NUCLEAR NORM PENALIZED ESTIMATION OF INTERACTIVE FIXED EFFECT MODELS

HYUNGSIK ROGER MOON AND MARTIN WEIDNER

Incomplete and Work in Progress

1. Introduction

Interactive fixed effects panel regression models have been widely studied in recent panel literature. The interactive fixed effect model parsimoniously represents heterogeneity in both dimensions of the panel and includes the conventional additive error component model as a special case.

One of widely used estimation methods of the interactive fixed effects panel regression is the fixed effect approach (or principal component approach in a linear model) that treats the interactive fixed effects in a factor form as parameters to estimate [1]. For example, Bai (2009), Moon and Weidner (2015a), and Moon and Weidner (2015b)) investigated asymptotic properties of the fixed effect least squares estimator for linear panel regression models with interactive fixed effects, and Fernández-Val and Weidner (2016) studied the fixed effect maximum likelihood estimator for nonlinear panel regressions with interactive fixed effects.

The main advantage of the fixed effect approach is that it does not restrict the relationship between the unobserved heterogeneity and the observed explanatory variables. On the other hand, the computation of the fixed effects estimator requires solving a non-convex optimization problem with respect to high dimensional fixed effects parameters. Also, establishing consistency of the fixed effect estimator as an M-estimator is quite challenging due to the presence of the incidental parameters in both directions of panel. In this paper, we investigate a nuclear (trace) norm penalized estimator of interactive fixed effects models that is motivated by the two difficulties of the fixed effects estimator of interactive fixed effects panel regressions.

This paper is incomplete and very preliminary. We provide asymptotic analysis of the nuclear norm penalized estimator with a baseline model where the relationship between the dependent variable and the regressors are linear, the panel is balanced, and the rank of the interactive fixed effects is finite and known.

Date: January 28, 2016.

[1] Other estimation methods in the interactive fixed effects literature include the quasi-difference approach in Holtz-Eakin, Newey, and Rosen (1988) and the common correlated random effect method (e.g., Pesaran (2006)).

Extensions are in progress now. In the extensions, we consider more general models that allow heterogeneous regression coefficients, nonlinearity, and unbalanced panel. We also work on the estimators whose penalty threshold is small, which is closely related with unknown dimension of the interactive fixed effects.

Potential empirical applications we consider include estimation of the students' education performance model with teacher and student specific effects (generalized value added models), or modeling college admission decision as binary choice (with college and student specific effect).

2. General Model

Let $m(w_{it}, z_{it})$ be the objective function for $i = 1, \dots, N$ and $t = 1, \dots, T$, where w_{it} is observed, and $z_{it} = g_{it}(x_{it}) + \lambda_i' f_t$ is a scalar single index that depends on the observed covariates x_{it} and the unknown $f_{it}(.)$ and $\lambda \in \mathbb{R}^{N \times R}$ and $f \in \mathbb{R}^{T \times R}$. The function $m(.,.)$ is known. We define $m_{it}(z) = m(w_{it}, z)$, which is a random function of the single index. We require that $m_{it}(z)$ is *convex* in z, and we assume that the true parameters solve the following population FOC for all i, t

$$\mathbb{E}\left[\partial_z m_{it}(z_{it}^0) \,\middle|\, z_{it}^0\right] = 0, \qquad\qquad z_{it}^0 = g_{it}^0(x_{it}) + \lambda_i^{0\prime} f_t^0.$$

The set of functions $g_{it}(x_{it})$ can either be linearly parameterized in terms of parameters β such as a linear single index.

Examples for $m_{it}(z)$ include

(i) (Weighted) Least Squares Estimation with Interactive Effects:

$$m_{it}(z) = s_{it}(y_{it} - z)^2,$$

$w_{it} = (y_{it}, s_{it})$. The weights s_{it} could be "population weights", or could be binary variables that indicate missing data.

(ii) MLE:

$$m_{it}(z) = -\log p(y_{it}|z),$$

such that $\log p(y_{it}|z)$ is concave in z.

(iii) (Smoothed) Quantile Regression. Without smoothing:

$$m_{it}(z) = \rho_\tau(y_{it} - z),$$

where $\rho_\tau(u) = u \cdot (\tau - 1(u < 0))$.

Examples for $g_{it}(x_{it})$ are

(i) Homogeneous Coefficients: $g_{it}(x_{it}) = g(x_{it}) = \beta' x_{it}$.

(ii) Heterogeneous Coefficient: $g_{it}(x_{it}) = (\alpha_i + \gamma_t)' x_{it}$, where $\beta_{it} = \alpha_i + \gamma_t$ and $\beta = [\beta_{it}]_{it}$.

