Nuclear Norm Regularized Estimation of Panel Regression Models

Hyungsik Roger Moon (USC & Yonsei)
Martin Weidner (UCL & CeMMaP)

ASSA, 2019

Introduction: Panel Regression Model with Factors

$$Y_{it} = \sum_{k=1}^{K} \beta_{0,k} X_{k,it} + \sum_{r=1}^{R_0} \lambda_{0,ir} f_{0,tr} + E_{it} , \quad i = 1 \dots N, \; t = 1 \dots T ,$$

where Y_{it} is the dependent variable, $X_{k,it}$ are regressors, $f_{0,tr}$ are factors, and $\lambda_{0,ir}$ are factor loadings.

- $\lambda_{0,ir}$ and $f_{0,tr}$ are unobserved and are treated as parameters (no distributional assumptions, interactive fixed effect model).
- R_0 is fixed but unknown.
- Object of interest: regression parameters β_0.
- Classic Example: *Holtz-Eakin, Newey & Rosen (1988)*.
  Study wage-dynamics using PSID data: Y_{it} is hours worked (log), $X_{k,it}$ is wage rate, lagged values of hours worked, f_t describes unobserved changes in working conditions, and λ_i unobserved earnings ability.
- Other applications: risk factors in asset pricing, controlling for global shocks in cross-country panels, etc.

Introduction: Model and Estimation Methods

$$Y_{it} = \sum_{k=1}^{K} \beta_{0,k}\, X_{k,it} + \sum_{r=1}^{R_0} \lambda_{0,ir}\, f_{0,tr} + E_{it}\,, \quad i = 1 \ldots N,\; t = 1 \ldots T\,,$$

- Quasi-Differencing ($R_0 = 1$): *Holtz-Eakin, Newey & Rosen (1988)*

  $$Y_{it} - \frac{f_{0,t}}{f_{0,t-1}}\, Y_{i,t-1} = \beta_0'\, X_{it} - \left(\frac{f_{0,t}}{f_{0,t-1}} \beta_0 \right)' X_{i,t-1} + \left(E_{it} - \frac{f_{0,t}}{f_{0,t-1}}\, E_{i,t-1} \right),$$

  then use appropriate IV (e.g. $Y_{i,t-2}$, etc.) to estimate this equation.
- Common Correlated Effects Estimator: *Pesaran (2006)*
  Use $\overline{Y}_t = N^{-1} \sum_i Y_{it}$ and $\overline{X}_{k,t} = N^{-1} \sum_i X_{k,it}$ as a proxys for $f_{0,t}$, estimate β including these proxys for $f_{0,t}$ in a linear regression.
- Least Squares Estimator:
  Kiefer (1980), *Bai (2009)*, *Moon & Weidner (2015, 2017)*
  Minimize the sum of squared residuals jointly over β, λ and f.
- Others: *Ahn, Lee & Schmidt (2001,2013)*, *Chamberlain & Moreira (2009)*, *Juodis & Sarafidis (2018)*, etc

Introduction: Least Squares Estimator

- Denote Y, X_k: $N \times T$ matrices,
  λ: $N \times R$,
  f: $T \times R$.
  Denote $\|A\|_2^2 = \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}^2$.
- Conventional way of writing the LS estimator:

$$\widehat{\beta}_{\mathrm{LS}} = \operatorname*{argmin}_{\beta} \min_{\lambda, f} \left\| Y - \underbrace{\beta \cdot X}_{:=\sum_k \beta_k X_k} - \lambda f' \right\|_2^2 .$$

- Equivalently this can be expressed as

$$\widehat{\beta}_{\mathrm{LS}} = \operatorname*{argmin}_{\beta} \min_{\Gamma} \left\| Y - \beta \cdot X - \Gamma \right\|_2^2 \quad \text{s.t.} \quad \mathrm{rank}(\Gamma) \leq R,$$

  where Γ is an $N \times T$ matrix, and the model in terms of Γ reads

$$Y_{it} = \sum_{k=1}^{K} \beta_k X_{k,it} + \Gamma_{it} + E_{it}$$

Introduction: Non-convexity of LS objective function

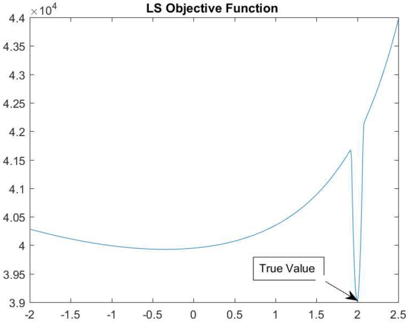
Example DGP:

$$y_{it} = \beta_0 x_{it} + \sum_{r=1}^{2} \lambda_{0,ir} f_{0,tr} + e_{it}, \qquad x_{it} = 0.04 e_{x,it} + \lambda_{0,i1} f_{0,t2} + \lambda_{x,i} f_{x,t},$$

where $\beta_0 = 2$, $\lambda_{0,i} = (\lambda_{0,i1}, \lambda_{0,i2})' \sim \text{i.i.d. } \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$,

