Asymptotic linear spectral statistics for spiked Hermitian random matrices

Damien Passemier[1] & Matthew R. McKay[2] & Yang Chen[3]

[1,2] *Department of Electronic and Computer Engineering*
Hong Kong University of Science and Technology (HKUST)
Clear Water Bay, Kowloon, Hong Kong
[1] *eepassemier@ust.hk* [2] *eemckay@ust.hk*

[3] *Faculty of Science and Technology, Department of Mathematics*
University of Macau, Avenue Padre Tomás Pereira, Taipa Macau, China
yangbrookchen@yahoo.co.uk

Résumé. Dans cet exposé, nous présentons des théorèmes centraux limites (CLT) pour statistiques spectrales linéaires (LSS) de trois ensembles « spikes » hermitiens de matrices aléatoires, en utilisant la méthode des fluides de Coulomb. Ces ensembles sont : le Wishart centré avec une variance hétérogène, le Wishart non-centré avec non-centralité de rang un, et une classe associée de matrices F non-centrales. Pour une LSS générale, nous montrons des CLT avec des expressions simples et explicites. Nous démontrons que l'effet du spike est d'introduire une correction en $O(1)$ à la moyenne asymptotique de la LSS, pour laquelle nous donnons des formules simples. Trois problèmes liés à des LSS illustrent l'utilité de ces résultats : le test du rapport de vraisemblance classique pour une matrice de covariance, l'analyse de la capacité dans un système de communication multi-antennes avec ligne de mire (LoS), et le test mutiple d'égalité de moyenne. Cet exposé est basé sur [8].

Mots-clés. Matrices aléatoires, grande dimension, modèle à variances isolées, ensembles de Wishart, F-matrix, systèmes MIMO, test d'hypothèses.

Abstract. Using the Coulomb Fluid method, this talk derives central limit theorems (CLTs) for linear spectral statistics of three "spiked" Hermitian random matrix ensembles. These include the central Wishart with spiked correlation, non-central Wishart with rank-one non-centrality, and a related class of non-central F matrices. For a generic linear statistic, we derive simple and explicit CLT expressions as the matrix dimensions grow large. We find that the primary effect of the spike is to introduce an $O(1)$ correction term to the asymptotic mean of the linear spectral statistic, which we characterize with simple formulas. The utility of our proposed framework is demonstrated through application to three different linear statistics problems: the classical likelihood ratio test for a population covariance, the capacity analysis of multi-antenna wireless communication systems with a line-of-sight transmission path, and a classical multiple sample significance testing problem. This talk is based on [8].

Keywords. Random matrix theory, high-dimensional statistics, spiked population model, Whisart ensembles, F-matrix, MIMO systems, hypothesis testing.

1 Matrix Models and Eigenvalue Distributions

We consider the following three "spiked" random matrix models:

- *Model A: Spiked central Wishart:*
  Matrices with distribution $\mathbb{C}\mathcal{W}_n(m, \mathbf{\Sigma}, \mathbf{0}_{n\times n})$ $(m \geq n)$, where $\mathbf{\Sigma}$ has one "spike" eigenvalue equal to $1+\delta$ with $\delta \geq 0$, and all other eigenvalues equal to 1.
- *Model B: Spiked non-central Wishart:*
  Matrices with distribution $\mathbb{C}\mathcal{W}_n(m, \mathbf{I}_n, \mathbf{\Theta})$ $(m \geq n)$, where $\mathbf{\Theta}$ is rank 1 (or zero) with "spike" eigenvalue $n\nu$ for $\nu \geq 0$.
- *Model C: Spiked multivariate F:*
  Matrices of the form $\mathbf{F} = \mathbf{W}_1\mathbf{W}_2^{-1}$, where $\mathbf{W}_1 \sim \mathbb{C}\mathcal{W}_n(m_1, \mathbf{\Sigma}, \mathbf{\Theta})$ $(m_1 > n)$, $\mathbf{W}_2 \sim \mathbb{C}\mathcal{W}_n(m_2, \mathbf{\Sigma}, \mathbf{0}_{n\times n})$ $(m_2 > n)$ are independent, with $\mathbf{\Theta}$ rank 1 (or zero) having "spike" eigenvalue $n\nu$ for $\nu \geq 0$.