The fixed effects estimator studied in the existing literature is an M-estimator that solves the following minimize problem:

$$(\widehat{\beta}^*,\, \widehat{\lambda}^*,\, \widehat{f}^*) \in \underset{\beta\in\mathcal{B},\lambda\in\mathbb{R}^{N\times R},f\in\mathbb{R}^{T\times R}}{\operatorname{argmin}} Q_{NT}(\beta,\lambda f'),$$

$$Q_{NT}(\beta,\lambda f') = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m_{it}((\beta\odot X_{NT})_{it} + \lambda_i' f_t).$$

There are two difficulties with the M-estimator $(\widehat{\beta}^*,\, \widehat{\lambda}^*,\, \widehat{f}^*)$. One is a computational issue and the other one is a theoretical issue.

(i) Computational Issue: Calculating $\widehat{\beta}^*,\, \widehat{\lambda}^*,\, \widehat{f}^*$ is complicated by the fact that the objective function $Q_{NT}(\beta_{NT}, \lambda_{NT} f_{NT}')$ is non-convex in the parameters and can have multiple local minima. For fixed β to optimization over λ and f only becomes easy for non-weighted balanced least-squares (principal components), but for weighted least squares (including missing data), MLE and quantile regression becomes challenging. In any case, after profiling out λ and f the resulting profile objective is in general still non-convex in β, so that high-dimensionality of β also causes potentially serious computational problems.

(ii) Theoretical Issue: Consistency of $\widehat{\beta}^*,\, \widehat{\lambda}^*,\, \widehat{f}^*$ is only shown in some cases, but for many of the above examples the consistency problem is unresolved. For the penalized estimator that we consider, we can show consistency for a much larger class of models. The consistency then carries over to the improved estimator that we consider in the second step — that improved estimator may often be (asymptotically) equivalent to $\widehat{\beta}^*,\, \widehat{\lambda}^*,\, \widehat{f}^*$, but we do not show this, and it may not always be the case.

3. Penalization Estimation

In the general model in Section 2, notice that only the $N \times T$ matrix $\Gamma = \lambda f'$ enters into the objective function $Q_{NT}(\beta, \Gamma)$. When the function $m_{it}(z)$ is convex in z and z is linear in β and Γ, the objective function $Q_{NT}(\beta, \Gamma)$ is now a convex function of the parameters (β, Γ).

Assumption 1 (Convexity of $Q_{NT}(\beta, \Gamma)$)**.** *We assume that the objective function $Q_{NT}(\beta, \Gamma)$ is convex in β and Γ.*

Even thought the objective function $Q_{NT}(\beta, \Gamma)$ is convex in all parameters, the optimization problem with the rank restriction,

$$\min_{\beta}\min_{\Gamma} Q_{NT}(\beta,\Gamma) \quad \text{s.t.} \quad \operatorname{rank}(\Gamma) = (\text{or } \leq)\ \mathrm{R} \tag{3.1}$$

is non-convex because the constraint $\operatorname{rank}(\Gamma) = (\text{or } \leq)\ \mathrm{R}$ is non-convex. This implies that the objective function $Q_{NT}(\beta, \lambda f')$ is non-convex (in (β, λ, f).

Notice that the rank-constraint in (3.1) can be considered as an L_0- penalty on the singular values of Γ. When replacing the L_0- penalty with the convex L_1- penalty, then we obtain the convex objective function

$$Q_{NT}^{\psi}(\beta,\Gamma) = Q_{NT}(\beta,\Gamma) + 2\,\psi \sum_{r=1}^{\min(N,T)} s_r(\Gamma),$$

where $\psi = \psi_{NT} > 0$ is a penalty parameter that needs to be chosen and $s_r(A)$ denotes the r'th largest singular value of Γ.

The penalty function

$$\|\Gamma\|_* := \sum_{r=1}^{\min(N,T)} s_r(\Gamma)$$

is a matrix norm called called trace norm, nuclear norm, Schatten 1-norm, or Ky Fan n-norm in literature. Notice that the nuclear norm can equivalently be defined as $\|\Gamma\|_* = \mathrm{Tr}\left[(\Gamma\Gamma')^{1/2}\right]$, or $\|\Gamma\|_* = \sup_{\|B\|\leq 1} |\mathrm{Tr}(B\Gamma)|$. It satisfies $\|A+B\|_* \leq \|A\|_* + \|B\|_*$ (norm) and $\|AB\|_* \leq \|A\|_*\|B\|_*$ (submultiplicative). [2]

By definition, we have $\|c_1\Gamma_1 + (1-c_1)\Gamma_2\|_* \leq \|c_1\Gamma_1\|_* + \|(1-c_1)\Gamma_2\|_* = c_1\|\Gamma_1\|_* + (1-c_1)\|\Gamma_2\|_*$. This implies that the penalty function $\|\Gamma\|_* := \sum_{r=1}^{\min(N,T)} s_r(\Gamma)$ is convex in Γ.

