Estimating the Number of Common Factors in Serially Dependent Approximate Factor Models*

Ryan Greenaway-McGrevy
Bureau of Economic Analysis
Washington, D.C.

Chirok Han
Department of Economics
Korea University

Donggyu Sul
School of Economics, Political and Policy Sciences
University of Texas at Dallas

First draft: April 2009; Revision: April 2011

Abstract

The Bai-Ng criteria are shown to overestimate the true number of common factors when panel data exhibit considerable serial dependence. We consider filtering the data before applying the Bai-Ng method as a practical solution to the problem. Despite possible bias and model misspecification, the AR(1) least squares dummy variable (LSDV) filtering method is shown to be consistent for panels with a wide range of serial correlation, and a combination of the LSDV and the first-difference filtering methods is shown to be consistent in both weakly and strongly serially correlated panels. According to simulations these filtering methods considerably improve the finite sample performance of the Bai-Ng selection methods. We illustrate the practical advantages of LSDV filtering by considering three economic panels that exhibit moderate serial dependence. In each case, LSDV filtering yields a reasonable factor number estimate, whereas conventional methods do not.

Keywords: Factor Model, Selection Criteria, Serial Correlation, Weakly Intergrated Process, Principal Components Estimator, Bai-Ng Criteria.

JEL Classification: C33

*The views expressed herein are those of the authors and not necessarily those of the Bureau of Economic Analysis or the Department of Commerce. Research for the paper was partially supported under a Marsden grant. The authors thank Donald Andrews, Yoosoon Chang, In Choi, Yoonseok Lee and H. Roger Moon for helpful comments. Corrresponding author: Ryan Greenaway-McGrevy, Bureau of Economic Analysis, 1441 L Street N.W. BE-40, Washington D.C., 20232, USA. tel. +1 202-606-9844, email ryan.greenaway-mcgrevy@bea.gov.

1 Introduction

The precise estimation of the number of common factors is a cornerstone of the factor model literature, and several recent studies have suggested various methods for selecting the factor number (e.g., Connor and Korajczyk, 1993, Forni et al., 2000; Bai and Ng, 2002; Bai, 2004; Stock and Watson, 2005; Amengual and Watson, 2007; and Onatski, 2007). The approximate factor model (Chamberlain and Rothschild, 1983; Bai and Ng, 2002) permits weak-form serial (and cross-sectional) dependence in the idiosyncratic component as long as N (cross section dimension) and T (time-series dimension) are large. This is because dependence due to the factor structure asymptotically dominates any weak dependence in the idiosyncratic component, and hence well-designed criteria (e.g., Bai and Ng, 2002; BN hereafter) can eventually determine the number of factors as both N and T go infinity.

However if the idiosyncratic component exhibits high serial dependence relative to the given sample size, or equivalently if the sample size is not large enough for the given degree of serial dependence, then the BN factor number estimate may be different from the truth with considerable probability. The problem is particularly acute for economic panel data, as many economic time series exhibit considerable dependence. To illustrate this point we consider three different panel datasets: disaggregate US personal consumption expenditure growth, US metropolitan consumer price inflation, and US industry employment growth. These panels exhibit both moderate serial dependence and a moderate time series dimension, and in each case the Bai-Ng criteria select an unreasonably large factor number. An upwardly biased factor number undercuts one of the main purposes of the factor model, namely to summarize a glut of data using a handful of factors (often referred to as "dimension reduction"; see e.g., Bai and Ng, 2008).

This paper has two purposes. First, we formally analyze the effect of serial dependence in the idiosyncratic component on factor number estimation. To do so we adopt a local alternative setting. For example, consider $X_{it} = \lambda_i' F_t + e_{it}$, $e_{it} = \rho e_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim iid\left(0, \sigma_\varepsilon^2\right)$. If $|\rho| < 1$ is constant the BN criteria estimates the factor number consistently as $N \to \infty$ and $T \to \infty$, but with finite N and T the BN criteria may overestimate the factor number with considerable probability. Hence it is difficult to analyze the effects of serial dependence under this fixed alternative framework, because the serial dependence is not related to the sample size. However, under a local alternative setting, ρ approaches unity as the sample size increases, so that the effect of a large ρ or a small sample size may be theoretically analyzed.

The second purpose of the paper is to propose and analyze simple methods to estimate the factor number for serially dependent panels. Ideally these methods must bridge the space between weak and strong form serial dependence: Regardless of whether the panel data exhibits weak or strong serial dependence, the proposed methods should estimate the factor number consistently.

To fulfil the first purpose of the paper we provide a couple of dedicated examples. Example 1 in Section 2 considers the case where $\rho \uparrow 1$ as the sample size increases at some controlled rates. The example shows that the BN criteria applied to X_{it} can overestimate the true factor number with probability approaching one

if $1-\rho = O(T^{-\alpha})$ for any $\alpha > 0$.[1] In this example, ρ is homogenous and $\varepsilon_{it} \sim iid\ \left(0, \sigma_{\varepsilon}^2\right)$, so that there is no heteroskedasticity across panels and only strong serial correlation is present. Example 2 in Section 2 considers the case of heterogenous panels (with ρ_i possibly different across i). In this example the BN criteria overestimate the factor number if strong idiosyncratic autocorrelation is present in a considerable number of cross sections.

To achieve the second purpose of the paper we consider linear filters to mitigate the serial dependence in the panel. We show that the BN criteria applied to the filtered data produce consistent estimates of the factor number for a wide-ranging degree of serial dependence in the idiosyncratic component. Of course designing a criterion which is valid under very mild regularity conditions (that is, regardless of whether the source of strong serial dependence comes from unobserved common factors or idiosyncratic components) would be a desirable alternative solution, but it is not straightforward to construct such a general criterion without knowledge of the source and degree of serial dependence.

The filter may be nonrandom or data-dependent. An obvious nonrandom filter is first differencing (FD), which is already widely used in practice. The first differenced data ΔX_{it} give a consistent estimate because ΔX_{it} would satisfy regularity in Bai and Ng (2002) in general. But as one may well guess, when using the first difference filter there is the risk of over-differencing and thus inducing negative serial correlation in the transformed data. This risk is particularly acute when e_{it} is close to white noise. Indeed, in our empirical examples, first differencing results in apparent over-estimation of the factor number. A more effective filter in this case is the pooled least squares dummy variable (LSDV) fitting. Despite possible model misspecification and estimator bias, the pooled LSDV AR(1) filter is shown to yield a consistent factor number estimate for a wide range of processes with serial dependence. We show the filter should be common to all cross sections in order to preserve the factor number in the transformed panel; thus in the LSDV setting, pooling is crucial regardless of whether the "true" parameters governing serial dependence in e_{it} are heterogenous. In our empirical examples the LSDV filter yields a more credible factor number estimate than either the first-differenced or unfiltered panels.

While FD filtering works well when the idiosyncratic autocorrelation is strong (as long as the signal from the differenced common factors is asymptotically preserved), the LSDV filtering method requires that the idiosyncratic serial correlation is sufficiently bounded away from unity (e.g., $T(1-\rho) \to \infty$ for homogeneous panels). So the validity of the LSDV filtering may depend on unknown parameters. However the combination of the FD filtering and the LSDV filtering by a minimum rule, which is explained later in Section 3.4, estimates the factor number consistently without the knowledge of the persistency in e_{it}. For example, if $e_{it} = \rho_i e_{it-1} + \varepsilon_{it}$, we need not observe the rates at which the autoregressive coefficients approach unity, if at all. By means of a Monte Carlo study, we show that these approaches work quite well in the finite sample. Thus the combination of data-dependent LSDV filtering with FD filtering through the

[1] For example, we can let $\rho_T = 1 - c/T^{\alpha}$ for some $c > 0$ and α (Phillips, 1987; Elliott et al., 1996; Cavanagh et al., 1995; Moon and Phillips, 2000 and 2004; Phillips et al., 2001; Giraitis and Phillips, 2006 and 2009; Phillips and Magdalinos, 2007; Park, 2003). The case with $\alpha < 1$ is said to be weakly integrated or nearly stationary.

minimum rule provides a useful tool for the practitioner to estimate the factor number for panels that exhibit serial dependence and a moderate time-series dimension.

The rest of the paper is organized as follows. In the following section we present three empirical examples that highlight the need for data-dependent filtering in practice. These three panel datasets all exhibit moderate serial dependence that is sufficiently strong for methods based on levels to fail, but sufficiently weak for methods based on first-differenced data to likewise fail. In section 3 we provide an intuitive explanation of why the BN criteria may not work well in the finite sample when there is serial dependence in the panel. Section 4 provides general theoretical results for the consistency of the estimates based on both the simple AR(1) and first differencing filters. This section also discusses the minimum rule and how to apply the method to some dynamic factor models. Section 5 provides Monte Carlo simulation results. In section 6 we return to the empirical examples, showing that the estimated factor numbers obtained from the LSDV-filtered data are more credible to those obtained from the panel in levels or first differences. Section 7 contains concluding remarks. Mathematical proofs and lemmas are gathered in the appendix.

2 Illustrative Empirical Examples

To better motivate the need for filtering in practice, we consider the problem of estimating the number of common factors in three different panel datasets. In each of the examples, the Bai-Ng criteria applied to panels in levels selects the maximum factor number permitted (we set the maximum factor number to 5 in all examples for clarity). As we have discussed in the introduction, and as we will formally demonstrate using a local asymptotic framework in the following sections, serial dependence in the idiosyncratic component causes over-estimation of the factor number. Hence we conjecture that the poor performance of the criteria in these examples is due to moderate serial dependence in the panel relative to the time series dimension of the sample. We go on to demonstrate that standard treatments for addressing possible overestimation, such as first-differencing and standardization, likewise do not give satisfactory results in these examples, either because they also pick an incredibly large factor number or because the estimated factor number is not robust to small changes in the sample selected.

2.1 Disaggregate PCE Growth

We estimate the number of common factors to annual growth in disaggregated personal consumption expenditure (PCE). We construct deflated consumption data for 182 PCE items using NIPA tables 2.4.4.U and 2.4.5.U available from the BEA website (www.bea.gov). Growth is calculated by multiplying log-differences of deflated nominal consumption by 100. Our sample spans 1983 to 2010. We focus on this period for two reasons. First, many newer technology items in the PCE consumption basket are not available for earlier time periods. For example, “Computer software and accessories,” “Personal Computers and peripheral software,” and “Video cassettes and discs, blank and prerecorded” only enter into PCE from the

late 1970s, whilst "Video media rental" enters into PCE from 1982 onwards. Second, the sample period approximately coincides with the so-called "Great Moderation" of output volatility documented by McConnell and Perez-Quiros (2000). For example, Stock and Watson (2002) estimate that the break in output volatility occurred between 1983Q2 and 1984Q3. Structural breaks in a factor model can result in breaks in the volatility of a cross sectional aggregate of the panel, and structural breaks in the factor structure leads to over-estimation of the factor number (see, e.g., Breitung and Eickmeier, 2009).

Figure 1 depicts the 5%, 50% (median) and 95% percentiles of the distribution of disaggregate consumption growth. There is a significant spread in consumption growth in any given year, but substantial covariation is evident, particularly around the 1990, 2000 and 2007-2009 recessions.

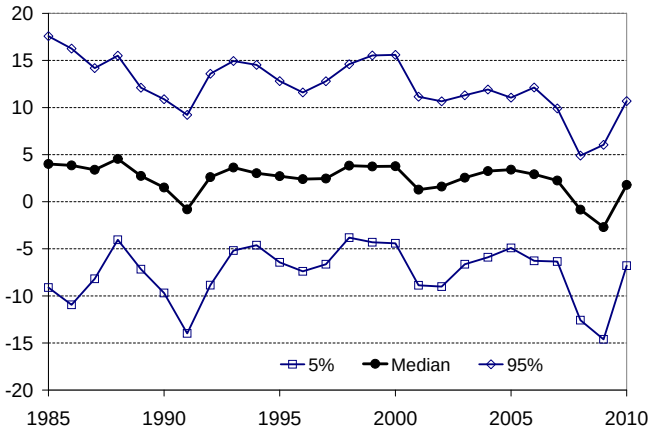

Figure 1: Disaggregate PCE growth

Table 1 exhibits the various factor number estimators based on the BN $IC_{p2}(k)$ criterion. When applied to consumption growth in levels, the criterion selects the maximum number of factors permitted. This result does not change if we standardize each cross section or time-series in the panel. Meanwhile first-differencing likewise produces a factor number estimate of five, indicating the possibility that the FD filter has over-differenced in this case, inducing negative serial dependence that likewise hampers factor number estimation. We also consider a small sample robustness check to see if this over-estimation is due to a particular time period in the sample. The main results for each method in general hold, although in the later subsamples the FD method produces a slightly smaller factor number estimate of four.

