

CONSTRAINED MAXIMUM LIKELIHOOD MODELING WITH GAUSSIAN DISTRIBUTIONS

R. A. Gopinath

IBM T. J. Watson Research Center, Yorktown Heights, NY.,
email: rameshg@watson.ibm.com, phone: (914)-945-2794

ABSTRACT

Maximum Likelihood (ML) modeling of multiclass data using gaussian distributions for classification often suffers from the following problems: a) data insufficiency implying over-trained or unreliable models b) large storage requirement c) large computational requirement and/or d) ML is not discriminating between classes. Sharing parameters across classes (or constraining the parameters) clearly tends to alleviate the first three problems. In this paper we show that in some cases it can also lead to better discrimination (as evidenced by reduced misclassification error). The parameters considered are the means and variances of the gaussians and linear transformations of the feature space (or equivalently the gaussian means). Some forms of sharing (either explicit or implicit via constraints) on the parameters are shown to lead to Linear Discrimination Analysis (a well-known result) while others (like diagonal, block-diagonal and factor analyzed covariances) are shown to lead to optimal feature spaces. The key idea is that in constrained ML modeling one may be able to better model the data after it is linearly transformed, perhaps in a class dependent fashion. If the constraints are invariant to linear transformations (ILT), then, the original feature space is as good as any to model the data. Results using optimal feature spaces for diagonal covariances is shown using the speech recognition problem as an example.

1. INTRODUCTION

Modeling data using Gaussian or Gaussian mixture distributions is very common in many applications. This popularity stems partially from the fact that any distribution can be approximated by gaussian mixtures and partially from the fact that a rich set of mathematical results and computational techniques are available for using gaussian distributions.

This paper considers modeling data using gaussians for classification applications. The basic problem is the following: Given *labeled* training data how does one model it “well” for classification applications. An implicit assumption here is that the *training* data and the *test* data have the same underlying statistical distributions. With this assumption, it is reasonable to try and model the training data as well as possible. The Maximum Likelihood (ML) Principle is the criterion of choice in this paper. Some dissimilarities between the training data and test data can be accounted for by parametrically adapting the trained models. In this case, the ML principle is invoked on the test data: adaptation parameters are chosen to maximize the likelihood of the test data; clearly, an example of constrained ML modeling with gaussians.

The focus of this paper is parametric modeling of training or test data with gaussian distributions using the ML principle. If the data is modeled with gaussian mixtures, then each data sample can probabilistically be assigned to the gaussians and a similar analysis as below can be carried out. Using the EM algorithm these assignment probabilities can be iteratively refined [6].

The main idea emphasized in this paper is that in constrained ML modeling (eg., diagonal covariances, factor-analyzed covariances [17, 18]) there are optimal feature spaces in which to model the classes. This author was first exposed to this idea in the context of diagonal covariances in [1] where the author was trying to generalize linear discriminant analysis. Independently, the same idea in the context of diagonal covariances is also explored in a slightly less general form as “semi-tied” covariances [2, 3], where the author tries to model data with covariances of the form AD_jA^T where A is shared between gaussians and D_j is a diagonal gaussian-dependent term. This paper generalizes this notion to factor-analyzed covariances and more by introducing the notion of covariance and mean structures that are invariant to linear transformations.

The *training data* is a collection of N independent labeled vectors (x_i, l_i) , $x_i \in \mathbb{R}^d$, $l_i \in \{1, 2, \dots, J\}$ and $i \in \{1, 2, \dots, N\}$. Each class $j \in \{1, 2, \dots, J\}$ has N_j samples and is modeled by a Gaussian distribution with mean μ_j and covariance Σ_j . The likelihood of the data is given by

$$\begin{aligned} p(x_1^N, \{\mu_j\}, \{\Sigma_j\}) &= \prod_{i=1}^N p(x_i, \{\mu_j\}, \{\Sigma_j\}) \\ &= \prod_{i=1}^N \frac{e^{-\frac{1}{2}(x_i - \mu_{l_i})^T \Sigma_{l_i}^{-1} (x_i - \mu_{l_i})}}{\sqrt{(2\pi)^d |\Sigma_{l_i}|}}. \end{aligned} \quad (1)$$

