Inférence statistique dans un modèle à variances isolées de grande dimension

Damien Passemier

ANNÉE 2012

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications

École doctorale Matisse

présentée par

Damien Passemier

préparée à l'unité de recherche 6625 CNRS - IRMAR
Institut de Recherche Mathématique de Rennes
UFR de Mathématiques

Inférence statistique dans un modèle à variances isolées de grande dimension

Thèse soutenue à Rennes
le 4 décembre 2012

devant le jury composé de :

Jamal NAJIM
Chargé de recherche à Télécom ParisTech / *rapporteur*

Mérouane DEBBAH
Professeur à SUPELEC / *examinateur*

Mylène MAÏDA
Maître de conférence à l'université Paris-Sud / *examinateur*

Valérie MONBET
Professeur à l'université de Rennes 1 / *examinateur*

Jian-Feng YAO
Professeur à l'université de Hong Kong (HKU) / *directeur de thèse*

Rapporteurs :

Jean-Marc AZAÏS – Professeur à l'université de Toulouse
Jamal NAJIM – Chargé de recherche à Télécom ParisTech

Remerciements

Voici venu le moment tant redouté de l'écriture de mes remerciements, et qui conclut toute thèse : ce n'est vraiment pas un exercice facile ! J'espère que je ne vais oublier personne...

Pour commencer, je tiens à remercier mon directeur de thèse, Jian-Feng Yao, pour m'avoir fait confiance et avoir accepté de diriger mes travaux. Malgré son départ à Hong Kong au début de ma deuxième année de thèse, il a su rester présent et a tenu ses promesses en me faisant venir 2 fois à Hong Kong. J'ai beaucoup appris de ses compétences en statistiques.

Je remercie mes rapporteurs, Jean-Marc Azaïs et Jamal Najim, d'avoir accepté de juger ce travail, et pour les remarques pertinentes qu'ils m'ont faites. Je remercie aussi mes examinateurs Mérouane Debbah, Mylène Maïda et Valérie Monbet d'avoir fait parti de mon jury de thèse. C'est un honneur pour moi de la part de chercheurs d'un tel renom.

Mes remerciements vont aussi à l'ensemble du personnel administratif de l'UFR mathématiques et de l'IRMAR, pour leur disponibilité, leur gentillesse et leur efficacité.

J'exprime aussi ma profonde reconnaissance à tous les chercheurs que j'ai pu côtoyer, que ce soit au sein de l'IRMAR ou lors de conférences, en autres Bernard Delyon et Lionel Truquet.

Mon intérêt pour les mathématiques ne date pas d'hier, et je ne peux que remercier mes parents de m'avoir laissé la liberté de choisir ma voie. Je remercie aussi les différents enseignants que j'ai rencontré au cours de ma scolarité, et qui m'ont transmis leur passion des mathématiques. Je pense en particulier à M. Nedelec, M. Tygat, Mme Feibel, ainsi que Marc Peigné, Emmanuel Lesigne et Marie-Françoise Bidaut-Véron à l'université de Tours.

Lors de ces trois années, j'ai eu l'occasion de rencontrer beaucoup de personnes, et de me faire de nouveaux amis. Cela a commencé par Matthieu, que j'ai connu pendant mon M2, qui m'a très vite présenté aux membres du bureau 434 : Maher, Alinette et Nirmoul en particulier. Ils m'ont très vite intégré dans leur groupe, et la pause-café et mots-croisés au 434 est devenu une habitude. De nombreuses personnes participaient à cette pause : cela m'a permis de connaitre Richou, Anjara, John et Jobinou. Je pense aussi à Cyril et Gaël qui étaient de passage. Ou encore Quentin et Sylvain de Ker-Lann, et Christophe maintenant à Nantes. Merci à eux.

Merci à Yiqing d'avoir partagé mon bureau pendant ces 3 années. J'ai beaucoup apprécié son ouverture d'esprit, et été très touché qu'il me choisisse comme témoin à son mariage.

Un événement particulier dans ma vie m'a fait prendre conscience que certaines personnes que je côtoyais tous les jours étaient de vrais amis. Je pense en particulier à Nicolas, qui m'a offert un toit temporaire et m'a soutenu dans cette période, ainsi qu'à Matthieu et plus récemment à Julie. J'ai aussi appris à connaître Guillaume et Thibaut, les rois de la dalouze !

Au cours de ces 3 années, des personnes sont parties, et de nouvelles têtes sont apparues et ont perduré la tradition de la pause-café : Kodjo, Christophe, Elise, Romain, Jeroen, Marie et Guillaume entre autres. Je vous remercie tous pour les bons moments passés en votre compagnie. J'ai sûrement oublié de citer certaines personnes : toutes celles qui se reconnaitront reçoivent toute ma gratitude.

En dehors du laboratoire, je souhaiterais remercier Jonathan (et Caroline), qui m'a permis de me changer les idées en m'entrainant dans ses sessions planche à voile. Je tiens aussi à remercier mes amis de plus longue date : Xavier et Aurélien, ainsi que Jérôme. Et bien entendu, deux amis en particulier que j'ai connu pendant les 4 premières années à l'université de Tours : Roland et Valentin. Merci d'être présents depuis toutes ces années, pour tout ces bons moments passés en votre compagnie, pour les innombrables soirées WA, CS, L4D, SC2, BF3...Merci de m'avoir soutenu pendant ma thèse et pendant les moments difficiles. Je vous en suis sincèrement reconnaissant (et merci à Valentin pour les nombreuses relectures de ce manuscrit, je crois que tu es celui qui l'a le plus lu après moi !). Merci aussi à Marianne et Magalie.

Enfin, je ne peux terminer sans remercier l'ensemble de ma famille, et en particulier mes parents, mes grands parents et mes frères Vincent et Clément. C'est grâce à vous si j'en suis arrivé là, tout cela n'aurait jamais été possible sans votre soutien sans faille. J'ai toujours été touché par votre admiration envers moi, admiration que je ne trouve pas justifiée. J'ai une pensée particulière pour Vincent qui, vivant à Melbourne, ne peut être présent aujourd'hui.

A tous ceux que j'aime, et qui m'aiment...

Avant-propos : résumé en français

Cette thèse se place dans le cadre général de l'inférence statistique des données de grande dimension. C'est un domaine relativement récent, dont l'intérêt est apparu grâce au développement de l'informatique moderne et la possibilité d'observer et de consigner une quantité importante de données. La théorie des matrices aléatoires de grande dimension permet de prendre en compte ce cadre, étant donné que la plupart des résultats limites considèrent que la taille de la matrice tend vers l'infini. Une part non négligeable de ces résultats concerne la matrice de covariance empirique $\mathsf{S}_n = \frac{1}{n}\mathsf{X}\mathsf{X}'$, où $\mathsf{X} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ est un n-échantillon de vecteurs aléatoires de dimension p (en général gaussien), possédant une matrice de variance-covariance (ou de population) Σ. En particulier, plusieurs théorèmes limites concernent les valeurs propres de S_n quand p et n tendent vers l'infini de manière proportionnelle (voir Anderson et al. (2010); Bai & Silverstein (2010)). Ces derniers fournissent des outils fondamentaux pour l'étude des statistiques usuelles, car la plupart d'entre-elles sont des fonctions des valeurs propres de la matrice de covariance empirique S_n.

Ce travail est divisé en six chapitres. Les trois premiers constituent une partie introductive, tandis que les trois suivants concernent les contributions originales de cette thèse.

Données de grande dimension et matrices aléatoires

Le premier chapitre commence par présenter le problème de la grande dimension, i.e. lorsque que le nombre p de variables est « grand » par rapport à la taille de l'échantillon n. En effet, le cadre asymptotique classique considère que p est « petit » et fixé tandis que n tend vers l'infini. Nous faisons ici l'hypothèse que p et n tendent ensemble vers l'infini de manière proportionnelle. Nous dressons ensuite un bref historique de la théorie des matrices aléatoires. Nous rappelons les résultats sur l'analyse spectrale des matrices aléatoires de grande dimension qui nous seront utiles pour la suite.

Comme indiqué au début de ce résumé, nous nous intéressons à la matrice de covariance empirique, et plus particulièrement au comportement de son spectre. La distribution spectrale empirique de S_n, notée F^{S_n} est donc naturellement considérée. Une des méthodes utilisée pour son étude est la transformée de Stieltjes (ou de Cauchy), qui permet de définir une mesure et qui caractérise leur convergence en loi. En effet, la transformée de Stieltjes s_n de la mesure spectrale empirique d'une matrice carrée A n'est autre que le résolvant de cette dernière à un facteur $1/n$ près :

$$s_n(z) = \int \frac{1}{x-z} F^{\mathsf{A}_n}(\mathrm{d}x) = \frac{1}{n}\mathrm{tr}(\mathsf{A} - z\mathsf{I})^{-1}.$$

Dans le cas de la matrice de covariance empirique S_n, la transformée de Stieltjes permet de démontrer que sa distribution spectrale empirique F^{S_n} converge en loi vers une distribution de Marčenko-Pastur si p et n tendent vers l'infini, avec $p/n \to c$. Lorsque $\Sigma = \sigma^2 \mathsf{I}_p$ par exemple, F^{S_n} converge en loi vers la distribution de Marčenko-Pastur standard de paramètres c et σ^2. Ce n'est évidemment pas le cas dans le cadre classique : si p reste petit et fixé, et $n \to \infty$, $F^{\mathsf{S}_n} \overset{\mathcal{L}}{\to} F^{\Sigma} = \delta_{\sigma^2}$. L'emploi de cette limite vers σ^2 à la place de la limite de Marčenko-Pastur constitue la raison fondamentale des mauvaises performances des méthodes classiques de statistique multivariée en grande dimension.

Nous énonçons ensuite le théorème central limite pour la mesure spectrale de la matrice de covariance empirique de Bai & Silverstein (2004). Ce dernier concerne les statistiques spectrales linéaires

$$\widehat{\theta}(f) \quad = \quad \int f(x)\,\mathrm{d}F^{\mathsf{B}_n}(x).$$

où B_n est une matrice de covariance empirique. Leur étude a un intérêt fondamental, car la plupart des statistiques de population en analyse multivariée peuvent s'écrire en fonction de la distribution spectrale empirique de F^{B_n}. Ce théorème est ensuite appliqué à un exemple issu de Bai et al. (2009), qui concerne le problème du test de covariance d'un échantillon.

Nous présentons enfin le modèle que nous considèrerons dans cette thèse, à savoir le modèle à variances isolées, ainsi que quelques résultats qui lui sont associés. Ce modèle a été introduit par Johnstone (2001), qui a remarqué que plusieurs valeurs propres extrêmes de certains échantillons de données s'écartent des autres valeurs propres, ces dernières restant confinées dans le support de la loi de Marčenko-Pastur. Pour expliquer ce phénomène, il proposa un « modèle à variances isolées » (spiked population model), où toutes les valeurs propres de T_p sont égales à un, sauf un nombre fixé relativement petit m d'entres elles appelées « spikes ». En d'autres termes, la matrice de population Σ a pour valeurs propres

$$\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \ldots, \underbrace{\alpha_K, \ldots, \alpha_K}_{n_K}, \underbrace{1, \cdots, 1}_{p-m},$$

où $n_1 + \cdots + n_K = m$ est le nombre de spikes. Bai & Yao (2012) ont ensuite étendu le modèle présenté ci-dessus à un modèle à variances isolées généralisé.

Plusieurs auteurs ont étudié ce modèle, et en particulier Baik & Silverstein (2006), qui ont montré la convergence presque-sûre des valeurs propres $\lambda_{n,1} \geq \cdots \geq \lambda_{n,m}$ de la matrice de covariance empirique S_n correspondant aux spikes. On note $\phi(\alpha_i)$ ces limites presque-sûres. Dans Bai & Yao (2012), les auteurs ont démontré un théorème limite central pour les vecteurs de dimension n_k

$$\sqrt{n}(\lambda_{n,j} - \psi(\alpha_k)),\ j \in J_k,$$

où J_k désigne l'ensemble des n_k indices de α_k.

Le modèle à facteurs

Le chapitre 2 présente le modèle à facteurs et la théorie de la vraisemblance classique qui lui est associée. On se place ici dans le cadre classique : la dimension p des données reste fixe, tandis que la taille de l'échantillon n tend vers l'infini. Le modèle à facteurs repose sur la modélisation suivante : soit p le nombre de variables étudiées, n le nombre de données observées x_i et m le nombre de facteurs communs. Le modèle à facteurs s'écrit

$$\begin{aligned} \mathsf{x}_i &= \sum_{k=1}^{m} \mathsf{f}_{ki}\Lambda_k + \mathsf{e}_i + \mu && (1) \\ &= \Lambda\mathsf{f}_i + \mathsf{e}_i + \mu, && (2) \end{aligned}$$

où

- $\mu \in \mathbb{R}^p$ représente la moyenne générale ;
- $\mathsf{f}_i = (\mathsf{f}_{1i}, \ldots, \mathsf{f}_{mi})'$ sont les m facteurs aléatoires, appelés facteurs communs ou facteurs scores $(m < p)$;
- $\Lambda = (\Lambda_1, \ldots, \Lambda_m)$ est une matrice $p \times m$ de rang plein, appelée matrice des pondérations (*factors loadings*) ;
- e_i est le vecteur de bruit de dimension p, centré, indépendant de f_i et de matrice de variance-covariance $\Psi = \mathbb{E}(\mathsf{e}_i\mathsf{e}_i')$.

Les éléments de e_i sont les facteurs spécifiques, appelés aussi facteurs uniques ou idiosyncratiques : la variabilité non expliquée par les facteurs communs est représentée par la variance de ce vecteur. Les hypothèses classiques de ce modèle sont :

- $\mathbb{E}(\mathsf{f}_i) = 0$ et $\mathbb{E}(\mathsf{f}_i\mathsf{f}_i') = \mathsf{I}_p$;
- $\Psi = \mathrm{cov}(\mathsf{e}_i)$ est diagonale ;
- $\Gamma = \Lambda'\Psi^{-1}\Lambda$ est diagonale, d'éléments diagonaux ordonnés et différents.

La dernière hypothèse permet d'éviter un problème d'identification. Par conséquent, nous pouvons exprimer le modèle à facteurs par une condition sur la matrice de population $\Sigma = \mathrm{cov}(\mathsf{x}_i)$

$$\Sigma = \Lambda\Lambda' + \Psi,$$

où les éléments diagonaux de $\Lambda\Lambda'$ sont appelés communalités, et les éléments de Ψ sont les spécificités ou unicités. Notons qu'il existe d'autres hypothèses permettant aussi résoudre le problème de l'identification.

Si les facteurs communs f_i et les facteurs uniques e_i sont Gaussiens, une théorie reposant sur la vraisemblance est connue depuis Lawley (1940) (voir aussi Lawley & Maxwell (1971)). On suppose ici que le nombre de facteurs communs m est donné. Dans ce cas, le vecteur des observations x suit une loi normale $\mathcal{N}(\mu, \Sigma)$, où $\Sigma = \Lambda\Lambda' + \Psi$. Nous détaillons cette théorie, et présentons le test du rapport de vraisemblance d'adéquation au modèle à facteurs.

Différents types de modèles à facteurs existent : celui présenté en début de paragraphe (voir (2)) est appelé « modèle à facteurs strict », dans lequel la matrice Ψ est supposée diagonale.

Le modèle à facteurs strict à variance homoscédastique est une simplification du modèle à facteurs strict : nous supposons en effet que $\Psi = \sigma^2 \mathsf{I}_p$. C'est le modèle que nous considérerons dans les chapitres 4 à 6. Dans ce cas, les équations définissant les estimateurs du maximum de vraisemblance se simplifient et ces derniers possèdent une solution explicite. La statistique du rapport de vraisemblance L^* se simplifie aussi.

Le modèle à facteurs strict à variance homoscédastique est en fait une reformulation du modèle à variances isolées. En effet, dans ce cas, $\Sigma = \Lambda\Lambda' + \mathsf{I}_p$ et a pour spectre

$$\text{spec}(\Sigma) = (\alpha_1 + \sigma^2, \ldots, \alpha_m + \sigma^2, \underbrace{\sigma^2, \ldots, \sigma^2}_{p-m}),$$

qui peut aussi être écrit sous la forme

$$\text{spec}(\Sigma) \quad = \quad \sigma^2(\alpha_1^*, \ldots, \alpha_m^*, \underbrace{1, \ldots, 1}_{p-m}), \tag{3}$$

avec $\alpha_i^* = \frac{\alpha_i}{\sigma^2} + 1$, pour tout $1 \leq i \leq m$.

Méthodes d'estimation du nombre de facteurs/spikes

Le chapitre 3 expose plusieurs méthodes pour l'estimation du nombre de facteurs/spikes : notre première contribution consiste en effet en la construction d'une nouvelle méthode d'estimation, et il est donc intéressant de connaitre les méthodes existantes.

Dans une première partie, nous considérons le cadre asymptotique classique. Nous présentons brièvement la méthode du diagramme des valeurs propres (ou scree plot), ainsi que les estimateurs AIC et BIC basés sur les critères d'information théorique introduits par Akaike (1973, 1974) (AIC) et Rissanen (1978) (BIC/MDL). La « méthode de Laplace » de Minka (2000), une autre méthode bayésienne, est aussi abordée.

La deuxième partie traite du contexte de la grande dimension, qui nous intéresse ici particulièrement. Nous commençons par décrire la méthode SURE de Ulfarsson & Solo (2008) qui est basée sur l'estimateur sans biais du risque de Stein (SURE). Nous abordons ensuite la méthode de Harding (2007), avec laquelle nous comparons notre méthode dans le chapitre 4. Son estimateur est basé sur une comparaison entre les moments du spectre de la matrice de covariance empirique S_n et ceux de cette même distribution spectrale empirique de S_n, mais sans les facteurs. Les plus grandes valeurs propres de S_n sont successivement retirées jusqu'à obtenir un écart « faible » entre les différents moments. Enfin, nous présentons la méthode de Kritchman & Nadler (2008, 2009), qui est basée sur le fait qu'en l'absence de facteur ($m = 0$), $n\mathsf{S}_n$ suit une loi de Wishart de paramètres n et p. Johnstone (2001) a donné la distribution asymptotique de la plus grande valeur propre dans ce cas. Nous comparons notre estimateur à cette méthode dans les chapitres 4 et 5.

Estimation du nombre de facteurs/spikes en grande dimension

Nous présentons chapitres 4 et 5 une nouvelle méthode d'estimation du nombre de facteurs (ou spikes) en grande dimension. Dans le chapitre 4, nous nous plaçons dans le cadre du modèle à variances isolées et nous considérons le cas où toutes les spikes sont distinctes (i.e. de multiplicité un). Les valeurs propres de la matrice de covariance empirique sont donc

$$\operatorname{spec}(\Sigma) = \sigma^2(\underbrace{\alpha_1^*, \ldots, \alpha_m^*}_{m}, \underbrace{1, \ldots, 1}_{p-m}), \text{ avec } \alpha_1^* > \cdots > \alpha_m^*.$$

Nous supposons de plus que les α^* sont plus grand que $1 + \sqrt{c}$, et donc que toutes les valeurs propres α des facteurs sont plus grandes que $\sigma^2\sqrt{c}$. Pour tout $\alpha \neq 1$, on définit la fonction suivante

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

Baik & Silverstein (2006) ont prouvé, sous une condition de moment sur x, que pour tout $k \in \{1, \ldots, m\}$ et presque-sûrement,

$$\lambda_{n,k} \quad \longrightarrow \quad \sigma^2\phi(\alpha_k^*). \tag{4}$$

Ils ont aussi prouvé, pour tout $1 \leq i \leq L$, où L est un rang prédéfini, que presque-sûrement,

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

Notre méthode d'estimation de m est basée sur une étude approfondie des différences entre deux valeurs propres consécutives

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

En effet, les résultats cités plus haut impliquent que $\delta_{n,j} \to 0$ p.s., pour tout $j \geq m$ tandis que pour $j < m$, $\delta_{n,j}$ tend vers une limite strictement positive. On pourra donc estimer m à partir de l'indice j où $\delta_{n,j}$ devient petit. Plus précisément, notre estimateur est défini par

$$\hat{m}_n \quad = \quad \min\{j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n\}, \tag{5}$$

où $s > m$ est un entier fixé suffisamment grand et d_n est un seuil à définir. En pratique, l'entier s doit être vu comme une borne préliminaire du nombre maximal de facteurs. Nous faisons de plus l'hypothèse suivante, que vérifient les vecteurs gaussiens

Hypothèse 1. Les coordonnées y_{ij} du vecteur aléatoire y possèdent une loi symétrique et ont une décroissance sous-exponentielle, i.e. il existe deux constantes positives C, C' telles que, pour tout $t \geq C'$,

$$\mathbb{P}(|\mathsf{y}_{ij}| \geq t^C) \leq e^{-t}.$$

Nous prouvons alors le théorème suivant

Théorème 1. *Soient* $(\mathsf{x}_i)_{1 \leq i \leq n}$ *n copies i.i.d. du vecteur* $\mathsf{x} = E\Sigma^{\frac{1}{2}}\mathsf{y}$, *où* $\mathsf{y} \in \mathbb{R}^p$ *est un vecteur aléatoire de moyenne nulle et de coordonnées qui vérifient l'hypothèse 1 et* E *une* *matrice orthogonale. On suppose que*

$$\Sigma = cov(\mathrm{x}) = \sigma^2 \begin{pmatrix} V_m & 0 \\ 0 & I_{p-m} \end{pmatrix}$$

où V_m possède m valeurs propres non nulles et différentes de un : $\alpha_1^* > \cdots > \alpha_m^* > 1 + \sqrt{c}$.
On suppose que $\frac{p}{n} \to c > 0$ quand $n \to +\infty$.
Soit $(d_n)_{n\geq 0}$ une suite réelle telle que $d_n \to 0$ et $n^{2/3}d_n \to +\infty$. Alors l'estimateur $\hat{m}_n$ est consistant, i.e $\mathbb{P}(\widehat{m}_n = m) \to 1$ quand $n \to +\infty$.

La démonstration utilise le fait que $\lambda_{n,j} - \phi(\alpha_j^*)$ possède une loi limite (Paul (2007); Bai & Yao (2008)), ainsi que la proposition 5.8 de Benaych-Georges et al. (2011), qui montre que la suite $n^{\frac{2}{3}}(\lambda_{n,m+i} - b)$ est tendue pour $i \geq 1$.

Dans un premier temps, nous supposons que σ^2 est connu et égal à un (dans le cas contraire, il suffit de diviser les valeurs propres $\lambda_{n,j}$ par σ^2). Nous effectuons plusieurs simulations, en prenant pour seuil d_n la suite $4n^{-2/3}\beta\sqrt{2\log\log n}$, où $\beta = (1+\sqrt{c})(1+c^{-1/2})^{1/3}$. Nous considérons ensuite le cas où σ^2 n'est pas connu et doit être estimé. L'estimateur considéré est

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

Comme m n'est pas connu, nous construisons un algorithme qui prend en compte ce fait. Nous effectuons ensuite des simulations pour comparer notre méthode à deux autres existantes : celles de Harding (2007) dans un contexte d'économétrie, et celle de Kritchman & Nadler (2008) en traitement du signal. Ces deux auteurs se placent dans le cadre équivalent du modèle à facteurs. En faisant varier plusieurs paramètres du modèle, notre méthode donne des résultats similaires et parfois meilleurs. Nous terminons le chapitre 4 en abordant la question du cas d'égalité (spikes multiples) qui sera développée dans le chapitre 5, puis par une discussion sur le choix de la suite d_n.

Le chapitre 5 se place dans le cadre du modèle à facteurs où toutes les spikes ne sont pas nécessairement simples. Supposons qu'il y ait K spikes différentes, chacune d'entres elles apparaissant n_k fois (i.e. de multiplicité n_k). Dans ce cas, le spectre de Σ est,

$$\begin{aligned}
\mathrm{spec}(\Sigma) &= (\underbrace{\alpha_1,\ldots,\alpha_1}_{n_1},\ldots,\underbrace{\alpha_K,\ldots,\alpha_K}_{n_K},\underbrace{0,\ldots,0}_{p-m}) + \sigma^2(\underbrace{1,\ldots,1}_{p}) \qquad (6)\\
&= \sigma^2(\underbrace{\alpha_1^*,\ldots,\alpha_1^*}_{n_1},\ldots,\underbrace{\alpha_K^*,\ldots,\alpha_K^*}_{n_K},\underbrace{1,\cdots,1}_{p-m}). \qquad (7)
\end{aligned}$$

avec $n_1 + \cdots + n_k = m$ et $\alpha_i^* = \frac{\alpha_i}{\sigma^2} + 1$. Quand toutes les spikes ne sont pas égales, les différences entre les valeurs propres de la matrice de covariance empirique correspondant aux spikes tendent vers une constante positive, tandis qu'avec deux spikes égales, cette différence va tendre vers zéro : cela crée une confusion avec les différences de valeurs propres qui ne sont pas perturbées, qui tendent aussi vers zéro. Cependant, la convergence des $\delta_{n,i}$ pour $i > m$ (bruit) est plus rapide (en $O_{\mathbb{P}}(n^{-2/3})$) que celle des $\delta_{n,i}$ provenant de spikes égales (en $O_{\mathbb{P}}(n^{-1/2})$) : ceci est une conséquence du théorème 3.1 de Bai & Yao (2008), et c'est l'élément clef pour l'adaptation de l'estimateur (5) à cette nouvelle utilisation, en utilisant un nouveau seuil d_n. Le résultat de consistance est le suivant

Théorème 2. *Soient* $(\mathsf{x}_i)_{1\leq i\leq n}$ *n copies i.i.d. de* x *qui suit le modèle à facteurs (2) et vérifie l'hypothèse 1. On suppose que la matrice de population Σ possède K valeurs propres non nulles et différentes de un :* $\alpha_1^* > \cdots > \alpha_K^* > 1+\sqrt{c}$, *de multiplicités respectives* $(n_k)_{1\leq k\leq K}$ $(n_1+\cdots+n_K = m)$*, et $p-m$ valeurs propres de valeur un. On suppose que* $\frac{p}{n} \to c > 0$ *quand $n \to +\infty$. Soit* $(d_n)_{n\geq 0}$ *une suite réelle telle que* $d_n = o(n^{-1/2})$ *et* $n^{2/3}d_n \to +\infty$. *Alors l'estimateur $\hat{m}_n$ est consistant, i.e $\hat{m}_n \to m$ en probabilité quand $n \to +\infty$.*

On peut remarquer que, comparé au théorème 1, la seule modification de l'estimateur porte sur la vitesse de convergence de d_n qui doit être en $o(n^{-1/2})$. La démonstration est similaire à celle du théorème 1, en utilisant en plus le théorème 3.1 de Bai & Yao (2008). Nous effectuons ensuite plusieurs simulations, en utilisant une version modifiée de l'estimateur $\hat{m}_n$, à savoir

$$\hat{m}_n^* \quad = \quad \min\{j \in \{1,\ldots,s\} : \delta_{n,j+1} < d_n \text{ et } \delta_{n,j+2} < d_n\}. \tag{8}$$

Au lieu de s'arrêter dès qu'une différence $\delta_{n,k}$ est en-dessous du seuil d_n, l'estimateur modifié s'arrête lorsque deux différences consécutives $\delta_{n,k}$ et $\delta_{n,k+1}$ sont toutes les deux plus petites que d_n. Il est facile de voir que cet estimateur est toujours consistant.

Nous modifions aussi la suite utilisée pour le seuil d_n par rapport au chapitre 4 : nous prenons une suite de la forme $Cn^{-2/3}\sqrt{2\log\log n}$, avec C un paramètre à ajuster. Nous effectuons ensuite de nombreuses simulations afin de vérifier la qualité de notre estimateur, et nous prenons le C qui donne de meilleurs résultats. Nous comparons ensuite notre estimateur $\hat{m}_n^*$ à la méthode KN de Kritchman & Nadler (2008) : notre algorithme présente de meilleures performances dans la plupart des cas. Cependant, il faut noter que l'estimateur KN a été construit de manière à minimiser la probabilité de surestimation tandis que le nôtre cherche à minimiser l'erreur globale. C'est pour cela que nous étudions ensuite l'influence de la constante C sur la probabilité de surestimation et que nous observons que cette dernière n'est pas constante et plus grande que celle de la méthode KN (fixée à $\gamma = 0.5\%$). Finalement, comme la constante C a été choisie au cas par cas, nous construisons une méthode afin que cette dernière soit déterminée de manière auto-adaptative. Des simulations montrent une légère dégradation des performances de notre estimateur avec ce nouveau choix.

Corrections de quelques statistiques basées sur la vraisemblance dans un modèle à facteurs strict de grande dimension

Le chapitre 6 considère le modèle à facteurs strict avec une variance homoscédastique. Comme pour les chapitres précédents, cela revient à prendre un échantillon gaussien $\mathsf{x}_1,\ldots,\mathsf{x}_n$ indépendant dont la matrice de population Σ possède la représentation spectrale (7).

Une théorie basée sur la vraisemblance est bien connue depuis Lawley (1940) dans le cadre asymptotique classique. Les estimateurs du maximum de vraisemblance sont les suivants (Anderson & Rubin (1956)) :

$$\widehat{\sigma}^2 = \frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_{n,i}, \tag{9}$$

$$\widehat{\Lambda}_k = \left(\lambda_{n,k} - \widehat{\sigma}^2\right)^{\frac{1}{2}} v_{n,k},\ 1 \leq k \leq m, \tag{10}$$

où $v_{n,k}$ est le vecteur propre normalisé de $\mathsf{S_n}$ correspondant à $\lambda_{n,k}$, pour $1 \leq k \leq p$.

Dans le cadre classique, où p est petit et fixé tandis que la taille de l'échantillon n tend vers l'infini, la convergence presque-sûre de ces estimateurs est bien établie, ainsi que leur normalité asymptotique (Anderson & Amemiya (1988)). Ce n'est plus le cas quand p est grand comparé à n. Des résultats de la théorie des matrices aléatoires permettent de résoudre ce problème.

Dans un premier temps, nous considérons l'estimateur (9) $\widehat{\sigma}^2$ de la variance commune σ^2. Nous démontrons sa normalité dans le cadre de la grande dimension, et exhibons un biais négatif, qui n'existe pas dans le cadre classique, mais qui a été observé dans Kritchman & Nadler (2008, 2009) par exemple. Le théorème est le suivant

Théorème 3. *Nous supposons que les composantes x_{ij} des vecteurs $(\mathsf{x}_i)_{1\leq i\leq n}$ sont des variables aléatoires centrées telles que $\mathbb{E}(|x_{ij}|^4) = 3\sigma^4$ et de matrice de covariance cov$(\mathsf{x}_i) = \Sigma$. Nous supposons de plus que $\frac{p}{n} \to c > 0$ quand $n \to +\infty$. Alors,*

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

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

Le biais est asymptotiquement nul. La variance asymptotique reste la même que dans le cadre classique. La démonstration utilise le théorème central limite pour statistiques spectrales linéaires de Bai & Silverstein (2010) (théorème 9.10), ainsi que le résultat de convergence presque-sûre (4) de Baik & Silverstein (2006). Nous illustrons ce résultat par des simulations numériques. Ensuite, nous évaluons les performances de l'estimateur plug-in

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

Cet estimateur sans biais donne de bien meilleurs résultats que l'estimateur $\hat{\sigma}^2$.

Dans un second temps, nous nous intéressons au test du rapport de vraisemblance d'adéquation à un modèle à facteurs strict à variance homoscédastique. L'hypothèse nulle est

$$\mathcal{H}_0 :\ \Sigma = \Lambda\Lambda' + \sigma^2\mathsf{I}_p,$$

où le nombre de facteurs m est donné. La statistique du rapport de vraisemblance est (Anderson & Rubin (1956))

$$T_n = -nL^*,$$

où

$$L^* = \sum_{j=m+1}^{p} \log \frac{\lambda_{n,j}}{\hat{\sigma}^2},$$

et $\hat{\sigma}^2$ est l'estimateur de la variance (9). En gardant p fixé et en faisant tendre n vers l'infini, la théorie classique montre que T_n converge vers une loi χ^2_q, avec $q = p(p+1)/2 + m(m-1)/2 - pm - 1$. Cette approximation n'est plus valide en grande dimension. En particulier, ce test devient biaisé puisque son niveau est beaucoup plus élevé que celui fixé.

En utilisant à nouveau le théorème 9.10 de Bai & Silverstein (2010), et les calculs effectués dans Bai et al. (2009) et Zheng (2012), nous construisons une version corrigée de la statistique T_n. Plus précisément, nous démontrons le théorème suivant

Théorème 4. *Sous les mêmes hypothèses que le théorème 3, mais avec $c < 1$, on a*

$$v(c)^{-\frac{1}{2}}(L^* - m(c) - ph(c_n, \tilde{H}_n) + \eta + (p-m)\log(\beta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où

- $m(c) = \frac{\log(1-c)}{2}$;
- $h(c_n, \tilde{H}_n) = \int \log(x)\, \mathrm{d}F_{c_n,\tilde{H}_n}(x)$, *avec* $\tilde{H}_n = \frac{p-m}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^m \delta_{\frac{\alpha_i}{\sigma^2}+1}$;
- $\eta = \sum_{i=1}^m \log((\alpha_i+1)(1+c\sigma^2\alpha_i^{-1}))$;
- $\beta = 1 - \frac{c}{p-m}(m + \sigma^2\sum_{i=1}^m \alpha_i^{-1})$;
- $v(c) = -2\log(1-c) + \frac{2c}{\beta}\left(\frac{1}{\beta} - 2\right)$.

En grande dimension, T_n ne converge plus vers une loi du χ^2, mais vers une loi normale. On utilisera la statistique $v(c)^{-\frac{1}{2}}(L^* - m(c) - ph(c_n, \tilde{H}_n) + \eta + (p-m)\log(\beta)))$ pour tester $\mathcal{H}_0$. Ce test sera asymptotiquement normal. En pratique, nous avons besoin d'une expression pour $h(c_n, \tilde{H}_n)$. Nous conjecturons la valeur de cette intégrale et l'utilisons ensuite dans les simulations numériques.

Notre test produit des niveaux proches du théorique, sauf quand la limite c du rapport p/n est plus petite que 0.1. Par contre, le test classique devient rapidement biaisé produisant des niveaux beaucoup plus élevés que celui de référence quand c se rapproche de un, ce qui fait que l'hypothèse nulle n'est jamais acceptée quand p est grand.

Dans la dernière partie de ce chapitre, nous construisons un test de l'égalité de deux spikes ou, de manière équivalente, de la norme de deux vecteurs de pondération. Le but est d'effectuer le test suivant

$$\mathcal{H}_0 : \alpha_i = \alpha_{i+1} \qquad v.s. \qquad \mathcal{H}_1 : \alpha_i \neq \alpha_{i+1}, \tag{11}$$

où $i \leq m-1$. Pour construire ce test, nous utilisons le théorème 3.1 de Bai & Yao (2008), qui donne la loi limite jointe de

$$\{\sqrt{n}(\lambda_{n,j} - \phi(\alpha_k^*)),\ j \in J_k\} \tag{12}$$

où $J_k = \{s_{k-1}+1, \ldots, s_k\}$, $s_i = n_1 + \cdots + n_i$ pour $1 \leq i \leq K$. Dans le cas gaussien réel, c'est la loi des valeurs propres d'une matrice de Wigner gaussienne réelle. Cela permet d'obtenir la distribution limite m_{n_2} de $\sqrt{n}(\lambda_{n,i} - \lambda_{n,i+1})$ pour $1 \leq i < m$ et dans le cas où les spikes correspondantes sont simples. Pour le cas multiple, nous utilisons la conjecture de Wigner sur l'espacement des valeurs propres d'une matrice de Wigner gaussienne réelle pour obtenir m_{n_k}. On peut donc utiliser la statistique $D_{n,i} = \sqrt{n}(\lambda_{n,i} - \lambda_{n,i+1})$ pour conduire le test : sous $\mathcal{H}_0$, $D_{n,i}$ a une densité m_{n_k} pour un certain n_k tandis que sous $\mathcal{H}_1$, $D_{n,i}$ est équivalente à $\sqrt{n}(\phi(\alpha_i) - \phi(\alpha_{i+1}))$, et donc tend vers l'infini quand $n \to +\infty$. Pour $t > 0$, la p-valeur de ce test est :

$$pv(t) = e^{-\frac{t^2}{4s_k^2}}.$$

Nous effectuons ensuite des simulations pour vérifier les performances de ce test : l'erreur de première espèce est proche du seuil théorique, cependant la puissance diminue quand les valeurs de deux spikes différentes sont proches, ce qui était prévisible.

Conclusion et perspectives

Nous concluons cette thèse par quelques perspectives de travail. La première d'entre-elles concerne l'extension du résultat de convergence de notre estimateur du nombre de facteurs au modèle à variances isolées généralisé, en utilisant les résultats de convergence presque-sûre énoncés dans Bai & Yao (2012). La difficulté concernera le cas d'égalité.

La seconde consiste à développer un estimateur de la multiplicité des spikes à partir du test d'égalité. En utilisant les p-valeurs des tests consécutifs, on pourrait établir une partition des spikes en fonction de leur multiplicité.

Enfin, on peut se poser la question de la correction de l'estimateur du maximum de vraisemblance des vecteurs de pondération. Cela revient à étudier le comportement des vecteurs propres de la matrice de covariance empirique en fonction de ceux de la matrice de population. Quelques résultats existent, comme ceux de Benaych-Georges & Nadakuditi (2011), mais il est difficile de trouver un meilleur estimateur que les vecteurs propres de la matrice de covariance empirique.

Contents

Introduction

Random matrix theory has been considerably developed over the past few years. This thesis will consider the large dimensional framework. High-dimensional random matrices allow this particularity to be taken into account, since most asymptotic results assume that the matrix size tends to infinity. Many of these results concern the empirical covariance matrix $\mathsf{S}_n = \frac{1}{n}\mathsf{X}\mathsf{X}'$, where $\mathsf{X} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ is a n-sample of random vectors of dimension p (generally Gaussians). In particular, several limiting theorems deal with the eigenvalues of S_n when p and n tend to infinity proportionally (see Anderson et al. (2010); Bai & Silverstein (2010)). The latter provides fundamental tools for the study of usual statistics, since most of them are functions of the eigenvalues of the sample covariance matrix S_n.

In this work, we are interested in the spiked population model, introduced by Johnstone (2001), where all the eigenvalues of the population covariance matrix Σ are equal, except for a relatively small number among them, called “spikes”. This model covers the strict factor model as a particular case.

This thesis is divided into six chapters. First three are introductory chapters. In the first chapter, we briefly present random matrix theory. More precisely, we recall general results regarding the spectral analysis of random matrices using the Stieltjes transform. We describe the Marčenko-Pastur distributions, and give the central limit theorem for linear spectral statistics of Bai & Silverstein (2004). Then we consider spiked population models and recall the associated results of Baik & Silverstein (2006) and Bai & Yao (2008). The second chapter presents the factor models and the associated maximum likelihood theory. We finally review, in chapter 3, several standard methods for the factors number estimation, in the classical framework as well as in the large dimensional one.

The remaining chapters present new contributions of this thesis. In particular, chapters 4 and 5 describe a new estimation method for the factors/spikes number in the high-dimensional setting, using the convergence results of Baik & Silverstein (2006) and Bai & Yao (2008). The considered estimator uses the eigenvalue behavior of the sample covariance matrix S_n which differs depending on whether they correspond to spikes or not. The estimator is based on differences between consecutive eigenvalues of S_n. Chapter 4 establishes the consistency of the estimator in the case where all the spikes are different and compares it to two existing methods through extensive simulations. The estimator depends on a threshold d_n which should satisfy some conditions. In chapter 5, we extend our result of consistency to the equality case and improve our estimator by changing the threshold.

