Serdica J. Computing **6** (2012), 437–450

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

A BIMODALITY TEST IN HIGH DIMENSIONS

Dean Palejev

Abstract. We present a test for identifying clusters in high dimensional data based on the k-means algorithm when the null hypothesis is spherical normal. We show that projection techniques used for evaluating validity of clusters may be misleading for such data. In particular, we demonstrate that increasingly well-separated clusters are identified as the dimensionality increases, when no such clusters exist. Furthermore, in a case of true bimodality, increasing the dimensionality makes identifying the correct clusters more difficult. In addition to the original conservative test, we propose a practical test with the same asymptotic behavior that performs well for a moderate number of points and moderate dimensionality.

1. Introduction and notations. Hartigan [3] develops an asymptotic distribution for clustering criteria in the one-dimensional case. In this case, the apparent bimodality of the sample can be used to infer bimodality in the population. For two clusters, one can project the data into the line determined by the cluster means. Hartigan notes an observation by Day [1] that apparent

ACM Computing Classification System (1998): I.5.3.
Key words: Clustering, bimodality, multidimensional space, asymptotic test.

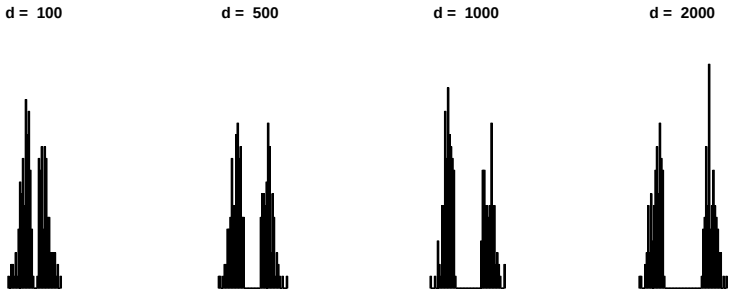

Fig. 1. Projections for two clusters, $n = 200$

bimodality on the projected data might not correspond to real bimodality on the population in a case of relatively few observations in many dimensions.

Starting with n observations in d dimensions from $N(0, I_d)$ where I_d is the $d \times d$ identity matrix, we can apply the k-means clustering method described by Hartigan [2]. As shown in Figure 1, the projection tests suggest the presence of more than one cluster. We also see that for a fixed n, when d increases, the separation becomes larger, which is misleading.

In many cases, we have multi-dimensional data for a relatively smaller number of observations, e.g., high-frequency stock price or return series, or microarray or high-throughput sequencing data. Therefore, we are interested in developing a test for determining whether the data came from a single mode. Our theoretical results are proven when n and $d \to \infty$. First, we find an asymptotic equation for the tail probability of the maximum of many independent χ^2 random variables. Then we develop a Slepian-type inequality for the probabilities of the maxima of independent and positively correlated χ^2 processes. Based on that inequality, we derive a conservative test for testing whether the data came from one multidimensional normal distribution vs. two normals. After that, we propose a practical test with the same asymptotic behavior. Although the theoretical results are asymptotic, the test can also be used for moderate values of n and d. We conclude that in the case of true bimodality, when sampling from a mixture of normals whose means are a fixed distance apart, increasing the dimensionality makes identifying the correct clusters more difficult. In particular, when the components are separated just enough for the distribution to be bimodal, the correct components are not identified when $d > n/10$.

We consider a set $X \equiv \{x_1, x_2, \ldots, x_n\}$ of n points from $N(0, I_d)$. The d-dimensional coordinates are denoted by superscripts, e.g. $x = (x^1, x^2, \ldots, x^d)$. Let S be a split of X into two clusters with means c_1 and c_2, consisting of n_1 and n_2 points respectively. The between cluster sum of squares $B(X, S)$ is given by

$$B(X,S) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \|c_1 - c_2\|^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \sum_{j=1}^{d} (c_1^j - c_2^j)^2$$

For $m = 1, 2$ and $j = 1, 2 \ldots d$, c_m^j has a $N(0, 1/n_m)$ distribution. Therefore $(1/n_1 + 1/n_2)^{-1/2}(c_1^j - c_2^j)$ is $N(0,1)$ and $(1/n_1 + 1/n_2)^{-1}(c_1^j - c_2^j)^2$ has a χ_1^2 distribution. The coordinates are independent, therefore $B(X, S)$ has a χ_d^2 distribution as a sum of d independent χ_1^2 random variables.

