HETEROSCEDASTIC DISCRIMINANT ANALYSIS WITH TWO-DIMENSIONAL CONSTRAINTS

Si-Bao Chen[1]*, Yu Hu*[1]*, Bin Luo*[2]*, Ren-Hua Wang*[1]

[1]Iflytek Speech Lab, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China
[2]School of Computer Science and Technology, Anhui University, Hefei 230039, China

ABSTRACT

Heteroscedastic discriminant analysis (HDA) with two - dimensional (2D) constraints is proposed in this paper. HDA suffers from the small sample size problem and instability when lack of training data or feature dimension is high, even when the number of dimension is in a suitable range. Two-dimensional HDA is first proposed, then we show that 2D methods are actually a kind of structure-constrained 1D methods, and lastly, HDA with 2D constraints is proposed. Experiments on TIMIT and WSJ0 show that the proposed method outperforms other methods.

Index Terms— linear transformation, 2DHDA, 2DLDA, HDA, dimensionality reduction

1. INTRODUCTION

Heteroscedastic discriminant analysis [1], which finds a linear transformation by maximizing individual-weighted Fisher-kind ratio, achieves great success in dimensionality reduction and other pattern recognition areas. In speech recognition, usually several successive frame features are concatenated, forming long span vectors, as inputs of HDA. The more features concatenated, the more coefficients in HDA transformation matrix should be estimated. When lack of training data or the number of dimension of data is too big, HDA suffers from the small sample size problem and instability of final recognition performance. We think some constraint on the structure of the transformation matrix, to reduce its degree of freedom, may solve this problem.

Recently, two-dimensional dimensionality reduction methods, such as 2DPCA [2], 2DLDA [3], 2DLPP [4], 2DLDA +PCA[5] and 2DADA [6], appeared in the literature and performed well in image recognition area. Now 2DLDA has been tested for speech recognition [7]. They make use of matrix form of features, greatly reduce the computational complexity and can solve the small sample size problem.

Inspired by these 2D methods, we propose two- dimensional HDA (2DHDA). 2D methods are actually a kind of structure-constrained 1D methods. We try to use this property to solve the small sample size problem and instability of HDA, and propose 2D-constrained HDA.

The rest of this paper is organized as follows. In Section 2, the HDA and 2DLDA are reviewed, and then 2DHDA is proposed. In Section 3, we analyse that 2D methods are a kind of structure-constrained 1D methods, and HDA with 2D constraints is presented. Experiment results are shown in Section 4 and Section 5 is the conclusions.

2. METHODS

2.1. Heteroscedastic Discriminant Analysis

HDA is an extension to LDA by considering the individual weighted contributions of the classes to the objective function. It removes the equal within-class covariance constraint of LDA, and achieves more discriminant information.

Consider a set of N independently sampled column feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{x}_i \in \mathbf{R}^p$, each of which belongs to one and only one class $j \in \{1, ..., c\}$ through the surjective mapping of indices $l : \{1, ..., N\} \to \{1, ..., c\}$. Assume class j has N_j samples, $\sum_{j=1}^{c} N_j = N$. The mean $\overline{\mathbf{x}}_j$ and covariance $\mathbf{S}_j$ of class j is defined as:

$$\overline{\mathbf{x}}_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} \mathbf{x}_i, \quad \mathbf{S}_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} (\mathbf{x}_i - \overline{\mathbf{x}}_j)(\mathbf{x}_i - \overline{\mathbf{x}}_j)^\top.$$

Within-class scatter matrix $\mathbf{S}_\omega$ and between-class scatter $\mathbf{S}_b$ matrix are defined as:

$$\mathbf{S}_\omega = \frac{1}{N} \sum_{j=1}^{c} N_j \mathbf{S}_j, \quad \mathbf{S}_b = \frac{1}{N} \sum_{j=1}^{c} N_j (\overline{\mathbf{x}}_j - \overline{\mathbf{x}})(\overline{\mathbf{x}}_j - \overline{\mathbf{x}})^\top,$$

where $\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$.

