Econometrica, Vol. 70, No. 1 (January, 2002), 191–221

DETERMINING THE NUMBER OF FACTORS IN APPROXIMATE FACTOR MODELS

BY JUSHAN BAI AND SERENA NG[1]

In this paper we develop some econometric theory for factor models of large dimensions. The focus is the determination of the number of factors (r), which is an unresolved issue in the rapidly growing literature on multifactor models. We first establish the convergence rate for the factor estimates that will allow for consistent estimation of r. We then propose some panel criteria and show that the number of factors can be consistently estimated using the criteria. The theory is developed under the framework of large cross-sections (N) and large time dimensions (T). No restriction is imposed on the relation between N and T. Simulations show that the proposed criteria have good finite sample properties in many configurations of the panel data encountered in practice.

KEYWORDS: Factor analysis, asset pricing, principal components, model selection.

1. INTRODUCTION

THE IDEA THAT VARIATIONS in a large number of economic variables can be modeled by a small number of reference variables is appealing and is used in many economic analyses. For example, asset returns are often modeled as a function of a small number of factors. Stock and Watson (1989) used one reference variable to model the comovements of four main macroeconomic aggregates. Cross-country variations are also found to have common components; see Gregory and Head (1999) and Forni, Hallin, Lippi, and Reichlin (2000b). More recently, Stock and Watson (1999) showed that the forecast error of a large number of macroeconomic variables can be reduced by including diffusion indexes, or factors, in structural as well as nonstructural forecasting models. In demand analysis, Engel curves can be expressed in terms of a finite number of factors. Lewbel (1991) showed that if a demand system has one common factor, budget shares should be independent of the level of income. In such a case, the number of factors is an object of economic interest since if more than one factor is found, homothetic preferences can be rejected. Factor analysis also provides a convenient way to study the aggregate implications of microeconomic behavior, as shown in Forni and Lippi (1997).

Central to both the theoretical and the empirical validity of factor models is the correct specification of the number of factors. To date, this crucial parameter

[1] We thank three anonymous referees for their very constructive comments, which led to a much improved presentation. The first author acknowledges financial support from the National Science Foundation under Grant SBR-9709508. We would like to thank participants in the econometrics seminars at Harvard-MIT, Cornell University, the University of Rochester, and the University of Pennsylvania for help suggestions and comments. Remaining errors are our own.

is often assumed rather than determined by the data.[2] This paper develops a formal statistical procedure that can consistently estimate the number of factors from observed data. We demonstrate that the penalty for overfitting must be a function of both N and T (the cross-section dimension and the time dimension, respectively) in order to consistently estimate the number of factors. Consequently the usual AIC and BIC, which are functions of N or T alone, do not work when both dimensions of the panel are large. Our theory is developed under the assumption that both N and T converge to infinity. This flexibility is of empirical relevance because the time dimension of datasets relevant to factor analysis, although small relative to the cross-section dimension, is too large to justify the assumption of a fixed T.

A small number of papers in the literature have also considered the problem of determining the number of factors, but the present analysis differs from these works in important ways. Lewbel (1991) and Donald (1997) used the rank of a matrix to test for the number of factors, but these theories assume either N or T is fixed. Cragg and Donald (1997) considered the use of information criteria when the factors are functions of a set of observable explanatory variables, but the data still have a fixed dimension. For large dimensional panels, Connor and Korajczyk (1993) developed a test for the number of factors in asset returns, but their test is derived under sequential limit asymptotics, i.e., N converges to infinity with a fixed T and then T converges to infinity. Furthermore, because their test is based on a comparison of variances over different time periods, covariance stationarity and homoskedasticity are not only technical assumptions, but are crucial for the validity of their test. Under the assumption that $N \to \infty$ for fixed T, Forni and Reichlin (1998) suggested a graphical approach to identify the number of factors, but no theory is available. Assuming $N, T \to \infty$ with $\sqrt{N}/T \to \infty$, Stock and Watson (1998) showed that a modification to the BIC can be used to select the number of factors optimal for forecasting a single series. Their criterion is restrictive not only because it requires $N \gg T$, but also because there can be factors that are pervasive for a set of data and yet have no predictive ability for an individual data series. Thus, their rule may not be appropriate outside of the forecasting framework. Forni, Hallin, Lippi, and Reichlin (2000a) suggested a multivariate variant of the AIC but neither the theoretical nor the empirical properties of the criterion are known.

We set up the determination of factors as a model selection problem. In consequence, the proposed criteria depend on the usual trade-off between good fit and parsimony. However, the problem is nonstandard not only because account needs to be taken of the sample size in both the cross-section and the time-series dimensions, but also because the factors are not observed. The theory we developed does not rely on sequential limits, nor does it impose any restrictions between N and T. The results hold under heteroskedasticity in both the time and

[2] Lehmann and Modest (1988), for example, tested the APT for 5, 10, and 15 factors. Stock and Watson (1989) assumed there is one factor underlying the coincident index. Ghysels and Ng (1998) tested the affine term structure model assuming two factors.

the cross-section dimensions. The results also hold under weak serial and cross-section dependence. Simulations show that the criteria have good finite sample properties.

The rest of the paper is organized as follows. Section 2 sets up the preliminaries and introduces notation and assumptions. Estimation of the factors is considered in Section 3 and the estimation of the number of factors is studied in Section 4. Specific criteria are considered in Section 5 and their finite sample properties are considered in Section 6, along with an empirical application to asset returns. Concluding remarks are provided in Section 7. All the proofs are given in the Appendix.

2. FACTOR MODELS

Let X_{it} be the observed data for the ith cross-section unit at time t, for $i = 1, \ldots, N$, and $t = 1, \ldots, T$. Consider the following model:

$$X_{it} = \lambda_i' F_t + e_{it}, \tag{1}$$

where F_t is a vector of common factors, λ_i is a vector of factor loadings associated with F_t, and e_{it} is the idiosyncratic component of X_{it}. The product $\lambda_i' F_t$ is called the common component of X_{it}. Equation (1) is then the factor representation of the data. Note that the factors, their loadings, as well as the idiosyncratic errors are not observable.

Factor analysis allows for dimension reduction and is a useful statistical tool. Many economic analyses fit naturally into the framework given by (1).

1. *Arbitrage pricing theory.* In the finance literature, the arbitrage pricing theory (APT) of Ross (1976) assumes that a small number of factors can be used to explain a large number of asset returns. In this case, X_{it} represents the return of asset i at time t, F_t represents the vector of factor returns, and e_{it} is the idiosyncratic component of returns. Although analytical convenience makes it appealing to assume one factor, there is growing evidence against the adequacy of a single factor in explaining asset returns.[3] The shifting interest towards use of multifactor models inevitably calls for a formal procedure to determine the number of factors. The analysis to follow allows the number of factors to be determined even when N and T are both large. This is especially suited for financial applications when data are widely available for a large number of assets over an increasingly long span. Once the number of factors is determined, the factor returns F_t can also be consistently estimated (up to an invertible transformation).

2. *The rank of a demand system.* Let p be a price vector for J goods and services, e_h be total spending on the J goods by household h. Consumer theory postulates that Marshallian demand for good j by consumer h is $X_{jh} = g_j(p, e_h)$. Let $w_{jh} = X_{jh}/e_h$ be the budget share for household h on the jth good. The

[3] Cochrane (1999) stressed that financial economists now recognize that there are multiple sources of risk, or factors, that give rise to high returns. Backus, Forsei, Mozumdar, and Wu (1997) made similar conclusions in the context of the market for foreign assets.

rank of a demand system holding prices fixed is the smallest integer r such that $w_j(e) = \lambda_{j1}G_1(e) + \cdots + \lambda_{jr}G_r(e)$. Demand systems are of the form (1) where the r factors, common across goods, are $F_h = [G_1(e_h) \cdots G_r(e_h)]'$. When the number of households, H, converges to infinity with a fixed J, $G_1(e) \cdots G_r(e)$ can be estimated simultaneously, such as by nonparametric methods developed in Donald (1997). This approach will not work when the number of goods, J, also converges to infinity. However, the theory to be developed in this paper will still provide a consistent estimation of r and without the need for nonparametric estimation of the $G(\cdot)$ functions. Once the rank of the demand system is determined, the nonparametric functions evaluated at e_h allow F_h to be consistently estimable (up to a transformation). Then functions $G_1(e) \cdots G_r(e)$ may be recovered (also up to a matrix transformation) from $\widehat{F}_h$ $(h = 1, \ldots, H)$ via nonparametric estimation.

3. *Forecasting with diffusion indices.* Stock and Watson (1998, 1999) considered forecasting inflation with diffusion indices ("factors") constructed from a large number of macroeconomic series. The underlying premise is that these series may be driven by a small number of unobservable factors. Consider the forecasting equation for a scalar series

$$y_{t+1} = \alpha' F_t + \beta' W_t + \epsilon_t.$$

The variables W_t are observable. Although we do not observe F_t, we observe $X_{it}, i = 1, \ldots, N$. Suppose X_{it} bears relation with F_t as in (1). In the present context, we interpret (1) as the reduced-form representation of X_{it} in terms of the unobservable factors. We can first estimate F_t from (1). Denote it by $\widehat{F}_t$. We can then regress y_t on $\widehat{F}_{t-1}$ and W_{t-1} to obtain the coefficients $\hat{\alpha}$ and $\hat{\beta}$, from which a forecast

$$\hat{y}_{T+1|T} = \hat{\alpha}' \widehat{F}_T + \hat{\beta} W_T$$

can be formed. Stock and Watson (1998, 1999) showed that this approach of forecasting outperforms many competing forecasting methods. But as pointed out earlier, the dimension of F in Stock and Watson (1998, 1999) was determined using a criterion that minimizes the mean squared forecast errors of y. This may not be the same as the number of factors underlying X_{it}, which is the focus of this paper.

2.1. *Notation and Preliminaries*

Let F_t^0, λ_i^0, and r denote the true common factors, the factor loadings, and the true number of factors, respectively. Note that F_t^0 is r dimensional. We assume that r does not depend on N or T. At a given t, we have

$$\underset{(N\times 1)}{X_t} = \underset{(N\times r)}{\Lambda^0}\ \underset{(r\times 1)}{F_t^0} + \underset{(N\times 1)}{e_t} \tag{2}$$

where $X_t = (X_{1t}, X_{2t}, \ldots, X_{Nt})'$, $\Lambda^0 = (\lambda_1^0, \lambda_2^0, \ldots, \lambda_N^0)'$, and $e_t = (e_{1t}, e_{2t}, \ldots, e_{Nt})'$. Our objective is to determine the true number of factors, r. In classical

factor analysis (e.g., Anderson (1984)), N is assumed fixed, the factors are independent of the errors e_t, and the covariance of e_t is diagonal. Normalizing the covariance matrix of F_t to be an identity matrix, we have $\Sigma = \Lambda^0 \Lambda^{0'} + \Omega$, where Σ and Ω are the covariance matrices of X_t and e_t, respectively. Under these assumptions, a root-T consistent and asymptotically normal estimator of Σ, say, the sample covariance matrix $\widehat{\Sigma} = (1/T)\sum_{t=1}^{T}(X_t - \overline{X})(X_t - \overline{X})'$ can be obtained. The essentials of classical factor analysis carry over to the case of large N but fixed T since the $N \times N$ problem can be turned into a $T \times T$ problem, as noted by Connor and Korajczyk (1993) and others.

Inference on r under classical assumptions can, in theory, be based on the eigenvalues of $\widehat{\Sigma}$ since a characteristic of a panel of data that has an r factor representation is that the first r largest population eigenvalues of the $N \times N$ covariance of X_t diverge as N increases to infinity, but the $(r+1)$th eigenvalue is bounded; see Chamberlain and Rothschild (1983). But it can be shown that all nonzero sample eigenvalues (not just the first r) of the matrix $\widehat{\Sigma}$ increase with N, and a test based on the sample eigenvalues is thus not feasible. A likelihood ratio test can also, in theory, be used to select the number of factors if, in addition, normality of e_t is assumed. But as found by Dhrymes, Friend, and Glutekin (1984), the number of statistically significant factors determined by the likelihood ratio test increases with N even if the true number of factors is fixed. Other methods have also been developed to estimate the number of factors assuming the size of one dimension is fixed. But Monte Carlo simulations in Cragg and Donald (1997) show that these methods tend to perform poorly for moderately large N and T. The fundamental problem is that the theory developed for classical factor models does not apply when both N and $T \to \infty$. This is because consistent estimation of Σ (whether it is an $N \times N$ or a $T \times T$ matrix) is not a well defined problem. For example, when $N > T$, the rank of $\widehat{\Sigma}$ is no more than T, whereas the rank of Σ can always be N. New theories are thus required to analyze large dimensional factor models.

In this paper, we develop asymptotic results for consistent estimation of the number of factors when N and $T \to \infty$. Our results complement the sparse but growing literature on large dimensional factor analysis. Forni and Lippi (2000) and Forni et al. (2000a) obtained general results for dynamic factor models, while Stock and Watson (1998) provided some asymptotic results in the context of forecasting. As in these papers, we allow for cross-section and serial dependence. In addition, we also allow for heteroskedasticity in e_t and some weak dependence between the factors and the errors. These latter generalizations are new in our analysis. Evidently, our assumptions are more general than those used when the sample size is fixed in one dimension.