$f_{0,t} = (f_{0,t1}, f_{0,t2})' \sim \text{i.i.d. } \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$,

$\lambda_{x,i} \sim \text{i.i.d. } 2\chi^2(1)$, $f_{x,t} \sim \text{i.i.d. } 2\chi^2(1)$, $e_{x,it}, e_{it} \sim \text{i.i.d. } \mathcal{N}(0,1)$, all mutually independent, choose $N = T = 200$. Plot $Q(\beta) = \min_{\lambda, f} \left\| Y - \beta \cdot X - \lambda f' \right\|_2^2$



Digression: Matrix Norms used in this paper

- For $N \times T$ matrix Γ, let $s_r(\Gamma)$ be the r^{th} largest singular value of Γ.

$$\|\Gamma\|_\infty := s_1(\Gamma) = \sup_{u:\|u\|=1} \sup_{v:\|v\|=1} u'\Gamma v.$$

$$\|\Gamma\|_2 := \left(\sum_r s_r^2(\Gamma)\right)^{1/2} = \mathrm{Tr}(\Gamma'\Gamma)^{1/2}.$$

$$\|\Gamma\|_1 := \sum_r s_r(\Gamma) = \sup_{\|A\|_\infty=1} \mathrm{Tr}(A'\Gamma).$$

- $\|\Gamma\|_1$ is called nuclear norm, trace norm, Schatten 1-norm, or Ky Fan n-norm.
- $\|\Gamma\|_\infty \leq \|\Gamma\|_2 \leq \|\Gamma\|_1 \leq \sqrt{\mathrm{rank}(\Gamma)}\,\|\Gamma\|_2 \leq \mathrm{rank}(\Gamma)\,\|\Gamma\|_1.$

Introduction: Nuclear norm regularization

- Constraint on unobserved error component Γ_{it}:

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

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

- Convex relaxation of this constraint:

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

Introduction: Nuclear norm penalization

- For some $\psi > 0$ we have

$$\widehat{\beta}_\psi = \underset{\beta}{\operatorname{argmin}} \, \min_{\Gamma} \left\| Y - \beta \cdot X - \Gamma \right\|_2^2 \quad \text{s.t.} \quad \|\Gamma\|_1 \leq \text{const.}$$

$$= \underset{\beta}{\operatorname{argmin}} \, \min_{\Gamma} \underbrace{\frac{1}{2NT} \left\| Y - \beta \cdot X - \Gamma \right\|_2^2 + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1}_{= Q_\psi(\beta, \Gamma)}$$

- Nuclear norm penalized estimation used in e.g.
  - Machine learning and statistical learning: e.g., *Fazel (2002)*, *Candes & Recht (2009)*, and for a recent survey see *Fazel & Parrilo (2010)*.
  - High dimensional low rank matrix estimation: e.g., *Rohde & Tsybakov (2011)*, *Negahbab & Wainwright (2011) Negahbab, Ravikumar, Wainwright & Yu (2012)*, *Athey, Bayati, Doudchenko, Imbens & Khosravi (2017)*, and many others.
  - Factor models without regressors: *Bai & Ng (2017)*

Introduction: Nuclear norm minimization

- Another estimator that we consider is

$$\widehat{\beta}_* = \underset{\beta}{\operatorname{argmin}} \left\| Y - \beta \cdot X \right\|_1 .$$

- One can show that

$$\widehat{\beta}_* = \lim_{\psi \to 0} \widehat{\beta}_\psi,$$

because $\lim_{\psi \to 0} \widehat{\Gamma}_\psi \to Y - \beta \cdot X$

Introduction: Contributions of this paper

- Study nuclear-norm regularizid estimator $\widehat{\beta}_\psi$ and its $\psi \to 0$ limit $\widehat{\beta}_*$.
- Show consistency of $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ as $N, T \to \infty$ and $\psi = \psi_{NT} \to 0$, under appropriate assumptions.
- Find that generically the convergence rate of $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ is at most $1/\sqrt{\min(N,T)}$, while the convergence rate of $\widehat{\beta}_{\mathrm{LS}}$ is $1/\min(N,T)$

  $\Rightarrow$ Therefore we suggest to use $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ as preliminary estimators (initial conditions), and obtain improved estimators that are asymptotically equivalent to $\widehat{\beta}_{\mathrm{LS}}$ in a finite number of simple LS iteration steps.
- Motivations to consider $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$:
  - Computational advantage of a convex objective function, in particular when $\dim \beta$ is large.
  - Identification of interactive fixed effect models when the true number of factors R is unknown, and there are low-rank regressors.

Introduction: Contributions of this paper (cont.)

- Post-nuclear-norm-regularized Estimation:
  - Use $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ as a preliminary consistent estimator.
  - Then iterate estimating β^0 and $\lambda^0 f^{0\prime}$.
  - After two iterations, we have an estimator that is asympotically equivalent to the LS estimator (QMLE).
- Extensions: Nonlinear single-index models of unbalanced panel. These include panel probit and quantile regressions.

