MINIMUM DISCRIMINATION INFORMATION CLUSTERING: MODELING AND QUANTIZATION WITH GAUSS MIXTURES

Robert M. Gray and John C. Young and Anuradha K. Aiyer

Information Systems Laboratory, Department of Electrical Engineering, Stanford, CA 94305
rmgray@stanford.edu

ABSTRACT

Gauss mixtures have gained popularity in statistics and statistical signal processing applications for a variety of reasons, including their ability to well approximate a large class of interesting densities and the availability of algorithms such as EM for constructing the models based on observed data. We here consider a different motivation and framework based on the information theoretic view of Gaussian sources as a "worst case" for compression developed by Sakrison and Lapidoth. This provides an approach for clustering Gauss mixture models using a minimum discrimination distortion measure and provides the intuitive support that good modeling is equivalent to good compression.

1. INTRODUCTION

Gauss mixtures have played an important role in modeling random processes for purposes of both theory and design. Although newly popular, they have been used in signal processing for many decades. For example, linear predictive coded speech (LPC) can be viewed as fitting Gauss mixture models to speech when the autoregressive (AR) models fit to segments of speech are excited by Gaussian residual processes. In this case the synthesized speech becomes a composite Gaussian process and hence locally a Gauss mixture. The most popular means of fitting a Gauss mixture model to data is the EM algorithm, but clustering techniques with suitable distortion measures between observed data and resulting model can be used, as was the Itakura-Saito distortion used for fitting AR models and quantizing them in [3]. The Itakura-Saito distortion is an example of a minimum discrimination information (MDI) distortion, a measure based on model fitting techniques of Kullback using relative entropies [4]. Potential advantages of clustering techniques over the EM algorithm are the use of minimum distortion rules for model selection and the formulas describing centroids for the distortion measures, formulas which when combined with quantization theory provide quantitative relations between minimum discrimination information distortion measures and the performance of optimized robust compression systems.

One of the key properties of the Gaussian model is its role as a "worst case" in compression/source coding problems, a characterization developed by Sakrison [6] and Lapidoth [5] and subsequently extended to show that a Gauss mixture model provides a "worst case" model for compression for any mixture source, including sources formed by classifying and conditioning [1]. Thus the use of Gauss mixture models provides a robust approach to classification and compression of nonGaussian sources with similar local second order properties.

The basic idea of MDI clustering to form Gauss mixture models is reviewed, its application to random fields is discussed, and a preliminary application to context based image retrieval is described. Further results will be presented at ICIP.

2. PRELIMINARIES

Suppose that $X = \{X_i;\ i \in \mathcal{Z}_N\}$, $\mathcal{Z}_N = \{0, 1, 2, \ldots, N-1\}$ is a k-dimensional Gaussian random vector with probability density function (pdf) g, mean vector m, and covariance matrix K with determinant $|K|$. Notational problems arise when using ordinary vectors and matrices to model images, e.g., for some purposes it is more useful to think of an image as a raster or random field $X = \{X(i,j);\ i,j \in \mathcal{Z}_N\}$ rather than as a single-indexed vector $X = \{X_i;\ i \in \mathcal{Z}_{N^2}\}$. In the latter case the covariance matrix is easily described in vector notation as $K = E[(X-EX)(X-EX)^t]$, but in the former case it is often more convenient to deal directly with the covariance function. For example, if an image is assumed to be spatially stationary, then the covariance function will be a Toeplitz operator, but if the raster is converted into a single indexed vector X to obtain the covariance matrix K, the matrix will not be a Toeplitz matrix.

Given a k-dimensional random vector X with a smooth pdf f, a Lloyd-optimal vector quantizer is described by (see, e.g., [2])

• an encoder α mapping input vectors x into an index set $\mathcal{I}$

This work was partially supported by by the National Science Foundation under grants NSF:MIP-9706284-001 and CCR-0073050.

- a decoder β mapping each index $i \in \mathcal{I}$ into a reproduction value $y_i \in \mathcal{C} = \{y_m;\ m \in \mathcal{I}\}$
- an overall quantizer mapping $Q(x) = \beta(\alpha(x))$
- a distortion measure $d(x, y_i)$ between input x and reproduction y_i.
- a measure of rate (in bits or nats) required to specify y_i.

