On Projections of Gaussian Distributions using Maximum Likelihood Criteria

Haolang Zhou*, Damianos Karakos*, Sanjeev Khudanpur*, Andreas G. Andreou* and Carey E. Priebe†
*Department of Electrical and Computer Engineering
and Center for Language and Speech Processing
† Department of Applied Mathematics and Statistics
Johns Hopkins University, Baltimore, MD, USA
{haolangzhou,damianos,khudanpur,andreou,cep}@jhu.edu

***Abstract*— Generative statistical models with a very large number of parameters are frequently used in real-world data applications, such as large-vocabulary speech recognition (LVCSR). Complex models are needed in order to capture the ubiquitous variability in the observed signal, but data sparsity causes significant problems in their training. One way of dealing with data sparsity is to perform dimensionality reduction of the observed features, with the goal of reducing the model parameter space without sacrificing performance. When the data are Gaussian distributed, the dimensionality reduction can be done efficiently using the maximum likelihood criterion; this leads to the Heteroscedastic Linear Discriminant Analysis (HLDA), which is a natural extension of Linear Discriminant Analysis (LDA) to the case where the class-conditional Gaussians have unequal covariance matrices. A further extension of HLDA to multiple transforms (MLDA) can also be tackled efficiently. This paper presents the theory behind HLDA and MLDA, and demonstrates their performance with synthetic data.**

I. Generative Models in Speech Recognition

Speech recognition is a complex classification task. The observed signal vector $\mathbf{A}$, which represents the *acoustics*, is the result of a cascade of signal processing operations, and is parameterized in a way that preserves information about the words uttered, while being invariant to certain kinds of irrelevant variations (e.g., microphone). Modeling the observations is usually done in a generative way, which assumes that $\mathbf{A}$ is the realization of a random process, whose parameters depend on the true "labels" (e.g., word identities) of the observations. Inference is usually done in a maximum-aposteriori fashion,

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{A}|\mathbf{W})P(\mathbf{W}),$$

where $P(\mathbf{A}|\mathbf{W})$ is the generative model of the acoustics given the word sequence $\mathbf{W}$ (acoustic model), and $P(\mathbf{W})$ is the language model [1].

A number of reasonable independence assumptions allow to factor $P(\mathbf{A}|\mathbf{W})$ into a number of components, each of which corresponds to a conditional distribution that models some aspect of the speech production. The (usually very large) collection of conditional distributions is assumed to belong to a parameterized model family, and training acoustic models amounts to estimating the parameters of these models. Mixtures of Gaussian distributions with diagonal covariance matrices are very frequently used as the underlying models, because of their expressiveness and efficiency in their training. (See [1], [2] for details about how acoustic models are trained.) Thus, expressing the acoustic vector $\mathbf{A}$ as a sequence of observations $\{\boldsymbol{a}_i\}$, each $\boldsymbol{a}_i$ is generated (given label w) by an underlying process

$$p(\boldsymbol{a}|w) = p_w(\boldsymbol{a}) = \sum_{m=1}^{M} \lambda_m \phi(\boldsymbol{a}; \boldsymbol{\mu}_m(w), \Sigma_m(w))$$

where M is the number of Gaussian components in the mixture, and $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M\}$, $\{\Sigma_1, \ldots, \Sigma_M\}$ are the means and covariance matrices of these components. These can be estimated with the EM algorithm [3], with the objective of maximizing the likelihood of the training data.

Estimation of covariance matrices suffers from high variance and is computationally intensive when the dimensionality of the Gaussian vectors is large (e.g., of the order of thousands). For this reason, projection into a space of lower dimensionality is performed before training complex acoustic models with many mixture components. The high dimensionality arises in speech recognition from concatenating together many features, e.g., PLP/MFCC features, PLP/MFCC features from neighboring speech frames, articulatory features, etc. Since many of these features are highly correlated, the lower-dimensional projection can be used to *decorrelate* them as well, keeping only those features which carry information for discrimination between the classes and discarding the rest. This paper is mainly a presentation of the mathematics behind two popular methods used for projecting data: Linear Discriminant Analysis (LDA) [4], [5] and Heteroscedastic Linear Discriminant Analysis (HLDA) [6] as well as its variant, Multiple LDA (MLDA) [7]. HLDA uses maximization of likelihood of the non-projected data as the criterion for estimating the transform, and it has been used with success in speech recognition. In fact, as is shown in [8] and reviewed here, LDA can *also* be derived as a maximum-likelihood solution, under a constraint of equal covariance matrices.