The estimator we consider in this paper is the following nuclear norm penalized estimator:

$$(\widehat{\beta}^{\psi},\, \widehat{\Gamma}^{\psi}) = \underset{\beta\in\mathcal{B}_{NT},\, \Gamma\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} Q_{NT}^{\psi}(\beta,\Gamma).$$

4. Asymptotics of the Nuclear Norm Panelized Estimator in a Baseline Model

4.1. **Baseline Model.** Before we discuss the general model, we consider the following baseline model:

$$Y_{it} = \beta' X_{it} + \lambda_i' f_t + e_{it},$$

which we write in $N\times T$ matrix notation as $Y = \beta\cdot X + \lambda f' + e$. The baseline model assumes the following:

(i) The panel is balanced, that is, Y_{it} and X_{it} are observed for all pairs i,t.
(ii) The model is linear, given by (4.1) above.
(iii) The penalization parameter $\psi = \psi_{NT}$ is chosen such that as $N,T\to\infty$, the probability of rank $\left[\widehat{\Gamma}(b,\psi_{NT})\right] = R_0$ goes to one, for all b in an appropriate shrinking neighborhood of β. This means that asymptotically we estimate the correct number of factors. In practice, this essentially means that R_0 should be known.

[2]The last inequality can be strengthened to $\|AB\|_* \leq \|A\|\|B\|_*$.

Defining $\Gamma := \lambda f'$ the model can equivalently be written as

$$Y = \beta \cdot X + \Gamma + e, \qquad \operatorname{rank}(\Gamma) \leq R_0. \tag{4.1}$$

Here, β and Γ are the true unknown parameters, and we usually use b and G for generic parameter values. R_0 denotes the true number of factors.

Notation: Matrix norms: $\|.\|_2$ spectral norm, $\|.\|_{\mathrm{F}}$ Frobenius norm. $s_r(.)$ for r'th largest singular value of a matrix. $A^{\dagger}$ for Moore-Penrose pseudo-inverse of A.

Least Squares (or Fixed Effects) Estimator. For given $R \in \{0, 1, 2, \ldots\}$ the least squares estimator is given by

$$\left[\widehat{\beta}^{\mathrm{LS}}(R), \widehat{\Gamma}^{\mathrm{LS}}(R)\right] := \underset{\{b \in \mathbb{R}^K,\, G \in \mathbb{R}^{N \times T} \,:\, \operatorname{rank}(G) \leq R\}}{\operatorname{argmin}} \|Y - b \cdot X - G\|_{\mathrm{F}}^2 .$$

Alternatively, we can obtain $\widehat{\beta}^{\mathrm{LS}}(R)$ as

$$\widehat{\beta}^{\mathrm{LS}}(R) = \underset{b \in \mathbb{R}^K}{\operatorname{argmin}}\, Q_{NT}(b, R),$$

$$Q_{NT}(b, R) := \min_{\{G \in \mathbb{R}^{N \times T} \,:\, \operatorname{rank}(G) \leq R\}} \frac{1}{NT} \|Y - b \cdot X - G\|_{\mathrm{F}}^2 .$$

Here, $Q_{NT}(b, R)$ is the profile least squares objective function. We also define $\widehat{\Gamma}^{\mathrm{LS}}(b, R) := \operatorname{argmin}_{\{G \in \mathbb{R}^{N \times T} \,:\, \operatorname{rank}(G) \leq R\}} \|Y - b \cdot X - G\|_{\mathrm{F}}^2$.

Nuclear Norm Penalized Estimator. For given $\psi > 0$ we define

$$\left[\widehat{\beta}(\psi), \widehat{\Gamma}(\psi)\right] := \underset{b \in \mathbb{R}^K,\, G \in \mathbb{R}^{N \times T}}{\operatorname{argmin}} \|Y - b \cdot X - G\|_{\mathrm{F}}^2 + 2\,\psi\,\|G\|_*.$$

The additional factor 2 in the penalty term turns out to be convenient below. Alternatively, we can obtain $\widehat{\beta}(\psi)$ as

$$\widehat{\beta}(\psi) = \underset{b \in \mathbb{R}^K}{\operatorname{argmin}}\, S_{NT}(b, \psi),$$

$$S_{NT}(b, \psi) := \min_{G \in \mathbb{R}^{N \times T}} \frac{1}{NT} \left(\|Y - b \cdot X - G\|_{\mathrm{F}}^2 + 2\,\psi\,\|G\|_*\right).$$

Here, $S_{NT}(b, \psi)$ is the profile penalized objective function. We also define

$$\widehat{\Gamma}(b, \psi) := \underset{G \in \mathbb{R}^{N \times T}}{\operatorname{argmin}} \|Y - b \cdot X - G\|_{\mathrm{F}}^2 + 2\,\psi\,\|G\|_*$$

.

