情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

VC理論とWishart行列の固有値の和の集中不等式

VC Theory and a Concentration Inequality for Sums of Eigenvalues of Wishart Matrix

上野 康隆 *　　赤間 陽二*
Yasutaka Uwano　　Yohji Akama

Abstract: Let d-dimensional column vectors $x_1, \ldots, x_n$ be an i.i.d. sample drawn from the d-dimensional standard normal distribution. Let S be $\sum_{i=1}^n x_i x_i^\top / n$. The left and the right tail probabilities for the sum of *any* k eigenvalues of S is uniformly evaluated *non-asymptotically* from above, by using upper bound of the VC dimensions of principal component analysis and by using a Vapnik's theorem of generalization errors in empirical risk minimization. For the right tail probability, we represent a subspace with the kernel of a linear mapping and then employ a concentration inequality for the chi square distributions.

1 Introduction

Let $x_1, \ldots, x_n$ be independently distributed, each subject to d-dimensional normal distribution $N(0, \Sigma)$. Then the distribution of $\sum_{i=1}^n x_i x_i^\top$ is defined to be *Wishart distribution*, denoted by $W(\Sigma, n)$. If $\Sigma = E_d$, the identity matrix E_d of size d, then so-called data covariance matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ of the sample is subject to $W(E_d/n, n)$.

Johnstone [6] proved that for a matrix subject to $W(E_d, n)$, if the largest eigenvalue is appropriately centered and scaled, then the distribution approaches to the Tracy-Widom law of order 1, as n, d goes to infinity with n/d fixed $\gamma \geq 1$. On the other hand, for the *data covariance matrix* $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$, as $n \to \infty$ with d being fixed, the sum of any k eigenvalues of S tends to k almost surely, because the law of large numbers guarantees that S converges to the identity matrix E_d almost surely. In terms of principal component analysis (PCA), the sum of the largest k eigenvalues of S is the sum of variances of principal component and the sum of the square distances of data and the approximate affine subspace.

Below, the left and the right tail probabilities for the sum of *any* k eigenvalues of the data covariance matrix S is uniformly evaluated *non-asymptotically* from above, by using upper bound of the VC dimensions of principal component analysis and by using a theorem [9, (5.43)] of Vapnik's statistical learning theory. For the right tail probability, we represent a subspace with the kernel of a linear mapping and then employ a concentration inequality [7, (5.1)] for the chi square distributions.

Johnstone's result [6] is satisfied if the number n of observations and the dimensions p are large enough with n/d fixed. On the other hand, our results are useful in the case that $n/d = \Omega(n^{\frac{1}{2}+\epsilon})$, i.e. $d/n = O(n^{-(\frac{1}{2}+\epsilon)})$, where ϵ is any positive number, especially, in the case that d is fixed and n is large.

For the largest eigenvalue λ_1 of a symmetric random matrix whose entries are independent random vari-

*東北大学理学研究科数学専攻, 980-8578 仙台市青葉区 荒巻字青葉, e-mail sa8m07@math.tohoku.ac.jp,
Mathematical Institute, Tohoku University Sendai Miyagi JAPAN, 980-8578

ables with *absolute value bounded by 1*, the sub-gaussian evaluation of the right tail of λ_1 is derived in [1] from Talagrand's inequality. Our result is for Gaussian random variables, which vary from $-\infty$ to ∞.

This paper is organized as follows: In the next section, we review VC theory. In Section 3, we relate the sum of eigenvalues of the data covariance matrix to the statistical learning that formulates PCA. In Section 4, we provide an upper bound of VC dimension of the PCA. In Section 5, we present a concentration inequality for the sum of eigenvalues of the data covariance matrix for an i.i.d. sample from the multi-dimensional standard normal distribution. In the final section, we mention future work, which are hopefully related to concentration inequality and model selection.