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

$$\underset{(T\times 1)}{\underline{X}_i} = \underset{(T\times r)}{F^0}\ \underset{(r\times 1)}{\lambda_i^0} + \underset{(T\times 1)}{\underline{e}_i}, \tag{3}$$

where $\underline{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iT})'$, $F^0 = (F_1^0, F_2^0, \ldots, F_T^0)'$, and $\underline{e}_i = (e_{i1}, e_{i2}, \ldots, e_{iT})'$. For the panel of data $X = (\underline{X}_1, \ldots, \underline{X}_N)$, we have

$$\underset{(T\times N)}{X} = \underset{(T\times r)}{F^0}\,\underset{(r\times N)}{\Lambda^{0'}} + \underset{(T\times N)}{e}, \tag{4}$$

with $e = (\underline{e}_1, \ldots, \underline{e}_N)$.

Let $\mathrm{tr}(A)$ denote the trace of A. The norm of the matrix A is then $\|A\| = [\mathrm{tr}(A'A)]^{1/2}$. The following assumptions are made:

ASSUMPTION A—*Factors*: $E\|F_t^0\|^4 < \infty$ *and* $T^{-1}\sum_{t=1}^T F_t^0 F_t^{0'} \to \Sigma_F$ *as* $T \to \infty$ *for some positive definite matrix* Σ_F.

ASSUMPTION B—*Factor Loadings*: $\|\lambda_i\| \le \bar{\lambda} < \infty$, *and* $\|\Lambda^{0'}\Lambda^0/N - D\| \to 0$ *as* $N \to \infty$ *for some* $r \times r$ *positive definite matrix* D.

ASSUMPTION C—*Time and Cross-Section Dependence and Heteroskedasticity*: *There exists a positive constant* $M < \infty$, *such that for all* N *and* T,

1. $E(e_{it}) = 0, E|e_{it}|^8 \le M$;
2. $E(e_s'e_t/N) = E(N^{-1}\sum_{i=1}^N e_{is}e_{it}) = \gamma_N(s,t), |\gamma_N(s,s)| \le M$ *for all* s, *and* $T^{-1}\sum_{s=1}^T\sum_{t=1}^T |\gamma_N(s,t)| \le M$;
3. $E(e_{it}e_{jt}) = \tau_{ij,t}$ *with* $|\tau_{ij,t}| \le |\tau_{ij}|$ *for some* τ_{ij} *and for all* t; *in addition*, $N^{-1}\sum_{i=1}^N\sum_{j=1}^N |\tau_{ij}| \le M$;
4. $E(e_{it}e_{js}) = \tau_{ij,ts}$ *and* $(NT)^{-1}\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^T\sum_{s=1}^T |\tau_{ij,ts}| \le M$;
5. *for every* (t,s), $E|N^{-1/2}\sum_{i=1}^N [e_{is}e_{it} - E(e_{is}e_{it})]|^4 \le M$.

ASSUMPTION D—*Weak Dependence between Factors and Idiosyncratic Errors*:

$$E\left(\frac{1}{N}\sum_{i=1}^N \left\| \frac{1}{\sqrt{T}}\sum_{t=1}^T F_t^0 e_{it}\right\|^2\right) \le M.$$

Assumption A is standard for factor models. Assumption B ensures that each factor has a nontrivial contribution to the variance of X_t. We only consider nonrandom factor loadings for simplicity. Our results still hold when the λ_i are random, provided they are independent of the factors and idiosyncratic errors, and $E\|\lambda_i\|^4 \le M$. Assumption C allows for limited time-series and cross-section dependence in the idiosyncratic component. Heteroskedasticity in both the time and cross-section dimensions is also allowed. Under stationarity in the time dimension, $\gamma_N(s,t) = \gamma_N(s-t)$, though the condition is not necessary. Given Assumption C1, the remaining assumptions in C are easily satisfied if the e_{it} are independent for all i and t. The allowance for some correlation in the idiosyncratic components sets up the model to have an *approximate factor structure.* It is more general than a *strict factor model,* which assumes e_{it} is uncorrelated across i, the framework in which the APT theory of Ross (1976) is based. Thus, the results to be developed will also apply to strict factor models. When the factors

and idiosyncratic errors are independent (a standard assumption for conventional factor models), Assumption D is implied by Assumptions A and C. Independence is not required for D to be true. For example, suppose that $e_{it} = \epsilon_{it}\|F_t\|$ with ϵ_{it} being independent of F_t and ϵ_{it} satisfies Assumption C; then Assumption D holds. Finally, the developments proceed assuming that the panel is balanced. We also note that the model being analyzed is static, in the sense that X_{it} has a contemporaneous relationship with the factors. The analysis of dynamic models is beyond the scope of this paper.

For a factor model to be an approximate factor model in the sense of Chamberlain and Rothschild (1983), the largest eigenvalue (and hence all of the eigenvalues) of the $N \times N$ covariance matrix $\Omega = E(e_t e_t')$ must be bounded. Note that Chamberlain and Rothschild focused on the cross-section behavior of the model and did not make explicit assumptions about the time-series behavior of the model. Our framework allows for serial correlation and heteroskedasticity and is more general than their setup. But if we assume e_t is stationary with $E(e_{it}e_{jt}) = \tau_{ij}$, then from matrix theory, the largest eigenvalue of Ω is bounded by $\max_i \sum_{j=1}^{N} |\tau_{ij}|$. Thus if we assume $\sum_{j=1}^{N} |\tau_{ij}| \leq M$ for all i and all N, which implies Assumption C3, then (2) will be an approximate factor model in the sense of Chamberlain and Rothschild.

3. ESTIMATION OF THE COMMON FACTORS

When N is small, factor models are often expressed in state space form, normality is assumed, and the parameters are estimated by maximum likelihood. For example, Stock and Watson (1989) used $N = 4$ variables to estimate one factor, the coincident index. The drawback is that because the number of parameters increases with N,[4] computational difficulties make it necessary to abandon information on many series even though they are available.

We estimate common factors in large panels by the method of asymptotic principal components.[5] The number of factors that can be estimated by this (nonparametric) method is $\min\{N, T\}$, much larger than permitted by estimation of state space models. But to determine which of these factors are statistically important, it is necessary to first establish consistency of all the estimated common factors when both N and T are large. We start with an arbitrary number k ($k < \min\{N, T\}$). The superscript in λ_i^k and F_t^k signifies the allowance of k factors in the estimation. Estimates of λ^k and F^k are obtained by solving the optimization problem

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

[4] Gregory, Head, and Raynauld (1997) estimated a world factor and seven country specific factors from output, consumption, and investment for each of the G7 countries. The exercise involved estimation of 92 parameters and perhaps stretched the state-space model to its limit.

[5] The method of asymptotic principal components was studied by Connor and Korajzcyk (1986) and Connor and Korajzcyk (1988) for fixed T. Forni et al. (2000a) and Stock and Watson (1998) considered the method for large T.

subject to the normalization of either $\Lambda^{k'}\Lambda^k/N = I_k$ or $F^{k'}F^k/T = I_k$. If we concentrate out Λ^k and use the normalization that $F^{k'}F^k/T = I_k$, the optimization problem is identical to maximizing $\mathrm{tr}(F^{k'}(XX')F^k)$. The estimated factor matrix, denoted by $\widetilde{F}^k$, is $\sqrt{T}$ times the eigenvectors corresponding to the k largest eigenvalues of the $T \times T$ matrix XX'. Given $\widetilde{F}^k$, $\widetilde{\Lambda}^{k'} = (\widetilde{F}^{k'}\widetilde{F}^k)^{-1}\widetilde{F}^{k'}X = \widetilde{F}^{k'}X/T$ is the corresponding matrix of factor loadings.

The solution to the above minimization problem is not unique, even though the sum of squared residuals $V(k)$ is unique. Another solution is given by $(\overline{F}^k, \overline{\Lambda}^k)$, where $\overline{\Lambda}^k$ is constructed as $\sqrt{N}$ times the eigenvectors corresponding to the k largest eigenvalues of the $N \times N$ matrix $X'X$. The normalization that $\overline{\Lambda}^{k'}\overline{\Lambda}^k/N = I_k$ implies $\overline{F}^k = X\overline{\Lambda}^k/N$. The second set of calculations is computationally less costly when $T > N$, while the first is less intensive when $T < N$.[6]

Define

$$\widehat{F}^k = \overline{F}^k(\overline{F}^{k'}\overline{F}^k/T)^{1/2},$$

a rescaled estimator of the factors. The following theorem summarizes the asymptotic properties of the estimated factors.

THEOREM 1: *For any fixed $k \geq 1$, there exists a $(r \times k)$ matrix H^k with* $\mathrm{rank}(H^k) = \min\{k, r\}$, *and* $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$, *such that*

$$C_{NT}^2\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{F}_t^k - H^{k'}F_t^0\right\|^2\right) = O_p(1). \tag{5}$$

Because the true factors (F^0) can only be identified up to scale, what is being considered is a rotation of F^0. The theorem establishes that the time average of the squared deviations between the estimated factors and those that lie in the true factor space vanish as $N, T \to \infty$. The rate of convergence is determined by the smaller of N or T, and thus depends on the panel structure.

Under the additional assumption that $\sum_{s=1}^{T}\gamma_N(s,t)^2 \leq M$ for all t and T, the result[7]

$$C_{NT}^2\left\|\widehat{F}_t - H^{k'}F_t^0\right\|^2 = O_p(1), \quad \text{for each } t, \tag{6}$$

can be obtained. Neither Theorem 1 nor (6) implies uniform convergence in t. Uniform convergence is considered by Stock and Watson (1998). These authors obtained a much slower convergence rate than C_{NT}^2, and their result requires $\sqrt{N} \gg T$. An important insight of this paper is that, to consistently estimate the number of factors, neither (6) nor uniform convergence is required. It is the average convergence rate of Theorem 1 that is essential. However, (6) could be useful for statistical analysis on the estimated factors and is thus a result of independent interest.

[6] A more detailed account of computation issues, including how to deal with unbalanced panels, is given in Stock and Watson (1998).

[7] The proof is actually simpler than that of Theorem 1 and is thus omitted to avoid repetition.

4. ESTIMATING THE NUMBER OF FACTORS

Suppose for the moment that we observe all potentially informative factors but not the factor loadings. Then the problem is simply to choose k factors that best capture the variations in X and estimate the corresponding factor loadings. Since the model is linear and the factors are observed, λ_i can be estimated by applying ordinary least squares to each equation. This is then a classical model selection problem. A model with $k+1$ factors can fit no worse than a model with k factors, but efficiency is lost as more factor loadings are being estimated. Let F^k be a matrix of k factors, and

$$V(k, F^k) = \min_{\Lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \lambda_i^{k'} F_t^k\right)^2$$

be the sum of squared residuals (divided by NT) from time-series regressions of $\underline{X}_i$ on the k factors for all i. Then a loss function $V(k, F^k) + kg(N, T)$, where $g(N, T)$ is the penalty for overfitting, can be used to determine k. Because the estimation of λ_i is classical, it can be shown that the BIC with $g(N, T) = \ln(T)/T$ can consistently estimate r. On the other hand, the AIC with $g(N, T) = 2/T$ may choose $k > r$ even in large samples. The result is the same as in Geweke and Meese (1981) derived for $N = 1$ because when the factors are observed, the penalty factor does not need to take into account the sample size in the cross-section dimension. Our main result is to show that this will no longer be true when the factors have to be estimated, and even the BIC will not always consistently estimate r.

Without loss of generality, we let

$$V(k, \widehat{F}^k) = \min_{\Lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \lambda_i^{k'} \widehat{F}_t^k\right)^2 \tag{7}$$

denote the sum of squared residuals (divided by NT) when k factors are estimated. This sum of squared residuals does not depend on which estimate of F is used because they span the same vector space. That is, $V(k, \widetilde{F}^k) = V(k, \overline{F}^k) = V(k, \widehat{F}^k)$. We want to find penalty functions, $g(N, T)$, such that criteria of the form

$$PC(k) = V(k, \widehat{F}^k) + kg(N, T)$$

can consistently estimate r. Let *kmax* be a bounded integer such that $r \leq kmax$.

THEOREM 2: *Suppose that Assumptions A–D hold and that the k factors are estimated by principal components. Let* $\hat{k} = \arg\min_{0 \leq k \leq kmax} PC(k)$. *Then* $\lim_{N,T\to\infty} \text{Prob}[\hat{k} = r] = 1$ *if* (i) $g(N, T) \to 0$ *and* (ii) $C_{NT}^2 \cdot g(N, T) \to \infty$ *as* N, $T \to \infty$, *where* $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$.

Conditions (i) and (ii) are necessary in the sense that if one of the conditions is violated, then there will exist a factor model satisfying Assumptions A–D, and

yet the number of factors cannot be consistently estimated. However, conditions (i) and (ii) are not always required to obtain a consistent estimate of r.

A formal proof of Theorem 2 is provided in the Appendix. The crucial element in consistent estimation of r is a penalty factor that vanishes at an appropriate rate such that under and overparameterized models will not be chosen. An implication of Theorem 2 is the following:

COROLLARY 1: *Under the Assumptions of Theorem* 2, *the class of criteria defined by*

$$IC(k) = \ln(V(k, \widehat{F}^k)) + kg(N, T)$$

will also consistently estimate r.

