Methodologies for Analysis of Random Matrices’s Spectral and LDA’ error in High-dimensional setting

Nguyen Hoang Huy

April, 2012
There are two three important methods employed in this area: the moment method, Stieltjes transform, and orthogonal polynomial decomposition of the exact density of eigenvalues.

We also present some basic result about spectral of Wigner matrices (semicircular law) and sample Covariance matrices (Marčenko-Pastur law) and Linear Discriminant Analysis error rate estimation in high-dimensional setting.
Definition 1

Suppose $A$ is an $p \times p$ matrix with eigenvalues $\lambda_j$, $j = 1, 2, \cdots, p$. If all these eigenvalues are real (e.g., if $A$ is Hermitian), we can define a one-dimensional distribution function:

$$F^A = \frac{1}{p} \# \{ j \leq p : \lambda_j \leq x \} \quad (1)$$

called the empirical spectral distribution of the matrix $A$.

If the eigenvalues $\lambda_j$’s are not all real, we can define a two-dimensional empirical spectral distribution of the matrix $A$:

$$F^A(x, y) = \frac{1}{p} \# \{ j \leq p : \Re(\lambda_j) \leq x, \Im(\lambda_j) \leq y \}. \quad (2)$$
**Spectral Analysis**

**Definition 1**

The limit distribution $F = \lim_{n \to \infty} F^{A_n}$ for a given sequence of random matrices $\{A_n\}$, which is usually nonrandom, is called the limiting spectral distribution of the sequence $\{A_n\}$.

- Many important statistics in multivariate analysis can be expressed as functions of the empirical spectral distribution of some random matrices.

- **Example 1.** Let $A$ be an $p \times p$ positive definite matrix. Then

$$\det(A) = \prod_{j=1}^{n} \lambda_j = \exp \left( n \int_{0}^{\infty} \log x F^{A}(dx) \right)$$
Suppose \( \{F_n\} \) denotes a sequence of distribution functions with finite moments \( \beta_k(F_n) \) of all orders \( k \):

\[
\beta_{n,k} = \beta_k(F_n) := \int x^k dF_n(x).
\]  

Lemma 1

(Unique limit). A sequence of distribution function \( \{F_n\} \) converges weakly to a limit if the following conditions are satisfied:

1. Each \( F_n \) has finite moments of all orders.
2. For each fixed integer \( k \geq 0 \), \( \beta_{n,k} \) converges to a finite limit \( \beta_k \) as \( n \to \infty \).
3. If two right-continuous nondecreasing functions \( F \) and \( G \) have same moment sequence \( \{\beta_k\} \), then \( F = G + \text{const.} \)
Lemma 2

(M. Riesz). Let \( \{ \beta_k \} \) be the sequence of moments of distribution function \( F \). If

\[
\liminf_{k \to \infty} \frac{1}{k} \beta_{2k}^2,
\]

then \( F \) is uniquely determined by the moment sequence \( \{ \beta_k, k = 0, 1, \cdots \} \).

Lemma 3

(Carleman). Let \( \{ \beta_k = \beta_k(F) \} \) be the sequence of moments of the distribution \( F \). If the Carleman condition

\[
\sum \beta_{2k}^{-1/2k} = \infty
\]

is satisfied, then \( F \) is uniquely determined by the moment sequence \( \{ \beta_k, k = 0, 1, \cdots \} \).
Spectral Analysis

- Let $\mathbf{A}$ be an $p \times p$ Hermitian matrix, the $k$-th moment of the empirical spectral distribution $F^\mathbf{A}$ can be written as

$$\beta_{p,k}(F^\mathbf{A}) = \int_{-\infty}^{\infty} x^k F^\mathbf{A}(dx) = \frac{1}{p} tr(\mathbf{A}^k).$$ (6)

- This expression plays a fundamental role in Random Matrices Theory.