2.2 Metropolitan CPI-U Inflation

In the US the headline consumer price index for urban workers (CPI-U) inflation published by the Bureau of Labor Statistics is calculated by taking the average inflation rate over 27 metropolitan areas. This measure of inflation may not reflect changes in the price level for all cities within the US, and hence practitioners may wish to consider a second method of measuring the "representative" inflation rate. One such measure could be to construct a handful of common factors underlying the panel of inflation rates. Of course as a

Table 1: Estimated factor number for disaggregate PCE growth

$IC_{p2}(k)$ with 5 factors maximum; $N = 182, T = 28$

sample	level	level + CSS	level + TSS	first-difference
1983-2010	5	5	5	5
sub-sample robustness				
1983-2008	5	5	5	5
1983-2009	5	5	5	5
1984-2010	5	5	5	4

"level + CSS" denotes Bai-Ng in levels with each cross section standardized; "level + TSS" denotes Bai-Ng in levels with each time series standardized.

first step the factor number must be determined.

We use annual CPI-U data spanning 1978 to 2010 for 23 metropolitan areas (in our application Washington-Baltimore, Tampa, Miami and Phoenix are omitted because sampling only begins in 1997, 1998, 1978 and 2002, respectively). This dataset has been used by, among others, Cecchetti, Mark and Sonora (2002) and Phillips and Sul (2007) to test for convergence in regional prices. We log and first-difference the indices and multiply by 100 to obtain the panel of inflation rates. Figure 2 below shows the minimum, maximum and median of the distribution over time, and demonstrates that there can be substantial differences in the inflations rates across cities in any given year.

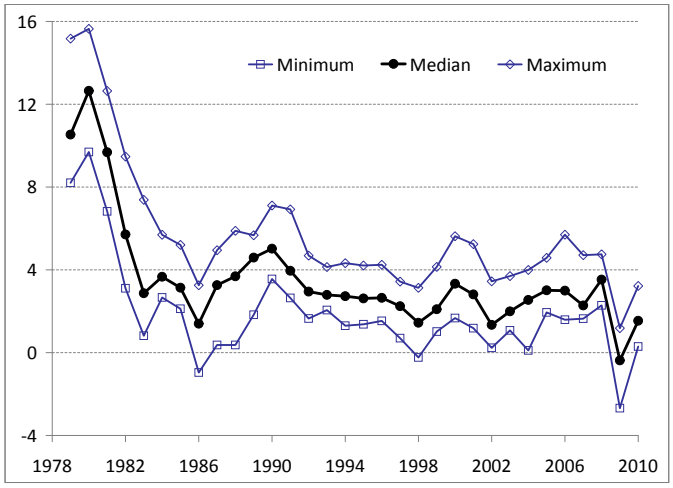

Figure 2: Annual percent change in metropolitian CPI-U

As shown in Table 2, the $IC_{p2}(k)$ criteria applied to inflation in levels selects the maximum number of factors permitted. This result does not change if we standardize each cross section. However, if we standardize each time series we get an estimate of three factors, and we also get three factors if we first

difference the data. We again also consider a small sample robustness check. It appears that the choice of subsample can reduce the factor number estimate based on levels and cross sectional standardization to four in some cases. Meanwhile the FD estimate of the factor number appears very sensitive to the sub-sample selected, with estimates ranging between two to four factors. In summary, none of the considered methods give a concrete estimate of the true factor number.

Table 2: Estimated factor number; Metropolitan CPI-U inflation.

Bai-Ng $IC_{p2}(k)$ with 5 maximum factors, $N = 23, T = 32$.

sample	level	level + CSS	level + TSS	first-difference
1979-2010	5	5	3	3
sub-sample robustness				
1981-2010	4	4	3	3
1982-2009	4	4	3	2
1979-2008	5	5	3	4

"level + CSS" denotes Bai-Ng in levels with each cross section standardized; "level + TSS" denotes Bai-Ng in levels with each time series standardized.

2.3 Industry Employment Growth

We estimate the number of common factors to annual growth in North American Industry Classification System (NAICS) employment. We obtain annual "wage and salary employment" figures from table SA27 on the BEA website (www.bea.gov). We use the highest degree of disaggregation possible (93 industries) and our sample spans 1990-2009. This represents the longest possible time series of NAICS employment currently available (NAICS classification was implemented in 1997). Employment growth is calculated as log-differences of annual wage and salary employees. The panel is also standardized to remove the excessive heteroskedasticity in the panel that is present at this granular level. (For example, employment in "ISPs, search portals, and data processing" grew by 90% over 2006-2007.) Figure 3 below depicts the 5%, 50% (median) and 95% percentiles of the distribution over time.

Table 3 shows that $IC_{p2}(k)$ applied to levels selects the maximum number of factors permitted, while Bai-Ng applied to first differences offers anything between three to five factors depending on the sub-sample. None of the considered methods give a concrete estimate of the true factor number.

The above examples demonstrate that when a panel exhibits moderate serial correlation and the time-series dimension of the panel is small, extant methods for estimating factor numbers tend to pick the maximum number of factors permitted. This may indicate that the factor number is over-estimated. As yet, however, there is no theory to verify this conjecture. We provide this theory in the following section. In

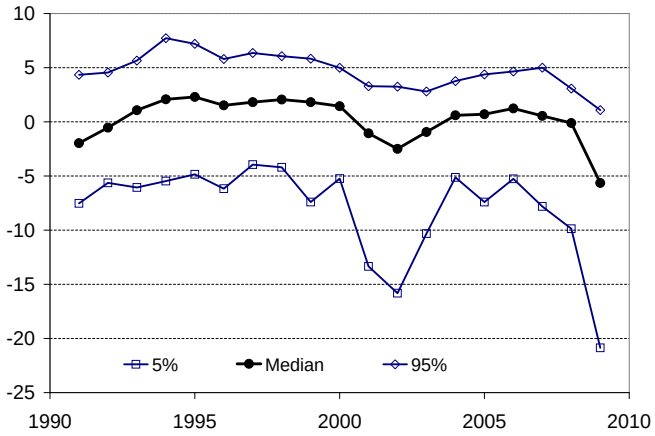

Figure 3: Annual percent change in NAICS inudstry employment

Table 3: Estimated factor number; Industry employment growth

Bai-Ng $IC_{p2}(k)$ with 5 maximum factors, $N = 92, T = 19$.

sample	level	first-difference
1990-2009	5	4
sub-sample robustness		
1992-2009	5	5
1990-2007	5	3

addition, in order to gain a more sensible estimate of the factor number, we suggest data-dependent filtering to mitigate the serial dependence in the panel before applying an eigenvalue-based criterion (such as Bai-Ng criteria). We provide the theoretical justification for this approach in section four.

3 Inconsistency under Strong Serial Dependence

For the approximate factor model $X_{it} = \lambda_i' F_t + e_{it}$ (see Chamberlain and Rothschild, 1983, and Bai and Ng, 2002, for the identification of components for this model with large N and large T), Bai and Ng (2002) propose estimating the factor number r by minimizing

$$PC(k) = V_{NT}(k) + kg(N,T), \quad V_{NT}(k) = \min_{\{\lambda_i^k\},\{F_t^k\}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \lambda_i^{k\prime} F_t^k)^2 \tag{1}$$

or some variant thereof, with respect to $k \in \{0, 1, \ldots, k_{\max}\}$ for some fixed $k_{\max}$, where $g(N,T)$ is a penalty function and $\lambda_i^k, F_t^k \in \mathbb{R}^k$. We hereafter refer to the above as the PC criteria following BN (2002). The eigenvalue-eigenvector decomposition of the covariance matrix of either $(X_{i1}, \ldots, X_{iT})'$ or

$(X_{1t}, \ldots, X_{Nt})'$ is usually used as a solution to the above minimization problem. The $V_{NT}(k)$ term is the average squared residuals from a least squares projection of X_{it} onto the column space of the first k eigenvectors, and it decreases monotonically as more eigenvectors are allowed (i.e., as k increases). The PC criteria balance $V_{NT}(k)$, decreasing in k, against the penalty function $kg(N,T)$, increasing in k, so that the criteria are minimized at the true factor number r asymptotically. To be more precise, according to BN (2002), $V_{NT}(k) - V_{NT}(r) \not\to 0$ for $k < r$, thus (1) is not (asymptotically) minimized at a $k < r$ if $g(N,T) \to 0$. On the other hand, if $k > r$, then $V_{NT}(k) - V_{NT}(r) \to 0$ at a certain rate, C_{NT}^2 say. So if $g(N,T)$ diminishes to zero at a rate slower than C_{NT}^2, then the penalty will eventually become dominant, and overfitting will be prohibited. Under the stationarity assumption (allowing only weak serial and cross-sectional dependence in e_{it}), BN (2002) show that $C_{NT} = \min\{N,T\}^{1/2}$, thus $g(N,T) = C_{NT}^{-2} \log C_{NT}^2$ (and variants thereof) give consistent estimates.

When e_{it} is stationary, $V_{NT}(k)$ has the right order such that $V_{NT}(k) - V_{NT}(r) \not\to 0$ for $k < r$ and $V_{NT}(k) - V_{NT}(r) = O_p(C_{NT}^{-2})$, thus the minimization of $PC(k)$ provides a consistent estimate of r. In practice the scale of $V_{NT}(k)$ matters**,** but the effect can be negated by either dividing $V_{NT}(k)$ by a consistent estimate of $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Ee_{it}^2$ (the PC_p criteria of BN, 2002) or by taking the natural logarithm of $V_{NT}(k)$ (their $IC\,(k)$ criteria). We consider the latter, i.e., minimizing

$$IC(k) = \ln V_{NT}(k) + kg(N,T), \tag{2}$$

as it is convenient and widely used.