In ML modeling the idea is to choose the parameters $\{\mu_j\}$ and $\{\Sigma_j\}$ so as to maximize $p(x_1^N, \{\mu_j\}, \{\Sigma_j\})$. For later use it is convenient to organize classes into K *class clusters* with the cluster identity $c_j \in \{1, 2, \dots, K\}$. By collecting together terms for each class in Eqn. 1, $p(x_1^N, \{\mu_j\}, \{\Sigma_j\})$ can be easily expressed in the following well-known fashion:

$$a(N, d) e^{-\frac{1}{2} \left[\sum_j N_j \left\{ (\bar{\mu}_j - \mu_j)^T \bar{\Sigma}_j^{-1} (\bar{\mu}_j - \mu_j) + \text{Tr}(\bar{\Sigma}_j^{-1} \bar{\Sigma}_j) + \log |\bar{\Sigma}_j| \right\} \right]}, \quad (2)$$

where $\bar{\mu}_j$ and $\bar{\Sigma}_j$ are the sample means and covariances respectively of the classes and $a(N, d) = (2\pi)^{-\frac{Nd}{2}}$.

Now consider linearly transforming the samples from each class: $y_i = A_j x_i$, where A_j is a non-singular $d \times d$ matrix. This gives a new dataset (y_i, l_i) which can also be modeled with gaussians. However, it is difficult to compare the likelihood of a test data sample coming from the classes when the classes are modeled in the transformed space. The problem is one of scaling: one can always choose A_j such that the likelihood of data from class j is arbitrarily large. Two obvious approaches to compare likelihoods suggest themselves. One is to ensure that $|A_j| = 1$ for every class, in which case the likelihood of the data corresponding to each class is the same in the original and transformed spaces (implying $p(x_1^N) = p(y_1^N)$). The second is to only consider the likelihood in the original

space (i.e., $p(x_1^N)$) even though the data is modeled in the transformed space. In this case it is easy to show that

$$p(x_1^N, \{\mu_j\}_x, \{\Sigma_j\}_x) = p(y_1^N, \{\mu_j\}_y, \{\Sigma_j\}_y) \prod_{j=1}^J |A_j|^{N_j},$$

which again shows that ensuring $|A_j| = 1$ ensures that the likelihoods are the same. Is there any advantage in modeling y_1^N rather than x_1^N ? If the data is modeled using full-covariance gaussians, then, it makes no difference. However, if one constrains the variances to be structured (block-diagonal or diagonal, for example), then, the transformations can be used to find the basis in which this structural constraint on the variances is "more valid" as evidenced from the data.

2. SINGLE CLASS

Consider ignoring the class labels and modeling the entire data with one gaussian: (μ, Σ) (with one class there is no longer a classification problem; however, the discussion, should bring out the key ingredients in the multi-class problem). Then from Eqn. 2, $p_{one}(x_1^N, \mu, \Sigma)$ can be expressed as

$$a(N, d) e^{-\frac{1}{2}[(N\bar{\mu} - \mu)^T \Sigma^{-1}(\bar{\mu} - \mu) + Tr(\Sigma^{-1}\bar{\Sigma}) + \log|\Sigma|]}, \quad (3)$$

where $\bar{\mu}$ and $\bar{\Sigma}$ are the global mean and covariance of the data. Clearly, $p_{one}(x_1^N, \mu, \Sigma)$ is maximized by the ML estimates $\hat{\mu} = \bar{\mu}$ and $\hat{\Sigma} = \bar{\Sigma}$, whence the ML value of the training data is

$$p_{one}^*(x_1^N) = p_{one}(x_1^N, \bar{\mu}, \bar{\Sigma}) = g(N, d) |\bar{\Sigma}|^{-\frac{N}{2}}, \quad (4)$$

where $g(N, d) = (2\pi e)^{-\frac{Nd}{2}}$. On average each sample contributes $\bar{\Sigma}^{-\frac{1}{2}}$ to the ML value $p_{one}^*(x_1^N)$, which, depends only on the training data.

2.1. Linear Transformations of the Data

Consider a global non-singular linear transformation of the data: $y_i = Ax_i$. If $(\bar{\mu}, \bar{\Sigma})$ and $(\bar{\mu}_y, \bar{\Sigma}_y)$ denote the sample mean and covariance respectively (abuse of notation!!) in the two spaces, then, $\bar{\mu}_y = A\bar{\mu}$ and $\bar{\Sigma}_y = A\bar{\Sigma}A^T$. The maximum likelihood values in the two spaces are related as expected:

$$p_{one}^*(y_1^N) = g(N, d) |A\bar{\Sigma}A^T|^{-\frac{N}{2}} = |A|^{-N} p_{one}^*(x_1^N). \quad (5)$$

If $|A| = 1$ then $p^*(y_1^N) = p^*(x_1^N)$. Essentially, the ML value is invariant to *unimodular* or *volume-preserving* linear transformations of the data.

