On consistency of eigenvalues for principal component analysis

Yohji Akama$^1$ Yasutaka Uwano$^1$

$^1$ Mathematical Institute, Tohoku University

Abstract: We study the empirical spectral distribution of so-called large dimensional random matrices. By empirical process theory and measure concentration inequalities, we provide a sufficient condition for the sum of the largest eigenvalues of the sample covariance matrix to be consistent, in the limit of the sample size $n$ with the dimension $d$ of data in the sample varying along $n$.

1 Introduction

Given an i.i.d. sample

$$x_1, \ldots, x_n \in \mathbb{R}^d,$$

one may wish, for the sake of analysis, to transform them into a space of lower dimension, say $k < d$, by considering that the true data are drawn from a (affine) space of $k$ dimension. A principal component analysis (PCA) is a statistical inference used very frequently for such purpose. It is based on the greatest $k$ among the eigenvalues

$$\lambda_1 \geq \cdots \geq \lambda_k \geq \cdots \geq \lambda_d$$

of the sample covariance matrix

$$S = \sum_{i=1}^{n} x_i x_i^T/n.$$

The PCA projects the sample to the space spanned by the eigenvectors corresponding to $\lambda_1, \ldots, \lambda_k$.

A PCA is adapted to various statistical models. For example, a PCA is used for processing DNA micro array data, which is a typical high dimension, low sample size (HDLSS) data. A kernel PCA seems to be a PCA regarding raw data as high dimensional data in the feature space.

When a statistical model of the problem is complicated, the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of the sample covariance matrix $S$ are not necessarily consistent to the eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

of the population covariance matrix $\Sigma$. In other words, $\lambda_i$ does not necessarily converge to $\lambda_i$ in probability.

For example, let us consider the case when we take the simultaneous limit of the sample size $n$ and the dimension $d$ with the ratio

$$d/n \to c > 0,$$

and the set of the components $(x_i)$ of the sample $(i = 1, \ldots, n; j = 1, \ldots, d)$ is i.i.d. with zero mean, unit variance and the fourth moment. Then the extreme eigenvalues of the sample covariance matrix satisfy almost surely that

$$\ell_1 \to (1 + \sqrt{c})^2, \quad \ell_d \to (1 - \sqrt{c})^2.$$

For the comprehensive survey on related topics, see [3]. For a particular case where the population is a standard normal distribution and $c < 1$, Johnstone [11] established that the limiting distribution of appropriately centered and scaled $\ell_1$ is subject to a Tracy-Widom distribution of order 1.

Taking the simultaneous limit of the dimension $d$ and the sample size $n$ under a certain condition on $d$ and $n$ is sometimes more useful in analyzing linear functional relation models (e.g. PCA), than taking the limit of $n$ but with $d$ fixed, when the population is normal, as we see below:

- for the distribution of the likelihood ratio statistic for equality of the smallest $m$ eigenvalues in high-dimensional PCA, Fujikoshi et al. [9] derived an asymptotic expansion of it, under a limit of $m$ with $m/n \to c \in (0, 1)$ by using [3]. Then they showed by a numerical experiment, that the asymptotic expansion is more accurate than the asymptotic expansion based on the limiting $n \to \infty$ with $d$ fixed.

- Kunitomo derived asymptotic expansions of the distributions of maximum likelihood estimator and the ordinary least square estimator in a linear functional relation model under the condition that a related natural statistics decays fast as $d$ increases [13].
As for the eigenvalues of the sample covariance matrix, we can find many sufficient conditions of their consistency, for the case when the approximate location of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$ of the population covariance matrix are known. For example, the existence of uniformly bounded fourth moment of the population distribution is a sufficient condition if the “order” of the eigenvalues satisfies a special condition, according to [23]. If the population is a kind of linear model with Gaussian errors, the (in)consistency results of eigenvalues and eigenvectors of the sample covariance matrix were obtained by Johnstone–Lu.

