Small Sample Size in High Dimensional Space - Minimum Distance Based Classification

Ewa Skubalska-Rafajłowicz

Institute of Computer Engineering, Automatics and Robotics,
Department of Electronics, Wrocław University of Technology, Poland
ewa.rafajlowicz@pwr.wroc.pl

Abstract. In this paper we present some new results concerning classification in small sample and high dimensional case. We discuss geometric properties of data structures in high dimensions. It is known that such data form in high dimension an almost regular simplex, even if covariance structure of data is not unity. We restrict our attention to two class discrimination problems. It is assumed that observations from two classes are distributed as multivariate normal with a common covariance matrix. We develop consequences of our findings that in high dimensions N Gaussian random points generate a sample covariance matrix estimate which has similar properties as a covariance matrix of normal distribution obtained by random projection onto subspace of dimensionality N. Namely, eigenvalues of both covariance matrices follow the same distribution. We examine classification results obtained for minimum distance classifiers with dimensionality reduction based on PC analysis of a singular sample covariance matrix and a reduction obtained using normal random projections. Simulation studies are provided which confirm the theoretical analysis.

Keywords: small sample, classification in high dimensions, eigenvalues of a sample covariance matrix, maximum likelihood ratio, normal random projection, minimum distance rule.

1 Introduction

Analysis of high-dimension low-sample size classification is one of most important problems both from theoretical and practical point of view. It often happens that the dimension d of data vectors is larger than the sample size N and this case is referred to as small sample size, high dimensional data. Microarrays, medical imaging, text recognition, finance and chemometrics are examples of such classification problems. On the other hand we know that in practice, statistical methods based on very small sample sizes might not be reliable. Many results in this area have been obtained in the asymptotic setting, when both dimension d of the vector observations and the size of the data sample N is very large, with d possibly much larger than N [2], [19], [4]. It is assumed that d and N grow at the same rate, i.e. $d/N \to \gamma$ as $d \to \infty$ [19]. Others focus their attention on

L. Rutkowski et al. (Eds.): ICAISC 2014, Part I, LNAI 8467, pp. 610–621, 2014.
© Springer International Publishing Switzerland 2014

the case that dimension d increases while the sample size N is fixed [1], [8], [10], [9]. In [13] and [8] it was observed that data in high dimensions form an almost regular simplex and distances between $N < d$ random points are very close to $c\sqrt{(2d)}$.

Here we restrict our attention to two class discrimination problems. It is assumed that observations from two classes are distributed as multivariate normal with a common covariance matrix. We assume that the number of available samples equals N and consists of independent vector observations from both classes, N_0 and N_1, respectively.

Using the very popular Fisher classifier LDA (linear discrimination analysis) [5], [21], [16], [19] when the data are from the normal distribution with a common covariance matrix: $X_l \sim N(m_l, \Sigma)$ for $l = 0, 1$, we can estimate a single, pooled covariance matrix as an estimate of the common covariance matrix:

$$S = \frac{1}{N-2}\left(\sum_{j=1}^{N_0}(X_{0j} - \bar{X}_0)(X_{0j} - \bar{X}_0)^T + \sum_{j=1}^{N_1}(X_{1j} - \bar{X}_1)(X_{1j} - \bar{X}_1)^T\right),$$

where

$$\bar{X}_0 = \sum_{j=1}^{N_0} X_{0j}, \quad \bar{X}_1 = \sum_{j=1}^{N_1} X_{1j},$$

and $N = N_0 + N_1$.

The Fisher classification rule is based on

$$D(X) = (X - M)S^{-1}(\bar{X}_0 - \bar{X}_1), \tag{1}$$

where $M = (\bar{X}_0 + \bar{X}_1)/2$. If $D(X) > 0$ classify X to class C_0 (labeled by 0), otherwise classify X to class C_1 (labeled by 1).