2.2. Constrained ML - Diagonal Covariance

If we are constrained to use a diagonal covariance model, Eqn. 3 is maximized by the estimates $\hat{\mu} = \bar{\mu}$ and $\hat{\Sigma} = \text{diag}(\bar{\Sigma})$. The ML value is given by

$$p_{diag}^*(x_1^N) = p(x_1^N, \bar{\mu}, \text{diag}(\bar{\Sigma})) = g(N, d) |\text{diag}(\bar{\Sigma})|^{-\frac{N}{2}}.$$

Because of the diagonal constraint on the covariances, $p_{diag}^*(x_1^N) \leq p^*(x_1^N)$, which interestingly gives a proof of Hadamard's inequality for symmetric non-negative definite matrices: $|\text{diag}(\bar{\Sigma})| \geq |\bar{\Sigma}|$.

If one linearly transforms the data ($y_i = Ax_i$) and models y_1^N using a diagonal gaussian then ML value is

$$p_{diag}^*(y_1^N) = g(N, d) |\text{diag}(A\bar{\Sigma}A^T)|^{-\frac{N}{2}}.$$

The best ML value is a function of the transformation A which is assumed to be unimodular. One can maximize this over A

to obtain the best feature space in which to model with the diagonal covariance constraint. By inspection it is easy to see *one* optimal choice of A : $A = U^T$, where $\bar{\Sigma}_x = U\Lambda U^T$ is the eigendecomposition of $\bar{\Sigma}_x$. With this choice

$$p_{diag}^*(y_1^N) = g(N, d) |\Lambda|^{-\frac{N}{2}} = g(N, d) |\bar{\Sigma}_x|^{-\frac{N}{2}} = p_{one}^*(x_1^N).$$

In other words, in the transformed space there is no loss in likelihood relative to full-covariance modeling.

3. MULTI-CLASS MODELING

In this case the training data is modeled with a Gaussian for each class: (μ_j, Σ_j) . One can split the data into J classes and model each one separately. Hence the ML estimates are $\hat{\mu}_j = \bar{\mu}_j$, $\hat{\Sigma}_j = \bar{\Sigma}_j$ and the ML value is

$$p^*(x_1^N) = p(x_1^N, \{\bar{\mu}_j\}, \{\bar{\Sigma}_j\}) = g(N, d) \prod_{j=1}^J |\bar{\Sigma}_j|^{-\frac{N_j}{2}}. \quad (6)$$

Notice that the ML estimates of the parameters for each are obtained solely based on the examples from the class. There is "no interaction" between the classes and therefore unconstrained ML modeling is not "discriminating" between the classes.

Each class can be modeled in its own feature space using unimodular transformations as discussed earlier. However, this does not change the ML value or help in better classification.

3.1. Constrained ML - Diagonal Covariance

In this case the ML estimates are $\hat{\mu}_j = \bar{\mu}_j$, $\hat{\Sigma}_j = \text{diag}(\bar{\Sigma}_j)$, and the ML value is

$$p_{diag}^*(x_1^N) = g(N, d) \prod_{j=1}^J |\text{diag}(\bar{\Sigma}_j)|^{-\frac{N_j}{2}}. \quad (7)$$

If one linearly transforms the data from each class with a matrix A_j , and then models it with a diagonal gaussian the ML value of likelihood is

$$p_{diag}^*(y_1^N) = g(N, d) \prod_{j=1}^J |\text{diag}(A_j \bar{\Sigma}_j A_j^T)|^{-\frac{N_j}{2}}.$$

Equivalently the likelihood of the data in the original space is

$$p_{diag}^*(x_1^N) = g(N, d) \prod_{j=1}^J |A_j|^{N_j} |\text{diag}(A_j \bar{\Sigma}_j A_j^T)|^{-\frac{N_j}{2}}.$$

By choosing A_j to be the eigenbasis of $\bar{\Sigma}_j$, $p_{diag}^*(x_1^N)$ achieves the value $p^*(x_1^N)$, the likelihood of full-covariance modeling.

