FLUCTUATIONS FOR LINEAR SPECTRAL STATISTICS OF LARGE RANDOM COVARIANCE MATRICES

Jamal Najim and Jianfeng Yao

CNRS and Université Paris Est - Marne La Vallée

ABSTRACT

The theory of large random matrices has proved to be an efficient tool to address many problems in wireless communication and statistical signal processing these last two decades.

We provide hereafter a central limit theorem (CLT) for linear spectral statistics of large random covariance matrices, improving Bai and Silverstein's celebrated 2004 result. This fluctuation result should be of interest to study the fluctuations of important estimators in statistical signal processing.

Index Terms— Large random matrices fluctuations.

1. INTRODUCTION

The theory of large random matrices has proved to be an efficient tool to address many problems in wireless communication and statistical signal processing; one may refer to [1] for an updated overview. Among the many results of regular use, the Central limit theorem developed by Bai and Silverstein [2] is a generic tool to study the performances of important estimators in statistical signal processing; see for instance [3] where the fluctuations of the estimators developed by Mestre [4, 5] in the context of population eigenvalue estimation are studied with the help of this CLT.

The purpose of this communication is to present an extended version of this CLT where two important and limitating assumptions in [2] are removed; as an important consequence, changes occur in the limiting variance and bias formulas. All the mathematical details can be found in the extended version of this paper [6].

Consider a $N \times n$ random matrix $\Sigma_n = (\xi_{ij}^n)$ given by:

$$\Sigma_n = \frac{1}{\sqrt{n}} R_n^{1/2} X_n \,, \tag{1.1}$$

where $N = N(n)$ and R_n is a $N \times N$ nonnegative definite deterministic hermitian matrix. The entries $(X_{ij}^n \,; i \leq N, j \leq n, n \geq 1)$ of matrices (X_n) are real or complex, independent and identically distributed (i.i.d.) with mean 0 and variance 1. Matrix $\Sigma_n \Sigma_n^*$ models a sample covariance matrix, formed from n samples of the random vector $R_n^{1/2} X_{\cdot 1}^n$, with the population covariance matrix R_n.

Since the seminal work of Marčenko and Pastur [7], the study of the spectrum of large covariance matrices of the type $X_n X_n^*$ under the asymptotic regime where $N, n \to \infty$ and:

$$0 \,<\, \boldsymbol{\ell}^- \stackrel{\triangle}{=} \liminf \frac{N}{n} \,\leq\, \boldsymbol{\ell}^+ \stackrel{\triangle}{=} \limsup \frac{N}{n} \,<\, \infty \,, \tag{1.2}$$

(a condition that will be simply referred as $N, n \to \infty$ in the sequel) has drawn a considerable interest.

This asymptotic regime models the case where the dimension of the data (N) is of the same order as the size of the sample (n), a context of particular interest in modern statistical signal processing.

In this article, we study the fluctuations of linear spectral statistics of the form:

$$\operatorname{tr} f(\Sigma_n \Sigma_n^*) = \sum_{i=1}^{N} f(\lambda_i) \,, \qquad \text{as} \quad N, n \to \infty \tag{1.3}$$

where $\operatorname{tr}(A)$ refers to the trace of A and the λ_i's are the eigenvalues of $\Sigma_n \Sigma_n^*$. In their '04 article [2], Bai and Silverstein established a CLT for the linear spectral statistics (1.3) under two important assumptions: (1) The entries (X_{ij}^n) are centered with unit variance and a finite fourth moment equal to the fourth moment of a (real or complex) gaussian standard variable. (2) Function f in (1.3) is analytic in a neighbourhood of the asymptotic spectrum of $\Sigma_n \Sigma_n^*$. Such a result proved to be highly useful in probability theory, statistics, wireless communication, statistical signal processing and various other fields.

The purpose of this article is to present a CLT for linear spectral statistics (1.3) for general entries X_{ij}^n and for non-analytic functions f with sufficient derivatives, hence to relax both Assumptions (1) and (2) in [2].