In chapter 6, we first consider the maximum likelihood estimator in a strict factor model with homoscedastic variance. We correct the estimator of the common variance in the large dimensional context by evaluating its bias and establishing its asymptotic distribution. Then we present a corrected version of the likelihood ratio statistic for the goodness-of-fit test and find its asymptotic distribution. Finally, we propose a test for the equality of two spikes or, equivalently, of the equality of the norm of two factor scores.

Chapter 1

Large dimensional data and random matrices

1.1 Large dimensional data

In multivariate statistics, we observe a random sample of p-dimensional observations $\mathsf{x}_1, \ldots, \mathsf{x}_n$. The statistical methods, such as the principal components analysis, were developed at the beginning of the 20th century. Although some non-asymptotic methods exist in the Gaussian case (Student or Fisher test for instance), results mostly consider an asymptotic framework, where the number of observations n grows to infinity.

Most of these results assume that the dimension p of the variables is fixed and "small" (less than ten generally), whereas the number of observations n tends to infinity. This is the classical asymptotic theory. This theory has been adopted by the practitioners, but recently they have been faced with a new problem, the analysis of high-dimensional data.

For a variety of reasons, these high-dimensional data appeared in a lot of scientific fields. In the genetic field, thanks to the micro-array techniques, it is possible to record the expression level of several thousands of genes from a single tissue. In finance, thanks to the constant evolution of computing and the generalization of the Internet, each day we can possibly have several gigabytes of data, from different markets around the world. Other examples include wireless communications which could have a large number of users or antennas, or the physics of mixture. In Table 1.1, the dimension of several types of data, as well as the commonly associated sample sizes, are presented. We remark that the dimension p of the data is quite far away from classical situations where p is lower than ten. This new type of data is called "large dimensional data".

Table 1.1: Examples of large dimensional data.

	Data size p	Sample size n	$c = p/n$
Portfolio	50	500	0.1
Climate surveys	320	600	0.21
Speech analysis	$a \times 10^2$	$b \times 10^2$	$\simeq 1$
ORL face data base	1440	320	4.5
Micro-arrays	1000	100	10

When the dimension of the data p becomes large, several well-known methods become inefficient or even misleading. A seminal example is provided in Dempster (1958), where he establishes the inefficiency of the Hottelling's T^2 test statistic in such cases and provides a remedy (named as non-exact test). However, by that time no statistician was able to discover the fundamental reasons for such break-down of the well-established methods.

A new area in asymptotic statistics has been since then developed, where the data dimension p is not fixed anymore but tends to infinity together with the sample size n. This is the scheme of large dimensional asymptotics. For multivariate analysis, the problem is therefore, which of the large sample size scheme and the large dimensional scheme is closer to reality? As explained in Huber (1973), some statisticians might say that five samples for each parameter in average are enough for using large sample asymptotic results. Now, suppose there are $p = 20$ parameters and we have a sample of size $n = 100$. We may consider the case as $p = 20$ being fixed and n tending to infinity (large sample asymptotics: classical setting), or $p = 2\sqrt{n}$ or $p = 0.2n$ for instance (large dimensional asymptotics: high-dimensional setting). So, we have at least three different options to choose for an asymptotic setup. The natural question is then, which setup is the best choice among the three? Huber (1973) strongly suggested to study the situation of increasing dimension together with the sample size in linear regression analysis.

This situation occurs in many cases. In parameter estimation for a structured covariance matrix, simulation results show that parameter estimation becomes very poor when the number of parameters is larger than four. Also, it has been found that in linear regression analysis, if the covariates are random (or have measurement errors) and the number of covariates is larger than six, the behavior of the estimates departs far away from the theoretical values, unless the sample size is very large. In signal processing, when the number of signals is two or three and the number of sensors is more than ten, the traditional MUSIC (Multivariate Signal Classification) approach provides very poor estimation of the number of signals, unless the sample size is larger than a thousand. Paradoxically, if we use only half of the data set, namely, we use the data set collected by only five sensors, the signal number estimation is almost 100% correct if the sample size is larger than two hundred. The underlying reason of this paradox is the following: if the number of sensors (the dimension of data) is p, then one has to estimate p^2 parameters. Therefore, when p increases, the number of parameters to be estimated increases proportional to p^2 while the number of observations $2np$ increases proportional to p. This suggests that one has to revise the traditional MUSIC method if the sensor number is large. For instance, it has been done in Hachem et al. (2012).

Bai & Saranadasa (1996) presented an interesting problem, where they theoretically prove that when testing the difference of means of two high dimensional populations, Dempster (1958) non-exact test is more powerful than Hotellings T^2 test, even when the T^2 statistic is well defined. It is well known that statistical efficiency will be significantly reduced when the dimension of data or number of parameters becomes large. Thus, several techniques of dimension reduction were developed in multivariate statistical analysis. As an example, let us consider a problem in principal component analysis. If the data dimension is ten, one may select three principal components so that more than 80% of the information is reserved in the principal components. However, if the data dimension is a thousand and three hundred principal components are selected, one would still have to face a large dimensional problem. If only three principal components are selected, 90% or even more of the information carried in the original data set could be lost. Now, let us consider another example[1]. Let $\mathsf{x}_1, \dots, \mathsf{x}_n$ be a Gaussian sample $\mathcal{N}(0, \mathsf{I}_p)$ of dimension p, centered and with identity covariance matrix (also called population covariance matrix). The associated sample covariance matrix S_n is defined by

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

An important statistic in multivariate analysis is

$$\mathsf{T}_n = \log(\det \mathsf{S}_n) = \sum_{i=1}^{p} \log \lambda_{n,i},$$

where $(\lambda_{n,j})_{1\leq j\leq p}$ are the eigenvalues of S_n. If p is kept fixed, then $\lambda_{n,j} \to 1$ almost surely as $n \to \infty$ and thus $\mathsf{T}_n \to 0$. Furthermore, by taking a Taylor expansion of $\log(1+x)$, one can show that, for any p fixed

$$\sqrt{\frac{n}{p}}\mathsf{T}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,2).$$

This suggests the possibility that T_n remains asymptotically normal for large p, assuming that $p = O(n)$. However, this is not the case: if we assume that $p/n \to c \in (0,1)$ as $n \to \infty$, using results on empirical spectral distribution of S_n (see Section 1.3.1), it can be proved that, almost surely,

$$\frac{1}{p}\mathsf{T}_n \to \int_a^b \frac{\log x}{2\pi c x}((b-x)(x-a))^{\frac{1}{2}}\,\mathrm{d}x = \frac{c-1}{c}\log(1-c) - 1 := d(c) < 0,$$

where $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$. Thus, almost surely;

$$\sqrt{\frac{n}{p}}\mathsf{T}_n \simeq d(c)\sqrt{np} \to -\infty.$$

Consequently, any test which assumes asymptotic normality of T_n will lead to a serious error.

These examples show that the classical large sample limits are no longer suitable for dealing with large dimensional data. Statisticians must seek out new limiting theorems instead. Thus, the theory of random matrices (RMT) might be one possible method for this aim, and hence, has received more attention among statisticians in recent years. For the same reason, the importance of random matrix theory has found applications in many research areas, such as signal processing, network security, image processing, genetic statistics, stock market analysis, etc.

[1] This example is inspired by the introduction of the book of Bai & Silverstein (2010).

1.2 Random matrix theory

Random matrix theory goes back to the development of quantum mechanics in the 1940s and early 1950s. In quantum mechanics, the energy levels of a system are described by the eigenvalues of an Hermitian operator A on an Hilbert space, called the Hamiltonian. To avoid working with an infinite dimensional operator, it is common to approximate the system by discretization, amounting to a truncation, keeping only the part of the Hilbert space that is important to the problem under consideration. Thus, A becomes a finite, but large dimensional random linear operator. Hence, the limiting behavior of large dimensional random matrices attracts special interest among people working in quantum mechanics and many laws were discovered during this period. For a more detailed review on applications of random matrix theory in quantum mechanics and other related areas, the reader is referred to the book by Mehta (2004).

Since the late 1950s, research on the limiting spectral analysis of large dimensional random matrices has attracted considerable interest among mathematicians, probabilists and statisticians. One pioneering work is due to Wigner (1955, 1958), and deals with the semicircular law for a Gaussian (or Wigner) matrix which states that the empirical spectral distribution of a high-dimensional Wigner matrix tends to a "semicircular law". This work was generalized by Arnold (1967, 1971) and Grenander (1963) in various aspects. Bai & Yin (1988) proved that the empirical spectral distribution of a sample covariance matrix (suitably normalized) tends to the semicircular law when the dimension of the data is small, compared to the sample size. Following the work of Marčenko & Pastur (1967) and Pastur (1972, 1973), the spectral analysis of large dimensional sample covariance matrices was developed by many researchers, including Bai et al. (1986), Grenander & Silverstein (1977), Jonsson (1982), Wachter (1978), Yin (1986) and Yin & Krishnaiah (1983). The following authors have also worked on the empirical spectral distribution of the multivariate Fisher matrix (or more generally of products of random matrices): Bai et al. (1986, 1987), Silverstein (1985), Wachter (1980), Yin (1986) and Yin & Krishnaiah (1983). In the early 1980s, major contributions on the existence of limiting spectral distributions and their explicit forms for certain classes of random matrices were made. In recent years, research on random matrix theory is turning toward second order limiting theorems, such as the central limit theorem for linear spectral statistics, the limiting distributions of spectral spacings and extreme eigenvalues.

Recently, these results have been widely used in statistics. In the signal processing field, the detection of a source by a sensor array is of particular interest. To cope with the high-dimensional setting, large random matrix theory has been applied to signal detection (Combettes & Silverstein (1992); Couillet & Debbah (2010)) and recently to hypothesis testing, see Kritchman & Nadler (2009); Nadakuditi et al. (2008); Nadakuditi & Silverstein; Bianchi et al. (2011); Onatski et al. (2012); Hachem et al. (2012). The book of Couillet & Debbah (2011) shows also how random matrix theory can be applied to a variety of problems in signal processing and wireless communications. In economics, we can cite Harding (2007); Onatski (2009, 2010). More generally, in El Karoui (2005), the author constructs a methodology of testing for white Gaussian noise in time series analysis using random matrix theory. El Karoui (2008) and Bai et al. (2010) deal with the problem of estimating the population spectral distribution from a high-dimensional sample covariance matrix. Ledoit & Wolf (2002); Srivastava (2005); Schott (2007); Bai et al. (2009) propose several procedures in the high-dimensional setting for testing that a covariance matrix is identity, or several covariance matrices are equal.

1.3 Spectral analysis of large dimensional random matrices

1.3.1 Fundamental tools

The aim of this section is to give an idea about the fundamental concepts and tools which will be used in the following chapters.

1.3.1.1 Empirical and limiting spectral distributions

Let $\mathcal{M}_p(\mathbb{C})$ be the set of squared matrices of size p with complex entries.

Definition 1. Let $\mathsf{A} \in \mathcal{M}_p(\mathbb{C})$ and $(\lambda_{n,j})_{1\leq j\leq p}$ be its eigenvalues. Its empirical spectral distribution (ESD) is given by

$$F^{\mathsf{A}} = \frac{1}{p}\sum_{j=1}^{p} \delta_{\lambda_{n,j}},$$

where δ_a is the Dirac mass at point a.

Generally, the empirical spectral distribution F^{A} is a probability measure on $\mathbb{C}$. Its support is included in $\mathbb{R}$ (resp. $\mathbb{R}^+$) if A is Hermitian (resp. Hermitian nonnegative definite). For example, the rotation matrix

$$\mathsf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

has eigenvalues $\pm i$, so we have $F^{\mathsf{A}} = \frac{1}{2}(\delta_i + \delta_{-i})$, which is a measure on $\mathbb{C}$. The symmetry

$$\mathsf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

has eigenvalues ± 1, so $F^{\mathsf{B}} = \frac{1}{2}(\delta_1 + \delta_{-1})$ has its support in $\mathbb{R}$. In the following section, we will often consider Hermitian and nonnegative definite covariance matrices: their empirical spectral distribution will have a support included in $\mathbb{R}^+$.

One of the main problems in random matrix is the study of the limiting behavior of a empirical spectral distributions sequence $(F^{\mathsf{M}_n})_{n\geq 1}$, for a given random matrix sequence $(\mathsf{M}_n)_{n\geq 1}$.

Definition 2. Let $(\mathsf{A}_n)_{n\geq 1}$ be a sequence of $\mathcal{M}_p(\mathbb{C})$. If the sequence of corresponding empirical spectral distributions $(F^{\mathsf{A}_n})_{n\geq 1}$ converges vaguely to a measure F (i.e. for all function ϕ continuous and compactly supported, $F^{\mathsf{A}_n}(\phi) \to F(\phi)$ as $n \to \infty$), F is called the limiting spectral distribution (LSD) of the matrices sequence $(\mathsf{A}_n)_{n\geq 1}$.

If the limiting spectral distribution F is of mass one, the vague convergence becomes the usual weak convergence (or in law), i.e. $F^{\mathsf{A}_n}(\phi) \to F(\phi)$ when $n \to \infty$ for all continuous and bounded functions ϕ.

We are especially interested in sequences of random matrices with dimension p growing to infinity: their study is called "theory of large dimensional random matrices". More precisely, we study the sample covariance matrices. Let $\mathsf{x}_1, \ldots, \mathsf{x}_n$ be a sample of random observations of dimension p. The population covariance matrix is denoted by $\Sigma = \mathrm{cov}(\mathsf{x}_i)$. The sample covariance matrix is defined by

$$\mathsf{S}_n^* = \frac{1}{n} \sum_{i=1}^{n} (\mathsf{x}_i - \bar{\mathsf{x}})(\mathsf{x}_i - \bar{\mathsf{x}})',$$

where $\bar{\mathsf{x}} = n^{-1} \sum_i \mathsf{x}_i$ is the empirical mean of the sample. Most multivariate statistical methods rely on this sample covariance matrix e.g., principle components analysis, multivariate regressions, one-sample or two-sample hypothesis testing, factor analysis, etc.

In spectral analysis of large dimensional random matrices, we generally define the sample covariance matrix as the following

$$\mathsf{S}_n = \frac{1}{n} \sum_{i=1}^{n} \mathsf{x}_i \mathsf{x}_i',$$

as S_n and S_n^* have the same limiting spectral distribution. Indeed, the difference between S_n^* and S_n is a matrix of rank one (see Theorem A.44 of Bai & Silverstein (2010)).

In the spectral analysis of S_n, it is usual to assume that the data size p tends to infinity proportionally to the sample size n, i.e. $\frac{p}{n} \to c \in (0, \infty)$ when p, $n \to \infty$. When we consider sample covariance matrices S_n, the eigenvalues are random variables, and the corresponding empirical spectral distributions $(F^{\mathsf{S}_n})_{n \geq 1}$ are random probability measures on $\mathbb{R}^+$ or, equivalently, a sequence of random variables of measures.

1.3.1.2 The Stieltjes transform

Eigenvalues of a matrix can be viewed as continuous functions of the matrix entries. Nevertheless, these functions do not have closed forms when the matrix size exceeds four. This is the reason why specific tools are needed for their study. There are three important methods employed in this area:

- Moment method;
- Stieltjes transform;
- Orthogonal polynomial decomposition of the exact density of the eigenvalues.

We will consider results obtained via the Stieltjes transform method. We denote by Γ_μ the support of a finite measure μ defined on $\mathbb{R}$. Let

$$\mathbb{C}^+ = \{z \in \mathbb{C} | \Im(z) > 0\}$$

be the open upper half complex plan with positive imaginary part.

Definition 3. Let μ be a finite measure on the real line. Its Stieltjes (or Cauchy) transform is defined as

$$s_\mu(z) = \int \frac{1}{x-z}\mu(\mathrm{d}x),\ z \in \mathbb{C}\backslash\Gamma_\mu.$$

The proofs of the results of this part can be found in Akhiezer (1965) and Kreĭn & Nudel′man (1977).

Proposition 1. *The Stieltjes transform has the following properties:*

1. *s_μ is holomorphic on $\mathbb{C}\backslash\Gamma_\mu$;*
2. *$z \in \mathbb{C}^+$ if and only if $s_\mu(z) \in \mathbb{C}^+$;*
3. *If $\Gamma_\mu \subset \mathbb{R}^+$ and $z \in \mathbb{C}^+$, then $zs_\mu(z) \in \mathbb{C}^+$;*
4. *$|s_\mu(z)| \leq \frac{\mu(1)}{d(z,\Gamma_\mu)\vee|\Im(z)|}$.*

The next result is an inversion result.

Proposition 2. *The total mass $\mu(1)$ can be recovered through the formula*

$$\mu(1) = \lim_{\nu\to\infty} - i\nu s_\mu(i\nu).$$

Moreover, for all continuous and compactly supported $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mu(\phi) = \int_{\mathbb{R}} \phi(x)\,\mu(\mathrm{d}x) = \lim_{\nu\to 0^+}\frac{1}{\pi}\int_{\mathbb{R}} \phi(x)\Im s_\mu(x+i\nu)\,\mathrm{d}x.$$

Especially, for all two continuity points $a < b$ of μ,

$$\mu([a,b]) = \lim_{\nu\to 0^+}\frac{1}{\pi}\int_{\mathbb{R}} \Im s_\mu(x+i\nu)\,\mathrm{d}x.$$

Thus, we can recover the initial measure from its Stieltjes transform.

Proposition 3. *We assume that the following conditions are satisfied for a complex function g:*

1. *g is holomorphic on $\mathbb{C}^+$;*
2. *$g(z) \in \mathbb{C}^+$ for all $z \in \mathbb{C}^+$;*
3. *$\limsup_{\nu\to\infty}|i\nu g(i\nu)| < \infty$.*

Then g is the Stieltjes transform of a real finite positive measure.

Stieltjes transform characterizes the vague convergence of finite measures. It is an important tool for the study of random matrices.

Proposition 4. *A sequence $(\mu_n)_{n\geq 1}$ of probability measures $\mathbb{R}$ converges vaguely to a positive measure μ if and only if their Stieltjes transform $(s_{\mu_n})_{n\geq 1}$ converge to s_μ on $\mathbb{C}^+$.*

In order to obtain the weak convergence of the sequence $(\mu_n)_{n\geq 1}$, one can check the vague convergence using the previous proposition and ensure that the limiting measure μ is a probability measure (i.e. $\mu(1) = 1$), using Proposition 2, or by direct calculation.

The link between Stieltjes transform and random matrix theory is the following: the Stieltjes transform of an empirical spectral distribution F^{A} of an Hermitian squared matrix $\mathsf{A} = (a_{ij})_{1\leq i,j\leq n}$ of size n is given by

$$s_n(z) = \int \frac{1}{x-z} F^{\mathsf{A}}(\mathrm{d}x) = \frac{1}{n}\mathrm{tr}(\mathsf{A} - z\mathsf{I})^{-1},$$

which is the resolvent of the matrix A times $1/n$. Using a formula for the trace of an inverse matrix (Bai & Silverstein (2010), Theorem A.4), we have

$$s_n(z) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{a_{kk} - z - \alpha_k'(\mathsf{A}_k - z\mathsf{I})^{-1}\alpha_k},$$

where A_k is the squared matrix of size $n-1$ obtained by removing the k-th row and the k-th column, and α_k is the k-th column of A without the element k. If the denominator $a_{kk} - z - \alpha_k'(\mathsf{A}_k - z\mathsf{I})^{-1}\alpha_k'$ can be proved to be equal to $g(z, s_n(z)) + o(1)$ for some function g, then a limiting spectral distribution F exists and its Stieltjes transform is given by the solution of the equation

$$s = \frac{1}{g(z,s)}.$$

1.3.2 Marčenko-Pastur distributions

The Marčenko-Pastur distribution F_{c,σ^2} (MP law) of index c and scale parameter σ^2 has the density

$$p_{c,\sigma^2}(x) = \begin{cases} \frac{1}{2\pi x c\sigma^2}\sqrt{(b(c)-x)(x-a(c))} & \text{when } a(c) \leq x \leq b(c), \\ 0 & \text{otherwise,} \end{cases}$$

with $a(c) = \sigma^2(1-\sqrt{c})^2$ and $b(c) = \sigma^2(1+\sqrt{c})^2$, and a supplementary point of mass $1 - 1/c$ at the origin when $c > 1$. The constant c is the ratio of the dimension over the sample size. This distribution has a mean σ^2 and a variance $c\sigma^4$. Its support is an interval of length $b(c) - a(c) = 4c\sigma^2$. When $\sigma^2 = 1$, this distribution is called the standard Marčenko-Pastur distribution, denoted by F_c. Figure 1.1 displays three densities of the standard Marčenko-Pastur law for $c \in \left\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\right\}$. We can notice a behavior close to the squared root function at the boundaries of these densities.

It is easy to see that when c tends to zero, the MP law F_c is reduced to the Dirac mass δ_1. Furthermore, if X_c follows the MP law F_c, then the sequence $\frac{1}{2\sqrt{c}}(\mathsf{X}_c - 1)$ converges weakly to the semicircular law of Wigner.

1.3.2.1 Marčenko-Pastur law for independent vectors without cross-correlations

Marčenko and Pastur have been the first to find the limiting spectral distribution of a high-dimensional sample covariance matrix. Their result has then been extended in various directions (see Marčenko & Pastur (1967)).

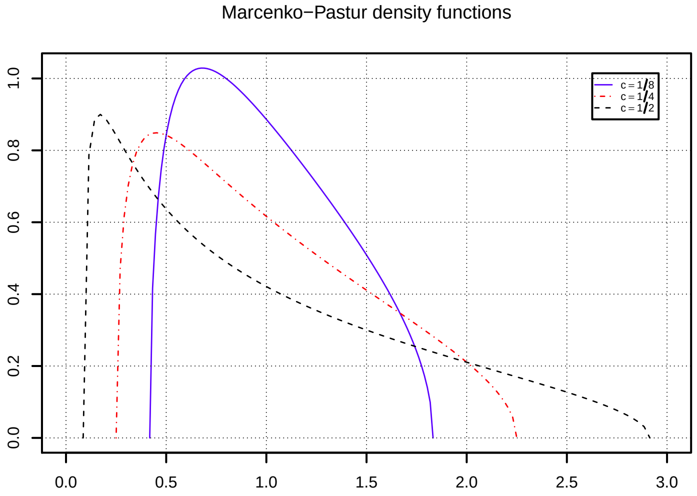

Figure 1.1: Density plots of the standard Marčenko-Pastur law with indexes 1/8, 1/4 and 1/2.

Proposition 5. *Suppose that the coordinates of x_i are complex i.i.d. with mean zero and variance σ^2. We assume that $p/n \to c \in (0, \infty)$. Then, with probability one, F^{S_n} tends to the MP law defined in (1.3.2).*

This theorem was found by the end of the 1960s (for the mean convergence), but its importance in high-dimensional statistics was only recognized at the begin of this century. To understand its deep influence in multivariate analysis, we plot in Figure 1.2 the eigenvalues of the sample covariance matrix of an independent Gaussian sample. We generated $n = 320$ realizations $(\mathsf{x}_i)_{1 \leq i \leq 320}$ of i.i.d. Gaussian random vectors $\mathcal{N}(0, \mathsf{I}_p)$ of size $p = 40$. The histogram of the $p = 40$ eigenvalues of S_n shows a large dispersion from the value one. If we refer to the classical asymptotic theory (assuming $n = 320$ is large enough), the sample covariance matrix should be close to $\Sigma = \mathsf{I}_p = \mathbb{E}(\mathsf{x}_i \mathsf{x}_i')$. As the eigenvalues are continuous functions of the entries of the matrix, the eigenvalues of S_n should converge to one, which is the unique eigenvalue of I_p. The plot clearly shows that this is far away from being the reality. We also draw on the same plot the Marčenko-Pastur density p_c, with $c = 40/320 = 1/8$. The closeness between this density and the histogram of the sample eigenvalues is striking.

Since the sample eigenvalues deviate significantly from the population eigenvalues, the sample covariance matrix S_n is no more a reliable estimator of population covariance matrix Σ anymore. This is the fundamental reason why classical multivariate methods perform poorly when the data size becomes large. As an example, consider the T^2 Hotelling's statistic, linked to S_n^{-1}. In a large dimensional framework (like $p = 40$ et $n = 320$), S_n^{-1} deviates significantly from Σ^{-1}. In this example, the wider spread of the sample eigenvalues can lead to a large number of small eigenvalues, even more if p/n is close to one. For instance, for $\Sigma = \sigma^2 \mathsf{I}_p$ and $c = 1/8$, the smallest eigenvalue of S_n is close to $a(c) = \sigma^2(1 - \sqrt{c})^2 = 0.42\sigma^2$, so the largest eigenvalue of S_n^{-1} is close to $a(c)^{-1}\sigma^{-2} = 1.55\sigma^{-2}$, a 55% over-spread to the population value σ^{-2}. When the ratio of the data size over the sample size increases to

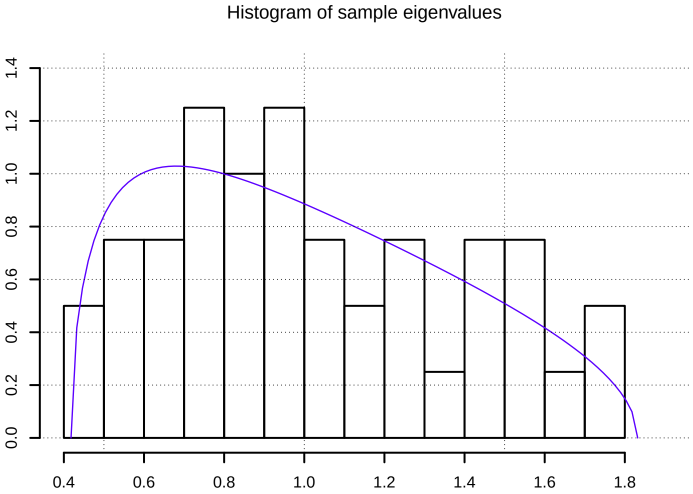

Figure 1.2: Eigenvalues of the sample covariance matrix issued from a Gaussian sample $\mathcal{N}(0, \mathsf{I}_p)$ of dimension $p = 40$ and with size $n = 320$. The curve is the standard Marčenko-Pastur density of index $1/8$.

$c = 0.9$, the largest eigenvalue of S_n^{-1} becomes close to $380\sigma^{-2}$. Therefore S_n^{-1} is clearly not a reliable estimator of Σ^{-1}.

1.3.2.2 Generalized Marčenko-Pastur distributions

In Proposition 5, the population covariance matrix has the simple form $\Sigma = \sigma^2 \mathsf{I}_p$, which is quite restrictive. In order to consider a general population covariance matrix Σ, we assume the following: the observed vectors $(\mathsf{y}_k)_{1 \le k \le n}$ can be expressed as $\mathsf{y}_k = \Sigma^{1/2} \mathsf{x}_k$, where x_k have i.i.d. components as in Proposition 5 and $\Sigma^{1/2}$ is any non-negative squared root of Σ. The associated sample covariance matrix is

$$\mathsf{B}_n = \frac{1}{n} \sum_{k=1}^{n} \mathsf{y}_k \mathsf{y}_k' = \Sigma^{1/2} \left(\frac{1}{n} \sum_{k=1}^{n} \mathsf{x}_k \mathsf{x}_k' \right) \Sigma^{1/2} = \Sigma^{1/2} \mathsf{S}_n \Sigma^{1/2}.$$

Here S_n denotes the sample covariance matrix (1.3.1.1) with i.i.d. components. Notice that the eigenvalues of B_n are the same as the product $\mathsf{S}_n \Sigma$.

The following result extends Proposition 5 to random matrices of type $\mathsf{B}_n = \mathsf{S}_n \Sigma$, for all non-negative matrices Σ.

Proposition 6 (Bai & Silverstein (2010)). *Let S_n be the sample covariance matrix defined in (1.3.1.1) with i.i.d. components and $(\Sigma_n)_{n \ge 1}$ be a sequence of nonnegative Hermitian squared matrices of size p. Let $\mathsf{B}_n = \mathsf{S}_n \Sigma_n$. We assume that:*

1. *The coordinates of x_i are complex i.i.d. with mean zero and variance one;*
2. *The ratio of the data dimension over the sample size $p/n \to c > 0$ as $n \to \infty$;*
3. *The sequence $(\Sigma_n)_{n \ge 0}$ is deterministic, or independent from $(\mathsf{S}_n)_{n \ge 1}$;*

4. *The sequence* $(H_n)_{n\geq 0} = (F^{\Sigma_n})_{n\geq 0}$ *of the empirical spectral distributions of* $(\Sigma_n)_{n\geq 0}$ *converges weakly to a fixed probability measure* H.

Then F^{B_n} *converges weakly to a fixed probability measure* $F_{c,H}$, *whose Stieltjes transform, denoted by* s, *is implicitly defined by the equation*

$$s(z) \quad = \quad \int \frac{1}{t(1-c-czs(z))}\,\mathrm{d}H(t), \tag{1.1}$$

where $z \in \mathbb{C}^+$.

The implicit equation given above has an unique solution in the space of functions from $\mathbb{C}^+$ to $\mathbb{C}^+$. Moreover, the solution s of this equation has no closed-form expression, and this is the unique information that we know about the limiting spectral distribution $F_{c,H}$.

There is, however, another way to present the fundamental equation (1.1). Take the squared matrix of size n

$$\underline{\mathsf{B}}_n = \frac{1}{n}\mathsf{X}'\mathsf{T}\mathsf{X},$$

where X is the matrix made by the vectors $(\mathsf{x}_i)_{1\leq i\leq n}$. The two matrices B and $\underline{\mathsf{B}}$ have the same positive eigenvalues and their empirical spectral distributions satisfy

$$nF^{\underline{\mathsf{B}}_n} - pF^{\mathsf{B}_n} = (n-p)\delta_0.$$

Assuming that $p/n \to c > 0$, F^{B_n} has a limit $F_{c,H}$ if, and only if, $F^{\underline{\mathsf{B}}_n}$ has a limit $\underline{F}_{c,H}$. In this case, the limits satisfy

$$\underline{F}_{c,H} - cF_{c,H} = (1-c)\delta_0,$$

and their respective Stieltjes transform $\underline{s}$ and s are linked to each other by

$$\underline{s}(z) = -\frac{1-c}{z} + cs(z).$$

Replacing s by $\underline{s}$ in (1.1), we find

$$\underline{s} = -\left(z - c\int \frac{t}{1+\underline{s}}\,\mathrm{d}H(t)\right)^{-1}.$$

Then solving this equation with respect to z leads to

$$z \quad = \quad -\frac{1}{\underline{s}} + c\int \frac{t}{1+t\underline{s}}\,\mathrm{d}H(t), \tag{1.2}$$

which indeed gives the inverse function of $\underline{s}$. The equations (1.1) and (1.2) are of fundamental importance in the methods of statistical estimation, and are called “Marčenko-Pastur equations”.

The limiting spectral distribution $F_{c,H}$ and its companion $\underline{F}_{c,H}$ are called “generalized Marčenko-Pastur distributions” with indexes c et H. In the case where $\mathsf{T}_n = \Sigma$, the limiting spectral distribution H of Σ is called “population spectral distribution”.

1.3.3 Central limit theorem for linear spectral statistics of the sample covariance matrix

In multivariate analysis, most of the population statistics can be written as a function of the empirical spectral distribution F_n of some random matrices, i.e.

$$\hat{\theta} \quad = \quad \int f(x)\,\mathrm{d}F_n(x).$$

$\hat{\theta}$ is called a "linear spectral statistic" (LSS), and can be considered as an estimator of $\theta = \int f(x)\,\mathrm{d}F(x)$, where F is the limiting spectral distribution of F_n.

If we consider the sample covariance matrix B_n, we saw in Section 1.3.2.2 that its empirical spectral distribution F_n converges weakly to a generalized Marčenko-Pastur distribution $F_{c,H}$. This consistency is not enough for a better statistical inference, for which a central limit theorem is often required. In this section, we will present the result of Bai & Silverstein (2004).

1.3.3.1 Statement of the theorem

We consider the following linear spectral statistic

$$\widehat{\theta}(f) \quad = \quad \int f(x)\,\mathrm{d}F^{\mathsf{B}_n}(x).$$

As the convergences $c_n \to c$ and $H_n \to H$ can be very slow, the difference

$$p\left(\widehat{\theta}(f) - \int f(x)\,\mathrm{d}F^{c,H}(x)\right)$$

could have no limit. Consequently, we have to consider the limiting distribution of the normalized difference

$$p\left(\widehat{\theta}(f) - \int f(x)\,\mathrm{d}F^{c_n,H_n}(x)\right).$$

In the sequel, we will denote

$$X_n(f) = \int f(x)\,\mathrm{d}G_n(x),$$

where:

$$G_n(x) = p(F^{\mathsf{B}_n}(x) - F_{c_n,H_n}(x)).$$

Proposition 7. *We denote by (x_{jk}) the entries of the vector x_j. We assume:*

(i) For all $\eta \geq 0$,

$$\frac{1}{np}\sum_{j,k}\mathbb{E}(|x_{jk}|^4 \mathbb{1}_{|x_{jk}|\geq \eta\sqrt{n}}) \to 0 \textit{ as } n \to \infty;$$

(ii) For all n, the $x_{ij} = x_{ij}^{(n)}$, $1 \leq i \leq p$, $0 \leq j \leq n$ are independent, and satisfy

$$\mathbb{E}|x_{ij}|^2 = 1,\ \max_{i,j,n}\ \mathbb{E}|x_{ij}|^4 < \infty,\ \frac{p}{n} \to y;$$

(iii) $\mathsf{T}_n \in \mathcal{M}_p(\mathbb{C})$ is nonnegative Hermitian, with bounded spectral norm in p, and there is a cumulative distribution function H such that

$$H_n \equiv F^{\mathsf{T}_n} \xrightarrow{\mathcal{L}} H.$$

Let $f_1, \ldots, f_k$ be analytic functions on an open set of $\mathbb{C}$ which includes the interval

$$[\liminf_n \lambda_{n,\min}^{\mathsf{T}_n} \mathbb{1}_{]0,1[}(y)(1-\sqrt{c})^2, \limsup_n \lambda_{n,\max}^{\mathsf{T}_n}(1+\sqrt{c})^2].$$

Then

(a) The random vectors $(X_n(f_1), \ldots, X_n(f_k))$ are a tight sequence in n;

(b) If x_{ij} and T_n are real, and $\mathbb{E}(x_{ij}^4) = 3$, then

$$(X_n(f_1), \ldots, X_n(f_k)) \xrightarrow{\mathcal{L}} (X_{f_1}, \ldots, X_{f_k}),$$

where $(X_{f_1}, \ldots, X_{f_k})$ is a k-dimensional Gaussian vector with the following mean and covariance

$$\mathbb{E}(X_f) = -\frac{1}{2\pi i}\oint_{\mathcal{C}} f(z) \frac{c\int \frac{\underline{s}(z)^3 t^2\,\mathrm{d}H(t)}{(1+t\underline{s}(z))^3}}{\left(1 - c\int \frac{\underline{s}(z)^3 t^2\,\mathrm{d}H(t)}{(1+t\underline{s}(z))^2}\right)^2}\,\mathrm{d}z,$$

$$cov(X_f, X_g) = -\frac{1}{2\pi^2}\oint_{\mathcal{C}_1}\oint_{\mathcal{C}_2} \frac{f(z_1)g(z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2}\underline{s}'(z_1)\underline{s}'(z_2)\,\mathrm{d}z_1\mathrm{d}z_2,$$

where $\underline{s}(z)$is the Stieltjes transform of $\underline{F}_{c,H} \equiv (1-c)\mathbb{1}_{[0,\infty)} + cF_{c,H}$ ($f, g \in \{f_1, \ldots, f_k\}$), and $\mathcal{C}, \mathcal{C}_1, \mathcal{C}_2$ are closed contours taken in the positive direction in the complex plane, each enclosing the support of $F_{c,H}$;

(c) If x_{ij} is complex with $\mathbb{E}(x_{ij}^2) = 0$ and $\mathbb{E}(|x_{ij}|^4) = 2$, then (b) also holds, except the mean is zero and the covariance function is a half of the function given in (b).

1.3.3.2 Example of application

We consider here an example from Bai et al. (2009), which deals with the problem of testing the covariance of a sample. Let $\mathsf{x} \in \mathbb{R}^p$ be a random variable such that

$$\mathsf{x} \sim \mathcal{N}(0_p, \Sigma_p).$$

We would like to test

$$\mathcal{H}_0 :\ \Sigma_p = \mathsf{I}_p \text{ versus } \mathcal{H}_1 :\ \Sigma_p \neq \mathsf{I}_p.$$

If we want to test $\Sigma_p = \mathsf{A}$, with a given $\mathsf{A} \in \mathcal{M}_p(\mathbb{C})$, we can go back to the null above by the transformation $\mathsf{A}^{-1/2}\mathsf{x}$. Let $(\mathsf{x}_1, \cdots, \mathsf{x}_n)$ be a n-sample of x such that $p < n$ and S_n the sample covariance matrix. We define

$$K^* \quad = \quad \operatorname{tr} \mathsf{S}_n - \log |\mathsf{S}_n| - p. \tag{1.3}$$

The likelihood ratio statistic is $K_n = n.K^*$. When p is fixed and $n \to \infty$, $K_n \xrightarrow{\mathcal{L}} \chi^2_{\frac{1}{2}p(p+1)}$ under $\mathcal{H}_0$. However, when p becomes large, K_n grows to infinity, which leads to a test with higher level than the given one. Thus it is necessary to construct a version of K_n suitable in large dimensional setting. Notice that

$$K^* = \sum_{i=1}^{p} (\lambda_{n,i} - \ln(\lambda_{n,i}) - 1),$$

where $(\lambda_{n,i})_{1 \le i \le p}$ are the eigenvalues of S_n: this is a linear spectral statistic. We will apply Proposition 7 to obtain the asymptotic distribution of K_n in large dimensional setting. By taking $\mathsf{T}_n = \mathsf{I}_p$, B_n becomes S_n. Moreover, we have $H_n = H = F^{\mathsf{T}_n} = \delta_1$, and also $F_{c,H} = F_c$, and $X_n(f) = \int_{\mathbb{R}} f(x)\, \mathrm{d}(F^{\mathsf{S}_n} - F_{c_n})(x)$.

Applying Proposition 7, we obtain the following result

Proposition 8. *We assume that the conditions in Proposition 7 hold, K^* is defined as in* (1.3) *and $g(x) = x - \ln(x) - 1$. Then, under $\mathcal{H}_0$ and when $n \to \infty$,*

$$\widetilde{K_n} = v(c)^{-1/2} \left(K^* - p \int_{\mathbb{R}} g(x)\, \mathrm{d}F_{c_n}(x) - m(c) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where $m(c) = -\frac{\log(1-c)}{2}$, and $v(c) = -2\log(1-c) - 2c$.

In large dimensions, the limiting distribution of K_n is not a χ^2 law anymore, but a Gaussian law. We reproduce here a table from Bai et al. (2009). For different values of p and n, type I errors has been calculated from 10000 independent replications of the real Gaussian distribution. The nominal Type I error is $\alpha = 0.05$. Computations are done for the traditional likelihood ratio test (LRT) and for the corrected likelihood ratio test (CLRT) defined above.

p	n	CLRT			LRT	
		Size	Difference with 5%	Power	Size	Power
5	500	0.0803	0.0303	0.6013	0.0521	0.5233
10	500	0.0690	0.0190	0.9517	0.0555	0.9417
50	500	0.0594	0.0094	1	0.2252	1
100	500	0.0537	0.0037	1	0.9757	1
300	500	0.0515	0.0015	1	1	1

Table 1.2: Sizes and powers of the traditional LRT compared to the corrected LRT.

