Discrete Comput Geom (2010) 44: 589–598
DOI 10.1007/s00454-009-9236-5

VC Dimensions of Principal Component Analysis

Yohji Akama · Kei Irie · Akitoshi Kawamura · Yasutaka Uwano

Received: 20 April 2009 / Revised: 17 July 2009 / Accepted: 11 December 2009 /
Published online: 30 December 2009

Abstract Motivated by statistical learning theoretic treatment of principal component analysis, we are concerned with the set of points in $\mathbb{R}^d$ that are within a certain distance from a k-dimensional affine subspace. We prove that the VC dimension of the class of such sets is within a constant factor of $(k+1)(d-k+1)$, and then discuss the distribution of eigenvalues of a data covariance matrix by using our bounds of the VC dimensions and Vapnik's statistical learning theory. In the course of the upper bound proof, we provide a simple proof of Warren's bound of the number of sign sequences of real polynomials.

Keywords VC dimensions · Principal component analysis · Warren's bound

1 Introduction

Given a set of data $x_1, \ldots, x_n \in \mathbb{R}^d$, one may wish, for the sake of analysis, to transform them into a space of lower dimension, say $k < d$. A commonly used transform

Y. Akama (✉) · Y. Uwano
Mathematical Institute, Tohoku University, Aoba, Sendai 980-8578, Japan
e-mail: akama@math.tohoku.ac.jp

Y. Uwano
e-mail: sa8m07@math.tohoku.ac.jp

K. Irie
Department of Mathematics, Kyoto University, Kyoto 606, Japan
e-mail: iriek@math.kyoto-u.ac.jp

A. Kawamura
Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3G4, Canada
e-mail: kawamura@cs.toronto.edu

is *principal component analysis* (PCA, [4, p. 111]), which projects the data into the k-dimensional affine subspace $H \subseteq \mathbb{R}^d$ that minimizes the mean square distance from the data

$$\frac{1}{n}\sum_{i=1}^{n} \operatorname{dist}(x_i, H)^2, \tag{1}$$

where dist denotes the Euclidean distance. The minimum squared distance r_{emp} is the sum of the k largest eigenvalues of a symmetric matrix, called data "covariance matrix," made from the data.

If the data are drawn independently from some fixed probability distribution, then, according to Vapnik's statistical learning theory [12], as the number l of data goes to infinity, the affine subspace H that minimizes the mean (1) does not necessarily converge, but r_{emp} does converge in probability to $\inf_H \mathrm{E}_x[\operatorname{dist}(x, H)^2]$, where the expectation is with respect to the distribution of the data x. In fact, the convergence can be slow, if the set system induced by PCA has large *Vapnik–Chervonenkis dimension* (*VC dimension*). The upper bound of the VC dimension gives non-asymptotic, distribution-independent evaluations both for the convergence rates, and for the sample complexity in the style of computational learning theory [3].

Definition 1.1 For sets $X \subseteq \mathbb{R}^d$ and $Y \subseteq X$, we say that a set $B \subseteq \mathbb{R}^d$ *cuts Y out of X* if $Y = X \cap B$. A class $\mathcal{C}$ of subsets of $\mathbb{R}^d$ is said to *shatter* a set $X \subseteq \mathbb{R}^d$ if every $Y \subseteq X$ is cut out of X by some $B \in \mathcal{C}$. The *VC dimension* of $\mathcal{C}$, denoted by $\operatorname{VCdim}(\mathcal{C})$, is defined to be the maximum n (or ∞ if no such maximum exists) for which some subset of $\mathbb{R}^d$ of cardinality n is shattered by $\mathcal{C}$.

The set system induced by PCA in the sense of [12, Theorem 5.4] is $\mathcal{C}_k^d$ defined below. For example, $\mathcal{C}_0^d, \mathcal{C}_1^2, \mathcal{C}_1^3$, and $\mathcal{C}_2^3$ are the classes of balls, usual bands, cylinders, and slabs, respectively.

