AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

heck for updates

The Stata Journal (2023)
23, Number 2, pp. 438–454 DOI: 10.1177/1536867X231175305

xtnumfac: A battery of estimators for the number of common factors in time series and panel-data models

Jan Ditzen
Free University of Bozen-Bolzano
Bozen, Italy
jan.ditzen@unibz.it

Simon Reese
Lund University
Lund, Sweden
simon.reese@nek.lu.se

Abstract. In this article, we introduce a new community-contributed command, `xtnumfac`, for estimating the number of common factors in time-series and panel datasets using the methods of Bai and Ng (2002, *Econometrica* 70: 191–221), Ahn and Horenstein (2013, *Econometrica* 81: 1203–1227), Onatski (2010, *Review of Economics and Statistics* 92: 1004–1016), and Gagliardini, Ossola, and Scaillet (2019, *Journal of Econometrics* 212: 503–521). Common factors are usually unobserved or unobservable. In time series, they influence all predictors, while in panel-data models, they influence all cross-sectional units at different degrees. Examples are shocks from oil prices, inflation, or demand or supply shocks. Knowledge about the number of factors is key for multiple econometric estimation methods, such as Pesaran (2006, *Econometrica* 74: 967–1012), Bai (2009, *Econometrica* 77: 1229–1279), Norkute et al. (2021, *Journal of Econometrics* 220: 416–446), and Kripfganz and Sarafidis (2021, *Stata Journal* 21: 659–686). This article discusses a total of 10 methods to estimate the number of common factors. Examples based on Kapetanios, Pesaran, and Reese (2021, *Journal of Econometrics* 221: 510–541) show that U.S. house prices are exposed to up to 10 common factors. Therefore, when one fits models with house prices as a dependent variable, the number of factors must be considered.

Keywords: st0715, xtnumfac, common factors, factor models, cross-section dependence, panel-data models, time-series models

1 Introduction

The notion that one or more common factors drive economic or social variables is predominant in the empirical macroeconomics and finance literature. For example, Ross (1976) argues that many asset returns are influenced by a few common factors. Stock and Watson (1989) propose to use three indexes to model the co-movements of several macroeconomic variables. Common factors are also a recurring phenomenon in modeling house prices (Holly, Pesaran, and Yamagata 2010, 2011), economic growth (Eberhardt, Helmers, and Strauss 2013; Ditzen 2018b; Chudik et al. 2017), and the effect of oil price shocks (El-Anshasy, Mohaddes, and Nugent 2017). In such models, common factors are often interpreted as shocks that affect all countries at the same point in time but with different magnitude.

© 2023 StataCorp LLC

An understanding of the number of factors is important for the practitioner. Common factors are modeled by principal components in macroeconomic time series. Knowing their number is key for forecasting (Stock and Watson 2002), break point estimation (Duan, Bai, and Han 2023), or in general for asset pricing models, where the number of common factors is used to verify whether all systematic determinants in the cross-section of returns are accounted for. Models with common factors are popular in panel data (Sul 2019). If not accounted for, common factors can bias regression results. Methods such as the common correlated effects (CCE) estimator (Pesaran 2006) do not require exact knowledge about the number of factors. However, the number of common factors has to be less than the number of observables used to approximate them (Karabiyik, Reese, and Westerlund 2017) to ensure the asymptotic theory of the estimator is valid. Other methods, such as the principal component-based estimators (Bai 2009) or the instrumental-variables estimator for large panel-data models (Norkute et al. 2021; Kripfganz and Sarafidis 2021), require knowledge about the number of factors. The number of factors is also necessary in the estimation of the exponent of cross-section dependence (Bailey, Kapetanios, and Pesaran 2016, 2019).

This article introduces the community-contributed command `xtnumfac`, which allows obtaining 10 different estimators for the number of common factors. The six information criteria by Bai and Ng (2002) are implemented, as well as the eigenvalue-based estimators of Ahn and Horenstein (2013), Onatski (2010), and Gagliardini, Ossola, and Scaillet (2019). The latter is particularly designed for the estimation of the number of common factors of residuals. All estimators are designed for high-dimensional time-series models or panel-data models with many observations over time and cross-section dimensions.

The community-contributed `baing` command (Núñez and Otero 2020) allows the estimation of the number of common factors following the methods in Bai and Ng (2002) of variables in wide format. That is, each cross-section is represented by a variable. `xtnumfac` offers more functionality. First, it includes the methods by Ahn and Horenstein (2013), Onatski (2010), and Gagliardini, Ossola, and Scaillet (2019), in addition to those by Bai and Ng. Second, `xtnumfac` can be directly applied to panel data without reshaping the data into wide format. Finally, unbalanced panels are supported using an expectation-maximization algorithm, as proposed in Stock and Watson (1998) and Bai, Liao, and Yang (2015).

The next section specifies the factor-model framework used in all four articles mentioned above and reviews all estimators that are implemented by `xtnumfac` from a theoretical perspective. Section 3 provides an overview of the command `xtnumfac`. The article closes with examples drawn from Kapetanios, Pesaran, and Reese (2021).

