Replication of "Estimating cross-section common stochastic trends in nonstationary panel data"

Xuanbin Yang

November 12, 2022

Contents

1 Introduction

This paper is the replication of "Estimating cross-section common stochastic trends in nonstationary panel data" (Bai, 2004).

2 Models

2.1 Restricted dynamic factor model

Consider the following models:

$$X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{jt} + e_{it} = \lambda_i' F_t + e_{it}, \quad (i = 1, 2, ..., N; t = 1, 2, ..., T), \tag{1}$$

where e_{it} is an I(0) error process, which can be serially corelated for each i, $\lambda_i = (\lambda_{i1}, ..., \lambda_{ir})'$, $F_t = (F_{1t}, ..., F_{rt})'$ is a vector of integrated process

$$F_t = F_{t-1} + u_t,$$

and u_t is a vector $(r \times 1)$ zero-mean I(0) process (not necessarily i.i.d.) that drive the stochastic trend F_t. For each given i, the process X_{it} is I(1) unless $\lambda_i \neq 0$.

We use F_t^0, λ_i^0 and r to denote the true common trends, the true factor loading coefficient, and the true number of trends, respectively. At a given t, we have

$$X_t = \Lambda^0 F_t^0 + e_t, \tag{2}$$

where $X_t = (X_{1t}, X_{2t}, ..., X_{Nt})'$, $\Lambda^0 = (\lambda_1^0, \lambda_2^0, ..., \lambda_N^0)'$, $e_t = (e_{1t}, e_{2t}, ..., e_{Nt})'$.

Let $\underline{X}_i$ be a $T \times 1$ vector of time series observations for the ith cross-section unit. For a given i, we have

$$\underline{X}_i = F^0 \lambda_i^0 + \underline{e}_i, \tag{3}$$

where $\underline{X}_i = (X_{i1}, X_{i2}, ..., X_{iT})'$, $F^0 = (F_1^0, F_2^0, ..., F_T^0)'$, $\underline{e}_i = (e_{i1}, e_{i2}, ..., e_{iT})'$.

2.2 Generalized dynamic factor model

Consider the following dynamic factor models:

$$X_{it} = \lambda_i(L)' F_t + e_{it} \tag{4}$$

$$F_t = F_{t-1} + u_t \tag{5}$$

where $\lambda_i(L)$ is a vector of polynomials of the lag operator. The relation between X_{it} and F_t is now dynamic.

We assume

$$\lambda_i(L) = \sum_{i=0}^{\infty} a_{ij} L^j,$$

where $\sum_{i=0}^{\infty} j \left| a_{ij} \right| < \infty$, and we assume $F_t = 0$ for $t < 0$.

For deriving the limiting distribution, we restrict $\lambda_i(L)$ to be a finite order polynomial. Consider

$$X_{it} = \lambda_{i0}' F_t + \lambda_{i1}' F_{t-1} + \cdots + \lambda_{ip}' F_{t-p} + e_{it}, \tag{6}$$

This can be rewritten as

$$X_{it} = \gamma_{i0}' F_t - \gamma_{i1}' \Delta F_{t-1} - \cdots - \gamma_{ip}' \Delta F_{t-p} + e_{it}, \tag{7}$$

where $\gamma_{ik} = \lambda_{ik} + \lambda_{ik+1} + \cdots + \lambda_{ip}$. Denoting

$$\gamma_i' = \left(\gamma_{i0}', -\gamma_{i1}', \ldots, -\gamma_{ip}'\right) \quad \underline{F}_t = \left(F_t', \Delta F_{t-1}', \ldots, \Delta F_{t-p}'\right)', \tag{8}$$

Eq.(7) can be rewritten as

$$\begin{aligned} X_{it} =& \gamma_i' \underline{F}_t + e_{it} & (9)\\ =& \gamma_{i0}' F_t + \gamma_{i0-}' G_t + e_{it}, & (10) \end{aligned}$$

where $G_t = \left(\Delta F_{t-1}', \ldots, \Delta F_{t-p}'\right)'$, γ_{i0-} is a sub-vector of γ_i other than γ_{i0}. This reparametrization implies that F_t is a vector of I(1) factors, G_t is a vector of I(0) factors.

3 Estimation

3.1 Restricted dynamic factor model

3.1.1 Estimating common stochastic trends and factor loading

Because the true dimension r is unknown, we start with an arbitrary number $k (k < min\{N, T\})$. The superscript in λ_i^k and F_t^k highlights the allowance for k stochastic trends in the estimation. Estimates of Λ^k and F^k are obtained by solving the optimization problem

$$V(k) = \min_{\Lambda^k, F^k} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \lambda_i^{k\prime} F_t^k\right)^2 \tag{11}$$

$$\text{s.t. } F^{k\prime} F^k / T^2 = I_k \text{ or } \Lambda^{k\prime} \Lambda^k / N = I_k \tag{12}$$

If we use the normalization that $F^{k\prime} F^k / T^2 = I_k$, the optimization problem is identical to maximizing $\text{tr}\left(F^{k\prime} \left(XX'\right) F^k\right)$, where $X = \left(\underline{X}_1, \ldots, \underline{X}_N\right)$ is $T \times N$. The estimated common-trend matrix, denoted by $\tilde{F}^k$, is T times the eigenvectors corresponding to the k largest eigenvalues of $T \times T$ matrix XX'; Given $\tilde{F}^k$, we have $\tilde{\Lambda}^{k\prime} = \left(\tilde{F}^{k\prime} \tilde{F}^k\right)^{-1} \tilde{F}^{k\prime} X = \tilde{F}^{k\prime} X / T^2$.

If we use the normalization that $\Lambda^{k\prime} \Lambda^k / N = I_k$, $\bar{\Lambda}^k$ is constructed as the $\sqrt{N}$ times the eigenvectors corresponding to the k largest eigenvalues of $N \times N$ matrix $X'X$ and $\bar{F}^k = X \bar{\Lambda}^k / N$.

The second solution is easier to compute when $N < T$ and the first is easier when $T < N$.

