The Central Limit Theorem for Linear Statistics of the Sum of Rank One Matrices with Log-Concave Distribution

Anna Lytova

B.I. Verkin Institute for Low Temperature Physics, Kharkiv, Ukraine

July 1, 2013

Joint works with Alain Pajor and Leonid Pastur

General Settings

We consider a sequence of real random $n \times n$ matrices

$$M_n = \sum_{\alpha=1}^{m} \tau_\alpha \mathbf{y}_\alpha \otimes \mathbf{y}_\alpha$$

- $\tau_\alpha > 0 \quad \sigma_m(\Delta) = \sharp\{\alpha = 1, .., m : \tau_\alpha \in \Delta\}/m \to \sigma$ weakly as $n \to \infty$
- $\mathbf{y}_\alpha = (y_{\alpha 1}, .., y_{\alpha n}) \in \mathbb{R}^n$, $\alpha = 1, .., m$ – i.i.d. random vectors
- $m = m_n : \; m_n/n \to c \in (0, \infty), \quad n \to \infty$

- This is analog of settings of paper by **Marchenko and Pastur (1967)** where vectors uniformly distributed over the unit sphere were considered.
- **In statistics** there arise such matrices known as the **Sample Covariance matrices**: $M_n = Y^T D Y, \quad Y = (\mathbf{y}_1 \; \mathbf{y}_2 \; ... \; \mathbf{y}_m)^T, \quad D = \{\tau_\alpha \delta_{\alpha\beta}\}_{\alpha,\beta=1}^m$, where now all $\{y_{\alpha k}\}_{\alpha,k=1}^{m,n}$ are i.i.d. random variables.
- The same random matrix appears in **asymptotic convex geometry**. Here i.i.d. random points $\mathbf{y}_\alpha$ are uniformly distributed over a convex body in $\mathbb{R}^n$.

Some Definitions. Linear Eigenvalue Statistic

Counting Measure of eigenvalues $\{\lambda_k^{(n)}\}_{k=1}^n$ of M_n:

$$\mathcal{N}_n(\Delta) = \sharp\{k = 1, ..., n : \lambda_k^{(n)} \in \Delta\}$$

Normalized Counting Measure (NCM) : $N_n(\Delta) := n^{-1}\mathcal{N}_n(\Delta).$

Linear Eigenvalue Statistic (LES) for a given test-function $\varphi : \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$\mathcal{N}_n[\varphi] := \sum_{j=1}^n \varphi(\lambda_j^{(n)}) = \mathrm{Tr}\varphi(M_n),$$

Stieltjes transform of Counting Measure $\mathcal{N}_n(\Delta)$ of eigenvalues of M_n is a particular case of LES corresponding to $\varphi(\lambda) = (\lambda - z)^{-1}$, $\mathrm{Im}\, z \neq 0$:

$$\gamma_n(z) := \int \frac{\mathcal{N}_n(d\lambda)}{\lambda - z} = \mathrm{Tr}G(z), \quad \mathrm{Im}\, z \neq 0,$$

Here $G(z) = (M_n - zI)^{-1}$, $\mathrm{Im}\, z \neq 0$, is the resolvent of M_n.

Convergence of the Normalized Counting Measure

Theorem (A. Marchenko, L. Pastur'67)

If $\{\mathbf{y}_\alpha\}_{\alpha=1}^m$ are uniformly distributed over the unit sphere of $\mathbb{R}^n$ ($\mathbb{C}^n$) then there exists a non-random probability measure N ($N(\mathbb{R} = 1)$) s.t. we have convergence in probability

$$\lim_{n\to\infty} N_n(\Delta) = N(\Delta), \quad \forall\Delta \subset \mathbb{R}.$$

The Stieltjes transform of N,

$$f(z) = \int \frac{N(d\lambda)}{\lambda - z}, \quad \mathrm{Im}\, z \neq 0,$$

is uniquely determined by the functional equation

$$zf(z) = c - 1 - c\int (1 + \tau f(z))^{-1}\sigma(d\tau)$$

considered in the class of functions analytic in $\mathbb{C}\backslash\mathbb{R}$ and such that $\mathrm{Im}\, z\, \mathrm{Im}\, f(z) \geq 0$, $\mathrm{Im}\, z \neq 0$.