In this paper, we prove the sum of the greatest $k(\leq d)$ eigenvalues $\ell_1, \ldots, \ell_k$ of the sample covariance matrix to be consistent in the limit of the sample size $n$ with the dimension $d$ of data depending on $n$, when the population is not necessarily a normal distribution. In other words, in such a situation, we establish

$$\lim_{n \to \infty} P \left( \sum_{i=1}^{k} \ell_i - \sum_{i=1}^{k} \lambda_i > \delta \right) = 0 \quad (\forall \delta > 0).$$

The sum $\sum_{i=1}^{k} \ell_i$ coincides with the supremum over any $(d-k)$-dimensional subspace $H$, of the average of squared distance $\text{dist}(x_1, H)^2$ ($i = 1, \ldots, n$) between the datum $x_i$ and $H$ (See Lemma 1.4). The population distribution under our consideration is a $d$-dimensional Latala distribution $P$, a slight generalization of a normal distribution. By it, we mean there exist $\eta_i \in [1, 2]$ and a real invertible matrix $B$ of size $d$ such that $dP(\xi_1, \ldots, \xi_d)$ is proportional to $\prod_{i=1}^{d} \exp \left( -\eta_i |r_i| \right) d\eta_1 \cdots d\eta_d$ with $(\xi_1, \ldots, \xi_d)^\top$ being $B(\eta_1, \ldots, \eta_d)^\top$. We will call each $r_i$ a rate and $\min_{1 \leq i \leq d} r_i$ the dominant rate. When all the rates are two, the distribution happens to be a normal distribution. The Latala distributions were considered by Latala [15] in relation to a dimension-free upper bound of the Csiszár-type divergence such that the upper bound causes the concentration (Lemma 1.2) of the probability measure.

Here is our main result:

**Theorem 1.1.** Suppose the population is a Latala distribution with the dimension $d = d(n)$ and

$$\liminf_{n \to \infty} d\lambda_1 > 0, \quad \lim_{n \to \infty} \frac{d}{n} = \lim_{n \to \infty} \frac{d^2 \lambda_1^2}{n} \log \frac{n}{d} = 0.$$

Then for every $k(\leq \liminf_{n \to \infty} d)$, the sum $\sum_{i=1}^{k} \ell_i$ converges to the expectation $\sum_{i=1}^{k} \lambda_i$ in probability.

In above mentioned work [12] of Johnstone–Lu, they assumed the largest eigenvalue converges, while we do not assume the convergence and our theorem actually covers such a situation. They used matrix perturbation theories [22] and Gaussian measure concentration inequality [14, 16].

Our proof method is based on empirical process theory [8, 10] (or VC theory [23]) and measure concentration inequalities, and might be novel.

For the proof of the theorem, we first provide non-asymptotic upper bounds of the left- and the right-tail probabilities of the sum $\sum_{i=1}^{k} \ell_i$ in terms of the size $n$ of the sample, the dimension $d$ of data, and the VC dimension of PCA. Then we provide an upper bound of the VC dimension in terms of $d$.

The upper bound of the right-tail probability is due to partly a following measure concentration inequality, which is independent from the dimension $d$.

**Lemma 1.2 (15).** There is $C > 0$ such that for every Latala distribution $P$, every $1$-Lipschitz function $h$ and every $t > 0$,

$$P \left( h - \int h \, dP \geq t \right) \leq \exp \left( -C \min(t, t^2) \right).$$

The remaining part of the proof of our theorem is due to an upper bound of so-called generalization error of statistical learning, which is studied in empirical process theory [8]. The upper bound is a kind of sample complexity of randomized algorithms [19, 7] and/or computational learning [8]. The upper bound is distribution-free, and depends not directly on the dimension of the sample datum but directly on so-called VC dimension of the problem, where VC dimension is a combinatorial number about the expressiveness of the concept class to learn. This learning-theoretic view of PCA may be useful because the concept classes PCA induces are also used as clusters in projective clustering of data [1], such as the class of balls (if $k = 0$), the class of cylinders (if $k = 1$), the class of slabs (if $k = 2$), and so on.