2 VC theory

In the framework of statistical learning theory [9], a learning model in general consists of (i) an unknown distribution $F(z)$ of training data z drawn from a space Z, (ii) a class Λ of hypotheses, (iii) a *loss function* $Q : Z \times \Lambda \to \mathbb{R}$. Here $Q(z, \alpha)$ stands for the loss of training data z against a hypothesis α. The *risk* of $\alpha \in \Lambda$ is defined to be $R(\alpha) = \mathrm{E}_z[Q(z, \alpha)]$. The goal of learning is to estimate

$$\alpha_0 \in \Lambda \text{ s.t. } R(\alpha_0) = \min_{\alpha \in \Lambda} R(\alpha)$$

from training data $z_1, \cdots, z_n$ which are independently drawn from the distribution F.

Proposition 1 ([9, (5.43)]). *Let $\{Q(x, \alpha) : \alpha \in \Lambda\}$ be any class of unbounded class of nonnegative, functions. Then for any $\alpha \in \Lambda$, with probability greater than or equal to $1 - \eta$, it holds that*

$$R(\alpha) - R_{\mathrm{emp}}(\alpha) < R(\alpha)\tau(p)\left(\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}\right)^{1/p}\varepsilon.$$

Here $p > 2$ is such that

$$\sup_{\alpha \in \Lambda} \frac{\mathrm{E}_x\left[Q(x, \alpha)^p\right]^{1/p}}{\mathrm{E}_x\left[Q(x, \alpha)\right]} < \tau(p),$$

and

$$\eta := 4 \exp\left\{\left(\frac{G_\Lambda(2n)}{n} - \frac{\varepsilon^2}{4}\right) n\right\}.$$

$G_\Lambda(n)$ is the so-called growth function *for Λ, and*

$$G_\Lambda(n) \begin{cases} = n \log 2 & (n \le v) \\ \le v(\log \frac{n}{v} + 1) & (n > v), \end{cases}$$

where v is the VC dimension of the class of $\{x \in Z : Q(x, \alpha) \ge r\}$ such that $\alpha \in \Lambda$ and $r \in \mathbb{R}$.

Let $\mathcal{C}$ be a nonempty class of subset of Z. We say a finite subset X of Z is *shattered* by $\mathcal{C}$, if $\{X \cap C : C \in \mathcal{C}\}$ is *the* class of subsets of X. By *the VC dimension* of the class $\mathcal{C}$, we mean the supremum of the cardinality of a set $X \subset Z$ shattered by $\mathcal{C}$. Important properties on the VC dimension in the study of empirical process(=statistical learning) are found in [5].

3 Eigenvalues of data covariance matrix and empirical risks

First, we relate the eigenvalues of the data covariance matrix S, to the empirical risk of a statistical learning. Put Λ to be the set of $d \times k$ real matrices T such that $T^\top T = E_k$. We represent a $(d-k)$-dimensional subspace H by any $T \in \Lambda$ such that $H = \ker T$. For any $x \in \mathbb{R}^d$ and any $T \in \Lambda$, we define a loss function $Q(x, T)$ to be $\mathrm{dist}(x, \ker T)^2 = \|T^\top x\|^2$. On the other hand, we represent a k-dimensional subspace K by any $T \in \Lambda$ such that $K = \mathrm{Im}\, T$. For any $x \in \mathbb{R}^d$ and any $T \in \Lambda$, we define another loss function $Q'(x, T)$ to be $\mathrm{dist}(x, \mathrm{Im}\, T)^2$.

The empirical risks caused by $T \in \Lambda$ are

$$\begin{aligned} R_{\mathrm{emp}}(T) &= \frac{1}{n}\sum_{i=1}^n \|T^\top x_i\|^2, \\ R'_{\mathrm{emp}}(T) &= \frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - R_{\mathrm{emp}}(T) \\ &= \frac{1}{n}\sum_{i=1}^n x_i^\top (E_d - TT^\top) x_i. \end{aligned}$$

If T consists of the k orthonormal eigenvectors of the data covariance matrix S, then $R_{\mathrm{emp}}(T)$ is the sum X

of the k corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$ of S and $R'_{\mathrm{emp}}(T)$ is $\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 - X$.