Definition 1.2 Let d and $k < d$ be nonnegative integers. For a k-dimensional affine subspace H of $\mathbb{R}^d$ and a nonnegative real number r, the set $\{\, x \in \mathbb{R}^d : \operatorname{dist}(x, H) \le r \,\}$ is called the *band* with *center H* and *radius r*. Define $\mathcal{C}_k^d$ to be the class of all such bands.

Since this set system is described uniformly by real polynomials, an upper bound of VC dimension is derivable by Milnor–Thom bound [10] for the number of connected components of algebraic varieties, or Warren's bound [13, Theorem 3] on the number of sign sequence of real polynomials. Unfortunately, a direct application of such theorems is not enough to yield our upper bound (2).

However, we will establish a sharp upper bound, which does not follow from a direct application of the Milnor–Thom bound or Warren's bound, and will additionally establish a lower bound from which $\operatorname{VCdim}(\mathcal{C}_k^d) = \Theta((d-k+1)(k+1))$ follows.

$$(d-k+1)(k+1) \le \operatorname{VCdim}\bigl(\mathcal{C}_k^d\bigr) \le (8.740\ldots)(d-k+1)(k+1). \tag{2}$$

The coefficient $8.740\ldots$ in the upper bound of $\operatorname{VCdim}(\mathcal{C}_k^d)$ comes from a refined version of Warren's bound which will be stated in Sect. 3 and proved more succinctly

with Sard's theorem rather than Warren's, where Sard's theorem is a basic and common tool for differential geometry. We will also try to make explicit methodological difference between our evaluation of the number of sign sequences and Basu–Pollack–Roy's homological approach [1, 2] to a more general problem.

For the better order $(k+1)(d-k+1)$ of the upper bound of $\mathrm{VCdim}(\mathcal{C}_k^d)$, we choose a more concise description for an affine subspace, based on a fact that any linear space is represented as the image of a linear mapping as well as the kernel of another.

The lower bound (2) of $\mathrm{VCdim}(\mathcal{C}_k^d)$ is derived in Sect. 2 from a recurrence formula for $\mathrm{VCdim}(\mathcal{C}_k^d)$, and is independently derived from a specific geometric configuration of $(k+1)(d-k+1)$ points in $\mathbb{R}^d$, shattered by $\mathcal{C}_k^d$.

Although these bounds are not tight because of an easy fact $(1+1)(2-1+1) = 4 < \mathrm{VCdim}(\mathcal{C}_1^2) = 5$ and so on, our asymptotically tight order $\mathrm{VCdim}(\mathcal{C}_k^d) = \Theta((k+1)(d-k+1))$ explains that approximating data by a k-dimensional affine subspace is equally hard as approximating data by an *orthogonal complement* of a k-dimensional affine subspace. This suggests a strategy of structural risk minimization [12] for PCA.

Another implication of our evaluation of $\mathrm{VCdim}(\mathcal{C}_k^d)$ is related to a study of Johnstone [6] on the distribution of the *largest* eigenvalue of the data "covariance matrix" where the data are drawn from the multi-dimensional standard normal distribution. In the final section, by using (2) with Vapnik's statistical learning theory [12], we will discuss the tail probabilities of the distribution of any k eigenvalues of the data "covariance matrix" for the same situation, and then we will mention possible future work.

2 The Lower Bounds

A simple observation comes in handy for the later use: if $\mathcal{C}_k^d$ shatters a finite set $X \subseteq \mathbb{R}^d$, then it does so *with margins*; more precisely, there is a margin $\delta > 0$ such that each $Y \subseteq X$ is cut out of X by some element of $\mathcal{C}_k^d$ whose boundary is at least δ apart from every point of X.