- The problem of showing that $\{F^\mathbf{A}_p\}$ of $\{\mathbf{A}_p\}$ (strongly or weakly or in another sense) tends to a limit reduces to showing that, for each fixed $k$, the $\{\frac{1}{p} tr(\mathbf{A}_p^k)\}$ tends to a limit $\beta_k$ in the corresponding sense and then verifying the Carleman condition (5).
If $G(x)$ is a function of bounded variation on the real line, then its Stieltjes transform is defined by

$$s_G(z) = \int \frac{1}{\lambda - z} dG(\lambda), \quad z \in D,$$

where $z \in D \equiv \{ z \in \mathbb{C} : \Re z > 0 \}$.  

Lemma 4

(Inversion formula). For any continuity points $a < b$ of $G$, we have

$$G\{[a, b]\} = \lim_{\varepsilon \to 0^+} \frac{1}{\pi} \int_a^b \Im s_G(x + i\varepsilon) dx.$$
Lemma 5

Assume that \( \{ G_n \} \) is a sequence of functions of bounded variation and \( G_n(-\infty) = 0 \) for all \( n \). Then,

\[
\lim_{n \to \infty} s_{G_n}(z) = s(z) \quad \forall z \in D
\]  

(9)

if and only if there is a function of bounded variation \( G \) with \( G(-\infty) = 0 \) and Stieltjes transform \( s(z) \) and such that \( G_n \to G \) vaguely.

Lemma 6

Let \( G \) be a function of bounded variation and \( x_0 \in \mathbb{R} \). Suppose that \( \lim_{z \in \mathbb{R} \to x_0} \Im s_G(z) \) exists. Call it \( \Im s_G(x_0) \). Then \( G \) is differentiable at \( x_0 \), and its derivative is \( \frac{1}{\pi} \Im s_G(x_0) \).
Compared with the Fourier transform, an important advantage of Stieltjes transforms is that one can easily find the density function of a signed measure via its Stieltjes transform.

The Stieltjes transform of the empirical spectral distribution $F_p$ of the $p \times p$ Hermitian matrix $A$:

$$s_p(z) = \int \frac{1}{x - z} dF_p(x) = \frac{1}{p} tr(A - zI)^{-1}. \quad (10)$$

$X = (A - zI)^{-1}$: the resolvent of the Hermitian matrix $A$. 

Nguyen Hoang Huy
Methodologies for Analysis of Random Matrices's Spectral and LDA error in High-dimensional setting
The semicircular law $F$ whose density is given by

$$F'(x) = \begin{cases} \frac{1}{2\pi} \sqrt{4 - x^2}, & \text{if } |x| \leq 2, \\ 0, & \text{otherwise} \end{cases}$$

(11)

Suppose that $X_p$ is an $p \times p$ Hermitian matrix. The entries above or on the diagonal of $X_p$ are independent but may be dependent on $p$ and may not necessarily be identically distributed. Assume that all the entries of $X_p$ are of mean zero and variance 1 and satisfy the condition that, for any constant $\nu > 0$,

$$\lim_{n \to \infty} \frac{1}{p^2} \sum_{jk} E|x_{jk}^{(p)}|^2 I(|x_{jk}^{(p)}| \geq \nu \sqrt{p}) = 0.$$  

Then, with probability 1, the empirical spectral distribution of $W_p = \frac{1}{\sqrt{p}} X_p$ tends to the semicircular law.
**Definition 4**

The Marčenko-Pastur (M-P) law $F_y(x)$ has a density function

$$p_y(x) = \begin{cases} \frac{1}{2\pi xy\sigma^2} \sqrt{(b - x)(x - a)}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

and has a point mass $1 - 1/y$ at the origin if $y > 1$, where $a = \sigma^2(1 - \sqrt{y})^2$ and $b = \sigma^2(1 + \sqrt{y})^2$.

