Inference for Low-Rank Models

Yuan Liao
Rutgers University

August 4, 2023
Pittsburgh Seminar

based on works from

- "Inference for low rank models", with Chernozhukov, Hansen and Zhu.
- "Inference for heterogeneous effects using low rank estimations", with Chernozhukov, Hansen and Zhu.
- "Inference for low-rank models without sample splitting with an application to treatment effect studies", with Kwon and Choi
- "Inference for low-rank models without knowing the rank", with Kwon and Choi

Econometric Model We Consider

$$y_{it} = x_{it}\theta_{it,1} + ... + x_{it}\theta_{it,d} + u_{it}, \quad i \leq N, \quad t \leq T.$$

- Each coefficient matrix is approximately low-rank.

$$\Theta_k = (\theta_{it,k}))_{N \times T}$$

- Goal: Inference about linear functionals of Θ_k

$$\frac{1}{N}\sum_{i=1}^{N}\theta_{it,k}g_i$$

Introduction

- Low-rank models refer to one (or multiple large) matrices, whose rank is much smaller than the dimension.
- The "low-rank" is an assumption, the key for dimension reduction.
- But in fact it holds (approximately) in many interesting econometric settings.
- Applications:
  - Statistics: missing data, "Netflix challenge" (classical)
  - Finance: factor models
  - Economics: treatment effect (most recently)

This talk

- Explain the setting of Spiked appproximate low rank model (SALR)
  - definition, assumptions ...
- Why the model covers a large number of Econometric settings
- This paper is about inference, e.g., the confidence interval
- Three key ingredients so far (hopefully relax them)
  - sample splitting / cross fitting (can be avoided)
  - rank is known or consistently estimable. (possibly avoidable)
  - strong signal/ beta-min/“spiked” (challenging, open question)
- As for the “spiked”:
  - Give a special case to relax it
  - Recently Armstrong, Weidner, Zeleneev have a paper, promising !

Examples I

- Example 1: Factor-model $y_{i,t} = \lambda_i' f_t + u_{it}$
- Example 2: Interactive fixed effect

$$y_{i,t} = x_{it}'\theta + \lambda_i' f_t + u_{it}$$

- Example 3: Heterogeneous coefficients

$$y_{i,t} = x_{it}\theta_{it} + \lambda_i' f_t + u_{it}$$

$$(\theta_{it})_{NT} \text{ is low-rank : } \theta_{it} = \alpha_i' g_t$$

  - Test about homogeneity: H_0: $\theta_{it} = \theta_i$

    simplified to tesing: $\mathrm{Var}(g_t) = 0$

  - Test about equal effect of two periods: H_0: $\theta_{i,t_1} = \theta_{i,t_2}$ for all i.

    simplified to tesing: $g_{t_1} = g_{t_2}$

Examples II

- Approaches to time-varying coefficients:

  1. slowly-varying $\theta_{i,t}$:

  $$\theta_{i,t} = \theta_i \left(\frac{t}{T} \right)$$

  2. additionally observe "characteristics" $w_{i,t}$:

  $$\theta_{i,t} = \beta' w_{i,t}$$

  3. Low-rank approach: a new approach

  $$\theta_{i,t} = \alpha_i' g_t$$

  can vary arbitrarily, and no-need additional observations

Examples III

- Example 4: Matrix completion: $Y_{i,t} = \theta_{i,t} + e_{i,t}$. But $Y_{i,t}$ is subject to missing.
  - Let $x_{i,t} = 1\{\text{observe}(i,t)\}$. Then
    $$y_{i,t} := Y_{i,t} x_{i,t} = \theta_{i,t} x_{i,t} + u_{i,t}$$
  - Netflix challenge: $\theta_{i,t} =$ Member t's expected review for movie i.
  - Can handle exogenous missing (but possibly non-random)
    $$\mathbb{E}(x_{i,t} e_{i,t}) = 0$$

Examples IV

- Example 5: Time-varying effects with latent variables

$$y_{i,t} = h_t(\eta_i) + u_{i,t}$$

  - Unknown function $h_t(\cdot)$, may arbitrarily vary over time.
  - unobserved latent variable η_i.
  - We will see in a minute that

$$\Theta := (h_t(\eta_i))_{N\times T} \text{ is approximately low-rank}$$

Examples V

- Example 6: Treatment effect studies:

$$\text{Treatment potential outcome: } Y_{i,t}(1) = h_{t,1}(\eta_i) + e_{i,t}(1)$$
$$\text{Control potential outcome: } Y_{i,t}(0) = h_{t,0}(\eta_i) + e_{i,t}(0)$$