  We show consistency of $\widehat{\beta}_\psi$: New in the literature.

Outline of the remaining talk

1. Motivation (convex relaxation / unique matrix separation)
2. Consistency and convergence rate results for $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$
3. Post-nuclear-norm regularized estimation
4. Monte Carlo Simulations
5. Extensions: Single Index Models with Unbalanced Panel

Two Main Motivations

Non-convex Least-Squares Objective Function

$$
\begin{aligned}
L_R(\beta) &= \min_{\lambda \in \mathbb{R}^{N \times R}} \min_{f \in \mathbb{R}^{T \times R}} \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \beta' X_{it} - \lambda_i' f_t \right)^2 \\
&= \frac{1}{2} \sum_{r=R+1}^{\min(N,T)} \left[s_r \left(\frac{Y - \beta \cdot X}{\sqrt{NT}} \right) \right]^2 \\
&= \sum_{r=1}^{\min(N,T)} \ell_\psi \left[s_r \left(\frac{Y - \beta \cdot X}{\sqrt{NT}} \right) \right],
\end{aligned}
$$

where

$$
\ell_\psi(s) := \begin{cases} \frac{1}{2}\, s^2, & \text{for } s < \psi, \\ 0, & \text{for } s \geq \psi, \end{cases}
$$

and

$$
s_{R+1} \left(\frac{Y - \beta \cdot X}{\sqrt{NT}} \right) < \psi(\beta, R) \leq s_R \left(\frac{Y - \beta \cdot X}{\sqrt{NT}} \right).
$$

(one-to-one relationship between $R \leftrightarrow \psi$)

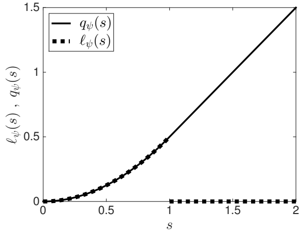
Motivation 1: Convex Relaxation

$$Q_\psi(\beta) = \min_{\Gamma \in \mathbb{R}^{N\times T}} \left[\frac{1}{2NT} \|Y - \beta \cdot X - \Gamma\|_2^2 + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1 \right]$$

$$= \sum_{r=1}^{\min(N,T)} q_\psi \left[s_r \left(\frac{Y - \beta \cdot X}{\sqrt{NT}} \right) \right],$$

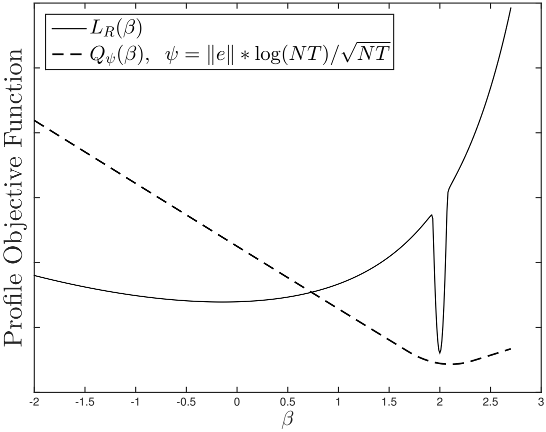
where

$$q_\psi(s) := \begin{cases} \frac{1}{2}\, s^2, & \text{for } s < \psi, \\ \psi s - \frac{\psi^2}{2}, & \text{for } s \geq \psi. \end{cases}$$



Plot of the functions $q_\psi(s)$ and $\ell_\psi(s)$ for $\psi = 1$.

Motivation 1: Convex Relaxation



Plot of $L_R(\beta)$ and $Q_\psi(\beta)$ for the example with $R = 2$ above and $\beta_0 = 2$.

Motivation 2: Unknown number of factors

(1) The LS estimator for β requires specifying the number of factors R.

(2) In order to estimate R one requires a preliminary consistent estimator for β — to apply e.g. *Bai & Ng (2002)*, *Onatski (2010)*, *Ahn & Horenstein (2013)* to $Y - \widehat{\beta} \cdot X$.

$\Rightarrow$ (1) and (2) can be circular (in particular for low-rank regressors).

$\Rightarrow$ Thus, $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ can be very useful here. In particular, $\widehat{\beta}_*$ requires neither to specify R nor to specify ψ.

Identification Problem for Low Rank X and R_0 Unknown.