3.1.2 Confidence intervals

Actually, the principal components method is estimating a rotation of the true F^0. To evaluate whether a given economic series is one of (or a linear combination) the underlying factors, consider a rotation of $\tilde{F}_t$ toward R_t via the regression

$$R_t = \alpha + \tilde{F}_t'\delta + error. \tag{13}$$

Let $(\hat{\alpha}, \hat{\delta})$ be the least-squares estimator, and define $\hat{R}_t = \hat{\alpha} + \tilde{F}_t'\hat{\delta}$. When $N, T \to \infty$, $N/T^3 \to 0$, we have

$$\frac{\sqrt{N}\left(\hat{R}_t - \alpha - \delta' F_t^0\right)}{\left[\hat{\delta}' V_{NT}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \tilde{e}_{it}^2 \tilde{\lambda}_i \tilde{\lambda}_i'\right) V_{NT}^{-1}\hat{\delta}\right]^{1/2}} \xrightarrow{\mathrm{d}} N(0,1), \tag{14}$$

where $\tilde{e}_{it} = X_{it} - \tilde{\lambda}_i'\tilde{F}_t$, and V_{NT} is a diagonal matrix consisting of the first r largest eigenvalues of $XX'/(T^2N)$. From this, the 95% confidence interval for $R_t = \alpha + \delta' F_t^0 (t = 1, 2, ..., T)$ is

$$\left(\hat{R}_t - 1.96 S_t N^{-1/2}, \hat{R}_t + 1.96 S_t N^{-1/2}\right), \tag{15}$$

where S_t is the denominator expression given in Eq.(14).

For the null hypothesis $R_t = \delta' F_t^0$ for all t, the constant regressor in Eq.(13,14) can be suppressed, the method above is still valid.

3.2 Generalized dynamic factor models

3.2.1 Estimating common stochastic trends and factor loading

Let $\tilde{F}$ be the r eigenvectors of XX' corresponding to the first r largest eigenvalues normalized such that $\tilde{F}'\tilde{F}/T^2 = I$ and let $\tilde{G}$ be the q eigenvectors corresponding to the next q largest eigenvalues, normalized such that $\tilde{G}'\tilde{G}/T = I$. Denote

$$\underline{\tilde{F}} = (\tilde{F}, \tilde{G}).$$

Let V_{NT}^r be the diagonal matrix of the first r eigenvalues of the matrix $XX'/(T^2N)$ and V_{NT}^q be the diagonal matrix of the $(r+1)$th to $(r+q)$th largest eigenvalues of the matrix $XX'/(TN)$. Denote $\underline{V}_{NT} = \mathrm{diag}\left(V_{NT}^r, V_{NT}^q\right)$. We use superscript 0 to represent the true quantities so that $\underline{F}^0 = \left(\underline{F}_1^0, \underline{F}_2^0, ..., \underline{F}_T^0\right)'$ is the $T \times (r+q)$ true factor matrix and $\Gamma^0 = \left(\gamma_1^0, \gamma_2^0, ..., \gamma_N^0\right)'$ is the $N \times (r+q)$ true factor loading matrix. We estimate $\underline{F}^0$ by $\underline{\tilde{F}}$ and estimate Γ^0 by

$$\tilde{\Gamma} = X'\underline{\tilde{F}}\Upsilon_T^{-2}$$

where $\Upsilon_T = \mathrm{diag}\left(TI_r, \sqrt{T}I_q\right)$.

3.2.2 Confidence intervals

Similarly, we can test the hypothesis that an observable sequence R_t is one of (or a linear combination) the underlying factors. Consider rotating the estimated factors toward R_t by running the regression:

$$R_t = \alpha + \delta' \underline{\tilde{F}}_t + error. \tag{16}$$

Let $(\hat{\alpha}, \hat{\delta})$be the least-squares estimator and define $\hat{R}_t = \hat{\alpha} + \hat{\delta}' \underline{F}_t$. When $N, T \to \infty$, $N/T^2 \to 0$, we have

$$\frac{\sqrt{N}\left(\hat{R}_t - \alpha - \delta' \underline{F}_t^0\right)}{\left[\hat{\delta}' \underline{V}_{NT}^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \tilde{e}_{it}^2 \tilde{\gamma}_i \tilde{\gamma}_i'\right) \underline{V}_{NT}^{-1} \hat{\delta}\right]^{1/2}} \xrightarrow{\mathrm{d}} N(0,1). \tag{17}$$

From this, the 95% confidence interval for $R_t = \alpha + \delta' F_t^0 (t = 1, 2, ..., T)$ is

$$\left(\hat{R}_t - 1.96 S_t N^{-1/2}, \hat{R}_t + 1.96 S_t N^{-1/2}\right), \tag{18}$$

where S_t is the denominator expression of Eq.(17).

If we test $R_t = \delta' \underline{F}_t^0$, the constant in the regression should be suppressed and the corollary continues to hold.

4 The number of common stochastic trends

4.1 Using data in differences

Model (1) under first differencing takes the form

$$\Delta X_{it} = \lambda_i' u_t + \Delta e_{it}.$$

Let

$$V(k) = \min_{\Lambda^k, U^k} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\Delta X_{it} - \lambda_i^k u_t^k\right)^2, \tag{19}$$

where $U^k = \left(u_1^k, u_2^k, ..., u_T^k\right)'$. Consider the criterion of the form:

$$PC(k) = V(k) + kg(N, T),$$

where $g(N, T)$ is a penalty function. Let $kmax$ be a positive integer such that $r < kmax$ and let

$$\hat{k} = \underset{0 \leqslant k \leqslant kmax}{\arg\min}\ PC(k). \tag{20}$$

Denote $C_{NT} = \min[\sqrt{N}, \sqrt{T}]$, let $\hat{\sigma}^2$ be a consistent estimate of $(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} E\left(\Delta e_{it}\right)^2$, say $\hat{\sigma}^2 = V(kmax)$. The criteria in (24-26) with $\alpha_T = 1$ can consistently estimate the number of common stochastic trends.

