

The High Dimension, Low Sample Size Geometric Representation Holds Under Mild Conditions

BY JEONGYOUN AHN

Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.

jyahn@stat.uga.edu

J. S. MARRON

Department of Statistics and Operations Research, University of North Carolina,

Chapel Hill, North Carolina 27599, U.S.A.

marron@email.unc.edu

KEITH M. MULLER

Department of Epidemiology and Health Policy Research, University of Florida,

Gainesville, Florida 32610, U.S.A.

Keith.Muller@biostat.ufl.edu

AND YUEH-YUN CHI

Department of Biostatistics, University of Washington, Seattle, Washington 98195,

U.S.A.

yychi@u.washington.edu

SUMMARY

High dimension, low small sample size datasets have different geometrical properties from those of traditional low dimensional data. In their asymptotic study regarding increasing dimensionality with a fixed sample size, Hall et al. (2005) showed that each data vector is approximately located on the vertices of a regular simplex in a high-dimensional

space. A perhaps unappealing aspect of their result is the underlying assumption which basically requires the variables, viewed as a time series, to be almost independent. We establish an equivalent geometric representation under much milder conditions using asymptotic properties of sample covariance matrices. We discuss implications of the results, such as the use of principal component analysis in a high-dimensional space, extension to the case of non-independent samples and also the binary classification problem.

Some key words: High dimension, low sample size; Large p small n ; Linear discrimination; Sample covariance matrix.

1. INTRODUCTION

Datasets with more variables than observations are now important in many fields, such as genetics (Golub et al., 1999; Furey et al., 2000), medical imaging and text recognition. Such data have surprising and often counter-intuitive geometrical structures (Hall et al., 2005; Donoho & Tanner, 2005). For example, zero-mean Gaussian random samples in a very high-dimensional space are hardly ever located near the population mean. Furthermore, they also tend to be further away from the mean as the dimension increases, which appears paradoxical since the standard univariate Gaussian density is largest at zero.

Related asymptotic studies assume that the dimension d increases, whereas the sample size n can be fixed or increases along with d . In this paper, we let d go to infinity with a fixed n , which we call the d -asymptotics. Hall et al. (2005) took a d -asymptotic approach and showed that, as d increases, under some regularity conditions, the geometrical structure of data becomes deterministic: pairwise distances between data vectors are approximately constant, so that each vector is located on the vertices of a regular n -simplex in \mathbb{R}^d . This

geometric representation essentially says that the randomness of data of this type only lies in random rotations of this simplex. Hall et al. (2005) also applied this result to the binary classification problem and discussed the asymptotic behaviour of some popular discrimination methods such as support vector machines (Cristianini & Shawe-Taylor, 2000) and distance weighted discrimination; see Benito et al. (2004) and a University of North Carolina technical report by J. S. Marron, M. Todd and J. Ahn.

The results in Hall et al. (2005) require variables to be ‘nearly independent’ in the sense that, when they are viewed as a time series, the variables must satisfy a ρ -mixing condition. This assumption has some obvious shortcomings. First, it is somewhat too strict because it is common to have severe collinearity among variables. Secondly, the condition depends on the order of the data entries, which can be arbitrary in many applications.

In this paper, we study d -asymptotic properties of sample covariance matrices to show the geometric representation in Hall et al. (2005) in more general settings. The new condition is imposed on the population eigenvalues and controls the departure from sphericity by restricting the relative sizes of dominating eigenvalues.

Some recent high dimension, low sample size asymptotic studies on sample covariance matrices, including Baik et al. (2005), Baik & Silverstein (2006) and a University of California, Davis technical report by D. Paul, found that, when the ratio d/n goes to a constant, the sample eigenvalues behave as if the underlying covariance matrix were the identity matrix. This holds if the underlying covariance matrix is not far from the identity. Section 2 shows that this so-called ‘phase transition’ phenomenon also occurs in the d -asymptotic case. The phenomenon makes principal component analysis with high dimension, low sample size data often unreliable; see a Stanford University technical report by I. M. Johnstone

and A. Y. Lu. In §4, an extremely non-spherical population model is presented for which the phenomenon no longer takes place. We show that the first principal component converges to the first true eigenvector as d increases, but its variance, i.e., the first sample eigenvalue, does not converge to the population counterpart.