If $\bar{x}$ is a mean of the points in X, the total sum of squares $T(X)$ has a $\chi_{d(n-1)}^2$ distribution and is given by $T(X) = \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$.

For a fixed n, there are $N = 2^{n-1} - 1$ different splits of X into two non-empty sets. We denote these splits by $S_1, S_2, \ldots, S_N$. Let $B_i = B(X, S_i)$ for $i = 1, 2, \ldots, N$ and let $B_{\max}(X) = \max(B_1, B_2, \ldots, B_N)$ be the maximum between clusters sum of squares among all possible splits of X into two clusters. The squared distances $B(X, S)$ over different splits are not independent.

One could use a test with rejection region $B_{\max}(X) > A$ for testing bimodality, however finding $B_{\max}(X)$ is computationally infeasible even for moderate values of n. Let $B_k(X)$ be the optimal among the sum of squares obtained by the k-means method. The method does not guarantee finding the global maximum, thus $B_k(X) \leq B_{\max}(X)$.

We say that $x_n \sim y_n$ if $x_n/y_n \to 1$ when $n \to \infty$.

2. Tail probabilities for one and the maximum of many independent χ^2 random variables

Lemma 2.1. $\int_0^\infty \exp\left(-x - \frac{x^2}{2\beta}\right) dx \to 1$ *when* $\beta \to \infty$.

Proof. For $\beta > 0$, $\exp\left(-\frac{x^2}{2\beta}\right) \leq 1$. For a fixed x, when $\beta \to \infty$,

$\exp\left(-\frac{x^2}{2\beta}\right) \to 1$ therefore by dominated convergence

$$\int_0^\infty \exp\left(-x - \frac{x^2}{2\beta}\right)\, dx = \int_0^\infty \exp(-x)\exp\left(-\frac{x^2}{2\beta}\right)\, dx \to \int_0^\infty \exp(-x)\, dx = 1$$

□

Lemma 2.2. *If Y has χ^2_d distribution, $A > d$, $d \to \infty$, and $(A-d)^2/d \to \infty$, then*

$$P(Y > A) \sim \frac{A^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right) 2^{\frac{d}{2}} e^{\frac{A}{2}} \left(\frac{A}{2} - \frac{d}{2} + 1\right)}$$

P r o o f. If Y has χ^2_d distribution, then

$$P(Y > A) = \frac{1}{\Gamma\left(\frac{d}{2}\right) 2^{\frac{d}{2}}} \int_A^\infty y^{\frac{d}{2}-1} e^{-\frac{y}{2}}\, dy$$

Substituting $u = y - A$, after basic transformations we get

$$P(Y > A) = \frac{A^{\frac{d}{2}-1}}{\Gamma\left(\frac{d}{2}\right) 2^{\frac{d}{2}} e^{\frac{A}{2}}} \int_0^\infty \exp\left\{\left(\frac{d}{2} - 1\right)\log\left(\frac{u}{A} + 1\right) - \frac{u}{2}\right\}\, du \tag{1}$$

For $x > 0$ we have

$$x - \frac{x^2}{2} < \log(x+1) < x \tag{2}$$

Using the right-hand side inequality, we get

$$\begin{aligned}\int_0^\infty \exp\left\{\left(\frac{d}{2} - 1\right)\log\left(\frac{u}{A} + 1\right) - \frac{u}{2}\right\}\, du &< \int_0^\infty \exp\left\{\left(\frac{d}{2} - 1\right)\frac{u}{A} - \frac{u}{2}\right\}\, du \\ &= \frac{A}{\frac{A}{2} - \frac{d}{2} + 1}\end{aligned} \tag{3}$$