- Estimation of treatment effects with interactive fixed effects is a widely applied "low-rank" regressor example: $X_{it} = v_i w_t$, where v_i is a binary treatment dummy and w_t is the time indicator of treatment. (e.g., *Kim & Oka (2014)*, *Gobillon & Magnac (2016)*, *Chan & Kwok (2016)*, *Powell (2017)*, *Gobillon & Wolff (2017)*, *Adams (2017)*, *Piracha, Tani, & Tchuente (2017)*, *Li (2018)*).
- Consider a simple case of rank 1 regressor:

$$Y = \beta_0 \underbrace{vw'}_{=X} + \lambda_0 f_0' + E,$$

  where $\mathrm{rank}(\lambda_0 f_0') = R_0$.
- Then, for any $\beta_\star$,

$$\beta_0 vw' + \lambda_0 f_0' = \beta_\star vw' + \lambda_0 f_0' + v(\beta - \beta_\star)w' = \beta_\star vw' + \lambda_\star f_\star',$$

  where $\lambda_\star = [\lambda_0, v]$, and $f_\star = [f_0, (\beta_0 - \beta_\star)w]$.
- Parameter values $(\beta_0, \lambda_0 f_0', R_0)$ and $(\beta_\star, \lambda_\star f_\star', R_\star)$, where $R_\star = R_0 + 1$, are observationally equivalent.

Motivation 2: Unique Matrix Separation Result

Question: How to estimate regression coefficients for low-rank regressors when R_0 is unknown?

We first want to answer this in a simplified setting, where the objective function is replaced by the expected objective function. Consider

$$\bar{\beta}_\psi := \underset{\beta}{\operatorname{argmin}} \min_{\Gamma} \left\{ \frac{1}{2NT} \mathbb{E}\left[\|Y - \beta \cdot X - \Gamma\|_2^2 \,\middle|\, X \right] + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1 \right\}.$$

Assumption

(i) $\mathbb{E}(E_{it}|X) = 0$ and $\mathbb{E}(E_{it}^2|X) < \infty$.

(ii) For all $\alpha \in \mathbb{R}^K \setminus \{0\}$,

$$\|\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\|_1 > \|\mathbf{P}_{\lambda_0}(\alpha \cdot X)\mathbf{P}_{f_0}\|_1.$$

Motivation 2: Unique Matrix Separation Result (cont.)

Proposition

$\|\bar{\beta}_\psi - \beta_0\| = O(\psi)$ as $\psi \to 0$.

- The proposition considers fixed N, T, with only $\psi \to 0$.
- The statement of the proposition implies that $\lim_{\psi \to 0} \bar{\beta}_\psi = \beta_0$.
- Thus, the proposition provides conditions under which the nuclear norm regularization approach identifies the true parameter β_0.
- For a single ($K = 1$) regressor with $X_{it} = v_i w_t$, the condition simply becomes $\|\mathbf{M}_{\lambda_0} v\| \|\mathbf{M}_{f_0} w\| > \|\mathbf{P}_{\lambda_0} v\| \|\mathbf{P}_{f_0} w\|$.
- It is possible to show that the weaker condition $\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0} \neq 0$ for any linear combination $\alpha \neq 0$ is sufficient for local identification of β in a sufficiently small neighborhood around β_0.

  However, that weaker condition is not sufficient for global identification of β_0.

Consistency and Convergence Rates

Consistency for only low-rank regressors

$$\widehat{\beta}_\psi = \operatorname*{argmin}_\beta \min_\Gamma \underbrace{\frac{1}{2NT}\left\| Y - \beta \cdot X - \Gamma \right\|_2^2 + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1}_{=Q_\psi(\beta,\Gamma)}$$

$$\widehat{\beta}_* = \lim_{\psi \to 0} \widehat{\beta}_\psi = \operatorname*{argmin}_\beta \|Y - \beta \cdot X\|_1$$

Assume $R_0 := \mathrm{rank}(\Gamma_0)$ is finite.

Theorem

Assume

$$\min_{\{\alpha \in \mathbb{R}^K \,:\, \|\alpha\|=1\}} \left\| \frac{\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}} \right\|_1 - \left\| \frac{\mathbf{P}_{\lambda_0}(\alpha \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}} \right\|_1 \geq c > 0,$$

and $\|E\|_\infty = O_P(\sqrt{\max(N,T)})$, *and* $\mathrm{rank}(X_k) = O_P(1)$. *Then,*

$$\left\| \widehat{\beta}_\psi - \beta_0 \right\| = O_P(\psi) + O_P\left(\frac{1}{\sqrt{\min(N,T)}} \right),$$

$$\left\| \widehat{\beta}_* - \beta_0 \right\| = O_P\left(\frac{1}{\sqrt{\min(N,T)}} \right).$$

Consistency for more general regressors (and for $\widehat{\Gamma}_\psi$)

- Want to show consistency of $(\widehat{\beta}_\psi, \widehat{\Gamma}_\psi) = \text{argmin}_{\beta,\Gamma}\, Q_\psi(\beta, \Gamma)$.
- Various equivalent ways to write the model:

$$
\begin{aligned}
y_{it} &= x_{it}'\beta_0 + \gamma_{0,it} + e_{it}, & \gamma_{0,it} &= \lambda_{0,i}' f_{0,t} \\
Y &= \sum_{k=1}^{K} X_k \beta_{0,k} + \Gamma_0 + E, & \Gamma_0 &= \lambda_0 f_0', \\
y &= x\beta_0 + \gamma_0 + e, & \gamma_0 &= (f_0 \otimes \lambda_0)\mathrm{vec}(\mathbf{I}_R),
\end{aligned}
$$

where y and γ_0 are NT-vectors, and x is an $NT \times K$ matrix.