Powers are estimated under the alternative $\Sigma = \operatorname{diag}(1, 0.05, 0.05, 0.05, \ldots)$. As showed by the Table 1.2, the traditional LRT always rejects $\mathcal{H}_0$ when p is large, for instance for $p = 100$ or 300, whereas the size calculated from the corrected LRT is close to the theoretical size. For intermediate dimensions such as $p = 50$, the corrected LRT still gives good results, whereas the traditional LRT has a size larger than 5%.

1.3.4 Limits of extreme eigenvalues

The smallest and largest eigenvalues of S_n give the spread of the sample eigenvalues. This is the reason why their properties will be important in multivariate analysis. We consider here the case $\Sigma = \mathsf{I}_p$. The following proposition is for the simple case where the components are i.i.d. as in Proposition 5.

Proposition 9. *Let $(x_{ij})_{1\leq i,j\leq n}$ be a double entries array with complex-valued random variables, with mean zero and variance one, and finite fourth moment. Consider the sample covariance matrix S_n defined in (1.3.1.1) where $\mathsf{x}_k = (x_{1k}, \dots, x_{pk})'$ and $\lambda_{n,1} \geq \cdots \geq \lambda_{n,p}$ denote its eigenvalues in a decreasing order. Then, when $p/n \to c > 0$,*

$$\lambda_{n,1} \overset{a.s.}{\to} b = (1+\sqrt{c})^2, \tag{1.4}$$

$$\lambda_{n,\min} \overset{a.s.}{\to} a = (1-\sqrt{c})^2, \tag{1.5}$$

where

$$\lambda_{n,\min} = \begin{cases} \lambda_{n,p} & \textit{for } p \leq n, \\ \lambda_{n,p-n+1} & \textit{otherwise.} \end{cases}$$

The existence of the fourth moment is also a necessary condition for the convergence (1.4), see Bai et al. (1988). What is necessary for the convergence (1.5) is still an open question.

1.4 Spiked population models

1.4.1 Definition of the model

In Section 1.3.2.2, we consider observations of the form $\mathsf{x}_i = \Sigma^{1/2}\mathsf{y}_i$, where y_i are i.i.d. vectors of size p, with mean zero, variance one, and i.i.d. components. $(\mathsf{x}_i)_{i\geq 1}$ is thus a random sequence of i.i.d. vectors with mean zero and population covariance matrix Σ. If we take $\Sigma = \mathsf{I}_p$, then this corresponds to the "null" case, and we saw in 1.3.2 that the limiting spectral distribution of S_n is the standard Marčenko-Pastur law. Nevertheless, as noticed in Johnstone (2001), there are examples of real data which are significantly different from this null case. Several extreme sample eigenvalues are separated from others that are confined in the support of the Marčenko-Pastur distribution. To explain this phenomenon, Johnstone (2001) proposed a "spiked population model", where all the population eigenvalues equal one, except a fixed and relatively small number among them, called "spikes". In other words, the population covariance matrix Σ has the following eigenvalues

$$\underbrace{\alpha_1, \dots, \alpha_1}_{n_1}, \dots, \underbrace{\alpha_K, \dots, \alpha_K}_{n_K}, \underbrace{1, \cdots, 1}_{p-m} \tag{1.6}$$

where $n_1 + \cdots + n_K = m$ is the number of spikes. The spiked population model can be viewed as a finite rank perturbation of the null case.

When $p/n \to c > 0$, it is easy to see that the empirical spectral distribution of S_n still converges to the standard Marčenko-Pastur law. However, the asymptotic behavior of the extreme eigenvalues of S_n will be different from the null case.

Several authors have studied this model: Baik & Silverstein (2006) showed the almost sure convergence of the extreme sample eigenvalues issued from a spiked population model. Paul (2007) established a central limit theorem for the extreme sample eigenvalues corresponding to simple spikes obtained from a Gaussian sample, and gave a result on the related eigenvectors. In the Gaussian Wishart matrices case, the asymptotics of the extreme eigenvalues have been established by Baik et al. (2005), and a transition phase, called "BBP transition" has been revealed: there is a difference in the behavior regarding the value of the perturbation. Benaych-Georges & Nadakuditi (2011) extended these results to other perturbation models, additive or multiplicative, which are more general than spiked population models, and showed the almost sure convergence of appropriate projections of the eigenvectors corresponding to the spikes. Moreover, Benaych-Georges et al. (2011) studied the deviations of the extreme eigenvalues of perturbed matrices.

Bai & Yao (2012) generalized the above model to a "generalized spiked population model". We assume that T_p has the following structure

$$\Sigma_p = \begin{pmatrix} \mathsf{V}_m & \mathsf{0} \\ \mathsf{0} & \mathsf{T}_{p-m} \end{pmatrix}.$$

Moreover, we assume

(i) V_m is squared matrix of size m, where m is a fixed integer. The eigenvalues of V_m are $\alpha_1 > \cdots > \alpha_K > 0$ with respective multiplicities $n_1, \ldots, n_K$ $(m = n_1 + \cdots + n_K)$. We denote by J_k the set of the n_k indexes of α_k in the matrix Σ;

(ii) The empirical spectral distribution H_p of T_{p-m} converges to a limiting non-random distribution H;

(iii) The sequence of the largest eigenvalues of Σ is bounded;

(iv) The eigenvalues $\beta_{n,j}$ of T_{p-m} verify

$$\sup_j d(\beta_{n,j}, \Gamma_H) = \varepsilon_p \to 0,$$

where $d(x, A)$ is the distance from x to a set A and Γ_H is the support of H.

Definition 4. An eigenvalue α of V_m is called a "generalized spike", or simply "spike", if $\alpha \notin \Gamma_H$.

Consequently, the spectrum of the population covariance matrix Σ is composed of a main part, the $\beta_{n,j}$, and a smaller part of m spiked eigenvalues that are well separated from the main part, in the form of Definition 4.

1.4.2 Limits of spiked eigenvalues

We denote by y_{ij} the components of y_j and also assume

(v) $\mathbb{E}y_{ij} = 0$, $\mathbb{E}|y_{ij}^2| = 1$ et $\mathbb{E}|y_{ij}|^4 < +\infty$;

(vi) $p/n \to c > 0$.

Bai & Yao (2012) proved the following result

Proposition 10. *Assume hypothesis (i)-(vi) holds true. Let $\lambda_{n,1} \geq \cdots \geq \lambda_{n,p}$ be the eigenvalues of S_n, and let*

$$\psi(\alpha) = \alpha + c\int \frac{t\alpha}{\alpha - t}\,\mathrm{d}H(t).$$

1. *For a spiked eigenvalue α_k which verifies $\psi'(\alpha_k) > 0$, we have*

$$\lambda_{n,k} \overset{a.s.}{\to} \psi(\alpha_k),\ \forall k \in J_k;$$

2. *We assume that $\psi'(\alpha_k) \leq 0$. Let I be the maximum interval of Γ_H^c including α_k. Then:*

   (a) If I has a sub-interval (u_k, v_k) on which $\psi' > 0$ (we take the larger interval), then

$$\lambda_{n,k} \overset{a.s.}{\to} \psi(w),\ \forall k \in J_k,$$

   where w is the bound of u_k or v_k closest to α_k;

   (b) If for all $\alpha \in I$, $\psi'(\alpha) \leq 0$, then

$$\lambda_{n,k} \overset{a.s.}{\to} \gamma_k,\ \forall k \in J_k,$$

   where γ_k is the γ-th quantile of H, with $\gamma = H(-\infty, \alpha_k)$, and H is the limiting spectral distribution of T_{p-m}.

This proposition distinguishes two different types of spikes, those with positive ψ', called "distant spikes" and the others with a negative ψ', called "closed spikes". Distant spikes are also characterized by $\psi'(\alpha) > 0$ if and only if $\psi(\alpha_j)$ is outside the support of the limiting spectral distribution $F_{c,H}$. Furthermore, this property is highly dependent on the value of c, since a spike can become distant (or stop being distant) if c is sufficiently large.

We now consider the special case where $\mathsf{T}_{p-m} = \mathsf{I}_{p-m}$. In this case we have:

$$\psi(\alpha) = \alpha + c\frac{\alpha}{\alpha - 1},$$

and

$$\psi'(\alpha) = 1 - \frac{c}{(\alpha-1)^2}.$$

Thus a spike α is distant if $\alpha > 1 + \sqrt{c}$ or $\alpha < 1 - \sqrt{c}$. The following result, proved by Baik & Silverstein (2006), is a corollary of Proposition 10 :

Corollary 1. *We assume that $\mathsf{T}_{p-m} = \mathsf{I}_{p-m}$. Under the same hypotheses of Proposition 10, we have*

1. *If $\alpha_k > 1 + \sqrt{c}$, then*

$$\lambda_{n,i} \overset{a.s.}{\to} \alpha_k + c\frac{\alpha_k}{\alpha_k - 1}(= \phi(\alpha_k)),\ \forall i \in J_k;$$

2. If $1 < \alpha_k \leq 1 + \sqrt{c}$*, then*

$$\lambda_{n,i} \overset{a.s.}{\to} (1 + \sqrt{c})^2, \; \forall i \in J_k;$$

3. If $\alpha_k < 1 - \sqrt{c}$ *and* $c < 1$*, then*

$$\lambda_{n,i} \overset{a.s.}{\to} (1 - \sqrt{c})^2, \; \forall i \in J_k;$$

4.

$$\lambda_{n,m+1} \overset{a.s.}{\to} b = (1 + \sqrt{c})^2.$$

Notice that when $c \geq 1$, a distant spike has to be larger than one.

Figure 1.3 illustrates the previous result. It shows the eigenvalues of the sample covariance matrix drawn from a normal law $\mathcal{N}(0, \Sigma_p)$, where $\Sigma_p = \operatorname{diag}(5, 4, 3, 1, \ldots, 1)$. We set $p = 100$ and the sample size is $n = 300$.

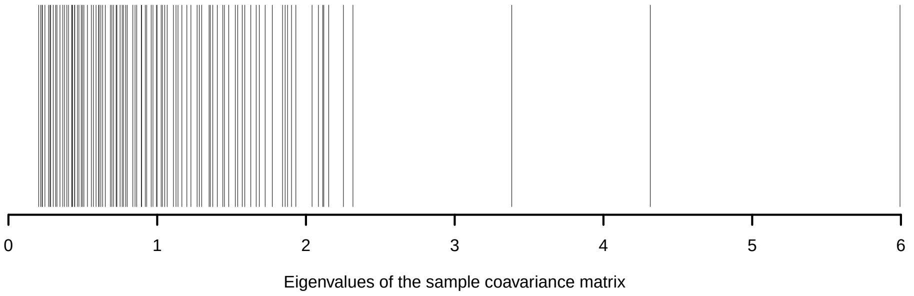

Figure 1.3: Eigenvalues of a sample covariance matrix drawn from a normal law with $\Sigma_p = \operatorname{diag}(5, 4, 3, 1, \ldots, 1)$ and $p = 100$, $n = 300$.

We observe three eigenvalues which stand out from the others. These are the three largest eigenvalues, and they correspond to the spikes. The other eigenvalues remain in the support $[a(c), b(c)]$ of the Marčenko-Pastur law F_c.

1.4.3 Central limit theorem for spiked eigenvalues

In Bai & Yao (2012), the authors proved a central limit theorem for the following vectors of dimension n_k

$$\sqrt{n}(\lambda_{n,j} - \psi(\alpha_k)), \; j \in J_k.$$

We rewrite the observed vectors as $\mathsf{x}_j = \Sigma^{1/2}\mathsf{y}_j$ where $\mathsf{y}_j = (w_{ij})_{1 \leq i \leq p}$, by blocs $\mathsf{x}_j = (\xi_j, \eta_j)'$, with $\xi_j = \mathsf{V}_m^{1/2}(w_{ij})_{1 \leq j \leq m}$ and $\eta_j = \mathsf{T}_{p-m}^{1/2}(w_{ij})_{m \leq j \leq p}$. Let

$$X_1 = \frac{1}{\sqrt{n}}(\xi_1, \ldots \xi_n) = \frac{1}{\sqrt{n}}\xi_{1:n} \text{ et } X_2 = \frac{1}{\sqrt{n}}(\eta_1, \ldots, \eta_n) = \frac{1}{\sqrt{n}}\eta_{1:n}.$$

For $\lambda \notin F_{c,H}$, we define the following random matrix

$$R_n(\lambda) = \frac{1}{\sqrt{n}} \left(\xi_{1:n}(I + A_n)\xi'_{1:n} - \mathsf{V}_m \mathrm{tr}(I + A_n) \right),$$

where $A_n = A_n(\lambda) = X'_2(\lambda I - X_2 X'_2)^{-1} X_2$ for $\lambda \notin F_{c,H}$. Bai & Yao (2008) gives the asymptotic distribution of the sequence $(R_n(\lambda))_{n \geq 1}$. For $\lambda \notin \Gamma_{F_{c,H}}$,

1. If the random variables (w_{ij}) are real valued, the random matrix $R_n(\lambda)$ converges weakly to a symmetric random matrix $R(\lambda) = (R_{ij}(\lambda))$ with centered Gaussian entries and covariance matrix with a known explicit expression;
2. If the random variables (w_{ij}) are complex, the random matrix $R_n(\lambda)$ converges weakly to a centered Hermitian random matrix $R(\lambda) = (R_{ij}(\lambda))$. Moreover, the joint distribution of the real and imaginary parts of the upper-triangular block $(R_{ij})_{1 \leq i\ \leq j \leq m}$ is a $2K$-dimensional Gaussian vector, whose covariance matrix has a known explicit expression.

Take the spectral decomposition of Σ,

$$\Sigma = U \begin{pmatrix} \alpha_1 \mathsf{I}_{n_1} & \cdots & 0 \\ 0 & \ddots & 0 \\ \cdots & 0 & \alpha_K \mathsf{I}_{n_K} \end{pmatrix} U',$$

where U is a unitary matrix. Let $\psi_k = \psi(\alpha_k)$ and $R(\psi_k)$ denote the limiting Gaussian distribution of the sequence of matrices of random forms $(R_n(\psi_k))_n$ described above. Let

$$\tilde{R}(\psi_k) = U' R(\psi_k) U$$

and

$$m_3(\psi_k) = \int \frac{x}{(\psi_k - x)^2} \, \mathrm{d}F_{c,H}(x).$$

Then we have the following result.

Proposition 11. *For each distant spike, the real vector of dimension n_k*

$$\sqrt{n}(\lambda_{n,j} - \psi(\alpha_k)), \; j \in J_k,$$

converges weakly to the distribution of the n_k eigenvalues of the Gaussian matrix

$$\frac{1}{1 + y m_3(\psi_k)\alpha_k} \tilde{R}_{kk}(\psi_k),$$

where $\tilde{R}_{kk}(\psi_k)$ is the k-th diagonal block of $\tilde{R}(\psi_k)$ corresponding to the indexes $\{u, v \in J_k\}$.

It is interesting to notice that the limiting distribution of the n_k sample eigenvalues is generally not Gaussian, and asymptotically dependent. Nevertheless, the limiting distribution of a single eigenvalue $\lambda_{n,i}$ is Gaussian if, and only if, the corresponding spike eigenvalue is simple. Especially, we can recover Theorem 3 of Paul (2007), which consider the real Gaussian case where $\mathsf{T}_{p-m} = \mathsf{I}_{p-m}$, and V_m diagonal with all its eigenvalues simple.

If $\alpha_k \notin [1 \pm \sqrt{c}]$, then

$$\sqrt{n}(\lambda_{n,k} - \phi(\alpha_k)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2_{\alpha_k}),$$

with $\sigma^2_{\alpha_k} = \frac{2\alpha_k^2((\alpha_k-1)^2-c)}{(\alpha_k-1)^2}$.

1.4.4 Fluctuations of the first non-spike eigenvalues when $\mathsf{T}_{p-m} = \mathsf{I}_{p-m}$

The previous result give the precise asymptotic behavior of sample eigenvalues corresponding to the spikes. In Benaych-Georges et al. (2011), the authors prove a result (Proposition 5.8) on fluctuations of the extreme sample eigenvalues for a spiked population model, including the first ones corresponding to non-spikes eigenvalues, i.e. $\lambda_{n,m+i}$, $1 \leq i \leq L$, where L is a prefixed range. The following result is issued from their proposition.

Proposition 12. *We assume the same assumptions as Proposition 11 with $\mathsf{T}_{p-m} = \mathsf{I}_{p-m}$ and that the entries w_{ij} of y_j have a symmetric law and a sub-exponential decay, that means there exists positive constants C, C' such that, for all $t \geq$ C', $\mathbb{P}(|w_{ij}| \geq t^C) \leq e^{-t}$. Then, for all $1 \leq i \leq L$ with a prefixed range L,*

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

where $b = (1 + \sqrt{c})^2$.

Consequently, the convergence of the $\lambda_{n,i}$, for $i > m$ (noise) is faster (in $O_{\mathbb{P}}(n^{-2/3})$) than that of the $\lambda_{n,i}$ from spikes (in $O_{\mathbb{P}}(n^{-1/2})$).

Chapter 2

The factor model

2.1 Introduction

Factor analysis began with the works of Spearman (1904) on the human behavior, completed then by Kelley (1928) and Thurstone (1931), who introduces, among others, the representation of the factorial space and the use of matrix calculation. This method has been developed and diversified thanks to Hotteling (1973) in particular. Then, factor analysis has been commonly used in social sciences, especially in psychology (see Cudeck & MacCallum (2007) for example). Recently, this method has become a tool widely used in macroeconomic where the APT (Arbitrage Pricing Theory) of Ross (1976) and its extension in Chamberlain & Rothschild (1983) rely deeply on the factor model. In wireless communications, the relation between a signal emitted by a source and the received one by the antennas is described by a factor model (see Tulino & Verdú (2004)). The first aim of factor model was to reduce high-dimensional data into a smaller number of common factors and variables are then described by linear combinations of these factors.

2.2 The model

In this section we consider the classical framework, where the size p of the data is kept fixed whereas the sample size n tends to infinity. A factor model is defined as follows. Let p denote the number of variables, n the number of data x_i observed and m the number of common factors. In a strict factor model, we have

$$\begin{aligned} \mathsf{x}_i &= \sum_{k=1}^{m} \mathsf{f}_{ki}\Lambda_k + \mathsf{e}_i + \mu & (2.1) \\ &= \Lambda\mathsf{f}_i + \mathsf{e}_i + \mu, & (2.2) \end{aligned}$$

where:

- $\mu \in \mathbb{R}^p$ represents the general mean;
- $\mathsf{f}_i = (\mathsf{f}_{1i}, \ldots, \mathsf{f}_{mi})'$ are the m random factors, called common factors or factors scores $(m < p)$;

- $\Lambda = (\Lambda_1, \dots \Lambda_m)$ is a $p \times m$ full rank matrix, called factors loadings;
- e_i is the p-dimensional noise vector, centered, independent from f_i and with population covariance matrix $\Psi = \mathbb{E}(\mathsf{e}_i \mathsf{e}_i')$.

The components of e_i are the specific factors, also called unique factors or idiosyncratics factors. The variability not explained by the common factors is represented by the variance of this vector. Classical assumptions on this model are:

- $\mathbb{E}(\mathsf{f}_i) = 0$ et $\mathbb{E}(\mathsf{f}_i \mathsf{f}_i') = \mathsf{I}_p$;
- $\Psi = \text{cov}(\mathsf{e}_i)$ is diagonal;
- $\Gamma = \Lambda' \Psi^{-1} \Lambda$ is diagonal, with diagonal elements ordered and different.

The last hypothesis is used to avoid an identification problem (see Section 2.3.1.1). Consequently, we can describe the factor model by a condition on the population covariance matrix $\Sigma = \text{cov}(\mathsf{x}_i)$

$$\Sigma = \Lambda\Lambda' + \Psi, \tag{2.3}$$

where the diagonal elements of $\Lambda\Lambda'$ are called commonalities and the elements of Ψ are the specificities.

The parameters to be estimated in a factor model are:

- The number of factors m;
- The population covariance matrix Ψ of the noise;
- The matrix of factors loadings Λ.

2.3 The identification constraints

The factor model 2.2 is characterized by an important number of parameters. For their identification and in order to avoid multiple solutions, we need to impose some restrictions on the correlations structure (2.3).

2.3.1 The problem of orthogonal rotations

Let C be an orthogonal squared matrix of size m. Let $\mathsf{f}_i^* = C^{-1}\mathsf{f}_i$ and $\Lambda^* = \Lambda C$. The model (2.2) can be rewritten as

$$\mathsf{x}_i = \Lambda^* \mathsf{f}_i^* + \mathsf{e}_i + \mu,$$

with $\mathbb{E}(\mathsf{f}_i^*) = 0$, $\mathbb{E}(\mathsf{f}_i^* \mathsf{f}_i^{*'}) = \mathsf{I}_p$, and $\Sigma = \Lambda^* \Lambda^{*'} + \Psi$. It is seen that any orthogonal transformation of one solution (Λ, f_i) is also a solution. To solve this identification problem, several solutions have been proposed in the literature. We briefly present three of them.

2.3.1.1 Diagonality of the matrix Γ

We have already outlined this hypothesis in the previous section. It consists of assuming that $\Gamma = \Lambda' \Psi^{-1} \Lambda$ is diagonal, with diagonal elements ordered and different, and it will fix the orthogonal rotation C. This is the widely used hypothesis in the study of maximum likelihood problems. It is invariant by scale changes and Λ will be unique apart from sign. With this constrain, we add $\frac{1}{2}m(m+1)$ conditions on the correlations structure.

2.3.1.2 Simple structure

These are conditions which were proposed in psychology by Thurstone (1931) and consist in choosing the matrix which has a maximum number of zero among the elements of ΛC. This matrix solution can be considered as the matrix which gives the simplest structure and the meaningful psychological interpretation.

2.3.1.3 Zero elements at specified positions

These conditions require prior information from the practitioner. In psychology, the tester needs to know that some specific tests do not depend on specific factors. The corresponding coordinates of these factors are then assumed to be zero. In this case, the hypothesis $\mathbb{E}(\mathsf{f}_i\mathsf{f}_i') = \mathsf{I}_p$ is no longer valid.

2.3.2 The number of parameters

The factor model structure has another identification problem. The number of distinct elements of Σ is $\frac{1}{2}p(p+1)$, and the number of free parameters is $m(1+p)$ for Ψ and Λ. From them we need to remove $\frac{1}{2}m(m+1)$ elements fixed by the diagonality constraint of Γ. The uniqueness of the solution is given when the difference $q = \frac{(p-m)^2-p-m}{2}$ between the number of equations and conditions, minus the number of unknown, is positive. When $q = 0$, we will have the same number of parameters and equations, whereas when $q > 0$, there will be more equations than parameters. In this case, the factor model will be a simplification compared to the unique observation of the population covariance matrix.

2.4 Maximum likelihood estimation

If the common factors f_i and the idiosyncratic factors e_i are Gaussian, a likelihood based theory is well-known since Lawley (1940) (see also Lawley & Maxwell (1971)). We assume here that the number of common factors m is known. In this case, the vector of observations x follows a normal distribution $\mathcal{N}(\mu, \Sigma)$, where $\Sigma = \Lambda\Lambda' + \Psi$. Le $\mathsf{x}_1, \ldots, \mathsf{x}_n$ be an n-sample of x. We review the calculus done in Anderson (2003). The likelihood of this sample is

$$\mathcal{L} \quad = \quad (2\pi)^{-\frac{pn}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathsf{x}_i-\mu)'\Sigma^{-1}(\mathsf{x}_i-\mu)\right).$$

The maximum likelihood estimator of μ is $\bar{\mathsf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i$. We denote by S_n the sample covariance matrix. Thus we have

$$\mathcal{L} \quad = \quad (2\pi)^{-\frac{pn}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\mathrm{tr}(\mathsf{S}_n\Sigma^{-1})\right).$$

In order to obtain the maximum likelihood estimator, we have to maximize the logarithm of $\mathcal{L}$

$$f(\Lambda, \Psi) = -\frac{n}{2}\left(\log|\Sigma| + \mathrm{tr}(\mathsf{S}_n\Sigma^{-1})\right) - \frac{pn}{2}\log(2\pi).$$

So it is the same as solving the following equations simultaneously

$$\begin{aligned} \frac{\partial f(\Lambda,\Psi)}{\partial\Lambda} &= -\frac{n}{2}\left(\Sigma^{-1}\Lambda - \Sigma^{-1}\mathsf{S}_n\Sigma^{-1}\Lambda\right) = 0, \\ \frac{\partial f(\Lambda,\Psi)}{\partial\Psi} &= -\frac{n}{2}\mathrm{diag}\left(\Sigma^{-1} - \Sigma^{-1}\mathsf{S}_n\Sigma^{-1}\right) = 0. \end{aligned}$$

These equations simplify as

$$\begin{aligned} \Lambda &= \mathsf{S}_n\Sigma^{-1}\Lambda, & (2.4) \\ \mathrm{diag}(\Lambda\Lambda' + \Psi) &= \mathrm{diag}(\mathsf{S}_n). & (2.5) \end{aligned}$$

By definition of Σ, we have $\Sigma^{-1}\Lambda = \Psi^{-1}\Lambda(\mathsf{I}_p + \Lambda'\Psi^{-1}\Lambda)^{-1}$. Finally we obtain

$$\begin{aligned} \Lambda(\Gamma + \mathsf{I}_p) &= \mathsf{S}_n\Psi^{-1}\Lambda, & (2.6) \\ \mathrm{diag}(\Lambda\Lambda' + \Psi) &= \mathrm{diag}(\mathsf{S}_n), & (2.7) \end{aligned}$$

and the diagonality constraint (2.3.1.1) of the matrix $\Gamma = \Lambda'\Psi^{-1}\Lambda$. Let $\tilde{\Lambda} = \Psi^{-\frac{1}{2}}\Lambda$ and $\tilde{\mathsf{S}}_n = \Psi^{-\frac{1}{2}}\mathsf{S}_n\Psi^{-\frac{1}{2}}$. Equation (2.6) can be rewritten as

$$\tilde{\Lambda}(\mathsf{I}_p + \tilde{\Lambda}'\tilde{\Lambda}) = \tilde{\mathsf{S}}_n\tilde{\Lambda}.$$

This equation shows that the column of $\tilde{\Lambda}$ are the eigenvectors of the matrix $\tilde{\mathsf{S}}_n$, and the diagonal elements of Γ are the corresponding eigenvalues. Let $\tilde{\lambda}_{n,1} \geq \cdots \geq \tilde{\lambda}_{n,p}$ be the eigenvalues of $\tilde{S}_n$ and $\tilde{u}_{n,1}, \ldots, \tilde{u}_{n,p}$ the corresponding eigenvectors. We denote by $\tilde{D} = \mathrm{diag}(\tilde{\lambda}_{n,1}, \ldots, \tilde{\lambda}_{n,m})$ et $\tilde{U} = (\tilde{u}_{n,1}, \ldots, \tilde{u}_{n,m})$. In this case, we have $\tilde{U}'\tilde{U} = \mathsf{I}_m$ and, if $\lambda_{n,m} > 1$,

$$\tilde{\Lambda} = \tilde{U}(\tilde{D} - \mathsf{I}_m)^{\frac{1}{2}}.$$

Thus the maximum likelihood estimator of Λ can be written as following

$$\hat{\Lambda} = \Psi^{\frac{1}{2}}\tilde{U}(\tilde{D} - \mathsf{I}_m)^{\frac{1}{2}}.$$

When $\lambda_{n,m} \leq 1$, this method will not give real solutions. Nevertheless, it has been observed that this problem appears only when the number of factors m is large. Practically speaking, Lawley (1940) proposed first to calculate Λ for a given Ψ, then to update Ψ using the equation (2.7).

Now we give the expression for the maximized likelihood. As $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$, we have

$$\begin{aligned} \operatorname{tr}(\mathsf{S}_n\hat{\Sigma}^{-1}) &= \operatorname{tr}(\mathsf{S}_n\hat{\Sigma}^{-1}(\hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}')\hat{\Psi}^{-1}) \\ &= \operatorname{tr}(\mathsf{S}_n\hat{\Psi}^{-1} - (\mathsf{S}_n\hat{\Sigma}^{-1}\hat{\Lambda})\hat{\Lambda}'\hat{\Psi}^{-1}) \\ &= \operatorname{tr}(\mathsf{S}_n\hat{\Psi}^{-1} - \hat{\Lambda}\hat{\Lambda}'\hat{\Psi}^{-1}) \\ &= \operatorname{tr}((\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})\hat{\Psi}^{-1} - \hat{\Lambda}\hat{\Lambda}'\hat{\Psi}^{-1}) \\ &= p, \end{aligned}$$

where the third equality uses (2.4), and the fourth (2.5) and the fact that $\hat{\Psi}$ is diagonal. Moreover,

$$\begin{aligned} |\Psi^{-\frac{1}{2}}\hat{\Sigma}\Psi^{-\frac{1}{2}}| &= |\Psi^{-\frac{1}{2}}(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})\Psi^{-\frac{1}{2}}| \\ &= |\hat{\Lambda}'\Psi^{-1}\hat{\Lambda} + \mathsf{I}_m| \\ &= \tilde{\lambda}_{n,1}\cdots\tilde{\lambda}_{n,m}, \end{aligned}$$

and

$$|\Psi^{-\frac{1}{2}}\mathsf{S}_n\Psi^{-\frac{1}{2}}| = \tilde{\lambda}_{n,1}\cdots\tilde{\lambda}_{n,p}.$$

Thus we obtain

$$\begin{aligned} \frac{|\hat{\Sigma}|}{|\mathsf{S}_n|} &= \frac{\Psi^{-\frac{1}{2}}\hat{\Sigma}\Psi^{-\frac{1}{2}}}{\Psi^{-\frac{1}{2}}\mathsf{S}_n\Psi^{-\frac{1}{2}}} \\ &= \frac{\tilde{\lambda}_{n,1}\cdots\tilde{\lambda}_{n,m}}{\tilde{\lambda}_{n,1}\cdots\tilde{\lambda}_{n,p}} \\ &= \frac{1}{\tilde{\lambda}_{n,m+1}\cdots\tilde{\lambda}_{n,p}}. \end{aligned}$$

The maximized likelihood is then:

$$f(\hat{\Lambda}, \hat{\Psi}) = -\frac{n}{2}\left(\log|\mathsf{S}_n| + \sum_{i=m+1}^{p}\log(\tilde{\lambda}_{n,i})\right) - \frac{pn}{2}\log(2\pi) - \frac{pn}{2}.$$

2.5 Goodness of fit test for the factor model

We will give a likelihood ratio test that the factor model fits, namely the population covariance matrix can be written as $\Sigma = \Lambda\Lambda' + \Psi$, with Ψ a positive-definite squared matrix of size p, and Λ a real matrix of size $p \times m$, where m is given. As without any constraint, the maximum likelihood estimator of Σ is S_n, the likelihood ratio test statistic is

$$\frac{\max\limits_{\mu,\Lambda,\Psi}\mathcal{L}(\mu, \Lambda\Lambda' + \Psi)}{\max\limits_{\mu,\Sigma}\mathcal{L}(\mu, \Sigma)} = \frac{|\mathsf{S}_n|^{\frac{n}{2}}}{|\hat{\Psi} + \hat{\Lambda}\hat{\Lambda}'|^{\frac{n}{2}}} = \prod_{i=m+1}^{p}\tilde{\lambda}_{n,i}^{\frac{n}{2}}.$$

Generally, we use -2 times the logarithm of the likelihood ratio statistic, which is

$$-n \sum_{i=m+1}^{p} \log \tilde{\lambda}_{n,i}, \tag{2.8}$$

and the null will be rejected if (2.8) is too large. In the case where $\hat{\Psi}$ and $\hat{\Lambda}$ are asymptotically Gaussian, the classical theory (i.e. p "small" is kept fixed, whereas $n \to \infty$) gives the limiting distribution of (2.8). This is a χ^2_q law, where the degree of freedom q corresponds to the number of elements of Σ plus the number of restrictions concerning the identification, minus the number of parameters in Ψ and Λ. Here $q = \frac{1}{2}((p-m)^2 - p - m)$. A variation, proposed par Bartlett (1950), is to replace n by $n - (2p+11)/6 - 2m/3$ in (2.8).

2.6 Asymptotic law of the maximum likelihood estimators

The results of this section consider the classical framework, where the dimension of the data p is "small" and fixed, whereas the sample size $n \to \infty$. We do not assume anymore that the common factors $(\mathsf{f}_i)_{1 \le i \le m}$ and the specific factors are Gaussian. We have the following Proposition 13.

Proposition 13 (Anderson (2003))**.** *If we assume that* $\Gamma = \Lambda' \Psi^{-1} \Lambda$ *is diagonal for the identification constraint, with diagonal elements different and ordered, and if* $\mathsf{S}_n \xrightarrow{\mathbb{P}} \Lambda\Lambda' + \Psi$*, then* $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$*, and* $\hat{\Psi} \xrightarrow{\mathbb{P}} \Psi$*.*

In order to let $\mathsf{S}_n \xrightarrow{\mathbb{P}} \Lambda\Lambda' + \Psi$, it is enough that $(\mathsf{f}_i, \mathsf{e}_i)'$ has a distribution with finite moments of order two. The asymptotic normality is given by the following proposition:

Proposition 14 (Anderson & Amemiya (1988))**.** *Let* $\Theta = (\theta_{ij})_{1 \le i,j \le p} = \Psi - \Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'$*. If* $(\theta_{ij}^2)_{1 \le i,j \le p}$ *is nonsingular, if* Λ *and* Ψ *are identified by the condition that* $\Lambda'\Psi\Lambda$ *is diagonal and the diagonal elements are different and ordered, if* $\mathsf{S}_n \to \Lambda\Lambda' + \Psi$ *in probability and if* $\sqrt{n}(\mathsf{S}_n - \Sigma)$ *has a limiting distribution, then* $\sqrt{n}(\hat{\Lambda} - \Lambda)$ *and* $\sqrt{n}(\hat{\Psi} - \Psi)$ *have a limiting distribution. The covariance of* $\sqrt{n}(\hat{\Psi}_{ii} - \Psi_{ii})$ *and* $\sqrt{n}(\hat{\Psi}_{jj} - \Psi_{jj})$ *in the limiting distribution is* $2\Psi_{ii}^2\Psi_{jj}^2\xi^{ij}$ $(1 \le i,j \le p)$*, where* $(\xi^{ij}) = (\theta_{ij}^2)^{-1}$*.*

It simply requires that $(\mathsf{f}_i, \mathsf{e}_i)'$ has a distribution with finite fourth moments so that $\sqrt{n}(\mathsf{S}_n - \Sigma)$ has a limiting distribution.

2.7 The different types of factor models

According to the different assumptions made on the parameters of the factor model, we can distinguish between several types of factor models, described in this section.

2.7.1 The strict factor model

This is the model described Section 2.2. In this model, the matrix Ψ is assumed to be diagonal.

2.7.2 The strict factor model with homoscedastic variance

This model is a simplification of the strict factor model. We assume that $\Psi = \sigma^2 \mathsf{I}_p$. This model is considered in chapters 5 and 6. In this case, the equations (2.4) and (2.5), which define the maximum likelihood estimators can be simplified and become

$$\Lambda(\Gamma + \mathsf{I}_m) = \mathsf{S}_n \left(\frac{1}{\sigma^2 \mathsf{I}_p}\right) \Lambda, \tag{2.9}$$

$$p\sigma^2 = \operatorname{tr}(\mathsf{S}_n - \Lambda\Lambda'), \text{ with } \Gamma = \Lambda' \left(\frac{1}{\sigma^2 \mathsf{I}_p}\right) \Lambda \text{ diagonal.} \tag{2.10}$$

The estimation of Ψ is reduced to those of σ^2. We will denote its estimator by $\hat{\sigma}^2$. These equations have now an explicit solution, given by:

$$\widehat{\sigma}^2 = \frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_{n,i}, \tag{2.11}$$

$$\widehat{\Lambda}_k = \left(\lambda_{n,k} - \widehat{\sigma}^2\right)^{\frac{1}{2}} u_{n,k},\ 1 \leq k \leq m, \tag{2.12}$$

where $u_{n,k}$ is the normalized eigenvector of $\mathsf{S_n}$ which corresponds to $\lambda_{n,k}$, for all $1 \leq k \leq p$. The previous theorems still apply to these estimators.

The likelihood ratio statistics can be simplified as

$$\begin{aligned} L^* &= n \sum_{i=m+1}^{p} \log\left(\frac{\lambda_{n,i}}{\hat{\sigma}^2}\right) \\ &= \log\left(\frac{\prod_{i=m+1}^{p} \lambda_{n,i}^{1/(p-m)}}{\frac{1}{p-m}\sum_{i=m+1}^{p} \lambda_{n,i}}\right)^{n(p-m)} . \end{aligned}$$

2.7.3 The approximate factor model

In this case, we do not assume the diagonality of Ψ anymore, we allow correlations between the different idiosyncratic factors. This hypothesis allows to take into account more cases, like cross-sectional and time dependences. We often find this model in finance (see Harding (2007) for example).

2.8 Link with the spiked population model

In the strict factor model, we have seen that the population covariance matrix is $\Sigma = \operatorname{cov}(\mathsf{x}_i) = \Lambda\Lambda' + \Psi$, where Ψ is diagonal. Thus the spectrum of Σ will have the following general form

$$\operatorname{spec}(\Sigma) = (\eta_1, \ldots, \eta_m, \underbrace{\beta_{m+1}, \ldots, \beta_p}_{p-m}),$$

where $\eta_1 \geq \cdots \geq \eta_m$ are the eigenvalues corresponding to noise plus the perturbation part $\Lambda\Lambda'$, and $\beta_{m+1} > \cdots > \beta_p$ are eigenvalues which arise only from the noise. So Σ has the structure of a generalized spiked population model with spikes $(\eta_i)_{1\leq i \leq m}$, under the hypotheses that $\eta_i \notin \Gamma_H$, where H is the limiting spectral distribution of the sub-matrix of Σ obtained by removing their first m rows and columns, and Γ_H is its support.

In the strict factor model case with homoscedastic variance, we have $\Sigma = \Lambda\Lambda' + \sigma^2 \mathbf{I}_p$ which has the spectrum:

$$\text{spec}(\Sigma) = (\alpha_1 + \sigma^2, \ldots, \alpha_m + \sigma^2, \underbrace{\sigma^2, \ldots, \sigma^2}_{p-m}),$$

and can be rewritten as

$$\text{spec}(\Sigma) \quad = \quad \sigma^2(\alpha_1^*, \ldots, \alpha_m^*, \underbrace{1, \cdots, 1}_{p-m}), \tag{2.13}$$

with $\alpha_i^* = \frac{\alpha_i}{\sigma^2} + 1$, for all $1 \leq i \leq m$. Thus, we recover the classical form of the spiked population model (1.6).

2.9 Contributions of the thesis

In chapter 6, we assume the high-dimensional framework, where the dimension p of the data tends to infinity together with the sample size n, and p/n tends to a positive constant c. Firstly, we study the maximum likelihood estimator $\hat{\sigma}^2$ in this new framework, and we give its asymptotic distribution. This allows us to obtain the expression of the bias of this estimator, which appears when we consider high-dimensional setting. Secondly, we correct the test that the factor model fits, by giving the asymptotic limit of the likelihood ratio statistic (2.8). We conclude chapter 4 by defining an equality test of the norm of two factors scores, or equivalently of two spikes.

Chapter 3

Existing methods for the estimation of the number of factors/spikes

3.1 Estimation of the number of factors in the classical framework

First we present estimators of the number of factors in the classical framework, where the dimension of the data p is kept fixed, whereas the number of observations n tends to infinity.