For these three models, expressions for the joint probability density functions of the eigenvalues x_k, $1 \leq k \leq n$ (taken in the following to be unordered) are well-known in various forms. Quite recently, however, it has been discovered that for Models A and B, the eigenvalue densities admit a particularly convenient contour integral representation

$$\frac{K_n[l]}{2\pi\imath}\oint_C l(z)\prod_{1\leq j<k\leq n}(x_k-x_j)^2\prod_{j=1}^{n}\frac{x_j^{m-n}e^{-x_j}}{z-x_j}\mathrm{d}z, \tag{1}$$

for $x_j \in (0,\infty), 1 \leq j \leq n$, where $K_n[l]$ is a normalization constant, and the contour C encloses $x_1, \ldots, x_n$ in its interior. The function $l(x)$ captures the effect of the spiked eigenvalue and is given by

$$l(z) = \begin{cases} \exp\left(\frac{\delta}{1+\delta}z\right), & \text{for Model A} \\ {}_0F_1(m-n+1, n\nu z), & \text{for Model B} \end{cases} \tag{2}$$

where ${}_pF_q(\cdot)$ represents a hypergeometric function. For Model C, it turns out that an analogous representation also exists:

Lemma 1. *Under Model C, let $x_j \in (0,\infty), 1 \leq j \leq n$ denote the eigenvalues of $\mathbf{F}$. Then, the joint density of $\mathsf{f}_j = x_j/(1+x_j) \in (0,1)$, $1 \leq j \leq n$ has the form*

$$K_n\oint_C {}_1F_1\left(m_1+m_2-n+1, m_1-n+1, \nu z\right)\prod_{j=1}^{n}\frac{\mathsf{f}_j^{m_1-n}(1-\mathsf{f}_j)^{m_2-n}}{z-\mathsf{f}_j}\prod_{1\leq j<k\leq n}(\mathsf{f}_k-\mathsf{f}_j)^2\,\mathrm{d}z,$$

where K_n is a normalization constant, and the contour C encloses $\mathsf{f}_1, \ldots, \mathsf{f}_n$ in its interior.

Based on these results, in the following we will compute the asymptotic distribution of linear spectral statistics for each of the three matrix models. In taking asymptotic, for Models A and B, we will be concerned with the following limits:

Assumption 1. $m, n \to \infty$ such that $m/n \to c \geq 1$.

For Model C, we will be concerned with:

Assumption 2. $m_1, m_2, n \to \infty$ such that $m_1/n \to c_1 \geq 1$ and $m_2/n \to c_2 \geq 1$.

2 Main Results

Theorem 1. *Consider Models A and B. Under Assumption 1, for f real and analytic,*

$$\sum_{k=1}^{n} f\left(\frac{x_k}{n}\right) - n\mu \xrightarrow{\mathcal{L}} \mathcal{N}\left(\bar{\mu}(z_0), \sigma^2\right), \tag{3}$$

where

$$\mu = \frac{1}{2\pi}\int_a^b f(x)\frac{\sqrt{(b-x)(x-a)}}{x}\,\mathrm{d}x \tag{4}$$

$$\sigma^2 = \frac{1}{2\pi^2}\int_a^b \frac{f(x)}{\sqrt{(b-x)(x-a)}}\left[\mathcal{P}\int_a^b \frac{f'(y)\sqrt{(b-y)(y-a)}}{x-y}\,\mathrm{d}y\right]\mathrm{d}x \tag{5}$$

with $a = (1-\sqrt{c})^2$, $b = (1+\sqrt{c})^2$. The spike-dependent term $\bar{\mu}(z_0)$ admits

$$\bar{\mu}(z_0) = \frac{1}{2\pi}\int_a^b \frac{f(x)}{\sqrt{(b-x)(x-a)}}\left(\frac{\sqrt{(z_0-a)(z_0-b)}}{z_0-x} - 1\right)\mathrm{d}x \tag{6}$$

where $z_0 = \frac{(1+c\delta)(1+\delta)}{\delta}$ for Model A, and $z_0 = \frac{(1+\nu)(c+\nu)}{\nu}$ for Model B.

