

ON THE DETECTION OF THE NUMBER OF SIGNALS WITH POSSIBLY EQUAL STRENGTHS IN THE HIGH-DIMENSIONAL CASE

DAMIEN PASSEMIER AND JIAN-FENG YAO

ABSTRACT. Detection of the number of signals impinging on an array of sensors is an important problem in signal and array processing. Most of the papers consider asymptotic expansions of the sample size n whereas the dimension p of the observations is kept small. In this paper, we consider the case of high dimension, when p is large compared to n , using recent results of random matrix theory. We extend our results obtained in the paper [13] to the case of equal signals, and compare our algorithm to the method of Kritchman & Nadler [9], [10].

1. INTRODUCTION

In the area of signal processing, detection of the number of sources impinging on an array of sensors in presence of noise is a known and well-investigated problem [5], [6], [10], [16]. This detection is generally a first step preliminary to any further study such as estimation of parameters.

The underlying statistical model also appears in other scientific fields. In economics and psychological literature, this model is called a factor model where the number of factors (signals) has a primary importance [1], [15]. Similar models can be found in physics of mixture [9], [11] or population genetics. More recently and in a slightly more general set-up, the model is introduced as a spiked population model [8].

Many methods for determining the number of signals have been developed, mostly based on the minimum description length (MDL), Bayesian model selection or Bayesian Information Criteria (BIC) [16]. 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 avoid this problem of high dimension, several methods have been recently proposed using the random matrix theory, such as Harding [7] or Onatski [12] in economics, and Kritchman & Nadler in chemometrics literature [9] and array processing [10]. In [13], we have also introduced a new method based on recent results of [2] and [14] in random matrix theory.

In all the cited references, signals are assumed to have distinct strengths. However, we observe that when some of these signals strengths become close, the detection problem is more difficult and most of existing algorithms need to be modified. We refer this new situation as the case with possibly equal signals and its precise formulation is given in Section 3.2. The aim of this work is to extend our method [13] to this new case and to compare it with the method of Kritchman & Nadler [10] which, surprisingly enough, still work here even though it was introduced initially for the standard situation of distinct signals.

The paper is organized as follows. In Section 2, we introduce the model. In Section 3, we define the detection problem of possibly equal signals and present our solution. We establish its asymptotic consistency. In Section 4, we recall the algorithm of Kritchman & Nadler [10]. Next we conduct simulations experiments to compare these two methods.

2. PROBLEM FORMULATION

Assuming an array of p sensors we consider the following standard model

$$(1) \quad \mathbf{x}(t) = \sum_{k=1}^{q_0} a_k s_k(t) + \sigma n(t)$$

$$(2) \quad = \mathbf{A} s(t) + \sigma n(t),$$

where

- $s(t) = (s_1(t), \dots, s_{q_0}(t))^* \in \mathbb{R}^{q_0}$ are q_0 random signals assumed to have zero mean, unit variance and be mutually uncorrelated;
- $\mathbf{A} = (a_1, \dots, a_{q_0})$ is the $p \times q_0$ steering matrix of q_0 linearly independent p -dimensional vectors;
- $\sigma \in \mathbb{R}$ is the unknown noise level, $n(t) \sim \mathcal{N}(0, \mathbf{I}_p)$ is a $p \times 1$ vector of additive noise, independent of $s(t)$.

In this case, the population covariance matrix $\Sigma = \text{cov}(\mathbf{x}(t))$ of $\mathbf{x}(t)$ takes the diagonal form

$$\mathbf{W}^* \Sigma \mathbf{W} = \sigma^2 \mathbf{I}_p + \text{diag}(\alpha_1, \dots, \alpha_{q_0}, 0, \dots, 0)$$

where \mathbf{W} is an unknown basis of \mathbb{R}^p and $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_{q_0} > 0$. The sample covariance matrix of the n p -dimensional i.i.d. signal vectors received at each time t , $(\mathbf{x}_i = \mathbf{x}(t_i))_{1 \leq i \leq n}$ is

$$\mathbf{S}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^*.$$

Denote by $\lambda_{n,1} \geq \lambda_{n,2} \geq \dots \geq \lambda_{n,p}$ its eigenvalues. Our aim is to estimate q_0 on the basis of \mathbf{S}_n . For the moment, we assume that the noise level σ^2 is known. If this is not the case, we will give a method in section 4.2 to estimate it.

3. DETECTION OF THE NUMBER OF SIGNALS

In this section, we first recall our previous result of [13] in the case of different signals. Next, we propose an extension of the detection algorithm to the case with possibly equal signals. The consistency of the extended algorithm is established.

