2013

Modern Multivariate Methods for Accurate Dialect Classification

Ernest Fokoue
Rochester Institute of Technology

Zichen Ma
Rochester Institute of Technology

Modern Multivariate Methods for Accurate Dialect Classification

Zichen Ma
Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive,
Rochester, NY 14623, USA
zxm7743@rit.edu

Ernest Fokoué
Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive,
Rochester, NY 14623, USA
ernest.fokoue@rit.edu

Technical Report - Draft 1

Abstract

We perform discriminant analysis together with principal component analysis on dialect and accent recognition. Since the data matrix exhibits high dimension low sample size feature, we calculate the principal components and the score matrix based on the dual space. Given the transformed score matrix, linear discriminant model does not fit the data well, while quadratic discriminant model, the superior model comparing to LDA, may fail sometimes when large number of principal components are required. Using the Gaussian radial basis function kernel, we calculate the kernel matrix and perform LDA directly on it. Comparing the LDA-PCA method, the in-sample prediction error rate of LDA reduces by more than 20% *on average.*

Keywords: *Dialect Recognition, Large p small n, Discriminant Analysis, Principal Component Analysis, Gram Matrix, Kernel Method, Confusion Matrix.*

I. Introduction

Dialect and accent recognition is an interesting and profound topic in both linguistics and statistics, in which the main task is to recognize if someone is a native speaker and further to predict the speaker's native language based on some samples of his/her voice. However, this is frequently considered to be difficult to perform, mainly because it involves the manipulation of high dimension low sample size data. Figure 1 and 2 provide plots of two 8-second sample, one from a native speaker and the other from a non-native speaker.

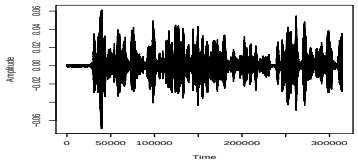

Figure 1: *Voice of a Native Speaker*

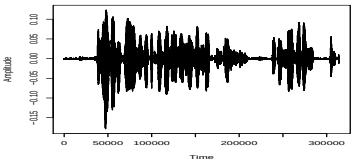

Figure 2: *Voice of a Non-native Speaker*

The methodology of this paper is to perform discriminant analysis on the sample data we collected, but we would perform the principal component analysis first for feature extraction since the dimension of the data matrix is too large to handle. Later, we would enhance the performance of discriminant analysis by taking kernel method into account. At last,we would examine some features of the eigenvoice.

II. Discriminant Analysis

Discriminant analysis is one of the standard approaches to classification problems. Let the data matrix $\mathbf{X}$, given every class k follow a Gaussian distribution

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right), \tag{1}$$

where p is the dimension and $\mathbf{\Sigma}_k$ is the covariance matrix for class k. Both vector $\mathbf{x}$ and mean vector $\boldsymbol{\mu}_k$ are column vectors.
In linear discriminant analysis (LDA), we assume that the covariance matrices in all classes are the same. That is, $\mathbf{\Sigma}_k = \mathbf{\Sigma}, \forall k$. By Bayesian theory, we have

$$\begin{aligned}\hat{Y}(x) &= \underset{k}{\text{argmax}}\, Pr(Y=k|X=x) \\ &= \underset{k}{\text{argmax}}\, f_k(\mathbf{x})\pi_k \\ &= \underset{k}{\text{argmax}}[\mathbf{x}^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + log(\pi_k)],\end{aligned}$$

where $\pi_k = Pr(Y=k)$ is the prior probability. Define the linear discriminant function as

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + log(\pi_k). \tag{2}$$

Then

$$\hat{Y}(x) = \underset{k}{\text{argmax}}\, \delta_k(\mathbf{x}). \tag{3}$$

In practice, μ_k and Σ can be estimated by the sample mean and sample covariance.
Quadratic discriminant analysis (QDA) is almost the same as LDA, except that we no longer assume that the covariance matrix is the same for all classes. Thus, we have to estimate Σ_k separately for each class k. The quadratic discriminant function is given by

$$\delta_k(\mathbf{x}) = -\frac{1}{2}log|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^\top\Sigma_k^{-1}(\mathbf{x}-\mu_k) + log(\pi_k). \tag{4}$$

In both LDA and QDA, the classification rule is to search for the class k which maximizes the discriminant function $\delta_k(\mathbf{x})$.