2 Econometric theory

The model to which estimators for the number of factors apply is the approximate factor model of Chamberlain and Rothschild (1983), which we can formally write as

$$\mathbf{X}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{e}_t, \qquad t = 1, 2, \ldots, T \tag{1}$$

Here $\mathbf{X}_t$ is a vector containing observations over N cross-sections at time t, $\mathbf{F}_t$ is an $r \times 1$ vector of common factors, $\mathbf{\Lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_N)'$ is an $N \times r$ matrix containing the corresponding factor loadings, and $\mathbf{e}_t = (e_{1t}, e_{2t}, \ldots, e_{Nt})'$ is a random noise component. None of the latter three components is observed. In appendix A, we provide a comprehensible description of the assumptions typically made on factors, loadings, and random noise to specify which properties they must have. r is the number of common factors, unknown to the researcher and also the main parameter of interest in this article. Equation (1) is often embedded in the following model:

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{X}_t\boldsymbol{\beta} + \mathbf{u}_t \\ \mathbf{u}_t &= \mathbf{\Gamma}\mathbf{F}_t + \mathbf{v}_t \end{aligned} \tag{2}$$

Here $\mathbf{Y}_t$ is the variable of interest, and $\mathbf{X}_t$ is a matrix with k explanatory variables and N observations over the cross-section dimension. $\mathbf{u}_t$ is a multifactor error term and consists of r common factors $\mathbf{F}_t$, an $N \times r$ matrix of associated loadings $\mathbf{\Gamma}$, and a random noise component $\mathbf{v}_t$. In time series and panel data, it is of interest to estimate the common factors in $\mathbf{Y}$ and $\mathbf{X}$. Panel model (2) collapses to a time-series model if $N = 1$. In this framework, common factors are often used to predict an individual macroeconomic time series $\mathbf{Y}_t$. In panel data, the parameter of interest is mostly $\boldsymbol{\beta}$. If the common factors in $\mathbf{X}_t$ and the composite error $\mathbf{u}_t$ are not accounted for, regressing $\mathbf{X}_t$ on $\mathbf{Y}_t$ can lead to a biased estimate of $\boldsymbol{\beta}$. It is popular to approximate the common factors by either principal components (Bai 2009) or cross-section averages (Pesaran 2006).[1] The principal component approach requires exact knowledge about the number of common factors for a consistent and efficient estimation of $\boldsymbol{\beta}$. The CCE estimator in Pesaran hinges on the rank condition of the common factors, which implies that the number of common factors should not exceed the number of cross-section averages (Karabiyik, Reese, and Westerlund 2017).

2.1 Bai and Ng's (2002) information criteria

Bai and Ng (2002) suggest estimating the number of factors using an information criterion where a loss function is penalized by a strictly increasing function of the number of cross-sections N and time periods T. In particular, Bai and Ng propose two different types of criteria: the panel criterion (PC) and the panel information criterion (IC). Formally, they are given by

$$\text{PC}(k) = V(k, F^k) - k\widehat{\sigma}^2 g(N, T) \tag{3}$$

$$\text{IC}(k) = \ln\left\{V\left(k, F^k\right)\right\} - kg(N, T) \tag{4}$$

1. The two estimation methods are implemented in Stata by `regife` (Gomez 2015) and `xtdcce2` (Ditzen 2018a, 2021).

Here $V(k, F^k)$ is further specified to be squared loss. That is, we have

$$V(k, F^k) = \min_{\Lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{i,t} - \boldsymbol{\lambda}_{k,i} \mathbf{F}_t^k\right)^2$$

where $\boldsymbol{\lambda}_{k,i}$ are the factor loadings with k factors and $\mathbf{F}_t^k$ are the common factors. The loss function depends on an estimate of the factor loadings and requires knowledge about the factors. If the factors are selected out of a set of known factors, one could compare all possible solutions for $\text{PC}(\widehat{k})$ and $\text{IC}(\widehat{k})$. In the more realistic case of unknown common factors, the factors can be estimated by principal components (Forni and Lippi 1997; Forni and Reichlein 1998). Thus, F_t is replaced by $\widehat{F}_t$. In this case, $V(k, \widehat{F}^k) = 1/N \sum_{i=1}^{N} \widehat{\sigma}_{i,k}^2 = 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{e}_{i,t,k}^2$, where $\widehat{e}_{i,t,k}$ is the residual from a regression of the k first principal components on $\mathbf{X}$.[2] The estimate of $\sigma^2 = N^{-1} \sum_{i=1}^{N} \sigma_i^2$ that explicitly enters the penalty term of $\text{PC}(k)$ is usually obtained from the model with the maximum number of factors considered, $k_{\max}$. This ensures an unbiased estimator of $E\left(e_{i,t}^2\right)$ and ensures that all components of the penalty, except the number of factors k itself, are unaffected by the value of k.