This paper is organized as follows. In Section 2, we recall an upper bound of so-called generalization error of statistical learning when the set of risk functions is unbounded. In Section 3, from the upper bound, we derive an upper bound for the left-tail probability of $\sum_{i=1}^{k} \ell_i$. In Section 4, from the two upper bounds for the left- and the right-tail probabilities, we derive a consistency theorem for the sum of eigenvalues of the sample covariance matrix, the main theorem.

This work was supported by Grant-in-Aid Scientific Research (C:21540105), and the first author thanks Dr. Y. Muroi of graduate school of economics and management, Tohoku university, for informing him about Kunitomo’s paper [13] as a related work.

## 2 Statistical learning with unbounded set of risk functions

The setting of statistical learning theory [23] consists of (i) an i.i.d. sample $z_1, \ldots, z_l \in Z$ subject to an

---

The theory is developed for the analysis of limit theorems of probability theory in Suslin spaces.

2The inequalities are studied by geometers in terms of Ricci curvature of Riemannian manifolds [15], and more generally in measure metric spaces.
unknown distribution $F(z)$, (ii) the class $A$ of hypotheses, and (iii) the loss function $Q : Z \times A \to \mathbb{R}$. The goal of the learning is to estimate an hypothesis $\alpha \in A$ that minimizes a risk $R(\alpha)$ ($\alpha \in A$), which is the expectation $E[Z, \alpha]$ with respect to the data $z$ subject to the distribution $F$. Instead of estimating the unknown risk $R(\alpha)$, we use an empirical risk $R_{\text{emp}}(\alpha) = \sum_{i=1}^n Q(z_i, \alpha)/n$ for the i.i.d. sample $z_1, \cdots, z_n$, and minimize it. This is the so-called empirical risk minimization. As for the consistency of the empirical risk, a following proposition is developed. Define

$$V_p(\varepsilon) := \left(1 - \frac{\log \varepsilon}{p - \sqrt{p(p-1)}}\right)^{1/p},$$

$$\mathbb{E}_z = 1 + 2/p, \quad q := 2(1 - 1/p). \quad (1)$$

**Proposition 2.1** (23 Theorem 5.4). Let $\{Q(\cdot, \alpha) : \alpha \in A\}$ be a not necessarily uniformly bounded class of nonnegative functions. Then for any $\varepsilon > 0$ and for any $p \in (1, 2]$, we have

$$\sup_{\alpha \in A} \mathbb{E}[Q(z, \alpha)]^{1/2} \geq \varepsilon V_p(\varepsilon),$$

with probability at most $4 \exp\left(-\mathcal{C}(A)(2n) - \varepsilon^2 n^q/2^u\right)$. Here $\mathcal{C}(A)$ is $\{\{z \in Z : Q(z, \alpha) < r\} : \alpha \in A, r \geq 0\}$, and $\mathcal{C}(A)(n)$ is called the growth function of $\mathcal{C}(A)$, defined as

$$G_{\mathcal{C}(A)}(n) := \log \sup_{X \in \mathcal{C}(A)} \#\{X \cap C : C \in \mathcal{C}(A)\}.$$

The growth function is bounded from above by using so-called VC dimension, a combinatorial dimension defined as follows: Let $C$ be a nonempty class of subsets of $Z$. Then for a finite subset $X$ of $Z$, a subset $X \subseteq Y$ is said to be cut out by $C \in C$, if $X \cap C = Y$. We say $X$ is shattered by $C$, if every subset of $X$ is cut out by some $C \in C$. The VC dimension of $C$ is the supremum of the cardinality of a subset $X$ of $Z$ such that $X$ is shattered by $C$, and denoted by $\text{VCdim}(C)$.