The average distortion is defined by

$$D_f(Q) = E_f[d(X, Q(X))].$$

Several notions of rate are used. The most common are $r(y_i) = \log ||\mathcal{C}||$ for fixed rate coding, $r(y_i)$ = the number of bits required by a noiseless code to specify i to the decoder, and $r(y_i) = -\log p(y_i)$, where $p(y_i)$ is the probability X is encoded into reproduction y_i. The latter definition is an approximation to the optimal rate when the codeword indices are optimally encoded, e.g., by a Huffman code. We use this definition of rate, which results in entropy-constrained vector quantization (ECVQ) and an average rate $R_f(q) = H_f(q(X))$, the entropy of the quantized output.

The operational distortion-rate function $\delta(R)$ is $\delta_f(R) = \inf_{Q:R_f(Q)\leq R} D_f(Q)$. Optimal codes must satisfy the generalized Lloyd conditions:

- The encoder is the minimum Lagrangian distortion mapping $\alpha(x) = \text{argmin}_i[d(x, y_i) + \lambda r(y_i)]$, where λ is a Lagrange multiplier.
- The reproduction codewords are centroids: $y_i = \inf_y E[d(X, y)|\alpha(X) = i)]$
- The indices are optimally losslessly encoded.

The Lloyd clustering algorithm iteratively applies these properties to improve a given code. The algorithm is well defined whenever both the minimum distortion rule and the centroid rule can be applied with reasonable complexity.

3. MINIMUM DISCRIMINATION INFORMATION QUANTIZATION

Consider now the problem of fitting a Gaussian mixture model to observed data as given by a learning or training set. The primary motivation here is that Gaussian models will provide a worst case for the actual source data that is mapped into the model. Because there are many such Gaussian models which will be chosen at random according to the observed source data, the overall model is a composite Gaussian source or, confining attention to a single vector, a Gauss mixture. We follow Kullback's approach as applied to low rate speech coding [3]. The method is simply an extension of the speech case to multiple dimensional sources such as images.

Since each Gaussian model is described by its mean and covariance matrix, say (m_l, K_l) for the lth model, the issue is how to measure the distortion between an observed vector x and each of the models in order to select the one with the smallest distortion. We assume that second order moments can be estimated from the observation x, that is, we have estimates $\hat{m}_x$ and $\hat{K}_x$. This effectively assumes that it is the second order characteristics which are important. Assuming local spatial stationarity, $\hat{m}_x$ and $\hat{K}_x$ might be estimated by a sample average, e.g.,

$$\hat{K}_{x,m}(n) = \frac{\sum_{i,j:|i-j|=n}(x_i - m)(x_j - m)}{N(n)};\ n \in \mathcal{I}^2 \qquad (1)$$

where, e.g., one might choose $N(n) = \#\{i, j : |i - j| = n\}$. Choosing $m = \hat{m}_x$ in particular yields a covariance estimate. This is a notoriously bad estimate since some values are based on very few pixels, but the estimates will be smoothed when computing centroids in the Lloyd clustering. Alternatively, one might use sample averages only for small lags where they are reasonably trustworthy, e.g., only for adjacent pixels, and then find a "maximum entropy" extension if it exists, e.g., estimate the full $\hat{K}$ as that agreeing with the trusted value and having the maximum determinant $|K|$ (which means the maximum differential entropy over all pdfs with the known second order moments). This is an example of the famous MAXDET algorithm [7].