3.1.1 The scree plot

This is an empirical method introduced by Cattell (1966), based on the analysis of the plot of the sample covariance eigenvalues, arranged in decreasing order. This plot generally shows an important decrease, followed by a stabilization of the eigenvalues, and it has been observed that the number of eigenvalues before the drop corresponds to the number of factors. It is a subjective criterion based solely on the analysis of a plot.

3.1.2 The estimators based on information theoretic criteria

We present here estimators based on information theoretic criteria introduced by Akaike (1973, 1974) (AIC), Schwarz (1978) and Rissanen (1978) (BIC/MDL). The principle is to take the number of factors which minimizes the criteria AIC or BIC/MDL. These criteria consider the problem of finding the model which fits best with the data, given a n-sample $\mathsf{x}_1, \ldots, \mathsf{x}_n$ with dimension p and a parametrized set of densities $f(\mathsf{x}_1, \ldots, \mathsf{x}_n, \Theta)$.

The model which gives the minimum AIC is selected, with

$$\mathrm{AIC} = -2 \log f(\mathsf{x}_1, \ldots, \mathsf{x}_n, \hat{\Theta}) + 2k,$$

where k is the number of free parameters in Θ, and $\hat{\Theta}$ is the maximum likelihood estimator of Θ. The AIC is an unbiased estimator of the Kulback-Liebler mean distance between the estimated density $f(\mathsf{x}_1, \ldots, \mathsf{x}_n, \hat{\Theta})$ and the modeled density $f(\mathsf{x}_1, \ldots, \mathsf{x}_n, \Theta)$. The second term is a bias correction, whereas the first term is the log-likelihood of the maximum likelihood estimator.

Schwarz and Rissanen were inspired by the work of Akaike (1973, 1974). Schwarz (1978) uses a Bayesian approach: he associates a prior density with each candidate model, and selects those which give a maximum posterior probability (BIC). The Rissanen's approach is based on information theoretic arguments. As each model can be used to encode the observed data, Rissanen proposes to select the model which leads to a minimal length code. Schwarz and Rissanen's approaches lead to the same criterion when the sample size becomes large,

$$\mathrm{BIC} = -\log f(\mathsf{x}_1, \ldots, \mathsf{x}_n, \hat{\Theta}) + \frac{1}{2} k \log n.$$

Except a factor 2, the first term is identical to the criterion AIC. The second term differs only from a factor $\frac{1}{2}\log n$.

In the strict factor model case with $\Psi = \sigma^2 \mathsf{I}_p$, we find $k = m(2p-m)+1$. For $f(\mathsf{x}_1, \ldots, \mathsf{x}_n, \hat{\Theta})$, we take the likelihood ratio statistic (2.8)

$$\begin{aligned} n \sum_{i=m+1}^{p} \log \tilde{\lambda}_{n,i} &= n \sum_{i=m+1}^{p} \log \left(\frac{\lambda_{n,i}}{\hat{\sigma}^2} \right) \\ &= \log \left(\frac{\prod_{i=m+1}^{p} \lambda_{n,i}^{1/(p-m)}}{\frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_{n,i}} \right)^{n(p-m)} . \end{aligned}$$

Estimators of the number of factors $\hat{m}_{\mathrm{AIC}}$ or $\hat{m}_{\mathrm{BIC}}$ are then given by the value of $m \in \{0, \ldots, p-1\}$, which minimize

$$\mathrm{AIC}(m) = -2 \log \left(\frac{\prod_{i=m+1}^{p} \lambda_{n,i}^{1/(p-m)}}{\frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_{n,i}} \right)^{n(p-m)} + 2m(2p-m),$$

or

$$\mathrm{BIC}(m) = -\log \left(\frac{\prod_{i=m+1}^{p} \lambda_{n,i}^{1/(p-m)}}{\frac{1}{p-m} \sum_{i=m+1}^{p} \lambda_{n,i}} \right)^{n(p-m)} + \frac{1}{2} m(2p-m) \log n.$$

Several works have analyzed the performance of these two estimators, such as Fishler et al. (2002), Liavas & Regalia (2001), Xu & Kaveh (1995) or Zhang et al. (1989). The latter proved that the BIC estimator is strongly consistent when $n \to \infty$. Associated with its simplicity, the BIC estimator is therefore the standard tool for detecting the number of signals (factors) in signal processing. For large samples, it has been empirically observed that the main source of error for this estimator is an under-estimation of the factors number of one.

In the AIC estimator case, it has been shown that it tends to over-estimate the factors number when the sample size tends to infinity. Expressions of this over-estimation probability can be found in Xu & Kaveh (1995) and Zhang et al. (1989). One could rely on Stoica & Sélen (2004) for a review of the rules based on the information criteria.

3.1.3 Another Bayesian method: the Laplace method

In a Bayesian framework, the general methodology is to maximize the probability of the data given m factors $p(\mathsf{x}_1, \dots, \mathsf{x}_n|m)$, which is called the evidence. This is the framework used for the BIC criterion. For this framework, the authors used a second order Taylor expansion, because it is generally difficult to obtain an analytic expression of the evidence, as it needs an integration over all the parameters of the model. Nevertheless, Minka (2000) gave a more precise expression, using a Laplace approximation of the integrals, which is

$$\int g(\theta)\,\mathrm{d}\theta \simeq g(\hat{\theta})(2\pi)^{\frac{\mathrm{col}(\mathsf{A})}{2}}|\mathsf{A}|^{-\frac{1}{2}},$$

where $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\theta)$, $\mathsf{A} = -\left[\frac{\mathrm{d}^2 \log f(\theta)}{\mathrm{d}\theta_i \mathrm{d}\theta_j}\right]_{\theta=\hat{\theta}}$ and $\mathrm{col}(\mathsf{A})$ is the column number of A. Using this approximation, the author obtains the following expression

$$-\log p(\mathsf{x}_1, \dots, \mathsf{x}_n|m) \simeq \mathcal{L} - \log p(P_m) - \frac{d+r}{2}\log 2\pi + \frac{1}{2}\log|A_z| + \frac{m}{2}\log n, \qquad (3.1)$$

where $d = pm - m(m+1)/2$, and

$$\begin{aligned} p(P_m) &= 2^{-m}\prod_{i=1}^{m}\Gamma\left(\frac{p-i+1}{2}\right)\pi^{-\frac{p-i+1}{2}}, \\ |A_z| &= \prod_{i=1}^{m}\prod_{j=i+1}^{p} n(l_{n,j}^{-1} - l_{n,i}^{-1})(\lambda_{n,i} - \lambda_{n,j}), \end{aligned}$$

where Γ is the gamma function, $p(P_m)$ is a prior non-informative distribution for $P_m = (v_{n,1}, \dots, v_{n,m})$ ($v_{n,i}$ being the eigenvector which corresponds to $\lambda_{n,i}$), $l_{n,j}$ equals to $\lambda_{n,j}$ if $j \leq m$ and $\hat{\sigma}^2$ otherwise. The m minimizing the expression (3.1) is the estimator of the number of factors.

The BIC method can be viewed as a simplification of this criterion, obtained by removing the terms which are not growing in n.

3.2 Estimation of the number of factors in high-dimension

When the data dimension p is large compared to the sample size n (not necessarily $p > n$), the classical methods described above are not effective anymore. One can consult the paper of Kritchman & Nadler (2009), in which the authors compare their algorithm described in Section 3.2.3 below with the AIC and BIC methods. As the classical methods are ineffective in these circumstances, it becomes necessary to develop new methods which deal with this high-dimensional framework. The aim of this section is to present some estimators built in this context.

3.2.1 Method SURE of Ulfarsson and Solo

The method of Ulfarsson & Solo (2008) is based on the Stein Unbiased Risk Estimator (SURE). We aim to find the number of factors m that minimizes the risk

$$R_m = \mathbb{E}\|\nu - \hat{\nu}\|^2,$$

where $\nu = \Lambda \mathsf{f}$ and $\hat{\nu} = \hat{\Lambda}\hat{\mathsf{f}}$, $\hat{\Lambda}$ being the maximum likelihood estimator of Λ, and

$$\hat{\mathsf{f}} = \mathbb{E}_{\hat{\Lambda},\hat{\sigma}^2}(\mathsf{f}|\mathsf{x}_1, \ldots \mathsf{x}_n).$$

The problem is that ν is generally unknown. The idea is to replace R_m by an unbiased estimator that we are able to calculate. Stein (1981) explained how to construct such a risk under Gaussian assumptions. This estimator is given by the following expression

$$\hat{R}_m = \frac{1}{n}\sum_{i=1}^{n}\|n_i\|^2 + \frac{2\sigma^2}{n}\sum_{i=1}^{n}\operatorname{tr}\left(\frac{\partial \hat{\nu}_i}{\partial \mathsf{x}_i'}\right) - m\sigma^2,$$

where $n_i = \mathsf{x}_i - \hat{\nu}_i$. Use of SURE is based on the fact that, since SURE is an unbiased estimator of the risk, then one expects that, on average, the minimizer of SURE is an unbiased estimator of the minimizer of the risk.

The main task is to compute the partial derivative in the previous expression. After some calculations, we obtain

$$\begin{aligned}\hat{R}_m &= (p-m)\hat{\sigma}^2 + \hat{\sigma}^4\sum_{i=1}^{m}\lambda_{n,i}^{-1} + 2\hat{\sigma}^2(1-n^{-1})m \\ &\quad -2\sigma^2\hat{\sigma}^2(1-n^{-1})\sum_{i=1}^{m}\lambda_{n,i}^{-1} + \frac{4(1-n^{-1})\sigma^2\hat{\sigma}^2}{n}\sum_{i=1}^{m}\lambda_{n,i}^{-1} + C_m,\end{aligned}$$

with $C_m = \frac{2(1-n^{-1}\sigma^2)}{n}\sum_{i=1}^{m}\left(1-\frac{\hat{\sigma}^2}{\lambda_{n,i}}\right)\sum_{i\neq j}\frac{\lambda_{n,j}+\lambda_{n,i}}{\lambda_{n,j}-\lambda_{n,i}}$. The variance σ^2 is assumed to be known. Otherwise, a natural choice is $\hat{\sigma}^2$, but the authors observed that it does not work well in practice. So they propose an other estimator which performs better. The estimator of m will be the minimizer of $\hat{R}_m$.

3.2.2 Method of Harding

In the work of Harding (2007), an estimation method is presented with a factor model where the idiosyncratic factors e_i can have a dependence in n (auto-regressive vectors in n for instance). Nevertheless, simulations are done for strict factor models. The general idea is to compare the spectral moments of S_n with the empirical spectral distribution of S_n without the factors, and to remove the largest eigenvalues one by one in S_n until the "distance" between the moments is minimum.

More precisely, the author considers the decomposition $\mathsf{S}_n = \Xi_n + \Psi_n$ (i.e. $\Xi_n = \Lambda\Lambda'$), where the rank of Ξ_n is m. Let $\Pi(\mathsf{S}_n)$ be the vector of the first s moments of the empirical spectral distribution of the covariance matrix S_n, $\Pi(\Psi_n)$ the equivalent for Ψ_n and $\Pi(\sigma^2)$

its limit as p and $n \to \infty$, $\frac{p}{n} \to c$. Harding's method is summarized as follows:

- First, compute the moments $\Pi(\sigma^2)$ of the asymptotic eigenvalue distribution of the covariance matrix of Ψ_n for a large (p, n) sample;
- By Proposition 7, we have that $p(\Pi(\Psi_n) - \Pi(\sigma^2)) \xrightarrow{\mathcal{L}} \mathcal{N}(\Delta, W)$;
- Consequently, estimate σ^2 by

  $$\hat{\sigma}_0^2 = \underset{\sigma^2}{\operatorname{argmin}}\ J(\sigma^2),$$

  where $J(\sigma^2) = (\Pi(\sigma^2) - \Pi(\mathsf{S}_n))'\hat{W}^{-1}(\Pi(\sigma^2) - \Pi(\mathsf{S}_n))$ and $\hat{W}$ is a consistent estimate of W, calculated by estimated σ^2 from a first step estimation with $W = \mathsf{I}_p$;
- Next, remove the largest eigenvalue of the spectrum of S_n and re-estimate the parameter σ^2, as previous, to get a new estimate $\hat{\sigma}_1^2$;
- The previous step is repeated by progressively removing largest eigenvalues and for a predetermined number of times to get a sequence of estimates $\hat{\sigma}_1^2, \hat{\sigma}_2^2$, etc.;
- Finally, among the minimized objective functions $J(\hat{\sigma}_i^2)$, choose the order corresponding to the smallest minimized value. This is the estimator of the number of factors

  $$\hat{m} = \underset{i}{\operatorname{argmin}}\ J(\hat{\sigma}_i^2).$$

Actually, we know that for m fixed and p, $n \to \infty$, $\Pi(\mathsf{S}_n) \to \Pi(\sigma^2)$. So the criterion is the minimization of the variance $W = W(\sigma^2)$. This decreases up to m (until we have removed the eigenvalues corresponding to the spikes), then it stays stable. The procedure of Harding leads to an underestimation of m, where p and n are fixed. As a result, Harding constrained the function J with a function of type $k\hat{\sigma}^2 g(p, n)$, where k is the number of eigenvalues removed, $\hat{\sigma}^2$ is the estimated variance at the step q and $g(p, n)$ is a function such that $g(p, n) \to 0$ when p, $n \to \infty$. The finally proposed choice for g is the following function given by Bai & Ng (2002) based on a BIC criterion:

$$g(p, n) = \left(\frac{p+n}{pn}\right) \ln \left(\frac{pn}{p+n}\right).$$

3.2.3 Method of Kritchman and Nadler

The method of Kritchman & Nadler (2008, 2009) is based on the fact that, in the absence of factors (m=0), $n\mathsf{S}_n$ follows a standard Wishart distribution with parameters n and p. In this case, Johnstone (2001) has provided the asymptotic distribution of the largest eigenvalue of S_n.

Proposition 15. *Let S_n be the sample covariance matrix of n vectors distributed as $\mathcal{N}(0, \sigma^2 \mathsf{I}_p)$, and $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ be its eigenvalues. Then, when $n \to \infty$, such that $\frac{p}{n} \to c > 0$*

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

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

From Corollary 1, we know that the eigenvalues of S_n which correspond to factors tend to a value greater than $b = (1+\sqrt{c})^2$. Consequently, we can distinguish a factor eigenvalue λ from a noise one at an asymptotic significance level γ by checking whether

$$\lambda_{n,k} > \sigma^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b \right) \tag{3.2}$$

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 $k = 1, 2, \ldots, \min(p, n) - 1$,

$$\mathcal{H}_0^{(k)}\colon\ m \leq k-1 \quad vs. \quad \mathcal{H}_1^{(k)}\colon\ m \geq k\ .$$

For each value of k, if (3.2) is satisfied, $\mathcal{H}_0^{(k)}$ is rejected and k is increased by one. The procedure stops once an instance of $\mathcal{H}_0^{(k)}$ is accepted and the number of factors is then estimated to be $\hat{m}_{\text{KN}} = k - 1$. Formally, the estimator KN is defined by

$$\hat{m}_{\text{KN}} = \operatorname*{argmin}_k \left(\lambda_{n,k} < \widehat{\sigma}^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b \right) \right) - 1.$$

The authors proved the strong consistency of their estimator as $n \to \infty$ with fixed p, by replacing the fixed confidence level γ with a sample-size dependent one γ_n, where $\gamma_n \to 0$ sufficiently slowly as $n \to \infty$. They also proved that $\lim\limits_{p,n\to\infty} \mathbb{P}(\hat{m}_{\text{KN}} \geq m) = 1$.

3.3 Contributions of the thesis

In chapters 4 and 5 we propose an estimator of the number of factors based on the analysis of the difference between two consecutive eigenvalues $\delta_{n,j} = \lambda_{n,j} - \lambda_{n,j+1}$ of the sample covariance matrix S_n, the latter being in decreasing order.

Chapter 4 considers the framework of the spiked population model described in Section 1.4, with $T_{p-m} = \mathsf{I}_{p-m}$, which is a formulation equivalent to the factor model. In this case, Corollary 1 shows that the eigenvalues which correspond to the spikes α_k tend to $\phi(\alpha_k)$ almost surely, whereas the following converge to b, which is the upper-bound of the Marčenko-Pastur law. If we assume that the spikes are all different (i.e. with multiplicity one), the difference will tend almost surely to a positive constant if there is a spike inside, and zero otherwise. Accordingly, we can detect the number of spikes by inspecting the index j where $\delta_{n,j}$ has a value below a given threshold. We compare this new estimator to the methods of Harding (2007) and Kritchman & Nadler (2008, 2009);

In chapter 5, we consider the framework of the strict factor model. We extend our previous method to the case where the factors/spikes can be equal, by using a difference in terms of convergence rate between the eigenvalues of S_n corresponding to the spikes (see Proposition 11), and the other eigenvalues. We further modify the threshold used in chapter 4. As this new threshold depends on a constant to be adjusted, we construct a procedure with an automatic calibration of this constant. We do simulation experiments and compare our method with the one of Kritchman & Nadler (2008, 2009).

Chapter 4

On determining the number of spikes in a high-dimensional spiked population model

Abstract: In a spiked population model, the population covariance matrix has all its eigenvalues equal to units except for a few fixed eigenvalues (spikes). Determining the number of spikes is a fundamental problem which appears in many scientific fields, including signal processing (linear mixture model) or economics (factor model). Several recent papers studied the asymptotic behavior of the eigenvalues of the sample covariance matrix (sample eigenvalues) when the dimension of the observations and the sample size both grow to infinity so that their ratio converges to a positive constant. Using these results, we propose a new estimator based on the difference between two consecutive sample eigenvalues.

Keywords: spiked population model, high-dimensional statistics, sample covariance matrices, factor model, extreme eigenvalues, Tracy-Widom laws.

AMS subject classification: 62F07, 62F12, 60B20.

This chapter was published in
Random Matrices: Theory and Applications
Volume 1, Number 1, January 2012, 115002
under the title:
“On determining the number of spikes in a high-dimensional spiked population model.”

It has been written in collaboration with Jian-Feng Yao.

4.1 Introduction

In a spiked population model, the population covariance matrix has all its eigenvalues equal to units except for a few fixed eigenvalues (spikes). This model appears in many scientific fields often with different names. In economics, it is called “factor model” within the Ross Arbitrage Pricing Theory (APT) and the aim is to relate observed data (assets) to a small dimensional set of unobserved variables which are then estimated (see Ross (1976)). In physics of mixture, “linear mixture models” are naturally considered for various phenomena (see Naes et al. (2002)). In wireless communication, a signal emitted by a source is modulated and received by an array of antennas which will permit the reconstruction of the original signal.

An important question to be addressed under this model is how many factors (or components, or signals) there are. It is generally a first step preliminary to any further study such as estimation and forecasting.

Many methods for determining the number of factors have been developed, based on the minimum description length (MDL), Bayesian model selection or Bayesian Information Criteria (BIC) (See Bai & Ng (2002)). 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 (2007) or Onatski (2009) in economics, and Kritchman & Nadler (2008) in array processing or chemometrics literature.

In this chapter, we present a new estimator for the number of spikes from high-dimensional data. Our approach is based on the results of Bai & Yao (2008) and Paul (2007) which give the limiting distributions of the extreme eigenvalues of a sample covariance matrix coming from a spiked population model, and a recent result of Benaych-Georges et al. (2011). The obtained results are presented in Section 4.3.

The remaining sections of the chapter are organized as follows. In Section 4.2, we introduce the spiked population model, and recall known results on the almost sure limits of extreme eigenvalues which lead to the idea of our estimator. In Section 4.3 we define precisely our estimator and prove its consistency in the case of simple spikes with known variance. Next we give a method of estimation in the case of simple spikes with unknown variance. In Section 4.5, we define the factor/linear mixture model that we link to the spiked population model and we compare our method to those of Harding (2007) and Kritchman & Nadler (2008). We consider the case of spikes with greater multiplicity in Section 4.6. Throughout the chapter, simulation experiments are conducted to access the quality of the proposed estimation.

4.2 Spiked population model

We consider $\mathsf{x} = E\Sigma^{\frac{1}{2}}\mathsf{y}$, where $\mathsf{y} \in \mathbb{R}^p$ is a zero-mean random vector of i.i.d. components, E is an orthogonal matrix and

$$\Sigma = \operatorname{cov}(\mathsf{x}) = \sigma^2 \begin{pmatrix} V_m & 0 \\ 0 & I_{p-m} \end{pmatrix},$$

where V_m has K non null and non unit eigenvalues $(\alpha_k^*)_{1\le k\le K}$ with respective multiplicity $(n_k)_{1\le k\le K}$ $(n_1 + \cdots + n_K = m)$. Therefore, the eigenvalues of the population covariance matrix Σ are unit except the α_j, called spike eigenvalues. Notice that, if the observations are Gaussian, we may assume that V is diagonal by using a suitable orthogonal transformation.

Let $(\mathsf{x}_i)_{1\le i\le n}$ be n independent copies of x. The sample covariance matrix is

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

It is assumed in the sequel that m is fixed, and p and n are related so that when $n \to \infty$, $\frac{p}{n} \to c > 0$. Moreover, we assumed that $\alpha_1^* > \cdots > \alpha_K^* > 1 + \sqrt{c}$. For $\alpha \neq 1$, we define the function

$$\phi(\alpha) = \alpha + \frac{c\alpha}{\alpha - 1}. \tag{4.1}$$

Let $\lambda_{n,1} \ge \lambda_{n,2} \ge \cdots \ge \lambda_{n,p}$ be the eigenvalues of the sample covariance matrix S_n. Let $s_i = n_1 + \cdots + n_i$ for $1 \le i \le K$. Baik & Silverstein (2006) proved that, under a moment condition on x, for each $k \in \{1, \ldots, K\}$ and $s_{k-1} < j \le s_k$ almost surely,

$$\lambda_{n,j} \longrightarrow \sigma^2\phi(\alpha_k^*).$$

In other words, with the hypotheses that $\alpha_k^* > 1 + \sqrt{c}$ for all k, and has multiplicity n_k, then $\phi(\alpha_k^*)$ is the limit of n_k packed sample eigenvalue $\{\lambda_{n,j},\ s_{k-1} + 1 \le j \le s_k\}$. They also prove that for all $1 \le i \le L$ with a prefixed range L almost surely,

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

Our aim is to estimate m when only S_n is known. The idea is to use, as suggested in Onatski (2009), differences between consecutive eigenvalues

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

Indeed, applying the results quoted above it is easy to see that a.s. if $j \ge m$, $\delta_{n,j} \to 0$ while when $j < m$, $\delta_{n,j}$ tends to a positive limit if the α_k^* are different. Thus it is possible to detect m from index-numbers j where $\delta_{n,j}$ becomes small.

4.3 Case of simple spikes with known variance σ^2

In this section, we suppose that σ^2 is known and that all the spikes are simple, i.e $n_1 = \cdots = n_K = 1$. Under these hypotheses the population eigenvalues are

$$\mathrm{spec}(\Sigma) = \sigma^2(\underbrace{\alpha_1^*, \ldots, \alpha_m^*}_{m}, \underbrace{1, \ldots, 1}_{p-m}).$$

We also need the following assumption:

Assumption 1. The entries y_{ij} of the random vector y have a symmetric law and a sub-exponential decay, that is there exists positive constants C, C' such that, for all $t \geq C'$,

$$\mathbb{P}(|\mathsf{y}_{ij}| \geq t^C) \leq e^{-t}.$$

Especially, the Gaussian vectors satisfy this hypothesis.

As stated previously, the main observation is that when one follows the sample eigenvalues in a descending order, the successive spacings $\delta_{n,j}$ shrink to small values when approaching non-spiked values. Therefore, our estimation method will use a carefully determined threshold d_n. We propose to estimate m by the following

$$\hat{m}_n = \max\{j \in \{1, \ldots, s\} : \forall k \in \{1, \ldots, j\},\ \delta_{n,j} \geq d_n \text{ and } \delta_{n,j+1} < d_n\},$$

where $s > m$ is a fixed number big enough, and d_n is a level to determine. In practice, the integer s should be thought as a preliminary bound on the number of possible spikes.

4.3.1 Consistency

Theorem 1. *Let $(\mathsf{x}_i)_{1 \leq i \leq n}$ be n copies i.i.d. of $\mathsf{x} = E\Sigma^{\frac{1}{2}}\mathsf{y}$, where $\mathsf{y} \in \mathbb{R}^p$ is a zero-mean random vector of i.i.d. components which satisfies Assumption 1 and E is an orthogonal matrix. Assume that*

$$\Sigma = cov(\mathsf{x}) = \sigma^2 \begin{pmatrix} V_m & 0 \\ 0 & \mathsf{I}_{p-m} \end{pmatrix}$$

where V has m non null, non unit and different eigenvalues $\alpha_1^* > \cdots > \alpha_m^* > 1 + \sqrt{c}$. Assume that $\frac{p}{n} \to c > 0$ when $n \to \infty$.
Let $(d_n)_{n \geq 0}$ be a real sequence such that $d_n \to 0$ and $n^{2/3}d_n \to \infty$. Then the estimator $\hat{m}_n$ is consistent, i.e $\mathbb{P}\left(\widehat{m}_n = m\right) \to 1$ when $n \to \infty$.

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 16, shows that the limiting law of $\lambda_{n,j} - \phi(\alpha_j^*)$ is Gaussian (Bai & Yao (2008) and Paul (2007)):

Proposition 16. *Assume that the entries x_{ij} of x satisfy $\mathbb{E}(|\mathsf{x}_{ij}|^4) < \infty$, $\alpha_j^* > 1 + \sqrt{c}$ for all $1 \leq j \leq m$ and have multiplicity 1. Then as p, $n \to \infty$ so that $\frac{p}{n} \to c$,*

$$\sqrt{n}(\lambda_{n,j} - \phi(\alpha_j^*)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\alpha_j^*))$$

where $\sigma^2(\alpha_j^*) = 2\alpha_j^{*2}\left(1 - \frac{c}{(\alpha_j^*-1)^2}\right)$.

The second, Proposition 17, is issued from the second part of Proposition 5.8 of Benaych-Georges et al. (2011):

Proposition 17. *Under the Assumption 1, for all* $1 \leq i \leq L$ *with a prefixed range* L,

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

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

We also need the following lemma:

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

$$\mathbb{P}(|X_n| \geq u_n) \to 0.$$

Proof. As $(X_n)_{n\geq 0}$ is a tight sequence, for all $\varepsilon > 0$, it exists a compact K such that, for all $n \in \mathbb{N}$, $\mathbb{P}(\mathsf{X}_n \notin K) < \varepsilon$. Furthermore, as $u_n \to \infty$, it exists $N \in \mathbb{N}$ such that for all $n \geq N$, $[-u_n, u_n] \supset K$. So $\mathbb{P}(|\mathsf{X}_n| > u_n) \leq \mathbb{P}(\mathsf{X}_n \notin K) < \varepsilon$. Consequently, $\mathbb{P}(|\mathsf{X}_n| > u_n) \to 0$. □

Proof. of Theorem 1. We have

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

Therefore

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

Convergence of $\mathbb{P}(\delta_{n,m+1} \geq d_n)$. In this case, $\delta_{n,m+1} = \lambda_{n,m+1} - \lambda_{n,m+2}$ (non-spike eigenvalues). We consider the following sequence of random variables

$$Y_n = \frac{n^{\frac{2}{3}}}{\beta}(\lambda_{n,m+i} - b).$$

By Proposition 17, $(Y_n)_{n\geq 1}$ is a tight sequence. So by using lemma 1, for any sequence $(a_n)_{n\geq 0}$, $a_n \to \infty$ we have

$$\mathbb{P}(|Y_n| \geq a_n) \to 0.$$

Therefore

$$\begin{aligned}\mathbb{P}(|Y_n| \leq a_n) &= \mathbb{P}\left(\frac{n^{\frac{2}{3}}}{\beta}(|\lambda_{n,m+i} - b| \leq a_n\right) \\ &= \mathbb{P}\left(|\lambda_{n,m+i} - b| \leq \frac{a_n}{n^{\frac{2}{3}}}\beta\right) \\ &\longrightarrow 1.\end{aligned}$$

We choose $d_n \to 0$ such that $n^{2/3}d_n \to \infty$. So we have

$$\mathbb{P}(\lambda_{n,m+i} \in \mathsf{J}_n) \to 1,$$

with

$$\mathsf{J}_n = [b \pm d_n/2].$$

It follows

$$\mathbb{P}(\delta_{n,m+1} \leq d_n) \geq \mathbb{P}(\{\lambda_{n,m+i} \in \mathsf{J}_n\} \cap \{\lambda_{n,m+i+1} \in \mathsf{J}_n\}) \to 1.$$

Therefore

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

Case of $1 \leq j \leq m$. These indexes correspond to the spike eigenvalues. By using Proposition 16 and the previous argument, it is easy to show that we can choose a real sequence $(v_n)_{n\geq 0}$, $v_n \to 0$ such that $\sqrt{n}v_n \to \infty$ and

$$\mathbb{P}(\lambda_{n,j} \in \mathsf{I}_{n,j}) \to 1,$$

where

$$\mathsf{I}_{n,j} = \left[\phi(\alpha_j^*) \pm v_n\right].$$

Therefore

- For all $1 \leq j < m$, we have

  $$\mathbb{P}\left(\delta_{n,j} \geq \phi(\alpha_j^*) - \phi(\alpha_{j+1}^*) - v_n\right) \geq \mathbb{P}(\{\lambda_{n,j} \in \mathsf{I}_{n,j}\} \cap \{\lambda_{n,j+1} \in \mathsf{I}_{n,j+1}\}) \to 1.$$

  Let

  $$w_{n,j} = \phi(\alpha_j^*) - \phi(\alpha_{j+1}^*) - v_n.$$

- For $j = m$, $\delta_{n,m} = \lambda_{n,m} - \lambda_{n,m+1}$. By using the first section of the proof, one can show that

  $$\mathbb{P}(\delta_{n,m} \geq \phi(\alpha_m^*) - b - (v_n + d_n)) \geq \mathbb{P}(\{\lambda_{n,m} \in \mathsf{I}_{n,m}\} \cap \{\lambda_{n,m+1} \in \mathsf{J}_n\}) \to 1.$$

Let

$$w_{n,m} = \phi(\alpha_m^*) - b - (v_n + d_n)\,.$$

– Therefore for all $0 \leq j \leq m$ we have

$$\mathbb{P}(\delta_{n,j} \geq w_{n,j}) \to 1 \quad \Rightarrow \quad \mathbb{P}(\delta_{n,j} < w_{n,j}) \to 0.$$

As $d_n \to 0$ and for all $1 \leq j \leq m$, $w_{n,j} \to w_j > 0$, it exists $N \in \mathbb{N} : \forall n \geq N$,

$$\mathbb{P}(\delta_{n,j} < d_n) \leq \mathbb{P}(\delta_{n,j} < w_{n,j}) \to 0.$$

So we have

$$\sum_{j=1}^{m} \mathbb{P}(\delta_{n,j} < d_n) \to 0.$$

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

$$\mathbb{P}(\hat{m}_n = m) \underset{n\to\infty}{\longrightarrow} 1.$$

□

4.3.2 Simulation experiments

Now we will illustrate the previous result by some simulations. First, we have to chose the sequence d_n to be used. Theoretically speaking, all the sequences satisfying the requirement $d_n \to 0$ such that $n^{2/3}d_n \to \infty$ are convenient. We tested several sequences and we decided to take one of the form $\frac{a_n}{n^{2/3}}\beta$, with a sequence $(a_n)_{n\geq 0}$ proportional to $\sqrt{2\log\log n}$: this idea came from that, as in the case of the mean of i.i.d random variables, the $\lambda_{n,j}$ corresponding to the spikes tend to a Gaussian law (Proposition 16). So we can conjecture a result analog to the law of the iterated logarithm [1] for the $\lambda_{n,j}$, $j \leq m$. Finally, we choose $a_n = 4\sqrt{2\log\log n}$ and simulate two different models: one with dispersed spikes which should lead to an easier estimation of m, and a more difficult case with closer spikes:

- **Model 1**: $m = 5$, $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*) = (259.72, 17.97, 11.04, 7.88, 4.82)$;
- **Model 2**: $m = 4$, $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*) = (7, 6, 5, 4)$.

Note that the values of model 1 have been chosen to be the same as in Harding (2007). For each model, two different values of c, 0.3 and 0.6, are considered. We give in Tables 4.1 and 4.2, respectively, the distribution of $\hat{m}_n$, its mean and mean squared error over 1000 independent replications. The frequency of $\hat{m}_n = m$ is given in Figure 4.1.

[1] If we consider an i.i.d. sequence of random variables $(x_i)_{1\leq i\leq n}$ with mean 0 and variance 1, the sum $\mathsf{S}_n = x_1 + \cdots + x_n$ has an almost-sure fluctuation of order $a_n = \sqrt{2\log\log n}$, i.e. $-\liminf_n \mathsf{S}_n/a_n = \limsup S_n/a_n = 1$, so the empirical mean has an a.s. fluctuation of order $\sqrt{2\log\log n}/\sqrt{n}$. The empirical mean has also Gaussian fluctuations in distribution of order $\sqrt{n}$. In the non-spike case, $n^{2/3}(\lambda_1 - b) \to F_1$, the Tracy-Widom law of order 1. Therefore, a law of the iterated logarithm for λ_1 would be that a.s., $\lambda_1 - b$ is of order of $\sqrt{2\log\log n}/n^{2/3}$ ($\sqrt{n}$ would be replaced by $n^{2/3}$).

Table 4.1: Mean, mean squared error and empirical distribution of $\hat{m}_n$ over 1000 independent replications for model 1.

			Distribution of $\hat{m}_n$						
(p, n)	Mean	MSE	1	2	3	4	**5**	6	7
(30,100)	5.057	0.212	0.001	0.007	0.009	0.0	**0.883**	0.1	0.002
(60,200)	5.081	0.107	0.001	0.001	0.0	0.0	**0.91**	0.088	0.0
(120,400)	5.079	0.073	0.0	0.0	0.0	0.0	**0.921**	0.079	0.0
(240,800)	5.069	0.064	0.0	0.0	0.0	0.0	**0.931**	0.069	0.0
(60,100)	5.056	0.139	0.001	0.004	0.003	0.002	**0.914**	0.076	0.0
(120,200)	5.08	0.098	0.0	0.001	0.002	0.0	**0.91**	0.087	0.0
(240,400)	5.072	0.079	0.002	0.0	0.0	0.0	**0.924**	0.075	0.0
(480,800)	5.072	0.069	0.0	0.0	0.0	0.0	**0.929**	0.07	0.001

Table 4.2: Mean, mean squared error and empirical distribution of $\hat{m}_n$ over 1000 independent replications for model 2.

			Distribution of $\hat{m}_n$					
(p, n)	Mean	MSE	0	1	2	3	**4**	5
(30,100)	3.718	1.086	0.0	0.001	0.059	0.0	**0.778**	0.085
(60,200)	3.925	0.582	0.013	0.024	0.019	0.0	**0.857**	0.087
(120,400)	4.005	0.331	0.01	0.01	0.001	0.0	**0.902**	0.077
(240,800)	4.062	0.110	0.002	0.001	0.0	0.0	**0.924**	0.073
(60,100)	3.478	1.655	0.053	0.086	0.059	0.001	**0.734**	0.067
(120,200)	3.818	0.823	0.025	0.033	0.024	0.0	**0.853**	0.065
(240,400)	3.969	0.394	0.009	0.015	0.011	0.0	**0.893**	0.072
(480,800)	4.051	0.108	0.003	0.0	0.0	0.0	**0.934**	0.063

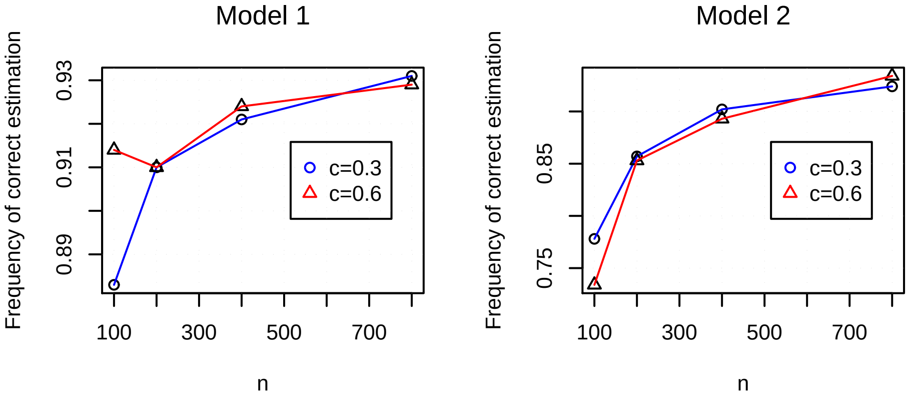

Figure 4.1: Frequency of $\hat{m}_n = m$ over 1000 independent replications.

In both cases, we can observe the asymptotic consistency of the estimator. Comparing the two models, except the last case $(p, n) = (480, 800)$, the estimator performs better in model 1 than in model 2. This phenomenon is due to the fact that the differences between consecutive eigenvalues are smaller in model 2 so that it is more difficult to distinguish spikes from non spikes.

Within a given model, the convergence is slower in the $c = 0.6$ case. We could explain this by the fact that the gap between two consecutive spike eigenvalues stays the same, and when c increases, the spectrum of S_n is more dispersed, so that the differences $\delta_{n,j}$ from non-spikes are larger and again our detection problem is more difficult to solve.

It is worth mentioning that the chosen constant $d_n = \frac{4\sqrt{2\log\log n}}{n^{2/3}}\beta$ leads to a slight overestimation of m for the tested sizes (p, n). This finite-sample behavior could be improved with a more sophisticated choice of d_n which however seems a difficult point to address.

4.4 Case of simple spikes with unknown variance

In practice, the scale parameter σ^2 is also unknown and we need to estimate it as well. First, we will explain how to do in the non-spikes (null) case, i.e. $\Sigma = \sigma^2 I_p$, and then in the case with spikes.

4.4.1 Estimation of the variance in the white case

We consider a zero-mean random vector $\mathsf{x} \in \mathbb{R}^p$ with population covariance matrix

$$\Sigma = \mathrm{cov}(\mathsf{x}) = \sigma^2 \mathsf{I}_p.$$

We keep the previous assumptions. We will use the law of large numbers to estimate the unknown variance σ^2. We have the following theorem (Marčenko & Pastur (1967), Bai & Silverstein (2004))

Proposition 18. *We denote by (x_{jk}) the entries of the vector x_j. Assume that, for any $\eta \geq 0$:*

$$\frac{1}{\eta^2 np} \sum_{j,k} \mathbb{E}(|x_{jk}|^2 \mathbb{1}_{|x_{jk}| \geq \eta\sqrt{n}}) \to 0 \text{ when } n \to \infty.$$

Then, with probability one, the empirical spectral distribution (ESD) F^{S_n} of S_n weakly converges to the Marčenko-Pastur distribution with ratio index c and scale parameter σ^2, denoted by $F_{c,\sigma^2}(x)$, which has a density function

$$p_{c,\sigma^2}(x) = \begin{cases} \frac{1}{2\pi x c \sigma^2}\sqrt{(b(c) - x)(x - a(c))} & \text{if } a(c) \leq x \leq b(c) \\ 0 & \text{otherwise} \end{cases},$$

where $a(c) = \sigma^2(1 - \sqrt{c})^2$ and $b(c) = \sigma^2(1 + \sqrt{c})^2$.