Using the left-hand side inequality from (2) and substituting $z = \left(\frac{A}{2} - \frac{d}{2} + 1\right) u/A$ in the right-hand side yields

$$(4)\quad \int_0^\infty \exp\left\{\left(\frac{d}{2}-1\right)\log\left(\frac{u}{A}+1\right)-\frac{u}{2}\right\}\,du > \frac{A}{\frac{A}{2}-\frac{d}{2}+1}\int_0^\infty \exp\left\{-z-\frac{z^2\left(\frac{d}{2}-1\right)}{2\left(\frac{A}{2}-\frac{d}{2}+1\right)^2}\right\}\,dz$$

Because of Lemma 2.1 when $(A-d)^2/d \to \infty$ we have

$$\int_0^\infty \exp\left\{-z-\frac{z^2\left(\frac{d}{2}-1\right)}{2\left(\frac{A}{2}-\frac{d}{2}+1\right)^2}\right\}\,dz \to 1$$

Combining the above limit, (3) and (4) gives us the desired convergence. □

Lemma 2.3. *Let $Y_1, Y_2, \ldots, Y_N$ be N independent χ^2_d random variables. If α is a constant, $A > d$, $d \to \infty$, $(A-d)^2/d \to \infty$ and N is such that $NP(Y_1 > A) = \alpha$ for some constant α, then*

$$P(\max\{Y_i\}_1^N \le A) \sim \exp\left\{-\frac{A^{\frac{d}{2}}N}{\Gamma\left(\frac{d}{2}\right)2^{\frac{d}{2}}e^{\frac{A}{2}}\left(\frac{A}{2}-\frac{d}{2}+1\right)}\right\}$$

P r o o f. Under the conditions of the Lemma $(A-d)/\sqrt{2d} \to \infty$. In addition $(Y_1-d)/\sqrt{2d}$ converges in distribution to $N(0,1)$, thus

$$P(Y_1 > A) = P\left(\frac{Y_1-d}{\sqrt{2d}} > \frac{A-d}{\sqrt{2d}}\right) \to 0$$

We have that $NP(Y_1 > A) = \alpha$, therefore $N \to \infty$. For N independent χ^2_d random variables:

$$P(\max\{Y_i\}_1^N \le A) = \{P(Y_1 \le A)\}^N = \{1-P(Y_1 > A)\}^N = \left(1-\frac{\alpha}{N}\right)^N \to$$

$$\to \exp(-\alpha) = \exp\{-NP(Y_1 > A)\} \sim \exp\left\{-\frac{A^{\frac{d}{2}}N}{\Gamma\left(\frac{d}{2}\right)2^{\frac{d}{2}}e^{\frac{A}{2}}\left(\frac{A}{2}-\frac{d}{2}+1\right)}\right\} \quad \square$$

Lemma 2.4. *Let $Y_1, Y_2, \ldots, Y_N$ be $N = 2^{n-1} - 1$ independent χ^2_d random variables, where $d \to \infty$. For a constant $\alpha \in (0,1)$, the cutoff value $A = A(\alpha, d, n)$ so that $P(\max\{Y_i\}_1^N \leq A) = \alpha$ is a solution of*

$$\log(-\log(\alpha)) - \log(2^{n-1} - 1) + \log(\sqrt{4\pi}) + o(1) =$$

$$= \frac{d}{2}\log(A) - \left(\frac{d}{2} - \frac{1}{2}\right)\log(d) + \frac{d}{2} - \frac{A}{2} - \log\left(\frac{A}{2} - \frac{d}{2} + 1\right) \tag{5}$$