The maximum likelihood ratio (MLR) rule [11], [19] classifies X to the class C_0 if

$$\frac{N_0+1}{N_0}(X - \bar{X}_0)^T S^{-1}(X - \bar{X}_0) \leq \frac{N_1+1}{N_1}(X - \bar{X}_1)^T S^{-1}(X - \bar{X}_1). \tag{2}$$

The MLR rule can be also used when covariances in the both classes differ. Firstly, one estimates the covariance matrix of each class, based on samples known to belong to each class. Then, given a new sample X, one computes the squared Mahalanobis distance [12] to each class, i.e.,

$$(X - \bar{X}_l)^T S_l^{-1}(X - \bar{X}_l), \; l = 0, 1$$

and classifies the new point as belonging to that class for which the (weighted by $\frac{Nl+1}{Nl}$) Mahalanobis distance is minimal. It is well known [22] (and easy to show) that the MLR rule coincides with the Fisher method when numbers of samples from both classes are equal to each other, i.e., $N_0 = N_1$. Mahalanobis distance is also closely related to Hotelling's T-square distribution used for multivariate statistical testing [15]. It is, however, hard to implement the covariance based

classification methods when dimensionality is high due to the difficulty of estimating the unknown covariance matrix. Even if number of samples N is greater than the dimension of the data it is advisable to reduce dimensionality due to some near zero eigenvalues of S. Thus, it is proposed to drop zero or near zero eigenvalues of S.

If the number of data samples is smaller than the dimension of the data space, the sample based estimate of the covariance matrix is singular with probability one. Srivastava [19] has proposed a sample-squared distance between the two groups, using the Moore–Penrose pseudoinverse of the singular sample covariance matrix S.

The Moore–Penrose pseudoinverse of a matrix A is unique and is defined as matrix A^+ satisfying the following four properties:

- $AA^+A = A$,
- $A^+AA^+ = A^+$
- $(AA^+)^T = AA^+$,
- $(A^+A)^T = A^+A$

The sample covariance matrix S is a symmetric positive semidefinite matrix. Thus, it can be written as

$$S = QDQ^T,$$

where D is diagonal with the positive eigenvalues of S, and $Q \in R^{d\times N}$ is orthogonal and consists of N eigenvectors of S connected to the positive eigenvalues of S. This orthogonal decomposition is often called principal components analysis (PCA). The sample covariance matrix provides the conventional estimator of principal component analysis (PCA) through the eigenvalue-eigenvector decomposition. For the covariance or correlation matrix, the eigenvectors correspond to principal components and the eigenvalues to the variance explained by the principal components. The MoorePenrose inverse of S is defined by

$$S^+ = QD^{-1}Q^T,$$

where $d - N$ principal components directions connected to zero (or close to zero) eigenvalues are removed. For numerical matrices computations of their pseudoinverses are based on singular value decomposition (SVD) [7].

In this paper we will show why in many cases it is more effective to use unit diagonal matrix I and the Euclidean distance instead of Mahalanobis distance based on the pseudoinverse of the sample covariance matrix S^+ in the context of small sample and high dimension classification problems. The analysis is followed by some simulation experiments which indicate that the same phenomena one can observe also in a relatively small dimension in comparison to many practical high dimensional problems. The most important condition is that the number of samples is smaller than the dimension of the data.

The starting point of the paper is the observation that in high dimensions N Gaussian random points generate a sample covariance matrix estimate which

has similar properties as a covariance matrix of normal distribution obtained by random projection onto subspace of dimensionality N. Namely, eigenvalues of both covariance matrices follow the same distribution. This property explains why PC analysis of singular sample covariance matrix (for $N < d$) leads to the similar results as dimensionality reduction made at random.

The next section describes geometric properties of small sample data in high dimensions. Section 3 is concentrated on properties of both mentioned earlier dimensionality reduction methods. Section 4. shows some simulation results which explain and confirm proposals and conclusions developed in the previous sections. In Section 5 we summarize the provided theoretical and simulation results.