When e_{it} shows considerable serial correlation but the sample size is not sufficiently large, the $IC\,(k)$ criteria may overfit. To see this, we consider a simple example with no common factors and weakly integrated idiosyncratic errors:

Example 1. Let there be no common factors so $r = 0$, and $e_{it} = \rho_T e_{it-1} + \varepsilon_{it}$, where $\varepsilon_{it} \sim iid\ N(0,1)$ for simplicity. Let $N \geq T$. Let $X_i = (X_{i1}, \ldots, X_{iT})'$ and $X = (X_1, \ldots, X_N)$. Let $\hat{\ell}_1, \ldots, \hat{\ell}_T$ be the eigenvalues of $N^{-1}\sum_{i=1}^{N} X_i X_i'$ ordered from largest to smallest, where $X_i = (X_{i1}, \ldots, X_{iT})'$. Then $V_{NT}(k) = \frac{1}{T}\sum_{j=k+1}^{T} \hat{\ell}_j$. When $r = 0$ we have $\ln V_{NT}(1) - \ln V_{NT}(0) = \ln(1 - \hat{\ell}_1/\sum_{j=1}^{T}\hat{\ell}_j)$, thus

$$IC(1) - IC(0) = \ln\left(1 - \frac{\hat{\ell}_1}{\sum_{j=1}^{T}\hat{\ell}_j}\right) + g(N,T). \tag{3}$$

Let $\Sigma_{ee} = E[e_i e_i']$ and $\hat{\Sigma}_{ee} = N^{-1}\sum_{i=1}^{N} e_i e_i'$. Let A be such that $AA' = \Sigma_{ee}$ and $v_i = A^{-1}e_i$ so that $v_i \sim N(0,I)$. Let $\hat{\Sigma}_{vv} = N^{-1}\sum_{i=1}^{N} v_i v_i'$. From $\hat{\Sigma}_{ee} = A\hat{\Sigma}_{vv}A'$, we have

$$\begin{aligned}\hat{\ell}_1 &= \max_x \frac{x'\hat{\Sigma}_{ee}x}{xx} = \max_x \left\{\frac{x'A\hat{\Sigma}_{vv}A'x}{x'AA'x} \cdot \frac{x'\Sigma_{ee}x}{x'x}\right\} \geq \max_y \frac{y'\hat{\Sigma}_{vv}y}{y'y} \cdot \min_x \frac{x'\Sigma_{ee}x}{x'x} \\ &= \text{eigval}_{\max}(\hat{\Sigma}_{vv}) \cdot \text{eigval}_{\min}(\Sigma_{ee}).\end{aligned}$$

(The inequality holds because when $f = gh$ with $h > 0$, we have $g = f/h$, so $\max g \leq (\max f)/\min h$, implying that $\max f \geq \max g \cdot \min h$.) But by Yin et al. (1988), $\text{eigval}_{\max}(\hat{\Sigma}_{vv})$ converges in probability to

$(1+\sqrt{\lim T/N})^2 \geq 1$. Also by the Perron-Frobenius Theorem, $\text{eigval}_{\min}(\Sigma_{ee})$ is no less than the minimal row sum of Σ_{ee} which is

$$\frac{\sum_{t=1}^{T}\rho_T^{t-1}}{1-\rho_T^2} = \frac{1-\rho_T^T}{(1-\rho_T)(1-\rho_T^2)}.$$

If $T(1-\rho_T)\to\infty$, we have $\rho_T^T \to 0$ because $\rho_T^T = \left\{[1-(1-\rho_T)]^{1/(1-\rho_T)}\right\}^{T(1-\rho_T)} \to 0$, so $\text{eigval}_{\min}(\Sigma_{ee})$ is eventually bounded from below by $0.5/(1-\rho_T)(1-\rho_T^2)$ as $T\to\infty$. Next, when $T(1-\rho_T)\to\infty$, we have

$$\frac{1-\rho_T^2}{T}\sum_{j=1}^{T}\hat{\ell}_j = \frac{1-\rho_T^2}{NT}\sum_{i=1}^{N}\sum_{i=1}^{N}e_{it}^2 \to_p 1$$

(Giraitis and Phillips, 2009), implying that for N and T large enough,

$$\frac{\hat{\ell}_1}{\sum_{j=1}^{T}\hat{\ell}_j} = \frac{T^{-1}(1-\rho_T^2)\hat{\ell}_1}{T^{-1}(1-\rho_T^2)\sum_{j=1}^{T}\hat{\ell}_j} \geq \frac{1}{4T(1-\rho_T)} \tag{4}$$

with arbitrarily high probability. Now because $\ln(1-x)\leq -x$, we have, from (3) and (4), that

$$IC(1) - IC(0) \leq -\frac{1}{4T(1-\rho_T)} + g(N,T)$$

for N and T large enough, with arbitrarily high probability. When $g(N,T) = (\ln T)/T$ as proposed by BN (2002), if $1-\rho_T$ is smaller than $0.25/\ln T$ (e.g., $\rho_T = 1-T^{-\alpha}$ for any $\alpha>0$), then we eventually have $IC(1)-IC(0)<0$ with arbitrarily high probability, and hence the $IC\,(k)$ criteria overestimates r. This inconsistency can easily be extended to all specific penalty functions considered by BN (2002). ∎

Technically, the overfitting arises because when $e_{it} = \rho_T e_{it-1} + \varepsilon_{it}$ with $\rho_T \to 1$, some of BN's (2002) regularity conditions are violated. In particular, the variance of e_{it} is of order $(1-\rho_T^2)^{-1}$ and autocorrelation $cor(e_{it}, e_{it-k}) = \rho_T^k$, both of which increase at the rate that ρ_T approaches unity (violating Assumption C2 in Bai and Ng, 2002). By taking the logarithm the $IC\,(k)$ criteria ensures that the large common variance in e_{it} does not adversely impact model selection. However the high autocorrelation in e_{it} causes the $IC\,(k)$ criteria to overestimate the factor number because it is mistaken as a signal for a common factor. Note that scaling X_{it} by its inverse sample standard deviation does not solve the problem, because the standardized idiosyncratic error $e_{it}/sd_i(e_{it})$ still exhibits strong autocorrelation.

In Example 1 the idiosyncratic components have homogeneous persistency. If the autoregressive coefficients are heterogenous across i, then overestimation still occurs if $\hat{\ell}_1/\sum_{j=1}^{T}\hat{\ell}_j > g(N,T)$. This may happen if a non-negligible portion of the cross sections have large ρ_i, so a considerable subset of cross sections have e_{it} that exhibit high serial correlation. Simulation results (e.g., the "LEV" columns of Tables 5–7 in Section 5) clearly demonstrate the $IC\,(k)$ criteria overestimate the factor number in that case. Also overestimation may result if one e_{it} has ρ_i growing fast as N and T increase and that particular cross section has overwhelming idiosyncratic variance, as the following example demonstrates.

Example 2. Let $e_{it} = \rho_i e_{it-1} + \varepsilon_{it}$, where ε_{it} is $iid\ (0,1)$. Suppose that ρ_i are uniformly strictly smaller than 1 for all $i \geq 2$**,** except ρ_1 which converges to 1 from below as N and T increase. Suppose that there are no factors, so $X_{it} = e_{it}$ and $r = 0$. Let $\hat{\ell}_j$ be the eigenvalues (sorted from the largest to the smallest) of $N^{-1}\sum_{i=1}^{N} X_i X_i'$ as in Example 1, where $X_i = (X_{i1}, \ldots, X_{iT})'$. Then $\hat{\ell}_1 \geq \tilde{\ell}_1^*$, where $\tilde{\ell}_1^*$ is the largest eigenvalue of $N^{-1}X_1X_1'$, i.e., $\tilde{\ell}_1^* = N^{-1}X_1'X_1$. Next, we have

$$\frac{1}{T}\sum_{j=1}^{T}\hat{\ell}_j = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 = \frac{1}{NT}\sum_{t=1}^{T} e_{1t}^2 + \frac{1}{NT}\sum_{i=2}^{N}\sum_{t=1}^{T} e_{it}^2 = \frac{\tilde{\ell}_1^*}{T} + [c + o_p(1)],$$

where $c < \infty$ is the probability limit of $(NT)^{-1}\sum_{i=2}^{N}\sum_{t=1}^{T} e_{it}^2$. So

$$\frac{\hat{\ell}_1}{\sum_{j=1}^{T}\hat{\ell}_j} \geq \frac{T^{-1}\tilde{\ell}_1^*}{T^{-1}\tilde{\ell}_1^* + c + o_p(1)}.$$

If $N(1-\rho_1^2) \to 0$, then we have $E[T^{-1}\tilde{\ell}_1^*] = 1/[N(1-\rho_1^2)] \to \infty$, so with a nonzero probability, $T^{-1}\tilde{\ell}_1^*$ diverges to infinity. In that event, $\hat{\ell}_1/\sum_{j=1}^{T}\hat{\ell}_j$ is supported by $\frac{1}{2}$ for N and T large enough, implying that $\hat{\ell}_1/\sum_{j=1}^{T}\hat{\ell}_j > g(N,T)$ for large enough N and T (because $g(N,T) \to 0$). Thus, by the same algebra as in Example 1, the $IC\,(k)$ criteria overestimates the factor number with a nonzero probability. ∎

4 Consistent Filtering Procedures

The whitening filter has been used in many areas of econometrics. The basic idea is to remove most of the temporal dependence in the data (usually by an autoregression) in order to make the transformed data closer to white noise. (See Andrews and Monahan, 1992, for references to the use of whitening filters in the existing literature.) We employ an autoregressive filtering, and as such we must first focus on two preliminary specification issues: (i) whether to perform an individual or a pooled filtering, and (ii) AR lag order.

To address the first issue, first consider the transformed data

$$Z_{it} = X_{it} - \textstyle\sum_{j=1}^{p}\phi_{ij}X_{it-j}, \quad X_{it} = \lambda_i'F_t + e_{it},$$

where the filter ϕ_{ij} is permitted to be different for each i. Let r be the true factor number. Writing Z_{it} as

$$Z_{it} = \lambda_i'\left(F_t - \textstyle\sum_{j=1}^{p}\phi_{ij}F_{t-j}\right) + \left(e_{it} - \textstyle\sum_{j=1}^{p}\phi_{ij}e_{it-j}\right),$$

we see that if $\phi_{ij} = \phi_j$ (i.e., homogeneous for all i), then the common factors of Z_{it} are $F_t - \sum_{j=1}^{p}\phi_j F_{t-j}$ and the dimension of factors is preserved under the transformation unless some coordinates of the filtered factors $F_t - \sum_{j=1}^{p}\phi_j F_{t-j}$ are wiped out. Without the homogeneity restriction in the filtering coefficients, the filtered common component $\lambda_i'(F_t - \sum_{j=1}^{p}\phi_{ij}F_{t-j})$ cannot generally be expressed as a factor structure with the same dimension as F_t, though we can write it in terms of $r(p+1)$ common factors $(F_t', F_{t-1}', \ldots, F_{t-p}')'$.

Thus the filter must have coefficients common to all i in order to preserve the number of factors in the residuals Z_{it}.

The second issue is the choice of the lag order p. Conveniently, an AR(1) fitting ($p = 1$)

$$Z_{it} = X_{it} - \phi X_{it-1}. \tag{5}$$

is sufficient for consistent factor number estimation for many common panel processes, as we show below. Of course other orders p can also be used but we do not see any particular advantage in using more lags unless e_{it} is more than once integrated. Hence we focus only on AR(1) filtering throughout the paper. Note that ϕ may not be a "true" AR(1) coefficient and $X_{it} - \phi X_{it-1}$ may be dependent over t.

We now turn to deriving pooled AR(1) filters that lead to consistent factor number estimates. We consider two methods for choosing ϕ. The first is a nonrandom filter with $\phi = 1$ (first difference); the second is a data-dependent filter that uses the LSDV estimator, $\hat{\phi}_{lsdv}$, obtained by regressing X_{it} on X_{it-1} and including individual intercepts. For this latter filter, we demonstrate below that consistent factor number estimation does not require $\hat{\phi}_{lsdv}$ to be consistent for a "true"AR(1) coefficient if one exists.

The rest of this section is organized as follows. We first derive the consistency conditions for the first differenced filter in section 3.1. We then consider other nonrandom filters in section 3.2 as an intermediate step toward deriving the conditions required for the consistency of the LSDV filtering method, which is addressed in the final subsection.

Our consistency results are built on the main findings of BN (2002). More precisely, we will show that properly transformed data satisfy their assumptions (Assumptions A–D of BN, 2002.) For completion we present these assumptions here as the definition of "regularity":

Definition (BN-regularity). A common factor process $\{F_t\}$, idiosyncratic error processes $\{e_{it}\}$, and factor loadings $\{\lambda_i\}$ are said to be *BN-regular* if

A. $E\|F_t\|^4 < \infty$ and both $T^{-1}\sum_{t=1}^{T} F_t F_t'$ and $T^{-1}\sum_{t=1}^{T} E(F_t F_t')$ converge to a nonrandom strictly positive definite matrix as $T \to \infty$;

B. $\|\lambda_i\| \leq \bar{\lambda} < \infty$ and $N^{-1}\sum_{i=1}^{N} \lambda_i \lambda_i'$ converges to a nonsingular matrix as $N \to \infty$;

C. There exists a positive constant $M < \infty$ such that for all N and T,

   1. $E(e_{it}) = 0$, $E|e_{it}|^8 \leq M$;
   2. $N^{-1}\sum_{i=1}^{N} Ee_{is}e_{it} = \gamma_N(s,t)$, $|\gamma_N(s,s)| \leq M$ for all s, and $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T} |\gamma_N(s,t)| \leq M$ for all s and t;
   3. $E(e_{it}e_{jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq |\tau_{ij}|$ for some τ_{ij} and for all t; in addition, $N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} |\tau_{ij}| \leq M$;
   4. $E(e_{it}e_{js}) = \tau_{ij,ts}$ and $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} |\tau_{ij,ts}| \leq M$;

5. for every (t,s), $E|N^{-1/2}\sum_{i=1}^{N}(e_{is}e_{jt}-Ee_{is}e_{jt})|^4 \leq M$;

D. $E\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t e_{it}\right\|^2\right) \leq M.$

Let $\hat{k}(\phi)$ denote the factor number estimate obtained by applying a BN criterion to the filtered data $X_{it}-\phi X_{it-1}$.