Theorem 2. *Consider Model C. Under Assumption 2, for f real and analytic,*

$$\sum_{k=1}^{n} f\left(x_k\right) - n\mu_F \xrightarrow{\mathcal{L}} \mathcal{N}\left(\bar{\mu}_F(z_0), \sigma_F^2\right) \tag{7}$$

where

$$\mu_F = \frac{c_1+c_2}{2\pi}\int_a^b f\left(\frac{x}{1-x}\right)\frac{\sqrt{(b-x)(x-a)}}{x(1-x)}\,\mathrm{d}x \tag{8}$$

$$\sigma_F^2 = \frac{1}{2\pi^2}\int_a^b \frac{f\left(\frac{x}{1-x}\right)}{\sqrt{(b-x)(x-a)}}\left[\mathcal{P}\int_a^b \frac{f'\left(\frac{y}{1-y}\right)\sqrt{(b-y)(y-a)}}{x-y}\,\mathrm{d}y\right]\mathrm{d}x \tag{9}$$

where $a = \frac{c_1(c_1+c_2-1)+c_2-2\sqrt{c_1c_2(c_1+c_2-1)}}{(c_1+c_2)^2}$, $b = \frac{c_1(c_1+c_2-1)+c_2+2\sqrt{c_1c_2(c_1+c_2-1)}}{(c_1+c_2)^2}$. Let $z_0 = \frac{(1+\nu)(c_1+\nu)}{\nu(c_1+c_2+\nu)}$. The spike-dependent term $\bar{\mu}_F(z_0)$ admits

$$\bar{\mu}_F(z_0) = \frac{1}{2\pi}\int_a^b \frac{f\left(\frac{x}{1-x}\right)}{\sqrt{(b-x)(x-a)}}\left(\frac{\sqrt{(z_0-a)(z_0-b)}}{z_0-x} - 1\right)\mathrm{d}x. \tag{10}$$

These theorem are proven using the framework in [5], derived based on Dyson's Coulomb fluid interpretation, and using saddle-point approximations.

These results show that the asymptotic contribution coming from the spiked eigenvalue contributes to a $O(1)$ correction to the mean of the linear statistic. In the absence of a spike, such $O(1)$ terms disappear, which is consistent with prior results in [5].

3 Some Example Applications

In this section, to illustrate the utility of our main results, we consider a specific application of relevance for each of the three random matrix models. For Model A, we will reproduce a known result, whilst for Models B and C we will present results which appear new.

3.1 Model A: LRT Test of $\mathbf{\Sigma} = \mathbf{I}_n$

As an application of Theorem 1 for Model A, we consider the classical likelihood ratio test (LRT) that the population covariance matrix is the identity, under a rank-one spiked population alternative. Consider the m samples $\mathbf{y}_1, \ldots, \mathbf{y}_m$, drawn from a n-dimensional complex Gaussian distribution with covariance matrix $\mathbf{\Sigma}$. We aim to test the hypothesis:

$$H_0 : \mathbf{\Sigma} = \mathbf{I}_n.$$

Denoting the sample covariance by $\mathbf{S}_m = m^{-1} \sum_{k=1}^{m} \mathbf{y}_k \mathbf{y}_k^\dagger$, the LRT is based on the linear statistic:

$$\mathrm{L} = \mathrm{tr}(\mathbf{S}_m) - \ln(\det \mathbf{S}_m) - n.$$

Under H_0, and in high-dimension, this test was presented in [2] using a CLT framework established in [4]. Under H_1: "$\mathbf{\Sigma}$ has a spiked covariance structure as in Model A", this problem was addressed only very recently in the independent works [7] and [9]. The result in [9] was again based on the CLT framework of [4], with their derivation requiring lengthy calculations involving contour integrals. The same result was presented in [7], in this case making use of sophisticated tools of contiguity and Lecam's first and third lemmas.