P r o o f. Let b be such that $\alpha = \exp(-\exp(b))$. Because of Lemma 2.3, we need to solve:

$$-e^b = -\frac{A^{\frac{d}{2}}(2^{n-1} - 1)}{\Gamma\left(\frac{d}{2}\right) 2^{\frac{d}{2}} e^{\frac{A}{2}} \left(\frac{A}{2} - \frac{d}{2} + 1\right)}, \text{ or}$$

$$b = \frac{d}{2}\log(A) + \log(2^{n-1} - 1) - \log\Gamma\left(\frac{d}{2}\right) - \frac{d}{2}\log 2 - \frac{A}{2} - \log\left(\frac{A}{2} - \frac{d}{2} + 1\right) \tag{6}$$

Substituting Binet's formula

$$\log\Gamma\left(\frac{d}{2}\right) = \log\sqrt{2\pi} + \left(\frac{d}{2} - \frac{1}{2}\right)\log\left(\frac{d}{2}\right) - \frac{d}{2} + o(1)$$

in (6), after transformations we get

$$b - \log(2^{n-1} - 1) + \log\sqrt{4\pi} + o(1) = \frac{d}{2}\log(A) - \left(\frac{d}{2} - \frac{1}{2}\right)\log(d) + \frac{d}{2} - \frac{A}{2} - \log\left(\frac{A}{2} - \frac{d}{2} + 1\right)$$

Expressing b in terms of α yields the result of the Lemma. □

3. A conservative test.

3.1. Proposed conservative test.

Lemma 3.1. *Let n be fixed and $Y_1, Y_2, \ldots, Y_N$ be $N = 2^{n-1} - 1$ independent χ^2_d random variables with maximum $Y_{\max}$. Then asymptotically in d for a fixed A we have*

$$P(B_{\max} > A) \leq P(Y_{\max} > A)$$

P r o o f. For $i = 1, 2 \ldots N$, we denote the standardized versions of B_i and Y_i by asterisks, $B_i^* = (B_i - d)/\sqrt{2d}$ and $Y_i^* = (Y_i - d)/\sqrt{2d}$. Then $E(B_i^*) = E(Y_i^*) = 0$ and $Var(B_i^*) = Var(Y_i^*) = 1$. Let $B^*_{\max} = (B_{\max} - d)/\sqrt{2d}$, $Y^*_{\max} = (Y_{\max} - d)/\sqrt{2d}$ and $A^* = (A - d)/\sqrt{2d}$.

Let S_i and S_j be two splits of X into two clusters. Let the cluster sizes for S_i be $n_{1,i}$ and $n_{2,i}$ and the respective centers be $c_{1,i}$ and $c_{2,i}$. Let the cluster sizes for S_j be $n_{1,j}$ and $n_{2,j}$ and the respective centers be $c_{1,j}$ and $c_{2,j}$. Therefore

$$\begin{aligned} Cor(B_i^*, B_j^*) &= Cor\left(\left(\frac{1}{n_{1,i}} + \frac{1}{n_{2,i}}\right)^{-1} \|c_{1,i} - c_{2,i}\|^2, \left(\frac{1}{n_{1,j}} + \frac{1}{n_{2,j}}\right)^{-1} \|c_{1,j} - c_{2,j}\|^2\right) \\ &= Cor(\|c_{1,i} - c_{2,i}\|^2, \|c_{1,j} - c_{2,j}\|^2) \\ &= \{Cor(c_{1,i} - c_{2,i}, c_{1,j} - c_{2,j})\}^2 \end{aligned}$$

The latter equality is true because $c_{1,i} - c_{2,i}$ and $c_{1,j} - c_{2,j}$ are normally distributed. Therefore $Cor(B_i^*, B_j^*) = Cor(B_i, B_j) \geq 0$. In addition, by independence $Cor(Y_i, Y_j) = 0$.

