EFWW
7/11/07
11:00 am

A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models*

Catherine Doz, Université Cergy-Pontoise
Domenico Giannone, Université Libre de Bruxelles, ECARES and CEPR
Lucrezia Reichlin, European Central Bank, ECARES and CEPR

April 12, 2007

Abstract

This paper considers maximum likelihood estimation of a dynamic approximate factor model when the panel of time series is large (n large). Maximum likelihood is analyzed under two sources of misspecification: omitted serial correlation and cross-sectional correlation of the idiosyncratic components. It is shown that the quasi maximum likelihood (QML) estimator is feasible for large approximate dynamic factor models and can be easily implemented using the Kalman smoother and the EM algorithm as in traditional factor analysis. The QML estimator is therefore a valid alternative to principal components studied by Bai (2003), Bai and Ng (2002), Forni, Hallin, Lippi, and Reichlin (2000, 2005b), Stock and Watson (2002a,b).

JEL Classification: C51, C32, C33.

Keywords: Factor Model, large cross-sections, Quasi Maximum Likelihood.

* We would like to thank Ursula Gather, Massimo Franchi and Marco Lippi for helpful suggestions and seminar participants at the International Statistical Institute in Berlin 2003, the European Central Bank, 2003, the Statistical Institute at the Catholic University of Louvain la Neuve, 2004, the Institute for Advanced Studies in Vienna, 2004, the Department of Statistics at Carlos III University, Madrid, 2004, the Federal Reserve Board of Governors, 2006, the Department of Economics at New York University, 2006, the Department of Economics, Columbia University, 2006. The opinions in this paper are those of the authors and do not necessarily reflect the views of the European Central Bank. Please address any comments to Catherine Doz doz@u-cergy.fr; Domenico Giannone dgiannon@ulb.ac.be; or Lucrezia Reichlin lucrezia.reichlin@ecb.int

1 Introduction

The idea that the dynamics of large panels of time series can be characterized as being driven by few common factors and a variable specific-idiosyncratic component is appealing for macroeconomic and finance applications where data are strongly collinear. Applications in macroeconomics date back to the seventies (Geweke, 1977; Sargent and Sims, 1977; Geweke and Singleton, 1980). In finance the factor model has also a long tradition since it relates closely to the CAPM model of asset prices.

In traditional factor analysis, for a given size of the cross-section n, the model can be consistently estimated by maximum likelihood. The literature has proposed both frequency domain (Geweke, 1977; Sargent and Sims, 1977; Geweke and Singleton, 1980) and time domain (Engle and Watson, 1981; Stock and Watson, 1989; Quah and Sargent, 1992) methods.

Identification is achieved by assuming that, for each series, the component driven by the common factors (common component) is orthogonal to the idiosyncratic component and the idiosyncratic component has cross-sectionally orthogonal elements. A factor model with orthogonal idiosyncratic elements is called an exact factor model.

Although the idea of factor analysis is appealing, the traditional approach presents some limitations. First of all, the assumption of an exact factor structure is an excessive straightjacket on the data, leading to potentially harmful misspecification problems. In particular, with large panels, the assumption of orthogonal idiosyncratic elements is likely to be less adequate than with panels including a small number of aggregate variables. Second, although the coefficients of the factor loadings can be consistently estimated for T large via maximum likelihood, the factors are indeterminate and one can only obtain their expected value (on this point, see Steiger, 1979). Third, many empirically interesting economic applications require the study of large panels, situation in which the properties of the maximum likelihood estimates are unknown and where maximum likelihood is generally considered not feasible (Bai, 2003; Bai and Ng, 2002).

As a response to these limitations, recent literature has generalized the idea of factor analysis to handle less strict assumptions on the covariance of the idiosyncratic elements (approximate factor structure) and proposed non-parametric estimators of the common factors based on principal components, which are feasible for n large (Forni, Hallin, Lippi, and Reichlin, 2000; Stock and Watson, 2002a,b).

A key feature of this approach is that consistency is analyzed as n, as well as T, go to infinity. It is shown that, under suitable assumptions, if the cross-sectional dimension n tends to infinity, the principal components of the observations become increasingly collinear with the common factors (Chamberlain, 1983; Chamberlain and Rothschild, 1983; Forni, Hallin, Lippi, and Reichlin, 2000; Forni and Lippi, 2001). Principal components are also proved to be n, T consistent estimators of the factor space (Bai, 2003; Bai and Ng, 2002; Forni, Hallin, Lippi, and Reichlin, 2000, 2005b; Stock and Watson, 2002a,b; Forni, Giannone, Lippi, and Reichlin, 2005a).

The approximate factor model presents several advantages with respect to the exact model. It is very flexible and suitable under general assumptions on measurement error, geographical clustering and, in general, local cross correlation. However, the maximum likelihood estimator has never been analyzed for this model. The reason is that, in order to estimate the model by maximum likelihood, it is necessary to impose a parametrization while retaining parsimony. In the exact factor model parsimony is achieved by restricting the cross-correlation among idiosyncratic components to be zero. Once this restriction is relaxed, there is no obvious way to model the cross-sectional correlation among idiosyncratic terms since in the cross-section there is no natural order.

This paper studies maximum likelihood estimation for the approximate factor model for large cross-sections. The central idea is to treat the exact factor model as a misspecified approximating model and analyze the properties, for n and T going to infinity, of the maximum likelihood estimator of the factors under misspecification, that is when the true probabilistic model is approximated by a more restricted model. This is a quasi maximum likelihood estimator (QML) in the sense of White (1982). We derive the n, T rates of convergence for the implied estimates of the common factors. We show that traditional factor analysis in large cross-section is feasible and that consistency is achieved even if the underlying data generating process is an approximate factor model rather than an exact one. More precisely, our consistency result shows that the expected value of the common factors converges to the true factors as $n, T \to \infty$ along any path (we also provide the consistency rates).

This result tells us that the misspecification error due to the approximate structure of the idiosyncratic component vanishes asymptotically, for n and T large, provided that the cross-correlation of the idiosyncratic processes is limited and that of the common components is pervasive throughout the cross section as n increases. These are conditions that have been introduced by Chamberlain and Rothschild (1983) and used, reinterpreted and extended by Connor and Korajczyk (1986, 1988, 1993); Forni, Hallin, Lippi, and Reichlin (2000); Forni and Lippi (2001); Stock and Watson (2002a,b).

Our result should be interpreted as a reconciliation of the classical factor analysis approach with the new generation of dynamic factor models with n large in which the common factors are estimated by principal components. We show that these two approaches are related in the sense that principal components estimators can be reinterpreted as quasi-maximum likelihood estimators, i.e. maximum likelihood under a misspecified model where data are assumed to be generated by a factor model with spherical idiosyncratic components and non serially correlated observations.

From the practical point of view we show that, unlike what sometimes claimed in the literature, classical likelihood based methods are feasible in the large n case. Under standard parameterizations, the factor model can in fact be cast in a state space form and the likelihood can be maximized via the EM algorithm which requires at each iteration only one run of the Kalman smoother (Engle and Watson, 1981). Under the exact factor structure restriction on the approximating model, the computational complexity of the smoother depends essentially on the number of common factors r which is typically small. The intuition of why this works was first suggested by Quah and Sargent (1992) who estimated a model with $n = 60$. Furthermore, principal components provide a good approximation of the common factors in a large cross-section, they can be used to obtain a good initial estimate of the parameters to initialize the numerical algorithm for maximum likelihood estimation.

There are many reasons why our result is a useful contribution to the literature of factor models in large panels. First, the n, T asymptotic properties of maximum likelihood estimation of factor models have never been studies. Second, maximum likelihood estimation is particularly attractive for economic applications since it provides a framework for incorporating restrictions deriving from economic theory in the statistical model. Indeed, an increasing number of studies in macroeconomics have used likelihood based Bayesian methods for extracting the common factors from a large panel of time series (Kose, Otrok, and Whiteman, 2003; Boivin and Giannoni, 2005; Bernanke, Boivin, and Eliasz, 2005). However, in that approach, the model does not allow for correlation amongst idiosyncratic components, that is an exact factor structure is imposed in estimation. One implication of the result of our paper is that misspecification is harmless in large panels since its effect vanishes asymptotically. Moreover, even assuming that the true model has an exact factor structure, the asymptotic properties of the estimates, when both the sample size and the cross-sectional dimension are large, have not been studied. Third, if the true data generating process (DGP) and the approximating model coincide, then maximum likelihood estimates are the most efficient.

The paper is organized as follows. Section two states the assumptions for the model generating the model and those for the approximating model we will use in estimation. Section three states the basic proposition showing consistency and rates for the quasi maximum likelihood estimator. Section four discusses the relations between quasi maximum likelihood and principal components estimator. Section five illustrates the empirical results and Section six concludes.

2 Models

2.1 Notation

For any positive definite square matrix M, we will denote by $\lambda_{max}(M)$ ($\lambda_{min}(M)$) its largest (smallest) eigenvalue. Moreover, for any matrix M we will denote by $\|M\|$ the spectral norm defined as $\|M\| = \sqrt{\lambda_{max}(M'M)}$. Given a stochastic process $\{X_{n,T}; T \in \mathbb{Z}, n \in \mathbb{Z}\}$, and a real sequence $\{a_{n,T}; T \in \mathbb{Z}, n \in \mathbb{Z}\}$ we will say that $X_{n,T} = O_P\left(\frac{1}{a_{nT}}\right)$ as $n, T \to \infty$, if the probability that $a_{n,T}X_{n,T}$ is bounded tends to one as $n, T \to \infty$.