3.2. Multi-class ML Modeling - Some Issues

Firstly, if the sample size for each class (N_j) is not large enough then the ML parameter estimates may have large variance and hence be unreliable. Secondly, the storage requirements for the model is $O(Jd^2)$ - either you have to store the full-covariance or the diagonal covariance and its associated optimal feature space transform. Thirdly, in order to compute the likelihood of some test data using this model the computational requirement is $O(Jd^2)$: either you have to transform the data samples for each class and evaluate a diagonal gaussian or you have to evaluate a full-covariance Gaussian for each sample. Finally, the parameters for each class are obtained independently: ML principle does not allow for discrimination between the classes.

If we share parameters across classes then it reduces a) the number of parameters b) storage requirements c) computational requirements and sometimes d) is more discriminating leading to better classifiers. Claim d) is hard to justify without quantifying what we mean by discrimination. However, in some cases we will appeal to the Fischer-heuristic of Linear Discrimination Analysis and a result of Campbell to argue that sometimes constrained ML modeling is discriminating between classes [5, 1].

We have already seen that by imposing diagonal Gaussian models in the original feature space the number of parameters and the storage and computational requirements are reduced substantially. However, this comes with a loss in likelihood. Moreover, it is not discriminatory since the model parameters for the classes are estimated independently. We can globally transform the data with a unimodular matrix A and model the transformed data with diagonal gaussians. In this case too there is a loss in likelihood. If, among all possible transformations A , we can choose the one that takes the least loss in likelihood, in essence we will be finding a linearly transformed (shared) feature space in which the diagonal gaussian assumption is most valid (in the sense of least loss in likelihood). This is the main idea emphasized in this paper. We now look at some examples of constrained ML estimation with sharing of parameters.

3.3. Constrained ML - Equal Covariances

Here all the covariances are assumed to be equal. The ML estimates are $\hat{\mu}_j = \bar{\mu}_j$ and $\hat{\Sigma} = W = \sum_j N_j \bar{\Sigma}_j$. W is the so-called within-class-covariance. The sample covariance of the entire data (i.e., all N samples) is the sum of the within-class-covariance and between-class-covariance:

$$\bar{\Sigma} = W + B = \sum_j N_j \bar{\Sigma}_j + \frac{1}{N} \sum_{j=1}^J N_j (\bar{\mu}_j - \bar{\mu})(\bar{\mu}_j - \bar{\mu})^T.$$

Each sample on average contributes $\frac{1}{\sqrt{|W|}}$ to the likelihood and the ML value is

$$p_{eq}^*(x_1^N) = g(N, d) |\hat{\Sigma}|^{-\frac{N}{2}} = g(N, d) |W|^{-\frac{N}{2}}. \quad (8)$$

Clearly $p^*(x_1^N) \geq p_{eq}^*(x_1^N)$ (since the later imposes the equal covariance constraint and constraints can only reduce likelihood) and this gives a proof of the fact that the log of the determinant of a symmetric non-negative-definite matrix is concave. Indeed from Eqn. 8 and Eqn. 7

$$\prod_{j=1}^N |\Sigma_j|^{\frac{N_j}{N}} \leq \frac{1}{N} \sum_{j=1}^J N_j |\Sigma_j|. \quad (9)$$

Also, since $p_{eq}^*(x_1^N) \geq p_{one}^*(x_1^N)$ we get the following inequality for non-negative definite matrices W and B :

$$|W| \leq |W + B|. \quad (10)$$

3.4. Equal Covariance Clusters

Classes are organized into clusters and each cluster modeled with a single mean or collection of means and a single covariance. In the former case the data can be relabeled using cluster labels ($m_i = c_i$) and ML estimates and ML values can be obtained as before for the full-covariance multiclass case. In the latter case (of per class mean but per cluster full-covariance), the data can be split into K groups; in which case this essentially becomes the ‘‘equal-covariance’’ problem for each group.