For splits S_i, $1 \leq i \leq N$, consisting of clusters with $n_{1,i}$ and $n_{2,i}$ elements and centers $c_{1,i}$ and $c_{2,i}$ respectively, we can write

$$B(X, S_i) = \left(\frac{1}{n_{1,i}} + \frac{1}{n_{2,i}}\right)^{-1} \sum_{j=1}^{d} (c_{1,i}^j - c_{2,i}^j)^2$$

Therefore for arbitrary N coefficients $\beta_1, \beta_2 \ldots \beta_N$, we have

$$\sum_{i=1}^{N} \beta_i B(X, S_i) = \sum_{i=1}^{N} \left\{ \beta_i \left(\frac{1}{n_{1,i}} + \frac{1}{n_{2,i}}\right)^{-1} \sum_{j=1}^{d} (c_{1,i}^j - c_{2,i}^j)^2 \right\}$$

$$= \sum_{j=1}^{d} \left\{ \sum_{i=1}^{N} \beta_i \left(\frac{1}{n_{1,i}} + \frac{1}{n_{2,i}} \right)^{-1} (c_{1,i}^j - c_{2,i}^j)^2 \right\}$$

The coordinates are independent, therefore $\sum_{i=1}^{N} \beta_i (1/n_{1,i} + 1/n_{2,i})^{-1} (c_{1,i}^j - c_{2,i}^j)^2$ are independent and identically distributed over the different values of j (i.e. coordinates). Therefore by the Central Limit Theorem, $\sum_{i=1}^{N} \beta_i B(X, S_i)$ approaches normal distribution when $d \to \infty$. Thus, asymptotically in d, $\{B_i\}_{i=1}^N \equiv \{B(X, S_i)\}_{i=1}^N$ are jointly normal. By the asymptotic behavior of χ_d^2 when $d \to \infty$, $\{Y_i\}_{i=1}^N$ are also jointly normal. The process $\{B_i^*\}_{i=1}^N$ is normal for being a linear transformation of $\{B_i\}_{i=1}^N$, similarly $\{Y_i^*\}_{i=1}^N$ is normal.

Asymptotically in d, the processes $\{B_i^*\}_{i=1}^N$ and $\{Y_i^*\}_{i=1}^N$ are normal, centered and each of the variables has unit variance. In addition, $Cor(B_i^*, B_j^*) \geq 0 = Cor(Y_i^*, Y_j^*)$. According to a generalized version, e.g., Rinott [6], of the Slepian inequality [7], for such processes, the less correlated one has a larger (in probability) maximum. Thus asymptotically in d for each fixed A^*, $P(B_{\max}^* > A^*) \leq P(Y_{\max}^* > A^*)$. Therefore $P(B_{\max} > A) \leq P(Y_{\max} > A)$. □

Theorem 3.2. *Asymptotically in d, the test with rejection region $B_k(X) > A(\alpha, d, n)$ is a conservative level $\alpha^* = 1-\alpha$ test for testing whether the data comes from $N(0, I_d)$ against two spherical normals.*

Proof. The k-means method does not guarantee a global maximum, thus $B_k(X) \leq B_{\max}(X)$ and $P(B_k(X) > A) \leq P(B_{\max}(X) > A)$. From Lemma 3.1 we have $P(B_{\max} > A) \leq P(Y_{\max} > A)$, where $Y_{\max}$ is defined in that Lemma. Therefore

$$P(B_{\max} > A) \leq P(Y_{\max} > A) = 1 - P(Y_{\max} < A) = 1 - \alpha = \alpha^*$$

The rejection probability is no more than $1 - \alpha$ thus the test is conservative at that level. □

3.2. Cutoff point asymptotic behavior.

Lemma 3.3. *If $n \to \infty$, $d \to \infty$ and $n/d \to 0$ then*

$$\sqrt{A} - \sqrt{d} \sim \sqrt{(n-1)\log(2)}$$

P r o o f. When $n \to \infty$, $\log(2^{n-1}-1) = (n-1)\log(2) + o(1)$. For a fixed α, (5) becomes

$$-(n-1)\log(2) + O(1) = \frac{d}{2}\log(A) - \left(\frac{d}{2} - \frac{1}{2}\right)\log(d) + \frac{d}{2} - \frac{A}{2} - \log\left(\frac{A}{2} - \frac{d}{2} + 1\right)$$

Let B and ϵ be such that $A = Bd$, $B = 1 + \epsilon$. Substituting them in the above equation, transforming it and absorbing some terms in $O(1)$ results in

$$-(n-1)\log(2) + O(1) = \frac{d}{2}\log(1+\epsilon) - \frac{d\epsilon}{2} - \log\left(\frac{\epsilon\sqrt{d}}{2} + \frac{1}{\sqrt{d}}\right) \tag{7}$$

The left-hand side of the above equation converges to $-\infty$. We will prove that under the conditions of the Lemma, $d\epsilon^2 \to \infty$ and $\epsilon \to 0$.

