Freie Universität Bozen
Libera Università di Bolzano
Università Liedia de Bulsan

Illuminating the factor and dependence structure in large panel models

28th UK Stata Conference 2022

Jan Ditzen

Free University of Bozen-Bolzano, Bozen, Italy
www.jan.ditzen.net, jan.ditzen@unibz.it

September 09, 2022

Motivation

- Large Panel models or Panel-Time Series models are a mix of time series and panel data models with a large number of observations over time (T) and cross-section units (N):
- What is large?
- In *theory* it means that N and T grow with the same speed to infinity $(N, T) \xrightarrow{j} \infty$ with $T/N \rightarrow \kappa,\ 0 < \kappa < \infty$.
- Panel-Time Series models often contain common factors (interactive fixed effects) and dependence across cross-sectional units.
- Common factors influence all or many cross-sectional units at the same time.
- Strong cross-sectional dependence occurs if common factors are part of the observed variables.

Econometric Model

- Most general model: dynamic panel model with heterogeneous slopes and interactive fixed effects:

$$\begin{aligned} y_{i,t} &= \lambda_i y_{i,t-1} + \beta_i x_{i,t} + u_{i,t} \\ x_{i,t} &= \gamma_{x,1,i} f_{1,t} + \gamma_{x,2,i} f_{2,t} + \xi_{i,t} \\ u_{i,t} &= \gamma_{u,1,i} f_{1,t} + \gamma_{u,3,i} f_{3,t} + \epsilon_{i,t} \end{aligned}$$

- We observe $y_{i,t}$ and $x_{i,t}$, the common factors $(f_{l,t})$ and the loadings $(\gamma_{k,l,i})$ are unobserved.
- $\xi_{i,t}$ and $\epsilon_{i,t}$ are both IID white noise.
- Potential dependence across units via the common factors $(f_{m,t}, m = 1, 2, 3)$.
- Understanding the factor structure important for estimation, two questions:
    - How many common factors are there?
    - How strong is the dependence across units?

Estimation of Number of Common Factors

Why Important?

- In many macroeconomic applications common factors play a key role. Examples: Asset returns, house prices, economic growth, oil price shocks,...
- The number of factors is necessary for:
  - Estimation of the exponent of cross-section dependence (Bailey et al., 2016, 2019).
  - The instrumental variables (IV) estimator for large panel data models **xtivdfreg** (Norkute et al., 2021; Kripfganz and Sarafidis, 2021).
  - The CCE (Pesaran, 2006; Karabiyik et al., 2020) or PCA (Bai, 2009) estimator require knowledge of the number of factors to account for cross-sectional dependence.
- **xtnumfac** (Reese and Ditzen) implements the methods to estimate the number of common factors proposed by Bai and Ng (2002); Ahn and Horenstein (2013); Onatski (2010); Gagliardini et al. (2019).

Estimation of Number of Common Factors

- Information Criteria by Bai and Ng (2002) ▸ details.
  - Idea: Penalise loss function $V()$ with a penalty $g()$ which increases with each additional factor: $PC(k) = V(k, F^k) - k\hat{\sigma}^2 g(N, T)$ and $IC(k) = ln(V(k, F^k)) - kg(N, T)$ with $V()$ are the squared sum of residuals of the variable of interest on the first k first principal components.
  - Bai and Ng (2002) suggest three different functions for $g(N, T)$.
  - Estimated number of factors is the number k which minimizes the statistics.
- Ahn and Horenstein (2013) Estimator ▸ details
  - Disadvantage of Bai and Ng (2002): functions are not data driven and number of factors sensitive to specified maximum.
  - Suggest two estimators: the "Eigenvalue Ratio" ($ER(k)$) and "Growth Rate" ($GR(k)$).
  - Both estimators take the ratio of residual variances when an additional common factor is added into account.