2.2 The approximate dynamic factor model

We suppose that an n-dimensional zero-mean stationary process $\mathbf{x}_t$ is the sum of two unobservable components:

$$\mathbf{x}_t = \Lambda_0 \mathbf{f}_t + \mathbf{e}_t \tag{2.1}$$

where $\mathbf{f}_t = (f_{1t}, ..., f_{rt})'$, the common factors, is an r-dimensional stationary process with mean zero; Λ_0, the factor loadings, is an $n \times r$ matrix; $\mathbf{e}_t = (e_{1t}, ..., e_{nt})'$, the idiosyncratic components, is an n-dimensional stationary process with mean zero and covariance matrix $\mathrm{E}(\mathbf{e}_t\mathbf{e}_t') = \Psi_0$, whose entries will be denoted by $\mathrm{E}(e_{it}e_{jt}) = \psi_{0,ij}$. The common factors $\mathbf{f}_t$ and the idiosyncratic component $\mathbf{e}_t$ are assumed to be uncorrelated at all leads and lags, that is $\mathrm{E}(f_{jt}e_{is}) = 0$ for all $j = 1, ..., r$, $i = 1, ..., n$ and $t, s \in \mathbb{Z}$. The number of common factors r is typically much smaller than the cross-sectional dimension n.

Given a sample of size T, we will denote by capital cases the matrices collecting all the variables, that is $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)'$ is the $T \times n$ matrix of observables, $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_T)'$ is the $T \times r$ matrix of common factors and $\mathbf{E} = (\mathbf{e}_1, ..., \mathbf{e}_T)'$. All these quantities depend on the size of the cross-section and on the sample size. For notational convenience we will not index them by n, T.

The following assumptions define an approximate dynamic factor model. "Approximate" stands for a model that allows for limited crosss-correlation among idiosyncratic components (Chamberlain and Rothschild, 1983). This is to be distinguished from the "exact factor model" whose idiosyncratic elements are restricted to be cross-sectionally orthogonal.[1] The model is dynamic since we allow for weak serial correlations of the common factor and the idiosyncratic components. Approximate factor models for dynamic panels have been studied, under similar assumptions, by Bai and Ng (2002, 2006); Forni, Giannone, Lippi, and Reichlin (2005a); Forni, Hallin, Lippi, and Reichlin (2000, 2005b); Stock and Watson (2002a,b).

Assumption A (Approximate factor model)

A1 $0 < \underline{\lambda} < \liminf_{n\to\infty} \frac{1}{n}\lambda_{min}\left(\Lambda_0'\Lambda_0\right) \leq \limsup_{n\to\infty} \lambda_{max}\frac{1}{n}\left(\Lambda_0'\Lambda_0\right) < \bar{\lambda} < \infty$

A2 $0 < \underline{\psi} < \liminf_{n\to\infty} \lambda_{min}\left(\Psi_0\right) \leq \limsup_{n\to\infty} \lambda_{max}\left(\Psi_0\right) < \bar{\psi} < \infty$

Assumption B
There exists a positive constant M such that for all $i, j \in \mathbb{N}$ and for all $T \in \mathbb{Z}$

i) $\mathrm{E}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(e_{it}e_{jt} - \psi_{0,ij})\right)^2 < M$

ii) $\mathrm{E}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t e_{jt}\right\|^2 < M$

iii) $\mathrm{E}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\mathbf{f}_t\mathbf{f}_t' - I_r)\right\|^2 < M$

[1] Exact factor model have been studied and applied in econometrics by Engle and Watson (1981); Geweke (1977); Kose, Otrok, and Whiteman (2003); Quah and Sargent (1992); Sargent and Sims (1977); Stock and Watson (1991), among others.

Assumption A1 entails that for n sufficiently large $\Lambda_0'\Lambda_0/n$ has full rank r. Under this assumption the common factors are required to remain pervasive as we increase the number of series in the data-set. Assumption A2 limits the cross-correlation of the idiosyncratic components. While it includes the case in which they are mutually orthogonal ("exact factor model"), it allows for a more general structure.

Assumption B requires insures that the sample covariance matrix of the common factors and the idiosyncratic component are $\sqrt{T}$ consistent to their population counterpart, uniformly with respect to the cross-sectional dimension. Precisely:

In order to estimate this model by maximum likelihood we need to impose a parameterizations that is sufficiently parsimonious. Parsimony is achieved in the exact factor model by restricting the cross-correlation among idiosyncratic components to be zero. Once this restriction is relaxed, as in Assumption A2, there is no obvious way to model the cross-sectional correlation among idiosyncratic terms since in the cross-section there is no natural order.

We proceed as follows. First, we will define an approximating models that restricts the idiosyncratic components to be neither cross-sectionally nor serially correlated. We will consider this "exact factor model" as a miss-specified approximation to model 2.1. Second, we will prove that the effects of missspecification due to the approximation vanishes as $n, T \to \infty$, under Assumptions A and B.

2.3 The approximating models

An approximating model is a possibly misspecified model that we will use to define the likelihood. A natural candidate is the model that has been used in traditional exact factor analysis for small cross-section (see, for example, Stock and Watson, 1991).

Approximating model: the exact factor model

R1 the common factors follow a finite order Gaussian VAR: $A(L)\mathbf{f}_t = \mathbf{u}_t$, with $A(L) = I - A_1L - ... - A_pL^p$ an $r \times r$ filter of finite length p with roots outside the unit circle, and $\mathbf{u}_t$ an r dimensional gaussian white noise, $\mathbf{u}_t \sim$ i.i.d $\mathcal{N}(0, Q)$.

R2 the idiosyncratic components are cross-sectionally independent gaussian white noises: $\mathbf{e}_t \sim$ i.i.d $\mathcal{N}(0, \Psi_d)$ where Ψ_d is a diagonal matrix.

Under assumptions R1 and R2, the model can be cast in a state space form with the number of states equal the number of common factors r. For any set of parameters the likelihood can then be evaluated using the Kalman filter.

The model is characterized by the quadruplet $\Lambda, \Psi_d, A(L), Q$. All the parameters will be collected into $\theta \in \Theta$, where Θ is the parameter space defined by R1 and R2.

Given the quasi maximum likelihood estimates of the parameters θ, the common factors can be approximated by their expected value, which can be computed using the Kalman smoother[2]:

$$\hat{\mathbf{F}}_{\hat{\theta}} = \mathrm{E}_{\hat{\theta}}\left[\mathbf{F}|\mathbf{X}\right]$$

where $\hat{\mathbf{F}}_{\hat{\theta}} = \left(\hat{\mathbf{f}}_{\hat{\theta}1}, ..., \hat{\mathbf{f}}_{\hat{\theta}T}\right)'$.

The idiosyncratic components are modelled as a cross-sectionally independent and non serially correlated Gaussian processes. The orthogonality restriction among the idiosyncratic components is key to maintain parsimony and identification.[3]

Asymptotic properties of the estimator are known to for n is fixed and $T \to \infty$ and under the assumption that data are generated from an "exact factor structure" (see Engle and Watson, 1981; Stock and Watson, 1991, for example). In what follows, we extend previous studies by considering joint n, T asymptotic and under the more general assumption that data are generated from an "approximate dynamic factor structure".

Heuristically, we will ask what is the price that one pays by using an estimation model which is misspecified in the way we have described.

We will now study the properties of a maximum likelihood estimator in which the data follow a factor model that is dynamic and approximate (Assumptions A), while we restrict the approximating model to be exact, with non serially correlated idiosyncratic component and autoregressive common factors (R1 and R2). This is a Quasi Maximum Likelihood (QML) estimator in the sense of White (1982).

3 The asymptotic properties of the QML estimator of the common factors

Let us know introduce some further technical assumptions. First, to avoid degenerate solutions for the maximum likelihood problem, we will impose the following constraints in the maximization of the likelihood:

Constraints in the maximization of the likelihood

i) $\underline{c} \leq \hat{\psi}_{ii} \leq \bar{c}$ for all $i \in \mathbb{N}$.

ii) $|\hat{A}(z)| \neq 0, \forall |z| \leq 1$

[2] We write $\mathrm{E}_{\hat{\theta}}\left[\mathbf{F}|\mathbf{X}\right]$ to denote $\mathrm{E}_{\theta}\left[\mathbf{F}|\mathbf{X}\right]$ computed at $\theta = \hat{\theta}$.

[3] We could also take into account, serial correlation of the idiosyncratic components without compromising the parsimony of the model by modelling it as cross-sectionally orthogonal autoregressive process. We do not consider this case in order not to compromise expositional simplicity.

Let $\hat{\theta}$ the parameters estimated by maximum likelihood under the constraints (i) and (ii). We write $\hat{\mathbf{F}}_{\hat{\theta}}$ for the implied estimates of the common factors.

Constraints (i) and (ii) define a new parameter space $\Theta^c \subseteq \Theta$. This constraint is necessary to avoid situations in which estimated parameters imply non-stationarity of the common factors and/or trivial situation in which the variance of the idiosyncratic noise is either zero or infinite. Then, with Assumption C below we will insure that the constraint on the size of the idiosyncratic component is never binding.

Assumption C
There exists $\delta > 0$ such that $\underline{c} \leq \psi_{ii} - \delta \leq \psi_{ii} + \delta \leq \bar{c}$ for all $i \in \mathbb{N}$, where $\underline{c}$ and $\bar{c}$ are the constant in Assumption A (ii).

We are now ready to prove our main result.