Lemma 2.1 *Let $X \subseteq \mathbb{R}^d$ be a set of affinely independent $d+1$ points. Then there is $r_0 > 0$ such that for all $r > r_0$, the class of closed balls with radius r shatters X with margins.*

Proof For any $Y \subseteq X$, it is easy to see that there is an open half-space H that cuts Y out of X. Let l be a line orthogonal to the boundary of H. For each $r > 1$, let B_r be the closed ball of radius $r - 1/r$ whose center lies on l at a distance r from H. Since $H = \bigcup_{r>1} B_r$, we have $Y \subseteq B_{r_Y}$ for some $r_Y > 0$. For each $r > r_Y$, the interior B_r° of B_r contains B_{r_Y}, and hence B_r cuts Y from X with margins. Thus, $r_0 = \max_{Y \subseteq X} r_Y$ satisfies the condition. □

Theorem 2.2 *For any nonnegative integers d and $k < d$, we have*

$$\mathrm{VCdim}\big(\mathcal{C}_{k+1}^{d+1}\big) \geq \mathrm{VCdim}\big(\mathcal{C}_k^d\big) + d - k + 1.$$

Proof Let $X \subseteq \mathbb{R}^d$ be any set of size $\mathrm{VCdim}(\mathcal{C}_k^d)$ that is shattered by $\mathcal{C}_k^d$. In the next paragraph, we will construct a set $Z \subseteq \mathbb{R}^d$ of $d-k+1$ points with the following property: for each $Y \subseteq X$ and $W \subseteq Z$, there are bands $A, B \in \mathcal{C}_k^d$ that are parallel, have equal radii, and cut out Y and W from X and Z, respectively. Once this is done, we choose a number $\lambda > 0$ and show that $\mathcal{C}_{k+1}^{d+1}$ shatters $(X \times \{0\}) \cup (Z \times \{\lambda\})$, thus proving the first bound.

Since X is shattered by $\mathcal{C}_k^d$, each $Y \subseteq X$ is cut out by some $A_Y \in \mathcal{C}_k^d$. Let N_Y be the $(d-k)$-dimensional linear subspace of $\mathbb{R}^d$ orthogonal to the center of A_Y. Let $Z \subseteq \mathbb{R}^d$ be a set of $d-k+1$ points that are mapped to distinct affinely independent points by each orthogonal projection $\pi_Y : \mathbb{R}^d \to N_Y$. By Lemma 2.1, each image $\pi_Y(Z)$ is shattered with a margin by closed balls of radius r, whenever r is greater than some r_0. Since there are only finitely many $Y \subseteq X$, this r_0 can be taken uniformly. By scaling Z down if necessary, we may assume that r_0 is less than the radius of A_Y for any $Y \subseteq X$. Then for each $W \subseteq Z$, the image $\pi_Y(W)$ is cut out of $\pi_Y(Z)$ by some $(d-k)$-dimensional closed ball $E_{Y,W} \subseteq N_Y$ with the same radius as A_Y. We have thus found bands $A = A_Y$ and $B = \pi_Y^{-1}(E_{Y,W})$ as desired above.

Now let $\lambda > 0$ be a large number to be determined later. To show that $(X \times \{0\}) \cup (Z \times \{\lambda\})$ is shattered by $\mathcal{C}_{k+1}^{d+1}$, let $(Y \times \{0\}) \cup (W \times \{\lambda\})$ be arbitrary subset of it. Take $A, B \in \mathcal{C}_k^d$ as above that correspond to these Y, W. Let H_A and H_B be their centers. Let $C \in \mathcal{C}_{k+1}^{d+1}$ be the band whose center passes through $H_A \times \{0\}$ and $H_B \times \{\lambda\}$ and which has the same radius as A and B. As λ increases, the sections $C_0 = \{x \in \mathbb{R}^d : (x, 0) \in C\}$ and $C_\lambda = \{z \in \mathbb{R}^d : (z, \lambda) \in C\}$ approach A and B, respectively. Since A and B cut out Y and W with margins, so do C_0 and C_λ, whence C cuts out $(Y \times \{0\}) \cup (W \times \{\lambda\})$, for large enough λ. Since there are only finitely many Y and W, such a λ can be taken uniformly. □

In [5], it is stated that the VC dimension of $\mathcal{C}_0^d$ is less than or equal to $d+1$. By Lemma 2.1, it is exactly $d+1$. Thus we have