4.2. **Asymptotic Results for the Baseline Case.**

4.2.1. *Least Squares Estimator.* Here, we briefly summarize some known results on the least squares profile objective $Q_{NT}(b, R)$, and on the corresponding estimators $\widehat{\beta}^{\mathrm{LS}}(R)$ and $\widehat{\Gamma}^{\mathrm{LS}}(R)$.

Those results will be used afterwards to derive the asymptotic properties of the penalized estimator. For the profile least squares objective it is well-known that

$$Q_{NT}(b,R) := \frac{1}{NT}\sum_{r=R+1}^{\min(N,T)} \left[s_r\left(Y-\beta\cdot X\right)\right]^2.$$

This representation of the profile objective in terms of singular values, or eigenvalues,[3] is useful both for the numerical evaluation of $Q_{NT}(b,R)$, and for the derivation of the asymptotic properties of the least squares estimator.

Assumption 2. *As $N,T\to\infty$ we have*

(i) $s_{R_0}(\Gamma/\sqrt{NT}) \to_P c > 0$.

(ii) $\|e\|_2 = \mathcal{O}_P\left(\sqrt{\max(N,T)}\right)$.

(iii) $\|X_k\|_2 = \mathcal{O}_P\left(\sqrt{NT}\right)$, *for all* $k=1,\ldots,K$.

For $k,\ell\in\{1,\ldots,K\}$ we define

$$\begin{aligned}
W_{NT,k\ell} &:= \frac{1}{NT}\mathrm{Tr}(M_\lambda X_k M_f X_\ell'),\\
C^{(1)}_{NT,k} &:= \frac{1}{\sqrt{NT}}\,\mathrm{Tr}(M_\lambda\, X_k\, M_f\, e')\,,\\
C^{(2)}_{NT,k} &:= -\frac{1}{\sqrt{NT}}\left[\mathrm{Tr}\left(eM_f\, e'\, M_\lambda\, X_k\,\Gamma^\dagger\right) + \mathrm{Tr}\left(e' M_\lambda\, e\, M_f\, X_k'\,(\Gamma')^\dagger\right)\right.\\
&\qquad\qquad \left. + \mathrm{Tr}\left(e' M_\lambda\, X_k\, M_f\, e'(\Gamma')^\dagger\right)\right].
\end{aligned}$$

Let W_{NT} be the $K\times K$ matrix with elements $W_{NT,k\ell}$, and let $C^{(1)}_{NT}$ and $C^{(2)}_{NT}$ be the K-vectors with elements $C^{(1)}_{NT,k}$ and $C^{(2)}_{NT,k}$, respectively. Remember that $\Gamma=\lambda f'$. We therefore have $M_\lambda = M_\Gamma$ and $M_f = M_{\Gamma'}$. We prefer to write M_λ and M_f, because we find it slightly more transparent. Notice also that $\Gamma^\dagger = f(f'f)^{-1}(\lambda'\lambda)^{-1}\lambda'$, and $(\Gamma')^\dagger = (\Gamma^\dagger)'$.

The results in the following are known from (Moon and Weidner 2015b).[4]

Lemma 3. *Let Assumption 2 hold. Consider $N,T\to\infty$ with $N/T\to a>0$. Then,*

$$\begin{aligned}
Q_{NT}(b,R_0) = Q_{NT}(\beta,R_0) &- \tfrac{2}{\sqrt{NT}}\left(b-\beta\right)'\left(C^{(1)}_{NT}+C^{(2)}_{NT}\right)\\
&+ \left(b-\beta\right)' W_{NT}\left(b-\beta\right) + Q^{(\mathrm{rem})}_{NT}(b),
\end{aligned}$$

[3] $s_r\left(Y-\beta\cdot X\right)^2$ is the r'th largest eigenvalue of $(Y-\beta\cdot X)(Y-\beta\cdot X)'$.