Non-Gaussian entries. The presence of matrix R_n yields interesting phenomena when considering non-gaussian entries: terms proportionnals to the fourth cumulant and to $|\mathbb{E}(X_{11}^n)^2|^2$ appear in the variance but their convergence is not granted under usual assumptions (roughly, under the convergence of R_n's spectrum), mainly because these extra-terms also depend on R_n's eigenvectors. A careful description of the asymptotic variance is provided in Section 2.3.

Denote by $L_n(f)$ the (approximately) centered version of the linear statistics (1.3), to be properly defined below. Instead of expressing the CLT in the usual way, i.e. ($\xrightarrow{\mathcal{D}}$ stands

for the convergence in distribution):

$$L_n(f) \xrightarrow[N,n\to\infty]{\mathcal{D}} \mathcal{N}(\mathcal{B}_\infty, \Theta_\infty)\,, \tag{1.4}$$

for some well-defined parameters $\mathcal{B}_\infty, \Theta_\infty$, we establish that there exists a family of Gaussian random variables $\mathcal{N}(\mathcal{B}_n^f, \Theta_n^f)$, such that

$$d_{\mathcal{L}P}\left(L_n(f), \mathcal{N}(\mathcal{B}_n^f, \Theta_n^f)\right) \xrightarrow[N,n\to\infty]{} 0\,, \tag{1.5}$$

where $d_{\mathcal{L}P}$ denotes the Lévy-Prohorov distance (and in particular metrizes the convergence of laws).

This framework may also prove to be useful for other interesting models such as large dimensional information-plus-noise type matrices [8] and more generally mixed models combining large dimensional deterministic and random matrices.

Non-analytic functions. In Section 3, we first present the CLT for the trace of the resolvent $\mathrm{tr}\,(\Sigma_n\Sigma_n^* - zI_N)^{-1}$. In order to transfer the CLT from the resolvent to the linear statistics of the eigenvalues $\mathrm{tr}\, f(\Sigma_n\Sigma_n^*)$, we use Helffer-Sjöstrand's representation formula for a function f of class C^{k+1} [9]. Denote by $\Phi(f) : \mathbb{C}^+ \to \mathbb{C}$ the function:

$$\Phi(f)(x+\boldsymbol{i}y) = \sum_{\ell=0}^{k} \frac{(\boldsymbol{i}y)^\ell}{\ell!} f^{(\ell)}(x)\chi(y)\,, \tag{1.6}$$

where $\chi : \mathbb{R} \to \mathbb{R}^+$ is smooth, compactly supported, with value 1 in a neighbourhood of 0 and let $\overline{\partial} = \partial_x + \boldsymbol{i}\partial_y$. Helffer-Sjöstrand's formula writes:

$$\mathrm{tr}\, f(\Sigma_n\Sigma_n^*) = \frac{1}{\pi}\mathrm{Re}\int_{\mathbb{C}^+} \overline{\partial}\Phi(f)(z)\mathrm{tr}\,(\Sigma_n\Sigma_n^* - zI_N)^{-1}\,\lambda_2(dz)\,, \tag{1.7}$$

where λ_2 stands for the Lebesgue measure over $\mathbb{C}^+$.

Representation formula (1.7) enables us to transfer the CLT to linear statistics for functions of class C^3, a lower regularity requirement than in [10].

2. VARIANCE AND BIAS FORMULAS

2.1. Assumptions

Recall the asymptotic regime (1.2) and denote by $c_n = \frac{N}{n}$.

Assumption A-1. The random variables $(X_{ij}^n\ ;\ 1 \le i \le N(n),\ 1 \le j \le n\,,\ n \ge 1)$ are independent and identically distributed. They satisfy: $\mathbb{E}X_{ij}^n = 0$, $\mathbb{E}|X_{ij}^n|^2 = 1$ and $\mathbb{E}|X_{ij}^n|^4 < \infty$.

