2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science

Random-matrix regularized discriminant analysis of high-dimensional dataset

Peng Liu, Bin Ye, Yangquan Guo, Hanyang Wang and Fei Chu
School of Information and Control Engineering
China University of Mining and Technology
Xuzhou, P.R. China
Email: yebin@cumt.edu.cn

***Abstract*—Linear discriminant analysis (LDA) is one of the most popular parametric classification methods in machine learning and data mining tasks. Although it performs well in many applications, LDA is impractical for high-dimensional data sets. A primary reason for it is that the sample covariance matrix is no longer a good estimator of the actual covariance matrix when the dimension of feature vector p is close to or even larger than the sample size n. Here we propose to regularize LDA classifier by employing a consistent estimator of high-dimensional covariance matrices. Using the theoretical tools from random matrix theory, the covariance matrices in high-dimensions are estimated in a linear or nonlinear shrinkage manner depending on the relationship between the dimension p and the sample size n. Numerical simulations demonstrate that the regularized discriminant analysis using random matrix theory yield higher accuracies than existing competitors for a wide variety of synthetic and real data sets.**

***Keywords*-linear discriminant analysis; high-dimensional data; random matrix theory; classification; covariance matrix**

I. Introduction

Linear discriminant analysis is a well-established supervised learning technique applicable in a variety of areas [1], [2]. As a model-based classifier, it aims to allocate a data point into one of the predefined classes on the basis of a number of feature variables. Compared with other classification algorithms such as random forests or support vector classifier, the model constructed in LDA is more interpretable and easy to make predictions.

In the present era of ”Big Data”, high-dimensional data sets are now generated and collected in almost all fields [3]. The most direct manifestation of high-dimensional data is that its dimension p is not fixed but becomes large together with the sample size n, which is called large n, large p asymptotics. Thus high-dimensional data will transcend the boundary of classical multivariate statistics where we implicitly assume that the dimension of feature vector p is fixed while the sample size n tends to infinity. High-dimensional data brings great challenges to statistical learning techniques, including LDA. The linear discriminant classifier becomes inefficient in high dimensional settings. One important reason is that the sample covariance matrix $\boldsymbol{S}$ in high dimensions is singular (noninvertible) or very close to being singular. It is no longer a good approximation to the population covariance matrix $\boldsymbol{\Sigma}$ under high dimensional asymptotics and leads to high misclassification error rates.

To cope with the singularity of sample covariance matrices, the procedure of ridge regression or diagonal loading is proposed in [4]. By artificially adding a positive diagonal matrix to the singular sample covariance matrix, it converts a singular sample covariance matrix into an invertible covariance. Similar modifications have been proposed by Friedman to regularize the covariance estimation in LDA, which bring forth the popular regularized discriminant analysis [5]. But how to choose the optimal regularization parameter is a long-standing research problem. Ledoit and Wolf derived an asymptotic optimal formula to estimate the regularization parameter and proposed an consistent estimator for the precision matrix, i.e., the inverse of the covariance matrix [6]. However, the method applies only to the situation that the number of features p is less than the sample size n.

Random matrix theory as a powerful theoretical framework is believed to meet the challenges of high-dimensional data, since the large p, large n settings in high-dimensional data analysis fall exactly into the realm of random matrix theory. Motivated by the recent developments in random matrix theory [7], [8], we propose to regularize the linear discriminant classifier by optimally shrinking the eigenvalues of the sample covariance matrix while keeping the eigenvectors unchanged. An extensive simulation analysis is conducted to test the performance of our algorithm in high-dimensional settings. The results show that our algorithm is more flexible and obtains lower misclassification rates for a variety of data sets.