The correction applied to loss $V(k, \widehat{\mathbf{F}}^k)$ or its logarithm ensures consistent estimation of the number of factors under general conditions on the penalty function $g(N, T)$ (see Bai and Ng [2002, theorem 2]). For a principal components-based estimator of the unobserved factors, Bai and Ng make three specific suggestions for $g(N, T)$. Together with the two general types of criteria (3) and (4), this results in the following six statistics:

$$\begin{aligned}
\text{PC}_{p1} &= V\left(k, \widehat{F}^k\right) + k\widehat{\sigma}_{k_{\max}}^2 \frac{N+T}{NT} \ln\left(\frac{NT}{N+T}\right) \\
\text{PC}_{p2} &= V\left(k, \widehat{F}^k\right) + k\widehat{\sigma}_{k_{\max}}^2 \frac{N+T}{NT} \ln\{\min(N, T)\} \\
\text{PC}_{p3} &= V\left(k, \widehat{F}^k\right) + k\widehat{\sigma}_{k_{\max}}^2 \frac{\ln\{\min(N, T)\}}{\min(N, T)} \\
\text{IC}_{p1} &= \ln\left\{V\left(k, \widehat{F}^k\right)\right\} + k\frac{N+T}{NT} \ln\left(\frac{NT}{N+T}\right) \\
\text{IC}_{p2} &= \ln\left\{V\left(k, \widehat{F}^k\right)\right\} + k\frac{N+T}{NT} \ln\{\min(N, T)\} \\
\text{IC}_{p3} &= \ln\left\{V\left(k, \widehat{F}^k\right)\right\} + k\frac{\ln\{\min(N, T)\}}{\min(N, T)}
\end{aligned}$$

For any of the six statistics above, estimating the number of factors requires obtaining the values of the chosen statistics within a user-specified range of factors $k \in \{1, 2, \ldots, k_{\max}\}$. The estimated number of factors is then the value of k that minimizes the chosen statistic.

2. Equivalently, $V(k, \widehat{F}^k)$ is equal to the sum of all but the first k eigenvalues of $\mathbf{X}'\mathbf{X}/(NT)$. This result allows representing the information criteria of Bai and Ng (2002) in a formal framework similar to those of the three articles referred to in the following three sections.

2.2 Ahn and Horenstein's (2013) estimators

Ahn and Horenstein (2013) point out two disadvantages of the approach by Bai and Ng (2002). First, the criteria functions are prespecified and not data driven. Second, the maximum number of possible factors, $k_{\max}$, has to be set ex ante, and the estimation method should not be sensitive to the choice of $k_{\max}$. To overcome these shortcomings, Ahn and Horenstein suggest two alternative estimators, the eigenvalue ratio (ER) and growth rate (GR) estimators. Both estimators account for the ratio of residual variances when an additional common factor is added. In detail, the estimators for a given number of common factors k are

$$\mathrm{ER}(k) = \frac{\widetilde{\mu}_{NT,k}}{\widetilde{\mu}_{NT,k+1}}, \quad k = 1, 2, \ldots, k_{\max}$$

$$\mathrm{GR}(k) = \frac{\ln\left(1 + \widetilde{\mu}^*_{NT,k}\right)}{\ln\left(1 + \widetilde{\mu}^*_{NT,k+1}\right)}$$

with $\widetilde{\mu}_{NT,k}$ being the kth largest eigenvalue of $\mathbf{X}'\mathbf{X}/(NT)$ and $\widetilde{\mu}^*_{NT,k} = \widetilde{\mu}_{NT,k}/V(k)$, where $V(k)$ is the mean of the squared residuals of a regression of $\mathbf{X}$ on the first k principal components of $\mathbf{X}'\mathbf{X}/(NT)$. The number of common factors is then

$$\widetilde{k}_{\mathrm{ER}} = \max_{1 \le k \le k_{\max}} \mathrm{ER}(k)$$

$$\widetilde{k}_{\mathrm{GR}} = \max_{1 \le k \le k_{\max}} \mathrm{GR}(k)$$

2.3 Onatski's (2010) estimator

Onatski (2010) exploits differences in the properties of the first r eigenvalues and a limited number of subsequent eigenvalues of $\mathbf{X}'\mathbf{X}/T$. The author emphasizes that the eigenvalues with order index slightly larger than k tend to cluster around a particular finite value. The differences between subsequent eigenvalues in this cluster constitute a threshold δ that differences between eigenvalues representing factors in the data should surpass. Accordingly, the magnitude by which each of the first $k_{\max}$ eigenvalues of $\mathbf{X}'\mathbf{X}/T$ exceeds the next smaller eigenvalue is investigated. The smallest eigenvalue with more than a δ distance to its next smaller successor indicates the true number of factors. The specific algorithm suggested by Onatski (p. 1008) is as follows:

1. Obtain the eigenvalues $\mu_1, \mu_2, \ldots, \mu_N$ of $T^{-1}\sum_{t=1}^{T} \mathbf{X}_t\mathbf{X}_t'$. Set $j = k_{\max} + 1$.
2. Estimate the slope coefficient $\widehat{\beta}$ in a linear regression of $\mu_j, \ldots, \mu_{j+4}$ on $(j-1)^{2/3}$, $\ldots, (j+3)^{2/3}$ and a constant, and set the threshold δ to $\delta = 2|\widehat{\beta}|$.
3. Obtain the estimated number of factors as $\widehat{r}(\delta) = \max(i \le r^N_{\max} : \mu_i - \mu_{i+1} \ge \delta)$ if there exists some $i \le r^N_{\max}$ subject to $\mu_i - \mu_{i+1} \ge \delta$. Otherwise, set $\widehat{r}(\delta) = 0$.
4. Set $j = \widehat{r}\,(\delta)$. Repeat steps 2 and 3 until $\widehat{r}(\delta)$ remains unchanged in two subsequent repetitions. Let this final estimator be denoted $\widetilde{r}_{\mathrm{ED}}$.