Assumption A-2. Consider a sequence (R_n) of deterministic, nonnegative definite hermitian $N \times N$ matrices, with $N = N(n)$. The sequence $(R_n, n \ge 1)$ is bounded for the spectral norm as $N, n \to \infty$, in particular:

$$0 \le \boldsymbol{\lambda}_R^- \triangleq \liminf_{N,n\to\infty} \|R_n\| \le \boldsymbol{\lambda}_R^+ \triangleq \limsup_{N,n\to\infty} \|R_n\| < \infty\,.$$

2.2. Resolvent, canonical equation and deterministic equivalents

Denote by $Q_n(z)$ and $\tilde{Q}_n$ the resolvents of $\Sigma_n\Sigma_n^*$ and $\Sigma_n^*\Sigma_n$:

$$Q_n(z) = (\Sigma_n\Sigma_n^* - zI_N)^{-1}\ ,\ \tilde{Q}_n(z) = (\Sigma_n^*\Sigma_n - zI_n)^{-1}\ , \tag{2.1}$$

and by $f_n(z)$ and $\tilde{f}_n(z)$ their normalized traces which are the Stieltjes transforms of the empirical distribution of $\Sigma_n\Sigma_n^*$'s and $\Sigma_n^*\Sigma_n$'s eigenvalues:

$$f_n(z) = \frac{1}{N}\mathrm{tr}\, Q_n(z)\,,\quad \tilde{f}_n(z) = \frac{1}{n}\mathrm{tr}\,\tilde{Q}_n(z)\,. \tag{2.2}$$

The following canonical equation admits a unique solution t_n in the class of Stieltjes transforms of probability measures:

$$t_n(z) = \frac{1}{N}\mathrm{tr}\ (-zI_N + (1-c_n)R_n - zc_nt_n(z)R_n)^{-1} \tag{2.3}$$

for $z \in \mathbb{C} \setminus \mathbb{R}^+$ where c_n stands for the ratio N/n (see for instance [2]). The function t_n being introduced, we can define the following $N \times N$ matrix

$$T_n(z) = (-zI_N + (1-c_n)R_n - zc_nt_n(z)R_n)^{-1}\ . \tag{2.4}$$

Matrix $T_n(z)$ can be thought of as a *deterministic equivalent* of the resolvent $Q_n(z)$ in the sense that it approximates the resolvent in various senses. For instance,

$$\frac{1}{N}\mathrm{tr}\, T_n(z) - \frac{1}{N}\mathrm{tr}\, Q_n(z) \xrightarrow[N,n\to\infty]{} 0$$

(in probability or almost surely). It is also true that

$$u_n^* Q_n v_n - u_n^* T_n v_n \xrightarrow[N,n\to\infty]{} 0 \tag{2.5}$$

where (u_n) and (v_n) are deterministic $N \times 1$ vectors with uniformly bounded euclidian norms in N. As a consequence of (2.5), not only T_n conveys information on the limiting spectrum of the resolvent Q_n but also on the eigenvectors of Q_n.

If $R_n = I_N$, then t_n is simply the Stieltjes transform of Marčenko-Pastur distribution [7] with parameter c_n.

2.3. Limiting covariance for the trace of the resolvent

In [2], the CLT for the trace of the resolvent is first studied. Let $\mathcal{V}$ be the second moment of the random variable X_{ij} and κ its fourth cumulant:

$$\mathcal{V} = \mathbb{E}(X_{ij}^n)^2 \quad \text{and} \quad \kappa = \mathbb{E}\left|X_{ij}^n\right|^4 - |\mathcal{V}|^2 - 2\,.$$

If the entries are real or complex standard Gaussian, then $\mathcal{V} = 1$ or 0 and $\kappa = 0$. Otherwise $\kappa \neq 0$ a priori and this induces extra-terms in the limiting variance, mainly due to the following $(\mathcal{V}, \kappa)$-dependent identity:

$$\begin{aligned}\mathbb{E}(X_{\cdot 1}^* A X_{\cdot 1} - \mathrm{tr}\, A)(X_{\cdot 1}^* B X_{\cdot 1} - \mathrm{tr}\, B)\\ = \mathrm{tr}\, AB + |\mathcal{V}|^2\ \mathrm{tr}\, AB^T + \kappa\sum_{i=1}^N A_{ii}B_{ii}\,,\end{aligned} \tag{2.6}$$