2. PROPERTIES OF THE SAMPLE COVARIANCE MATRIX

In this section we examine the asymptotic properties of sample covariance matrices when the dimension d tends to infinity while the sample size n is fixed.

Suppose we have a $d \times n$ data matrix $X = [x_1, \dots, x_n]$ with $d > n$, where $x_j = (x_{1j}, \dots, x_{dj})^T, j = 1, \dots, n$, are independent and identically distributed from a d -dimensional multivariate distribution with mean zero and nonnegative definite covariance matrix Σ . The eigenvalue decomposition of Σ is $\Sigma = V\Lambda V^T$, where Λ is a diagonal matrix of eigenvalues $\lambda_1 \geq \dots \geq \lambda_d > 0$ and V is the matrix of corresponding eigenvectors. A ‘factor matrix’, which is essentially the square root of Σ , is defined as $F \equiv V\Lambda^{1/2}$, so that $\Sigma = FF^T$. We can write $X = FZ$, where $Z = \Lambda^{-\frac{1}{2}}V^T X$ is a $d \times n$ random data matrix from a distribution with the identity covariance matrix. Note that, if X is Gaussian, the elements of Z are independent standard univariate normal variables.

The sample covariance matrix $S = n^{-1}XX^T$ is decomposed as $S = n^{-1}FZZ^TF^T$. Note that we do not subtract the sample mean vector in order to avoid unnecessary complexity and also because the population mean is a zero vector. A ‘dual’ approach switches the roles of columns and rows of a data matrix, by replacing X by X^T . The $n \times n$ dual sample covariance matrix is defined as $S_D = n^{-1}X^T X$. Note that S_D has the same eigenvalues as

S . If we write X as FZ and use the fact that $V^T V$ is the identity,

$$nS_D = (Z^T F^T)(FZ) = Z^T \Lambda Z = \sum_{i=1}^d \lambda_i W_i. \quad (1)$$

Here the $n \times n$ matrix W_i is $Z_i^T Z_i$ where $Z_i, i = 1, \dots, d$, are the row vectors of Z . If X is Gaussian, each W_i follows independently the Wishart distribution $\mathcal{W}_n(1, I_n)$.

The following theorem states that, under some mild conditions on population eigenvalues, S_D approximately becomes a scaled identity matrix as d increases with a fixed n . Thus all the eigenvalues of S_D are approximately the same, and so are those of S . In a sense, extreme data of this type behave as if the underlying distribution was spherical. An analogous result when d/n approaches a constant in $(0, \infty)$ and the population covariance matrix is the identity exists in references such as Bai & Silverstein (1998).

The assumption for this theorem involves a well-known measure of sphericity,

$$\epsilon \equiv \frac{\text{tr}^2(\Sigma)}{d \text{tr}(\Sigma^2)} = \frac{\left(\sum_{i=1}^d \lambda_i\right)^2}{d \sum_{i=1}^d \lambda_i^2}. \quad (2)$$

The empirical version of (2), with Σ replaced by S , is a locally most powerful invariant test statistic of sphericity of multivariate Gaussian distributions (John, 1972). Note that ϵ is always between d^{-1} and 1. Perfect sphericity of the distribution occurs only when $\epsilon = 1$, and $\epsilon = d^{-1}$ corresponds to the ‘most singular’ case in which only a single eigenvalue is nonzero. Our key assumption concerns the singular end of the ϵ spectrum: we need ϵ to be not too close to d^{-1} for large d , in the sense that $\epsilon^{-1} = o(d)$. In other words, the underlying distribution needs to be not too close to the singular case.