Note that $V(k, \widehat{F}^k)$ is simply the average residual variance when k factors are assumed for each cross-section unit. The IC criteria thus resemble information criteria frequently used in time-series analysis, with the important difference that the penalty here depends on both N and T.

Thus far, it has been assumed that the common factors are estimated by the method of principle components. Forni and Reichlin (1998) and Forni et al. (2000a) studied alternative estimation methods. However the proof of Theorem 2 mainly uses the fact that $\widehat{F}_t$ satisfies Theorem 1, and does not rely on principal components per se. We have the following corollary:

COROLLARY 2: *Let* $\widehat{G}^k$ *be an arbitrary estimator of* F^0. *Suppose there exists a matrix* $\widetilde{H}^k$ *such that* $\operatorname{rank}(\widetilde{H}^k) = \min\{k, r\}$, *and for some* $\widetilde{C}_{NT}^2 \leq C_{NT}^2$,

$$\widetilde{C}_{NT}^2 \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \widehat{G}_t^k - \widetilde{H}^{k\prime} F_t^0 \right\|^2 \right) = O_p(1). \tag{8}$$

Then Theorem 2 *still holds with* $\widehat{F}^k$ *replaced by* $\widehat{G}^k$ *and* C_{NT} *replaced by* $\widetilde{C}_{NT}$.

The sequence of constants $\widetilde{C}_{NT}^2$ does not need to equal $C_{NT}^2 = \min\{N, T\}$. Theorem 2 holds for any estimation method that yields estimators $\widehat{G}_t$ satisfying (8).[8] Naturally, the penalty would then depend on $\widetilde{C}_{NT}^2$, the convergence rate for $\widehat{G}_t$.

5. THE PC_p AND THE IC_p

In this section, we assume that the method of principal components is used to estimate the factors and propose specific formulations of $g(N, T)$ to be used in

[8] We are grateful for a referee whose question led to the results reported here.

practice. Let $\hat{\sigma}^2$ be a consistent estimate of $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}E(e_{it})^2$. Consider the following criteria:

$$PC_{p1}(k) = V(k,\widehat{F}^k) + k\hat{\sigma}^2\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right);$$

$$PC_{p2}(k) = V(k,\widehat{F}^k) + k\hat{\sigma}^2\left(\frac{N+T}{NT}\right)\ln C_{NT}^2;$$

$$PC_{p3}(k) = V(k,\widehat{F}^k) + k\hat{\sigma}^2\left(\frac{\ln C_{NT}^2}{C_{NT}^2}\right).$$

Since $V(k,\widehat{F}^k) = N^{-1}\sum_{i=1}^{N}\hat{\sigma}_i^2$, where $\hat{\sigma}_i^2 = \hat{e}_i'\hat{e}_i/T$, the criteria generalize the C_p criterion of Mallows (1973) developed for selection of models in strict time-series or cross-section contexts to a panel data setting. For this reason, we refer to these statistics as Panel C_p (PC_p) criteria. Like the C_p criterion, $\hat{\sigma}^2$ provides the proper scaling to the penalty term. In applications, it can be replaced by $V(kmax,\widehat{F}^{kmax})$. The proposed penalty functions are based on the sample size in the smaller of the two dimensions. All three criteria satisfy conditions (i) and (ii) of Theorem 2 since $C_{NT}^{-2} \approx ((N+T)/NT) \to 0$ as $N,T \to \infty$. However, in finite samples, $C_{NT}^{-2} \leq (N+T)/NT$. Hence, the three criteria, although asymptotically equivalent, will have different properties in finite samples.[9]

Corollary 1 leads to consideration of the following three criteria:

$$IC_{p1}(k) = \ln(V(k,\widehat{F}^k)) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right);$$

$$IC_{p2}(k) = \ln(V(k,\widehat{F}^k)) + k\left(\frac{N+T}{NT}\right)\ln C_{NT}^2; \tag{9}$$

$$IC_{p3}(k) = \ln(V(k,\widehat{F}^k)) + k\left(\frac{\ln C_{NT}^2}{C_{NT}^2}\right).$$

The main advantage of these three panel information criteria (IC_p) is that they do not depend on the choice of *kmax* through $\hat{\sigma}^2$, which could be desirable in practice. The scaling by $\hat{\sigma}^2$ is implicitly performed by the logarithmic transformation of $V(k,\widehat{F}^k)$ and thus not required in the penalty term.

The proposed criteria differ from the conventional C_p and information criteria used in time-series analysis in that $g(N,T)$ is a function of both N and T. To understand why the penalty must be specified as a function of the sample size in

[9] Note that PC_{p1} and PC_{p2}, and likewise, IC_{p1} and IC_{p2}, apply specifically to the principal components estimator because $C_{NT}^2 = \min\{N,T\}$ is used in deriving them. For alternative estimators satisfying Corollary 2, criteria PC_{p3} and IC_{p3} are still applicable with C_{NT} replaced by $\widetilde{C}_{NT}$.

both dimensions, consider the following:

$$AIC_1(k) = V(k, \widehat{F}^k) + k\hat{\sigma}^2\left(\frac{2}{T}\right);$$

$$BIC_1(k) = V(k, \widehat{F}^k) + k\hat{\sigma}^2\left(\frac{\ln T}{T}\right);$$

$$AIC_2(k) = V(k, \widehat{F}^k) + k\hat{\sigma}^2\left(\frac{2}{N}\right);$$

$$BIC_2(k) = V(k, \widehat{F}^k) + k\hat{\sigma}^2\left(\frac{\ln N}{N}\right);$$

$$AIC_3(k) = V(k, \widehat{F}^k) + k\hat{\sigma}^2\left(2\frac{(N+T-k)}{NT}\right);$$

$$BIC_3(k) = V(k, \widehat{F}^k) + k\hat{\sigma}^2\left(\frac{(N+T-k)\ln(NT)}{NT}\right).$$

The penalty factors in AIC_1 and BIC_1 are standard in time-series applications. Although $g(N,T) \to 0$ as $T \to \infty$, AIC_1 fails the second condition of Theorem 2 for all N and T. When $N \ll T$ and $N\log(T)/T \not\longrightarrow \infty$, the BIC_1 also fails condition (ii) of Theorem 2. Thus we expect the AIC_1 will not work for all N and T, while the BIC_1 will not work for small N relative to T. By analogy, AIC_2 also fails the conditions of Theorem 2, while BIC_2 will work only if $N \ll T$. The next two criteria, AIC_3 and BIC_3, take into account the panel nature of the problem. The two specifications of $g(N,T)$ reflect first, that the effective number of observations is $N \cdot T$, and second, that the total number of parameters being estimated is $k(N+T-k)$. It is easy to see that AIC_3 fails the second condition of Theorem 2. While the BIC_3 satisfies this condition, $g(N,T)$ does not always vanish. For example, if $N = \exp(T)$, then $g(N,T) \to 1$ and the first condition of Theorem 2 will not be satisfied. Similarly, $g(N,T)$ does not vanish when $T = \exp(N)$. Therefore BIC_3 may perform well for some but not all configurations of the data. In contrast, the proposed criteria satisfy both conditions stated in Theorem 2.

6. SIMULATIONS AND AN EMPIRICAL APPLICATION

We first simulate data from the following model:

$$X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$$

$$= c_{it} + \sqrt{\theta} e_{it},$$

where the factors are $T \times r$ matrices of $N(0,1)$ variables, and the factor loadings are $N(0,1)$ variates. Hence, the common component of X_{it}, denoted by c_{it}, has variance r. Results with λ_{ij} uniformly distributed are similar and will not

be reported. Our base case assumes that the idiosyncratic component has the same variance as the common component (i.e. $\theta = r$). We consider thirty configurations of the data. The first five simulate plausible asset pricing applications with five years of monthly data ($T = 60$) on 100 to 2000 asset returns. We then increase T to 100. Configurations with $N = 60, T = 100$ and 200 are plausible sizes of datasets for sectors, states, regions, and countries. Other configurations are considered to assess the general properties of the proposed criteria. All computations were performed using Matlab Version 5.3.

Reported in Tables I to III are the averages of $\hat{k}$ over 1000 replications, for $r = 1, 3$, and 5 respectively, assuming that e_{it} is homoskedastic $N(0, 1)$. For all cases, the maximum number of factors, *kmax*, is set to 8.[10] Prior to computation of the eigenvectors, each series is demeaned and standardized to have unit variance. Of the three PC_p criteria that satisfy Theorem 2, PC_{p3} is less robust than PC_{p1} and PC_{p2} when N or T is small. The IC_p criteria generally have properties very similar to the PC_p criteria. The term $NT/(N+T)$ provides a small sample correction to the asymptotic convergence rate of C_{NT}^2 and has the effect of adjusting the penalty upwards. The simulations show this adjustment to be desirable. When $\min\{N, T\}$ is 40 or larger, the proposed tests give precise estimates of the number of factors. Since our theory is based on large N and T, it is not surprising that for very small N or T, the proposed criteria are inadequate. Results reported in the last five rows of each table indicate that the IC_p criteria tend to underparameterize, while the PC_p tend to overparameterize, but the problem is still less severe than the AIC and the BIC, which we now consider.

The AIC and BIC's that are functions of only N or T have the tendency to choose too many factors. The AIC_3 performs somewhat better than AIC_1 and AIC_2, but still tends to overparameterize. At first glance, the BIC_3 appears to perform well. Although BIC_3 resembles PC_{p2}, the former penalizes an extra factor more heavily since $\ln(NT) > \ln C_{NT}^2$. As can be seen from Tables II and III, the BIC_3 tends to underestimate r, and the problem becomes more severe as r increases.

Table IV relaxes the assumption of homoskedasticity. Instead, we let $e_{it} = e_{it}^1$ for t odd, and $e_{it} = e_{it}^1 + e_{it}^2$ for t even, where e_{it}^1 and e_{it}^2 are independent $N(0, 1)$. Thus, the variance in the even periods is twice as large as the odd periods. Without loss of generality, we only report results for $r = 5$. PC_{p1}, PC_{p2}, IC_{p1}, and IC_{p2} continue to select the true number of factors very accurately and dominate the remaining criteria considered.

We then vary the variance of the idiosyncratic errors relative to the common component. When $\theta < r$, the variance of the common component is relatively large. Not surprisingly, the proposed criteria give precise estimates of r. The results will not be reported without loss of generality. Table V considers the case $\theta = 2r$. Since the variance of the idiosyncratic component is larger than the

[10] In time-series analysis, a rule such as $8\ \text{int}[(T/100)^{1/4}]$ considered in Schwert (1989) is sometimes used to set *kmax*, but no such guide is available for panel analysis. Until further results are available, a rule that replaces T in Schwert's rule by $\min\{N, T\}$ could be considered.

TABLE I

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $r = 1$; $\theta = 1$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	1.02	1.00	2.97	1.00	1.00	1.00	8.00	2.97	8.00	8.00	7.57	1.00
100	60	1.00	1.00	2.41	1.00	1.00	1.00	8.00	2.41	8.00	8.00	7.11	1.00
200	60	1.00	1.00	1.00	1.00	1.00	1.00	8.00	1.00	8.00	8.00	5.51	1.00
500	60	1.00	1.00	1.00	1.00	1.00	1.00	5.21	1.00	8.00	8.00	1.57	1.00
1000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
2000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
100	100	1.00	1.00	3.24	1.00	1.00	1.00	8.00	3.24	8.00	3.24	6.68	1.00
200	100	1.00	1.00	1.00	1.00	1.00	1.00	8.00	1.00	8.00	8.00	5.43	1.00
500	100	1.00	1.00	1.00	1.00	1.00	1.00	8.00	1.00	8.00	8.00	1.55	1.00
1000	100	1.00	1.00	1.00	1.00	1.00	1.00	1.08	1.00	8.00	8.00	1.00	1.00
2000	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
40	100	1.01	1.00	2.69	1.00	1.00	1.00	8.00	8.00	8.00	2.69	7.33	1.00
60	100	1.00	1.00	2.25	1.00	1.00	1.00	8.00	8.00	8.00	2.25	6.99	1.00
60	200	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	8.00	1.00	5.14	1.00
60	500	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	4.67	1.00	1.32	1.00
60	1000	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00	1.00	1.00
60	2000	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00	1.00	1.00
4000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
4000	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
8000	60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
8000	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00
60	4000	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00	1.00	1.00
100	4000	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00	1.00	1.00
60	8000	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00	1.00	1.00
100	8000	1.00	1.00	1.00	1.00	1.00	1.00	8.00	8.00	1.00	1.00	1.00	1.00
10	50	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.18
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	5.88
20	100	4.73	3.94	6.29	1.00	1.00	1.00	8.00	8.00	8.00	6.29	8.00	1.00
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
100	20	5.62	4.81	7.16	1.00	1.00	1.00	8.00	7.16	8.00	8.00	8.00	1.00

Notes: Table I–Table VIII report the estimated number of factors ($\hat{k}$) averaged over 1000 simulations. The true number of factors is r and $kmax = 8$. When the average of $\hat{k}$ is an integer, the corresponding standard error is zero. In the few cases when the averaged $\hat{k}$ over replications is not an integer, the standard errors are no larger than .6. In view of the precision of the estimates in the majority of cases, the standard errors in the simulations are not reported. The last five rows of each table are for models of small dimensions (either N or T is small).

common component, one might expect the common factors to be estimated with less precision. Indeed, IC_{p1} and IC_{p2} underestimate r when $\min\{N, T\} < 60$, but the criteria still select values of k that are very close to r for other configurations of the data.