where $X_{.1}$ stands for the first column (of dimension $N \times 1$) of matrix X_n and where A, B are deterministic $N \times N$ matrices. As a consequence, there will be three terms in the limiting covariance of the quantity (1.3). Let:

$$\tilde{t}_n(z) \quad = \quad -\frac{1-c_n}{z} + c_n t_n(z) \, . \tag{2.7}$$

The quantity $\tilde{t}_n(z)$ is the deterministic equivalent associated to $n^{-1}\mathrm{tr}\,(\Sigma_n^* \Sigma_n - zI_n)^{-1}$. Denote by R_n^T the transpose matrix of R_n (notice that $R_n^T = \bar{R}_n$) and by T_n^T, the transpose matrix of T_n (beware that T_n^T is not the entry-wise conjugate of T_n, due to the presence of z):

$$T_n^T(z) = \left(-zI_N + (1-c_n)\bar{R}_n - zc_n t_n(z)\bar{R}_n\right)^{-1} \; ; \tag{2.8}$$

notice that the definition of $t_n(z)$ in (2.3) does not change if R_n is replaced by $\bar{R}_n$ since the spectrum of both matrices R_n and $\bar{R}_n$ is the same. We can now describe the limiting covariance of the trace of the resolvent:

$$\begin{aligned} &\mathrm{cov}\left(\mathrm{tr}\, Q_n(z_1), \mathrm{tr}\, Q_n(z_2)\right) \\ &\quad = \quad \Theta_{0,n}(z_1,z_2) + |\mathcal{V}|^2 \Theta_{1,n}(z_1,z_2) + \kappa \Theta_{2,n}(z_1,z_2) + o(1) \\ &\quad \triangleq \quad \Theta_n(z_1,z_2) + o(1) \, , \end{aligned} \tag{2.9}$$

where $o(1)$ is a term that converges to zero as $N, n \to \infty$ and

$$\Theta_{0,n}(z_1,z_2) \quad \triangleq \quad \left\{ \frac{\tilde{t}_n'(z_1)\tilde{t}_n'(z_2)}{(\tilde{t}_n(z_1) - \tilde{t}_n(z_2))^2} - \frac{1}{(z_1-z_2)^2} \right\} \tag{2.10}$$

$$\Theta_{1,n}(z_1,z_2) \quad \triangleq \quad \frac{\partial}{\partial z_2} \left\{ \frac{\partial \mathcal{A}_n(z_1,z_2)}{\partial z_1} \frac{1}{1-|\mathcal{V}|^2 \mathcal{A}_n(z_1,z_2)} \right\} \tag{2.11}$$

$$\begin{aligned} \Theta_{2,n}(z_1,z_2) \quad &\triangleq \quad \frac{z_1^2 z_2^2 \tilde{t}_n'(z_1)\tilde{t}_n'(z_2)}{n} \\ &\times \sum_{i=1}^N \left(R_n^{1/2} T_n^2(z_1) R_n^{1/2} \right)_{ii} \left(R_n^{1/2} T_n^2(z_2) R_n^{1/2} \right)_{ii} \end{aligned} \tag{2.12}$$

with

$$\begin{aligned} \mathcal{A}_n(z_1,z_2) \quad &= \quad \frac{z_1 z_2}{n} \tilde{t}_n(z_1)\tilde{t}_n(z_2) \\ &\quad \times \mathrm{tr}\left\{ R_n^{1/2} T_n(z_1) R_n^{1/2} \bar{R}_n^{1/2} T_n^T(z_2) \bar{R}_n^{1/2} \right\} \, . \end{aligned} \tag{2.13}$$

At first sight, these formulas may seem complicated; however, much information can be inferred from them.