3.5. Diagonal Covariances and Class Cluster Transformations

Again classes are grouped into clusters. Each cluster is modeled with a diagonal gaussian in a transformed feature space. That is $y_i = A_{c_i} x_i$ and y_1^N is modeled with a diagonal gaussian. The ML estimates in the original feature space are given by $\hat{\mu}_j = \bar{\mu}_j$, $\hat{\Sigma}_j = A_{c_j}^{-1} \text{diag}(A_{c_j} \bar{\Sigma}_j A_{c_j}^T) A_{c_j}^T$ and the ML value in the original feature space is

$$p_{diag}^*(x_1^N) = g(N, d) \prod_{j=1}^J |A_{c_j}|^{N_j} |\text{diag}(A_{c_j} \bar{\Sigma}_j A_{c_j}^T)|^{-\frac{N_j}{2}}. \quad (11)$$

One can choose the best feature space for each class cluster by maximizing over the A_k 's, $k \in \{1, 2, \dots, K\}$. Notice that the A_k for each class cluster is obtained independently. In the extreme case where the number of clusters is one (i.e., $K = 1$), there is single global transformation and the classes are modeled as diagonal gaussians in this feature space. The optimal A can be obtained by optimization as follows:

$$A = \text{argmax}_A |A|^N \prod_{j=1}^J |\text{diag}(A \bar{\Sigma}_j A^T)|^{-\frac{N_j}{2}}. \quad (12)$$

Differentiating the log of the objective function with respect to A and setting it to zero we get

$$\sum_j N_j (\text{diag}(A \bar{\Sigma}_j A^T))^{-1} A \bar{\Sigma}_j = N(A^T)^{-1}.$$

Either one can numerically optimize the objective function or solve the above nonlinear equation numerically. For efficient (time or memory) algorithms see [4].

3.6. Equal Covariances and Reduced-Rank Means - LDA

An interesting connection between ML modeling and Linear Discriminant Analysis was noticed by Campbell [5]. If the class covariances are equal and the means lie in a p -dimensional affine subspace $S \subset \mathbb{R}^d$ (obviously $p \leq \min(J - 1, d)$) the estimates of the means and the common covariance are projections of the sample means and the within class-covariance onto the top p LDA directions. In this case, the parameters are Σ and μ_j , with $\text{Span}\{\mu_j\}$ p -dimensional. The ML estimates are given by [5] $\hat{\mu}_j = W L L^T (\bar{\mu}_j - \bar{\mu}) + \bar{\mu}$ and $\hat{\Sigma}_j = W + \sum_j \frac{N_j}{N} (\bar{\mu}_j - \hat{\mu}_j)(\bar{\mu}_j - \hat{\mu}_j)^T$, where L is the matrix of p leading eigenvectors of $W^{-1}B$ (or LDA directions). This suggests that a formulation of ML with unequal covariances should, being a generalization of LDA, lead to better discrimination; an idea explored by Kumar in [1] where the development can easily be seen to imply the results of the previous section as a special case.

4. CONSTRAINED MEAN ESTIMATION - MLLR ADAPTATION

Consider the constrained estimation of the means of the form $A m_j + b$, where $\{m_j\}$ are known. The variances Σ_j are assumed to be known. From Eqn. 2 by substituting for μ_j and Σ_j we get a quadratic expression in A and b . Maximizing the likelihood of the data is equivalent to minimizing this quadratic expression: $\sum_j N_j (\bar{\mu}_j - A m_j - b)^T \Sigma_j^{-1} (\bar{\mu}_j - A m_j - b)$. The optimal values of A and b are obtained by solving a set of linear equations. This is essentially the MLLR technique for adaptation of gaussian means which is widely used in speaker/environment adaptation [8]. $\{m_j, \Sigma_j\}$ can be thought of as prior means and variances and the idea is to adapt m_j to $A m_j + b$ by choosing parameters that maximize the likelihood

of the adaptation data. Typically in standard implementations of MLLR, there are several (A, b) pairs shared across class clusters that are independently estimated. Also typically the Σ_j 's are assumed to be diagonal in the above problem.

5. CONSTRAINED VARIANCE ESTIMATION - FULL-VARIANCE ADAPTATION TRANSFORM

Given a prior model with means $\{\mu_j\}$ and diagonal variances $\{D_j\}$ and some adaptation data one can ask what is ML estimate of constrained variances of the $\Sigma_j = AD_jA^T$ for some parameter A . Such a constrained form for the covariance has the advantage that it can be implemented using diagonal covariances with a feature space transformation and transformation of the means [4]. In this case, from Eqn. 2 one sees that this corresponds to minimizing the following expression over A :

$$\sum_j N_j \{ \log |\Sigma_j| + \text{Tr}(A^{-T} D_j^{-1} A^{-1} (\bar{\Sigma}_j + (\bar{\mu}_j - \mu_j)(\bar{\mu}_j - \mu_j)^T)) \}$$

where μ_j and D_j are prior information about the means and variances and $\bar{\mu}_j$ and $\bar{\Sigma}_j$ are sample means and covariances from the test data (see [4], where an efficient algorithm to compute A is given).