Proposition 1 Under assumptions A, B and C we have:

$$\text{trace}\left(\frac{1}{T}(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H})\right) = O_p\left(\frac{1}{\Delta_{nT}}\right) \text{ as } n, T \to \infty$$

where $\hat{H} = \left(\hat{\mathbf{F}}_{\hat{\theta}}'\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}\hat{\mathbf{F}}_{\hat{\theta}}'\mathbf{F}$ is the coefficient of the OLS projection of $\mathbf{F}$ on $\hat{\mathbf{F}}_{\hat{\theta}}$ and $\Delta_{nT} = \min\left\{\sqrt{T}, \frac{n}{\log(n)}\right\}$ is the consistency rate.

Proof See the appendix.

The result above tells us that the common factors extracted using the Quasi Maximum Likelihood estimates of the parameters converge to the true common factors as the sample size T and the cross-sectional dimension n go to infinity. No restriction of the relative path of divergence of T and n is needed in order to achieve consistency. In this sense the estimates are viable also when the size of the cross-section n is much larger than the sample size T. Notice that since factors are identified only up to a rotation, we converge to a rotation of the common factors.

The result of Proposition 1 holds also in case of miss-specification of the number of factors and of the parameters governing the dynamics of the common factors.

Remark 1: The result of Proposition 1 holds if the likelihood if maximized under any additional restrictions on $A(L)$ and Q.

Remark 2: The result of Proposition 1 still holds if the approximating model has more than r common factors.

The proof of Remark 1 and Remark 2 is in the appendix.

4 Quasi Maximum Likelihood and Principal Components

Factor in large cross-sections have been traditionally estimated by principal components. The latter are closely connected with the QML estimator we propose here. Replace R1 and R2 by the stronger restrictions:

Approximating model: the exact factor model

R1* $\mathbf{f}_t \sim$ i.i.d. $\mathcal{N}(0, I_r)$

R2* $\mathbf{e}_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2 I_n)$.

In this case the log likelihood takes the form:

$$\mathcal{L}_{\mathbf{X}}(\mathbf{X};\theta) = -\frac{nT}{2}\log 2\pi - \frac{T}{2}\log|\Lambda\Lambda' + \sigma^2 I_n| - \frac{T}{2}\mathrm{Tr}\left(\Lambda\Lambda' + \sigma^2 I_n\right)^{-1} S$$

where $S = \frac{1}{T}\mathbf{X}'\mathbf{X}$ is the sample covariance matrix of the observation. Under the normalization that $\Lambda'\Lambda$ is a diagonal matrix with diagonal entries in decreasing order of magnitude, the maximum likelihood solution is[4]:

$$\hat{\Lambda} = \mathcal{V}(\mathcal{D} - \hat{\sigma}^2 I_r)^{1/2} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\mathrm{Trace}(S - \hat{\Lambda}\hat{\Lambda}')$$

where $\mathcal{D}$ is the $r \times r$ diagonal matrix containing the r largest eigenvalues of sample covariance matrix and $\mathcal{V}$ is the $n \times r$ matrix whose columns are the corresponding normalized eigenvalues ($\mathcal{V}'\mathcal{V} = I_r$), that is $S\mathcal{V} = \mathcal{V}\mathcal{D}$. The estimator for the common factors is given by

$$\hat{\mathbf{F}}_{\hat{\theta}} = \mathrm{E}_{\hat{\theta}}[\mathbf{F}|\mathbf{X}] = \mathbf{X}\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\sigma}^2 I_n\right)^{-1}\hat{\Lambda} = \mathbf{X}\hat{\Lambda}\left(\hat{\Lambda}'\hat{\Lambda} + \hat{\sigma}^2 I_n\right)^{-1} = \mathbf{X}\mathcal{V}(\mathcal{D} - \hat{\sigma}^2 I_r)^{1/2}\mathcal{D}^{-1}.$$

which are proportional to the sample principal components $\hat{\mathbf{Z}} = (\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_T)'$ which are defined as $\hat{\mathbf{Z}} = \mathbf{X}\mathcal{V}\mathcal{D}^{-1/2}$.

Result of Proposition 1 still holds in this case. Consistency of the principal components estimates is a particular case of Proposition 1 which provides an alternative proof of the result in Bai and Ng (2002) under a different set of assumptions. The proof of this result is in the appendix.[5]

[4] See, for instance, Lawley and Maxwell (1963), Chap. 4.

[5] Further, traditional factor analysis with non serially correlated data corresponds to the case $A(L) = I_r$, $Q = I_r$. Also under this restriction we have consistency of the common factors estimates.

5 Monte Carlo study

In this section we run a simulation study to asses the performances of our estimator.

The model from which we simulate is standard in the literature. A similar model has been used, for example, in Stock and Watson (2002a).

Let us define it below.

$$\mathbf{x}_t = \Lambda \mathbf{f}_t + \mathbf{E}_t$$

$$A(L)\mathbf{f}_t = \mathbf{u}_t, \text{ with } \mathbf{u}_t \text{ i.i.d. } \mathcal{N}(0, I_r);$$

$$D(L)\mathbf{E}_t = \mathbf{v}_t \text{ with } \mathbf{v}_t \text{ i.i.d. } \mathcal{N}(0, \mathcal{T})$$

$$A_{ij}(L) = \begin{cases} 1-\rho L & \text{if} \quad i=j \\ 0 & \text{if} \quad i \neq j \end{cases} ; i,j = 1,...,r$$

$$D_{ij}(L) = \begin{cases} \sqrt{\alpha_i}(1-dL) & \text{if} \quad i=j \\ 0 & \text{if} \quad i \neq j \end{cases} ; i,j = 1,...,n$$

$$\Lambda_{ij} \text{ i.i.d. } \mathcal{N}(0,1), i = 1,...,n; j = 1,..,r$$

$$\alpha_i = \frac{\beta_i}{1-\beta_i}\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{r}\Lambda_{ij}f_{jt}\right)^2 \text{ with } \beta_i \text{ i.i.d. } \mathcal{U}([u, 1-u])$$

$$\mathcal{T}_{ij} = \tau^{|i-j|}\frac{1}{1-d^2}, i,j = 1,...,n$$

Notice that we allow for cross-correlation between idiosyncratic elements. Since $\mathcal{T}$ is a Toeplitz matrix the cross-correlation among idiosyncratic elements is limited and it is easily seen that Assumption A (ii) is satisfied. The coefficient τ controls for the amount of cross-correlation. The exact factor model correspond to $\tau = 0$.

The coefficient β_i is the ratio between the variance of the idiosyncratic component, e_{it}, and the variance of the common component, $\sum_{j=1}^{r}\Lambda_{ij}f_{jt}$ (the inverse of the signal to noise ratio. In our simulation this ratio is uniformly distributed with an average of 50%. If $u = .5$ then the standardized observations have cross-sectionally homoscedastic idiosyncratic components.

Notice that if $\tau = 0, d = 0$, our approximating model is well specified and hence Maximum Likelihood provides the most efficient estimates. If $\tau = 0, d = 0, \rho = 0$, we have a static exact factor model and iteratively reweighed principal components provide the most efficient estimates. Finally, if $\tau = 0, d = 0, u = 1/2$, then we have a static factor models with spherical idiosyncratic components on standardized variables. In this case principal components on standardized variables provide the most efficient estimates.

We generate the model for different sizes of the cross-section: $n = 5, 10, 25, 50, 100$, and for sample size $T = 50, 100$.

Maximum likelihood estimates are computed using the EM algorithm as in Engle and Watson (1981) and Quah and Sargent (1992).

This algorithm has the advantage of requiring only one run of the Kalman smoother at each iteration. The computational complexity of the Kalman smoother depends mainly on the number of states which in our approximating model corresponds to the number of factors r and hence is independent of the size of the cross-section n.

To initialize the algorithm, we compute the first r sample principal components $\mathbf{f}_{pc,t}$ and estimate the parameters $\hat{\Lambda}^{(0)}\hat{A}^{(0)}(L), \hat{\Psi}_d^{(0)}$ by OLS, treating the principal components as if they were the true common factors. Since these estimates have been proved to be consistent for large cross-sections (Bai, 2003; Forni, Giannone, Lippi, and Reichlin, 2005a; Doz, Giannone, and Reichlin, 2005), the initialization is quite good provided that the cross-section dimension is large. We hence expect the number of iterations required for consistency to decrease as the cross-sectional dimension increases.

The two features highlighted above – small number of state variables and good initialization – make the algorithm feasible in a large cross-section.

To get the intuition of the EM algorithm, let us collect the initial values of the parameters in $\hat{\theta}^{(0)}$. We obtain a new value of the common factors by applying the Kalman smoother:

$$\hat{\mathbf{f}}_{\theta^{(0)},t} = \mathrm{E}_{\hat{\theta}^{(0)}}(\mathbf{f}_t|\mathbf{x}_1, ..., \mathbf{x}_T).$$

If we stop here we have the two-step estimates of the common factors proposed by Doz, Giannone, and Reichlin (2005); Giannone, Reichlin, and Sala (2004); Giannone, Reichlin, and Small (2005).

A new estimate of the parameters, to be collected in $\hat{\theta}^{(1)}$, can then be computed by OLS regression treating $\hat{\mathbf{f}}_{\theta^{(0)},t}$ as if they were the true common factors. If the OLS regressions are modified in order to take into account the fact that the common factors are estimated[6], then we have the EM algorithm which converges to the local maximum of the likelihood[7].

We control convergence by looking at $c_m = \frac{\mathcal{L}_{\mathbf{X}}(\mathbf{X};\hat{\theta}^{(m)}) - \mathcal{L}_{\mathbf{X}}(\mathbf{X};\hat{\theta}^{(m-1)})}{(\mathcal{L}_{\mathbf{X}}(\mathbf{X};\hat{\theta}^{(m)}) + \mathcal{L}_{\mathbf{X}}(\mathbf{X};\hat{\theta}^{(m-1)}))/2}$. We stop after M iterations if $c_M < 10^{-4}$.