3.1. Previous work: detection in the case of different signals. We consider the case where the $(\alpha_i)_{1 \leq i \leq q_0}$ are all different, so there is q_0 distinct signals. According to [13], the population covariance matrix Σ has the spectral representation

$$W^* \Sigma W = \sigma^2 \text{diag}(\alpha'_1, \dots, \alpha'_{q_0}, 1, \dots, 1),$$

with the α'_i 's equals to

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

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

Moreover, we assumed that $\alpha'_1 > \dots > \alpha'_{q_0} > 1 + \sqrt{c}$, i.e all the signal strength α 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 \mathbf{x} , for each $k \in \{1, \dots, q_0\}$ and almost surely,

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

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

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

The estimation method of the number of signals q_0 in [13] is based on a close inspection of the following differences between consecutive eigenvalues

$$\delta_{n,j} = \lambda_{n,j} - \lambda_{n,j+1}, \quad j \geq 1.$$

Indeed, from the results quoted above it is easy to see that a.s. if $j \geq q_0$, $\delta_{n,j} \rightarrow 0$ while when $j < q_0$, $\delta_{n,j}$ tends to a positive limit. Thus it is possible to detect q_0 from index-numbers j where $\delta_{n,j}$ become small. More precisely, the estimator is

$$(3) \quad \hat{q}_n = \min\{j \in \{1, \dots, s\} : \delta_{n,j+1} < d_n\},$$

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 signals. In [13], we proved the consistency of \hat{q}_n providing that the threshold satisfies $d_n \rightarrow 0$, $n^{2/3}d_n \rightarrow +\infty$, 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 is there exists positive constants C, C' such that, for all $t \geq C'$,

$$\mathbb{P}(|x^i| \geq t^C) \leq e^{-t}.$$

3.2. Detection in the case with possibly equal signals. As said in Introduction, when some of signals have close strengths, detection algorithms need to be modified. More precisely, we adopt the following theoretic model where we have K different signal strengths $\alpha_1, \dots, \alpha_K$ and for each signal strength α_k , we have n_k signals to be detected, that is

$$\begin{aligned} \text{spec}(\Sigma) &= \underbrace{(\alpha_1, \dots, \alpha_1)}_{n_1}, \underbrace{(\alpha_2, \dots, \alpha_2)}_{n_2}, \dots, \underbrace{(\alpha_K, \dots, \alpha_K)}_{n_K}, \underbrace{(0, \dots, 0)}_{p-q_0} + \sigma^2 \underbrace{(1, \dots, 1)}_p \\ &= \sigma^2 \underbrace{(\alpha'_1, \dots, \alpha'_1)}_{n_1}, \underbrace{(\alpha'_2, \dots, \alpha'_2)}_{n_2}, \dots, \underbrace{(\alpha'_K, \dots, \alpha'_K)}_{n_K}, \underbrace{(1, \dots, 1)}_{p-q_0}. \end{aligned}$$

with $n_1 + \dots + n_K = q_0$. When the signal strengths are different, the difference between the corresponding eigenvalues of the sample covariance matrix will tends to a positive constant, whereas with two signals of equal strength this difference will tends to zero. This fact creates an ambiguity with those differences corresponding to the noise eigenvalues which also tend to zero. Nevertheless, the convergence of the $\delta_{n,i}$, for $i > q_0$ (noise) is faster (in $O_{\mathbb{P}}(n^{-2/3})$) than that of the $\delta_{n,i}$'s corresponding to two identical signals (in $O_{\mathbb{P}}(n^{-1/2})$) as a consequence of Theorem 3.1 of Bai & Yao [2]. This allows us to use the same estimator (3), provided we use 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 (2) and satisfies Assumption 1. We suppose that the population covariance matrix Σ has K non null and non unit eigenvalues $\alpha_1 > \dots > \alpha_K > \sigma^2 \sqrt{c}$ with respective multiplicity $(n_k)_{1 \leq k \leq K}$ ($n_1 + \dots + n_K = q_0$), and $p - q_0$ unit eigenvalues. Assume that $\frac{p}{n} \rightarrow c > 0$ when $n \rightarrow +\infty$. Let $(d_n)_{n \geq 0}$ be a real sequence such that $d_n = o(n^{-1/2})$ and $n^{2/3}d_n \rightarrow +\infty$. Then the estimator \hat{q}_n is strongly consistent, i.e $\hat{q}_n \rightarrow q_0$ almost surely when $n \rightarrow +\infty$.*

Notice that the only modification of our estimator comparing to the different signals case 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.

4. METHOD OF KRITCHMAN & NADLER WITH A NOISE ESTIMATION

4.1. Algorithm of Kritchman & Nadler. In their paper [9] and [10], these authors develop a method based on another theorem from random matrix theory to detect the number of signals.