If $d\epsilon^2$ has a subsequence bounded from above, then $\epsilon \to 0$, because $d \to \infty$. Therefore for that subsequence we can expand $\log(1+\epsilon)$ and (7) becomes

$$-(n-1)\log(2) + O(1) = -\frac{d\epsilon^2}{4}(1 + o(1)) - \log\left(\frac{\epsilon\sqrt{d}}{2} + \frac{1}{\sqrt{d}}\right)$$

For the subsequence bounded from above, the right-hand side of the latter equation is bounded from below, whereas the left-hand side converges to $-\infty$. Therefore $d\epsilon^2 \to \infty$ and we can write

$$\log\left(\frac{\epsilon\sqrt{d}}{2} + \frac{1}{\sqrt{d}}\right) = \log\left(\frac{\epsilon\sqrt{d}}{2}\right) + o(1) \tag{8}$$

Substituting the latter equality in (7), dividing by $d/2$, absorbing $\log(\sqrt{d}/2)/d$ in $o(1)$, and $o(1)$ in $O(1)$ yields

$$\frac{-(n-1)\log(2) + O(1)}{d/2} = \log(1+\epsilon) - \epsilon - \frac{2\log(\epsilon)}{d} + o(1)$$

Under the conditions of the Lemma, the left-hand side converges to 0, therefore

$$\log(1+\epsilon) - \epsilon - \frac{2\log(\epsilon)}{d} \to 0 \tag{9}$$

If $\epsilon \to \infty$ when $d \to \infty$ then $-2\log(\epsilon)/d \le 0$ (in limit) and $\log(1+\epsilon) - \epsilon \to -\infty$, therefore (9) does not hold.

If ϵ is bounded and away from 0 (i.e., there are constants c and C so that $0 < c < \epsilon < C$), then $2\log(\epsilon)/d \to 0$ when $d \to \infty$. In addition, $\log(1+\epsilon) - \epsilon$ is bounded from above away from 0 (i.e. $\log(1+\epsilon) - \epsilon < \log(1+c) - c < 0$) and (9) does not hold.

Therefore $\epsilon \to 0$ and $d\epsilon^2 \to \infty$. Using (8), transferring $O(1)$ to the right-hand side, expanding $\log(1+\epsilon)$ and absorbing minor-order terms into $d\epsilon^2$, turns (7) into

$$-(n-1)\log(2) = -\frac{d\epsilon^2}{4}(1+o(1))$$

Therefore

$$\frac{d\epsilon^2}{4} = (n-1)\log(2)(1+o(1)) \text{ and } \epsilon = 2\sqrt{\frac{(n-1)\log(2)}{d}}(1+o(1))$$

Finally,

$$\begin{aligned} A = Bd = (1+\epsilon)d &= d + 2\sqrt{d(n-1)\log(2)}(1+o(1)) \\ &= d + 2\sqrt{d}\sqrt{(n-1)\log(2)}(1+o(1)) + (n-1)\log(2)(1+o(1)) \\ &\quad - (n-1)\log(2)(1+o(1)) \end{aligned}$$

Under the conditions of the Lemma $n/d \to 0$, therefore $(n-1)\log(2)(1+o(1))$ gets absorbed in $\sqrt{d}\sqrt{(n-1)\log(2)}o(1)$, thus

$$\begin{aligned} A &= d + 2\sqrt{d}\sqrt{(n-1)\log(2)}(1+o(1)) + (n-1)\log(2)(1+o(1)) \\ &= \{\sqrt{d} + \sqrt{(n-1)\log(2)}(1+o(1))\}^2 \end{aligned}$$

