Estimation robuste dans le régime des grandes matrices aléatoires

Romain COUILLET[1],Abla KAMMOUN[2],Frédéric PASCAL[1]

[1]L2S, CentraleSupélec, CNRS, Université PSud, Gif-sur-Yvette, France

[2]King Abdullah's University of Science and Technology, Arabie Saoudite

romain.couillet@centralesupelec.fr,abla.kammoun@kaust.edu.sa,frederic.pascal@centralesupelec.fr

Résumé − Cet article propose un bref aperçu de résultats récents des auteurs sur l'analyse d'estimateurs robustes de matrices de covariance dans le régime dit des grandes matrices aléatoires, où le nombre d'échantillons et la taille des données augmentent au même rythme. Les principaux résultats montrent que ces estimateurs *implicites* peuvent être approximés asymptotiquement par des matrices aléatoires *explicites* plus faciles à étudier. De nombreuses nouvelles applications en découlent.

Abstract − This article proposes a short survey on recent results by the authors on the performance analysis of robust estimators of scatter in the so-called random matrix regime, where the number of population and sample dimensions grow large at the same rate. The main finding consists in exhibiting an asymptotically tight *explicit* approximation for those estimators that often take an *implicit* form. Many probabilistic corollaries and statistical applications unfold.

1 Introduction

La récente impulsion du BigData remet au goût du jour le domaine de l'estimation robuste, né il y a plus de cinquante ans maintenant sous l'impulsion d'auteurs tels que Tukey et Huber (Huber [1964], Donoho [2000]). Un des points important levé par Tukey et Huber à l'époque est la capacité d'effectuer de l'inférence statistique automatisée en dépit de l'intrusion possible de données erronées ou de bruits impulsifs.

Nous nous intéressons ici aux estimateurs robustes de matrices de dispersion (ou de manière équivalente, de matrices de covariance). Initialement introduits par Huber (Huber [1964]), puis généralisés par la suite par Maronna (Maronna [1976]), ces estimateurs permettent de mieux appréhender les deux scénarios évoqués plus haut, en particulier pour des données indépendantes de loi elliptique ou des données contaminées par des échantillons aberrants. Si on appelle ces données $x_1, \ldots, x_n \in \mathbb{C}^N$ (ou $\mathbb{R}^N$), avec $n > N$, l'estimateur en question, noté $\hat{C}_N$, est la matrice définie comme l'unique solution de

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^{n} u\left(\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i\right) x_i x_i^* \qquad (1)$$

où $u(t)$ est une fonction continue de $\mathbb{R}^+$ dans $\mathbb{R}^+$, avec $t \mapsto \phi(t) \triangleq tu(t)$ croissante et bornée. On assurera également que $\phi_\infty \triangleq \lim_{t\to\infty} \phi(t) > 1$. Dans le cas de Huber, $u(t)$ est constante sur un intervalle puis décroissante en $1/t$ au delà d'un seuil. Dans le cas de Maronna, $u(t)$ prend une forme très générale et s'adapte, pour ce qui est des lois elliptiques, aux paramètres de la loi.

Certains résultats sur $\hat{C}_N$ sont connus dans la littérature des statistiques robustes, particulièrement lorsque n tend vers l'infini. Cependant, les difficultés liées à la définition implicite de $\hat{C}_N$ sont longtemps demeurées problématiques et la théorie des statistiques robustes est tombée un moment en désuétude. La résurgence des modèles de traitement d'antennes puis le domaine du BigData ont relancé récemment l'intérêt pour ces méthodes d'estimation robuste sous l'angle du régime où N et n sont tous deux larges et tels que $N \simeq n$.

