Principal components based estimation of multilevel factor models

... and the `gretl` package: `GlobalFactors`

Ioannis Venetis[α] and Paraskevi Salamaliki[β]

[α]University of Patras
Dep. of Economics

[β]University of Ioannina
Dep. of Economics

June 16, 2023

GC2023, 8th Gretl Conference, Gdańsk, Poland

Multilevel factor model set up (**two** levels here)

✓ Parametric sparse structure. The model:

$$\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{M,t} \end{pmatrix} = \begin{pmatrix} \Lambda_1 & L_1 & 0 & \cdots & 0 \\ \Lambda_2 & 0 & L_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_M & 0 & 0 & \vdots & L_M \end{pmatrix} \cdot \begin{pmatrix} f_t \\ g_{1,t} \\ \vdots \\ g_{M,t} \end{pmatrix} + \begin{pmatrix} u_{1,t} \\ \vdots \\ u_{M,t} \end{pmatrix}$$

$$\underset{N\times 1}{y_t} = \underset{N\times r}{\Lambda} \cdot \underset{r\times 1}{K_t} + u_t$$

✓ total number of series: $N = \sum_{m=1}^{M} N_m$, total number of factors: $r = r_0 + \sum_{m=1}^{M} r_m$,

✓ $\underset{N_m\times r_m}{L_m}$, $\underset{N_m\times r_0}{\Lambda_m}$, $\underset{N_m\times 1}{y_{m,t}}$

(1a). Package Methods-in-brief. Choi et al. (2018)

Choi, Kim, Kim, Kwark, 2018. A multilevel factor model: Identification, asymptotic theory and applications. Journal of Applied Econometrics

- ✓ f_t, $g_1, ..., g_M$ are uncorrelated; weak correlation (within group) between f, g_m, u_m
- ✓ idiosyncratic error terms belonging to different groups are uncorrelated
- ✓ global factors make all of the observed data correlated
- ✓ idiosyncratic errors: (mild) serial and cross-sectional correlation within blocks
- ✓ no group should have a dominantly large or small number of series N_m
- ✓ (good) consistency rates: **(i)** $\hat{f}_t$ at $\min\{\sqrt{N}, T\}$ **(ii)** $\hat{g}_{m,t}$ at $\min\{\sqrt{N_m}, T\}$
- ✓ **Pre-step. Estimation of the number of group factors.** A sufficiently large and common upper bound r^*_{max}, satisfying

$$r^*_{max} \geq \max\{r_0 + r_1, ..., r_0 + r_M\} \quad (1)$$

must be selected in order to estimate the total number of factors in each group $r_0 + r_m$ given that r_0 is considered to be known. Given a choice for r^*_{max}

(1a). Package Methods-in-brief. Choi et al. (2018)

Choi, Kim, Kim, Kwark, 2018. A multilevel factor model: Identification, asymptotic theory and applications. Journal of Applied Econometrics

- ✓ **Pre-step (A).** Either use various static factor selection procedures (package offers: $IC_{p1}, IC_{p2}, IC_{p3}, BIC_3, PC_{p1}, PC_{p2}, PC_{p3}, ER, GR, ED$), to estimate the number of static factors in each group and then subtract the assumed r_0 value **or**
- ✓ **Pre-step (B).** or eliminate the (pre)estimated global factors $\hat{f}_t^{(1)}$ and apply static factor selection procedures to the residuals in each group. The initial estimator $\hat{f}_t^{(1)}$ is based on canonical correlation analysis using - in theory - any two groups m, h. In particular, it relies on PC estimates $\hat{K}_{m,t}, \hat{K}_{h,t}$ from any two groups.
- ✓ **Proposal:** adopt the block-pair m, h that yields the maximum canonical correlation amongst $\hat{K}_{m,t}, \hat{K}_{h,t}$. **In practice**, however, the pairwise identification strategy would not always lead to consistent estimation of the global factors, in particular, if local factors are cross-correlated or the block-pair shares the same (regional?) local factor (misspecification of the two-level model).

- ✓ **Steps 1-2-3. Estimation of the global/local factors and factor loadings.**
- ✓ **(1)** Select two groups m and h and obtain the initial estimator of f_t, denoted $\hat{f}_t^{(1)}$,
- ✓ **(2)** project $\hat{f}_t^{(1)}$ out of the data, then get a first estimate of L_m, $g_{m,t}$ by principal components denoted $\hat{L}_m^{(1)}$, $\hat{g}_{m,t}^{(1)}$
- ✓ **(3)** concentrate $\hat{L}_m^{(1)}$, $\hat{g}_{m,t}^{(1)}$ out of the model and obtain the final two-step estimates of global loadings and factors $\hat{\Lambda}_m^{(2)}$, $\hat{f}_t^{(2)}$
- ✓ **(4)** concentrate out the final global estimates and obtain the second and final estimation of $\hat{L}_m^{(2)}$, $\hat{g}_{m,t}^{(2)}$.

