Wiley StatsRef:
Statistics Reference Online

Factor Modeling for High-Dimensional Time Series

By *Chun Yip Yau*

Keywords: *maximum likelihood, Kalman filter, Whittle likelihood, principle component, information criteria*

Abstract: This article investigates the analysis of high-dimensional time series via factor modeling, with a focus on the methodological and computational aspects of estimating factor models. In particular, we review various estimation methods for the factor loadings and factor process, covering time- and frequency-domain, likelihood- and principle-component-based procedures. Moreover, determination of the number of factors is briefly discussed.

1 Introduction

Consider a high-dimensional time series $\{Y_t\}_{t=1,\dots,n}$, where $Y_t = (Y_{1,t}, Y_{2,t} \dots, Y_{p,t})' \in \mathbb{R}^p$ is a p-dimensional vector observed at time t. We call $\mathbf{y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,n})'$ the ith component of the high-dimensional time series. For example, if $Y_{i,t}$ represents the price of stock i at time t, then Y_t collects the prices of all p stocks at time t, and $\mathbf{y}_i$ collects the historical prices of the ith stock from time 1 to n. The whole dataset is denoted by the $n \times p$-dimensional matrix $Y = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p)$.

When p is large, it is difficult to simultaneously analyze the cross-correlations among the p components and the serial dependence across time. For example, the simplest vector autoregressive (VAR) model of order 1,

$$Y_t = \Phi Y_{t-1} + \epsilon_t\ ,\ \epsilon_t \sim N(0, \Sigma_\epsilon)$$

already involves p^2 unknown parameters in the coefficient matrix $\Phi \in \mathbb{R}^{p\times p}$. Therefore, some structure has to be imposed for feasible estimation. For a small integer r, the *factor model* of the form

$$Y_t = \Lambda F_t + \epsilon_t \tag{1}$$

where $\Lambda \in \mathbb{R}^{p\times r}$ and $F_t \in \mathbb{R}^r$ is a latent low-dimensional stationary process, which provides a parsimonious option for modeling high-dimensional data. Intuitively, (1) states that the serial and cross-sectional dependence of the p-dimensional time series Y_t is captured by the r-dimensional latent common factor F_t. The error term $\epsilon_t \in \mathbb{R}^p$, also known as the *idiosyncratic component*, is an orthogonal white noise, that is,

Chinese University of Hong Kong, Shatin, Hong Kong

DOI: 10.1002/9781118445112.stat08291

$\mathrm{Cov}(\epsilon_t, \epsilon_s) = \Delta 1_{\{s=t\}}$, where Δ is a diagonal matrix. Also, $\mathrm{Cov}(F_t, \epsilon_s) = 0$ for all t, s. Equivalently, (1) can be expressed as

$$Y' = \Lambda F + \boldsymbol{\epsilon} \tag{2}$$

where $F = (F_1\ F_2\ \cdots\ F_n) \in \mathbb{R}^{r\times n}$ and $\boldsymbol{\epsilon} = (\epsilon_1\ \epsilon_2\ \cdots\ \epsilon_n) \in \mathbb{R}^{p\times n}$.

Factor models can be classified as *static* or *dynamic*, according to whether time dependence is described in the model. For example, (1) is a static model since all variables appear at the same time t. On the other hand, dynamic factor models capture the dependence of Y_t on previous lags of the factor process and take the form

$$Y_t = \Lambda_0 F_t + \Lambda_1 F_{t-1} + \cdots + \Lambda_s F_{t-s} + \epsilon_t \tag{3}$$

where $\Lambda_j \in \mathbb{R}^{p\times r}$, $j = 1, \ldots, s$. Indeed, by defining $F_t = (F_t', F_{t-1}', \ldots, F_{t-s}')'$ and $\Lambda = (\Lambda_0, \Lambda_1, \ldots, \Lambda_s)$, the dynamic factor model (3) with an r-dimensional factor and lag order s can be expressed as a static factor model (1) with an $r(1+s)$-dimensional factor. Although static models cover dynamic models, dynamic model has a merit in explicitly describing the time dependence of the factor process on the data.

Factor model can also be classified as *exact* or *approximate*. If the noise ϵ_t has no cross and serial correlation, that is, Δ is diagonal and $\mathrm{Cov}(\epsilon_t, \epsilon_s) = 0$ if $s \neq t$, then the model is exact. Relaxing this assumption to allow for cross or serial-correlation results in the approximate factor model. Since the factor loadings and factor process are usually of primary interest, for simplicity, we focus on the exact factor model unless otherwise specified. Indeed, many methods reviewed in this article also apply to the approximate factor model.

This article reviews the methodological and computational aspects of various estimation methods of factor models for high-dimensional time series. We introduce the identifiability issue of factor modeling in Section 2. In Section 3, we discuss estimation methods for the factor loadings and factor process under a known number of factors. Finally, determination of the number of factors is briefly discussed in Section 4.

2 Identifiability

In (1), as both Λ and F are unobserved, there is an issue of identifiability. Specifically, for any invertible matrix $H \in \mathbb{R}^{r\times r}$, (2) can be expressed as $Y' = (\Lambda H)(H^{-1}F) + \epsilon$. Thus, the sets (Λ, F) and $(\Lambda H, H^{-1}F)$ correspond to the same model. Since it requires r^2 parameters to completely specify an $r \times r$ invertible matrix, to ensure identifiability, we need r^2 restrictions on Λ and F.

Some common identifiability conditions include $\Lambda'\Lambda = I_r$, $\Lambda'\Lambda/p = I_r$, $FF' = I_r$, or $FF'/n = I_r$, which assert that either the columns of Λ or rows of F are orthonormal; see Bai and Ng[1]. However, ensuring the orthonormality of r vectors only requires $r(r+1)/2$ constraints. Thus, an additional set of $r(r-1)/2$ constraints has to be imposed to completely ensure the identifiability. For simplicity, we consider the following identifiability constraints:

$$\text{(i) } \Lambda'\Lambda = I_r, \qquad \text{(ii) } \lambda_{ij} = 0 \text{ for } 1 \le i < j \le r, \qquad \text{(iii) } \lambda_{ii} > 0 \text{ for } i = 1, \ldots, r \tag{4}$$

Although (ii) in (4) provides the required $r(r-1)/2$ constraints for identifiability, the sign of each column of Λ is still unidentified since (i) only ensures that the column norm is 1. Therefore, (iii) is needed to fix the signs of the columns. Given an arbitrary Λ satisfying (i), one can multiply Λ by a suitable rotation matrix to ensure that (ii) holds.

For other choices of identifying restrictions, see Bai and Li[2] for details.

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

3 Estimation of High-Dimensional Factor Model

3.1 Least-Squares or Principal Component Estimation

One way to estimate Λ is to minimize the least-squares criterion

$$S(\Lambda, F) = \sum_{t=1}^{n} |Y_t - \Lambda F_t|^2 = \text{tr}((Y' - \Lambda F)'(Y' - \Lambda F)) \tag{5}$$

with, respectively, to Λ and F, subject to the identifiability constraints. Let $(\hat{\Lambda}, \hat{F})$ be the minimizer of S in (5). Differentiating S with respect to F, we have $-2\hat{\Lambda}'Y' + 2\hat{\Lambda}'\hat{\Lambda}\hat{F} = 0$. Combining with the identifiability constraint $\hat{\Lambda}'\hat{\Lambda} = I_r$ gives

$$\hat{F} = \hat{\Lambda}'Y' \tag{6}$$

Substituting (6) back to (5), we have $S(\hat{\Lambda}, \hat{F}) = \text{tr}(YY') - \text{tr}(\hat{\Lambda}'Y'Y\hat{\Lambda})$. In other words, $\hat{\Lambda}$ maximizes the quantity $\text{tr}(\hat{\Lambda}'Y'Y\hat{\Lambda})$. From standard matrix algebra, we have

$$\hat{\Lambda} = (V_1 \; V_2 \; \cdots \; V_r) \tag{7}$$

where $V_i \in \mathbb{R}^p$ is the eigenvector corresponding to the ith largest eigenvalue of $Y'Y$. Note that as $Y'Y$ is a nonnegative definitive symmetric matrix, $V_i'V_j = 1_{\{i=j\}}$. Thus, the constraint $\hat{\Lambda}'\hat{\Lambda} = I_r$ is automatically satisfied.