III. Methodology of PCA

Principal component analysis (PCA) is a well-known method to reduce dimensionality and extract features of data matrices. Mathematically, let $\mathbf{X}$ denote an $n \times p$ matrix of standardized data where n is the sample size and p is dimensionality, or the number of predictors, and $n > p$. PCA is performed by first decomposing the $p \times p$ full-rank matrix $\mathbf{X}^\top\mathbf{X}$, which indicates the total variation of the data matrix. We have

$$\mathbf{X}^\top\mathbf{X} = \mathbf{WAW}^\top, \tag{5}$$

where $\mathbf{A}$ is a $p \times p$ diagonal matrix of the eigenvalues, i.e. $\mathbf{A} = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$, and $\mathbf{W}$ is a $p \times p$ orthogonal matrix in which the columns are the corresponding eigenvectors. Therefore, the full principal component decomposition is given by

$$\mathbf{Z} = \mathbf{XW}. \tag{6}$$

By keeping only the first q components based on the magnitude of the eigenvalues, where $q \ll p$, we essentially performed dimension reduction.
However, this method may fail when $p \gg n$. In this case, the matrix $\mathbf{X}^\top\mathbf{X}$ is a $p \times p$ matrix with rank $n < p$. It is shown that we can deal with this kind of data based on the dual matrix $\mathbf{XX^T}$ rather than $\mathbf{X}^\top\mathbf{X}$ by transforming equation (1). Notice again that $\mathbf{W}$ is an orthogonal matrix, which leads to $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. We have

$$\mathbf{X}^\top\mathbf{XW} = \mathbf{WAW}^\top\mathbf{W} = \mathbf{WA}. \tag{7}$$

Then by pre-multiplying both sides of equation (3) by $\mathbf{X}$ and plugging in equation (2), we have

$$\mathbf{XX}^\top\mathbf{XW} = \mathbf{XWA} \tag{8}$$

and

$$\mathbf{XX}^\top\mathbf{Z} = \mathbf{ZA}. \tag{9}$$

$\mathbf{Z}$ is an $n \times p$ matrix with independent columns, and $\mathbf{A}$ is a $p \times p$ diagonal matrix with n non-zero eigenvalues. The zero eigenvalues contribute nothing to the performance of PCA. Thus, we can keep the non-zero part of matrix $\mathbf{A}$, which is the top-left $n \times n$ block, and truncate off the other parts. For matrix $\mathbf{Z}$, we only keep the first n columns and truncate all the following $n - p$ columns. This truncation gives

$$\mathbf{XX}^\top\mathbf{V} = \mathbf{VA_D}, \tag{10}$$

where $\mathbf{V}$ and $\mathbf{A_D}$ are the truncated matrices of $\mathbf{V}$ and $\mathbf{A}$. Also notice that columns in $\mathbf{V}$ are uncorrelated, thus $\mathbf{V}$ is invertible, which leads to

$$\mathbf{XX}^\top = \mathbf{VA_DV^{-1}}. \tag{11}$$

Equation (7) indicates that instead of dealing with matrix $\mathbf{X}^\top\mathbf{X}$, we can perform eigenvalue decomposition on the dual matrix $\mathbf{XX^T}$ and obtain the component scores directly from matrix $\mathbf{V}$.
In this project we applied this method to dialect recognition, where we classified if a person is a native English speaker by analyzing his/her voice of reading certain phrases.

IV. Data Description

A total of 117 people's voices were recorded, of which 60 were native speakers and 57 were non-native speakers. Each person was required to read 5 certain phrases. Data were read into R based on time domain. The data matrices $X_i, i = 1, 2, \ldots, 5$, contained data of all the voices for the i^{th} phrase, with rows denoting different people and columns the amplitude at certain time. Each matrix has $n = 117$ observations. Since the length of each record is unique, we simplified the dimensionality of each data matrix to

$$p = min(p_j), j = 1, 2, \ldots, 117.$$

The responses are binary, with 0 indicating non-native speakers and 1 native speakers.

V. Main Results

There are multiple methods of doing binary classification. Here we compared generalized linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). All three models were trained on the whole data set, and then the in-sample prediction error rate was calculated. Table 1 and 2 provide the fitted accuracy and error rate of each model.

