Abstract. Estimation of the number of factors in a factor model is an important problem in many areas such as economics or signal processing. Most of classical approaches assume a large sample size \( n \) whereas the dimension \( p \) of the observations is kept small. In this paper, we consider the case of high dimension, where \( p \) is large compared to \( n \). The approach is based on recent results of random matrix theory. We extend our previous results to a more difficult situation when some factors are equal, and compare our algorithm to an existing benchmark method.

1. Introduction

The factor model appears in many scientific fields, such as economics and psychology literature, where the number of factors has a primary importance ([1], [19]). Similar models can be found in physics of mixture [9], [15] or population genetics. In wireless communications, a signal (factor) emitted by a source is modulated and received by an array of antennas which will permit the reconstruction of the original signal. More recently spiked population models has been introduced in [8] that encompass factor models.

A fundamental problem here is the determination of the number of factors. Many methods have been developed, mostly based on information theoretic criteria, such as the minimum description length (MDL) estimator, Bayesian model selection or Bayesian Information Criteria (BIC) estimators, see [21] for a review. Nevertheless, these methods are based on asymptotic expansions for large sample size and may not perform well when the dimension of the data \( p \) is large compared to the sample size \( n \). To our knowledge, this problem in the context of high-dimension appears for the first time in [5]. Recent advances have been made using the random matrix theory by [7] or Onatski [16] in economics, and Kritchman & Nadler in chemometrics literature [9].

Several studies have also appeared in the area of signal processing from high-dimensional data. Everson & Roberts [6] proposed a method using both RMT and bayesian inference, while Ulfarsson & Solo combined RMT and Stein’s unbiased risk estimator (SURE) in [20]. In [13] and [14], the authors improved estimators based on information theoretic criteria.

Date: February 29, 2012.
and in [10], Kritchman & Nadler constructed an estimator based on the distribution of the largest eigenvalue (hereafter refereed as the KN estimator). In [17], we have also introduced a new method based on recent results of [2] and [18] in random matrix theory. It is worth mentioning that for high-dimensional time series, an empirical method for the estimation of factor number has been recently proposed in [11] and [12].

In most cited references, factors are assumed to be distinct. However, we observe that when some of these factors become close, the estimation problem becomes more difficult and these algorithms need to be modified. We refer this new situation as the case with possibly equal factors and its precise formulation will be given in Section 3.2. The aim of this work is to extend our method [17] to this new case and to compare it with the KN estimator, that is known in the literature as one of best estimation method.

The rest of the paper is organized as follows. Section 2 introduces the model. In Section 3, we define the estimation problem of the number of possibly equal factors and present our solution. We establish its asymptotic consistency. Section 4 provides simulation experiments to access the quality of our estimator. Next, we recall the KN estimator and conduct simulation experiments to compare these two methods. In Section 6, we analyze the influence of a tuning parameter \( C \) used in our estimator. Finally, Section 7 concludes with discussions. All proofs are given in the appendix.

2. Problem Formulation

We consider the following strict factor model

\[
\begin{align*}
\mathbf{x}(t) &= \sum_{k=1}^{q_0} f_k(t)a_k + \sigma n(t) \\
&= A f(t) + \sigma n(t),
\end{align*}
\]

where

- \( f(t) = (f_1(t), \ldots, f_{q_0}(t))^* \in \mathbb{R}^{q_0} \) are \( q_0 \) random factors (\( q_0 < p \)) assumed to have zero mean, unit variance and be mutually uncorrelated;
- \( A = (a_1, \ldots, a_{q_0}) \) is the \( p \times q_0 \) full rank matrix of factor loadings;
- \( \sigma \in \mathbb{R} \) is the unknown noise level, \( n(t) \sim \mathcal{N}(0, I_p) \) is a \( p \times 1 \) vector of additive noise, independent of \( f(t) \).

The population covariance matrix \( \Sigma = \text{cov}(\mathbf{x}(t)) \) of \( \mathbf{x}(t) \) equals \( AA^* + \sigma^2 I_p \) and has the spectral decomposition

\[
W^* \Sigma W = \sigma^2 I_p + \text{diag}(\alpha_1, \ldots, \alpha_{q_0}, 0, \ldots, 0)
\]
where $W$ is an unknown basis of $\mathbb{R}^p$ and $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{q_0} > 0$. The sample covariance matrix of the $n$ $p$-dimensional i.i.d. vectors considered at each time $t$, $(x_i = x(t_i))_{1 \leq i \leq n}$ is

$$S_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^*.$$ 

Denote by $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ its eigenvalues. Our aim is to estimate $q_0$ on the basis of $S_n$. To start with, we assume that the noise level $\sigma^2$ is known. If this is indeed not the case, we will give a method in Section 3.3 to estimate it.

### 3. Estimation of the number of factors

In this section, we first recall our previous result of [17] in the case of different factors. Next, we propose an extension of the algorithm to the case with possibly equal factors. The consistency of the extended algorithm is established.

#### 3.1. Previous work: estimation with different factors

We consider the case where the $(\alpha_i)_{1 \leq i \leq q_0}$ are all different, so there are $q_0$ distinct factors. According to [17], let us rewrite the spectral representation of $\Sigma$ as

$$W^* \Sigma W = \sigma^2 \text{diag}(\alpha'_1, \ldots, \alpha'_q, 1, \ldots, 1),$$

with

$$\alpha'_i = \frac{\alpha_i}{\sigma^2} + 1.$$

It is assumed in the sequel that $p$ and $n$ are related so that when $n \to +\infty$, $p/n \to c > 0$. Therefore, $p$ can be large compared to the sample size $n$ (high-dimensional case).