La nouveauté de cette hypothèse réside dans la possibilité d'exploiter des résultats de moyennage de variables aléatoires à la fois dans la direction du temps mais aussi dans la direction de l'espace, nous amenant ainsi dans le domaine plus puissant des matrices aléatoires. Grâce aux matrices aléatoires, un phénomène nouveau va apparaître qui n'est aucunement valable dans le cas où N est supposé fixe. Nous allons en effet pouvoir montrer, plus spécifiquement dans l'hypothèse où les x_i sont des vecteurs aléatoires elliptiques (mais aussi lorsque des observations erronées et considérées arbitraires sont effectuées), que $\hat{C}_N$ se comporte asymptotiquement (i.e., lorsque $N, n \to \infty$) de la même manière qu'une matrice aléatoire $\hat{S}_N$ de définition explicite (et non plus solution d'une équation implicite) dont le modèle est très facile à analyser. Précisément, sous des hypothèses appropriées mais en général légères, on montrera que $\|\hat{C}_N - \hat{S}_N\| \to 0$ presque sûrement, où la norme considérée est la norme spectrale. Ce résultat permet de transférer de nombreuses propriétés de $\hat{S}_N$ vers $\hat{C}_N$ et ainsi de permettre leur étude

de manière bien plus précise que dans le cas classique où $n \gg N$.

Le reste de l'article expose ces résultats clairement dans le cas de données elliptiques ou de données gaussiennes contaminées par des données absurdes intermittentes.

Quelques notations: Dans la suite, les matrices seront notées en majuscules, les vecteurs en minuscules. Pour C hermitienne, nous écrirons $C \succeq 0$ pour indiquer qu'elle est définie positive. La norme $\|\cdot\|$ est la norme spectrale pour les matrices et euclidienne pour les vecteurs. La convergence presque sûre est notée "$\xrightarrow{\text{p.s.}}$".

2 Cas elliptique

Notre étude concerne l'estimateur de matrice de dispersion de Maronna, introduit brièvement en (1). Nous considérons tout d'abord que les dimensions N des échantillons $x_1, \ldots, x_n$ et leur nombre n sont telles que, lorsque $n \to \infty$, $c_N \triangleq N/n$ vérifie l'hypothèse suivante.

Hypothèse 1. *Pour chaque n, $c_N < 1$ et*

$$c_- < \liminf_n c_N \leq \limsup_n c_N < c_+$$

où $0 < c_- < c_+ < 1$.

Nous supposons ici que les $x_i \in \mathbb{C}^N$ sont indépendants et de distribution similaire (en fait plus générale dans un certain sens) à une loi elliptique.

Hypothèse 2. *Les vecteurs $x_i = \sqrt{\tau_i} C_N^{\frac{1}{2}} w_i$, $i \in \{1, \ldots, n\}$, satisfont les propriétés suivantes:*

1. *la distribution empirique des amplitudes aléatoires $\tilde{\nu}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\tau_i}$ vérifie $\int \tau \tilde{\nu}_n(d\tau) \xrightarrow{\text{p.s.}} 1$*
2. *il existe $\varepsilon < 1 - \phi_\infty^{-1} < 1 - c_+$ et $m > 0$ tel que, pour tout n large presque sûrement, $\tilde{\nu}_n([0, m)) < \varepsilon$*
3. *$C_N \succ 0$ et $\limsup_N \|C_N\| < \infty$*
4. *$w_1, \ldots, w_n \in \mathbb{C}^N$ sont i.i.d., complexes unitairement invariants, de moyenne nulle telle que, pour tout i, $\|w_i\|^2 = N$, et sont indépendants des $\tau_1, \ldots, \tau_n$.*

Il est important de noter ici que x_i peut être essentiellement vu comme un vecteur unitairement invariant de norme contrôlée par le paramètre τ_i. Les τ_i peuvent être arbitrairement définis mais s'ils sont indépendants et de même distribution, alors les observations x_i sont de loi elliptique centrée, paramétrée par la distribution des τ_i. Ainsi, en choisissant les τ_i de sorte que $2N\tau_i$ suive une loi du chi-carré à $2N$ degrés de libertés, alors les x_i sont gaussiens et sont à queue légère (comme en témoigne le fait que $\tau_i \xrightarrow{\text{p.s.}} 1$ dans ce cas). Si on choisit pour les τ_i des lois à support (asymptotique) non compact, la nature des x_i devient alors impulsive. Par exemple, si $1/\tau_i$ est de loi chi-carré de degré de liberté fixe, indépendant de N, alors x_i suit une loi de Student centrée de paramètre adapté au degré de liberté de la loi du chi-carré. Il est ainsi possible de paramétrer les observations pour les rendre plus ou moins impulsives.