A number of experiments with synthetic Gaussian data demonstrate the performance of the above schemes, under a variety of conditions (amount of training data, amount of

This work was supported by National Science Foundation grant No CCF-0728931.

"overlap" of the classes). It is observed that HLDA always outperforms LDA when the class-conditional distributions have unequal covariance matrices, and MLDA always outperforms HLDA.

II. Mathematical Preliminaries

It is assumed that a training corpus exists, consisting of N observations (column vectors) $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, each belonging to $\mathbb{R}^n$. There are C labels (classes) $\{1, \ldots, C\}$, and the label of observation $\boldsymbol{x}_j$ is denoted by c_j. The number of observations of class c is N_c, and hence $N_1 + \cdots + N_C = N$. The sample mean of class c is denoted by $\boldsymbol{\mu}_c$, while the sample covariance of class c (the "within-class" covariance) is denoted by Σ_c. Specifically,

$$\boldsymbol{\mu}_c \triangleq \frac{1}{N_c} \sum_{j:c_j=c} \boldsymbol{x}_j, \quad \Sigma_c \triangleq \frac{1}{N_c} \sum_{j:c_j=c} (\boldsymbol{x}_j - \boldsymbol{\mu}_c)(\boldsymbol{x}_j - \boldsymbol{\mu}_c)^{\mathrm{T}},$$

where a^{T} is the transpose of matrix (or vector) a. The global mean and variance are denoted by $\boldsymbol{\mu}$ and Σ, respectively. The k-dimensional Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ is denoted by $\phi_{(k)}(\cdot; \boldsymbol{\mu}, \Sigma)$, or just $\phi(\cdot; \boldsymbol{\mu}, \Sigma)$ when the dimensionality is clear from the context.

Projecting a vector $\boldsymbol{x}$ into $\mathbb{R}^p$ is done by multiplying it with a $p \times n$ matrix Θ ($p < n$). Thus,

$$\boldsymbol{y}_j = \Theta \boldsymbol{x}_j, \quad j = 1, \ldots, N$$

is the projection of the j-th observation.

III. Linear Projections for Classification

Two projection methods are especially popular in speech recognition: (i) Linear Discriminant Analysis (LDA), and (ii) Heteroscedastic Linear Discriminant Analysis (HLDA) [6]. In addition, a natural extension of HLDA is presented in [7] as (iii) Multiple LDA. These techniques are reviewed in the rest of this section.

A. Classical LDA

For the 2-class problem, Linear Discriminant Analysis was introduced by Fisher [4] and Rao [5] as a method for finding the "most discriminant" projection direction which maximizes the ratio between the average between-class squared Euclidean distance and the average within-class squared Euclidean distance. That is, the goal is to estimate an $1 \times n$ matrix $\Theta^{(1)}_{\text{LDA}}$, which represents the projection direction, such that the ratio

$$J(\Theta^{(1)}_{\text{LDA}}) \triangleq \frac{\sum_c N_c \left(\Theta^{(1)}_{\text{LDA}} (\boldsymbol{\mu}_c - \boldsymbol{\mu}) \right)^2}{\sum_c \left(\sum_{j:c_j=c} \left(\Theta^{(1)}_{\text{LDA}} (\boldsymbol{x}_j - \boldsymbol{\mu}_c) \right)^2 \right)}$$

is maximized. More generally, the matrix $\Theta^{(p)}_{\text{LDA}}$ that gives the projection to the p most discriminant directions can be determined by first projecting the data to the $p-1$ most discriminant directions, and then finding the next most discriminant direction of the difference between the original and the projected data. Then, the LDA objective function becomes