If the noises $\{\epsilon_t\}$ are heteroscedastic or autocorrelated, Choi[3] suggests a weighted least-squares approach, which minimizes $\text{tr}(\hat{\Delta}^{-1}(Y' - \Lambda F)'(Y' - \Lambda F))$, where the estimator $\hat{\Delta}$ of Δ may be obtained from the residuals of the ordinary least-squares estimator.

3.2 Factor Loading Space Estimation

This approach, developed by Lam *et al.*[4], considers estimating the column space of the factor loading matrix Λ. Specifically, since $\{\epsilon_t\}$ is a sequence of white noises uncorrelated with $\{F_t\}$, we have, for $k \geq 1$,

$$\Sigma_y(k) := \text{Cov}(Y_{t+k}, Y_t) = \Lambda \text{Cov}(F_{t+k}, F_t)\Lambda' =: \Lambda\Sigma_f(k)\Lambda' \tag{8}$$

Clearly, (8) implies that the columns of $\Sigma_y(k)$ lie in the column space of Λ. To pool together the information from different time lags without cancellations, one can consider

$$M := \sum_{k=1}^{k_0} \Sigma_y(k)\Sigma_y'(k) = \Lambda \left(\sum_{k=1}^{k_0} \Sigma_f(k)\Sigma_f'(k) \right) \Lambda' \tag{9}$$

where k_0 is a prespecified constant. By construction, the column spaces of M and Λ are the same. Since the column spaces of a matrix can be captured by the eigenvectors of the matrix, one can estimate Λ by

$$\hat{\Lambda} = (V_1 \; V_2 \; \cdots \; V_r) \tag{10}$$

where $V_i \in \mathbb{R}^p$ is the eigenvector corresponding to the ith largest eigenvalue of $\hat{M} = \sum_{k=1}^{k_0} \hat{\Sigma}_y(k)\hat{\Sigma}_y'(k)$, the sample version of M. Note that as $\hat{M}$ is a nonnegative definitive symmetric matrix, $V_i'V_j = 1_{\{i=j\}}$. Thus, the constraint $\hat{\Lambda}'\hat{\Lambda} = I_r$ is automatically satisfied. Given $\hat{\Lambda}$, one can use (6) to estimate the factor process by $\hat{F} = \hat{\Lambda}'Y'$. Simulation evidences in Lam *et al.*[4] suggest that the choice of k_0 is not sensitive, and usually $k_0 \leq 5$ provides satisfactory results.

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

We remark that capturing the information of factor loading space is not limited to $\hat{M}$. Gao and Tsay[5] consider an alternative of $\hat{M}$ based on canonical correlation analysis; Pan and Yao[6] propose to estimate the factor loading space using an innovation expansion algorithm which requires solving a sequence of nonlinear optimization problems.

3.2.1 Improved Estimation of Factor Process

From (2), the estimated factor can be expressed as $\hat{F} = \hat{\Lambda}'Y' = \hat{\Lambda}'\Lambda F + \hat{\Lambda}'\boldsymbol{\epsilon}$. Even if $\hat{\Lambda} \approx \Lambda$, the term $\hat{\Lambda}'\boldsymbol{\epsilon}$ could contribute to a large bias on the estimation of F_t, especially when p is large. To tackle this problem, Gao and Tsay[7] propose a procedure to estimate the factor process by first eliminating the effect of $\boldsymbol{\epsilon}$. Specifically, assume that

$$Y' = \Lambda F + \boldsymbol{\epsilon} = \Lambda F + \Gamma \mathbf{e} \tag{11}$$

where $\Gamma \in \mathbb{R}^{p\times(p-r)}$ satisfies $\Gamma'\Gamma = I_{p-r}$, $\mathbf{e} = (e_1\ e_2\ \cdots\ e_n) \in \mathbb{R}^{(p-r)\times n}$, $e_t \in \mathbb{R}^{p-r}$ is a white noise with covariance matrix Σ_e, and the largest K eigenvalues of Σ_e are diverging.

To remove the effect of $\boldsymbol{\epsilon} = \Gamma\mathbf{e}$ in estimating F, the idea is to construct a matrix $\hat{B}$ such that

$$\text{(i)}\ \ \hat{B}'\Gamma\mathbf{e}\ \text{ is negligible;} \qquad \text{(ii)}\ \ \hat{B}'\hat{\Lambda}\ \text{ is invertible.}$$

If (i) and (ii) hold, then multiplying $\hat{B}'$ on both sides of (11) gives $\hat{B}'Y' \approx \hat{B}'\Lambda F$, and thus F can be estimated by

$$(\hat{B}'\hat{\Lambda})^{-1}\hat{B}'Y' \tag{12}$$

To achieve (i), one can estimate the orthogonal complement of $\Gamma\mathbf{e}$, which is equivalent to the orthogonal complement of $\Gamma\mathbf{e}\mathbf{e}'\Gamma' \approx \Gamma\Sigma_e\Gamma'$. Following the idea in (8) and (9), define

$$\Sigma_y := \mathrm{Cov}(Y_t, Y_t) = \Lambda\mathrm{Cov}(F_t, F_t)\Lambda' + \Gamma\mathrm{Cov}(e_t, e_t)\Gamma' =: \Lambda\Sigma_f\Lambda' + \Gamma\Sigma_e\Gamma' \tag{13}$$

To extract the information about Γ, eliminate Λ in (13) by defining

$$S := \Sigma_y\Lambda_c\Lambda_c'\Sigma_y = \Gamma\Sigma_e\Gamma'\Lambda_c\Lambda_c'\Gamma'\Sigma_e\Gamma \tag{14}$$

where $\Lambda_c \in \mathbb{R}^{p\times(p-r)}$ is the orthogonal complement of Λ. By construction, S is symmetric, nonnegative definite with columns lying in the column space of $\Gamma\Sigma_e\Gamma'$. As Σ_e contains K diverging eigenvalues, the orthogonal complement of $\Gamma\Sigma_e\Gamma'$ can be estimated by the eigenvectors corresponding to the $p-K$ smallest eigenvalues of S.

Next, we discuss the estimation of S. In view of (10), the orthogonal complement of Λ, that is, Λ_c, can be estimated by the eigenvectors corresponding to the $p-r$ smallest eigenvalues of $\hat{M}$, that is,

$$\hat{\Lambda}_c = (V_{r+1}\ V_{r+2}\ \cdots\ V_p) \tag{15}$$

Thus, the sample version of S is given by $\hat{S} = \hat{\Sigma}_y\hat{\Lambda}_c'\hat{\Lambda}_c\hat{\Sigma}_y$. Given $\hat{S}$, one can define $B_* \in \mathbb{R}^{p\times(p-K)}$ as the matrix containing the eigenvectors corresponding to the $p-K$ smallest eigenvalues of $\hat{S}$, so that $B_*'\Gamma\mathbf{e}$ is negligible.

However, setting $\hat{B} = B_*$ does not fulfill ii) since $B_*'\hat{\Lambda} \in \mathbb{R}^{(p-K)\times r}$ is not a square matrix. This can be fixed by constructing a matrix $\hat{R} \in \mathbb{R}^{(p-K)\times r}$ so that $\hat{R}'B_*'\hat{\Lambda} \in \mathbb{R}^{r\times r}$ is invertible. This $\hat{R}$ can be defined as the matrix whose columns are the r eigenvectors corresponding to the r largest eigenvalues of $B_*'\hat{\Lambda}\hat{\Lambda}'B_*$. In conclusion, with $\hat{B} := B_*\hat{R}$, i) and ii) are fulfilled, and the improved estimator of F can be computed by (12).