**Proposition 2.2** (23). For $v = \text{VCdim}(\mathcal{C}(A))$, if $n > v$ then $G_{\mathcal{C}(A)}(n)$ is no more than $v(\log(n/v) + 1)$, else $G_{\mathcal{C}(A)}(n)$ is $n \log 2$.

### 3 Generalization error of PCA

A PCA is formulated as a statistical learning problem with the set of hypotheses being the class of $k$-dimensional affine subspaces $H$ and the loss function $Q(x, H)$ being the squared distance of $x$ from $H$.

**Definition 3.1.** For nonnegative integers $k < d$, $\mathcal{C}_d^k$ is, by definition, the class of $\{x \in \mathbb{R}^d : \text{dist}(x, H) < r\}$ such that $r \geq 0$ and $H$ is any $k$-dimensional affine subspace of $\mathbb{R}^d$. By restricting $H$ to be linear, we obtain a subclass $\mathcal{D}_d^k$ (“manifold”) induced by PCA (2). It is by Milnor–Thom upper bound (20) of the number of the connected components of algebraic varieties.

**Theorem 3.2** (2). There exists $c > 0$ such that for any nonnegative integers $k < d$, we have $\text{VCdim}(\mathcal{C}_d^k)$ is at most $c(k + 1)(d - k + 1)$.

By $\mathcal{D}_d^k \subseteq \mathcal{C}_d^k$, we have $\text{VCdim}(\mathcal{D}_d^k) \leq \text{VCdim}(\mathcal{C}_d^k) \leq c(k + 1)(d - k + 1)$.

Of independent interest in discrete geometry is that the upper bound $(k + 1)(d - k + 1)$ of VCdim($\mathcal{C}_d^k$) is asymptotically tight.

**Definition 3.3.** Fix $0 < k < d$. Let $V_{d,k}$ be the set $\{T \in \mathbb{R}^{d \times k} : T^T T = I_k\}$, and for the random vector $x \in \mathbb{R}^d$ subject to the population distribution, let

$$M_k(p) := \sup_{T \in V_{d,k}} \mathbb{E}[\|T^T x\|^p]^{1/2} \quad (p \in (1, 2]).$$

**Lemma 3.4.** If each component of a data $x_1$ in the sample has the finite fourth moment, then there is $p \in (1, 2]$ such that

$$M_k(p) < \infty, \quad M_{d-k}(p) < \infty, \quad p \in (1, 2]. \quad (2)$$

**Definition 3.5.** Choose $p$ such that (2) holds. Let $V_p(\varepsilon), u, q$ be as in (1).

For any integers $d' \leq d$ and any $d'$-dimensional subspace $H$ of $\mathbb{R}^d$, there is $T \in V_{d,k}$ such that $H = \text{Im} T$ if $k = d'$, while $H = \ker T^T$ if $k = d - d'$. These give rise to two statistical learning problems for each PCA problem that approximates $d$-dimensional data by $k$-dimensional space.

**Definition 3.6.** The class of hypotheses is $V_{d,k}$. The loss functions are $Q, Q'$ such that

$$Q(x, T) = \text{dist}(x, \ker T)^2 = \|T^T x\|^2,$$

$$Q'(x, T) = \text{dist}(x, \text{Im} T)^2 = \|x\|^2 - \|T^T x\|^2.$$

Then, the corresponding risks $R(T), R'(T)$ and corresponding empirical risks $R_{\text{emp}}(T), R'_{\text{emp}}(T)$ are:

$$R(T) := \mathbb{E}[\|T^T x\|^2] = \text{tr}(T^T \Sigma T),$$

$$R_{\text{emp}}(T) := \frac{1}{n} \sum_{i=1}^n \|T^T x_i\|^2 = \text{tr}(T^T \Sigma T),$$

$$R'(T) := \mathbb{E}[\|x\|^2 - \|T^T x\|^2] = \mathbb{E}[\|x\|^2] - \text{tr}(T^T \Sigma T),$$

$$R'_{\text{emp}}(T) := \frac{1}{n} \sum_{i=1}^n \|x_i\|^2 - \|T^T x_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|x_i\|^2 - \text{tr}(T^T \Sigma T).$$