In the absence of signals, nS_n follows a Wishart distribution with parameters n, p . In this case, Johnstone [8] gave the asymptotic distribution of the largest eigenvalue of S_n .

Proposition 1. Let \mathbf{S}_n be the sample covariance matrix of n vectors distributed as $\mathcal{N}(0, \sigma^2 \mathbf{1}_p)$, and $\lambda_{n,1} \geq \lambda_{n,2} \geq \dots \geq \lambda_{n,p}$ be its eigenvalues. Then, when $n \rightarrow +\infty$, such that $\frac{p}{n} \rightarrow c > 0$

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

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

Assuming the variance σ^2 is known. To distinguish a signal eigenvalue λ from a noise one at an asymptotic significance level γ , their idea is to check whether

$$(4) \quad \lambda_{n,k} > \sigma^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b \right)$$

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. Their estimator is based on a sequence of nested hypothesis tests of the following form: for $k = 1, 2, \dots, \min(p, n) - 1$,

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

For each value of k , they test the likelihood of the k -th eigenvalue $\lambda_{n,k}$ as arising from a signal or from noise as (4). If (4) is satisfied, \mathcal{H}_0 is accepted and k is increased by one. The procedure stops once an instance of \mathcal{H}_0 is rejected and the number of signals is estimated to be $\tilde{q}_n = k - 1$. Formally, their estimator is defined by

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

We refer this as the KN estimator. The authors proved the strong consistency of their algorithm as $n \rightarrow +\infty$ with fixed p , by replacing the fixed confidence level γ with a sample-size dependent one γ_n , where $\gamma_n \rightarrow 0$ sufficiently slow as $n \rightarrow +\infty$. They also proved that $\lim_{p,n \rightarrow +\infty} \mathbb{P}(\tilde{q}_n \geq q_0) = 1$.

4.2. Estimation of the noise level. When the noise level σ^2 is unknown, an estimation is needed. In [13], 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 it is 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 \hat{q}_n and \tilde{q}_n .

5. SIMULATION EXPERIMENTS AND COMPARISON

In this section we compare by simulation our signal detector (PY) to the Kritchman & Nadler's one (KN). In their papers [9] and [10], the authors compare the estimator KN with some other standard estimators in the signal processing literature, based on the minimum description length (MDL), Bayesian information criterion (BIC) and Akaike information criterion (AIC) [16]. In mostly studied cases, the estimator KN performs better. Thus we decided to consider only this detector for comparison.

After several experiments, we found that the following version of our detector significantly improves the detection performance. Indeed, instead of stop once one difference δ_k is below the threshold d_n (see (3)), the modified estimator now stops when two consecutive differences δ_k and δ_{k+1} are both below d_n . More precisely, we set

$$(5) \quad \hat{q}_n^* = \min\{j \in \{1, \dots, s\} : \delta_{n,j+1} < d_n \text{ and } \delta_{n,j+2} < d_n\}.$$

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.

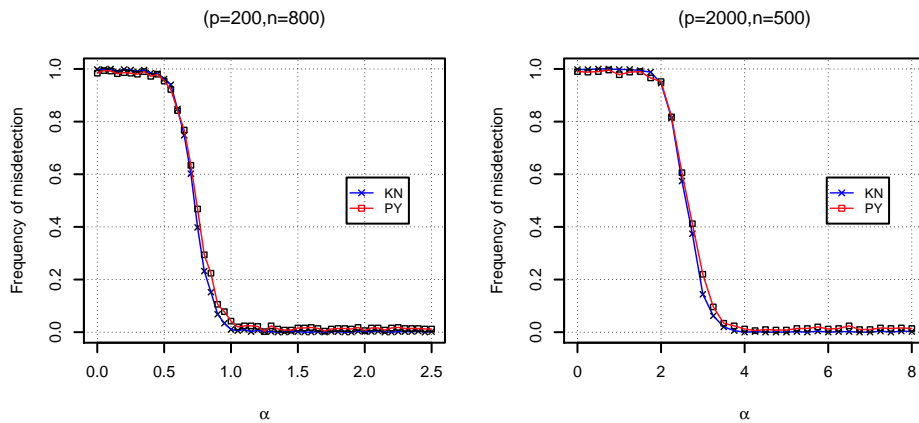
It remains to choose a threshold sequence d_n to be used for our estimator \hat{q}_n^* . As argued in [13], we use a sequence d_n of the form $Cn^{-2/3}\sqrt{2\log\log n}$, where C is a constant to be adjusted for each case. In all simulations, we consider 500 independent replications and take $\sigma^2 = 1$. 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.