Without step 4 in the algorithm above, the estimator of Onatski would merely amount to going through a list of sorted eigenvalues and picking the smallest such eigenvalue with sufficient distance to its next smaller neighbor. However, the iterations of step 4 update the threshold value for a sufficient distance and hence make the entire procedure somewhat less intuitive.

One disadvantage with Onatski's (2010) edge distribution (ED) estimator is that the error component may be only weakly correlated over time or over cross-sections but not over both dimensions of the panel dataset. Only under the additional assumption of normally distributed noise is it possible to allow for weak dependence over both time and cross-sections. However, an interesting advantage of the estimator is that the theoretical framework, under which its consistency is proven, allows for factors that have an integration order of 1.

2.4 Gagliardini, Ossola, and Scaillet's (2019) estimator

The method of Gagliardini, Ossola, and Scaillet (2019) focuses on residuals from linear models and can be used as a postestimation criterion. It is based on two rival models, one with a weakly cross-section dependence structure and the second with a factor structure. To establish the first model, one calculates the difference between the largest eigenvalue and a penalty depending on N and T. If the penalty is larger than 0, at least one factor exists. The selection rule picks the number of factors. The number equals the kth largest eigenvalue for which the difference between the eigenvalue and the penalty is negative for the first time.

The difference between the kth largest eigenvalue, $\mu(k)$, and the penalty $g(N,T)$ is

$$\begin{aligned}\xi(k) &= \mu(k) - g(N,T)\\ g(N,T) &= \frac{\left(\sqrt{N}+\sqrt{T}\right)^2}{NT}\ln\left\{\frac{NT}{\left(\sqrt{N}+\sqrt{T}\right)^2}\right\}\end{aligned}$$

The number of factors is then selected according to

$$\begin{aligned}&\text{no factors} : \xi(1) < 0\\ &\widehat{k}\text{ factors} : \widehat{k} = \min\{0,\dots,T-1 : \xi(k) < 0\}\end{aligned}$$

If $\xi(k_i) > 0 \;\forall\; k_i = 0,\dots,T-1$, then $\widehat{k} = T$.

The method relies on standardized variables or residuals to ensure that the eigenvalues are measured with a common scale. Gagliardini, Ossola, and Scaillet (2019) show that the method consistently estimates the number of factors if the cross-section dimension is comparable or much larger than the time dimension, that is, $N = O(T^{1/\underline{\gamma}})$ and $T = O(N^{\overline{\gamma}})$ with $\underline{\gamma} > 0$ and $\overline{\gamma} \in (0,1]$.

The estimator of Gagliardini, Ossola, and Scaillet (2019) is proven to be consistent under considerably stricter assumptions than all other estimators. However, this may

be due to the authors' explicitly allowing for missing data. By contrast, the theoretical properties of all other estimators discussed above rely on the availability of a balanced panel dataset.

3 The xtnumfac command

3.1 Syntax

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

Data must be `tsset` or `xtset` (see [TS] **tsset** or [XT] **xtset**) before using `xtnumfac`. *varlist* may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

3.2 Options

`kmax(#)` specifies how many factors to consider at most when estimating its true number. The default is `kmax(8)`. The choice of `kmax()` mostly affects the length of the reported table of results. Additionally, the values of the `PC_{p1}`, `PC_{p2}`, and `PC_{p3}` statistics may be slightly affected because they are functions of an estimated error variance in the idiosyncratic component that is obtained from the most general model (that is, the one with `kmax()` factors).

`detail` reports exact values for the IC, PC, ER, GR, and GOS for every possible number of factors. The estimator of Onatski (2010) is left out from this representation to avoid confusion about how the estimated number of factors is obtained from a list of potential values.

`standardize(#)` specifies how to transform variables prior to factor estimation. Standardization is important when using the criterion in Gagliardini, Ossola, and Scaillet (2019) because it requires standardized data. The default is `standardize(1)`. # may be one of the following:

#	Description
1	No transformations
2	Remove individual fixed effects
3	Remove individual fixed effects, and standardize variance of each cross-section to 1
4	Remove individual and time fixed effects
5	Remove individual and time fixed effects, and standardize variance of each cross-section to 1

3.3 Stored results

`xtnumfac` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(N_g)`	number of cross-sections
`e(T)`	number of time periods
`e(kmax)`	maximum number of factors considered, $k_{\max}$
`e(missnum)`	number of missing values that are imputed

Matrices

`e(best_numfac)`	a (1×10) matrix containing the number of factors estimated by any of the 10 criteria; the order is as in the reported function output
`e(allICs)`	a $(k_{\max} \times 10)$ matrix containing the value of all measures for all numbers of factors under consideration; corresponds to the values in the reported function output (albeit without asterisks)

3.4 Unbalanced panel data

Estimation of the number of common factors requires a balanced dataset. In practice, many datasets are unbalanced, and selecting a balanced subset might be either impossible or undesirable. To accommodate practitioners' needs, `xtnumfac` imputes data in the case of unbalanced datasets using an expectation-maximization algorithm as proposed in Stock and Watson (1998); Bai, Liao, and Yang (2015); and Kripfganz and Sarafidis (2021). This algorithm is itself based on repeated estimation of a factor model. The number of factors chosen here is set to $k_{\max}+5$ to avoid interference with the estimation of the number of factors in the range $\{1, 2, \ldots, k_{\max}\}$.

