Factor models in microeconometrics: An Overview

Thierry Magnac Toulouse School of Economics

Panel Data Methods: Recent advances
Advanced EEE 1
Fall 2021

Outline

- Set up: Cross-section and serial dependence in panels
- Factor models: T fixed, N large
- Factor models: T and N large

References

- Ahn, Lee and Schmidt, 2001 & 2013
- Bai, 2009
- Pesaran, 2006
- Westerlund and Urbain, 2015
- Sul, 2019, *Panel Data Econometrics,* Routledge
- Chudik and Pesaran, 2015, "Large Panel Data Models with Cross-Sectional Dependence: A Survey" in Baltagi (Ed.), *The Oxford Handbook on Panel Data*, New York: Oxford University Press.

Serial and cross-section dependence

Write :

$$y_{it} = \mu + \alpha_i + \delta_t + u_{it},$$

in which u_{it}, δ_t and α_i are i.i.d. and independent between them, and:

$$E(\alpha_i) = E(\delta_t) = E(u_{it}) = 0.$$

Compute

$$\begin{aligned} E(\frac{1}{T-1}\sum_t (y_{it} - y_{i.})^2 &= \frac{1}{T-1} E(\sum_t (\delta_t - \delta_.)^2 + \sum_t (u_{it} - u_{i.})^2) \\ &= V(\delta_t) + V(u_{it}), \end{aligned}$$

and:

$$\begin{aligned} E(\frac{1}{T-1}\sum_t (y_{it} - y_{i.})(y_{jt} - y_{j.}) &= \frac{1}{T-1} E(\sum_t (\delta_t - \delta_.)^2) \\ &= V(\delta_t). \end{aligned}$$

Variance $V(\delta_t)$ is standing for (strong) cross-section dependence.
Remark: The same goes for serial dependence given by $V(\alpha_i)$.

Weak serial dependence

Replace the iid assumption for u_{it} which is a stationary process whose Wold decomposition is:

$$u_{it} = \sum_{t=0}^{+\infty} a_{it} u_{it}^{(0)}, Var(u_{it}) = \sum_{t=0}^{+\infty} a_{it}^2 = \sigma_i^2 < \infty$$

so that $Cov(u_{it}, u_{it'}) \longrightarrow 0$ with $t - t'$ at a fast rate.
Slitghtly less easy for weak cross-section dependence since there is no ordering.
Remark: Below, weak dependence is associated with weak factors while strong dependence is associated with strong factors.

Example: Bai's assumptions

Assumption C: serial and cross-sectional weak dependence and heteroskedasticity

1. $E(\varepsilon_{it}) = 0$, $E|\varepsilon_{it}|^8 \leq M$;
2. $E(\varepsilon_{it}\varepsilon_{js}) = \sigma_{ij,ts}$, $|\sigma_{ij,ts}| \leq \bar{\sigma}_{ij}$ for all (t, s) and $|\sigma_{ij,ts}| \leq \tau_{ts}$ for all (i, j) such that
$$\frac{1}{N}\sum_{i,j=1}^{N} \bar{\sigma}_{ij} \leq M, \ \frac{1}{T}\sum_{t,s=1}^{T} \tau_{ts} \leq M, \text{ and } \frac{1}{NT}\sum_{i,j,t,s=1} |\sigma_{ij,ts}| \leq M$$
The largest eigenvalue of $\Omega_i = E(\varepsilon_i\varepsilon_i')$ $(T \times T)$ is bounded uniformly in i and T.
3. For every (t, s), $E|N^{-1/2}\sum_{i=1}^{N}\left[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})\right]|^4 \leq M$.
4.
$$T^{-2}N^{-1}\sum_{t,s,u,v}\sum_{i,j} |\text{cov}(\varepsilon_{it}\varepsilon_{is}, \varepsilon_{ju}\varepsilon_{jv})| \leq M$$
$$T^{-1}N^{-2}\sum_{t,s}\sum_{i,j,k,\ell} |\text{cov}(\varepsilon_{it}\varepsilon_{jt}, \varepsilon_{ks}\varepsilon_{\ell s})| \leq M$$

Time varying fixed effects

Write now:

$$y_{it} = \mu + \alpha_i \delta_t + u_{it},$$

under the same assumptions.
Interpretation: individuals are diversely affected by the common shock and:

$$\begin{aligned} E(\frac{1}{T-1}\sum_t (y_{it} - y_{i.})^2) &= \alpha_i^2 V(\delta_t) + V(u_{it}), \\ E(\frac{1}{T-1}\sum_t (y_{it} - y_{i.})(y_{jt} - y_{j.}) &= \alpha_i \alpha_j V(\delta_t). \end{aligned}$$

General factor models: R factors

$$
\begin{aligned}
y_{it} &= \mu + \sum_{r=1}^{R} \lambda_i^{(r)} f_t^{(r)} + u_{it}, \\
&= \mu + \lambda_i' f_t + u_{it} = \mu + f_t' \lambda_i + u_{it},
\end{aligned}
$$

in which λ_i a.k.a. *factor loadings* and f_t are $[R, 1]$ vectors a.k.a. *factors.*
Stacking across time the equations:

$$Y_i = \mu e_T + F \lambda_i + U_i,$$

in which e_T is the constant vector and F is the $[T, R]$ matrix with row f_t'.
More generally :

$$\underset{[T,N]}{Y} = \mu J_{NT} + \underset{[T,R]}{F} \underset{[R,N]}{\Lambda} + U.$$

Reintroducing fixed effects and time dummies

Rewrite:

$$y_{it} = \mu + \sum_{r=1}^{R} \lambda_i^{(r)} f_t^{(r)} + u_{it},$$

and set $f_t^{(1)} = 1$, $\lambda_i^{(1)} = \alpha_i$ and $\lambda_i^{(2)} = 1$, $f_t^{(2)} = \delta_t$ so that:

$$y_{it} = \mu + \alpha_i + \delta_t + \sum_{r=3}^{R} \lambda_i^{(r)} f_t^{(r)} + u_{it}.$$

The first factor is known as well as the second factor loading. We can then always demean (recommended in linear models) by subtracting $y_{i.}$ and $y_{.t}$ and adding $y_{..}$ i.e.

$$dm(y_{it}) = y_{it} - y_{i.} - y_{.t} + y_{..},$$

and get:

$$dm(y_{it}) = \sum_{r=3}^{R} (\lambda_i^{(r)} - \lambda_.^{(r)})(f_t^{(r)} - f_.^{(r)}) + dm(u_{it}).$$

Fixed factors or factor loadings

Write:

$$\underset{[T,N]}{Y} = \mu J_{NT} + \underset{[T,R_0]}{F_0}\underset{[R_0,N]}{\Lambda_0} + \underset{[T,R_1]}{F_1}\underset{[R_1,N]}{\Lambda_1} + \underset{[T,R]}{F}\underset{[R,N]}{\Lambda} + U,$$

in which F_0 and Λ_1 are known (i.e. observed). The other factors and factor loadings are unobserved.