We simulate the model 500 times and, at each repetition, we apply the algorithm to standardized data since the principal components used for initialization are not scale invariant.

We compute the following estimates of the common factors:

- principal components: $\hat{\mathbf{f}}_{pc,t} := \hat{\mathbf{z}}_t$;

- two-step estimates: $\hat{\mathbf{f}}_{2s,t} = \hat{\mathbf{f}}_{\hat{\theta}^{(0)},t}$

- maximum likelihood estimates: $\hat{\mathbf{f}}_{ml,t} := \hat{\mathbf{f}}_{\theta^{(M)},t}$.

[6] This requires the computation of $\mathrm{E}_{\theta^{(m)}}(\hat{\mathbf{f}}_{\theta^{(m)},t} - \mathbf{f}_t)(\hat{\mathbf{f}}_{\theta^{(m)},t-k} - \mathbf{f}_{t-k})', k = 0, ..., p$, which are also computed by the Kalman smoother. See for example Engle and Watson (1981).

[7] A detailed derivation of the EM algorithm for dynamic factor model is provided by Ghahramani and Hinton (1996).

We measure the performance of the different estimators by means of the following trace statistics:

$$\frac{\mathrm{Tr}\left(\mathbf{F}'\hat{\mathbf{F}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'\mathbf{F}\right)}{\mathrm{Tr}\left(\mathbf{F}'\mathbf{F}\right)}$$

where $\hat{\mathbf{F}} = \left(\hat{\mathbf{f}}_1, ..., \hat{\mathbf{f}}_T\right)'$, and $\hat{\mathbf{f}}_t$ is any of the three estimates of the common factors. This statistics is a multivariate version of the R^2 of the regression of the observed factors on the estimated factors. This is an appropriate measure since the common factors are identified only up to a rotation. This statistics is also closely related to the empirical canonical correlation between the true factors and their estimates. A number close to one indicates a good approximation of the true common factors. Denoting by TR_{pc}, TR_{2s} TR_{ml} the trace statistics for, respectively, principal component, two-step and maximum likelihood estimates of the common factors, we compute the relative trace statistics TR_{ml}/TR_{pc} and TR_{ml}/TR_{2s}. Numbers higher than one indicates that Maximum Likelihood estimates of the common factors are more accurate than principal components and two-step estimates.

Table 1: Simulation results for the model: $\rho = .9$, $d = .5$, $\tau = .5$, $u = .1$, $r = 1$

	TR_{ml}				
	$n = 5$	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.52	0.68	0.74	0.75	0.76
$T = 100$	0.64	0.78	0.84	0.85	0.86
	Number of iterations				
	$n = 5$	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	13	9	5	4	3
$T = 100$	13	7	4	4	3
	Computation time: seconds				
	$n = 5$	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.53	0.25	0.20	0.33	1.07
$T = 100$	0.66	0.37	0.33	0.61	2.13
	TR_{ml}/TR_{pc}				
	$n = 5$	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.11	1.04	1.00	1.00	1.00
$T = 100$	1.09	1.02	1.01	1.00	1.00
	TR_{ml}/TR_{2s}				
	$n = 5$	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.03	1.01	1.00	1.00	1.00
$T = 100$	1.02	1.00	1.00	1.00	1.00

Table 1 reports the results of the Montecarlo experiment for one common factor, $r = 1$, with serial correlation in both common factors, $\rho = .9$, and idiosyncratic components, $d = .5$. The model is approximat because of the weak cross-sectional correlation among idiosyncratic components, $\tau = .5$. Finally the idiosyncratic component is cross-sectionally heteroscedastic, $u = .1$. The numbers in the table refer to the average across experiments.

We would like to stress the following results:

1. The precision of the common factors estimated by Maximum Likelihood increases with the size of the cross-section n.

2. The number of iterations required for convergence is small and decreases with the size of the cross-section. As remarked above this is explained by the fact that, as n increases, the initialization provided by principal components are increasingly accurate and hence the computation time for convergence does not increase too much with the cross-sectional dimension.

3. The Maximum Likelihood estimates always dominate simple principal components and to a less extent the two-step procedure. As both n, T become large, the precision of the estimated common factors increases and all methods tend to perform similarly. This is not surprising, given that both methods provide consistent estimates for n and T large. Improvement of the ML estimates are significant for $n = 5$ and the improvement is of the order of 10% with respect to principal components and less than 5% for the two-step estimates. This suggests that the two-step Kalman smoother estimates already take appropriately into account the dynamics of the common factors and the cross-sectional heteroscedasticity of the idiosyncratic component. Hence the gains from further iterations are small.[8]

Table 2 reports the results for $r = 3$ while the remaining parameters are the same as those used the Table 1: $\rho = .9$, $d = .5$, $\tau = .5$, $u = .1$. The simulations have been run for $n \geq 10$ only, because an exact factor model with $n = 5$ and $r = 3$ is not identified. Notice that as expected, although the main features outlined above are still present, the estimates of the common factors are less precise with respect to the case of only one common factors (given the same a set of data, it is more difficult to extract additional factors). Improvements by the maximum likelihood are more sizable in this case. This indicates that efficiency improvements are larger, the harder is the factor extraction. We finally study a case in which the approximating model is well specified, that is the idiosyncratic components is neither serially nor cross-sectionally correlated ($d = 0, \tau = 0$). The remaining parameters are set as for the experiments reported in Table 1 and 2. In this case, as one can see from Table 3 below, the efficiency gains from QML estimates over principal components and two-step estimates are more relevant.

Summarizing, QML estimates of approximate factor models work well in finite sample. Because of the explicit modelling of the dynamics and the cross-sectional heteroscedasticity, the maximum likelihood estimates dominate the principal components and, to a less extent, the two two-step procedure. Efficiency improvements are relevant when the factor extraction is difficult, that is, when there are more common factors to estimate.

[8] See also Doz et al. (2005).

Table 2: Simulation results for the model: $\rho = .9$, $d = .5$, $\tau = .5$, $u = .1$, $r = 3$

	TR_{ml}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.48	0.59	0.65	0.67
$T = 100$	0.58	0.75	0.80	0.82

	Number of iterations			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	26	12	7	5
$T = 100$	20	9	5	4

	Computation time: seconds			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.72	0.46	0.56	1.44
$T = 100$	1.08	0.68	0.87	2.31

	TR_{ml}/TR_{pc}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.08	1.05	1.03	1.01
$T = 100$	1.10	1.06	1.02	1.01

	TR_{ml}/TR_{2s}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.05	1.02	1.01	1.00
$T = 100$	1.07	1.03	1.00	1.00

Table 3: Simulation results for the model: $\rho = .9$, $d = 0$, $\tau = 0$, $u = .1$, $r = 3$

	TR_{ml}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.54	0.65	0.68	0.70
$T = 100$	0.66	0.78	0.81	0.82

	Number of iterations			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	21	9	6	5
$T = 100$	15	7	5	4

	Computation time: seconds			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.58	0.36	0.49	1.30
$T = 100$	0.83	0.54	0.84	2.29

	TR_{ml}/TR_{pc}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.14	1.06	1.03	1.01
$T = 100$	1.19	1.06	1.02	1.01

	TR_{ml}/TR_{2s}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.07	1.02	1.01	1.00
$T = 100$	1.10	1.01	1.00	1.00

6 Summary and conclusions

The paper has studied quasi maximum likelihood (QML) estimation of the factors for an approximate factor model. Consistency under different sources of miss-specification is shown for n and T going to infinity.

As principal components the QML estimator is feasible for n large and can be easily implemented using the Kalman smoother and the EM algorithm as in traditional factor analysis. One desirable characteristic of this approach is that it can potentially produce efficiency improvements with respect to PC because it exploits factor dynamics and non sphericity of the idiosyncratic component.

Simulation results illustrate in what empirical conditions we can expect improvement with respect to simple principle components.

The importance of this result, beside the potential efficiency improvements, is that our parametric approach provides a natural framework for structural analysis since it allows for imposing restrictions on the loadings as done, for example, in Bernanke, Boivin, and Eliasz (2005); Boivin and Giannoni (2005); Kose, Otrok, and Whiteman (2003); Forni and Reichlin (2001) and extracting shocks. These features are not studied in this paper but they are natural extensions to explore in further work.

References

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71 (1):135–171, 2003.

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

Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor augmented regressions. *Econometrica*, 74(1):1133–1150, 2006.

Ben Bernanke, Jean Boivin, and Piotr Eliasz. Measuring monetary policy: A factor augmented autoregressive (favar) approach. *Quarterly Journal of Economics*, 120: 387–422, 2005.

Jean Boivin and Marc P. Giannoni. Dsge models in a data-rich environment. Manuscript, Columbia University, 2005.

Gari Chamberlain. Funds, factors, and diversification in arbitrage pricing models. *Econometrica*, 51:1281–1304, 1983.

Gari Chamberlain and Michael Rothschild. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, 51:1305–1324, 1983.

Gregory Connor and Robert A. Korajczyk. Performance measurement with arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15: 373–394, 1986.

Gregory Connor and Robert A. Korajczyk. Risk and return in an equilibrium apt: Application to a new test methodology. *Journal of Financial Economics*, 21:255–289, 1988.

Gregory Connor and Robert A. Korajczyk. A test for the number of factors in an approximate factor model. *Journal of Finance*, 48:1263–1291, 1993.

Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A two-step estimator for large approximate dynamic factor models based on kalman filtering. Manuscript, ECARES-Université Libre de Bruxelles, 2005.

Robert. F. Engle and Mark Watson. A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76:774–781, 1981.

Mario Forni and Marco Lippi. The generalized dynamic factor model: representation theory. *Econometric Theory*, 17:1113–1141, 2001.