Moreover, we assumed that $\alpha'_1 > \cdots > \alpha'_q > 1 + \sqrt{c}$ for all $i \in \{1, \ldots, q_0\}$, i.e all the factors $\alpha$ are greater than $\sigma^2 \sqrt{c}$. For $\alpha \neq 1$, we define the function

$$\phi(\alpha) = \alpha + \frac{c\alpha}{\alpha - 1}.$$ 

Baik and Silverstein [3] proved that, under a moment condition on $x$, for each $k \in \{1, \ldots, q_0\}$ and almost surely,

$$\lambda_{n,k} \to \sigma^2 \phi(\alpha'_k).$$

They also proved that for all $1 \leq i \leq L$ with a prefixed range $L$ and almost surely,

$$\lambda_{n,q_0+i} \to b = \sigma^2 (1 + \sqrt{c})^2.$$ 

The estimation method of $q_0$ in [17] is based on a close inspection of differences between consecutive eigenvalues

$$\delta_{n,j} = \lambda_{n,j} - \lambda_{n,j+1}; \ j \geq 1.$$
Indeed, the results quoted above imply that a.s. \( \delta_{n,j} \to 0 \), for \( j \geq q_0 \) whereas for \( j < q_0 \), \( \delta_{n,j} \) tends to a positive limit. Thus it becomes possible to estimate \( q_0 \) from index-numbers \( j \) where \( \delta_{n,j} \) become small. More precisely, the estimator is
\[
(3) \quad \hat{q}_n = \min\{j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n\},
\]
where \( s > q_0 \) is a fixed number big enough, and \( d_n \) is a threshold to be defined. In practice, the integer \( s \) should be thought as a preliminary bound on the number of possible factors. In [17], we proved the consistency of \( \hat{q}_n \) providing that the threshold satisfies \( d_n \to 0 \), \( n^{2/3}d_n \to +\infty \) and under the following assumption on the entries of \( \mathbf{x} \).

**Assumption 1.** The entries \( x^i \) of the random vector \( \mathbf{x} \) have a symmetric law and a sub-exponential decay, that means there exists positive constants \( C, C' \) such that, for all \( t \geq C' \),
\[
P(|x^i| \geq t^C) \leq e^{-t}.
\]

### 3.2. Estimation with possibly equal factors.

As said in Introduction, when some factors have close values, estimation algorithms need to be modified. More precisely, we adopt the following theoretic model with \( K \) different factor strengths \( \alpha_1, \ldots, \alpha_K \), each of them can appear \( n_k \) times (equal factors), respectively. In other words,
\[
\text{spec}(\Sigma) = (\alpha_1, \ldots, \alpha_1, \ldots, \alpha_K, \ldots, \alpha_K, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)
\]
\[
= \sigma^2(\alpha_1', \ldots, \alpha_1', \ldots, \alpha_K', \ldots, \alpha_K', 1, \ldots, 1),
\]
with \( n_1 + \cdots + n_K = q_0 \). When all the factors are unequal, differences between sample factor eigenvalues tend to a positive constant, whereas with two equal factors, such difference will tend to zero. This fact creates an ambiguity with those differences corresponding to the noise eigenvalues which also tend to zero. However, the convergence of the \( \delta_{n,i} \)'s, for \( i > q_0 \) (noise) is faster (in \( O_p(n^{-2/3}) \)) than that of the \( \delta_{n,j} \) from equal factors (in \( O_p(n^{-1/2}) \)) as a consequence of Theorem 3.1 of Bai & Yao [2]. This is the key feature we use to adapt the estimator (3) to the current situation with a new threshold \( d_n \). The precise asymptotic consistency is as follows.

**Theorem 1.** Let \( (x_i)_{1 \leq i \leq n} \) be \( n \) copies i.i.d. of \( \mathbf{x} \) which follows the model (3) and satisfies Assumption 7. Suppose that the population covariance matrix \( \Sigma \) has \( K \) non null and non unit eigenvalues \( \alpha_1 > \ldots > \alpha_K > \sigma^2 \sqrt{c} \) with respective multiplicity \( (n_k)_{1 \leq k \leq K} \) \((n_1 + \cdots + n_K = q_0)\), and \( p - q_0 \) unit eigenvalues. Assume that \( \frac{c}{n} \to c > 0 \) when \( n \to +\infty \). Let \( (d_n)_{n \geq 0} \) be a real sequence such that \( d_n = o(n^{-1/2}) \) and \( n^{2/3}d_n \to +\infty \). Then the estimator \( \hat{q}_n \) is consistent, i.e \( \hat{q}_n \to q_0 \) in probability when \( n \to +\infty \).

Notice that, compare to the previous situation, the only modification of our estimator is a new condition \( d_n = o(n^{-1/2}) \) on the convergence rate of \( d_n \). The proof of Theorem 1 is postponed to Appendix.
3.3. **Estimation of the noise level.** When the noise level $\sigma^2$ is unknown, an estimation is needed. In [17], we used an algorithm based on the maximum likelihood estimate

$$\hat{\sigma}^2 = \frac{1}{p - q_0} \sum_{i=q_0+1}^{p} \lambda_{n,i}.$$ 

As explained in [9] and [10], this estimator has a negative bias. Hence the authors developed an improved estimator with a smaller bias. We will use this improved estimator of noise level in our simulations for both estimator $\tilde{q}_n$ and the estimator $\tilde{q}_n$ (see Section 5).