Key Assumption: Restricted Strong Convexity

- Let $\mathbf{M}_A = \mathbb{I} - A(A'A)^{-1}A'$, $\quad \theta = \gamma - \gamma_0$, $\quad \Theta = \Gamma - \Gamma_0$.

Restricted Strong Convexity

Let $\mathbb{C} = \left\{ \Theta \in \mathbb{R}^{N \times T} \ : \ \|\mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 \leq 3 \|\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 \right\}$. Let there exists $\mu > 0$, independent from N and T, such that for any $\theta \in \mathbb{R}^{NT}$ with $\mathrm{mat}(\theta) \in \mathbb{C}$ we have $\theta' \mathbf{M}_x \theta \geq \mu \, \theta' \theta$, for all N, T.

- $\mathbb{C}$ is a cone of possible values for $\Theta = \Gamma - \Gamma_0$ that are close to $\lambda_0 f_0'$.
- Require that the quadratic term $\frac{1}{2NT}(\gamma - \gamma_0)' \mathbf{M}_x (\gamma - \gamma_0)$ of LS-objective function after profiling out β is bounded below by a strictly convex function, $\frac{\mu}{2NT}(\gamma - \gamma_0)'(\gamma - \gamma_0)$, if $\Gamma - \Gamma_0 \in \mathbb{C}$.
- corresponds to the restricted strong convexity condition in *Negahbab & Wainwright (2011)* and *Negahban, Ravikumar, Wainwright & Yu (2012)*, and it plays the same role as the restricted eigenvalue condition in recent LASSO literature.
- Can show that restricted strong convexity holds under low-level assumption on X_k, λ and f. (see below)

First show consistency of $\widehat{\Gamma}_\psi$

Bound on $\widehat{\Gamma}_\psi - \Gamma_0$

Let RSC hold, and assume that

$$\psi \geq \frac{2}{\sqrt{NT}} \|\mathrm{mat}(\mathbf{M}_x e)\|_\infty .$$

Then we have

$$\frac{1}{\sqrt{NT}} \left\| \widehat{\Gamma}_\psi - \Gamma_0 \right\|_2 \leq \frac{3\sqrt{2R_0}}{\mu} \, \psi .$$

- Proof analogous to arguments in machine learning literature.

Consistency of $\widehat{\Gamma}_\psi$ and $\widehat{\beta}_\psi$

Additional Regularity Conditions

(i) $\|E\|_\infty = \mathcal{O}_p\left(\max(N,T)^{1/2}\right)$,

(ii) $\frac{1}{\sqrt{NT}}\, e'x = \mathcal{O}_p(1)$,

(iii) $\frac{1}{NT}\, x'x \;\to_p\; \Sigma_x > 0$,

(iv) $\psi = \psi_{NT} \to 0$ such that $\sqrt{\min(N,T)}\,\psi_{NT} \to \infty$.

Theorem

Under RSC and above regularity conditions we have, as $N, T \to \infty$,

$$\frac{1}{\sqrt{NT}}\left\|\widehat{\Gamma}_\psi - \Gamma_0\right\|_2 \leq \mathcal{O}_p(\psi).$$

$$\left\|\widehat{\beta}_\psi - \beta_0\right\| \leq \mathcal{O}_p(\psi).$$

Regarding proof of (b), note that $\widehat{\beta}_\psi - \beta_0 = (x'x)^{-1}x'[e - (\widehat{\gamma}_\psi - \gamma_0)]$

Sufficient Conditions for Restricted Strong Convexity

- For $K = 1$ with $x'x = 1$ (normalized), the SRC condition is satisfied if

$$\liminf_{N,T} \min_{\theta \in \mathbb{C}} \|x - \theta\| \geq \mu > 0.$$

- A further set of sufficient conditions are as follows.
  - For simplicity consider $K = 1$ (one regressor X only)

Lemma

Let $s_1 \geq s_2 \geq s_3 \geq \ldots \geq 0$ be the singular values of the $N \times T$ matrix $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$. Assume that there exists a sequence q_{NT} such that

(i) $\frac{1}{\sqrt{NT}} \|X\|_2 = \mathcal{O}_p(1)$.

(ii) $\frac{1}{NT} \sum_{r=q_{NT}}^{\min(N,T)} s_r^2 \geq \mu > 0$ wpa1.

(iii) $\frac{1}{\sqrt{NT}} \sum_{r=1}^{q_{NT}-1} (s_r - s_{q_{NT}}) \to_P \infty$.

Then the above RSC assumption is satisfied.

Sufficient Conditions for Restricted Strong Convexity (cont.)

- This can be verified for explicit DGP's using random matrix theory.
  e.g.:
- $X_{it} \sim \mathrm{i.i.d.}\, \mathcal{N}(0, \sigma^2)$
- $X = \lambda_x f_x' + e_x$, where $e_{x,ij} \sim \mathrm{i.i.d.}\, \mathcal{N}(0, \sigma^2)$, and $\lambda_x f_x'$ describe a finite number of factors.