Convergence of the NCM. "Good" Vectors.

Definition. A random real vector $\mathbf{y} = (y_1, ..., y_n)$ is called **isotropic** if

$$\mathbf{E}\{y_j\} = 0, \quad \mathbf{E}\{y_j y_k\} = \delta_{jk} n^{-1}.$$

Definition. A random isotropic vector $\mathbf{y} = (y_1, ..., y_n)$ is called **"good"** if for any $n \times n$ complex matrix A_n we have

$$\mathbf{Var}\{(A_n\mathbf{y}, \mathbf{y})\}(= \mathbf{E}\{(A_n\mathbf{y}, \mathbf{y})^2\} - (n^{-1}\mathrm{Tr}A_n)^2) \leq ||A_n||^2\delta_n, \quad \delta_n = o(1), \; n \to \infty.$$

Theorem (Pajor, Pastur'07).

The Marchenko-Pastur theorem remains valid for $M_n = \sum_{\alpha=1}^{m} \tau_\alpha \mathbf{y}_\alpha \otimes \mathbf{y}_\alpha$, where $\{\mathbf{y}_\alpha\}_{\alpha=1}^m$ are good vectors.

Theorem (B.Klartag'07). There exist $C > 0$, $\beta \in (0,1)$ s. t. for any isotropic random vector $\mathbf{y} \in \mathbb{R}^n$ with a log-concave distribution $\mathbf{Var}\{|\mathbf{y}|^2\} \leq C/n^\beta$.

Theorem (Pajor, Pastur'07). Random isotropic vectors with log-concave distribution are good.

So for good vectors we have in probability for any bounded continuous φ:

$$\lim_{m,n\to\infty, m/n\to\infty} n^{-1}\mathcal{N}_n[\varphi] = \int \varphi(\lambda)N(d\lambda), \quad (\mathcal{N}_n[\varphi] = \mathrm{Tr}\varphi(M_n))$$

What about the limiting probability law for fluctuations of $\mathcal{N}_n^\circ[\varphi] = \mathcal{N}_n[\varphi] - \mathbf{E}\{\mathcal{N}_n[\varphi]\}$?

$$\mathbf{y}_\alpha - ?: \ \mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi]}\} \underset{m,n\to\infty, m/n\to c}{\longrightarrow} ?$$

Proposition (M. Shcherbina'11). For any $s > 0$ and any M_n

$$\mathbf{Var}\{\mathcal{N}_n[\varphi]\} \le C_s||\varphi||_s^2 \int_0^\infty dy e^{-y} y^{2s-1} \int_{-\infty}^\infty \mathbf{Var}\{\gamma_n(x+iy)\}dx,$$

where $\varphi \in H_s = \{\psi : ||\psi||_s^2 = \int (1+2|\xi|)^{2s}|\widehat{\psi}(\xi)|^2 d\xi < \infty\}$, $\gamma_n(z) = \mathrm{Tr}G(z)$.

$\mathbf{Var}\{\gamma_n(z)\} \leq ?$

- $\mathbf{E}_{\leq \alpha}$ – averaging w.r. to $\{\mathbf{y}_1, ..., \mathbf{y}_\alpha\}$, $\mathbf{E}_\alpha$ – averaging w.r. to $\mathbf{y}_\alpha$,
- $(\xi)^\circ_\alpha = \xi - \mathbf{E}_\alpha\{\xi\}$ – centering w.r. to $\mathbf{y}_\alpha$

According to the standard martingale method

$$\mathbf{Var}\{\gamma_n\} \leq \sum_{\alpha=1}^{m} \mathbf{E}\{|\mathbf{E}_{\leq \alpha-1}\{\gamma_n\} - \mathbf{E}_{\leq \alpha}\{\gamma_n\}|^2\} \leq \sum_{\alpha=1}^{m} \mathbf{E}\{|(\gamma_n)^\circ_\alpha|^2\} \qquad (1)$$

