MAXIMUM LIKELIHOOD MODELING WITH GAUSSIAN DISTRIBUTIONS FOR CLASSIFICATION

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 for classification often suffers from the following problems: a) data insufficiency implying overtrained 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. It 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 constraints on the parameters are shown to lead to Linear Discrimination Analysis (a well-known result) while others are shown to lead to optimal feature spaces (a relatively new result). Applications of some of these ideas to the speech recognition problem are also given.

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.

In this paper we consider 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 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.

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 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 [5].

The main idea emphasized in this paper is that in constrained ML modeling (eg., diagonal covariances) there are optimal feature spaces in which to model the classes. The author was first exposed to this idea in [1]; it is also explored in a less general form with a set of efficient algorithms in [2].

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

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

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 $c_j \in \{1, 2, \ldots, K\}$. Notice that $p(x_1^N, \{\mu_j\}, \{\Sigma_j\})$ can be expressed as follows [1, 2, 3]:

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

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: $y_i = A_{l_i} x_i$, where A_j is a non-singular $d \times d$ matrix. This gives an 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 multiclass problem). Then from Eqn. 2, $p_{one}(x_1^N, \mu, \Sigma)$ can be expressed as

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

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

$$p^\star_{one}(x_1^N) = p_{one}(x_1^N, \bar{\mu}, \bar{\Sigma}) = g(N,d)\left|\bar{\Sigma}\right|^{-\frac{N}{2}}, \tag{4}$$

where $g(N,d) = (2\pi e)^{-\frac{Nd}{2}}$. On average each sample contributes $\bar{\Sigma}^{-\frac{1}{2}}$ to the ML value $p^\star_{one}(x_1^N)$, 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:

$$p^\star_{one}(y_1^N) = g(N,d)\left|A\bar{\Sigma}A^T\right|^{-\frac{N}{2}} = |A|^{-N} p^\star_{one}(x_1^N). \tag{5}$$

If $|A| = 1$ then $p^\star(y_1^N) = p^\star(x_1^N)$. 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 $\hat{\Sigma} = diag(\bar{\Sigma})$. The ML value is given by

$$p^\star_{diag}(x_1^N) = p(x_1^N, \bar{\mu}, diag(\bar{\Sigma})) = g(N,d)\left|diag(\bar{\Sigma})\right|^{-\frac{N}{2}}.$$

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

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

$$p^\star_{diag}(y_1^N) = g(N,d)\left|diag(A\bar{\Sigma}A^T)\right|^{-\frac{N}{2}}.$$

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 $\bar{\Sigma}$. With this choice

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

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

3. MULTICLASS 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

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

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 $\hat{\mu}_j = \bar{\mu}_j$, $\hat{\Sigma}_j = diag(\bar{\Sigma}_j)$, and the ML value is

$$p^\star_{diag}(x_1^N) = g(N,d)\prod_{j=1}^{J}\left|diag(\bar{\Sigma}_j)\right|^{-\frac{N_j}{2}}. \tag{7}$$

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

$$p^\star_{diag}(y_1^N) = g(N,d)\prod_{j=1}^{J}\left|diag(A_j\bar{\Sigma}_jA_j^T)\right|^{-\frac{N_j}{2}}.$$

Equivalently the likelihood of the data in the original space is

$$p^\star_{diag}(x_1^N) = g(N,d)\prod_{j=1}^{J}|A_j|^{N_j}\left|diag(A_j\bar{\Sigma}_jA_j^T)\right|^{-\frac{N_j}{2}}.$$

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

3.2. Multiclass 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 [4, 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:

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

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

$$p^{\star}_{eq}(x_1^N) = g(N,d) \left|\hat{\Sigma}\right|^{-\frac{N}{2}} = g(N,d) \left|W\right|^{-\frac{N}{2}} . \qquad (8)$$

Clearly $p^{\star}(x_1^N) \geq p^{\star}_{eq}(x_1^N)$ (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

$$\prod_{j=1}^{N} \left|\bar{\Sigma}_j\right|^{\frac{N_j}{N}} \leq \frac{1}{N} \sum_{j=1}^{j=J} N_j \left|\bar{\Sigma}_j\right| . \qquad (9)$$