The term $\Theta_{0,n}$. This term is familiar as it already appears in Bai and Silverstein's CLT [2]. Notice that the quantities $\tilde{t}_n$ and $\tilde{t}_n'$ only depend on the spectrum of matrix R_n. Hence, under the additional assumption that:

$$c_n \xrightarrow[N,n\to\infty]{} c \in (0,\infty) \quad \text{and} \quad F^{R_n} \xrightarrow[N,n\to\infty]{\mathcal{L}} F^{\mathbf{R}} \, , \tag{2.14}$$

where F^{R_n} denotes the empirical distribution of R_n's eigenvalues and $F^{\mathbf{R}}$ is a probability measure, it can easily be proved that as $N, n \to \infty$, the term $\Theta_{0,n}(z_1,z_2)$ converges to

$$\Theta_0(z_1,z_2) = \left\{ \frac{\tilde{t}'(z_1)\tilde{t}'(z_2)}{(\tilde{t}(z_1) - \tilde{t}(z_2))^2} - \frac{1}{(z_1-z_2)^2} \right\} \, ,$$

where $\tilde{t}, \tilde{t}'$ are the limits of $\tilde{t}_n, \tilde{t}_n'$ under (2.14).

The term $\Theta_{1,n}$. The interesting phenomenon lies in the fact that this term involves products of matrices $R_n^{1/2}$ and its conjugate $\bar{R}_n^{1/2}$. These matrices have the same spectrum but conjugate eigenvectors. If R_n is not real, the convergence of $\Theta_{1,n}$ is not granted, even under (2.14).

The term $\Theta_{2,n}$. This term involves quantities of the type $(R_n^{1/2} T_n R_n^{1/2})_{ii}$ which not only depend on the spectrum of matrix R_n but also on its eigenvectors. As a consequence, the convergence of such terms does not follow from an assumption such as (2.14), except in some particular cases.

2.4. Representation of the linear statistics and limiting bias

Recall that $t_n(z)$ is the Stieltjes transform of a probability measure $\mathcal{F}_n$:

$$t_n(z) = \int_{\mathcal{S}_n} \frac{\mathcal{F}_n(d\lambda)}{\lambda - z} \tag{2.15}$$

with support $\mathcal{S}_n$ included in a compact set. The purpose of this article is to describe the fluctuations of the linear statistics

$$L_n(f) \; = \; \sum_{i=1}^N f(\lambda_i) - N \int f(\lambda) \, \mathcal{F}_n(d\lambda) \tag{2.16}$$

$$= \quad \frac{1}{\pi} \mathrm{Re} \int_{\mathbb{C}^+} \overline{\partial} \Phi(f)(z) \left\{ \mathrm{tr}\, Q_n(z) - N t_n(z) \right\} \lambda_2(dz) \, . \tag{2.17}$$

as $N, n \to \infty$, and where the last equality follows form Helffer-Sjöstrand's formula and the fact that

$$\int f(\lambda) \mathcal{F}_n(d\lambda) = \frac{1}{\pi} \mathrm{Re} \int_{\mathbb{C}^+} \overline{\partial} \Phi(f)(z) t_n(z) \lambda_2(dz)$$

(recall that $\Phi(f)$ is defined in (1.6)).

Based on (2.17), we shall first study the fluctuations of:

$$\begin{aligned} &\mathrm{tr}\, Q_n(z) - N t_n(z) \\ &\quad = \left\{ \mathrm{tr}\, Q_n(z) - \mathbb{E}\mathrm{tr}\, Q_n(z) \right\} + \left\{ \mathbb{E}\mathrm{tr}\, Q_n(z) - N t_n(z) \right\} \end{aligned}$$

for $z \in \mathbb{C}^+$. The first difference in the r.h.s. will yield the fluctuations with a covariance $\Theta_n(z_1,z_2)$ described in (2.9) while the second difference, deterministic, will yield the bias:

$$\begin{aligned} \mathbb{E}\mathrm{tr}\, Q_n(z) - N t_n(z) \quad &= \quad |\mathcal{V}|^2 \mathcal{B}_{1,n}(z) + \kappa \mathcal{B}_{2,n}(z) + o(1) \\ &\triangleq \quad \mathcal{B}_n(z) + o(1) \end{aligned} \tag{2.18}$$