Note that σ^2 represents the mean of the limiting distribution. Moreover, it is well-known that under the condition of Proposition 18, it holds almost surely,

$$\widehat{\sigma}^2 = \frac{1}{p}\sum_{i=1}^{p} \lambda_{n,i} \to \sigma^2.$$

4.4.2 Determining the number of spikes with an unknown variance

As we notice in the first section, when the variance is known and different from one, we only have to divide the consecutive differences $\delta_{i,n}$ by this variance. As the variance is unknown, we will replace it by the estimator $\widehat{\sigma}^2 = \frac{1}{p}\sum_{i=1}^{p}\lambda_{n,i}$, which converges almost surely to σ^2 when $p \to \infty$. Nevertheless, because of the spikes, the variance of $\widehat{\sigma}^2$ will be greater than the one in the null case. The variance will be minimum if we only take the mean of the non-spike eigenvalues i.e. those that have an index $i \geq m+1$. The problem is that we do not know m. By consequence, the idea is to make a first estimation $\hat{m}_n^0$ of m with $\widehat{\sigma}_0^2 = \frac{1}{p}\sum_{i=1}^{p}\lambda_{n,i}$. Then, if $\hat{m}_n^0 > 0$, we set $\widehat{\sigma}_1^2 = \frac{1}{p-\hat{m}_n^0}\sum_{i=\hat{m}_n^0+1}^{p}\lambda_{n,i}$ (so we have $\widehat{\sigma}_0^2 \geq \widehat{\sigma}_1^2$), and we reestimate m by $\hat{m}_n^1$ using this new estimation. We repeat it until we find an index k such that $\hat{m}_n^k = \hat{m}_n^{k+1}$. If such an index does not exist, the algorithm will stop at the preliminary bound $k = s$ fixed initially. To sum up, here is the algorithm:

```
m1=0
sigma2=1/p*(lambda_1+...+lambda_p)
m2="estimator of the known variance case with division by sigma2"

while m2~=m1 do
 m1:=m2
 sigma2=1/(p-m1)*(lambda_(m1+1)+...+lambda_p)
 m2="estimator of the known variance case with division by sigma2"
end

result=(m1,sigma2)
```

4.4.3 Simulation experiments

We conduct the simulations with two values of the variance $\sigma^2 = 1$, and $\sigma^2 = 500$ to see if a high variance will influence the estimation. We keep the same other parameters as in the previous simulation study of Section 4.3 and estimate σ^2 and the number of spikes with the method explained above. Additional to the statistics about the spikes number estimator $\hat{m}_n$, we provide also those about the final estimate $\widehat{\sigma}^2$ of the unknown variance. The results are displayed in Tables 4.3 to 4.6. The frequency of $\hat{m}_n = m$ is given in Figure 4.2 and 4.3, and the mean of $\widehat{\sigma}^2$ in Figure 4.4.

Table 4.3: Mean, mean squared error and empirical distribution of $\hat{m}_n$, mean and mean squared error of $\widehat{\sigma}^2$ over 1000 independent replications for model 1 and $\sigma^2 = 1$.

			Distribution of $\hat{m}_n$							$\widehat{\sigma}^2$	
(p, n)	Mean	MSE	1	2	3	4	**5**	6	7	Mean	MSE
(30,100)	5.052	0.338	0.003	0.015	0.008	0.0	**0.849**	0.0	0.125	0.955	0.015
(60,200)	5.108	0.112	0.0	0.001	0.0	0.0	**0.89**	0.107	0.002	0.97	0.0
(120,400)	5.069	0.076	0.0	0.001	0.0	0.0	**0.927**	0.072	0.0	0.986	0.0
(240,800)	5.084	0.077	0.0	0.0	0.0	0.0	**0.916**	0.084	0.0	0.993	0.0
(60,100)	5.087	0.236	0.001	0.009	0.004	0.0	**0.865**	0.122	0.002	0.943	0.003
(120,200)	5.095	0.092	0.0	0.0	0.001	0.0	**0.902**	0.097	0.0	0.971	0.0
(240,400)	5.07	0.065	0.0	0.0	0.0	0.0	**0.93**	0.07	0.0	0.985	0.0
(480,800)	5.067	0.063	0.0	0.0	0.0	0.0	**0.933**	0.067	0.0	0.993	0.0

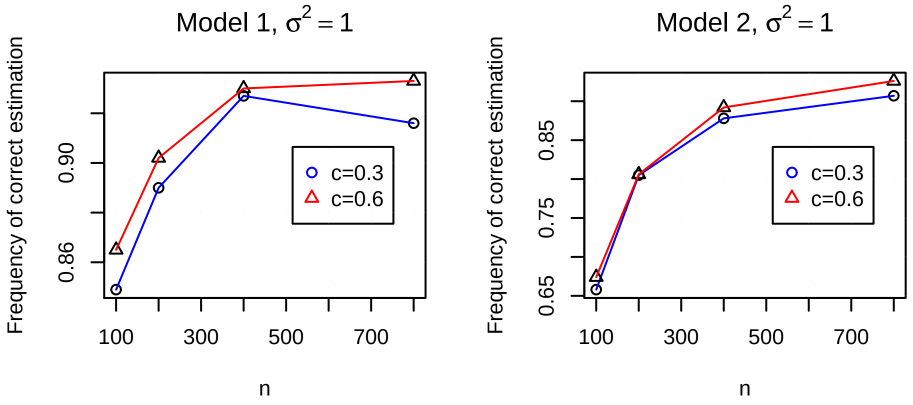

Figure 4.2: Frequency of $\hat{m}_n = m$ over 1000 independent replications with $\sigma^2 = 1$.

Table 4.4: Mean, mean squared error and empirical distribution of $\hat{m}_n$, mean and mean squared error of $\widehat{\sigma}^2$ over 1000 independent replications for model 2 and $\sigma^2 = 1$.

			Distribution of $\hat{m}_n$							$\widehat{\sigma}^2$	
(p, n)	Mean	MSE	0	1	2	3	**4**	5	6	Mean	MSE
(30,100)	3.362	2.019	0.079	0.078	0.091	0.0	**0.658**	0.094	0.0	1.052	0.043
(60,200)	3.806	1.023	0.032	0.038	0.026	0.0	**0.805**	0.098	0.001	0.994	0.005
(120,400)	3.983	0.483	0.019	0.008	0.004	0.0	**0.878**	0.091	0.0	0.991	0.001
(240,800)	4.071	0.144	0.003	0.001	0.001	0.0	**0.907**	0.088	0.0	0.994	0.0
(60,100)	3.367	1.898	0.069	0.081	0.096	0.001	**0.674**	0.079	0.0	1.003	0.012
(120,200)	3.781	1.04	0.034	0.034	0.036	0.0	**0.806**	0.089	0.001	0.986	0.002
(240,400)	3.965	0.472	0.015	0.015	0.007	0.0	**0.892**	0.071	0.0	0.99	0.0
(480,800)	4.052	0.125	0.002	0.003	0.0	0.0	**0.926**	0.069	0.0	0.994	0.0

Table 4.5: Empirical distribution of $\hat{m}_n$, mean and mean squared error of $\widehat{\sigma}^2$ over 1000 independent replications for model 1 and $\sigma^2 = 500$.

	Distribution of $\hat{m}_n$							$\widehat{\sigma}^2$	
(p, n)	1	2	3	4	**5**	6	7	Mean	MSE
(30,100)	0.003	0.012	0.005	0.0	**0.823**	0.155	0.002	474.909	3281.714
(60,200)	0.0	0.001	0.0	0.0	**0.904**	0.094	0.001	485.019	99.558
(120,400)	0.0	0.001	0.0	0.0	**0.918**	0.080	0.001	492.608	21.244
(240,800)	0.0	0.0	0.0	0.0	**0.914**	0.086	0.0	496.316	3.519
(60,100)	0.002	0.008	0.006	0.001	**0.870**	0.113	0.0	472.816	688.994
(120,200)	0.0	0.002	0.0	0.0	**0.898**	0.099	0.001	485.49	55.489
(240,400)	0.0	0.0	0.0	0.0	**0.928**	0.071	0.001	492.699	7.242
(480,800)	0.0	0.0	0.0	0.0	**0.933**	0.067	0.0	496.377	1.654

Table 4.6: Empirical distribution of $\hat{m}_n$, mean and mean squared error of $\widehat{\sigma}^2$ over 1000 independent replications for model 2 and $\sigma^2 = 500$.

	Distribution of $\hat{m}_n$							$\widehat{\sigma}^2$	
(p, n)	0	1	2	3	**4**	5	6	Mean	MSE
(30,100)	0.079	0.088	0.090	0.0	**0.649**	0.093	0.001	528.651	11223.872
(60,200)	0.037	0.037	0.029	0.0	**0.794**	0.103	0.0	498.032	1478.184
(120,400)	0.009	0.01	0.005	0.0	**0.880**	0.096	0.0	494.613	107.355
(240,800)	0.003	0.0	0.002	0.0	**0.918**	0.075	0.002	496.813	8.770
(60,100)	0.071	0.104	0.059	0.001	**0.687**	0.078	0	501.754	3126.083
(120,200)	0.036	0.038	0.043	0.0	**0.809**	0.074	0.0	493.687	438.063
(240,400)	0.013	0.007	0.009	0.0	**0.900**	0.071	0.0	494.445	39.686
(480,800)	0.004	0.001	0.0	0.0	**0.941**	0.054	0.0	496.836	3.576

First, we can see the asymptotic consistency of the estimator of m in all the four cases. If we compare these simulations with the known variance case, we can see that the estimation is less accurate in the small (p, n). Furthermore, as in the previous case, the convergence is slower in the $c = 0.6$ case and the estimator performs better in model 1 than in model 2, for both values of σ^2. The estimation of m is more accurate with an unknown variance of $\sigma^2 = 500$. This is due to the fact that the difference between the eigenvalues of S_n

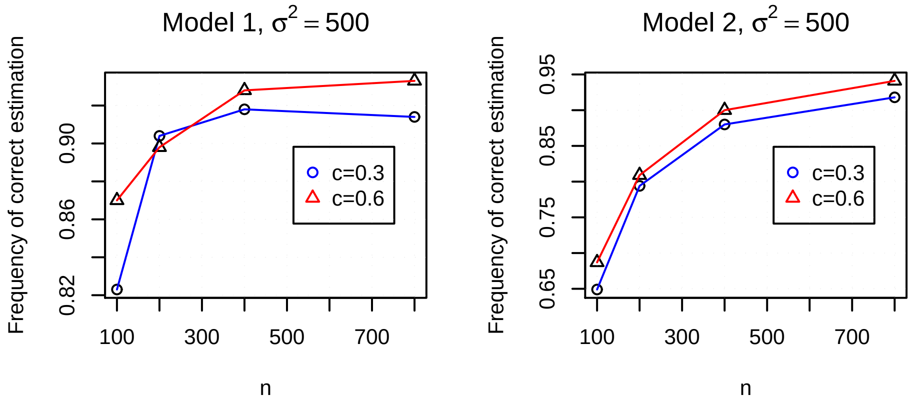

Figure 4.3: Frequency of $\hat{m}_n = m$ over 1000 independent replications with $\sigma^2 = 500$.

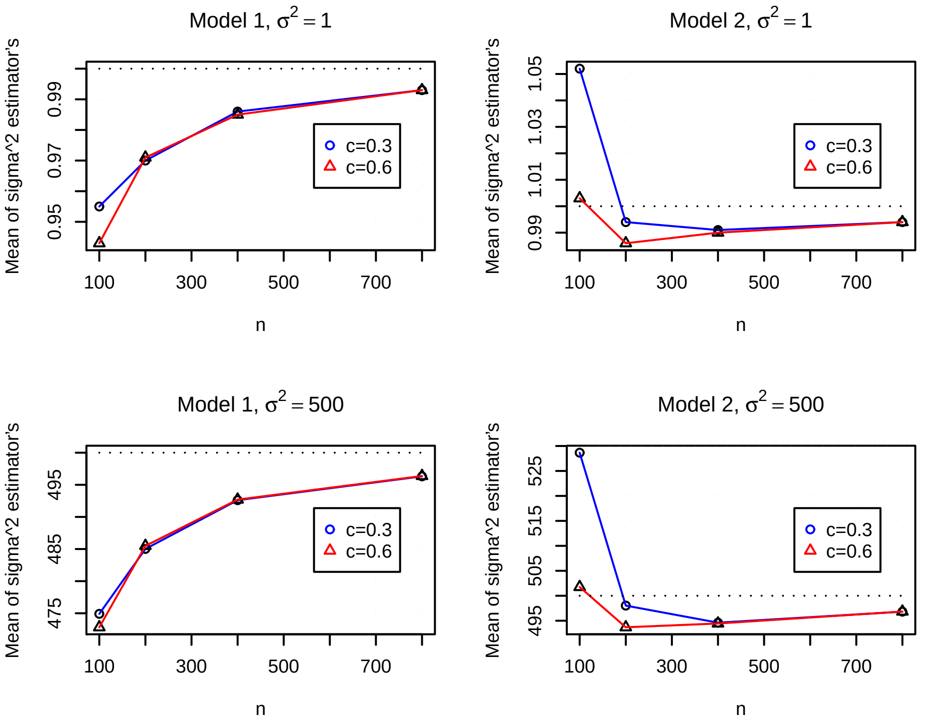

Figure 4.4: Mean of $\widehat{\sigma}^2$ over 1000 independent replications.

corresponding to the spikes is higher in this case, because the spikes are multiplied by σ^2.

We also give the mean and mean squared error of $\hat{m}_n$ in the $\sigma^2 = 1$ case (Tables 4.2 and 4.3) to compare with Table 4.1, where $\sigma^2 = 1$ also, to see the effect of its estimation. The variance and the bias are higher especially for small values of (p, n) in this case with unknown variance.

The estimation of σ^2 performs well, but it seems to be underestimated. There is no particular difference between the two values of c in model 1 but in model 2, contrary to the estimation of m, the convergence seems to be faster in the $c = 0.6$ case for $\widehat{\sigma}^2$. The variance of the estimator decreases if n and p increase, and is lower in the $c = 0.6$ case. As expected, the mean squared error is lower in the $\sigma^2 = 1$ case.

4.5 Comparison with two related methods

In signal processing or econometric literature, the factor model (or linear mixture model) is often used. This model is defined as follows: let $(\mathsf{x}_i = \mathsf{x}(t_i))_{1\leq i \leq n}$ be an i.i.d n-sample of

p-dimensional random vectors satisfying

$$\begin{aligned} \mathsf{x}_i &= \sum_{k=1}^{m} \mathsf{f}_{ki}\Lambda_k + \mathsf{e}_i \\ &= \Lambda \mathsf{f}_i + \mathsf{e}_i, \end{aligned}$$

where

- $\mathsf{f}_i = (\mathsf{f}_{1i}, \dots, \mathsf{f}_{mi})' \in \mathbb{R}^m$ are m random factors (or signals) assumed to have zero mean, unit variance, and mutually independent;
- $\Lambda = (\Lambda_1, \dots, \Lambda_m)$ is a $p \times m$ fixed unknown matrix of rank m (response vectors or factor loadings);
- $\mathsf{e}_i \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_p)$, $\sigma^2 \in \mathbb{R}$ is the noise level.

It is easy to show that in this case, the population covariance matrix takes the form of a spiked population model: the spikes are only slightly modified. If we denote by α the vector of spikes in the factor model, we have the following relationship with our original vector α^*:

$$\alpha^* = \frac{\alpha}{\sigma^2} + 1.$$

Here determining the number of spikes m means the detection of the number of factors/signals m. We will explain and compare two methods from econometrics (Harding (2007)) and signal processing (Kritchman & Nadler (2008)), respectively.

4.5.1 Method of Harding and comparison

In his paper, Harding (2007) uses less restrictive assumptions as the sequence e_i is not necessarily independent, but he simulates a Gaussian model. His general idea is to compare the spectral moments of S_n with the empirical spectral distribution of S_n without the factors (or spikes), and to remove the largest eigenvalues one by one in S_n until a "distance" between the moments is minimum.

More precisely, the variance of the noise is seen as a parameter θ and his idea is to write $\mathsf{S}_n = \Xi_n + \Omega_n$ ($\text{rank}(\Xi_n) = m$) as a sum of a finite rank perturbation Ξ_n of the noise covariance Ω_n. Let $\Pi(\mathsf{S}_n)$ be the vector of the first s moments of the empirical spectral distribution of the covariance matrix S_n, $\Pi(\Omega_n)$ the equivalent for Ω_n and $\Pi(\theta)$ its limit as p and $n \to \infty$, $\frac{p}{n} \to c$. Here is the procedure of Harding:

- First, compute the moments $\Pi(\theta)$ of the asymptotic eigenvalue distribution of the covariance matrix of Ω_n for a large (p, n) sample;
- By Bai & Silverstein (2004), we have that $p\left(\Pi(\Omega_n) - \Pi(\theta)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\Delta, W)$. Consequently, estimate θ by:

$$\hat{\theta}_0 = \underset{\theta}{\text{argmin}} \underbrace{(\Pi(\theta) - \Pi(\mathsf{S}_n)' \, \hat{W}^{-1} \, (\Pi(\theta) - \Pi(\mathsf{S}_n)}_{J(\theta)},$$

where $\hat{W}$ is a consistent estimate of W, computed by estimating θ from a first step estimation with $W = \mathsf{I}_p$;

- Next, remove the largest eigenvalue of the spectrum of S_n and re-estimate the parameter θ as previously to get a new estimate $\hat{\theta}_1$;
- This step is repeated by progressively removing large eigenvalues and for prefixed number of times to get a sequence of estimates $\hat{\theta}_2$, $\hat{\theta}_3$, ...etc;
- Finally, among the minimized objective functions $J(\hat{\theta}_i)$ choose the order one which corresponds to the smallest minimized value

$$\hat{m}_0 = \underset{i}{\operatorname{argmin}} J(\hat{\theta}_i).$$

Actually, we know that for m fixed and p, $n \to \infty$, $\Pi(\mathsf{S}_n) \to \Pi(\theta)$. So the criterion is the minimization of the variance $W = W(\theta)$: it decreases until m (until we have removed the eigenvalues corresponding to the spikes), then it stays stable. The procedure of Harding leads to an underestimation of m, at p and n fixed. That is why he penalized the function J with a function of type $k\hat{\theta}g(p,n)$, where k is the number of eigenvalues removed, $\hat{\theta}$ is the estimated variance at the step q and $g(p,n)$ is a function such that $g(p,n) \to 0$ when p, $n \to \infty$. The finally proposed choice for g is the following function given by Bai & Ng (2002) based on a BIC criterion

$$g(p,n) = \left(\frac{p+n}{pn}\right) \ln\left(\frac{pn}{p+n}\right).$$

For his simulation experiments, he tested four different "distances" but we only keep the one based on the BIC criterion which is the best. Furthermore, we do not give all cases he tested. The simulation design was a little bit different, indeed Harding does not choose the spikes directly, but he generates e_i as a Gaussian law $\mathcal{N}(0, \mathsf{I}_p)$ and Λ in a deterministic way. We calculate the corresponding spikes and it leads to the following values:

- $(p,n) = (30,100)$: $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*) = (258.719, 16.973, 10.038, 6.877, 3.817)$;
- $(p,n) = (90,100)$: $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*) = (259.010, 18.101, 10.785, 7.276, 3.692)$;
- $(p,n) = (210,300)$: $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*) = (259.083, 18.418, 10.992, 7.377, 3.649)$;
- $(p,n) = (250,500)$: $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*) = (259.005, 18.453, 11.057, 7.448, 3.634)$.

Nonetheless, these cases stay very close. Below we compare his results to ours. We only give in Table 4.7 the mean and mean squared errors of the estimator as reported in Harding's paper.

Both methods perform well and their results are overall very close except that Harding's estimation yields a slightly smaller MSE for $\hat{m}_n$. However, one should have in mind that this estimation has a very complex construction and a rigorous justification of its different steps is still open. Moreover, the spikes in Table 4.7 are large and well-separated one from each other; it remains unclear how this method will perform in a case where the spikes are much smaller and closer like in model 2, considered in sections 4.3 and 4.4. By contrast, our estimator has a very simple construction and we proved its consistency under reasonable assumptions.

Table 4.7: Compared mean and mean squared error of our $\widehat{m}_n$ and $\widehat{\sigma}^2$ and those of Harding over 5000 independent replications and $\sigma^2 = 1$.

	$\widehat{m}_n$				$\widehat{\sigma}^2$			
	Harding's estimator		PY estimator		Harding's estimator		PY estimator	
(p, n)	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
(30,100)	5.028	0.028	5.087	0.266	0.942	0.004	0.946	0.008
(90,100)	5.040	0.048	5.049	0.232	0.944	0.001	0.943	0.0
(210,300)	5.004	0.004	5.087	0.082	0.982	0.0	0.980	0.0
(250,500)	5.002	0.002	5.077	0.072	0.989	0.0	0.988	0.0

4.5.2 Method of Kritchman & Nadler and comparison

These authors assume the Gaussian case. In the absence of spikes, $n\mathsf{S}_n$ follows a Wishart distribution with parameters n, p. In this case, Johnstone (2001) gave the asymptotic distribution of the largest eigenvalue of S_n.

Proposition 19. *Let S_n be the sample covariance matrix of n vectors distributed as $\mathcal{N}(0, \sigma^2 \mathsf{I}_p)$, and $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ be its eigenvalues. Then, when $n \to \infty$, such that $\frac{p}{n} \to c > 0$*

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

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

We assume that the variance σ^2 is known. To distinguish a spike eigenvalue λ from a non-spike one at an asymptotic significance level γ, their idea is to check whether

$$\lambda_{n,k} > \sigma^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b\right) \tag{4.2}$$

where the value of $s(\gamma)$ 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 $k = 1, 2, \ldots, \min(p, n) - 1$,

$$\mathcal{H}_0\colon\ m \geq k \ vs.\ \mathcal{H}_1\colon\ m \leq k - 1\ .$$

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 (4.2). If (4.2) 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 spikes is estimated to be $\widehat{m}_{n,2} = k - 1$. Formally, their estimator is defined by

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

When σ^2 is unknown, they estimate it by the same method we used. For their simulations, they use four different settings, with $\sigma^2 = 1$

- A1: $\alpha = (200, 50)$, $c = 4$ (i.e. $\alpha^* = (201, 51)$);
- A2: $\alpha = (200, 50)$, $c = 1$;
- B1: $\alpha = (200, 50, 10, 5)$, $c = 4$ (i.e. $\alpha^* = (201, 51, 11, 6)$);
- B2: $\alpha = (200, 50, 10, 5)$, $c = 1$;

with $p = 64$ and $p = 1024$. Notice that, contrary to ours and those of Harding, in their simulation, $c > 1$ and the difference between two consecutive spikes is higher. We add two settings with different variance

- A2': $\alpha = (200, 50)$, $c = 1$, $\sigma^2 = 20$ (i.e. $\alpha^* = (11, 3.5)$);
- B2': $\alpha = (200, 50, 10, 5)$, $c = 1$, $\sigma^2 = 2$ (i.e. $\alpha^* = (101, 26, 6, 3.5)$);

and $p = 64$. The results are displayed in Tables 4.8 and 4.9.

Table 4.8: Summary for $p = 64$ showing the frequency of $\hat{m} = m$.

Setting	Our estimator	Estimator KN
A1 ; $(p, n) = (64, 16)$	0.943	0.994
A2 ; $(p, n) = (64, 64)$	0.966	0.993
A2'; $(p, n) = (64, 64)$	0.602	0.513
B1 ; $(p, n) = (64, 16)$	0.348	0.238
B2 ; $(p, n) = (64, 64)$	0.947	0.995
B2'; $(p, n) = (64, 64)$	0.734	0.682

With small p and n, both estimators performs well, except for the A2', B1, and B2' cases where the spikes α^* are closer to $1 + \sqrt{c}$ than in the other cases.

Table 4.9: Summary for $p = 1024$ showing the frequency of $\hat{m} = m$.

Setting	Our estimator	Estimator KN
A1; $(p, n) = (1024, 256)$	0.995	0.994
A2; $(p, n) = (1024, 1024)$	0.986	0.993
B1; $(p, n) = (1024, 256)$	0.999	0.999
B2; $(p, n) = (1024, 1024)$	0.986	0.994

With larger p and n, the results from both methods are comparable. Nevertheless, theoretical properties remain unclear for the KN estimator: it is proved that

$$\lim_{p,n\to\infty} \mathbb{P}\left(\widehat{m}_{n,2} \geq m\right) = 1,$$

and, in the one factor case ($m = 1$) that

$$\lim_{p,n\to\infty} \mathbb{P}\left(\widehat{m}_{n,2} > m\right) = \gamma.$$

That is by construction, the proposed estimator cannot be fully consistent but nearly consistent with an incompressible asymptotic error of γ. Actually the authors are using a very small test level $\gamma = 0.005$ in their experiments. Whether this property remains true for general case with more than one spike stays open and even so, this near-consistency is a bit unsatisfactory from a theoretical point a view.

4.6 Case of spikes with multiplicity greater than one

The problem with two identical spikes is that the difference between the corresponding eigenvalues of the sample covariance matrix will tend to zero, as the non-spike ones. Nevertheless, we tried to estimate the number of spikes with the same procedure, and our method still works: we can explain it by the fact that the convergence of the $\lambda_{n,i}$, for $i > m$ (non-spikes) is in $O_{\mathbb{P}}\left(n^{-2/3}\right)$, whereas that of the difference corresponding of two identical spikes is in $O_{\mathbb{P}}\left(n^{-1/2}\right)$ (consequence of Theorem 3.1 of Bai & Yao (2008)). Consequently, for finite n, the difference $\delta_{n,i}$ corresponding to two equals spikes will be still higher than the $\delta_{n,i}$ corresponding to non-spikes. Furthermore, the variance in the convergence of this difference is $2\alpha^{*2}\left(1 - \frac{c}{(\alpha^*-1)^2}\right) \underset{\infty}{\sim} 2\alpha^{*2}$, which is quite high for high spikes. A complete justification of our method in this case with multiple spikes is described in chapter 5. Here we provide some simulation results in order to have a first idea about its performance.

We will only consider the known variance case. If it is not the case, the procedure explained before will apply without any problem. Here are the results with the same simulation design as previously, except that we introduce multiple spikes. We consider two models:

- **Model 3**: $m = 6$, $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*) = (259.7, 259.7, 18, 11.1, 7.9, 4.8)$;
- **Model 4**: $m = 6$, $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*) = (7, 6, 6, 6, 5, 4)$.

For each model, two different values of c, 0.3 and 0.6, are considered, and we give in Figure 4.5 the frequency of $\hat{m}_n = m$ and in Table 4.10 the mean and the mean squared error of our estimator over 1000 independent replications.

Table 4.10: Mean and mean squared error of $\hat{m}_n$ over 1000 independent replications for model 1 and 2.

	Model 3, $m = 6$		Model 4, $m = 6$	
(p, n)	Mean	MSE	Mean	MSE
(30,100)	6.085	0.168	4.529	4.393
(60,200)	6.077	0.121	4.86	4.199
(120,400)	6.088	0.082	5.31	3.061
(240,800)	6.073	0.068	5.597	2.051
(60,100)	6.043	0.151	4.118	4.797
(120,200)	6.092	0.108	4.614	4.453
(240,400)	6.081	0.074	5.159	3.447
(480,800)	6.079	0.073	5.562	2.058

In both cases, we can observe the asymptotic consistency of the estimator, but the convergence is slower in model 4. Indeed, the eigenvalue spacings are smaller. Furthermore, the values of the spikes are small, so that the variance in the convergence of the spikes is not very high and the fluctuations of the difference are smaller than in model 3.

4.7 Complement: on the choice of the sequence d_n

In this section, we present a complement on the choice of the sequence d_n, which has not been published in Passemier & Yao (2012b).

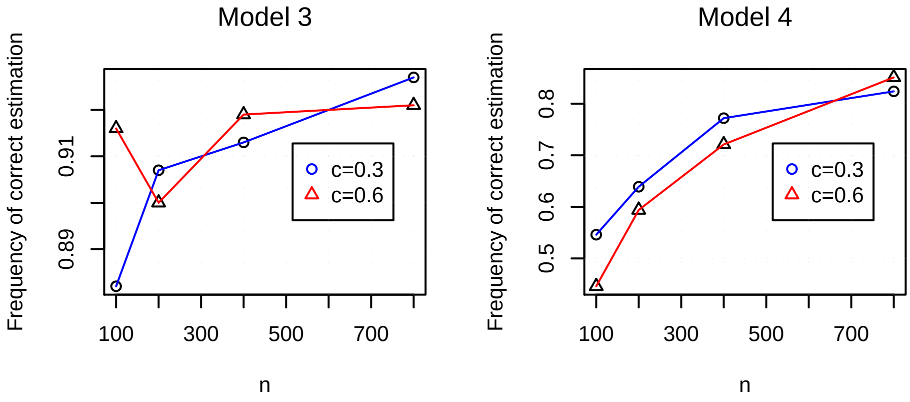

Figure 4.5: Frequency of $\hat{m}_n = m$ over 1000 independent replications.

4.7.1 Introduction

In Theorem 1, we can choose any sequence satisfying the requirement $d_n \to 0$ such that $n^{2/3} d_n \to \infty$. In practice, the choice of d_n is an important but also difficult question to address. In our simulations, we used a sequence of the form $\frac{a_n}{n^{2/3}}\beta$, where β is the variance in Proposition 3.2 and $a_n = 4\sqrt{2 \log \log n}$: this choice came when thinking about a result analog to the law of the iterated logarithm. With this chosen sequence d_n, we saw that our estimator performed well. Nevertheless, we mentioned a slight over-estimation of m.

In this complement, we present a new sequence which performs better despite the fact that it requires to know α_m.

4.7.2 The sequence

To avoid the over-estimation of m, we need to increase slightly the sequence d_n. Proposition 16 shows that the Gaussian convergence of the eigenvalues $\lambda_{n,j}$ of S_n to $\phi(\alpha_j^*)$ is in $1/\sqrt{n}$, that is why we choose a new sequence $\tilde{d}_n$ in $o(1/\sqrt{n})$. As the variance $\sigma^2(\alpha_j^*)$ in this Gaussian convergence is an increasing function of α_j^*, we decide to take into account $\sigma(\alpha_m)$: if α_m^* is large, $\tilde{d}_n$ is increased so we reduce the overestimation of m. We define:

$$\tilde{d}_n = C\frac{\sigma(\alpha_m^*)}{\sqrt{n}}$$

where C is a constant to determine. After several experiments, we set $C = 0.75$.

4.7.3 Simulation experiments

4.7.3.1 Comparison with our previous simulations experiments

We keep the same parameter as in the previous simulation studies. As the estimation of the variance σ^2 has only a slightly influence, we do not estimate it and we use $\sigma^2 = 1$:

- **Model 1**: $m = 5$, $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*) = (259.72, 17.97, 11.04, 7.88, 4.82)$;
- **Model 2**: $m = 4$, $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*) = (7, 6, 5, 4)$.

For each model, two different values of c, 0.3 and 0.6, are considered. We give in Tables 4.11 and 4.12, respectively, the distribution of $\hat{m}_n$, its mean and mean squared error over 1000 independent replications.

Table 4.11: Mean, mean squared error and empirical distribution of $\hat{m}_n$ over 1000 independent replications for model 1.

			Distribution of $\hat{m}_n$					
(p, n)	Mean	MSE	1	2	3	4	**5**	6
(30,100)	4.964	0.153	0.002	0.011	0.003	0.0	**0.973**	0.11
(60,200)	4.991	0.0	0.005	0.001	0.0	0.0	**0.989**	0.60
(120,400)	5.0	0.0	0.0	0.0	0.0	0.0	**1**	0.0
(240,800)	5.0	0.0	0.0	0.0	0.0	0.0	**1**	0.0
(60,100)	4.939	0.245	0.002	0.019	0.007	0.0	**0.954**	0.018
(120,200)	5.002	0.038	0.0	0.003	0.0	0.0	**0.986**	0.011
(240,400)	5.004	0.0	0.0	0.0	0.0	0.0	**0.996**	0.004
(480,800)	5.072	0.069	0.0	0.0	0.0	0.0	**1**	0.0

Table 4.12: Mean, mean squared error and empirical distribution of $\hat{m}_n$ over 1000 independent replications for model 2.

			Distribution of $\hat{m}_n$					
(p, n)	Mean	MSE	0	1	2	3	**4**	5
(30,100)	3.408	1.519	0.045	0.086	0.088	0.0	**0.759**	0.02
(60,200)	3.648	1.059	0.038	0.044	0.041	0.0	**0.863**	0.014
(120,400)	3.894	0.363	0.009	0.019	0.012	0.0	**0.949**	0.011
(240,800)	4.062	0.110	0.002	0.001	0.0	0.0	**1**	0.0
(60,100)	3.428	1.604	0.056	0.069	0.092	0.001	**0.758**	0.044
(120,200)	3.724	0.885	0.024	0.037	0.052	0.0	**0.852**	0.035
(240,400)	3.969	0.394	0.013	0.014	0.011	0.0	**0.947**	0.015
(480,800)	3.899	0.383	0.003	0.0	0.0	0.0	**0.99**	0.01

In both cases, we can observe an improvement of the results, especially in model 1: the convergence is faster than in model 2. The aim of reduce the overestimation is reached. We can still notice a slower convergence in the $c = 0.6$ case.

4.7.3.2 Comparison with the method of Kritchman and Nadler

We recall the previous settings used by Kritchman & Nadler (2008). For their simulations, they use four different settings, with $\sigma^2 = 1$:

- A1: $\alpha = (200, 50)$, $c = 4$ (i.e. $\alpha^* = (201, 51)$;
- A2: $\alpha = (200, 50)$, $c = 1$;
- B1: $\alpha = (200, 50, 10, 5)$, $c = 4$ (i.e. $\alpha^* = (201, 51, 11, 6)$;
- B2: $\alpha = (200, 50, 10, 5)$, $c = 1$;

with $p = 64$ and $p = 1024$. Notice that in these simulations, $c > 1$ and the difference between two consecutive spikes is higher. We still do not estimate the variance σ^2. The results are displayed in Tables 4.13 and 4.14.

Table 4.13: Summary for $p = 64$ showing the frequency of $\hat{m} = m$.

Setting	With d_n	With $\tilde{d}_n$	Estimator KN
A1; $(p, n) = (64, 16)$	0.943	**0.987**	0.994
A2; $(p, n) = (64, 64)$	0.966	**1**	0.993
B1; $(p, n) = (64, 16)$	0.348	**0.528**	0.238
B2; $(p, n) = (64, 64)$	0.947	**0.986**	0.995

With small p and n, there is an improvement with the use of $\tilde{d}_n$: it is close to the results of the estimator KN for A1 and B2 cases, but better in the two other cases.

Table 4.14: Summary for $p = 1024$ showing the frequency of $\hat{m} = m$.

Setting	With d_n	With $\tilde{d}_n$	Estimator KN
A1; $(p, n) = (1024, 256)$	0.995	**1**	0.994
A2; $(p, n) = (1024, 1024)$	0.986	**1**	0.993
B1; $(p, n) = (1024, 256)$	0.999	**0.98**	0.999
B2; $(p, n) = (1024, 1024)$	0.986	**1**	0.994

With larger p and n, the new $\tilde{d}_n$ performs better than the other two, except for the B1 case, where the spikes are closer than in the A case.

4.7.3.3 Comparison of d_n and $\tilde{d}_n$

We draw the plot of the two sequences showing the difference between them in Figure 4.6. We consider two cases:

- Case 1: c=0.3, $\alpha_m = 4,82$;
- Case 2: c=1, $\alpha_m = 6$;

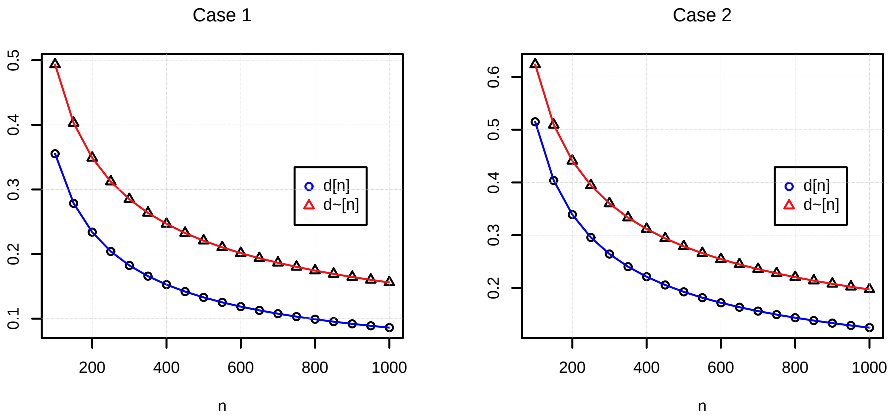

Figure 4.6: Comparison of d_n and $\tilde{d}_n$.

The two sequences are close in both cases: a small difference can change the performance of the estimation.

4.7.4 Conclusion

The new sequence performs better in most cases. The problem is now how to estimate $\sigma(\alpha_m^*)$. One can estimate α_m^* by inverting the function ϕ (4.1), but here m is assumed to be unknown. This requires further work to construct a procedure which take into account this hypothesis. Nevertheless, despite this new sequence can not be use in practice, this study underlines the difficulty of choosing the sequence d_n and shows that we can choose a sequence which performs better.

Chapter 5

Estimation of the number of factors, possibly equal, in the high-dimensional case

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 chapter, 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.

Keywords: Factor model, covariance matrix, random matrix theory, high-dimensional statistics, Tracy-Widom laws.

AMS subject classification: 62F07, 62F12, 60B20.

This chapter is submitted in
Journal of Multivariate Analysis
under the title:
"Estimation of the Number of Factors, Possibly Equal, in the High-Dimensional Case."

It has been written in collaboration with Jian-Feng Yao.

5.1 Introduction

The factor model appears in many scientific fields, such as economics and psychology literature, where the number of factors has a primary importance (Anderson (2003), Ross (1976)). Similar models can be found in physics of mixture (see Kritchman & Nadler (2008), Naes et al. (2002)) 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 have been introduced in Johnstone (2001) that encompass factors 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 Wax & Kailath (1985) 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 Combettes & Silverstein (1992). Recent advances have been made using random matrix theory by Harding (2007) or Onatski (2009) in economics, and Kritchman & Nadler (2008) in chemometrics literature.

Several studies have also appeared in the area of signal processing from high-dimensional data. Everson & Roberts (2000) proposed a method using both RMT and bayesian inference, while Ulfarsson & Solo (2008) combined random matrix theory and Stein's Unbiased Risk Estimator (SURE). Nadakuditi & Edelman (2008) and Nadler (2010) improved estimators based on information theoretic criteria and Kritchman & Nadler (2009) constructed an estimator based on the distribution of the largest eigenvalue (hereafter refereed as the KN estimator). In Passemier & Yao (2012b), we have also introduced a new method based on recent results of Bai & Yao (2008) and Paul (2007) 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 Lam et al. (2011) and Lam & Yao (2012).

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 5.3.2. The aim of this work is to extend our method Passemier & Yao (2012b) 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 chapter is organized as follows. Section 5.2 introduces the model. In Section 5.3, we define the estimation problem of the number of possibly equal factors and present our solution. We establish its asymptotic consistency. Section 5.4 provides simulation experiments to assess the quality of our estimator. Next, we recall the KN estimator and conduct simulation experiments to compare these two methods. In Section 5.6, we analyze the influence of a tuning parameter C used in our estimator. Finally, Section 5.7 concludes with discussions. All proofs are given in the appendix.

5.2 Problem formulation

We consider the following strict factor model

$$\begin{aligned} \mathsf{x}_i &= \sum_{k=1}^{m} \mathsf{f}_{ki}\Lambda_k + \mathsf{e}_i & (5.1)\\ &= \Lambda \mathsf{f}_i + \mathsf{e}_i, & (5.2) \end{aligned}$$

where

- $\mathsf{f} = (\mathsf{f}_{1i}, \ldots, \mathsf{f}_{mi})' \in \mathbb{R}^m$ are m random factors ($m < p$) assumed to have zero mean, unit variance and be independent;
- $\Lambda = (\Lambda_1, \ldots, \Lambda_m)$ is the $p \times m$ full rank matrix of factors loadings;
- $\mathsf{e} \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_p)$ is a $p \times 1$ vector of additive noise, independent from f_i, $\sigma^2 \in \mathbb{R}$ is the unknown noise level.