Denote $M_n^\alpha = M_n|_{\tau_\alpha=0}$, $G^\alpha(z) = (M_n^\alpha - zI_n)^{-1}$, $\gamma_n^\alpha(z) = \mathrm{Tr}\ G^\alpha$.
It follows from the rank-one perturbation formula that

$$\gamma_n - \gamma_n^\alpha = \frac{-\tau_\alpha(G^{\alpha 2}\mathbf{y}_\alpha, \mathbf{y}_\alpha)}{1 + \tau_\alpha(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)} \qquad (2)$$

Note that $\mathbf{E}\{(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)\} = n^{-1}\mathbf{E}\{\mathrm{Tr} G^\alpha(z)\} \to f(z)!$
After subtracting γ_n^α in the r.h.s. of (1) one can get with the help of (2):

$$\mathbf{Var}\{\gamma_n\} \leq \frac{4}{|\operatorname{Im} z|^2} \sum_{\alpha=1}^{m} \tau_\alpha^2 \mathbf{E}\left\{ \frac{\mathbf{E}_\alpha\{|(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)^\circ_\alpha|^2\}}{|\mathbf{E}_\alpha\{1 + \tau_\alpha(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)\}|^2} \right\}$$

$\mathbf{Var}\{\gamma_n(z)\} \leq ?$

We have for isotropic vectors with log-concave distribution

$$\mathbf{E}_\alpha\{|(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)^\circ_\alpha|^2\} \leq |\operatorname{Im} z|^{-2} n^{-\beta}, \quad \beta < 1$$

Let us denote

$$m_\sigma^p = \int_0^\infty \tau^p d\sigma(\tau)$$

So if $m_\sigma^2 < \infty$ and $\{\mathbf{y}_\alpha\}_{\alpha=1}^m$ are isotropic with log-concave distribution then there exists constant $C = C(z)$ such that

$$\mathbf{Var}\{\gamma_n(z)\} \leq C n^{1-\beta}.$$

Note that for Sample Covariance Matrices we have

$$\mathbf{Var}\{\gamma_n(z)\} \leq C$$

We see that to get this bound in general case it is not enough for vectors to be just good.

$\mathbf{Var}\{\gamma_n(z)\} = O(1), \quad n \to \infty$

Definition The joint density of $\mathbf{y} \in \mathbb{R}^n$ is **unconditional** if the random vector $(y_1, ..., y_n)$ has the same distribution as $(\pm y_1, ..., \pm y_n)$ for any choice of signs.

Theorem (B.Klartag'08) There exists a positive universal constant $C \leq 16$ s. t.

$$\mathbf{Var}\{|\mathbf{y}|^2\} \leq C/n$$

for any random $\mathbf{y} \in \mathbb{R}^n$ with a log-concave unconditional distribution and $\mathbf{E}\{y_j^2\} = 1/n$.

It follows from the theorem by B.Klartag that for random vectors $\mathbf{y} \in \mathbb{R}^n$ with a log-concave unconditional distribution and $\mathbf{E}\{y_j^2\} = 1/n$ we have

$$\mathbf{Var}\{(A_n\mathbf{y}, \mathbf{y})\} = O(1/n), \quad n \to \infty$$

for any A_n, $||A_n|| = 1$. For such vectors

$$\mathbf{Var}\{\gamma_n(z)\} \leq C|z||\operatorname{Im} z|^{-4}.$$

$\mathbf{Var}\{\mathcal{N}_n[\varphi]\} \le C_s||\varphi||^2_{5/2+\varepsilon}$

Recall that

$$\mathbf{Var}\{\mathcal{N}_n[\varphi]\} \le C_s||\varphi||_s^2 \int_0^\infty dy e^{-y} y^{2s-1} \int_{-\infty}^\infty \mathbf{Var}\{\gamma_n(x+iy)\}dx,$$

$$\mathbf{Var}\{\gamma_n\} \le \frac{4}{|\operatorname{Im} z|^2} \sum_{\alpha=1}^m \tau_\alpha^2 \mathbf{E}\left\{\frac{\mathbf{E}_\alpha\{|(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)^\circ_\alpha|^2\}}{|\mathbf{E}_\alpha\{1+\tau_\alpha(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)\}|^2}\right\}$$