(1b). Package Methods-in-brief. Choi et al. (2021)

Choi, Lin, Shin, 2021. Canonical correlation-based model selection for the multilevel factors. Journal of Econometrics

- ✓ develop **two** consistent selection criteria to determine the number of global factors r_0
- ✓ Schematically, both criteria are described by the maximization procedure

$$\hat{r}_0 = \underset{r=0,1,\ldots,r^*_{max}}{\arg\max} \; CCD\,(r) \quad , \quad \hat{r}_0 = \underset{r=0,1,\ldots,r^*_{max}}{\arg\max} \; MCC\,(r)$$

based on a sufficiently large and common upper bound r^*_{max},

- ✓ robust to the presence of serially correlated and weakly cross-sectionally correlated idiosyncratic errors

(1b). Package Methods-in-brief. Choi et al. (2021)

Choi, Lin, Shin, 2021. Canonical correlation-based model selection for the multilevel factors. Journal of Econometrics

- ✓ based on average canonical correlations among all $M(M-1)/2$ block-pairs $\hat{K}_{m,t}$, $\hat{K}_{h,t}$
- ✓ first criterion, canonical correlation difference (**CCD**), **does not** allow for correlation among the local factors
- ✓ second criterion, modified canonical correlation (**MCC**), **does** allow for correlation among the local factors
- ✓ Focus on (practical case) fixed number of blocks M, still valid as $M \to \infty$.

Chen, 2022. Circularly Projected Common Factors for Grouped Data. Journal of Business & Economic Statistics

- ✓ Proposes **two** selection criteria based on the average residual sum of squares (ARSS) from a regression of (estimated) global factors on the factor spaces in each block
- ✓ allows for non-zero correlations between local factors, but does not cover the case of no-global factors $r_0 = 0$.
- ✓ Global factors are estimated using **two** alternative methods. Circular projection estimation (**CPE**) and Augmented circular projection estimation (**ACPE**)
- ✓ Schematically, both criteria are described by the maximization procedure

$$\hat{r}_{0,CPE} = \underset{r=1,\ldots,r^*_{max}}{\arg\max} \; f\left(ARSS^{CPE}_{r+1}\right) - f\left(ARSS^{CPE}_{r}\right)$$

$$\tilde{r}_{0,ACPE} = \underset{r=1,\ldots,r^{m_0}_{max}}{\arg\max} \; f\left(ARSS^{ACPE}_{r+1}\right) - f\left(ARSS^{ACPE}_{r}\right)$$

where the $f(\dot)$ is the logistic function, with ARSS being suitable scaled, and - of course - r^*_{max} is present

- ✓ $\hat{r}_{0,CPE}$ and $\hat{r}_{0,ACPE}$ do work under non-zero correlations between local factors - as mentioned above
- ✓ but do not cover the case of no global factors $r_0 = 0$

(2). Package Methods-in-brief. Chen (2022)

Chen, 2022. Circularly Projected Common Factors for Grouped Data. Journal of Business & Economic Statistics

- ✓ The first proposed method: circular projection estimation (CPE), based on the matrix

$$\left(\prod_{m=1}^{M} P\left(K_m\right)\right)' \cdot \left(\prod_{m=1}^{M} P\left(K_m\right)\right)$$

where $P\left(K_m\right) = K_m\left(K_m' K_m\right)^{-1} K_m'$.

- ✓ The second method, augmented circular projection estimation (ACPE) uses a *reference group*, say m_0

$$\left(K_{m_0}' K_{m_0}\right)^{-1/2} K_{m_0}' \left(\prod_{m=1}^{M} P\left(K_m\right)\right)' \left(\prod_{m=1}^{M} P\left(K_m\right)\right) K_{m_0} \left(K_{m_0}' K_{m_0}\right)^{-1/2}$$

- ✓ consistency rates: **(i)** $\hat{f}_t$ at $\min\{\sqrt{N_*}, \sqrt{T}\}$, $N_* = \min\{N_1, ..., N_M\}$ **(ii)** $\hat{g}_{m,t}$ at $\min\{\sqrt{N_*}, \sqrt{T}\}$

(3). Package Methods-in-brief. Lin and Shin (2022)

Lin and Shin, Nov 2022 Working Paper. Generalised Canonical Correlation Estimation of the Multilevel Factor.