[4] The result for $\widehat{\Gamma}^{\mathrm{LS}}(b,R_0)$ is not explicit in (Moon and Weidner 2015b), but can easily derived, because $\widehat{\Gamma}^{\mathrm{LS}}(b,R_0) = \Gamma - (b-\beta)\cdot X + e - \widehat{e}_{R_0}(b)$, and results for $\widehat{e}_{R_0}(b)$ are given.

and

$$\widehat{\Gamma}^{\mathrm{LS}}(b, R_0) = \Gamma - \sum_{k=1}^{K} (b_k - \beta_k) (X_k - M_\lambda X_k M_f) + e - M_\lambda e M_f + \widehat{\Gamma}^{(\mathrm{rem})}(b),$$

where the remainder terms $Q_{NT}^{(\mathrm{rem})}(b)$ *and* $\widehat{\Gamma}^{(\mathrm{rem})}(b)$ *satisfy for any sequence* $c_{NT} \to 0$,

$$\sup_{\{b:\|b-\beta\| \leq c_{NT}\}} \frac{\left|Q_{NT}^{(\mathrm{rem})}(b)\right|}{\left(1 + \sqrt{NT}\ \|b - \beta\|\right)^2} = o_P\left(\frac{1}{NT}\right),$$

$$\sup_{\{b:\|b-\beta\| \leq c_{NT}\}} \frac{\left\|\widehat{\Gamma}^{(\mathrm{rem})}(b)\right\|_2}{\left(1 + \sqrt{N}\, \|b - \beta\|\right)^2} = \mathcal{O}_P(1).$$

Notice that Assumption 2(i) is slightly weaker than the strong factor assumption in (Moon and Weidner 2015b), but sufficient to get all the results. We can give a more precise expansion of $\widehat{\Gamma}^{\mathrm{LS}}(b, R_0)$, but it is not required for the following.

If $\widehat{\beta}^{\mathrm{LS}}(R_0) \to_P \beta$ and $W_{NT} \to_P W > 0$ and $C_{NT}^{(1)} = \mathcal{O}_P(1)$, then using the expansion of $Q_{NT}(b, R_0)$ in Lemma 3 one can show that $\sqrt{NT}\left[\widehat{\beta}^{\mathrm{LS}}(R_0) - \beta\right] \to_P W_{NT}^{-1}\left(C_{NT}^{(1)} + C_{NT}^{(2)}\right)$. Using this one can derive the asymptotic distribution of $\widehat{\beta}^{\mathrm{LS}}(R_0)$.

4.2.2. *Nuclear Norm Penalized Estimator.* For $\psi > 0$ we define $g_\psi : \mathbb{R}_0^+ \to \mathbb{R}_0^+ : s \mapsto \min[\psi^2, s^2] + 2\psi \max[0, s - \psi]$. Thus, for $s \in [0, \psi]$ we have $g_\psi(s) = s^2$, and for $s > \psi$ we have $g_\psi(s) = 2\psi s - \psi^2$. Notice that $g_\psi(s)$ is continuous and convex. We then have

$$S_{NT}(b, \psi) := \frac{1}{NT} \sum_{r=1}^{\min(N,T)} g_\psi\left[s_r\left(Y - \beta \cdot X\right)\right]. \tag{4.2}$$

The derivation of equation (4.2) is given in the appendix.

Define $R_{NT}(b, \psi) = \sum_{r=1}^{\min(N,T)} \mathbb{I}\{s_r(Y - b \cdot X) > \psi\}$, the number of singular values of $Y - \beta \cdot X$ that is larger than ψ. Notice that $R_{NT}(b, \psi) = \mathrm{rank}[\widehat{\Gamma}(b, \psi)]$. Comparing those formulas for $S_{NT}(b, \psi)$ and $Q_{NT}(b, R)$, we find that for small singular values, namely for $s_r(Y - \beta \cdot X) \leq \psi$ and $r > R$, respectively, the contribution of the singular value $s_r(Y - \beta \cdot X)$ to both profile objective functions is given by the square of the singular value.

We therefore have[5]

$$S_{NT}(b,\psi) = Q_{NT}(b, R_{NT}(b,\psi)) + \frac{2\,\psi}{NT} \sum_{r=1}^{R_{NT}(b,\psi)} s_r\left(Y - \beta\cdot X\right) - \frac{R_{NT}(b,\psi)\,\psi^2}{NT}$$
$$= Q_{NT}(b, R_{NT}(b,\psi)) + \frac{2\,\psi}{NT} \left\| \widehat{\Gamma}^{\mathrm{LS}}(b, R_{NT}(b,\psi)) \right\|_{tr} - \frac{R_{NT}(b,\psi)\,\psi^2}{NT}.$$

In the last step we used that the non-zero singular values of $\widehat{\Gamma}^{\mathrm{LS}}(b,R)$ are equal to $s_r\left(Y-\beta\cdot X\right)$, $r=1,\ldots,R$, which is also well-known from principal components analysis.