2 Geometric Structure of Data in High Dimensions

It is well known that high dimensional data concentrates close to the surface of a hypersphere. For example, if samples are drawn according to multidimensional normal distribution in a high-dimensional space, the center region where the value of the density function is largest is empty and almost all samples are located close to the sphere of radius $\sqrt{(Trace(\Sigma))} = c\sqrt{(d)}$ (for detailed assumptions see [8] or [1], [6]). Figure 1 (left panel) illustrates this phenomenon, where diagram of the ordered lengths of 100 iid random observations taken from normal distribution $N_d(0, I)$ of dimensionality $d = 20000$ is depicted. Furthermore, as is indicated in [8], [3], the data in high dimension form an almost regular simplex. The distances between random points are very close to $c\sqrt{(2d)}$. Figure 1, right panel depicts a diagram of the Euclidean distances between the same set of 100 points. Mean Euclidean distance between these points (iid normal $N_d(0, 1)$, for $d = 20000$) equals 199.69 with standard deviation 0.78.

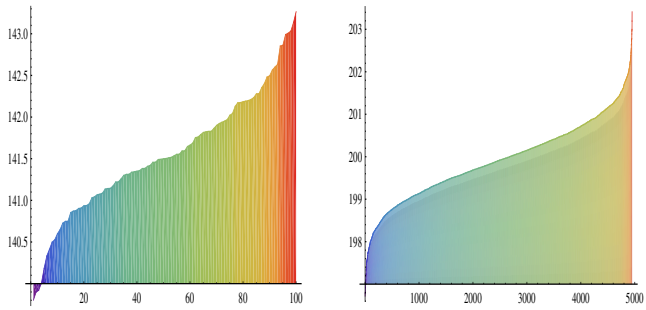

Fig. 1. Left) Diagram of the length of 20000 dimensional random vectors generated independently according to multivariate normal distribution (100 samples). Right) Diagram of the Euclidean distances between the same set of points.

This means that the variability of the small sample of high-dimensional data is contained only in the random rotation of this simplex. For random vector X being the multivariate Gaussian distribution with identity covariance matrix I it is known that $||X||^2$ has χ^2 distribution with d degree of freedom and

furthermore $||X||$ follows χ distribution with the same degree of freedom. The mean of χ distribution is

$$\sqrt{(2)}\frac{\Gamma[(d+1)/2]}{\Gamma[d/2]}$$

and $\mu = E||X|| \to \sqrt{(d)}$ as $d \to \infty$ and its variance equals $d - \mu^2$.

In general the Euclidean norm of a multivariate normally distributed random vector follows a noncentral χ distribution.

Ahn et al [1] have shown, using asymptotic properties of sample covariance matrices, that the conditions for forming by a small data sample a regular simplex are rather mild. More precisely, they have shown, that if eigenvalues of the true covariance matrix Σ are all distinct and positive

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0,$$

and if

$$\frac{\sum_{i=1}^{d} \lambda_i^2}{(\sum_{i=1}^{d} \lambda_i)^2} \to 0 \tag{3}$$

as $d \to \infty$, then the nonzero eigenvalues of the sample covariance matrix behave as if they are from diagonal matrix $\frac{Trace(\Sigma)}{N} I_N$.

The similar phenomenon one can obtain for a uniform distribution. Figure 2 shows a diagram of the lengths of 20000 dimensional random vectors generated independently according to the uniform distribution from $[0,1]^d$ cube (100 samples) and of the Euclidean distances between the same set of points. Mean Euclidean distance between 100 points iid uniform from unit cube $[0,1]^d$, for $d = 20000$ equals 57.74 with standard deviation 0.236. Mean vector‘s length (averaged over 100 observations) equals 81.60 with standard deviation equal to 0.236.