4.2 New criteria for data in levels

As for data in levels, let

$$V(k) = V\left(k, \hat{F}^k\right) = \min_{\Lambda^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \lambda_i^{k\prime} \hat{F}_t^k\right)^2 \tag{21}$$

denote the sum of squared residuals (divided by NT) when k trends are estimated. Consider the criteria

$$IPC(k) = V(k) + kg(N,T), \tag{22}$$

to consitently estimate r, where the label "IPC" refers to "Integrated Panel Criterion". Again, assume $r < kmax$ and let

$$\hat{k} = \underset{0 \leqslant k \leqslant kmax}{\arg\min}\, IPC(k). \tag{23}$$

Let $\alpha_T = T/[4 \log\log(T)]$, consider the following criteria:

$$IPC_1(k) = V(k) + k\hat{\sigma}^2 \alpha_T \left(\frac{N+T}{NT}\right) \log\left(\frac{NT}{N+T}\right); \tag{24}$$

$$IPC_2(k) = V(k) + k\hat{\sigma}^2 \alpha_T \left(\frac{N+T}{NT}\right) \log C_{NT}^2; \tag{25}$$

$$IPC_3(k) = V(k) + k\hat{\sigma}^2 \alpha_T \left(\frac{N+T-k}{NT}\right) \log(NT). \tag{26}$$

These criteria can consitently estimate the number of common stochastic trends.

Because there are r of I(1) factors, the r I(1) factors can be consistently estimated by the data in levels, and the differenced data approach leads to consistent estimation of the total number of factors $(r+q)$. Thus, q can also be consistently estimated.

5 Simulation results

5.1 The dimension of common trends

5.1.1 Restricted dynamic factor model

We first consider standard (restricted) dynamic factor models (no lags of F_t entering into X_{it}). Data are generated according to

$$X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{jt} + e_{it} \tag{27}$$

$$F_{jt} = F_{jt-1} + u_{jt} \tag{28}$$

$$e_{it} = \rho e_{it-1} + v_{it} + \theta v_{it-1} \tag{29}$$

where λ_{ij}, u_{ij} and v_{it} are i.i.d. $N(0,1)$ for all (i,j,t), and are independent of each other. The parameter values are $r = 2$, $\rho = 0.5$, $\theta = 0.5$. Thirteen combinations of N and T of various sizes are considered. In

Table 1: Estimated number of factors averaged over 1000 repetitions (Restricted dynamic factor model)

		Differenced data			Level data		
N	T	PC1	PC2	PC3	IPC1	IPC2	IPC3
100	40	3.776	2.843	2	1.998	1.997	1.920
100	60	2.137	2.002	2	1.999	1.996	1.915
200	60	2.000	2.000	2	2.000	2.000	1.925
500	60	2.000	2.000	2	2.000	2.000	1.932
1000	60	2.000	2.000	2	2.000	2.000	1.930
40	100	2.375	2.035	2	1.990	1.983	1.832
60	100	2.007	2.000	2	1.998	1.991	1.873
60	200	2.000	2.000	2	1.995	1.993	1.862
60	500	2.000	2.000	2	1.998	1.998	1.861
60	1000	2.000	2.000	2	1.997	1.997	1.851
50	50	4.243	2.605	2	1.994	1.982	1.876
100	100	2.000	2.000	2	2.000	1.996	1.927
200	200	2.000	2.000	2	2.000	2.000	1.976

Note:
The true number of I(1) factors is 2.

all cases, $kmax = 8$. Both the differenced data and level data methods are used and evaluated. Table 1 reports the average $\hat{k}$ over 1000 simulations. The differenced and level methods are both estimating $r = 2$. All criteria perform reasonably well, except the first two criteria for the differenced data with $T = 40$ and with $T = N = 50$.

5.1.2 Generalized dynamic factor models

We next consider generalized dynamic factor models, Eq.(24) is replaced by

$$X_{it} = \sum_{j=1}^{r} \sum_{k=0}^{p} \lambda_{ijk} F_{jt-k} + e_{it}, \tag{30}$$

where the λ_{ijk} are i.i.d. $N(0,1)$. The parameter are $r = 2$ and $p = 1$ while the other parameters are consistent with the previous model. Given $r = 2$, Eq.(30) can be rewritten as

$$X_{it} = \lambda_{i10} F_{1t} + \lambda_{i11} F_{1t-1} + \lambda_{i20} F_{2t} + \lambda_{i21} F_{2t-1} + e_{it}. \tag{31}$$

With data in levels, $r = 2$ factors should be identified; with data in differences, $r(p + 1) = 4$ factors should be identified. The results are reported in Table 2, with each entry representing the average of $\hat{k}$ over 1000 repetitions. The simulation results are consistent with the theory that the number of factors in a generalized dynamic factor model can be identified.

Table 2: Estimated number of factors averaged over 1000 repititions (Generalized dynamic factor model)

		Differenced data			Level data		
N	T	PC1	PC2	PC3	IPC1	IPC2	IPC3
100	40	4.648	4.136	3.999	2.062	2.026	1.965
100	60	4.005	4.000	4.000	1.999	1.998	1.978
200	60	4.000	4.000	4.000	2.000	2.000	1.984
500	60	4.000	4.000	4.000	2.000	2.000	1.980
1000	60	4.000	4.000	4.000	2.000	2.000	1.982
40	100	4.031	4.001	4.000	2.000	2.000	1.960
60	100	4.000	4.000	4.000	2.000	2.000	1.973
60	200	4.000	4.000	4.000	2.000	1.999	1.983
60	500	4.000	4.000	4.000	2.000	2.000	1.975
60	1000	4.000	4.000	4.000	2.000	2.000	1.982
50	50	5.014	4.068	4.000	2.000	1.998	1.962
100	100	4.000	4.000	4.000	2.000	2.000	1.990
200	200	4.000	4.000	4.000	2.000	2.000	2.000

Note:
The level-data method gives an estimate of r (true value r = 2), and the differenced-data method gives an estimate of r(p + 1) (true value is 4).