4. **Simulation experiments**

To access the quality of our estimator, we first make the following modification: instead of making a decision once some difference $\delta_k$ is below the threshold $d_n$ (see (3)), the modified estimator stops when two consecutive differences $\delta_k$ and $\delta_{k+1}$ are both below $d_n$. More precisely, we set

$$\hat{q}_n^* = \min \{ j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n \text{ and } \delta_{n,j+2} < d_n \}.$$ (4)

It is easy to see that the proof for the consistency of $\hat{q}_n$ applies equally to $\hat{q}_n^*$ under the same conditions as in Theorem 1.

It remains to choose a threshold sequence $d_n$ to be used for our estimator $\hat{q}_n^*$. As argued in [17], we use a sequence $d_n$ of the form $Cn^{-2/3}/\sqrt{2\log \log n}$, where $C$ is a “tuning” parameter to be adjusted. In all simulations, we consider 500 independent replications and take $\sigma^2 = 1$.

Table 1 gives a summary of parameters in our simulation experiments. There are two sets of experiments. In the first one (Figures 1, 2 and Models A, B, C and D in Table 1), factors are different and these experiments extend and complete results already reported in [17]. The second set of experiments (Figures 3, 4 and Models E, F, G, H and J in Table 1) addresses the new situation where some factors are equal. Figure 7 considers the case of no factor. (Figures 5 and 6 report comparison results developed in Section 5).

4.1. **Case of different factors.** In Figure 1, we consider the case of a single factor of strength $\alpha$, and we analyze the probability of misestimation as a function of factor strength $\alpha$, for $(p, n) = (200, 800)$, $c = 0.25$ and $(p, n) = (2000, 500)$, $c = 4$. We set $C = 5.5$ for the first case and $C = 9$ for the second case. The noise level $\sigma^2 = 1$ is given.

Our estimator performs well: we recover the threshold from which the behavior of the factor eigenvalues differ from the noise ones ($\sqrt{c} = 0.5$ for the first case, and 2 for the second).
Table 1. Summary of parameters used in the simulation experiments. (L: left, R: right)

| No. | Fig. | Factors Mod. | Factor values | Fixed parameters | Var. |
|-----|------|--------------|---------------|------------------|------|
|     | No.  | No.          |               |                  |      |
| 1   | Different | (α) | (200, 800) | 1/4 | Given | 5.5 | α |
|     |       |                | (2000, 500) | 4 |       |     |
| 2L  | Different | A | (6,5) | 10 | Given | 11 | n |
|     |       | B | (10,5) | Estimated |       |     |
| 2R  | Different | C | (1.5) | 1 | Given | 5 | n |
|     |       | D | (2.5,1.5) | Estimated |       |     |
| 3   | Possibly | E | (α,α,5) | (200, 800) | 1/4 | Given | 6 | α |
|     | equal  | F | (α,α,15) | (2000, 500) | 4 |       |     |
| 4L  | Possibly | G | (6,5,5) | 10 | Given | 9.9 | n |
|     | equal  | H | (10,5,5) | Estimated |       |     |
| 4R  | Possibly | I | (1.5,1.5) | 1 | Given | 5 | n |
|     | equal  | J | (2.5,1.5,1.5) |       |       |     |
| 5   | Models A and D | | | | | |
| 6   | Models G and J | | | | | |
| 7   | No factor | K | No factor | 1 | 10 | Given | 8 | 15 | n |
| 8L  | Models A and G | | | | | |
| 8R  | Models B and H | | | | | |
| 9L  | Models C and I, with C automatically chosen | | | | | |
| 9R  | Models D and J, with C automatically chosen | | | | | |

Figure 1. Misestimation rates as a function of factor strength for \((p, n) = (200, 800)\) and \((p, n) = (2000, 500)\).
In Figure 2 left panel, we consider two models with two factors \((q_0 = 2)\), in three situations:

- **Model A**: \((\alpha_1, \alpha_2) = (6, 5)\) and \(\sigma^2 = 1\) is given;
- **Model B**: \((\alpha_1, \alpha_2) = (10, 5)\) and \(\sigma^2 = 1\) is given;
- **Model B**: \((\alpha_1, \alpha_2) = (10, 5)\) and \(\sigma^2 = 1\) is to be estimated;

The estimation is harder in Model A as the factor have closer strengths. We fix \(c = 10\) \((p \gg n)\), and we plot the misestimation rates against the sample size \(n\). Here \(C = 11\).

**Figure 2.** Misestimation rates as a function of \(n\) for Models A, B (left) and Model C, D (right).

As expected, our estimator performs better in Model B than in Model A. In both cases, we observe the asymptotic consistency. Compared to Model B with \(\sigma^2\) given, the estimation of \(\sigma^2\) does not affect our estimator significantly, which seems robust against the unknown noise level.

Figure 2 right panel considers two cases with \(c = 1\) and a given noise level \(\sigma^2 = 1\):

- **Model C**: \((\alpha) = (1.5)\);
- **Model D**: \((\alpha_1, \alpha_2) = (2.5, 1.5)\).