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

3.3 Frequency-Domain Approach

The frequency-domain approach, developed by Forni *et al*.[8] and Forni *et al*.[9], addresses the dynamic factor model (3), with the assumption $\mathrm{Var}(\epsilon_t) = \Delta$ relaxed to a symmetric matrix instead of a diagonal matrix. The idea is based on the following result on frequency-domain time series:

Theorem 1. (Theorem 9.3.1 of Brillinger[9]). *Consider a zero-mean p-dimensional stationary time series* $\{Y_t\}$ *with absolutely summable autocovariance function* $\Sigma_y(k)$ *and spectral density matrix*

$$f_Y(\omega) := \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} e^{-i\omega k}\Sigma_y(k)$$

Let $\{\mathbf{b}(u)\}$ *and* $\{\mathbf{c}(u)\}$, *respectively, be* $r \times p$- *and* $p \times r$-*dimensional filters such that*

$$E\left[\left(Y_t - \sum_{u=-\infty}^{\infty}\mathbf{c}(t-u)\boldsymbol{\xi}_u\right)'\left(Y_t - \sum_{u=-\infty}^{\infty}\mathbf{c}(t-u)\boldsymbol{\xi}_u\right)\right] \tag{16}$$

achieves its minimum value among all possible $r \times p$- *and* $p \times r$-*dimensional filters, where*

$$\boldsymbol{\xi}_t = \sum_{u=-\infty}^{\infty}\mathbf{b}(t-u)Y_t$$

Then, $\{\mathbf{b}(u)\}$ *and* $\{\mathbf{c}(u)\}$ *are, respectively, given by*

$$\mathbf{b}(u) = \frac{1}{2\pi}\int_0^{2\pi}\mathbf{B}(\alpha)e^{iu\alpha}\,d\alpha, \qquad \mathbf{c}(u) = \frac{1}{2\pi}\int_0^{2\pi}\mathbf{C}(\alpha)e^{iu\alpha}\,d\alpha$$

where $\mathbf{B}(\omega) = (V_1(\omega)\ V_2(\omega)\cdots\ V_r(\omega))$, $\mathbf{C}(\omega) = \mathbf{B}(\omega)'$, *and* $V_j(\omega)$ *is the eigenvector corresponding to the jth largest eigenvalue of* $f_Y(\omega)$.

From Theorem 1, the r-dimensional filtered process $\boldsymbol{\xi}_t$ can be viewed as the factors of Y_t in the sense that $Y_t - \sum_{u=-\infty}^{\infty}\mathbf{c}(t-u)\boldsymbol{\xi}_u$ is small. However, $\{\mathbf{c}(u)\}$ is a "two-sided" filter, and it requires future ξ_u's to approximate Y_t. Thus, $\boldsymbol{\xi}_t$ cannot serve as the factor F_t in model (3) since only past F_u's are required to approximate Y_t. To tackle this problem, Forni *et al*.[9] return to the time domain using the spectral densities of the process $X_t = \sum_{u=-\infty}^{\infty}\mathbf{c}(t-u)\boldsymbol{\xi}_u$ and the idiosyncratic component $\varepsilon_t = Y_t - X_t$, denoted by $f_X(\omega)$ and $f_\varepsilon(\omega)$, respectively. Specifically, the covariance matrices $\Sigma_X(k)$ and $\Sigma_\varepsilon(k)$ are first computed from $f_X(\omega)$ and $f_\varepsilon(\omega)$. Then, for $j = 1, \ldots, p$, consider the generalized eigenvalue λ_j, which satisfies

$$\lambda_j = \arg\max_{\lambda\in\mathbb{R}^p}\lambda'\Sigma_X(0)\lambda \text{ s.t. } \lambda'\Sigma_\epsilon(0)\lambda = 1 \text{ and } \lambda'\Sigma_\epsilon(0)\lambda_i = 0 \tag{17}$$

for $i = 1, \ldots, j-1$. Intuitively, $\lambda_j Y_t = \lambda_j X_t + \lambda_j\varepsilon_t$ has $\lambda_j X_t$ maximized with $\lambda_j\varepsilon_t$ bounded and orthogonal to $\lambda_i\varepsilon_t$ for $i = 1, \ldots, j-1$. Thus, the linear combination $\lambda_j Y_t$ is close to the factor space, and $(\lambda_1 Y_t, \lambda_2 Y_t, \ldots, \lambda_r Y_t)'$ can be used as an estimate for the common factor F_t.

The computation of the frequency-domain estimation is summarized as follows:

Step 1: Spectral densities estimation

a) Estimate the spectral density matrix of Y_t by

$$\hat{f}_Y(\omega_h) = \frac{1}{2\pi}\sum_{k=-M}^{M}\left(1 - \frac{|k|}{M+1}\right)\hat{\Sigma}_y(k)e^{-ik\omega_h}$$

where $\omega_h = \frac{2\pi h}{2M+1}$, $h = 0, 1, \ldots, 2M$ are frequencies. The tuning parameter M has to satisfy $M \to \infty$ and $M/n \to \infty$ as $n \to \infty$, and a rule-of-thumb choice is $M = \frac{2}{3}n^{1/3}$.

DOI: 10.1002/9781118445112.stat08291

b) For $h = 0, 1, \ldots, 2M$, compute the eigenvectors $\hat{V}_j(\omega_h)$ corresponding to the j-largest eigenvalues $\hat{a}_j$ of $\hat{f}_Y(\omega_h)$, $j = 1, \ldots, r$.
c) Estimate the spectral densities of X_t and ε_t by

$$\hat{f}_X(\omega_h) = \sum_{j=1}^{r} \hat{a}_j \hat{V}_j(\omega_h)\hat{V}_j^*(\omega_h) \ , \text{ and } \hat{f}_\varepsilon(\omega_h) = \hat{f}(Y)(\omega_h) - \hat{f}(X)(\omega_h)$$

where v^* stands for the transpose of the complex conjugate of a vector v.
d) Compute the sample covariance matrices of X_t and ε_t by the inverse Fourier transform

$$\hat{\Sigma}_X(k) = \frac{1}{2M+1} \sum_{k=-M}^{M} \hat{f}_X(\omega_h) e^{ik\omega_h},$$

$$\hat{\Sigma}_\varepsilon(k) = \frac{1}{2M+1} \sum_{k=-M}^{M} \hat{f}_\varepsilon(\omega_h) e^{ik\omega_h},$$

evaluated at $k = 0$

Step 2: Generalized eigenvalue estimation

a) Compute the generalized eigenvalue $\hat{\lambda}_j$ (see Theorem A.2.4 of Anderson[11]), which satisfies

$$\hat{\lambda}_j = \arg\max_{\lambda \in \mathbb{R}^p} \lambda' \hat{\Sigma}_X(0)\lambda \text{ s.t. } \lambda' \hat{\Sigma}_\varepsilon(0)\lambda = 1 \text{ and } \lambda' \hat{\Sigma}_\varepsilon(0)\lambda_i = 0$$

for $j = 1, \ldots, p$ and $i = 1, \ldots, j-1$.
b) Set $\hat{F}_t = (\hat{\lambda}_1 Y_t, \hat{\lambda}_2 Y_t, \ldots, \hat{\lambda}_r Y_t)'$. The estimated factor loading $\hat{\Lambda}_j, j = 0, \ldots, s$ can be obtained by regressing Y_t against $\hat{F}_t$ using (3).