II. Preliminaries to Linear Discriminant Analysis

In linear discriminant analysis, one or more new data points (observations) are classified into one of the predefined classes (groups) based on the observed features (variables). LDA is based on the assumption that every probability density within the kth class is following a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, i.e., the p-dimensional joint probability density function for the kth class can be modelled as:

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)\,\boldsymbol{\Sigma}_k^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}} \qquad (1)$$

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00060

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k, k = 1, 2, \ldots, K$, are the mean vector and the covariance matrix for the kth class, respectively. It is assumed further in LDA that the variables for each class share the same covariance matrix, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma},\ k = 1, \ldots, K$. For a new observed data vector $\boldsymbol{x} \in \mathbb{R}^{1\times p}$, the posterior probability $P(G = k|\boldsymbol{x})$ that $\boldsymbol{x}$ belongs to class k can be obtained by using Bayes' rule

$$P(G = k|\boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{K} f_l(\boldsymbol{x})\pi_l}, \tag{2}$$

where π_k is the prior probability of class k. The optimal classification is obtained by selecting the class k which maximize the class posteriors $P(G = k|\boldsymbol{x})$,

$$\hat{G}_1(\boldsymbol{x}) = \arg\max_k\ P(G = k|\boldsymbol{x}). \tag{3}$$

Another equivalent, yet simple, description of the decision rule in (3) is

$$\hat{G}_2(\boldsymbol{x}) = \arg\max_k\ \{\boldsymbol{x}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k^{\mathrm{T}} - \frac{1}{2}\boldsymbol{\mu}_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k^{\mathrm{T}} + \log\pi_k\}. \tag{4}$$

In practice, the mean vector $\boldsymbol{\mu}_k$, the covariance matrix $\boldsymbol{\Sigma}$ and the prior probability π_k in (4) are estimated using the training data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ which consists of n labelled observations of the p-dimensional feature vectors. In particular, the covariance matrix $\boldsymbol{\Sigma}$ can be set equal to the overall sample covariance $\boldsymbol{K} = \frac{1}{n-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}(\boldsymbol{X} - \overline{\boldsymbol{X}})$ with $\overline{\boldsymbol{X}}$ denoting the sample mean.

III. Random matrix regularized linear discriminant analysis

The sample covariance matrix $\boldsymbol{K}$ converges almost surely to $\boldsymbol{\Sigma}$ only in the case where $p << n$. In the cases where p is close to, or even larger than n, the sample covariance matrix $\boldsymbol{K}$ will become ill-conditioned or even singular. So the precision matrix $\boldsymbol{\Sigma}^{-1}$ in (4) is badly estimated and result in inefficient classifications. In this section, we propose to regularize the linear discriminant analysis by using a consistent estimator of the covariance $\boldsymbol{\Sigma}$ from random matrix theory.

A. Estimation of the covariance based on random matrix theory

By considering the number of variables p relative to the sample size n in the high-dimensional setting, two different methods from random matrix theory, namely, the rotational invariant estimation method and the eigenvalues clipping method, are employed to estimate the covariance.

1) The case of $n \geq p$: As above, we denote the $p\times p$ population covariance matrix by $\boldsymbol{\Sigma}$. And the sample covariance matrix which is obtained from the training data matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{K}$. In the case where the number of variables p is close to the sample size n, the sample covariance matrix $\boldsymbol{K}$ is ill-conditioned or near singular.

To overcome the near singularity of $\boldsymbol{K}$, the rotational invariant estimator is proposed, which can be seen as a optimal nonlinear shrinkage procedure. Before we go into the rotational invariant estimation procedure, we first shift the sample vectors in $\boldsymbol{X}$ to zero mean, to eliminate the effect of different scales. By doing so, we are actually handling the empirical correlation matrix $\boldsymbol{C}$. It has been demonstrated in [7] that $\boldsymbol{C}$ and $\boldsymbol{K}$ share identical statistically properties when $n \to \infty$, $p \to \infty$ up to a rank one perturbation. So we shall work with $\boldsymbol{K}$ henceforth.