Estimation of Number of Common Factors

- Onatski (2010) Estimator ▸ details
  - Idea is to take differences between ordered eigenvalues into account.
  - Number of factors is estimated as the first difference which is larger than a threshold.
  - Allows for weak dependence across time and cross-sections and integrated factors.
- Gagliardini et al. (2019) Estimator ▸ details
  - Method focuses on residuals from linear models and can be used as a post estimation criterion.
  - Sequential method: first step checks if there are common factor structure, second step how many factors.
  - Difference between largest eigenvalue $\mu(k)$ and penalty used $g()$:

Cross-Section Dependence (CSD) I

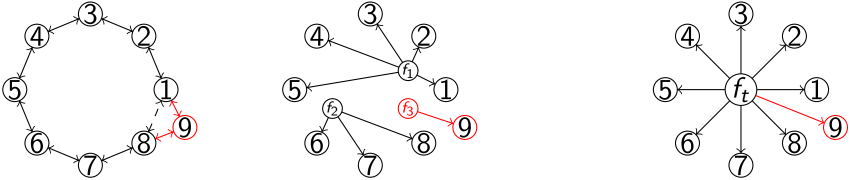
$$x_{i,t} = \gamma_{x,1,i} f_{1,t} + \gamma_{x,2,i} f_{2,t} + \xi_{i,t}$$
$$u_{i,t} = \gamma_{u,1,i} f_{1,t} + \gamma_{u,3,i} f_{3,t} + \epsilon_{i,t}$$

- Cross-section dependence occurs if the factor loadings are not equal to zero.
- It implies that all units are exposed to the same common factor (or shock).
- If it is not accounted for, then CSD potentially leads to:
  1. Omitted variable bias if $\gamma_{x,1,i} \neq 0$ and $\gamma_{u,1,i} \neq 0$
  2. Residuals can be correlated across units if $\gamma_{u,1,i} \neq 0$ and $\gamma_{u,3,i} \neq 0$
- If $\gamma_{x,1,i} = \gamma_{x,2,i} = 0$ then no first order problem for estimator.

Cross-Section Dependence (CSD) II

- Dependence is measured by constant α (Chudik et al., 2011)

$$\lim_{N\to\infty} N^{-\alpha} \sum_{i=1}^{N} |\gamma_{k,i,l}| = K < \infty$$



(Semi-) Weak CSD: $0 \leq \alpha < 0.5$

(Semi-) Strong CSD: $0.5 \leq \alpha \leq 1$

Weak and strong cross-sectional dependence with additional unit 9 as $N \to \infty$

- We can a) estimate the number of factors (**xtnumfac**), b) estimate the exponent of cross-section dependence (**xtcse2**) and c) test for weak cross-section dependence (**xtcd2**)

Estimating exponent of CSD

- Bailey et al. (2016, 2019) propose an estimator for the exponent of CSD.
- Estimation only possible for $\alpha > 1/2$.
- In Stata implemented by `xtcse2` (Ditzen, 2021).

Testing for weak cross-sectional dependence

CD Test

- Pesaran (2015, 2021) proposes a test for weak cross-section dependence, the CD-test:

  H_0 weak dependence vs. H_1 strong dependence
- Test statistic:

$$CD = \sqrt{\frac{2T}{N(N-1))}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho_{i,j},$$

  with $\rho_{i,j}$ is the correlation coefficient.
- Under the null hypothesis asymptotically: $CD \sim N(0,1)$.
- Problem of the CD test: tends to over reject.
- Further developments: CDw (Juodis and Reese, 2021), CDw with power enhancement (Fan et al., 2015) and CD* (Pesaran and Xie, 2021b)

Testing for weak cross-sectional dependence I

Recent Developments

- Weighted CD Test (Juodis and Reese, 2021)
  - Under certain conditions CD test diverges if applies to FE or CCE residuals.
  - Juodis and Reese (2021) propose to reweight the correlation coefficients: $\rho_{i,j} = \sum_{t=1}^{T} w_i \epsilon_{i,t} \epsilon_{j,t} w_j$
    $\epsilon_{i,t}$ are residuals and w_i are Rademacher weights.
  - Under the null hypothesis: $CD_w \xrightarrow{d} N(0,1)$.
- Power Enhanced CD_w test (Fan et al., 2015; Juodis and Reese, 2021)
  - CD_w can be underpowered if N is very large, bias corrected version by Juodis and Reese (2021) based on Fan et al. (2015):
    $CD_{w+} = CD_w + \sum_{i=2}^{N} \sum_{j=1}^{i-1} |\hat{\rho}_{i,j}| 1 \left(|\hat{\rho}_{i,j}| > 2\sqrt{ln(N)T} \right)$