Nous appellerons $\hat{C}_N$ la matrice définie, lorsqu'elle existe, comme la solution unique de l'équation (1) où u satisfait les propriétés suivantes:

(i) $u : [0, \infty) \to (0, \infty)$ est continue et décroissante

(ii) $\phi : x \mapsto xu(x)$ est croissante et bornée, et telle que $\lim_{x\to\infty} \phi(x) \triangleq \phi_\infty > 1$

(iii) $\phi_\infty < c_+^{-1}$.

Une dernière hypothèse technique s'avère importante (et vraisemblablement nécessaire) pour l'obtention de notre résultat central.

Hypothèse 3. *Pour chaque $a > b > 0$, presque sûrement,*

$$\limsup_{t\to\infty} \frac{\limsup_n \tilde{\nu}_n((t, \infty))}{\phi(at) - \phi(bt)} = 0.$$

Cette hypothèse dit que le poids de la queue de la distribution empirique des τ_i doit rester faible relativement au taux de croissance de ϕ. Par exemple, pour $u(t) = (1 + t)/(t + \alpha)$, $\alpha > 0$, souvent utilisée dans la littérature, et τ_i i.i.d., il est suffisant que $\mathrm{E}[\tau_1^{(1+\varepsilon)}] < \infty$ pour un $\varepsilon > 0$ quelconque, ce est une contrainte légère.

Nous pouvons alors introduire notre résultat principal.

Théorème 1. *Sous les Hypothèses 1–3,*

$$\left\| \hat{C}_N - \hat{S}_N \right\| \xrightarrow{\text{p.s.}} 0$$

où

$$\hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^n v(\tau_i \gamma_N) x_i x_i^*$$

avec γ_N l'unique solution positive de

$$1 = \frac{1}{n} \sum_{i=1}^n \frac{\psi(\tau_i \gamma_N)}{1 + c_N \psi(\tau_i \gamma_N)}$$

et avec $v : x \mapsto (u \circ g^{-1})(x)$, $\psi : x \mapsto xv(x)$, et $g : \mathbb{R}_+ \to \mathbb{R}_+$, $x \mapsto x/(1 - c_N\phi(x))$.

Le résultat stipule donc que $\hat{C}_N$ devient asymptotiquement bien approximée par $\hat{S}_N$ dont la définition est explicite, si ce n'est pour le paramètre γ_N; cependant, γ_N dépend uniquement de $\tau_1, \ldots, \tau_n$ et est indépendant des vecteurs $w_1, \ldots, w_n$. Conditionnellement aux τ_i, $\hat{S}_N$ est un modèle matriciel aléatoire bien connu et étudié (essentiellement) déjà dans l'article de Marčenko–Pastur en 1967 (Marčenko and Pastur [1967]) pour $C_N = I_N$, puis plus tard dans (Silverstein and Bai [1995], Silverstein and Choi [1995]), et par la suite dans (Couillet and Hachem [2014]).

Du fait de la convergence $\|\hat{C}_N - \hat{S}_N\| \xrightarrow{\text{p.s.}} 0$, un certain nombre des propriétés de $\hat{S}_N$ déterminées dans ces articles se transfère immédiatement à $\hat{C}_N$. En particulier, le Théorème 1 implique la propriété suivante: les