Asymptotic expansions of $Q_{NT}(b,R_0)$ and $\Gamma^{\mathrm{LS}}(b,R_0)$ are given in Lemma 3. Using the expansion of $\Gamma^{\mathrm{LS}}(b,R_0)$ one can show that under the assumptions of Lemma 3, and also assuming that $\|P_\lambda e P_f\|_2 = \mathcal{O}_P(1)$, we have, uniformly in any $\sqrt{N}$-shrinking neighborhood of β, that

$$\left\| \widehat{\Gamma}^{\mathrm{LS}}(b,R_0) \right\|_* = \|\Gamma\|_* - \sqrt{NT}(b-\beta)' D_{NT} + \mathcal{O}_P(1),$$

where D_{NT} is the K-vector with components

$$D_{NT,k} := \frac{1}{\sqrt{NT}} \mathrm{Tr}\left[(\lambda'\lambda)^{-1/2}\lambda' X_k\, f(f'f)^{-1/2}\right].$$

Notice that $\mathrm{Tr}\left[(\lambda'\lambda)^{-1/2}\lambda' X_k\, f(f'f)^{-1/2}\right]$ can also be written as $\mathrm{Tr}\left[X_k B\right]$, where $B := f(f'f)^{-1/2}(\lambda'\lambda)^{-1/2}\lambda'$ satisfies $\|B\|_2 \leq 1$ and $\|\Gamma\|_* = \mathrm{Tr}(\Gamma B)$. This explains why $\|\Gamma+\Delta\|_* = \|\Gamma\|_* + \mathrm{Tr}(\Delta B) + \mathcal{O}(\|\Delta\|_2^2)$. See also (Watson 1992). From these, it is possible to deduce the following theorem:

Theorem 4. *Let Assumption 2 hold. Furthermore, assume that* $\|P_\lambda e P_f\|_2 = \mathcal{O}_P(1)$ *and* $\frac{1}{NT}\mathrm{Tr}(eX') = \mathcal{O}_P(N^{-1/2})$. *Consider* $N,T\to\infty$ *with* $N/T \to a > 0$. *Consider a sequence* $\psi_{NT} > 0$ *such that* $R_{NT}(\beta,\psi_{NT}) = R_0$, *wpa1. Then,*

$$S_{NT}(b,\psi_{NT}) = S_{NT}(\beta,\psi_{NT}) - \frac{2\,\psi_{NT}}{\sqrt{NT}}\left(b-\beta\right)' D_{NT} + \left(b-\beta\right)' W_{NT}\left(b-\beta\right) + S_{NT}^{(\mathrm{rem})}(b),$$

where the remainder term $S_{NT}^{(\mathrm{rem})}(b)$ *satisfies for any sequence* $c_{NT}\to 0$,

$$\sup_{\{b:\|b-\beta\|\leq c_{NT} \text{ and } R_{NT}(b,\psi_{NT})=R_0\}} \frac{\left|S_{NT}^{(\mathrm{rem})}(b)\right|}{\frac{1+\psi_{NT}\log N}{NT} + \frac{\sqrt{N}+\psi_{NT}}{\sqrt{NT}}\|b-\beta\| + \|b-\beta\|^2} = o_P(1).$$

From the expansion in Theorem 4, we want to conclude that

$$\widehat{\beta}(\psi_{NT}) - \beta = \frac{\psi_{NT}}{\sqrt{NT}} W_{NT}^{-1} D_{NT} + o_P\left(\frac{\psi_{NT}}{\sqrt{NT}}\right). \tag{4.3}$$

[5]One can also show that the nonzero singular values of $\widehat{\Gamma}(b,\psi)$ are equal to $s_r\left(Y-\beta\cdot X\right) - \psi$, for $r=1,\ldots,R_{NT}(b,\psi)$. We therefore have $S_{NT}(b,\psi) = Q_{NT}(b,R_{NT}(b,\psi)) + 2\,\psi\,\|\widehat{\Gamma}(b,\psi)\|_*/NT + R_{NT}(b,\psi)\,\psi^2/NT$.

However, the bound on the remainder term of the expansion is only applicable if $R_{NT}(b, \psi_{NT}) = R_0$. Therefore, we can only conclude (4.3) if $R_{NT}(b, \psi_{NT}) = R_0$ holds within a sufficiently large neighborhood of $b = \beta + \frac{\psi_{NT}}{\sqrt{NT}} W_{NT}^{-1} D_{NT}$. This can be achieved by choosing ψ_{NT} appropriately, which will now be discussed.

The strong factor Assumption 2(i) guarantees that for $\psi_{NT} = o_P(\sqrt{NT})$ we have, in any shrinking neighborhood of β, that $s_{R_0}(Y - b \cdot X) > \psi_{NT}$, implying that $R_{NT}(b, \psi_{NT}) \geq R_0$.