The goal: ATE: $$\tau_t = \frac{1}{N}\sum_{i=1}^{N} h_{t,1}(\eta_i) - h_{t,0}(\eta_i)$$

Athey et al. (2022, *JASA*), Chernozhukov et al (2023, *Ann. Stats.*):

- Fix $m \in \{0, 1\}$, let $x_{i,t} = 1\{Y_{i,t}(m) \text{ is actually observed}\}$

$$y_{i,t} = Y_{i,t}(m)x_{i,t} = \theta_{i,t}(m)x_{i,t} + u_{i,t}, \quad \theta_{i,t}(m) := h_{t,m}(\eta_i)$$

- Hence can estimate $\theta_{i,t}(m)$ using the low-rank approach, separately for $m = 0, 1$.
- Then ATE is:

$$\tau_t = \frac{1}{N}\sum_i \theta_{i,t}(1) - \theta_{i,t}(0)$$

- As said, can handle exogenous treatment assignments:

$$\mathbb{E}x_{i,t}e_{i,t} = 0$$

Statistical Model: Spiked Approx Low-Rank (SALR)

$$\Theta = \Theta_0 + R. \quad N \times T$$

It is *SALR* if satisfies four conditions:

1. Low rank: $\text{rank}(\Theta_0) = J$ is either fixed or grows slowly
2. Spiked singular values:

$$\psi_1(\Theta_0) > ... > \psi_J(\Theta_0) \geq \psi_{NT} \to \infty$$

3. Incoherent singular vectors: SVD of $\Theta_0 = U_0 D_0 V_0'$,

$$U_0 = \begin{pmatrix} u_1' \\ \vdots \\ u_N' \end{pmatrix}_{N \times J}, \quad V_0 = \begin{pmatrix} v_1' \\ \vdots \\ v_T' \end{pmatrix}_{T \times J}$$

   We require

$$\max_{i \leq N} \|u_i\| = o_P(1), \quad \max_{t \leq T} \|v_t\| = o_P(1)$$

4. R is sufficiently small

Remarks about what we can/cannot achieve I

- Under these assumptions, no need to explicitly debias for inference
- If incoherent singular vectors is not satisfied, it is the sparse PCA setting.
  - eigenvectors are sparse, need debias
- If spiked singular values is not satisfied, it is the weak factors setting.
  - It all depends on "how weak" ψ_{np} is.
  - If $\psi_{np} \to \infty$ mildly fast, fine
  - If $\psi_{np} \to \infty$ very slowly, or does not grow: open question

Remarks about what we can/cannot achieve II

- if ψ_{np} is bounded, depends on the goal:
  - The entire matrix: rates derived in the stats. literature, (no inference)
  - $\frac{1}{N}\sum_i \theta_{i,t}$: we solved the univariate case, with inference.
  - A particular $\theta_{i,t}$: impossible. (Onatski 2003, weak factor)
  - Armstrong, Weidner, Zeleneev's recent work: promising
  - weak factor literature...

Remarks about what we can/cannot achieve III

So this talk focuses on SALR (strong factors, incoherent eigenvectors)

- But even in this setting, the inference problem is complicated.
- By far we do not have a beautiful unified approach.
- Approach separately:

$$\text{univariate} \begin{cases} \text{sample splitting: done} \\ \text{no sample splitting: done} \end{cases}$$

$$\text{multivariate} \begin{cases} \text{Neyman orthogonality+ sample splitting: done} \\ \text{other cases: ??} \end{cases}$$

Low-rank model is quite broad I

The well-known example is factor models (with strong factors).
Here we present another model:

$$y_{it} = h_t(\eta_i) + u_{i,t}.$$

- η_i is individual's unobserved state; $h_t(.)$ is time-varying function
- Let sieve-representation

$$\begin{aligned} h_t(\eta_i) &= \sum_{k=1}^{J} \lambda_{t,k}\phi_k(\eta_i) + r_{it} \\ &= \lambda_t'\Phi_i + r_{it} \\ \Theta &= \Phi\Lambda' + R \end{aligned}$$