4 VC dimension of PCA formulated as statistical learning

Put $\mathcal{C}_k^d$ be the class of $\{x \in \mathbb{R}^d : \mathrm{dist}(x, H) < r\}$ such that H is any k-dimensional affine subspace and the r is any positive real number.

Theorem 1. *There exists a positive constant c such that the VC dimension of $\mathcal{C}_k^d$ is less than $c(k+1)(d-k+1)$.*

Corollary 1. *Let $\mathcal{D}_k^d$ denote the class of $\{x \in \mathbb{R}^d : \mathrm{dist}(x, H) < r\}$ such that H is any k-dimensional subspace and the r is any positive real number. Then the VC dimension of $\mathcal{D}_k^d$ is less than $c(k+1)(d-k+1)$, where c is an absolute positive constant.*

The proof of the theorem uses a fact that any linear subspace is represented as a kernel and an image, as well as rather a standard evaluation of the number of sign sequences arising from an algebraic variety.

We prove this proposition by following Basu-Pollack-Roy's argument [2]:

For an element $a \in \mathbb{R}$,

$$\mathrm{sgn}(a) := \begin{cases} 0 & \text{if } a = 0, \\ 1 & \text{if } a > 0, \\ -1 & \text{if } a < 0. \end{cases}$$

Let $\mathcal{Q}$ and $\mathcal{P}$ be finite subsets of $\mathbb{R}[x_1, \ldots, x_m]$. A *sign condition on* $\mathcal{P}$ is an element of $\{0, 1, -1\}^{\mathcal{P}}$.

The *realization of the sign condition* σ *over* $\mathcal{Q}$, $\mathcal{R}(\sigma, \mathcal{Q})$, is the real semi-algebraic set

$$\{x \in \mathbb{R}^m : g(x) = 0 \text{ for all } g \in \mathcal{Q}, \text{ and } \mathrm{sgn}(P(x)) = \sigma(P) \text{ for all } P \in \mathcal{P}\}.$$

Let $b_i(\sigma, \mathcal{Q})$ denote the i-th Betti number of $\mathcal{R}(\sigma, \mathcal{Q})$, i.e., the dimension of the i-th singular homology group of $\mathcal{R}(\sigma, \mathcal{Q})$ as a $\mathbb{Q}$ vector space, and let $b_i(\mathcal{Q}, \mathcal{P}) = \sum_\sigma b_i(\sigma, \mathcal{Q})$. Especially, $b_0(\sigma, \mathcal{Q})$ is the total number of semi-algebraically connected components of the realizations of all realizable sign conditions of $\mathcal{P}$ over $\mathcal{Q}$.

We write $b_i(d, m, L, s)$ for the maximum of $b_i(\mathcal{Q}, \mathcal{P})$ over all $\mathcal{Q}$, $\mathcal{P}$ where $\mathcal{Q}$ and $\mathcal{P}$ are finite subsets of $\mathbb{R}[x_1, \ldots, x_m]$, whose elements have degree at most $d \geq 1$, the cardinality of $\mathcal{P}$ is s, and the algebraic set $\{x \in \mathbb{R}^m : g(x) = 0 \text{ for all } g \in \mathcal{Q}\}$ has real dimension L.

Proposition 2 ([2])**.**

$$b_i(d, m, L, s) \leq d(2d-1)^{m-1} \sum_{j=0}^{L-i} 4^j \binom{s}{j}.$$

Let $(\mathcal{C}_k^d)^\flat$ be the class of open sets $\{x \in \mathbb{R}^d : \mathrm{dist}(x, H) < r\} \in \mathcal{C}_k^d$ such that the k-dimensional affine subspace H intersects with the $(d-k)$-dimensional subspace $x_1 = \cdots = x_k = 0$ at exactly one point. Note that $\mathrm{VCdim}(\mathcal{C}_k^d) = \mathrm{VCdim}((\mathcal{C}_k^d)^\flat)$, because if $\mathcal{C}_k^d$ shatters a finite set then $(\mathcal{C}_k^d)^\flat$ does the set by appropriate perturbation.