The models considered thus far have idiosyncratic errors that are uncorrelated across units and across time. For these strict factor models, the preferred criteria are PC_{p1}, PC_{p2}, IC_1, and IC_2. It should be emphasized that the results reported are the averages of $\hat{k}$ over 1000 simulations. We do not report the standard deviations of these averages because they are identically zero except for a few

TABLE II

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $r = 3$; $\theta = 3$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	3.00	3.00	3.90	3.00	3.00	3.00	8.00	3.90	8.00	8.00	7.82	2.90
100	60	3.00	3.00	3.54	3.00	3.00	3.00	8.00	3.54	8.00	8.00	7.53	2.98
200	60	3.00	3.00	3.00	3.00	3.00	3.00	8.00	3.00	8.00	8.00	6.14	3.00
500	60	3.00	3.00	3.00	3.00	3.00	3.00	5.95	3.00	8.00	8.00	3.13	3.00
1000	60	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00
2000	60	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00
100	100	3.00	3.00	4.23	3.00	3.00	3.00	8.00	4.23	8.00	4.23	7.20	3.00
200	100	3.00	3.00	3.00	3.00	3.00	3.00	8.00	3.00	8.00	8.00	6.21	3.00
500	100	3.00	3.00	3.00	3.00	3.00	3.00	8.00	3.00	8.00	8.00	3.15	3.00
1000	100	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.00	8.00	8.00	3.00	3.00
2000	100	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00
40	100	3.00	3.00	3.70	3.00	3.00	3.00	8.00	8.00	8.00	3.70	7.63	2.92
60	100	3.00	3.00	3.42	3.00	3.00	3.00	8.00	8.00	8.00	3.42	7.39	2.99
60	200	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	8.00	3.00	5.83	3.00
60	500	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	5.44	3.00	3.03	3.00
60	1000	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00	3.00	3.00
60	2000	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00	3.00	3.00
4000	60	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	2.98
4000	100	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00
8000	60	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	2.97
8000	100	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00
60	4000	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00	3.00	2.99
100	4000	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00	3.00	3.00
60	8000	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00	3.00	2.98
100	8000	3.00	3.00	3.00	3.00	3.00	3.00	8.00	8.00	3.00	3.00	3.00	3.00
10	50	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.21
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.01
20	100	5.22	4.57	6.62	2.95	2.92	2.98	8.00	8.00	8.00	6.62	8.00	2.68
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
100	20	6.00	5.29	7.39	2.95	2.91	2.99	8.00	7.39	8.00	8.00	8.00	2.72

cases for which the average itself is not an integer. Even for these latter cases, the standard deviations do not exceed 0.6.

We next modify the assumption on the idiosyncratic errors to allow for serial and cross-section correlation. These errors are generated from the process

$$e_{it} = \rho e_{it-1} + v_{it} + \sum_{j \neq 0,\, j=-J}^{J} \beta v_{i-jt}.$$

The case of pure serial correlation obtains when the cross-section correlation parameter β is zero. Since for each i, the unconditional variance of e_{it} is $1/(1-\rho^2)$, the more persistent are the idiosyncratic errors, the larger are their variances relative to the common factors, and the precision of the estimates can be expected to fall. However, even with $\rho = .5$, Table VI shows that the estimates provided by the proposed criteria are still very good. The case of pure cross-

TABLE III

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $r = 5$; $\theta = 5$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	4.99	4.98	5.17	4.88	4.68	4.99	8.00	5.17	8.00	8.00	7.94	3.05
100	60	5.00	5.00	5.07	4.99	4.94	5.00	8.00	5.07	8.00	8.00	7.87	3.50
200	60	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	6.91	3.80
500	60	5.00	5.00	5.00	5.00	5.00	5.00	6.88	5.00	8.00	8.00	5.01	3.88
1000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	3.82
2000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	3.59
100	100	5.00	5.00	5.42	5.00	5.00	5.01	8.00	5.42	8.00	5.42	7.75	4.16
200	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	7.06	4.80
500	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	5.02	4.97
1000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	4.98
2000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	4.98
40	100	5.00	4.99	5.09	4.86	4.69	5.00	8.00	8.00	8.00	5.09	7.86	2.96
60	100	5.00	5.00	5.05	4.99	4.94	5.00	8.00	8.00	8.00	5.05	7.81	3.46
60	200	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	8.00	5.00	6.71	3.83
60	500	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	6.44	5.00	5.00	3.91
60	1000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.79
60	2000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.58
4000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	3.37
4000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	4.96
8000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	3.10
8000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	4.93
60	4000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.35
100	4000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	4.96
60	8000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.12
100	8000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	4.93
10	50	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.28
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.30
20	100	5.88	5.41	6.99	4.17	3.79	4.68	8.00	8.00	8.00	6.99	8.00	2.79
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
100	20	6.49	5.94	7.62	4.24	3.87	4.81	8.00	7.62	8.00	8.00	8.00	2.93

section dependence obtains with $\rho = 0$. As in Chamberlain and Rothschild (1983), our theory permits some degree of cross-section correlation. Given the assumed process for e_{it}, the amount of cross correlation depends on the number of units that are cross correlated ($2J$), as well as the magnitude of the pairwise correlation (β). We set β to .2 and J to $\max\{N/20, 10\}$. Effectively, when $N \leqq 200$, 10 percent of the units are cross correlated, and when $N > 200$, $20/N$ of the sample is cross correlated. As the results in Table VII indicate, the proposed criteria still give very good estimates of r and continue to do so for small variations in β and J. Table VIII reports results that allow for both serial and cross-section correlation. The variance of the idiosyncratic errors is now $(1 + 2J\beta^2)/(1 - \rho^2)$ times larger than the variance of the common component. While this reduces the precision of the estimates somewhat, the results generally confirm that a small degree of correlation in the idiosyncratic errors will not affect the properties of

TABLE IV

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $e_{it} = e_{it}^1 + \delta_t e_{it}^2$ ($\delta_t = 1$ FOR t EVEN, $\delta_t = 0$ FOR t ODD); $r = 5$; $\theta = 5$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	4.96	4.86	6.09	4.09	3.37	4.93	8.00	6.09	8.00	8.00	8.00	1.81
100	60	4.99	4.90	5.85	4.69	4.18	5.01	8.00	5.85	8.00	8.00	8.00	2.08
200	60	5.00	4.99	5.00	4.93	4.87	5.00	8.00	5.00	8.00	8.00	8.00	2.22
500	60	5.00	5.00	5.00	4.99	4.98	5.00	8.00	5.00	8.00	8.00	7.91	2.23
1000	60	5.00	5.00	5.00	5.00	5.00	5.00	7.97	5.00	8.00	8.00	6.47	2.02
2000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.51	5.00	8.00	8.00	5.03	1.72
100	100	5.00	4.98	6.60	4.98	4.79	5.24	8.00	6.60	8.00	6.60	8.00	2.56
200	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	3.33
500	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	7.94	3.93
1000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	6.13	3.98
2000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.36	5.00	8.00	8.00	5.00	3.85
40	100	4.94	4.80	5.39	4.04	3.30	4.90	8.00	8.00	8.00	5.39	7.99	1.68
60	100	4.98	4.88	5.41	4.66	4.14	5.00	8.00	8.00	8.00	5.41	7.99	2.04
60	200	5.00	4.99	5.00	4.95	4.87	5.00	8.00	8.00	8.00	5.00	7.56	2.14
60	500	5.00	5.00	5.00	4.99	4.98	5.00	8.00	8.00	7.29	5.00	5.07	2.13
60	1000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.90
60	2000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.59
4000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	1.46
4000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	3.67
8000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	1.16
8000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	3.37
60	4000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.30
100	4000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.62
60	8000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.08
100	8000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.29
10	50	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.27
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.34
20	100	6.13	5.62	7.23	2.85	2.23	3.93	8.00	8.00	8.00	7.23	8.00	1.86
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
100	20	7.52	6.99	7.99	3.31	2.64	6.17	8.00	7.99	8.00	8.00	8.00	2.30

the estimates. However, it will generally be true that for the proposed criteria to be as precise in approximate as in strict factor models, N has to be fairly large relative to J, β cannot be too large, and the errors cannot be too persistent as required by theory. It is also noteworthy that the BIC_3 has very good properties in the presence of cross-section correlations (see Tables VII and VIII) and the criterion can be useful in practice even though it does not satisfy all the conditions of Theorem 2.

6.1. *Application to Asset Returns*

Factor models for asset returns are extensively studied in the finance literature. An excellent summary on multifactor asset pricing models can be found in Campbell, Lo, and Mackinlay (1997). Two basic approaches are employed. One

TABLE V

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $r = 5$; $\theta = r \times 2$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	4.63	4.29	5.14	2.79	1.91	4.47	8.00	8.00	8.00	5.14	7.93	0.82
100	60	4.78	4.41	5.06	3.73	2.61	4.96	8.00	8.00	8.00	5.06	7.86	0.92
200	60	4.90	4.80	5.00	4.42	4.03	4.94	8.00	8.00	8.00	5.00	6.92	0.93
500	60	4.96	4.94	4.99	4.77	4.68	4.92	8.00	8.00	6.88	4.99	5.01	0.77
1000	60	4.97	4.97	4.98	4.88	4.86	4.93	8.00	8.00	5.00	4.98	5.00	0.56
2000	60	4.98	4.98	4.99	4.91	4.89	4.92	8.00	8.00	5.00	4.99	5.00	0.34
100	100	4.96	4.67	5.42	4.64	3.61	5.01	8.00	5.42	8.00	5.42	7.74	1.23
200	100	5.00	4.99	5.00	4.98	4.90	5.00	8.00	8.00	8.00	5.00	7.05	1.80
500	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	8.00	5.00	5.02	2.19
1000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	2.17
2000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	2.06
40	100	4.61	4.25	5.07	2.65	1.84	4.48	8.00	5.07	8.00	8.00	7.83	0.74
60	100	4.76	4.38	5.05	3.66	2.60	4.97	8.00	5.05	8.00	8.00	7.81	0.92
60	200	4.90	4.78	5.00	4.43	4.07	4.95	8.00	5.00	8.00	8.00	6.70	0.88
60	500	4.97	4.95	4.99	4.78	4.71	4.93	6.44	4.99	8.00	8.00	5.00	0.74
60	1000	4.98	4.97	4.99	4.87	4.84	4.92	5.00	4.99	8.00	8.00	5.00	0.51
60	2000	4.99	4.98	4.99	4.89	4.88	4.92	5.00	4.99	8.00	8.00	5.00	0.32
4000	60	4.99	4.99	4.99	4.92	4.92	4.93	8.00	8.00	5.00	4.99	5.00	0.18
4000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.72
8000	60	4.99	4.99	4.99	4.92	4.92	4.93	8.00	8.00	5.00	4.99	5.00	0.08
8000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.40
60	4000	4.99	4.99	4.99	4.93	4.92	4.95	5.00	4.99	8.00	8.00	5.00	0.15
100	4000	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	1.70
60	8000	4.99	4.99	4.99	4.92	4.92	4.93	5.00	4.99	8.00	8.00	5.00	0.08
100	8000	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	1.40
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.24
100	20	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.18
10	50	5.73	5.22	6.90	1.67	1.33	2.79	8.00	6.90	8.00	8.00	8.00	1.12
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
20	100	6.39	5.79	7.57	1.85	1.44	3.04	8.00	8.00	8.00	7.57	8.00	1.31

is statistical factor analysis of unobservable factors, and the other is regression analysis on observable factors. For the first approach, most studies use grouped data (portfolios) in order to satisfy the small N restriction imposed by classical factor analysis, with exceptions such as Connor and Korajczyk (1993). The second approach uses macroeconomic and financial market variables that are thought to capture systematic risks as observable factors. With the method developed in this paper, we can estimate the number of factors for the broad U.S. stock market, without the need to group the data, or without being specific about which observed series are good proxies for systematic risks.