- spiked eigenvalues

$$\psi_J(\Theta) \geq \sqrt{J^{-a}\sum_{i=1}^{N}\sum_{t=1}^{T} h_t(\eta_i)^2}$$

Low-rank model is quite broad II

- Eigen-gap:

$$\psi_k(\Theta) - \psi_{k+1}(\Theta) \geq cJ^{-b}$$

- Incoherent eigenvectors

$$\max_{t \leq T} \|v_t\| \leq L\psi_{\min}^{-1/2}(\Lambda'\Lambda)$$

$$\max_{i \leq N} \|u_i\| \leq C\psi_{\min}^{-1/2}(\Phi'\Phi)$$

- Sieve approximation error: suppose $h_t(.)$ belongs to a Holder class.

$$\max_{it} |r_{it}| = O_P(J^{-d}), \quad d = \text{smoothness}, \ \dim(\eta)$$

Low-rank model is quite broad III

- For proof: Reproducing Kernel Hilbert Space
  - Suppose $h_t \sim GassianProcess$, covariance kernel
    $$\mathrm{Cov}(h_t(\eta_1), h_t(\eta_2)) = K(\eta_1, \eta_2)$$
  - The Covariance Operator:
    $$\mathcal{T}(f)(\cdot) = \int K(\cdot, \eta) f(\eta) d\eta$$
    has eigenvalues and eigenfunctions $\{\lambda_i, \phi_i\}_{i=1}^{\infty}$.
  - Mercer's theorem:
    $$K(\eta_1, \eta_2) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\eta_1)\phi_i(\eta_2) \approx \sum_{i=1}^{J} \lambda_i \phi_i(\eta_1)\phi_i(\eta_2)$$
  - Then the matrix
    $$\frac{1}{T}\Theta\Theta' = \left[\frac{1}{T}\sum_t h_t(\eta_k)h_t(\eta_j)\right]_{N\times N} \approx \Phi_J \Lambda_J \Phi_J'$$
  - Hence $\{\lambda_i, \phi_i\}_{i=1}^{N}$ can characterize the eigenvalues / eigenvectors of Θ.

The Literature

- Reduced rank for dimension reduction.
  Anderson et al. (1949), Geweke (1996)

- (low-rank) matrix completion

  Negahban and Wainwright (2011); Recht et al. (2010); Sun and Zhang (2012); Candes and Tao (2010); Koltchinskii et al. (2011)....

  inference: Chen et al (2019), Xia and Yuan (2019)

- missing data for factor models and PCA

  Stock and Watson (2002), Cho et al (2015), Su et al (2019), Zhu et al (2019), Bai and Ng (2019),

  finance: Giglio et al (2021)

- ATE/Synthetic control using factors

  Athey et al (2022), and some recent works...

The Proposed Inference

We separately consider two cases:

1. Single low-rank Θ
2. Multiple low-rank $\Theta_1, ..., \Theta_d$.

- In the multivariate case, the effect of estimating Θ_j's affects each other.
- So the estimation steps are more complicated.
- Let us start with the univariate case.

Outline of Inference

$$Y = X \circ \Theta + U$$

SVD:

$$\Theta \approx UDV' = \Gamma V'$$

First step: Initially estimate by Nuclear-norm penalization

Second step: Take $\widetilde{V}$ as the eigenvector of the initial estimate.

Third step: Iterative least squares.

$$\widehat{\Gamma} = \arg\min_{\Gamma} \| Y - X \circ \Gamma \widetilde{V}' \|_F^2$$

$$\widehat{V} = \arg\min_{V} \| Y - X \circ \widehat{\Gamma} V' \|_F^2$$

$$\widehat{\Theta} = \widehat{\Gamma} \widehat{V}'$$

key technical arguments: (1) $\widehat{\Gamma}$ is "unbiased". (2) sample-splitting

First step: Nuclear-norm regularization

$$\widetilde{\Theta} = \arg\min_{\Theta} \|Y - X \circ \Theta\|_F^2 + \nu \|\Theta\|_{(n)}$$

- $\|\Theta\|_{(n)} = \sum$ all singular values.
- Gain insights from pure factor models, $X := 1$, then:
  - $$\widetilde{\Theta} = \text{soft-thresholding singular values of } Y$$

    Compared to PCA:

    $$PCA = \text{hard-thresholding singular values of } Y$$
  - $\widetilde{\Theta}$ is not ready for inference, due to shrinkage bias in soft-thresholding.
  - But, a nice insight: $\widetilde{\Theta}$ and PCA have the same eigenvectors.
- General X: the eigen-space of $\widetilde{\Theta}$ is unbiased.