Mario Forni and Lucrezia Reichlin. Federal policies and local economies: Europe and the us. *European Economic Review*, 45:109–134, 2001.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics*, 82: 540–554, 2000.

Mario Forni, Domenico Giannone, Marco Lippi, and Lucrezia Reichlin. Opening the black box: Structural factor models with large cross-sections. Manuscript, Université Libre de Bruxelles, 2005a.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic factor model: one-sided estimtion and forecasting. *Journal of the American Statistical Association*, 100:830–840, 2005b.

John F. Geweke. The dynamic factor analysis of economic time series models. In D. Aigner and A. Goldberger, editors, *Latent Variables in Socioeconomic Models*, pages 365–383. North-Holland, 1977.

John F. Geweke and Kenneth J. Singleton. Maximum likelihood "confirmatory" factor analysis of economic time series. *International Economic Review*, 22:37–54, 1980.

Zoubin Ghahramani and Geoffrey E. Hinton. Parameter estimation for linear dynamical systems. Technical report, Manuscript, University of Toronto, available at http://www.gatsby.ucl.ac.uk/ zoubin, 1996.

Domenico Giannone, Lucrezia Reichlin, and Luca Sala. Monetary policy in real time. In Mark Gertler and Kenneth Rogoff, editors, *NBER Macroeconomics Annual*, pages 161–200. MIT Press, 2004.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting gdp and inflation: the real-time informational content of macroeconomic data releases. Finance and Economics Discussion Series 2005-42, Board of Governors of the Federal Reserve System (U.S.), 2005.

M. Ayhan Kose, Christopher Otrok, and Charles H. Whiteman. International business cycles: World, region, and country-specific factors. *American Economic Review*, 93: 1216–1239, 2003.

D. N. Lawley and A. E. Maxwell. *Factor Analysis as a Statistical Method.* Butterworths, 1963.

Danny Quah and Thomas J. Sargent. A dynamic index model for large cross-section. In James Stock and Mark Watson, editors, *Business Cycle*, pages 161–200. Univeristy of Chicago Press, 1992.

Thomas J. Sargent and Christopher Sims. Business cycle modelling without pretending to have to much a-priori economic theory. In Christopher Sims, editor, *New Methods in Business Cycle Research*. Federal Reserve Bank of Minneapolis, 1977.

James H. Steiger. Factor indeterminacy in the 1930s and the 1970s some interesting parallels. *Psychometrika*, 40:157–167, 1979.

James H. Stock and Mark W. Watson. A probability model of the coincident economic indicators. In G. Moore and K. Lahiri, editors, *The Leading Economic Indicators: New Approaches and Forecasting Records*, pages 63–90. Cambridge University Press, 1991.

James. H. Stock and Mark. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97: 147–162, 2002a.

James. H. Stock and Mark. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economics Statistics*, 20:147–162, 2002b.

James. H. Stock and Mark. W. Watson. New indexes of coincident and leading economic indicators. In Olivier J. Blanchard and Stanley Fischer, editors, *NBER Macroeconomics Annual*, pages 351–393. MIT Press, 1989.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

7 Appendix

We adopt the following notations to define the pseudo likelihood under the approximating model which is completely characterized by the parameter θ:

- $f_{(\mathbf{X},\mathbf{F})}(X,F;\theta)$ is the joint density of the common factors and the observables, depending on the parameter θ,
- $f_{\mathbf{X}}(X;\theta)$ and $f_{\mathbf{F}}(F;\theta)$ are the corresponding marginal densities,
- $f_{\mathbf{X}|\mathbf{F}=F}(X;\theta)$ and $f_{\mathbf{F}|\mathbf{X}=X}(F;\theta)$ are the corresponding conditional densities

where $F \in \mathbb{R}^{(T\times r)}$ and $X \in \mathbb{R}^{(T\times n)}$. We know that, for any (X,F):

$$\begin{aligned} f_{(\mathbf{X},\mathbf{F})}(X,F;\theta) &= f_{\mathbf{X}|\mathbf{F}=F}(X;\theta) f_{\mathbf{F}}(F;\theta) \\ &= f_{\mathbf{F}|\mathbf{X}=X}(F;\theta) f_{\mathbf{X}}(X;\theta) \end{aligned}$$

so that:

$$f_{\mathbf{X}}(X;\theta) = \frac{f_{\mathbf{X}|\mathbf{F}=F}(X;\theta) f_{\mathbf{F}}(F;\theta)}{f_{\mathbf{F}|\mathbf{X}=X}(F;\theta)}.$$

The log-likelihood of the data $\mathcal{L}_{\mathbf{X}}(X;\theta) = \log f_{\mathbf{X}}(X;\theta)$ can then be decomposed in the following way:

$$\mathcal{L}_{\mathbf{X}}(X;\theta) = \mathcal{L}_{\mathbf{X}|\mathbf{F}}(X|F;\theta) + \mathcal{L}_{\mathbf{F}}(F;\theta) - \mathcal{L}_{\mathbf{F}|\mathbf{X}}(F|X;\theta)$$

where $\mathcal{L}_{\mathbf{X}|\mathbf{F}}(X|F;\theta) = \log f_{\mathbf{X}|\mathbf{F}=F}(X;\theta)$, $\mathcal{L}_{\mathbf{F}}(F;\theta) = \log f_{\mathbf{F}}(F;\theta)$ and $\mathcal{L}_{\mathbf{F}|\mathbf{X}}(F|X;\theta) = \log f_{\mathbf{F}|\mathbf{X}=X}(F;\theta)$.

Under the normality assumption, and denoting by $\mathbf{X}$ the actual observed values of the underlying process, we can write, for any value of F:

$$\mathcal{L}_{\mathbf{X}|\mathbf{F}}(\mathbf{X}|F;\theta) = -\tfrac{nT}{2}\log(2\pi) - \tfrac{T}{2}\log|\Psi_d| - \tfrac{1}{2}\mathrm{Tr}(\mathbf{X} - F\Lambda')\Psi_d^{-1}(\mathbf{X} - F\Lambda')'$$

$$\mathcal{L}_{\mathbf{F}}(F;\theta) = -\tfrac{rT}{2}\log(2\pi) - \tfrac{1}{2}\log|\mathbf{\Phi}_\theta| - \tfrac{1}{2}(\mathrm{vec}F')'\mathbf{\Phi}_\theta^{-1}(\mathrm{vec}F')$$

$$\mathcal{L}_{\mathbf{F}|\mathbf{X}}(F|\mathbf{X};\theta) = -\tfrac{rT}{2}\log(2\pi) - \tfrac{1}{2}\log|\mathbf{\Omega}_\theta| - \tfrac{1}{2}(\mathrm{vec}(F - \hat{\mathbf{F}}_\theta)')'\mathbf{\Omega}_\theta^{-1}(\mathrm{vec}(F - \hat{\mathbf{F}}_\theta)')$$

with
$\mathbf{\Phi}_\theta = \mathrm{E}_\theta\left[(\mathrm{vec}\mathbf{F}')(\mathrm{vec}\mathbf{F}')'\right]$,
$\hat{\mathbf{F}}_\theta = \mathrm{E}_\theta\left[\mathbf{F}|\mathbf{X}\right] = (\hat{\mathbf{f}}_{\theta,1}, ..., \hat{\mathbf{f}}_{\theta,T})'$
and
$\mathbf{\Omega}_\theta = \mathrm{E}_\theta\left[(\mathrm{vec}(\mathbf{F} - \hat{\mathbf{F}}_\theta)')(\mathrm{vec}(\mathbf{F} - \hat{\mathbf{F}}_\theta)')'\right]$.

We hence have, for any value of F:

$$\mathcal{L}_{\mathbf{X}}(\mathbf{X};\theta) = -\tfrac{nT}{2}\log(2\pi) - \tfrac{T}{2}\log|\Psi_d| - \tfrac{1}{2}\mathrm{Tr}(\mathbf{X} - F\Lambda')\Psi_d^{-1}(\mathbf{X} - F\Lambda')'$$
$$-\tfrac{1}{2}(\mathrm{vec}F')'\mathbf{\Phi}_\theta^{-1}(\mathrm{vec}F') - \tfrac{1}{2}\log|\mathbf{\Omega}_\theta| + \tfrac{1}{2}(\mathrm{vec}(F - \hat{\mathbf{F}}_\theta)')'\mathbf{\Omega}_\theta^{-1}(\mathrm{vec}(F - \hat{\mathbf{F}}_\theta)') \tag{7.2}$$

If we consider the likelihood computed by using $F = \hat{\mathbf{F}}_\theta$, (7.2) the above expression becomes:

$$\begin{aligned} \mathcal{L}(\mathbf{X};\theta) = \quad & -\tfrac{nT}{2}\log(2\pi) - \tfrac{T}{2}\log|\Psi_d| - \tfrac{1}{2}\mathrm{vec}(\hat{\mathbf{F}}_\theta')'\mathbf{\Phi}_\theta^{-1}\mathrm{vec}(\hat{\mathbf{F}}_\theta') - \tfrac{1}{2}\log|\mathbf{\Omega}_\theta| \\ & -\tfrac{1}{2}\mathrm{Tr}(\mathbf{F}\Lambda_0' - \hat{\mathbf{F}}_\theta\Lambda' + \mathbf{E})\Psi_d^{-1}(\mathbf{F}\Lambda_0' - \hat{\mathbf{F}}_\theta\Lambda' + \mathbf{E})' \end{aligned} \tag{7.3}$$

Let us now evaluate the likelihood at the following set of parameters:

$$\theta_0^c := \{A(L) = I_r; Q = I_r; \Lambda = \Lambda_0; \Psi = \Psi_{0,d}\}$$

where $\Psi_{0,d}$ is the diagonal matrix obtained by setting equal to zero all the out of diagonal elements of Ψ_0.