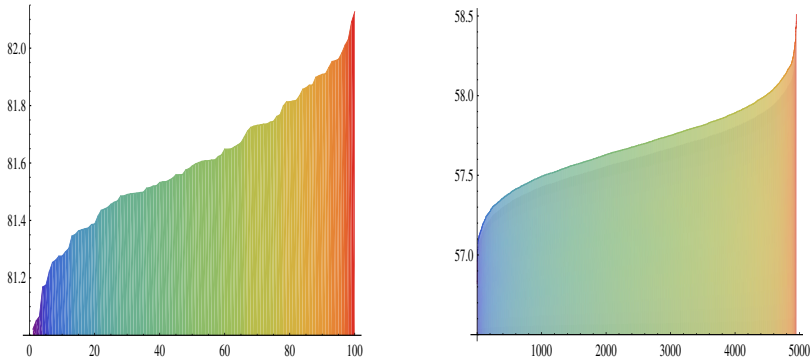

Fig. 2. Left) Diagram of the lengths of 20000 dimensional random vectors generated independently according to the uniform distribution from $[0,1]^d$ cube (100 samples). Right) Diagram of the Euclidean distances between the same set of points.

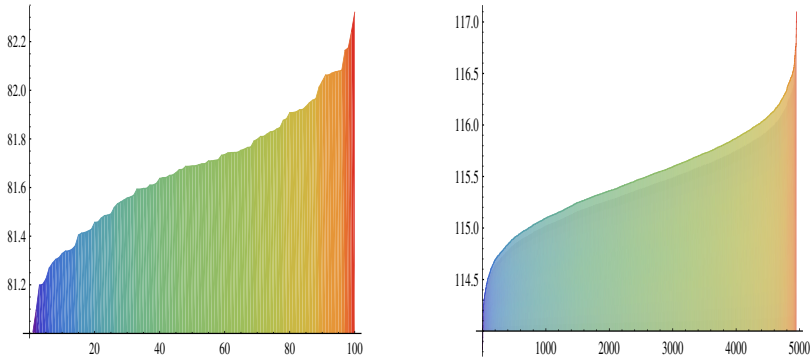

Fig. 3. Left) Diagram of the lengths of 20000 dimensional random vectors generated independently according to the uniform distribution from $[-1,1]^d$ cube (100 samples). Right) Diagram of the Euclidean distances between the same set of points.

Further examples are given in Figures 3, 4 and 5.

Mean Euclidean distance between 100 points iid uniform unit cube $[-1,1]^d$, for $d = 20000$ equals 115.49 with standard deviation 0.473. Mean vector's length (averaged over 100 observations) equals 81.64 with standard deviation equal to 0.236 (see Figure 3).

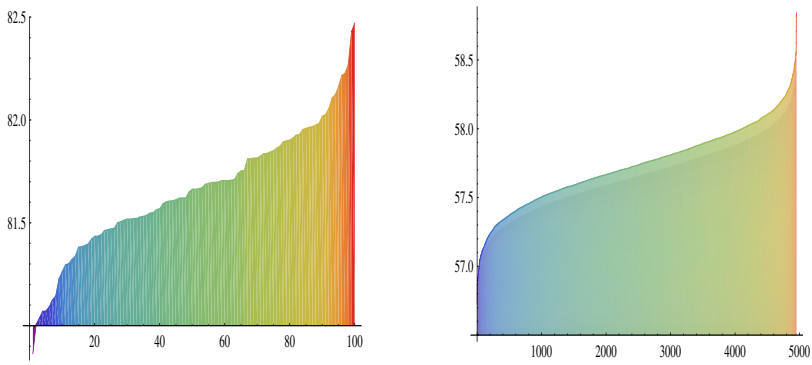

Fig. 4. Diagram of the normalized lengths of 20000 dimensional random vectors generated independently according to the uniform distribution from $[0, \lambda_i^{0.5}]^d$ cube (100 samples). Right) Diagram of the Euclidean distances between the same set of points with scaling factor $1/\sqrt{(\sum \lambda_i^2/d)}$ with $\lambda_i^2 = i$.

Mean vector's length (averaged over 100 observations) equals 8165 with standard deviation equal to 30.28. The adequate mean with rescaling factor

$$1/\sqrt{(\sum \lambda_i^2/d)} \approx 100$$

equals 81.65. Mean Euclidean distance between 100 points iid uniform from unit cube $[0, i^{0.5}]^d$, for $d = 20000$ equals 57.73 with standard deviation 0.284 (see Figure 4).

