ON THE GOODNESS-OF-FIT TEST IN A FACTOR MODEL WITH HIGH-DIMENSIONAL DATA

Damien Passemier\(^1\) & Jian-Feng Yao\(^2\)

\(^1\)Department of Electronic and Computer Engineering  
Hong Kong University of Science and Technology (HKUST)  
Clear Water Bay, Kowloon, Hong Kong  
eepassemier@ust.hk  

\(^2\)Department of Statistics and Actuarial Science  
The University of Hong Kong (HKU)  
Pokfulam Road, Hong Kong  
jeffyao@hku.hk

Abstract. Factor models appear in many areas, such as economics or signal processing. If the factors and errors are Gaussian, a likelihood-based theory is well-known since Lawley (1940). However, these results are obtained in the classical scheme where the data dimension \(p\) is kept fixed while the sample size \(n\) tends to infinity. This point of view is not valid anymore for large-dimensional data, and usual statistics have to be modified. In this talk, we consider the strict factor model with homoscedastic variance. First, we give the bias of the maximum likelihood estimator of the noise variance by giving a CLT. We then give a bias-corrected estimator. Secondly, we present a corrected likelihood ratio test of the hypothesis that the factor model fits. Throughout the talk, simulation experiments are conducted to access the quality of our results.

Keywords. Covariance matrix, factor model random matrix theory, high-dimensional statistics, likelihood ratio test, maximum-likelihood estimation.

1 Introduction

In a factor model, variables are described as linear combinations of factors with added noise. This model, which first appears in psychology, is now widely used and appears in many scientific fields: in finance, the Arbitrage Pricing Theory (APT) of [15] heavily rely on factor analysis model. Similar models can be found in physics of mixture, see [9, 12], population genetics or wireless communications [6, 7, 16]. More recently, spiked population models have been introduced in [8] that encompass factor models.

A statistical theory for the maximum likelihood estimation is well-known since [11]. [2] also gives a likelihood ratio test for model fit which has an asymptotic \(\chi^2\) distribution under the null. However, these results are developed from a classical point of view where the data dimension \(p\) is kept fixed while the sample size \(n\) tends to infinity. This scheme is not valid anymore for large-dimensional data.

In the strict factor model case, [9] observed that the maximum likelihood estimator of the homoscedastic variance has a negative bias, and proposed an empirical correction. In Section 3, we give the bias and propose an unbiased estimator. Section 4 considers the goodness-of-fit test for the strict factor model: we propose a corrected likelihood ratio test to cope with the high-dimensional effects.
In the remaining Section 2, we introduce the definition of the strict factor model and the related maximum likelihood theory. Throughout the paper, simulation experiments are conducted to access the quality of the proposed estimation.

2 Problem formulation

2.1 The model

Let \( p \) denote the number of variables and \( n \) the sample size. In a general factor analysis model, the \( p \)-dimensional observation vectors \((x_i)_{1 \leq i \leq n}\) are of the form

\[
x_i = \sum_{k=1}^{m} f_{ki} \Lambda_k + e_i + \mu \quad (1)
\]

\[
= \Lambda f_i + e_i + \mu, \quad (2)
\]

where

- \( \mu \in \mathbb{R}^p \) represents the general mean;
- \( f_i = (f_{1i}, \ldots, f_{mi})' \) are \( m \) random factors \( (m < p) \);
- \( \Lambda = (\Lambda_1, \ldots, \Lambda_m) \) is the \( p \times m \) full rank matrix of factor loadings;
- \( e_i \) is a \( p \)-dimensional centered vector of noise, independent from \( f_i \) and with covariance matrix \( \Psi = \mathbb{E}(e_i e_i') \).

In order to remove indeterminacy and avoid identification problem in the model, commonly used restrictions are

- \( \mathbb{E}(f_i) = 0 \) and \( \mathbb{E}(f_i f_i') = I_p \);
- \( \Psi = \text{cov}(e_i) \) is diagonal;
- \( \Gamma = \Lambda' \Psi^{-1} \Lambda \) is diagonal.