$$J(\Theta) = \frac{|\Theta \mathrm{B} \Theta^{\mathrm{T}}|}{|\Theta \mathrm{W} \Theta^{\mathrm{T}}|}, \tag{1}$$

where

$$\mathrm{B} \triangleq \sum_c \frac{N_c}{N} (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} \quad \text{and} \quad \mathrm{W} \triangleq \sum_c \frac{N_c}{N} \Sigma_c$$

are the average "between-class" and "within-class" covariance matrices, respectively. The solution of the maximization of (1) is a matrix Θ_{LDA} which is computed by post-multiplying $\mathrm{W}^{-1/2}$ with a matrix, whose rows are the eigenvectors corresponding to the p largest eigenvalues of

$$\mathrm{M} = \mathrm{W}^{-1/2} \mathrm{B} \mathrm{W}^{-1/2},$$

provided that W is non-singular. (If it is singular, a lower-dimensional subspace can be identified using Principal Components Analysis.)

B. Maximum-Likelihood Projection of Gaussian Data: Heteroscedastic Linear Discriminant Analysis

Projection of the data in $\mathbb{R}^p$ can be viewed as the process of (i) first transforming the data into $\mathbb{R}^n$, and (ii) keeping only p dimensions in the transformed space. The transformation plays the role of making the classes as separable as possible through p dimensions only, allowing the "safe" removal of the remaining $n - p$ dimensions; under a Gaussian class-conditional distribution assumption, this amounts to giving the same class-conditional mean and covariance matrices to these $n - p$ dimensions. Moreover, the projection is computed so that the likelihood of the original data is as high as possible.

A summary of the maximum-likelihood approach appears below.

- Each class-conditional distribution in the original space is assumed to be Gaussian.
- The data are transformed as $\boldsymbol{y}_j = \Theta \boldsymbol{x}_j, \quad j = 1, \ldots, N$, where Θ is an $n \times n$ invertible matrix.
- The class-conditional distributions in the transformed space are Gaussians with parameters

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_c &= (\widetilde{\boldsymbol{\mu}}_c^{(p)}, \widetilde{\boldsymbol{\mu}}^{(n-p)}) \triangleq (\widetilde{\mu}_{c,1}, \ldots, \widetilde{\mu}_{c,p}, \widetilde{\mu}_{p+1}, \ldots, \widetilde{\mu}_n)^{\mathrm{T}}, \\ \widetilde{\Sigma}_c &= \begin{pmatrix} \widetilde{\Sigma}_c^{(p)} & \mathbf{0} \\ \mathbf{0} & \widetilde{\Sigma}^{(n-p)} \end{pmatrix}. \end{aligned}$$

  Note that only the first p dimensions are useful in discriminating between the classes.
- The objective in the estimation of Θ is the maximization of the log-likelihood of the *original* data:

$$L(\{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}) = \sum_c \sum_{j:c_j=c} \log \phi(\boldsymbol{x}_j; \boldsymbol{\mu}_c, \Sigma_c).$$

  Conditioned on class c, the relationship between the pdfs of $\boldsymbol{x}$ and $\boldsymbol{y} = \Theta \boldsymbol{x}$ can be easily established

$$\begin{aligned} \phi(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c) &= |\Theta| \, \phi(\Theta \boldsymbol{x}; \widetilde{\boldsymbol{\mu}}_c, \widetilde{\Sigma}_c) \\ &= |\Theta| (\phi_{(p)}(\Theta^{(p)} \boldsymbol{x}; \widetilde{\mu}_c^{(p)}, \widetilde{\Sigma}_c^{(p)}) \times \\ &\quad \phi_{(n-p)}(\Theta^{(n-p)} \boldsymbol{x}; \widetilde{\mu}^{(n-p)}, \widetilde{\Sigma}^{(n-p)})) \end{aligned}$$