3.4 Likelihood-Based Estimation

Maximum-likelihood estimation (MLE) is a popular approach in statistics as it is efficient in many classical statistics problems[12]. To conduct maximum-likelihood estimation, a parametric model has to be imposed. Therefore, in view of (1) and (3), one has to impose a model for the factor process $\{F_t\}$, and a distribution for ϵ_t. The multivariate normal distribution $N(\mathbf{0}, \Delta)$ is a natural candidate for ϵ_t. Also, since $\{F_t\}$ is a low-dimensional process, it is natural to further assume that $\{F_t\}$ follows a vector ARMA (VARMA) model, that is,

$$\Phi(B)F_t = \Theta(B)z_t \tag{18}$$

where $z_t \sim N(\mathbf{0}, \Sigma_z)$, $\Phi(B) = I - \Phi_1 B - \cdots - \Phi_{\tilde{p}} B^{\tilde{p}}$, $\Theta(B) = I + \Theta_1 B + \cdots + \Theta_{\tilde{q}} B^{\tilde{q}}$, Σ_z, Φ_j, and Θ are $r \times r$-dimensional matrices. On the other hand, the likelihood function can be formulated conditional on $\{F_t\}$, that is, treating F_t as fixed numbers[2]. In this section, we review several likelihood-based methods for the estimation of factor models.

3.4.1 Exact Likelihood via Kalman filtering

Consider the factor model (1) with VARMA$(\tilde{p}, \tilde{q})$ factor process (18). In Jungbacker and Koopman[13], the noise is also allowed to be a VAR process

$$\epsilon_t = \Psi_1 \epsilon_{t-1} + \cdots + \Psi_{\tilde{s}} \epsilon_{t-\tilde{s}} + e_t \tag{19}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

where $e_t \sim N(\mathbf{0}, \Sigma_e)$ is the white noise. Without loss of generality, assume $\tilde{s} \geq \tilde{p}$. Otherwise, one can regard the factor process as a VARMA$(\tilde{s}, \tilde{q})$ process with the last $\tilde{s} - \tilde{p}$ autoregressive coefficient matrices equaling zero. Denote $\Psi(B) = (I_p - \Psi_1 B - \cdots - \Psi_{\tilde{s}} B^{\tilde{s}})$, where B is the backshift operator, so that $\Psi(B)\epsilon_t = e_t$. Multiplying $\Psi(B)$ to both sides of (1), together with the VAR(1) representation of the VARMA process

$$\alpha_t := \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-\tilde{p}+1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{\tilde{p}} \\ I_r & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & I_r & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & I_r \end{pmatrix} \begin{pmatrix} F_{t-1} \\ F_{t-2} \\ \vdots \\ F_{t-\tilde{p}} \end{pmatrix} + \begin{pmatrix} I_r & \Theta_1 & \cdots & \Theta_{\tilde{q}} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \begin{pmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-\tilde{q}} \end{pmatrix} =: H\alpha_{t-1} + R\eta_t \tag{20}$$

one obtains the *state space representation*

$$\begin{aligned} \Psi(B)Y_t &= Z\alpha_t + e_t \\ \alpha_t &= H\alpha_{t-1} + R\eta_t \ , \ \eta_t \sim N(\mathbf{0}, Q) \end{aligned} \tag{21}$$

where $Z = (\Lambda, -\Psi_1\Lambda, \ \ldots \ , -\Psi_{\tilde{s}}\Lambda, \mathbf{0}) =: (\overline{\Lambda}, \mathbf{0}) \in \mathbb{R}^{p \times \tilde{p}r}$, with $\overline{\Lambda} \in \mathbb{R}^{p \times (\tilde{s}+1)r}$, $\mathbf{0} \in \mathbb{R}^{p \times (\tilde{p}-\tilde{s}-1)r}$ is a zero matrix, and Q is defined by the Kronecker product $Q = I_{\tilde{q}+1} \bigotimes \Sigma_z$. To enhance computations, Jungbacker and Koopman[13] construct a full rank matrix $A = \begin{pmatrix} A_L \\ A_H \end{pmatrix} \in \mathbb{R}^{p \times p}$, where $A_L = (\overline{\Lambda}' \Sigma_e^{-1} \overline{\Lambda})^{-1} \overline{\Lambda}' \Sigma_e^{-1} \in \mathbb{R}^{(\tilde{s}+1)r \times p}$ and $A_H \in \mathbb{R}^{(\tilde{p}-\tilde{s}-1)r \times p}$, such that

$$\text{(i) } A_L \Sigma_e A_H' = 0, \ \text{(ii) } A_H Z = 0, \ \text{(iii) } |A_H \Sigma_e A_H'| = I_{p-(\tilde{s}+1)r} \tag{22}$$

With (22), multiplying A on both sides of the first equation of (21) yields

$$Y_t^L := A_L \Psi(B) Y_t = A_L Z \alpha_t + A_L e_t =: \tilde{Z}\alpha_t + e_t^L \tag{23}$$

$$Y_t^H := A_H \Psi(B) Y_t = A_H e_t =: e_t^H \tag{24}$$

$$\alpha_t = H\alpha_{t-1} + R\eta_t \tag{25}$$

where $e_t^L = A_L e_t$, and $e_t^H = A_H e_t$ satisfies $\begin{pmatrix} e_t^L \\ e_t^H \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} (\overline{\Lambda}' \Sigma_e^{-1} \overline{\Lambda})^{-1} & \mathbf{0} \\ \mathbf{0} & A_H \Sigma_e A_H' \end{pmatrix}\right) =: \begin{pmatrix} \Sigma_H & \mathbf{0} \\ \mathbf{0} & \Sigma_L \end{pmatrix}$. Since A is of full rank, the likelihood functions of $\{Y_1, \ \ldots \ , Y_n\}$ and $\{AY_1, \ \ldots \ , AY_n\}$ only differ by a Jacobian term $\log |A|^n$. Together with the independence of e_t^H and e_t^L, the log-likelihood function is

$$\ln L(Y_1, \ \ldots \ , Y_n) = \ln L(Y_1^L, \ \ldots \ , Y_n^L) + \ln L(Y_1^H, \ \ldots \ , Y_n^H) + n \ln |A| \tag{26}$$

Now, we evaluate each of the three terms on the right side of (26). First, as $\{Y_t^{(L)}\}_{t=1, \ \ldots \ , n}$ follows a low-dimensional state space model (23) and (25), its likelihood can readily be computed by

$$\ln L(Y_1^L, \ \ldots \ , Y_n^L) = -\frac{(\tilde{s}+1)rn}{2} \ln 2\pi - \frac{1}{2} \sum_{t=1}^{n} \ln |D_t| - \frac{1}{2} \sum_{t=1}^{n} \nu_t' D_t \nu_t \tag{27}$$

where the quantities ν_t and D_t are computed via Kalman filtering,

$$\begin{aligned} \nu_t &= y_t^L - \tilde{Z} a_{t|t-1} \\ D_t &= \tilde{Z} P_{t|t-1} \tilde{Z}' + \Sigma_L \\ K_t &= H P_{t|t-1} \tilde{Z}' D_t^{-1} \\ a_{t+1|t} &= H a_{t|t-1} + K_t \nu_t \\ P_{t+1|t} &= H P_{t|t-1} H' - K_t D_t K_t' + RQR' \end{aligned}$$

DOI: 10.1002/9781118445112.stat08291

with initial values $a_{1|0} = E(\alpha_1) = \mathbf{0}$ and $P_{1|0} = \mathrm{Var}(\alpha_1)$; see, for example, Brockwell and Davis[14] for details. Second, from (24),

$$\begin{aligned} \ln L(Y_1^H, \ldots, Y_n^H) &= -\frac{(\tilde{p}-\tilde{s}-1)rn}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{n}(Y_t^H)'\Sigma_H^{-1}Y_t^H \\ &= -\frac{(\tilde{p}-\tilde{s}-1)rn}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{n}\tilde{e}_t'\Sigma_e^{-1}\tilde{e}_t \end{aligned} \tag{28}$$

where $\tilde{e}_t = (I_p - \Sigma_e A_L'(A_L\Sigma_e A_L')^{-1}A_L)\Psi(B)Y_t$, and the last equality follows from (22) and the fact that

$$A_H'(A_H\Sigma_e A_H')^{-1}A_H\Sigma_e + A_L'(A_L\Sigma_e A_L')^{-1}A_L\Sigma_e = I_p$$

since the two terms on the left side are orthogonal projection matrices spanning $\mathbb{R}^p$. Note that explicit formula of A_H is not required to compute (28). Finally, note from (22) (iii) that

$$|A|^2 = |\Sigma_e|^{-1}|A\Sigma_e A'| = |\Sigma_e|^{-1}|A_L\Sigma_e A_L'||A_H\Sigma_e A_H'| = |\Sigma_e|^{-1}|\Sigma_L|$$

which implies

$$n\ln|A| = -\frac{1}{2}\ln\frac{|\Sigma_e|}{|\Sigma_L|} \tag{29}$$

Combining (27), (28), and (29), the likelihood function can be computed efficiently. Jungbacker and Koopman[13] developed an EM algorithm to optimize the likelihood function.