Table 1 gives a summary of parameters in our simulation experiments. There are two sets of experiments for the comparison of the KN estimator with the ours \hat{q}_n^* (PY). In the first one (Figures 1-4 in the Table 1), signals have different strengths and these experiments extend and complete few results already reported in [13]. The second set of experiments (Figures 5-8 in Table 1) addresses the new situation where some signals have equal strengths. The last experiment (Figure 9) considers the case of no signal.

5.1. Case of different signals. In Figure 1, we consider the case of a single signal with strength α , and we analyze the probability of misdetection as a function of the signal α , 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 to both estimators.

TABLE 1. Summary of parameters used in the simulation experiments.

Figure number	Signal strengths	Model number	Signals strengths	Fixed parameters			Varying parameters	
				(p, n)	c	$\sigma^2 = 1$		C
1	Different		(α)	$(200, 800)$ $(2000, 500)$	0.25 4	Given	5.5 9	α
2	Different	A B	$(6, 5)$ $(10, 5)$		10	Given	11	n
3	Different	B	$(6, 5)$		10	To be estimated	11	n
4	Different	C D	(1.5) $(2.5, 1.5)$		1	Given	5	n
5	Possibly equal	E F	$(\alpha, \alpha, 5)$ $(\alpha, \alpha, 15)$	$(200, 800)$ $(2000, 500)$	0.25 4	Given	6 9.9	α
6	Possibly equal	G H	$(6, 5, 5)$ $(10, 5, 5)$		10	Given	9.9	n
7	Pos. equal	H	$(10, 5, 5)$		10	To be estimated	9.9	n
8	Possibly equal	I J	$(1.5, 1.5)$ $(2.5, 1.5, 1.5)$		1 10	Given	4 15	n
9	No signals	K	No signals		1 10	Given	8 15	n

FIGURE 1. Misdetection rates as a function of signal strength for $(p, n) = (200, 800)$ (left) and $(p, n) = (2000, 500)$ (right).

The two estimators have similar performance: the levels of detection are the same, and fit with the theory ($\sqrt{c} = 0.5$ for the first case, and 2 for the second).

In Figure 2, we consider two models with two signals ($q_0 = 2$):

- Model A: $(\alpha_1, \alpha_2) = (6, 5)$;
- Model B: $(\alpha_1, \alpha_2) = (10, 5)$.

The detection is harder in Model A as the signal strengths are closer. We fix $c = 10$ ($p \gg n$), and we plot the misdetection rates against the sample size n . Here $C = 11$. Again, $\sigma^2 = 1$ is given to both estimators.

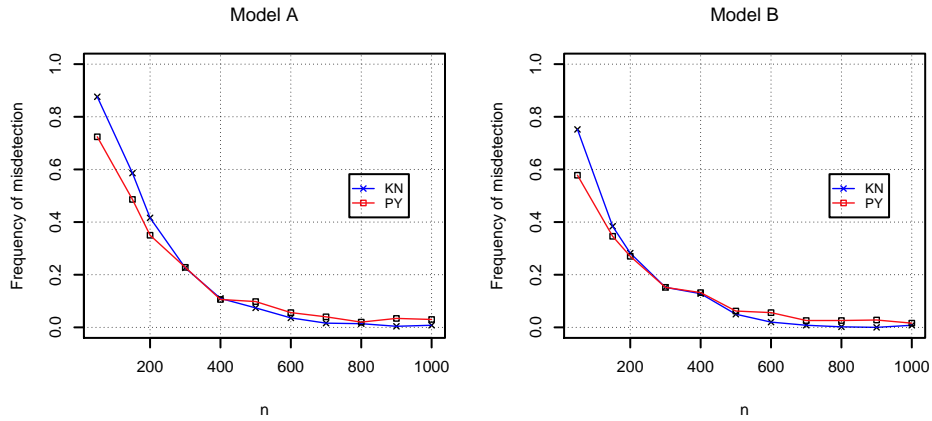


FIGURE 2. Misdetection rates as a function of n for $\alpha = (6, 5)$ (left) and $\alpha = (10, 5)$ (right).

As in Figure 1, 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 .

In Figure 3, we keep the same settings as in the previous simulation for Model B but with an unknown noise level σ^2 to be estimated either.

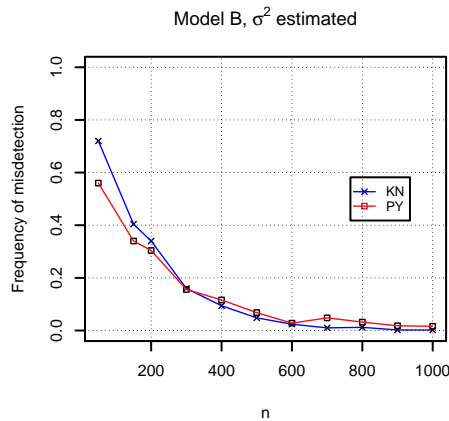


FIGURE 3. Misdetection rates as a function of n for $\alpha = (10, 5)$ with σ^2 estimated.