THEOREM 1. *For a fixed n , consider a sequence of $d \times n$ random data matrices X_1, \dots, X_d, \dots from multivariate distributions with dimension d , with zero means and covariance matrices*

$\Sigma_1, \dots, \Sigma_d, \dots$. Assume that the fourth moments of each variable are uniformly bounded and also the representation in (1) holds for each X_i . Let $\lambda_{1,d} > \dots > \lambda_{d,d}$ be the eigenvalues of the covariance matrix Σ_d , and let $S_{D,d}$ be the corresponding dual sample covariance matrix. Suppose the eigenvalues of Σ_d are sufficiently diffused, in the sense that

$$(d\epsilon)^{-1} = \frac{\sum_{i=1}^d \lambda_{i,d}^2}{\left(\sum_{i=1}^d \lambda_{i,d}\right)^2} \rightarrow 0 \quad \text{as } d \rightarrow \infty. \quad (3)$$

Then the sample eigenvalues behave as if they are from an identity covariance matrix, in the sense that $c_d^{-1} S_{D,d} \rightarrow I_n$ as $d \rightarrow \infty$, where $c_d = n^{-1} \sum_{i=1}^d \lambda_{i,d}$.

Proof. By (1), any diagonal element of $nS_{D,d}$ can be expressed as $\sum_{i=1}^d \lambda_{i,d} z_i^2$, where the z_i 's are independent and identically distributed with zero mean and unit variance. Define the relative eigenvalues by $\tilde{\lambda}_{i,d} = \lambda_{i,d} / (\sum_{i=1}^d \lambda_{i,d})$. The condition in (3) is equivalent to $\sum_{i=1}^d \tilde{\lambda}_{i,d}^2 \rightarrow 0$ as $d \rightarrow \infty$. Then, by Chebyshev's inequality, for any $\tau > 0$ and the uniform bound M for the fourth moments condition, $\text{pr}(|\sum_{i=1}^d \tilde{\lambda}_{i,d} z_i^2 - 1| > \tau) \leq \tau^{-2} \text{var}(\sum_{i=1}^d \tilde{\lambda}_{i,d} z_i^2) \leq \tau^{-2} M \sum_{i=1}^d \tilde{\lambda}_{i,d}^2 \rightarrow 0$ as $d \rightarrow \infty$. Thus a diagonal element $\sum_{i=1}^d \tilde{\lambda}_{i,d} z_i^2$ converges to 1 almost surely. The off-diagonal elements of $nS_{D,d}$ can be expressed as $\sum_{i=1}^d \lambda_{i,d} z_i z_{i'}$, where z_i and $z_{i'}$ are independent. Then, similarly, $\text{pr}(|\sum_{i=1}^d \tilde{\lambda}_{i,d} z_i z_{i'}| > \tau) \leq \tau^{-2} \text{var}(\sum_{i=1}^d \tilde{\lambda}_{i,d} z_i z_{i'}) = \tau^{-2} \sum_{i=1}^d \tilde{\lambda}_{i,d}^2 \rightarrow 0$ as $d \rightarrow \infty$. Thus the off-diagonal elements converge to 0 almost surely. \square

The condition in (3) holds for quite general settings, including the following cases: (a) all the eigenvalues are the same; (b) the first k eigenvalues are moderately larger than the rest, an example being $\lambda_{1,d} = \dots = \lambda_{k,d} = c_1 d^\alpha$, $\lambda_{k+1,d} = \dots = \lambda_{d,d} = c_2$, where $k < d, \alpha < 1, c_1, c_2 > 0$; (c) the eigenvalues decrease in any polynomial order. Cases in which (3) fails to hold include the following: (d) the first k eigenvalues are much larger

than the rest, an example being $\lambda_{1,d} = \dots = \lambda_{k,d} = c_1 d^\alpha$, $\lambda_{k+1,d} = \dots = \lambda_{d,d} = c_2$, where $k < d, \alpha \geq 1, c_1, c_2 > 0$; (e) the eigenvalues decrease exponentially; (f) only the first k eigenvalues are nonzero. The case in (f) has a singular covariance structure, and (d) and (e) are examples which become nearly singular as the dimension tends to infinity.