4 Examples

As an example, we use quarterly house price changes for 48 mainland states of the United States over the years 1975 to 2014 from Kapetanios, Pesaran, and Reese (2021). Changes in house prices in the United States are well known to be exposed to common factors; see, for example, Holly, Pesaran, and Yamagata (2010) and Bailey, Holly, and Pesaran (2016a). The variable of interest is `d_lrhp`, which is the rate of change of real house prices after seasonal adjustment and nominal price deflation. We search for up to 10 common factors by setting `kmax(10)`.

The criteria from Bai and Ng (2002) are labeled as PC_1 to IC_3. The two criteria from Ahn and Horenstein (2013) are displayed as ER and GR; the one from Gagliardini, Ossola, and Scaillet (2019) as GOS; and the criterion from Onatski (2010) as ED. The results are the following:

```
. use kpr2021_hpdata
. xtnumfac d_lrhp, kmax(10)
 N   =     7632                T      =      159
 N_g =       48                vars.  =        1
```

IC	# factors	IC	# factors
PC_{p1}	10	IC_{p1}	8
PC_{p2}	9	IC_{p2}	8
PC_{p3}	10	IC_{p3}	10
ER	1	GR	1
GOS	0	ED	3

```
10 factors maximally considered.
PC_{p1},...,IC_{p3} from Bai and Ng (2002)
ER, GR from Ahn and Horenstein (2013)
ED from Onatski (2010)
GOS from Gagliardini, Ossola, Scaillet (2019)
```

The criteria of Bai and Ng (2002) find between 8 and 10 common factors, while the ratio tests by Ahn and Horenstein (2013) find 1 common factor. The estimator by Onatski (2010) estimates 3 common factors. The estimator by Gagliardini, Ossola, and Scaillet (2019) finds no factors. However, in the default setting, the data are not transformed. The GOS estimator requires standardized data and thus leads to invalid results. Onatski suggests in the empirical application to standardize the data as well. To remove potential state individual effects and standardize the variance of each cross-section to unity, we specify `standardize(3)`:

```
. xtnumfac d_lrhp, kmax(10) standardize(3)
 N   =     7632                T      =      159
 N_g =       48                vars.  =        1
```

IC	# factors	IC	# factors
PC_{p1}	10	IC_{p1}	9
PC_{p2}	9	IC_{p2}	5
PC_{p3}	10	IC_{p3}	10
ER	1	GR	1
GOS	1	ED	2

```
10 factors maximally considered.
PC_{p1},...,IC_{p3} from Bai and Ng (2002)
ER, GR from Ahn and Horenstein (2013)
ED from Onatski (2010)
GOS from Gagliardini, Ossola, Scaillet (2019)
```

Once again, the criteria by Ahn and Horenstein (2013) point to 1 common factor that drives changes in house prices. The result is in line with the estimate from the Gagliardini, Ossola, and Scaillet (2019) estimator. The estimator by Onatski (2010) now finds 2 common factors. The information criteria from Bai and Ng (2002) vary between 5 to 10 common factors. A reason for the difference is that the criteria by information criteria rely on a penalty, while the estimators by Onatski, Ahn and Horenstein, and Gagliardini, Ossola, and Scaillet rely on functions on the eigenvalues.

The economic implication of the findings is that house prices in the United States are exposed to at least one common factor. The common factor can be observed in things such as changes in the interest rate, shocks to labor markets, or unobserved nationwide factors. If house prices are estimated, the factor structure needs to be accounted for. If the principal component analysis estimator in Bai (2009) is used, the number of principal components equals the number of common factors for a consistent and efficient estimation; see Bai (2009, remark 4). If instead the CCE estimator (Pesaran 2006) is used, then the number of common factors should not exceed the number of cross-section averages (Karabiyik, Reese, and Westerlund 2017).

As a final exercise, we want to display the values of IC, PC, ER, GR, and GOS for every possible number of factors. We do so by using the `detail` option:

```
. xtnumfac d_lrhp, kmax(10) standardize(3) detail
Statistics for number of common factors in d_lrhp
   Number of obs        =     7632              Obs per group       =      159
   Number of groups     =       48              Number of variables =        1
```

# factors	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}
0	1.000	1.000	1.000	0.000	0.000	0.000
1	0.497	0.498	0.494	-0.632	-0.625	-0.649
2	0.406	0.409	0.401	-0.783	-0.768	-0.817
3	0.373	0.377	0.365	-0.823	-0.801	-0.874
4	0.352	0.357	0.342	-0.842	-0.813	-0.910
5	0.333	0.338	0.319	-0.871	-0.836*	-0.957
6	0.324	0.330	0.308	-0.874	-0.831	-0.977
7	0.316	0.324	0.297	-0.879	-0.829	-0.999
8	0.312	0.321	0.290	-0.880	-0.823	-1.018
9	0.308	0.319*	0.284	-0.884*	-0.820	-1.039
10	0.308*	0.320	0.282*	-0.880	-0.808	-1.052*

# factors	ER	GR	GOS
0	0.499	0.315	0.368
1	4.891*	2.940*	-0.044*
2	2.190	1.802	-0.101
3	1.341	1.181	-0.114
4	1.033	0.915	-0.115
5	1.429	1.275	-0.125
6	1.073	0.969	-0.127
7	1.151	1.040	-0.130
8	1.078	0.975	-0.131
9	1.202	1.090	-0.135
10	1.139	1.039	-0.136

```
10 factors maximally considered.
PC_{p1},...,IC_{p3} from Bai and Ng (2002)
ER, GR from Ahn and Horenstein (2013)
ED from Onatski (2010)
GOS from Gagliardini, Ossola, Scaillet (2019)
```

Because the number of factors is directly estimated in Onatski (2010), the ED criterion is not displayed. The column GOS, denoting the Gagliardini, Ossola, and Scaillet (2019) estimator, shows the difference between the eigenvalue and the penalty. We note that the first difference is slightly above 0, indicating that the factor structure identified by the estimator is very weak.