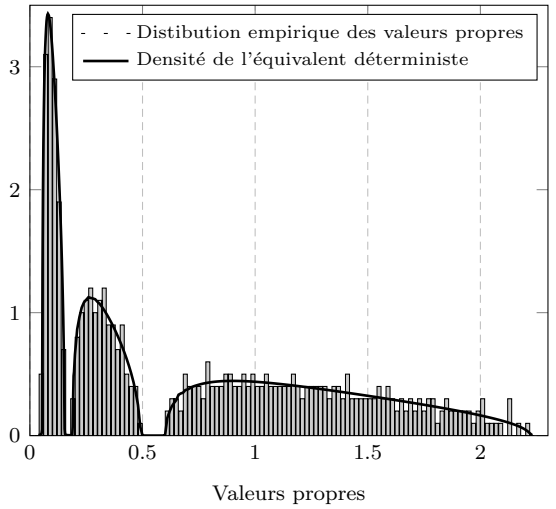

FIG. 1: Histogramme des valeurs propres de $\hat{C}_N$ fpour $n = 2500$, $N = 500$, $C_N = diag(I_{125}, 3I_{125}, 10I_{250})$, τ_1 de loi $\Gamma(.5, 2)$. Comparaison avec la loi limite des valeurs propres de la matrice $\hat{S}_N$.

valeurs propres ordonnées $\lambda_1(\hat{C}_N) \geq \ldots \geq \lambda_N(\hat{C}_N)$ et $\lambda_1(\hat{S}_N) \geq \ldots \geq \lambda_N(\hat{S}_N)$ vérifient

$$\max_{1\leq i\leq N} \left|\lambda_i(\hat{C}_N) - \lambda_i(\hat{S}_N)\right| \xrightarrow{\text{p.s.}} 0$$

et donc nous avons ici une caractérisation très précise des valeurs propres individuelles de $\hat{C}_N$.

La Figure 1 présente un exemple graphique de la distribution des $\lambda_i(\hat{C}_N)$ vis-à-vis de la loi limite des valeurs propres de $\hat{S}_N$ pour des τ_i i.i.d. de loi Gamma. Notons de manière fondamentale ici qu'en écrivant

$$\frac{1}{n}\sum_{i=1}^{n} v(\tau_i\gamma_N) x_i x_i^* = \frac{1}{\gamma_N}\frac{1}{n}\sum_{i=1}^{n} \psi(\tau_i\gamma_N) C_N^{\frac{1}{2}} w_i w_i^* C_N^{\frac{1}{2}}$$

alors, comme ψ est borné, on trouve que $\limsup_N \|\hat{C}_N\| < \infty$. Ce résultat implique qu'alors que la matrice de covariance empirique $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^*$ a un support asymptotique souvent non borné, les valeurs propres de $\hat{C}_N$ restent quant à elles bornées. Ce résultat est fondamental pour les applications en estimation sous-espace où il est crucial que des valeurs propres puissent s'isoler.

3 Présence de données erronées

Nous étudions maintenant le comportement de $\hat{C}_N$ lorsqu'une partie des données observées est déterministe et inconnue. L'idée est ici de déterminer les quantités fondamentales régissant l'effet de rejet de données absurdes par $\hat{C}_N$.

Le modèle que nous considérerons ici est celui d'une observation matricielle $X \in \mathbb{C}^{N\times n}$ formées des n observations de dimension N en colonnes

$$X = \left[x_1, \ldots, x_{(1-\varepsilon_n)n}, a_1, \ldots, a_{\varepsilon_n n}\right]$$

où ε_n est la proportion de données erronées (ou absurdes) observées, où $x_1, \ldots, x_{(1-\varepsilon_n)n} \in \mathbb{C}^N$ sont des vecteurs de "données correctes" modélisées par $x_i = C_N^{\frac{1}{2}} w_i$, $C_N \in \mathbb{C}^{N\times N}$ déterministe et définie positive et $w_1, \ldots, w_{(1-\varepsilon_n)n}$ i.i.d. gaussiens de moyenne nulle et de covariance I_N tandis que $a_1, \ldots, a_{\varepsilon_n n} \in \mathbb{C}^N$ sont déterministes et telles que

$$\limsup_n \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} \frac{1}{N} a_i^* C_N^{-1} a_i < \infty.$$

Nous ferons également l'hypothèse ici que $\varepsilon_n \to \varepsilon \in [0, 1)$ et que $N/n = c_N \to c \in (0, 1-\varepsilon)$.