Compared to Figure 2 (Model B), the estimation of σ^2 does not affect the two estimators significantly. Both estimators seem robust against the unknown noise level. Note that we again observe a different hierarchy before and after $n \simeq 400$.

Figure 4 considers two cases with $c = 1$ and a given noise level $\sigma^2 = 1$:

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

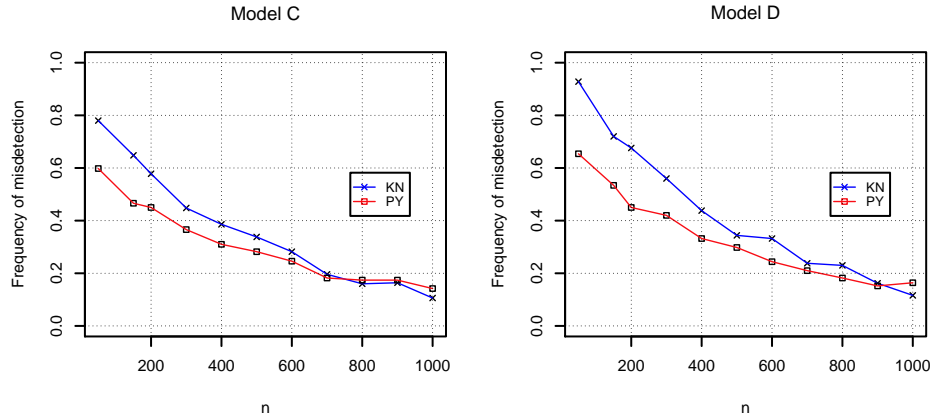


FIGURE 4. Misdetecion rates as a function of n for $\alpha = (1.5)$ (left) and $\alpha = (2.5, 1.5)$ (right).

This experiment is designed with signal strengths getting closer to the critical value $\sqrt{c} = 1$. This detecion becomes more difficult and as expected both methods will have higher misdetecion rates. Here we used $C = 5$. Meanwhile as displayed in Figure 4, our algorithm have a lower misdetecion rate in almost all cases in both models, with an improvement ranging from 10% to 30% for moderate sample sizes $n \leq 400$.

5.2. Case with equal signals. We keep the same parameters as in the previous section and only change the signal strengths. In Figure 5, 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.

This figure is to be compared to Figure 1 for different signals. Adding multiplicity only slightly increases the level of detecion. As previously, the two estimators have similar performance.

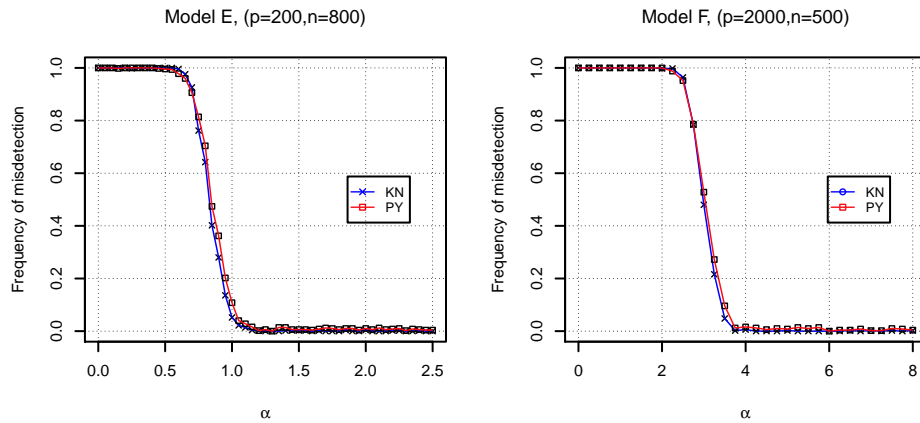


FIGURE 5. Misdetction rates as a function of signal strength for $(p, n) = (200, 800)$ (left) and $(p, n) = (2000, 500)$ (right).

In Figure 6, we consider two models analog to Model A and B with three signals ($q_0 = 3$):

- Model G: $(\alpha_1, \alpha_2, \alpha_3) = (6, 5, 5)$;
- Model H: $(\alpha_1, \alpha_2, \alpha_3) = (10, 5, 5)$.