How to choose ψ?

- Choice of ψ essentially equivalent to choosing number of factors R.
- (Cross-validation?)
- For R the recommendation from *Moon & Weidner (2015)* is to choose larger R in case of doubt.
- Similarly, here the recommendation is to rather choose a smaller ψ, in particular since $\widehat{\beta}_* = \lim_{\psi \to 0} \widehat{\beta}_\psi$ also has good properties (albeit under stronger assumptions, and more difficult to prove).

Nuclear norm minimizing estimator: $\widehat{\beta}_\psi$

- Consider

$$\widehat{\beta}_* = \underset{\beta}{\operatorname{argmin}} \left\| Y - \beta \cdot X \right\|_1 .$$

$$= \underset{\beta}{\operatorname{argmin}} \sum_{r=1}^{\min(N,T)} s_r \left(Y - \beta \cdot X \right)$$

- Convex objective function, neither R nor ψ needs to be chosen.

Nuclear norm minimizing estimator: $\widehat{\beta}_\psi$

For simplicity consider again $K = 1$.

Theorem

As $N, T \to \infty$ with $N > T$, the following conditions are satisfied;

(i) $\|E\|_\infty = \mathcal{O}_p(\sqrt{N})$ *and* $\frac{1}{T\sqrt{N}}\|E\|_1 \leq \frac{1}{2} c_{\mathrm{up}}$, *wpa1.*

(ii) $\|X\|_\infty = \mathcal{O}_p(\sqrt{NT})$.

(iii) *Let $U_E S_E V_E'$ be the singular value decomposition of $\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0}$. We assume*

$$\mathrm{Tr}\left(X' U_E V_E'\right) = \mathcal{O}_p(\sqrt{NT}).$$

(iv) $T^{-1} N^{-1/2} \|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_1 \geq c_{\mathrm{low}} > 0$, *wpa*1.

(v) *Let $U_x S_x V_x' = \mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$ be the singular value decomposition of the matrix $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$. We assume that there exists $c_x \in (0, 1)$ such that wpa1*

$$\mathrm{Tr}\left(U_E' U_x S_x U_x' U_E\right) \leq (1 - c_x)\mathrm{Tr}(S_x).$$

Then $\sqrt{T}\left(\widehat{\beta}_* - \beta_0\right) = \mathcal{O}_p(1)$.

Nuclear norm minimizing estimator: $\widehat{\beta}_\psi$

- We consider a limit with $N > T$ here. Alternatively, we could consider a limit with $T < N$, but then we also need to replace N by T, and X by X' in the assumptions.
- Here, we not only need conditions on the singular values of e and X, but also assumptions involving the singular vectors. Much less results in random matrix theory on this.
- Condition (iv) rules out "low-rank regressors", for which we typically have $\|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_1 = \mathcal{O}_p(\sqrt{NT})$, but is satisfied generically for "high-rank regressors", for which $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$ has T singular values of order $\sqrt{N}$, so that $\|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_1$ is of order $T\sqrt{N}$.
- Example where all assumptions can be verified:

  $e_{it} \sim \text{i.i.d.}\,\mathcal{N}(0,\sigma^2)$,
  $X = \lambda_x f_x' + e_x$, where $e_{x,ij} \sim \text{i.i.d.}\,\mathcal{N}(0,\sigma^2)$,
  and $\lambda_x f_x'$ describe a finite number of factors.

Post-nuclear-norm regularized estimation

Post Nuclear Norm Regularized Estimation

Consider the case where R is known.

Updating procedure for β:

- For $s = 0$ set $\widehat{\beta}^{(s)} := \widehat{\beta}_{\psi}$ or $\widehat{\beta}_{*}$.

Step 1: We estimate the factor loadings and the factors of the $s-$step residuals $Y - \widehat{\beta}^{(s)} \cdot X$ by the principle component method:

$$(\widehat{\lambda}^{(s+1)}, \widehat{f}^{(s+1)}) := \operatorname*{argmin}_{\lambda \in \mathbb{R}^{N \times R}, f \in \mathbb{R}^{T \times R}} \left\| Y - \widehat{\beta}^{(s)} \cdot X - \lambda f' \right\|_2^2 .$$

Step 2: We update the s-stage estimator $\widehat{\beta}^{(s)}$ by

$$\widehat{\beta}^{(s+1)} := \operatorname*{argmin}_{\beta} \min_{g,h} \left\| Y - X \cdot \beta - \widehat{\lambda}^{(s+1)} g' - h \, \widehat{f}^{(s+1)\prime} \right\|_2^2$$

$$= \left(x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} \right) x \right)^{-1} x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} \right) y .$$

- Iterate steps 1,2 a finite number of times.