Idéalement, si un oracle pouvait prédire la position des données erronées, un estimateur optimal de C_N serait défini comme

$$\frac{1}{(1-\varepsilon_n)n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^*.$$

Dans cette section, nous allons évaluer la proximité entre $\hat{C}_N$ et cet estimateur optimal. Nous allons observer que la puissance de rejet de $\hat{C}_N$ est intimement lié au conditionnement de la matrice C_N.

Sans autre hypothèse, nous pouvons dès lors présenter notre résultat. Rappelons tout d'abord que $\hat{C}_N$ est définie ici comme l'unique solution de l'équation

$$\hat{C}_N = \frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} u\left(\frac{x_i^*\hat{C}_N^{-1}x_i}{N}\right) x_i x_i^* + \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} u\left(\frac{a_i^*\hat{C}_N^{-1}a_i}{N}\right) a_i a_i^*.$$

Nous avons alors le résultat central suivant, établi dans (Morales-Jimenez et al. [2015]).

Théorème 2. *Sous ces hypothèses, lorsque $N, n \to \infty$,*

$$\left\|\hat{C}_N - \hat{S}_N\right\| \xrightarrow{\text{p.s.}} 0$$

où

$$\hat{S}_N \triangleq \frac{1}{n}\sum_{i=1}^{(1-\varepsilon_n)n} v(\gamma_n)\, x_i x_i^* + \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} v(\alpha_{i,n})\, a_i a_i^*$$

avec γ_n et $\alpha_{1,n}, \ldots, \alpha_{\varepsilon_n n, n}$ les solutions uniques du système de $\varepsilon_n n + 1$ équations ($i = 1, \ldots, \varepsilon_n n$)

$$\gamma_n = \frac{1}{N} tr C_N \left(\frac{(1-\varepsilon)v(\gamma_n)}{1 + cv(\gamma_n)\gamma_n} C_N + \frac{1}{n}\sum_{i=1}^{\varepsilon_n n} v(\alpha_{i,n})\, a_i a_i^*\right)^{-1}$$

$$\alpha_{i,n} = \frac{1}{N} a_i^* \left(\frac{(1-\varepsilon)v(\gamma_n)}{1 + cv(\gamma_n)\gamma_n} C_N + \frac{1}{n}\sum_{j\neq i}^{\varepsilon_n n} v(\alpha_{j,n})\, a_j a_j^*\right)^{-1} a_i$$

et $v(x) = u\left(g^{-1}(x)\right)$, $g(x) = x/(1 - c\phi(x))$.

Ainsi, $\hat{C}_N$ est équivalent à la somme d'une matrice de covariance empirique pour les $x_i x_i^*$, pondérée par $v(\gamma_n)$, et de la somme pondérée par $v(\alpha_{i,n})$ des données erronées $a_i a_i^*$. Il est intéressant d'analyser les constantes $v(\alpha_{i,n})/v(\gamma_n)$

pour comprendre l'effet de $\hat{C}_N$ sur les données erronées. Le système implicite qui régit les γ_n et $\alpha_{i,n}$ est cependant peu interprétable et nous allons nous contenter de l'analyse de cas simplifiés.

Prenons tout d'abord $\varepsilon_n = 1/n$ (une seule donnée erronée). Alors, $\gamma_n \to \gamma$ où $\gamma = \frac{1+cv(\gamma)\gamma}{v(\gamma)}$ qui se résout explicitement en $\gamma = \frac{\phi^{-1}(1)}{1-c}$. Quant à $\alpha_{1,n}$, il est donné par

$$\alpha_{1,n} = \left(\frac{\phi^{-1}(1)}{1-c} + o(1) \right) \frac{1}{N} a_1^* C_N^{-1} a_1.$$

Ainsi, si $\liminf_N \frac{1}{N} a_1^* C_N^{-1} a_1 > 1$, nous avons $v(\alpha_{1,n})/v(\gamma_n) \leq 1$ pour tout n large et donc l'impact de a_1 sera amorti. Au contraire, si $\limsup_N \frac{1}{N} a_1^* C_N^{-1} a_1 < 1$, son impact sera renforcé. En particulier, à norme $\|a_1\|$ donnée, $\frac{1}{N} a_1^* C_N^{-1} a_1$ est d'autant plus grand (et donc a_1 d'autant plus atténué) que a_1 ne s'aligne pas aux espace-propres dominants de C_N.