Monthly data between 1994.1–1998.12 are available for the returns of 8436 stocks traded on the New York Stock Exchange, AMEX, and NASDAQ. The data include all lived stocks on the last trading day of 1998 and are obtained from the CRSP data base. Of these, returns for 4883 firms are available for each

TABLE VI

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $e_{it} = \rho e_{it-1} + v_{it} + \sum_{j=-J,\, j\neq 0}^{J} \beta v_{i-jt}$; $r = 5$; $\theta = 5$, $\rho = .5$, $\beta = 0$, $J = 0$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	7.31	6.59	8.00	5.52	4.53	8.00	8.00	8.00	8.00	8.00	8.00	2.97
100	60	6.11	5.27	8.00	5.00	4.76	8.00	8.00	8.00	8.00	8.00	8.00	3.09
200	60	5.94	5.38	7.88	5.01	4.99	7.39	8.00	7.88	8.00	8.00	8.00	3.31
500	60	5.68	5.39	6.79	5.00	5.00	5.11	8.00	6.79	8.00	8.00	8.00	3.41
1000	60	5.41	5.27	6.02	5.00	5.00	5.00	8.00	6.02	8.00	8.00	8.00	3.27
2000	60	5.21	5.14	5.50	5.00	5.00	5.00	8.00	5.50	8.00	8.00	8.00	3.06
100	100	5.04	5.00	8.00	5.00	4.97	8.00	8.00	8.00	8.00	8.00	8.00	3.45
200	100	5.00	5.00	7.75	5.00	5.00	7.12	8.00	7.75	8.00	8.00	8.00	4.26
500	100	5.00	5.00	5.21	5.00	5.00	5.00	8.00	5.21	8.00	8.00	8.00	4.68
1000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	4.73
2000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	4.69
40	100	5.37	5.05	7.30	4.58	4.08	5.82	8.00	8.00	8.00	7.30	8.00	2.45
60	100	5.13	4.99	7.88	4.93	4.67	7.40	8.00	8.00	8.00	7.88	8.00	2.80
60	200	5.00	5.00	5.02	4.99	4.96	5.00	8.00	8.00	8.00	5.02	8.00	2.84
60	500	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	8.00	5.00	7.53	2.72
60	1000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.72	5.00	5.04	2.54
60	2000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	2.28
4000	60	5.11	5.08	5.22	5.00	5.00	5.00	8.00	5.22	8.00	8.00	8.00	2.81
4000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	4.62
8000	60	5.05	5.05	5.08	5.00	5.00	5.00	8.00	5.08	8.00	8.00	8.00	2.55
8000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	4.37
60	4000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.92
100	4000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	4.21
60	8000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	1.64
100	8000	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00	5.00	3.97
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.47
100	20	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.69
10	50	7.16	6.68	7.89	3.57	2.92	5.70	8.00	8.00	8.00	7.89	8.00	2.42
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
20	100	8.00	7.99	8.00	7.93	7.58	8.00	8.00	8.00	8.00	8.00	8.00	3.92

of the 60 months. We use the proposed criteria to determine the number of factors. We transform the data so that each series is mean zero. For this balanced panel with $T = 60$, $N = 4883$ and $kmax = 15$, the recommended criteria, namely, PC_{p1}, PC_{p2}, IC_{p1}, and IC_{p2}, all suggest the presence of two factors.

7. CONCLUDING REMARKS

In this paper, we propose criteria for the selection of factors in large dimensional panels. The main appeal of our results is that they are developed under the assumption that $N, T \to \infty$ and are thus appropriate for many datasets typically used in macroeconomic analysis. Some degree of correlation in the errors is also allowed. The criteria should be useful in applications in which the number of factors has traditionally been assumed rather than determined by the data.

TABLE VII

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $e_{it} = \rho e_{it-1} + v_{it} + \sum_{j=-J,\, j\neq 0}^{J} \beta v_{i-jt}$; $r = 5$; $\theta = 5$, $\rho = 0.0$, $\beta = .20$, $J = \max\{N/20, 10\}$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	5.50	5.27	6.02	5.09	5.01	5.63	8.00	6.02	8.00	8.00	7.98	4.24
100	60	5.57	5.24	6.03	5.15	5.02	5.96	8.00	6.03	8.00	8.00	7.96	4.72
200	60	5.97	5.94	6.00	5.88	5.76	5.99	8.00	6.00	8.00	8.00	7.63	4.89
500	60	5.01	5.01	5.10	5.00	5.00	5.01	7.44	5.10	8.00	8.00	6.00	4.93
1000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.98	5.00	8.00	8.00	5.93	4.93
2000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.05	5.00	8.00	8.00	5.01	4.88
100	100	5.79	5.30	6.31	5.43	5.04	6.03	8.00	6.31	8.00	6.31	7.95	4.98
200	100	6.00	6.00	6.00	6.00	5.98	6.00	8.00	6.00	8.00	8.00	7.84	5.00
500	100	5.21	5.11	5.64	5.06	5.03	5.41	8.00	5.64	8.00	8.00	6.02	5.00
1000	100	5.00	5.00	5.00	5.00	5.00	5.00	6.00	5.00	8.00	8.00	6.00	5.00
2000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.72	5.00	8.00	8.00	5.41	5.00
40	100	5.17	5.06	5.95	5.00	4.98	5.30	8.00	8.00	8.00	5.95	7.96	4.22
60	100	5.30	5.06	6.01	5.03	5.00	5.87	8.00	8.00	8.00	6.01	7.94	4.69
60	200	5.35	5.16	5.95	5.04	5.01	5.65	8.00	8.00	8.00	5.95	7.39	4.89
60	500	5.43	5.29	5.83	5.05	5.02	5.35	8.00	8.00	7.49	5.83	6.04	4.94
60	1000	5.55	5.45	5.79	5.08	5.05	5.25	8.00	8.00	6.01	5.79	6.00	4.93
60	2000	5.64	5.59	5.76	5.07	5.04	5.17	8.00	8.00	6.00	5.76	6.00	4.91
4000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	4.84
4000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00
8000	60	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	4.72
8000	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8.00	8.00	5.00	5.00
60	4000	5.65	5.63	5.72	5.05	5.04	5.09	8.00	8.00	6.00	5.72	6.00	4.85
100	4000	6.00	6.00	6.00	6.00	6.00	6.00	8.00	8.00	6.14	6.00	6.02	5.00
60	8000	5.67	5.66	5.71	5.04	5.04	5.05	8.00	8.00	6.00	5.71	6.00	4.77
100	8000	6.00	6.00	6.00	6.00	6.00	6.00	8.00	8.00	6.00	6.00	6.00	5.00
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.34
100	20	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.49
10	50	6.23	5.84	7.18	4.82	4.67	5.14	8.00	8.00	8.00	7.18	8.00	3.72
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
20	100	6.75	6.27	7.75	4.97	4.73	5.71	8.00	7.75	8.00	8.00	8.00	3.81

Our discussion has focused on balanced panels. However, as discussed in Rubin and Thayer (1982) and Stock and Watson (1998), an iterative EM algorithm can be used to handle missing data. The idea is to replace X_{it} by its value as predicted by the parameters obtained from the last iteration when X_{it} is not observed. Thus, if $\lambda_i(j)$ and $F_t(j)$ are estimated values of λ_i and F_t from the jth iteration, let $X_{it}^*(j-1) = X_{it}$ if X_{it} is observed, and $X_{it}^*(j-1) = \lambda_i'(j-1) \times F_t(j-1)$ otherwise. We then minimize $V^*(k)$ with respect to $F(j)$ and $\Lambda(j)$, where $V^*(k) = (NT)^{-1} \sum_{i=1}^{T} \sum_{t=1}^{T} (X_{it}^*(j-1) - \lambda_i^k(j) F_t^k(j))^2$. Essentially, eigenvalues are computed for the $T \times T$ matrix $X^*(j-1)X^*(j-1)'$. This process is iterated until convergence is achieved.

Many issues in factor analysis await further research. Except for some results derived for classical factor models, little is known about the limiting distribution

TABLE VIII

DGP: $X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta} e_{it}$; $e_{it} = \rho e_{it-1} + v_{it} + \sum_{j=-J,\, j\neq 0}^{J} \beta v_{i-jt}$; $r = 5$; $\theta = 5$, $\rho = 0.50$, $\beta = .20$, $J = \max\{N/20, 10\}$.

N	T	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	AIC_1	BIC_1	AIC_2	BIC_2	AIC_3	BIC_3
100	40	7.54	6.92	8.00	6.43	5.52	8.00	8.00	8.00	8.00	8.00	8.00	4.14
100	60	6.57	5.93	8.00	5.68	5.28	8.00	8.00	8.00	8.00	8.00	8.00	4.39
200	60	6.52	6.15	7.97	6.00	5.91	7.84	8.00	7.97	8.00	8.00	8.00	4.68
500	60	6.16	5.97	7.12	5.40	5.30	5.92	8.00	7.12	8.00	8.00	8.00	4.76
1000	60	5.71	5.56	6.20	5.03	5.02	5.08	8.00	6.20	8.00	8.00	8.00	4.76
2000	60	5.33	5.26	5.61	5.00	5.00	5.00	8.00	5.61	8.00	8.00	8.00	4.69
100	100	5.98	5.71	8.00	5.72	5.27	8.00	8.00	8.00	8.00	8.00	8.00	4.80
200	100	6.01	6.00	7.95	6.00	5.99	7.78	8.00	7.95	8.00	8.00	8.00	5.03
500	100	5.89	5.81	6.06	5.59	5.46	5.94	8.00	6.06	8.00	8.00	8.00	5.00
1000	100	5.13	5.09	5.37	5.01	5.01	5.09	8.00	5.37	8.00	8.00	8.00	5.00
2000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	5.00
40	100	5.88	5.46	7.55	5.07	4.93	6.57	8.00	8.00	8.00	7.55	8.00	3.76
60	100	5.84	5.45	7.96	5.24	5.05	7.79	8.00	8.00	8.00	7.96	8.00	4.25
60	200	5.67	5.44	5.99	5.20	5.07	5.83	8.00	8.00	8.00	5.99	8.00	4.42
60	500	5.59	5.47	5.88	5.13	5.08	5.48	8.00	8.00	8.00	5.88	7.91	4.50
60	1000	5.61	5.54	5.81	5.13	5.08	5.34	8.00	8.00	6.91	5.81	6.15	4.40
60	2000	5.64	5.60	5.74	5.11	5.08	5.22	8.00	8.00	6.00	5.74	6.00	4.27
4000	60	5.12	5.10	5.24	5.00	5.00	5.00	8.00	5.24	8.00	8.00	8.00	4.56
4000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	5.00
8000	60	5.05	5.05	5.08	5.00	5.00	5.00	8.00	5.08	8.00	8.00	8.00	4.37
8000	100	5.00	5.00	5.00	5.00	5.00	5.00	8.00	5.00	8.00	8.00	8.00	5.00
60	4000	5.63	5.61	5.70	5.07	5.06	5.12	8.00	8.00	6.00	5.70	6.00	4.04
100	4000	6.00	6.00	6.00	6.00	6.00	6.00	8.00	8.00	6.44	6.00	6.17	5.00
60	8000	5.63	5.62	5.68	5.06	5.05	5.07	8.00	8.00	6.00	5.68	6.00	3.83
100	8000	6.00	6.00	6.00	6.00	6.00	6.00	8.00	8.00	6.08	6.00	6.02	5.00
100	10	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	7.54
100	20	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	6.85
10	50	7.34	6.87	7.93	4.84	4.37	6.82	8.00	8.00	8.00	7.93	8.00	3.41
10	100	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00
20	100	8.00	8.00	8.00	7.99	7.84	8.00	8.00	8.00	8.00	8.00	8.00	4.54

of the estimated common factors and common components (i.e., $\hat{\lambda}_i' \widehat{F}_t$). But using Theorem 1, it may be possible to obtain these limiting distributions. For example, the rate of convergence of $\widehat{F}_t$ derived in this paper could be used to examine the statistical property of the forecast $\hat{y}_{T+1|T}$ in Stock and Watson's framework. It would be useful to show that $\hat{y}_{T+1|T}$ is not only a consistent but a $\sqrt{T}$ consistent estimator of y_{T+1}, conditional on the information up to time T (provided that N is of no smaller order of magnitude than T). Additional asymptotic results are currently being investigated by the authors.

The foregoing analysis has assumed a static relationship between the observed data and the factors. Our model allows F_t to be a dependent process, e.g, $A(L)F_t = \epsilon_t$, where $A(L)$ is a polynomial matrix of the lag operator. However, we do not consider the case in which the dynamics enter into X_t directly. If the

method developed in this paper is applied to such a dynamic model, the estimated number of factors gives an upper bound of the true number of factors. Consider the data generating process $X_{it} = a_i F_t + b_i F_{t-1} + e_{it}$. From the dynamic point of view, there is only one factor. The static approach treats the model as having two factors, unless the factor loading matrix has a rank of one.

The literature on dynamic factor models is growing. Assuming N is fixed, Sargent and Sims (1977) and Geweke (1977) extended the static strict factor model to allow for dynamics. Stock and Watson (1998) suggested how dynamics can be introduced into factor models when both N and T are large, although their empirical applications assumed a static factor structure. Forni et al. (2000a) further allowed X_{it} to also depend on the leads of the factors and proposed a graphic approach for estimating the number of factors. However, determining the number of factors in a dynamic setting is a complex issue. We hope that the ideas and methodology introduced in this paper will shed light on a formal treatment of this problem.

Dept. of Economics, Boston College, Chestnut Hill, MA 02467, U.S.A.; jushan.bai@bc.edu

and

Dept. of Economics, Johns Hopkins University, Baltimore, MD 21218, U.S.A.; serena.ng@jhu.edu

Manuscript received April, 2000; final revision received December, 2000.

APPENDIX

To prove the main results we need the following lemma.

