EDGEWORTH CORRECTION FOR THE LARGEST EIGENVALUE IN A SPIKED PCA MODEL : Supplementary Materials

Jeha Yang and Iain Johnstone

Stanford University

S1. Identities

S1.1. Expectations with respect to Marchenko-Pastur distribution

Let $\gamma \in (0, \infty), \ell > 1 + \sqrt{\gamma}$ and $\rho_n = \ell + \ell\gamma_n/(\ell - 1)$. Then

$$\frac{\partial \rho_n}{\partial \ell} = \frac{(\ell - 1)^2 - \gamma_n}{(\ell - 1)^2}. \tag{S1.1}$$

Also, the Stieltjes transform of the companion Marchenko-Pastur distribution is given by

$$\mathsf{F}_\gamma(f_z) = (-z + \gamma - 1 + \sqrt{(z - \gamma - 1)^2 - 4\gamma})/(2z), \quad \forall z \in (b(\gamma), +\infty)$$

where $f_z(\lambda) := (\lambda - z)^{-1}$, from equantion (2.8) of Yao, Zheng, and Bai (2015). Substituting γ_n into γ and ρ_n into z(which is possible since $\rho_n > (1 + \sqrt{\gamma_n})^2$), it follows that

$$\mathsf{F}_{\gamma_n}(g_n) = \ell^{-1}. \tag{S1.2}$$

Taking partial derivatives of (S1.2) with respect to ℓ, along with (S1.1), gives

$$\mathsf{F}_{\gamma_n}(g_n^2) = (1 - \ell^{-1})^2((\ell - 1)^2 - \gamma_n)^{-1} = 2\sigma_n^{-2} \tag{S1.3}$$

and

$$\mathsf{F}_{\gamma_n}(g_n^3) = (1 - \ell^{-1})^3((\ell - 1)^3 + \gamma_n)((\ell - 1)^2 - \gamma_n)^{-3}, \tag{S1.4}$$

as desired.

S1.2. Explicit expressions of $\mu(g)$ and $\mu(g_n)$

We use the formula (5.13) in Bai and Silverstein (2004). First, by $x = 1+\gamma+2\sqrt{\gamma}\cos\theta$,

$$\begin{aligned}\int_{a(\gamma)}^{b(\gamma)} \frac{g(x)}{\sqrt{4\gamma-(x-1-\gamma)^2}}dx &= \int_{-\pi}^{0} \frac{g(1+\gamma+2\sqrt{\gamma}\cos\theta)}{\sqrt{1-\cos^2\theta}}(-\sin\theta)d\theta \\ &= \frac{1}{2}\int_{-\pi}^{\pi} g(1+\gamma+2\sqrt{\gamma}\cos\theta)d\theta.\end{aligned}$$

Then, letting $z = \exp(i\theta)$ gives

$$\begin{aligned}\int_{-\pi}^{\pi} g(1+\gamma+2\sqrt{\gamma}\cos\theta)d\theta &= \oint_{|z|=1} g(1+\gamma+\sqrt{\gamma}(z+z^{-1}))(iz)^{-1}dz \\ &= i\oint_{|z|=1} (\sqrt{\gamma}z^2-(\ell-1+\gamma(\ell-1)^{-1})z+\sqrt{\gamma})^{-1}dz \\ &= i\oint_{|z|=1} (z-\sqrt{\gamma}(\ell-1)^{-1})^{-1}\left(\sqrt{\gamma}z-(\ell-1)\right)^{-1}dz \\ &= -2\pi(\gamma(\ell-1)^{-1}-(\ell-1))^{-1} \\ &= 2\pi(\ell-1)\left(\ell-1-\sqrt{\gamma}\right)^{-1}\left(\ell-1+\sqrt{\gamma}\right)^{-1}\end{aligned}$$

by Cauchy integral formula with the assumption $\ell-1>\sqrt{\gamma}$. Meanwhile, $g((1\pm\sqrt{\gamma})^2) = (\rho(\ell,\gamma)-(1\pm\sqrt{\gamma})^2)^{-1} = (\ell-1)\left(\ell-1\mp\sqrt{\gamma}\right)^{-2}$, hence

$$\mu(g) = (\ell-1)((\ell-1-\sqrt{\gamma})^{-1}-(\ell-1+\sqrt{\gamma})^{-1})^2/4 = \gamma(\ell-1)((\ell-1)^2-\gamma)^{-2},$$

as desired. The corresponding expression for $\mu(g_n)$

$$\mu(g_n) = \gamma_n(\ell-1)((\ell-1)^2-\gamma_n)^{-2}, \tag{S1.5}$$

is available when $\ell-1>\sqrt{\gamma_n}$ i.e. for large enough n.