Un autre cas pertinent est celui où les a_i sont gaussiens, centrés et de covariance $D_N \neq C_N$. Ce cas donne lieu au corollaire suivant.

Corollaire 1. *Sous les hypothèses ci-dessus, ainsi que sous la condition* $\limsup_N \|D_N\| < \infty$*, lorsque* $N, n \to \infty$*,*

$$\left\| \hat{C}_N - \hat{S}_N^{\text{rnd}} \right\| \xrightarrow{\text{p.s.}} 0$$

où

$$\hat{S}_N^{\text{rnd}} \triangleq \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} v(\gamma_n)\, x_i x_i^* + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v(\alpha_n)\, a_i a_i^*$$

avec γ_n *et* α_n *les solutions uniques de*

$$\gamma_n = \frac{1}{N} tr C_N \left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n} D_N \right)^{-1}$$

$$\alpha_n = \frac{1}{N} tr D_N \left(\frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n} D_N \right)^{-1}.$$

Comme précédemment, considérons que $\varepsilon_n \to \varepsilon = 0$, où $\gamma_n \to \gamma$ défini plus haut. Alors α_n est donné par

$$\alpha_n = \frac{\phi^{-1}(1)}{1-c} \frac{1}{N} tr D_N C_N^{-1}.$$

Le paramètre d'intérêt est ici la quantité $\frac{1}{N} tr D_N C_N^{-1}$. Les rôles de C_N et D_N ne sont pas symétriques. Si $C_N \simeq I_N$, le rejet des données erronées sera faible et surtout indépendant de D_N de trace donnée. A contrario, pour C_N mal conditionné, le pouvoir de rejet sera plus important. Pour illustrer cela, nous comparons en Figure 2 la loi limite des valeurs propres de $\hat{C}_N$ à celles de la matrice de covariance empirique $\frac{1}{n} XX^*$ et de la matrice de covariance empirique des bonnes données seules. Nous observons un net gain de performance de $\hat{C}_N$ comparativement à $\frac{1}{n} XX^*$, en particulier dans la queue de la distribution.

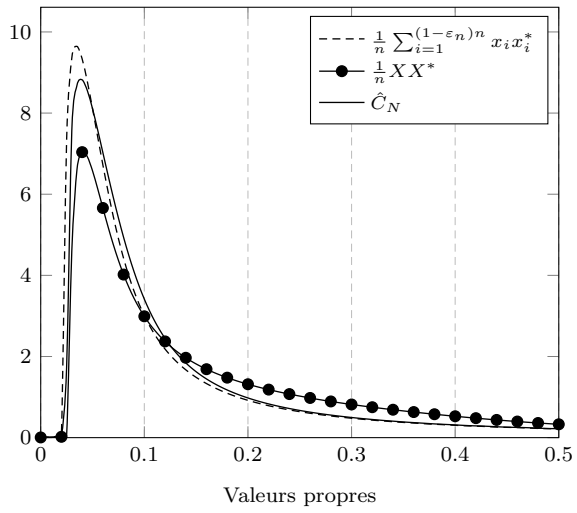

FIG. 2: Distribution limite des valeurs propres, pour $[C_N]_{ij} = .9^{|i-j|}$, $D_N = I_N$, $\varepsilon = .05$.

References

R. Couillet and W. Hachem. Analysis of the limit spectral measure of large random matrices of the separable covariance type. *Random Matrix Theory and Applications*, 3(4):1–23, 2014.

D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67, 1976.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math USSR-Sbornik*, 1(4):457–483, 1967.

D. Morales-Jimenez, R. Couillet, and M. McKay. Large dimensional analysis of Maronna's M-estimator with outliers. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*, Brisbane, Australia, 2015.

J. W. Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54 (2):175–192, 1995.

J. W. Silverstein and S. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.