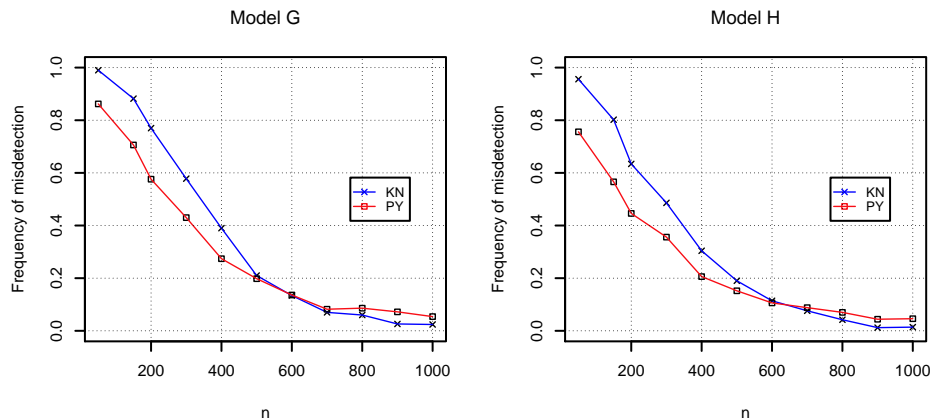


FIGURE 6. Misdetction rates as a function of n for $\alpha = (6, 5, 5)$ (left) and $\alpha = (10, 5, 5)$ (right).

Again we fix $c = 10$ and we plot the probability of misdetection against the sample size n . Here $C = 9.9$ and σ^2 is given. Comparing to the different signal strengths case (Figure 2), the two estimators have significantly higher error rates. Nevertheless, the estimator PY shows superior detection performance for $n \leq 500$ (up to 20% less error): adding an equal signal affect more the performance of the estimator KN, but both estimators remain asymptotically consistent. If we compare Model G and Model H, a smaller spacing between the two first signals gives only a slightly degradation of correct detection.

In Figure 7, we keep the same settings as in the previous simulation for Model H but with an unknown noise level σ^2 to be estimated either.

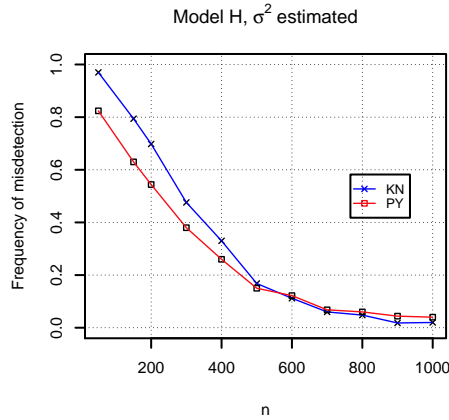


FIGURE 7. Misdetection rates as a function of n for $\alpha = (10, 5, 5)$ with σ^2 estimated.

Compared to Figure 6 (Model H), the estimation of σ^2 does not affect the two estimators significantly. Both estimators seem robust against the unknown noise level. Note that we again observe a different hierarchy before and after $n \simeq 500$.

Figure 8 considers two cases with $c = 1$, and again $\sigma^2 = 1$ is given:

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

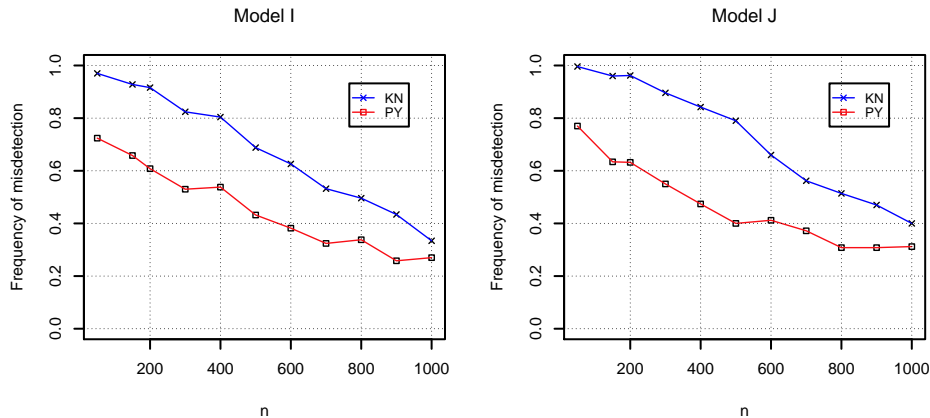


FIGURE 8. Misdetection rates as a function of n for $\alpha = (1.5, 1.5)$ (left) and $\alpha = (2.5, 1.5, 1.5)$ (right).