Remark. Although the formula (5.13) is derived only for $\gamma\le 1$ in Bai and Silverstein (2004), the following identity

$$G_n(f) = \sum_{i=1}^{p} f(\lambda_i) - pF_{\gamma_n}(f) = \sum_{i=1}^{n} \tilde{f}_n(\tilde{\lambda}_i) - nF_{\gamma_n^{-1}}(\tilde{f}_n) =: \mathsf{G}_p(\tilde{f}_n),$$

where $\tilde{f}_n(\lambda) := f(\gamma_n \lambda)$ and $\tilde{\lambda}_i := \gamma_n^{-1}\lambda_i$, turns the setting

$$n, \quad p, \quad \gamma_n, \quad n^{-1}Z_2'Z_2, \quad f$$

into

$$p, \quad n, \quad \gamma_n^{-1}, \quad p^{-1}Z_2Z_2', \quad \tilde{f}_n.$$

Thus, along with Lemma 7(which is proved below), this correspondence gives the same formula for $\gamma > 1$.

S2. Tail bounds propositions

S2.1. Proposition 5

We can prove and use results in Example 2.4 of Wainwright (2015) : the moment generating function of $(z_0^2 - 1)$ where $z_0 \sim N(0,1)$ is given by

$$\mathbb{E}\left[\exp(\theta(z_0^2-1))\right] = \exp(-\theta)(1-2\theta)^{-1/2} = \exp(\sum_{k=2}^{\infty} 2^{k-1}\theta^k/k)$$

for $\theta < 1/2$, and is bounded by $\exp(2\theta^2)$ for $\theta \in [-1/4, 1/4]$, because

$$2\theta^2 - \sum_{k=2}^{\infty} 2^{k-1}\theta^k/k = \theta^2(1 - \sum_{k=3}^{\infty} 2^{k-1}\theta^{k-2}/k) \geq \theta^2(1 - \sum_{k=3}^{\infty} 2^{-k+3}/k) = \theta^2(6 - 8\log 2) > 0.$$

Combining this, Markov inequality and independence of z and Λ, it follows that

$$\begin{aligned}\mathbb{P}(E_{1,n}^c \cap E_{2,n}(f,M) \mid \Lambda) &\leq \mathbb{E}\left[I(E_{1,n}^c)\left(\exp(S_n(f/U_f)) + \exp(-S_n(f/U_f))\right) \mid \Lambda\right]\exp(-M/U_f) \\ &\leq 2\exp(2\mathsf{F}_n(f^2/U_f^2))\exp(-M/U_f) \leq 15\exp(-M/U_f),\end{aligned}$$

which directly implies $\mathbb{P}(E_{1,n}^c \cap E_{2,n}(f,M)) \leq 15\exp(-M/U_f)$, as desired.

S2.2. Proposition 6

Let $\mathsf{f}_n(\lambda) := f_n\left((\lambda \vee 0) \wedge (b(\gamma) + \delta)\right)$, so that $\mathsf{f}_n(x^2), n \in \mathbb{N}$ share a Lipschitz constant L, and $G_n(f_n) = G_n(\mathsf{f}_n)$ on $E_{1,n}^c$ for all $n \in \mathbb{N}$. Hence, $\mathbb{P}(E_{1,n}^c \cap E_{3,n}(f_n, M)) \leq \mathbb{P}(E_{3,n}(\mathsf{f}_n, M))$.

Meanwhile, we have

$$\mathbb{P}\left(|p(F_n(\mathsf{f}_n) - \mathbb{E}\left[F_n(\mathsf{f}_n)\right])| > M\right) \leq 2\exp(-M^2/(2L^2))$$

for $M > 0, n \in \mathbb{N}$ from the Corollary 1.8 of Guionnet and Zeitouni (2000)(or Lemma A.4 of Paul (2007)). For all $p \geq 1$, from the identity $\mathbb{E}\left[|X|^p\right] = p\int_0^\infty y^{p-1}\mathbb{P}\left(|X| > y\right)dy$, it follows that $\{p(F_n(\mathsf{f}) - \mathbb{E}\left[F_n(\mathsf{f})\right])\}_{n\in\mathbb{N}}$ is bounded in L_p. i.e. is uniformly integrable and thus tight. But we assume that $\{G_n(f_n)\}_{n\in\mathbb{N}} = \{G_n(\mathsf{f}_n)\}_{n\in\mathbb{N}} = \{p\left(F_n(\mathsf{f}_n) - F_{\gamma_n}(\mathsf{f}_n)\right)\}_{n\in\mathbb{N}}$ is also tight, hence by triangle inequality $M(\{f_n\}_{n\in\mathbb{N}}) = \sup_{n\in\mathbb{N}}|p\left(\mathbb{E}\left[F_n(\mathsf{f}_n)\right] - F_{\gamma_n}(\mathsf{f}_n)\right)|$ is finite. Consequently, for $M > 2M(\{f_n\}_{n\in\mathbb{N}})$,

$$\begin{aligned}\mathbb{P}(E_{4,n}(\mathsf{f}_n, M)) &\leq \mathbb{P}\left(|p\left(F_n(\mathsf{f}_n) - \mathbb{E}\left[F_n(\mathsf{f}_n)\right]\right)| > M - M(\{f_n\}_{n\in\mathbb{N}})\right)\\ &\leq \mathbb{P}\left(|p\left(F_n(\mathsf{f}_n) - \mathbb{E}\left[F_n(\mathsf{f}_n)\right]\right)| > M/2\right) \leq 2\exp(-M^2/(8L^2)),\end{aligned}$$

as desired.