The goal of HDA is to find a linear transformation $f : \mathbf{R}^p \to \mathbf{R}^q$, $\mathbf{y} = f(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$, with $\mathbf{W}$ a $(p \times q)$ matrix of rank q $(q \leq p)$, such that the following objective function is maximized:

$$\prod_{j=1}^{c} \left(\frac{|\mathbf{W}^\top \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^\top \mathbf{S}_j \mathbf{W}|} \right)^{N_j} = \frac{|\mathbf{W}^\top \mathbf{S}_b \mathbf{W}|^N}{\prod_{j=1}^{c} |\mathbf{W}^\top \mathbf{S}_j \mathbf{W}|^{N_j}} \qquad (1)$$

By taking log, we get:

Thanks to the China National 863 Program (No. 2004AA114030) and National Natural Science Foundation of China (No.60772122) for funding.

$$H(\mathbf{W}) = N \log |\mathbf{W}^\top \mathbf{S}_b \mathbf{W}| - \sum_{j=1}^{c} N_j \log |\mathbf{W}^\top \mathbf{S}_j \mathbf{W}|. \quad (2)$$

When diagonal variance modeling constraints are present in the final feature space, the objective function to be maximized is:

$$G(\mathbf{W}) = N \log |\mathbf{W}^\top \mathbf{S}_b \mathbf{W}| - \sum_{j=1}^{c} N_j \log |diag(\mathbf{W}^\top \mathbf{S}_j \mathbf{W})|. \quad (3)$$

There are no analytical solutions for the above two optimization problems, and numerical optimization routines (such as: steepest descent, quasi-Newton) are adopted.

HDA solution is invariant to full rank linear transformations of the data in the original space, and subsequent feature space full rank transformations will not affect the value of the objective function [1].

When doing experiments of speech recognition with HDA, usually several successive column frame features $\mathbf{o}_t$ are concatenated together, forming long span vectors $\mathbf{x}_t = (\mathbf{o}_{t-k}^\top, ..., \mathbf{o}_{t-1}^\top, \mathbf{o}_t^\top, \mathbf{o}_{t+1}^\top, ..., \mathbf{o}_{t+k}^\top)^\top$. If the dimension of feature $\mathbf{o}_t$ is m, then the dimension of spanned vector $\mathbf{x}_t$ is $p = (2k+1)m$. It should estimate transformation matrix $\mathbf{W}$ of size $((2k+1)m \times q)$. The more features concatenated, the more coefficients in $\mathbf{W}$ should be estimated. Experiments shown that it will degrade the final recognition performance when more features are concatenated for HDA, even when k is in a suitable range. We think it may be too more coefficients needed to be estimated, which may lead to small sample size problem and instability as the training data unchanged. Some constraint on the structure of transformation matrix $\mathbf{W}$ may solve the problem.

2.2. Two-Dimensional Linear Discriminant Analysis

2DLDA [3] makes use of the matrix structure of features and achieved great success in image recognition area.

Now let's consider features taking the form of matrix. Consider a set of N feature matrices $\{\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_N\}$ taken from an $(m \times n)$-dimensional feature matrix space. Assume that each feature belongs to one and only one class $j \in \{1, ..., c\}$ through the surjective mapping of indices $l : \{1, ..., N\} \to \{1, ..., c\}$. Class j has N_j samples, $\sum_{j=1}^{c} N_j = N$.