In the rotational invariant estimator, the spectral decomposition of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \sum_{i=1}^{p} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^{\dagger} \tag{5}$$

where $\mu_i, i = 1, \ldots, p$, are the real eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{v}_i, i = 1, \ldots, p$ are the corresponding eigenvectors. Similarly, the sample covariance matrix $\boldsymbol{K}$ can be decomposed as

$$\boldsymbol{K} = \sum_{i=1}^{p} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger} \tag{6}$$

with the eigenvalues λ_i and the corresponding eigenvectors $\boldsymbol{u}_i$ of $\boldsymbol{K}$. The rotational invariant estimator is expected to find an estimator $\boldsymbol{\Xi}(\boldsymbol{K})$ of the population covariance matrix $\boldsymbol{\Sigma}$ from $\boldsymbol{K}$ in a rotationally invariant way. More formally, the estimator $\boldsymbol{\Xi}(\boldsymbol{K})$ satisfies:

$$\boldsymbol{\Omega}\ \boldsymbol{\Xi}(\boldsymbol{K})\ \boldsymbol{\Omega}^{\dagger} = \boldsymbol{\Xi}(\boldsymbol{\Omega}\,\boldsymbol{K}\,\boldsymbol{\Omega}^{\dagger}) \tag{7}$$

for any rotation matrix $\boldsymbol{\Omega}$. It has been shown that any rotational invariant estimator $\boldsymbol{\Xi}(\boldsymbol{K})$ shares the same eigenbasis as $\boldsymbol{K}$[9], that is,

$$\boldsymbol{\Xi}(\boldsymbol{K}) = \sum_{i=1}^{p} \xi_i \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger} \tag{8}$$

where the eigenvalues $[\xi_i]_{i\in[[1,p]]}$ are the quantities that the rotational invariant estimator wish to estimate.

Given the sample covariance $\boldsymbol{K}$ with the eigenvalues λ_i and the corresponding eigenvectors $\boldsymbol{u}_i, i \in \{1, \ldots, p\}$, an optimal rotational invariant estimator is

$$\hat{\boldsymbol{\Xi}}(\boldsymbol{K}) = \sum_{i=1}^{p} \hat{\xi}(\lambda_i)\ \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger}. \tag{9}$$

Here $\hat{\xi}(\lambda_i)$ can be found using the Marchenko-Pastur law in random matrix theory and it is given by the following nonlinear mapping [10]

$$\hat{\xi}(\lambda_i) = \frac{\lambda_i}{|1 - q + q\, z_i \mathfrak{g}_{\boldsymbol{K}}^{p}(z_i)|^2} \tag{10}$$

here $q = \frac{p}{n}$, $\mathfrak{g}_{\boldsymbol{K}}^{p}(z)$ is the Stieltjes transform in random matrix theory and the complex number z_i is set to be $z_i = \lambda_i - \mathrm{i}p^{-1/2}$.

Together with (6), (9) and (10), one obtains a complete procedure for the optimal rotational invariant estimator of the covariance in high dimensions. It is rather simple and works perfectly when the sample size n is larger than the number of variables p.

2) The case of $n < p$: In the case of $n < p$, the method of eigenvalues clipping is used to correct the sample eigenvalues. This method is different to the rotational invariant estimator and is an intuitive application of the Marchenko-Pastur law in random matrix theory.

Consider an $n \times p$ random matrix $\boldsymbol{R}$ whose elements come from an independent standard Gaussian distribution. The Marchenko-Pastur law describes the asymptotic behavior of the eigenvalues of the $p \times p$ Wishart matrix $\boldsymbol{W} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$ when both n and p tend to infinity [10]. For $q = \frac{p}{n} \in (0, \infty)$, the largest eigenvalue λ_+ of $\boldsymbol{W}$ converges in probability to $(1 + \sqrt{q})^2$.

In the eigenvalues clipping method, all its eigenvalues beyond the largest expected eigenvalue λ_+ are interpreted as signal while the others are noise. To infer the covariance matrix $\boldsymbol{\Sigma}$ from the sample covariance matrix $\boldsymbol{K}$, we first decompose the matrix $\boldsymbol{K}$ and keep eigenvectors unchanged. Then apply the following scheme to correct the eigenvalues

$$\Xi^{\text{clip}} = \sum_{i=1}^{p} \xi_i \, \boldsymbol{u}_i \boldsymbol{u}_i^{\dagger}, \qquad \xi_i = \begin{cases} \lambda_i, & \text{if } \lambda_i \geq (1+\sqrt{q})^2 \\ \bar{\lambda}, & \text{otherwise} \end{cases} \tag{11}$$

here $\bar{\lambda}$ is set to be a constant such that the trace of Ξ^{clip} is equal to that of $\boldsymbol{K}$ [11].

