizontal regression:**

$$\widehat{Y}_{\mathsf{CA},1989} = \beta_0 + \widehat{\beta}_{1988} \times Y_{\mathsf{CA},1988} + \widehat{\beta}_{1987} \times Y_{\mathsf{CA},1987} + \dots$$

with $\widehat{\beta}$ choosen to ensure that for $i = 1, \dots, N-1$

$$Y_{i,1989} \approx \beta_0 + \beta_{1988} \times Y_{i,1988} + \beta_{1987} \times Y_{i,1987} + \dots$$

versus **synthetic controls / vertical regression:**

$$\widehat{Y}_{\mathsf{CA},1989} = \alpha_0 + \widehat{\alpha}_{\mathsf{UT}} \times Y_{\mathsf{UT},1989} + \widehat{\alpha}_{\mathsf{NY}} \times Y_{\mathsf{NY},1989} + \dots$$

with $\widehat{\alpha}$ choosen to ensure that for $t = 1, \dots, T-1$

$$Y_{\mathsf{CA},t} \approx \alpha_0 + \alpha_{\mathsf{UT}} \times Y_{\mathsf{UT},t} + \alpha_{\mathsf{NY}} \times Y_{\mathsf{NY},t} + \dots$$

In both approaches we end up with prediction that is a linear combination of lagged outcomes, and a linear combination of contemporaneous outcomes:

$$\widehat{Y}_{\text{CA},1989} = \beta_0 + \widehat{\beta}_{1988} \times Y_{\text{CA},1988} + \widehat{\beta}_{1987} \times Y_{\text{CA},1987} + \dots$$

$$= \alpha_0 + \widehat{\alpha}_{\text{UT}} \times Y_{\text{UT},1989} + \widehat{\alpha}_{\text{NY}} \times Y_{\text{NY},1989} + \dots$$

but the weights/coefficients are very different for two methods.

How do we choose between these methods?

- If $N >> T$ use unconfoundedness / horizontal regression.
- If $N << T$ use synthetic control / vertical regression.
- What if in between, $N \approx T$?

Focus on problem of imputing missing in $N \times T$ matrix $\mathbf{Y} = \mathbf{Y}(0)$

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

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

- Now the problem is a **Matrix Completion Problem**.

Differences-In-Differences Literature and Econometric Panel Data literature

Models developed here for complete-data matrices $\mathbf{Y}$:
DID: (Bertrand, Dulfo & Mullainathan, 2003)

$$Y_{it} = \delta_i + \gamma_t + \tau W_{it} + \varepsilon_{it}$$

Generalized Fixed Effects (Bai 2003, 2009, Bai & Ng 2002)

$$Y_{it} = \sum_{r=1}^{R} \delta_{ir}\gamma_{rt} + \tau W_{it} + \varepsilon_{it}$$

Fix R (or test hypotheses on R), and estimate δ and γ conditional on R by least squares and use $\widehat{\delta}$ and $\widehat{\gamma}$ to impute missing values.

Matrix completion literature (Candés & Recht, 2009; Candés & Plan (2010), Keshavan, Montanari & Oh, 2010)

– Focus on setting with many missing entries: $|\mathcal{O}|/|\mathcal{M}| \approx 0$. E.g., netflix problem, with units corresponding to individuals, and time periods corresponding to movie titles, or image recovery from limited information, with i and t corresponding to different dimensions.

– Focus on computational feasibility, with both N and T large.

– Focus on randomly missing entries: $\mathbf{W} \perp\!\!\!\perp \mathbf{Y}$, $W_{it} \perp\!\!\!\perp W_{js}$ for all i, j, t, s.

Like interactive fixed effect, focus on low-rank structure underlying observations.

Set Up

General Case: $N >> T$, $N << T$, or $N \approx T$, possibly stable patterns over time, possibly stable patterns within units.