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

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

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 ($m_i = c_{l_i}$) 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 $y_i = A_{c_{l_i}} x_i$ and y_1^N is modeled with a diagonal gaussians. The ML estimates *in the original feature space* are given by $\hat{\mu}_j = \bar{\mu}_j$, $\hat{\Sigma}_j = A_{c_j}^{-1} diag(A_{c_j} \bar{\Sigma}_j A_{c_j}^T) A_{c_j}^T$ and the ML value in the original feature space is

$$p^{\star}_{diag}(x_1^N) = g(N,d) \prod_{j=1}^{J} \left|A_{c_j}\right|^{N_j} \left|diag(A_{c_j} \bar{\Sigma}_j A_{c_j}^T)\right|^{-\frac{N_j}{2}} . \qquad (11)$$

One can choose the best feature space for each class cluster by maximizing over the A_k’s, $k \in \{1, 2, \ldots, K\}$. 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:

$$A = arg\max_A |A|^N \prod_{j=1}^{J} \left|diag(A \bar{\Sigma}_j A^T)\right|^{-\frac{N_j}{2}} . \qquad (12)$$

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

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

Either one can numerically optimize the objective function or solve the above nonlinear equation numerically. For efficient (time or memory) algorithms see [3].

3.6. Equal Covariances and Reduced-Rank Means - LDA

An interesting connection between ML modeling and Linear Discriminant Analysis was noticed by Campbell [4]. 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 $Span\,\{\mu_j\}$ p-dimensional. The ML estimates are given by [4] $\hat{\mu}_j = W L L^T (\bar{\mu}_j - \bar{\mu}) + \bar{\mu}$ 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. SOME ADDITIONAL CONSTRAINTS

Eqn. 2 allows one to readily see the expressions for several additional constraints. For example, consider ML estimation of (A, b), where the means are assumed to be of the form $A\mu_j + b$ and the variances are known. Clearly from Eqn. 2, (A, b) can be obtained by solving the linear equations corresponding to minimizing the quadratic $\sum_j N_j(\bar{\mu}_j - A\mu_j - b)^T \Sigma_j^{-1}(\bar{\mu}_j - A\mu_j - b)$. This is the basic idea in the MLLR technique of adaptation of gaussian means which is widely used in speaker/environment adaptation in speech recognition [7]. As another example, consider ML estimates of variances of the form $\Sigma_j = AD_jA^T$ (with known means and known diagonal matrices D_j). This is useful in adapting the diagonal gaussian model variances to test data, for instance. From Eqn. 2 this corresponds to minimizing the following expression over A:

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

where μ_j and D_j are prior information about the means and variances (from training data, say), and $\bar{\mu}_j$ and $\bar{\Sigma}_j$ are sample means and covariances from the test data (see [3], where efficient algorithms are also given).

5. 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 ([9]) 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 [8]. 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 5. 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 dipthongs 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.

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.

6. CONCLUSION

This paper describes several issues in ML modeling with gaussians. In particular it shows that constrained gaussian modeling can (depending on the constraint) lead to LDA or optimal feature spaces for modeling classes. Sharing parameters leads to advantages in robustness, computation, storage, and perhaps discrimination. Well-known matrix inequalities are introduced in the context of ML modeling. Some forms of constrained ML estimation of gaussians are essentially methods for adapting means and/or variances of *trained* models to maximize the likelihood of *test* data. An application of the optimal feature space idea to the speech recognition problem is shown to give significant improvements to baseline word error rate.

Acknowledgment The author thanks the members of the signal processing interest group at IBM T. J. Watson Research for several fruitful discussions; particularly C. Neti who introduced the author to the work [1] that started this investigation.

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, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", Tech. Report, CUED/FINFENG/TR291, Cambridge Univ., 1997.

[4] N. A. Campbell, "Canonical Variate Analysis - A General Model Formulation", Austral. J. Statist., 1984, 86-96.

[5] A. P. Dempster et al., "Maximum Likelihood from Incomplete data via the EM Algorithm", Journal of the Royal Statistical Society, 39:1-38, 1977.

[6] A. Ljolje, "The Importance of cepstral parameter correlations in speech recognition", Comp. Speech. and Language, 8:223-232, 1994.

[7] 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.

[8] 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.

[9] L. Polymenakos et al., "Transcription of Broadcast News - Some Recent Improvements to IBM's LVCSR System", submitted to ICASSP-98.