This eigenvalues clipping method for covariance estimation has also been found in a number of applications such as gas identification and immunogen design etc [12], [13].

B. Random matrix regularized discriminant analysis (RMRDA) algorithm

Combined the linear discriminant analysis with the consistent covariance estimator given above, we have the regularized linear discriminant classifier based on random matrix theory. The pseudocode for our algorithm are shown in Algorithm 1.

IV. Analysis of the synthetic data

In this section, we use the synthetic data to compare the classification performance of our proposed method with other existing methods including DLDA [1], MDMP [14] and smDLDA [15]. We will consider the simulated data generated from three multivariate normal distributions: $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma})$. And the mean value of the 1st class is set to be $\boldsymbol{\mu}_1 = \boldsymbol{0}$, while for $\boldsymbol{\mu}_2$ its first 100 values are set to $\boldsymbol{0.5}$ and the rest are $\boldsymbol{0}$. For the 3rd class, its mean value is $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$. The covariance matrix Σ is constructed as the block diagonal matrix and this covariance model has been widely used to mimic the real world data sets [16]. The $(i,j)_{th}$ entry in each block matrix is $\sigma_{ij} = \rho^{|i-j|}$. Without loss of generality, we will set $\rho = 0.1, 0.3, 0.6$ and 0.8, respectively. The number of variables is set to $p = 1000$. Each of the three classes has the same number of training samples n_k. We also generate additional 1200 samples as test dataset. The average correct classification rate (ACCR) for each algorithm is obtained by averaging over 100 runs and the standard deviation is also calculated.

Algorithm 1 Random matrix regularized discriminant analysis (RMRDA)

```
Input: The training data $X_{train}$ and the test data $X_{test}$
Output: The average correct classification rate(ACCR)
Divide the labelled samples in $X_{train}$ to $K$ groups
for $k = 1 : K$ do
    Compute $\mu_k$ and $\pi_k$ in (4)
end for
Compute the sample covariance matrix $\Sigma$ in (4)
if $n \geq p$ then
    Estimate $\Sigma$ using rotational invariant estimator in subsection III-A1
else
    Estimate $\Sigma$ using eigenvalues clipping method in subsection III-A2
end if
for each data vector $x$ in $X_{test}$ do
    for $k = 1 : K$ do
        Compute the discriminant function in (4)
    end for
    Classify $x$ into the $k$-th class according to (4)
end for
Compute the average correct classification rate
return
```

Table I
ACCR for different algorithms $(n = 1200, p = 1000)$

Methods	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
LDA	0.776	0.790	0.887	0.972
DLDA	0.979	0.972	0.934	0.824
MDMP	0.922	0.896	0.825	0.705
smDLDA	0.987	0.978	0.941	0.837
RMRDA	0.984	0.978	0.993	0.999

Table II
ACCR for different algorithms $(n = 900, p = 1000)$

Methods	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
LDA	0.361	0.293	0.312	0.328
DLDA	0.980	0.974	0.917	0.823
MDMP	0.876	0.881	0.784	0.653
smDLDA	0.985	0.981	0.933	0.849
RMRDA	0.979	0.977	0.944	0.966

The average correct classification rates for different settings are shown in Tables I and II. It can be seen from Tables I and II that our algorithm works better than most of the competitors and is only worse than smDLDA in few cases. When the correlations between the variables become stronger, our algorithm is superior to other classifiers.

Table III
SUMMARY OF TRAINING AND TESTING DATASETS

Dataset	Class	Dimension	Training set	Testing set
pix	10	240	260	600
fac	10	216	260	600

Table IV
ACCR FOR THE MFEAT DATASETS

Dataset	LDA	DLDA	MDMP	smDLDA	RMRDA
pix	0.402	0.918	0.940	0.918	0.932
fac	0.071	0.887	0.884	0.887	0.951

V. ANALYSIS OF REAL WORLD DATA SETS

The Multiple Feature(Mfeat) dataset is a multi-class dataset [17], which consists of features of handwritten numerals (from 0 to 9). The dataset includes six different feature sets of the same data, such as Fourier coefficients of the character shapes (fou), profile correlations (fac), pixel averages (pix), etc. We have chosen the fac and pix feature sets in our experiments. The sample size n and the dimension p of the training and testing datasets are summarized in Table III.