Set up without covariates, connecting to the interactive fixed effect literature (Bai, 2003), and the matrix completion literature (Candés, and Recht, 2009):

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

- Assumption (may be able to relax a bit):

$$\mathbf{W}_{N\times T} \perp\!\!\!\perp \varepsilon_{N\times T}, \mathbf{L}_{N\times T}$$

- Possible dependence in $\mathbf{W}$, e.g., staggered entry, $W_{it+1} \geq W_{it}$

- $\mathbf{L}$ has low rank relative to N and T.

More general case, with unit-specific P-component covariate X_i, time-specific Q-component covariate Z_t, and unit-time-specific covariate V_{it}:

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

• We do not necessarily need the fixed effects γ_i and δ_t, these can be subsumed into $\mathbf{L}$. It is 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 standard Lasso-type element-wise ℓ_1 norm: defined as $\|\mathbf{H}\|_{1,e} = \sum_{p=1}^{P} \sum_{q=1}^{Q} |H_{pq}|$, could use ridge.

Choice for regularization for L is more important.

$$\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\mathbf{\Sigma}_{N\times T}\mathbf{R}_{T\times T}$$

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

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

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

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

Following Candés & Recht, 2009; Candés & Plan (2010) we regularize using using nuclear norm:

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

For the general case we estimate $\mathbf{H}$, $\mathbf{L}$, δ, γ, and β as

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

$$+\lambda_L \|\mathbf{L}\|_* + \lambda_H \|\mathbf{H}\|_{1,e}$$

We choose λ_L and λ_H through cross-validation.

Algorithm (Mazumder, Hastie, & Tibshirani 2010, direct minimization is computationally difficult).

Given any $N \times T$ matrix $\mathbf{A}$, define the two $N \times T$ matrices $\mathbf{P}_{\mathcal{O}}(\mathbf{A})$ and $\mathbf{P}_{\mathcal{O}}^{\perp}(\mathbf{A})$ with typical elements:

$$\mathbf{P}_{\mathcal{O}}(\mathbf{A})_{it} = \begin{cases} A_{it} & \text{if } (i,t) \in \mathcal{O}\,, \\ 0 & \text{if } (i,t) \notin \mathcal{O}\,, \end{cases}$$

and

$$\mathbf{P}_{\mathcal{O}}^{\perp}(\mathbf{A})_{it} = \begin{cases} 0 & \text{if } (i,t) \in \mathcal{O}\,, \\ A_{it} & \text{if } (i,t) \notin \mathcal{O}\,. \end{cases}$$

Let $\mathbf{A} = \mathbf{S}\mathbf{\Sigma}\mathbf{R}^\top$ be the Singular Value Decomposition for $\mathbf{A}$, with $\sigma_1(\mathbf{A}), \ldots, \sigma_{\min(N,T)}(\mathbf{A})$, denoting the singular values.

Then define the matrix shrinkage operator

$$\text{shrink}_\lambda(\mathbf{A}) = \mathbf{S}\tilde{\mathbf{\Sigma}}\mathbf{R}^\top,$$

where $\tilde{\mathbf{\Sigma}}$ is equal to $\mathbf{\Sigma}$ with the i-th singular value $\sigma_i(\mathbf{A})$ replaced by $\max(\sigma_i(\mathbf{A}) - \lambda, 0)$.

Given these definitions, the algorithm proceeds as follows.

– Start with the initial choice $\mathbf{L}_1(\lambda) = \mathbf{P}_{\mathcal{O}}(\mathbf{Y})$, with zeros for the missing values.

– Then for $k = 1, 2, \ldots,$ define,

$$\mathbf{L}_{k+1}(\lambda) = \text{shrink}_{\lambda}\Big\{\mathbf{P}_{\mathcal{O}}(\mathbf{Y}) + \mathbf{P}_{\mathcal{O}}^{\perp}(\mathbf{L}_k(\lambda))\Big\},$$

until the sequence $\{\mathbf{L}_k(\lambda)\}_{k\geq 1}$ converges.

– The limiting matrix $\mathbf{L}^*$ is our estimator for the regularization parameter λ, denoted by $\widehat{\mathbf{L}}(\lambda, \mathcal{O})$.

Illustrations

• We take complete matrices $\mathbf{Y}_{N\times T}$,

• We pretend some entries are missing.