Here, we adopt our general framework to recover the same result as [7] and [9] very efficiently, simply by calculating a few integrals. Under H_1, as before we denote by $1 + \delta$ the spiked eigenvalue of $\mathbf{\Sigma}$. Since $m\mathbf{S}_m \sim \mathbb{C}\mathcal{W}_n(m, \mathbf{\Sigma}, \mathbf{0}_{n\times n})$, we now apply Theorem 1 for the case of Model A to the function

$$f_{\mathrm{L}}(x) = \frac{x}{c} - \ln\left(\frac{x}{c}\right) - 1.$$

Let x_k, $1 \leq k \leq n$, be the eigenvalues of $m\mathbf{S}_m$. Then, under Assumption 1,

$$\mathrm{L} = \sum_{k=1}^{n} f_{\mathrm{L}}\left(\frac{x_k}{n}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(n\left[1 + (c-1)\ln\left(1 - c^{-1}\right)\right] + \bar{\mu}_{\mathrm{L}}, -c^{-1} - \ln\left(1 - c^{-1}\right)\right),$$

with the spike-dependent term $\bar{\mu}_{\mathrm{L}} = \delta - \ln(1 + \delta)$.

3.2 Model B: Capacity of MIMO Systems with Line-of-Sight

We consider a MIMO wireless communication system with n_t transmit and n_r receive antennas. The linear model relating the transmitted signal vector $\mathbf{x}$ of size n_t and received signal vector $\mathbf{y}$ of size n_r takes the form

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n},$$

where $\mathbf{n}$ is a complex Gaussian vector of size n_t with zero mean and $\mathbb{E}(\mathbf{n}\mathbf{n}^\dagger) = \mathbf{I}_{n_r}$. The $n_r \times n_t$ matrix $\mathbf{H}$ represents the wireless fading coefficients. We consider a communication scenario in which there is a direct line-of-sight (LoS) path between the transmitter and the receiver, thus

$$\mathbf{H} = \sqrt{\frac{K}{K+1}}\mathbf{M} + \sqrt{\frac{1}{K+1}}\mathbf{H}_w,$$

where $\mathbf{H}_w$ is an i.i.d. matrix with zero mean, unit variance complex Gaussian entries, $\mathbf{M}$ is deterministic such that $\mathrm{tr}(\mathbf{M}\mathbf{M}^\dagger) = n_r n_t$ and $K < \infty$. We assume that the sole

non-null eigenvalue of $\mathbf{M}^\dagger\mathbf{M}$ is $n_r n_t$, and that $K = K_0/\max(n_r, n_t)$ for fixed K_0. The transmitted signals obey $\mathbf{x} \sim \mathbb{C}\mathcal{N}\left(\mathbf{0}_{n_t\times 1}, \frac{P}{n_t}\mathbf{I}_{n_t}\right)$, where P is the total transmit power. In terms of performance evaluation, we are interested in the distribution of the quantity:

$$\mathrm{C} = \ln\det\left(\mathbf{I}_{n_r} + \frac{P}{n_t}\mathbf{H}\mathbf{H}^\dagger\right).$$

The asymptotic distribution of this quantity has been studied in e.g. [6], considering large numbers of antennas, but such results were not explicit. Here, we will find an explicit expression which, to the best of our knowledge, is new.