5 Conclusions

Understanding the number of common factors is important for empirical analysis. We introduced a community-contributed command called `xtnumfac`, which implements a total of 10 estimation methods of the number of factors. The methods are based on

Bai and Ng (2002), Ahn and Horenstein (2013), Onatski (2010), and, for residuals, Gagliardini, Ossola, and Scaillet (2019). We discussed the use of `xtnumfac` with an empirical example and showed that house prices in the United States are exposed to multiple common factors. `xtnumfac` can be applied to balanced and unbalanced panel data and to a single or multiple variables.

The methods presented here and, thus, `xtnumfac` are limited to static factor models. Accordingly, none of the methods presented in this article is capable of estimating the number of factors in a dynamic factor model while accounting for their dynamic impact on the observed data. Only the number of factors in a static representation of the dynamic model can be estimated. However, this identifies the contemporaneous and lagged impact of a single factor erroneously as several individual factors.

Furthermore, `xtnumfac` is limited to presenting a smorgasbord of different estimators for the number of factors. The choice of selecting an appropriate criterion is left to the researcher. For example, while the criterion in Bai and Ng (2002) can be applied to a wide range of models, the criterion in Gagliardini, Ossola, and Scaillet (2019) is particularly designed for residuals.

Lastly, in a panel-data setting, dominant units or units with pervasive effects can appear. Such units influence all other units and mimic common factors. While the literature (Kapetanios, Pesaran, and Reese 2021; Brownlees and Mesters 2021) identifying those units is closely related to the methods presented here, `xtnumfac` cannot distinguish between common factors and such units. Such differentiation is left to the researcher.

6 Acknowledgments

We are grateful for comments from an anonymous referee, Stephen Jenkins, and Sebastian Kripfganz. Jan Ditzen acknowledges financial support from the Italian Ministry of Education, University and Research under the PRIN project Hi-Di NET—Econometric Analysis of High Dimensional Models with Network Structures in Macroeconomics and Finance (grant 2017TA7TYC).