For $\theta = \theta_0^c$, we have $\mathbf{\Phi}_{\theta_0^c} = I_{rT}$ and $\mathbf{\Omega}_{\theta_0^c} = I_T \otimes \left(I_r - \Lambda_0'\left(\Lambda_0\Lambda_0' + \Psi_{0,d}\right)^{-1}\Lambda_0\right)$.

It can be easily checked that

$$\left(\Lambda_0\Lambda_0' + \Psi_{0,d}\right)^{-1} = \Psi_{0,d}^{-1} - \Psi_{0,d}^{-1}\Lambda_0\left(I_r + \Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right)^{-1}\Lambda_0'\Psi_{0,d}^{-1} \tag{7.4}$$

so that: $\mathbf{\Omega}_{\theta_0^c} = I_T \otimes \left(I_r + \Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right)^{-1}$.

We then have:

$$\begin{aligned} \mathcal{L}(\mathbf{X};\theta_0^c) = \quad & -\tfrac{nT}{2}\log(2\pi) - \tfrac{T}{2}\log|\Psi_{0,d}| - \tfrac{1}{2}\mathrm{Tr}\hat{\mathbf{F}}_{\theta_0^c}'\hat{\mathbf{F}}_{\theta_0^c} - \tfrac{T}{2}\log\left|I_r + \Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right| \\ & -\tfrac{1}{2}\mathrm{Tr}\left((\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c})\Lambda_0' + \mathbf{E}\right)\Psi_{0,d}^{-1}\left((\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c})\Lambda_0' + \mathbf{E}\right)'. \end{aligned} \tag{7.5}$$

As n and T go to infinity (7.5) simplifies drastically since some of the terms are asymptotically negligible. This is shown as a corollary of the following Lemma.

Lemma 1 Under assumptions A, B, we have

1. $\left\|\frac{\mathbf{E}'\mathbf{E}}{nT}\right\| = O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{\sqrt{T}}\right)$ as $n, T \to \infty$
2. $\frac{1}{T}\mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c})'(\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c}) = O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{\sqrt{T}}\right)$ as $n, T \to \infty$
3. $\frac{1}{nT}\mathrm{Tr}(\mathbf{E}'\Psi_{0,d}^{-1}\mathbf{E}) = 1 + O_p\left(\frac{1}{\sqrt{T}}\right)$ as $n, T \to \infty$
4. $\frac{1}{nT}\mathrm{Tr}\hat{\mathbf{F}}_{\theta_0^c}'\hat{\mathbf{F}}_{\theta_0^c} = O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{\sqrt{T}}\right)$ as $n, T \to \infty$
5. $\frac{1}{n}\log\left|I_r + \Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right| = \left(\frac{\log(n)}{n}\right)$ as $n \to \infty$

Proof We have:

$$\left\|\frac{\mathbf{E}'\mathbf{E}}{nT}\right\| \leq \frac{1}{n}\|\Psi_0\| + \frac{1}{n}\left\|\frac{\mathbf{E}'\mathbf{E}}{T} - \Psi_0\right\|$$

$$\left\|\frac{1}{n}\left(\frac{\mathbf{E}'\mathbf{E}}{T} - \Psi_0\right)\right\|^2 \leq \frac{1}{n^2}\mathrm{trace}\left[\left(\frac{\mathbf{E}'\mathbf{E}}{T} - \Psi_0\right)'\left(\frac{\mathbf{E}'\mathbf{E}}{T} - \Psi_0\right)\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T} e_{it}e_{jt} - \psi_{0,ij}\right)^2$$

Taking expectations, from assumption B we obtain:

$$\frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T} e_{it}e_{jt} - \psi_{0,ij}\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} e_{it}e_{jt} - \psi_{0,ij}\right)^2\right] \leq \frac{M}{T}$$

Result 1 hence follows from the Markov inequality.

Let us turn now to result 2. First, we have: $\mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c})'(\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c}) \leq r\|\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c}\|^2$. Then, using (7.4), we have:

$$\hat{\mathbf{F}_{\theta_0^c}} = \mathbf{X}\Psi_{0,d}^{-1}\Lambda_0(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1} = \mathbf{F}\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1} + \mathbf{E}\Psi_{0,d}^{-1}\Lambda_0(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1}$$

so that:

$$\frac{1}{\sqrt{T}}\|\mathbf{F} - \hat{\mathbf{F}}_{\theta_0^c}\| \leq \left\|\frac{1}{\sqrt{T}}\mathbf{F}\right\|\|\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1} - I_r\| + \left\|\frac{1}{\sqrt{nT}}\mathbf{E}\right\|\|\sqrt{n}\Lambda_0\Psi_{0,d}^{-1}(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1}\|$$

Assumptions A implies:

$$\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1} - I_r = (\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1} = O\left(\frac{1}{n}\right) \text{ as } n \to \infty$$

Further, we have: $\|\Lambda_0\Psi_{0,d}^{-1}(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1}\| \leq \|\Lambda_0'\Psi_{0,d}^{-1/2}\|\|\Psi_{0,d}^{-1/2}\|\|(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0 + I_r)^{-1}\|$.

As assumptions A also imply:

$$\|\Psi_{0,d}^{-1/2}\| \leq \frac{1}{\sqrt{\lambda_{min}\left(\Psi_0\right)}} = O\left(1\right) \text{ as } n \to \infty$$

and

$$\|\Lambda_0'\Psi_{0,d}^{-1/2}\| = \|\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\|^{1/2} \leq \frac{1}{\lambda_{min}\left(\Psi_0\right)}\|\Lambda_0'\Lambda_0\|^{1/2} = O\left(\sqrt{n}\right) \text{ as } n \to \infty$$

Result 2 then follows from the previous result of this lemma and the fact that by assumption B we have $\left\|\frac{1}{\sqrt{T}}\mathbf{F}\right\| = O_p(1)$.

Result 3 is a direct consequence of Assumption B (i) and the Markov inequality. In fact:

$$\frac{1}{nT}\text{Tr}\left(\mathbf{E}\Psi_{0,d}^{-1}\mathbf{E}'\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\frac{1}{T}\sum_{t=1}^{T} e_{it}^2}{\psi_{0,ii}}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{\psi_{0,ii}}{\hat{\psi}_{0,ii}} + O_p\left(\frac{1}{\sqrt{T}}\right)$$

To obtain result 4, notice that:

$$\frac{1}{nT}\text{Tr}\hat{\mathbf{F}}_{\theta_0^c}'\hat{\mathbf{F}}_{\theta_0^c} \leq \frac{r}{nT}\left\|\hat{\mathbf{F}}_{\theta_0^c}\right\|^2 = \frac{r}{nT}\left\|\mathbf{F}+\hat{\mathbf{F}}_{\theta_0^c}-\mathbf{F}\right\|^2 \leq \frac{2r}{n}\left(\|\frac{1}{\sqrt{T}}\mathbf{F}\|^2 + \frac{1}{T}\|\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}\|^2\right)$$

As $\|\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}\|^2 \leq \text{Tr}\left(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}\right)'\left(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}\right)$, the desired rate follows from Assumption B (iii) and result 2.

Concerning result 5, notice that, by assumptions A:

$\log\left|I_r + \Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right| = \log(n) + \log\left|\frac{I_r}{n} + \frac{\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0}{n}\right|$, with:

$\log\left|\frac{I_r}{n} + \frac{\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0}{n}\right| \simeq \log\left|\frac{\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0}{n}\right| \leq r\log\frac{\lambda_{max}\left(\frac{\Lambda_0'\Lambda_0}{n}\right)}{\lambda_{min}(\Psi_0)} = O(1)$ as $n \to \infty$. Q.E.D.

Corollary Under the same assumptions of Lemma 1, we have:

$$\frac{1}{nT}\mathcal{L}(\mathbf{X};\theta_0^c) = -\frac{1}{2n}\log(2\pi) - \frac{1}{2}\log|\Psi_{0,d}| - \frac{1}{2} + O_p\left(\frac{\log(n)}{n}\right) + O_p\left(\frac{1}{\sqrt{T}}\right), \text{ as } n,T \to \infty$$

Proof

The only term for which the asymptotic behavior is not a direct consequence of Lemma 1 is the following:

$\frac{1}{nT}\text{Tr}\left((\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c})\Lambda_0' + \mathbf{E}\right)\Psi_{0,d}^{-1}\left((\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c})\Lambda_0' + \mathbf{E}\right)'$

$$= \frac{1}{nT}\text{Tr}\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c})'(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}) - 2\frac{1}{nT}\text{Tr}\Lambda_0'\Psi_{0,d}^{-1}\mathbf{E}'(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}) + \frac{1}{nT}\text{Tr}\Psi_{0,d}^{-1}\mathbf{E}'\mathbf{E}$$

Let us analyze the three terms in the summation separately.

The asymptotic behavior of the third term in the summation is a direct consequence on Lemma 1 (3).

The asymptotic behavior of the first term follows from Assumption A and Lemma 1 (2):

$$\frac{1}{nT}\text{Tr}\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c})'(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}) \leq \frac{1}{nT}\lambda_{max}\left(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right)\text{Tr}(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c})'(\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c})$$

We know (see the proof of lemma 1) that $\frac{1}{n}\lambda_{max}\left(\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\right) = \frac{1}{n}\|\Lambda_0'\Psi_{0,d}^{-1}\Lambda_0\| = O(1)$ so that the result directly follows from lemma 1 (2).