Consider M_{F_0} the projector on the orthogonal to F_0 i.e. $I_T - F_0(F_0'F_0)^{-1}F_0'$ of dimension T. Then :

$$\underset{[T,N]}{M_{F_0}Y} = \mu M_{F_0}J_{NT} + \underset{[T,R_1]}{M_{F_0}F_1}\underset{[R_1,N]}{\Lambda_1} + \underset{[T,R]}{M_{F_0}F}\underset{[R,N]}{\Lambda} + M_{F_0}U.$$

Similarly, let $M_{\Lambda_1'}$ the projector on the orthogonal to Λ_1 i.e. $I_N - \Lambda_1'(\Lambda_1\Lambda_1')^{-1}\Lambda_1$ of dimension N. Then:

$$\underset{[T,N]}{M_{F_0}YM_{\Lambda_1'}} = \mu M_{F_0}J_{NT}M_{\Lambda_1'} + \underset{[T,R]}{M_{F_0}F}\underset{[R,N]}{\Lambda M_{\Lambda_1'}} + M_{F_0}UM_{\Lambda_1'}.$$

This suggests two-step methods.

Observational equivalence

As known as the "rotation" problem ($\ntriangleright$! not rotations only).
The decomposition of the matrix $M = \underset{[T,R]}{F}\underset{[R,N]}{\Lambda}$ is not unique. Let Q an arbitrary invertible square matrix of dimension R, $QQ^{-1} = Q^{-1}Q = I_R$ then:

$$F\Lambda = FQQ^{-1}\Lambda = F_Q\Lambda_Q,$$

in which $F_Q = FQ$ and $\Lambda_Q = Q^{-1}\Lambda$ are observationally equivalent characterization of factors and factor loadings.

Normalizations:

$$\frac{F'F}{T} = I_R, \text{ (R*(R+1)/2 restrictions),}$$

$$\frac{\Lambda\Lambda'}{N} \text{ is diagonal, (R*(R-1)/2 restrictions.}$$

Normalizations: Interpretation

- All columns of $F = (f^{(1)}, ., f^{(R)})$ (R vectors of $\mathbb{R}^T$) are orthogonal between them and their norms are equal to T (i.e. the elements of $\frac{F'F}{T}$ are $\frac{1}{T}\sum_t f^{(r)\prime} f^{(p)}$ are equal to zero except if $p = r$.
- All columns of $\Lambda' = (\lambda^{(1)}, ., \lambda^{(R)})$ (R vectors of $\mathbb{R}^N$) are orthogonal between them.

Remark: Immaterial in terms of explanation or prediction. Only the subspaces generated by columns of F, in $\mathbb{R}^T$, or columns of Λ', in $\mathbb{R}^N$, matters.

Remark 2: The demeaning wrt time and individual means implies additional constraints.

- $f^{(r)}$ for all r, are orthogonal to a constant if are considered deviations from individual means i.e. $\sum_t f_t^{(r)} = 0$.
- $\lambda^{(r)}$ for all r, are orthogonal to a constant if are considered deviations from time means i.e. $\sum_i \lambda_i^{(r)} = 0$.

Introducing covariates

Replace μ by an index function $\mu(x_{it}) = x_{it}\beta$ in which x_{it} are covariates (among which a constant generally):

$$\begin{aligned} y_{it} &= x_{it}\beta + f_t'\lambda_i + u_{it}, \\ \underset{[T,1]}{Y_i} &= \underset{[T,K]}{X_i}\beta + \underset{[T,R]}{F}\underset{[R,1]}{\lambda_i} + U_i \text{ or } \underset{[1,N]}{Y_t} = \beta' \underset{[K,N]}{X_t} + f_t \underset{[R,N]}{\Lambda} + U_t. \end{aligned}$$

Arising is the issue of **correlated regressors** as in the standard fixed effect/time dummies framework.

- X_i and λ_i can be correlated, $E(X_i \otimes \lambda_i) \neq 0$
- X_t and f_t can be correlated, $E(X_t \otimes f_t) \neq 0$.

and OLS becomes inconsistent.

Strong and weak factors

Consider:

$$z_{it} = \sum_{l=1}^{r} f_{lt}\lambda_{il} + u_{it}$$

in which r is finite (there are extensions, see Chudik et al., 2011). A factor f_{lt} is **strong** when:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} |\lambda_{il}| = K > 0$$

while they are weak, semi-weak and semi-strong if respectively $\alpha = 0, \alpha \in (0, 1/2], \alpha \in (1/2, 1)$ in:

$$\lim_{N\to\infty} \frac{1}{N^{\alpha}} \sum_{i=1}^{N} |\lambda_{il}| = K < \infty$$

Other topics (not treated)

- Dynamic factors: an additional time series model for f_t e.g.

$$f_t = af_{t-1} + \varepsilon_{it}$$

- Weak exogeneity of regressors

Ahn, Lee and Schmidt, 2001 and 2013

Models

- Ahn, Lee and Schmidt, 2001: a single interaction between individual and time effects (Holtz-Eakin, Newey and Rosen, 1989)

  $$y_{it} = x_{it}\beta + z_i\gamma + \lambda_i f_t + u_{it},$$

  in which λ_i and f_t are scalars.
  Remark: Parameter γ is identified if f_t is not constant over time. Time varying γ_t would not be identified. We drop z_i in the following, by acknowledging that it is kept "hidden" in λ_i (as with standard fixed effects).
- Ahn, Lee and Schmidt, 2013: more than one factor model

  $$Y_i = X_i\beta + F\lambda_i + U_i,$$

  in which $\theta = (F, \beta)$ are parameters and λ_i are random variables

Suppose, in the next slides, that we know the true number of factors $R = R_0$.

Assumptions

1. ε_i is iid over i (no residual cross section dependence) and $E(\varepsilon_{it}) = 0$
2. $(X_i, \lambda_i, \varepsilon_i)$ have finite moments up to order 4 (usual CLTs)
3. $E((X_i, \lambda_i)'(X_i, \lambda_i))$ has full rank (i.e. if λ_i are known, β and F are identified)
4. $E(\varepsilon_i \mid X_i) = 0$ (could be weakened into non correlation, see ALS-01)
5. $Rank(E(X_i \lambda_i')) = R < T$ (only correlated effects matter, see **Discussion**)
6. $Rank(E(M_F X_i \otimes X_i)) \geq K$. (identification condition of β, see below)

An assumption on the correlation between λ_i and ε_i seems to be missing (See **Discussion**)

An Example

ALS-2001: A single x_{it}.

$$y_{it} = x_{it}\beta + \lambda_i f_t + u_{it},$$

in which the moments we use are $E(x_{is}u_{it}) = 0$.
Project λ_i on x_i to get:

$$\lambda_i = \sum_s \gamma_s x_{is} + \eta_i, E(\eta_i x_{is}) = 0.$$

Set $a_{ts} = E(y_{it}x_{is})$ a $T \times T$ matrix. Write a_{ts} as a function of β, f_t and γ_s (i.e. $1 + 2T$). Normalize for some t, $f_t = 1$. Even the 2 period case is identified.