4.1 Consistency of the First Differencing Filter

We first state the consistency of $\hat{k}(1)$, i.e., the factor number estimate applied to first differenced data. This is rather straightforward once we assume that $\{\lambda_i\}$, $\{\Delta F_t\}$ and $\{\Delta e_{it}\}$ are BN-regular. Note that the BN-regularity of the differenced processes allows for the original processes in levels to be integrated.

We consider panel data such that the first difference filter provides a consistent factor number estimate. For future reference, we state the following:

Assumption 1 *$\{\lambda_i\}$, $\{\Delta F_t\}$ and $\{\Delta e_{it}\}$ are BN-regular.*

This assumption is generally satisfied if the common factors and the idiosyncratic errors are at most once integrated. Related conditions and assumptions on the levels X_{it} are found in Bai and Ng (2004), where conditions are imposed to variables in levels and notably X_{it} can contain a linear trend. (See equations (1)–(3) of Bai and Ng, 2004.) When Assumption 1 is violated, higher order autoregressive filtering or differencing may solve the problem, though extension to this case is not considered in the present paper. The following result is known (Bai and Ng, 2004).

Theorem 1 (FD filter) *Under Assumption 1, $\hat{k}(1) \to_p r$.*

This theorem is proven directly from BN (2002). Bai (2004) and Bai and Ng (2004) note and make use of the same result for integrated common factors based on regularity assumptions for λ_i, F_t and e_{it} in levels.

According to Theorem 1, applying the BN criteria to the differenced data produces a consistent factor number estimator. However, there is a possibility of over-differencing, which may create negative serial correlation in the filtered panels. This over-differencing causes overestimation of the factor number in finite samples when the serial dependence is weak in e_{it}.

To avoid this problem we will consider filtering based on an LSDV fitting. However, as an intermediate step, we next consider nonrandom but parameter dependent filtering methods. This step is instructive because it allows us to discern how much bias and misspecification is permissible in a data-dependent procedure if the BN criteria are to be consistent.

4.2 Nonrandom Filters

Let ϕ_T be a nonrandom number, possibly dependent on the time series dimension. We ask under what conditions this nonrandom filter can yield a factor number estimate $\hat{k}(\phi_T)$ that is consistent, where $\hat{k}(\phi)$ is the BN factor number estimate using $X_{it} - \phi X_{it-1}$ as defined previously. To investigate this issue, we start by rewriting the filtered data $Z_{it} := X_{it} - \phi_T X_{it-1}$ as

$$Z_{it} = (X_{it} - X_{it-1}) + (1 - \phi_T) X_{it-1} = \Delta X_{it} + (1 - \phi_T) X_{it-1}. \tag{6}$$

For stationary and once integrated processes, we may assume the first term satisfies BN-regularity (i.e. λ_i, ΔF_t, and Δe_{it} are BN-regular as Assumption 1 states). Next we will outline the required regularity for the second term in (6). Let $\sigma^2_{e,T} = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T E e_{it}^2$ and $\Sigma_{FF,T} = T^{-1} \sum_{t=1}^T E(F_t F_t')$. Note that $\sigma^2_{e,T}$ may depend on N as well as T but the N subscript is omitted for notational brevity. (It does not depend on N if the random variables are iid across i.) Also $\sigma^2_{e,T}$ and $\Sigma_{FF,T}$ may diverge as $T \to \infty$. Further define $e^*_{it} = \sigma^{-2}_{e,T} e_{it}$ and $F^*_t = \Sigma^{-1}_{FF,T} F_t$. It is worth noting that e_{it} and F_t are divided by their variances rather than their standard deviations in the definition of e^*_{it} and F^*_t, so $(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T E e^{*2}_{it} = \sigma^{-2}_{e,T}$ and $T^{-1} \sum_{t=1}^T F^*_t F^{*\prime}_t = \Sigma^{-1}_{FF,T}$. The reason for this normalization is both to ensure that the variables e^*_{it} and F^*_t behave regularly when the original processes e_{it} and F_t are stationary, and to ensure that e^*_{it} and F^*_t are negligible (so they do not endanger the validity of the BN method) when e_{it} and F_t are integrated. Now Z_{it} of (6) can be rewritten as

$$Z_{it} = \lambda_i'[\Delta F_t + (1 - \phi_T)\Sigma_{FF,T} F^*_{t-1}] + [\Delta e_{it} + (1 - \phi_T)\sigma^2_{e,T} e^*_{it-1}], \tag{7}$$

so the common factors of the transformed series Z_{it} are $\Delta F_t + (1 - \phi_T)\Sigma_{FF,T} F^*_{t-1}$ and the idiosyncratic component is $\Delta e_{it} + (1 - \phi_T)\sigma^2_{e,T} e^*_{it-1}$. If ϕ_T is chosen such that $(1 - \phi_T)\Sigma_{FF,T}$ and $(1 - \phi_T)\sigma^2_{e,T}$ are bounded, then those new common factors and idiosyncratic components are likely to satisfy BN-regularity. For a rigorous treatment along this line, we make the following regularity assumption and present several remarks to show when it is satisfied.

Assumption 2 *For any constant b_1 and b_2, $\{\lambda_i\}$, $\{\Delta F_t + b_1 F^*_{t-1}\}$ and $\{\Delta e_{it} + b_2 e^*_{it-1}\}$ are BN-regular.*

Remark 1. If $\{F_t\}$ itself is BN-regular, then $\Delta F_t + b_1 F^*_{t-1}$ would also be BN-regular for any given b_1 as long as $T^{-1} \sum_{t=1}^T (F_t - \phi F_{t-1})(F_t - \phi F_{t-1})'$ has eigenvalues bounded sufficiently away from zero for all ϕ, which is the case as long as the generating shocks (e.g., v_t in $F_t = \phi F_{t-1} + v_t$) have non-negligible variation. If $\{e_{it}\}$ is BN-regular itself, then usually $\Delta e_{it} + b_2 e^*_{it}$ would also be BN-regular for any given b_2. ∎

Remark 2. When F_t is highly serially correlated, F_t itself may violate Condition A of the BN-regularity. However, in this case as well, $\{\Delta F_t + b_1 F^*_{t-1}\}$ is likely to satisfy the condition for any constant b_1. To

see this, let F_t be a scalar. Let $T^{-1}\sum_{t=1}^{T}(\Delta F_t)^2$ follow a law of large numbers and $F_t \sim I(1)$ such that $T^{-1/2}F_t$ follows an invariance principle. Let $\tilde{F}_t = \Delta F_t + b_1 F^*_{t-1}$. Then

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{F}_t^2 = \frac{1}{T}\sum_{t=1}^{T}(\Delta F_t)^2 + \frac{2b_1}{\Sigma_{FF,T}}\cdot\frac{1}{T}\sum_{t=1}^{T}\Delta F_t F_{t-1} + \frac{Tb_1^2}{\Sigma_{FF,T}^2}\cdot\frac{1}{T^2}\sum_{t=1}^{T}F_{t-1}^2,$$

where $T^{-1}\sum_{t=1}^{T}\Delta F_t F_{t-1}$ and $T^{-2}\sum_{t=1}^{T}F_{t-1}^2$ are $O_p(1)$. But $\Sigma_{FF,T}^{-1} = O(T^{-1})$, so

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{F}_t^2 = \frac{1}{T}\sum_{t=1}^{T}(\Delta F_t)^2 + O_p(T^{-1}),$$

thus condition A of the BN-regularity is satisfied by $\tilde{F}_t$. ∎

Remark 3. The key condition that would be violated by e_{it} when it is strongly serially correlated is Condition C2 of the BN-regularity. But $\Delta e_{it} + b_2 e^*_{it}$ still satisfies this condition for any constant b_2 when e_{it} is quite persistent as well. Let $\tilde{e}_{it} = \Delta e_{it} + b_2 e^*_{it-1}$, where $e^*_{it-1} = \sigma_{e,T}^{-2} e_{it-1}$ and $\sigma_{e,T}^2 = T^{-1}\sum_{t=1}^{T}Ee_{it}^2$. We want to see if condition C2 of the BN-regularity is satisfied, i.e.,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\left|\frac{1}{N}\sum_{i=1}^{N}Ee^*_{it}e^*_{is}\right| = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}|Ee^*_{it}e^*_{is}| \leq M < \infty.$$

Suppose that $e_{it} = \sum_{j=0}^{\infty} c_{Tj}\varepsilon_{it-j}$ where ε_{it} are iid and c_{Tj} may depend on T so local asymptotic analysis can be included. Then $Ee_{it}^2 = \sum_{j=0}^{\infty} c_{Tj}^2$, and

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}|Ee^*_{it}e^*_{is}| &\leq \frac{\sum_0^\infty c_j^2}{(\sum_0^\infty c_{Tj}^2)^2} + \frac{2\sum_{k=1}^{\infty}\sum_{j=0}^{\infty}|c_{Tj}c_{T,j+k}|}{(\sum_0^\infty c_{Tj}^2)^2} \\ &\leq \frac{1}{\sum_0^\infty c_j^2}\left[1 + \frac{2\sum_{j=0}^{\infty}|c_{Tj}|\sum_{k=1}^{\infty}|c_{T,j+k}|}{\sum_0^\infty c_{Tj}^2}\right].\end{aligned}$$

This is bounded if $\sum_0^\infty |c_{Tj}|$ and $\sum_0^\infty c_{Tj}^2$ are of the same order (as T increases), which is so for general square-summable processes. (Nonsummable processes cannot be handled by this argument.) More importantly, this converges to zero if $\sum_0^\infty c_{Tj}^2 \to \infty$ as $T \to \infty$ (though $\sum_{j=0}^{\infty} c_{Tj}^2$ is finite for all T) and if $\sum_{k=1}^{\infty}|c_{T,j+k}| \leq M^* \sup_{k\geq 0}|c_{T,j+k}|$ for some $M^* < \infty$, which happens if, for example, e_{it} is weakly-integrated. Once the regularity of e^*_{it} is shown, it is straightforward to show that $\Delta e_{it} + b_2 e^*_{it}$ satisfies condition C2 of the BN-regularity. ∎

Remark 4. If $e_{it} = e_{i0} + \sum_{s=1}^{t}\varepsilon_{is}$ (i.e., integrated), where ε_{it} are iid and independent of e_{i0}, then $\sigma_{e,T}^2 = O(\sigma_{0,T}^2) + O(T^2)$ (and it is the exact order) and $T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}|Ee_{it}e_{is}|$ is $O(T\sigma_{0,T}^2) + O(T^2)$ where $\sigma_{0,T}^2 = Ee_{i0}^2$. So

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}|Ee^*_{it}e^*_{is}| = O\left(\frac{T\sigma_{0,T}^2 + T^2}{[\sigma_{0,T}^2 + T^2]^2}\right) = O\left(\frac{\sigma_{0,T}^2/T + 1}{[\sigma_{0,T}^2/T + T]^2}\right) \to 0,$$

thus validating condition C2 of the BN-regularity for $\Delta e_{it} + b_2 e^*_{it-1}$ for any constant b_2. ∎

Remark 5. Assumption 2 can be satisfied if the persistency is different across i as well. For example, suppose that $e_{it} = \rho_i e_{it-1} + \varepsilon_{it}$, where $\rho_i = 1 - c_i/T^\alpha$ with $0 < \underline{c} \leq c_i \leq \bar{c} < \infty$ and $\varepsilon_{it} \sim iid\ (0, \sigma^2_\varepsilon)$. Also suppose that c_i is iid. Then $\sigma^2_{e,T} = \sigma^2_\varepsilon E[(1-\rho_i^2)^{-1}] = \sigma^2_\varepsilon T^\alpha E[c_i^{-1}(1+\rho_i)^{-1}]$. When $\rho_i > 0$ for all i, we furthermore have

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}|Ee_{it}e_{is}| \leq \sum_{r=0}^{\infty}|Ee_{it}e_{it-r}| = E\left[\frac{\sigma^2_\varepsilon}{(1-\rho_i)(1-\rho_i^2)}\right] = \sigma^2_\varepsilon T^{2\alpha} E[c_i^{-2}(1+\rho_i)^{-1}].$$

Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}|Ee^*_{it}e^*_{is}| \leq \frac{E[c_i^{-2}(1+\rho_i)^{-1}]}{E[c_i^{-1}(1+\rho_i)^{-1}]^2} \leq \frac{4E(c_i^{-2})}{(Ec_i^{-1})^2} \leq \frac{4\underline{c}^{-2}}{\bar{c}^{-2}} = 4(\bar{c}/\underline{c})^2 < \infty.$$

and condition C2 of the BN-regularity is satisfied. ∎

Given Assumption 2, the following is true.