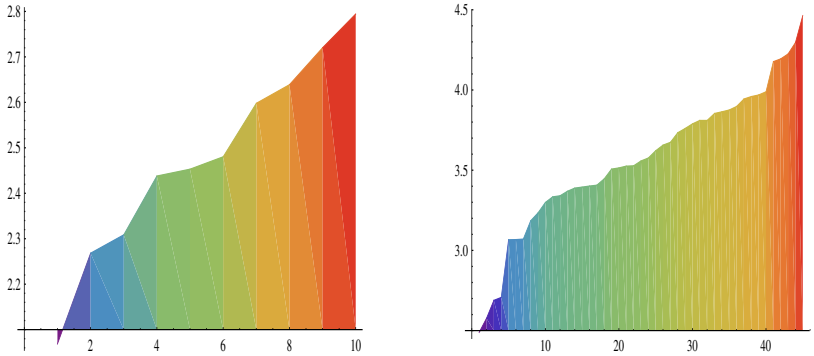

Fig. 5. Left) Diagram of the lengths of 20 dimensional random vectors generated independently according to the uniform distribution from $[-1, 1]^d$ cube (10 samples). Right) Diagram of the Euclidean distances between the same set of points.

Even if a dimension of data is relatively small, the random observations form a rather regular structure. The mean Euclidean distance between 10 points iid uniform from unit cube $[-1, 1]^d$, for $d = 20$ equals 3.56 with standard deviation 0.447 (see Figure 5).

Regular, close to simplex structure of the data indicate that the sample covariance estimate for small sample size $N < d$ will be in most cases rather regular with rather uniform nonzero eigenvalues and corresponding eigenvectors will form random subspace of the original data space. This problem will be analyzed in more detail in the next section.

3 Random Projections and Eigenvalues of a Sample Covariance Matrix

As previously we consider Gaussian data with positive definite covariance matrix Σ, i.e., such that its eigenvalues are all distinct and strictly positive

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0.$$

Suppose we have a data matrix $Y \in R^{d \times N}$, where $N \leq d$. For the sake of simplicity we will assume that the class means are known. Let's say that $m_0 = 0$ and $m_1 = \mu$. Thus, without loss of generality we can assume that each column of Y is iid normal $N_d(0, \Sigma)$, since every observation from the class labeled by 1, let say X, is replaced by $X - \mu$.

It is well known [7] that the nonzero eigenvalues of $S = YY^T/N$ are also the eigenvalues of Y^TY/N. Srivastava observed [19] that if columns of $d \times N$ matrix

Y are independent identically distributed observations of normal random vector $\sim N_d(0, \Sigma)$ then the eigenvalues of $Y^T Y$ have the same distribution as the eigenvalues of

$$U^T \Sigma U,$$

where $U \in R^{d \times N}$ is a random matrix with entries iid $N(0,1)$ (see also [1]).

When number of samples N is smaller than a dimension of a data, then randomly chosen vector observations Y and vector 0 span a random subspace of the original data space of dimensionality N. If means are not given and should be estimated from data N centered vectors span (with probability one) $N-1$ dimensional space.

The same effect of choosing random subspace of the original space one can obtain using normal random projections [20], [17], [18], [3].

Let $Z \in R^{d \times N}$ be a random matrix with entries iid $N(0,1)$. We can project matrix of observations onto N dimensional subspace using transformation $V = Z^T Y$.