S2.3. Lemma 7

First, note that in view of the Vitali-Porter and Weierstrass theorems(e.g. Schiff (2013, Ch. 1.4, 2.4)), there exists a neighborhood Ω_1 of I with compact closure $\bar{\Omega}_1 \subset \Omega$ such that f_n and f_n' converge uniformly to f and f' respectively on $\bar{\Omega}_1$ and so in particular $\{f_n\}_{n\in\mathbb{N}}$ and $\{f_n'\}_{n\in\mathbb{N}}$ are each uniformly bounded on $\bar{\Omega}_1$.

The truncation and centralization step runs parallel to Bai and Silverstein (2004, pp. 559-560), [BS] below. Let $\tilde{G}_n(\cdot)$ denote the analog of $G_n(\cdot)$ with matrix B_n – which does not depend on f, f_n – replaced by $\tilde{B}_n$. Then the argument there shows that $\tilde{G}_n(f) - G_n(f)$ and $\tilde{G}_n(f_n) - G_n(f_n) \xrightarrow{p} 0$ because $f, \{f_n'\}_{n\in\mathbb{N}}$ are uniformly bounded on $\bar{\Omega}_1$. Therefore, it suffices to consider when $G_n(\cdot)$ denotes the centered linear spectral statistic based on the truncated and centered variables.

Now we argue as on [BS] p.563. Let $M_n(z)$ be the normalized Stieltjes transform difference and $\hat{M}_n(z)$ be its modification on $\mathcal{C}$ as defined on [BS, p.561] – none of these depend on f, f_n. For all large n, we have

$$G_n(f_n) - G_n(f) = -\frac{1}{2\pi i}\int [f_n(z) - f(z)]M_n(z)dz$$

almost surely. In addition, by arguing as shown on [BS] p. 563,

$$\int [f_n(z) - f(z)][M_n(z) - \hat{M}_n(z)]dz \xrightarrow{p} 0$$

as $n \to \infty$ because f_n are uniformly bounded on $\bar{\Omega}_1$ which contains the contour of integration.

Finally,

$$\left|\int [f_n(z) - f(z)]\hat{M}_n(z)dz\right| \leq \|f_n - f\|_\infty \int |\hat{M}_n(z)|dz \xrightarrow{p} 0,$$

since $f_n \to f$ uniformly on $\bar{\Omega}_1$ and, crucially, $\{\hat{M}_n(\cdot)\}$ is a tight sequence on $C(\mathcal{C}, \mathbb{R}^2)$ as shown in Lemma 1 of [BS], and hence so is $\int |\hat{M}_n(z)|dz$.

S2.4. Corollary 10

Let $k \in \mathbb{N}$. From the proof of Proposition 6, $\{(G_n(\mathsf{f}_n))^k\}_{n\in\mathbb{N}}$ is uniformly integrable by $\mathbb{E}\left[|X|^p\right] = p\int_0^\infty y^{p-1}\mathbb{P}\left(|X| > y\right)dy, p \geq 1$ again. Also, from Lemma 7 and continuous mapping theorem, $(G_n(\mathsf{f}_n))^k \xrightarrow{d} (N(\mu(f), \sigma^2(f))^k$. Therefore, by Theorem 2.20 of Van der Vaart (2000), a combination of Skorokhod representation theorem and Vitali's convergence theorem, we obtain $\lim_{n\to\infty}\mathbb{E}\left[(G_n(\mathsf{f}_n))^k\right] = \tau_k(f)$. Also, $\left|\mathbb{E}\left[I(E_n^c)(G_n(\mathsf{f}_n))^k\right]\right| \leq \mathbb{P}\left(E_n^c\right)\mathbb{E}\left[(G_n(\mathsf{f}_n))^{2k}\right] = o(1)$ by Cauchy and the assumption $\lim_{n\to\infty}\mathbb{P}\left(E_n\right) = 1$, hence it follows from another assumption $E_n \subset E_{1,n}^c$ and $G_n(f_n) = G_n(\mathsf{f}_n)$ on $E_{1,n}^c$ that

$$\lim_{n\to\infty}\mathbb{E}\left[I(E_n)(G_n(f_n))^k\right] = \lim_{n\to\infty}\left(\mathbb{E}\left[(G_n(\mathsf{f}_n))^k\right] - \mathbb{E}\left[I(E_n^c)(G_n(\mathsf{f}_n))^k\right]\right) = \tau_k(f),$$

as desired.

References

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32(1A)**, 553-605.

Guionnet, A. and Zeitouni, O. (2000). Concentration of the spectral measure for large matrices. *Electron. Comm. Probab.* **5**, 119-136.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617-1642.

Schiff, J. L. (2013). *Normal families.* Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics.* Vol.**3.** Cambridge University Press.

Wainwright, M. (2015). *Basic tail and concentration bounds.* `http://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`

Yao, J., Zheng, S. and Bai, Z. D. (2015). *Sample covariance matrices and high-dimensional data analysis.* Cambridge University Press.