Consequently, \textit{the population covariance matrix} \( \Sigma = \text{cov}(x_i) \) is \( \Sigma = \Lambda \Lambda' + \Psi \). In a strict factor model with homoscedastic variance, we assume in addition that \( \Psi = \sigma^2 I_p \), where \( \sigma^2 \in \mathbb{R} \) is the common variance of the noise \( e_i \). In this case, \( \Sigma = \Lambda \Lambda' + \sigma^2 I_p \) and has the spectral decomposition

\[
W' \Sigma W = \sigma^2 I_p + \text{diag}(\alpha_1, \ldots, \alpha_m, 0, \ldots, 0)
\]

where \( W \) is an unknown basis of \( \mathbb{R}^p \) and \( \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_m > 0 \). Let \( \bar{x} \) be the sample mean. The sample covariance matrix of the \( n \) \( p \)-dimensional i.i.d. vectors \((x_i)_{1 \leq i \leq n}\) is

\[
S_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'.
\]

We denote by \( \lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p} \) its eigenvalues.

2.2 Maximum likelihood estimators

If the \( f_i \) and \( e_i \) are Gaussian, a likelihood-based theory has been developed by [11]. The maximum likelihood estimator of \( \mu \) is \( \bar{x} \) and those of \( \Lambda \) and \( \sigma^2 \) are given by (see [3]):

\[
\hat{\sigma}^2 = \frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_i \quad \text{and} \quad \hat{\Lambda}_k = (\lambda_{n,k} - \hat{\sigma}^2)^{1/2} v_{n,k}, \ 1 \leq k \leq m,
\]

where \( v_{n,k} \) are the eigenvectors of the covariance matrix.
where $v_{n,k}$ is the normalized eigenvector of $S_n$ corresponding to $\lambda_{n,k}$, for $1 \leq k \leq p$.

In the classical setting where $p$ is kept fixed and small whereas the sample size $n \to \infty$, the almost sure convergence of these estimators is well-established. Nevertheless, this is no longer the case when $p$ is large compared to $n$.

### 2.3 CLT for LSS of a high-dimensional covariance matrix

We recall a proposition which will be useful in the sequel. Let $F_n = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_{n,i}}$ be the empirical spectral distribution (ESD) of $S_n$ and $F_{c,H}$ be the generalized Marchenko-Pastur distribution with indexes $(c,H)$. We consider the following empirical process

$$G_n(f) = p \int_{\mathbb{R}} f(x) [F_n - F_{c,H_n}] (dx), \; f \in \mathcal{A},$$

where $\mathcal{A}$ is the set of analytic functions $f : \mathbb{U} \to \mathbb{C}$, with $\mathbb{U}$ an open set of $\mathbb{C}$ such that $[1 \times (0,1) \times \mathbb{C}] \subset \mathbb{U}$. As $H_n = \Sigma \to \delta_{\sigma^2}$ and following [4], we have the following proposition which is a specialization of Theorem 9.10 of [5] (which covers more general matrices).

**Proposition 1.** Assume that $f_1, \ldots, f_k \in \mathcal{A}$ and the entries $x_{ij}$ of the vectors $(x_i)_{1 \leq i \leq n}$ are i.i.d. real random variables with mean 0, $\mathbb{E}(|x_{ij}|^4) = 3 \sigma^4$ and $\text{cov}(x_{ij}) = \Sigma = \Lambda \Lambda' + \sigma^2 I_p$. Then the random vector $(G_n(f_1), \ldots, G_n(f_k))$ converges to a $k$-dimensional Gaussian vector with given mean vector $m(f_j), j = 1, \ldots, k$ and covariance function $v(f_j, f_l), j, l = 1, \ldots, k$.