Each column of matrix $V \in R^{N \times N}$ is iid zero mean normal random vector with covariance matrix $Z^T \Sigma Z$. A normal random projection with projection matrix Z transforms $N_d(0, \Sigma)$ distribution onto $N_N(0, Z^T \Sigma Z)$. Thus, instead of $x^T S^+ x$ we can analyze a random quadratic form in normal variables $N_N(0, Z^T \Sigma Z)$:

$$(Z^T x)^T (Z^T S Z)^{-1} (Z^T x). \tag{4}$$

It is easy to show that matrix Z can be decomposed into $Z = R\tilde{Q}$, where

$$\tilde{Q} \in St(d, N) = \{A \in R^{d \times N} : A^T A = I_N\}$$

consists of N columns of rotation matrix obtained from Z by the Gram–Schmidt orthogonalization process and $R \in R^{N \times N}$ is an adequate scaling matrix. So, (4) equals

$$x^T \tilde{Q} (\tilde{Q} S \tilde{Q}^T)^{-1} \tilde{Q}^T x. \tag{5}$$

This formula is very similar to the previous one, proposed by Srivastava [19], i.e.,

$$x^T Q D^{-1} Q^T x, \tag{6}$$

where $Q^T x$ follows the normal distribution with zero mean and the covariance matrix $Q^T \Sigma Q$, and where D is random with respect to the matrix of learning samples Y.

The precise connections between (6) and (5) are still open question. However it is known, that the mean value of

$$\tilde{Q} (\tilde{Q} S \tilde{Q}^T)^{-1} \tilde{Q}^T$$

with respect to $\tilde{Q}$ (taken uniformly, i.e., according to the Haar measure on the compact Stiefel manifold $St(d, N)$) is equal to a pseudoinverse of S [14]. It is hard, for the numerical reasons, to apply this property in practice.

Notice that eigenvalues of covariance matrix $Z^T \Sigma Z$ of a normal distribution $N_N(0, Z^T \Sigma Z)$ obtained after random normal projection have the same distribution as nonzero eigenvalues of YY^T. This property is not asymptotic and it holds for every $N < d$.

In the next section we will show experimentally that using only one N dimensional projection at a time leads to the same mean classification error as the method MLR proposed by Srivastava when results are averaged over many different learning samples.

4 Numerical Results

We had performed the following experiment. A two class Gaussian problem was examined with different numbers of learning samples in dimensions $d = 20$. The classes differ in the mean, i.e., $m_0 = (0, \ldots, 0)^T$ and $m_1 = (1, \ldots, 1)^T$. Eigenvalues of the common covariance matrix equal $1, 2, \ldots, 20$. It is easy to check that such covariance structure fulfills assumption (3). For simplicity, we have also assumed that there are available $N_0 = N_1$ samples from both classes $N_0 = N_1 = 5,\ 8,\ 10,\ 20,\ 40,\ 100,\ 200$. 100 different learning samples were drawn and both methods were performed on 2×10000 testing samples (10000 for each class). Class means are also estimated from the data.

Figure 6 demonstrates a comparison of averaged classification accuracy obtained using different classification methods, namely: Srivastava modification of the MLR rule and the MLR rule applied to randomly projected data (using normal random projections) onto subspaces of dimension $k = 5,\ 10$ and 20. When the number of samples was greater than the dimension of the problem, the original version of the MLR rule was used and a pseudoinverse was replaced by the inverse of the sample covariance matrix estimate. For projection of dimensionality $k = 20$ which is in fact nonsingular the eigenvalues of the sample covariance matrix were almost the same. Small differences occur due to numerical errors. For $N_0 = N_1 = 5$ one of the estimated covariance matrices has the following nonzero eigenvalues:

$$\{56.8024, 41.7466, 30.6846, 28.181, 15.2784, 12.4807, 5.73685, 3.17399\}$$

and the trace of these covariance matrices equals ≈ 194.1. When the dimension of a projection was lower than 20, the eigenvalues of the estimated covariance matrix were more spread. For example, for $k = 10$ (and the same learning sample) we have obtained the following set of nonzero eigenvalues:

$$\{49.1553, 39.4426, 31.963, 13.7017, 9.29589, 3.33462, 1.41191, 0.115814\}.$$

Using unity covariance matrix I instead of estimated ones, i.e., the minimum Euclidean distance classification rule (MEDC), results in $0.61 - -0.0.74$ mean classification accuracy. The true covariance matrix allows to obtain accuracy changing from 0.684425 (means are estimated from only 5 samples) to 0.82 (for $N_0 = N_1 = 200$). These results indicate that for small data samples estimation of the covariance matrix was useless.