where

$$\Theta = \begin{pmatrix} \Theta^{(p)} \\ \Theta^{(n-p)} \end{pmatrix}.$$

Thus, the log-likelihood of the training data is given by

$$\begin{aligned} & N \log|\Theta| - \frac{N}{2}\log(2\pi)^n - \sum_c [\frac{N_c}{2}\log|\widetilde{\Sigma}_c^{(p)}| \\ & + \frac{1}{2}\sum_{j:c_j=c} (\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})^{\mathrm{T}} (\widetilde{\Sigma}_c^{(p)})^{-1} (\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})] \\ & - \frac{N}{2}\log|\widetilde{\Sigma}^{(n-p)}| \\ & - \frac{1}{2}\sum_{j=1}^{N} (\Theta^{(n-p)}\boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})^{\mathrm{T}} (\widetilde{\Sigma}^{(n-p)})^{-1} (\Theta^{(n-p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(n-p)}) \end{aligned}$$

which is maximized when

$$\widetilde{\boldsymbol{\mu}}_c^{(p)} = \frac{1}{N_c}\sum_{j:c_j=c} \Theta^{(p)}\boldsymbol{x}_j = \Theta^{(p)}\boldsymbol{\mu}_c \quad (2)$$

$$\begin{aligned} \widetilde{\Sigma}_c^{(p)} &= \frac{1}{N_c}\sum_{j:c_j=c} (\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})(\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})^{\mathrm{T}} \\ &= \Theta^{(p)}\Sigma_c(\Theta^{(p)})^{\mathrm{T}} \end{aligned} \quad (3)$$

$$\widetilde{\boldsymbol{\mu}}^{(n-p)} = \frac{1}{N}\sum_{j=1}^{N} \Theta^{(n-p)}\boldsymbol{x}_j = \Theta^{(n-p)}\boldsymbol{\mu} \quad (4)$$

$$\begin{aligned} \widetilde{\Sigma}^{(n-p)} &= \\ & \frac{1}{N}\sum_{j=1}^{N} (\Theta^{(n-p)}\boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})(\Theta^{(n-p)}\boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})^{\mathrm{T}} \\ &= \Theta^{(n-p)}\Sigma(\Theta^{(n-p)})^{\mathrm{T}}, \end{aligned} \quad (5)$$

where $\boldsymbol{\mu}, \Sigma$ are the global mean and covariance of the data, respectively. Substituting these values in the expression for the log-likelihood of the training data, it becomes

$$\begin{aligned} L^*(\{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}) &= N\log|\Theta| - \frac{N}{2}\log(2\pi)^n \\ & - \sum_c \frac{N_c}{2}\log|\Theta^{(p)}\Sigma_c(\Theta^{(p)})^{\mathrm{T}}| \\ & - \sum_c \frac{pN_c}{2} - \frac{N}{2}\log|\Theta^{(n-p)}\Sigma(\Theta^{(n-p)})| - \frac{(n-p)N}{2} \\ = \; & N\log|\Theta| - \sum_c \frac{N_c}{2}\log|\Theta^{(p)}\Sigma_c(\Theta^{(p)})^{\mathrm{T}}| \\ & - \frac{N}{2}\log|\Theta^{(n-p)}\Sigma(\Theta^{(n-p)})| - \frac{nN}{2}\log(2\pi e) \end{aligned} \quad (6)$$

Expression (6) is the objective function of Heteroscedastic Linear Discriminant Analysis (HLDA), introduced by Kumar and Andreou [6]. The maximizing Θ cannot be given in closed form, and a steepest-descent algorithm is needed for its computation. However, as Gales points out in [9], in the special case where the projected per-class Gaussians are constrained to have diagonal covariance matrices, there is a very efficient algorithm for the maximization of (6). Furthermore, as is shown in the next subsection, when the per-class Gaussian distributions are constrained to have *equal* covariance matrices, maximization of (6) is equivalent to maximization of (1).