For vectors with unconditional distribution we have:

$$\mathbf{E}\{y_j y_k y_p y_q\} = a_{2,2}(\delta_{jk}\delta_{pq} + \delta_{jp}\delta_{kq} + \delta_{jq}\delta_{kp}) + \kappa_4 \delta_{jk}\delta_{jp}\delta_{jq}, \quad \kappa_4 = a_4 - 3a_{2,2}.$$

where $a_{2,2} = a_{2,2}(n) := \mathbf{E}\{y_j^2 y_k^2\}$, $a_4 = a_4(n) := \mathbf{E}\{y_j^4\}$. Hence

$$\mathbf{E}_\alpha\{|(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)^\circ_\alpha|^2\} = (a_{2,2} - n^{-2})|\mathrm{Tr}\, G^\alpha|^2 + 2a_{2,2}\mathrm{Tr}|G^\alpha|^2 + \kappa_4 \sum_{j=1}^n |G^\alpha_{jj}|^2.$$

If $(a_{2,2} - n^{-2}) = O(n^{-3})$ and $\kappa_4 = O(n^{-2})$, then

$$\mathbf{E}_\alpha\{|(G^\alpha \mathbf{y}_\alpha, \mathbf{y}_\alpha)^\circ_\alpha|^2\} \le Cn^{-2}\mathrm{Tr}|G^\alpha|^2.$$

$\mathbf{Var}\{\mathcal{N}_n[\varphi]\} \le C_s||\varphi||^2_{5/2+\varepsilon}$

Lemma. Let $\mathbf{y}_\alpha \in \mathbb{R}^n$, $\alpha = 1, .., m$ be i.i.d. random vectors satisfying

(i) the joint distribution of $\mathbf{y}_\alpha$ is log-concave and unconditional,

(ii) $\mathbf{E}\{y_j^2\} = 1/n$, $j = 1, .., n$,

(iii) $a_{2,2} := \mathbf{E}\{y_j^2 y_k^2\} = n^{-2} + O(n^{-3})$, $\forall j \ne k$,
$a_4 := \mathbf{E}\{y_j^4\} = O(n^{-2})$.

Then

$$\mathbf{Var}\{\gamma_n(x+iy)\} \le \frac{C}{ny^4} \sum_{\alpha=1}^{m} \tau_\alpha^4 \mathbf{E}\{n^{-1}\mathrm{Tr}|G^\alpha(x+iy)|^2\}. \tag{3}$$

Theorem

If i.i.d. vectors $\mathbf{y}_\alpha \in \mathbb{R}^n$, $\alpha = 1, .., m$ satisfy (i) – (iii), $\varphi \in H_{5/2+\varepsilon}$, $\varepsilon > 0$, and $m_\sigma^4 < \infty$ then

$$\mathbf{Var}\{\mathcal{N}_n[\varphi]\} \le C||\varphi||^2_{5/2+\varepsilon}, \quad \forall \varepsilon > 0.$$

Approximation Procedure

For any $\varphi \in H_s$ put

$$\varphi_\eta = \varphi * P_\eta, \quad P_\eta = \frac{1}{\pi}\frac{\eta}{x^2+\eta^2}.$$