LEMMA 1: *Under Assumptions A–C, we have for some $M_1 < \infty$, and for all N and T,*

(i) $$T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\gamma_N(s,t)^2 \le M_1,$$

(ii) $$E\left(T^{-1}\sum_{t=1}^{T}\|N^{-1/2}e_t'\Lambda^0\|^2\right) = E\left(T^{-1}\sum_{t=1}^{T}\left\|N^{-1/2}\sum_{i=1}^{N}e_{it}\lambda_i^0\right\|^2\right) \le M_1,$$

(iii) $$E\left(T^{-2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(N^{-1}\sum_{i=1}^{N}X_{it}X_{is}\right)^2\right) \le M_1,$$

(iv) $$E\left\|(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\lambda_i^0\right\| \le M_1.$$

PROOF: Consider (i). Let $\rho(s,t) = \gamma_N(s,t)/[\gamma_N(s,s)\gamma_N(t,t)]^{1/2}$. Then $|\rho(s,t)| \le 1$. From $\gamma_N(s,s) \le M$,

$$\begin{aligned}
T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\gamma_N(s,t)^2 &= T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\gamma_N(s,s)\gamma_N(t,t)\rho(s,t)^2 \\
&\le MT^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,s)\gamma_N(t,t)|^{1/2}|\rho(s,t)| \\
&= MT^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \le M^2
\end{aligned}$$

by Assumption C2. Consider (ii).

$$E\left\|N^{-1/2}\sum_{i=1}^{N}e_{it}\lambda_i^0\right\|^2=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}E(e_{it}e_{jt})\lambda_i^{0\prime}\lambda_j^0\le\bar{\lambda}^2\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}|\le\bar{\lambda}^2M$$

by Assumptions B and C3. For (iii), it is sufficient to prove $E|X_{it}|^4\le M_1$ for all (i,t). Now $E|X_{it}|^4\le 8E(\lambda_i^{0\prime}F_t^0)^4+8E|e_{it}|^4\le 8\bar{\lambda}^4E\|F_t^0\|^4+8E|e_{it}|^4\le M_1$ for some M_1 by Assumptions A, B, and C1. Finally for (iv),

$$\begin{aligned}E\left\|(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\lambda_i^0\right\|^2&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E(e_{it}e_{js})\lambda_i^{0\prime}\lambda_j^0\\&\le\bar{\lambda}^2\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}|\tau_{ij,ts}|\le\bar{\lambda}^2M\end{aligned}$$

by Assumption C4.

PROOF OF THEOREM 1: We use the mathematical identity $\widehat{F}^k=N^{-1}X\widetilde{\Lambda}^k$ and $\widetilde{\Lambda}^k=T^{-1}X'\widetilde{F}^k$. From the normalization $\widetilde{F}^{k\prime}\widetilde{F}^k/T=I_k$, we also have $T^{-1}\sum_{t=1}^{T}\|\widetilde{F}_t^k\|^2=O_p(1)$. For $H^{k\prime}=(\widetilde{F}^{k\prime}F^0/T)(\Lambda^{0\prime}\Lambda^0/N)$, we have

$$\widehat{F}_t^k-H^{k\prime}F_t^0=T^{-1}\sum_{s=1}^{T}\widetilde{F}_s^k\gamma_N(s,t)+T^{-1}\sum_{s=1}^{T}\widetilde{F}_s^k\zeta_{st}+T^{-1}\sum_{s=1}^{T}\widetilde{F}_s^k\eta_{st}+T^{-1}\sum_{s=1}^{T}\widetilde{F}_s^k\xi_{st}\quad\text{where}$$

$$\zeta_{st}=\frac{e_s'e_t}{N}-\gamma_N(s,t),$$

$$\eta_{st}=F_s^{0\prime}\Lambda^{0\prime}e_t/N,$$

$$\xi_{st}=F_t^{0\prime}\Lambda^{0\prime}e_s/N=\eta_{ts}.$$

Note that H^k depends on N and T. Throughout, we will suppress this dependence to simplify the notation. We also note that $\|H^k\|=O_p(1)$ because $\|H^k\|\le\|\widetilde{F}^{k\prime}\widetilde{F}^k/T\|^{1/2}\|F^{0\prime}F^0/T\|^{1/2}\|\Lambda^{0\prime}\Lambda/N\|$ and each of the matrix norms is stochastically bounded by Assumptions A and B. Because $(x+y+z+u)^2\le 4(x^2+y^2+z^2+u^2)$, $\|\widehat{F}_t^k-H^{k\prime}F_t^0\|^2\le 4(a_t+b_t+c_t+d_t)$, where

$$a_t=T^{-2}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k\gamma_N(s,t)\right\|^2,$$

$$b_t=T^{-2}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k\zeta_{st}\right\|^2,$$

$$c_t=T^{-2}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k\eta_{st}\right\|^2,$$

$$d_t=T^{-2}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k\xi_{st}\right\|^2.$$

It follows that $(1/T)\sum_{t=1}^{T}\|\widehat{F}_t^k-H^{k\prime}F_t^0\|^2\le 1/T\sum_{t=1}^{T}(a_t+b_t+c_t+d_t)$.

Now $\|\sum_{s=1}^{T}\widetilde{F}_s^k\gamma_N(s,t)\|^2\le(\sum_{s=1}^{T}\|\widetilde{F}_s^k\|^2)\cdot(\sum_{s=1}^{T}\gamma_N^2(s,t))$. Thus,

$$\begin{aligned}T^{-1}\sum_{t=1}^{T}a_t&\le T^{-1}\left(T^{-1}\sum_{s=1}^{T}\left\|\widetilde{F}_s^k\right\|^2\right)\cdot T^{-1}\left(\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_N(s,t)^2\right)\\&=O_p(T^{-1})\end{aligned}$$

by Lemma 1(i).

For b_t, we have that

$$\begin{aligned}\sum_{t=1}^{T} b_t &= T^{-2}\sum_{t=1}^{T}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k\zeta_{st}\right\|^2\\ &= T^{-2}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{u=1}^{T}\widetilde{F}_s^{k\prime}\widetilde{F}_u^k\zeta_{st}\zeta_{ut}\\ &\le \left(T^{-2}\sum_{s=1}^{T}\sum_{u=1}^{T}(\widetilde{F}_s^{k\prime}\widetilde{F}_u^k)^2\right)^{1/2}\left[T^{-2}\sum_{s=1}^{T}\sum_{u=1}^{T}\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut}\right)^2\right]^{1/2}\\ &\le \left(T^{-1}\sum_{s=1}^{T}\|\widetilde{F}_s^k\|^2\right)\cdot\left[T^{-2}\sum_{s=1}^{T}\sum_{u=1}^{T}\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut}\right)^2\right]^{1/2}.\end{aligned}$$

From $E(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut})^2 = E(\sum_{t=1}^{T}\sum_{v=1}^{T}\zeta_{st}\zeta_{ut}\zeta_{sv}\zeta_{uv}) \le T^2\max_{s,t}E|\zeta_{st}|^4$ and

$$E|\zeta_{st}|^4 = \frac{1}{N^2}E\left|N^{-1/2}\sum_{i=1}^{N}(e_{it}e_{is}-E(e_{it}e_{is})\right|^4 \le N^{-2}M$$

by Assumption C5, we have

$$\sum_{t=1}^{T} b_t \le O_p(1)\cdot\sqrt{\frac{T^2}{N^2}} = O_p\left(\frac{T}{N}\right),$$

$T^{-1}\sum_{t=1}^{T} b_t = O_p(N^{-1})$. For c_t, we have

$$\begin{aligned}c_t &= T^{-2}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k\eta_{st}\right\|^2 = T^{-2}\left\|\sum_{s=1}^{T}\widetilde{F}_s^k F_s^{0\prime}\Lambda^{0\prime}e_t/N\right\|^2\\ &\le N^{-2}\|e_t'\Lambda^0\|^2\left(T^{-1}\sum_{s=1}^{T}\|\widetilde{F}_s^k\|^2\right)\left(T^{-1}\sum_{s=1}^{T}\|F_s^0\|^2\right)\\ &= N^{-2}\|e_t'\Lambda^0\|^2 O_p(1).\end{aligned}$$

It follows that

$$T^{-1}\sum_{t=1}^{T} c_t = O_p(1)N^{-1}T^{-1}\sum_{t=1}^{T}\left\|\frac{e_t'\Lambda^0}{\sqrt{N}}\right\|^2 = O_p(N^{-1}),$$

by Lemma 1(ii). The term $d_t = O_p(N^{-1})$ can be proved similarly. Combining these results, we have $T^{-1}\sum_{t=1}^{T}(a_t+b_t+c_t+d_t) = O_p(N^{-1})+O_p(T^{-1})$.

To prove Theorem 2, we need additional results.

LEMMA 2: *For any k, $1\le k\le r$, and H^k defined as the matrix in Theorem* 1,

$$V(k,\widehat{F}^k) - V(k,F^0H^k) = O_p(C_{NT}^{-1}).$$

PROOF: For the true factor matrix with r factors and H^k defined in Theorem 1, let $M_{FH}^0 = I - P_{FH}^0$ denote the idempotent matrix spanned by null space of F^0H^k. Correspondingly, let $M_{\widehat{F}}^k = I_T - \widehat{F}^k(\widehat{F}^{k\prime}\widehat{F}^k)^{-1}\widehat{F}^{k\prime} = I - P_{\widehat{F}}^k$. Then

$$V(k,\widehat{F}^k) = N^{-1}T^{-1}\sum_{i=1}^{N}\underline{X}_i'M_{\widehat{F}}^k\underline{X}_i,$$

$$V(k,F^0H^k) = N^{-1}T^{-1}\sum_{t=1}^{N}\underline{X}_i'M_{FH}^0\underline{X}_i,$$

$$V(k,\widehat{F}^k) - V(k,F^0H^k) = N^{-1}T^{-1}\sum_{i=1}^{N}\underline{X}_i'(P_{FH}^0 - P_{\widehat{F}}^k)\underline{X}_i.$$

Let $D_k = \widehat{F}^{k\prime}\widehat{F}^k/T$ and $D_0 = H^{k\prime}F^{0\prime}F^0H^k/T$. Then

$$\begin{aligned}
P^k_{\widehat{F}} - P^0_{FH} &= T^{-1}\widehat{F}^k\left(\frac{\widehat{F}^{k\prime}\widehat{F}^k}{T}\right)^{-1}\widehat{F}^{k\prime} - T^{-1}F^0H^k\left(\frac{H^{k\prime}F^{0\prime}F^0H^k}{T}\right)^{-1}H^{k\prime}F^{0\prime} \\
&= T^{-1}[\widehat{F}^{k\prime}D_k^{-1}\widehat{F}^k - F^0H^kD_0^{-1}H^{k\prime}F^{0\prime}] \\
&= T^{-1}[(\widehat{F}^k - F^0H^k + F^0H^k)D_k^{-1}(\widehat{F}^k - F^0H^k + F^0H^k)' - F^0H^kD_0^{-1}H^{k\prime}F^{0\prime}] \\
&= T^{-1}[(\widehat{F}^k - F^0H^k)D_k^{-1}(\widehat{F}^k - F^0H^k)' + (\widehat{F}^k - F^0H^k)D_k^{-1}H^{k\prime}F^{0\prime} \\
&\qquad + F^0H^kD_k^{-1}(\widehat{F}^k - F^0H^k)' + F^0H^k(D_k^{-1} - D_0^{-1})H^{k\prime}F^{0\prime}].
\end{aligned}$$

Thus, $N^{-1}T^{-1}\sum_{i=1}^N \underline{X}_i'(P^k_{\widehat{F}} - P^0_{FH})\underline{X}_i = I + II + III + IV$. We consider each term in turn.

$$\begin{aligned}
I &= N^{-1}T^{-2}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T(\widehat{F}^k_t - H^{k\prime}F^0_t)'D_k^{-1}(\widehat{F}^k_s - H^{k\prime}F^0_s)X_{it}X_{is} \\
&\le \left(T^{-2}\sum_{t=1}^T\sum_{s=1}^T[(\widehat{F}^k_t - H^{k\prime}F^0_t)'D_k^{-1}(\widehat{F}^k_s - H^{k\prime}F^0_s)]^2\right)^{1/2}\cdot\left[T^{-2}\sum_{t=1}^T\sum_{s=1}^T\left(N^{-1}\sum_{i=1}^N X_{it}X_{is}\right)^2\right]^{1/2} \\
&\le \left(T^{-1}\sum_{t=1}^T\|F^k_t - H^{k\prime}F^0_t\|^2\right)\cdot\|D_k^{-1}\|\cdot O_p(1) = O_p(C_{NT}^{-2})
\end{aligned}$$

by Theorem 1 and Lemma 1(iii). We used the fact that $\|D_k^{-1}\| = O_p(1)$, which is proved below.