This experiment is designed with factor strengths close to the critical value \(\sqrt{c} = 1\). Thus the problem becomes more difficult and misestimation rates are higher than in the left panel. Here we used \(C = 5\).
4.2. **Case with equal factors.** We keep the same parameters as in the previous section while adding some equal factors. This leads to results reported in Figures 3 and 4 which are to be compared to Figures 1 and 2. In Figure 3, we consider

- Model E: \((\alpha_1, \alpha_2, \alpha_3) = (\alpha, \alpha, 5), 0 \leq \alpha \leq 2.5;\)
- Model F: \((\alpha_1, \alpha_2, \alpha_3) = (\alpha, \alpha, 15), 0 \leq \alpha \leq 8;\)

with \((p, n) = (200, 800)\) for the Model E and \((p, n) = (2000, 500)\) for the Model F. Here \(q_0 = 3, C = 6\) for Model E and \(C = 9.9\) for Model F.

![Figure 3](image-url)  
**Figure 3.** Misestimation rates as a function of factor strength for \((p, n) = (200, 800),\) Model E and \((p, n) = (2000, 500),\) Model F.

In Figure 4 left panel, we consider two models, analog to Model A and B, with three factors \((q_0 = 3):\)

- Model G: \((\alpha_1, \alpha_2, \alpha_2) = (6, 5, 5)\) and \(\sigma^2 = 1\) is given;
- Model H: \((\alpha_1, \alpha_2, \alpha_2) = (10, 5, 5)\) and \(\sigma^2 = 1\) is given;
- Model H: \((\alpha_1, \alpha_2, \alpha_2) = (10, 5, 5)\) and \(\sigma^2 = 1\) is to be estimated.

Again we fix \(c = 10 (p \gg n),\) and we plot misestimation rates against the sample size \(n.\) Here \(C = 9.9\) and \(\sigma^2\) is given. Comparing to the case of different factors (Figure 2), these rates are significantly higher with however a clear and rapidly decreasing trend. If we compare Model G and Model H, a smaller spacing between two first factors deteriorates the algorithm only slightly. Moreover in Model H and similar to Figure 2, estimation of an unknown variance \(\sigma^2\) does not affect our estimator significantly.
Figure 4. Misestimation rates as a function of $n$ for Model G, H (left) and Model I, J (right).

Figure 4 right panel considers two cases with $c = 1$, $\sigma^2 = 1$ given and factor strengths close to the critical value $\sqrt{c}$:

- Model I: $(\alpha, \alpha) = (1.5, 1.5)$;
- Model J: $(\alpha_1, \alpha_2, \alpha_2) = (2.5, 1.5, 1.5)$.

Here we use $C = 5$. In this more difficult situation, misestimation rates vanish much more slowly than in the left panel.

In summary, these experiments have demonstrated the proposed estimator is able to find the number of factors in all the considered situations. In particular, when factor strengths are close or even equal, or close to the critical value, the algorithm remains consistent although the convergence rate becomes slower.

5. Method of Kritchman & Nadler and comparison

5.1. Algorithm of Kritchman & Nadler. In their paper [9] and [10], these authors develop a different method also based on random matrix theory to estimate the number of factors. In this section we compare by simulation our estimator (PY) to the Kritchman & Nadler’s one (KN). The authors have compared their estimator KN with existing estimators in the signal processing literature, based on the minimum description length (MDL), Bayesian information criterion (BIC) and Akaike information criterion (AIC) [21]. In most of the studied cases, the estimator KN performs better. Furthermore in [14], this estimator
is compared to an improved AIC estimator and it still has a better performance. Thus we
decide to consider only this estimator KN for the comparison here.

In the absence of factors, $nS_n$ follows a Wishart distribution with parameters $n, p$. In
this case, Johnstone \cite{Johnstone2001} has provided the asymptotic distribution of the largest eigenvalue
of $S_n$.

**Proposition 1.** Let $S_n$ be the sample covariance matrix of $n$ vectors distributed as $N(0, \sigma^2 I_p)$, and $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ be its eigenvalues. Then, when $n \to +\infty$, such that $\frac{p}{n} \to c > 0$

$$P \left( \frac{\lambda_{n,1}}{\sigma^2} < \frac{\beta_{n,p}}{n^{2/3}} s + b \right) \to F_1(s), \ s > 0$$

where $b = (1+\sqrt{c})^2$, $\beta_{n,p} = (1 + \sqrt{\frac{p}{n}}) \left( 1 + \sqrt{\frac{p}{n}} \right)^{\frac{1}{3}}$ and $F_1$ is the Tracy-Widom distribution
of order 1.

Assume the variance $\sigma^2$ is known. To distinguish a factor eigenvalue $\lambda$ from a noise one
at an asymptotic significance level $\gamma$, their idea is to check whether

\begin{equation}
\lambda_{n,k} > \sigma^2 \left( \frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b \right)
\end{equation}

where $s(\gamma)$ verifies $F_1(s(\gamma)) = 1 - \gamma$ and can be found by inverting the Tracy-Widom distribution. This distribution has no explicit expression, but can be computed from a solution of a second order Painlevé ordinary differential equation. The estimator KN is based on a
sequence of nested hypothesis tests of the following form: for $k = 1, 2, \ldots, \min(p, n) - 1$,

$$H_0^{(k)}: q_0 \leq k - 1 \quad \text{vs.} \quad H_1^{(k)}: q_0 \geq k$$

For each value of $k$, if \cite{5} is satisfied, $H_0^{(k)}$ is rejected and $k$ is increased by one. The
procedure stops once an instance of $H_0^{(k)}$ is accepted and the number of factors is then
estimated to be $\hat{q}_n = k - 1$. Formally, their estimator is defined by

$$\hat{q}_n = \arg\min_k \left( \lambda_{n,k} < \hat{\sigma}^2 \left( \frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b \right) \right) - 1.$$ 