5.2 Estimating common trends

5.2.1 Restricted dynamic factor model

Data are generated according to Eqs. (27)–(29) with $r = 2, \rho = 0.5, \theta = 0.5$. The true factors are denoted by $F^0(T \times 2)$. We fix T at $T = 30$. We examine the behavior of the factor estimates as N varies from $N = 25$ to 50 and then to 100. For each (T, N), we simulate a sample of observations, denoted by X, a $T \times N$ matrix. We use the estimator $\tilde{F}(T \times N)$, which is equal to the eigenvectors of the first two largest eigenvalues of XX' multiplied by T. To see that $\tilde{F}$ is estimating a transformation of F^0, we rotate $\tilde{F}$ toward each of the true factor process via the following regression

$$F^0_{kt} = \delta'_k \tilde{F}_t + error \tag{32}$$

for $k = 1, 2$. Let $\hat{\delta}_k$ be the least-squares estimate of δ_k. Then $\hat{\delta}'_k \tilde{F}_t$ is the predicted value of F^0_{kt} using the predictor $\tilde{F}_t$. The precision of the factor estimates increases as N becomes larger (for example, see the sample correlation coefficient reported in Fig.1). A plot of $\hat{\delta}'_k \tilde{F}_t$ along with F^0_{kt} would show that they track each other extremely well, but instead, we have plotted the confidence intervals. These 95% confidence intervals together with the true factor process are plotted in Fig.1. The left panels are for the first factor and the right panels for the second factor. The true factor processes are indeed located inside the confidence intervals with the exception of a small number of data points.

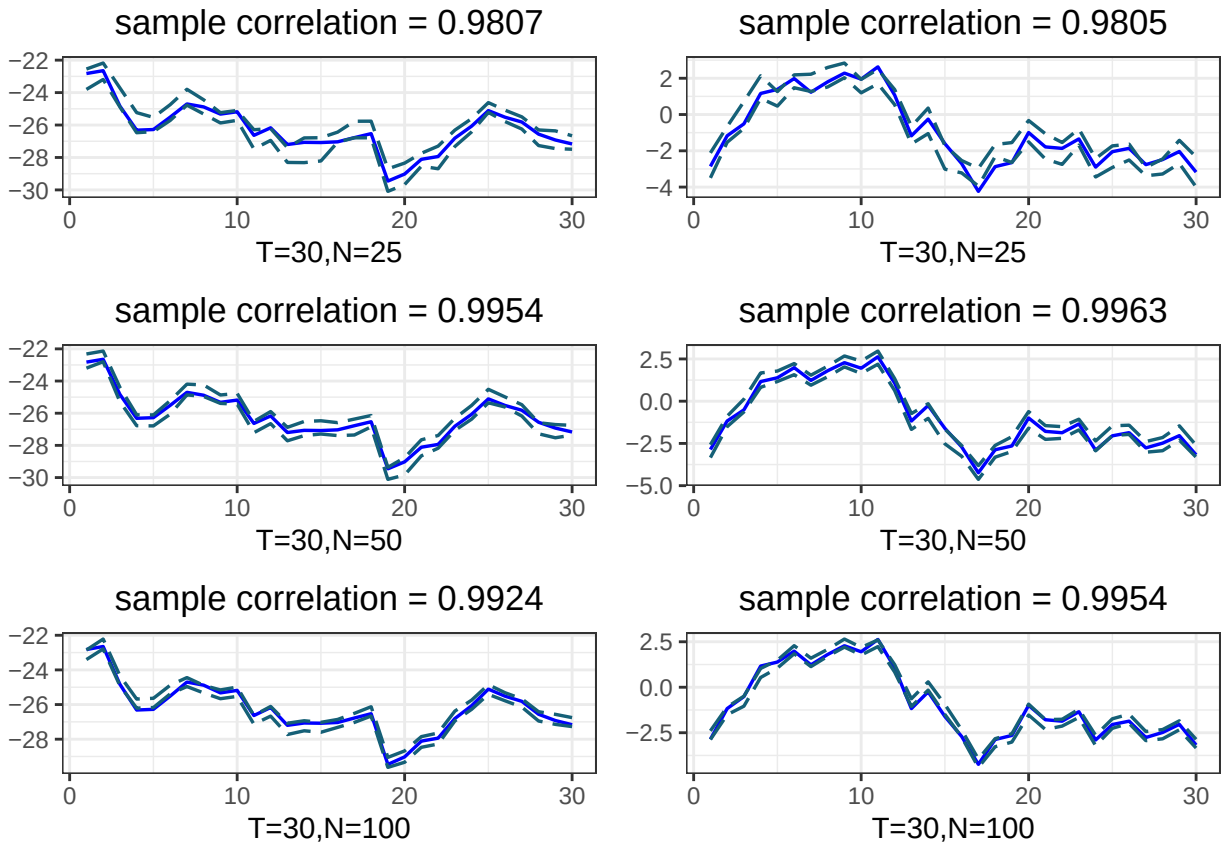

Figure 1: Confidence intervals for the true factor process. Data are generated according to the model specified in Table 1. The left three panels are the confidence intervals (dashed line) for the first true factor along with the true factor process, and the intervals are estimated from N=25

5.2.2 Generalized dynamic factor models

The data are generated according to Eqs. (30, 28, 29) with $r = 2, p = 1, \rho = 0.5$ and $\theta = 0.5$, Again T is fixed at 30 and N takes on the values $25, 50$, and 100. In this case, 4 factors need to be estimated, with 2 being I(1) and 2 being I(0). Let $\underline{\tilde{F}}$ be the $T \times 4$ factor estimate described in Section 3.2.1. For $k = 1, 2$, we consider the rotation

$$F_{kt}^0 = \delta_k' \underline{\tilde{F}}_t + error.$$

These 95% intervals along with the true factor process $\{F_{kt}^0\}$ and sample correlation coefficient are plotted in Fig.2. Again, $\hat{\delta}_k' \underline{\tilde{F}}_t$ tracks F_{kt}^0 extremely well.