Define *between-class image scatter matrix* to be:

$$\mathbf{G}_{\mathbf{A}b} = \frac{1}{N} \sum_{j=1}^{c} N_j (\overline{\mathbf{A}}_j - \overline{\mathbf{A}})^\top (\overline{\mathbf{A}}_j - \overline{\mathbf{A}}), \quad (4)$$

and *within-class image scatter matrix* to be:

$$\mathbf{G}_{\mathbf{A}\omega} = \frac{1}{N} \sum_{j=1}^{c} \sum_{i \in l^{-1}(j)} (\mathbf{A}_i - \overline{\mathbf{A}}_j)^\top (\mathbf{A}_i - \overline{\mathbf{A}}_j), \quad (5)$$

then 2DLDA is to find a transformation matrix $\mathbf{V}$ of size $(n \times d)$ which maximizes the following objective function:

$$\arg\max_{\mathbf{V}} \frac{|\mathbf{V}^\top \mathbf{G}_{\mathbf{A}b} \mathbf{V}|}{|\mathbf{V}^\top \mathbf{G}_{\mathbf{A}\omega} \mathbf{V}|}, \quad (6)$$

which is just a generalized eigenvalue problem, and the maximum is obtained by the generalized eigenvectors between $\mathbf{G}_{\mathbf{A}b}$ and $\mathbf{G}_{\mathbf{A}\omega}$ corresponding to the first d largest generalized eigenvalues.

2DLDA makes use of matrix structure of features, and greatly reduces the computational complexity. As it will be shown later, 2DLDA actually is a kind of structure-constrained traditional 1D linear transformation, which may be useful for the problem of transformation matrix structure encountered in performing HDA described above.

2.3. Two-Dimensional Heteroscedastic Discriminant Analysis

Before we add some constraint on the transformation matrix of HDA, let's first extend HDA to two-dimensional HDA (2DHDA).

Still consider a set of N feature matrices $\{\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_N\}$ taken from an $(m \times n)$-dimensional feature matrix space as described above. In image recognition area, $\mathbf{A}_i$ can just be image matrix. In speech recognition area, $\mathbf{A}_i$ at time t can be spliced by several successive column frame features $\mathbf{o}_t$ in row direction: $\mathbf{A}_i = [\mathbf{o}_{t-k}, ..., \mathbf{o}_{t-1}, \mathbf{o}_t, \mathbf{o}_{t+1}, ..., \mathbf{o}_{t+k}]$. Each $\mathbf{A}_i$ belongs to one and only one class $j \in \{1, ..., c\}$ through the surjective mapping of indices $l : \{1, ..., N\} \to \{1, ..., c\}$. Assume class j has N_j samples, $\sum_{j=1}^{c} N_j = N$.

The between-class image scatter $\mathbf{G}_{Ab}$ is defined as in (4) and image scatter $\mathbf{G}_j$ of class j is defined as:

$$\mathbf{G}_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} (\mathbf{A}_i - \overline{\mathbf{A}}_j)^\top (\mathbf{A}_i - \overline{\mathbf{A}}_j). \quad (7)$$

The goal of 2DHDA is to find a linear transformation $g : \mathbf{R}^{(m \times n)} \to \mathbf{R}^{(m \times d)}$, $\mathbf{B} = g(\mathbf{A}) = \mathbf{AV}$, with $\mathbf{V}$ a $(n \times d)$ matrix of rank d $(d \leq n)$, such that the following objective function is maximized:

$$\prod_{j=1}^{c} \left(\frac{|\mathbf{V}^\top \mathbf{G}_{\mathbf{A}b} \mathbf{V}|}{|\mathbf{V}^\top \mathbf{G}_j \mathbf{V}|} \right)^{N_j} = \frac{|\mathbf{V}^\top \mathbf{G}_{\mathbf{A}b} \mathbf{V}|^N}{\prod_{j=1}^{c} |\mathbf{V}^\top \mathbf{G}_j \mathbf{V}|^{N_j}} \quad (8)$$

By taking log, and we get:

$$H(\mathbf{V}) = N \log |\mathbf{V}^\top \mathbf{G}_{\mathbf{A}b} \mathbf{V}| - \sum_{j=1}^{c} N_j \log |\mathbf{V}^\top \mathbf{G}_j \mathbf{V}|. \quad (9)$$

There is no analytical solution for this optimization problem. We adopt quasi-Newton numerical optimization routine to solve the optimal transformation matrix $\mathbf{V}$.

Like HDA, 2DHDA solution is invariant to full rank linear transformations of the data in the original space, and subsequent feature space full rank transformations will not affect the value of the objective function.