Then it is ready to derive relations between (empirical) risks and eigenvalues of the (sample) covariance matrix.
Lemma 3.7. For the first statistical learning problem involving the risk function \( Q \), we have
\[
\sup_{T \in V_{d,k}} R(T) = \sum_{i=1}^{k} \lambda_i, \quad \sup_{T \in V_{d,k}} R_{\text{emp}}(T) = \sum_{i=1}^{k} \ell_i. \tag{3}
\]

Then, for the second statistical learning problem involving the risk function \( Q' \), two infima \( \inf_{T \in V_{d,k}} R(T) \) and \( \inf_{T \in V_{d,k}} R'_{\text{emp}}(T) \) are \( E[|x|^2] - \sum_{i=1}^{k} \lambda_i \) and \( \sum_{i=1}^{n} ||x_i||^2/n - \sum_{i=1}^{k} \ell_i \), respectively.

By using Proposition 2.1 we can prove following Theorems. Recall that \( p \) is defined in Definition 5.6, which is well-defined if each component of a random vector subject to the population distribution has the finite fourth moment (Lemma 3.4).

Theorem 3.8 (left-tail probability of \( \sum_{i=1}^{k} \ell_i \)). Let \( \varepsilon > 0 \). Then

\[
\sum_{i=1}^{k} \lambda_i - \sum_{i=1}^{k} \ell_i \varepsilon \geq V_p(\varepsilon) M_k(p)
\]

with probability at most \( 4 \exp \left( G_{D_d}^q(2n) - \varepsilon^2 n^q/2^q \right) \).

Theorem 3.9 (right-tail probability of \( \sum_{i=1}^{k} \ell_i \)). Let \( \varepsilon > 0 \) and \( \delta > 0 \). Then \( \sum_{i=1}^{k} \ell_i - \sum_{i=1}^{k} \lambda_i \) is greater than or equal to \( \varepsilon V_p(\varepsilon) M_{d-k}(p)+\delta \) with probability at most \( 4 \exp \left( G_{D_d}^q(2n) - \varepsilon^2 n^q/2^q \right) \) plus the probability of \( \sum_{i=1}^{n} ||x_i||^2/n - E[||x||^2] > \delta \).

4 Measure concentration

We will provide an upper bound of the last probability in Theorem 3.9. Let us denote the operator norm of an invertible matrix \( B \) by \( ||B|| \).

Lemma 4.1. Suppose

- \( y \) is any \( d \)-dimensional random vector consisting of independent components, each having variance greater than \( c > 0 \) and zero mean.
- \( B \) is an invertible matrix of size \( d \), and the covariance matrix of \( x = By \) has \( \lambda_1 \) as the greatest eigenvalue.

Then \( ||B^{-1}|| \leq \sqrt{\lambda_1/c} \).

By a measure concentration inequality (Lemma 4.2) and Jensen’s inequality, we have a following:

Theorem 4.2. For any Latala population with dominant rate \( r \), there are positive constants \( C, K \) not depending on \( d \) or the rates such that for each \( \delta > 0 \), \( \sum_{i=1}^{n} ||x_i||^2/n - E[||x||^2] \) is greater than \( \delta \) with probability at most \( 4 \exp \left( -K \left( \frac{\alpha(\varepsilon)}{C} \right)^{r/2} \rho^r \right) \) where

\[
\rho := \sqrt{\frac{n}{d\lambda_1}} \frac{\delta}{\sqrt{1 + \frac{d}{\lambda_1} + 1}}.
\]

Combined with Theorem 3.9 we can provide an upper bound for the right-tail probability of \( \sum_{i=1}^{k} \ell_i \).