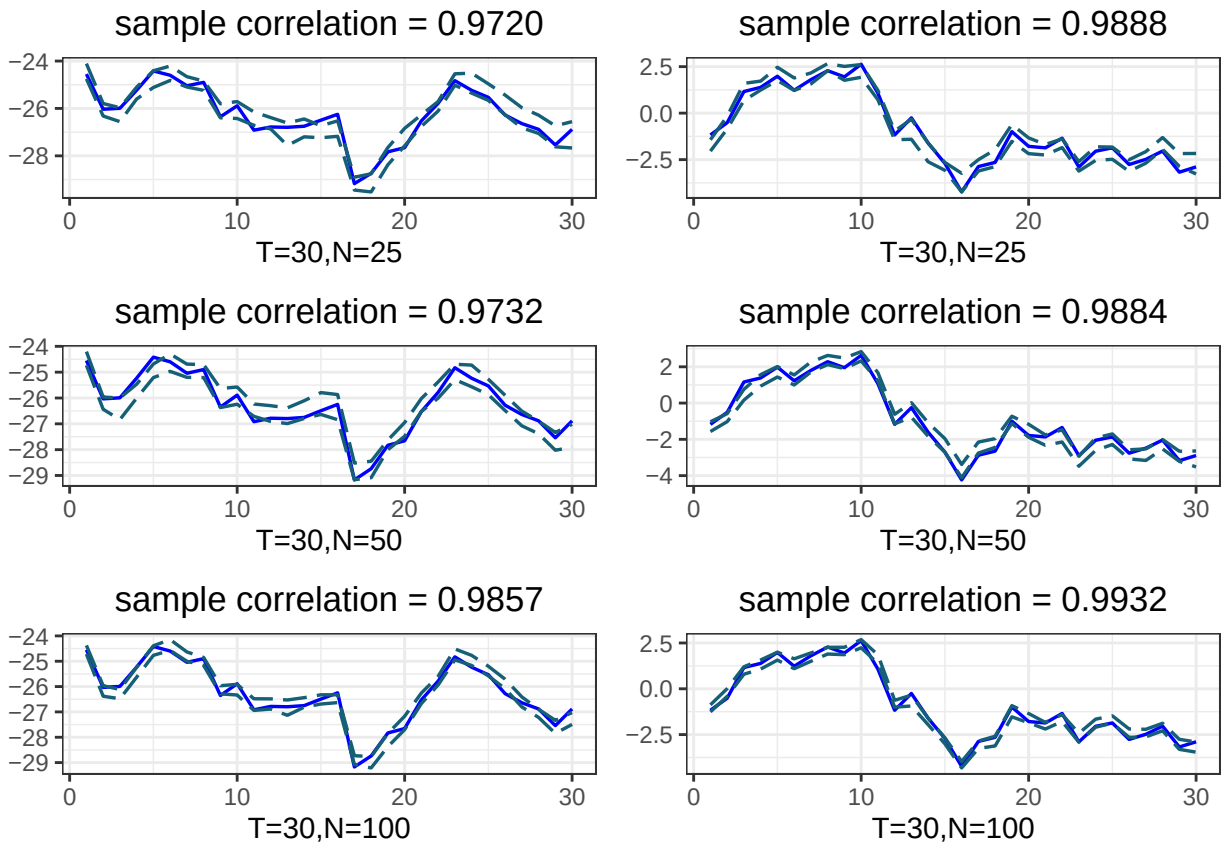

Figure 2: Confidence intervals for the true factor process. Data are generated according to the model specified in Table 2 (generalized dynamic factor models). The left three panels are the confidence intervals (dashed lines) for the first factor and the intervals are estimated from N = 25

(The simulation results are nearly the same with the original ones.)

6 Application: sectoral employment

In this section we study fluctuations in employment across 60 industries for the U.S. We examine the hypothesis that these fluctuations can be explained by a small number of aggregate factors.

The Bureau of Economic Analysis (BEA) reports the number of full-time equivalent (FTE) workers across various industries (NIPA, Tables 6.5b and 6.5c). There are a total of sixty private sector industries. A list of them is provided in Appendix. The data are annual frequency, ranging from 1948 to 2000.

The two sectors, "Social services" and "Membership organizations" under category "Social services and membership organizations" miss data during 1948-1974, after analysing the data trends, I interpolate the missing data such that "Social services" equals floor(1/3* "Social services and membership organizations") and "Membership organizations" equals ceiling(2/3* "organizations Social services and membership organizations").