We furthermore have

$$\begin{aligned} s_{R_0+1}(Y - b \cdot X) &= s_1\left[\widehat{e}^{\mathrm{LS}}(b, R_0)\right] \\ &= s_1\left[M_\lambda (Y - b \cdot X) M_f\right] + \text{lower order terms} \\ &\leq \|e\| + \|M_\lambda[(b - \beta) \cdot X] M_f\| + \text{lower order terms}. \end{aligned}$$

The last bound is crude and needs to be improved if we want to consider $\psi_{NT} = \mathcal{O}_P(\sqrt{N})$. However, if $\psi_{NT}/\sqrt{N} \to \infty$, then $\|e\|$ in the last expression can be neglected for the question whether $s_{R_0+1}(Y - b \cdot X) < \psi_{NT}$. Thus, if $\psi_{NT} = o_P(\sqrt{NT})$ and $\psi_{NT}/\sqrt{N} \to \infty$ and $\|M_\lambda[(b_{NT} - \beta) \cdot X] M_f\| / \psi_{NT} < 1 - \epsilon$, wpa1, for some $\epsilon > 0$, then $R_{NT}(b_{NT}, \psi_{NT}) = R_0$, wpa1. We thus obtain the following.

Theorem 5. *Let Assumption 2 hold. Furthermore, assume that* $\|P_\lambda e P_f\|_2 = \mathcal{O}_P(1)$ *and* $\frac{1}{NT}\mathrm{Tr}(eX') = \mathcal{O}_P(N^{-1/2})$. *Consider* $N, T \to \infty$ *with* $N/T \to a > 0$. *Consider a sequence* $\psi_{NT} > 0$ *such that* $\psi_{NT} = o_P(\sqrt{NT})$ *and* $\psi_{NT}/\sqrt{N} \to \infty$. *Assume furthermore that* $W_{NT} \to_P W > 0$ *and that there exists* $\epsilon > 0$ *such that*

$$\left\| \sum_{k=1}^{K} \left(W^{-1} D_{NT}\right)_k \frac{M_\lambda X_k M_f}{\sqrt{NT}} \right\| < 1 - \epsilon, \qquad \textit{wpa1.} \tag{4.4}$$

Then we have

$$\frac{\sqrt{NT}}{\psi_{NT}} \left[\widehat{\beta}(\psi_{NT}) - \beta\right] \to_P W^{-1} D_{NT}.$$

4.3. **Bias Corrected Estimattor.** For the baseline case (linear model, balanced panel, known R_0) a good way to calculate an improved estimator is to start from $\widehat{\beta}^{(0)} = \widehat{\beta}(\psi_{NT})$ and then iterate the following:

(1) Given $\widehat{\beta}^{(j)}$, calculate $\widehat{\lambda}^{(j)}$ and $\widehat{f}^{(j)}$ as the R_0 principal components of $Y - \widehat{\beta}^{(j)} \cdot X$.
(2) Given $\widehat{\lambda}^{(j)}$ and $\widehat{f}^{(j)}$ calculate the $K \times K$ matrix $\widehat{W}_{k\ell}^{(j)} = \frac{1}{NT}\mathrm{Tr}(M_{\widehat{\lambda}^{(j)}} X_k M_{\widehat{f}^{(j)}} X_\ell')$ and the K-vector $\widehat{A}_k^{(j)} = \frac{1}{NT}\mathrm{Tr}(M_{\widehat{\lambda}^{(j)}} X_k M_{\widehat{f}^{(j)}} Y')$, and update the estimator for β as

$$\widehat{\beta}^{(j+1)} = \left(\widehat{W}^{(j)}\right)^{-1} \widehat{A}^{(j)}.$$

Starting with $\psi_{NT} = a\sqrt{N}\log N$, for some constant a, ander the above assumptions, as N, T grow at the same rate, this will give an estimator that is asymptotically equivalent to $\widehat{\beta}^{\mathrm{LS}}$ after very few iterations.

Alternatively, we could bias correct $\widehat{\beta}(\psi_{NT})$ directly, e.g. by modifying the objective function

$$S_{NT}^{\mathrm{BC}}(b, \psi_{NT}) = S_{NT}(b, \psi_{NT}) + \frac{2\,\psi_{NT}}{\sqrt{NT}}\,(b-\beta)'\,\widehat{D}_{NT},$$

where the estimator $\widehat{D}_{NT}$ uses the first stage estimator for λ and f. While this method seems overly complicated in the baseline case, it might be a very convenient method for non-linear models or unbalanced panels.

5. Extension to General Model

Extensions under progress include models that allow heterogeneous regression coefficients, some nonlinear regression model, unbalanced panel. We also work on the estimators whose penalty threshold is small, which is closely related with unknown dimension of the interactive fixed effects.