Corollary 2.3 *For any nonnegative integers d and $k < d$,*

$$\mathrm{VCdim}\big(\mathcal{C}_k^d\big) \geq (k+1)(d-k+1).$$

A concrete set of size $(k+1)(d-k+1)$ that is shattered by $\mathcal{C}_k^d$ is given below:

Theorem 2.4 *For any nonnegative integers d and $k < d$, the class $\mathcal{C}_k^d$ shatters the set $R \times E \subseteq \mathbb{R}^d$ of $(d-k+1)(k+1)$ points, for the vertex set $R \subset \mathbb{R}^{d-k}$ of some sufficiently small $(d-k)$-dimensional regular simplex, and for some set $E \subset \mathbb{R}^k$ consisting of k pairwise orthogonal unit vectors in $\mathbb{R}^k$ and the origin.*

3 A Bound on the Number of Sign Sequences

It is a standard way to employ the Milnor–Thom bound [10], Warren's bound [13, Theorem 3] on the number of sign sequences for algebraic varieties, and so on, to evaluate from above the VC dimension of a set system such that each set is

uniformly described by real polynomials. The following theorem is an improvement of Theorem 2 of Warren [13], in a sense that the coefficient is slightly better than his. By this theorem, we prove our upper bound (2) in the next section.

Define $\operatorname{sgn} x$ to be $+$, $-$ or 0 when $x \in \mathbb{R}$ is positive, negative or zero, respectively.

Theorem 3.1 *Let m, d be positive integers, and $f_1, \dots, f_s, g_1, \dots, g_t$ be real polynomials in m variables, each of degree $d \geq 1$. Suppose that equation $g_1 = \cdots = g_t = 0$ define an L-dimensional smooth submanifold V of $\mathbb{R}^m$. If $s > m$, then the number of elements of $\{+,-\}^s$ that arise as $(\operatorname{sgn} f_1(x), \dots, \operatorname{sgn} f_s(x))$ for some $x \in V$ does not exceed*

$$d(2d-1)^{m-1} \sum_{k=0}^{L} 2^k \binom{s}{k}. \tag{3}$$

We remark that this bound (3) is smaller than $(4d)^m (\mathrm{e}s/L)^L$, because $d(2d-1)^{m-1} \leq (2d)^m$ and $\sum_{k=0}^{L} 2^k \binom{s}{k} \leq 2^m \sum_{k=0}^{L} \binom{s}{k} \leq 2^m (\mathrm{e}s/L)^L$. The last inequality is due to, for example, Blumer et al. [3, Proposition A2.1].

Warren [13, Theorem 3] has proved Theorem 3.1 with this looser bound for the special case where $t = 0$ (and thus $V = \mathbb{R}^m$ and $L = m$). His proof was based on the observation that the number in question is bounded by the number of connected components of $\mathbb{R}^m \setminus \bigcup_{i=1}^{s} \{x \in \mathbb{R}^m : f_i(x) = 0\}$. In general, $\{x \in \mathbb{R}^m : f_i(x) = 0\}$ are not smooth manifolds, and they can intersect badly, we need technically intricate arguments to estimate the number of these components. Here we present a simpler proof by directly estimating the number of sign sequences.

Proof of Theorem 3.1 Let $\Sigma \subseteq \{+,-\}^s$ denote the set of such sequences. For each $I \subseteq \{1, \dots, s\}$, let V_I be the set of $x \in V$ for which $f_i(x) = 0$ for all $i \in I$.

We first claim that for almost every $a = (a_1, \dots, a_s) \in \mathbb{R}^s$, replacing each f_i by $f_i - a_i$ makes V_I empty for all sets $I = \{i_1, \dots, i_{L+1}\}$ of size $L+1$. To see why, note that for V_I to be nonempty, the vector a must belong to the inverse image $\pi_I^{-1}[N_I]$ of $N_I = \{(f_{i_1}(x), \dots, f_{i_{L+1}}(x)) : x \in V\}$ under the canonical projection $\pi_I : \mathbb{R}^s \to \mathbb{R}^{L+1}$. By Sard's Theorem [11], N_I is a null set, and thus so is the union of $\pi_I^{-1}[N_I]$ over all I. Hence the claim follows. Now, if each a_i is sufficiently close to 0, replacing f_i by $f_i - a_i$ removes no element from Σ. We may therefore assume that V_I is empty for all I of size greater than L.