Here $\hat{\sigma}$ is some estimator of the noise level (as discussed in Section \cite{3.3}). The authors
proved the strong consistency of their estimator as $n \to +\infty$ with fixed $p$, by replacing the
fixed confidence level $\gamma$ with a sample-size dependent one $\gamma_n$, where $\gamma_n \to 0$ sufficiently
slow as $n \to +\infty$. They also proved that $\lim_{p, n \to +\infty} P(\hat{q}_n \geq q_0) = 1$.

It is important to notice here that the construction of the KN estimator differs form
ours, essentially because of the fixed alarm rate $\gamma$. We will discuss the issue of the false
alarm rate in the last section.
5.2. **Comparison with our method.** We give a value of $\gamma = 0.5\%$ to the false alarm rate of the estimator KN, as suggested in [10] and use their algorithm available at the author’s homepage.

In Figure 5, we consider Model A and Model D as previously:

- Model A: $(\alpha_1, \alpha_2) = (6, 5)$;
- Model D: $(\alpha_1, \alpha_2) = (2.5, 1.5)$.

We keep the same constant $C$ and $\sigma^2 = 1$ is given to both estimators.

![Figure 5](image)

**Figure 5.** Misestimation rates as a function of $n$ for Model A (left) and Model D (right).

For Model A, the performances of the two estimator are close. However the estimator PY is slightly better for moderate values of $n$ ($n \leq 400$) while the estimator KN has a slightly better performance for larger $n$. For Model B, our algorithm has a lower misestimation rate in almost all cases in both models, with an improvement ranging from 10% to 30% for moderate sample sizes $n \leq 400$.

Figure 6 considers Model G and J, two models analogous to Model A and D but with two equal factors:

- Model G: $(\alpha_1, \alpha_2, \alpha_2) = (6, 5, 5)$;
- Model J: $(\alpha_1, \alpha_2, \alpha_2) = (2.5, 1.5, 1.5)$.

Again we keep the same constant $C$ and $\sigma^2 = 1$ is given to both estimators.
For Model G, the estimator PY shows superior performance for $n \leq 500$ (up to 20% less error): adding an equal factor affects more the performance of the estimator KN. The difference between the two algorithms for Model J is higher than in the previous cases: the estimator PY performs better, up to 25%.

In Figure 7 we examine a special case with no factor at all (Model K). The estimation rates become the so-called false-alarm rate, a concept widely used in signal processing literature. The cases of $c = 1$ and $c = 10$ with $\sigma^2 = 1$ given are considered.

We chose $C = 8$ for the first case and $C = 15$ for the second case. In both situations, false alarm rates of two estimators are similar and low (less than 1%), and the KN one has a slightly better performance.

In summary, in most of situations reported here, our algorithm compares quite favorably to an existing benchmark method (the KN estimator). It is also important to notice a fundamental difference between these two estimators: the KN estimator is designed to keep the false alarm rate as a very low level while our estimator attempts to minimize an overall misestimation rate. We develop more in details these issues in next section.

6. **On the tuning parameter $C$**

6.1. **Influence of $C$ on the misestimation and false alarm rate.** In the simulation experiments, we chose the constant $C$ “by hand” to have the lowest misestimation rate. However, to have a fair comparison to either the KN estimator or any other method of the
Figure 7. Misestimation rates as a function of $n$ in the case of no factor for $c = 1$ (left) and $c = 10$ (right).

number of factors, the different methods should have comparable false alarm probabilities. This section is devoted to an analysis of possible relationship between the constant $C$ and the implied false alarm rate. Following [10], the false alarm rate $\gamma$ of such an algorithm can be viewed as the type I error of the following test

$$\mathcal{H}_0: q_0 = 0 \quad \text{vs.} \quad \mathcal{H}_1: q_0 > 0,$$

that is the probability of overestimation in the white case. Recall the step $k$ of the algorithm KN tests

$$\mathcal{H}^{(k)}_0: q_0 \leq k - 1 \quad \text{vs.} \quad \mathcal{H}^{(k)}_1: q_0 \geq k.$$ 

In [10], the authors argue that their threshold is determined such that

$$\mathbb{P}(\text{reject } \mathcal{H}^{(k)}_0 | \mathcal{H}^{(k)}_0) \approx \gamma.$$ 

More precisely, they give an asymptotic bound of the overestimation probability: they show that for $n = 500$ and $p > 10$, this probability is close to $\gamma$.

Since for our method, we do not know explicitly the corresponding false alarm rate, we evaluate it by simulation. We choose two typical situations among previously reported ones, namely Figure 2 (see Table 1). Table 2 gives the results with 500 independent replications.