Remark 1. Alternative to the exact likelihood approach, two-step procedures, which estimate the factor F_t first and then the model of the factor process (18), have been developed. Specifically, Doz *et al.*[15] obtain principle component estimates, denoted as $\{\hat{F}_t^{(1)}\}$ and $\hat{\Lambda}^{(1)}$, and fit a VAR model on $\{\hat{F}_t^{(1)}\}$ in the first step. In the second step, using $\hat{\Lambda}^{(1)}$ and the estimated parameters of the VAR model, Kalman filter is employed to update the estimate for the factor $\{F_t\}$. Bai and Li[16] use a similar second step, but with the first step replaced by the estimation method in Section 3.4.3. On the other hand, Doz *et al.*[17] employ the same first step as in Doz *et al.*[15] and propose a different second step, which estimates the model of factor process using quasi-likelihood.

3.4.2 Exact Likelihood via Matrix Decomposition

Instead of using Kalman filter and EM algorithm, Ng *et al.*[18] employ matrix decomposition techniques to efficiently compute the log-likelihood, score function, and Fisher information, so that Newton–Raphson method can be directly employed to compute the maximum-likelihood estimator. They consider the model

$$\begin{aligned} Y_t &= \Lambda F_t + \epsilon_t \\ F_t &= \Phi F_{t-1} + z_t \end{aligned} \tag{30}$$

where $\eta_t \sim N(0, \Delta)$, $\epsilon_t \sim N(0, \Sigma_z)$, and Δ is a diagonal matrix with diagonal elements $\overline{\Delta} = (\Delta_1, \ldots, \Delta_p)'$. The parameter set is defined at $\theta = (\Lambda, \Delta, \Phi)$. Define

$$F = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad L = \begin{pmatrix} \Lambda & 0 & \ldots & 0 \\ 0 & \Lambda & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Lambda \end{pmatrix}, \quad D = \begin{pmatrix} \Delta & 0 & \ldots & 0 \\ 0 & \Delta & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Delta \end{pmatrix}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

$S = YY^T$, $\Psi = \text{Var}(F)$, and $\Omega = \text{Var}(Y) = L\Psi L^T + D$. The negative log-likelihood function (ignoring the constant term) is

$$Q(\theta) = \frac{1}{2}[\text{tr}(\Omega^{-1}S) + \log|\Omega|] \tag{31}$$

The major challenge in (31) is the computation of the inverse and determinant of the $np \times np$-dimensional matrix Ω. To tackle this problem, Ng *et al*.[18] employ the matrix identities

$$\Omega^{-1} = D^{-1} - D^{-1}L(\Psi^{-1} + J)^{-1}L^T D, \text{ and } |\Omega| = |\Delta|^n|\Psi^{-1}|^{-1} \cdot |\Psi^{-1} + L^T DL|$$

to express $Q(\theta)$ as

$$Q(\theta) = \frac{1}{2}(\log|\Delta|^n \cdot |\Psi| \cdot |\Upsilon|^{-1} + Y^T D^{-1}Y - (LD^{-1}Y)^T \Upsilon (L^T D^{-1}Y)) \tag{32}$$

where $\Upsilon := (\Psi^{-1} + L^T D^{-1}L)^{-1}$. Since Ψ is the covariance matrix of the VAR process, it is a block Toeplitz matrix, and it can be shown that Ψ^{-1} is a tridiagonal block matrix. This observation substantially simplifies the computations involving Υ and thus $Q(\theta)$. Moreover, the score function and the Fisher information matrix can also be expressed in terms of Υ. This allows the whole estimation procedure to be completed in $O(np)$ steps in each iteration of Newton–Ralphson algorithm.

3.4.3 Bai and Li's Quasi-Likelihood Estimation

If the factor process $\{F_t\}$ and the noise $\{\epsilon_t\}$ are normally distributed, then we have from (1) that $Y_t \sim N(\mathbf{0}, \Sigma_y)$, where[2]

$$\Sigma_y := \text{Var}(Y_t) = \Lambda \text{Var}(F_t)\Lambda' + \Delta =: \Lambda\Sigma_F\Lambda' + \Delta \tag{33}$$

Thus, the marginal distribution of Y_t is

$$f(Y_t) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_y|}} e^{-\frac{1}{2}Y_t'\Sigma_y^{-1}Y_t} \tag{34}$$

Based on (34), Bai and Li[2] formulate the log-likelihood function as

$$\begin{aligned} \ln L &= -\frac{1}{2n}\ln|M_y| - \frac{1}{2n}\sum_{t=1}^{n} Y_t' M_y^{-1} Y_t \\ &= -\frac{1}{2n}\ln|M_y| - \frac{1}{2n}\text{tr}(\hat{\Sigma}_y M_y^{-1}) \end{aligned} \tag{35}$$

where $\hat{\Sigma}_y = \sum_{t=1}^{T} Y_t Y_t'/n$, $M_y := \Lambda M_F \Lambda' + \Delta$, and $M_F = FF'/n$. Note that (35) is a quasi-likelihood since it ignores the serial dependence of Y_t and approximates the true variance Σ_y of Y_t by M_y. The use of M_y, which involves the sample moment $M_F = FF'/n$ instead of the population moment Σ_F, can gain computational efficiency under certain identifiability conditions. For example, Bai and Li[2] consider the identifiability conditions

$$FF'/n = I_r, \text{ and } \frac{1}{N}\Lambda'\Delta^{-1}\Lambda \text{ is a diagonal matrix,}$$

so that $M_F = I_r$, and the parameters to be estimated reduce to Λ and Δ.

Differentiating (35) with respect to Λ and Δ, the maximum-likelihood estimates $\hat{\Lambda}$ and $\hat{\Delta}$ satisfy

$$\left.\frac{\partial \ln L}{\partial \Lambda}\right|_{\Lambda=\hat{\Lambda}} = \hat{\Lambda}'\hat{M}_y^{-1}(\hat{S}_y - \hat{M}_y) = 0 \tag{36}$$

$$\left.\frac{\partial \ln L}{\partial \text{diag}(\Delta)}\right|_{\Delta=\hat{\Delta}} = \text{diag}(\hat{M}_y^{-1}) - \text{diag}(\hat{M}_y^{-1}\hat{\Sigma}_y\hat{M}_y^{-1}) = 0 \tag{37}$$

DOI: 10.1002/9781118445112.stat08291

where $\hat{M}_y = \hat{\Lambda}'\hat{\Lambda} + \hat{\Delta}$, and diag($A$) is a vector containing the diagonal elements of a square matrix A. The high-dimensional system of equations (36) and (37) can be readily solved by EM algorithm, see Rubin and Thayer[19] and Bai and Li[2] for details. Recently, Bai and Liao[20] consider an extension to cover sparse covariance matrix for ϵ by introducing a penalty term in the quasi-likelihood function.