Therefore $\sqrt{A} - \sqrt{d} = \sqrt{(n-1)\log(2)}(1+o(1))$ or $\sqrt{A} - \sqrt{d} \sim \sqrt{(n-1)\log(2)}$. □

4. Suggested test and separation detection.

4.1. Suggested general test. Using numerical simulations for $n = 100$ and $d = 10$ we found that less than 0.1% of the empirical 95th percentiles of B_k are higher than $A(0.95, 10, 100)$. For larger values of d and n, the fraction of 95th percentiles that is larger than $A(0.95, d, n)$ is even lower. This is consistent with Theorem 3.2, but the test is too conservative and is not very useful.

$T(X)$ has a $\chi^2_{(n-1)d}$ distribution with mean $(n-1)d$. Standardizing $\sqrt{B_k}$ by $\sqrt{T(X)/(n-1)d}$ results in the proposed test $R(X) = \sqrt{B_k(X)(n-1)d/T(X)}$.

Theorem 4.1. *Let $B_k(X)$ be the between cluster sum of squares found by the $k-means$ method and $T(X)$ be the total sum of squares. Also let $R = \sqrt{B_k(X)(n-1)d/T(X)}$. Then the test with rejection region*

$$R(X) > \mu_{n,d} + z_{(1-\alpha^*)}\sigma_{n,d}$$

where $z_{(1-\alpha^*)}$ is the $1-\alpha^*$ quantile of the standard normal distribution,

$$\mu_{n,d} = -0.493 + \sqrt{d} + 0.789\sqrt{n-1} - 0.00018\sqrt{d(n-1)} \text{ and}$$

$$\sigma_{n,d} = 0.159 + 0.604/\sqrt{d} + 0.697/\sqrt{n} - 0.9595/\sqrt{d(n-1)}$$

is approximately $\alpha^* = 1-\alpha$ significance level test for testing whether the data came from one multidimensional normal distribution vs. two normal distributions.

P r o o f. We use numerical simulations for several values of d between 10 and 5000 and n between 10 and 200. For each pair of values of d and n, we use the k-means method to simulate 10000 observations from B_k. For each observation, the k-means method was used with 100 choices of initial cluster means and 100 iterations within each attempt. The results for confidence level = 0.90 (significance level $\alpha^* = 10\%$) and confidence level = 0.95 (significance level $\alpha^* = 5\%$) are given in Tables 1 and 2. Each table entry shows the proportion of the simulated B_k values that are within the proposed test rejection region.

Table 1. Comparison with simulation results for confidence level = 0.90, significance level = 10%

n\d	10	20	50	100	200	500	1000	2000	5000
10	0.0239	0.0355	0.0470	0.0539	0.0595	0.0610	0.0604	0.0660	0.0810
20	0.0330	0.0446	0.0628	0.0733	0.0717	0.0780	0.0842	0.0918	0.1010
50	0.0425	0.0603	0.0764	0.0790	0.0839	0.0912	0.0854	0.1000	0.1240
100	0.0521	0.0730	0.0873	0.0892	0.0882	0.0820	0.0854	0.0952	0.1148
200	0.0683	0.0894	0.1052	0.1054	0.0906	0.0805	0.0595	0.0672	0.0829

Table 2. Comparison with simulation results for confidence level = 0.95, significance level = 5%

n\d	10	20	50	100	200	500	1000	2000	5000
10	0.0090	0.0137	0.0212	0.0242	0.0280	0.0299	0.0300	0.0318	0.0360
20	0.0137	0.0199	0.0297	0.0381	0.0346	0.0407	0.0431	0.0440	0.0520
50	0.0169	0.0300	0.0377	0.0428	0.0465	0.0491	0.0446	0.0548	0.0674
100	0.0239	0.0367	0.0467	0.0475	0.0466	0.0435	0.0418	0.0488	0.0622
200	0.0307	0.0481	0.0568	0.0554	0.0477	0.0422	0.0298	0.0340	0.0483