### 3 Estimation of the homoscedastic variance

#### 3.1 Central limit theorem for the estimator of the variance

As observed in [9, 10], in high-dimensional setting, the m.l.e. $\hat{\sigma}^2$ in (2.2) has a negative bias. The following theorem give this bias and show its asymptotic normality:

**Theorem 1.** We assume the same conditions of Proposition 1. Then, we have

$$\frac{(p-m)}{\sigma^2 \sqrt{2c}} (\hat{\sigma}^2 - \sigma^2) + b(\sigma^2) \overset{\mathcal{L}}{\to} \mathcal{N}(0,1),$$

where $b(\sigma^2) = \sqrt{c \sum_{i=1}^{m} \frac{1}{\alpha_i}}$.

#### 3.2 A bias-corrected estimator

The previous theory recommends to correct the negative bias of $\hat{\sigma}^2$. However, the bias $b(\sigma^2)$ depends on the number $m$ and the values $\alpha_i$ of the spikes. These parameters could not be known in real-data applications and they need to be first estimated. In the literature, consistent estimators of $m$ have been proposed, e.g. in [13, 14] and [9]. For the values of the spikes $\alpha_i$, it can be done by inverting their almost sure limit at the corresponding eigenvalues $\lambda_j$.

As the bias depends on $\sigma^2$ which we want to estimate, a natural correction is to use the plug-in estimator

$$\hat{\sigma}^2 = \frac{b(\hat{\sigma}^2) + \hat{\sigma}^2}{p-m}.$$

Using Theorem 1 and the delta-method, we obtain the following CLT
Theorem 2. We assume the same conditions of Proposition 1. Then, we have

\[ \tilde{v}(c)^{-\frac{1}{2}}(\tilde{\sigma}_c^2 - \sigma^2 + \tilde{b}(\sigma^2)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \]

where

- \( \tilde{b}(\sigma^2) = \frac{c\sqrt{2c\sigma^2}}{(p-m)^2}(mb(\sigma^2) + 2\sigma^2b(\sigma^2)\sum_{i=1}^{m} \alpha_i^{-1}) - \frac{2c^2\sigma^4b(\sigma^2)^2\sum_{i=1}^{m} \alpha_i^{-1}}{(p-m)^3} = O\left(\frac{1}{p^2}\right); \]

- \( \tilde{v}(c) = \frac{c^2\sigma^4}{(p-m)^2}\left(1 + \frac{cm}{p-m} + \frac{4c^2\sigma^4}{(p-m)^2}\sum_{i=1}^{m} \alpha_i^{-1}\right)^2 = v(c)\left(1 + O\left(\frac{1}{p}\right)\right)\).

\( \sigma_c^2 \) is still a biased estimator, but with a bias of order \( O\left(\frac{1}{p^2}\right) \) instead of \( O\left(\frac{1}{p}\right) \) for \( \hat{\sigma}^2 \).

3.3 Simulation experiments

We conduct some simulation experiments in three different settings and compare with two existing estimators of the common variance. Our estimator performs well.

4 Corrected likelihood ratio test of the hypothesis that the factor model fits

In this section we consider the following goodness-of-fit test for the strict factor model. The null hypothesis is then

\[ \mathcal{H}_0 : \Sigma = \Lambda\Lambda' + \sigma^2I_p, \]

where the number of factors \( m \) is specified.

Following [3], the likelihood ratio test (LRT) statistic is

\[ T_n = -nL^*, \text{ where } L^* = \sum_{j=m+1}^{p} \log \frac{\hat{\lambda}_{n,j}}{\sigma^2}, \]

and \( \hat{\sigma}^2 \) is the variance estimator (2.2). Keeping \( p \) fixed while letting \( n \to \infty \), then the classical theory states that \( T_n \) converges to \( \chi_q^2 \), where \( q = p(p+1)/2 + m(m-1)/2 - pm - 1 \), see [3]. However, this classical approximation is no more valid in the large-dimensional setting. Indeed, we will prove that this criterion leads to a high rate of false-alarm. In particular, the test becomes biased since the size will be much higher than the nominal level (see Table 1).