3.4.4 Breitung and Tenhofen's Quasi-Likelihood Estimation

Breitung and Tenhofen[21] consider model (1) with the noise process satisfying the autoregressive model

$$\epsilon_{it} = \sum_{j=1}^{q_i} \rho_{j,i}\epsilon_{i,t-j} + e_{it} \tag{38}$$

where $e_{it} \sim WN(0, \sigma_i^2)$ for $i = 1, \ldots, p$ and $t \in \mathbb{Z}$. Considering the distribution of ϵ_{it}, the quasi-log-likelihood function is given by

$$\ln L(\theta) = -\sum_{i=1}^{p} \frac{n - q_i}{2} \log \sigma_i^2 - \sum_{i=1}^{p} \sum_{t=p_i+1}^{n} \frac{(e_{it} - \rho_{1,i}e_{i,t-1} - \cdots - \rho_{q_i,i}e_{i,t-q_i})^2}{2\sigma_i^2} \tag{39}$$

where $e_{it} = Y_{it} - \sum_{j=1}^{p} \Lambda_{ij}F_{jt}$ and $\theta = (F, \Lambda, \rho_{1,1}, \ldots, \rho_{1,q_1}, \ldots, \rho_{p,q_p}, \sigma_1^2, \ldots, \sigma_p^2)$ are unknown parameters. The $\ln L(\theta)$ is a quasi-likelihood in the sense that (38) is only a "working" model, and misspecification is allowed. In contrast to Bai and Li[2], the quasi-likelihood here involves F as unknown parameters and thus induces higher computational burden. Therefore, it is infeasible to obtain the maximum-likelihood estimator by simultaneously solving the score functions

$$\frac{\partial \ln L}{\partial \Lambda} = \frac{1}{\sigma_i^2} \sum_{t=q_i+1}^{n} e_{it} \left[F_t - \sum_{j=1}^{q_i} \rho_{i,j} F_{t-j} \right] = \mathbf{0} \tag{40}$$

$$\frac{\partial \ln L}{\partial F_t} = \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \left(e_{it} - \sum_{j=1}^{q_i} \rho_{i,j} e_{i,t+j} \right) \Lambda_{i\cdot}' = \mathbf{0} \tag{41}$$

$$\frac{\partial \ln L}{\partial \rho_{k,i}} = \frac{1}{\sigma_i^2} \sum_{t=q_i+1}^{n} e_{it}(Y_{i,t-k} - \Lambda_{i\cdot}F_{t-k}) = 0 \tag{42}$$

$$\frac{\partial \ln L}{\partial \sigma_i^2} = \frac{1}{\sigma_i^4} \sum_{t=q_i+1}^{n} e_{it}^2 - \frac{n - q_i}{2\sigma_i^2} = 0$$

where $\Lambda_{i\cdot}$ is the ith column of Λ, and $e_{is} = 0$ for $s > n$.

To tackle this problem, Breitung and Tenhofen[21] suggest a two-step estimation as follows. In the first step, the principle component estimators $\hat{\Lambda}^{(1)}$ and $\hat{F}^{(1)}$ are obtained based on (6) and (7). Then, in the second step, some of the score functions are employed to compute the estimated parameters. Specifically, with the estimates from the first step, define $\hat{e}_{i,t} = Y_{i,t} - \hat{\Lambda}_{i\cdot}^{(1)}\hat{F}_t^{(1)}$. Then, using (42), for each $i = 1, \ldots, p$, one can solve for $(\hat{\rho}_{1,i}, \ldots, \hat{\rho}_{q_i,i})$ from

$$\sum_{t=q_i+1}^{n} (\hat{e}_{i,t} - \hat{\rho}_{1,i}\hat{e}_{i,t-1} - \cdots - \hat{\rho}_{q_i,i}\hat{e}_{i,t-q_i})\hat{e}_{i,t-k} = 0, \text{ for } k = 1, \ldots, q_i \tag{43}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

Note that solving (43) is equivalent to computing the least-squares estimator from the regression model $\hat{\epsilon}_{i,t} = \rho_{1,i}\hat{\epsilon}_{i,t-1} + \cdots + \rho_{q_i,i}\hat{\epsilon}_{i,t-q_i} + e_t$. Next, using (40), we solve for $\hat{\Lambda}_{i\cdot}$ from

$$\sum_{t=q_i+1}^{n} \left[(Y_{it} - \hat{\Lambda}_{i\cdot}\hat{F}_t^{(1)}) - \sum_{j=1}^{q_i} \hat{\rho}_{i,j}(Y_{it} - \hat{\Lambda}_{i\cdot}\hat{F}_t^{(1)}) \right] \left[\hat{F}_t^{(1)} - \sum_{j=1}^{q_i} \hat{\rho}_{i,j}\hat{F}_{t-j}^{(1)} \right] = 0 \tag{44}$$

Finally, to gain computational efficiency, (41) is modified as

$$\sum_{i=1}^{p} \frac{1}{\hat{\omega}_i^2}(Y_{it} - \hat{\Lambda}_{i\cdot}\hat{F}_t)\hat{\Lambda}_{i\cdot}' = 0 \tag{45}$$

to solve for $\hat{F}_t$, where $\hat{\omega}_i^2 := \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_{i,t}^2$. Intuitively, solving for $\hat{F}_t$ in (45) is equivalent to minimizing the weighted sum of squares $\sum_{i=1}^{p}(Y_{it} - \hat{\Lambda}_{i\cdot}F_t)^2/\hat{\omega}_i^2$ in which $\hat{\omega}_i^2$ is estimating the variance of $Y_{it} - \hat{\Lambda}_{i\cdot}F_t$. Although Breitung and Tenhofen[21] do not consider the estimation of σ_i^2 since (38) is only a working model, the estimator can be defined by

$$\hat{\sigma}_i^2 = \frac{1}{n - q_i} \sum_{t=q_i+1}^{n} (\hat{\epsilon}_{i,t} - \hat{\rho}_{1,i}\hat{\epsilon}_{i,t} - \cdots - \hat{\rho}_{q_i,i}\hat{\epsilon}_{i,t-q_i})^2$$

Note that each of (43), (44), and (45) involves low-dimensional root solving and thus can be computed efficiently.

3.4.5 Frequency-Domain (Whittle) Likelihood

Fiorentini *et al.*[22] propose a frequency-domain likelihood for the estimation of dynamic factor model (3) with the factor process following a VARMA model (18). Moreover, each component of the noise process $\{\epsilon_{it}\}_{t=1,\ldots}$ follows a univariate ARMA model

$$\alpha_i(B)\epsilon_{it} = \beta_i(B)e_{it}, \text{ where } e_{it} \sim N(0, \phi_i) \tag{46}$$

Denote the parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\phi}, \theta_f, \boldsymbol{\theta}_\epsilon, \Lambda)$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, θ_f is the parameters associated with the VARMA model (18), and $\boldsymbol{\theta}_\epsilon$ is the parameters associated with the ARMA models (46).

Denote $f_{\epsilon_i}(\omega)$ and $f_F(\omega)$ as the spectral density matrices of $\{\epsilon_{it}\}_{t=1,\ldots}$ and $\{F_t\}_{t=1,\ldots}$, respectively, evaluated at frequency ω. Assuming that the latent factors process $\{F_t\}$ are observed, then the independence of $\{\epsilon_{it}\}_{t=1,\ldots}$ across $i = 1, \ldots, p$ implies that the components $\mathbf{y}_1, \ldots, \mathbf{y}_p$ are independent given the factor process. Thus, the "complete data" frequency-domain log-likelihood has a simple decomposition

$$\begin{aligned} \ln L_{\boldsymbol{\theta}}(Y, F) &= \ln L_{\boldsymbol{\theta}}(Y|F) + \ln L_{\boldsymbol{\theta}}(F) = \sum_{i=1}^{p} \ln L_{\boldsymbol{\theta}}(\mathbf{y}_i|F) + \ln L(F) \\ &= \sum_{i=1}^{p} \mathrm{WL}_{\boldsymbol{\theta}}(\{\mathbf{y}_{it} - \Lambda_0 F_t - \cdots \Lambda_s F_{t-s}\}_{t=1,\ldots,n}; f_{\epsilon_i}) + \mathrm{WL}_{\boldsymbol{\theta}}(\{F_t\}_{t=1,\ldots,n}; f_F) \end{aligned} \tag{47}$$

where

$$WL_{\boldsymbol{\theta}}(\{z_t\}_{t=1,\ldots,n}; f_z) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{j=0}^{n-1}\ln|f_z(\omega_j)| - \frac{2\pi}{2}\sum_{j=0}^{n-1}\mathrm{tr}(f_z^{-1}(\omega_j)I_z(\omega_j))$$

is the Whittle likelihood (see, e.g., Brockwell and Davis[14]) of a time series $z_1, \ldots, z_n$ with spectral density f_z and periodogram I_z, and $\omega_j = \frac{2\pi j}{n}, j = 1, \ldots, n-1$ are the Fourier frequencies.