Here we used $C = 4$. As explained in the previous section, this situation is more difficult and this causes a degradation in detection performance. The difference between the two

algorithms is higher than in the previous cases: the estimator PY performs better, up to 40%. However, the convergence of both algorithms is quite slow.

5.3. Case of no signal. In Figure 9 we examine the performance of the two algorithms in the case of no signal (Model K). The cases of $c = 1$ and $c = 10$ with $\sigma^2 = 1$ given are considered.

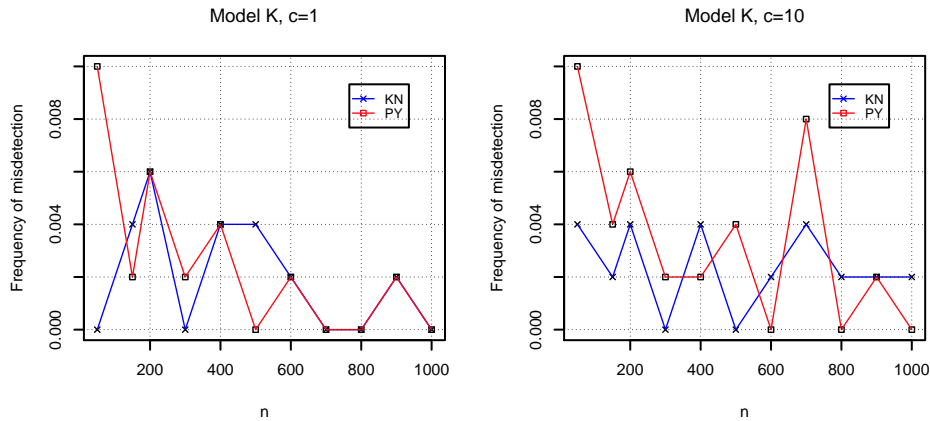


FIGURE 9. Misdetection rates as a function of n in the case of no signals for $c = 1$ (left) and $c = 10$ (right).

We chose $C = 8$ for the first case and $C = 15$ for the second case. In both situations, the misdetection rates of the two estimators are similar and low (less than 1%).

6. CONCLUDING REMARKS

In this paper we have considered the signal detection problem in the high-dimensional case. When some signals have close or even equal strengths, the detection 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 detector has smaller misdetection rates, especially in cases with close and relatively low signal strengths (Figure 4 and 8) or more generally for almost all the cases provided that the sample size n is moderately large ($n \leq 400$ or 500).

Another important conclusion concerns the influence of the tuning parameter, C for our algorithm and the false alarm rate γ for KN. The main drawback of our algorithm is its lack of robustness with regard to the value of C . The experiments reported here are obtained with a finely-turned value of C and this value varies from case to case. How to overcome

this drawback, e.g by a data-adapted C remains an open problem. By comparison, the KN detector is remarkably robust and a single value of $\gamma = 0.5\%$ was used in all the experiments.

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 a CLT for the n_k -packed eigenvalues

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

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

Proposition 2. *Assume that the entries x^i of \mathbf{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, \dots, n_K respectively. Then as $p, n \rightarrow +\infty$ so that $\frac{p}{n} \rightarrow c$, 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 depend of α'_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 \mathbf{x} have a symmetric law and a sub-exponential decay, that is 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 ,*

$$\frac{n^{\frac{2}{3}}}{\beta}(\lambda_{n,q_0+i} - b) = O_{\mathbb{P}}(1),$$

where $\beta = (1 + \sqrt{c})(1 + \sqrt{c^{-1}})^{\frac{1}{3}}$.

We also need the following lemma:

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

$$\mathbb{P}(\mathbf{X}_n \leq u_n) \rightarrow 0.$$

Proof. As $(\mathbf{X}_n)_{n \geq 0}$ converges weakly, it exists a function G such that, for all $v \geq 0$, $\mathbb{P}(\mathbf{X}_n \leq v) \rightarrow G(v)$. Furthermore, as $u_n \rightarrow 0$, it exists $N \in \mathbb{N}$ such that for all $n \geq N$, $u_n \leq v$. So $\mathbb{P}(\mathbf{X}_n \leq u_n) \leq \mathbb{P}(\mathbf{X}_n \leq v)$, and $\overline{\lim}_{n \rightarrow +\infty} \mathbb{P}(\mathbf{X}_n \leq u_n) \leq \overline{\lim}_{n \rightarrow +\infty} \mathbb{P}(\mathbf{X}_n \leq v) = G(v)$. Now we can take $v \rightarrow 0$: as $(\mathbf{X}_n)_{n \geq 0}$ is positive, $G(v) \rightarrow 0$. Consequently, $\mathbb{P}(\mathbf{X}_n \leq u_n) \rightarrow 0$. \square