3. CONSTRAINED HDA

3.1. 2D methods: constrained 1D methods

Suppose there is a two-dimensional (2D) transformation matrix $\mathbf{V}$ of size $(n \times d)$, which maps matrix-structured feature $\mathbf{A}$ of size $(m \times n)$ to a smaller matrix-structured feature $\mathbf{B}$

by $\mathbf{B} = \mathbf{AV}$. Then the size of $\mathbf{B}$ is $(m \times d)$. The vectorized representation of this transformation is:

$$\begin{aligned} \mathbf{z} &\hat{=} Vec(\mathbf{B}) = Vec(\mathbf{AV}) = (\mathbf{V}^\top \otimes \mathbf{I}_m) Vec(\mathbf{A}) \\ &\hat{=} (\mathbf{V}^\top \otimes \mathbf{I}_m)\mathbf{x} = (\mathbf{V} \otimes \mathbf{I}_m)^\top \mathbf{x} \\ &\hat{=} \mathbf{W}_2^\top \mathbf{x}, \end{aligned} \tag{10}$$

where $Vec(\cdot)$ is vectorization operator of matrix, which splices all the columns of the matrix one after another to form a long column vector. Operator $\otimes$ is the Kronecker product of matrices. $\mathbf{I}_m$ is the identity matrix of order m.

From equation (10), we can see that 2D linear transformation methods are actually kinds of traditional one-dimensional (1D) methods with some structural constraint on transformation matrix $\mathbf{W}_2$, which should be in the form of the Kronecker product between a smaller transformation matrix and an identity matrix: $(\mathbf{V} \otimes \mathbf{I}_m)$.

This constraint on the structure of transformation matrix greatly reduces the degree of freedom of transformation matrix. The traditional 1D methods should estimate matrix $\mathbf{W}_2$ of size $(nm \times dm)$, while 2D methods only need estimate matrix $\mathbf{V}$ of size $(n \times d)$. This property contributes to the stability and computational simplicity of 2D methods, especially when lack of training data.

3.2. HDA with 2D constraints

Inspired by the property that 2D methods are actually constrained 1D methods, now we try to add constraints on the transformation matrix of HDA.

Suppose each data feature can be easily represented in matrix $\mathbf{A}_i$ form and in column vector $\mathbf{x}_i$ form. In image recognition area, $\mathbf{A}_i$ can just be image matrix and let $\mathbf{x}_i = Vec(\mathbf{A}_i)$; In speech recognition area, let speech feature $\mathbf{o}_t$ at time t be column vector, then $\mathbf{A}_i$ at time t can be $\mathbf{A}_i = [\mathbf{o}_{t-k}, ..., \mathbf{o}_{t-1}, \mathbf{o}_t, \mathbf{o}_{t+1}, ..., \mathbf{o}_{t+k}]$, and $\mathbf{x}_i = Vec(\mathbf{A}_i)$. Suppose the size of $\mathbf{A}_i$ is $(m \times n)$, then the size of $\mathbf{x}_i$ is $(mn \times 1)$. Now we try to perform HDA and reduce the dimension to q $(q < mn)$.

Traditional HDA can be performed directly on features in $\mathbf{x}_i$ form to estimate transformation matrix $\mathbf{W}$ of size $(mn \times q)$, and do dimensionality reduction by $\mathbf{y} = \mathbf{W}^\top \mathbf{x}$. However, when we did experiments of speech recognition with HDA, experiments shown that it will degrade the final recognition performance when more feature are concatenated, even when frame range $2k + 1$ is in a suitable range. What's more, when k becomes larger, instability of HDA begins to increase.