• We use different estimators to impute the "missing" entries and compare them to actual values, and calculate the mean-squared-error (averaged over missing entries, and draws of which entries are missing).

We focus mainly on five estimators:

1. Matrix Completion Nuclear Norm, **MC-NNM**

2. Vertical Regr with Abadie-Diamond-Hainmueller restrictions (nonnegative weights, weights summing to one), **SC-V** (original Abadie-Diamond-Hainmueller estimator)

3. Vertical Regr with Elastic Net Regularization, **EN-V** (Synthetic Control type regression, Imbens-Doudchenko)

4. Horizontal Regr with Elastic Net Regularization, **EN-H** (close to common estimator in causal literature)

5. Horizontal Regr with ADH restrictions, **SC-H** (Program evaluation regr with restrictions – not used previously)

Questions:

1. How does matrix estimator compare to elastic net regression, either vertical or horizontal, as a function of the ratio N/T? Does the matrix completion estimator adapt well to the shape of the matrix?

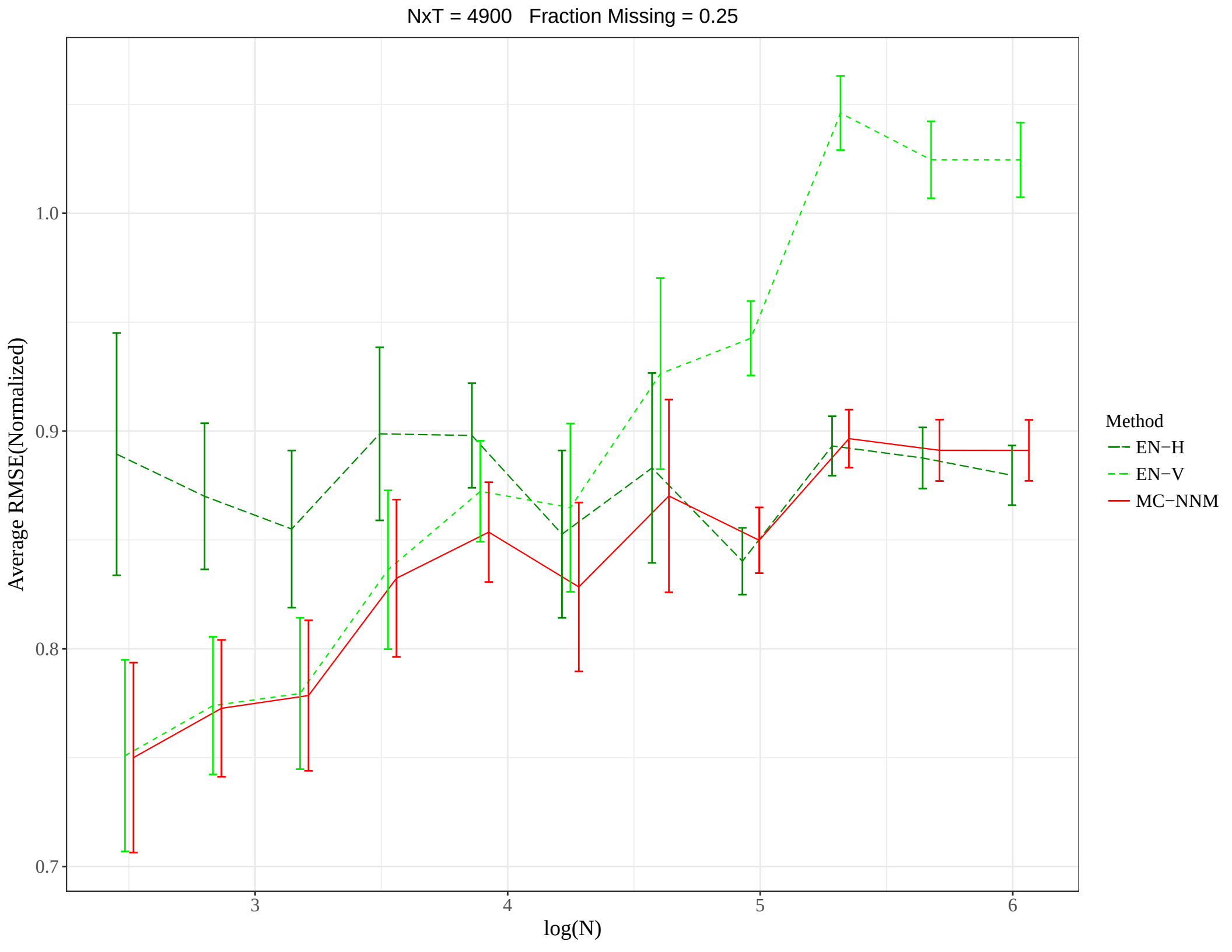
2. How does matrix estimator compare to vertical regressions (synthetic control regressions) as a function of the ratio N/T? Do the ADH restrictions improve the mean-squared-error in the synthetic control setting, at least when N is large?