**Theorem 2**

Suppose that, for each $n$, the entries of $X$ are independent complex variables with a common mean $\mu$ and variance $\sigma^2$. Assume that $p/n \to y \in (0, \infty)$ and that, for any $\nu > 0$,

$$\frac{1}{\nu^2 np} \sum_{jk} E|X_{jk}^{(n)}|^2 I(|X_{jk}^{(n)}| \geq \nu \sqrt{n}) \to 0.$$  

Then, with probability 1, $F_S$ tends to M-P law with ratio index $y$ and scale index $\sigma^2$, $S = \frac{1}{n}XX^*$. 
Theorem 3

Suppose that \( \{x_1, \cdots, x_n\} \), \( x_j = (x_{1j}, \cdots, x_{pj}) \) drawn from the population satisfies \( \text{cov}(x, x) = \Sigma_p \), \( Ex = 0 \), the four moments of all components of \( x \) exist,

\[
0 < M = \sup_{|e|=1} E(e^T x)^4 < c_0 \text{ with } c_0 \text{ does not depend on } p,
\]

\[
\gamma = \nu / M \to 0 \text{ as } p \to \infty \text{ where } \nu = \sup_{\|\Omega\|=1} \text{var}(x^T \Omega x / n),
\]

\[
p/n \to \lambda, \ 0 < c_1 \leq \lambda(\Sigma) \leq c_2 \text{ with } c_1, c_2 \text{ do not depend on } p,
\]

and as \( p \to \infty \), \( F_{\Sigma_p}(u) \to F(u) \) for each \( u \geq 0 \).

1. If \( \lambda = 0 \), then \( F_{\hat{\Sigma}_p} \to F \) almost everywhere.

2. If \( 0 < \lambda \neq 1 \) then \( F_{\hat{\Sigma}_p} \to \hat{F} \) satisfies \( \hat{F}(0) = \hat{F}(u_1 - 0) = \max(0, 1 - \lambda^{-1}), \hat{F}(u_2) = 1 \), where \( u_1 = c_1(1 - \sqrt{\lambda})^2 \), \( u_2 = c_2(1 + \sqrt{\lambda})^2 \), moreover it derivative exits and \( \hat{F}'(u) \leq \pi^{-1}(c_1 \lambda u)^{-1/2} \).
LDA in High-dimensional setting

- We consider a sequence \( \mathcal{B} = \{ \mathcal{B}_p \} \) of \( p \)-dimensional discrimination problems

\[
\mathcal{B}_p = (\mu_1, \mu_2, \Sigma, n_1, n_2, \hat{\delta}(x), \bar{W}(\hat{\delta}))_p, \quad p = 1, 2, \ldots
\]

- We restrict \( \mathcal{B} \) with the following conditions:
  1. \( 0 < c_1 \leq \lambda(\Sigma) \leq c_2 \) with \( c_1, c_2 \) do not depend on \( p \)
  2. \( \text{Exits } y_k = \lim_{p \to \infty} \frac{p}{n_k} \geq 0, \quad k = 1, 2 \) such that
     \[
y = \lim_{p \to \infty} \frac{p}{n_1 + n_2} = \frac{y_1y_2}{y_1 + y_2} < 1
\]
  3. \( F\Sigma_p \to F \) and \( B_p(u) \to B(u) \) almost everywhere,
     \[
     B_p(u) = \sum_{i=1}^{p} \frac{\alpha_i^2}{\lambda_i} \#(\lambda_i \leq u),
     \]
     where \( \alpha_i \) are components of \( \alpha = \mu_1 - \mu_2 \) in a system of coordinates where \( \Sigma \) is diagonal.
  4. \( \hat{\delta}(x) = (x - (\hat{\mu}_1 + \hat{\mu}_2)/2)^T \Gamma(\hat{\Sigma})(\hat{\mu}_1 - \hat{\mu}_2) \),

Nguyen Hoang Huy
Methodologies for Analysis of Random Matrices's Spectral and LDA's error in High-dimensional setting
LDA in High-dimensional setting

- \( \Gamma(\hat{\Sigma}) \) is diagonalized together with \( \hat{\Sigma} \) and has eigenvalues \( \Gamma(\lambda) \) for the eigenvalues \( \lambda \) of \( \hat{\Sigma} \); it is defined by the scalar function \( \Gamma : \mathbb{R} \rightarrow \mathbb{R} \)

\[
\Gamma = \Gamma(u) = \int_{t \geq 0} (1 + ut)^{-1} d\eta(t),
\]

where \( \eta(t) \) is a function of finite variation on \([0, \infty)\) not depending on \( p \).