Now we do 2D methods (such as: 2DLDA or 2DHDA) on features in $\mathbf{A}_i$ form. Suppose we've gotten the 2D transformation matrix $\mathbf{V}$ of size $(n \times d)$. Transfer it into 1D form and we get transformation matrix $\mathbf{W}_2 = (\mathbf{V} \otimes \mathbf{I}_m)$ for features in $\mathbf{x}_i$ form. The size of $\mathbf{W}_2$ is $(nm \times dm)$, and after projection $\mathbf{z} = \mathbf{W}_2^\top \mathbf{x}$, we get new features $\{\mathbf{z}_i\}$ of size $(dm \times 1)$.

Then we perform another HDA on the new features $\{\mathbf{z}_i\}$. Suppose we've gotten the 1D transformation matrix $\mathbf{W}_1$ of size $(dm \times q)$, which maps $\mathbf{z}_i$ to final feature $\mathbf{y}_i$ by $\mathbf{y}_i = \mathbf{W}_1^\top \mathbf{z}_i$. $\mathbf{y}_i$ is the required q-dimensional column vector.

In fact, this combined method is a kind of structure- constrained 1D method. As we can see,

$$\begin{aligned} \mathbf{y}_i &= \mathbf{W}_1^\top \mathbf{z}_i = \mathbf{W}_1^\top \mathbf{W}_2^\top \mathbf{x}_i = \mathbf{W}_1^\top (\mathbf{V} \otimes \mathbf{I}_m)^\top \mathbf{x}_i \\ &= [(\mathbf{V} \otimes \mathbf{I}_m)\mathbf{W}_1]^\top \mathbf{x}_i. \end{aligned} \tag{11}$$

The total structure-specific transformation matrix is $[(\mathbf{V} \otimes \mathbf{I}_m)\ \mathbf{W}_1]$ of size $(mn \times q)$, while its degree of freedom is only $nd + dmq$ $(d < n)$. This constraint on the structure of transformation matrix may alleviate the small sample size problem and instability encountered by HDA.

4. EXPERIMENTS

To evaluate the performance of constrained HDA and other methods, we test them on two databases: TIMIT [8] and WSJ0 [9]. When performing feature transformations, we chose 39 -dimensional MFCCs (12 static MFCCs, log energy, and their first- and second-order time derivatives) as inputs. All methods also chose 39 as the final dimension of features for comparison. For 1D methods HLDA and HDA, usually three to five successive frames are concatenated for input. For 2D methods 2DLDA+HLDA, 2DLDA +HDA and 2DHDA+HDA, usually five to seven successive frames are spliced, and the mid-reduced column dimension d in 2D methods was set to be three.

When performing 2DLDA and 2DHDA, we first compressed features of size $(39 \times (2k + 1))$ to (39×1) size. But this kind of transformation fails to consistently improve the final recognition performance. This may be due to the fact that this time each dimensions share only one transformation. Therefore, we adopted 13-dimensional static MFCCs as inputs for 2DLDA and 2DHDA, compressed features of size $(13 \times (2k + 1))$ to (13×3) size, and concatenated them to form 39-dimensional final features.

4.1. Tests on TIMIT

We first applied these methods for continuous phoneme recognition on the TIMIT database. The details of the database can be found in [8]. The standard training set (3,696 utterances) and coreTest set (192 utterances) were used. 48 phones were used to create context dependent triphone models. Eight Gauss components per tied state were trained in the final models. The features in the baseline (maximum likelihood, ML) are the conventional 39-dimensional MFCCs.

When counting the results, only 39 effective phones were counted as described in [8]. Table 1 gives the top recognition performance in phone error rate (PER) for different methods. Relative reduction (R.R.) is computed based on baseline (ML). "Dimension" in the table lists the dimension changes of different methods when achieving the top recognition accuracy. From the table we can see that HDA with 2D constraints achieved great improvement than other methods.