For the second term:

$$\begin{aligned} \frac{1}{nT}\text{Tr}\Lambda_0'\Psi_{0,d}^{-1}\mathbf{E}'(\mathbf{F}-\hat{\mathbf{F}}_{\theta}) &\leq r\left\|\frac{\mathbf{E}'\mathbf{E}}{nT}\right\|^{1/2}\left\|\frac{\Lambda_0'\Lambda_0}{n}\right\|\frac{1}{\left(\lambda_{min}\Psi_{0,d}\right)^2}\frac{1}{\sqrt{T}}\|\mathbf{F}-\hat{\mathbf{F}}_{\theta_0^c}\| \\ &= O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{\sqrt{T}}\right) \end{aligned}$$

where the last equality follows for Lemma 1 (1-2) and Assumptions A and B.

The asymptotic simplification of the likelihood, in the Corollary above, is due to the fact that under the simple approximating model the expected common factor converge to the true ones (Lemma 1 (i)). The expected values of the common factors, $\hat{\mathbf{F}}_{\theta_0^c}$, are essentially the coefficients of an OLS regression of the observation, $\mathbf{X}$, on the factor loadings, Λ_0. If data are gaussian and the restrictions in θ_0^c are satisfied, then such estimates of the common factors are the most efficient. However, the estimates are still consistent under the weaker assumptions A (i) and A (ii). This result also tells us that a large cross-section solves the common factors indeterminacy we have with a finite cross-section dimension.

Consider now the likelihood evaluated at its maximum where $\hat{\theta} := \left\{\hat{A}(L); \hat{H}; \hat{\Lambda}; \hat{\Psi}_d\right\}$ are the Maximum Likelihood estimates of the parameters, with $\hat{\theta} \in \Theta^c$. We will denote by $\hat{\mathbf{F}}_{\hat{\theta}}$ the corresponding estimates of the common factors.

The likelihood at its maximum takes the form, see equation (7.2):

$$\begin{aligned} \mathcal{L}(\mathbf{X};\hat{\theta}) = & -\frac{nT}{2}\log(2\pi) - \frac{T}{2}\log|\hat{\Psi}_d| - \frac{1}{2}\text{Tr}(\mathbf{X}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')\hat{\Psi}_d^{-1}(\mathbf{X}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')' \\ & -\frac{1}{2}\text{vec}(\hat{\mathbf{F}}_{\hat{\theta}}')'\mathbf{\Phi}_{\hat{\theta}}^{-1}\text{vec}(\hat{\mathbf{F}}_{\hat{\theta}}') - \frac{1}{2}\log|\mathbf{\Omega}_{\hat{\theta}}| \end{aligned}$$

Assumption C insures that the constraints on the size of the idiosyncratic variance that is imposed in the maximization is not binding, that is $\theta_0^c \in \Theta^c$. Consequently,

$\mathcal{L}(\mathbf{X};\hat{\theta}) \geq \mathcal{L}(\mathbf{X};\theta_0)$. Using the Corollary, this implies:

$$\begin{aligned} 0 \geq \tfrac{2}{nT}\left(\mathcal{L}(\mathbf{X};\theta_0^c) - \mathcal{L}(\mathbf{X};\hat{\theta})\right) &= \tfrac{1}{n}\log|\hat{\Psi}_d| + \tfrac{1}{nT}\mathrm{Tr}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')\hat{\Psi}_d^{-1}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')' \\ &+ \tfrac{1}{nT}\mathrm{vec}(\hat{\mathbf{F}}'_{\hat{\theta}})'\mathbf{\Phi}_{\hat{\theta}}^{-1}\mathrm{vec}(\hat{\mathbf{F}}'_{\hat{\theta}}) + \tfrac{1}{nT}\log|\mathbf{\Omega}_{\hat{\theta}}| \\ &- \tfrac{1}{n}\log|\Psi_{0,d}| - 1 + O_p\left(\tfrac{1}{\sqrt{T}}\right) + O_p\left(\tfrac{\log(n)}{n}\right) \end{aligned}$$

Lemma 2 Under assumptions A, B, and C, we have:

$$\begin{aligned} \tfrac{1}{nT}\mathrm{Tr}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')\hat{\Psi}_d^{-1}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')' \geq \; & \tfrac{1}{nT}\mathrm{Tr}(\Lambda_0'\hat{\Psi}_d^{-1}\Lambda_0)'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H}) \\ & -2\sqrt{\tfrac{1}{T}\mathrm{Tr}((\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H}))}\sqrt{O_p\left(\tfrac{1}{\sqrt{T}}\right) + O_p\left(\tfrac{1}{n}\right)} \\ & + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\tfrac{\psi_{0,ii}}{\hat{\psi}_{ii}} + O_p\left(\tfrac{1}{\sqrt{T}}\right) + O_p\left(\tfrac{1}{n}\right) \end{aligned}$$

where $\hat{H} = \left(\hat{\mathbf{F}}'_{\hat{\theta}}\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}\hat{\mathbf{F}}'_{\hat{\theta}}\mathbf{F}$ is the coefficient of the OLS projection of $\mathbf{F}$ on $\hat{\mathbf{F}}_{\hat{\theta}}$

Proof Consider the coefficients of the OLS projection of $\mathbf{X}$ on $\hat{\mathbf{F}}_{\hat{\theta}}$:

$$\hat{\hat{\Lambda}} = \mathbf{X}'\hat{\mathbf{F}}_{\hat{\theta}}\left(\hat{\mathbf{F}}'_{\hat{\theta}}\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}$$

Least squares properties imply that:

$$\tfrac{1}{nT}\mathrm{Tr}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')\hat{\Psi}_d^{-1}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\Lambda}')' \geq \tfrac{1}{nT}\mathrm{Tr}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\hat{\Lambda}}')\hat{\Psi}_d^{-1}(\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\hat{\Lambda}}')'$$

Notice that:

$$\begin{aligned} (\mathbf{X} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{\hat{\Lambda}}') &= \left(\mathbf{F}\Lambda_0' + \mathbf{E} - \hat{\mathbf{F}}_{\hat{\theta}}\left(\hat{\mathbf{F}}'_{\hat{\theta}}\hat{\mathbf{F}}_{\hat{\theta}}\right)^{\mathbf{-1}}\hat{\mathbf{F}}'_{\hat{\theta}}\mathbf{F}\mathbf{\Lambda_0'} - \hat{\mathbf{F}}_{\hat{\theta}}\left(\hat{\mathbf{F}}'_{\hat{\theta}}\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}\hat{\mathbf{F}}'_{\hat{\theta}}\mathbf{E}\right) \\ &= (\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}}\hat{H})\Lambda_0' + (I_T - P_{\hat{F}_{\hat{\theta}}})\mathbf{E} \end{aligned}$$

where $\hat{H} = \left(\hat{\mathbf{F}}'_{\hat{\theta}}\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}\hat{\mathbf{F}}'_{\hat{\theta}}\mathbf{F}$ is the coefficient of the OLS projection of $\mathbf{F}$ on $\hat{\mathbf{F}}_{\hat{\theta}}$ and $P_{\hat{F}_{\hat{\theta}}} = \hat{\mathbf{F}}_{\hat{\theta}}\left(\hat{\mathbf{F}}'_{\hat{\theta}}\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}\hat{\mathbf{F}}'_{\hat{\theta}}$ is the projection matrix associated with $\hat{\mathbf{F}}_{\hat{\theta}}$.

Consequently:

$$
\begin{aligned}
\tfrac{1}{nT}\mathrm{Tr}(\mathbf{X}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{\hat{\Lambda}}')\hat{\Psi}_d^{-1}(\mathbf{X}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{\hat{\Lambda}}')' &= \tfrac{1}{nT}\mathrm{Tr}(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})\Lambda_0'\hat{\Psi}_d^{-1}\Lambda_0(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})' \\
&+\tfrac{1}{nT}\mathrm{Tr}(I_T-P_{\hat{F}_{\hat{\theta}}})\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'(I_T-P_{\hat{F}_{\hat{\theta}}}) \\
&+2\tfrac{1}{nT}\mathrm{Tr}(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})\Lambda_0'\hat{\Psi}_d^{-1}\mathbf{E}'(I_T-P_{\hat{F}_{\hat{\theta}}})
\end{aligned}
$$

We have:

$$
\begin{aligned}
\tfrac{1}{nT}\mathrm{Tr}(I_T-P_{\hat{F}_{\hat{\theta}}})\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'(I_T-P_{\hat{F}_{\hat{\theta}}}) &= \tfrac{1}{nT}\mathrm{Tr}\left(\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'(I_T-P_{\hat{F}_{\hat{\theta}}})\right) \\
&= \tfrac{1}{nT}\mathrm{Tr}\left(\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'\right)-\tfrac{1}{nT}\mathrm{Tr}\left(\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'P_{\hat{F}_{\hat{\theta}}}\right)
\end{aligned}
$$

By assumption B (ii):

$$
\frac{1}{nT}\mathrm{Tr}\left(\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\frac{1}{T}\sum_{t=1}^{T}e_{it}^2}{\hat{\psi}_{ii}}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_{0,ii}}{\hat{\psi}_{ii}}+O_p\left(\frac{1}{\sqrt{T}}\right)
$$

Furthermore:

$$
\frac{1}{nT}\mathrm{Tr}\left(\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'P_{\hat{F}_{\hat{\theta}}}\right)=\frac{1}{nT}\mathrm{Tr}\left(\hat{\mathbf{F}}_{\hat{\theta}}'\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'\hat{\mathbf{F}}_{\hat{\theta}}\left(\hat{\mathbf{F}}_{\hat{\theta}}'\hat{\mathbf{F}}_{\hat{\theta}}\right)^{-1}\right)\leq r\frac{1}{nT}\lambda_{max}\left(\mathbf{E}\hat{\Psi}_d^{-1}\mathbf{E}'\right)=O_p\left(\frac{1}{\sqrt{T}}\right)+O_p\left(\frac{1}{n}\right)
$$

Finally,

$$
\begin{aligned}
\tfrac{1}{nT}\left|\mathrm{Tr}(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})\Lambda_0'\hat{\Psi}_d^{-1}\mathbf{E}'(I_T-P_{\hat{F}_{\hat{\theta}}})\right| &\leq \sqrt{\tfrac{1}{T}\mathrm{Tr}(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})'(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})}\sqrt{\tfrac{1}{n^2T}\mathrm{Tr}(\Lambda_0'\hat{\Psi}_d^{-1}\mathbf{E}'\mathbf{E}\hat{\Psi}_d^{-1}\Lambda_0)} \\
&= \sqrt{\tfrac{1}{T}\mathrm{Tr}(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})'(\mathbf{F}-\hat{\mathbf{F}}_{\hat{\theta}}\hat{H})}\sqrt{O_p\left(\tfrac{1}{\sqrt{T}}\right)+O_p\left(\tfrac{1}{n}\right)}
\end{aligned}
$$

The desired result follows. Q.E.D.