$$\mathcal{B}_{1,n}(z) \triangleq -z^3\tilde{t}_n^3 \frac{\frac{1}{n}\mathrm{tr}\, R_n^{1/2}T_n^2(z)R_n^{1/2}\bar{R}_n^{1/2}T_n^T(z)\bar{R}_n^{1/2}}{\left(1-z^2\tilde{t}_n^2\,\frac{1}{n}\mathrm{Tr}\, R_n^2T_n^2\right)\left(1-|\mathcal{V}|^2z^2\tilde{t}_n^2\,\frac{1}{n}\mathrm{Tr}\, R_n^{1/2}T_n(z)R_n^{1/2}\bar{R}_n^{1/2}T_n^T(z)\bar{R}_n^{1/2}\right)} \tag{2.19}$$

$$\mathcal{B}_{2,n}(z) \triangleq -z^3\tilde{t}_n^3 \frac{\frac{1}{n}\sum_{i=1}^N \left(R_n^{1/2}T_nR_n^{1/2}\right)_{ii}\left(R_n^{1/2}T_n^2R_n^{1/2}\right)_{ii}}{1-z^2\tilde{t}_n^2\,\frac{1}{n}\mathrm{tr}\, R_n^2T_n^2} \tag{2.20}$$

where $\mathcal{B}_{1,n}$ and $\mathcal{B}_{2,n}$ are defined in (2.19) and (2.20). The discussions on the terms $\Theta_{1,n}$ and $\Theta_{2,n}$ also apply to the terms $\mathcal{B}_{1,n}$ and $\mathcal{B}_{2,n}$ which are also likely not to converge.

2.5. Gaussian processes and the central limit theorem

Due to the a priori absence of convergence of $\mathcal{B}_n$ and Θ_n, we shall express the Gaussian fluctuations of the linear statistics (2.16) in the following way: consider a family $(N_n(z), z \in \mathcal{C})_{n\in\mathbb{N}}$ of tight Gaussian processes with mean and covariance:

$$\begin{aligned}\mathbb{E}N_n(z) &= \mathcal{B}_n(z)\,,\\ \mathrm{cov}(N_n(z_1),N_n(z_2)) &= \Theta_n(z_1,z_2)\,.\end{aligned}$$

We then express the fluctuations of the centralized trace as

$$d_{\mathcal{L}P}\left(\left(\mathrm{tr}\, Q_n(z)-Nt_n(z)\right),N_n(z)\right) \xrightarrow[N,n\to\infty]{} 0\,.$$

with $d_{\mathcal{L}P}$ the Lévy-Prohorov distance between P and Q probability measures over borel sets of $\mathbb{R},\mathbb{R}^d,\mathbb{C}$ or $\mathbb{C}^d$:

$$d_{\mathcal{L}P}(P,Q) = \inf\left\{\varepsilon>0,\ P(A)\le Q(A^\varepsilon)+\varepsilon\,,\right.$$
$$\left.\text{for all Borel sets } A\right\}\,, \tag{2.21}$$

where A^ε is an ε-blow up of A (cf. [11, Section 11.3] for more details). If X is a random variable and $\mathcal{L}(X)$ its distribution, denote similarly by $d_{\mathcal{L}P}(X,Y) \triangleq d_{\mathcal{L}P}(\mathcal{L}(X),\mathcal{L}(Y))$. We will express the fluctuations of $L_n(f)$ in the same way:

$$d_{\mathcal{L}P}\left(L_n(f),\mathcal{N}_n(f)\right) \xrightarrow[N,n\to\infty]{} 0\,,$$

where $\mathcal{N}_n(f)$ is a well-identified gaussian random variable.

3. STATEMENT OF THE CLT FOR THE TRACE OF THE RESOLVENT

3.1. Further notations

Recall the definition of $\mathcal{F}_n$ in (2.15) and let similarly $\tilde{\mathcal{F}}_n$ be the probability distribution associated to $\tilde{t}_n$. The central object of study is the Stieltjes transform

$$M_n(z) = N(f_n(z)-t_n(z)) = n\left(\tilde{f}_n(z)-\tilde{t}_n(z)\right)\,. \tag{3.1}$$