Let Γ be the set of all triples (I, S, W) consisting of a set $I \subseteq \{1, \dots, s\}$ of size L or smaller, a mapping S from I to $\{+,-\}$ and a connected component W of V_I. V_I has at most $d(2d-1)^{m-1}$ components by the Milnor–Thom bound [10], because V_I is the common real zeros of m-variate real polynomials of degree at most d. So the size of Γ is bounded by (3). It therefore suffices to give an injection φ from Σ into Γ.

For each $\sigma = (\sigma_1, \dots, \sigma_s) \in \Sigma$, let $I \subseteq \{1, \dots, s\}$ be a maximal set (with respect to set inclusion) for which there is $x \in V_I$ that satisfies $\operatorname{sgn} f_i(x) = \sigma_i$ for all $i \notin I$. Choose such an x and call it x_σ. As assumed above, I is of size at most L. This justifies defining $\varphi(\sigma) = (I, S, W)$, where $S(i) = \sigma_i$ for each $i \in I$ and $W \subseteq V_I$ is the component containing x_σ.

To see that φ is injective, suppose $\varphi(\sigma) = \varphi(\sigma') = (I, S, W)$. Since W is connected, there is a continuous path $c\colon [0,1] \to W$ from $x_\sigma = c(0)$ to $x_{\sigma'} = c(1)$. Let T be the set of $t \in [0,1]$ such that $\operatorname{sgn} f_i(c(t)) \neq \sigma_i$ for some $i \notin I$. Since T is closed, it has a minimum element t_0 unless it is empty. But then $f_i(c(t_0))$ would be 0 for one or more $i \notin I$ and have sign σ_i for all other $i \notin I$, contradicting the maximality of I. Hence, T must be empty, and thus $\sigma_i = \sigma'_i$ for all $i \notin I$. Since $\sigma_i = S(i) = \sigma'_i$ for $i \in I$ as well, $\sigma = \sigma'$. □

To end this section, we put some remark. In [1], Basu–Roy–Pollack evaluate from above the total number of the ith Betti numbers of the realizations of *all* realizable sign conditions of $f_1, \ldots, f_s$ over the set $\{x \in \mathbb{R}^m : g_1(x) = \cdots = g_t(x) = 0\}$ by the Oleinik–Petrovski–Milnor–Thom bound and an elaborated inductive argument with the Mayer–Vietoris long exact sequence (for a more homological treatment, see another paper [2] of theirs.) Since a realizable sign sequence contributes to the 0th Betti numbers of the realizations, their evaluation implies Theorem 3.1, however, with a larger base, which results in a looser upper bound of the VC dimensions we are concerned with. But by counting directly the number of sign sequences with the help of Sard's Theorem, we derive a slightly refined version of Warren's theorem with a simple proof which is more elementary than Warren's.

4 The Upper Bounds

In [8, Proposition 10.3.2], it is stated that the class $\mathcal{Q}_d$ of sets $\{x \in \mathbb{R}^d : p(x) \geq 0\}$ such that p is any real polynomial of total degree at most 2, has the VC dimension less than or equal to $(d+1)(d+2)/2$. Because $\mathcal{C}_k^d \subseteq \mathcal{Q}_d$, we have a trivial upper bound $(d+1)(d+2)/2$ of $\mathrm{VCdim}(\mathcal{C}_k^d)$. The following upper bound is improvement when k is close to either 0 or d.

Theorem 4.1 $\mathrm{VCdim}(\mathcal{C}_k^d) \leq (8.740\ldots)(k+1)(d-k+1)$.

This evaluation is used in the next section. First, we prove technical lemmas. For terminology about manifolds, see [11].