DOI: 10.1002/9781118445112.stat08291

In practice, $\{F_t\}$ are not observed. Nevertheless, parameter estimates can be obtained using the generalized EM principle, which asserts that for a given $\boldsymbol{\theta}^{(n)}$, any increase in $E(\ln L_{\boldsymbol{\theta}}(Y,F)|Y,\boldsymbol{\theta}^{(n)})$ must represent an increase in $\ln L_{\boldsymbol{\theta}}(Y,F)$. In other words, the sequence $\{\boldsymbol{\theta}^{(n)}\}_{n=1,\ldots,}$ where

$$\boldsymbol{\theta}^{(n+1)} = \arg\max_{\boldsymbol{\theta}} E(\ln L_{\boldsymbol{\theta}}(Y,F)|Y,\boldsymbol{\theta}^{(n)}) \tag{48}$$

guarantees that $\ln L_{\boldsymbol{\theta}^{(n)}}(Y,F)$ increases with n. To compute $\{\boldsymbol{\theta}^{(n)}\}_{n=1,\ldots,}$, Fiorentini *et al.*[22] derive an expression of $E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(n)}) := E(\ln L_{\boldsymbol{\theta}}(Y,F)|Y,\boldsymbol{\theta}^{(n)})$ (E-step) and conduct the maximization (M-step) in (48) by a zig-zag procedure, which solves $\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(n)})}{\partial \boldsymbol{\phi}} = 0$, $\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(n)})}{\partial \boldsymbol{\theta}_f} = 0$, $\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(n)})}{\partial \boldsymbol{\theta}_\epsilon} = 0$, and $\frac{\partial E_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{(n)})}{\partial \Lambda} = 0$ successively until convergence and sets the resulting parameter vector as $\boldsymbol{\theta}^{(n+1)}$.

4 Determining the Number of Factors

The estimation methods discussed in Section 3 require a prespecified number of factors, r. In this section we briefly summarize the existing methods for estimating r.

4.1 Information Criterion

As in many model selection problems, using information criterion is a popular approach to select r. Under this approach, the estimated number of factor $\hat{r}$ is the minimizer of an information criterion over a range of values of r, say $r = 0, 1, \ldots, r_{\max}$ for some prespecified $r_{\max}$. Typically, this choice of r strikes a good balance between a certain lack of fit measure and a model complexity penalty in the criterion. For example, Bai and Ng[1], Alessi *et al.*[23], and Li *et al.*[24] consider information criterion of the form

$$IC(r) = \ln\left[\frac{1}{np}\sum_{t=1}^{n}(Y_t - \hat{\Lambda}\hat{F}_t)^2\right] + r \times P(n,p) \tag{49}$$

where $P(n,p)$ is a function depending on n and p. Some examples include $P(n,p) = c\frac{np}{n+p}, c\frac{np}{n+p}\ln\left(c\frac{np}{n+p}\right)$, $\frac{\ln\min(\sqrt{n},\sqrt{p})}{\min(\sqrt{n},\sqrt{p})}$, where c is a positive constant. Choi and Jeong[25] systematically compare the empirical performance of the above ICs with some classical information criteria such as AIC, BIC, and Hannan and Quinn's criterion. Other classical information criterion such as final prediction error is also studied in Chan *et al.*[26].

4.2 Eigenvalues Difference/Ratio Estimators

As we have seen in Sections 3.1 and 3.2, estimation of factor models is highly connected to the largest eigenvalues of the sample covariance matrix. In particular, many estimators are developed based on the fact that if the number of factor is r, then the r-largest eigenvalues of the sample covariance matrix would be substantially greater than the rest in magnitude. Therefore, r corresponds to the index where a large value is observed in the ratio or difference of adjacent eigenvalues. For example, Lam and Yao[27] suggest that

$$\hat{r} = \arg\min_{1\le i\le R}\frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i} \tag{50}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

where the upper bound R may be taken as $p/2$ or $p/3$. Independently, Ahn and Horenstein[28] considered

$$\hat{r} = \arg\max_{1\le i\le R} \frac{\hat{\lambda}_i}{\hat{\lambda}_{i+1}} \text{ and } \hat{r} = \arg\max_{1\le i\le R} \frac{\ln(1+\hat{\lambda}_i/\sum_{k=i+1}^{M}\hat{\lambda}_k)}{\ln(1+\hat{\lambda}_{i+1}/\sum_{k=i+2}^{M}\hat{\lambda}_k)} \tag{51}$$

Xia *et al*.[29] modify (50) as the *contribution ratio* (CR) estimator

$$\hat{r} = \arg\min_{1\le i\le R} \frac{\hat{\lambda}_{i+1}/\sum_{k=i+1}^{M}\hat{\lambda}_k}{\hat{\lambda}_i/\sum_{k=i}^{M}\hat{\lambda}_k} \tag{52}$$

where $M = \min\{n, p\}$. Alternatively, Li *et al*.[30] propose

$$\hat{r} = \left\{ \text{The first } i \ge 1 \text{such that } \frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i} > 1 - d_n \right\} - 1 \tag{53}$$

where d_n is a threshold parameter that can be calibrated by simulating Gaussian vectors.

Besides the ratios, differences of eigenvalues can be employed to determine r. Onatski[31] proposes the estimator $\hat{r} = \max\{i \le r^n_{\max} : \hat{\lambda}_i - \hat{\lambda}_{i+1} \ge \delta\}$, where $r^n_{\max} \to \infty$, $r^n_{\max}/n \to 0$, and δ is a constant that needs to be calibrated based on the eigenvalues.

4.3 Testing Approaches

Gao and Tsay[7] adopted methods in testing high-dimensional white noise from Chang *et al*.[32] and Tsay[33] for estimating the number of factors. The idea is as follows. Recall in Section 3.2 that the estimated factor is $\hat{F} = \hat{\Lambda}'Y'$, where $\hat{\Lambda} = (V_1\ V_2\ \cdots\ V_r)$ defined in (10) are the eigenvectors corresponding to the rth largest eigenvalues of the matrix $\hat{M}$. In other words, denoting $\hat{G} = (V_1\ V_2\ \cdots\ V_p)$ and $\hat{\mathbf{u}}_t = (\hat{u}_{1t}, \ldots, \hat{u}_{pt})' := \hat{G}'Y_t$, the first r component of $\hat{\mathbf{u}}_t$ is $\hat{F}_t$, and the remaining components $\hat{\mathbf{w}}_{r,t} = (\hat{u}_{r+1,t}, \ldots, \hat{u}_{pt})$ should behave like a high-dimensional white noise if r is greater than the true number of factor. Therefore, one can test the null hypothesis that $\{\hat{\mathbf{w}}_{i,t}\}_{t=1,\ldots,n}$ is a white noise sequentially for $i = 1, 2, \ldots$, and set $\hat{r} = i$ if the ith test is the first one that does not reject the null hypothesis. Note that when $p > n$, the eigenvectors $V_{n+1}, \ldots, V_p$ are degenerate. In this case, Gao and Tsay[7] suggest using $\hat{G} = (V_1\ V_2\ \cdots\ V_{p_*})$ with $p_* = \epsilon n$ for a small number $\epsilon \in (0, 1)$.