Estimating equation

Consider

$$Y_i = X_i\beta + F\lambda_i + U_i,$$

and the projector on the orthogonal to F, M_F:

$$M_F = I_T - F(F'F)^{-1}F' = I - \frac{FF'}{T}.$$

Derive:

$$M_F Y_i = M_F X_i\beta + M_F U_i.$$

The moment restrictions are :

$$E(M_F(Y_i - X_i\beta) \otimes vec(X_i)) = E(M_F U_i \otimes vec(X_i))) = 0.$$

Remark: Identification condition (see **Discussion**)
Remark 2: The variance of the right hand side can be written as (under homoskedasticity wrt X_i):

$$M_F V(U_i) M_F \otimes E(vec(X_i)vec(X_i)')).$$

Estimation method

Non-linear GMM since M_F is a non linear function. Define the T^2K vector:

$$h_N(\beta, F) = \frac{1}{N}\sum_{i=1}^{N}(M_F(Y_i - X_i\beta) \otimes vec(X_i)).$$

The GMM estimates are given by, indexing them by any square matrix A of dimension $T \times T$:

$$\min_{\beta,F} h_N'(\beta, F)\left[M_F A M_F \otimes E(vec(X_i)vec(X_i)'))\right]^{-1} h_N(\beta, F).$$

Note that F appears in h_N but also in the weighting matrix (i.e. continuously updated GMM, one of the empirical likelihood methods see Newey and Smith, 2004).

Iterative procedure: Start with $F^{(0)}$ and construct the estimate of β,say $\hat{\beta}^{(0)}$. Fix $\hat{\beta}^{(0)}$ and then solve for F (see paper for a simplification) etc etc

Asymptotic properties

T fixed, $N \to \infty$

- The GMM estimate is consistent
- The optimal GMM is efficient and is given by choosing $A = V(U_i)$.

Extensions:

- Using higher order moments : related to homoskedasticity for instance

Number of factors

Using the following property:

- If $R < R_0$: The moment restriction above $E(M_F(Y_i - X_i\beta) \otimes vec(X_i)) = 0$ is generically not true. $M_F F_0 \neq 0$.
- If $R = R_0$: The moment restriction identifies β_0
- If $R > R_0$: Some elements of F are not identified although β_0 is identified.

Those translate into:

- If $R = R_0$: The optimal GMM criterion, a.k.a. the Hansen-Sargan test statistic is distributed chi-square.
- If $R < R_0$: The optimal GMM criterion diverges to ∞

so that the test based on the Hansen-Sargan statistic is consistent. A possible procedure: start with $R = 1$ and increase progressively until not rejecting the null. (Beware that tests are nested so that some corrections are needed, see paper).

A likelihood procedure using fixed effects

Also known as concentrated least squares (Kiefer, 1980, Lee, 1991).
Assume $T = 2$, a one factor setting and no regressor i.e.:

$$y_{i1} = \lambda_i + u_{i1}, y_{i2} = f_2\lambda_i + u_{i2}.$$

Consider the least-square criterion, or pseudo-likelihood criterion, associated with a diagonal matrix for $V(u_i)$:

$$\sum_i \left[(y_{i1} - \lambda_i)^2 + (y_{i1} - f_2\lambda_i)^2\right]$$

Maximizing with respect to λ_i and replacing leads to a concentrated or profiled likelihood, which maximised wrt to f_2 results in (see **Discussion**):

$$\sum_i (y_{i1} + \hat{f}_2 y_{i2})(y_{i2} - \hat{f}_2 y_{i1}) = 0.$$

Inconsistencies

This estimate is consistent if and only if (see **Discussion**):

$$V(u_i) = \sigma^2 I$$

and even if consistent, is not efficient because it fails to use the second restriction that :

$$E(y_{i2} - f_2 y_{i1}) = 0$$

Summary

- For T fixed, the GMM estimate is consistent and if optimal, is efficient
- For T fixed, the fixed effect estimate obtained by concentrated OLS or pseudo-likelihood, is consistent if and only if errors are homoskedastic and non serially correlated because of the incidental parameter issue.

Extensions:

- Robertson and Sarafidis (2015): extension of GMM methods to dynamic panel data.

Bai, 2009

A full fixed effect approach

Consider that factor loadings and factors are fixed and treated as parameters.

From ALS (2001), estimates are not consistent for T fixed and this "contaminates" all other parameters to estimate including common effects i.e. parameters β.

Assume that N and T are large. Assume also that no factors or factor loadings are known (e.g. take deviations from individual and time series means first.

Write the *interactive effect* model as:

$$\underset{[T,1]}{Y_i} = \underset{[T,K]}{X_i}\beta + \underset{[T,r]}{F}\lambda_i + U_i.$$

Remark: If F and λ_i are not correlated with X_i then an OLS estimate of β is consistent and $\sqrt{NT}$ asymptotically normal (albeit inefficient).

Properties

Even under weak serial and cross-section dependence and heteroskedasticity (a.k.a. an approximate factor structure, Chamberlain and Rotschild, 1983) :

- Factors and factor loadings are estimated by Principal Component Analysis (PC or PCA)
- The OLS estimate of β is is consistent and $\sqrt{NT}$ asymptotically normal (albeit asymptocally biased) (Looks like the within estimator which consistency is restored in dynamic models when $T \rightarrow \infty$).
- These biases can be corrected and bias-corrected estimators are consistent and asymptotically normal and unbiased

Estimation

The least squares objective function is:

$$SSR(\beta, F, \Lambda) = \sum_{i=1}^{N} (Y_i - X_i\beta - F\lambda_i)'(Y_i - X_i\beta - F\lambda_i)$$

under the normalization constraints, $F'F/T = I$ and $\Lambda'\Lambda$ being diagonal.
Define the projection matrix on the orthogonal to factors :

$$M_F = I - FF'/T.$$

If F were known:

$$\hat{\beta}(F) = \left(\sum_{i=1}^{N} X_i' M_F X_i \right)^{-1} \sum_{i=1}^{N} X_i' M_F Y_i.$$

Estimation of factors

Given β and $W_i = Y_i - X_i\beta$,this is a factor model and:

$$SSR(\beta, F, \Lambda) = \sum_{i=1}^{N}(W_i - F\lambda_i)'(W_i - F\lambda_i)$$

Concentrating out λ_i i.e. using $\hat{\lambda}_i(F) = (F'F)^{-1}F'W_i$ and $W_i - F\lambda_i = M_F W_i$ we get:

$$SSR(\beta, F, \hat{\Lambda}) = \sum_{i=1}^{N} W_i' M_F W_i = \sum_{i=1}^{N} W_i' W_i - \sum_{i=1}^{N} W_i' F F' W_i.$$

Matrix $\sum_{i=1}^{N} W_i W_i'$ (dim$=T \times T$) is symmetric s.d.p. so diagonalizable with non negative eigenvalues.
A solution consists in choosing F as the eigenvectors of $\sum_{i=1}^{N} W_i W_i' = \sum_{i=1}^{N}(Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})'$ associated with its r largest eigenvalues (See **Discussion**).

Summary

The least-square or interactive effect estimate solves two equations:

$$\hat{\beta}_{NT} = \hat{\beta} = \left(\sum_{i=1}^{N} X_i' M_{\hat{F}} X_i\right)^{-1} \sum_{i=1}^{N} X_i' M_{\hat{F}} Y_i,$$

$$\sum_{i=1}^{N} (Y_i - X_i\hat{\beta})(Y_i - X_i\hat{\beta})'\hat{F} = D_{NT}^{(r)}\hat{F}$$

in which $D_{NT}^{(r)}$ are the r largest eigenvalues ordered in a decreasing way.
Remark: The algorithm suggested by Bai (2009) is to replace the first equation by an equation for a given F and Λ and iterate: (see **Discussion)**

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \sum_{i=1}^{N} X_i'(Y_i - F\lambda_i)$$

Remark 2: It is not a convex program and any algorithm might converge to a local minimum (see below). Assume that we have an initially converging estimate then this one is converging (Hsiao, 2018)

Random effect GLS

If λ_i is correlated with X_i, we can always "control" the correlation by linearly projecting λ_i on X_i :

$$\underset{[r,1]}{\lambda_i} = \underset{[r,TK]}{L} \underset{[TK,1]}{vec(X_i)} + \eta_i, \quad E(vec(X_i)\eta_i') = 0.$$

The model becomes:

$$Y_i = X_i\beta + FLvec(X_i) + F\eta_i + U_i.$$

in which

$$E((F\eta_i + U_i).vec(X_i')) = 0.$$

If in addition X_{it} and f_t are uncorrelated, then random effect GLS is consistent and asymptotically normal.