6. CONSTRAINED MEAN AND VARIANCE ESTIMATION - SPEAKER ADAPTED TRAINING

The Speaker Adapted Training (SAT) technique described in [9] can also be viewed as a constrained ML estimation problem. In this case, each speaker (or environmental condition) has its associated set of gaussians. So the parameters are the means and covariances of all the classes for all the speakers. However, SAT postulates the existence of a "canonical" speaker with means $\{m_j\}$ and covariances $\{\Sigma_j\}$ such that the means of class j for speaker s is estimated in the form $\mu_{s,j} = A_s m_j + b_s$ and the variances are Σ_j . The parameters in this estimation are $\{m_j, \Sigma_j\} \cup \{A_s, b_s\}$.

7. FACTOR ANALYZED COVARIANCES - SHARED FACTORS

In earlier sections we saw that the covariance structure of gaussians can be captured better with diagonal gaussians if the classes are modeled in an optimal feature space. Equivalently, one can say that covariances are represented in the original space as AD_jA^T , where D_j are the diagonal covariances in the optimal space and A maps the optimal space to the original space. For the covariance, this structure is one of many structures that is more flexible than a diagonal structure. A classic method of modeling covariance structure with a reduced number of parameters is factor analysis [14]. Recently, factor analysis has been successfully used to model the covariance structure for the speech recognition problem [17, 18]. In factor analysis each class covariance is assumed to have the following form $\Lambda_j \Lambda_j^T + \Psi_j$, where Ψ_j is a diagonal "specific" covariance, and $\Lambda_j \Lambda_j^T$ is the "commonality" variance. Λ_j is the "factor loading matrix", which usually has fewer columns (each column corresponding to a factor) than rows. Factor analysis corresponds to modeling the data x from a gaussian process using zero-mean, unit variance uncorrelated factors z and zero-mean uncorrelated "noise" u with variance Ψ .

$$x = \Lambda z + u + \mu. \quad (13)$$

Clearly, the mean of the gaussian process is μ and the covariance is $\Lambda \Lambda^T + \Psi$. From Eqn. 2 one can directly compute the ML estimates of the factors numerically. In fact for gaussian mixture factor analysis an EM algorithm the explicitly uses factor variables has been available for several years [15, 16].

One can imagine scenarios where either the Ψ_j 's, Λ_j 's or both are shared across class clusters to reduce the number of parameters. In all these cases, in a straightforward algebraic fashion (writing down the so-called Q function and differentiating it with respect to the parameters) one can obtain an EM algorithm for the factor loading matrices and specific variances within. The latent variables in this EM algorithm are the factors [13].

8. FACTOR ANALYZED COVARIANCES AND OPTIMAL FEATURE SPACES

Factor analyzed covariances represent specific assumptions about the structure of the covariance matrix. This begs the questions whether these assumptions are "more valid" in a linearly transformed space. For example consider the single gaussian case. In this case, for a given number of factors Λ and Ψ can be obtained. However, if we linearly transform the data into the eigenbasis of its sample covariance matrix, then, in that space the optimal Λ is zero and Ψ is the diagonal matrix of eigenvalues of $\bar{\Sigma}$. Moreover, there is no loss in likelihood relative to full covariance modeling. Such optimal transformed spaces can, as we have seen before for the diagonal constraint, be chosen in a class-cluster dependent fashion. In the original space, the covariance is being represented in the form $\Lambda \Lambda^T A^T + A \Psi A^T$. Since ΛA can be re-labeled as Λ , factor analysis in optimal feature spaces corresponds to having covariances of the form $\Lambda_j \Lambda_j^T + A \Psi_j A^T$ in the original space. These A 's could be shared across a cluster of classes and be labeled A_k . In a fairly straightforward fashion one can concoct a numerical scheme, which, given $\{A_k\}$, computes the optimal Λ_j 's and Ψ_j 's. A descent algorithm can therefore be used to find the optimal $\{A_k\}$. A more efficient way to compute the Λ_j 's, Ψ_j 's and A_k 's is possible by explicitly using the hidden factor variables and the EM algorithm. In this case, the Λ_j 's, Ψ_j 's, and A_k 's (each of which could be individually shared across classes albeit with minor constraints) are obtained in the M-step and guaranteed to increase the likelihood [13].