3. GEOMETRIC REPRESENTATION OF HIGH DIMENSION, LOW SAMPLE SIZE DATA

In this section we establish the data geometric representation using Theorem 1 and show that our assumption (3) is more general than that of Hall et al. (2005).

Let $x_j = (x_{1j}, \dots, x_{dj})^T$, $j = 1, \dots, n$, be the j th column of the data matrix X , of which underlying distribution satisfies the conditions in Theorem 1. The squared distance between x_k and x_ℓ is

$$\|x_k - x_\ell\|^2 = \sum_{i=1}^d (x_{ik} - x_{i\ell})^2 = \sum_{i=1}^d x_{ik}^2 + \sum_{i=1}^d x_{i\ell}^2 - 2 \sum_{i=1}^d x_{ik} x_{i\ell}. \quad (4)$$

Note that the first two terms in (4) are the k th and ℓ th diagonal entries of nS_D respectively. Then, by the theorem, for a sufficiently large d , both terms become similar to $\sum_{i=1}^d \lambda_i$. Also since the third term is the (k, ℓ) th entry of nS_D , it diminishes as d grows. Thus, after scaling with $\sum_{i=1}^d \lambda_i$, $\|x_k - x_\ell\|^2$ becomes 2; that is, for sufficiently large d , (4) is approximately $2 \sum_{i=1}^d \lambda_i$, so that the pairwise distances between the n data vectors are approximately the same and the data form a regular n -simplex in \mathbb{R}^d .

Now we compare the ρ -mixing condition specified in Hall et al. (2005) with the condition in (3). The ρ -mixing condition for the geometric representation states that, for $i, j = 1, \dots, d$ with $|i - j| \geq r$, $\sup_{|i-j| \geq r} |E(x_{i,d} x_{j,d})| \leq \rho(r) \rightarrow 0$, as $r \rightarrow \infty$. Suppose this mixing condition is satisfied, and additionally assume that $\text{tr}(\Sigma_d) \asymp d$, which means that

$\text{tr}(\Sigma_d)/d$ is bounded away from both 0 and ∞ . This condition is equivalent to the variance condition in Hall et al. (2005). Then the left-hand side of (3) becomes

$$\frac{\sum_{i=1}^d \lambda_{i,d}^2}{\left(\sum_{i=1}^d \lambda_{i,d}\right)^2} = \frac{\sum_{i,j=1}^d E(x_{i,d}x_{j,d})^2}{\sum_{i,j=1}^d E(x_{i,d}^2)E(x_{j,d}^2)} = \frac{o(d^2)}{O(d^2)} \rightarrow 0, \quad \text{as } r \rightarrow \infty,$$

which implies that our condition in (3) is at least as mild as their mixing condition. Also, by a random permutation of the data entries, it is easy to construct an example that satisfies (3) but not the mixing condition. Hence our conditions are strictly milder than their original conditions.

4. AN EXTREMELY SPIKED POPULATION MODEL

4.1. *The problem setting*

For high-dimensional data, principal component analysis often fails to estimate the true eigen-directions and their variances. The condition in (3) characterizes an underlying structure of the data that leads to the failure of principal component analysis in a high dimension, low sample size limit. In this section we consider an extremely non-spherical distribution, in which the first eigenvalue dominates the others so strongly so that (3) does not hold, and furthermore there is a possibility for principal component analysis to estimate correctly important eigen-structures.

We consider Gaussian data from the population covariance matrix $\Sigma_d = \Lambda_d = \text{diag}(d^\alpha, 1, \dots, 1)$, $\alpha > 1$. This model can be considered as an extreme case of the spiked population model (Johnstone, 2001) and as a special version of case (d) in §2. Note that we already looked at the case where $0 < \alpha < 1$ in case (b) of that section. Note also that $\lambda_{1,d} = d^\alpha, \lambda_{2,d} = \dots = \lambda_{d,d} = 1$ are the eigenvalues of Λ_d , and the eigenvectors of Λ_d are the

d -dimensional unit vectors. In the following two subsections and in §5, all the quantities depend on d , but the subscript ‘ d ’ will be omitted for the sake of simpler notation.