Let $m = \max\{n_r, n_t\}$, $n = \min\{n_r, n_t\}$ and define $\mathbf{W} = (K+1)\mathbf{H}\mathbf{H}^\dagger$ when $n_r < n_t$ and $\mathbf{W} = (K+1)\mathbf{H}^\dagger\mathbf{H}$ when $n_r \geq n_t$. We have $\mathbf{W} \sim \mathbb{C}\mathcal{W}_n(m, \mathbf{\Sigma}, \mathbf{\Theta})$, where the sole non-null eigenvalue of $\mathbf{\Theta}$ is $Knm = K_0 n$. Thus, in accordance with Model B, we set $\nu = K_0$. We will apply Theorem 1 for the case of Model B with the function

$$f_{\mathrm{C}}(x) = \ln\left(1 + \frac{x}{T}\right), \quad T = \frac{n_t}{nP}\left(\frac{K_0}{m} + 1\right).$$

Let $(x_k)_{1\leq k\leq n}$, be the eigenvalues of $\mathbf{W}$. Then, under Assumption 1,

$$\mathrm{C} = \sum_{k=1}^{n} f_{\mathrm{C}}\left(\frac{x_k}{n}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(n\mu_{\mathrm{C}} + \bar{\mu}_{\mathrm{C}}, \sigma_{\mathrm{C}}^2\right) \tag{11}$$

where the mean and the variance are explicitly calculated. We perform some simulation experiments to illustrate the previous results.

3.3 Model C: High-Dimensional Multiple Sample Significance Test

As an application of Theorem 2 for Model C, we consider the multiple significance test problem: consider q Gaussian populations $\mathbb{C}\mathcal{N}(\mathbf{u}^{(j)}, \mathbf{\Sigma})$ of dimension n and that each population has a sample of size p_j: $\mathbf{y}_1^{(j)}, \ldots, \mathbf{y}_{p_j}^{(j)}$. We aim to test the hypothesis

$$H_0 : \mathbf{u}^{(1)} = \cdots = \mathbf{u}^{(q)} = \mathbf{0}_{n\times 1}.$$

Define $p = \sum_{j=1}^{q} p_j$. The likelihood ratio statistic can be written as

$$\mathbf{\Lambda} = \det\left(\mathbf{I}_n + \mathbf{F}\right)^{-1}, \tag{12}$$

with $\mathbf{F} = \hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{B}}\mathbf{A}\hat{\mathbf{B}}^\dagger$, $\hat{\mathbf{B}} = \left(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(q)}, \ldots, \bar{\mathbf{y}}^{(q-1)} - \bar{\mathbf{y}}^{(q)}\right)$, $\bar{\mathbf{y}}^{(j)} = p_j^{-1}\sum_{k=1}^{p_j}\mathbf{y}_k^{(j)}$ is the empirical mean of each population, $\hat{\mathbf{\Sigma}} = \sum_{j,k}(\mathbf{y}_k^{(j)} - \bar{\mathbf{y}}_k^{(j)})(\mathbf{y}_k^{(j)} - \bar{\mathbf{y}}_k^{(j)})^\dagger$ and $\mathbf{A} = (a_{ij})_{1\leq i,j\leq q-1}$ is symmetric, with $a_{ii} = p_i(p-p_i)/p$ and $(a_{ij})_{i<j} = -p_i p_j/p$. As seen from [1], $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{B}}\mathbf{A}\hat{\mathbf{B}}^\dagger$ are independent, and

$$\hat{\mathbf{\Sigma}} \sim \mathbb{C}\mathcal{W}_n(p-q, \mathbf{\Sigma}, \mathbf{0}_{(p-q)\times(p-q)}), \qquad \hat{\mathbf{B}}\mathbf{A}\hat{\mathbf{B}}^\dagger \sim \mathbb{C}\mathcal{W}_n(q-1, \mathbf{\Sigma}, \mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{A}\mathbf{B}^\dagger), \tag{13}$$

where $\mathbf{B} = \left(\mathbf{u}^{(1)} - \mathbf{u}^{(q)}, \ldots, \mathbf{u}^{(q-1)} - \mathbf{u}^{(q)}\right)$. Under H_0 and Assumption 2, the distribution of $-\ln\mathbf{\Lambda}$ has been calculated in [3]. Here, we will find the distribution of $-\ln\mathbf{\Lambda}$ under the specific alternative

$$H_1 : \mathbf{u}^{(1)} \neq 0 \text{ and } \mathbf{u}^{(2)} = \cdots = \mathbf{u}^{(q)} = \mathbf{0}_{n\times 1}. \tag{14}$$