Table 1: *A Summary of the Performance of LDA Model*

Phrase	Percent	N of PC's	TP+TN	FP	FN	Error Rate
1	90%	47	90	27	0	0.2308
($p = 299008$)	95%	61	94	22	1	0.1966
2	90%	47	89	27	1	0.2393
($p = 310272$)	95%	62	92	22	3	0.2137
3	90%	40	85	31	1	0.2735
($p = 270134$)	95%	55	92	25	0	0.2137
4	90%	40	85	31	1	0.2735
($p = 283648$)	95%	55	92	25	0	0.2137
5	90%	42	85	30	2	0.2735
($p = 314368$)	95%	56	94	22	1	0.1966

Table 2: *A Summary of the Performance of QDA Model*

Phrase	Percent	N of PC's	TP+TN	FP	FN	Error Rate
1	90%	47	111	6	0	0.0513
($p = 299008$)	95%	61	-	-	-	-
2	90%	47	106	11	0	0.0940
($p = 310272$)	95%	62	-	-	-	-
3	90%	40	100	17	0	0.1453
($p = 270134$)	95%	55	112	5	0	0.0427
4	90%	40	101	16	0	0.1368
($p = 283648$)	95%	55	115	2	0	0.0171
5	90%	42	102	15	0	0.1282
($p = 314368$)	95%	56	115	2	0	0.0171

For phrase 1 and 2, there were not enough data to perform QDA when 95% of the variation was kept, since Σ_k could not be estimated under that condition. Apart from these, QDA made a good performance, especially when 95% of the variation was kept in the model. Comparing to QDA, LDA had relatively high fitted error.

It is also of interest to look at this summary result from the aspect of linguistics. If we focus on LDA, for instance, most of the false predictions were false positive, which may indicate that some non-native accents are similar to the native accent, or that some non-native speakers do have native accents. Both may lead to the failure of a linear discriminant method.

We also analysed the pattern of the precision rate based on different models as number of principal components preserved is increasing. A plot was created based on Phrase 5 and was shown below.

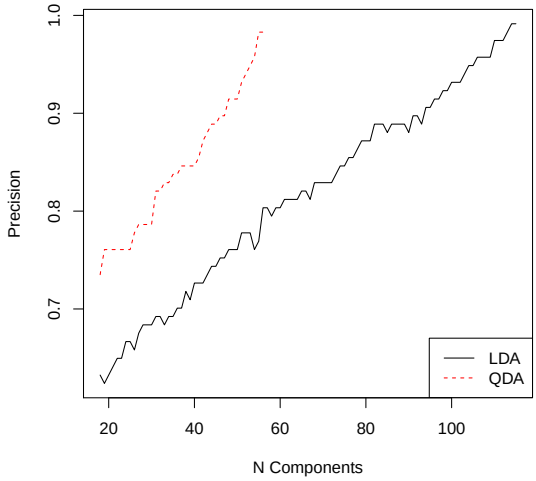

Figure 3: *Precision Rate vs. Number of Principal Components*

Based on the plot, QDA has a better performance than the other two models in general, but it cannot be performed when $N \geq 57$, which is the size of class 0. The performance of LDA increases as the number of principal components preserved gets larger, but it is not as good as QDA.

VI. Analysis with Kernel Method

An alternative analysis was performed using kernel method and was shown to have a better performance than using the original data matrix $\mathbf{X}$. Instead of working on the p dimensional space, we can create an $n \times n$ kernel matrix $\mathbf{K}$, which represents the inner product space of row vectors in $\mathbf{X}$.

Let $\mathbf{x_i}, i = 1, 2, \ldots, n$, be the row vectors that represent the feature the i^{th} person's sound track on time domain, $k(\mathbf{x_i}, \mathbf{x_j})$ be some kernel function, indicating a function of the distance between the two vectors. We have

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x_1}, \mathbf{x_1}) & k(\mathbf{x_1}, \mathbf{x_2}) & \cdots & k(\mathbf{x_1}, \mathbf{x_n}) \\ k(\mathbf{x_2}, \mathbf{x_1}) & k(\mathbf{x_2}, \mathbf{x_2}) & \cdots & k(\mathbf{x_2}, \mathbf{x_n}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x_n}, \mathbf{x_1}) & k(\mathbf{x_n}, \mathbf{x_2}) & \cdots & k(\mathbf{x_n}, \mathbf{x_n}) \end{pmatrix} \tag{12}$$

$\mathbf{K}$ is an $n \times n$ symmetric matrix and there exists some function k which makes $\mathbf{K}$ semi-definite. Some popular kernel functions include the Gaussian radial basis function (RBF) kernel and the hyperbolic tangent kernel. The RBF kernel, which was used in our project, was given as

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{||\mathbf{x_i} - \mathbf{x_j}||_2^2}{2\sigma^2}\right). \tag{13}$$

$||\mathbf{x_i} - \mathbf{x_j}||_2^2$ is the squared Euclidean distance between two feature vectors and σ is a free parameter. After the calculation of $\mathbf{K}$, we may perform classification, like GLM or LDA, directly on the kernel matrix.