4.2. The first sample eigenvalue

Let $\hat{\lambda}_1 > \dots > \hat{\lambda}_n$ be nonzero eigenvalues of the sample covariance matrix S , or of S_D . By (1), the dual sample covariance matrix S_D can be expressed as $n^{-1}(d^\alpha W_1 + \sum_{i=2}^d W_i)$, where W_i 's are independently $\mathcal{W}_n(1, I_n)$. Let $U = W_1$ and $V = \sum_{i=2}^d W_i$. Then $U \sim \mathcal{W}_n(1, I_n)$ and $V \sim \mathcal{W}_n(d-1, I_n)$, independently. Dividing S_D by d^α gives $d^{-\alpha} S_D = n^{-1}U + n^{-1}d^{-\alpha}V$. Note that, as d increases, V becomes close to $(d-1)I_n$, and thus the second term tends to the zero matrix. Also note that U can be expressed as the outer product of an n -dimensional random vector from $\mathcal{N}_n(0, I_n)$ with itself, so that its only nonzero eigenvalue is its inner product with itself, which is a χ_n^2 random variable. Thus $\hat{\lambda}_1$ is approximately distributed as $n^{-1}d^\alpha \chi_n^2$ when d is large. Note that, with a reasonably large n , $\hat{\lambda}_1$ can estimate $\lambda_1 = d^\alpha$ fairly well since $n^{-1}\chi_n^2$ will be near 1. The rest of the sample eigenvalues tend to 0 as d increases since U has rank one.

4.3. The first sample eigenvector

Consider the eigenvalue decomposition of $S = GLG^T$, where $G = \{\hat{g}_{ij} : i, j = 1, \dots, d\}$ is the matrix of corresponding eigenvectors, $\hat{v}_j = (\hat{g}_{1j}, \dots, \hat{g}_{dj})^T, j = 1, \dots, d$, are the eigenvectors, and $L = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_n, 0, \dots, 0)$. Now writing Λ_d as Λ to simplify the notation, define a standardized version of S as

$$\tilde{S} = \Lambda^{-\frac{1}{2}} S \Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}} G L G^T \Lambda^{-\frac{1}{2}}. \quad (5)$$

Since $S = n^{-1}FZZ^T F^T = n^{-1}\Lambda^{\frac{1}{2}}ZZ^T\Lambda^{\frac{1}{2}}$,

$$\tilde{S} = n^{-1}\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}ZZ^T\Lambda^{\frac{1}{2}}\Lambda^{-\frac{1}{2}} = n^{-1}ZZ^T \sim n^{-1}\mathcal{W}_d(n, I_d), \quad (6)$$

where Z is a $d \times n$ data matrix whose elements are independently $N(0, 1)$. By (5), the i th diagonal entry of \tilde{S} is $\tilde{s}_{ii} = \lambda_i^{-1}(\hat{g}_{i1}^2 \hat{\lambda}_1 + \cdots + \hat{g}_{in}^2 \hat{\lambda}_n)$. Substituting our underlying eigenvalues gives $\tilde{s}_{11} = d^{-\alpha}(\hat{g}_{11}^2 \hat{\lambda}_1 + \cdots + \hat{g}_{1n}^2 \hat{\lambda}_n)$, and $\tilde{s}_{jj} = (\hat{g}_{j1}^2 \hat{\lambda}_1 + \cdots + \hat{g}_{jn}^2 \hat{\lambda}_n)$, $2 \leq j \leq d$. From the results about sample eigenvalues in §4.2, for a large d , $\tilde{s}_{11} \simeq \hat{g}_{11}^2 Q$, and $\tilde{s}_{jj} \simeq \hat{g}_{j1}^2 d^\alpha Q$, $2 \leq j \leq d$, where $nQ \sim \chi_n^2$. However, each \tilde{s}_{jj} , $1 \leq j \leq d$, is independently distributed as χ_n^2/n by (6). Thus it can be seen that $\hat{g}_{11}^2 \rightarrow 1$ in probability, and also since $\hat{g}_{j1} = O_p(d^{-\alpha})$, $2 \leq j \leq d$, we have $\max_{2 \leq j \leq d} \hat{g}_{j1} = O_p(d^{1-\alpha})$ as $d \rightarrow \infty$; that is, the first sample eigenvector converges to the population counterpart as d increases.