Lemma 4.2 *Let $a \leq b$ be positive integers. Let V be the set of $b \times a$ real matrices F such that*

(1) *The column vectors of F are orthonormal; and*
(2) *The (i,j)-component F_{ij} of F is 0 for any $1 \leq i < j \leq a$.*

Then V is a nonempty $a(b-a)$-dimensional smooth submanifold of $\mathbb{R}^{b \times a}$.

Proof The set V is nonempty, as it contains the $b \times a$ matrix $(\delta_{ij})_{1 \leq i \leq b, 1 \leq j \leq a}$, where δ_{ij} is Kronecker's delta. Let M be the set of all $b \times a$ real matrices F satisfying (2). Clearly, M can be identified with $\mathbb{R}^{ab - a(a-1)/2}$. Define $\varphi\colon M \to \mathbb{R}^{a(a+1)/2}$ by $\varphi = (\varphi_\nu)_\nu$, where ν ranges over all pairs (l, m) such that $1 \leq l \leq m \leq a$, and

$$\varphi_\nu(F) = \sum_u F_{ul} F_{um} - \delta_{lm}.$$

Then $V = \varphi^{-1}(\{0\})$. So the proof will be complete by implicit function theorem, if we show that the Jacobian matrix J_φ at F is full-rank, i.e., $\operatorname{rank} J_\varphi = a(a+1)/2$. Here the Jacobian matrix is

$$(J_\varphi)_{\nu\mu} = \frac{\partial \varphi_\nu}{\partial F_\mu} = \frac{\partial}{\partial F_{uv}}\left(\sum_w F_{wl}F_{wm} - \delta_{lm}\right) = \delta_{vl}F_{um} + \delta_{vm}F_{ul},$$

where μ is any pair (u,v) such that $1 \le v \le a, v \le u \le b$, and F_μ is F_{uv}. Let $F \in \varphi^{-1}(\{0\})$. For each $\lambda = (i,j)$ with $1 \le i \le j \le a$, define $p^{(\lambda)} \in \mathbb{R}^{ab-a(a-1)/2}$ by $(p^{(\lambda)})_\mu = \delta_{vi}F_{uj}$. Then $\sum_\mu (J_\varphi)_{\nu\mu}(p^{(\lambda)})_\mu = \delta_{il}\sum_u F_{um}F_{uj} + \delta_{im}\sum_u F_{ul}F_{uj}$, where u ranges over $1 \le u \le b$. It is $\delta_{il}\delta_{mj} + \delta_{im}\delta_{lj}$ because $F \in \varphi^{-1}(\{0\})$. Thus $\sum_\mu (J_\varphi)_{\nu\mu}(p^{(\lambda)})_\mu$ is $1 + \delta_{ij} > 0$ for $\lambda = \nu$, and 0 otherwise. Hence, $\operatorname{rank} J_\varphi = a(a+1)/2$. □

Let $(\mathcal{C}_k^d)^\flat$ be the set of elements of $\mathcal{C}_k^d$ whose center 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 some set of points is shattered by $\mathcal{C}_k^d$ then it is shattered with margins by $\mathcal{C}_k^d$, which implies that it is shattered by $(\mathcal{C}_k^d)^\flat$ by appropriate perturbation.

Lemma 4.3 *Let $L = (k+1)(d-k)+1$. Then, there exist a positive integer $m \le 2L$, an L-dimensional smooth submanifold V in $\mathbb{R}^m$ defined by $m-L$ quadratic equations in m variables, and $\Phi : V \to \mathcal{C}_k^d$ 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 polynomial in m variables f_p such that for all $x \in V$, $f_p(x) > 0$ if p belongs to the interior of $\Phi(x)$, while $f_p(x) < 0$ if $p \notin \Phi(x)$.*

Proof First, we consider the case $k \ge d/2$. Let $m = (d-k)(d+1)+1$. Then it is indeed $m \le 2L$. 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

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

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

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

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 band with center $\{z \in \mathbb{R}^d : (F^\top)z = b\}$ and radius $|r|$. Here $\top$ is the transpose of a matrix. Then Φ satisfies (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

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

This satisfies (b) because $\|(F^\top)p - b\|^2$ is equal to the square of the distance from p to the center of $\Phi(F,b,r)$.