The *population covariance matrix* $\Sigma = \mathrm{cov}(\mathsf{x}_i)$ of x_i equals $\Lambda\Lambda' + \sigma^2 \mathsf{I}_p$ and has the spectral decomposition

$$\mathsf{W}'\Sigma\mathsf{W} = \sigma^2 \mathsf{I}_p + \mathrm{diag}(\alpha_1, \ldots, \alpha_m, 0, \ldots, 0)$$

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

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

Denote by $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ its eigenvalues. Our aim is to estimate m on the basis of S_n. To start with, we assume that the noise level σ^2 is known. If this is indeed not the case, we will give a method in Section 5.3.3 to estimate it.

5.3 Estimation of the number of factors

In this section, we first recall our previous result of Passemier & Yao (2012b) 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.

5.3.1 Previous work: estimation with different factors

We consider the case where the $(\alpha_i)_{1 \leq i \leq m}$ are all different, so there are m distinct factors. According to Passemier & Yao (2012b), let us rewrite the spectral representation of Σ as

$$\mathsf{W}'\Sigma\mathsf{W} = \sigma^2 \mathrm{diag}(\alpha_1^*, \ldots, \alpha_m^*, 1, \ldots, 1),$$

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 $\alpha_1^* > \cdots > \alpha_m^* > 1 + \sqrt{c}$, i.e. all the factors strengths $(\alpha_i)_{1 \leq i \leq m}$ are greater than $\sigma^2\sqrt{c}$. For $\alpha \neq 1$, we define the function

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

Baik & Silverstein (2006) proved that, under a moment condition on x, for each $k \in \{1, \ldots, m\}$ and almost surely,

$$\lambda_{n,k} \longrightarrow \sigma^2 \phi(\alpha_k^*).$$

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

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

The estimation method of m in Passemier & Yao (2012b) 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 $j \geq m$ whereas for $j < m$, $\delta_{n,j}$ tends to a positive limit. Thus it becomes possible to estimate m from index-numbers j where $\delta_{n,j}$ becomes small. More precisely, the estimator is

$$\hat{m}_n \quad = \quad \min\{j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n\}, \tag{5.3}$$

where $s > m$ is a fixed sufficiently large number, 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 Passemier & Yao (2012b), we proved the consistency of $\hat{m}_n$ 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 x:

Assumption 2. The entries x_{ij} of the random vector x have a symmetric law and a sub-exponential decay, that means there exist positive constants C, C' such that, for all $t \geq C'$,

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

5.3.2 Estimation with possibly equal factors

As said in the introduction, when some factors have close values, estimation algorithms need to be modified. More precisely, we adopt the following theoretical 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,

$$\begin{aligned}\mathrm{spec}(\Sigma) &= (\underbrace{\alpha_1,\ldots,\alpha_1}_{n_1},\ldots,\underbrace{\alpha_K,\ldots,\alpha_K}_{n_K},\underbrace{0,\ldots,0}_{p-m}) + \sigma^2(\underbrace{1,\ldots,1}_{p}) \\ &= \sigma^2(\underbrace{\alpha_1^*,\ldots,\alpha_1^*}_{n_1},\ldots,\underbrace{\alpha_K^*,\ldots,\alpha_K^*}_{n_K},\underbrace{1,\cdots,1}_{p-m}).\end{aligned}$$

with $n_1+\cdots+n_K = m$. When all the factors are unequal, differences between sample factor eigenvalues tend to a positive constant, whereas with two equal factors, such differences 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 $i > m$ (noise) is faster (in $O_{\mathbb{P}}(n^{-2/3})$) than that of the $\delta_{n,i}$ from equal factors (in $O_{\mathbb{P}}(n^{-1/2})$) as a consequence of Theorem 3.1 of Bai & Yao (2008). This is the key feature we use to adapt the estimator (5.3) to the current situation with a new threshold d_n. The precise asymptotic consistency is as follows:

Theorem 2. *Let $(\mathsf{x}_i)_{1\leq i\leq n}$ be n copies i.i.d. of x which follows the model (5.2) and satisfies Assumption 2. Suppose that the population covariance matrix Σ has K non null and non unit eigenvalues $\alpha_1 > \cdots > \alpha_K > \sigma^2\sqrt{c}$ with respective multiplicity $(n_k)_{1\leq k\leq K}$ ($n_1+\cdots+n_K = m$), and $p-m$ unit eigenvalues. Assume that $\frac{p}{n} \to c > 0$ 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 $\widehat{m}_n$ is consistent, i.e $\widehat{m}_n \to m$ in probability when $n \to \infty$.*

Notice that, compared 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 2 is postponed to the appendix.

5.3.3 Estimation of the noise level

When the noise level σ^2 is unknown, an estimation is needed. In Passemier & Yao (2012b), we used an algorithm based on the maximum likelihood estimate

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

As explained in Kritchman & Nadler (2008, 2009), 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 $\widehat{m}_n$ and estimator $\tilde{m}_n$ (see Section 5.5).

5.4 Simulation experiments

To assess the quality of our estimator, we first make the following modification: instead of making a decision once some difference $\delta_{n,k}$ is below the threshold d_n (see (5.3)), the modified estimator stops when two consecutive differences $\delta_{n,k}$ and $\delta_{n,k+1}$ are both below

d_n. More precisely, we set

$$\hat{m}_n^* \quad = \quad \min\{j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n \text{ and } \delta_{n,j+2} < d_n\}. \tag{5.4}$$

It is easy to see that the proof for the consistency of $\hat{m}_n$ applies equally to $\hat{m}_n^*$ under the same conditions as in Theorem 2.

It remains to choose a threshold sequence d_n to be used for our estimator $\hat{m}_n^*$. As argued in Passemier & Yao (2012b), 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 5.1 gives a summary of parameters in our simulation experiments. There are two sets of experiments. In the first one (Figures 5.1, 5.2 and models A, B, C and D in Table 5.1), factors are different and these experiments extend and complete results already reported in Passemier & Yao (2012b). The second set of experiments (Figures 5.3, 5.4 and models E, F, G, H and J in Table 5.1) addresses the new situation where some factors are equal. Figure 5.7 considers the case of no factor. (Figures 5.5 and 5.6 report comparison results developed in Section 5.5).

Table 5.1: Summary of parameters used in the simulation experiments. (L: left, R: right)

Fig.	Factors	Mod.	Factor	Fixed parameters				Var.
No.		No.	values	p, n	c	σ^2	C	par.
5.1	Different		(α)	$(200, 800)$ $(2000, 500)$	1/4 4	Given	5.5 9	α
5.2L	Different	A B B	$(6, 5)$ $(10, 5)$ $(10, 5)$		10	Given Estimated	11	n
5.2R	Different	C D	(1.5) $(2.5, 1.5)$		1	Given	5	n
5.3	Possibly equal	E F	$(\alpha, \alpha, 5)$ $(\alpha, \alpha, 15)$	$(200, 800)$ $(2000, 500)$	1/4 4	Given	6 9.9	α
5.4L	Possibly equal	G H H	$(6, 5, 5)$ $(10, 5, 5)$ $(10, 5, 5)$		10	Given Estimated	9.9	n
5.4R	Possibly equal	I J	$(1.5, 1.5)$ $(2.5, 1.5, 1.5)$		1	Given	5	n
5.5	Models A and D							
5.6	Models G and J							
5.7	No factor	K	No factor		1 10	Given	8 15	n
5.8L	Models A and G							
5.8R	Models B and H							
5.9L	Models C and I, with C automatically chosen							
5.9R	Models D and J, with C automatically chosen							

5.4.1 Case of different factors

In Figure 5.1, we consider the case of a single factor of strength α, and we analyze the probability of misestimation as a function of factor strength 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.

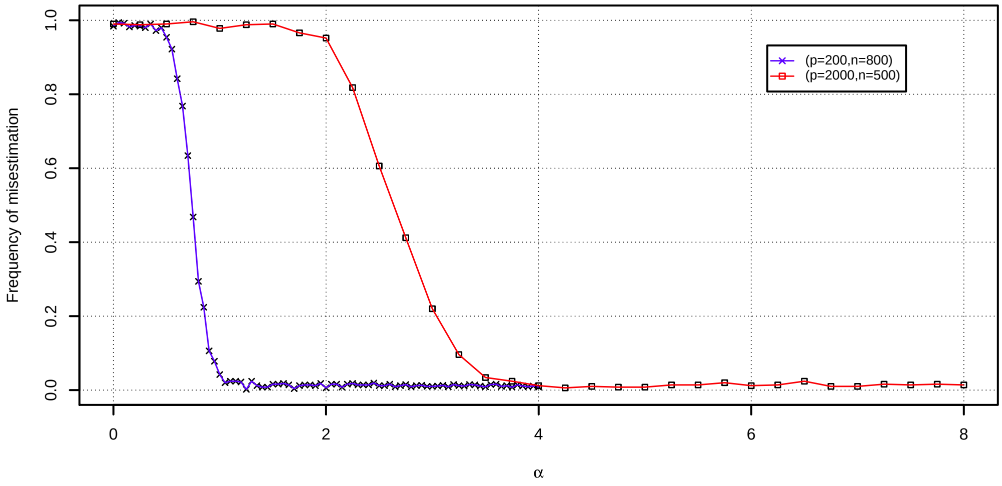

Figure 5.1: Misestimation rates as a function of factor strength for $(p, n) = (200, 800)$ and $(p, n) = (2000, 500)$.

Our estimator performs well: we recover the threshold from which the behavior of the factor eigenvalues differs from the noise ones ($\sqrt{c} = 0.5$ for the first case, and 2 for the second).

In Figure 5.2 left panel, we consider two models with two factors ($m = 2$), 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 $C = 11$.

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 σ^2 given, the estimation of σ^2 does not affect our estimator significantly, which seems robust against the unknown noise level.

Figure 5.2 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 $C = 5$.

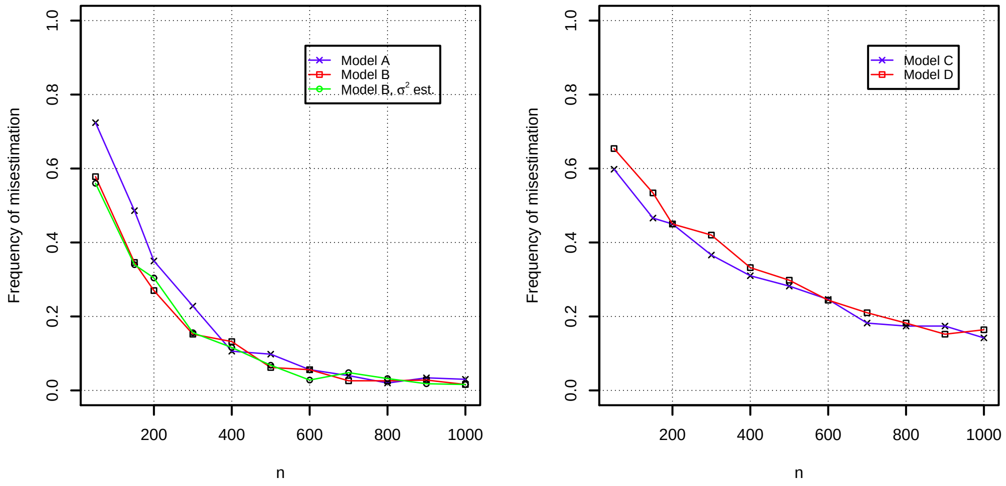

Figure 5.2: Misestimation rates as a function of n for models A, B (left) and model C, D (right).

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 5.3 and 5.4 which are to be compared to Figures 5.1 and 5.2. In Figure 5.3, 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 $m = 3$, $C = 6$ for model E and $C = 9.9$ for model F.

In Figure 5.4 left panel, we consider two models, analog to model A and B, with three factors ($m = 3$):

- 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 σ^2 is given. Comparing to the case of different factors (Figure 5.2), 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 5.2, estimation of an unknown variance σ^2 does not affect our estimator significantly.

Figure 5.4 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: $(\alpha_1, \alpha_2, \alpha_2) = (2.5, 1.5, 1.5)$.

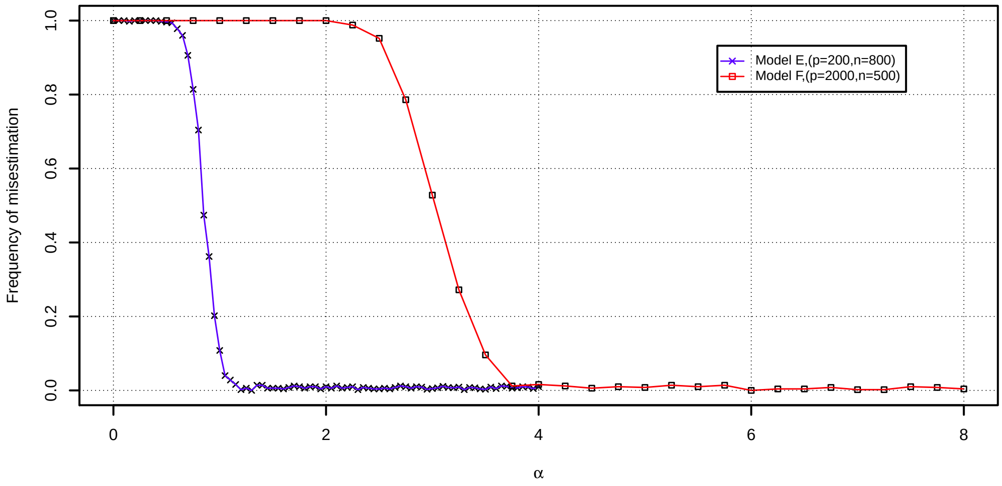

Figure 5.3: Misestimation rates as a function of factor strength for $(p, n) = (200, 800)$, model E and $(p, n) = (2000, 500)$, model F.

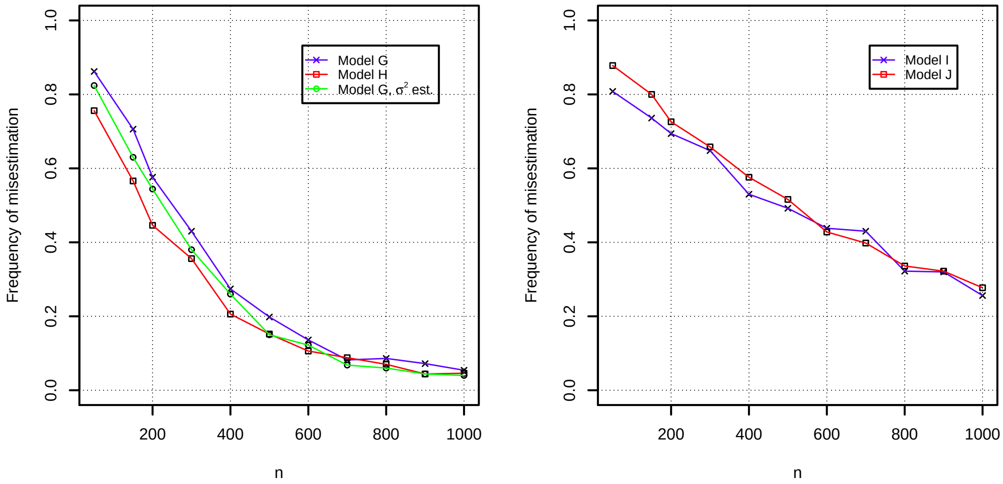

Figure 5.4: Misestimation rates as a function of n for model G, H (left) and model I, J (right).

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.

5.5 Method of Kritchman & Nadler and comparison

5.5.1 Algorithm of Kritchman & Nadler

In their paper, Kritchman & Nadler (2008, 2009) 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), see Wax & Kailath (1985). In most of the studied cases, the estimator KN performs better. Furthermore, in Nadler (2010) 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, $n\mathsf{S}_n$ follows a Wishart distribution with parameters n, p. In this case, Johnstone (2001) has provided the asymptotic distribution of the largest eigenvalue of S_n.

Proposition 20. *Let S_n be the sample covariance matrix of n vectors distributed as $\mathcal{N}(0, \sigma^2 \mathsf{I}_p)$, and $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$ be its eigenvalues. Then, when $n \to \infty$, such that $\frac{p}{n} \to c > 0$*

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

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

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

$$\lambda_{n,k} > \sigma^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b\right) \tag{5.5}$$

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 $k = 1, 2, \ldots, \min(p, n) - 1$,

$$\mathcal{H}_0^{(k)}\colon m \leq k - 1 \quad vs. \quad \mathcal{H}_1^{(k)}\colon m \geq k.$$

For each value of k, if (5.5) is satisfied, $\mathcal{H}_0^{(k)}$ is rejected and k is increased by one. The procedure stops once an instance of $\mathcal{H}_0^{(k)}$ is accepted and the number of factors is then estimated to be $\tilde{m}_n = k - 1$. Formally, their estimator is defined by

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

Here $\widehat{\sigma}$ is some estimator of the noise level (as discussed in Section 5.3.3). The authors proved the strong consistency of their estimator as $n \to \infty$ with fixed p, by replacing the fixed confidence level γ with a sample-size dependent one γ_n, where $\gamma_n \to 0$ sufficiently slow as $n \to \infty$. They also proved that $\lim_{p,n\to\infty} \mathbb{P}(\tilde{m}_n \geq m) = 1$.

It is important to notice here that the construction of the KN estimator differs from ours, essentially because of the fixed alarm rate γ. We will discuss the issue of the false alarm rate in the last section.

5.5.2 Comparison with our method

We give a value of $\gamma = 0.5\%$ to the false alarm rate of the estimator KN, as suggested in Kritchman & Nadler (2009) and use their algorithm available at the author's homepage.

In Figure 5.5, 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.

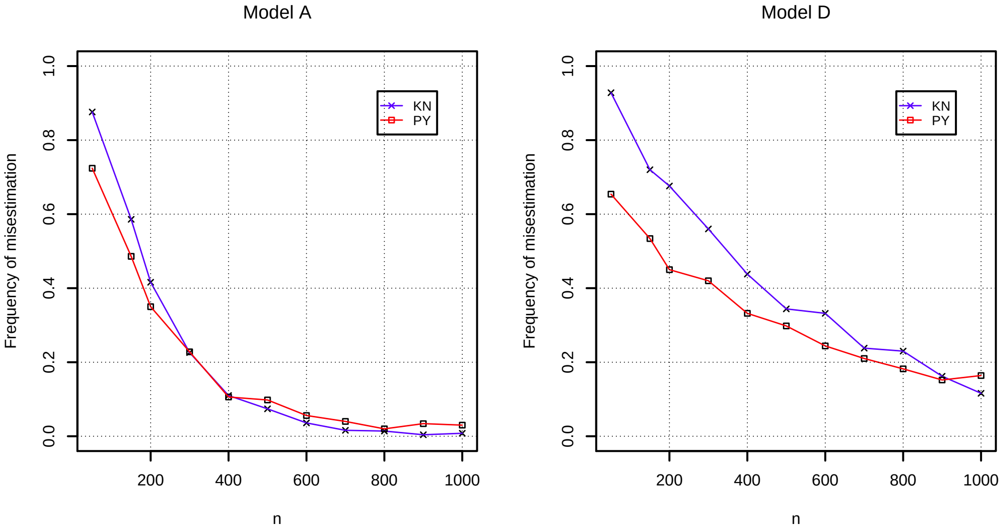

Figure 5.5: Misestimation rates as a function of n for model A (left) and model D (right).

For model A, the performance of the two estimators 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 D, 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 5.6 considers model G and J, two models analog to model A and D but with two equal factors:

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

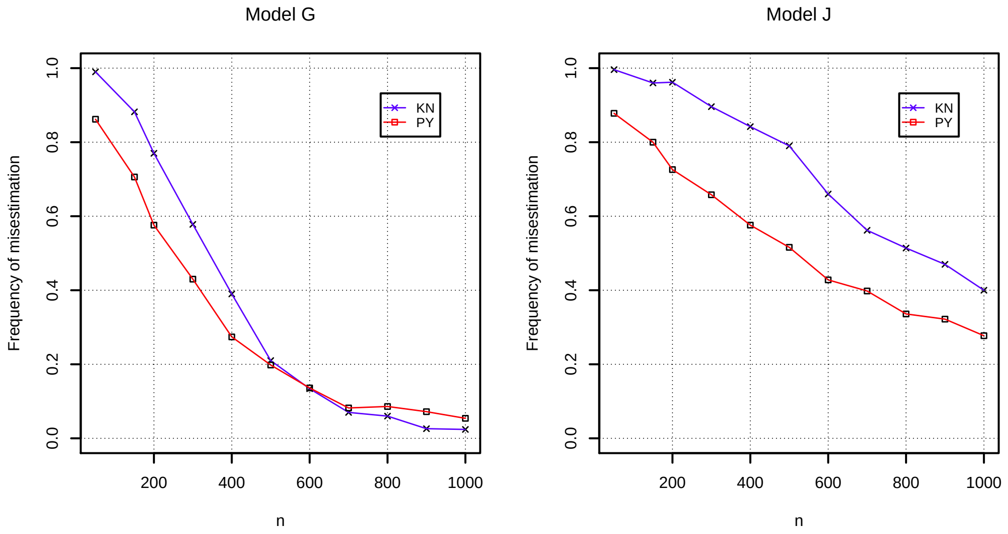

Figure 5.6: Misestimation rates as a function of n for model G (left) and model J (right).

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 5.7 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.

5.6 On the tuning parameter C

5.6.1 Influence of C on the misestimation and false alarm rate

In the simulation experiments, we choose 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 determining the number of factors, the different methods should have comparable false alarm probabilities. This section is devoted to an analysis of possible

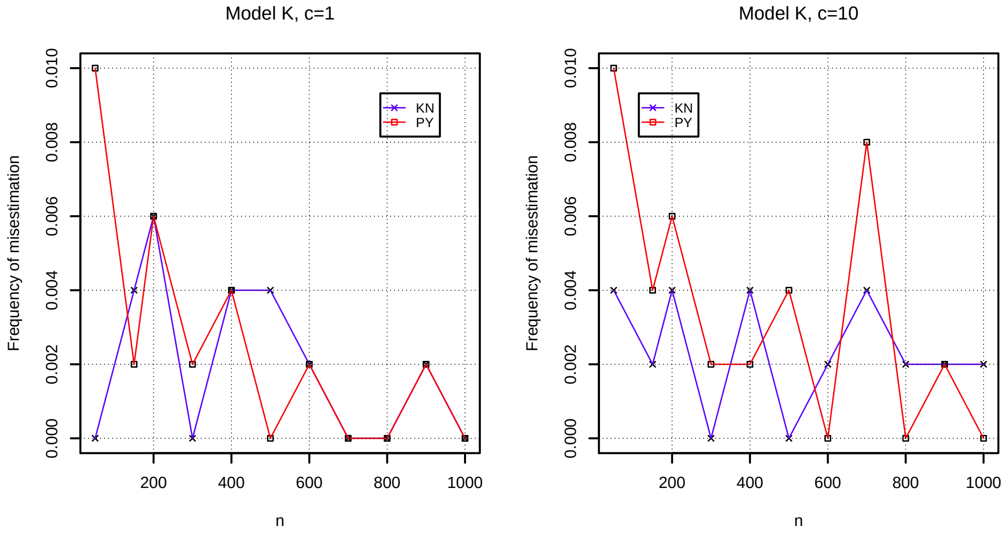

Figure 5.7: Misestimation rates as a function of n in the case of no factor for $c = 1$ (left) and $c = 10$ (right).

relationship between the constant C and the implied false alarm rate. Following Kritchman & Nadler (2009), the false alarm rate γ of such an algorithm can be viewed as the type I error of the following test

$$\mathcal{H}_0\colon\ m = 0 \quad vs. \quad \mathcal{H}_1\colon\ m > 0,$$

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

$$\mathcal{H}_0^{(k)}\colon\ m \leq k-1 \quad vs. \quad \mathcal{H}_1^{(k)}\colon\ m \geq k.$$

In Kritchman & Nadler (2009), the authors argue that their threshold is determined such that

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

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 γ.

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 5.3 (see Table 5.1). Table 5.2 gives the results with 500 independent replications.

Table 5.2: False alarm rates in case of $C = 5$, $c = 1$ (Figure 2R) and $C = 11$, $c = 10$ (Figure 2L).

(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

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 $\mathbb{P}(\hat{m}_n > m | m = q)$, for $q > 1$ will be close to $\mathbb{P}(\hat{m}_n > 0 | m = 0)$. To illustrate this phenomenon, we use the settings of models B ($m = 2$) and J ($m = 3$) and we evaluate the overestimation rate using 500 independent replications (note that the corresponding false alarm rates are those in Table 5.2). The results are displayed in Table 5.3.

Table 5.3: Empirical overestimation rates from model B ($\alpha = (10, 5)$, $c = 10$, $C = 11$) and model J ($\alpha = (2.5, 1.5, 1.5)$, $c = 1$, $C = 5$).

(p,n)	(150,150)	(300,300)	(500,500)	(700,700)
Model B	0.028	0.024	0.028	0.018

(p,n)	(1500,150)	(3000,300)	(5000,500)	(7000,700)
Model J	0.012	0.026	0.032	0.027

We observe that these overestimation rates are lower than the false alarm rates given in Table 5.2: this confirms that no obvious relationship exists between the false alarm rate γ and the overestimation rates for our algorithm.

Furthermore, we can easily see that when C increases, 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, since our estimator is constructed to minimize the overall misestimation rate.

5.6.2 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 corresponds to the case of no factor). Indeed, our algorithm stops 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 $\tilde{\lambda}_{n,1} - \tilde{\lambda}_{n,2}$ 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 $\mathbb{P}(\tilde{\lambda}_{n,1} - \tilde{\lambda}_{n,2} \leq s) = 0.98$: 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 results for various (p, n), with $c = 1$ and $c = 10$ are displayed in Table 5.4.

Table 5.4: Approximation of the threshold s such that $\mathbb{P}(\tilde{\lambda}_{n,1} - \tilde{\lambda}_{n,2} \leq s) = 0.98$.

(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 simulation experiments ($C = 5$ 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 5.8 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 dashed lines are the previous results with C manually chosen.

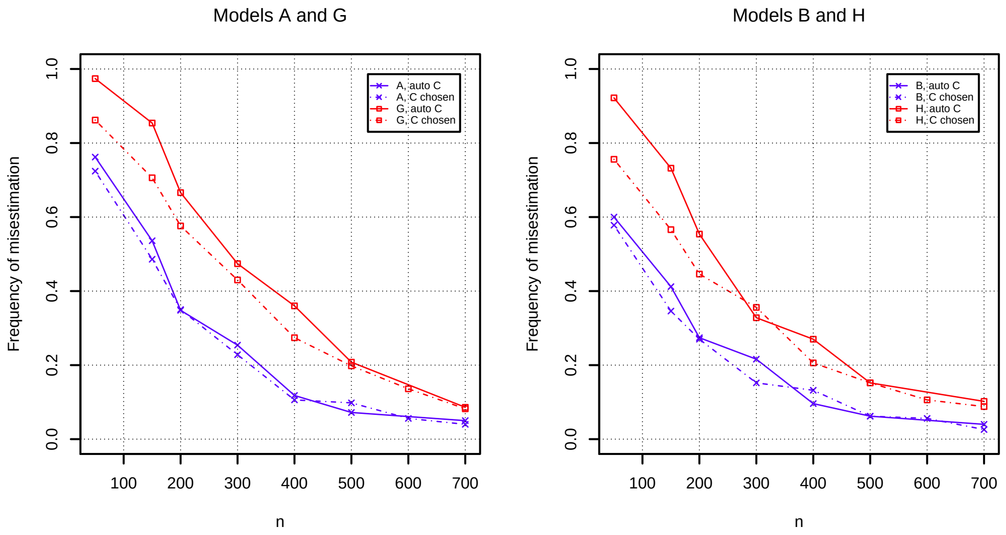

Figure 5.8: Misestimation rates as a function of n for models A, G (left) and models B, H (right).

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 5.9 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 where $c = 10$, using the automatic value $\tilde{C}$ 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,

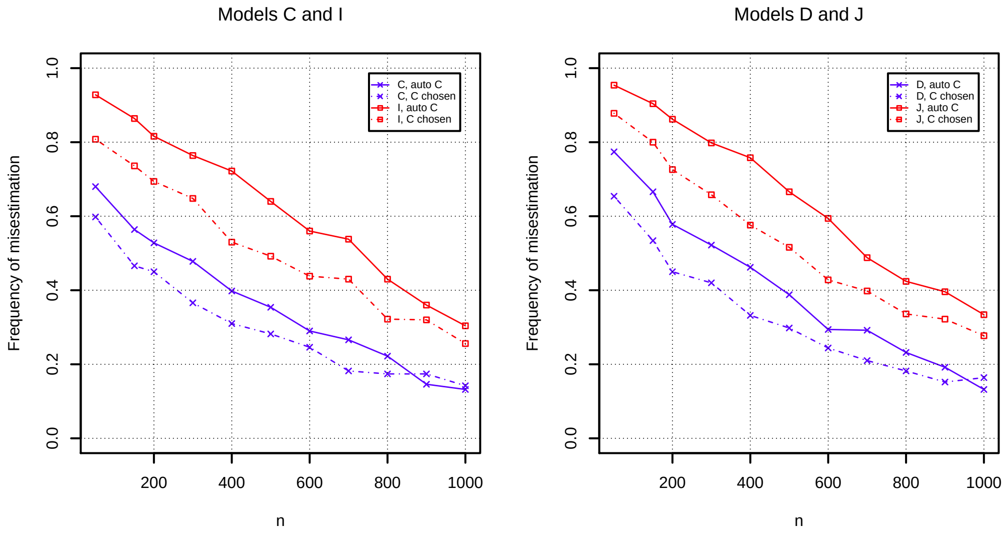

Figure 5.9: Misestimation rates as a function of n for models C, I (left) and models D, J (right).

theses factors are more separated so that the influence of C will be less important.

5.7 Concluding remarks

In this chapter 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 algorithms 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 (2008, 2009) (KN). 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 5.2 and 5.4) 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 depends on a tuning parameter C. Most of the experiments reported here are obtained with a finely-turned value of C and its 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 5.6, 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 21, is a result of Bai & Yao (2008) which gives a CLT for the n_k-packed eigenvalues

$$\sqrt{n}[\lambda_{n,j} - \phi(\alpha_k^*)],\ j \in J_k$$

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

Proposition 21. *Assume that the entries* x_{ij} *of* x *satisfy* $\mathbb{E}(|\mathsf{x}_{ij}|^4) < \infty$, $\alpha_j^* > 1 + \sqrt{c}$ *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* α_k^* *and* c.

The second, Proposition 22, is issued from the Proposition 5.8 of Benaych-Georges et al. (2011):

Proposition 22. *Assume that the entries* x_{ij} *of* x *have a symmetric law and a sub-exponential decay, that means there exists positive constants C, C' such that, for all* $t \geq C'$, $\mathbb{P}(|\mathsf{x}_{ij}| \geq t^C) \leq e^{-t}$. *Then, for all* $1 \leq i \leq L$ *with a prefixed range* L,

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

We also need the following lemma:

Lemma 2. *Let* $(\mathsf{X}_n)_{n\geq 0}$ *be a sequence of positive random variables which weakly converges to a probability distribution with a continuous cumulative distribution function. Then for all real sequence* $(u_n)_{n\geq 0}$ *which converges to 0,*

$$\mathbb{P}(\mathsf{X}_n \leq u_n) \to 0.$$

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

Proof. of Theorem 2. The proof is essentially the same as Theorem 3.1 in Passemier & Yao

(2012b) (Theorem 1, chapter 4), except when the factors are equal. We have

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

Therefore

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

Case of $j = m + 1$. In this case, $\delta_{n,m+1} = \lambda_{n,m+1} - \lambda_{n,m+2}$ (noise eigenvalues). As $d_n \to 0$ such that, $n^{2/3} d_n \to \infty$, and by using Proposition 22 in the same manner as in the proof of Theorem 3.1 in Passemier & Yao (2012b), we have

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

Case of $1 \leq j \leq m$. These indexes correspond to the factor eigenvalues.

- Let $I_1 = \{1 \leq l \leq m | \mathrm{card}(J_l) = 1\}$ (simple factor) and $I_2 = \{l - 1 | l \in I_1 \text{ and } l - 1 > 1\}$. 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 21 and the proof of Theorem 3.1 in Passemier & Yao (2012b) to show that

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

- Let $I_3 = \{1 \leq l \leq m - 1 | l \notin (I_1 \cup I_2)\}$. For all $j \in I_3$, it exists $k \in \{1, \dots, K\}$ such that $j \in J_k$.
  - If $j + 1 \in J_k$ then, by Proposition 21, $\mathsf{X}_n = \sqrt{n}\delta_{n,j}$ converges weakly to a limit which has a density function on $\mathbb{R}^+$. So by using Lemma 2 and that $d_n = o(n^{-1/2})$, we have
  $$\mathbb{P}\left(\delta_{n,j} < d_n\right) = \mathbb{P}\left(\sqrt{n}\delta_{n,j} < \sqrt{n} d_n\right) \to 0;$$
  - Otherwise, $j + 1 \notin J_k$, so $\alpha_j \neq \alpha_{j+1}$. Consequently, as previously, $\delta_{n,j}$ corresponds to a consecutive difference of $\lambda_{n,j}$ issued from two different factors, so we can still use Proposition 21 and the proof of Theorem 3.1 in Passemier & Yao (2012b) to

show that

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

- The case of $j = m$ is considered as in Passemier & Yao (2012b).

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

$$\mathbb{P}(\hat{m}_n = m) \underset{n\to\infty}{\longrightarrow} 1.$$

□

Chapter 6

Corrections of some likelihood statistics in a high-dimensional strict factor model

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 chapter, we consider the strict factor model with homoscedastic variance. First, we give the bias of the estimator of the noise variance. Then we present a corrected likelihood ratio test of the hypothesis that the factor model fits. Finally, we define a test of equality of the norm of two factor loadings vectors.

Keywords: Factor model, covariance matrix, random matrix theory, high-dimensional statistics, hypothesis testing, maximum-likelihood estimation, likelihood ratio test.

AMS subject classification: 62F03, 62F12, 60B20.

This chapter is a preprint
which have the title:
"Corrections of some likelihood statistics in a high-dimensional strict factor model."

It has been written in collaboration with Jian-Feng Yao.

6.1 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 Ross (1976) and its extension in Chamberlain & Rothschild (1983) heavily rely on factor analysis model. Similar models can be found in physics of mixture, see Kritchman & Nadler (2008); Naes et al. (2002) or population genetics. In wireless communications, a signal emitted by a source is modulated and received by an array of antennas which will permit the reconstruction of the original signal, using a factor model (Bianchi et al. (2011); Hachem et al. (2012); Vallet et al. (2012)). More recently, spiked population models have been introduced in Johnstone (2001) that encompass factor models.

A statistical theory for the maximum likelihood estimation is well-known since Lawley (1940), see also Lawley & Maxwell (1971). Furthermore, the asymptotic normality of the maximum likelihood estimators is established in Anderson & Amemiya (1988). Amemiya & Anderson (1990) also gives a likelihood ratio test for model fit which has an asymptotic χ^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, Kritchman & Nadler (2008) observed that the maximum likelihood estimator of the homoscedastic variance has a negative bias, and proposed an empirical correction. In Section 6.4, we give the bias and propose an unbiased estimator. Section 6.5 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. Next we define a test of the equality of the norm of two consecutive vectors of factor scores, or equivalently of two consecutive spikes.

The remaining sections of the chapter are organized as follows. In Section 6.2, we introduce the definition of the strict factor model and the related maximum likelihood theory. In Section 6.3, we recall some results from random matrix theory which will be useful in the following. Throughout the chapter, simulation experiments are conducted to access the quality of the proposed estimation.

6.2 Strict factor model

6.2.1 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 $(\mathsf{x}_i)_{1\leq i\leq n}$ are of the form

$$\begin{aligned} \mathsf{x}_i &= \sum_{k=1}^{m} \mathsf{f}_{ki}\Lambda_k + \mathsf{e}_i + \mu & (6.1)\\ &= \Lambda \mathsf{f}_i + \mathsf{e}_i + \mu, & (6.2) \end{aligned}$$

where

- $\mu \in \mathbb{R}^p$ represents the general mean;
- $\mathsf{f}_i = (\mathsf{f}_{1i}, \dots, \mathsf{f}_{mi})'$ are m random factors ($m < p$);
- $\Lambda = (\Lambda_1, \dots \Lambda_m)$ 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 $\Psi = \mathbb{E}(\mathsf{e}_i \mathsf{e}_i')$.

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

- $\mathbb{E}(\mathsf{f}_i) = 0$ and $\mathbb{E}(\mathsf{f}_i \mathsf{f}_i') = \mathsf{I}_p$;
- $\Psi = \mathrm{cov}(\mathsf{e}_i)$ is diagonal;
- $\Gamma = \Lambda' \Psi^{-1} \Lambda$ is diagonal.

Consequently, *the population covariance matrix* $\Sigma = \mathrm{cov}(\mathsf{x}_i)$ is

$$\Sigma = \Lambda\Lambda' + \Psi.$$

In a strict factor model with homoscedastic variance, we assume in addition

$$\Psi = \sigma^2 \mathsf{I}_p,$$

where $\sigma^2 \in \mathbb{R}$ is the common variance of the noise e_i. In this case, $\Sigma = \Lambda\Lambda' + \sigma^2 \mathsf{I}_p$ and has the spectral decomposition

$$\mathsf{W}' \Sigma \mathsf{W} = \sigma^2 \mathsf{I}_p + \mathrm{diag}(\alpha_1, \dots, \alpha_m, 0, \dots, 0)$$

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

$$\mathsf{S}_n = \frac{1}{n} \sum_{i=1}^{n} (\mathsf{x}_i - \bar{\mathsf{x}})(\mathsf{x}_i - \bar{\mathsf{x}})'.$$

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

6.2.2 Maximum likelihood estimators

If the f_i and e_i are Gaussian, a likelihood-based theory has been developed by Lawley (1940). The maximum likelihood estimator of μ is $\bar{\mathsf{x}}$ and those of Λ and Ψ are obtained by solving the following implicit equations

$$\Lambda(\Gamma + \mathsf{I}_m) = \mathsf{S}_n \Psi^{-1} \Lambda, \tag{6.3}$$

$$\mathrm{diag}(\Lambda\Lambda' + \Psi) = \mathrm{diag}(\mathsf{S}_n), \text{ with } \Gamma \text{ diagonal.} \tag{6.4}$$

These equations can be solved using EM-type algorithms, see Zhao et al. (2008) for a review. The asymptotic normality of the maximum likelihood estimators $\hat{\Lambda}$ (resp. $\hat{\Psi}$) of Λ (resp. Ψ) is established in Anderson & Amemiya (1988) (actually under a more general setting than assuming normal distributions):

Proposition 23. *Let $\Theta = (\theta_{ij})_{1\le i,j\le p} = \Psi - \Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'$. If $(\theta^2_{ij})_{1\le i,j\le p}$ is nonsingular, if Λ and Ψ are identified by the condition that $\Lambda'\Psi\Lambda$ is diagonal and the diagonal elements are different and ordered, if $\mathsf{S}_n \to \Lambda\Lambda' + \Psi$ in probability and if $\sqrt{n}(\mathsf{S}_n - \Sigma)$ has a limiting distribution, then $\sqrt{n}(\hat\Lambda - \Lambda)$ and $\sqrt{n}(\hat\Psi - \Psi)$ have a limiting distribution. The covariance of $\sqrt{n}(\hat\Psi_{ii} - \Psi_{ii})$ and $\sqrt{n}(\hat\Psi_{jj} - \Psi_{jj})$ in the limiting distribution is $2\Psi^2_{ii}\Psi^2_{jj}\xi^{ij}$ $(1 \le i, j \le p)$, where $(\xi^{ij}) = (\theta^2_{ij})^{-1}$.*

In the strict factor model case, the estimation of $\Psi = \sigma^2\mathsf{I}_p$ is simplified to that of σ^2. The equations (6.3) and (6.4) defining the maximum likelihood estimates (m.l.e.) become

$$\Lambda(\Gamma + \mathsf{I}_m) = \mathsf{S}_n\left(\frac{1}{\sigma^2\mathsf{I}_p}\right)\Lambda, \tag{6.5}$$

$$p\sigma^2 = \mathrm{tr}(\mathsf{S}_n - \Lambda\Lambda'), \text{ with } \Gamma = \Lambda'\left(\frac{1}{\sigma^2\mathsf{I}_p}\right)\Lambda \text{ diagonal.} \tag{6.6}$$

In Anderson & Rubin (1956), the authors give the explicit solutions of (6.5) and (6.6):

$$\widehat{\sigma}^2 = \frac{1}{p-m}\sum_{i=m+1}^{p}\lambda_i, \tag{6.7}$$

$$\widehat{\Lambda}_k = \left(\lambda_{n,k} - \widehat{\sigma}^2\right)^{\frac{1}{2}} v_{n,k},\ 1 \le k \le m, \tag{6.8}$$

where $v_{n,k}$ is the normalized eigenvector of $\mathsf{S_n}$ corresponding to $\lambda_{n,k}$, for $1 \le k \le p$.