C. Interpretation of LDA as Maximum-Likelihood Projection under a Constraint of Equal Per-class Covariances

Under a constraint of equal per-class covariances, $\widetilde{\Sigma}_c^{(p)}$ is constant (equal, say, to $\widetilde{\Sigma}_1^{(p)}$). The log-likelihood of the training data then becomes

$$\begin{aligned} & N\log|\Theta| - \frac{N}{2}\log(2\pi)^n - \frac{N}{2}\log|\widetilde{\Sigma}_1^{(p)}| \\ & \quad - \sum_c \frac{1}{2}\sum_{j:c_j=c} (\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})^{\mathrm{T}} (\widetilde{\Sigma}_1^{(p)})^{-1} (\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)}) \\ & - \frac{N}{2}\log|\widetilde{\Sigma}^{(n-p)}| \\ & - \frac{1}{2}\sum_{j=1}^{N} (\Theta^{(n-p)}\boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})^{\mathrm{T}} (\widetilde{\Sigma}^{(n-p)})^{-1} (\Theta^{(n-p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(n-p)}) \end{aligned}$$

which is maximized by the same expressions (2), (4) and (5), but with (3) replaced by

$$\begin{aligned} \widetilde{\Sigma}_1^{(p)} &= \frac{1}{N}\sum_c \sum_{j:c_j=c} (\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})(\Theta^{(p)}\boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})^{\mathrm{T}} \\ &= \Theta^{(p)}\mathrm{W}(\Theta^{(p)})^{\mathrm{T}} \end{aligned} \quad (7)$$

instead of (3), where W is the average "within-class" covariance, defined earlier. Then, the log-likelihood becomes

$$\begin{aligned} L^*(\{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}) &= N\log|\Theta| - \frac{nN}{2}\log(2\pi e) \\ & - \frac{N}{2}\log|\Theta^{(p)}\mathrm{W}(\Theta^{(p)})^{\mathrm{T}}| - \frac{N}{2}\log|\Theta^{(n-p)}\Sigma(\Theta^{(n-p)}) \end{aligned} \quad (8)$$

Differentiating (8) with respect to Θ (e.g., using formulas found in [10] for computing derivatives with respect to matrices) and setting the (matrix) result to zero, it turns out that the Θ which maximizes (8) satisfies the conditions

$$\Theta^{(p)}\mathrm{W}(\Theta^{(n-p)})^{\mathrm{T}} = \mathbf{0} \;\text{ and }\; \Theta^{(n-p)}\Sigma(\Theta^{(p)})^{\mathrm{T}} = \mathbf{0} \quad (9)$$

Assuming that W is non-singular, and setting $\Theta = \Psi\mathrm{W}^{-1/2}$, it can be proved that conditions (9) are equivalent to the following two conditions

$$\Psi^{(p)} \text{ and } \Psi^{(n-p)} \text{ are orthogonal} \quad (10)$$

$$\Psi^{(n-p)}\mathrm{W}^{-1/2}\Sigma\mathrm{W}^{-1/2}(\Psi^{(p)})^{\mathrm{T}} = \mathbf{0}, \quad (11)$$

which are simultaneously satisfied when the rows of Ψ consist of the orthogonal eigenvectors of $\mathrm{W}^{-1/2}\Sigma\mathrm{W}^{-1/2}$ (or, equivalently, the orthogonal eigenvectors of $\mathrm{W}^{-1/2}\mathrm{B}\mathrm{W}^{-1/2}$, by virtue of the fact that $\Sigma = \mathrm{W} + \mathrm{B}$). Substituting this value of Ψ into (8), the log-likelihood becomes

$$-\frac{N}{2}\log(2\pi e)^n|\mathrm{W}| - \frac{N}{2}\sum_{i=p+1}^{n}\log(1+\nu_i) \quad (12)$$

where ν_i is the i-th eigenvalue of $\mathrm{W}^{-1/2}\mathrm{B}\mathrm{W}^{-1/2}$. Thus, to *maximize* the likelihood, it suffices to choose as the p rows of Ψ the eigenvectors corresponding to the maximum eigenvalues of $\mathrm{W}^{-1/2}\mathrm{B}\mathrm{W}^{-1/2}$, and then multiply by $\mathrm{W}^{-1/2}$ to obtain Θ. It is now obvious that the maximum likelihood estimate of $\Theta^{(p)}$ is equal to the LDA solution Θ_{LDA} given earlier.