Theorem 2 (Nonrandom filtering) *Given Assumption 2, if (i)* $(1-\phi_T)\Sigma_{FF,T}$ *converges to a finite limit, and (ii)* $(1-\phi_T)\sigma^2_{e,T} = O(1)$*, then* $\hat{k}(\phi_T) \to_p r$.

An intuitive explanation of the result is as follows. According to (7), the common factor of $X_{it} - \phi_T X_{it-1}$ is $\Delta F_t + (1-\phi_T)\Sigma_{FF,T} \cdot F^*_{t-1}$ and the idiosyncratic error is $\Delta e_{it} + (1-\phi_T)\sigma^2_{e,T} \cdot e^*_{it-1}$. Assumption 2 states that $\Delta F_t + b_1 F^*_{t-1}$ and $\Delta e_{it} + b_2 e^*_{it-1}$ are BN-regular for any constant b_1 and b_2. Conditions (i) and (ii) in the theorem impose the necessary restrictions on ϕ_T such that the terms corresponding to b_1 and b_2 behave as required under Assumption 2 in the limit as $T \to \infty$.

If e_{it} is integrated, then $\sigma^2_{e,T}$ increases at an $O(T)$ rate. So if $\phi_T = 1 - cT^{-1}$, for example, then conditions (i) and (ii) of the theorem are satisfied, and the BN method applied to $X_{it} - \phi_T X_{it-1}$ will yield a consistent factor number estimate. Though this approach is not practically useful (because in practice T is given, and the performance would depend on the constant c), it informs us of how much bias is allowed when we consider a data-dependent filtering procedure. This is the topic we investigate next.

4.3 Consistency of the LSDV Filter

In this subsection we consider a specific data-dependent filter, namely the LSDV filter. AR(1) LSDV filtering involves the following issues. Firstly, the AR(1) model may be misspecified in the sense that there is no ϕ such that $X_{it} - \phi X_{it-1}$ is independent over time; secondly, the LSDV estimator is biased (toward zero) even when X_{it} follows an AR(1); and lastly, the LSDV estimator is random. The problems of misspecification and bias were discussed in the previous subsection in a general context. In this subsection, we first show that the center of the LSDV estimator satisfies the conditions given in section 3.2 provided the degree of serial correlation is not stronger than weak integration. We then proceed to addressing the issue of randomness

in the LSDV filter $\hat{\phi}_{lsdv}$. By this method we will verify that the LSDV estimator $\hat{\phi}_{lsdv}$ satisfies all the required conditions in order for the filtered data $\hat{Z}_{it} := X_{it} - \hat{\phi}_{lsdv} X_{it-1}$ to satisfy BN-regularity unless the idiosyncratic errors are more persistent than permitted by the regularity assumptions as stated later.

Let us rewrite the filtered data as

$$\hat{Z}_{it} = \Delta X_{it} + (1 - \hat{\phi}_{lsdv}) X_{it-1} = \Delta X_{it} + (1 - \phi_T) X_{it-1} - (\hat{\phi}_{lsdv} - \phi_T) X_{it-1}, \tag{8}$$

where ϕ_T can be understood as the center of $\hat{\phi}_{lsdv}$.

Comparing (8) to (6), we see that in the data dependent framework the filtered data involves an additional term $(\hat{\phi}_{lsdv} - \phi_T) X_{it-1}$ compared to the nonrandom framework. Hence in order for the BN criteria to be consistent when applied to $\hat{Z}_{it}$, we would require not only that the center ϕ_T satisfies conditions (i) and (ii) of Theorem 2, but also that the variability of $\hat{\phi}_{lsdv}$ around ϕ_T is limited. The third term in the right hand side of (8) can be written as $(\hat{\phi}_{lsdv} - \phi_T)\sigma_{X,T}^2 \cdot \sigma_{X,T}^{-2} X_{it-1}$, where $\sigma_{X,T}^2 = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} E X_{it}^2$ (with the N subscript again suppressed for notational brevity). Because $\sigma_{X,T}^{-2} X_{it-1}$ is bounded and behaves regularly, we can imagine that if $(\hat{\phi}_{lsdv} - \phi_T)\sigma_{X,T}^2 = o_p(1)$, then this third term has negligible impact on the behavior of $\hat{Z}_{it}$ (See Theorem 3 and Remark 7.)

As discussed above the filter is employed to reduce the serial dependence in e_{it}. However, in order for this method to work, any strong serial correlation in F_t and e_{it} should be explained by an AR(1) structure. This is satisfied by a wide class of processes including integrated or nearly integrated processes. But we also note that this is not always possible, in particular if the process is I(2). Formally, we make the following assumption.

Assumption 3 *The common factors F_t and idiosyncratic errors e_{it} satisfy*

$$\frac{1}{T} \sum_{t=1}^{T} E[F_{t-1} \Delta F_t'] = O(1) \text{ and } \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E[e_{it-1} \Delta e_{it}] = O(1).$$

Remark 6. Assumption 3 is satisfied by rather general stationary and integrated (of order one) processes. If e_{it} are stationary for all i, then e_{it} obviously satisfies the assumption if the second moments are uniformly finite. If e_{it} is integrated for some i, such that $\Delta e_{it} = \sum_{j=0}^{\infty} c_{ij} \varepsilon_{it-j}$ and $\sup_{t>0,i} |E(e_{i0} \Delta e_{it})| < \infty$, where ε_{it} is $iid\ (0, \sigma^2)$ and $\sup_i \sum_{j=0}^{\infty} |c_{ij}| < \infty$, then $E[e_{it-1} \Delta e_{it}] = E[e_{i0} \Delta e_{it}] + \sum_{s=1}^{t-1} E[\Delta e_{is} \Delta e_{it}] = E[e_{i0} \Delta e_{it}] + \sum_{r=1}^{t-1} E[\Delta e_{it-r} \Delta e_{it}]$ and $E[\Delta e_{it-r} \Delta e_{it}] = \sigma^2 \times \sum_{j=0}^{\infty} c_{ij} c_{ij+r}$ (for $r > 0$). Thus

$$A_i := \frac{1}{T} \sum_{t=1}^{T} E[e_{it-1} \Delta e_{it}] = \frac{1}{T} \sum_{t=1}^{T} E[e_{i0} \Delta e_{it}] + \frac{\sigma^2}{T} \sum_{t=2}^{T} \sum_{r=1}^{t-1} \sum_{j=0}^{\infty} c_{ij} c_{ij+r},$$

so

$$|A_i| \leq \sup_{t>0,i} |E(e_{i0} \Delta e_{it})| + \sigma^2 \sum_{r=1}^{\infty} \sum_{j=0}^{\infty} |c_{ij} c_{ij+r}| \leq \sup_{t>0,i} |E(e_{i0} \Delta e_{it})| + \sigma^2 \left(\sum_{j=0}^{\infty} |c_{ij}| \right)^2,$$

which is uniformly bounded. The average of A_i is therefore bounded, so Assumption 3 holds. The common factors F_t are similarly treated (but without minding individual heterogeneity, of course). Note that Assumption 3 is not satisfied if e_{it} or F_t is more than once integrated. ∎

We have the following result.

Theorem 3 (LSDV filtering) *Given Assumptions 2 and 3, if* $\sigma^2_{e,T} = o(T)$ *and* $\Sigma_{FF,T} = o(T)$*, then* $\hat{k}(\hat{\phi}_{lsdv}) \to_p r$.

The theorem states that the Bai-Ng factor number estimator based on the AR(1) LSDV-filtered data $X_{it} - \hat{\phi}_{lsdv} X_{it-1}$ is consistent under suitable regularity. Unlike Assumptions 2 and 3, the regularity that $\sigma^2_{e,T} = o(T)$ and $\Sigma_{FF,T} = o(T)$ is binding. See the following remark.

Remark 7. Suppose that $e_{it} = \rho_T e_{it-1} + \varepsilon_{it}$ where ε_{it} is stationary. Let $\gamma_k = E[\varepsilon_{it}\varepsilon_{it-k}]$ and further assume $\varepsilon_{it} = \sum_{j=0}^{\infty} c_j w_{it}$ with w_{it} being iid and $\sum_0^\infty |c_k| < \infty$. Then we have $\sum_{k=0}^{\infty} |\gamma_k| < \infty$. If

$$T(1 - \rho_T) \to \infty, \tag{9}$$

then $\sigma^2_{e,T} = o(T)$. Note that $Ee^2_{it} = O((1-\rho_T)^{-1})$, so $\sigma^2_{e,T} = O((1-\rho_T)^{-1}) = o(T)$ if $T(1-\rho_T) \to \infty$. If $1 - \rho_T = O(T^{-\alpha})$, then this condition holds if $\alpha < 1$. The process for the common factor F_t is treated similarly. Note the importance of (9). ∎

Finally we note that filtering may impair the small sample performance of the BN criteria in some cases. For example, if the signal in the filtered factors is considerably smaller than the factors in levels and if the idiosyncratic component does not exhibit much serial correlation (that is, for some ϕ the variance of $F_t - \phi F_{t-1}$ is negligible compared to variance of $e_{it} - \phi e_{it-1}$), then the selection criteria may perform worse when applied to the filtered data than when applied to the levels data. However, the filtered data will still contain enough variation in common factors under Assumption 2 for consistency of the filtering method.[2]

4.4 Practical Issues and Dynamic Factors

While first differencing works well when the process is closer to unit root (e.g. $1 - \rho_T = O(T^{-1})$) or even integrated, the LSDV filtering typically performs better than first differencing if the process does not exhibit strong serial correlation. In practice the strength of the dependence is unknown, so it will be useful to provide a method which combines the two filtering methods and which is at least as good as the two filtering methods separately.

A simple way to enhance the small sample performance is to choose the minimum factor number estimate from the first differencing and the LSDV filtering, i.e.,

$$\hat{k}_{min} = \min\left\{\hat{k}(1), \hat{k}(\hat{\phi}_{lsdv})\right\}. \tag{10}$$

This "minimum rule" is justified by the fact that serial correlation usually causes overestimation rather than underestimation of the factor number, and should perform well provided that the differenced or quasi-differenced factors exhibit sufficient signal in the transformed data.

[2]The same role is played by Assumption B(ii) of Bai and Ng (2004).

The filtering procedures proposed above are useful for detecting the number of static factors. This method can also be applied to the restricted dynamic models (Amengual and Watson, 2007; Bai and Ng, 2007; Hallin and Liska, 2007) based on $F_t = \sum_{j=1}^{p} \Pi_j F_{t-j} + G\eta_t$, where η_t is $q \times 1$ and G is $r \times q$ with full column rank. (See Hallin and Liska, 2007, for a discussion of the distinction between static, restricted dynamic and dynamic factor models.) For this model, filtering the data using the methods suggested above preserves the dimensions of both the static and dynamic factors because

$$F_t - \phi F_{t-1} = \sum_{j=1}^{p} \Pi_j (F_{t-j} - \phi F_{t-j-1}) + G(\eta_t - \phi \eta_{t-1}),$$

where the transformed static factors $F_t - \phi F_{t-1}$ are still $r \times 1$ and the transformed primitive shocks $\eta_t - \phi \eta_{t-1}$ are still $q \times 1$. Further analysis of dynamic factor models is reserved for future research.