For a pdf estimate $\hat{f}$ consistent with the moment constraints the distortion from the input to g_l is given by the relative entropy $H(\hat{f}||g_l) = \int dx\, \hat{f}(x) \ln \hat{f}(x)/g_l(x)$ Choose the pdf $\hat{f}$ as the density consistent with the moment constraints which minimizes the relative entropy between $\hat{f}$ and the fixed g_l. This is the *minimum discrimination information (MDI) density estimate* of $\hat{f}$ given g_l and the second order constraints. If g is assumed to be Gaussian, then the minimizing $\hat{f}$ will also be Gaussian and

$$\begin{aligned} & d_{MDI}(x, (m_l, K_l)) \\ &= H(\hat{f}||g_l) \\ &= \frac{1}{2}[\log \frac{|K_l|}{|\hat{K}_x|} + \text{Tr}(\hat{K}_x K_l^{-1}) \\ &\quad +(\hat{m}_x - m_l)^t K_l^{-1}(\hat{m}_x - m_l) - k]. \end{aligned}$$

This can be rewritten by reverting from the matrix form to the raster form:

$$\begin{aligned} & d_{MDI}(x, (m_l, K_l)) \\ &= \frac{1}{2}[\log \frac{|K_l|}{|\hat{K}|} + \sum_{i,j\in\mathcal{I}} \hat{K}_x(i,j) K_l^{-1}(i,j) + \sum_{i,j\in\mathcal{I}} \\ &\quad (\hat{m}_x(i) - m_l(i))(\hat{m}_x(j) - m_l(j)) K_l^{-1}(i,j) - k] \\ &= \frac{1}{2}[\log \frac{|K_l|}{|\hat{K}_x|} + \sum_{i,j\in\mathcal{I}} K_l^{-1}(i,j)[\hat{K}_x(i,j \\ &\quad +(\hat{m}(i) - m_l(i))(\hat{m}(j) - m_l(j))] - k] \\ &= \frac{1}{2}[\log \frac{|K_l|}{|\hat{K}_x|} + \sum_{i,j\in\mathcal{I}} K_l^{-1}(i,j) \hat{K}_{x,m_l}(i,j) - k] \end{aligned}$$

Itakura and Saito originally derived their "error matching measure" by an approximate maximum likelihood argument. A similar informal argument can be used here. An alternative view of matching a model to an observed vector x is to assume that x was produced by one of the Gaussian sources g_l and to choose an l according to the maximum likelihood rule, which is equivalent to choosing l to minimize the maximum-likelihood (ML) or log-likelihood (LL) distortion

$$\begin{aligned}
&d_{\mathrm{LL}}(x,(m_l,K_l)) \\
&\quad= \ln|K_l| + (x-m_l)^t K_l^{-1}(x-m_l) \\
&\quad= \ln|K_l| + \mathrm{Tr}(K_l^{-1}(x-m_l)(x-m_l)^t)
\end{aligned}$$

(this is not strictly speaking a distortion measure since it is not necessarily nonnegative). Suppose for the moment that the inverse covariance operator K_l^{-1}, i.e., the function satisfying $\sum_{j\in\mathcal{I}} K_l(i,j)K_l^{-1}(j,m) = \delta_{i-m}$ (δ is the Kronecker delta), is approximately Toeplitz, i.e., that $K_l^{-1}(j,m) \approx K_l^{-1}(j-m)$ for $j,m\in\mathcal{I}$. Assume also that the means m_l are constant vectors, e.g., $m_l = \overline{m}_l(1,\ldots,1)$. By analogy with the properties for ordinary scalar random processes, it is conjectured that this is the case for stationary random fields when the dimension k is large. By analogy with the speech case, it is also conjectured that this is the case when autoregressive modelling methods are used and the dimension is large, e.g., when the underlying model is assumed to have the form $X_n = Z_n - \sum_{k\in\mathcal{N}} X_{n-k}$ where Z_n are iid Gaussian random variables the set $\mathcal{N}$ is suitably ordered so that the random field is a Markov mesh. Then the ML rule is equivalent to the minimization of

$$\begin{aligned}
&d_{LL}(x,(m_l,K_l)) \\
&\quad= \ln|K_l| + (x-m_l)^t K_l^{-1}(x-m_l) \\
&\quad= \ln|K_l| + \sum_{i,j}(x(i)-\overline{m}_l)(x(j)-\overline{m}_l)K_l^{-1}(i,j) \\
&\quad\approx \ln|K_l| + \sum_n K_l^{-1}(n)\times \\
&\qquad\qquad \sum_{i,j:|i-j|=n}(x(i)-\overline{m}_l)(x(j)-\overline{m}_l)) \\
&\quad= \ln|K_l| + \sum_n N(n)K_l^{-1}(n)\hat{K}_{x,m_l}(n),
\end{aligned}$$

where $\hat{K}_{x,m_l}$ is the second order estimator of (1). Thus

$$\begin{aligned}
d_{\mathrm{LL}}(x,(m_l,K_l)) &\approx \ln|K_l| + \sum_{i,j} K_l^{-1}(i,j)\hat{K}_{x,m_l}(i,j) \\
&= d_{\mathrm{MDI}}(x,(m_l,K_l)) + \ln|\hat{K}_x| + k
\end{aligned}$$