5. DISCUSSION

5.1. Non-independent samples

Let X be the data matrix from the generalized multivariate normal distribution $\mathcal{N}_{n,d}(0, B, \Sigma)$, i.e., $\text{vec}(X) \sim \mathcal{N}_d(0, B \otimes \Sigma)$. We assume that $\Sigma_{d \times d}$ and $B_{n \times n}$ are positive definite matrices and, especially, that $\Sigma = \Sigma_d$ satisfies the conditions in Theorem 1. Let F_1 and F_2 be the factor matrices for B and Σ respectively, and suppose that $Z \sim \mathcal{N}_{n,d}(0, I_n, I_d)$. Then $ZF_1 \sim \mathcal{N}_{n,d}(0, B \otimes I_d)$ and $F_2 Z \sim \mathcal{N}_{n,d}(0, I_n \otimes \Sigma)$. Using similar algebra to §2, we can show that $nS_D = \sum_{i=1}^d \lambda_i W_i$, where λ_i 's are the eigenvalues of Σ and each W_i independently has the Wishart distribution $\mathcal{W}_n(1, B)$. Note that $E(W_i) = B$ and each W_i is the outer product with itself of a Gaussian random vector $y_i = (y_{1i}, \cdots, y_{ni})^T, i = 1, \cdots, d$, from $\mathcal{N}_n(0, B)$. Let $B = \{b_{k\ell}\}$. Then the k th diagonal element of nS_D can be expressed as $\sum_{i=1}^d \lambda_i y_{ki}^2$,

which approximately becomes $b_{kk} \sum_{i=1}^d \lambda_i$ when d is large. Also the (k, ℓ) th off-diagonal element is $\sum_{i=1}^d \lambda_i y_{ki} y_{\ell i}$, $k \neq \ell$, which is approximately $b_{k\ell} \sum_{i=1}^d \lambda_i$. Then, when d is large,

$$\|x_k - x_\ell\|^2 \simeq (b_{kk} + b_{\ell\ell} - 2b_{k\ell}) \sum_{i=1}^d \lambda_i,$$

which implies that the geometrical structure again becomes deterministic.

5.2. *Unstability of binary classification with high dimension, low sample size data*

The associated asymptotic properties of the binary classification problem, in particular the behaviour of some simple linear classifiers, have been studied by Hall et al. (2005), Bickel & Levina (2004), and J. Ahn's unpublished 2006 Ph.D. thesis from the University of North Carolina. Here our interest lies in how sensitive classifiers are to the sample size.

Suppose we have independent samples from two different classes and the underlying covariance structures for each class satisfy the conditions in Theorem 1. Let X be the $d \times m$ data matrix for class +1, and let Y be the $d \times n$ data matrix for class -1 and let N be $m+n$, the combined sample size. Also let Σ_X and Σ_Y be the underlying covariance matrices for each class. Denote the traces of Σ_X and Σ_Y , i.e., the sums of population eigenvalues, by σ^2 and τ^2 , respectively. Then, as d increases, the data vectors from each class approximately form two regular simplexes, with the number of vertices being m and n and the lengths of edges being $\sqrt{2}\sigma$ and $\sqrt{2}\tau$, respectively. Also, the entire dataset asymptotically forms a convex N -polyhedron with N vertices by the analogous result of Hall et al. (2005). Let μ^2 be the squared distance between the population means of the two classes. Then, since the squared distances from each data vector to its class mean are approximately σ^2 and τ^2 , respectively, the squared distance between x_i and y_j is approximately $\sigma^2 + \tau^2 + \mu^2$. This is

easier to understand when $m = 1, n = 2$ or $m = 2, n = 1$, in which case the N -polyhedron is an isosceles triangle and we use Pythagoras' theorem.