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 23-2
. net install st0715     (to install program files, if available)
. net get st0715         (to install ancillary files, if available)
```

8 References

Ahn, S. C., and A. R. Horenstein. 2013. Eigenvalue ratio test for the number of factors. *Econometrica* 81: 1203–1227. https://doi.org/10.3982/ECTA8968.

Bai, J. 2009. Panel data models with interactive fixed effects. *Econometrica* 77: 1229–1279. https://doi.org/10.3982/ECTA6135.

Bai, J., Y. Liao, and J. Yang. 2015. Unbalanced panel data models with interactive effects. In *The Oxford Handbook of Panel Data*, ed. B. H. Baltagi, 149–170. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199940042.013.0005.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221. https://doi.org/10.1111/1468-0262.00273.

Bailey, N., S. Holly, and M. H. Pesaran. 2016a. A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence. *Journal of Applied Econometrics* 31: 249–280. https://doi.org/10.1002/jae.2468.

Bailey, N., G. Kapetanios, and M. H. Pesaran. 2016b. Exponent of cross-sectional dependence: Estimation and inference. *Journal of Applied Econometrics* 31: 929–960. https://doi.org/10.1002/jae.2476.

———. 2019. Exponent of cross-sectional dependence for residuals. *Sankhyā* 81: 46–102. https://doi.org/10.1007/s13571-019-00196-9.

Brownlees, C., and G. Mesters. 2021. Detecting granular time series in large panels. *Journal of Econometrics* 220: 544–561. https://doi.org/10.1016/j.jeconom.2020.04.013.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* 51: 1281–1304. https://doi.org/10.2307/1912275.

Chudik, A., K. Mohaddes, M. H. Pesaran, and M. Raissi. 2017. Is there a debt-threshold effect on output growth? *Review of Economics and Statistics* 99: 135–150. https://doi.org/10.1162/REST_a_00593.

Chudik, A., and M. H. Pesaran. 2015. Large panel data models with cross-sectional dependence: A survey. In *The Oxford Handbook of Panel Data*, ed. B. H. Baltagi, 3–45. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199940042.013.0001.

Ditzen, J. 2018a. Estimating dynamic common-correlated effects in Stata. *Stata Journal* 18: 585–617. https://doi.org/10.1177/1536867X1801800306.

———. 2018b. Cross-country convergence in a general Lotka–Volterra model. *Spatial Economic Analysis* 2: 191–211. https://doi.org/10.1080/17421772.2018.1397285.

———. 2021. Estimating long run effects and the exponent of cross-sectional dependence: An update to xtdcce2. *Stata Journal* 21: 687–707. https://doi.org/10.1177/1536867X211045560.

Duan, J., J. Bai, and X. Han. 2023. Quasi-maximum likelihood estimation of break point in high-dimensional factor models. *Journal of Econometrics* 233: 209–236. https://doi.org/10.1016/j.jeconom.2021.12.011.

Eberhardt, M., C. Helmers, and H. Strauss. 2013. Do spillovers matter when estimating private returns to R&D? *Review of Economics and Statistics* 95: 436–448. https://doi.org/10.1162/REST_a_00272.

El-Anshasy, A., K. Mohaddes, and J. B. Nugent. 2017. Oil, volatility and institutions: Cross-country evidence from major oil producers. Federal Reserve Bank of Dallas, Globalization and Monetary Policy Institute Working Paper No. 310. https://doi.org/10.24149/gwp310.

Forni, M., and M. Lippi. 1997. *Aggregation and the Microfoundations of Dynamic Macroeconomics*. Oxford: Oxford University Press.

Forni, M., and L. Reichlein. 1998. Let's get real: A factor-analytic approach to disaggregated business cycle dynamics. *Review of Economic Studies* 65: 453–473. https://doi.org/10.1111/1467-937X.00053.

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

Gomez, M. 2015. regife: Stata module to estimate linear models with interactive fixed effects. Statistical Software Components S458042, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458042.html.

Holly, S., M. H. Pesaran, and T. Yamagata. 2010. A spatio-temporal model of house prices in the USA. *Journal of Econometrics* 158: 160–173. https://doi.org/10.1016/j.jeconom.2010.03.040.

———. 2011. The spatial and temporal diffusion of house prices in the UK. *Journal of Urban Economics* 69: 2–23. https://doi.org/10.1016/j.jue.2010.08.002.

Kapetanios, G., M. H. Pesaran, and S. Reese. 2021. Detection of units with pervasive effects in large panel data models. *Journal of Econometrics* 221: 510–541. https://doi.org/10.1016/j.jeconom.2020.05.001.

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: 60–64. https://doi.org/10.1016/j.jeconom.2016.10.006.

Kripfganz, S., and V. Sarafidis. 2021. Instrumental-variable estimation of large-T panel-data models with common factors. *Stata Journal* 21: 659–686. https://doi.org/10.1177/1536867X211045558.

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: 416–446. https://doi.org/10.1016/j.jeconom.2020.04.008.

Núñez, H. M., and J. Otero. 2020. baing: Stata module to determine and estimate the number of common factors following Bai and Ng. Statistical Software Components S458851, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458851.html.

Onatski, A. 2010. Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics* 92: 1004–1016. https://doi.org/10.1162/REST_a_00043.

Pesaran, M. H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74: 967–1012. https://doi.org/10.1111/j.1468-0262.2006.00692.x.

Ross, S. A. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 341–360. https://doi.org/10.1016/0022-0531(76)90046-6.

Stock, J. H., and M. W. Watson. 1989. New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual* 4: 351–394. https://doi.org/10.1086/654119.

———. 1998. Diffusion indexes. NBER Working Paper No. 6702, The National Bureau of Economic Research. https://doi.org/10.3386/w6702.

———. 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20: 147–162. https://doi.org/10.1198/073500102317351921.

Sul, D. 2019. *Panel Data Econometrics: Common Factor Analysis for Empirical Researchers*. Abingdon, U.K.: Routledge.

About the authors

Jan Ditzen is an assistant professor at the Free University of Bozen—Bolzano, Italy. His research interests are in the field of applied econometrics and spatial econometrics, particularly cross-sectional dependence in large panels.

Simon Reese is an associate senior lecturer in economics at Lund University, Sweden. His primary areas of research are panel-data econometrics, econometric theory, and macroeconometrics.

A Assumptions on the approximate factor model (1)