D. Multiple LDA

A natural extension to the HLDA projection method is to generate multiple transforms instead of a single global transform. This section describes one way of doing that, called Multiple LDA (MLDA) [7].

To motivate the need for such an extension, consider the case where some of the class-conditional Gaussians have means which are arbitrarily *far* from each other, while some "confusable" classes are sufficiently *close* to each other. In the simpler case where the covariance matrices are all the same, it is easy to construct an example such that the closed-form solution of LDA yields a projection to a lower-dimensional space which *does not offer any discriminability* between the confusable classes; the "between" matrix is just dominated by the statistics of the well-separated classes. On the other hand, having multiple transforms, each computed from a group of classes, can mitigate this problem.

In MLDA, a class grouping has to be specified first: C classes are divided into S groups ($S \leq C$), with each class being assigned a group label $s \in \{1, ..., S\}$. Next, the objective is to estimate a transformation Θ_s for each group s in such a way that all the discrimination information is kept in the first p dimensions. Using the notation introduced earlier, we have

$$\Theta_s = \begin{pmatrix} \Theta_s^{(p)} \\ \Theta^{(n-p)} \end{pmatrix},$$

where the last $n-p$ dimensions are transformed independently of the class grouping.

As before, the projections are computed so that the likelihood of the original Gaussian data is as high as possible. The maximum-likelihood approach can be summarized as follows:

- Each class-conditional distribution in the original space is assumed to be Gaussian.
- The data are transformed through $\boldsymbol{y}_j = \Theta_s \boldsymbol{x}_j$, $j = 1, \ldots, N$, where Θ_s is a $n \times n$ invertible matrix and $s = s(c_j)$.
- As with HLDA, the class-conditional distributions in the transformed space are Gaussians with parameters $\widetilde{\boldsymbol{\mu}}_c$ and $\widetilde{\Sigma}_c$, which are dependent on the class c only through the first p components.
- The objective in the estimation of Θ_s is the maximization of the log-likelihood of the original data:

$$L(\{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}) = \sum_s \sum_{c:s(c)=s} \sum_{j:c_j=c} \log \phi(\boldsymbol{x}_j; \boldsymbol{\mu}_c, \Sigma_c). \quad (13)$$

Conditioning on class c, the relationship between the pdfs of the original data $\boldsymbol{x}$ and the transformed data $\boldsymbol{y} = \Theta_s \boldsymbol{x}$, $s = s(c)$, can be established as

$$\begin{aligned} \phi(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c) &= |\Theta_s| \, \phi(\Theta_s \boldsymbol{x}; \widetilde{\boldsymbol{\mu}}_c, \widetilde{\Sigma}_c) \\ &= |\Theta_s| (\phi_{(p)}(\Theta_s^{(p)} \boldsymbol{x}; \widetilde{\mu}_c^{(p)}, \widetilde{\Sigma}_c^{(p)}) \times \\ &\quad \phi_{(n-p)}(\Theta_s^{(n-p)} \boldsymbol{x}; \widetilde{\mu}^{(n-p)}, \widetilde{\Sigma}^{(n-p)})). \end{aligned}$$

Thus in the case of multiple transforms, the log-likelihood of the training data is given by

$$\begin{aligned} &-\frac{N}{2} \log(2\pi)^n + \sum_c [N_c \log |\Theta_{s(c)}| - \frac{N_c}{2} \log |\widetilde{\Sigma}_c^{(p)}| \\ &-\frac{1}{2} \sum_{j:c_j=c} (\Theta_{s(c)}^{(p)} \boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})^{\mathrm{T}} (\widetilde{\Sigma}_c^{(p)})^{-1} (\Theta_{s(c)}^{(p)} \boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})] \\ &-\frac{N}{2} \log |\widetilde{\Sigma}^{(n-p)}| \\ &-\frac{1}{2} \sum_{j=1}^{N} (\Theta^{(n-p)} \boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})^{\mathrm{T}} (\widetilde{\Sigma}^{(n-p)})^{-1} (\Theta^{(n-p)} \boldsymbol{x}_j - \widetilde{\mu}_c^{(n-p)}) \end{aligned}$$

Note that $\Theta^{(n-p)}$ does not have a subscript s, meaning that it is restricted to be the same for all s.