5 Monte Carlo Studies

We conduct simulations to illustrate our results. We consider the following data generating process:

$$\begin{aligned} X_{it} &= \textstyle\sum_{j=1}^{r} \lambda_{ji} F_{jt} + e_{it}, \;\; F_{jt} = \theta F_{jt-1} + v_{jt} \text{ for } j = 1, \ldots, r; \\ e_{it} &= \rho_i e_{it-1} + \varepsilon_{it}, \;\; \varepsilon_{it} = \sum_{k=-J, k\neq 0}^{J} \beta u_{i-k,t} + u_{it} \text{ for } J = \lfloor N^{1/3} \rfloor. \end{aligned}$$

where $\lfloor \cdot \rfloor$ denotes the largest integer not exceeding the argument. The DGP for e_{it} is similar to that employed in the Monte-Carlo study of BN (2002). Note that e_{it} can exhibit cross sectional and time series dependence through β and ρ_i, respectively. The random variables u_{it} are independently drawn from $N(0, s_i^2)$ for some s_i for each i. The v_{jt} are drawn from $N(0,1)$, and we draw λ_{ji} from $N\left(0, r^{-1/2}\right)$ for all i and $j = 1, \ldots, r$. This ensures that the variance of the common component relative to the variance of the idiosyncratic component is invariant to the number of factors. (Note BN, 2002, make a similar normalization by adjusting the variance of e_{it}.) Throughout we set $r = 2$, so there are two factors. We consider the $IC_{p2}(k)$ criterion only because it uses the largest penalty among $IC_{p1}(k)$, $IC_{p2}(k)$ and $IC_{p3}(k)$, so the probability of overestimation is the smallest.

We consider the following four cases to investigate the finite sample performance of the proposed methods.

Case 1: Asymptotic justification of filtering methods. We first consider a DGP design where $\beta = 0$ (no cross section dependence in e_{it}), $\rho_i = 0$ (no serial correlation in e_{it}), $s_i = 1$ for all i (no heteroskedasticity in the primitive idiosyncratic shock), and $\theta = 0.9$ (high serial correlation in F_t). This is a "bad case" for filtering methods compared to the BN method using the levels data because filtering can considerably reduce the signal in the transformed common factors while common factors in levels contains a relatively larger signal. The purpose of the present simulation is to demonstrate that filtering methods work asymptotically

Table 4: Case 1. $\rho_i = 0$, $s_i = 1$, $\theta = 0.9$, $\beta = 0$, $\sigma_v^2 = 1$, $r = 1$

		LEV			FD			AR1			MIN		
N	T	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
25	25	3.3	96.7	0.0	42.5	53.8	3.7	19.0	79.8	1.2	43.1	55.8	1.1
50	25	0.7	99.3	0.0	15.2	80.7	4.1	5.0	94.0	1.0	15.4	83.7	0.9
100	25	0.0	100	0.0	4.9	93.5	1.6	1.0	98.7	0.3	4.9	94.8	0.3
25	50	0.0	100	0.0	16.9	82.9	0.2	3.6	96.3	0.1	16.9	83.0	0.1
50	50	0.0	100	0.0	3.2	96.8	0.0	0.3	99.7	0.0	3.2	96.8	0.0
100	50	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0
25	100	0.0	100	0.0	4.8	95.2	0.0	0.5	99.5	0.0	4.8	95.2	0.0
50	100	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0
100	100	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0

$<$, $=$, $>$: underestimation, correct estimation, and overestimation, respectively.

if there is enough variation in v_{jt}. We set $s_i = 1$ here in order to avoid any complication arising from cross section heteroskedasticity, which typically results in overestimation.

We expect that the BN method in levels outperforms the filtering methods and filtering may lead to underestimation in small samples, but as long as the primitive common shocks v_{jt} have enough variation, filtering methods still estimate the factor numbers consistently. This expectation is verified by Table 4, which reports correct and incorrect selection frequencies in percentage for the BN method in levels (LEV), the first difference filtering (FD), the LSDV filtering (AR1), and the minimum rule (MIN) applied to FD and AR1. For small samples, the probability of underestimation by the filtering methods (FD, AR1 and MIN) is higher than that for LEV, but as N and T increase, the filtering methods achieve consistency. Results do not change in any considerable way when weak serial and cross-sectional dependence and mild heteroskedasticity is introduced.

Case 2: Positive idiosyncratic serial dependence. In this case both the common factors and idiosyncratic components exhibit moderate serial dependence; the AR(1) parameters ρ_i are independently and identically distributed as $U(0.5, 0.7)$, $\theta = 0.5$ and we set $\beta = 0.1$ to allow cross section dependence in the idiosyncratic errors. We let $s_i \sim U(0.5, 1.5)$ in the present and following settings. Both ρ_i and s_i are not re-drawn for each replication of the simulation.

The results are reported in Table 5. The BN method using levels data considerably overestimate the factor number, while both filtering methods show good finite sample properties. Using the minimum rule performs best in this example. If we increase the mean of the AR1 parameters ρ_i, while maintaining the same dispersion, the FD filter begins to outperform the AR1 filter. However in this case the minimum rule

Table 5: Case 2. $\rho_i \sim U(0.5, 0.7)$, $s_i \sim U(0.5, 1.5)$, $\theta = 0.5$, $\beta = 0.1$, $\sigma_v^2 = 1$, $r = 2$.

		LEV			FD			AR1			MIN		
N	T	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
25	25	2.2	13.6	84.2	12.9	73.8	13.3	16.1	72.1	11.8	18.5	76.1	5.4
50	25	0.0	0.4	99.6	1.1	90.0	8.9	2.0	88.7	9.3	2.1	95.2	2.7
100	25	0.0	0.0	100	0.3	99.0	0.7	0.5	97.6	1.9	0.6	99.3	0.1
25	50	0.1	14.2	85.7	1.1	84.9	14.0	2.3	89.4	8.3	2.4	92.1	5.5
50	50	0.0	13.8	86.2	0.0	99.5	0.5	0.0	99.7	0.3	0.0	99.9	0.1
100	50	0.0	1.0	99.0	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0
25	100	0.4	22.1	77.5	0.2	85.8	14.0	0.3	90.4	9.3	0.3	93.1	6.6
50	100	0.0	20.3	79.7	0.0	99.6	0.4	0.0	99.9	0.1	0.0	100	0.0
100	100	0.0	46.7	53.3	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0

$<, =, >$: underestimation, correct estimation, and overestimation, respectively.

continues to perform best. Likewise if we reduce the mean of ρ_i while maintaining the same degree of dispersion, the AR1 filter begins to out-perform the FD filter, but again the minimum rule performs best.

Case 3: More heterogenous idiosyncratic serial dependence. As more heterogeneity in ρ_i is introduced, the performance of FD deteriorates compared to AR1. For example, when ρ_i is drawn independently from the $U(-0.1, 0.9)$ distribution (with other settings the same as in Case 2), the AR1 filter outperforms the FD filter by a wide margin in small samples (with $N = 25$ or $T = 25$), as shown in in Table 6. This would be explained by the fact that the FD filter induces large negative autocorrelation in the idiosyncratic series with small ρ_i values, whereas the data-dependent filter leaves less residual correlation in the treated cross sections. The minimum rule improves on the AR1 filter slightly, while $IC_{p2}(k)$ in levels performs poorly. (In unreported simulations LEV performs well when T is very large relative to N.)

Case 4: Extremely heterogenous autoregressive parameters. In this set of simulations we draw ρ_i from $iid\ U(-0.1, 0.1)$ for $i = 1, \ldots, \frac{1}{2}N$ and $iid\ U(0.7, 0.9)$ for $i = \frac{1}{2}N + 1, \ldots, N$.[3] Other settings are as in Case 2. As in Case 3 above, in this framework there is a marked disparity between the degree of serial dependence in the AR1 parameter. But within each subgroup the degree of heterogeneity is low. Table 7 exhibits the results. Notably the AR1 filter performs much better than the FD or LEV methods. As in case 3, this result is attributable to the FD filter inducing large negative correlation in many of the cross sections. The effect is more noticeable in this DGP because more cross sections have an AR(1) coefficient close to zero. (As in Table 6, in unreported simulations we find that the $IC_{p2}(k)$ in levels performs well when T is

[3] We thank an anonymous referee from an earlier submission for suggesting this DGP.

Table 6: Case 3. $\rho_i \sim U(-0.1, 0.9)$, $s_i \sim U(0.5, 1.5)$, $\theta = 0.5$, $\beta = 0.1$, $\sigma_v^2 = 1$, $r = 2$.

		LEV			FD			AR1			MIN		
N	T	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
25	25	4.8	26.3	68.9	16.8	57.9	25.3	23.7	68.9	7.4	27.4	67.6	5.0
50	25	0.0	1.9	98.1	2.3	82.8	14.9	2.7	88.1	9.2	3.2	93.9	2.9
100	25	0.0	0.7	99.3	0.8	94.6	4.6	1.3	95.6	3.1	1.5	98.2	0.3
25	50	0.3	9.9	89.8	2.4	69.3	28.3	5.2	88.2	6.6	5.5	89.7	4.8
50	50	0.0	2.8	97.2	0.2	99.0	0.8	0.0	99.0	1.0	0.2	99.8	0.0
100	50	0.0	0.7	99.3	0.0	100	0.0	0.0	99.8	0.2	0.0	100	0.0
25	100	0.1	4.9	95.0	0.7	68.6	30.7	0.6	95.2	4.2	0.9	95.6	3.5
50	100	0.0	0.3	99.7	0.0	98.6	1.4	0.0	98.9	1.1	0.0	99.9	0.1
100	100	0.0	1.2	98.8	0.0	100	0.0	0.0	100	0.0	0.0	100	0.0

$<, =, >$: underestimation, correct estimation, and overestimation, respectively.

very large relative to N.)

We also considered other parameters settings and DGPs, as well as other estimation methods such as the Hallin and Liska (2007) cross validation approach. In the interests of brevity we do not report the results here.[4] The Hallin-Liska method (applied to the data in levels) indeed improves upon the BN criteria in most considered cases, but it does not resolve the overestimation problem created by high idiosyncratic autocorrelation in small samples to the same extent as the filtering methods. (For example, in the case 4 setting with $N = T = 100$, the correct selection frequency is 61% for the Hallin-Liska method applied to levels data, the $IC_{p2}(k)$ applied to levels never selects the correct factor number in the 1000 replications, while both filtering methods have a correct selection frequency of over 99%.) Interestingly we found that a simple combination of the filtering approach with the Hallin-Liska method does not improve the accuracy of the filtering approach. A further and more extensive comparison is worthy of further research.

6 Revisiting the Empirical Examples

We now consider the performance of LSDV filtering in each of the examples introduced in section 2 above. In each case the LSDV-filtered panels yields a more credible factor number estimate than the methods

[4]These additional results are available from the author upon request. The Hallin-Liska procedure requires the practitioner to specify both a sequence of penalty functions and a sequence of sub-samples. We mimicked the sequences adopted by Hallin and Liska (2007) in their MC study. We first set the penalty functions $c = (0.01, 0.02, ..., 3)$. Next, each of the the sub-samples begin at $(i, t) = (1, 1)$ and end at $(i, t) = (bN, bT)$, where $b = (0.7, 0.8, 0.9, 1)$, such that there are three sub-samples used in the cross validation procedure.

Table 7: Case 4. $\rho_i \sim U(-0.1, 0.1)$ for $i \leq \frac{1}{2}N$ and $\rho_i \sim U(0.7, 0.9)$ for $i > \frac{1}{2}N$, $s_i \sim U(0.5, 1.5)$, $\theta = 0.5$, $\beta = 0.1$, $\sigma_v^2 = 1$, $r = 2$.