The false alarm rates of our algorithm are much higher than the false alarm rate $\gamma = 0.5\%$ of the KN estimator, especially for the case with $C = 5$ and $c = 1$. Nevertheless, and contrary to the KN estimator, the overestimation rate of our estimator will be different from the false alarm rate, and will depend on the number of factors and their values.
Indeed, we use the gaps between two eigenvalues, instead of each eigenvalue separately. Consequently, there is no justification to claim that the probability \( P(\hat{q}_n > q | q = q_0) \), for \( q_0 > 1 \) will be close to \( P(\hat{q}_n > 0 | q = 0) \). To illustrate this phenomenon, we use the settings of Model B \((q_0 = 2)\) and J \((q_0 = 3)\) and we evaluate the overestimation rate using 500 independent replications (note that the corresponding false alarm rates are those in Table 2). The results are displayed in Table 3.

| (p,n)       | (150,150) | (300,300) | (500,500) | (700,700) |
|------------|-----------|-----------|-----------|-----------|
| \( C = 5, c = 1 \) (Fig. 2R) | 0.124     | 0.098     | 0.078     | 0.086     |
| (p,n)      | (1500,150) | (3000,300) | (5000,500) | (7000,700) |
| \( C = 11, c = 10 \) (Fig. 2L) | 0.046     | 0.04      | 0.048     | 0.024     |

TABLE 3. Empirical overestimation rates from Model B \((\alpha = (10,5), c = 10, C = 11)\) and Model J \((\alpha = (2.5,1.5,1.5), c = 1, C = 5)\).

We observe that these overestimation rates are lower than the false alarm rates given in Table 2: this confirms that no obvious relationship exists between the false alarm rate \( \gamma \) and the overestimation rates for our algorithm.

Furthermore, we can easily see that when \( C \) increase, overestimation rates will decrease but underestimation rates will then increase. It explains also why we had to use in Model K (no factor) a constant \( C \) greater than in the other model with the same ratio \( c_n = p/n \).

In summary, if the goal is to keep overestimation rates at a constant and low level, one should employ the KN estimator without hesitation (since by construction, the probability of overestimation is kept to a very low level). Otherwise, if the goal is also to minimize the overall misestimation rates i.e. including underestimation errors, our algorithm can be a good substitute to the KN estimator. One could think of choosing \( C \) in each case to have a probability of overestimation kept fixed at a low level, but in this case the probability of underestimation would be high and the performance of the estimation would be poor: our estimator is constructed to minimize the overall misestimation rate.

6.2. On the choice of \( C \). The tuning parameter \( C \) was chosen from case to case in previous experiments. We now provide an automatic calibration of this parameter. The idea is to use the difference of the two largest eigenvalues of a Wishart matrix (which correspond to the case of no factor): indeed, our algorithm stop once two consecutive eigenvalues are below the threshold \( d_n \) corresponding to a noise eigenvalue. As we do not
know precisely the distribution of the difference between eigenvalues of a Wishart matrix, we approximate the distribution of the difference between the two largest eigenvalues $\hat{\lambda}_1 - \hat{\lambda}_2$ by simulation under 500 independent replications. We then take the mean $s$ of the 10th and the 11th largest spacings, so $s$ has the empirical probability $\mathbb{P}(\hat{\lambda}_1 - \hat{\lambda}_2 \leq s) = 0.98$: this value will give reasonable results. We calculate a $\tilde{C}$ by multiplying this threshold by $n^{2/3} / \sqrt{2 \times \log \log(n)}$. The result for various $(p, n)$, with $c = 1$ and $c = 10$ are displayed in Table 4.

**Table 4.** Approximation of the threshold $s$ such that $\mathbb{P}(\hat{\lambda}_1 - \hat{\lambda}_2 \leq s) = 0.98$.

| $(p,n)$       | (200,200) | (400,400) | (600,600) | (2000,200) | (4000,400) | (7000,700) |
|---------------|-----------|-----------|-----------|------------|------------|------------|
| Value of $s$  | 0.340     | 0.223     | 0.170     | 0.593      | 0.415      | 0.306      |
| $\tilde{C}$   | 6.367     | 6.398     | 6.277     | 11.106     | 11.906     | 12.44      |

The values of $\tilde{C}$ are quite close to the values used in previous simulations experiment ($C = 5$ for $c = 1$ and $C = 9.9$ or 11 for $c = 10$), although they are slightly higher. Therefore, this automatic calibration of $\tilde{C}$ can be used in practice for any data and sample dimensions $p$ and $n$.

To assess the quality of this automatic calibration procedure, we run again a part of the previous simulation experiments this time using $\tilde{C}$. Figure 8 considers the case where $c = 10$. On the left we consider Model A ($\alpha = (6, 5)$) and Model G ($\alpha = (6, 5, 5)$) (upper curve). On the right we have Model B ($\alpha = (10, 5)$) and Model H ($\alpha = (10, 5, 5)$) (upper curve). The dots lines are the previous results with $C$ manually chosen.

**Figure 8.** Misestimation rates as a function of $n$ for Models A, G (left) and Models B, H (right).
Using the new automatically method causes only a slight deterioration of the estimation performance. We again observe significantly higher error rates in the case of equal factors for moderate sample sizes.

Figure 9 considers the case where \( c = 1 \), with Models C (\( \alpha = 1.5 \)) and I (\( \alpha = (1.5,1.5) \)) (upper curve) on the left and Model D (\( \alpha = (2.5,1.5) \)) and J (\( \alpha = (2.5,1.5,1.5) \)) (upper curve) on the right.

Compared to the previous situation of \( c = 10 \), using the automatic value \( \hat{C} \) affects a bit more our estimator (up to 10% of degradation). Nevertheless, the estimator remains consistent. Furthermore, we have to keep in mind that our simulation experiments have considered critical cases where factors eigenvalues are close: in many of practical applications, theses factors are more separated so that the influence of \( C \) will be less important.