Table 1. Top Recognition performances of different methods on TIMIT database

methods	ML	HDA	2DLDA	2DHDA	2DLDA+HLDA	2DLDA+HDA	2DHDA+HDA
Dimension	39	3*39→39	13x13→13x3	13x25→13x3	39x5→39	39x7→39	39x7→39
PER(%)	37.52	31.05	34.47	33.50	32.81	30.31	29.98
R.R.(%)	-	17.24	8.13	10.71	12.55	19.22	20.10

Table 2. Top Recognition performances of different methods on WSJ0 database

methods	ML	HDA	2DLDA	2DHDA	2DLDA+HLDA	2DLDA+HDA	2DHDA+HDA
Dimension	39	3*39→39	13x7→13x3	13x7→13x3	39x5→39	39x7→39	39x5→39
WER(%)	4.89	4.43	5.04	4.48	4.63	4.43	4.24
R.R.(%)	-	9.41	-3.07	8.38	5.32	9.41	13.29

4.2. Tests on WSJ0

Next we carried out constrained HDA and other methods on the Wall Street Journal speech corpus WSJ0 [9]. We used the standard SI-84 training set for training the acoustic models. All methods were tested on the standard nov92 5K nonverbalized test set using a trigram language model. Eight to sixteen Gauss components per tied state were trained in the final models. The features in the baseline (ML) are the conventional 39-dimensional MFCCs processed by CMN.

Table 2 gives the top recognition performance of word error rate (WER) for different methods. Again, HDA with 2DHDA constraint outperformed other methods.

4.3. Discussions

We adopted 39-dimensional MFCCs (incorporating derivatives) rather than 13-dimensional static MFCCs for our experiments. The reason is based on an initial contrast test: training LDA using long span vectors concatenated by three successive 39-dimensional MFCCs, or using long span vectors concatenated by nine successive 13-dimensional static MFCCs. The size of LDA transformation matrices are both (117×39). The computational complexities of training procedures and the final model complexities are both the same. However, the former concatenating method outperforms the latter in the recognition test with final models. This can be explained from the first principle of statistics. In statistics, all observations are assumed to be sampled independently. Concatenating methods obviously disobey this rule. The former concatenating method shares 2/3 of total elements between neighbor long span vectors and the latter shares 8/9. The former seems more closer to the rule. Due to coarticulation of speech, consecutive frames are dependent. We still can make full use of information existed in neighbor frames. Maybe other concatenating methods should be investigated.

5. CONCLUSIONS

To solve the small sample size problem and instability of HDA when concatenated feature dimension is high, we proposed 2D-constrained HDA. 2DLDA were extended to 2DHDA. We also showed that 2D methods are actually a kind of structure-constrained 1D methods, and finally, HDA with 2D constraints is described. Experiments on TIMIT and WSJ0 show that the proposed method outperformed other methods in recognition performance when model complexities are the same.

Acknowledgement The authors thank anonymous reviewers for their helpful comments, which greatly improve the quality of our paper.

6. REFERENCES

[1] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *ICASSP*. IEEE, 2000, vol. II, pp. 1129–1132.

[2] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional PCA: a new approach to appearance–based face representation and recognition," *IEEE Trans. Pattern Anal. & Mach. Intell.*, vol. 26, no. 1, pp. 131–137, 2004.

[3] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527–532, 2005.

[4] S. B. Chen, H. F. Zhao, M. Kong, and B. Luo, "2D-LPP: A two-dimensional extension of locality preserving projections," *Neurocomputing*, vol. 70, no. 4-6, pp. 912–921, 2007.

[5] P. Sanguansat, W. Asdornwised, S. Jitapunkul, and S. Marukatat, "Two-dimensional linear discriminant analysis of principle component vectors for face recognition," in *ICASSP*. IEEE, 2006, vol. II, pp. 345–348.

[6] Y. J. Lu, J. Yu, N. Sebe, and Q. Tian, "Two-dimensional adaptive discriminant analysis," in *ICASSP*. IEEE, 2007, vol. I, pp. 985–988.

[7] X. B. Li and D. O. Shaughnessy, "Clustering-based two-dimensional linear discriminant analysis for speech recognition," in *INTERSPEECH*, 2007, pp. 1126–1129.

[8] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, November 1989.

[9] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *DARPA Speech and Natural Language Workshop*, San Mateo, CA, 1992, Morgan Kaufmann Publishers.