`xtnumfac` implements methods from four different scientific articles whose theoretical properties have been proven under equally many different sets of assumptions. In particular, assumptions are often stated in terms of conditions on matrix properties such

as the largest or smallest singular value. However, most econometricians have only elementary training in matrix analysis, which makes an interpretation of the assumptions mentioned above difficult. Thus, we specify a single set of assumptions that follows more closely the modeling conventions in econometrics. This set of assumptions also serves as a least common denominator for properties of the data-generating process under which consistency of all 10 estimators implemented by `xtnumfac` has been shown.

Assumption 1. *For some finite fixed number M, $\|\mathbf{F}_t\| \leq M$ holds almost surely. Additionally, $T^{-1}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t' \xrightarrow{p} \mathbf{\Sigma}_F$, where $\mathbf{\Sigma}_F$ is a deterministic, positive definite $r \times r$ matrix, and there exist $\eta, \overline{\eta} \in (0,1]$ and $C_1, C_2, C_3, C_4 > 0$ such that*

$$\Pr\left\{\left\|T^{-1}\sum_{t=1}^{T}\left(\mathbf{F}_t\mathbf{F}_t' - \mathbf{\Sigma}_F\right)\right\| > \delta\right\} \leq C_1 T \exp\left(-C_2\delta^2 T^{\eta}\right) + C_3\delta^{-1}\exp\left(C_4 T^{\overline{\eta}}\right)$$

Assumption 2. *For every $i = 1, 2, \ldots, N$, $\boldsymbol{\lambda}_i$ is deterministic and satisfies $\|\boldsymbol{\lambda}_i\| < M$. Furthermore, $N^{-1}\sum_{i=1}^{T}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i' \to \mathbf{\Sigma}_\Lambda$, where $\mathbf{\Sigma}_\Lambda$ is positive definite.*

Assumption 3.

1. *The $T \times N$ matrix $\boldsymbol{\mathcal{E}} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T)'$ is defined as $\boldsymbol{\mathcal{E}} = \mathbf{RUG}$, where $\mathbf{G} = (g_{ij})$ is $N \times N$ and $\mathbf{R} = (r_{ts})$ is $T \times T$.*
2. *The elements of $\mathbf{U} = (u_{it})$ are independent and identically distributed random variables satisfying $E(u_{it}) = 0$, $E\left(u_{it}^2\right) = 1$, and $E\left(u_{it}^8\right) < \infty$.*
3. *For $i, j = 1, 2, \ldots, N$, (g_{ij}) are nonstochastic and satisfy $\max_i \sum_{j=1}^{N}|g_{ij}| \leq M$ and $\max_j \sum_{i=1}^{N}|g_{ij}| \leq M$.*
4. *For $t, s = 1, 2, \ldots, T$, (r_{ts}) are nonstochastic and satisfy $r_{ts} = 0$ for all $t \neq s$ as well as $\max_t |r_{tt}| \leq M$.*
5. *For all N, T, the Tth largest eigenvalue of $\mathbf{RR}'$ and the Nth largest eigenvalue of $\mathbf{G}'\mathbf{G}$ equal some small, positive number δ.*

Assumption 4. $E\left(N^{-1}\sum_{i=1}^{N}\|T^{-1/2}\sum_{t=1}^{T}\mathbf{F}_t e_{it}\|^2\right) \leq M$.

Assumption 5. *The asymptotic approximation considered is one where sample dimensions N and T diverge subject to the restrictions $N/T \to c \in (0, \infty)$ as $N, T \to \infty$.*

Assumption 1 requires factors to be realized from a distribution whose density has bounded support. Furthermore, the tail probabilities of the difference between factor sample covariance matrices and their limits are assumed to disappear at an exponential rate. This restricts the degree of serial dependence in $\mathbf{F}_t$. More specifically, dependence must be weak enough to satisfy the conditions of a concentration inequality with much tighter bounds than that of Chebyshev's inequality. Our assumptions on unobserved factors are strictly required by Gagliardini, Ossola, and Scaillet (2019). The estimators of Bai and Ng (2002) as well as of Ahn and Horenstein (2013) are developed under much weaker assumptions; namely, i) the fourth moments of $\mathbf{F}_t$ are bounded,

and ii) $T^{-1}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t'$ converges to its positive definite limit $\mathbf{\Sigma}_F$ at an arbitrary rate. Onatski (2010) even allows for factors with stationary differences.

In assumption 2, the perspective on factor loadings $\boldsymbol{\lambda}_i$ as nonstochastic model components is more restrictive than necessary but allows us to circumvent additional assumptions about the relation between loadings, factors, and idiosyncratic noise. Convergence of $N^{-1}\sum_{i=1}^{T}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$ to a positive definite limit implies "[...] that the cumulative effect of the least influential factor rises proportionally to [the number of cross-sections]" (Onatski 2010, 1005). This assumption is required by all estimators except for that of Onatski, who explicitly accounts for what Chudik and Pesaran (2015) refer to as "semistrong" factors.

The structure imposed on model errors $e_{i,t}$, turning them into linear combinations of independent and identically distributed random variables, follows Ahn and Horenstein (2013) and Onatski (2010). Furthermore, our conditions in assumption 3 on the properties of the matrices $\mathbf{R}$ and $\mathbf{G}$ impose conditions for the degree of serial and cross-section correlation in $e_{i,t}$. These conditions are sufficient for corresponding assumptions in all four articles considered by `xtnumfac`. Gagliardini, Ossola, and Scaillet (2019) and Onatski require the absence of serial correlation, requiring $\mathbf{R}$ to be a diagonal matrix. However, as emphasized by Onatski, consistency of the ED estimator can be shown even in the presence of weak serial correlation at the expense of assuming $e_{i,t}$ to be normally distributed. Under this assumption, an absolute summability condition on the rows and columns of $\mathbf{R}$, analogous to those of assumption 3.3, covers the theoretical frameworks of all estimators except that of Gagliardini, Ossola, and Scaillet.

Assumption 4 is required by Bai and Ng (2002) and Ahn and Horenstein (2013). Sufficient conditions for this assumption are given by mutual independence of $\mathbf{F}_t$ and e_{is} for all i, t, s or the applicability of a central limit theorem to $T^{-1/2}\sum_{t=1}^{T}\mathbf{F}_t e_{it}$.

Lastly, assumption 5 is strictly required only for the estimator of Onatski (2010) because the theoretical properties of the ED estimator are established using well-known results from random matrix theory. These results require that the rate of divergence of time periods and cross-sections be proportional. The remaining estimators make weaker assumptions (Gagliardini, Ossola, and Scaillet 2019) or no such assumptions (Bai and Ng 2002; Ahn and Horenstein 2013), meaning that their asymptotic results may be a better approximation of finite-sample behavior in panel datasets with highly unequal dimensions.