Then

$$\lim_{\eta \downarrow 0} ||\varphi - \varphi_\eta||_s = 0.$$

We have for the characteristic function:

$$\mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi]}\} = \mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi]} - e^{ix\mathcal{N}_n^\circ[\varphi_\eta]}\} + \mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi_\eta]}\},$$

$$|\mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi]} - e^{ix\mathcal{N}_n^\circ[\varphi_\eta]}\}| \le |x|\mathbf{Var}\{\mathcal{N}_n[\varphi - \varphi_\eta]\}^{1/2} \le |x|||\varphi - \varphi_\eta||_s.$$

Hence uniformly in $|x| \le C$:

$$\lim_{n\to\infty} \mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi]}\} = \lim_{\eta\downarrow 0}\lim_{n\to\infty} \mathbf{E}\{e^{ix\mathcal{N}_n^\circ[\varphi_\eta]}\}.$$

Besides

$$\mathcal{N}_n[\varphi_\eta] = \frac{1}{\pi}\int \varphi(\mu)\,\mathrm{Im}\,\gamma_n(\mu + i\eta)d\mu.$$

Very Good Vectors. Covariance of Traces of Resolvent

Definition. A random vector $\mathbf{y} = (y_1, ..., y_n)$ is called **very good** if

(i) the joint distribution of $\mathbf{y}$ is log-concave and unconditional,

(ii) $\mathbf{E}\{y_j^2\} = 1/n$, $j = 1, .., n$,

(iii) $a_{2,2} := \mathbf{E}\{y_j^2 y_k^2\} = n^{-2} + an^{-3} + o(n^{-3})$,
$\kappa_{\mathbf{4}} := \mathbf{E}\{y_j^4\} - 3a_{2,2} = bn^{-2} + o(n^{-2})$,

(iv) $\mathbf{E}\{|(A_n\mathbf{y}, \mathbf{y})^\circ|^4\} = o(n^{-1}), \quad \forall A_n,\ ||A_n|| \le 1$

It follows from (iii) that
$\mathbf{E}\{|(A_n\mathbf{y}, \mathbf{y})^\circ|^2\} = an^{-3}|\mathrm{Tr}A_n|^2 + 2n^{-2}\mathrm{Tr}A_n\overline{A_n} + bn^{-2}\sum_{j=1}^n A_{njj}\overline{A_{njj}} + O(n^{-2})$.

Theorem

Let $\{\mathbf{y}_\alpha\}_{\alpha=1}^m$ be very good random vectors, and $m_\sigma^4 < \infty$. Then

$$\lim_{m,n\to\infty, m/n\to c} \mathbf{E}\{\gamma_n(z_1)\gamma_n(z_2)^0\}$$
$$= \frac{\partial^2}{\partial z_1 \partial z_2}\left(-(a+b)f(z_1)f(z_2) + 2\ln\frac{f(z_1) - f(z_2)}{z_1 - z_2}\right) =: C(z_1, z_2),$$

where $zf(z) = c - 1 - c\int(1 + \tau f(z))^{-1}\sigma(d\tau)$.

CLT

Theorem (CLT)

Let

- $\{\mathbf{y}_\alpha\}_{\alpha=1}^m$ be very good random vectors,
- $m_\sigma^4 < \infty$,
- $\varphi \in H_{5/2+\varepsilon}$, $\varepsilon > 0$.

Then $\mathcal{N}_n^\circ[\varphi] = \mathcal{N}_n[\varphi] - \mathbf{E}\{\mathcal{N}_n[\varphi]\}$ converges in distribution to the Gaussian random variable with zero mean and the variance

$$V[\varphi] = \lim_{\eta \downarrow 0} \frac{1}{2\pi^2} \int \varphi(\mu) d\mu \int \varphi(\lambda) d\lambda \, \mathrm{Re}\left[C(z_\mu, \overline{z_\lambda}) - C(z_\mu, z_\lambda)\right],$$

where $z_\lambda = \lambda + \eta$, $z_\mu = \mu + \eta$,

$$C(z_1, z_2) = \frac{\partial^2}{\partial z_1 \partial z_2}\Big(-(a+b) f(z_1) f(z_2) + 2 \ln \frac{f(z_1) - f_(z_2)}{z_1 - z_2} \Big),$$

$$zf(z) = c - 1 - c \int (1 + \tau f(z))^{-1} \sigma(d\tau)$$

Examples of Good Vectors. 1. Vectors with i.i.d. components

Definition. A random vector $\mathbf{y} = (y_1, ..., y_n)$ is called **very good** if