3.2. The Central Limit Theorem for the resolvent

Recall that $\mathcal{S}_n$ is the support of the probability measure $\mathcal{F}_n$. Due to Assumption (A-2), it is clear that

$$\mathcal{S}_n \subset \mathcal{S}_\infty \triangleq \left[0, \boldsymbol{\lambda}_R^+\left(1+\sqrt{\boldsymbol{\ell}^+}\right)^2\right]\,, \tag{3.2}$$

uniformily in n. Let A be large enough, say

$$A > \boldsymbol{\lambda}_R^+\left(1+\sqrt{\boldsymbol{\ell}^+}\right)^2\,.$$

Denote by D^+ and D_ε the domains:

$$D^+ = [0,A]+\boldsymbol{i}(0,1]\,,\ D_\varepsilon = [0,A]+\boldsymbol{i}[\varepsilon,1]\ (\varepsilon>0)\,. \tag{3.3}$$

Theorem 3.1. Assume that (A-1) and (A-2) hold true , then

1. There exists a sequence $(N_n(z),\ z\in D^+)$ of two-dimensional Gaussian processes with mean

$$\mathbb{E}N_n(z) = |\mathcal{V}|^2\mathcal{B}_{1,n}(z)+\kappa\mathcal{B}_{2,n}(z) \tag{3.4}$$

where $\mathcal{B}_{1,n}(z)$ and $\mathcal{B}_{2,n}(z)$ are defined in (2.19) and (2.20), and covariance:

$$\begin{aligned}&\mathrm{cov}\left(N_n(z_1),N_n(z_2)\right)\\ &= \mathbb{E}\left(N_n(z_1)-\mathbb{E}N_n(z_1)\right)\left(N_n(z_2)-\mathbb{E}N_n(z_2)\right)\\ &= \Theta_{0,n}(z_1,z_2)+|\mathcal{V}|^2\,\Theta_{1,n}(z_1,z_2)+\kappa\Theta_{2,n}(z_1,z_2)\,.\end{aligned}$$

Moreover $(N_n(z),\ z\in D_\varepsilon)$ is tight.

2. For any functional F from $C\left(D_\varepsilon;\mathbb{C}\right)$ to $\mathbb{C}$,

$$\mathbb{E}F(M_n)-\mathbb{E}F\left(N_n\right) \xrightarrow[N,n\to\infty]{} 0$$

Remark 3.1. Differences between Theorem 3.1 and [2, Lemma 1.1] appear in the bias and in the covariance where there are respectively two terms instead of one and three terms instead of one in [2, Lemma 1.1].

4. STATEMENT OF THE CLT FOR NON-ANALYTIC FUNCTIONALS

In order to lift the CLT from the trace of the resolvent to a smooth function f, the key ingredient is Helffer-Sjöstrand's

formula (1.7). We introduce the notations:

$$\begin{aligned} L_n^1(f) &= \operatorname{Tr} f(\Sigma_n\Sigma_n^*) - \mathbb{E}\operatorname{Tr} f(\Sigma_n\Sigma_n^*) \\ L_n^2(f) &= \mathbb{E}\operatorname{Tr} f(\Sigma_n\Sigma_n^*) - N\int f(\lambda)\mathcal{F}_n(d\lambda) \end{aligned}$$

Then $L_n(f) = L_n^1(f) + L_n^2(f)$. We first describe the fluctuations of $L_n^1(f)$ for non-analytic functions f in Section 4.1 and study the bias $L_n^2(f)$ in Section 4.2.

4.1. Fluctuations for the linear spectral statistics

Assumption A-3. Function $f : \mathbb{R} \to \mathbb{R}$ is measurable and C^3 in a η-neighborhood of $\mathcal{S}_\infty$ defined in (3.2).

Theorem 4.1. Assume that (A-1) and (A-2) hold true, and let $f_1, \cdots, f_k$ satisfy (A-3). Consider the centered Gaussian random vector $Z_n^1(\mathbf{f}) \stackrel{\triangle}{=} (Z_n^1(f_1), \cdots, Z_n^1(f_k))$ with covariance