Testing for weak cross-sectional dependence II

Recent Developments

- Bias Corrected CD Test (Pesaran and Xie, 2021a)
  - Alternative to CD_w to overcome over rejection.
    $CD^* = \frac{CD+\sqrt{\frac{T}{2}}\theta}{1-\theta}$
  - θ is the bias correction and a function of the estimated factor loadings via principal components.

xtcd2 and **xtnumfac** Syntax

xtcd2 (Ditzen, 2018, 2021, on SSC and GitHub[1])

```
xtcd2 [varlist][if][pesaran cdw pea cdstar rho pca(integer)
reps(integer) kdensity name(string) heatplot[(...)]
contour[(...)]  noadjust]
```

xtnumfac (Reese and Ditzen, on SSC)

```
xtnumfac [varlist] [if] [in] [, kmax(#) detail standardize(#)]
```

[1] To install in Stata: `net install xtdcce2 , from("https://janditzen.github.io/xtdcce2/")`

Introduction

- We want to estimate a simple Solow-style growth model.
- Data: Penn World Tables with 93 countries over years 1960 - 2007
- Dynamic model:

$$\begin{aligned}\Delta log\_rgdpo_{i,t} = \beta_{0,i} + \alpha_i log\_rgdpo_{i,t-1} + \beta_{1,i} log\_hc_{i,t} \\ + \beta_{2,i} log\_ck_{i,t} + \beta_{3,i} log\_ngd_{i,t} + u_{i,t}\end{aligned}$$

- Variables:
  - `log_rgdpo`: Real GDP per capita
  - `log_hc`: human capital
  - `log_ngd`: population growth rate, misses observations for first period
  - `log_ck`: capital stock
- Let's start with investigating the number of factors of the dependent variable

Empirical Application

Estimation of number of factors

```
. xtnumfac log_rgdpo
 N    =     4464                 T      =      48
 N_g =        93                 vars.  =       1
-------------------------------------------------
 IC        |  # factors   |  IC        |  # factors
-------------------------------------------------
 PC_{p1}   |      8       |  IC_{p1}   |     8
 PC_{p2}   |      8       |  IC_{p2}   |     8
 PC_{p3}   |      8       |  IC_{p3}   |     8
 ER        |      1       |  GR        |     1
 GOL       |      1       |  ED        |     4
-------------------------------------------------
8 factors maximally considered.
PC_{p1},...,IC_{p3} from Bai and Ng (2002)
ER, GR from Ahn and Horenstein (2013)
ED from Onatski (2010)
GOL from Gagliardini, Ossola, Scaillet (2019)
```

- Ignore estimate from GOL, criteria from Bai and Ng (2002) often overselects.
- Number of factors somewhere around 1-4.

Empirical Application

Estimation of number of factors

```
. xtnumfac log_rgdpo log_hc log_ck log_ngd, stand(3)
 N    =     4464                 T      =      48
 N_g =        93                 vars.  =       4
-------------------------------------------------
 IC        | # factors    | IC        | # factors
-------------------------------------------------
 PC_{p1}   |      8       | IC_{p1}   |    8
 PC_{p2}   |      8       | IC_{p2}   |    8
 PC_{p3}   |      8       | IC_{p3}   |    8
 ER        |      1       | GR        |    1
 GOL       |      1       | ED        |    3
-------------------------------------------------
8 factors maximally considered.
PC_{p1},...,IC_{p3} from Bai and Ng (2002)
ER, GR from Ahn and Horenstein (2013)
ED from Onatski (2010)
GOL from Gagliardini, Ossola, Scaillet (2019)

93 missing values imputed before estimating number of factors.
```

- All are variables added, standardized and fixed effects removed.
- *log_ngd* has missings in the first period, `xtnumfac` uses an EM algorithm to balance the panel (Stock and Watson, 1998; Bai et al., 2015).