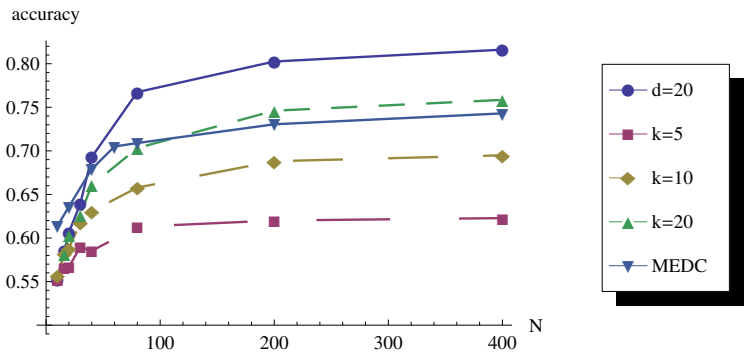

Fig. 6. Averaged classification accuracy for two 20-dimensional Gaussian classes and different number of learning samples $N = N_0 + N_1$. $N = 10,\ 16,\ 20,\ \ldots, 400$.

Figure 7 shows magnificated part of Figure 6, which contains results obtained when number of samples N was smaller than the dimension d. This figure demonstrates that both dimension reduction methods give the same mean recognition errors providing that the dimension of the random projection is larger than the number of samples.

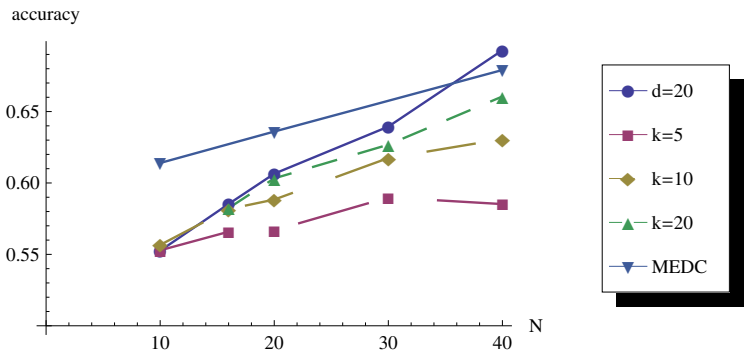

Fig. 7. Averaged classification accuracy for two 20-dimensional Gaussian classes and different number of learning samples $N = N_0 + N_1$. $N = 10,\ 16,\ 20,\ 30,\ 40$.

The similar behavior we have observed for $d = 80$ and $N_0 = N_1 = 30$. As previously the class means were $m_0 = (0, \ldots, 0)^T$ and $m_1 = (1, \ldots, 1)^T$ and eigenvalues of the common covariance matrix equal $1, 2, \ldots, 80$. The averaged classification accuracy was equal to 0.579 for Srivastava's method and equals 0.583 for random projection of dimensionality $k = 60$. Using I (unity) covariance matrix, i.e. MEDC rule, allows us to obtain better results of classification with accuracy 0.63.

5 Conclusions

The work in this paper is focused on minimum distance classification in high dimension in the small sample context.

We have presented a classification scheme based on random projections and have compared its efficiency to the recognition results obtained for the MLR type rule introduced by Srivastava [19]. Both approaches are outperformed by the minimum Euclidean distance classifier, when $N < d$. It is clear that even if the number of learning samples is larger than the dimension of the classification problems, the minimum Euclidean distance rule can perform better than the MLR (or Fisher discriminant method). It should be noted that both methods apply principal component analysis. The Srivastava method uses it for dimensionality reduction. The random projection method starts from reducing the dimension of the problem. It allows us to diminish computational costs for PCA. If small values of principal components occur, it is possible to reduce further the dimension of the data.

It is an open question how large should be the learning sample in high dimension taking into account the regular structure of data in high dimension. In other words one may ask where is the transition point between spherical and non-spherical structure of data when a covariance matrix is not unity.