A testing procedure based on eigenvalues is developed by Kapetanios[34]. The test statistic is given by $\hat{T}(r) = \hat{\tau}^r_{n,p}(\hat{\lambda}_{r+1} - \hat{\lambda}_{r_{\max}+1})$, where $\hat{\tau}^r_{n,p}$ is a normalizing constant determined by subsampling, and $r_{\max}$ is a prespecified positive integer. The critical value is also estimated by sumsampling. The test is applied for $i = 1, 2, \ldots$ sequentially, and the estimator $\hat{r}$ is defined as the first r such that $\hat{T}(r)$ does not exceed the critical value.

4.4 Estimation of Dynamic Factors

Since dynamic factor models (3) contain additional dynamic structure compared to (1), more delicate methods are required to estimate the dimension of the factor process, r. In Bai and Ng[35], Amengual and Watson[36], and Breitung and Pigorsch[37], a static factor model is first fitted to the data to obtain a factor process. Then, a VAR model is fitted to the factor process, and the number of dynamic factors is estimated based on some information criteria involving the estimated factor and the fitted VAR model. On the other hand, Hallin and Liska[38] and Onatski[39] directly estimate the number of dynamic factors using the eigenvalues of the periodogram, in the form of information criteria and a testing procedure, respectively.

DOI: 10.1002/9781118445112.stat08291

Acknowledgment

Supported in part by HKSAR-RGC Grants CUHK 14308218, 14305517, 14302719.

Related Articles

Factor Analysis; **Whittle Likelihood**; **History of Factor Analysis: A Statistical Perspective**; **Factor Analysis: Exploratory**; **Factor Score Estimation**; **Factor Analysis, Dynamic**; **State-Space Methods**

References

[1] Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70** (1), 191–221.
[2] Bai, J. and Li, K. (2012) Statistical analysis of factor models of high dimension. *Ann. Stat.*, **40**, 436–465.
[3] Choi, I. (2012) Efficient estimation of factor models. *Econometric Theory*, **28**, 274–308.
[4] Lam, C., Yao, Q. and Bathia, N. (2011) Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, 901–918.
[5] Gao, Z. and Tsay, R.S. (2019a) A structural-factor approach for modeling high-dimensional time series and space-time data. *J. Time Ser. Anal.*, **40**, 343–362.
[6] Pan, J. and Yao, Q. (2008) Modelling multiple time series via common factors. *Biometrika*, **95**, 365–379.
[7] Gao, Z., and Tsay, R.S. (2019b) Structural-factor modeling of high-dimensional time series: Another look at factor models with diverging eigenvalues. *arXiv:1808.07932*, 1–38.
[8] Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2000) The generalized dynamic factor model: identification and estimation. *Rev. Econ. Stat.*, **82** (4), 540–554.
[9] Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2005) The generalized dynamic factor model: one-sided estimation and forecasting. *J. Am. Stat. Assoc.*, **100**, 830–840.
[10] Brillinger, D.R. (1981) *Time Series: Data Analysis and Theory*, Rinehart and Winston, Inc., Holt.
[11] Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.
[12] Shao, J. (2003). *Mathematical Statistics*, 2nd edn, Springer Science & Business Media, Springer-Verlag, New York.
[13] Jungbacker, B. and Koopman, S.J. (2014) Likelihood-based dynamic factor analysis for measurement and forecasting. *Econ. J.*, **18** (2), 1–21.
[14] Brockwell, P.J. and Davis, R.A. (1991) *Time Series: Theory and Method*, Springer, New York.
[15] Doz, C., Giannone, D., and Reichlin, L. (2011) A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, **164**, 188–205.
[16] Bai, J. and Li, K. (2016) Maximum likelihood estimation and inference for approximate factor models of high dimension. *Rev. Econ. Stat.*, **98** (2), 298–309.
[17] Doz, C., Giannone, D., and Reichlin, L. (2012) A quasi maximum likelihood approach for large approximate dynamic models. *Rev. Econ. Stat.*, **94**, 1014–1024.
[18] Ng, C.T., Yau, C.Y., and Chan, N.H. (2015) Likelihood inferences for high dimensional dynamic factor analysis with applications in finance. *J. Comput. Graph. Stat.*, **24** (3), 866–884.
[19] Rubin, D.B. and Thayer, D.T. (1982) EM algorithms for ML factor analysis. *Psychometrika*, **47**, 69–76.
[20] Bai, J. and Liao, Y. (2016) Efficient estimation of approximate factor models via penalized maximum likelihood. *J. Econom.*, **191**, 1–18.
[21] Breitung, J. and Tenhofen, J. (2011) GLS estimation of dynamic factor model. *J. Am. Stat. Assoc.*, **106**, 1150–1166.
[22] Fiorentini, G., Galesi, A., and Sentana, E. (2018) A spectral EM algorithm for dynamic factor models. *J. Econom.*, **205**, 249–279.
[23] Alessi, L., Barigozzi, M., and Capasso, M. (2010) Improved penalization for determining the number of factors in approximate factor models. *Stat. Probab. Lett.*, **80**, 1806–1813.
[24] Li, H., Li, Q., and Shi, Y. (2017) Determining the number of factors when the number of factors can increase with sample size. *J. Econ.*, **197** (1), 76–86.
[25] Choi, I. and Jeong, H. (2019) Model selection for factor analysis: some new criteria and performance comparisons. *Econom. Rev.*, **38** (6), 577–596.

Wiley StatsRef: Statistics Reference Online, © 2014–2021 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08291

[26] Chan, N.H., Lu, Y., and Yau, C.Y. (2017) Factor modelling for high-dimensional time series: inference and model selection. *J. Time Ser. Anal.*, **38** (2), 285–307.
[27] Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Stat.*, **40** (2), 694–726.
[28] Ahn, L. and Horenstein, A.R. (2013) Eigenvalue ratio test for the number of factors. *Econometrica*, **81**, 1203–1227.
[29] Xia, Q., Liang, R., Wu, J., and Wong, H. (2018) Determining the number of factors for high-dimensional time series. *Stat. Interface.*, **11**, 307–316.
[30] Li, Z., Wang, Q., and Yao, Q. (2017) Identifying the number of factors from singular values of a large sample auto-covariance matrix. *Ann. Stat.*, **45** (1), 257–288.
[31] Onatski, A. (2010) Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.*, **92** (4), 1004–1016.
[32] Chang, J., Yao, Q., and Zhou, W. (2017) Testing for high-dimensional white noise using maximum cross-correlations. *Biometrika*, **104** (1), 111–127.
[33] Tsay, R. (2020) Testing for serial correlations in high-dimensional time series via extreme value theory. *J. Econom.*, **216**(1), 106–117.
[34] Kapetanios, G. (2010) A testing procedure for determining the number of factors in approximate factor models with large dataset. *J. Bus. Econ. Stat.*, **28** (3), 397–409.
[35] Bai, J. and Ng, S. (2007) Determining the number of primitive shocks in factor models. *J. Bus. Econ. Stat.*, **25**, 52–60.
[36] Amengual, D. and Watson, M.W. (2007) Consistent estimation of the number of dynamic factors in a large N and T panel. *J. Bus. Econ. Stat.*, **25** (1), 91–96.
[37] Breitung, J. and Pigorsch, U. (2011) A canonical correlation approach for selecting the number of dynamic factors. *Oxford B. Econ. Stat.*, **75**, 23–36.
[38] Hallin, M. and Liska, R. (2007) Determining the number of factors in the general dynamic factor model. *J. Am. Stat. Assoc.*, **102**, 603–617.
[39] Onatski, A. (2009) Testing hypotheses about the number of factors in large factor models. *Econometrica*, **77** (5), 1447–1479.

DOI: 10.1002/9781118445112.stat08291