$$\begin{aligned} &\operatorname{cov}\left(Z_n^1(f), Z_n^1(g)\right) = \\ &\frac{1}{2\pi^2}\operatorname{Re}\int_{D^2} \overline{\partial}\Phi(f)(z_1)\overline{\overline{\partial}\Phi(g)(z_2)}\Theta_n(z_1,\bar{z}_2)\lambda_2(dz_1)\lambda_2(dz_2) \\ &+\frac{1}{2\pi^2}\operatorname{Re}\int_{D^2} \overline{\partial}\Phi(f)(z_2)\overline{\partial}\Phi(g)(z_2)\Theta_n(z_1,z_2)\lambda_2(dz_1)\lambda_2(dz_2)\,, \end{aligned} \tag{4.1}$$

where $\Phi(f)$ and $\Phi(g)$ are defined as in (1.6). Then, the following convergence holds true:

$$d_{\mathcal{L}P}\left(L_n^1(\mathbf{f}), Z_n^1(\mathbf{f})\right) \xrightarrow[N,n\to\infty]{} 0\,,$$

or equivalently for every continuous bounded function $h : \mathbb{R}^k \to \mathbb{C}$,

$$\mathbb{E}\,h(L_n^1(\mathbf{f})) - \mathbb{E}\,h(Z_n^1(\mathbf{f})) \xrightarrow[N,n\to\infty]{} 0\,.$$

4.2. First-order expansions for the bias in the case of non-analytic functionals

Assumption A-4. Function $f : \mathbb{R} \to \mathbb{R}$ is measurable and C^{18} in a η-neighborhood of $\mathcal{S}_\infty$ defined in (3.2).

Theorem 4.2. Assume (A-1) and (A-2) hold true and let function f satisfy (A-4). Denote by

$$Z_n^2(f) = \frac{1}{\pi}\operatorname{Re}\int_D \overline{\partial}\Phi(f)(z)\mathcal{B}_n(z)\lambda_2(\,dz)\,,$$

where $\mathcal{B}_n$ is defined in (2.18). Then

$$\mathbb{E}\operatorname{Tr} f(\Sigma_n\Sigma_n^*) - N\int f(\lambda)\mathcal{F}_n(d\lambda) - Z_n^2(f) \xrightarrow[N,n\to\infty]{} 0\,.$$

REFERENCES

[1] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*, Cambridge University Press, 2011.

[2] Z. D. Bai and J. W. Silverstein, "CLT for linear spectral statistics of large-dimensional sample covariance matrices," *Ann. Probab.*, vol. 32, no. 1A, pp. 553–605, 2004.

[3] Jianfeng Yao, Romain Couillet, Jamal Najim, and Mérouane Debbah, "Fluctuations of an improved population eigenvalue estimator in sample covariance matrix models," *IEEE Trans. Inform. Theory*, vol. 59, no. 2, pp. 1149–1163, 2013.

[4] X. Mestre, "On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices," *Signal Processing, IEEE Transactions on*, vol. 56, no. 11, pp. 5353 –5368, nov. 2008.

[5] X. Mestre, "Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates," *Information Theory, IEEE Transactions on*, vol. 54, no. 11, pp. 5113 –5129, nov. 2008.

[6] J. Najim and Yao. J., "Gaussian fluctuations for linear spectral statistics of large random covariance matrices," Tech. Rep., `http://arxiv.org/abs/1309.3728`, 2013.

[7] V. A. Marcenko and L. A. Pastur, "Distribution of eigenvalues in certain sets of random matrices," *Mat. Sb. (N.S.)*, vol. 72 (114), pp. 507–536, 1967.

[8] R. Brent Dozier and J. W. Silverstein, "On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices," *J. Multivariate Anal.*, vol. 98, no. 4, pp. 678–694, 2007.

[9] B. Helffer and J. Sjöstrand, "Équation de Schrödinger avec champ magnétique et équation de Harper," in *Schrödinger operators (Sønderborg, 1988)*, vol. 345 of *Lecture Notes in Phys.*, pp. 118–197. Springer, Berlin, 1989.

[10] A. Lytova and L. Pastur, "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries," *Ann. Probab.*, vol. 37, no. 5, pp. 1778–1840, 2009.

[11] R. M. Dudley, *Real analysis and probability*, vol. 74 of *Cambridge Studies in Advanced Mathematics*, Cambridge University Press, Cambridge, 2002, Revised reprint of the 1989 original.