Unbiased eigen-space of $\widetilde{\Theta}$ I

- To explain the intuition, consider a simpler product parameter model:

$$\theta = \gamma\beta$$

Identified as

$$\theta_{true} = \arg\min_{\theta} Q(\theta)$$

  - Suppose initial $\widetilde{\beta}$ is available (consistent, but possibly biased). Consider iteration:

$$\widehat{\gamma} = \arg\min_{\gamma} Q_n(\gamma\widetilde{\beta})$$

$$\widehat{\beta} = \arg\min_{\beta} Q_n(\widehat{\gamma}\beta)$$

$$\widehat{\theta} = \widehat{\gamma}\widehat{\beta}$$

  - Taylor expansion:

$$\widehat{\gamma} - \gamma = G^{-1}\beta\dot{Q}_n(\theta) + G^{-1}\partial^2_{\gamma,\beta}Q(\gamma\beta)(\widetilde{\beta} - \beta) + o(\|\widehat{\gamma} - \gamma\|)$$

  - The usual approach: orthognalize Q, so that

$$\partial^2_{\gamma\beta}Q(\gamma\beta) = 0.$$

Unbiased eigen-space of $\widetilde{\Theta}$ II

- Our approach:

$$\partial^2_{\gamma\beta} Q(\gamma\beta) = \partial^2_\theta Q(\theta)\beta\gamma + \underbrace{\partial_\theta Q(\theta)}_{score=0} = \partial^2_\theta Q(\theta)\beta\gamma$$

- Therefore,

$$\widehat{\gamma} - \gamma = G^{-1}\beta\dot{Q}_n(\theta) + \underbrace{G^{-1}\partial^2_\theta Q(\theta)(\widetilde{\beta} - \beta)\beta}_{A} \cdot\gamma + o(\|\widehat{\gamma} - \gamma\|)$$

- Let $H = I + A$, where $A \to^P 0$,

$$\widehat{\gamma} - H\gamma = G^{-1}\beta\dot{Q}_n(\theta) + o(\|\widehat{\gamma} - \gamma\|)$$

  So $\widehat{\gamma}$ is unbiased for a rotated γ.

- For product parameters $\theta = \beta\gamma$, up-to- rotation is sufficient.

$$\widehat{\beta} \approx H^{-1}\beta$$

- Together

$$\widehat{\beta}\widehat{\gamma} \approx \beta H^{-1} H\gamma = \beta\gamma$$

Unbiased eigen-space of $\widetilde{\Theta}$ III

- Back to our model

$$\Theta_0 = \Gamma_0 V_0'$$

  - Initialize $\widetilde{V}$.

$$\widehat{\gamma}_i = \arg\min_{\gamma} Q_i(\gamma, \widetilde{V}), \quad Q_i(\gamma, V) = \sum_{t=1}^{T} [y_{it} - x_{it}\gamma' \widetilde{v}_j]^2$$

  - Then we have

$$\partial^2_{\gamma, V} Q_i(\gamma_i, V)(\widetilde{V} - V) = A\gamma_i + \text{higher order}$$

- If "higher order is indeed higher order", then done.

  This requires "sample splitting" (either actual or auxiliary)

Sample splitting, either "actual" or "auxiliary"

Sample Splitting

For each fixed t,

$$\text{higher order } = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\widetilde{v}_j - v_j)u_{j,t}x_{j,t} = o_P(1)?$$

where $\widetilde{V} = (\widetilde{v}_1, ..., \widetilde{v}_N)$ is eigenvector from the initial $\widetilde{\Theta}$.

- If $u_{j,t}$ were independent of $\widetilde{V}$, easy. Hence sample splitting
- actual sample splitting: Chernozhukov et al (2023, *Ann. Stats.*)
  - When estimate $\widetilde{\Theta}$, use sample excluding "t".
  - For example, split the sample across $t = 1, ..., T$ into
    $$\{1...T\} = \mathcal{I} \cup \mathcal{I}^c \cup \{t\}$$
  - Initial estimate only using $\mathcal{I}$ or $\mathcal{I}^c$, leaving out t. Then average
- While pervasive in "double-ML inference", cross-fitting is not practically elegant.