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.

Notice that using Proposition 23, one can calculate (see appendix for details) the asymptotic variance of the m.l.e. $\widehat{\sigma}^2$, which is

$$\sigma^2_{\mathrm{MLE}} = \frac{2\sigma^4}{p-m}. \tag{6.9}$$

6.3 Results from random matrix theory

6.3.1 Results about spiked population model

Let us rewrite the spectral representation of Σ as a *spiked population model*:

$$\begin{aligned}\mathrm{spec}(\Sigma) &= (\underbrace{\alpha_1,\ldots,\alpha_1}_{n_1},\ldots,\underbrace{\alpha_K,\ldots,\alpha_K}_{n_K},\underbrace{0,\ldots,0}_{p-m}) + \sigma^2(\underbrace{1,\ldots,1}_{p}) && (6.10)\\ &= \sigma^2(\underbrace{\alpha^*_1,\ldots,\alpha^*_1}_{n_1},\ldots,\underbrace{\alpha^*_K,\ldots,\alpha^*_K}_{n_K},\underbrace{1,\cdots,1}_{p-m}), && (6.11)\end{aligned}$$

with $n_1 + \cdots + n_K = m$ and

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

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

Moreover, we assumed that $\alpha_1^* \geq \cdots \geq \alpha_m^* > 1 + \sqrt{c}$, i.e all the eigenvalues α_i are greater than $\sigma^2\sqrt{c}$. For $\alpha \neq 1$, we define the function

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

In Baik & Silverstein (2006) it is proved that, under a moment condition on x_i, for each $k \in \{1, \ldots, m\}$ and almost surely,

$$\begin{aligned} \lambda_{n,k} &\longrightarrow \sigma^2\phi(\alpha_k^*) && (6.12) \\ &= \alpha_k + \sigma^2 + \sigma^2 c\left(1 + \frac{\sigma^2}{\alpha_k}\right). && (6.13) \end{aligned}$$

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

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

Furthermore, Bai & Yao (2008) give the joint distribution of

$$\{\sqrt{n}(\lambda_{n,j} - \phi(\alpha_k^*)),\, j \in J_k\} \tag{6.14}$$

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

6.3.2 Empirical spectral distribution and Marčenko-Pastur distributions

Let H be a probability measure on $\mathbb{R}^+$ and $c > 0$ a constant. We define the map

$$g(s) = g_{c,H}(s) = \frac{1}{s} + c\int \frac{t}{1 + ts}\,\mathrm{d}H(t) \tag{6.15}$$

in the set $\mathbb{C}^+ = \{z \in \mathbb{C} : \Im z > 0\}$. The map g is a one-to-one map from $\mathbb{C}^+$ onto itself (see Bai & Silverstein (2010), chapter 6), and the inverse map $m = g^{-1}$ satisfies all the requirements of the Stieltjes transform of a probability measure on $[0, \infty)$. We call this measure $\underline{F}_{c,H}$. Next, a companion measure $F_{c,H}$ is introduced by the equation $cF_{c,H} = (c-1)\,\delta_0 + \underline{F}_{c,H}$ (note that in this equation, measures can be signed). The measure $F_{c,H}$ is referred as the generalized Marčenko-Pastur distribution with indexes (c, H).

Let $F_n = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_{n,i}}$ be the empirical spectral distribution (ESD) of S_n. It is well known that when $\Sigma = \sigma^2\mathsf{I}_p$, F_n converges to the Marčenko-Pastur distribution of indexes (c, δ_{σ^2}), denoted as F_{c,σ^2}, with the following density function

$$p_{c,\sigma^2}(x) = \begin{cases} \frac{1}{2\pi x c\sigma^2}\sqrt{(b(c) - x)(x - a(c))} & \text{if } a(c) = \sigma^2(1 - \sqrt{c})^2 \leq x \leq b(c) = \sigma^2(1 + \sqrt{c})^2 \\ 0 & \text{otherwise.} \end{cases}$$

This limit still holds for the spiked population model (6.11).

Let $H_n = F^{\Sigma}$ be the ESD of Σ. We have

$$H_n = \frac{p-m}{p}\delta_{\sigma^2} + \frac{1}{p}\sum_{i=1}^{m}\delta_{\alpha_i+\sigma^2}$$

and $H_n \to \delta_{\sigma^2}$.

6.3.3 CLT for LSS of a high-dimensional covariance matrix

We consider the following empirical process

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

where $\mathcal{A}$ is the set of analytic functions $f : \mathcal{U} \to \mathbb{C}$, with $\mathcal{U}$ an open set of $\mathbb{C}$ such that $[\mathbb{1}_{(0,1)}(c)a(c), b(c)] \subset \mathcal{U}$. As $H_n = F^{\Sigma} \to \delta_{\sigma^2}$ and following Bai et al. (2009), we have the following proposition which is a specialization of Theorem 9.10 of Bai & Silverstein (2010) (which covers more general matrices).

Proposition 24. *Assume that $f_1, \dots, f_k \in \mathcal{A}$ and the entries x_{ij} of the vectors $(\mathsf{x}_i)_{1\le i\le n}$ are i.i.d. real random variables with mean 0, $\mathbb{E}(|x_{ij}|^4) = 3\sigma^4$ and cov$(\mathsf{x}_i) = \Sigma = \Lambda\Lambda' + \sigma^2\mathsf{I}_p$. Then the random vector $(G_n(f_1), \dots, G_n(f_k))$ converges to a k-dimensional Gaussian vector with mean vector*

$$m(f_j) = \frac{f_j(a(c)) + f_j(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)} \frac{f_j(x)}{\sqrt{4c\sigma^4 - (x - \sigma^2 - c\sigma^2)^2}}\,\mathrm{d}x,\ j = 1, \dots, k,$$

and covariance function

$$v(f_j, f_l) \quad = \quad -\frac{1}{2\pi^2}\oint\oint \frac{f_j(z_1)f_l(z_2)}{(\underline{m}(z_1) - \underline{m}(z_2))^2}\,\mathrm{d}\underline{m}(z_1)\mathrm{d}\underline{m}(z_2),\ j, l = 1, \dots, k, \quad (6.16)$$

where $\underline{m}(z)$ is the Stieltjes transform of $\underline{F}_{c,\delta_{\sigma^2}} = (1-c)\delta_0 + cF_{c,\delta_{\sigma^2}}$. The contours are non overlapping and both contain the support of $F_{c,\delta_{\sigma^2}}$.

6.4 Estimation of the homoscedastic variance

As observed in Kritchman & Nadler (2008, 2009), in high-dimensional setting, the m.l.e. $\widehat{\sigma}^2$ in (6.7) has a negative bias. In this section, we will give this bias and show its asymptotic normality.

Theorem 3. *We assume the same conditions of Proposition 24. Then, we have*

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

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

Proof. We have

$$\begin{aligned}(p-m)\widehat{\sigma}^2 &= \sum_{i=m+1}^{p} \lambda_{n,i} \\ &= \sum_{i=1}^{p} \lambda_{n,i} - \sum_{i=1}^{m} \lambda_{n,i}.\end{aligned}$$

By (6.13), we have

$$\sum_{i=1}^{m} \lambda_{n,i} \quad \longrightarrow \quad \sum_{i=1}^{m} \left(\alpha_i + \frac{c\sigma^4}{\alpha_i} \right) + \sigma^2 m(1+c) \text{ a.s.} \tag{6.17}$$

For the first term, we write

$$\begin{aligned}\sum_{i=1}^{p} \lambda_i &= p \int x dF_n(x) \\ &= p \int x \, \mathrm{d}(F_n - F_{c_n,H_n})(x) + p \int x \, \mathrm{d}F_{c_n,H_n}(x) \\ &= G_n(x) + p \int x \, \mathrm{d}F_{c_n,H_n}(x).\end{aligned}$$

By Proposition 24, the first term is asymptotically normal

$$G_n(x) = \sum_{i=1}^{p} \lambda_{n,i} - p \int x \, \mathrm{d}F_{c_n,H_n}(x) \xrightarrow{\mathcal{L}} \mathcal{N}(m(x), v(x)),$$

with the mean

$$m(x) \quad = \quad 0 \tag{6.18}$$

and the variance

$$v(x) \quad = \quad 2c\sigma^4 \tag{6.19}$$

are calculated in the appendix. Furthermore, by lemma 1 of Bai et al. (2010) (which can be proved using (6.15)), we have

$$\begin{aligned}\int x \, \mathrm{d}F_{c_n,H_n}(x) &= \int t \, \mathrm{d}H_n(t) \\ &= \frac{p-m}{p}\sigma^2 + \frac{1}{p}\sum_{i=1}^{m} \left(\alpha_i + \sigma^2 \right) \\ &= \sigma^2 + \frac{1}{p}\sum_{i=1}^{m} \alpha_i.\end{aligned}$$

So we get

$$\sum_{i=1}^{p} \lambda_{n,i} - p\sigma^2 - \sum_{i=1}^{m} \alpha_i \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0, 2c\sigma^4). \tag{6.20}$$

By (6.17) and (6.20) and using the Slutsky lemma, we obtain

$$(p-m)(\widehat{\sigma}^2 - \sigma^2) + c\sigma^2 \left(m + \sigma^2 \sum_{i=1}^{m} \frac{1}{\alpha_i} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2c\sigma^4).$$

□

6.4.1 Simulation experiments

We consider an i.i.d. Gaussian sample of size n in three different settings:

- Model 1: $\mathrm{spec}(\Sigma) = (25, 16, 9, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 4$, $c = 1$;
- Model 2: $\mathrm{spec}(\Sigma) = (4, 3, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 2$, $c = 0.2$;
- Model 3: $\mathrm{spec}(\Sigma) = (12, 10, 8, 8, 0, \ldots, 0) + \sigma^2(1, \ldots, 1)$, $\sigma^2 = 3$, $c = 1.5$.

Figure 6.1 presents the histograms of 1000 replications of

$$\frac{(p-m)}{\sigma^2\sqrt{2c}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2)$$

for the three models above, with different sample size n and $p = c \times n$, compared to the density of the standard normal probability law. Even when the sample size is moderate like $n = 100$, the distribution is almost normal.

In Table 6.1, we compare the empirical bias of $\widehat{\sigma}^2$ (i.e. the empirical mean of $\sigma^2 - \widehat{\sigma}^2 = \sigma^2 - \frac{1}{p-m}\sum_{i=m+1}^{p} \lambda_{n,i}$) over 1000 replications with the theoretical one $b(\sigma^2)/(p-m)$ in different settings.

Table 6.1: Comparison between the empirical and the theoretical bias in various settings.

Settings			Empirical bias	Theoretical bias	|Difference|
Model 1	$p = 100$	$n = 100$	-0.1556	-0.1589	0.0023
	$p = 400$	$n = 400$	-0.0379	-0.0388	0.0009
	$p = 800$	$n = 800$	-0.0189	-0.0193	0.0004
Model 2	$p = 20$	$n = 100$	-0.0654	-0.0704	0.0050
	$p = 80$	$n = 400$	-0.0150	-0.0162	0.0012
	$p = 200$	$n = 1000$	-0.0064	-0.0063	0.0001
Model 3	$p = 150$	$n = 100$	-0.0801	-0.0795	0.0006
	$p = 600$	$n = 400$	-0.0400	-0.0397	0.0003
	$p = 1500$	$n = 1000$	-0.0157	-0.0159	0.0002

The empirical bias is close to the theoretical one. As expected, the difference between the two bias decreases when p and n increase. There is no particular difference between the three models.

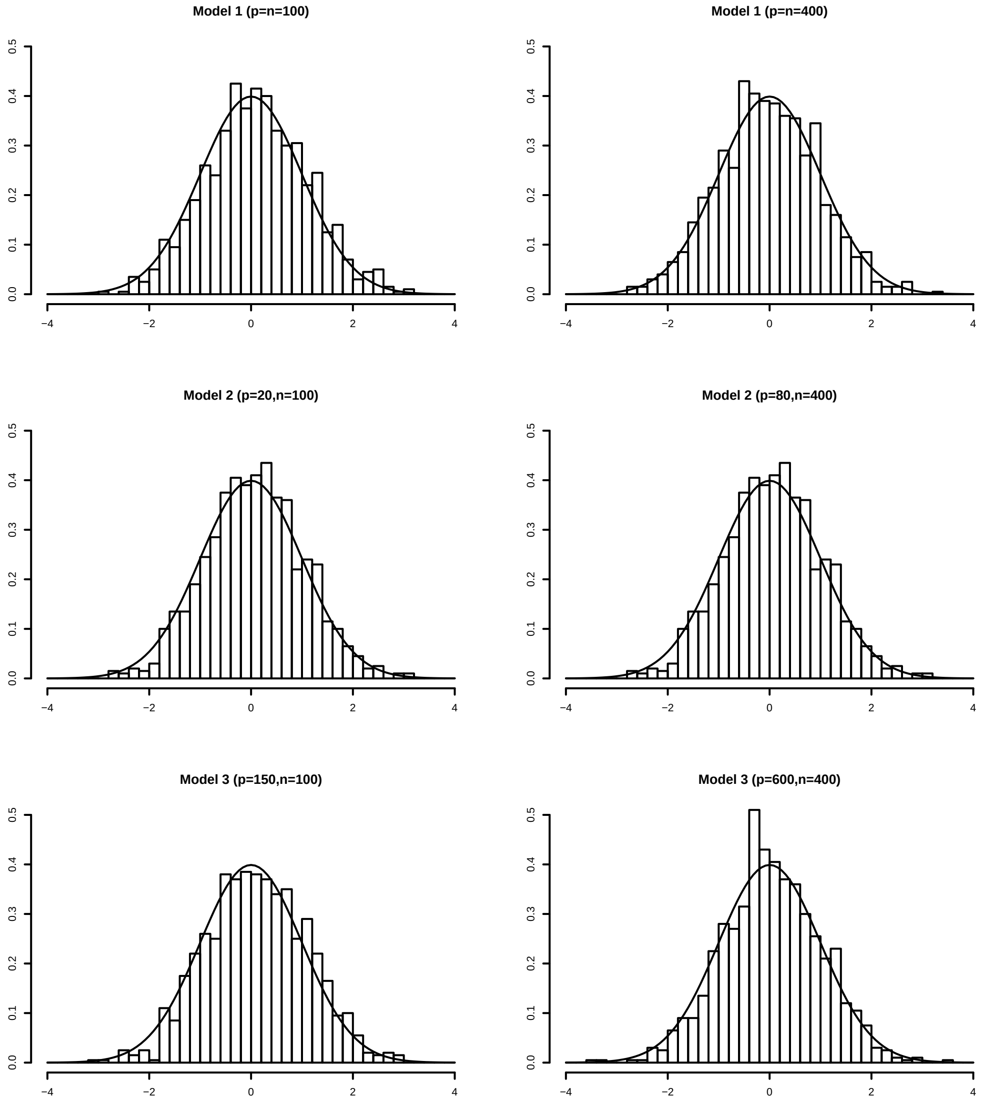

Figure 6.1: Histogram of $\frac{(p-m)}{\sigma^2\sqrt{2c}}(\widehat{\sigma}^2 - \sigma^2) + b(\sigma^2)$ compared with the density of a standard Gaussian law.

6.4.2 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 α_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 Passemier & Yao (2012b,a) and Kritchman & Nadler (2008). For for the values of the spikes α_i, it is easy to see that it can be done by inverting the function ϕ in (6.13) at the corresponding eigenvalues λ_j. Moreover, by applying the delta-method to (6.14), we can obtain the asymptotic distribution of this estimator.

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

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

To assess the quality of this bias-corrected estimator $\hat{\sigma}_*^2$, we conduct some simulation experiments using the previous settings: in Table 6.2, we give the empirical mean of $\hat{\sigma}_*^2$ over 1000 replications compared with the empirical mean of $\widehat{\sigma}^2$, as well as the mean absolute deviations.

Table 6.2: Comparison between $\hat{\sigma}^2$ and $\hat{\sigma}_*^2$ in various settings.

	Settings						
Model No.	p	n	σ^2	$\widehat{\sigma}^2$	$\lvert\sigma^2 - \widehat{\sigma}^2\rvert$	$\hat{\sigma}_*^2$	$\lvert\sigma^2 - \hat{\sigma}_*^2\rvert$
1	100	100	4	3.8441	0.1559	3.9966	0.0034
	400	400		3.9617	0.0383	4.0000	$< 10^{-5}$
	800	800		3.9806	0.0194	3.9998	0.0002
2	20	100	2	1.9321	0.0679	1.9993	0.0007
	80	400		1.9846	0.0154	2.0007	0.0007
	200	1000		1.9937	0.0063	2.0000	$< 10^{-5}$
3	150	100	3	2.8413	0.1587	2.9940	0.0060
	600	400		2.9599	0.0401	2.9992	0.0008
	1500	1000		2.9842	0.0158	3.0000	$< 10^{-5}$

The bias-corrected estimator is far much better than $\hat{\sigma}^2$: here mean absolute deviations are reduced by 95% at least. The proposed correction of the bias performs well in the three models.

6.5 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

$$\mathcal{H}_0 : \ \Sigma = \Lambda\Lambda' + \sigma^2 \mathsf{I}_p,$$

where the number of factors m is specified.

Following Anderson & Rubin (1956), the likelihood ratio test (LRT) statistic is

$$T_n = -nL^*,$$

where

$$L^* = \sum_{j=m+1}^{p} \log \frac{\lambda_{n,j}}{\widehat{\sigma}^2},$$

and $\hat{\sigma}^2$ is the variance estimator (6.7). Keeping p fixed while letting $n \to \infty$, then the classical theory states that T_n converges to χ_q^2, where $q = p(p+1)/2 + m(m-1)/2 - pm - 1$, see Anderson & Rubin (1956). 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 6.3).

In a way similar to Section 6.4, we will construct a corrected version of T_n using Proposition 24 and calculus done in Bai et al. (2009) and Zheng (2012). 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 4. *We assume the same conditions of Proposition 24, with $c < 1$, i.e. the entries x_{ij} of the vectors $(\mathsf{x}_i)_{1\le i\le n}$ are i.i.d. real random variables with mean 0, $\mathbb{E}(|x_{ij}|^4) = 3\sigma^4$ and cov$(\mathsf{x}_i) = \Sigma = \Lambda\Lambda' + \sigma^2\mathsf{I}_p$. Then, we have*

$$v(c)^{-\frac{1}{2}}(L^* - m(c) - ph(c_n, \tilde{H}_n(\sigma^2)) + \eta + (p-m)\log(\beta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where

- $m(c) = \frac{\log(1-c)}{2}$;
- $h(c_n, \tilde{H}_n(\sigma^2)) = \int \log(x)\,\mathrm{d}F_{c_n,\tilde{H}_n(\sigma^2)}(x)$, *with* $\tilde{H}_n(\sigma^2) = \frac{p-m}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^m \delta_{\alpha_i/\sigma^2+1}$;
- $\eta = \sum_{i=1}^m \log((\alpha_i\sigma^{-2}+1)(1+c\sigma^2\alpha_i^{-1}))$;
- $\beta = 1 - \frac{c}{p-m}(m + \sigma^2\sum_{i=1}^m \alpha_i^{-1})$;
- $v(c) = -2\log(1-c) + \frac{2c}{\beta}\left(\frac{1}{\beta} - 2\right)$.

Proof. We have

$$\begin{aligned} L^* &= \sum_{i=m+1}^{p} \log\frac{\lambda_{n,i}}{\hat{\sigma}^2} \\ &= \sum_{i=m+1}^{p} \log\frac{\lambda_{n,i}}{\sigma^2} - \sum_{i=m+1}^{p} \log\frac{\hat{\sigma}^2}{\sigma^2} \\ &= \sum_{i=m+1}^{p} \log\frac{\lambda_{n,i}}{\sigma^2} - (p-m)\log\left(\frac{1}{p-m}\sum_{i=m+1}^{p}\frac{\lambda_{n,i}}{\sigma^2}\right) \\ &= L_1 - (p-m)\log\left(\frac{L_2}{p-m}\right), \end{aligned}$$

where we have defined a two-dimensional vector $(L_1, L_2) = (\sum_{i=m+1}^p \log\frac{\lambda_{n,i}}{\sigma^2}, \sum_{i=m+1}^p \frac{\lambda_{n,i}}{\sigma^2})$.

CLT when $\sigma^2 = 1$. To start with, we consider the case $\sigma^2 = 1$. We have

$$\begin{aligned} L_1 &= \sum_{i=m+1}^{p} \log\lambda_{n,i} \\ &= \sum_{i=1}^{p} \log\lambda_{n,i} - \sum_{i=1}^{m} \log\lambda_{n,i} \end{aligned}$$

$$\begin{aligned} &= p\int \log(x)\,\mathrm{d}F_n(x) - \sum_{i=1}^{m}\log\lambda_{n,i} \\ &= p\int \log(x)\,\mathrm{d}(F_n - F_{c_n,H_n})(x) + p\int \log(x)\,\mathrm{d}F_{c_n,H_n}(x) - \sum_{i=1}^{m}\log\lambda_{n,i}. \end{aligned}$$

Similarly, we have

$$L_2 = p\int x\,\mathrm{d}(F_n - F_{c_n,H_n})(x) + p\int x\,\mathrm{d}F_{c_n,H_n}(x) - \sum_{i=1}^{m}\lambda_{n,i}.$$

By Proposition 24, we find that

$$p\begin{pmatrix} \int \log(x)\,\mathrm{d}(F_n - F_{c_n,H_n})(x) \\ \int x\,\mathrm{d}(F_n - F_{c_n,H_n})(x) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} m_1(c) \\ m_2(c) \end{pmatrix}, \begin{pmatrix} v_1(c) & v_{1,2}(c) \\ v_{1,2}(c) & v_2(c) \end{pmatrix}\right) \tag{6.21}$$

with $m_2(c) = 0$ and $v_2(c) = 2c$ and

$$m_1(c) = \frac{\log(1-c)}{2}, \tag{6.22}$$

$$v_1(c) = -2\log(1-c), \tag{6.23}$$

$$v_{1,2}(c) = 2c. \tag{6.24}$$

Formulas m_2 and v_2 have been established in the proof of Theorem 3 and the remaining ones are proved in Section 6.8.

In Theorem 3, with $\sigma^2 = 1$, we found that

$$\int x\,\mathrm{d}F_{c_n,H_n}(x) = 1 + \frac{1}{p}\sum_{i=1}^{m}\alpha_i,$$

and

$$\sum_{i=1}^{m}\lambda_{n,i} \xrightarrow{\text{a.s.}} \sum_{i=1}^{m}\left(\alpha_i + \frac{c}{\alpha_i}\right) + m(1+c).$$

For the last term of L_1, by (6.13), we have

$$\log\lambda_{n,i} \longrightarrow \log(\phi(\alpha_i+1)) = \log\left((\alpha_i+1)(1+c\alpha_i^{-1})\right) \text{ a.s.}$$

Summarizing, we have obtained that

$$L_1 - m_1(c) - ph(c_n, H_n) + \eta(c,\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}(0, v_1(c)),$$

where $h(c_n, H_n) = \int \log(x)\,\mathrm{d}F_{c_n,H_n}(x)$ and $\eta(c,\alpha) = \sum_{i=1}^{m}\log((\alpha_i+1)(1+c\sigma^2\alpha_i^{-1}))$.

Similarly, we have

$$L_2 - (p-m) + \rho(c,\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, v_2(c)\right),$$

where $\rho(c,\alpha) = c(m + \sum_{i=1}^m \alpha_i^{-1})$.

Using (6.21) and the Slutsky lemma, we finally get the following CLT for $(L_1, L_2)'$

$$\begin{pmatrix} L_1 \\ L_2 \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} m_1(c) + ph(c_n, H_n) - \eta(c,\alpha) \\ p - m - \rho(c,\alpha) \end{pmatrix}, \begin{pmatrix} v_1(c) & v_{1,2}(c) \\ v_{1,2}(c_n) & v_2(c_n) \end{pmatrix}\right),$$

with $h(c_n, H_n) = \int \log(x)\, \mathrm{d}F_{c_n,H_n}(x)$, $\eta(c,\alpha) = \sum_{i=1}^m \log((\alpha_i+1)(1+c\sigma^2\alpha_i^{-1}))$ and $\rho(c,\alpha) = c(m + \sum_{i=1}^m \alpha_i^{-1})$.

CLT with general σ^2. When $\sigma^2 = 1$,

$$\operatorname{spec}(\Sigma) = (\alpha_1 + 1, \ldots, \alpha_m + 1, 1, \ldots, 1),$$

whereas in the general case

$$\begin{aligned} \operatorname{spec}(\Sigma) &= (\alpha_1 + \sigma^2, \ldots, \alpha_m + \sigma^2, \sigma^2, \ldots, \sigma^2) \\ &= \sigma^2 \left(\frac{\alpha_1}{\sigma^2} + 1, \ldots, \frac{\alpha_m}{\sigma^2} + 1, \ldots, 1\right). \end{aligned}$$

Thus, if we consider λ_i/σ^2, we will find the same CLT by replacing the $(\alpha_i)_{1\le i\le m}$ by α_i/σ^2. Furthermore, we divide L_2 by $p-m$ to find

$$\begin{pmatrix} L_1 \\ \frac{L_2}{p-m} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} m_1(c) + ph(c_n, \tilde{H}_n(\sigma^2)) - \eta(c,\alpha/\sigma^2) \\ 1 - \frac{\rho(c,\alpha/\sigma^2)}{p-m} \end{pmatrix}, \begin{pmatrix} \frac{2c}{(p-m)^2} & \frac{2c}{p-m} \\ \frac{2c}{p-m} & -2\log(1-c) \end{pmatrix}\right),$$

with $\eta(c,\alpha/\sigma^2) = \sum_{i=1}^m \log((\alpha_i\sigma^{-2} + 1)(1 + c\sigma^2\alpha_i^{-1}))$, $\rho(c,\alpha/\sigma^2) = c(m + \sigma^2 \sum_{i=1}^m \alpha_i^{-1})$ and $\tilde{H}_n(\sigma^2) = \frac{p-m}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^m \delta_{\alpha_i/\sigma^2+1}$.

Asymptotic law of L^*. We have $L^* = g(L_1, L_2/(p-m))$, with $g(x,y) = x - (p-m)\log(y)$. We will apply the multivariate delta-method on (6.25) with the function g. We have $\nabla g(x,y) = \left(1, -\frac{p-m}{y}\right)$ and

$$L^* \xrightarrow{\mathcal{L}} \mathcal{N}(\beta_1 - (p-m)\log(\beta_2), \nabla g(\beta_1,\beta_2)\ \operatorname{cov}(L_1, L_2/(p-m)) \nabla g(\beta_1,\beta_2)'),$$

with $\beta_1 = m_1(c) + ph(c_n, \tilde{H}_n(\sigma^2)) - \eta(c,\alpha/\sigma^2)$ and $\beta_2 = 1 - \frac{\rho(c,\alpha/\sigma^2)}{p-m}$. After calculation we finally get

$$L^* \xrightarrow{\mathcal{L}} \mathcal{N}\left(m_1(c) + ph(c_n, \tilde{H}_n(\sigma^2)) - \eta(c, \frac{\alpha}{\sigma^2}) - (p-m)\log(\beta_2), -2\log(1-c) + \frac{2c}{\beta_2}\left(\frac{1}{\beta_2} - 2\right)\right).$$

□

To test $\mathcal{H}_0$, we then can use the statistic

$$v(c_n)^{-\frac{1}{2}}(L^* - m(c) - ph(c_n, \tilde{H}_n(\sigma^2)) + \eta + (p-m)\log(\beta)) \quad ,$$

This test is asymptotically normal. In practice, we need an accurate expression for $h(c_n, \tilde{H}_n(\sigma^2)) = \int \log(x)\,\mathrm{d}F_{c_n,\tilde{H}_n(\sigma^2)}(x)$. However, this is missing at the moment and we conjecture the following formula.

Conjecture 1. *For* $\tilde{H}_n(\sigma^2) = \frac{p-m}{p}\delta_1 + \frac{1}{p}\sum_{i=1}^m \delta_{\alpha_i/\sigma^2+1}$,

$$p\int \log(x)dF_{c_n,\tilde{H}_n}(x) \;=\; \sum_{j=1}^m \log\left(\frac{\alpha_j}{\sigma^2}+1\right) + p\int \log(x)dF_{c_n,\delta_1}(x) + o(1). \quad (6.25)$$

The second term can be calculated using the density of the Marčenko-Pastur law (see appendix): we have

$$h(c_n) \;=\; \int \log(x)dF_{c_n,\delta_1}(x) \quad (6.26)$$

$$\;=\; \frac{c_n-1}{c_n}\log(1-c_n) - 1. \quad (6.27)$$

Simulation results given in appendix suggest that the conjectured formula (6.25) is quite accurate.

6.5.1 Simulation experiments

For the simulation experiments, we will use the following statistic, obtained from (6.25) by changing $h(c_n, \tilde{H}_n(\sigma^2))$ to its conjectured value:

$$v(c_n)^{-\frac{1}{2}}(L^* - m(c) - ph(c_n) + \vartheta + (p-m)log(\beta)),$$

where

- $\beta = 1 - \frac{c}{p-m}(m + \sigma^2\sum_{i=1}^m \alpha_i^{-1})$ as previously and;
- $\vartheta = \eta - \sum_{j=1}^m \log(\alpha_j\sigma^{-2}+1) = \sum_{i=1}^m \log(1 + c\sigma^2\alpha_i^{-1})$.

The corresponding test will be hereafter referred as the corrected likelihood ratio test (CLRT in short).

We consider again the models 1 and 2 described in Section 6.4.1, and a new one (model 4):

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

Table 6.3 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. For the LRT, we use the correction proposed by Bartlett (1950), that is replacing $T_n = -nL^*$ by $\tilde{T}_n = -(n - (2p+11)/6 - 2m/3)L^*$. The computations are done under 10000 independent replications and the nominal test level is 0.05.

Table 6.3: Comparison of the realized size of the classical likelihood ratio test (LRT) and the corrected likelihood ratio test (CLRT) in various settings.

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

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.

6.6 Testing the equality of two spiked eigenvalues

In this section, the aim is to conduct the following test

$$\mathcal{H}_0\colon\ \alpha_i = \alpha_{i+1} \quad vs. \quad \mathcal{H}_1\colon\ \alpha_i \neq \alpha_{i+1},$$

where $i \leq m-1$ is fixed. From the point of view of the factor model, it is the same to test the equality of the norm of two vectors of factor scores. To this end, we will begin by detailing the result (6.14) of Bai & Yao (2008). Without any loss of generality, we consider here that σ^2 is known and equal to one.

6.6.1 Law of the spacings of two consecutive eigenvalues

Proposition 25. *Assume that the entries x^i of x satisfy $\mathbb{E}(|\mathsf{x}^i|^4) < \infty$, $\alpha_j^* > 1 + \sqrt{c}$ 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*

$$\beta_{n,j} = \sqrt{n}\{\lambda_{n,j} - \phi(\alpha_k^*),\ j \in J_k\}$$

converges weakly to the distribution of the n_k eigenvalues of a Gaussian random matrix G whose covariance depends on α_k and c.

In the real Gaussian case, G is a $n_k \times n_k$ real Gaussian-Wigner matrix (with independent entries) of the form

$$M = s_k \begin{pmatrix} W_{1,1} & \cdots & W_{1,n_k} \\ \vdots & \ddots & \vdots \\ W_{n_k,1} & \cdots & W_{n_k,n_k} \end{pmatrix},$$

where $s_k^2 = \frac{2\alpha_k^{*2}((\alpha_k^*-1)^2-c)}{(\alpha_k^*-1)^2}$.

Since the joint distribution of eigenvalues of a Gaussian-Wigner matrix is known, we get the following (unordered) density for the limiting distribution of the n_k-dimensional vector $\beta_{n,j}$:

$$g(x_1, \cdots, x_{n_k}) = C\text{exp}\left(-\frac{1}{2s_k^2}\sum_{i=1}^{n_k} x_i^2\right)\prod_{j<k}|x_j - x_k| \tag{6.28}$$

where $C^{-1} = s_k^{\frac{n_k(n_k+1)}{2}} 2^{\frac{3n_k}{2}} \prod_{j=1}^{n_k} \Gamma\left(1+\frac{j}{2}\right)$.

6.6.1.1 Case of multiplicity two

When $n_k = 2$, this expression becomes

$$g(x_1, x_2) = \frac{1}{4s_k^3\sqrt{\pi}}|x_1 - x_2|\text{exp}\left(-\frac{1}{2s_k^2}(x_1^2 + x_2^2)\right).$$

From this, one can get the limiting distribution of $\beta_{n,i} - \beta_{n,i+1} = \sqrt{n}(\lambda_{n,i} - \lambda_{n,i+1}) = \sqrt{n}U_2$ when $n_k = 2$ (case of multiplicity two)

$$m_2(x) = \frac{1}{2s_k^2}|x|e^{-\frac{x^2}{4s_k^2}},$$

so

$$\mathbb{P}(\sqrt{n}U_2 \leq \varepsilon) = 1 - e^{-\frac{\varepsilon^2}{4s_k^2}}.$$

6.6.1.2 Case of multiplicity greater than two

When the multiplicity is greater than two, we can not directly compute the law of $\beta_{n,i} - \beta_{n,i+1} = \sqrt{n}(\lambda_{n,i} - \lambda_{n,i+1})$ from (6.28). Nevertheless, we can use the Wigner Surmise (see Mehta (2004)) to approximate it. Wigner considered normalized spacings. Based on arguments from the analysis of a model for nuclear energy levels, he conjectures an expression for the nearest neighbor spacings probability density function, when the size of the matrix goes to infinity. When the spacings are normalized by $\sqrt{2\pi/n_k}$, the expression is

$$m_W(x) = \frac{\pi}{2s_k^2}|x|e^{-\frac{\pi x^2}{4s_k^2}}.$$

It was in fact an approximation and Gaudin (1961) demonstrated that this approximation is good, even for small n_k. Actually, the exact expression is known but complicated to calculate (see Section 7.4 of Mehta (2004)). If we remove the normalization, we find

$$m_{n_k}(x) = \frac{n_k}{4s_k^2}|x|e^{-\frac{n_k x^2}{8s_k^2}},$$

so

$$\mathbb{P}(\sqrt{n}U_{n_k} \leq \varepsilon) = 1 - e^{-\frac{n_k \varepsilon^2}{8s_k^2}}.$$

6.6.2 Definition of the test

We recall the test (6.6):

$$\mathcal{H}_0\colon\ \alpha_i = \alpha_{i+1} \quad vs. \quad \mathcal{H}_1\colon\ \alpha_i \neq \alpha_{i+1}$$

To conduct this test, we will use the statistic $D_{n,i} = \sqrt{n}(\lambda_{n,i} - \lambda_{n,i+1})$: under $\mathcal{H}_0$, $D_{n,i}$ have the density function m_{n_k} for some n_k whereas under $\mathcal{H}_1$, $D_{n,i}$ is equivalent to $\sqrt{n}(\phi(\alpha_i) - \phi(\alpha_{i+1}))$, thus tends to infinity when $n \to \infty$. For $t > 0$, the p-value function of the test is

$$\begin{aligned} pv(t) &= \sup_{\mathcal{H}_0} \mathbb{P}(D_{n,k} > t) \\ &= \sup_{n_k \geq 2} e^{-\frac{n_k t^2}{8s_k^2}} \\ &= e^{-\frac{n_k t^2}{8s_k^2}}|_{n_k=2} \\ &= e^{-\frac{t^2}{4s_k^2}}. \end{aligned}$$

Let d_{obs} be the observed value of $D_{n,i}$. Then, at a significance level γ, $\mathcal{H}_0$ will be rejected if $pv(d_{obs}) < \gamma$, i.e.

$$d_{obs} > \sqrt{-4s_k^2 \log(\gamma)}.$$

6.6.3 Simulation experiments

We consider an i.i.d. Gaussian sample of size n in three different models

- Model 5: $\text{Spec}(\Sigma) = (10, 10, 5, 5, 1, \ldots, 1)$, $c = 0.3$ and $c = 0.6$;
- Model 6: $\text{Spec}(\Sigma) = (10, 10, 10, 5, 5, \ldots, 1)$, $c = 0.3$ and $c = 0.6$;
- Model 7: $\text{Spec}(\Sigma) = (10, 10, 9, 9, 1, \ldots, 1)$, $c = 0.3$ and $c = 0.6$.

We performs two different tests: one where the null $\mathcal{H}_0$ is theoretically verified, and an other one with a null $\tilde{\mathcal{H}}_0$ not true. We ran 1000 independent replications and took a nominal level of 0.1. As the law of the statistic $D_{n,i}$ depends on s_k^2, which itself depends on the unknown α_i, we display the results of the simulation experiments using the real α_i, and using the estimated one obtained by inverting the function ϕ in (6.13).

Furthermore, when $\alpha_i = \alpha_{i+1}$, $s_i^2 = s_{i+1}^2$: this is no longer the case when $\alpha_i \neq \alpha_{i+1}$. As the p-value is an increasing function of s_i^2, we take s_{i+1}^2 in the calculation of the realized size, in order to maximize the power.

Table 6.4: Realized size of the test $\mathcal{H}_0$: $\alpha_1 = \alpha_2$ and $\tilde{\mathcal{H}}_0$: $\alpha_2 = \alpha_3$ for model 5, with $c = 0.3$ and $c = 0.6$.