When the approximation is valid, the two distortion measures yield approximately the same minimum distortion rule since they differ by a constant and by a term that depends only on the observed input x. Note in particular that the actual covariance estimate of the input $\hat{K}_x$ need not be calculated to find a minimum distortion codeword, it is the covariance of the models that is important.

4. MDI AND ML CENTROIDS

As in the analogous speech case [3], this distortion measure is amenable to the Lloyd clustering algorithm, i.e., there is a well defined minimum distortion encoder using d_{MDI} and the distortion has well defined Lloyd centroids. In particular, the centroids m_l and K_l must minimize the conditional expected distortion.

$$\begin{aligned}
&E[d_{\mathrm{MDI}}(X,g_l)\mid\alpha(X)=l] \\
&\quad= \frac{1}{2}E[\ln\frac{|K_l|}{|\hat{K}_X|} + \mathrm{Tr}(\hat{K}_X K_l^{-1}) \\
&\qquad +(\hat{m}_X-m_l)^t K_l^{-1}(\hat{m}_X-m_l) - k\mid\alpha(X)=l]
\end{aligned}$$

where $\hat{m}_X$ and $\hat{K}_X$ are the mean and the covariance estimates for observation X. The mean centroids are given by $m_l = E[\hat{m}_X\mid\alpha(X)=l]$ regardless of K_l since this choice minimizes the quadratic term in the mean as 0 (the centroid with respect to a weighted quadratic measure is the mean). With this choice of m_l need K_l to minimize

$$\begin{aligned}
&E[\ln\frac{|K_l|}{|\hat{K}_X|} + \mathrm{Tr}(\hat{K}_X K_l^{-1}) - k\mid\alpha(X)=l] \\
&\quad= \ln\frac{|K_l|}{|\overline{K}_l|} + \mathrm{Tr}(\overline{K}_l K_l^{-1}) - k + E[\ln\frac{|\overline{K}_l|}{|\hat{K}_X|}\mid\alpha(X)=l] \\
&\quad\geq E[\ln\frac{|\overline{K}_l|}{|\hat{K}_X|}\mid\alpha(X)=l]
\end{aligned}$$

with equality if $K_l = \overline{K}_l$ (since the first three terms are just the Kullback-Leibler distortion between two Gaussian distributions with the given covariances and 0 means).

The centroids for the ML distortion measure can be similarly found. Now the goal is to find m_l and K_l to minimize the conditional average distortion

$$\begin{aligned}
&E[d_{\mathrm{LL}}(X,g_l)\mid\alpha(X)=l] \\
&\quad= E[\ln|K_l| + (X-m_l)^t K_l^{-1}(X-m_l)|\alpha(X)=l]
\end{aligned}$$

As before, the optimal mean regardless of the covariance is given by $m_l = E[X\mid\alpha(X)=l]$. Define the average $\overline{K}_l = E[(X-m_l)(X-m_l)^t]$. Then

$$\begin{aligned}
&E[d_{\mathrm{LL}}(X,g_l)\mid\alpha(X)=l] \\
&\quad= [\ln\frac{|K_l|}{|\overline{K}_l|} + \mathrm{Tr}(K_l^{-1}\overline{K}_l) - k] + k + \ln|\overline{K}_l| \\
&\quad\geq k + \ln|\overline{K}_l|
\end{aligned}$$

with equality if $K_l = \overline{K}_l = E[(X-m_l)(X-m_l)^t\mid\alpha(X)=l]$, so that once again centroids are computed by averaging.