9. CONSTRAINTS AND OPTIMAL FEATURE SPACES

Constraints on the covariance structure sometimes implies a better model in an optimal feature space. Equivalently, this implies a modified (and more flexible) covariance structure in the original space. The former interpretation is useful, for example, in classifier implementation, while the latter is sometimes more useful for simplifying the problem. One wonders, then, what constraints on the mean and covariance can be better modeled in linearly transformed spaces. The answer is simple: if a covariance constraint is invariant to linear transformations (ILT) (i.e., if Σ is of a particular form then $A \Sigma A^T$ is also of the same form), then there is no gain in going to any other linearly transformed feature space; similarly, if a constraint on the means is invariant to linear transformations, there is no gain in going to a linearly transformed space for modeling. Some examples will illustrate this point. Take LDA for example: the constraint on the variance is ILT (since its a shared covariance), and the constraint on the means is also ILT (since its a geometric constraint on the means - their being reduced rank). Therefore, LDA is invariant to linear transformations of the data; one does not talk about optimal feature spaces for LDA! As another example, consider diagonal or block diagonal covariances: clearly this is not ILT. However, if the constraints are of the form AD_jA^T , with D_j diagonal, then they are ILT. Therefore linear transformations should give no gain in likelihood. The same goes with with factor analysis: covariances of the form $\Lambda \Lambda^T + \Psi$, with diagonal Ψ are not ILT, while those of the form $\Lambda \Lambda^T + A \Psi A^T$ are ILT. Finally consider MLLR adaptation where the means

are of the form $Am_j + b$. This is invariant to linear transformations. Indeed, if transformed by B , the mean in the new space has the same form: $BAB^{-1}Bm_j + Bb$. In summary, if one imposes constraints on means and variances, then, either they have to be ILT or one can model better in an linearly transformed optimal (perhaps class-dependent) feature space.

10. THE CONSTRAINED ML RECIPE

The main idea in this paper can be summarized as follows:

```

if (constraints on means and covariances are ILT)
  solve standard constrained ML problem
else
  either
    design optimal feature spaces using a
    generic numerical algorithm or one specialized
    to the problem
  or
    reformulate an associated ILT constraints problem
    and solve it numerically
fi

```

The optimal feature space viewpoint is useful from an implementation point of view because once the optimal feature space is designed any available code for the constrained ML problem can be reused. Often, the equivalent ILT problem is easier to work with to derive a numerical algorithm for designing the optimal feature space. In this section we tabulate several situations in which gaussian parameters are estimated in a constrained ML fashion (some of which we've already seen) and state whether the constraints are ILT or not.

11. SPEECH RECOGNITION EXPERIMENTS

A study of optimal feature spaces for diagonal gaussian modeling was carried out in the context of the ARPA Hub4 Broadcast News (BN) speech recognition task. The baseline recognition system ([11]) had 3500 classes (HMM states) modeled by gaussian mixtures (a total of 90K gaussians) in \mathbb{R}^{60} obtained by double-rotation (a variant of LDA) of cepstral features derived from the speech data [10]. The training data consisted of $N \approx 24M$ labeled samples. Because of data insufficiency and storage cost, sample covariances were computed only at the HMM state level. In other words, for computing the optimal feature spaces the classes were assumed to be modeled by gaussians (rather than gaussian mixtures). The optimal spaces were obtained by numerically optimizing Eqn. 11 using a conjugate gradient method with analytic gradient supplied. Once the spaces are known, using standard techniques, the classes were modeled by gaussian mixtures. The test data consisted of the planned speech (F0) and spontaneous speech (F1) portions of the 1996 DARPA Hub4 evaluation test. Results of two experiments are shown in Table 1 showing a significant gain in accuracy. The first experiment used a single feature space transform (i.e., single cluster), while the second used four class clusters; one each for the HMM states of the following sounds a) stop-consonants and flaps, b) fricatives, c) vowels and diphthongs d) nasals, glides and silence. The single cluster case performs better than the four cluster case. In fact several experiments with phonetic unit as clusters (51 clusters) and sub-phonetic units as clusters (153 clusters) were attempted with marginal gains at best over the single cluster case. This example seems to suggest that sharing between classes (in this case feature spaces for class clusters) seems to lead to better classification and hence discrimination.