Auxiliary Sample Splitting I

- Choi et al (2023) adopt "auxiliary sample splitting": directly proving

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\widetilde{v}_j - v_j)u_{j,t}x_{j,t} = o_P(1)$$

- Based on an elegant argument from Chen et al (2019 *PNAS*), but very technical.

- Do not do any sample splitting,
  1. hypothetically, if $\widetilde{v}_j$ were obtained by sample splitting, $\widetilde{v}_j^{-t}$
  2. show $\widetilde{v}_j \approx \widetilde{v}_j^{-t}$

- Potentially widely applicable to avoid cross-fitting in double ML under the panel setting.

Auxiliary Sample Splitting II

- We have

$$
\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_{jt} u_{jt} \left(\widetilde{v}_j - H' v_j\right) \\
&= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_{jt} u_{jt} \left(\widetilde{v}_j - \breve{v}_j^{(-t)}\right) + \frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_{jt} u_{jt} \left(\breve{v}_j^{(-t)} - H' v_j\right)
\end{aligned}
$$

  - where $\breve{v}_j^{(-t)}$ is the “hypothetical” leave t out estimator.
  - The main challenge is to show the first term is small.
  - $\breve{v}_j^{(-t)}$ is NOT trivially defined as $\tilde{v}_j$ but dropping t.

Auxiliary Sample Splitting III

- **Illustrating how leave-one-out defined here**
  - Consider estimating the mean

$$\min_{\beta} L^{full}(\beta), \quad L^{full}(\beta) := \sum_{s=1}^{T}(y_s-\beta)^2 = \sum_{s\neq t}(y_s-\beta)^2 + (y_t-\beta)^2$$

  - Leave-one-out defined as: replace $(y_t-\beta)^2$ by its expectation

$$L^{(-t)}(\beta) = \sum_{s\neq t}(y_s-\beta)^2 + \mathbb{E}(y_t-\beta)^2$$

Then:

* The solution $\min L^{(-t)}(\beta)$ is independent of y_t:

* The solution $\min L^{(-t)}(\beta)$ is close to the solution $\min_{\beta} L^{full}(\beta)$:

$$\breve{\beta}^{(-t)} = \frac{1}{T}[\sum_{s\neq t} y_s + \mathbb{E}y_t] = \bar{y} + O_P(T^{-1})$$

- Why defined in this way ?
  * Motivated by the EM algorithm: simply dropping t is less efficient.

- In the low-rank setting, much more sophisticated.
- Following Chen et al. (2019), consider following nonconvex problems:

Full sample:

$$L^{full}(\Gamma, V) = \frac{1}{2}\left\|X \circ (Y - \Gamma V')\right\|_F^2 + \frac{\lambda}{2}\|\Gamma\|_F^2 + \frac{\lambda}{2}\|V\|_F^2$$
$$= \frac{1}{2}\left\|X \circ (Y - \Gamma V')\right\|_{F,(-t)}^2 + \frac{1}{2}\left\|X \circ (Y - \Gamma V')\right\|_{F,t}^2 + \frac{\lambda}{2}\|\Gamma\|_F^2 + \frac{\lambda}{2}\|V\|_F^2 .$$

Leave-'t'-out:

$$L^{(-t)}(\Gamma, V) = \frac{1}{2}\left\|X \circ (Y - \Gamma V')\right\|_{F,(-t)}^2 + \frac{1}{2}\left\|\Theta^\star - \Gamma V'\right\|_{F,t}^2 + \frac{\lambda}{2}\|\Gamma\|_F^2 + \frac{\lambda}{2}\|V\|_F^2 .$$

- $||\cdot||_{F,(-t)}$ is computed ignoring the t-th column.
- $||\cdot||_{F,t}$ is only for the t-th column.
- $\Theta^\star = [\beta_i' F_t]_{N\times T}$ is the true low-rank matrix.

Auxiliary Sample Splitting V

- Showing $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_{jt} u_{jt} \left(\widetilde{v}_j - \breve{v}_j^{(-t)} \right) = o_p(1)$ proceeds by two steps:

  1. Both problems are iteratively solved after 's' steps via gradient descent :

  $$\breve{V}^{full} \Longleftarrow L^{full}(\Gamma, V), \quad \breve{V}^{(-t)} \Longleftarrow L^{(-t)}(\Gamma, V)$$

  2. Then, we show

     1) $\breve{V}^{full} \approx \widetilde{V}$

     2) $\breve{V}^{full} \approx \breve{V}^{(-t)}$

- About the stopping time "s":
  - s should be independent of obser at t because we want $\breve{V}^{(-t)}$ to be independent of observations at t.