Corollary 4.3. Suppose the population distribution is a Latala distribution with dominant rate \( r \). Let \( \varepsilon > 0 \) and \( \delta > 0 \). Then there are positive constants \( C, K \) not depending on \( d \) or the rates such that an inequality

\[
\sum_{i=1}^{k} \ell_i - \sum_{i=1}^{k} \lambda_i \geq \varepsilon V_p(\varepsilon) M_{d-k}(p)+\delta
\]

holds with probability at most

\[
4 \exp \left( G_{D_d}^q(2n) - \varepsilon^2 n^q/2^q \right) + \exp \left[ -K \left( \frac{\alpha(\varepsilon)}{C} \right)^{r/2} \rho^r \right].
\]

By a similar method with introducing another parameter other than \( \delta \), we can derive an upper bound for the left- and the right-tail probabilities of the condition numbers of the random matrices, which are useful in analyzing optimization algorithms for randomized/noisy data [21].

5 Consistency of eigenvalues of sample covariance matrix

For the upper bounds of the left- and the right-tail probabilities of \( \sum_{i=1}^{k} \ell_i \) to decrease, we consider a following lemma:

Lemma 5.1. Suppose the dimension \( d = d(n) \) of a sample datum depends on the sample size \( n \),

\[
\lim_{n \to \infty} \frac{d}{n} = \lim_{n \to \infty} \frac{d^2 \lambda_1^2}{n^{q}} \log \frac{n}{d} = 0.
\]

Moreover assume

\[
(\ast) \text{ there exists } M(p) > 0 \text{ depending only on } p \text{ such that for all nonnegative integer } \ell < d \text{ it holds that } M_{\ell}(p) < M(p) \lambda_d d.
\]

Then, as \( n \to \infty \), \( G_{D_d}^q(2n) - \varepsilon^2 n^q/2^q \to -\infty \) and \( \rho \to \infty \) \( (\varepsilon > 0, 0 \leq \ell < d) \).

As for a sufficient condition for (\ast) of the above Lemma, we have a following:

Lemma 5.2. Under the assumptions of Lemma 4.2 and \( \sup_{j=1,2,...} (E[|y_j|^2])^{1/p} < \infty \), there exists a positive \( M(p) \) such that for any nonnegative integer \( k \leq d \), we have \( M_k(p) \leq M(p) \lambda_d d \).

Because every Latala distribution has any moment, the two lemmas implies our main theorem (Theorem 4.1).
6 Concluding remark

Easy modification and approximation of statistical inferences in multivariate analysis may depend on estimators which is biased and inconsistent. But statistical inferences in multivariate analysis, from classical ones to modern ones, are employed in various critical fields such as medicine, pharmacy, and bioinformatics. So it is important to prove the consistencies of the inferences. We see that there are two methods to provide upper bounds for the error of such estimators.

The first method is empirical process theory [5], which supplies a distribution-free, non-asymptotic upper bound of the estimator’s error in terms of VC dimension and the sample size. Once we can realize that a statistical inference is very much related to discrete geometry, then we can relatively easily give an upper bound for the VC dimension of the inference in terms of the dimension of data and so on, because the VC dimension is not smaller than the number of connected components of some real algebraic varieties, and the upper bound of the latter is easily given by [5].

The second useful method is measure concentration theory [15, 14, 18], which supplies a non-asymptotic upper bound of random vectors, which is independent of the dimension of the data. If the probability density function of the population has a compact support, then we can use many concentration inequalities, which are survey in [17].

For the research on random matrix, the graph theoretical gadgets used in Geman [10] to study the great- est eigenvalue of Wishart matrix is recently developed impressively by Kuriki-Numata [14].

According to [3], the problem of convergence rates of empirical spectral distributions of large dimensional random matrices (LDRM) had been open for decades since no suitable tools were found, and there is a good deal of evidence that the behavior of LDRM is asymptotically distribution-free. We hope our use of empirical process theory, which is distribution-free, is related to the problem.

References