7. Concluding remarks

In this paper we have considered the problem of the estimation of the number of factors in the high-dimensional case. When some factors have close or even equal values, the estimation becomes harder and existing algorithm need to be re-examined or corrected. In this spirit, we have proposed a new version of our previous algorithm. Its asymptotic consistency is established. It becomes unavoidable to compare our algorithm to an existing competitor proposed by Kritchman & Nadler (KN, [9], [10]). From our extensive simulation experiments in various scenarios, we observe that overall our estimator could have smaller
misestimation rates, especially in cases with close and relatively low factor values (Figures 2 and 4) or more generally for almost all the cases provided that the sample size $n$ is moderately large ($n \leq 400$ or 500). Nevertheless, if the primary aim is to fix the false alarm rate and the overestimation rates at a very low level, the KN estimator is preferable.

However, our algorithm depend on a tuning parameter $C$. Most of the experiments reported here are obtained with a finely-turned value of $C$ and this value varies from case to case. By comparison, the KN estimator is remarkably robust and a single value of $\gamma = 0.5\%$ was used in all the experiments. In Section 6, we have provided a first approach to an automatic calibration of $C$ which is quite satisfactory. However, more investigation is needed in the future on this issue.

Appendix

In the sequel, we will assume that $\sigma^2 = 1$ (if it is not the case, we consider $\frac{\lambda_{n,j}}{\sigma^2}$). For the proof, we need two theorems. The first, Proposition 2, is a result of Bai and Yao [2] which derives from a CLT for the $n_k$-packed eigenvalues

$$\sqrt{n}(\lambda_{n,j} - \phi(\alpha'_k)), j \in J_k$$

where $J_k = \{s_k-1 + 1, \ldots, s_k\}$, $s_i = n_1 + \ldots + n_i$ for $1 \leq i \leq K$.

**Proposition 2.** Assume that the entries $x^i$ of $x$ satisfy $\mathbb{E}(||x^i||^4) < +\infty$, $\alpha'_j > 1 + \sqrt{c}$ for all $1 \leq j \leq K$ and have multiplicity $n_1, \ldots, n_K$ respectively. Then as $p, n \to +\infty$ so that $\frac{p}{n} \to c$, the $n_k$-dimensional real vector

$$\sqrt{n}\{\lambda_{n,j} - \phi(\alpha'_k), j \in J_k\}$$

converges weakly to the distribution of the $n_k$ eigenvalues of a Gaussian random matrix whose covariance depends on $\alpha'_k$ and $c$.

The second Proposition 3 is issued from the Proposition 5.8 of [4]:

**Proposition 3.** Assume that the entries $x^i$ of $x$ have a symmetric law and a sub-exponential decay, that means there exists positive constants $C$, $C'$ such that, for all $t \geq C'$, $\mathbb{P}(|x^i| \geq t^C) \leq e^{-t}$. Then, for all $1 \leq i \leq L$ with a prefixed range $L$,

$$n^2(\lambda_{n,q} - b) = O_P(1).$$

We also need the following lemma:

**Lemma 1.** Let $(X_n)_{n \geq 0}$ be a sequence of positive random variables which converges weakly. Then for all real sequence $(u_n)_{n \geq 0}$ which converges to 0,

$$\mathbb{P}(X_n \leq u_n) \to 0.$$
Proof. As \((X_n)_{n \geq 0}\) converges weakly, it exists a function \(G\) such that, for all \(v \geq 0\), 
\[ P(X_n \leq v) \to G(v). \]
Furthermore, as \(u_n \to 0\), it exists \(N \in \mathbb{N}\) such that for all \(n \geq N\), 
\(u_n \leq v\). So 
\[ P(X_n \leq u_n) \leq \lim_{n \to +\infty} P(X_n \leq u_n) \leq \lim_{n \to +\infty} P(X_n \leq v) = G(v). \]
Now we can take \(v \to 0\): as \((X_n)_{n \geq 0}\) is positive, \(G(v) \to 0\). Consequently, 
\[ P(X_n \leq u_n) \to 0. \]
□

Proof. of Theorem 1. The proof is essentially the same as Theorem 3.1 in [17], except when the factors are equal. We have

\[
\{q_n = q_0\} = \{q_0 = \min\{j : \delta_{n,j+1} < d_n\}\}
\]

\[
= \{\forall j \in \{1, \ldots, q_0\}, \delta_{n,j} \geq d_n\} \cap \{\delta_{n,q_0+1} < d_n\}.
\]

Therefore

\[
P(q_n = q_0) = P\left(\bigcap_{1 \leq j \leq q_0} \{\delta_{n,j} \geq d_n\} \cap \{\delta_{n,q_0+1} < d_n\}\right)
\]

\[
= 1 - P\left(\bigcup_{1 \leq j \leq q_0} \{\delta_{n,j} < d_n\} \cup \{\delta_{n,q_0+1} \geq d_n\}\right)
\]

\[
\geq 1 - \sum_{j=1}^{q_0} P(\delta_{n,j} < d_n) - P(\delta_{n,q_0+1} \geq d_n).
\]

Case of \(j = q_0 + 1\). In this case, \(\delta_{n,q_0+1} = \lambda_{n,q_0+1} - \lambda_{n,q_0+2}\) (noise eigenvalues). As \(d_n \to 0\) such that, \(n^{2/3}d_n \to +\infty\), and by using Proposition 3 in the same manner as in the proof of Theorem 3.1 in [17], we have

\[ P(\delta_{n,q_0+1} \geq d_n) \to 0. \]

Case of \(1 \leq j \leq q_0\). These indices correspond to the factor eigenvalues.