- When \( \pi_1 = \pi_2 = 1/2 \), the misclassification error \( \overline{W}(\hat{\delta}) \) is minimized at:

\[
c = c_{n \text{opt}} = 1/2(\mathbf{E}_1 \hat{\delta}(x) - \mathbf{E}_2 \hat{\delta}(x)) \quad (13)
\]

and is equal to

\[
\overline{W}(\hat{\delta})|_{c=c_{n \text{opt}}} = \Phi \left( -\frac{\alpha^T \Gamma \hat{\alpha}}{2 \sqrt{\hat{\alpha}^T \Gamma \Sigma \Gamma \hat{\alpha}}} \right). \quad (14)
\]
Theorem 6

Under assumption 1, 2, 3, 4, there exist the limits

\[ \lim_{n \to \infty} c_n^{opt} = c^{opt}, \]
\[ \lim_{n \to \infty} \alpha \Gamma(\hat{\Sigma}) \hat{\alpha} = 2M, \quad \lim_{n \to \infty} \hat{\alpha} \Gamma(\hat{\Sigma}) \Sigma \Gamma(\hat{\Sigma}) \hat{\alpha} = V \]

where

\[ h(z) = \lim_{p \to \infty} \frac{1}{p} \text{tr}((I - z\Sigma)^{-1}), \quad s(z) = 1 + y(h(z) - 1), \quad (15) \]

\[ b(z) = \int (1 - zs(z)u)^{-1} u dB(u), \quad (16) \]

\[ k(z) = \begin{cases} 
    b(z) + (y_1 + y_2)(h(z) - 1)/(zs(z)) & \text{if } z \neq 0, \\
    b(0) + (y_1 + y_2)(h(z) - 1)/\Lambda_1 & \text{if } z = 0. 
\end{cases} \quad (17) \]
Theorem 6

\[ \Lambda_1 = \lim_{p \to \infty} p^{-1} \text{tr}(\Sigma) \quad (18) \]

\[ c^{opt} = \frac{1}{2} (y_2 - y_1) \int \frac{1 - h(-t)}{ts(-t)} d\eta(t), \quad (19) \]

\[ M = \frac{1}{2} \int b(-t) d\eta(t), \quad (20) \]

\[ V = \int \int \frac{k(-t) - k(-t')}{s(-t)s(-t')(t - t')} d\eta(t)d\eta(t'), \quad (21) \]

and the last integrand is extended by continuity to \( t = t' \) and to \( t = 0 \).
Suppose assumption 1, 2, 3, 4 hold, $D > 0$, the discrimination function is used, and the threshold $c_{n}^{opt}$ is chosen for each $n$ so that it minimizes $\overline{W}(\hat{\delta})$. Then

$$\lim_{n \to \infty} \overline{W}(\hat{\delta})|_{c=c_{n}^{opt}} = \Phi(\sqrt{d^{\text{eff}}}/2),$$

(22)

where $d^{\text{eff}} = 4M^2/D$.

**Example.** LDA corresponds to the case

\[ \eta(t') = \text{ind}(t' < t), \text{ where } t \to \infty. \]

At that time

\[ h(-t) \to 0, \quad s(-t) \to 1 - y, \text{ and } tb(-t) \to J/(1 - y) \]

where

\[ J = \lim_{p \to \infty} \alpha^T \Sigma^{-1} \alpha \quad \text{and then we have} \]

\[ G \to J/2(1 - y)^{-1}, \quad D \to (J + y_1 + y_2)(1 - y)^{-3}, \quad \text{and} \]

\[ J^{\text{eff}} \to J^2(1 - y)/(J + y_1 + y_2). \]