	α_i^* known		α_i^* estimated	
(p, n)	$\mathcal{H}_0$: $\alpha_1 = \alpha_2$	$\tilde{\mathcal{H}}_0$: $\alpha_2 = \alpha_3$	$\mathcal{H}_0$: $\alpha_1 = \alpha_2$	$\tilde{\mathcal{H}}_0$: $\alpha_2 = \alpha_3$
(30,100)	0.082	0.903	0.183	0.841
(60,200)	0.087	0.995	0.165	0.992
(120,400)	0.099	1	0.150	1
(240,800)	0.103	1	0.136	1
(60,100)	0.101	0.873	0.188	0.831
(120,200)	0.083	0.997	0.154	0.992
(240,400)	0.102	1	0.144	1
(480,800)	0.100	1	0.132	1

For the model 5 (see Table 6.4), the realized sizes of the test are close to the theoretical ones when we considered the α_i^* known. For the first test ($\alpha_1 = \alpha_2$), $\mathcal{H}_0$ is accepted with an empirical probability of 0.1. For the second test ($\alpha_2 = \alpha_3$), $\tilde{\mathcal{H}}_0$ is not accepted in most of the cases, as expected. When α_i^* is estimated, the realized sizes are slightly higher than the theoretical one, but seems to tend to the correct value. From the last column, we can see that the powers are close to one, even if they are lower than in the case where α_i^* is known. When c increases, the realized sizes of the test ($\alpha_1 = \alpha_2$) are closer to the theoretical value 0.1. The use of estimated α_i in s_k^2 influences more the realized sizes than the powers: the realized sizes increase but tend to the correct value.

Model 6 is a modification of model 5: here we consider that the first spike is of multiplicity three. The results are displayed in Table 6.5.

Table 6.5: Realized size of the test $\mathcal{H}_0$: $\alpha_2 = \alpha_3$ and $\tilde{\mathcal{H}}_0$: $\alpha_3 = \alpha_4$ for model 6, with $c = 0.3$ and $c = 0.6$.

	α_i^* known		α_i^* estimated	
(p, n)	$\mathcal{H}_0$: $\alpha_2 = \alpha_3$	$\tilde{\mathcal{H}}_0$: $\alpha_3 = \alpha_4$	$\mathcal{H}_0$: $\alpha_2 = \alpha_3$	$\tilde{\mathcal{H}}_0$: $\alpha_3 = \alpha_4$
(30,100)	0.023	0.766	0.109	0.708
(60,200)	0.024	0.987	0.100	0.971
(120,400)	0.034	1	0.086	1
(240,800)	0.029	1	0.066	1
(60,100)	0.015	0.781	0.094	0.722
(120,200)	0.026	0.993	0.102	0.983
(240,400)	0.018	1	0.061	1
(480,800)	0.031	1	0.065	1

As expected by the construction of the test, we have lower realized sizes but lower powers too compared to model 5. As in this last one, the realized sizes are increased by using an estimator of α_i, but remain under the nominal level.

Model 7 considers a more difficult case, where the spacing between the spikes is small: it will be harder to distinguish α_2 from α_3, so the powers of the test will be lower. The results are displayed in Table 6.6.

Table 6.6: Realized size of the test $\mathcal{H}_0$: $\alpha_1 = \alpha_2$ and $\tilde{\mathcal{H}}_0$: $\alpha_2 = \alpha_3$ for model 7, with $c = 0.3$ and $c = 0.6$.

	α_i^* known		α_i^* estimated	
(p, n)	$\mathcal{H}_0$: $\alpha_1 = \alpha_2$	$\tilde{\mathcal{H}}_0$: $\alpha_2 = \alpha_3$	$\mathcal{H}_0$: $\alpha_1 = \alpha_2$	$\tilde{\mathcal{H}}_0$: $\alpha_2 = \alpha_3$
(30,100)	0.062	0.039	0.109	0.133
(60,200)	0.064	0.050	0.100	0.123
(120,400)	0.057	0.106	0.086	0.154
(240,800)	0.068	0.220	0.066	0.254
(480,1600)	0.087	0.354	0.106	0.372
(60,100)	0.058	0.041	0.099	0.151
(120,200)	0.058	0.057	0.094	0.127
(240,400)	0.064	0.092	0.085	0.141
(480,800)	0.075	0.199	0.100	0.238
(960,1600)	0.083	0.389	0.096	0.406

The powers are much lower than in the case where the spikes are well-separated. Nevertheless, the realized sizes remain good. The convergence is slower in this case.

6.7 About the estimation of Λ_k

In section 6.2.2 we give the maximum likelihood estimators of the factor scores Λ_k, $\widehat{\Lambda}_k = \left(\lambda_{n,k} - \widehat{\sigma}^2\right)^{1/2} v_{n,k}$, where $v_{n,k}$ is a normalized eigenvector of $\mathsf{S_n}$ corresponding to $\lambda_{n,k}$, for $1 \leq k \leq p$. Therefore, $\text{spec}(\widehat{\Lambda}'_k \widehat{\Lambda}_k) = (\lambda_{n,1}, \ldots, \lambda_{n,m})$, and the eigenvalues α_i are estimated by $\lambda_{n,i}$. However, by (6.13), this estimation is no longer accurate in large dimensional setting. As explained in the previous section, we will estimate α_i^* by $\phi^{-1}(\lambda_{n,i}/\sigma^2)$, i.e. α_i by $\widehat{\alpha}_i = \sigma^2 \phi^{-1}(\lambda_{n,i}/\sigma^2) - \sigma^2$ where

$$\phi^{-1}(x) = \frac{1}{2}\left(x + 1 - c + \sqrt{(x + 1 - c)^2 - 4x}\right).$$

For the eigenvectors $v_{n,k}$, $1 \leq k \leq m$, we know that they do not tend to the corresponding eigenvector u_k of the population covariance Σ when both p and n tend to infinity. More precisely, in Benaych-Georges & Nadakuditi (2011) the authors give the almost sure limit of $|\langle v_{n,k}, \ker(\alpha_i \mathsf{I}_p - \Lambda\Lambda')\rangle|^2$, $i \in J_k$, which is a function of α_k: when α_k is simple, this means that $v_{n,k}$ will, with high probability, lie on a cone around u_k. Furthermore, when the multiplicity of α_k is greater than two, the equations followed by the corresponding eigenvectors will have an infinity of solutions. Consequently, we keep the sample eigenvectors $v_{n,k}$ as an estimate of the population eigenvectors u_k, and we will estimate Λ_k by $\widetilde{\Lambda}_k = \widehat{\alpha}_i^{1/2} v_{n,k}$. The difference compared to $\widehat{\Lambda}$ is the estimation of its norm $\alpha_i^{1/2}$. We can notice that, by applying the delta-method and using (6.14), one can derive the asymptotic distribution of $\{\sqrt{n}(\widehat{\alpha}_j - \alpha_k),\, j \in J_k\}$.

6.8 Appendix: complementary proofs

Proof of (6.9) By Proposition 23, we know that the inverse of the Fisher Information Matrix (FIM) is $\mathcal{I}^{-1}(\psi_{11},\dots,\psi_{pp}) = (2\psi_{ii}^2\psi_{jj}^2\xi^{ij})_{ij}$. We have to change the parametrization: in our case, we have $\psi_{11} = \cdots = \psi_{pp}$. Let $g : \mathbb{R} \to \mathbb{R}^p$, $a \mapsto (a,\dots,a)$. The FIM in this new parametrization becomes

$$\mathcal{I}(\sigma^2) = J'\mathcal{I}(g(\sigma^2))J,$$

where J is the Jacobian matrix of g. As

$$\mathcal{I}(g(\sigma^2)) = \frac{1}{2\sigma^8}(\theta_{ij}^2)_{ij},$$

we have

$$\mathcal{I}(\sigma^2) = \frac{1}{2\sigma^8}\sum_{i,j=1}^{p}\theta_{ij}^2,$$

and

$$\begin{aligned}\Theta = (\theta_{ij})_{ij} &= \Psi - \Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda' \\ &= \sigma^2(\mathsf{I}_p - \Lambda(\Lambda'\Lambda)^{-1}\Lambda').\end{aligned}$$

By hypothesis, we have $\Lambda'\Lambda = \operatorname{diag}(d_1^2,\dots,d_m^2)$. Consider the Singular Value Decomposition of Λ, $\Lambda = \mathsf{UDV}$, where U is a $p\times p$ matrix such that $\mathsf{UU}' = \mathsf{I}_p$, V is a $m\times m$ matrix such that $\mathsf{V}'\mathsf{V} = \mathsf{I}_m$, and D is a $p\times m$ diagonal matrix with $d_1,\dots,d_m$ as diagonal elements. As $\Lambda'\Lambda$ is diagonal, $\mathsf{V} = \mathsf{I}_m$, so $\Lambda = \mathsf{UD}$. By elementary calculus, one can find that

$$\Lambda(\Lambda'\Lambda)^{-1}\Lambda' = \operatorname{diag}(\underbrace{1,\dots,1}_{m},\underbrace{0,\dots,0}_{p-m}),$$

so

$$\Theta = \sigma^2\operatorname{diag}(\underbrace{0,\dots,0}_{m},\underbrace{1,\dots,1}_{p-m}).$$

Finally,

$$\mathcal{I}(\sigma^2) = \frac{1}{2\sigma^8}(p-m)\sigma^4 = \frac{p-m}{2\sigma^4}$$

and

$$\sigma^2_{\text{MLE}} = \mathcal{I}^{-1}(\sigma^2) = \frac{2\sigma^4}{p-m}.$$

Proof of (6.18) By Proposition 24, for $g(x) = x$, by using the variable change $x = \sigma^2(1 + c - 2\sqrt{c}\cos\theta)$, $0 \leq \theta \leq \pi$, we have

$$\begin{aligned}
m(g) &= \frac{g(a(c)) + g(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)} \frac{x}{\sqrt{4c\sigma^4 - (x - \sigma^2 - c\sigma^2)^2}}\,\mathrm{d}x,\ j = 1, \ldots, k \\
&= \frac{\sigma^2(1+c)}{2} - \frac{\sigma^2}{2\pi}\int_0^{\pi} (1 + c - 2\sqrt{c}\cos\theta)\,\mathrm{d}\theta \\
&= 0.
\end{aligned}$$

Proof of (6.19) Let $\underline{s}(z)$ be the Stieltjes transform of $(1-c)\mathbb{1}_{[0,\infty)} + cF_{c,\delta_1}$. One can show that

$$\underline{m}(z) = \frac{1}{\sigma^2}\underline{s}\left(\frac{z}{\sigma^2}\right).$$

Then, in Proposition 24, we have

$$v(f_j, f_l) = -\frac{1}{2\pi^2}\oint\oint \frac{f_j(\sigma^2 z_1) f_l(\sigma^2 z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2),\ j, l = 1, \ldots, k. \tag{6.29}$$

For $g(x) = x$, we have

$$\begin{aligned}
v(g) &= -\frac{1}{2\pi^2}\oint\oint \frac{g(\sigma^2 z_1) g(\sigma^2 z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) \\
&= -\frac{\sigma^4}{2\pi^2}\oint\oint \frac{z_1 z_2}{(\underline{s}(z_1) - \underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) \\
&= 2c\sigma^4,
\end{aligned}$$

where $-\frac{1}{2\pi^2}\oint\oint \frac{z_1 z_2}{(\underline{s}(z_1) - \underline{s}(z_2))^2}\,\mathrm{d}\underline{s}(z_1)\,\mathrm{d}\underline{s}(z_2) = 2c$ is calculated in Bai et al. (2009) (it corresponds to $v(z_1, z_2)$, Section 5, proof of (3.4)).

Proof of (6.22) By Proposition 24, for $\sigma^2 = 1$ and $g(x) = \log(x)$, by using the variable change $x = 1 + c - 2\sqrt{c}\cos\theta$, $0 \leq \theta \leq \pi$, we have

$$\begin{aligned}
m(g) &= \frac{g(a(c)) + g(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)} \frac{x}{\sqrt{4c - (x - 1 - c)^2}}\,\mathrm{d}x,\ j = 1, \ldots, k \\
&= \frac{\log(1-c)}{2} - \frac{1}{2\pi}\int_0^{\pi} \log(1 + c - 2\sqrt{c}\cos\theta)\,\mathrm{d}\theta \\
&= \frac{\log(1-c)}{2} - \frac{1}{4\pi}\int_0^{2\pi} \log|1 - \sqrt{c}e^{i\theta}|^2\,\mathrm{d}\theta \\
&= \frac{\log(1-c)}{2},
\end{aligned}$$

where $\int_0^{2\pi} \log|1 - \sqrt{c}e^{i\theta}|^2\,\mathrm{d}\theta = 0$ is calculated in Bai & Silverstein (2010).

Proof of (6.23) By Proposition 24 and (6.29), for $\sigma^2 = 1$ and $g(x) = x$, we have

$$\begin{aligned} v(g) &= -\frac{1}{2\pi^2} \oint \oint \frac{g(z_1)g(z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2} \, \mathrm{d}\underline{s}(z_1) \, \mathrm{d}\underline{s}(z_2) \\ &= -\frac{1}{2\pi^2} \oint \oint \frac{\log(z_1)\log(z_2)}{(\underline{s}(z_1) - \underline{s}(z_2))^2} \, \mathrm{d}\underline{s}(z_1)\mathrm{d}\underline{s}(z_2) \\ &= -2\log(1 - c_n), \end{aligned}$$

where the last integral is calculated in Bai & Silverstein (2010).

Proof of (6.27) F_{c_n,δ_1} is the Marčenko-Pastur distribution of index c_n. By using the variable change $x = 1 + c_n - 2\sqrt{c_n}\cos\theta$, $0 \leq \theta \leq \pi$, we have

$$\begin{aligned} \int \log(x) dF_{c_n,\delta_1}(x) &= \int_{a(c_n)}^{b(c_n)} \frac{\log x}{2\pi x c_n} \sqrt{(b(c_n) - x)(x - a(c_n))} \, \mathrm{d}x \\ &= \frac{1}{2\pi c_n} \int_0^{\pi} \frac{\log(1 + c_n - 2\sqrt{c_n}\cos\theta)}{1 + c_n - 2\sqrt{c_n}\cos\theta} 4c_n \sin^2\theta \, \mathrm{d}\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} \frac{2\sin^2\theta}{1 + c_n - 2\sqrt{c_n}\cos\theta} \log|1 - \sqrt{c_n}e^{i\theta}|^2 \, \mathrm{d}\theta \\ &= \frac{c_n - 1}{c_n} \log(1 - c_n) - 1, \end{aligned}$$

where the last integral is calculated in Bai & Silverstein (2010).

Proof of (6.24) In the normal case with $\sigma^2 = 1$, Zheng (2012) gives the following equivalent expression of (6.16):

$$v(f_j, f_l) = -\lim_{r \to 1^+} \frac{\kappa}{4\pi^2} \oint \oint_{|\xi_1| = |\xi_2| = 1} f_j(|1 + h\xi_1|^2) f_l(|1 + h\xi_2|^2) \frac{1}{(\xi_1 - r\xi_2)^2} \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2,$$

where $\kappa = 2$ in the real case and $h = \sqrt{c}$ in our case. We take $f_j(x) = \log(x)$ and $f_l(x) = x$, so we need to calculate

$$v(\log(x), x) = -\lim_{r \to 1^+} \frac{1}{2\pi^2} \oint \oint_{|\xi_1| = |\xi_2| = 1} |1 + \sqrt{c}\xi_2|^2 \frac{\log(|1 + \sqrt{c}\xi_1|^2)}{(\xi_1 - r\xi_2)^2} \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2.$$

We follow the calculations done in Zheng (2012): when $|\xi| = 1$, $|1+\sqrt{c}\xi|^2 = (1+\sqrt{c}\xi)(1+\sqrt{c}\xi^{-1})$, so $\log(|1+\sqrt{c}\xi|^2) = \frac{1}{2}\left(\log(1+\sqrt{c}\xi)^2 + \log(1+\sqrt{c}\xi^{-1})^2\right)$. Consequently,

$$\begin{aligned}
\oint_{|\xi_1|=1} \frac{\log(|1+\sqrt{c}\xi_1|^2)}{(\xi_1 - r\xi_2)^2}\,\mathrm{d}\xi_1 &= \frac{1}{2}\oint_{|\xi_1|=1} \frac{\log(1+\sqrt{c}\xi_1)^2}{(\xi_1 - r\xi_2)^2}\,\mathrm{d}\xi_1 + \frac{1}{2}\oint_{|\xi_1|=1} \frac{\log(1+\sqrt{c}\xi_1^{-1})^2}{(\xi_1 - r\xi_2)^2}\,\mathrm{d}\xi_1 \\
&= \frac{1}{2}\oint_{|\xi_1|=1} \log(1+\sqrt{c}\xi_1)^2 \left(\frac{1}{(\xi_1 - r\xi_2)^2} + \frac{1}{(1 - r\xi_1\xi_2)^2}\right)\mathrm{d}\xi_1 \\
&= 0 + i\pi \left(\frac{1}{(r\xi_2)^2}\frac{2\sqrt{c}}{1+\frac{\sqrt{c}}{r\xi_2}}\right) \\
&= 2i\pi \frac{\sqrt{c}}{r\xi_2(r\xi_2+\sqrt{c})}.
\end{aligned}$$

Thus,

$$\begin{aligned}
v(\log(x), x) &= \frac{1}{i\pi}\oint_{|\xi_2|=1} |1+\sqrt{c}\xi_2|^2 \frac{\sqrt{c}}{\xi_2(\xi_2+\sqrt{c})}\,\mathrm{d}\xi_2 \\
&= \frac{1}{i\pi}\oint_{|\xi|=1} \left(1+c+c(\xi+\xi^{-1})\right)\frac{\sqrt{c}}{\xi(\xi+\sqrt{c})}\,\mathrm{d}\xi \\
&= \frac{1}{i\pi}\oint_{|\xi|=1} \left(\frac{\sqrt{c}(1+c)}{\xi(\xi+\sqrt{c})} + \frac{c}{\xi+\sqrt{c}} + \frac{c}{\xi^2(\xi+\sqrt{c})}\right)\mathrm{d}\xi \\
&= 2(1+c-(1+c)+c+1-1) \\
&= 2c.
\end{aligned}$$

Simulation experiments for the conjecture (6.25) We assume that $\sigma^2 = 1$. We have

$$\sum_{i=1}^{p} \log\lambda_i = p\int \log(x)\,\mathrm{d}(F_n - F_{c_n,H_n})(x) + p\int \log(x)\,\mathrm{d}F_{c_n,H_n}(x).$$

We know that

$$p\int \log(x)\,\mathrm{d}(F_n - F_{c_n,H_n})(x) \to \mathcal{N}\left(m_1(c), v_1(c)\right),$$

where $m_1(c) = \frac{\log(1-c)}{2}$ and $v_1(c) = -2\log(1-c)$. We consider three different settings

- Model A: $\mathrm{spec}(\Sigma) = (4,3,0,\ldots,0) + (1,\ldots,1)$, $c = 0.2$;
- Model B: $\mathrm{spec}(\Sigma) = (25,16,9,0,\ldots,0) + (1,\ldots,1)$, $c = 0.9$;
- Model C: $\mathrm{spec}(\Sigma) = (8,7,0,\ldots,0) + (1,\ldots,1)$, varying c.

In Table 6.7, we give the empirical mean of $\sum_{i=1}^{p}\log\lambda_i - \frac{\log(1-c)}{2}$ over 1000 replications, as well as the value of

$$\sum_{i=1}^{m} \log(\alpha_i + 1) + p\int \log(x)\,\mathrm{d}F_{c_n,\delta_1}(x),$$

where

$$\int \log(x)\,\mathrm{d}F_{c_n,\delta_1}(x) = \frac{c_n - 1}{c_n}\log(1-c_n) - 1.$$

Table 6.7: Empirical mean over 1000 replications, compared to the conjectured expression.

	Settings		Conjecture	Empirical mean
Model A	$p = 20$	$n = 100$	0.8472	0.8507
	$p = 80$	$n = 400$	-5.5983	-5.6063
	$p = 200$	$n = 1000$	-18.4894	-18.5231
Model B	$p = 90$	$n = 100$	-58.5803	-58.5959
	$p = 180$	$n = 200$	-125.5544	-125.5406
	$p = 720$	$n = 800$	-527.3993	-527.3442
Model C	$p = 10$	$n = 500$	-6.5567	-6.5572
	$p = 50$	$n = 500$	1.6889	1.6860
	$p = 200$	$n = 500$	-42.4756	-42.4657
	$p = 300$	$n = 500$	-112.4652	-112.4810

We see that the empirical mean is close to the value of the conjecture, in all the cases.

Conclusion and perspectives

Statistical analysis of large dimensional data is a relatively recent field, which interest appeared with the possibility of the observation and the storage of a high amount of data. Random matrix theory is a theoretical framework which allows to solve some problems in this specific context.

Spiked population model, which can be viewed as a reformulation of the factor model, appears in several scientific fields. This is the reason why it has been studied in the large dimensional contest using, in particular, random matrix theory.

Determination of the number of factors (or spikes) is a fundamental problem which is often a first step before a more complete study. In the first part of this thesis, we constructed a new method for the estimation of the number of spikes in a spiked population model, by analyzing the behavior of the difference of two consecutive ordered eigenvalues of the sample covariance matrix. We proved its consistency in the simple spikes case, as well as in the case where their multiplicity is greater than one. We proposed several thresholds verifying the consistency criterion, including an auto-adaptive one. The main advantages of this method are its simplicity and its ease of implementation, as well as its good performance.

The second part of our work considers the strict factor model with homoscedastic variance. We studied the maximum likelihood estimator of the variance by establishing its limiting distribution, when the sample and data sizes both tend to infinity. Thus, we can give its bias expression, which does not appear in the classical framework. Then we corrected the goodness-of-fit test to a factor model based on the likelihood ratio statistic. We finally construct an equality test of two consecutive spikes.

The current state of our work shows us several perspectives for further works.

(i) **Extension of the consistency result of the factor number estimator to the generalized spiked population model.** Almost sure convergence results have been proved by Bai & Yao (2012) for these models, which generalized the results of Baik & Silverstein (2006). It will be still possible to distinguish the spikes from the noise by analyzing the consecutive differences between two eigenvalues of S_n. The equality case will be more difficult to study, since there is no existing results on the speed of convergence for the non-spiked eigenvalues.

(ii) **Construction of an estimator for the spikes multiplicity from the equality test.** Assuming that the number of spikes is known, the result will be a partition of the spikes depending on their multiplicity. We could, for example, calculate the p-value of the equality test of two consecutive eigenvalues. From these p-values, we would search the partition that best fits, by minimizing a function of the p-value suitably chosen.

(iii) **Correction of the maximum likelihood estimator of the factor scores.** Actually, it means to study the eigenvectors behavior of the sample covariance matrix, as a function of those of the population covariance matrix. Few results exist (as Benaych-Georges & Nadakuditi (2011)), but it is difficult to find a better estimator than the eigenvectors of the sample covariance matrix.

Bibliography

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973. Cited pages x, 31 and 32.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis. Cited pages x, 31 and 32.

Naum Ilyich Akhiezer. *The classical moment problem and some related questions in analysis.* Translated by N. Kemmer. Hafner Publishing Co., New York, 1965. Cited page 9.

Yasuo Amemiya and Theodore W. Anderson. Asymptotic chi-square tests for a large class of factor analysis models. *Ann. Statist.*, 18(3):1453–1463, 1990. Cited page 82.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 2010. Cited pages vii and 1.

Theodore W. Anderson. *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third édition, 2003. Cited pages 25, 28 and 61.

Theodore W. Anderson and Yasuo Amemiya. The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.*, 16(2):759–771, 1988. Cited pages xiv, 28, 82 and 83.

Theodore W. Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. V*, pages 111–150, Berkeley and Los Angeles, 1956. University of California Press. Cited pages xiv, 84 and 90.

Ludwig Arnold. On the asymptotic distribution of the eigenvalues of random matrices. *J. Math. Anal. Appl.*, 20:262–268, 1967. Cited page 6.

Ludwig Arnold. On Wigner's semicircle law for the eigenvalues of random matrices. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 19:191–198, 1971. Cited page 6.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002. Cited pages 35, 40 and 53.

Zhidong D. Bai, Jiaqi Chen, and Jian-Feng Yao. On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.*, 52 (4):423–437, 2010. Cited pages 6 and 87.

Zhidong D. Bai, Dandan Jiang, Jian-Feng Yao, and Shurong Zheng. Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37(6B):3822–3840, 2009. Cited pages viii, xv, 7, 15, 16, 86, 91 and 101.

Zhidong D. Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, 6(2):311–329, 1996. Cited page 4.

Zhidong D. Bai and Jack W. Silverstein. Clt for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Prob.*, 32(1A):553–605, 2004. Cited pages viii, 1, 14, 47 and 52.

Zhidong D. Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices.* Springer Series in Statistics. Springer, New York, second édition, 2010. Cited pages vii, xiv, xv, 1, 5, 8, 10, 12, 85, 86, 101 and 102.

Zhidong D. Bai, Jack W. Silverstein, and Yong Q. Yin. A note on the largest eigenvalue of a large-dimensional sample covariance matrix. *J. Multivariate Anal.*, 26(2):166–168, 1988. Cited page 17.

Zhidong D. Bai and Jian-Feng Yao. Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(3):447–474, 2008. Cited pages xii, xiii, xv, 1, 21, 40, 42, 56, 62, 65, 77, 85 and 95.

Zhidong D. Bai and Jian-Feng Yao. On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.*, 106:167–177, 2012. Cited pages viii, xvi, 18, 19, 20 and 105.

Zhidong D. Bai and Yong Q. Yin. Convergence to the semicircle law. *Ann. Probab.*, 16(2): 863–875, 1988. Cited page 6.

Zhidong D. Bai, Yong Q. Yin, and Paruchuri R. Krishnaiah. On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic. *J. Multivariate Anal.*, 19(1):189–200, 1986. Cited page 6.

Zhidong D. Bai, Yong Q. Yin, and Paruchuri R. Krishnaiah. On limiting empirical distribution function of the eigenvalues of a multivariate F matrix. *Teor. Veroyatnost. i Primenen.*, 32(3):537–548, 1987. Cited page 6.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005. Cited page 18.

Jinho Baik and Jack W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, 97(6):1382–1408, 2006. Cited pages viii, xi, xiv, 1, 18, 19, 41, 64, 85 and 105.

Maurice S. Bartlett. Test of significance in factor analysis. *Brit. Jour. Psych.*, 3:97–104, 1950. Cited pages 28 and 94.

Florent Benaych-Georges, Alice Guionnet, and Mylène Maida. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.*, 16:no. 60, 1621–1662, 2011. Cited pages xii, 18, 22, 40, 43 and 77.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, 227(1):494–521, 2011. Cited pages xvi, 18, 99 and 106.

Pascal Bianchi, M'erouane Debbah, Mylène Maïda, and Jamal Najim. Performance of statistical tests for single source detection using random matrix theory. *IEEE Transactions on Information Theory*, 57(4):2400–2419, 2011. Cited pages 6 and 82.

Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 2(1):245–276, 1966. Cited page 31.

Florent Chamberlain, Gary and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983. Cited pages 23 and 82.

Patrick L. Combettes and Jack W. Silverstein. Signal detection via spectral theory of large dimensional random matrices. *IEEE Trans. Signal Process.*, 8(40):2100–2105, 1992. Cited pages 6 and 62.

Romain Couillet and Mérouane Debbah. A bayesian framework for collaborative multi-source signal sensing. *IEEE Trans. Signal Process.*, 58(10):5186–5195, 2010. Cited page 6.

Romain Couillet and Mérouane Debbah. *Random matrix methods for wireless communications.* Cambridge university press édition, 2011. Cited page 6.

Robert Cudeck and Robert C. MacCallum. *Factor analysis at 100: historical developments and future directions.* Mahwah, N.J., lawrence erlbaum associates édition, 2007. Cited page 23.

Arthur P. Dempster. A high dimensional two sample significance test. *Ann. Math. Statist.*, 29:995–1010, 1958. Cited page 4.

Noureddine El Karoui. Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Physica Polonica B*, 36(9):2681–2697, 2005. Cited page 6.

Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.*, 36(6):2757–2790, 2008. Cited page 6.

Richard Everson and Stephen Roberts. Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Trans. Signal Process.*, 48(7):2083–2091, 2000. Cited page 62.

Eran Fishler, Michael Grosmann, and Hagit Messer. Detection of signals by information theoretic criteria: general asymptotic performance analysis. *IEEE Trans. Signal Process.*, 50(5):1027–1036, 2002. Cited page 32.

Michel Gaudin. Sur la loi limite de l'espacement des valeurs propres d'une matrice aléatoire. *Nucl. Phys.*, 25:2083–2091, 1961. Cited page 97.

Ulf Grenander. *Probabilities on algebraic structures.* John Wiley & Sons Inc., New York, 1963. Cited page 6.

Ulf Grenander and Jack W. Silverstein. Spectral analysis of networks with random topologies. *SIAM J. Appl. Math.*, 32(2):499–519, 1977. Cited page 6.

Walid Hachem, Philippe Loubaton, Xavier Mestre, Jamal Najim, and Pascal Vallet. Large information plus noise matrix models and consistent subspace estimation in large sensor networks. *Random Matrices: Theory and Applications*, 1(2):1150006, 2012. Cited pages 4, 6 and 82.

Matthew C. Harding. Structural estimation of high-dimensional factor models. *Econometrica, r&r*, 2007. Cited pages x, xii, 6, 29, 34, 36, 40, 45, 52 and 62.

Harold Hotteling. Analysis of a complex of statistical variables into principal components. *J. Edu. Psycho.*, 24(6–7):417–441 ; 498–520, 1973. Cited page 23.

Peter J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, 1:799–821, 1973. Cited page 4.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001. Cited pages viii, x, 1, 17, 35, 54, 62, 70 and 82.

Dag Jonsson. Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.*, 12(1):1–38, 1982. Cited page 6.

John L. Kelley. *Crossroads in the Mind of Mind.* Standford University Press, 1928. Cited page 23.

Mark G. Kreĭn and Adolf A. Nudel′man. *The Markov moment problem and extremal problems.* American Mathematical Society, Providence, R.I., 1977. Ideas and problems of P. L. Čebyšev and A. A. Markov and their further development, Translated from the Russian by D. Louvish, Translations of Mathematical Monographs, Vol. 50. Cited page 9.

Shira Kritchman and Boaz Nadler. Determining the number of components in a factor model from limited noisy data. *Chem. Int. Lab. Syst.*, 94, 2008. Cited pages x, xii, xiii, xiv, 35, 36, 37, 40, 52, 58, 61, 62, 65, 70, 76, 82, 86 and 89.

Shira Kritchman and Boaz Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.*, 57(10): 3930–3941, 2009. Cited pages x, xiv, 6, 33, 35, 36, 37, 62, 65, 70, 71, 73, 76 and 86.

Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Statist.*, 40(2):694–726, 2012. Cited page 62.

Clifford Lam, Qiwei Yao, and Niel Bathia. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918, 2011. Cited page 62.

D. N. Lawley. The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinburgh*, 60:64–82, 1940. Cited pages ix, xiii, 25, 26, 81, 82 and 83.

D. N. Lawley and A. E. Maxwell. *Factor analysis as a statistical method.* American Elsevier Publishing Co., Inc., New York, second édition, 1971. Cited pages ix, 25 and 82.

Olivier Ledoit and Michael Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.*, 30(4):1081–1102, 2002. Cited page 7.

Athanasios P. Liavas and Phillip A. Regalia. On the behavior of information theoretic criteria for model order selection. *IEEE Trans. Signal Process.*, 49(8):1689–1695, 2001. Cited page 32.

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967. Cited pages 6, 10 and 47.

Madan L. Mehta. *Random matrices*, volume 142 of *Pure and Applied Mathematics (Amsterdam).* Elsevier/Academic Press, Amsterdam, third édition, 2004. Cited pages 6, 96 and 97.

Thomas T. Minka. Automatic of dimensionality for pca. Rapport technique, Massachusetts Institute of Technology, 2000. Cited pages x and 33.

Raj R. Nadakuditi, Mingo James A., Roland Speicher, and Alan Edelman. Statistical eigen-inference from large wishart matrices. *Ann. Statist.*, 36(6):2850–2885, 2008. Cited page 6.

Raj R. Nadakuditi and Alan Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans. Signal Process.*, 56(7, part 1):2625–2638, 2008. Cited page 62.

Raj R. Nadakuditi and Jack W. Silverstein. Cited page 6.

Boaz Nadler. Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. *IEEE Trans. Signal Process.*, 58(5):2746–2756, 2010. Cited pages 62 and 70.

Tormod Naes, Tomas Isaksson, Tom Fearn, and Tony Davies. *User-friendly guide to multivariate calibration and classification.* NIR Publications, Chichester, 2002. Cited pages 40, 62 and 82.

Alexei Onatski. Testing hypotheses about the numbers of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009. Cited pages 6, 40, 41 and 62.

Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4):1004–1016, 2010. Cited page 6.

Alexei Onatski, Marcelo J. Moreira, and Marc Hallin. Signal detection in high dimension: the multispiked case. Rapport technique, University of Cambridge, Fundação Getulio Vargas and Université libre de Bruxelles and Princeton, 2012. Cited page 6.

Damien Passemier and Jian-Feng Yao. Estimation of the number of factors, possibly equal, in the high-dimensional case. Rapport technique, IRMAR, Université de Rennes 1 and SAAS, The Unversity of Hong Kong, 2012a. Cited page 89.

Damien Passemier and Jian-Feng Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1(1): 1150002, 2012b. Cited pages 56, 62, 63, 64, 65, 66, 77, 78, 79 and 89.

Leonid A. Pastur. The spectrum of random matrices. *Theoret. and Math. Phys.*, 10(1): 67–74, 1972. Cited page 6.

Leonid A. Pastur. Spectra of random selfadjoint operators. *Russian Math. Surveys*, 28(1): 1–67, 1973. Cited page 6.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, 17(4):1617–1642, 2007. Cited pages xii, 18, 21, 40, 42 and 62.

Jorma Rissanen. Modeling by shortest data description. *Automatica—J. IFAC*, 14(5): 465–471, 1978. Cited pages x and 31.

Stephen A. Ross. The arbitrage theory of capital asset pricing. *J. Econom. Theory*, 13(3): 341–360, 1976. Cited pages 23, 40, 61 and 82.

James R. Schott. A test for the equality of covariance matrices when the dimension is large relative to the sample size. *Comput. Statist. Data Anal.*, 51:6535–6542, 2007. Cited page 7.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. Cited pages 31 and 32.

Jack W. Silverstein. The limiting eigenvalue distribution of a multivariate F matrix. *SIAM J. Math. Anal.*, 16(3):641–646, 1985. Cited page 6.

Charles Spearman. General intelligence objectively determined and measured. *Am. J. Psychol.*, 15(2):201–293, 1904. Cited page 23.

Muni S. Srivastava. Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.*, 35:251–272, 2005. Cited page 7.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981. Cited page 34.

Petre Stoica and Yngve Sélen. Model-order selection: a review of information criterion rules. *Signal Processing Magazine, IEEE*, 21(4):36–47, 2004. Cited page 32.

Louis L. Thurstone. Multiple factor analysis. *Psychological Review*, 38(4):406–427, 1931. Cited pages 23 and 25.

Antonio M. Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Foundations and Trends in Commun. and Inf. Theory*, 1(1):1–182, 2004. Cited page 23.

Magnus O. Ulfarsson and Victor Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.*, 56(12):5804–5816, 2008. Cited pages x, 34 and 62.

Pascal Vallet, Philippe Loubaton, and Xavier Mestre. Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case. *IEEE Transactions on Information Theory*, 58(2):1043–1068, 2012. Cited page 82.

Kenneth W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Prob.*, 6(1):1–18, 1978. Cited page 6.

Kenneth W. Wachter. The limiting empirical measure of multiple discriminant ratios. *Ann. Statist.*, 8(5):937–957, 1980. Cited page 6.

Mati Wax and Thomas Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Signal Process.*, 33(2):387–392, 1985. Cited pages 62 and 70.

Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)*, 62:548–564, 1955. Cited page 6.

Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. of Math. (2)*, 67:325–327, 1958. Cited page 6.

Wenyuan Xu and Mostafa Kaveh. Analysis of the performance and sensitivity of eigendecomposition-based detectors. *IEEE Trans. Signal Process.*, 43(6):1413–1426, 1995. Cited page 32.

Yong Q. Yin. Limiting spectral distribution for a class of random matrices. *J. Multivariate Anal.*, 20(1):50–68, 1986. Cited page 6.

Yong Q. Yin and Paruchuri R. Krishnaiah. A limit theorem for the eigenvalues of product of two random matrices. *J. Multivariate Anal.*, 13(4):489–507, 1983. Cited page 6.

Qi-Tu Zhang, Max W. Wong, Patrick C. Yip, and James P. Reilly. Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing. *IEEE Trans. Acoust. Speech Signal Process.*, 37(10):1557–1567, 1989. Cited page 32.

Jian-Hua Zhao, Philip L. H. Yu, and Qibao Jiang. ML estimation for factor analysis: EM or non-EM? *Stat. Comput.*, 18(2):109–123, 2008. Cited page 83.

Shurong Zheng. Central limit theorems for linear spectral statistics of large dimensional f-matrices. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(2):444–476, 2012. Cited pages xv, 91, 102 and 103.

Inférence statistique dans un modèle à variances isolées de grande dimension

Résumé : Cette thèse s'intéresse à l'estimation statistique dans un modèle à variances isolées (modèle spike) de grande dimension. La théorie des matrices aléatoires permet de prendre en compte cette spécificité, puisque la plupart des résultats limites s'appliquent aux matrices dont la taille tend vers l'infini. Une part importante de ces résultats concerne la matrice de covariance empirique.

Dans un premier temps, nous nous intéressons à l'estimation du nombre de facteurs/spikes. La différence de comportement des valeurs propres de la matrice de covariance empirique, selon que l'on considère celles correspondant aux spikes ou non, nous permet de construire un estimateur. Ce dernier correspond à la différence de deux valeurs propres consécutives ordonnées. Nous établissons la consistance de l'estimateur dans le cas où toutes les spikes sont distinctes, et le comparons à deux méthodes existantes à travers des simulations. L'estimateur dépend d'un seuil qui doit remplir certaines conditions. Dans la suite, nous étendons le résultat de consistance au cas d'égalité et améliorons l'estimateur en changeant de seuil.

Dans un second temps, nous considérons les estimateurs du maximum de vraisemblance d'un modèle à facteurs strict à variance homoscédastique. En utilisant un théorème limite pour les statistiques spectrales linéaires, nous corrigeons l'estimateur de la variance commune en grande dimension en donnant l'expression de son biais et en établissant sa loi limite. Nous présentons une version corrigée du test du rapport de vraisemblance d'adéquation à un modèle à facteurs. Finalement, nous construisons un test d'égalité de deux spikes.

Mots clefs : Matrices aléatoires, grande dimension, modèle à facteurs, modèle à variances isolées, mesure spectrale, matrice de covariance, test d'hypothèses, valeurs propres extrêmes, estimation paramétrique, maximum de vraisemblance.

Statistical inference in a high-dimensional spiked population model

Abstract: This thesis deals with the statistical inference of large dimensional data. The random matrix theory allows to take into account this framework, since most asymptotic results apply to large-dimensional random matrices. A large number of these results concerns the population covariance matrix.

First, we are interested in estimating the number of factors/spikes in large dimension. To construct our estimator, we use the fact that the eigenvalue behavior of the sample covariance matrix differs depending on whether they correspond to spikes or not. The estimator is based on differences between consecutive ordered eigenvalues. We establish the consistency of the estimator in the case where all the spikes are different, and compare it to two existing methods through simulation experiments. The estimator depends on a threshold which should satisfy some conditions. Furthermore, we extend our result of consistency to the equality case and improve our estimator by using a dimension-adapted threshold.

Secondly, we consider the maximum likelihood estimator in a strict factor model with homoscedastic variance. Using a central limit theorem for linear spectral statistics, we correct the estimator of the common variance in high-dimensional setting by evaluating its bias and establishing its limiting law. We present a corrected version of the goodness-of-fit test for a factor model. Finally, we propose a test for the equality of two spikes.

Keywords: Random matrices, large dimension, factor model, spiked population model, spectral distribution, sample covariance matrix, hypothesis testing, extreme eigenvalues, parametric estimation, maximum-likelihood estimation.