Remark 1: The number of terms explodes with T.

Remark 2: This might be slightly inefficient if the structure of the correlation between λ_i on X_i is specific.

Identification

Define for all (i,k), $a_{ik} = \lambda_i'(\frac{\Lambda'\Lambda}{N})^{-1}\lambda_k$ and note that $a_{ik} = a_{ki}$ and $\frac{1}{N}\sum_j a_{ij}a_{jk} = a_{ik}$.
Define:

$$Z_i = M_F X_i - \frac{1}{N}\sum_{k=1}^{N} M_F X_k a_{ik} \Longrightarrow \sum_{i=1}^{N} a_{ij} Z_i = 0$$

Then define:

$$\begin{aligned} D_{NT}(F) &= \frac{1}{NT}\sum_i Z_i' Z_i = \frac{1}{NT}\sum_i \left(\frac{1}{T}\sum_t Z_{it}' Z_{it}\right) \\ &= \frac{1}{NT}\sum_i X_i' M_F X_i - \frac{1}{T}\left[\frac{1}{N^2}\sum_{i,k} X_i' M_F X_k a_{ik}\right]. \end{aligned}$$

Define $D(F)$ as the limit in probability of $D_{NT}(F)$ when N and $T \to \infty$.

Identification (2)

Rank condition:

$$\text{For all } F, D_{NT}(F) \text{ has full rank, } K.$$

Remark: $D_{NT}(F)$ is semi definite positive.
Remark 2: Rules out time-invariant and individual-invariant regressors.
Remark 3: If ε_{it} is time and cross section stationary white noise with variance σ^2 then we can show that when N and $T \to \infty$:

$$\sqrt{NT}(\hat{\beta} - \beta) \rightsquigarrow N(0, \sigma^2 D^{-1})$$

Remark 4: the rank condition can be weakened into D has full rank at the true value F^0 only.

Regularity assumptions

Boundedness conditions on the data generating process:

$$E(\|X_{it}\|^4) < M < \infty,$$

$$E(\|F_t\|^4) < M, \frac{1}{T}\sum F_t F_t' \xrightarrow[T\to\infty]{P} \Sigma_F,$$

$$E(\|\lambda_i\|^4) < M, \frac{1}{T}\Lambda'\Lambda \xrightarrow[T\to\infty]{P} \Sigma_\Lambda.$$

+ Assumption C on weak cross section and time series dependence (see first pages).
Assumption D: ε_{it} is independent of all X_{js}, λ_j and F_s.

Consistency

Under those Assumptions:

$$\hat{\beta} - \beta_0 \xrightarrow[N,T\to\infty]{P} 0$$

$$F^0 \hat{F} / T \text{ is invertible and } \left\| P_{\hat{F}} - P_{F^0} \right\| \xrightarrow[N,T\to\infty]{P} 0.$$

Remark: $\hat{F}$ cannot converge to F^0,only subspaces converge.
Remark 2: $(\hat{\beta}, \hat{F})$ are the minimizer of the least squares criterion, $S_{NT}(\beta, F)$. The proof proceeds in showing that (Newey and McFadden, 1994):
(i) $S_{NT}(\beta, F)$ converges uniformly to $S(\beta, F)$ on some bounded set.
(ii) Show that that β_0, F_0 is the unique minimum of $S(\beta, F)$.
Difficulty: The dimension of F is growing with T so the proof needs to be adapted (see **Discussion**).

Asymptotic expansion

After tedious algebraic manipulations (proposition A3, p1274) and if $T/N^2 \rightarrow 0$ and $N/T^2 \rightarrow 0$:

$$\begin{aligned}\sqrt{NT}(\hat{\beta}-\beta) &= D(\hat{F})^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\left[X_i' M_{F^0} - \frac{1}{N}\sum_{k=1}^{N} a_{ik} X_k' M_{F^0}\right]\varepsilon_i \\ &\quad + \sqrt{\frac{T}{N}}B + \sqrt{\frac{N}{T}}C + o_P(1)\end{aligned}$$

If $\frac{T}{N} \rightarrow \rho \neq 0$, then the rate of convergence of $\hat{\beta}$ is $\sqrt{NT}$ (Theorem 1, p1244).
B and C are asymptotic biases.

- $B = 0$: no cross-sectional correlation nor heteroskedasticity
- $C = 0$: no time series correlation nor heteroskedasticity

Asymptotic distribution

Theorem 3: if $N/T \rightarrow \rho \neq 0$ then:

$$\sqrt{NT}(\hat{\beta} - \beta) \underset{N,T\rightarrow\infty}{\xrightarrow{d}} N(\sqrt{\rho}B + \frac{C}{\sqrt{\rho}}, D_0^{-1} D_Z D_0^{-1})$$

in which $D_0 = \sigma^2 p\lim\left[\frac{1}{NT}\sum_i \sum_j \sum_t Z_{it} Z'_{jt}\right]$ and D_Z defined in (16).

Number of factors

Bai and Ng (2002) write the most used IC_2 criterion as (assuming $\beta = 0$):

$$IC_2(r) = \log \frac{1}{NT} \sum (y_{it} - \hat{F}^{(r)} \hat{\lambda}_i)^2 + r \frac{N+T}{NT} \log(\min(N,T))$$

and minimize it.
Under iid shocks, good performance for N, T bigger than 20 (Sul, 2019).
Remark 1: If strong serial correlation take first differences
Remark 2: Normalize by each individual standard deviation:

$$\hat{\sigma}_i^2 = \frac{1}{T-1} \sum_t (y_{it} - y_{i.})^2.$$

Remark 3: Robustness checks using different subsamples
See survey in Choi and Jeong, 2018, EconReviews

Pesaran, 2006

Intuition

Write a single factor model:

$$y_{it} = \lambda_i f_t + u_{it},$$

which by averaging yields:

$$y_{.t} = \bar{\lambda} f_t + u_{.t}$$

if $\bar{\lambda} = \frac{1}{N}\sum \lambda_i$. Provided that $plim_{N\to\infty}\bar{\lambda} \neq 0$ and $plim_{N\to\infty}(u_{.t}) = 0$, normalize $f_1 = 1$ and form estimates:

$$\hat{f}_t = \frac{y_{.t}}{y_{.1}}.$$

Generalize to multiple factors by solving linear systems under constraints. Consider for instance weighted cross-section averages

$$y_{wt} = \sum w_i y_{it}.$$

in which weights w_i might depend on explanatory variables.