The result which we present in the following is new, and will permit the calculation of the asymptotic power of this test under the alternative (14).

Let $\mathbf{u}^{(1)} = (u_1, \ldots, u_n)^T$. Under H_1 above, the non-centrality matrix $\mathbf{\Sigma}^{-1}\mathbf{BAB}^\dagger$ in (13) has only one non-null eigenvalue $\nu = \xi(p_1 - p_1^2/p)\sum_{k=1}^n u_k^2$, where ξ is the top-left entry of $\mathbf{\Sigma}^{-1}$. We set $m_1 = q-1$, $m_2 = p-q$ and assume that $q > n+1$. For consistency with Model C, we will assume that p_1 is fixed, and either $\sum_{k=1}^n u_k^2 = K_1 n$ and $\xi = K_2$, or $\sum_{k=1}^n u_k^2 = K_1$ and $\xi = K_2 n$, where K_1, $K_2 > 0$. We can derive an explicit asymptotic characterization of the statistic $-\ln \mathbf{\Lambda}$ (12) by applying Theorem 2 with the function

$$f_{\mathrm{R}}(x) = \ln(1+x).$$

Let $(x_k)_{1 \leq k \leq n}$ be the eigenvalues of $\mathbf{F}$. Then, under Assumption 2,

$$-\ln \mathbf{\Lambda} = \sum_{k=1}^{n} f_{\mathrm{R}}\left(x_k\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(n\mu_{\mathrm{R}} + \bar{\mu}_{\mathrm{R}}, \sigma_{\mathrm{R}}^2\right), \tag{15}$$

where the mean and the variance are explicitly calculated. We perform some simulation experiments to illustrate the previous results.

References

[1] Anderson, T.W. (2003), *An introduction to multivariate statistical analysis*, Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.

[2] Bai, Z. D., Jiang, D., Yao, J.-F. and Zheng, S. (2009), *Corrections to LRT on large-dimensional covariance matrix by RMT*, Annals of Statistics, 37(6B), 3822–3840.

[3] Bai, Z. D., Yao, J.-F. and Zheng, S. (2013), *Testing linear hypotheses in high-dimensional regressions*, Statistics, 47(6), 1207–1223.

[4] Bai, Z. D. and Silverstein, J. W. (2004), *CLT for linear spectral statistics of large-dimensional sample covariance matrices*, Annals of Probability, 32(1A), 553–605.

[5] Chen, Y. and Lawrence, N. (1998), *On the linear statistics of Hermitian random matrices*, Journal of Physics. A. Mathematical and General, 31(4), 1141–1152.

[6] Kammoun, A., Kharouf, M., Hachem, W., Najim, J. and El Kharroubi, A. (2010), *On the fluctuations of the mutual information for non centered MIMO channels: The non gaussian case*, SPAWC Workshop.

[7] Onatski, A., Moreira, M.J. and Hallin, M (2013), *Asymptotic power of sphericity tests for high-dimensional data*, Annals of Statistics, 41(3), 1204–1231.

[8] Passemier, D., McKay, M. R. and Chen, Y. (2014), *Asymptotic Linear Spectral Statistics for Spiked Hermitian Random Matrix Models*, Preprint arXiv:1402.6419.

[9] Wang, Q., Silverstein, J. W. and Yao, J.-F. (2013), *A note on the CLT of the LSS for sample covariance matrix from a spiked population model*, Preprint arXiv:1304.6164.