We applied this method to phrase 5 the voice data, using Gaussian RBF as the kernel function (with $\sigma = 0.05$).Then LDA was performed based on the kernel matrix. The summary results of the in-sample error rate are shown below.

Table 3: *Summary of Classification with Kernel Method*

Phrase	TP+TN	FP	FN	Error Rate
1	115	1	1	0.0171
2	114	1	2	0.0256
3	111	4	2	0.0513
4	115	1	1	0.0171
5	116	1	0	0.0085

LDA, which was considered as the inferior method to QDA, yields little mistake with kernel matrix.

VII. Analysis on Eigenvoice

Beyond the pattern recognition based on discriminant analysis, we also analysed the eigenvoice and it turned out that the eigenvoices had some interesting features. Recall that in equation (7) we established the PCA based on the dual space $\mathbf{X}\mathbf{X}^\top$. An eigenvoice matrix was calculated based on the matrix $\mathbf{V}$:

$$\mathbf{EV}_{n\times p} = \mathbf{V}_{n\times n}\mathbf{X}_{n\times p}. \tag{14}$$

The matrix $\mathbf{EV}$ is an $n \times p$ matrix in which each row vector indicates a weighted linear combination of row vectors in the matrix $\mathbf{X}$. In other words, it is an average of different voices and every row vector in the eigenvoice matrix sounds like a crowded and noisy market. Moreover, each eigenvoice has a single dominating voice inside, but this feature of single dominating voice damps down gradually from the first row to the n^{th} row. One can tell that there might be a dominating person with significant Arabian or Indian accent in the first eigenvoice, but the last eigenvoice is completely noise without any dominating voice. Figure 4 and 5 shows the plots of the first and the last eigenvoice.

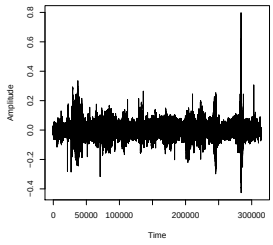

Figure 4: *Eigenvoice No.1*

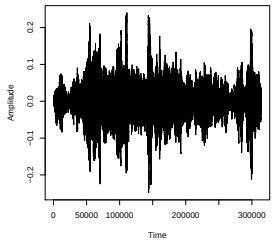

Figure 5: *Eigenvoice No.117*

Although it is hard to describe voice or sound based on graph, we can still see some feature here. That is, the amplitude in Figure 3 seems to be more stable than in Figure 2. We would expect that a plot of the complete noise exhibit the feature of stability rather than having extreme values.

VIII. Conclusion

We have shown that discriminant analysis together is a good approach to dialect and accent recognition on the time domain, but both LDA and QDA had some problems dealing with principal components. That is, LDA did not perform good while QDA may totally fail when the number of principal components preserved exceeds $min(N_k)$. Fortunately, we remedied these problems by introducing the kernel matrix. The performance of LDA improved tremendously based on the kernel method. Some future tasks include the multi-class classification rather than simply binary and perform classification based on the frequency domain by transforming the data.

References

Altun, Y., T. Hofmann, and A. Smola (2004). Gaussian process classifcation for segmenting and annotating sequences. In *Proceedings of the 21st International Conference on Machine Learning*.

Ansari, Z. and A. F. (2012). Implementing KPCA-based speaker adaptation methods with different optimization algorithms in a persian asr system. *Procedia - Social and Behavioral Sciences* (32), 117–127.

Herbig, T., F. Gerl, and W. Minker (2012). Self-learning speaker identification for enhanced speech recognition. *Computer Speech and Language* (26), 210–227.

Levow, G. (2005, 9). Context in multi-lingual tone and pitch accent recognition. *Interspeech*, 4–8.

Levow, G. (2006, 6). Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 224–231. Association for Computational Linguistics.

Li, C., J. Liu, and S. Xia (2007). English sentence stress detection system based on hmm framework. *Applied Mathematics and Computation* (185), 759–768.

Shih, P., P. Lin, J. Wang, and Y. Lin (2011). Robust several-speaker speech recognition with highly dependable online speaker adaptation and identification. *Journal of Network and Computer Applications* (34), 1459–1467.

Vieru, B., P. Mareüi, and M. Adda-Decker (2011). Characterisation and identification of non-native french accents. *Speech Communication* (53), 292–310.

Yata, K. and M. Aoshima (2010, 2). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*.