$$\begin{aligned}
II &= N^{-1}T^{-2}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T(\widehat{F}^k_t - H^{k\prime}F^0_t)'D_k^{-1}H^{k\prime}F^0_sX_{it}X_{is} \\
&\le \left(T^{-2}\sum_{t=1}^T\sum_{s=1}^T\|\widehat{F}^k_t - H^{k\prime}F^0_t\|^2\cdot\|H^{k\prime}F^0_s\|^2\cdot\|D_k^{-1}\|^2\right)^{1/2}\cdot\left[T^{-2}\sum_{t=1}^T\sum_{s=1}^T\left(N^{-1}\sum_{i=1}^N X_{it}X_{is}\right)^2\right]^{1/2} \\
&\le \left(T^{-1}\sum_{t=1}^T\|\widehat{F}^k_t - H^{k\prime}F^0_t\|^2\right)^{1/2}\cdot\|D_k^{-1}\|\cdot\left(T^{-1}\sum_{s=1}^T\|H^{k\prime}F^0_s\|^2\right)^{1/2}\cdot O_p(1) \\
&= \left(T^{-1}\sum_{t=1}^T\|\widehat{F}^k_t - H^{k\prime}F^0_t\|^2\right)^{1/2}\cdot O_p(1) = O_p(C_{NT}^{-1}).
\end{aligned}$$

It can be verified that *III* is also $O_p(C_{NT}^{-1})$.

$$\begin{aligned}
IV &= N^{-1}T^{-2}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T F^{0\prime}_tH^k(D_k^{-1} - D_0^{-1})H^{k\prime}F^0_sX_{it}X_{is} \\
&\le \|D_k^{-1} - D_0^{-1}\|N^{-1}\sum_{i=1}^N\left(T^{-1}\sum_{t=1}^T\|H^{k\prime}F^0_t\|\cdot|X_{it}|\right)^2 \\
&= \|D_k^{-1} - D_0^{-1}\|\cdot O_p(1),
\end{aligned}$$

where $O_p(1)$ is obtained because the term is bounded by $\|H^k\|^2(1/T)\sum_{t=1}^T\|F^0_t\|^2(1/NT)\times\sum_{i=1}^N\sum_{t=1}^T|X_{it}|^2$, which is $O_p(1)$ by Assumption A and $E|X_{it}|^2 \le M$. Next, we prove that $\|D_k - D_0\| =$

$O_p(C_{NT}^{-1})$. From

$$\begin{aligned}
D_k - D_0 &= \frac{\widehat{F}^{k\prime}\widehat{F}^k}{T} - \frac{H^{k\prime}F^{0\prime}F^0H^k}{T} \\
&= T^{-1}\sum_{t=1}^{T}[\widehat{F}_t^k\widehat{F}_t^{k\prime} - H^{k\prime}F_t^0F_t^{0\prime}H^k] \\
&= T^{-1}\sum_{t=1}^{T}(\widehat{F}_t^k - H^{k\prime}F_t^0)(\widehat{F}_t^k - H^{k\prime}F_t^0)' \\
&\quad + T^{-1}\sum_{t=1}^{T}(\widehat{F}_t^k - H^{k\prime}F_t^0)F_t^{0\prime}H^k + T^{-1}\sum_{t=1}^{T}H^{k\prime}F_t^0(\widehat{F}_t^k - H^{k\prime}F_t^0)', \\
\|D_k - D_0\| &\le T^{-1}\sum_{t=1}^{T}\|\widehat{F}_t^k - H^{k\prime}F_t^0\|^2 + 2\left(T^{-1}\sum_{t=1}^{T}\|\widehat{F}_t^k - H^{k\prime}F_t^0\|^2\right)^{1/2}\cdot\left(T^{-1}\sum_{t=1}^{T}\|H^{k\prime}F_t^0\|^2\right)^{1/2} \\
&= O_p(C_{NT}^{-2}) + O_p(C_{NT}^{-1}) = O_p(C_{NT}^{-1}).
\end{aligned}$$

Because $F^{0\prime}F^0/T$ converges to a positive definite matrix, and because $\operatorname{rank}(H^k) = k \le r$, $D_0(k\times k)$ converges to a positive definite matrix. From $\|D_k - D_0\| = O_p(C_{NT}^{-1})$, D_k also converges to a positive definite matrix. This implies that $\|D_k^{-1}\| = O_p(1)$. Moreover, from $D_k^{-1} - D_0^{-1} = D_k^{-1}(D_0 - D_k)D_0^{-1}$ we have $\|D_k^{-1} - D_0^{-1}\| = \|D_k - D_0\|O_p(1) = O_p(C_{NT}^{-1})$. Thus $IV = O_p(C_{NT}^{-1})$.

LEMMA 3: *For the matrix H^k defined in Theorem* 1, *and for each k with $k < r$, there exists a $\tau_k > 0$ such that*

$$\operatorname{plim}\inf_{N,T\to\infty} V(k, F^0H^k) - V(r, F^0) = \tau_k.$$

PROOF:

$$\begin{aligned}
V(k, F^0H^k) - V(r, F^0) &= N^{-1}T^{-1}\sum_{i}^{N}\underline{X}_i'(P_F^0 - P_{FH}^0)\underline{X}_i \\
&= N^{-1}T^{-1}\sum_{i-1}^{N}(F^0\lambda_i^0 + \underline{e}_i)'(P_F^0 - P_{FH}^0)(F^0\lambda_i^0 + \underline{e}_i) \\
&= N^{-1}T^{-1}\sum_{i=1}^{N}\lambda_i^{0\prime}F^{0\prime}(P_F^0 - P_{FH}^0)F^0\lambda_i^0 \\
&\quad + 2N^{-1}T^{-1}\sum_{i=1}^{N}\underline{e}_i'(P_F^0 - P_{FH}^0)F^0\lambda_i^0 \\
&\quad + N^{-1}T^{-1}\sum_{i=1}^{N}\underline{e}_i'(P_F^0 - P_{FH}^0)\underline{e}_i \\
&= I + II + III.
\end{aligned}$$

First, note that $P_F^0 - P_{FH}^0 \ge 0$. Hence, $III \ge 0$. For the first two terms,

$$\begin{aligned}
I &= \operatorname{tr}\left[T^{-1}F^{0\prime}(P_F^0 - P_{FH}^0)F^0N^{-1}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\right] \\
&= \operatorname{tr}\left(\left[\frac{F^{0\prime}F^0}{T} - \frac{F^{0\prime}F^0H^k}{T}\left(\frac{H^{k\prime}F^{0\prime}F^0H^k}{T}\right)^{-1}\frac{H^{k\prime}F^{0\prime}F^0}{T}\right]\cdot N^{-1}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\right) \\
&\to \operatorname{tr}\left(\left[\Sigma_F - \Sigma_FH_0^k\left(H_0^{k\prime}\Sigma_FH_0^k\right)^{-1}H_0^{k\prime}\Sigma_F\right]\cdot D\right) \\
&= \operatorname{tr}(A\cdot D),
\end{aligned}$$

where $A = \Sigma_F - \Sigma_F H_0^k (H_0^{k'} \Sigma_F H_0^k)^{-1} H_0^{k'} \Sigma_F$ and H_0^k is the limit of H^k with $\text{rank}(H_0^k) = k < r$. Now $A \neq 0$ because $\text{rank}(\Sigma_F) = r$ (Assumption A). Also, A is positive semi-definite and $D > 0$ (Assumption B). This implies that $\text{tr}(A \cdot D) > 0$.

REMARK: Stock and Watson (1998) studied the limit of H^k. The convergence of H^k to H_0^k holds jointly in T and N and does not require any restriction between T and N.

Now

$$II = 2N^{-1}T^{-1}\sum_{i=1}^{N} \underline{e}_i' P_F^0 F^0 \lambda_i^0 - 2N^{-1}T^{-1}\sum_{i=1}^{N} \underline{e}_i' P_{FH}^0 F^0 \lambda_i^0.$$

Consider the first term.

$$\begin{aligned}\left| N^{-1}T^{-1}\sum_{i=1}^{N} \underline{e}_i' P_F^0 F^0 \lambda_i^0 \right| &= \left| N^{-1}T^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it} F_t^{0'} \lambda_i^0 \right| \\ &\leq \left(T^{-1}\sum_{t=1}^{T} \|F_t^0\|^2 \right)^{1/2} \cdot \frac{1}{\sqrt{N}} \left(T^{-1}\sum_{t=1}^{T} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_{it}\lambda_i^0 \right\|^2 \right)^{1/2} \\ &= O_p\left(\frac{1}{\sqrt{N}}\right).\end{aligned}$$

The last equality follows from Lemma 1(ii). The second term is also $O_p(1/\sqrt{N})$, and hence $II = O_p(1/\sqrt{N}) \to 0$.

LEMMA 4: *For any fixed k with $k \geq r$, $V(k, \widehat{F}^k) - V(r, \widehat{F}^r) = O_p(C_{NT}^{-2})$.*

PROOF:

$$\begin{aligned}|V(k, \widehat{F}^k) - V(r, \widehat{F}^r)| &\leq |V(k, \widehat{F}^k) - V(r, F^0)| + |V(r, F^0) - V(r, \widehat{F}^r)| \\ &\leq 2 \max_{r \leq k \leq kmax} |V(k, \widehat{F}^k) - V(r, F^0)|.\end{aligned}$$

Thus, it is sufficient to prove for each k with $k \geq r$,

$$V(k, \widehat{F}^k) - V(r, F^0) = O_p(C_{NT}^{-2}). \tag{10}$$

Let H^k be as defined in Theorem 1, now with rank r because $k \geq r$. Let H^{k+} be the generalized inverse of H^k such that $H^k H^{k+} = I_r$. From $\underline{X}_i = F^0 \lambda_i^0 + \underline{e}_i$, we have $\underline{X}_i = F^0 H^k H^{k+} \lambda_i^0 + \underline{e}_i$. This implies

$$\begin{aligned}\underline{X}_i &= \widehat{F}^k H^{k+} \lambda_i^0 + \underline{e}_i - (\widehat{F}^k - F^0 H^k) H^{k+} \lambda_i^0 \\ &= \widehat{F}^k H^{k+} \lambda_i^0 + \underline{u}_i,\end{aligned}$$

where $\underline{u}_i = \underline{e}_i - (\widehat{F}^k - F^0 H^k) H^{k+} \lambda_i^0$.

Note that

$$V(k,\widehat{F}^k)=N^{-1}T^{-1}\sum_{i=1}^{N}\underline{u}_i'M_{\widehat{F}}^k\underline{u}_i,$$

$$V(r,F^0)=N^{-1}T^{-1}\sum_{i=1}^{N}\underline{e}_i'M_{F}^0\underline{e}_i,$$

$$\begin{aligned}V(k,\widehat{F}^k)&=N^{-1}T^{-1}\sum_{i=1}^{N}(\underline{e}_i-(\widehat{F}^k-F^0H^k)H^{k+}\lambda_i^0)'M_{\widehat{F}}^k(\underline{e}_i-(\widehat{F}^k-F^0H^k)H^{k+}\lambda_i^0)\\&=N^{-1}T^{-1}\sum_{i=1}^{N}\underline{e}_i'M_{\widehat{F}}^k\underline{e}_i-2N^{-1}T^{-1}\sum_{i=1}^{N}\lambda_i^{0'}H^{k+\prime}(\widehat{F}^k-F^0H^k)'M_{\widehat{F}}^k\underline{e}_i\\&\quad+N^{-1}T^{-1}\sum_{i=1}^{N}\lambda_i^{0'}H^{k+\prime}(\widehat{F}^k-F^0H^k)'M_{\widehat{F}}^k(\widehat{F}^k-F^0H^k)H^{k+}\lambda_i^0\\&=a+b+c.\end{aligned}$$

Because $I-M_{\widehat{F}}^k$ is positive semi-definite, $x'M_{\widehat{F}}^kx\le x'x$. Thus,

$$\begin{aligned}c&\le N^{-1}T^{-1}\sum_{i=1}^{N}\lambda_i^{0'}H^{k+\prime}(\widehat{F}^k-F^0H^k)'(\widehat{F}^k-F^0H^k)H^{k+}\lambda_i^0\\&\le T^{-1}\sum_{t=1}^{T}\|\widehat{F}_t^k-H^{k\prime}F_t^0\|^2\cdot\left(N^{-1}\sum_{i=1}^{N}\|\lambda_i^0\|^2\|H^{k+}\|^2\right)\\&=O_p(C_{NT}^{-2})\cdot O_p(1)\end{aligned}$$

by Theorem 1. For term b, we use the fact that $|\mathrm{tr}(A)|\le r\|A\|$ for any $r\times r$ matrix A. Thus

$$\begin{aligned}b&=2T^{-1}\mathrm{tr}\left(H^{k+}(\widehat{F}^k-F^0H^k)'M_{\widehat{F}}^k\left(N^{-1}\sum_{i=1}^{N}\underline{e}_i\lambda_i^0\right)\right)\\&\le 2r\|H^{k+}\|\cdot\left\|\frac{\widehat{F}^k-F^0H^k}{\sqrt{T}}\right\|\cdot\left\|\frac{1}{\sqrt{T}N}\sum_{i=1}^{N}\underline{e}_i\lambda_i^0\right\|\\&\le 2r\|H^{k+}\|\cdot\left(T^{-1}\sum_{t=1}^{T}\|\widehat{F}_t^k-H^{k\prime}F_t^0\|^2\right)^{1/2}\cdot\frac{1}{\sqrt{N}}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}e_{it}\lambda_i^0\right\|^2\right)^{1/2}\\&=O_p(C_{NT}^{-1})\cdot\frac{1}{\sqrt{N}}=O_p(C_{NT}^{-2})\end{aligned}$$

by Theorem 1 and Lemma 1(ii). Therefore,

$$V(k,\widehat{F}^k)=N^{-1}T^{-1}\sum_{i=1}^{N}\underline{e}_i'M_{\widehat{F}}^k\underline{e}_i+O_p(C_{NT}^{-2}).$$