To prepare the proof of Proposition 1, notice first that $\mathrm{vec}(\hat{\mathbf{F}}_{\hat{\theta}}')'\mathbf{\Phi}_{\hat{\theta}}^{-1}\mathrm{vec}(\hat{\mathbf{F}}_{\hat{\theta}}')\geq 0$.

Moreover, it can be shown that: $\log|\mathbf{\Omega}_{\hat{\theta}}|>0$. Indeed, if we denote $\mathbf{\Sigma}_\theta=\mathrm{E}_\theta\left[(\mathrm{vec}\mathbf{X}')(\mathrm{vec}\mathbf{X}')'\right]$, we have:

$$
\mathbf{\Sigma}_\theta=(I_T\otimes\Lambda)\,\mathbf{\Phi}_\theta\,(I_T\otimes\Lambda)'+(I_T\otimes\Psi_d)\,.
$$

Hence, it can the be checked that

$$
\Sigma_\theta^{-1}=\left(I_T\otimes\Psi_d^{-1}\right)-\left(I_T\otimes\Psi_d^{-1}\Lambda\right)\left(\Phi_\theta^{-1}+I_T\otimes\Lambda'\Psi_d^{-1}\Lambda\right)^{-1}\left(I_T\otimes\Lambda'\Psi_d^{-1}\right)
$$

and $\Omega_\theta = I_{rT} + \left(I_T \otimes \Lambda' \Psi_d^{-1} \Lambda \right) \Phi_\theta$.

It then follows that $\Omega_\theta > I_{rT}$, so that $\log |\boldsymbol{\Omega}_{\hat{\theta}}| > 0$. This property holds for all $A(L)$ and Q satisfying R1.

Finally:

$$\frac{1}{n} \log |\hat{\Psi}_d| + \frac{1}{n} \sum_{i=1}^{n} \frac{\psi_{0,ii}}{\hat{\psi}_{ii}} - \frac{1}{n} \log |\hat{\Psi}_{0d}| - 1 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\psi_{0i}}{\hat{\psi}_i} - \log \left(\frac{\psi_{0i}}{\hat{\psi}_i} \right) - 1 \right) \geq 0$$

Using the fact that $\frac{n}{\log(n)} = O(n)$, we then obtain:

$$\begin{aligned}
0 &\geq \tfrac{2}{nT} \left(\mathcal{L}(\mathbf{X}; \theta_0^c) - \mathcal{L}(\mathbf{X}; \hat{\theta}) \right) \\
&\geq \tfrac{1}{nT} \mathrm{Tr}(\Lambda_0' \hat{\Psi}_d^{-1} \Lambda_0)(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H}) \\
&\quad -2 \sqrt{\tfrac{1}{T} \mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})} O_p \left(\sqrt{\tfrac{1}{\Delta_{nT}}} \right) + O_p \left(\tfrac{1}{\Delta_{nT}} \right)
\end{aligned}$$

where

$$\Delta_{nT} = \min \left\{ \sqrt{T}, \frac{n}{\log(n)} \right\}$$

We can now prove our main result.

Proof of Proposition 1

$$\begin{aligned}
0 &\geq \tfrac{1}{nT} \mathrm{Tr}(\Lambda_0' \hat{\Psi}_d^{-1} \Lambda_0)(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H}) \\
&\quad -2 \sqrt{\tfrac{1}{T} \mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})} O_p \left(\sqrt{\tfrac{1}{\Delta_{nT}}} \right) + O_p \left(\tfrac{1}{\Delta_{nT}} \right) \\
&\geq \lambda_{min} \left(\tfrac{\Lambda_0' \hat{\Psi}_d^{-1} \Lambda_0}{n} \right) \tfrac{1}{T} \mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H}) \\
&\quad -2 O_p \left(\sqrt{\tfrac{1}{\Delta_{nT}}} \right) \sqrt{\tfrac{1}{T} \mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})} + O_p \left(\tfrac{1}{\Delta_{nT}} \right) \\
&= \lambda_{min} \left(\tfrac{\Lambda_0' \hat{\Psi}_d^{-1} \Lambda_0}{n} \right) V_{nT} - 2 \sqrt{V_{nT}} O_p \left(\sqrt{\tfrac{1}{\Delta_{nT}}} \right) + O_p \left(\tfrac{1}{\Delta_{nT}} \right)
\end{aligned}$$

where $V_{nT} = \frac{1}{T} \mathrm{Tr}(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})'(\mathbf{F} - \hat{\mathbf{F}}_{\hat{\theta}} \hat{H})$.

Since $\liminf_{n,T \to \infty} \lambda_{min} \left(\frac{\Lambda_0' \hat{\Psi}_d^{-1} \Lambda_0}{n} \right) > 0$, we have:

$$V_{nT} - \sqrt{V_{nT}} O_p \left(\sqrt{\frac{1}{\Delta_{nT}}} \right) + O_p \left(\frac{1}{\Delta_{nT}} \right) \leq 0 \qquad (7.6)$$

which implies that: $V_{nT} = O_p\left(\frac{1}{\Delta_{nT}}\right)$

In order to proves the latter, it is actually sufficient to notice that for any T and n we have a second order polynomial: $y^2 + by + c$ with $y := \sqrt{V_{nT}}$, $b = O_p\left(\sqrt{\frac{1}{\Delta_{nT}}}\right)$, $c = O_p\left(\frac{1}{\Delta_{nT}}\right)$ which is supposed to take a negative value in y. This is possible only if the following conditions are satisfied:

a) the discriminant is positive, i.e. $c < \frac{1}{4}b^2$ (which is possible since $b^2 = O_p\left(\frac{1}{\Delta_{nT}}\right)$)

b) y is between the two roots of the polynomial, i.e.

$$\frac{1}{2}\left(b - \sqrt{b^2 - 4c}\right) \leq y \leq \frac{1}{2}\left(b + \sqrt{b^2 + 4c}\right)$$

The conditions a) and b) imply that $y = O_p\left(\sqrt{\frac{1}{\Delta_{nT}}}\right)$ and hence $V_{nT} := y^2 = O_p\left(\frac{1}{\Delta_{nT}}\right)$.

Q.E.D.

Proof of Remark 1

The fact that Proposition 1 holds for any $A(L)$ and Q is easily proved by noticing that:

a) $A(L), Q$ only enter in $\text{vec}(\hat{\mathbf{F}}'_{\hat{\theta}})'\mathbf{\Phi}_{\hat{\theta}}^{-1}\text{vec}(\hat{\mathbf{F}}'_{\hat{\theta}})$ and $\log\left(I_{rT} + \mathbf{\Phi}_{\hat{\theta}}\mathbf{\Gamma}_{\hat{\theta}}\right)$ and the proof only requires these quantities to be positive.

b) imposing restrictions on $A(L)$ and Q in the approximating model, we define a parameter space $\tilde{\Theta}^c \subseteq \Theta^c$ for which we still have $\theta_0^c \in \Theta_c$ and hence $\mathcal{L}(\mathbf{X}; \hat{\theta}) \geq \mathcal{L}(\mathbf{X}; \theta_0)$.

Proof of Remark 2

If the maximization is run for a number of common factors $\tilde{r} > r$ the new model will encompass the previous one and hence $\mathcal{L}(\mathbf{X}; \hat{\theta}) \geq \mathcal{L}(\mathbf{X}; \theta_0)$. This is all we need for Proposition 1 to hold.

Consistency of Principal Components

This case does not follow immediately from the proof of Proposition 1. In fact, under the approximating model of the principal components we have a restricted parameter space, say Θ_{pc}^c, that does not necessarily contains θ_0^c defined above for which the idiosyncratic component is left unrestricted. However, if we replace in the proof of Proposition 1 θ_0^c with

$$\theta_0^{pc} := \left\{A(L) = I_r; Q = I_r, \Lambda = \Lambda_0; \Psi_d = \sigma_0^2 I_n\right\}$$

where $\sigma_0^2 = \frac{1}{n}\text{Tr}\Psi_0$, the result will follow along the same lines since we would have $\theta_0^{pc} \in \Theta_{pc}^c$ and hence $\mathcal{L}(\mathbf{X}; \hat{\theta}) \geq \mathcal{L}(\mathbf{X}; \theta_0^{pc})$. In addition it is possible to show that $\mathbf{F}_{\theta_0^{pc}}$ have the same asymptotic properties of $\mathbf{F}_{\theta_0^c}$. A detailed proof is available under request.