In a way similar to Section 3, we will construct a corrected version of \( T_n \) using Proposition 1 and calculus done in [4] and [17]. As we consider the logarithm of the eigenvalues of the sample covariance matrix, we will assume in the sequel that \( c < 1 \) to avoid null eigenvalues. We have the following theorem

Theorem 3. We assume the same conditions of Proposition 1, with \( c < 1 \), i.e. the entries \( x_{ij} \) of the vectors \( (x_i)_{1 \leq i \leq n} \) are i.i.d. real random variables with mean 0, \( \mathbb{E}(|x_{ij}|^4) = 3\sigma^4 \) and \( \text{cov}(x_i) = \Sigma = \Lambda\Lambda' + \sigma^2I_p \). Then, we have

\[ v(c)^{-\frac{1}{2}}(L^* - m(c) - ph(c_n) - \eta + (p-m)\log(\beta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \]

where \( m(c) = \frac{\log(1-c)}{2}, h(c_n) = \frac{c_n-1}{c_n} \log(1-c_n) - 1, \eta = \sum_{i=1}^{m} \log(1 + c\sigma^2\alpha_i^{-1}), \beta = 1 - \frac{c}{p-m}(m + \sigma^2\sum_{i=1}^{m} \alpha_i^{-1}) \) and \( v(c) = -2\log(1-c) + \frac{2c}{\beta}\left(\frac{1}{\beta} - 2\right) \).
To test $H_0$, we then can use the statistic
\[ v(c_n)^{-\frac{1}{2}} (L^* - m(c) - ph(c_n) - \eta + (p - m) \log(\beta)) \]
This test is asymptotically normal and will be hereafter referred as the corrected likelihood ratio test (CLRT in short).

4.1 Simulation experiments

We consider the following models:
- Model 1: $\text{spec}(\Sigma) = (25, 16, 9, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 4$, $c = 0.9$;
- Model 2: $\text{spec}(\Sigma) = (4, 3, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 2$, $c = 0.2$;
- Model 4: $\text{spec}(\Sigma) = (8, 7, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 1$, varying $c$.

Table 1 gives the realized sizes (i.e. the empirical probability of rejecting the null hypothesis) of the classical likelihood ratio test (LRT) and the corrected likelihood ratio test (CLRT) proposed above. The computations are done under 10000 independent replications and the nominal test level is 0.05.

<table>
<thead>
<tr>
<th>Settings</th>
<th>Realized size of CLRT</th>
<th>Realized size of LRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 90$</td>
<td>$n = 100$</td>
<td>0.0497 0.9995</td>
</tr>
<tr>
<td>$p = 180$</td>
<td>$n = 200$</td>
<td>0.0491 1</td>
</tr>
<tr>
<td>$p = 720$</td>
<td>$n = 800$</td>
<td>0.0496 1</td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 20$</td>
<td>$n = 100$</td>
<td>0.0324 0.0294</td>
</tr>
<tr>
<td>$p = 80$</td>
<td>$n = 400$</td>
<td>0.0507 0.0390</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>$n = 1000$</td>
<td>0.0541 0.0552</td>
</tr>
<tr>
<td>Model 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>$n = 500$</td>
<td>0.0108 0.0483</td>
</tr>
<tr>
<td>$p = 10$</td>
<td>$n = 500$</td>
<td>0.0190 0.0465</td>
</tr>
<tr>
<td>$p = 50$</td>
<td>$n = 500$</td>
<td>0.0424 0.0445</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>$n = 500$</td>
<td>0.0459 0.0461</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>$n = 500$</td>
<td>0.0491 0.2212</td>
</tr>
<tr>
<td>$p = 250$</td>
<td>$n = 500$</td>
<td>0.0492 0.7395</td>
</tr>
<tr>
<td>$p = 300$</td>
<td>$n = 500$</td>
<td>0.0509 0.9994</td>
</tr>
</tbody>
</table>

The sizes of our new CLRT are close to the theoretical one, except when the ratio $c = p/n$ is small (less than 0.1). On the contrary, the sizes produced by the classical LRT are much higher than the nominal level when $c$ is going close to one, and the test will always be rejected when $p$ is large.

References