The log-likelihood of the training data is maximized when

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_c^{(p)} &= \frac{1}{N_c} \sum_{j:c_j=c} \Theta_{s(c)}^{(p)} \boldsymbol{x}_j = \Theta_{s(c)}^{(p)} \boldsymbol{\mu}_c \\ \widetilde{\Sigma}_c^{(p)} &= \frac{1}{N_c} \sum_{j:c_j=c} (\Theta_{s(c)}^{(p)} \boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})(\Theta_{s(c)}^{(p)} \boldsymbol{x}_j - \widetilde{\mu}_c^{(p)})^{\mathrm{T}} \\ &= \Theta_{s(c)}^{(p)} \Sigma_c (\Theta_{s(c)}^{(p)})^{\mathrm{T}} \\ \widetilde{\boldsymbol{\mu}}^{(n-p)} &= \frac{1}{N} \sum_{j=1}^{N} \Theta^{(n-p)} \boldsymbol{x}_j = \Theta^{(n-p)} \boldsymbol{\mu} \\ \widetilde{\Sigma}^{(n-p)} &= \\ &\quad \frac{1}{N} \sum_{j=1}^{N} (\Theta^{(n-p)} \boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})(\Theta^{(n-p)} \boldsymbol{x}_j - \widetilde{\mu}^{(n-p)})^{\mathrm{T}} \\ &\quad = \Theta^{(n-p)} \Sigma (\Theta^{(n-p)})^{\mathrm{T}}, \end{aligned}$$

where $\boldsymbol{\mu}, \Sigma$ are the global mean and covariance of the data, respectively. Substituting these values in the expression for the log-likelihood of the training data gives

$$\begin{aligned} &L^*(\{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}) \\ &= \sum_c N_c \log |\Theta_{s(c)}| - \frac{N}{2} \log(2\pi)^n \\ &\quad - \sum_c \frac{N_c}{2} \log |\Theta_{s(c)}^{(p)} \Sigma_c (\Theta_{s(c)}^{(p)})^{\mathrm{T}}| \\ &\quad - \sum_c \frac{pN_c}{2} - \frac{N}{2} \log |\Theta^{(n-p)} \Sigma (\Theta^{(n-p)})| - \frac{(n-p)N}{2} \\ &= \sum_c N_c \log |\Theta_{s(c)}| - \sum_c \frac{N_c}{2} \log |\Theta_{s(c)}^{(p)} \Sigma_c (\Theta_{s(c)}^{(p)})^{\mathrm{T}}| \\ &\quad - \frac{N}{2} \log |\Theta^{(n-p)} \Sigma (\Theta^{(n-p)})| - \frac{nN}{2} \log(2\pi e) \quad (14) \end{aligned}$$

Thus expression (14) is the objective function for Multiple LDA. Comparing with the objective function of HLDA (6), the difference lies in having multiple transforms for the p dimension of the transformed data.

Once again, the maximizing Θ_s cannot be given in closed form. Even in the special case where the projected per-class Gaussians are constrained to have diagonal covariance matrices, the simplification used in [9] cannot be used

to estimate $\Theta^{(n-p)}$; instead a Newton-based optimization scheme can be used.

IV. Experimental Results

Here we aim to test the projection schemes described above under various conditions. For each specified condition, 100 data sets of 15 dimensional full covariance Gaussian data are generated for 5 classes, with each data set containing 1000 training samples and 2000 testing samples for each class.

For each data set a projection is trained using LDA, HLDA or MLDA to project the original 15 dimensional data into 3 dimensional space, and then the resulting lower dimensional test data are classified by the statistics obtained from the corresponding training data. In the case of MLDA when there is more than one possible grouping of the classes, we chose the grouping that gives the best performance on the training data. The average error rate of the 100 data sets are then reported for each condition.