Proof. of Theorem 1. The proof is essentially the same as Theorem 3.1 in [13], except when the spikes are equal. We have

$$\begin{aligned}\{\hat{q}_n = q_0\} &= \{q_0 = \min\{j : \delta_{n,j+1} < d_n\}\} \\ &= \{\forall j \in \{1, \dots, q_0\}, \delta_{n,j} \geq d_n\} \cap \{\delta_{n,q_0+1} < d_n\}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{P}(\hat{q}_n = q_0) &= \mathbb{P}\left(\bigcap_{1 \leq j \leq q_0} \{\delta_{n,j} \geq d_n\} \cap \{\delta_{n,q_0+1} < d_n\}\right) \\ &= 1 - \mathbb{P}\left(\bigcup_{1 \leq j \leq q_0} \{\delta_{n,j} < d_n\} \cup \{\delta_{n,q_0+1} \geq d_n\}\right) \\ &\geq 1 - \sum_{j=1}^{q_0} \mathbb{P}(\delta_{n,j} < d_n) - \mathbb{P}(\delta_{n,q_0+1} \geq d_n).\end{aligned}$$

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 $d_n \rightarrow 0$ such that, $n^{2/3}d_n \rightarrow +\infty$, and by using Proposition 3 in the same manner as in the proof of Theorem 3.1 in [13], we have

$$\mathbb{P}(\delta_{n,q_0+1} \geq d_n) \rightarrow 0.$$

Case of $1 \leq j \leq q_0$. These indices correspond to the signal eigenvalues.

- Let $I_1 = \{1 \leq l \leq q_0 | \text{card}(J_l) = 1\}$ (simple signal) 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 signals, so we can still use Proposition 2 and the proof of Theorem 3.1 in [13] to show that

$$\mathbb{P}(\delta_{n,j} < d_n) \rightarrow 0, \forall j \in I_1.$$

- 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 $k \in \{1, \dots, K\}$ 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

$$\mathbb{P}(\delta_{n,j} < d_n) = \mathbb{P}(\sqrt{n}\delta_{n,j} < \sqrt{n}d_n) \rightarrow 0.$$

- The case of $j = q_0 + 1$ is considered as in [13].

Conclusion. $\mathbb{P}(\delta_{n,q_0+1} \geq d_n) \rightarrow 0$ and $\sum_{j=1}^{q_0} \mathbb{P}(\delta_{n,j} < d_n) \rightarrow 0$, therefore

$$\mathbb{P}(\hat{q}_n = q_0) \xrightarrow{n \rightarrow +\infty} 1.$$

□

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, *Preprint*.
- [5] W. Chen, K. M. Wong, and J. P. Reilly, Detection of the number of signals: A predicted eigen-threshold approach, *IEEE Trans. Signal Process.* **39(5)** (1991) 1088–1098.
- [6] E. Fishler, M. Grossmann, and H. Messer, Detection of signals by information theoretic criteria: General asymptotic performance analysis, *IEEE Trans. Signal Process.* **50(5)** (2002) 1027–1036.
- [7] M.C. Harding, Structural estimation of high-dimensional factor models, *Econometrica* r&r.
- [8] I.M. Johnstone, On the distribution of the largest eigenvalue in principal component analysis, *Ann. Stat.* **29** (2001) 295–327.
- [9] S. Kritchman 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. Kritchman 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] T. Naes, T. Isaksson, T. Fearn and T. Davies, User-friendly guide to multivariate calibration and classification, *NIR Publications, Chichester* (2002).
- [12] A. Onatski, Testing hypotheses about the number of factors in large factors models, to appear in *Econometrica*, (2008).
- [13] D. Passemier and J.F Yao, On determining the number of spikes in a high-dimensional spiked population model, *To appear in Random Matrices: Theory and Applications* **1(1)** (2012).
- [14] D. Paul, Asymptotic of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica* **17** (2007) 1617–1642.
- [15] S.A. Ross, The arbitrage theory of capital asset pricing, *J. Economic Theory*, **13** (1977) 341–360.
- [16] M. Wax and T. Kailath, Detection of signals by information theoretic criteria, *IEEE Trans. Acoust., Speech, Signal Process.* **33(2)** (1985) 387–392.

(Damien Passemier) IRMAR, UNIVERSITÉ DE RENNES 1, CAMPUS DE BEAULIEU, 35042 RENNES CEDEX, FRANCE

E-mail address, Damien Passemier: damien.passemier@univ-rennes1.fr

(Jian-Feng Yao) DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, THE UNIVERSITY OF HONG KONG, POKFULAM ROAD, HONG KONG

E-mail address, Jian-Feng Yao: jeffiyao@hku.hk