Main Result

Theorem (Asymptotic normality of group average)

$$\mathcal{V}_{\mathcal{G}}^{-\frac{1}{2}} \left(\frac{1}{|\mathcal{G}|_o} \sum_{(i,t)\in\mathcal{G}} \widehat{\theta}_{it} - \frac{1}{|\mathcal{G}|_o} \sum_{(i,t)\in\mathcal{G}} \theta_{it} \right) \to^d \mathcal{N}(0,1),$$

where

$$\mathcal{V}_{\mathcal{G}} = \sigma^2 \left(\frac{1}{|\mathcal{T}|_o^2} \sum_{t\in\mathcal{T}} \bar{V}_{\mathcal{I}}' \left(\sum_{j=1}^{N} x_{jt} v_j v_j' \right)^{-1} \bar{V}_{\mathcal{I}} + \frac{1}{|\mathcal{I}|_o^2} \sum_{i\in\mathcal{I}} \bar{\Gamma}_{\mathcal{T}}' \left(\sum_{s=1}^{T} x_{is} \gamma_s \gamma_s' \right)^{-1} \bar{\Gamma}_{\mathcal{T}} \right)$$

- Rate of convergence:

$$\max \left\{ \frac{1}{\sqrt{N|\mathcal{T}|_o}}, \frac{1}{\sqrt{T|\mathcal{I}|_o}} \right\}$$

The Multivariate Case

Multivariate case I

$$Y = X_1 \circ \Theta_1 + ... + X_d \circ \Theta_d + U.$$

$$\Theta_j = \Gamma_j V_j, \quad j = 1, ..., d$$

- Iterative least squares.

$$(\widehat{\Gamma}_1, .., \widehat{\Gamma}_d) = \arg\min_{\Gamma} \| Y - (X_1 \circ \Gamma_1 \widetilde{V}_1' + ... + X_1 \circ \Gamma_d \widetilde{V}_d') \|_F^2$$

$$(\widehat{V}_1, ..., \widehat{V}_d) = \arg\min_{V} \| Y - (X_1 \circ \widehat{\Gamma}_1 V_1' + ... + X_d \circ \widehat{\Gamma}_d V_d') \|_F^2$$

- The effect of other $\widetilde{V}_j$:
    - on Γ_j: rotation argument, same
    - on Γ_k, $k \neq j$: challenging, new
    - Need orthogonality

- Suppose
$$X_k = \mu_k + E_k$$
Then the model is equivalent to
$$\widetilde{Y} = E_1 \circ \Theta_1 + ... + E_d \circ \Theta_d + U.$$
where
$$\widetilde{Y} = Y - \sum_k \mu_k \circ \Theta_k$$
- The moment condition is orthogonal wrt E, but not so wrt to X: suppose $d = 2$:
  - wrt X: not orthogonal:
  $$\partial_{V_2} \mathbb{E}(Y - X_1 \circ \Gamma_1 V_1' - X_2 \circ \Gamma_2 V_2') V_1 \circ X_1 = \mathbb{E}\Gamma_2 V_1 \circ X_1 \neq 0$$
  - wrt E: orthogonal:
  $$\partial_{V_2} \mathbb{E}(\widetilde{Y} - E_1 \circ \Gamma_1 V_1' - E_2 \circ \Gamma_2 V_2') V_1 \circ E_1 = \mathbb{E}\Gamma_2 V_1 \circ E_1 = 0$$

Multivariate case III

- Need to estimate $(\widetilde{Y}, E)$ first, using the initial $\widetilde{\Theta}$. (additional steps)
- We have not yet figured out the "auxiliary sample splitting" in the multivariate case, so adopt the actual sample splitting (additional steps)
- Therefore, the implementation of the multivariate case is not as elegant. (open question)

Extensions

Relaxing "spikedness"

- Consider $y_{i,t} = x_{i,t}\theta_{i,t} + u_{i,t}$
- If the goal is $\frac{1}{N}\sum_i \theta_{i,t}$, can use inverse probability weighting

$$\widehat{\tau} = \frac{1}{N}\sum_i \frac{y_{i,t}x_{i,t}}{\widehat{\mu}_i}, \quad \widehat{\mu}_i = \frac{1}{T}\sum_t x_{i,t}^2$$