(i) the joint distribution of $\mathbf{y}$ is log-concave and unconditional,

(ii) $\mathbf{E}\{y_j^2\} = 1/n,\ j = 1, .., n,$

(iii) $a_{2,2} := \mathbf{E}\{y_j^2 y_k^2\} = n^{-2} + an^{-3} + o(n^{-3}),$
$\kappa_{\mathbf{4}} := \mathbf{E}\{y_j^4\} - 3a_{2,2} = bn^{-2} + o(n^{-2}),$

(iv) $\mathbf{E}\{|(A_n\mathbf{y}, \mathbf{y})^\circ|^4\} = o(n^{-1}), \quad \forall A_n,\ ||A_n|| \le 1$

$\mathbf{y} = n^{-1/2}\mathbf{x}$: $\{x_j\}_{j=1}^n$ are i.i.d., $\mathbf{E}\{x_j\} = 0$, $\mathbf{E}\{x_j^2\} = 1$, $\mathbf{E}\{x_j^8\} < \infty$.

- $\mathbf{E}\{y_j\} = 0, \quad \mathbf{E}\{y_j y_k\} = \delta_{jk} n^{-1},$
- $a_{2,2} = n^{-2}, \quad (a = 0),$
- $\kappa_4 = -2n^{-2}, \quad (b = -2),$
- $\mathbf{E}\{|(A_n\mathbf{y}, \mathbf{y})^\circ|^4\} \le C||A_n||^4 n^{-2},$

If $\mathbf{E}\{x_j^{4+\varepsilon}\} = m_{4+\varepsilon} < \infty$, then we have for truncated $\tilde{y}_j = y_j I_{|y_j| \le C}$:

$$\mathbf{E}\{|(A_n\widetilde{\mathbf{y}}, \widetilde{\mathbf{y}})^\circ|^4\} \le C||A_n||^4 n^{-(1+\varepsilon/2)}.$$

Vectors Uniformly Distributed over l_p^n-Ball

Let $\mathbf{x}$ be uniformly distributed over $B_p^n = \{\mathbf{x} \in \mathbb{R}^n : \ ||\mathbf{x}||_p = (\sum |x_j|^p)^{1/p} \le 1\}$,
According to **F. Barthe, O. Guedon, S. Mendelson, and A. Naor**

$$\mathbf{E}_{B_p^n}\{f\} := \frac{1}{|B_p^n|}\int_{B_p^n} f dx$$
$$= \frac{1}{(2\Gamma(1+1/p))^n}\int_0^\infty dz e^{-z}\int_{\mathbb{R}^n} d\mathbf{x} e^{-||\mathbf{x}||_p^p} f(\mathbf{x}(||\mathbf{x}||_p^p + z)^{-1/p}).$$

This formula allows to calculate moments of $\mathbf{x}$ and get after renormalization:

$$\mathbf{y} = \left(\frac{1}{n}\frac{B(1/p, 2/p)}{B(n/p+1, 2/p)}\right)^{1/2}\mathbf{x}$$

is such that as $n \to \infty$

- $\mathbf{E}\{y_j\} = 0, \quad \mathbf{E}\{y_j y_k\} = \delta_{jk} n^{-1}$,
- $a_{2,2} = n^{-2} + an^{-3} + O(n^{-4})$,
- $\kappa_4 = bn^{-2} + O(n^{-3})$,
- $\mathbf{E}\{|(A_n\mathbf{y}, \mathbf{y})^\circ|^4\} \le C||A_n||^4 n^{-2}$,

where

$$a = -\frac{8}{p}, \quad b = \frac{\Gamma(1/p)\Gamma(5/p)}{\Gamma(3/p)^2} - 3.$$