Estimation approach

Form an auxiliary regression to which cross-section (possibly weighted) averages are added. OLS estimates are shown to have good properties.
Two different sort of issues:

- The coefficients of the individual specific regressors
- The means of the individual specific coefficients assumed random (Swamy, 1970).

Both are CCE (common correlated effects) estimators.

Set up

Allow for slope heterogeneity and observed (d_t) and unobserved common factors (f_t):

$$y_{it} = d_t\alpha_i + x_{it}\beta_i + f_t\gamma_i + u_{it}.$$

There could be correlation between factors and regressors and we assume that the regressors have a factor structure:

$$\underset{[1,K]}{x_{it}} = d_t A_i + \underset{[1,r]}{f_t}\underset{[r,K]}{\Gamma_i} + v_{it}$$

Combining (See **Discussion)**:

$$z_{it} = (y_{it}, x_{it}) = d_t B_i + f_t \underset{[r,K+1]}{C_i} + \varepsilon_{it}.$$

in which u_{it}, v_{it} are independent of all other variables.
Parameter of interest $\beta = E(\beta_i)$ and we are not particularly interested in getting consistent estimates for factors ($\neq$ Bai).

Assumptions

1. Factors are stationary
2. Errors u_{it} and v_{js} are independent and stationary.
3. Factor loadings γ_i and Γ_i are independent of everything and iid across i and $\gamma_i = \gamma + \eta_i$.
4. Random slope independent of everything (including γ_i) and $\beta_i = \beta + v_i$.

Remark: Note that γ_i is supposed to be independent of x_{it} and thus of Γ_i.

Identification: rank restrictions

Define $H_w = (D, \bar{Z}_w)$ where $\bar{Z}_w$ are weighted (by a system of weights w_i) cross section averages of z_{it} (dimension $T, K+1$) and:

$$M_w = I_T - H_w(H_w' H_w)^{-1} H_w'$$

Similarly for $G = (D, F)$ and M_G.

- Identification of β_i: The $K-$dimensional matrices, $\frac{1}{T} X_i' M_w X_i$ and $\frac{1}{T} X_i' M_G X_i$ are non singular and their inverses have finite second-order moments.
- Identification of β : The $\theta-$ weighted pooled matrix

$$\sum_i \theta_i \frac{1}{T} X_i' M_w X_i$$

is non singular for some scalar weights θ_i.

Construction

Consider weighted cross-section averages:

$$\bar{z}_{wt} = d_t \bar{B}_w + f_t \bar{C}_w + \bar{\varepsilon}_{wt}$$

and suppose $rank(\bar{C}_w) = r \leq K + 1$ (**less factors than the number of regressors+outcome**). Invert and get:

$$f_t = (\bar{z}_{wt} - d_t \bar{B}_w - \bar{\varepsilon}_{wt}) \bar{C}'_w (\bar{C}_w \bar{C}'_w)^{-1}.$$

As $plim_{N\to\infty} \bar{\varepsilon}_{wt} = plim_{N\to\infty} \sum_i w_i \varepsilon_{it} = 0$ if $E(\varepsilon_{it}) = 0$ and w_i has some regularity. Bai (2009) claims that this is key in distinguishing this model from his set up in which it is not clear that factors exhaust all the time series variation of regressors (as they do for outcomes).

We also have:

$$\bar{C}_w \xrightarrow[N\to\infty]{P} C = (\gamma, \Gamma) \begin{pmatrix} 1 & 0 \\ \beta & I \end{pmatrix}$$

Construction (ct'd)

.... so that:

$$f_t - (\bar{z}_{wt} - d_t \bar{B}_w) C'(C'C)^{-1} \xrightarrow[N \to \infty]{P} 0$$

which suggests that $\bar{z}_{wt}$ and d_t are good proxies for unobserved factors.

Remark: We should have $rank(C) = r$ and in particular $rank(\Gamma) = r$. But the replacement continues to work even if C is rank deficient e.g. $C = 0$.

Remark 2: YET: If γ_i is correlated with x_{it}, additional controls are needed under the form of individual averages of y_{it} and x_{it}. See Westerlund and Urbain below.

Individual specific coefficients

$$\hat{\beta}_i = (X_i' M_w X_i)^{-1} (X_i' M_w Y_i)$$

where M_w is defined above as the projector on the orthogonal to the observed factors and cross-section averages.
Theorem: Suppose $\|\beta_i\| < M$, as well as $\|A_i\|$ and $\|\Gamma_i\|$, and that $T/N^2 \longrightarrow 0$, then:

$$\sqrt{T}(\hat{\beta}_i - \beta_i) \xrightarrow[N,T\rightarrow\infty]{d} N(0, \Sigma_{b_i})$$

in which $\Sigma_{b_i} = \Sigma_i^{-1} plim_{T\rightarrow\infty} \left[\frac{1}{T} X_i' M_G E(\varepsilon_i' \varepsilon_i) M_G X_i'\right] \Sigma_i^{-1}$ and $\Sigma_i = plim_{T\rightarrow\infty} \left[\frac{1}{T} X_i' M_G X_i'\right]$.

Estimate of the mean

Consider the Mean Group estimator:

$$\hat{\beta}_{MG} = \frac{1}{N} \sum_i \hat{\beta}_i$$

Remark: We could consider weighting the individual elements by their relative variance. It is shown in Hsiao, Pesaran and Tahmiscioglu (1999) that the two estimators are asymptotically equivalent when $N, T \longrightarrow \infty$.
We obtain:

$$\sqrt{N}(\hat{\beta}_{MG} - \beta) \underset{N,T\rightarrow\infty}{\xrightarrow{d}} N(0, \Sigma_{MG})$$

Pooled estimate

An efficient estimate is obtained by waiting the individual elements, $\hat{\beta}_i$, exploiting the fact that they all estimate the same parameter under slope homogeneity (an assumption which might be incorrect) .

$$\hat{\beta}_P = \frac{1}{N}\sum_i v_i \hat{\beta}_i$$

in which v_i are weights which are equal to the weights w_i used in the construction of cross section averages. This estimate is asymptoticaly normal even if the .

Assume slope homogeneity, $\beta_i = \beta$. The same kind of asymptotic properties also holds under stronger assumptions and in particular if $T/N \rightarrow 0$. See below for an evaluation.

Monte Carlo experiments

TABLE I

SMALL SAMPLE PROPERTIES OF COMMON CORRELATED EFFECTS TYPE ESTIMATORS IN THE CASE OF EXPERIMENT 1A (FULL RANK + HETEROGENEOUS SLOPES)