Lemma 1. *Let $L = (k+1)(d-k)+1$. Then, there exist a positive integer $m \leq 2L$, an L-dimensional smooth submanifold V in $\mathbb{R}^m$ defined by $m-L$ quadratic equations in m variables, and $\Phi\colon V \to \mathcal{C}_k^d$ with the following properties:*

(a) $\mathrm{VCdim}(\mathcal{C}_k^d) = \mathrm{VCdim}(\Phi(V))$*; and*

(b) *for each $p \in \mathbb{R}^d$, there exists a quadratic m-variate real polynomial f_p such that for all $x \in V$, $f_p(x) > 0$ if p is in $\Phi(x)$, while $f_p(x) < 0$ if p is not in the closure of $\Phi(x)$.*

Proof. First, we consider the case $k \geq d/2$. Let $m = (d-k)(d+1)+1$. Then it is indeed $m \leq 2L$. For $(F, b, r) \in \mathbb{R}^m$ where F is a $d \times (d-k)$ real matrix, $b \in \mathbb{R}^{d-k}$ and $r \in \mathbb{R}$, we consider a system of $(d-k)^2 = m - L$ quadratic equations

$$\sum_u F_{ui}F_{uj} - \delta_{ij} = 0 \, (1 \leq i \leq j \leq d-k),$$

$$F_{i+k,\,j} = 0 \, (1 \leq i < j \leq d-k).$$

This defines an L-dimensional smooth submanifold V of $\mathbb{R}^m$ by the implicit function theorem.

For $(F, b, r) \in V$ where $F \in \mathbb{R}^{d\times(d-k)}, b \in \mathbb{R}^{d-k}$ and $r \in \mathbb{R}$, define $\Phi(F, b, r) \in \mathcal{C}_k^d$ to be the set of points whose distance from a k-dimensional affine space

$$\{\, z \in \mathbb{R}^d : (F^\top)z = b \,\} \tag{1}$$

is less than $|r|$. Then Φ satisfies the property (a), since $\Phi(V) \subseteq \mathcal{C}_k^d$ contains $(\mathcal{C}_k^d)^\flat$. Moreover, for $p \in \mathbb{R}^d$, define f_p by

$$f_p(F, b, r) = r^2 - \|(F^\top)p - b\|^2.$$

This satisfies the property (b), because $\|(F^\top)p-b\|^2$ is equal to the square of the distance from p to the affine subspace (1).

Next we consider the case $k < d/2$. Let $m = dk + d + 1$. Then it is indeed $m \le 2L$. For $(E, t, r) \in \mathbb{R}^m$ where E is a $d \times k$ real matrix, $t \in \mathbb{R}^d$ and $r \in \mathbb{R}$, we consider a system of $k + k^2 = m - L$ quadratic equations, consisting of k equations

$$\sum_u t_u E_{uj} = 0 \,(1 \le j \le k) \tag{2}$$

and k^2 equations

$$\sum_u E_{ui}E_{uj} - \delta_{ij} = 0 \quad (1 \le i \le j \le k),$$
$$E_{ij} = 0 \quad (1 \le i < j \le k).$$