Empirical Application

Testing for strong cross-sectional dependence

```
. set seed 09092022

. xtcd2 log_rgdpo log_hc log_ck log_ngd if log_ngd != .

Testing for weak cross-sectional dependence (CSD)
   H0: weak cross-section dependence
   H1: strong cross-section dependence
------------------------------------------------------------------------
               |      CD           CDw          CDw+          CD*
---------------+--------------------------------------------------------
log_rgdpo      |    139.22        -3.46     17774.47         4.35
               |   (0.000)      (0.001)      (0.000)      (0.000)
log_hc         |    423.05        -3.67     27666.79         2.22
               |   (0.000)      (0.000)      (0.000)      (0.027)
log_ck         |    412.21        -2.99     26963.84         9.83
               |   (0.000)      (0.003)      (0.000)      (0.000)
log_ngd        |     75.03         0.67     11228.57         3.30
               |   (0.000)      (0.501)      (0.000)      (0.001)
------------------------------------------------------------------------
p-values in parenthesis.
References
  CD:        Pesaran (2015, 2021)
  CDw:       Juodis, Reese (2021)
  CDw+:      CDw with power enhancement from Fan et. al. (2015)
  CD*:       Pesaran, Xie (2021) with 4 PC(s)
```

- We find strong cross-section dependence for all variables.
- The seed is set to ensure the CDw produces the same results.
- The CD* test does not support unbalanced panels, so the if statement is needed to drop all missing observations and create a balanced panel.

Empirical Application

Estimation and post-estimation

- Established: several common factors and strong cross-sectional dependence.
- Estimation method which accounts for strong cross-sectional dependence required, either CCE (Pesaran, 2006; Chudik and Pesaran, 2015) or PCA (Bai, 2009).
- Use `xtdcce2` (Ditzen, 2018, 2021) which implements the CCE estimator and then predict residuals:

```
. xtdcce2 d.log_rgdpo L.log_rgdpo log_hc log_ck log_ngd ///
>                    , cr(log_rgdpo log_hc log_ck log_ngd) cr_lags(3)
(output omitted)
. predict residuals, residuals
```

- (We should ensure that the number of cross-section averages does not exceed the common factors (Karabiyik et al., 2017; Juodis, 2022) and that the common factors are actually correlated with the regressors (Vos et al., 2022), but this is left for another talk....)

Empirical Application

Post-estimation - Number of Factors

```
. xtnumfac residuals if e(sample)
 N    =     4092                    T       =       44
 N_g =       93                    vars.   =        1
-------------------------------------------------
 IC        |  # factors   |  IC        |  # factors
-------------------------------------------------
 PC_{p1}   |      8       |  IC_{p1}   |      8
 PC_{p2}   |      8       |  IC_{p2}   |      8
 PC_{p3}   |      8       |  IC_{p3}   |      8
 ER        |      0       |  GR        |      4
 GOL       |      0       |  ED        |      0
-------------------------------------------------
8 factors maximally considered.
PC_{p1},...,IC_{p3} from Bai and Ng (2002)
ER, GR from Ahn and Horenstein (2013)
ED from Onatski (2010)
GOL from Gagliardini, Ossola, Scaillet (2019)
```

- The GOL indicates that all common factors are removed.

Empirical Application

Post-estimation - Test for Strong CSD

```
. xtcd2 residuals if e(sample),  cdw cdstar pca(1) reps(100)

Testing for weak cross-sectional dependence (CSD)
   H0: weak cross-section dependence
   H1: strong cross-section dependence
--------------------------------------------
               |     CDw           CD*
---------------+----------------------------
residuals      |    -1.55        -1.50
               |   (0.121)      (0.133)
--------------------------------------------

p-values in parenthesis.
References
  CDw:       Juodis, Reese (2021)
  CD*:       Pesaran, Xie (2021) with 1 PC(s)
```

- We only want to use the CD^* and CD_w test, specified by the options. The CD_w test is repeated 100 times with varying weights to ensure stable results.
- Only 1 PCA (minimum) used for CD^*.
- Both tests confirm, no strong cross-section dependence left.