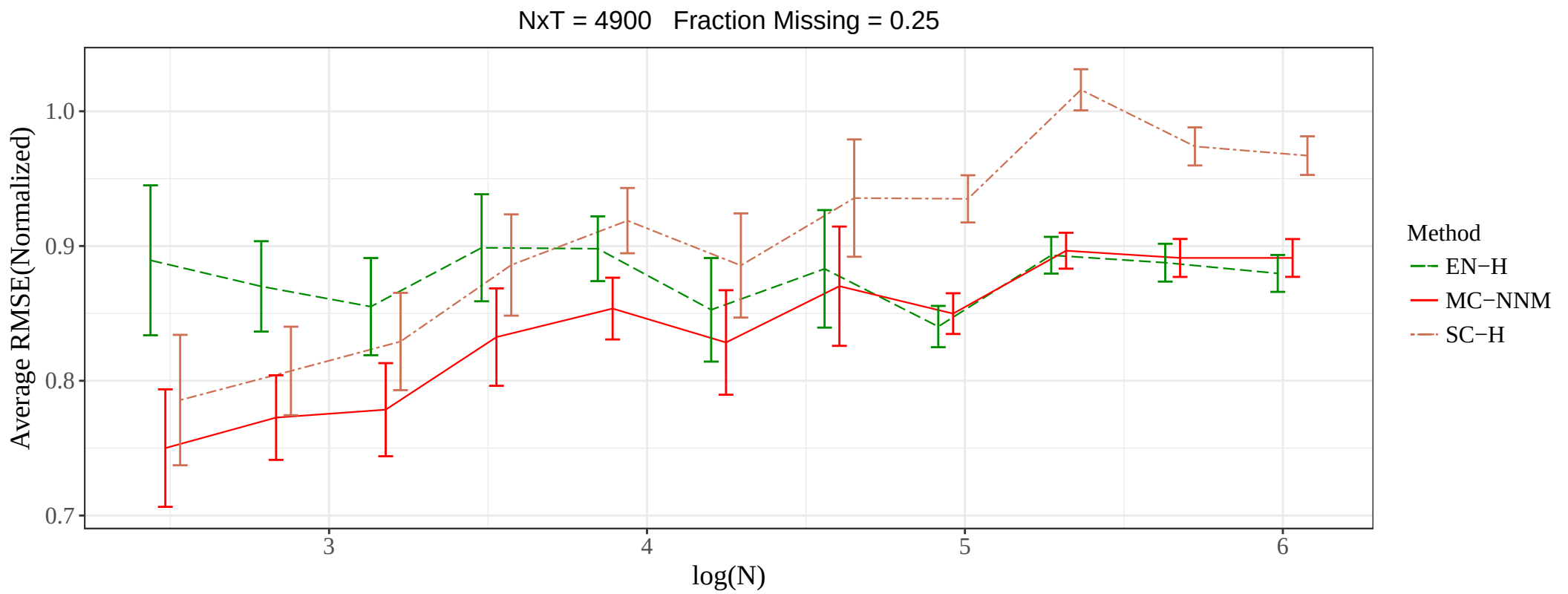
Illustration I: Stock Market Data

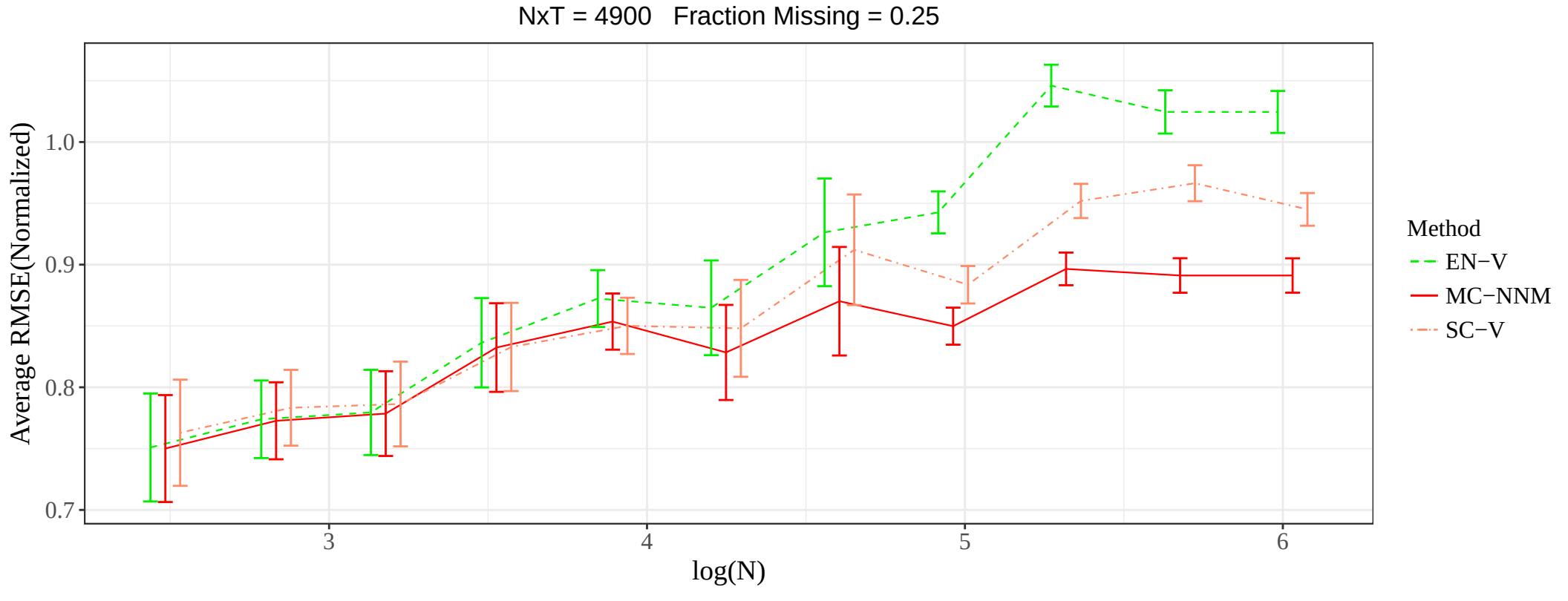
We use daily returns for 2453 stocks over 10 years (3082 days). We create sub-samples by looking at the first T daily returns of N randomly sampled stocks for pairs of (N, T) such that $N \times T = 4900$, ranging from fat to thin:
$(N, T) = (10, 490), \ldots, (70, 70), \ldots, (490, 10)$.

Given the sample, we pretend that half the stocks are treated at the mid point over time, so that 25% of the entries in the matrix are missing in a particular block.