Post Nuclear Norm Regularized Estimation

- Define the local LS estimator obtained from optimizing the LS objective function with R factor $L_R(\beta)$ in a shrinking neighborhood around β_0

$$\widehat{\beta}_{\mathrm{LS},R}^{\mathrm{local}} := \underset{\{\beta \in \mathbb{R}^K \,:\, \|\beta - \beta_0\| \leq r_{NT}\}}{\operatorname{argmin}} L_R(\beta),$$

  where r_{NT} is a sequence of positive numbers such that $r_{NT} \to 0$ and $\sqrt{NT}\, r_{NT} \to \infty$.

- We consider $\widehat{\beta}_{\mathrm{LS},R}^{\mathrm{local}}$ instead of the original LS estimator $\widehat{\beta}_{\mathrm{LS},R}$, because we do not want impose the conditions needed for consistency of $\widehat{\beta}_{\mathrm{LS},R}$.

Post Nuclear Norm Regularized Estimation

Theorem

Assume that N and T grow to infinity at the same rate, and that

(i) $\operatorname{plim}_{N,T\to\infty}(\lambda_0'\lambda_0/N) > 0$, *and* $\operatorname{plim}_{N,T\to\infty}(f_0'f_0/T) > 0$.

(ii) $\|E\|_\infty = \mathcal{O}_p\left(\max(N,T)^{1/2}\right)$, *and* $\|X_k\|_\infty = \mathcal{O}_p\left((NT)^{1/2}\right)$.

(iii) $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} x' \left(\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}\right) x > 0$.

(iv) $\frac{1}{\sqrt{NT}} x' \left(\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}\right) e = \mathcal{O}_p(1)$.

Then,

$$\sqrt{NT}\left(\widehat{\beta}^{\text{local}}_{\text{LS},R_0} - \beta_0\right) = \mathcal{O}_p(1).$$

Assume furthermore that that $\|\widehat{\beta}^{(0)} - \beta_0\| = \mathcal{O}_p(c_{NT})$, *for a sequence* $c_{NT} > 0$ *such that* $c_{NT} \to 0$. *For* $s \in \{1,2,3,\ldots\}$ *we then have*

$$\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| = \mathcal{O}_p\left\{ c_{NT}\left(c_{NT} + \frac{1}{\sqrt{\min(N,T)}}\right)^s \right\}.$$

Post Nuclear Norm Regularized Estimation

Corollary

Let the assumptions of Theorem 4 hold, and assume that $c_{NT} = o((NT)^{-1/6})$. For $s \in \{2, 3, 4, \ldots\}$ we then have

$$\sqrt{NT}\left(\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right) = o_P(1), \qquad \sqrt{NT}\left(\widehat{\beta}^{(s)} - \beta_0\right) = \mathcal{O}_p(1).$$

Post Nuclear Norm Regularized Estimation

- EITHER: Apply well-known methods for "pure factor models" (without regressors) to the matrix $Y - \widehat{\beta}^{(0)} \cdot X$, e.g. *Bai & Ng (2002)*, *Onatski (2010)*, *Ahn & Horenstein (2013)*.

- OR: In the paper we consider:

$$\widehat{R}_{\psi^*} := \sum_{r=1}^{\min(N,T)} \mathbb{1}\left\{ s_r\left(\frac{Y - \widehat{\beta}^{(0)} \cdot X}{\sqrt{NT}} \right) \geq \psi^* \right\},$$

example

MC Simulation (very simple illustration)

- Consider the linear model with one regressor and two factors:

$$Y_{it} = \beta^0 X_{it} + \sum_{r=1}^{2} \lambda_{ir}^0 f_{ir}^0 + e_{it},$$

$$X_{it} = 1 + \tilde{X}_{it} + \sum_{r=1}^{2} (\lambda_{ir}^0 + \chi_{ir})(f_{tr}^0 + f_{t-1,r}^0),$$

  where $f_{tr}^0 \sim iidN(0,1)$ and $\lambda_{ir}^0, \chi_{ir} \sim iidN(1,1)$, and $\tilde{X}_{it}, e_{it} \sim iidN(0,1)$, and mutually independent.

- $(N, T) = (50, 50), (200, 200)$.
- $\psi_{NT} = (\log(N)^{1/2} \frac{\sqrt{\max(N,T)}}{NT}$

MC Simulation Result

(N,T)	POLS	$\widehat{\beta}_{LS}$	$\widehat{\beta}_{\psi}$	$\widehat{\beta}_{\psi}^{(1)}$	$\widehat{\beta}_{\psi}^{(2)}$	$\widehat{\beta}_{\psi}^{(3)}$
(50,50)						
bias	0.229	-0.007	0.135	0.014	-0.006	-0.007
s.d.	(0.017)	(0.011)	(0.015)	(0.011)	(0.011)	(0.011)
(200,200)						
bias	0.229	-0.0017	0.099	0.008	-0.0015	-0.0017
s.d.	(0.008)	(0.003)	(0.007)	(0.003)	(0.003)	(0.003)