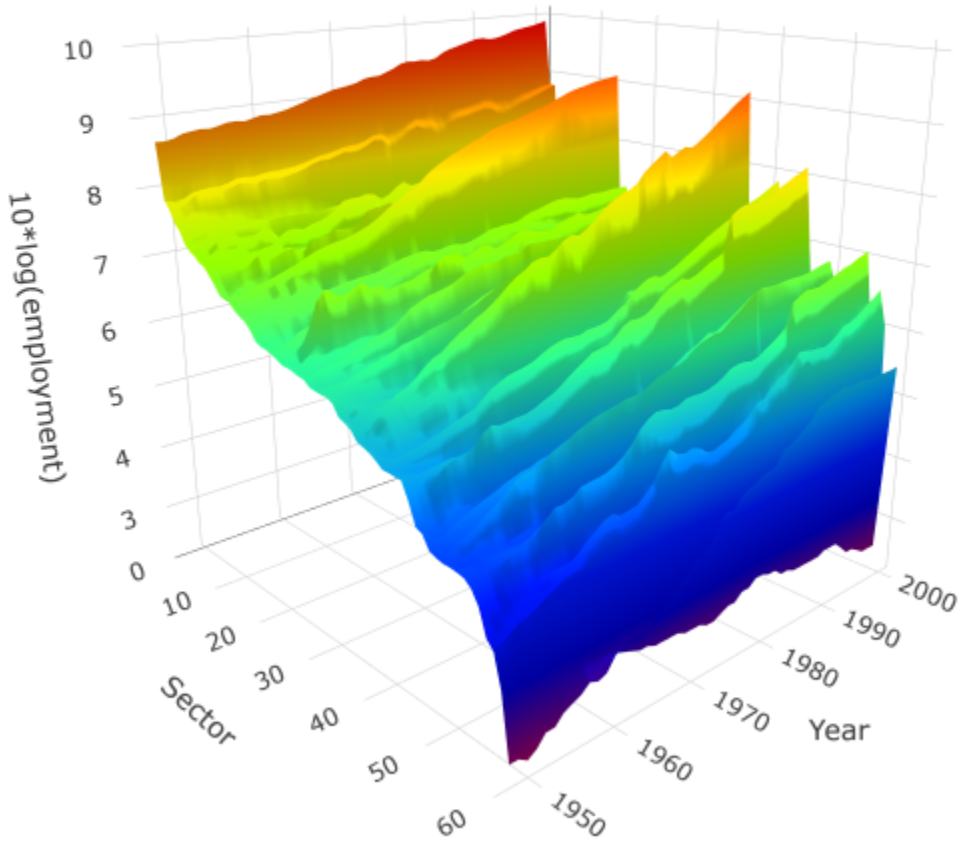

Figure 3: The number of full-time equivalent employees across 60 sectors. The sectors are arranged in ascending order according to their 1948 values.

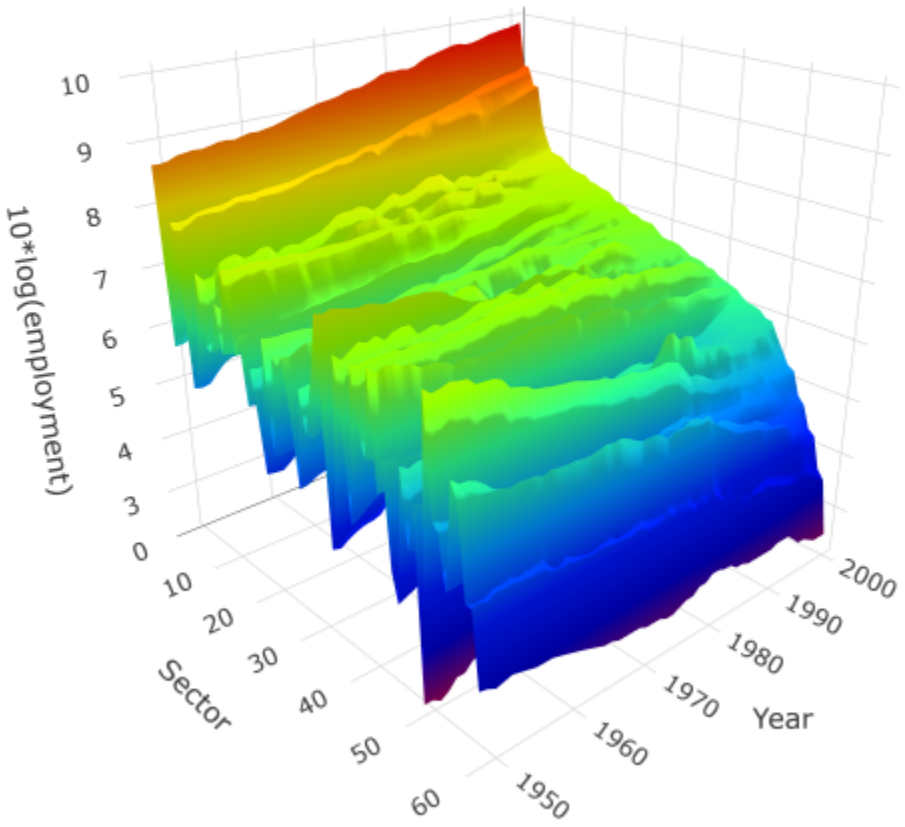

Figure 4: The number of full-time equivalent employees across 60 sectors. The sectors are arranged in ascending order according to their 2000 values.

Our analysis is based on the log-valued data. For graphical display, the series are ordered cross-sectionally to have a better view of the data. Two plots are given according to different methods of ordering. In Fig.3, we order the cross-sections according to their 1948 values in ascending order. In Fig.4, the cross-sections are ordered according to their 2000 values also in ascending order. The vertical axis represents the log-valued employment in each sector. The statistical analysis below does not depend on the ordering of cross-sections, and any permutation will give the same results.

The number of factors. For data in levels, we estimate the number of factors using the three criteria in Eq. (24-26). With $kmax = 6$, the first two criteria suggest four factors and the last criterion gives three factors. If we set $kmax = 4$, the first two criteria yield three factors and the last criterion gives two

factors. If we choose $kmax$ to 2, all criteria give two factors. These results provide evidence in support of two nonstationary common factors.

For data in differences, we start with $kmax = 6$, and then set $kmax$ at the estimated value in the first around as in the previous paragraph. With $kmax = 6$, the first two criteria suggest six factors and the last criterion gives four factors. If we set $kmax = 4$, all criteria give four factors. If we choose $kmax$ to 2, all criteria give two factors. We follow the decision in the original paper that there is one I(0) factor in the system. The estimated residuals resulting from a three-factor model is plotted in Fig.5. No discernable pattern is found in the residuals, indicating a reasonable fit.

(The results are a little different from the original paper. Since the simulation is consistant with the original paper, then the difference can be attributed to the data consistency. Our data have missing values and we don't know how the author deal with this issue, even more, we don't know how much the rest of our data differs from the author's, though the Fig 3 and Fig 4 is similar to the original ones.)

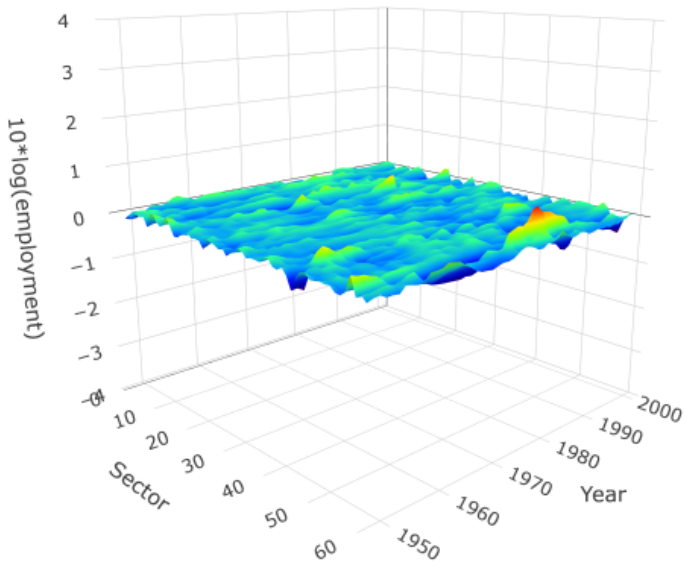

Figure 5: The number of full-time equivalent employees across 60 sectors. Estimated residuals from a three-factor model.