Denote the centroids of each simplex by C_X^m and C_Y^n . As shown in Hall et al. (2005), the ideal separating hyperplane for this case has the normal direction vector which connects the centroids C_X^m and C_Y^n . The 'basic' support vector machines, distance weighted discrimination and the nearest centroid rule approximately satisfy this. Note that, as shown in Hall et al. (2005), in the large- d limit, the expected squared distance between C_X^m and C_Y^n is $\mu^2 + \sigma^2/m + \tau^2/n$. If we consider one more data vector for class +1, by a similar calculation, the expected squared distance between the centroid C_X^{m+1} of the new $(m + 1)$ -simplex and C_Y^n is $\mu^2 + \sigma^2/(m + 1) + \tau^2/n$. Also the expected squared distance between C_X^m and C_X^{m+1} can be shown to be $\{m(m + 1)\}^{-1}\sigma^2$, by calculation of the distance between the two $(m + 1)$ -dimensional vectors, $\sigma^2(m^{-1}, \dots, m^{-1}, 0)^T$ and $\sigma^2((m + 1)^{-1}, \dots, (m + 1)^{-1})^T$. Then as d increases the angle between the hyperplanes based on m and n samples from each class and the one based on $(m + 1)$ and n samples becomes approximately

$$\theta = \cos^{-1} \left(\frac{\mu^2 + \sigma^2/(m + 1) + \tau^2/n}{\mu^2 + \sigma^2/m + \tau^2/n} \right)^{\frac{1}{2}}.$$

This implies that, when σ^2 , the variation of class +1, is large and m is small, the small change in the sample size can make a significant difference in the resulting classifiers.

The v -fold crossvalidation method, which is a very common way of tuning classifiers, uses $N - [N/v]$ samples to fit the classifier and evaluate it using the other $[N/v]$ samples. Let $k_1 = [m/v]$ and $k_2 = [n/v]$. By a similar calculation, in the large- d limit, the angle between the hyperplanes based on $m - k_1$ and $n - k_2$ samples and the hyperplane based on

the whole m and n samples is approximately

$$\theta = \cos^{-1} \left(\frac{\mu^2 + \sigma^2/m + \tau^2/n}{\mu^2 + \sigma^2/(m - k_1) + \tau^2/(n - k_2)} \right)^{\frac{1}{2}}.$$

This angle can also be large when the the sample sizes are small and the variances are large, and this arouses a serious concern about the stability of the crossvalidation method. More rigorous investigation is proposed for a future work.

ACKNOWLEDGEMENT

This research was partly supported by the U.S. National Institutes for Health and the U.S. National Science Foundation. The authors are thankful to Debashis Paul for helpful comments.

REFERENCES

- BAI, Z. D. & SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Prob.* **26**, 316–45.
- BAIK, J., BEN AROUS, G. & PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for non-null complex covariance matrices. *Ann. Prob.* **33**, 1643–97.
- BAIK, J. & SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Mult. Anal.* **97**, 1382–408.
- BENITO, M., PARKER, J., DU, Q., WU, J., XIANG, D., PEROU, C. M. & MARRON, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–44.

- BICKEL, P. & LEVINA, E. (2004). Some theory for Fisher’s linear discriminant function, “naive bayes”, and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.
- CRISTIANINI, N. & SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.
- DONOHO, D. L. & TANNER, J. (2005). Neighborliness of randomly-projected simplices in high dimensions. *Proc. Nat. Acad. Sci.* **102**, 9452–7.
- FUREY, T. S., CHRISTIANINI, N., DUFFY, N., BEDNARSKI, D. W., SCHUMMER, M. & HAUSSLER, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–14.
- GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. & LANDER, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–7.
- HALL, P., MARRON, J. S. & NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc. B* **67**, 427–44.
- JOHN, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika* **59**, 169–73.
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327.