- In the ATE, $\tau = \frac{1}{N}\sum_i \theta_{i,t}(1) - \frac{1}{N}\sum_i \theta_{i,t}(0)$, it leads to

$$\widehat{\tau} = \frac{1}{N(1)}\sum_{i: x_{i,t}(1)=1} \frac{y_{i,t}(1)}{\widehat{\mu}_i} - \frac{1}{N(0)}\sum_{i: x_{i,t}(0)=1} \frac{y_{i,t}(0)}{1-\widehat{\mu}_i}$$

- equivalent to propensity score weighting.
- Normality:

$$\sqrt{N}(\widehat{\tau} - \tau) \to^d \mathcal{N}(0, var)$$

- CLT arise from both $u_{i,t}$ and $x_{i,t}^2 - \mathbb{E}x_{i,t}^2$. So this approach requires random assignments.

Other extensions I

- New developments on Factor models provide other possibilities

$$y_{i,t} = \beta_i' f_t + u_{i,t}$$

- Example 1: Diversified Projection (Fan and Liao 2016, *JASA*), Pesaran (2003, *Ecma*), *Pesaran called it "CCE"*
  - Let w_i be a set of "weights" correlated with β_i, but independent of noises
  - e.g., $w_i = \phi(y_{i,0})$, initial period is independent of later.
  - Then

$$\widehat{f}_t = \frac{1}{N}\sum_i \phi(w_i) y_{i,t} = \frac{1}{N}\sum_i \phi(w_i)\beta_i f_t + \frac{1}{N}\sum_i \phi(w_i) u_{i,t}$$

$$\widehat{f}_t \to^P Hf_t$$

  - Advantage:
    1. No need to know the rank.
    2. More robust to weak factors than PCA.

Other extensions II

- In the low-rank framework,
  - The initial $\widetilde{\Theta}$
  - Introduce DP weighting matrices $W_F, \quad W_\beta$.

$$\tilde{\beta} = \frac{1}{T}\widetilde{\Theta} W_\beta, \quad \tilde{F} = \frac{1}{N}\widetilde{\Theta}' W_F$$

  - Then

$$\widehat{\Theta} = P_{\tilde{\beta}} \widetilde{\Theta} P_{\tilde{F}} - bias$$

  - Works as long as rank(W_β), rank(W_F) $\geq$ true rank (Choi et al 23).
  - Netflix Challenge: $W_\beta =$ customers' demographic characteristics.

    $W_F =$ films' genre. Film Studies Department at Yale reported over 40 film genres, styles, categories and series in their research catalog.

Other extensions III

- Example 2: Projected PCA (Fan et al 2016, *Ann. Stats*)

  Additionally observe individual-level (firm) characteristics z_i:

  $$\beta_i = z_i' \beta$$

  - Then
    $$y_{i,t} = \underbrace{z_i' \beta f_t}_{\widehat{y}_{i,t}} + u_{i,t}$$

    $\widehat{y}_{i,t}$ is "idiosyncratic-free".

  - Estimation steps:
    1. Cross-sectional regress $y_{i,t}$ onto $w_i \Rightarrow \widehat{y}_{i,t}$.
    2. PCA on $\widehat{y}_{i,t}$
  - Advantage: Fast rate of convergence of β_i; weaker factors

Optimality

- Semiparametric efficiency:

  under SALR conditions, the achieved variance is the semi. efficiency bound

- Minimaxity:

  WITHOUT SALR conditions,

  (1) It is impossible to consistently estimate $\theta_{i,t}$ for a fixed element
  (2) The minimax rate for estimating $\frac{1}{N}\sum_i \theta_{i,t}$ is $1/\sqrt{N}$

Simulations

Simulations I

- Matrix-completion setting, random missing
- Missing probability p_i: hetero or homo.
- Compare 4 methods:
  1. IPW (inverse prob weight)
     Replace missing $y_{i,t}$ by zero, and run PCA. (Pelger and Xiong 2019)
  2. UR: undebiased regularization
     Just nuclear-norm penalization
  3. EM
  4. Proposed
     UR + iterative LS

Simulations II

Table: MSE of estimated eigenvectors

N	T	IPW	UR	Proposed	EM
		Homogeneous missing			
100	200	0.176	0.116	0.109	0.109
200	100	0.252	0.171	0.161	0.161
		Heterogeneous missing			
100	200	0.263	0.211	0.119	0.119
200	100	0.369	0.304	0.204	0.203