The first set of experiments is designed to compare the performance of LDA, HLDA and MLDA under different degrees of class "overlap" in the original 15 dimensional space, which is reflected by the Bayes error and approximated by the classification error rate in the original 15 dimensional space. Five conditions are designated with the degree of overlap ranging from almost complete "overlap" to well separated, with condition 1 corresponding to the most confusable dataset. The error rate for each condition and projection scheme are presented in Table I.

	Average Error Rate (%)				
	Condition				
	1	2	3	4	5
Approximate Bayes Error	72.71	55.48	36.96	19.57	1.66
LDA projection	79.85	76.53	72.80	69.86	16.15
HLDA projection	79.48	71.39	64.25	54.69	16.03
Best MLDA projection					
$S = 2$ groups	79.33	69.01	59.12	45.48	10.77
$S = 3$ groups	79.16	67.36	55.66	39.57	8.22
$S = 4$ groups	79.10	66.04	52.88	35.24	6.99
$S = 5$ groups	79.03	64.99	50.83	32.27	6.40
$S \in \{1, \ldots, 5\}$ groups	79.06	65.04	50.86	32.27	6.41

TABLE I

Comparison under different "overlap" conditions.

As seen in the table, MLDA always outperforms HLDA, which always outperforms LDA. Also note that MLDA gives a more significant relative improvement over HLDA when the original data are well separated (condition 5). But even at the data-set level, MLDA is superior: Figure 1 shows that MLDA results in a lower error rate for each one of the 100 experiments (condition 2).

Another set of experiments is constructed to investigate how the size of training data affects the performance of the projection schemes. The results are shown in Table II with each column corresponding to the percentage of the 1000 samples per class used for training: while MLDA still performs best, its performance is the most affected by the lack of training data, while LDA is the least affected.

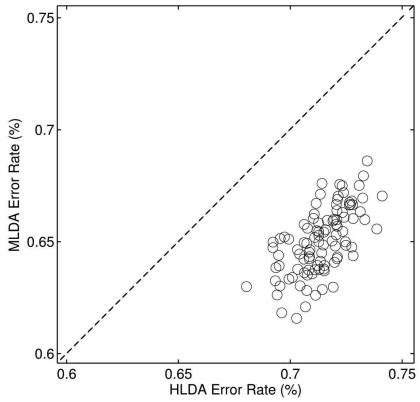

Fig. 1. HLDA and MLDA Error Rates for all data sets of Condition 2.

	Average Error Rate (%)		
	Condition 2 Training Data Size		
	20%	50%	100%
Approximate Bayes Error	55.48	55.48	55.48
LDA projection	77.73	77.01	76.53
HLDA projection	73.41	71.83	71.39
Best MLDA projection			
$S = 2$ groups	71.90	69.87	69.01
$S = 3$ groups	70.73	68.43	67.36
$S = 4$ groups	69.60	67.26	66.04
$S = 5$ groups	68.84	66.21	64.99
$S \in \{1, \ldots, 5\}$ groups	68.96	66.38	65.04

TABLE II

Comparison with different amounts of training data.

V. Concluding Remarks

From our experiments we observed that under the various conditions we constructed, MLDA always gives the best performance. Higher dimensional projections were also examined, and the same trend holds. At the same time, how much MLDA can improve over HLDA and LDA is determined by the characteristics of the original data.

References

[1] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1997.

[2] M. Gales and S. Young, *The Application of Hidden Markov Models in Speech Recognition*, NOW Publishers, 2008.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[5] C. R. Rao, *Linear statistical inference and its applications*, Wiley, New York, 1965.

[6] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[7] M. Gales, "Maximum likelihood multiple subspace projections for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 37–47, February 2002.

[8] N. Campbell, "Canonical variate analysis – a general formulation," *Australian Journal of Statistics*, vol. 26, pp. 86–96, 1984.

[9] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, May 1999.

[10] S. R. Searle, *Matrix Algebra useful for Statistics*, Wiley, New York, 1982.