Extensions to Some Nonlinear and/or Unbalanced Panel

- The model is a single index model.
- Let $m_{it}(z) := m(W_{it}, z)$ be a known convex function of the single index $z \in \mathbb{R}$, which also depends on the observed variables W_{it}. The single index is $X_{it}'\beta + \mathit{\Gamma}_{it}$.
- In the linear model, $W_{it} = Y_{it}$ and $m_{it}(z) = \frac{1}{2}(Y_{it} - z)^2$.
- The estimator is

$$\left(\widehat{\beta}_{\psi}, \widehat{\mathit{\Gamma}}_{\psi}\right) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^K,\, \mathit{\Gamma} \in \mathbb{R}^{N \times T}} Q_{\psi}(\beta, \mathit{\Gamma}),$$

$$Q_{\psi}(\beta, \mathit{\Gamma}) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} m_{it}\left(X_{it}'\beta + \mathit{\Gamma}_{it}\right) + \frac{\psi}{\sqrt{NT}} \left\| \mathit{\Gamma} \right\|_1 .$$

- We assume
  (i) W_{it} is independently distributed across i and over t, conditional on X.

Extensions to Some Nonlinear and/or Unbalanced Panel (cont.)

(ii) $m(w, z)$ is convex in z, once continuously differentiable in z almost everywhere in $\mathcal{W} \times \mathcal{Z}$. For any function $z_{it} = z_{it}(X) \in \mathcal{Z}$ the first derivative $\partial_z m_{it}(z_{it})$ exists almost surely, and satisfies $\max_{i,t,N,T} \mathbb{E}\left\{ [\partial_z m_{it}(z_{it})]^4 \,\middle|\, X \right\} < \infty$.

(iii) $\overline{m}_{it}(z)$ is twice continuously differentiable in $\mathcal{Z}$, with derivatives bounded uniformly over $i, t, N, T, \mathcal{Z}$. There exists $b > 0$ such that $\min_{i,t,N,T} \min_{z \in \mathcal{Z}} \partial_{z^2} \overline{m}_{it}(z) \geq b$.

(iv) $\partial_z \overline{m}_{it}(z^0_{it}) = 0$, for all i, t.

Examples

Let $z_{it}^0 = \beta_0' X_{it} + \Gamma_{0,it}$ be the true single index.

(a) Maximum likelihood: Let $p(y|z_{it}^0)$ is the conditional density function of Y_{it} on X.

- $W_{it} = Y_{it}$.
- $m_{it}(z) = -\log p(Y_{it}|z)$.
- Assume that $m_{it}(z)$ is strictly convex in z and three times continuously differentiable.
- A concrete example is a binary choice probit model, where $p(y|z) = \mathbb{1}(y = 1)\Phi(z) + \mathbb{1}(y = 0)[1 - \Phi(z)]$, and $\Phi(.)$ is the cdf of $\mathcal{N}(0, 1)$.

(b) Weighted Least Squares: Let $Y_{it} = z_{it}^0 + E_{it}$ with $\mathbb{E}(E_{it}|X_{it}, S_{it}) = 0$.

- $m_{it}(z) = \frac{1}{2} S_{it}(Y_{it} - z)^2$.
- $W_{it} = (Y_{it}, S_{it})$.
- $S_{it} \geq 0$ are observed weights for each observation. A special case is $S_{it} \in \{0, 1\}$, where S_{it} is an indicator of a missing outcome Y_{it}.

(c) Quantile Regression: Let $Y_{it} = z_{it}^0 + E_{it}$ with $\mathbb{E}[\mathbb{1}(E_{it} \leq 0)|X_{it}] = \tau$.

Examples (cont.)

- $m_{it}(z) = \rho_\tau(Y_{it} - z)$, where $\rho_\tau(u) = u \cdot [\tau - \mathbb{1}(u < 0)]$.
- $W_{it} = Y_{it}$.

Assumptions for Nonlinear Extensions

For simplicity, consider $K = 1$ (single regressor).

Assumptions

We assume the following.

(i) Assume $\psi \to 0$ as $\sqrt{NT}\psi \to \infty$.

(ii) Assume that $\|\Gamma_0\|_1 = O(\sqrt{NT})$.

(iii) The regressor X can be decomposed as $X = X^{(1)} + X^{(2)}$ such that $\|X^{(1)}\|_1 = o_P(\sqrt{NT}\,\psi^{-1/2})$, and $\|X^{(2)}\|_\infty = o_P(\sqrt{NT}\,\psi^{1/2})$.

(iv) $W := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}^{(2)})^2$ satisfies $W \to_P W_\infty > 0$.

Consistency

Theorem

Under the above assumptions,

$$\widehat{\beta}_\psi - \beta_0 = O_P(\psi^{1/2}).$$

Conclusion

- Nuclear norm penalized / minimized estimation of an interactive fixed effect regressions.
- Computational advantage: objective function is a convex function of the parameters.
- Identification: unique matrix separation through regularization.
- Extensions to single index models - probit, quantile, unbalanced panel.

Thank you