Appendix A. Derivation of Equation (4.2)

$$
\begin{aligned}
S_{NT}(b,\psi) &= \min_{\Gamma\in\mathbb{R}^{N\times T}} Q_{NT}(\beta,\Gamma) \\
&= \min_{\Gamma\in\mathbb{R}^{N\times T}} \frac{1}{NT}\sum_{r=1}^{\min(N,T)} \left\{\left[s_r\left(Y-\beta\cdot X-\Gamma\right)\right]^2+2\psi s_r(\Gamma)\right\} \\
&= \min_{\left\{\gamma\in\mathbb{R}^{\min(N,T)}\middle|\gamma\geq 0\right\}} \min_{\left\{U\in\mathbb{R}^{N\times\min(N,T)}\middle|U'U=\mathbb{I}\right\}} \min_{\left\{V\in\mathbb{R}^{T\times\min(N,T)}\middle|V'V=\mathbb{I}\right\}} \\
&\qquad\qquad \frac{1}{NT}\sum_{r=1}^{\min(N,T)} \left\{\left[s_r\left(Y-\beta\cdot X-U\mathrm{diag}(\gamma)V'\right)\right]^2+2\psi\gamma_r\right\} \\
&= \min_{\left\{\gamma\in\mathbb{R}^{\min(N,T)}\middle|\gamma\geq 0\right\}} \frac{1}{NT}\sum_{r=1}^{\min(N,T)} \left\{\left[s_r\left(Y-\beta\cdot X-U_{Y-\beta\cdot X}\mathrm{diag}(\gamma)V'_{Y-\beta\cdot X}\right)\right]^2+2\psi\gamma_r\right\} \\
&= \min_{\left\{\gamma\in\mathbb{R}^{\min(N,T)}\middle|\gamma\geq 0\right\}} \frac{1}{NT}\sum_{r=1}^{\min(N,T)} \left\{\left[s_r\left(Y-\beta\cdot X\right)-\gamma_r\right]^2+2\psi\gamma_r\right\} \\
&= \frac{1}{NT}\sum_{r=1}^{\min(N,T)} \min_{\{\alpha\in\mathbb{R}|\alpha\geq 0\}} \left\{\left[s_r\left(Y-\beta\cdot X\right)-\alpha\right]^2+2\psi\alpha\right\} \\
&= \frac{1}{NT}\sum_{r=1}^{\min(N,T)} \left\{\min\left[\psi^2, s_r\left(Y-\beta\cdot X\right)^2\right]+2\psi\max\left[0, s_r\left(Y-\beta\cdot X\right)-\psi\right]\right\}.
\end{aligned}
$$

Here, in the first step we rewrote the sum of squared residuals as the sum of squared singular values. In the second step we introduced the singular value decomposition $\Gamma = U\mathrm{diag}(\gamma)V'$, thus replacing the minimization over Γ by a minimization over the singular vector matrices U and V and the singular value vector γ. In the third step we use that for any given γ the minimizing U and V are equal to the corresponding singular vector matrices of $Y-\beta\cdot X$, denoted by $U_{Y-\beta\cdot X}$ and $V_{Y-\beta\cdot X}$ (This is the only step in the derivation that is not straightforward, we might want to provide a lemma in the appendix for this. Notice also that the minimizing singular vectors might not be unique, in particular they are not unique for $\gamma_r = 0$.). In the fourth step we use that the singular values of $Y-\beta\cdot X-U_{Y-\beta\cdot X}\mathrm{diag}(\gamma)V'_{Y-\beta\cdot X}$ are equal to $s_r\left(Y-\beta\cdot X\right)-\gamma_r$ (This assumes $s_r\left(Y-\beta\cdot X\right)-\gamma_r\geq 0$, which is always satisfied for the optimal γ.). The fifth step interchanges the sum and the minimization. The final step solves the minimization problem over α (the optimal α is $\alpha^* = \max[0, s_r\left(Y-\beta\cdot X\right)-\psi]$).

References

Bai, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77(4), 1229–1279.

Fernández-Val, I., and M. Weidner (2016): "Individual and time effects in nonlinear panel data models with large N, T," *forthcoming in Journal of Econometrics*.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56(6), 1371–95.

Moon, H., and M. Weidner (2015a): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," *Forthcoming in Econometric Theory*.

Moon, H. R., and M. Weidner (2015b): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83(4), 1543–1579.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

Watson, G. A. (1992): "Characterization of the subdifferential of some matrix norms," *Linear Algebra and its Applications*, 170, 33–45.

Department of Economics and USC Dornsife INET, University of Southern California, Los Angeles, CA 90089, U.S.A.

Department of Economics, University College London, Gower Street, London WC1E 6BT, U.K., and CeMMaP.