		LEV			FD			AR1			MIN		
N	T	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
25	25	4.8	23.4	71.8	17.1	40.6	42.3	25.0	60.6	14.4	28.4	60.7	10.9
50	25	0.1	4.0	95.9	2.3	51.3	46.4	4.1	80.2	15.7	4.8	84.3	10.9
100	25	0.0	0.1	99.9	0.7	79.1	20.2	1.5	86.6	11.9	1.8	95.4	2.8
25	50	0.0	5.5	94.5	2.0	43.4	54.6	5.6	80.8	13.6	5.9	82.7	11.4
50	50	0.0	1.7	98.3	0.3	84.0	15.7	0.4	97.0	2.6	0.5	98.8	0.7
100	50	0.0	0.0	100	0.0	96.7	3.3	0.0	98.4	1.6	0.0	99.8	0.2
25	100	0.0	1.3	98.7	0.5	36.9	62.6	1.3	85.0	13.7	1.4	86.7	11.9
50	100	0.0	0.3	99.7	0.0	63.6	36.4	0.0	97.6	2.4	0.0	98.5	1.5
100	100	0.0	0.0	100	0.0	99.6	0.4	0.0	100	0.0	0.0	100	0.0

$<, =, >$: underestimation, correct estimation, and overestimation, respectively.

based on either first-differencing or the data in levels. In order to assess the impact of serial dependence on the various factor number selection methods, we also report fitted AR(1) coefficients for the estimated idiosyncratic components of the panels.[5] We report X-differenced estimates of the AR(1) coefficient, which exhibit less small sample bias than LSDV (Han, Phillips and Sul, 2011). In order to give an indication of the amount of heterogeneity in the idiosyncratic panel, we report the minimum and maximum AR(1) coefficient across N separate univariate regressions. In order to give an indication of the average degree of persistence, we report the median coefficient across the N separate univariate regressions, as well as a pooled AR(1) coefficient.

Table 8 exhibits the results for the disaggregate PCE growth dataset. It reports the estimated factor numbers using the LSDV filter (AR1), as well as two of the standard methods from section 2 above, namely first-differencing (FD) and cross sectional standardization of the panel in levels (level + CSS). The LSDV filter yields a single factor, while the other standard methods select five factors. Evidently there is a very large degree of heterogeneity in the idiosyncratic component, as well as a moderate degree of persistence on average. This suggests that serial dependence is the reason that the Bai-Ng criteria applied to either the data in levels or first differences leads to the maximum number of factors being selected. In contrast, by filtering the panel we mitigate the adverse effect of the serial dependence in the idiosyncratic component on the factor number estimate, thereby gaining the more plausible factor number estimate of one.

[5]The AR(1) coefficients are fitted using the estimated idiosyncratic components of the panel, which are obtained by (1) applying principal components to the LSDV-filtered panel, (2) selecting the factor number using $IC_{p2}(k)$, and (3) recoloring the first-stage idiosyncratic components using the LSDV-estimated AR(1) coefficient.

Table 8: Serial dependence and estimated factor number; PCE consumption growth

$IC_{p2}(k)$ with 5 factors maximum; $N = 182, T = 28$

	estimated factor number			LSDV filter	X-differencing fitted univariate AR(1)			
sample	level + CSS	FD	AR1	$\hat{\phi}_{\text{lsdv}}$	pooled	minimum	median	maximum
1983-2010	5	5	1	0.386	0.378	-0.430	0.402	0.993
sub-sample robustness								
1983-2008	5	5	1	0.387	0.384	-0.435	0.401	1.003
1983-2009	5	5	1	0.406	0.392	-0.531	0.428	0.967
1984-2010	5	4	1	0.385	0.378	-0.430	0.420	1.006

"level + CSS" denotes Bai-Ng in levels with each cross section standardized

Table 9 reports the results for the Metropolitan CPI-U panel. There is not much heterogeneity in the degree of persistency in this example over the 1978-2010 period compared with PCE consumption growth panel, however the heterogeneity increases, and the degree of persistency decreases, for samples that begin after 1980. Applying the $IC_{p2}(k)$ criterion to the LSDV-filtered panel yields a factor number estimate of two. This result holds for both the full sample, as well as all the sub-samples considered. Again, the results from the LSDV-filtered data are more credible that those of the conventional methods, both because the selected factor number is small relative to the maximum number of factors permitted, and because the factor number is robust to different subsamples considered.

Table 9: Serial dependence and estimated factor number; Metropolitan CPI-U inflation

$IC_{p2}(k)$ with 5 maximum factors; $N = 23, T = 32$.

	estimated factor number			LSDV filter	X-differencing fitted univariate AR(1)			
sample	level + CSS	FD	AR1	$\hat{\phi}_{\text{lsdv}}$	pooled	minimum	median	maximum
1979-2010	5	3	2	0.702	0.639	-0.062	0.560	0.983
sub-sample robustness								
1981-2010	4	3	2	0.415	0.544	-0.232	0.443	0.847
1982-2009	4	2	2	0.375	0.611	0.017	0.528	0.826
1979-2008	5	4	2	0.716	0.783	0.195	0.704	1.103

"level + CSS" denotes Bai-Ng in levels with each cross section standardized

Table 10 shows the results for the NAICs industry growth example. There is substantial heterogeneity in the degree of serial dependence between cross sections, while the average amount of persistence is moderate

(note the pooled AR(1) coefficient is approximately 0.53). Applying the $IC_{p2}(k)$ criterion to the LSDV-filtered panel yields a factor number estimate of one. Again, this result holds for both the full sample and the subsamples. Evidently there is much less persistency in the estimated idiosyncratic components, and there is a substantial degree of heterogeneity in the estimated idiosyncratic AR(1). The moderate degree of persistency in the idiosyncratic component is likely to be the reason that the Bai-Ng criteria applied to either the data in levels or first differences leads to the a large number of factors being selected.

Table 10: Serial dependence and estimated factor number; Industry employment growth

$IC_{p2}(k)$ with 5 maximum factors; $N = 92, T = 19$.

	estimated factor number			LSDV filter	X-differencing fitted univariate AR(1)			
sample	level	FD	AR1	$\hat{\phi}_{\text{lsdv}}$	pooled	minimum	median	maximum
1991-2009	5	4	1	0.528	0.447	-0.293	0.566	1.003
sub-sample robustness								
1993-2009	5	5	1	0.558	0.457	-0.302	0.542	1.052
1991-2009	5	3	1	0.472	0.415	-0.511	0.499	1.108

7 Conclusion

Factor models are increasingly being used in empirical econometrics, and are often employed to summarize comovements in a glut of data using a handful of estimated factors. An integral part of estimating the factors is estimating the dimension of the factor space, i.e., the number of common factors underlying the panel. Existing factor selection criteria require large N and T for consistency, and may be inaccurate when one or both of the dimensions of the panel is moderate to small. Using a local alternative approach we analyze the impact of serial correlation on the popular Bai and Ng factor number selection criteria. We demonstrate that even a moderate degree of serial correlation in the idiosyncratic errors (relative to the given sample size) can cause the Bai-Ng criteria to overestimate the true number of factors. To overcome this problem, we suggest filtering the panel prior to applying Bai and Ng's method. We theoretically analyze how the filtering method can work for general processes with serial correlation and verify the applicability of the method by simulations. Using several different empirical examples we demonstrate how LSDV filtering yields reasonable factor number estimates when conventional methods yield estimates that are too large to be credible.

A Mathematical Proofs

Proof of Theorem 1. See Theorem 2 of Bai and Ng (2002). ∎

Proof of Theorem 2. Let $Z_{it} = X_{it} - \phi_T X_{it-1}$ as before. Let $b_{1,T} = (1-\phi_T)\Sigma_{FF,T}$ and $b_{2,T} = (1-\phi_T)\sigma^2_{e,T}$. By (7), we have $Z_{it} = \lambda_i'(\Delta F_t + b_{1,T}F^*_{t-1}) + (\Delta e_{it} + b_{2,T}e^*_{it})$. Assumption 2 and condition (i) of the theorem imply that $\{\Delta F_t + b_{1,T}F^*_{t-1}\}$ is BN-regular, and Assumption 2 and condition (ii) of the theorem imply that $\{\Delta e_{it} + b_{2,T}e^*_{it-1}\}$ is also BN-regular. The result follows from Bai and Ng (2002, Theorem 2) again. ∎

Before proving Theorem 3, we present a slightly more general result on data-dependent filtering. Let $\hat{\phi}$ be a random variable and ϕ_T a nonrandom quantity. The result to be presented below states that if $\hat{\phi}$ and ϕ_T are sufficiently close, then filtering based on $\hat{\phi}$ and filtering based on ϕ_T give the same probability limit. Let $\hat{Z}_{it} = X_{it} - \hat{\phi}X_{it-1}$. For panel data X_{it}, let $V_{NT}(k; X_{it}) = \min_{F\in\mathbb{R}^{T\times k}}(NT)^{-1}\sum_{i=1}^N X_i'M_F X_i$, where $X_i = (X_{i1}, \ldots, X_{iT})'$ and $M_F = I - F\left(F'F\right)^{-1}F'$. Let $h_{NT}(k; X_{it}) = V_{NT}(k; X_{it}) - V_{NT}(r; X_{it})$.

Lemma A.1 *Under the assumptions for Theorem 2, if* $(\hat{\phi} - \phi_T)\sigma^2_{X,T} \to_p 0$, *then* $\hat{k}(\hat{\phi}) \to_p r$.

Proof. Let $\hat{a} = (\hat{\phi} - \phi_T)\sigma^2_{X,T}$ for notational brevity, such that $\hat{a} \to_p 0$ under the supposition of the lemma. Also let $\hat{h}(k) = h_{NT}(k; \hat{Z}_{it})$ and $h(k) = h_{NT}(k; Z_{it})$. The goal is to show that (i) $\hat{h}(k)$ does not shrink to zero for $k < r$, and (ii) $\hat{h}(k) = O(C_{NT}^{-2})$ for $k > r$. (See Bai and Ng, 2002, proof of Theorem 2.) Note that $\hat{Z}_{it} = Z_{it} + \hat{a}X^*_{it-1}$, where $X^*_{it} := X_{it}/\sigma^2_{X,T}$.

(i) When $k < r$, $h(k)$ does not shrink to zero by Assumption 2, so it suffices to show that $\hat{h}(k) - h(k) \to_p 0$. But $\hat{h}(k) - h(k) = \hat{\xi}_r - \hat{\xi}_k$, where

$$\hat{\xi}_j = \max_{F\in\mathbb{R}^{T\times j}} \frac{1}{NT}\sum_{i=1}^N \hat{Z}_i'P_F\hat{Z}_i - \max_{F\in\mathbb{R}^{T\times j}} \frac{1}{NT}\sum_{i=1}^N Z_i'P_F Z_i.$$

So the proof can be done by showing that $\hat{\xi}_j \to_p 0$ for every $j \le r$. This part is easy: Because $|\max f - \max g| \le \max|f - g|$, we have

$$|\hat{\xi}_j| \le \max_{F\in\mathbb{R}^{T\times j}} \left|\frac{1}{NT}\sum_{i=1}^N (\hat{Z}_i'P_F\hat{Z}_i - Z_i'P_F Z_i)\right| \to_p 0,$$

where we used the fact that $\hat{a} \to_p 0$ and all the averages are stochastically bounded.

(ii) For the case with $k > r$, write $\hat{Z}_{it}$ as

$$\hat{Z}_{it} = \lambda_i'(F_t - \hat{\phi}F_{t-1}) + (e_{it} - \phi_T e_{it-1}) - \hat{a}e^*_{it-1}.$$

The common factors $F_t - \hat{\phi}F_{t-1}$ can be written as $(F_t - \phi_T F_{t-1}) - \hat{a}F^*_{t-1}$, which satisfies Assumption A of Bai and Ng (2002) because $F_t - \phi_T F_{t-1}$ satisfies it and $\hat{a} = o_p(1)$. Next, both $u_{it} := e_{it} - \phi_T e_{it-1}$ and e^*_{it-1} satisfy the assumptions of Bai and Ng (2002), where the idiosyncratic error of $\hat{Z}_{it}$ is $u_{it} - \hat{a}e^*_{it-1}$.