We can see that the proposed test performs very well, especially for moderate and large values of d. Moreover, the cutoff points for both the practical and conservative tests have the same asymptotic behavior with the main term $\sqrt{d}$ being the same. The term $\sqrt{d(n-1)}$ in $\mu_{m,d}$ takes care of moderate values of n and d. $\square$

4.2. Separation detection. If the data consists of $n/2$ observations from each of two normal distributions with covariance matrices I_d and distance c between their means, then the interclusters sum of squares for the genuine bimodality direction is given by $B_{real} = nc^2/4$, therefore $\sqrt{B_{real}} = c\sqrt{n}/2$. According to our asymptotic results, in other directions $\sqrt{B}$ behaves like $\sqrt{d}+\sqrt{(n-1)\log(2)}$. For bimodality in one dimension (the genuine bimodality dimension), the distance between the means c should be at least 2. When the clusters in the real bimodality directions are just enough separated for bimodality, $c = 2$. If $\sqrt{d} + \sqrt{(n-1)\log(2)} > c\sqrt{n}/2 = 2\sqrt{n}/2 = \sqrt{n}$, then the between clusters sum of squares is larger in a direction different than the genuine bimodality direction. The latter condition is (approximately) equivalent to $\sqrt{d/n} + \log(2) > 1$, or $d > n/10$. Therefore, for a fixed number of points n, the genuine bimodality is hard to discover even in a relatively small number of dimensions.

5. Conclusion remarks. The asymptotic result $\sqrt{A} - \sqrt{d} \sim \sqrt{(n-1)\log(2)}$ is similar to the expression for the square root of the shift parameter (except for the $\log(2)$ term) in an article by Johnstone [5] about the asymptotic distribution of the largest principal component variance of the data covariance matrix. It is possible that the maximum of many χ^2 random variables is also related to the Tracy-Widom distribution.

Cover's theorem (e.g. [4]) deals with similar behavior, according to it, when increasing the dimensionality, the probability of linear separation increases.

Finding A or its asymptotic behavior is equivalent to inverting the incomplete gamma distribution. Temme [8] discusses the topic, but there is no close form asymptotic solution in his article.

Acknowledgments. This article contains the theoretical part of the author's Ph.D. dissertation with the same title, presented to the Faculty of the Graduate School of Yale University in 2006. The author would like to thank John A. Hartigan for the guidance, thoughtful advice and patience, William N. Goetzmann and Hannes Leeb for their careful reviews and comments on the dissertation, and also the anonymous reviewers of this manuscript for their helpful suggestions.

REFERENCES

[1] Day N. E. Estimating the components of a mixture of normal distributions. *Biometrica,* **56** (1969), No 3, 463–474.

[2] Hartigan J. A. Clustering Algorithms. John Wiley and Sons, New York, NY, 1975.

[3] Hartigan J. A. Asymptotic distributions for clustering criteria. *The Annals of Statistics,* **6** (1978), No 1, 117–131.

[4] Haykin, S. Neural networks: A comprehensive foundation. Macmillan College Company, Inc, Englewood, NJ, 1994.

[5] Johnstone I. M. On The Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics,* **29** (2001), No 2, 295–327.

[6] Rinott Y. On two-stage selection procedures and related probability inequalities. *Communications in Statistics: Theory and Methods,* **A7** (1978), No 8, 799–81.

[7] Slepian D. The one-sided barrier problem for Gaussian noise. *The Bell System Technical Journal,* **41** (1962), 463–501.

[8] Temme N. Asymptotic inversion of incomplete gamma functions. *Mathematics of Computation,* **58** (1992), No 198, 755–764.

Dean Palejev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Block 8
1113 Sofia, Bulgaria
e-mail: palejev@math.bas.bg

Received December 7, 2012
Final Accepted January 14, 2013