Conclusion

or take aways

- Panel-Time series models offer a lot of flexibility and insights into data.
- The dependence structure, strong cross-section dependence and common factors play an important role for unbiased and consistent estimates.
- `xtnumfac`, `xtcd2` and `xtcse2` (not presented today) offer plenty of options to shed light on it.
- CCE and PCA estimator both offer ways to account for strong cross-section dependence.
- CCE easy to implement, but (!) developing literature on caveats of CCE and common factors.

References I

Ahn, S. C., and A. R. Horenstein. 2013. Eigenvalue Ratio Test for the Number of Factors. Econometrica 81(3): 1203–1227.

Bai, J. 2009. Panel Data Models With Interactive Fixed Effects. Econometrica 77(4): 1229–1279.

Bai, J., Y. Liao, and Jisheng Yang. 2015. Unbalanced Panel Data Models with Interactive Effects. In The Oxford Handbook of Panel Data, 149–170.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. Econometrica 70(1): 191–221.

Bailey, N., G. Kapetanios, and M. H. Pesaran. 2016. Exponent of Cross-Sectional Dependence: Estimation and Inference. Journal of Applied Econometrics 31: 929–960.

———. 2019. Exponent of Cross-sectional Dependence for Residuals. Sankhya B 81: 46–102.

References II

Chudik, A., and M. H. Pesaran. 2015. Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. Journal of Econometrics 188(2): 393–420.

Chudik, A., M. H. Pesaran, and E. Tosetti. 2011. Weak and strong cross-section dependence and estimation of large panels. The Econometrics Journal 14(1): C45–C90.

Ditzen, J. 2018. Estimating dynamic common-correlated effects in Stata. The Stata Journal 18(3): 585 – 617.

———. 2021. Estimating long run effects and the exponent of cross-sectional dependence: an update to xtdcce2. The Stata Journal 21(3): 687–707. URL https://ideas.repec.org/p/bzn/wpaper/bemps81.html.

Fan, J., Y. Liao, and J. Yao. 2015. Power Enhancement in High-Dimensional Cross-Section Tests. Econometrica 83(4): 1497–1541.

References III

Gagliardini, P., E. Ossola, and O. Scaillet. 2019. A diagnostic criterion for approximate factor structure. Journal of Econometrics 212(2): 503–521. URL `https://doi.org/10.1016/j.jeconom.2019.06.001.`

Juodis, A. 2022. A regularization approach to common correlated effects estimation. Journal of Applied Econometrics (October 2021): 788–810.

Juodis, A., and S. Reese. 2021. The Incidental Parameters Problem in Testing for Remaining Cross-Section Correlation. Journal of Business & Economic Statistics .

Karabiyik, H., S. Reese, and J. Westerlund. 2017. On the role of the rank condition in CCE estimation of factor-augmented panel regressions. Journal of Econometrics 197(1): 60–64. URL `http://dx.doi.org/10.1016/j.jeconom.2016.10.006.`

Karabiyik, H., J. Westerlund, and A. Juodis. 2020. On the Robustness of the Pooled CCE Estimator. Journal of Econometrics Forthcomin(xxxx).

References IV

Kripfganz, S., and V. Sarafidis. 2021. Instrumental-variable estimation of large-T panel-data models with common factors. Stata Journal 21(3): 659–686.

Norkute, M., V. Sarafidis, T. Yamagata, and G. Cui. 2021. Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure. Journal of Econometrics 220(2): 416–446.

Onatski, A. 2010. Determining the Number of Factors from Empirical Distribution of Eigenvalues. The Review ofEconomics and Statistics 92(4): 1004–1016.

Pesaran, M. H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74(4): 967–1012.