Macroeconomic factors. We test whether or not total employment and total output are the underlying factors.

To test whether total employment (log of value) is a true underlying factor, we rotate the three statistical factors toward E_t by running the regression $E_t = \delta' \underline{\tilde{F}}_t + error$. We then compute and plot the confidence intervals for the true underlying factor. Also plotted is the observable total employment. It is seen that total employment lies inside the confidence intervals throughout the most periods in 1948–2000, see Fig.6 (a little different from Bai's due to the using of different data). This suggests that we can accept the hypothesis that total employment is one of the underlying factors.

To test whether GNP (log of value) is one of the true factors, we rotate the three statistical factors toward Y_t by running the regression $Y_t = \delta' \underline{\tilde{F}}_t + error$. Since there are many periods for which GNP stays outside the confidence intervals, the evidence in supporting GNP as a factor is dubious, see Fig.7. It remains an open question as to which economic variable constitutes the second nonstationary factor.

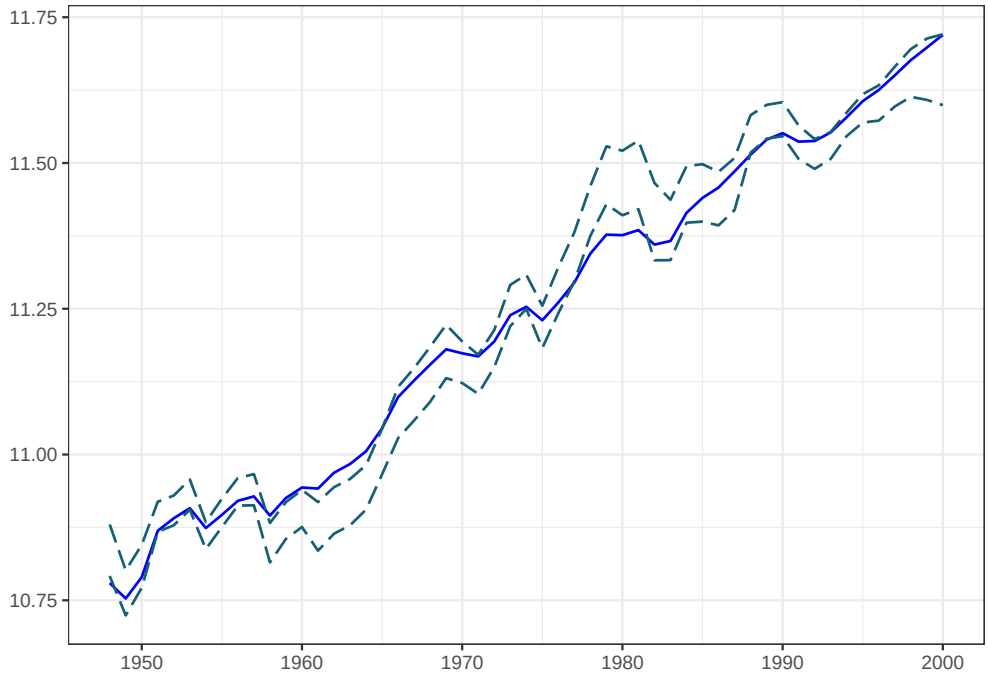

Figure 6: Confidence intervals for testing total employment as a factor. Confidence intervals—dashed line, log-valued total employment—solid line.

(It can be seen that the ranges of y axes in Fig.6 and Fig.7 are a little different from the original ones in Bai (2004), so our data is not totally the same with the author's. The Bureau of Economic Analysis (BEA) have changed their statistical method and report different data of total employment and GNP on their website, thus, some difference of the results may occur).

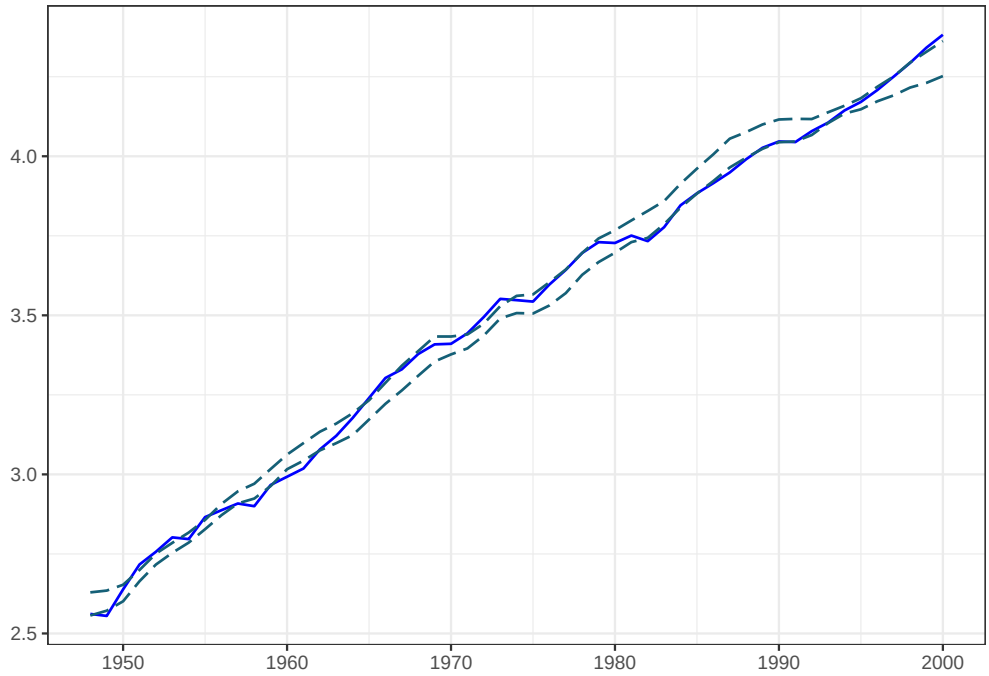

Figure 7: Confidence intervals for testing GNP as a factor. Confidence intervals—dashed line, log-valued GNP—solid line.