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

$$\sum_u t_u E_{uj} = 0 \quad (1 \le j \le k) \tag{4}$$

and k^2 equations

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

We can prove that the system defines an L-dimensional smooth submanifold V of $\mathbb{R}^m$, as we proved the previous lemma. 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 band with center $\{Ex + t : x \in \mathbb{R}^k\}$ and radius $|r|$. Then Φ satisfies (a), since $\Phi(V)$ contains $(\mathcal{C}_k^d)^\flat$. Moreover, for $p \in \mathbb{R}^d$, define f_p by

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

Then f_p is clearly quadratic. By (4), 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 center of $\Phi(E, t, r)$. Thus we have (b). □

Now we will complete the proof of the upper bound.

Proof of Theorem 4.1 Take m, L, V, Φ as in the previous lemma. Take quadratic polynomials $g_1, \ldots, g_{m-L}$ in m variables so that equations $g_1 = \cdots = g_{m-L} = 0$ define V. Let $\{p_1, \ldots, p_s\} \subseteq \mathbb{R}^d$ be a set shattered by $\mathcal{C}_k^d$. By (a) of the previous lemma, it is shattered by $\Phi(V)$ with margins. Suppose $s \le m$. Then because the previous lemma implies $m \le 2L$, we have $s \le m \le 2L < (8.740\ldots)L$ as desired. If $s > m$, then $f_{p_1}, \ldots, f_{p_s}, g_1, \ldots, g_{m-L}$ satisfy the assumption of Theorem 3.1 (here we put $m - L$ for t), and all 2^s elements of $\{+, -\}^s$ must appear as a sign pattern of $f_{p_1}, \ldots, f_{p_s}$ over V. As V is an L-dimensional submanifold of $\mathbb{R}^m$, we have $L \le m$ from which $L \le s$ follows by the premise $s \le m$. Hence, by Theorem 3.1, $2^s \le 2 \cdot 3^{m-1} \sum_{j=0}^L 2^j \binom{s}{j} \le 2 \cdot 3^{2L-1} \cdot 2^L \sum_{j=0}^L \binom{s}{j} \le 18^L (\frac{\mathrm{e}s}{L})^L$, where the last inequality is again by Blumer et al. [3, Proposition A2.1]. This gives $2^{s/L} \le 18\mathrm{e}(s/L)$, or $s/L \le 8.740\ldots$, as desired. □

5 Discussion and Future Work—Distributions of Eigenvalues of Data Covariance Matrix

Let $x_1, \ldots, x_n$ be any i.i.d. sample from the d-dimensional standard normal distribution $N_d(0, I_d)$, where the mean is the column zero vector, the covariance matrix is the identity matrix I_d of size d, and each x_i is a d-dimensional column vector. John-

stone [6] proved that if the largest eigenvalue of a $d \times d$ real matrix $(\sum_{i=1}^n x_i x_i^\top)$ is appropriately centered and scaled, then the distribution approaches to the Tracy–Widom law of order 1, as n/d tends to a 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 I_d almost surely.

Below, 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 our upper bound (Theorem 4.1) of the VC dimensions and a theorem [12, (5.27)] of Vapnik's statistical learning theory. But for the right tail probability, we represent a subspace with the kernel of a linear mapping and then employ a concentration inequality [7, (4.17)] for the chi-square distributions.

Recall that the chi-square distribution of degree k of freedom has the pth noncentral moment $m(k, p) = k(k+2)(k+4)\cdots(k+2p-2)$.