	Bias (×100)					RMSE (×100)					Size (5% level, $H_0:\beta_1=1.00$)					Power (5% level, $H_1:\beta_1=0.95$)				
(N,T)	20	30	50	100	200	20	30	50	100	200	20	30	50	100	200	20	30	50	100	200
CCE type estimators																				
CCEMG																				
20	0.18	−0.16	−0.08	0.06	−0.10	9.73	7.84	6.52	5.59	5.15	7.95	6.90	7.15	7.25	7.10	11.60	13.85	15.70	18.55	20.65
30	−0.18	0.02	−0.02	−0.09	0.09	7.42	6.02	5.11	4.41	4.10	6.85	6.05	6.50	6.50	5.90	11.60	15.90	19.10	22.75	26.80
50	−0.05	0.15	−0.07	0.15	−0.04	5.78	4.62	3.96	3.41	3.11	6.25	5.90	6.75	6.30	5.90	15.10	21.45	25.10	34.60	37.10
100	0.02	0.02	0.03	0.04	0.03	4.06	3.48	2.83	2.33	2.26	5.05	5.90	5.65	5.15	6.35	24.90	33.20	43.45	55.95	62.40
200	−0.08	−0.03	−0.01	0.05	0.00	3.07	2.44	1.96	1.71	1.51	5.75	5.60	5.50	5.35	5.05	37.15	52.90	70.55	84.70	89.95
CCEP																				
20	0.26	−0.13	−0.03	0.02	−0.12	8.70	7.42	6.44	5.70	5.36	8.00	7.75	7.45	7.65	7.10	12.45	14.05	16.00	18.15	20.20
30	−0.23	−0.04	0.01	−0.09	0.11	6.99	5.91	5.21	4.52	4.24	6.45	6.35	7.20	6.45	6.80	13.15	15.80	19.15	21.75	27.75
50	−0.05	0.16	−0.04	0.13	−0.01	5.27	4.52	3.98	3.43	3.19	6.25	6.15	6.30	6.00	6.25	17.00	21.90	26.25	33.50	37.15
100	0.08	0.03	0.01	0.01	0.02	3.73	3.31	2.84	2.35	2.28	4.90	6.00	5.30	5.20	6.15	28.70	35.45	44.00	54.10	61.60
200	−0.05	−0.05	−0.04	0.04	0.01	2.69	2.34	1.95	1.70	1.53	4.80	5.55	4.85	5.10	4.60	45.20	56.55	70.85	83.80	89.35
Infeasible estimators (including f_{1t} and f_{2t})																				
Mean group																				
20	−0.07	−0.15	−0.15	0.15	−0.10	7.58	6.60	5.76	5.11	4.78	6.85	6.90	7.00	6.50	6.45	13.20	14.70	16.65	20.15	18.80
30	−0.11	−0.03	0.00	−0.03	0.12	5.87	5.00	4.53	4.01	3.88	6.00	5.10	5.70	5.40	5.75	15.70	18.60	22.40	24.45	27.45
50	0.05	0.09	−0.06	0.13	−0.03	4.47	3.82	3.46	3.15	2.98	6.55	5.55	6.20	5.25	5.35	22.30	26.95	30.40	37.50	39.95
100	0.02	0.03	0.03	0.01	0.04	3.15	2.86	2.50	2.17	2.17	4.80	5.50	4.85	5.05	5.45	35.15	45.40	53.00	60.15	65.70
200	−0.05	0.02	−0.04	0.06	0.01	2.26	1.98	1.71	1.59	1.46	4.80	5.25	4.45	5.75	4.80	58.00	71.60	80.55	89.40	91.85
Pooled																				
20	0.15	−0.19	−0.20	−0.05	−0.09	7.22	6.71	6.44	5.98	5.76	6.60	7.25	7.20	7.40	7.20	13.30	14.40	16.60	18.55	17.45
30	−0.13	−0.10	0.07	0.03	0.13	6.02	5.39	5.16	4.66	4.56	6.90	5.10	6.80	5.50	5.65	15.65	16.80	19.70	20.60	22.80
50	0.16	0.15	−0.05	0.14	−0.01	4.50	4.11	3.81	3.62	3.50	5.95	6.25	6.05	5.60	5.85	23.05	25.80	27.70	31.55	31.60
100	−0.06	0.03	0.03	0.01	0.05	3.15	3.06	2.78	2.57	2.56	5.15	6.00	4.95	4.85	5.35	34.75	41.25	44.00	48.85	53.25
200	−0.08	0.00	−0.06	0.06	−0.01	2.29	2.12	1.90	1.86	1.72	5.00	5.45	4.65	5.05	4.60	58.55	66.55	72.70	77.80	81.05

PANELS WITH MULTIFACTOR ERROR STRUCTURE

Monte Carlo experiments: Naïve estimators

TABLE I—*Continued*

	Bias (×100)					RMSE (×100)					Size (5% level, $H_0 : \beta_1 = 1.00$)					Power (5% level, $H_1 : \beta_1 = 0.95$)				
(N, T)	20	30	50	100	200	20	30	50	100	200	20	30	50	100	200	20	30	50	100	200
Naïve estimators (excluding f_{1t} and f_{2t})																				
Mean group																				
20	14.73	14.21	14.05	14.11	13.90	19.45	18.05	17.08	16.51	16.02	31.95	34.20	39.25	44.45	48.10	47.00	51.30	58.35	66.65	70.15
30	15.64	15.23	15.35	15.07	15.06	19.44	18.00	17.57	16.79	16.50	43.15	47.80	56.00	60.70	65.55	60.95	68.75	76.75	83.05	87.30
50	14.85	14.58	13.94	14.14	14.02	18.10	17.08	15.86	15.40	15.03	58.20	64.25	66.75	76.80	81.55	76.15	82.25	86.50	94.30	96.50
100	14.64	15.08	14.79	14.61	14.43	17.01	16.90	16.04	15.44	15.03	72.90	81.35	88.45	94.85	97.35	89.50	94.45	98.65	99.55	99.85
200	14.91	14.89	14.62	14.54	14.49	17.08	16.45	15.65	15.12	14.88	85.30	92.00	96.10	99.20	99.85	95.05	98.65	99.70	100.00	100.00
Pooled																				
20	14.93	14.55	14.75	14.79	14.77	19.74	18.49	17.88	17.21	16.93	38.80	40.10	45.35	47.45	50.50	52.85	58.00	63.50	68.25	71.05
30	16.81	16.64	17.05	17.06	16.97	20.83	19.65	19.29	18.88	18.41	51.05	55.95	62.45	68.70	72.00	66.70	73.60	80.50	85.70	89.60
50	16.47	16.36	15.83	16.30	16.25	20.20	19.19	17.95	17.64	17.29	65.95	70.75	73.75	82.80	87.60	79.95	85.05	89.45	95.80	97.15
100	15.81	16.67	16.56	16.52	16.48	18.82	18.89	18.02	17.48	17.15	77.15	84.75	91.15	95.85	98.10	90.50	94.50	98.30	99.45	99.85
200	16.08	16.44	16.37	16.51	16.57	18.89	18.53	17.75	17.24	17.04	85.50	91.75	95.70	99.20	99.90	95.05	98.20	99.65	100.00	100.00

Derivatives: Common Correlated Effects

Comparison: CCE and PC

Westerlund and Urbain (2015)
Set up: make it common for CCE and PC

- Parameter of interest β ("slope homogeneity")
- x_{it} has a factor structure
- *Principal components*: not as in Bai (2009) ("residuals") but extract factors directly from $z_{it} = (y_{it}, x_{it})$. See Greenaway-McGrevy et al. (2012)
- The remaining assumptions close to Pesaran (2006)
- Define

$$\hat{\beta}(\hat{F}_p) = (\sum_i X_i' M_{\hat{F}_p} X_i)^{-1} \sum_i X_i' M_{\hat{F}_p} Y_i$$

in which $\hat{F}_p$ is obtained by PC or by Cross-section Averages

Asymptotic properties

If $\frac{T}{N^2}$ and $\frac{N}{T^2} \longrightarrow 0$, then:

- $\hat{\beta}(\hat{F}_p)$ is consistent at the $\sqrt{NT}$ rate but have asymptotic biases except if $N/T \rightarrow 0$.
- the asymptotic distribution without bias is the same as when F is known. Bias corrected estimators are asymptotically equivalent.
- Both estimators are asymptotically equivalent when there are no biases (e.g. when $\Lambda_i = 0$).
- Relative biases depend on:
  - the extent of heterogenity in λ_i
  - the number of factors
  - relative magnitudes of σ_u^2 and Σ_η.
- If β close to zero, biases are smaller for PC than CA in plausible configurations since PC is more efficient for the estimation of factors

Monte Carlo experiment

Table 1
Bias, 5% size and MSE.