12. CONCLUSION

This paper describes several issues in ML modeling with gaussians. In particular it shows that constrained gaussian modeling generally implies an optimal feature space in which the constraint is more valid. Several examples of this phenomenon are given. Sometimes constrained ML modeling can also

Expt	F0 (planned)	F1(spontaneous)
Baseline	21.1	29.1
1 Transform	19.3	28.4
4 Transforms	19.4	29.0

Table 1. % Word Error Rate Using Optimal Feature Spaces for Diagonal Gaussian Modeling of HMM state clusters: a) Baseline b) Single feature space c) Four class cluster feature spaces.

lead to LDA, a classical result that seems to be relatively unknown in the speech recognition community. Constraints such as sharing parameters leads to advantages in robustness, computation, storage, and perhaps discrimination. Some well-known matrix inequalities are introduced in the context of ML modeling. Several model adaptation algorithms can also be viewed as constrained ML modeling of adaptation data. An application of the optimal feature space idea in the context of diagonal gaussian constraint for the speech recognition problem is shown to give significant improvements to baseline word error rate. Covariance modeling using factor analysis in optimal feature spaces was introduced. This is being currently investigated for the speech recognition problem [13].

Acknowledgment The author thanks members of the signal processing interest group at IBM T. J. Watson Research. C. Neti introduced the author to [1] that started this investigation. This work was supported by DARPA in part under contract #DABT63-94-C0042 and in part under contract #MDA972-97-C-0012.

REFERENCES

- [1] N. Kumar. "Investigation of Silicon-Auditory Models and Generalization of LDA for Improved Speech Recognition", PhD Thesis, Johns Hopkins Univ., 1997.
- [2] M. J. F. Gales, "Semi-tied Full-covariance matrices for hidden Markov Models", Tech. Report, CUED/FINFENG/TR287, Cambridge Univ., 1997.
- [3] M. J. F. Gales, "Semi-tied Covariance Matrices", Proceedings of ICASSP 1998.
- [4] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", Tech. Report, CUED/FINFENG/TR291, Cambridge Univ., 1997.
- [5] N. A. Campbell, "Canonical Variate Analysis - A General Model Formulation", Austral. J. Statist., 1984, 86-96.
- [6] A. P. Dempster et al., "Maximum Likelihood from Incomplete data via the EM Algorithm", Journal of the Royal Statistical Society, 39:1-38, 1977.
- [7] A. Ljolje, "The Importance of cepstral parameter correlations in speech recognition", Comp. Speech. and Language, 8:223-232, 1994.
- [8] C. J. Legetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171-186.
- [9] C. Anastasakos et al., "A Compact Model for Speaker Adapted Training", Proc. ICSLP, 1996, pp. 1137-1140.
- [10] L. Bahl et. al., "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA NAB News Task", Proc. SLT Workshop, Austin, TX, 1995.
- [11] L. Polymenakos et al., "Transcription of Broadcast News - Some Recent Improvements to IBM's LVCSR System", Proceedings of ICASSP-98.
- [12] R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions for Classification", Proceedings of ICASSP 1998.
- [13] R. A. Gopinath, S. Dharanipragada and B. Ramabhadran, "Covariance Modeling using Factor Analysis in Optimal Feature Spaces", submitted to IEEE Trans. in SP.

- [14] K. V. Mardia, J. T. Kent and J. M. Bibby, "Multivariate Analysis", Academic Press, 1979.
- [15] D. B. Rubin and D. T. Thayer, "EM Algorithms for ML Factor Analysis". *Psychometrika*, Vol 47, No. 1, March, 1982.
- [16] Z. Ghahramani and G. E. Hinton, "The EM Algorithm for Mixtures of Factor Analyzers", CRG-TR-96-1, May 1997, University of Toronto.
- [17] M. Rahim and L. Saul, "Minimum Classification Error Factor Analysis for Automatic Speech Recognition", Proceedings of 1997 IEEE ASRU Workshop, Santa Barbara, CA, Dec 1997.
- [18] L. Saul and M. Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition", Submitted to *Trans. in SP*, 1997. Also an AT&T Tech. Report available from the authors: lsaul,mazin@research.att.com.