The system defines an L-dimensional smooth submanifold V of $\mathbb{R}^m$, by the implicit function theorem. For any $(E, t, r) \in V$ with $E \in \mathbb{R}^{d\times k}, t \in \mathbb{R}^d, r \in \mathbb{R}$, define $\Phi(E, t, r)$ to be the set of points whose distance from

$$\{\, Ex + t : x \in \mathbb{R}^k \,\} \tag{3}$$

is less than $|r|$. Then Φ satisfies the property (a), since $\Phi(V)$ contains $(\mathcal{C}_k^d)^\flat$. Moreover, for $p \in \mathbb{R}^d$, define f_p by

$$f_p(E, t, r) = r^2 - \|p - t\|^2 + \|(p^\top)E\|^2.$$

Then f_p is clearly quadratic. By (2), we have $\|p - t\|^2 - \|(p^\top)E\|^2 = \|p - t\|^2 - \|(p - t)^\top E\|^2$, which is equal to the square of the distance from p to the affine subspace (3). Thus we have the property (b). □

Now we will complete the proof of the upper bound.

Proof of Theorem 1. Let m, L, V, Φ be as in the previous lemma. Take a set $\mathcal{Q}$ consisting of quadratic m-variate real polynomials $g_1, \dots, g_{m-L}$ so that equations $g_1 = \cdots = g_{m-L} = 0$ define V. Let $\{p_1, \dots, p_s\} \subseteq \mathbb{R}^d$ be a set shattered by $\mathcal{C}_k^d$. By (a) of the previous lemma, it is shattered by $\Phi(V)$. If $s \le m$, then because the previous lemma implies $m \le 2L$, we have $s \le m \le 2L$ as desired. If $s > m$, then put $\mathcal{P} := \{f_{p_1}, \dots, f_{p_s}\}$. Because $\{p_1, \dots, p_s\}$ is shattered, $2^s \le \#\{\sigma \in \{-1, 1\}^{\mathcal{P}} : \mathcal{R}(\sigma, \mathcal{Q}) \ne \emptyset\}$. Then $2^s \le b_0(\mathcal{Q}, \mathcal{P}) \le b_0(2, m, L, s)$ by the definition. From Proposition 2, we have $2^s \le d(2d-1)^{m-1}\sum_{j=0}^{L} 4^j \binom{s}{j}$ which is less than or equal to $2 \cdot 3^{2L-1} \cdot 4^L \sum_{j=0}^{L} \binom{s}{j} \le 36^L \left(\frac{\mathrm{e}s}{L}\right)^L$. This gives $2^{s/L} \le 36\mathrm{e}(s/L)$, or $s/L \le c$ where c is large enough. □

5 The concentration inequalities

Let $x_1, \dots, x_n$ be an i.i.d. sample drawn from $N(0, E_d)$ and let $\lambda_1, \dots, \lambda_k$ be eigenvalues of the data covariance $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Let $T \in \mathbb{R}^{d\times k}$ consist of the corresponding orthonormal eigenvectors, as in Section 3. For the loss functions given there, the risks caused by T are $R(T) = \mathrm{E}[\|T^\top x\|^2]$ and $R'(T) = \mathrm{E}[x^\top(E_d - TT^\top)x]$. Because $TT^\top$ is an orthogonal projection of rank k, the loss functions are random variables subject to chi square distributions:

$$\|T^\top x\|^2 \sim \chi_k^2, \quad \|x\|^2 - \|T^\top x\|^2 \sim \chi_{d-k}^2, \tag{4}$$

where χ_m^2 is the chi square distribution with degree m of freedom. So $R(T) = k$ and $R'(T) = d - k$. By this and the last paragraph of Section 3,

$$k - (\lambda_1 + \cdots + \lambda_k) = R(T) - R_{\mathrm{emp}}(T), \tag{5}$$

and

$$(\lambda_1 + \cdots + \lambda_k) - k$$
$$= R'(T) - R'_{\mathrm{emp}}(T) + \left(\frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - d\right). \tag{6}$$

By applying Proposition 1 to R and R', we have inequalities for left and right tail probabilities of the sum of any k eigenvalues of S. But for the last term in the

inequality (6), we use a following inequality [7, (5.1)] for right tail probability of the chi square distribution:

$$P\left(Y \geq \left(\sqrt{d}+\sqrt{2y}\right)^2\right) \leq \mathrm{e}^{-y} \quad (Y \sim \chi^2_d,\ y>0), \tag{7}$$

which is proved by using Gaussian logarithmic Sobolev inequality [7, Theorem 3.4].