Another important problem is how to test classification results in a low sample size context. It is known that in such a case the cross validation methods are very unstable [1]. We can definitely not avoid a randomness introduced by a small sample size.

References

1. Ahn, J., Marron, J.S., Müller, K.M., Chi, Y.-Y.: The High Dimension. Low-Sample Size Geometric Representation Holds Under Mild Conditions. Biometrika 94, 760–766 (2007)
2. Bickel, P.J., Levina, E.: Some Theory for Fisher's Linear Discriminant Function, Naive Bayes, and Some ASlternatives when There Are Many More Variables than Observations. Bernoulli 10, 989–1010 (2004)
3. Donoho, D.L., Tanner, J.: Neighborliness of randomly-projected simplices in high dimensions. Proc. Nat. Acad. Sci. 102, 9452–9457 (2005)
4. Fan, J., Fan, Y., Wu, Y.: High-dimensional Classification. In: Cai, T.T., Shen, X. (eds.) High-dimensional Data Analysis. Frontiers of Statistics, vol. 2, pp. 3–37. World Scientific, Singapore (2011)
5. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugenics 7, 179–188 (1936)
6. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego (1990)
7. Golub, G., Van Loan, C.F.: Matrix Computations. The Johns Hopkins University Press, Baltimore (1996)
8. Hall, P., Marron, J.S., Neeman, A.: Geometric Representation of High-Dimension Low-Sample Size Data. Journal of the Royal Statistical Society, Ser. B 67, 427–444 (2005)

9. Jung, S., Senb, A., Marron, J.S.: Boundary Behavior in High Dimension, Low Sample Size Asymptotics of PCA. Journal of Multivariate Analysis 109, 190–203 (2012)
10. Jung, S., Marron, J.S.: PCA Consistency in High Dimension, Low Sample Size Context. Ann. Statist. 37, 4104–4130 (2009)
11. Kiefer, J., Schwartz, R.: Admissible Bayes Character of $T^2 - -R^2$ and Other Fully Invariant Tests for Classical Multivariate Normal Problems. Ann. Math. Statist. 36, 747–770 (1965)
12. Mahalanobis, P.C.: On the Generalised Distance in Statistics. Proceedings of the National Institute of Sciences of India 2, 49–55 (1936)
13. Marron, J.S., Todd, M.J., Ahn, J.: Distance-Weighted Discrimination. Journal of the American Statistical Association 102, 1267–1271 (2007)
14. Marzetta, T.L., Tucci, G.H., Simon, S.H.: A Random Matrix-Theoretic Approach to Handling Singular Covariance Estimates. IEEE Transactions on Information Theory 57, 6256–6271 (2011)
15. Rao, C.R.: Linear Statistical Inference and Its Applications, 2nd edn. Wiley, New York (1973)
16. Saranadasa, H.: Asymptotic Expansion of the Misclassification Probabilities of D- and A-criteria for Discrimination from the Two High dimensional Populations Using the Theory of Large Dimensional Metrices. J. Multivariate Anal. 46, 154–174 (1993)
17. Skubalska-Rafajłowicz, E.: Clustering of Data and Nearest Neighbors Search for Pattern Recognition with Dimensionality Reduction Using Random Projections. In: Rutkowski, L., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2010, Part I. LNCS, vol. 6113, pp. 462–470. Springer, Heidelberg (2010)
18. Skubalska-Rafajowicz, E.: Random Projections and Hotelling's T^2 Statistics for Change Detection in High–dimensional Data Streams. International Journal of Applied Mathematics and Computer Science 23, 447–461 (2013)
19. Srivastava, M.S.: Minimum Distance Classification Rules for High Dimensional Data. Journal of Multivariate Analysis 97, 2057–2070 (2006)
20. Vempala, S.: The Random Projection Method. American Mathematical Society, Providence (2004)
21. Wald, A.: On the statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statist. 15, 145–162 (1944)
22. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer, New York (2004)