Theorem 5.1 *Let $x_1, \ldots, x_n$ be an i.i.d. sample from the d-dimensional standard normal distribution $N_d(0, I_d)$, $\lambda_1, \ldots, \lambda_k$ $(k \leq d)$ be any eigenvalues of the data covariance $d \times d$ matrix $(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)$, $p > 2$, and $\varepsilon, \delta > 0$. Then, the left tail probability of $\sum_{i=1}^k \lambda_i$ satisfies the following*:

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

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

$$\begin{aligned} &P\left(\sum_{i=1}^k \lambda_i - k \geq \varepsilon\left(\frac{m(d-k,p)}{2}\left(\frac{p-1}{p-2}\right)^{p-1}\right)^{\frac{1}{p}} + \delta\right) \\ &\quad \leq 4\exp\left\{\left(\frac{G_{\mathcal{C}_k^d}(2n)}{n} - \frac{\varepsilon^2}{4}\right)n\right\} + \exp\left(-\frac{1}{2}nd\left(\sqrt{1+\frac{\delta}{d}} - 1\right)^2\right). \end{aligned}$$

Here $G_{\mathcal{C}_{d-k}^d}$ and $G_{\mathcal{C}_k^d}$ are the so-called growth functions of $\mathcal{C}_{d-k}^d$ and $\mathcal{C}_k^d$, respectively. In particular, if $n > v/2$ with v being $(8.740\ldots)(k+1)(d-k+1)$, then the inequalities can be made concrete by replacing the two growth functions $G_{\mathcal{C}_{d-k}^d}(2n)$ and $G_{\mathcal{C}_k^d}(2n)$ in the inequalities with $v(\log\frac{2n}{v} + 1)$.

We put some remark on related work and future work. We are suggested by an anonymous referee to refer papers of the local theory of Banach spaces on concentration of measure that is directly relevant (e.g., [9]), and to add a comment on Talagrand's work on concentration of measure. Talagrand's inequalities for concentration of measure are recently employed in [7, Chap. 8] for statistical learning problems with the class of loss function being bounded, as follows:

(1) Bousquet's version of Talagrand's concentration inequality for empirical processes 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 [12], 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 for similar concentration inequalities which improve: (1) previous Theorem for PCA and (2) model selection (i.e., selecting k) for PCA. Anyway, it may be one of interesting directions for future work.

Acknowledgements The authors are grateful to anonymous referees, Saugata Basu, Akito Sakurai, Jiří Matoušek, Yuya Matsumoto, and Yoshio Okamoto for helpful comments.

References

1. Basu, S., Pollack, R., Roy, M.-F.: On the Betti numbers of sign conditions. Proc. Am. Math. Soc. **133**(4), 965–974 (2005) (electronic)
2. Basu, S., Pollack, R., Roy, M.-F.: An asymptotically tight bound on the number of connected components of realizable sign conditions. Combinatorica (to appear)
3. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the Vapnik–Chervonenkis dimension. J. Assoc. Comput. Mach. **36**, 929–965 (1989)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, New York (2001)
5. Dudley, R.M.: Balls in $\mathbf{R}^k$ do not cut all subsets of $k+2$ points. Adv. Math. **31**(3), 306–308 (1979)
6. Johnstone, I.M.: On the distribution of the largest eigenvalue in principal components analysis. Ann. Stat. **29**(2), 295–327 (2001)
7. Massart, P.: Concentration Inequalities and Model Selection. Lecture Notes in Mathematics, vol. 1896. Springer, Berlin (2007). Lectures from the 33rd Summer School on Probability Theory Held in Saint-Flour, 6–23 July 2003, With a Foreword by Jean Picard
8. Matoušek, J.: Lectures on Discrete Geometry. Graduate Texts in Mathematics, vol. 212. Springer, New York (2002)
9. Mendelson, S., Vershynin, R.: Entropy and the combinatorial dimension. Invent. Math. **152**(1), 37–55 (2003)
10. Milnor, J.W.: On the Betti numbers of real varieties. Proc. Am. Math. Soc. **15**, 275–280 (1964)
11. Milnor, J.W.: Topology from the Differentiable Viewpoint. Princeton Landmarks in Mathematics. Princeton University Press, Princeton (1997)
12. Vapnik, V.N.: Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, New York (1998)
13. Warren, H.E.: Lower bounds for approximation by nonlinear manifolds. Trans. Am. Math. Soc. **133**, 167–178 (1968)