Then Theorem 1 of Bai and Ng (2002) still holds with this idiosyncratic error, but some part of the proof of Bai and Ng's Lemma 4 should be redone. More precisely, we need to show that

$$\max_{F\in\mathbb{R}^{T\times k}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_i - \hat{a}e^*_{i,-1})' M_F (u_i - \hat{a}e^*_{i,-1}) = O_p(C_{NT}^{-2}),$$

which corresponds to (1) of Bai and Ng (2006). But this holds because both u_{it} and and e^*_{it-1} satisfy the assumptions of Bai and Ng (2002, 2006) and $\hat{a} \to_p 0$. ∎

Now we prove that the LSDV estimator $\hat{\phi}_{lsdv}$ obtained by regressing X_{it} on X_{it-1} satisfies the assumptions for Theorem 2 and Lemma A.1 under suitable assumptions. Let

$$\hat{\Gamma}_0 = \frac{\sigma_{X,T}^{-2}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it-1}^2, \quad \hat{\Gamma}_1 = \frac{\sigma_{X,T}^{-2}}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it-1}\tilde{X}_{it},$$

where the "˜" notation stands for the within-group transformation. Note that $E\hat{\Gamma}_0$ is nonsingular. The AR(1) LSDV estimator is $\hat{\phi}_{lsdv} = \hat{\Gamma}_0^{-1}\hat{\Gamma}_1$. Let $\phi_{lsdv} = \Gamma_{0,T}^{-1}\Gamma_{1,T}$, where $\Gamma_{j,T} = E\hat{\Gamma}_j$. We will show that ϕ_{lsdv} and $\hat{\phi}_{lsdv}$ satisfy the conditions for Theorem 2 and Lemma A.1 under regularity.

Lemma A.2 *If $T^{-1}\sigma_{X,T}^2 = O(1)$, then under Assumption 3, $(1-\phi_{lsdv})\sigma_{X,T}^2 = O(1)$.*

Proof. We have $(1-\phi_{lsdv})\sigma_X^2 = \Gamma_{0,T}^{-1}(\Gamma_{0,T} - \Gamma_{1,T})\sigma_{X,T}^2$. Because $\Gamma_{0,T}^{-1}$ is finite, it suffices to show that $(\Gamma_{0,T} - \Gamma_{1,T})\sigma_{X,T}^2 = O(1)$, i.e.,

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E\tilde{X}_{it-1}\Delta\tilde{X}_{it} = O(1).$$

We use

$$\frac{1}{T}\sum_{t=1}^{T} E\tilde{X}_{it-1}\Delta\tilde{X}_{it} = \frac{1}{T}\sum_{t=1}^{T} EX_{it-1}\Delta X_{it} - \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} EX_{it-1}\Delta X_{is}.$$

This is bounded by Assumption 3. ∎

Let $X^*_{it} = X_{it}/\sigma_{X,T}^2$ as before.

Lemma A.3 *Suppose that (i) $\mathrm{var}(X_{it-1}^2) \le M\sigma_{X,T}^4$, and (ii) $\left|\sum_{k=1}^{\infty}\mathrm{cov}(X_{it}^2, X_{it+k}^2)\right| \le M\sigma_{X,T}^4$ for all i and t for some $M < \infty$. If $T^{-1}\sigma_{X,T}^2 = o(1)$, then $(\hat{\phi}_{lsdv} - \phi_{lsdv})\sigma_{X,T}^2 = o_p(1)$.*

Proof. Let $\hat{\phi} = \hat{\phi}_{lsdv}$ and $\phi = \phi_{lsdv}$ for notational simplicity. We have $\hat{\phi} - \phi = \hat{\Gamma}_0^{-1}(\hat{\Gamma}_1 - \phi\hat{\Gamma}_0)$. Note that $\Gamma_1 - \phi\Gamma_0 = 0$. Because $\hat{\Gamma}_0^{-1}$ is $O_p(1)$, we shall show that $(\hat{\Gamma}_1 - \phi\hat{\Gamma}_0)\sigma_X^2 = o_p(1)$, i.e.,

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it-1} - \bar{X}_{i,-1})(U_{it} - \bar{U}_i) = o_p(1), \quad U_{it} = X_{it} - \phi X_{it-1}. \tag{11}$$

Because $U_{it} = \Delta X_{it} + aX^*_{it-1}$, where $a = (1-\phi)\sigma_X^2 = O(1)$ by Lemma A.2, (11) can be proved by showing that

$$Y_a := \frac{1}{N}\sum_{i=1}^{N} Y_{ai} = o_p(1), \quad Y_{ai} = \frac{1}{T}\sum_{t=1}^{T}(X_{it-1}\Delta X_{it} - EX_{it-1}\Delta X_{it}),$$
$$Y_b := \frac{a}{N}\sum_{i=1}^{N} Y_{bi} = o_p(1), \quad Y_{bi} = \frac{1}{T}\sum_{t=1}^{T}(X_{it-1}X^*_{it-1} - EX_{it-1}X^*_{it-1}),$$
$$Y_c := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{cit} = o_p(1), \quad Y_{cit} = \frac{1}{T}\sum_{s=1}^{T}(X_{it-1}\Delta X_{is} - EX_{it-1}\Delta X_{is}),$$
$$Y_d := \frac{a}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{dit} = o_p(1), \quad Y_{dit} = \frac{1}{T}\sum_{s=1}^{T}(X_{it-1}X^*_{is-1} - EX_{it-1}X^*_{is-1}).$$

Because Y_a, Y_b, Y_c and Y_d are averages over i, we will show that $Y_{ji} = o_p(1)$ for all $j = a, b, c, d$, where the convergence holds uniformly over all i. Furthermore, because $EY_{ji} = 0$ for $j = a, b, c, d$, we will show that $EY_{ji}^2 \to 0$ for $j = a, b, c, d$, where the convergence and boundedness are uniform in i.

For Y_{ai}, we have $Y_{ai} = T^{-1}\left[X_{it-1}(X_{iT} - X_{i0}) - EX_{it-1}(X_{iT} - X_{i0})\right]$, so

$$EY_{ai}^2 \leq T^{-2}\mathrm{var}(X_{it-1}^2) \leq M(T^{-1}\sigma_X^2)^2 \to 0.$$

Next

$$EY_{bi}^2 = \frac{1}{T^2}\sum_{t=1}^{T}\mathrm{var}(X_{it-1}X^*_{it-1}) + \frac{2}{T^2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathrm{cov}(X_{it-1}X^*_{it-1}, X_{is-1}X^*_{is-1}).$$

But $\mathrm{var}(X_{it}X^*_{it}) = \mathrm{var}(X_{it}^2)/\sigma_X^4 = O(\sigma_X^{-2}) = O(1)$, so the first term is $O(T^{-1})$, and the second term is also $O(T^{-1})$ by (iii). Next, $Y_{cit} = T^{-1}[X_{it-1}(X_{iT} - X_{i0}) - EX_{it-1}(X_{iT} - X_{i0})]$, and the proof is similar to that for Y_{ai}. Finally, Y_{dit} is handled similar to Y_{bi}. Note that the convergences are uniform in i and t. ∎

Proof of Theorem 3. The first differenced process ΔX_{it} clearly gives a consistent estimate. For $X_{it} - \hat{\phi}_{lsdv}X_{it-1}$, we note that the assumptions that $T^{-1}\sigma_{e,T}^2 = o(1)$ and $T^{-1}\Sigma_{F,T} = o(1)$ imply that $T^{-1}\sigma_{X,T}^2 = o(1)$. Then it is straightforward to see that conditions for Lemma A.2 and A.3 are satisfied under the regularity Assumptions 1–3. The result follows from Lemmas A.2 and A.3. ∎

References

Amengual, D, Watson, M. 2007. Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business and Economic Statistics* **25**: 91–96.

Andrews D. W. K. and J. C. Monahan, 1992, An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, Econometrica 60, 953–966.

Bai, J., 2004, Estimating cross section common stochastic trends in non-stationary panel data, Journal of Econometrics 122, 137–183.

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

Bai, J. and S. Ng, 2004, A PANIC attack on unit roots and cointegration, Econometrica 72, 1127–1177.

Bai, J. and S. Ng, 2006, Determining the number of factors in approximate factor models, Errata, available at http://www.columbia.edu/˜sn2294/papers/correctionEcta2.pdf. Accessed April 2011.

Bai, J. and S. Ng, 2007, Determining the number of primitive shocks in factor models, Journal of Business and Economic Statistics 26, 52–60.

Bai, J. and S. Ng, 2008, Large Dimensional Factor Analysis, Foundations and Trends in Econometrics 3, 89–163.

Breitung, J. and S. Eickmeier, 2009, Testing for structural breaks in dynamic factor models, Deutsche Bundesbank Economic Studies Discussion Paper 05/2009.

Cavanagh, C. L., G. Elliot and J. H. Stock, 1995, Inference in models with nearly integrated regressors, Econometric Theory 11, 1131–1147.

Cecchetti, S.G., Mark, N. C. and R. Sonora, 2002, Price index convergence in United States cities, International Economic Review 43, 1081–1099.

Chamberlain, G. and M. Rothschild, 1983, Arbitrage, factor structure and mean-variance analysis in large asset markets, Econometrica 51, 1305–1324.

Connor, G. and R. A. Korajczyk, 1993, A test for the number of factors in approximate factor models, Journal of Finance 48, 1263–1291.

Elliot, G., T. J. Rothenberg and J. H. Stock, 1996, Efficient tests for an autoregressive unit root, Econometrica 64, 813–36.

Forni, M., M. Hallin, M. Lippi and L. Reichlin, 2000, The generalized dynamic-factor model: Identification and estimation, Review of Economics and Statistics 82, 540–554.

Giraitis, L. and P. C. B. Phillips, 2006, Uniform limit theory for stationary autoregression, Journal of Time Series Analysis 26, 51–60.

Giraitis, L. and P. C. B. Phillips, 2009, Mean and autocovariance function estimation near the boundary of stationarity, Cowles Foundation Discussion Paper 1690.

Hallin, M. and R. Liska, 2007, Determining the number of factors in the generalized dynamic factor model, Journal of the American Statistical Association 102, 603–617.

Han, C., Phillips, P. C. B. and D. Sul, 2007, 2011, X-Differencing and Dynamic Panel Model Estimation, mimeo, University of Texas at Dallas.

McConnell, M., and G. Perez-Quiros, 2000, Output Fluctuations in the United States: What Has Changed since the Early 1980s?, American Economic Review, 90, 1464–76.

Moon, R. H. and P.C.B. Phillips, 2000, Estimation of autoregressive roots near unity using panel data, Econometric Theory 16, 927–88.

Moon, R. H. and P.C.B. Phillips, 2004, GMM estimation of autoregressive roots near unity with panel data, Econometrica 72, 467–522.

Onatski, A., 2007, Determining the number of factors from empirical distribution of eigenvalues, Working paper, Columbia University.

Park, J. Y., 2003, Weak unit roots, mimeo, Rice University.

Phillips, P. C. B., 1987, Towards a unified asymptotic theory for autoregression, Biometrika 74, 535–47

Phillips, P. C. B. and T. Magdalinos, 2007, Limit theory for moderate deviations from a unit root, Journal of Econometrics 136, 115–130.

Phillips, P. C. B., H. R. Moon and Z. Xiao, 2001, How to estimate autoregressive roots near unity, Econometric Theory 17, 29–69.

Phillips, P. C. B. and D. Sul, 2007, Transition modelling and econometric convergence tests, Econometrica 75. 1771-1855.

Stock, J. H. and M. W. Watson, 2002, Has the Business Cycle Changed and Why?, NBER Macroeconomics Annual 2002, Mark Gertler and Ken Rogoff (eds), MIT Press.

Stock, J. H. and M. W. Watson, 2005, Implications of dynamic factor models for VAR analysis, NBER Working Paper No. W11467.

Yin, Y. Q., Z. D. Bai and P. R. Krishnaiah, 1988, On the limit of the largest eigenvalue of large dimensional sample covariance matrix, Probability Theory and Related Fields 78, 509–521.