The classification results for the handwritten digit dataset are presented in Table IV. We can see that the classification accuracy of MDMP slightly exceeds that of RMRDA on pix dataset, but RMRDA still outperform other competitors. For the fac dataset, RMRDA shows the best classification performance and maintains high classification accuracy.

VI. CONCLUSION

Linear discriminant analysis is a widely used method for classification. However, it may fail when the number of the features is close to or larger than the sample size. We propose a regularized discriminant analysis method based on random matrix theory. It can handle the high-dimensional data sets, regardless of the relative magnitudes of n and p. Compared with other popular classifiers, it shows competitive and satisfying performance when evaluated on both the synthetic data sets and the real world data sets. The regularization process can further extend to quadratic discriminant analysis after very slight modification to deal with datasets in high dimensions.

REFERENCES

[1] O. C. Hamsici and A. M. Martinez, “Bayes optimality in linear discriminant analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 647–657, Apr. 2008.

[2] S. Dudoit, J. Fridlyand and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, Mar. 2002.

[3] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski and M. E. Martone, “Big data from small data: data-sharing in the ’long tail’ of neuroscience,” *Nature Neuroscience*, vol. 17, no. 11, pp. 1442–1447, May 2014.

[4] B. D. Carlson, “Covariance matrix estimation errors and diagonal loading in adaptive arrays,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 397–401, Jul. 1988.

[5] J. H. Friedman, “Regularized discriminant analysis,” *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, Mar. 1989.

[6] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” Journal of Multivariate Analysis, vol. 88, no. 2, pp. 365–441, Feb. 2004.

[7] J. Bun, J. P. Bouchaud and M. Potters, “Cleaning large correlation matrices: tools from random matrix theory,” *Physics Reports*, vol. 666, pp. 1–109, Jan. 2017.

[8] J. Bai and S. Shi, “Estimating high dimensional covariance matrices and its applications,” *Annals of Economics and Finance*, vol. 12, no. 2, pp. 199–215, Nov. 2011.

[9] J. Bun, R. Allez and J. P. Bouchaud, “Rotational invariant estimator for general noisy matrices,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7475–7490, Dec. 2016.

[10] A. Edelman and N. R. Rao, “Random matrix theory,” *Acta Numerica*, vol. 14, pp. 233–297, May 2005.

[11] L. Laloux, P. Cizeau and M. Potters, “Random matrix theory and financial correlations,” *International Journal of Theoretical and Applied Finance*, vol. 03, no. 03, pp. 391–397, Jul. 2000.

[12] M. Hassan and A. Bermak, “Robust Bayesian inference for gas identification in electronic nose applications by using random matrix theory,” *IEEE Sensors Journal*, vol. 16, no. 7, pp. 2036–2045, Dec. 2016.

[13] A. A. Quadeer, R. H. Y. Louie, K. Shekhar, A. K. Chakraborty, I. Hsing and M.R. McKay, “Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design,” *Journal of Virology*, vol. 88, no. 13, pp. 7628–7644, Jul. 2014.

[14] M. S. Srivastava and T. Kubokawa, “Comparison of discrimination methods for high dimensional data,” *Journal of the Japan Statistical Society*, vol. 37, no. 1, pp. 123–134, 2007.

[15] T. Tong, L. Chen and H. Zhao, “Improved mean estimation and its application to diagonal discriminant analysis,” *Bioinformatics*, vol. 28, no. 4, pp. 531–537, Feb. 2012.

[16] Y. Guo, T. Hastie and R. Tibshirani, “Regularized linear discriminant analysis and its application in microarrays,” *Biostatistics*, vol. 8, no. 1, pp. 86–100, Jan. 2007.

[17] D. Dua and E. Karra Taniskidou, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2017.