DGP	$\beta = 0$							
	Bias				5% size		MSE	
	PC	Theory	CA	Theory	PC	CA	PC	CA
$N = T = 50$								
1	−0.129	−0.122	1.630	1.813	6.2	37.6	1.061	3.817
2	−0.130	−0.092	0.225	0.266	5.6	5.8	1.071	1.107
3	−0.113	−0.122	1.638	1.813	5.1	37.2	1.027	3.797
4	−0.093	0.000	2.578	2.750	5.8	15.1	10.311	17.647
5	−0.093	0.000	0.232	0.275	6.7	12.1	0.127	0.167
6	0.030	0.030	1.408	1.588	6.1	94.5	0.117	2.187
7	−0.114	−0.107	0.037	0.041	7.3	5.8	0.121	0.108
8	−0.846	0.000	2.280	2.750	6.7	12.5	10.150	15.661
$N = T = 100$								
1	−0.124	−0.122	1.715	1.813	6.2	40.6	1.079	4.047
2	−0.114	−0.092	0.242	0.266	5.4	5.6	1.018	1.062
3	−0.104	−0.122	1.728	1.813	5.9	42.2	1.061	4.084
4	−0.067	0.000	2.635	2.750	6.8	14.1	10.871	18.097
5	−0.033	0.000	0.257	0.275	5.0	12.2	0.103	0.169
6	0.017	0.030	1.469	1.588	5.1	98.4	0.100	2.308
7	−0.111	−0.107	0.039	0.041	6.5	6.0	0.118	0.107
8	−0.342	0.000	2.488	2.750	6.0	13.8	10.327	16.955
$N = T = 200$								
1	−0.136	−0.122	1.742	1.812	6.2	41.3	1.073	4.128
2	−0.102	−0.092	0.256	0.266	5.5	6.2	1.046	1.102
3	−0.125	−0.122	1.765	1.812	4.8	41.8	0.979	4.100
4	−0.021	0.000	2.711	2.750	5.5	14.5	10.206	17.685
5	−0.012	0.000	0.268	0.275	5.5	13.6	0.102	0.175
6	0.019	0.030	1.521	1.587	4.9	99.5	0.104	2.446
7	−0.097	−0.107	0.051	0.041	6.0	4.9	0.110	0.104
8	−0.092	0.000	2.672	2.750	5.1	13.5	9.972	17.263

A first correction

What happens if the number of cross-sections averages is too small wrt to the number of factors i.e. $r > k+1$?
Pesaran (2006) claims that it does not matter for consistency. This is true if and only if factor loadings are uncorrelated with covariates (see Westerlund and Urbain, 2013). If they are not, the CCE becomes inconsistent.
Intuition: cross-section averages are not spanning the full space of factors. The "remaining" directions are uncorrelated with covariates if and only factor loadings are uncorrelated with covariates.

Example

Write a model with two-way fixed effects and factors

$$y_{it} = x_{it}\beta + \alpha_i + \delta_t + f_t\lambda_i + u_{it}.$$

To get rid of α_i and δ_t, multiply by the within estimator and denote $y^*_{it} = y_{it} - y_{i.} - y_{.t} + y_{..}$:

$$y^*_{it} = x^*_{it}\beta + (f_t - f_.)(\lambda_i - \lambda_.) + u^*_{it}.$$

All cross section averages are equal to zero because $\lambda^*_i = \lambda_i - \lambda_.$ are centered. The Pesaran cross section averages are uninformative and do not control for $(\lambda_i - \lambda_.)$ if those are correlated with x^*_{it}. To be consistent the pooled estimator needs the assumption of uncorrelation.

A second correction

Karabiyik, Reese and Westerlund, 2017.
Proofs of asymptotic distribution of CCE are incorrect when $r < k + 1$. Only valid when $r = k + 1$.
When $r = k + 1$, estimate $\hat{F}$ is consistent for the space spanned by F. When $r > k + 1$,this is still true but the variance of the estimated factors become singular. This is the result which necessitates correction when $r < k + 1$ and there are additional bias terms appearing contradicting existing results (including Pesaran, 2006, Westerlund and Urbain, 2015). So the estimation of F matters for the asymptotic distribution.

Other cross-section averages

When $r > k + 1$, we have seen above that some uncorrelatedness assumption beteen factor loadings and covariates is needed.
Another solution is to augment the number of cross-section averages.
Instead of one set of weights, we can consider m sets of weights $w_i^{(m)}$ and construct cross section averages accordingly, $\bar{z}^{(m)}$. In this case, the number of cross section averages becomes $(k+1)m$ that might be larger than r. Yet, a rank condition is necessary.
Idea: Use individual observed variables, $\xi_i^{(m)}$ to construct the weights $w_i^{(m)}$. They act as "instruments".
See Karabiyik, Urbain and Westerlund (2017)

Maximum Likelihood Approach

Bai and Li (2014).
Key feature: x_i correlated with (λ_i, f_t) but this gives rise to a problem with too many parameters and an incidental parameter issue.
Frame the common shock model (i.e. Pesaran) differently in a ML set-up.

$$(I_N \otimes B)z_t = \mu + \Gamma f_t + \varepsilon_t$$

in which B depends on β. Suppose β and Γ are fixed.
Assumptions:

- No cross section dependence but heteroskedasticity, Σ_{ii} within a block. $V(\varepsilon)$ is a block diagonal matrix.
- Strict exogeneity of x_{it} conditional on factors.
- All parameters are bounded.

(Pseudo)-ML properties

Write the pseudo-likelihood function.
The MLE has no asymptotic bias even if cross section heteroskedasticity.
Furthermore, the consistency rate is:

$$\hat{\beta} - \beta = O_P(1/\sqrt{NT}) + O_P(T^{-1})$$

and teh asymptotic development yields:

$$\sqrt{NT}(\hat{\beta} - \beta) = A + O_P(T^{-1}N^{1/2}) + O_P(N^{-1/2}) + O_P(T^{-1/2}).$$

Remark: Same type of results if there are factors excluded from the outcome equation.
Remark 2: Same type of results if non time varying regressors are interacted with time varying coefficients or macro variables interacted with individual coefficients.
Algorithm: Expectation/Conditional Maximization (ZigZag in β and Γ, F)

Other readings

- Greenaway-McGrevy et al. (2012): factors estimated by PC but in a Pesaran framework.
- Chudik, Pesaran and Tosetti (2011): definition of strong and weak cross-section dependence; extension of Pesaran to the presence of infinitely many weak dependent factors
- Westerlund (2018): extension of CCE to factors of any kind: deterministic, non stationary etc.
- Chudik, Pesaran and Yang (2019): weak exogeneity of regressors and the application of Dhaene and Jochmans (2015) jackknife correction.
- Westerlund, Petrova and Norkute (2019): CCE in a fixed T setting: consistency arguments
- Westerlund (2019a): CCE and PC in a fixed T setting
- Westerlund (2019b): The zero sum estimators among which CCE and two way fixed effect: conditions for consistency.