- IPW is worst
- Proposed and EM very close, and favorably under hetero missing
- EM took much longer time.
- have done many more simulations in the ATE setting

Empirical Study

President allocation of the U.S. federal budget

$$y_{it} = \text{federal grants received by state } i \text{ year } t \text{ (detrended)}$$

- "treated": the state supported the president in the election (may receive higher grants?) (we also assume treatment is exogenous)
- treatment effects:

$$\Gamma_{it} := h_{t,1}(\eta_i) - h_{t,0}(\eta_i)$$

  - State effects: $\frac{1}{T}\sum_t \Gamma_{it}$
  - Loyal effects: $\frac{1}{\mathcal{S}}\sum_{i\in\mathcal{S}} \frac{1}{T}\sum_t \Gamma_{it}$

$$\mathcal{S} = \text{states who do not "swing"}$$

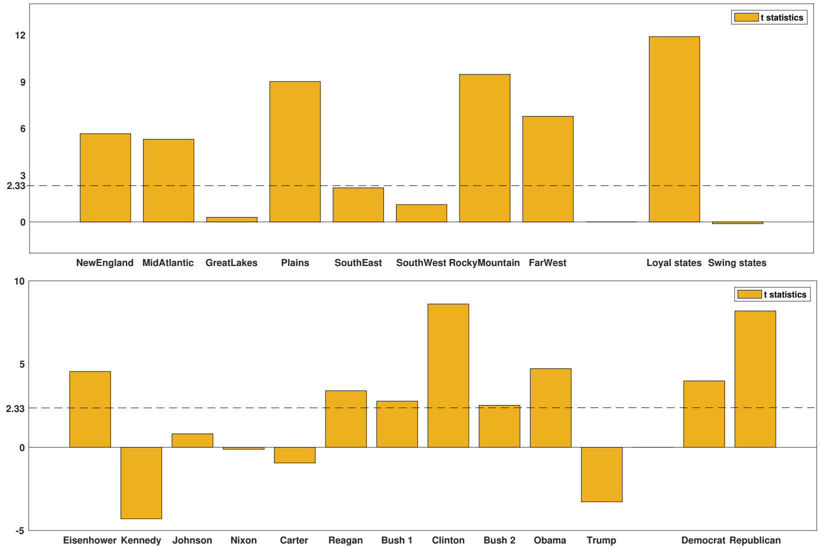
  - President effects: $\frac{1}{\mathcal{T}(j)}\sum_{t\in\mathcal{T}(j)} \frac{1}{N}\sum_i \Gamma_{it}$

$$\mathcal{T}(j) = \text{years when President } j \text{ was in office}$$

  - Party effects:

$$\frac{1}{\mathcal{J}}\sum_{j\in Party} \text{President effect}(j)$$

Figure: Region and Loyal effects/ President and Party effects



References

[Athey et al., 2021, Su et al., 2019, Choi et al., 2023b, Choi et al., 2023a, Chernozhukov et al., 2018, Chernozhukov et al., 2023, Chen et al., 2020, Chen et al., 2019, Xia and Yuan, 2021]

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021).
Matrix completion methods for causal panel data models.
Journal of the American Statistical Association, pages 1–15.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020).
Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization.
SIAM journal on optimization, 30(4):3098–3121.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019).
Inference and uncertainty quantification for noisy matrix completion.
Proceedings of the National Academy of Sciences, 116(46):22931–22937.

Chernozhukov, V., Hansen, C., Liao, Y., and Zhu, Y. (2018).
Inference for heterogeneous effects using low-rank estimation of factor slopes.
arXiv preprint arXiv:1812.08089.

Chernozhukov, V., Hansen, C., Liao, Y., and Zhu, Y. (2023).
Inference for low-rank models.
Annals of Statistics.

Choi, J., Kwon, H., and Liao, Y. (2023a).

Inference for low-rank completion without sample splitting with application to treatment effect estimation.
Working paper.

Choi, J., Kwon, H., and Liao, Y. (2023b).
Inference for low-rank models without estimating the rank.
Working paper.

Su, L., Miao, K., and Jin, S. (2019).
On factor models with random missing: Em estimation, inference, and cross validation.
Journal of Econometrics.

Xia, D. and Yuan, M. (2021).
Statistical inferences of linear forms for noisy matrix completion.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(1):58–77.