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


NxT = 4900 Fraction Missing = 0.25
Average RMSE(Normalized)
1.0
0.9
0.8
0.7
3
4
5
6
log(N)
Method
EN−H
EN−V
MC−NNM


NxT = 4900 Fraction Missing = 0.25
Average RMSE(Normalized)
1.0
0.9
0.8
0.7
3
4
5
6
log(N)
Method
EN−H
MC−NNM
SC−H


NxT = 4900 Fraction Missing = 0.25
Average RMSE(Normalized)
1.0
0.9
0.8
0.7
3
4
5
6
log(N)
Method
EN−V
MC−NNM
SC−V

Results

• MC-NNM does better than EN-H and EN-V, adapts to shape of matrix

• ADH restrictions (non-negativity of weights, and summing to one, and no intercept) sometimes improve things relative to Elastic-Net estimator, more so for the vertical regressions than for the horizontal regressions.

Result I (Informal Version) Consider, $\mathbf{Y} = \mathbf{L} + \varepsilon$,

- ε_{it} are independent
- Adoption times t_i are indep r.v. (may depend on $\mathbf{L}$)

Theorem 1 *For optimal λ_L, with high probability,*

$$\sqrt{\frac{\|\mathbf{L} - \hat{\mathbf{L}}\|_F^2}{NT}} \leq \text{constant} \times \frac{\sqrt{N\text{rank}(\mathbf{L})\log^3(N)}}{\#\text{ofcontrolunits}}$$

Result I (Formal Version) Consider, $\mathbf{Y} = \mathbf{L} + \varepsilon$,

- ε_{it} are independent σ-sub-Gaussian
- Adoption times are indep r.v. (may depend on $\mathbf{L}$)
- $L_{\max} = \max_{it}(|L_{it}|)$ and $p_c =$ fraction of control units

Theorem 2 *There is a constant C such that with probability greater than $1 - 2(N+T)^{-2}$,*

$$\sqrt{\frac{\|\mathbf{L}^* - \widehat{\mathbf{L}}\|_F^2}{NT}} \leq C \max\left[L_{\max}\sqrt{\frac{\log(N+T)}{N\,p_c^2}}, \sigma\sqrt{\frac{R\log(N+T)}{T\,p_c^2}}, \sigma\sqrt{\frac{R\log^3(N+T)}{N\,p_c^2}}\right.$$

when λ_L is a constant times $\sigma \max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right]/|\mathcal{O}|$.

Results II: Adaptive Properties of Matrix Regression

Suppose N is large, $T = 2$, $W_{N2} = 1$, $W_{it} = 0$ for all other (i, t), many control units (treatment effect setting)

– In that case the natural imputation is $\widehat{L}_{i2} = Y_{i1} \times \rho$, where ρ is the within unit correlation between periods for the control units.

– The program-evaluation / horizontal-regression approach would lead to this solution, $\widehat{L}_{N2} = \overline{Y}_2 + (Y_{N1} - \overline{Y}_1) \times \rho\sigma_2/\sigma_1$

– The elastic net synthetic-control / vertical-regression approach would lead to $\widehat{L}_{i2} = 0$ as long as the regularization is sufficiently strong, and otherwise the results would not be stable.

– How does the matrix completion estimator impute the missing values in this case?

Suppose for control units

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and suppose the pairs (Y_{i1}, Y_{i2}) are jointly independent.