- ✓ Local factors are allowed to be correlated or even identical across some blocks.
- ✓ consistent estimation of the global factors, does not even require orthogonality between global and local factors
- ✓ Develop a generalised canonical correlation approach (GCC)
- ✓ extends standard CCA by constructing a system-wide matrix, denoted $\mathbf{\Phi}$, that contains all $\mathbf{K}_m$ for $m = 1, \cdots, M$.
- ✓ A core computational element in their analysis, is the following matrix,

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{K_1} & -\mathbf{K_2} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{K_1} & \mathbf{0} & -\mathbf{K_3} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ & & & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K_{M-1}} & -\mathbf{K_M} \end{pmatrix}$$

with dimension $TM(M-1)/2 \times \sum_{m=1}^{M} (r_0 + r_m)$.

- ✓ **Estimation of the global/local factors and factor loadings.** Given $\hat{r_0}$, global factors $\hat{f}_t$ are estimated using the singular value decomposition $\mathbf{\Phi} = \mathbf{P} \cdot \mathbf{\Delta} \cdot \mathbf{Q}'$ of $\hat{\mathbf{\Phi}}$.
- ✓ estimation of local factors $\hat{g}_{m,t}$ and loadings $\hat{L}_m$, follows based on principal component analysis of the "residuals" $y_{m,t} - \hat{\Lambda}_m \cdot \hat{f}_t$.
- ✓ consistency rates: **(i)** $\hat{f}_t$ at $\min\{\sqrt{N_*}, \sqrt{T}\}$, $N_* = \min\{N_1, ..., N_M\}$ **(ii)** $\hat{g}_{m,t}$ at $\min\{\sqrt{N_*}, \sqrt{T}\}$
- ✓ **Estimation of the number of global factors.** Schematically, GCC criterion is described by the maximization procedure

$$\hat{r}_{0,GCC} = \underset{r=0,\ldots,r^*_{max}}{\arg\max} \frac{\hat{\delta}^2_{k+1}}{\hat{\delta}^2_k}$$

where the ratios of (squared) adjacent singular values of matrix $\mathbf{\Phi}$ are evaluated.

- ✓ no tuning parameters (except the sufficiently large and common upper bound r^*_{max}). In addition, it includes the boundary case of $r_0 = 0$.
- ✓ Simulations show that performs better than all aforementioned criteria (some of which overestimate when local factors are correlated)

Example 1 (part of package). Choi et al. (2018). $M = 25$ groups but small N_m

- ✓ an empirical application on international business cycles using a balanced quarterly panel dataset overing $M = 25$ OECD countries from 1981:Q1 to 2013:Q2. The total number of variables is "large" $N = 315$, and each variable has $T = 130$ observations but, within groups, the number of series N_m is limited. For example, there are $N_m = 17$ series for the US, $N_m = 9$ series for Greece, $N_m = 8$ series for Iceland and "only" $N_m = 6$ series for Turkey.
- ✓ **let's have a look at the code...**

Example 2 (part of package). Chen (2022). $M = 2$ groups, $T = 89$, $N_1 = 37, N_2 = 51, N = 88$

- ✓ The first empirical example in Chen (2022) employs a dataset for 22 developed countries ($m = 1$) and 33 emerging countries ($m = 2$). It covers the period from the 4th quarter of 1996 to the 4th quarter of 2018, a total of $T = 89$ time series observations
- ✓ it contains 22 series of real gross domestic product and 15 series of industrial production for developed countries, thus $N_1 = 37$ series, and 30 series of gross domestic product and 21 series of industrial production for emerging countries, thus $N_2 = 51$ series. In total, the dataset contains $N = N_1 + N_2 = 88$ series and it offers the opportunity to evaluate all methods under only two groups $M = 2$ and a sufficient number of group series N_1, N_2.
- ✓ **let's have a look at the code...**

Example 3 (part of package). Chen (2022). $M = 16$ groups, $T = 134$, $N_m = 32$, $N = 32 \cdot 16 = 512$

- ✓ The second empirical example in Chen (2022) employes monthly retail prices of 16 categories of commodities (by purpose) over the period from January 2010 to February 2021. The prices in each category are collected from 31 provinces in China, and one national price is calculated as an index reflecting the overall price level across the country.
- ✓ Hence, there are two ways to group the data: it can be divided into 16 groups, with each group corresponding to one category by purpose; it can also be classified into 32 groups, with each group corresponding to one province in China
- ✓ Because the new methods favor large N_m and small M, Chen (2022) groups the data by category.
- ✓ **let's have a look at the code...**

Thanks for your time

Thank you