Derivatives: Principal Components

Misspecification of the number of factors

Moon and Weidner (2015): both authors extended Bai (2009) to the case of predetermined regressors.
Remark: Asymptotic properties derived when the number of factors, r_0, is known.
If r in the model is strictly lower than r_0, the LS estimator is generically inconsistent.
Suppose $r \geq r_0$. What do we lose? Nothing.
Provided certain conditions are satisfied then $\hat{\beta}(r)$ is asymptotically equivalent to $\hat{\beta}(r_0)$.
Conditions:

- Errors are iid normal regressors (assumed for tractability reasons)
- Factors are strong
- Identification: $E(vec(X')(M_F \otimes M_{\Lambda_0})vec(X)) > 0$

Ranks of regressors

Regressors are composed of a "low rank" strictly stationary component, a "high rank" strictly stationary component and a "high rank" predetermined component. (See **Discussion**)
A regressor, X_k, of dimension $[T, N]$ is :

- of low rank if $rank(X_k)$ is bounded when N and $T \to \infty$.
- of high rank if $rank(X_k)$ diverges when N and $T \to \infty$.

Asymptotic properties

If $N/T \to \tau, 0 < \tau < \infty$ then:

$$\sqrt{NT}(\hat{\beta}_R - \beta) = \sqrt{NT}(\hat{\beta}_{R_0} - \beta) + o_P(1).$$

The asymptotic distribution of $\sqrt{NT}(\hat{\beta}_{R_0} - \beta)$ is e.g. given in Bai (2009) (asymptotic normality, bias and variance).
Additional results on the estimation of biases and variances in this asymptotic distribution. Then construct bias corrected estimators.

Bai's estimator: convexity issues

It is a Least Squares estimator (or PMLE/QMLE) minimizing wrt $\beta = (\beta_1, ., \beta_K), \Lambda$ and F

$$\left\| Y - \sum_{k=1}^{K} X_k \beta_k - \Lambda F' \right\|_2^2$$

where Y, X_k are $N \times T$ matrices while Λ is $[N, r]$ and F is $[T, r]$.The norm of a matrix, A, is $\|A\|_2 = \left[\sum \sum a_{ij}^2\right]^{1/2}$. This program would be convex in β and $\Gamma = \Lambda F'$ but it is not in Λ and F. It thius can have muliple minima and Bai's algorithm might not converge to the true value.

We can rewrite the minimization program as:

$$LS(\beta) = Min_{\Gamma, rank(\Gamma) \le r} \left\| Y - \sum_{k=1}^{K} X_k \beta_k - \Gamma \right\|_2^2$$

and the issue of convexity is related to the constraint, $rank(\Gamma) \le r$.

Singular values and norms

We can consider other norms.

Def 1: The singular values of a matrix A are the square roots of the eigenvalues of AA' (ie always sdp).

Def 2: $\|A\|_0 = rank(A)$ =number of non zero singular values

Def 3: *Nuclear norm:* $\|A\|_1 = \sum_{r=1}^{\min(N,T)} s_r(A)$ if $s_r(A)$ are the singular values of A (ranked in a decreasing order)

Prop: The least squares criterion above is (Moon and Weidner, 2017)

$$LS(\beta) = \sum_{r=r_0}^{\min(N,T)} (s_r(Y - \sum_{k=1}^{K} X_k\beta_k))^2$$

Convexity relaxation

Instead of imposing the constraint $\|\Gamma\|_0 = rank(\Gamma) \leq r$ which is difficult to deal with, we relax it using the nuclear norm and we penalize the distance to a low-rank matrix:

$$Q_\psi(\beta) = \left\| Y - \sum_{k=1}^{K} X_k \beta_k - \Gamma \right\|_2^2 + \psi \, \|\Gamma\|_1 .$$

Moon and Weidner (2017) show that this is a convex program which have a unique solution $\hat{\beta}_\psi$.

They also show that $\hat{\beta}_* = \lim_\psi \hat{\beta}_\psi$ exists and that:

$$\hat{\beta}_* = \arg\min_\beta \left\| Y - \sum_{k=1}^{K} X_k \beta_k - \Gamma \right\|_1 .$$

Properties

Under Bai's assumptions and $\psi = \psi_{NT} \to 0$ while $\sqrt{\min(N,T)}\psi_{NT} \longrightarrow \infty$,we have if N and $T \to \infty$:

$$\begin{aligned} \frac{1}{\sqrt{NT}} \left\| \hat{\Gamma}_{\psi} - \Gamma_0 \right\|_2 &\leq O_P(\psi), \left\| \hat{\beta}_{\psi} - \beta_0 \right\| \leq O_P(\psi), \\ \sqrt{\min(N,T)} \left\| \hat{\beta}_{*} - \beta_0 \right\| &\leq O_P(1). \end{aligned}$$

The rate is lower than the LS estimator. However, as $\hat{\beta}_{\psi}$ or $\hat{\beta}_{*}$ are consistent, we can use them as starting values for the algorithm of Bai (2009). By a result by Moon and Weidner, the result is $\sqrt{NT}$ consistent.

Low rank regressors and factors

It is difficult to identify simultaneously the number of factors and parameters of low rank regressors. Start with a single regressor, $x_{it} = \phi_t l_i$ and write:

$$y_{it} = \beta x_{it} + f_t \lambda_i + u_{it},$$

in which f_t has r elements. then we can rewrite for any β_0:

$$\begin{aligned} y_{it} &= \beta_0 x_{it} + (f_t, \phi_t)(\lambda_i', l_i(\beta - \beta_0))' + u_{it}, \\ &= \beta_0 x_{it} + f_t^* \lambda_i^* + u_{it}. \end{aligned}$$

That means that β, r, f_t, λ_i is observationally equivalent to $\beta_0, r+1, f_t^*, \lambda_i^*$ in which β_0 is arbitrary.
Impose the usual Bai's conditions for local identification. See Moon and Weidner (2018) for global identification.

Further reading

- Hsiao, 2018,
- Jiang, Yang, Gao and Hsiao, 2017
- Moon and Weidner, 2018
- Beyhum and Gautier, 2019

Conclusion

Characteristics of the models

Various dimensions, Hsiao (2018):

- Random or fixed effects for λ_i and f_t?
- N, T : fixed or ∞?
- Number of factors known or unknown?
- Structure of cross-section and serial dependence & heteroskedasticity
- Predetermined or strictly exogenous regressors (Not done in this Chapter, see Dynamic models M2 presentation)

Microeconometrics

- Number of factors unknown.
- N large, T small or moderately large: $T/N \to 0$, T/N^2, T/N^3?
- Correlated randon effects, i.e. fixed effects λ_i while f_t are fixed or random.
- Serial dependence, time and cross-section heteroskedasticity. Less clear for cross-section additional dependence (excepit if panel of countries, regions or other geographic units. Clusters are different.)

Pros and cons

- ALS: N large, T small. Efficient but costly to estimate.
- Bai: N large, T large. Very general in terms of correlated random effects.
- Pesaran: Restrictive for the correlation structure.