Then for small λ, large N, for the treated unit, the imputed value is

$$\hat{L}_{N2} = Y_{N1} \times \frac{\rho}{1 + \sqrt{(1 - \rho^2)}}$$

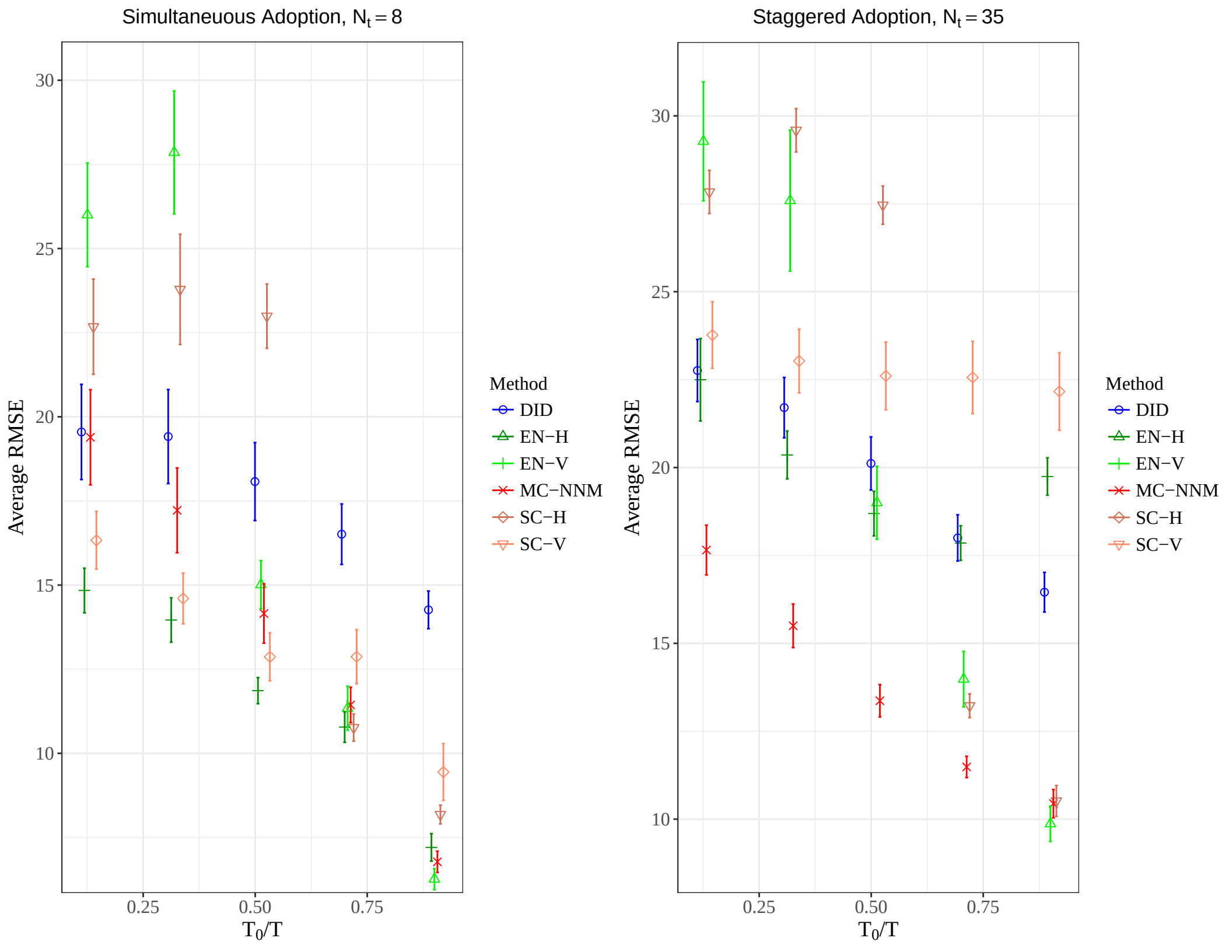
The regularization leads to a small amount of shrinkage towards zero relative to optimal imputation.

Matrix Completion method adapts well to shape of matrix and correlation structure.

Illustrations II: California Smoking Rate Data

The outcome here is per capita smoking rates by state, for 38 states, 31 years.

We compare both simultaneous adoption and staggered adoption.


Simultaneuous Adoption, $N_t = 8$
Staggered Adoption, $N_t = 35$
Average RMSE
T_0/T
30
25
20
15
10
0.25
0.50
0.75
Method
DID
EN−H
EN−V
MC−NNM
SC−H
SC−V

Results

• MC-NNM does substantially better than other estimators with staggered adoption.

• With simultaneous adoption EN-V does better than MC-NNM, if T_0/T is small (few data to impute missing values)

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:

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

and then give more weight to control (i,t) pairs that are "like" treated (i,t) pairs:

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

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.

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.

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.

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

Imbens, G., and D. Rubin *Causal Inference* Cambridge University Press.

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.

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.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani, (2010) "Spectral regularization algorithms for learning large incomplete matrices," *Journal of machine learning research*, 11(Aug):2287–2322.

Benjamin Recht, "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.