The p-th noncentral moment of the chi square distribution of degree k of freedom is written as $m(k,p)$, which is $k(k+2)(k+4)\cdots(k+2p-2)$.

Theorem 2. *Let $x_1,\ldots,x_n$ be an i.i.d. sample drawn from the d-dimensional standard normal distribution, $\lambda_1,\ldots,\lambda_k$ $(k \leq d)$ be any eigenvalues of the data covariance $d\times d$ matrix $\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)$, $p>2$, $\varepsilon>0$ and $\delta>0$. Then, the left tail probability of $\sum_{i=1}^k \lambda_i$ satisfies the following:*

$$P\left(k-\sum_{i=1}^k \lambda_i \geq \varepsilon\left(\frac{m(k,p)}{2}\left(\frac{p-1}{p-2}\right)^{p-1}\right)^{1/p}\right) \leq 4\exp\left\{\left(\frac{G_{\mathcal{D}^d_{d-k}}(2n)}{n}-\frac{\varepsilon^2}{4}\right)n\right\}.$$

The right tail probability of $\sum_{i=1}^k \lambda_i$ satisfies the following:

$$P\left(\sum_{i=1}^k \lambda_i - k \geq \varepsilon\left(\frac{m(d-k,p)}{2}\left(\frac{p-1}{p-2}\right)^{p-1}\right)^{1/p}+\delta\right) \leq 4\exp\left\{\left(\frac{G_{\mathcal{D}^d_k}(2n)}{n}-\frac{\varepsilon^2}{4}\right)n\right\} + \exp\left(-\frac{1}{2}nd\left(\sqrt{1+\frac{\delta}{d}}-1\right)^2\right).$$

In particular, if $n > v/2$ with v being $c(k+1)(d-k+1)$ where c is an absolute positive constant, then the inequalities can be made concrete by replacing the two growth functions $G_{\mathcal{D}^d_{d-k}}(2n)$ and $G_{\mathcal{D}^d_k}(2n)$ in the inequalities with $v(\log\frac{2n}{v}+1)$.

Proof. As for the left tail probability, in Proposition 1, as the loss function $Q(T,x)$ is subject to χ^2_k by (4), we can take $\tau(p) = m(k,p)^{1/p}/k$ and thus $k-\sum_{i=1}^k\lambda_i$, which is $R(T)-R_{\mathrm{emp}}(T)$ by (5), exceeds

$$\varepsilon\left(\frac{m(k,p)}{2}\left(\frac{p-1}{p-2}\right)^{p-1}\right)^{1/p},$$

with probability at most

$$4\exp\left(\left(\frac{G_{\mathcal{D}^d_{d-k}}(2n)}{n}-\frac{\varepsilon^2}{4}\right)n\right). \tag{8}$$

As for the right tail probability, in Proposition 1, as the loss function $Q'(T,x)$ is subject to χ^2_{d-k} by (4), we can take $\tau(p) = m(d-k,p)^{1/p}/(d-k)$ and thus $\left(\sum_{i=1}^k\lambda_i - k\right) - \left(\frac{1}{n}\sum_{i=1}^n\|x_i\|^2 - d\right)$, which is $R'(T) - R'_{\mathrm{emp}}(T)$ by (6), exceeds

$$a := \varepsilon\left(\frac{m(d-k,p)}{2}\left(\frac{p-1}{p-2}\right)^{p-1}\right)^{1/p},$$

with probability at most $4\exp\left(\left(G_{\mathcal{D}^d_k(2n)}/n-\varepsilon^2/4\right)n\right)$.