7 My Comments

This paper proposes three criteria for consistently estimating cross-section common stochastic trends in nonstationary panel data which are widely used and extended in both econometric theoretical and empirical study.

However, these criteria' accuracy depend on the choice of $kmax$ to some extent. If the $kmax$ is chosen to be too large, for example, $kmax = 20$ of $kmax = 30$ in section 5.1.2, then we will see the wrong results as follows.

Table 3: Estimated number of factors averaged over 1000 repititions (Generalized dynamic factor model) when kmax=20

		Differenced data			Level data		
N	T	PC1	PC2	PC3	IPC1	IPC2	IPC3
100	40	19.477	18.334	9.265	7.584	6.884	3.876
100	60	14.362	12.214	4.003	3.786	3.550	2.032
200	60	11.749	10.259	4.000	3.764	3.627	2.004
500	60	8.551	7.739	4.000	3.812	3.766	2.000
1000	60	6.146	5.709	4.000	3.840	3.824	2.000
40	100	17.184	15.672	5.151	2.427	2.247	2.000
60	100	12.440	10.210	4.000	2.053	2.018	2.000
60	200	6.442	5.240	4.000	2.000	2.000	2.000
60	500	4.000	4.000	4.000	2.000	2.000	2.000
60	1000	4.000	4.000	4.000	2.000	2.000	1.998
50	50	18.994	16.810	11.884	5.164	4.282	3.251
100	100	8.326	5.451	4.000	2.012	2.000	2.000
200	200	4.000	4.000	4.000	2.000	2.000	2.000

Note:

The level-data method gives an estimate of r (true value r = 2), and the differenced-data method gives an estimate of r(p + 1) (true value is 4).

Table 4: Estimated number of factors averaged over 1000 repititions (Generalized dynamic factor model) when kmax=30

		Differenced data			Level data		
N	T	PC1	PC2	PC3	IPC1	IPC2	IPC3
100	40	30.000	30.000	29.999	19.546	18.894	14.830
100	60	27.962	25.950	16.892	7.693	6.566	3.977
200	60	26.765	25.326	7.748	5.907	5.244	3.918
500	60	26.696	25.965	4.000	4.840	4.525	3.789
1000	60	27.152	26.725	4.000	4.364	4.227	3.674
40	100	30.000	30.000	29.301	5.544	4.896	3.696
60	100	26.123	23.862	12.712	3.912	3.771	2.111
60	200	20.245	18.351	4.000	2.000	2.000	2.000
60	500	9.969	8.925	4.000	2.000	2.000	2.000
60	1000	4.011	4.001	4.000	2.000	2.000	2.000
50	50	30.000	29.972	30.000	16.931	15.464	12.468
100	100	19.854	15.889	4.053	3.511	3.051	2.000
200	200	4.005	4.000	4.000	2.000	2.000	2.000

Note:

The level-data method gives an estimate of r (true value r = 2), and the differenced-data method gives an estimate of r(p + 1) (true value is 4).

According to my analysis to the simulation results, in practical application, some error may occur in the following three conditions:

1. T and N is too small. According to Table 3 and Table 4, small T and N tend to estimate the number of factors incorrectly.
2. The high dimensional dataset contains I(d)(d$\geq$ 2) series. Since this paper only consider I(0) and I(1) process, if X contains I(d)(d$\geq$ 2) series, the criteria tend to over estimate the number of factors. For example, I test the dataset in McCracken & Ng (2016) without transforming to ensure stationarity, the estimated number equals to $kmax$ by these criteria in this paper, no matter what $kmax$ is set (in this example, $T = 730$ and $N = 136$, obviously this error is not caused by data size).
3. Some special high dimensional data. $kmax = 20$ or $kmax = 30$ seems to be too large in this simulation even if we don't know the true number of factors, however, we cannot determine how large the $kmax$ is can be regarded as "too large". It is entirely possible that some datasets are extremely sensitive to $kmax$, a usually considered small $kmax$ may still be "too large" for such datasets.

Bibliography

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, *122*(1), 137–183.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Appendix

Index	Sector	Index	Sector
1	Farms	31	Trucking and warehousing1
2	Agricultural services, forestry, and fishing	32	Water transportation
3	Metal mining	33	Transportation by air1
4	Coal mining	34	Pipelines, except natural gas
5	Oil and gas extraction	35	Transportation services
6	Nonmetallic minerals, except fuels	36	Telephone and telegraph
7	Construction	37	Radio and television
8	Lumber and wood products	38	Electric, gas, and sanitary services
9	Furniture and fixtures	39	Wholesale trade
10	Stone, clay, and glass products	40	Retail trade
11	Primary metal industries	41	Banking
12	Fabricated metal products	42	Credit agencies other than banks
13	Machinery, except electrical	43	Security and commodity brokers
14	Electric and electronic equipment	44	Insurance carriers
15	Motor vehicles and equipment	45	Insurance agents, brokers, and service
16	Other transportation equipment	46	Real estate
17	Instruments and related products	47	Holding and other investment offices
18	Miscellaneous manufacturing industries	48	Hotels and other lodging places
19	Food and kindred products	49	Personal services
20	Tobacco manufactures	50	Business services
21	Textile mill products	51	Auto repair, services, and parking
22	Apparel and other textile products	52	Miscellaneous repair services
23	Paper and allied products	53	Motion pictures
24	Printing and publishing	54	Amusement and recreation services
25	Chemicals and allied products	55	Health services
26	Petroleum and coal products	56	Legal services
27	Rubber and miscellaneous plastics products	57	Educational services
28	Leather and leather products	58	Social services
29	Railroad transportation	59	Membership organizations
30	Local and interurban passenger transit	60	Miscellaneous professional services