- Let \(I_1 = \{1 \leq l \leq q_0|\text{card}(J_l) = 1\}\) (simple factor) and \(I_2 = \{l-1|l \in I_1\}\). For all \(j \in I_1 \cup I_2\), \(\delta_{n,j}\) corresponds to a consecutive difference of \(\lambda_{n,j}\) issued from two different factors, so we can still use Proposition 2 and the proof of Theorem 3.1 in [17] to show that

\[ P(\delta_{n,j} < d_n) \to 0, \forall j \in I_1. \]
• Let \( I_3 = \{1 \leq l \leq q_0 - 1 | l \notin (I_1 \cup I_2)\} \). For all \( j \in I_3 \), it exists \( k \in \{1, \ldots, K\} \) such that \( j \in J_k \). By Proposition 2, \( \mathbf{X}_n = \sqrt{n} \delta_{n,j} \) converges weakly. So by using Lemma 1 and that \( d_n = o(n^{-1/2}) \), we have
  \[
P(\delta_{n,j} < d_n) = P\left(\sqrt{n} \delta_{n,j} < \sqrt{n} d_n\right) \to 0.
  \]

• The case of \( j = q_0 + 1 \) is considered as in [17].

Conclusion. \( P(\delta_{n,q_0+1} \geq d_n) \to 0 \) and \( \sum_{j=1}^{q_0} P(\delta_{n,j} < d_n) \to 0 \), therefore
  \[
P(\hat{q}_n = q_0) \xrightarrow{n \to +\infty} 1.
  \]

\[ \square \]

References

[1] T.W. Anderson, An introduction to multivariate statistical analysis, *Wiley Series in Probability and Statistics* (2003).
[2] Z.D. Bai and J.F. Yao, Central limit theorems for eigenvalues in a spiked population model, *Ann. Inst. H. Poincaré Probab. Statist.* **44**(3) (2008) 447–474.
[3] J. Baik and J.W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, *J. Multivariate Anal.* **97** (2006) 1382–1408.
[4] F. Benaych-Georges, A. Guionnet and M. Maida, Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices, *Electron. J. Prob.* **16**(60) (2011) 1621–1662.
[5] P.L. Combettes and J. Silverstein, Signal detection via spectral theory of large dimensional random matrices, *IEEE Trans. Signal Process.* **40**(8) (1992) 2100–2105.
[6] R. Everson and S. Roberts, Inferring the eigenvalues of covariance matrices from limited, noisy data, *IEEE Trans. Signal Process.* **48**(7) (2000) 2083–2091.
[7] M.C. Harding, Structural estimation of high-dimensional factor models, *Econometrica* 65(1) (2007) 165–190.
[8] I.M. Johnstone, On the distribution of the largest eigenvalue in principal component analysis, *Ann. Stat.* **29** (2001) 295–327.
[9] S. Krčková and B. Nadler, Determining the number of components in a factor model from limited noisy data, *Chem. Int. Lab. Syst.* **94** (2008) 19–32.
[10] S. Krčková and B. Nadler, Non-parametric detection of the number of signals: hypothesis testing and random matrix theory, *IEEE Trans. Signal Process.* **57**(10) (2009) 3930–3941.
[11] C. Lam, Q. Yao and N. Bathia, Estimation of latent factors for high-dimensional time series, *Biometrika* **98** (2011) 901–918.
[12] C. Lam and Q. Yao, Factor modeling for high-dimensional time series: inference for the number of factors, *Ann. Stat.* (2011) to appear.
[13] R.R. Nadakuditi and A. Edelman, Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples, *IEEE Trans. Signal Process.* **56**(7) (2008) 2625–2638.
[14] B. Nadler, Non-parametric detection of signals by information theoretic criteria: performance analysis and an improved estimator, *IEEE Trans. Signal Process.* **58**(5) (2010) 2746–2756.
[15] T. Naes, T. Isaksson, T. Fearn and T. Davies, User-friendly guide to multivariate calibration and classification, *NIR Publications, Chichester* (2002).
[16] A. Onatski, Testing hypotheses about the number of factors in large factors models, to appear in *Econometrica*, (2008).

[17] D. Passemier and J.F Yao, On determining the number of spikes in a high-dimensional spiked population model, *Random Matrices: Theory and Applications* 1(1) (2012) 1150002.

[18] D. Paul, Asymptotic of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica* 17 (2007) 1617–1642.

[19] S.A. Ross, The arbitrage theory of capital asset pricing, *J. Economic Theory*, 13 (1977) 341–360.

[20] M.O. Ulfarsson and V. Solo, Dimension estimation in noisy PCA with SURE and random matrix theory, *IEEE Trans. Signal Process.* 56(12) (2008) 5804–5816.

[21] M. Wax and T. Kailath, Detection of signals by information theoretic criteria, *IEEE Trans. Acoust., Speech, Signal Process.* 33(2) (1985) 387–392.

**Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong**

_E-mail address_, Damien Passemier: _eepassemier@ust.hk_

(Jian-Feng Yao) **Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong**

_E-mail address_, Jian-Feng Yao: _jeffyao@hku.hk_