Using the fact that $V(k,\widehat{F}^k)-V(r,F^0)\le 0$ for $k\ge r$,

$$0\ge V(k,\widehat{F}^k)-V(r,F^0)=\frac{1}{NT}\sum_{i=1}^{N}\underline{e}_i'P_{\widehat{F}}^k\underline{e}_i-\frac{1}{NT}\sum_{i=1}^{N}\underline{e}_i'P_{F}^0\underline{e}_i+O_p(C_{NT}^{-2}).\tag{11}$$

Note that

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\underline{e}_i'P_F^0\underline{e}_i&\le\|(F^{0\prime}F^0/T)^{-1}\|\cdot N^{-1}T^{-2}\sum_{i=1}^{N}\underline{e}_i'F^0F^{0\prime}\underline{e}_i\\&=O_p(1)T^{-1}N^{-1}\sum_{i=1}^{N}\left\|T^{-1/2}\sum_{t=1}^{N}F_t^0e_{it}\right\|^2=O_p(T^{-1})\le O_p(C_{NT}^{-2})\end{aligned}$$

by Assumption D. Thus

$$0 \geq N^{-1}T^{-1}\sum_{i=1}^{N} \underline{e}_i' P_{\widehat{F}}^k \underline{e}_i + O_p(C_{NT}^{-2}).$$

This implies that $0 \leq N^{-1}T^{-1}\sum_{i=1}^{N} \underline{e}_i' P_{\widehat{F}}^k \underline{e}_i = O_p(C_{NT}^{-2})$. In summary

$$V(k, \widehat{F}^k) - V(r, F^0) = O_p(C_{NT}^{-2}).$$

PROOF OF THEOREM 2: We shall prove that $\lim_{N,T\to\infty} P(PC(k) < PC(r)) = 0$ for all $k \neq r$ and $k \leq kmax$. Since

$$PC(k) - PC(r) = V(k, \widehat{F}^k) - V(r, \widehat{F}^r) - (r-k)g(N,T),$$

it is sufficient to prove $P[V(k, \widehat{F}^k) - V(r, \widehat{F}^r) < (r-k)g(N,T)] \to 0$ as $N, T \to \infty$. Consider $k < r$. We have the identity:

$$\begin{aligned} V(k, \widehat{F}^k) - V(r, \widehat{F}^r) = {} & [V(k, \widehat{F}^k) - V(k, F^0H^k)] + [V(k, F^0H^k) - V(r, F^0H^r)] \\ & + [V(r, F^0H^r) - V(r, \widehat{F}^r)]. \end{aligned}$$

Lemma 2 implies that the first and the third terms are both $O_p(C_{NT}^{-1})$. Next, consider the second term. Because F^0H^r and F^0 span the same column space, $V(r, F^0H^r) = V(r, F^0)$. Thus the second term can be rewritten as $V(k, F^0H^k) - V(r, F^0)$, which has a positive limit by Lemma 3. Hence, $P[PC(k) < PC(r)] \to 0$ if $g(N,T) \to 0$ as $N, T \to \infty$. Next, for $k \geq r$,

$$P[PC(k) - PC(r) < 0] = P[V(r, \widehat{F}^r) - V(k, \widehat{F}^k) > (k-r)g(N,T)].$$

By Lemma 4, $V(r, \widehat{F}^r) - V(k, \widehat{F}^k) = O_p(C_{NT}^{-2})$. For $k > r$, $(k-r)g(N,T) \geq g(N,T)$, which converges to zero at a slower rate than C_{NT}^{-2}. Thus for $k > r$, $P[PC(k) < PC(r)] \to 0$ as $N, T \to \infty$.

PROOF OF COROLLARY 1: Denote $V(k, \widehat{F}^k)$ by $V(k)$ for all k. Then

$$IC(k) - IC(r) = \ln[V(k)/V(r)] + (k-r)g(N,T).$$

For $k < r$, Lemmas 2 and 3 imply that $V(k)/V(r) > 1 + \epsilon_0$ for some $\epsilon_0 > 0$ with large probability for all large N and T. Thus $\ln[V(k)/V(r)] \geq \epsilon_0/2$ for large N and T. Because $g(N,T) \to 0$, we have $IC(k) - IC(r) \geq \epsilon_0/2 - (r-k)g(N,T) \geq \epsilon_0/3$ for large N and T with large probability. Thus, $P[IC(k) - IC(r) < 0] \to 0$. Next, consider $k > r$. Lemma 4 implies that $V(k)/V(r) = 1 + O_p(C_{NT}^{-2})$. Thus $\ln[V(k)/V(r)] = O_p(C_{NT}^{-2})$. Because $(k-r)g(N,T) \geq g(N,T)$, which converges to zero at a slower rate than C_{NT}^{-2}, it follows that

$$P[IC(k) - IC(r) < 0] \leq P\big[O_p\big(C_{NT}^{-2}\big) + g(N,T) < 0\big] \to 0.$$

PROOF OF COROLLARY 2: Theorem 2 is based on Lemmas 2, 3, and 4. Lemmas 2 and 3 are still valid with F^k replaced by $\widehat{G}^k$ and C_{NT} replaced by $\widetilde{C}_{NT}$. This is because their proof only uses the convergence rate of $\widehat{F}_t$ given in (5), which is replaced by (8). But the proof of Lemma 4 does make use of the principle component property of $\widehat{F}^k$ such that $V(k, \widehat{F}^k) - V(r, F^0) \leq 0$ for $k \geq r$, which is not necessarily true for $\widehat{G}^k$. We shall prove that Lemma 4 still holds when $\widehat{F}^k$ is replaced by $\widehat{G}^k$ and C_{NT} is replaced by $\widetilde{C}_{NT}$. That is, for $k \geq r$,

$$V(k, \widehat{G}^k) - V(r, \widehat{G}^r) = O_p\big(\widetilde{C}_{NT}^{-2}\big). \tag{12}$$

Using arguments similar to those leading to (10), it is sufficient to show that

$$V(k, \widehat{G}^k) - V(r, F^0) = O_p\big(\widetilde{C}_{NT}^{-2}\big). \tag{13}$$

Note that for $k \geq r$,

$$V(k, \widehat{F}^k) \leq V(k, \widehat{G}^k) \leq V(r, \widehat{G}^r). \tag{14}$$

The first inequality follows from the definition that the principal component estimator gives the smallest sum of squared residuals, and the second inequality follows from the least squares property that adding more regressors does not increase the sum of squared residuals. Because $\widetilde{C}_{NT}^2 \leq C_{NT}^2$, we can rewrite (10) as

$$V(k, \widehat{F}^k) - V(r, F^0) = O_p(\widetilde{C}_{NT}^{-2}). \tag{15}$$

It follows that if we can prove

$$V(r, \widehat{G}^r) - V(r, F^0) = O_p(\widetilde{C}_{NT}^{-2}), \tag{16}$$

then (14), (15), and (16) imply (13). To prove (16), we follow the same arguments as in the proof of Lemma 4 to obtain

$$V(r, \widehat{G}^r) - V(r, F^0) = \frac{1}{NT}\sum_{i=1}^{N} \underline{e}_i' P_{\widehat{G}}^r \underline{e}_i - \frac{1}{NT}\sum_{i=1}^{N} \underline{e}_i' P_F^0 \underline{e}_i + O_p(\widetilde{C}_{NT}^{-2}),$$

where $P_{\widehat{G}}^r = \widehat{G}^r(\widehat{G}^{r'}\widehat{G}^r)^{-1}\widehat{G}^{r'}$; see (11). Because the second term on the right-hand side is shown in Lemma 4 to be $O_p(T^{-1})$, it suffices to prove the first term is $O_p(\widetilde{C}_{NT}^{-2})$. Now,

$$\frac{1}{NT}\sum_{i=1}^{N} \underline{e}_i' P_{\widehat{G}}^r \underline{e}_i \leq \|(\widehat{G}^{r'}\widehat{G}^r/T)^{-1}\| \frac{1}{N}\sum_{i=1}^{N} \|\underline{e}_i'\widehat{G}^r/T\|^2.$$

Because $\widetilde{H}^r$ is of full rank, we have $\|(\hat{G}^{r'}\hat{G}^r/T)^{-1}\| = O_p(1)$ (follows from the same arguments in proving $\|D_k^{-1}\| = O_p(1)$). Next,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \|\underline{e}_i'\hat{G}^r/T\|^2 &\leq \left(\frac{1}{NT}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} F_t^0 e_{it}\right\|^2\right)\|\widetilde{H}^r\|^2 + \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2\right)\frac{1}{T}\sum_{t=1}^{T}\|\widehat{G}_t^r - \widetilde{H}^{r'}F_t^0\|^2 \\ &= O_p(T^{-1})O_p(1) + O_p(1)O_p(\widetilde{C}_{NT}^{-2}) = O_p(\widetilde{C}_{NT}^{-2})\end{aligned}$$

by Assumption D and (8). This completes the proof of (16) and hence Corollary 2.

REFERENCES

ANDERSON, T. W. (1984): *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

BACKUS, D., S. FORSEI, A. MOZUMDAR, AND L. WU (1997): "Predictable Changes in Yields and Forward Rates," Mimeo, Stern School of Business.

CAMPBELL, J., A. W. LO, AND A. C. MACKINLAY (1997): *The Econometrics of Financial Markets*. Princeton, New Jersey: Princeton University Press.

CHAMBERLAIN, G., AND M. ROTHSCHILD (1983): "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets," *Econometrica*, 51, 1305–1324.

COCHRANE, J. (1999): "New Facts in Finance, and Portfolio Advice for a Multifactor World," NBER Working Paper 7170.

CONNOR, G., AND R. KORAJZCYK (1986): "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, 15, 373–394.

——— (1988): "Risk and Return in an Equilibrium APT: Application to a New Test Methodology," *Journal of Financial Economics*, 21, 255–289.

——— (1993): "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48, 1263–1291.

CRAGG, J., AND S. DONALD (1997): "Inferring the Rank of a Matrix," *Journal of Econometrics*, 76, 223–250.

DHRYMES, P. J., I. FRIEND, AND N. B. GLUTEKIN (1984): "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory," *Journal of Finance*, 39, 323–346.

DONALD, S. (1997): "Inference Concerning the Number of Factors in a Multivariate Nonparameteric Relationship," *Econometrica*, 65, 103–132.

FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2000a): "The Generalized Dynamic Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82, 540–554.

——— (2000b): "Reference Cycles: The NBER Methodology Revisited," CEPR Discussion Paper 2400.

FORNI, M., AND M. LIPPI (1997): *Aggregation and the Microfoundations of Dynamic Macroeconomics*. Oxford, U.K.: Oxford University Press.

——— (2000): "The Generalized Dynamic Factor Model: Representation Theory," Mimeo, Universitá di Modena.

FORNI, M., AND L. REICHLIN (1998): "Let's Get Real: a Factor-Analytic Approach to Disaggregated Business Cycle Dynamics," *Review of Economic Studies*, 65, 453–473.

GEWEKE, J. (1977): "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio Economic Models*, ed. by D. J. Aigner and A. S. Goldberger. Amsterdam: North Holland.

GEWEKE, J., AND R. MEESE (1981): "Estimating Regression Models of Finite but Unknown Order," *International Economic Review*, 23, 55–70.

GHYSELS, E., AND S. NG (1998): "A Semi-parametric Factor Model for Interest Rates and Spreads," *Review of Economics and Statistics*, 80, 489–502.

GREGORY, A., AND A. HEAD (1999): "Common and Country-Specific Fluctuations in Productivity, Investment, and the Current Account," *Journal of Monetary Economics*, 44, 423–452.

GREGORY, A., A. HEAD, AND J. RAYNAULD (1997): "Measuring World Business Cycles," *International Economic Review*, 38, 677–701.

LEHMANN, B. N., AND D. MODEST (1988): "The Empirical Foundations of the Arbitrage Pricing Theory," *Journal of Financial Economics*, 21, 213–254.

LEWBEL, A. (1991): "The Rank of Demand Systems: Theory and Nonparametric Estimation," *Econometrica*, 59, 711–730.

MALLOWS, C. L. (1973): "Some Comments on C_p," *Technometrics*, 15, 661–675.

ROSS, S. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Finance*, 13, 341–360.

RUBIN, D. B., AND D. T. THAYER (1982): "EM Algorithms for ML Factor Analysis," *Psychometrika*, 57, 69–76.

SARGENT, T., AND C. SIMS (1977): "Business Cycle Modelling without Pretending to Have too much a Priori Economic Theory," in *New Methods in Business Cycle Research*, ed. by C. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

SCHWERT, G. W. (1989): "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 7, 147–160.

STOCK, J. H., AND M. WATSON (1989): "New Indexes of Coincident and Leading Economic Indications," in *NBER Macroeconomics Annual 1989*, ed. by O. J. Blanchard and S. Fischer. Cambridge: M.I.T. Press.

——— (1998): "Diffusion Indexes," NBER Working Paper 6702.

——— (1999): "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293–335.