But by taking $Y = \sum_{i=1}^n\|x_i\|^2 \sim \chi^2_{nd}$ and $y = \left(\sqrt{nd+n\delta}-\sqrt{nd}\right)^2/2$ in (7), we have $\frac{1}{n}\sum_{i=1}^n\|x_i\|^2 - d \geq \delta$ with probability at most

$$b := \exp\left(-\frac{nd}{2}\left(\sqrt{1+\frac{\delta}{d}}-1\right)^2\right).$$

Therefore either $\sum_{i=1}^k\lambda_i - k \geq a$ or $\frac{1}{n}\sum_{i=1}^n\|x_i\|^2 - d \geq \delta$ holds with probability at most

$$P\left(\sum_{i=1}^k\lambda_i - k - \left(\frac{1}{n}\sum_{i=1}^n\|x_i\|^2 - d\right) \geq a\right) + P\left(\frac{1}{n}\sum_{i=1}^n\|x_i\|^2 - d \geq \delta\right).$$

But the former summand is less than or equal to (8) with k replaced by $d-k$, while the latter is less than or equal to b. □

6 Future work: concentration and model selection

Some mathematicians may be interested in how our approach is related to (1) papers of the local theory of Banach spaces on concentration of measure that is directly relevant (e.g. [8]), and (2) to the work on Talagrand's work on concentration of measure. Talagrand's inequalities for concentration of measure are recently employed in [7, Chapter 8], for statistical learning problems with the class of loss functions being uniformly bounded, as follows:

1. Bousquet's version of Talagrand's concentration inequality for empirical process is used to derive a new general upper bound of the difference between the expected risk and the empirical risk.

2. A concentration inequality is used to analyze Vapnik's structural risk minimization [9], a model selection method in terms of VC dimensions.

PCA has the *unbounded* class of loss functions $x \in \mathbb{R}^d \mapsto \mathrm{dist}(x, H)^2$ where H is any k-dimensional affine subspace. We hope similar concentration inequalities which improves (1) previous Theorem for PCA and (2) model selection (i.e., selecting k) for PCA.

This research is encouraged by a researcher who majors in concentration inequality and/or consistency of principal component analysis. We read in [7],

> "Since the impressive works of Talagrand, concentration inequalities have been recognized as fundamental tools in several domains such as geometry of Banach spaces or random combinatorics. They also turn out to be essential tools to develop a non-asymptotic theory in statistics, exactly as the central limit theorem and large deviations are known to play a central part in the asymptotic theory. An overview of a non-asymptotic theory for model selection is given here and some selected applications to variable selection, change points detection and statistical learning are discussed. ...."

We hope our work is connected to such applications and so on.

References

[1] Noga Alon, Michael Krivelevich, and Van H. Vu. On the concentration of eigenvalues of random symmetric matrices. *Israel J. Math.*, Vol. 131, pp. 259–267, 2002.

[2] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. On the Betti numbers of sign conditions. *Proc. Amer. Math. Soc.*, Vol. 133, No. 4, pp. 965–974 (electronic), 2005.

[3] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. An asymptotically tight bound on the number of connected components of realizable sign conditions. To appear in *Combinatorica*.

[4] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth, *Learnability and the Vapnik-Chervonenkis dimension*, J. Assoc. Comput. Mach. **36** (1989), 929–965. MR MR1072253 (91f:68178)

[5] R. M. Dudley, *Uniform central limit theorems*, Cambridge Studies in Advanced Mathematics, vol. 63, Cambridge University Press, 1999. MR MR1720712 (2000k:60040)

[6] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, Vol. 29, No. 2, pp. 295–327, 2001.

[7] Pascal Massart. *Concentration inequalities and model selection*, Vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[8] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, Vol. 152, No. 1, pp. 37–55, 2003.

[9] Vladimir N. Vapnik, *Statistical learning theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons Inc., New York, NY, 1998. MR MR1641250 (99h:62052)