References V

———. 2015. Testing Weak Cross-Sectional Dependence in Large Panels. Econometric Reviews 34(6-10): 1089–1117.

———. 2021. General diagnostic tests for cross-sectional dependence in panels. Empirical Economics 60(1): 13–50. URL `https://doi.org/10.1007/s00181-020-01875-7`.

Pesaran, M. H., and Y. Xie. 2021a. A Bias-Corrected CD Test for Error Cross-Sectional Dependence in Panel Data Models with Latent Factors. SSRN Electronic Journal .

———. 2021b. A Bias-Corrected CD Test for Error Cross- Sectional Dependence in Panel Data Models with Latent Factors. Cambridge Working Papers in Economics 2158.

Reese, S., and J. Ditzen. A battery of estimators for the number of common factors in time series and panel data models. The Stata Journal .

References VI

Stock, J. H., and M. W. Watson. 1998. Diffusion Indexes.

Vos, D., I. D. Vos, and V. Sarafidis. 2022. A method for evaluating the rank condition for CCE estimators (112305).

Estimation of Number of Common Factors ▸ back

Information Criteria by Bai and Ng (2002)

- Idea: Penalise loss function $V()$ with a penalty $g()$ which increases with each additional factor:

  $PC(k) = V(k, F^k) - k\hat{\sigma}^2 g(N, T)$ and
  $IC(k) = ln(V(k, F^k)) - kg(N, T)$
- with $V(k, \hat{F}^k) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{i,t}^2$ and $\hat{e}_{i,t}$ are residuals of the variable of interest on the first k first principal components.
- Three different functions for $g(N, T)$: $\frac{N+T}{NT} ln\left(\frac{NT}{N+T}\right)$, $\frac{N+T}{NT} ln\left(\min(N, T)\right)$ and $\frac{\ln(\min(N,T))}{\min(N,T)}$.
- Estimated number of factors is the number k which minimizes the statistics.
- Method very popular and well established, however tends to overselect the number of common factors.

Estimation of Number of Common Factors ▸ back

Ahn and Horenstein (2013) Estimator

- Disadvantage of Bai and Ng (2002): functions are not data driven and number of factors sensitive to specified maximum.
- Suggest two estimators: the "Eigenvalue Ratio" ($ER(k)$) and "Growth Rate" ($GR(k)$).
- Both estimators take the ratio of residual variances when an additional common factor is added into account.
- The $ER(k)$ uses the kth largest eigenvalues of $X'X/(NT)$ while the $GR(k)$ is based on the mean of the squared residuals of a regression of X on the first k principal components of $X'X/(NT)$.
- Number of factors is selected as:

$$\tilde{k}_{ER} = max_{1 \leq k \leq k_{max}} ER(k)$$
$$\tilde{k}_{GR} = max_{1 \leq k \leq k_{max}} GR(k)$$

Estimation of Number of Common Factors ▸ back

Onatski (2010) Estimator

- Least intuitive method.
- Idea is to take differences between ordered eigenvalues into account.
- Number of factors is estimated as the first difference which is larger than a threshold.
- Estimation contains an iteration with varying thresholds.
- Allows for weak dependence across time and cross-sections and integrated factors.

Estimation of Number of Common Factors ▸ back

Gagliardini et al. (2019) Estimator

- Method focuses on residuals from linear models and can be used as a post estimation criterion.
- Sequential method: first step checks if there are common factor structure, second step how many factors.
- Difference between largest eigenvalue $\mu(k)$ and penalty used $g()$:

$$\xi(k) = \mu(k) - g(N, T)$$

$$g(N, T) = \frac{\left(\sqrt{N} + \sqrt{T}\right)^2}{NT} ln\left(\frac{NT}{\left(\sqrt{N} + \sqrt{T}\right)^2}\right)$$

- Number of factors are then:

$$\text{no factors} : \xi(1) < 0$$

$$\hat{k}\ \text{factors} : \hat{k} = min(0, ..., T - 1 : \xi(k) < 0)$$