5. MDI AND ML CLUSTERING

Application of the Lloyd algorithm to the MDI or ML distortion measures yields a model VQ, a mapping of input vectors X (e.g., image blocks) into a model. Under reasonably general conditions, the Lloyd algorithm converges. If the algorithm converges to a stationary point, the centroid formulas provide a formula for the resulting MDI distortion in terms of the model covariances of the codebook and their probabilities of occurence, i.e., in terms of the Gauss mixture model produced by the Lloyd algorithm.

Since we are considering variable rate systems, it is natural to consider an entropy constrained VQ for the models as well: $d_{\mathrm{ECMDI}}(x, g_l) = d_{\mathrm{MDI}}(x, g_l) - \lambda \ln p_l$. Applying the MDI centroid formula provides a simple formula for the average ECMDI distortion:

$$D_{\mathrm{ECMDI}} = \frac{1}{2} \sum_{l} p_l \ln |K_l| - E[\ln |\hat{K}_X|] + \lambda H(p) \quad (2)$$

The ML centroid yields a similar result except that the $E[\ln |\hat{K}_X|]$ is absent.

It is shown in [1] using high rate asymptotic quantization theory that if one designs a classified VQ by first designing a classifier, e.g., the MDI VQ just considered, and then optimally designs VQs to minimize mean squared error for the resulting Gaussian models, and then applies the code by first classifying the input and then applying the optimal code for the class chosen, then the average distortion at rate R (assuming high rate and optimal bit allocation across the classes) is $D_{\mathrm{MSE}} = b_k (2\pi e) e^{\frac{2}{k}(D_{\mathrm{ECMDI}} + E[\ln |\hat{K}_X|] - R)}$, where b_k is a constant depending only on the dimension and not on the underlying pdfs and the MDI Lagrangian is chosen as $\lambda = 1$. This relates the MSE in the resulting classified VQ to the ECMDI distortion used to design the classifier, providing a new relation between modeling accuracy and the resulting performance in a quantizer based on the model.

6. AN IMAGE RETRIEVAL APPLICATION

A simple example of a clustered Gauss mixture model to image archiving and querying is presented and and compared with the common color histogram method. Comparisons to other methods are in progress. An annotated test database with one hundred 96 by 128 color images of fifteen different "types" (e.g., satellite images, indoor images) was constructed of which five images from five different types were used as a training set to produce a Gauss mixture codebook of 64 components as described. The block size was 8×8 and the dimensions of the partial covariance matrix were 2×2. Signatures for both query and target images were formed by encoding an image using the minimum distortion (MDI) encoder to obtain a histogram for the components. A simple decision tree was designed to decide whether or not a "match" occured between the query image (representing its type) and the target image based on the component histogram of each. The average accuracy results from 15 queries (one of each type) to the test database are presented along with the results for the color histogram method [8] in Table 1, where *Precision* is the fraction of the retrieved images that are relevant to the query and *Recall* is the fraction of the total number of relevant images that are retrieved.

Accuracy	Gauss Mixture	Color Histogram
Precision	0.9523	0.8928
Recall	0.9259	0.8064

Table 1. Comparison of GM-based and histogram-based image retrieval

7. REFERENCES

[1] R.M. Gray, "Gauss Mixture Vector Quantization," *Proceedings ICASSP 2001*.

[2] R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, Vol. 44, pp. 2325–2384, October 1998.

[3] R.M. Gray, A.H. Gray, Jr., G. Rebolledo, and J.E. Shore, "Rate distortion speech coding with a minimum discrimination information distortion measure,"*IEEE Transactions on Information Theory*, vol. IT–27, no. 6, pp. 708–721, Nov. 1981.

[4] S. Kullback. *Information Theory and Statistics*, Dover,New York, 1968. (Reprint of 1959 edition published by Wiley.)

[5] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.

[6] D. J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. 21, pp. 301–309, May 1975.

[7] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM Journal on Matrix Analysis and Applications*, Vol. 19, 499-533, 1998.

[8] M. Flickner and H. Shawney and W. Niblack and J. Ashley and Q. Huang and B. Dom and M. Gorkani and J. Hafner and D. Lee and D. Petkovic and D. Steel and P. Yonker, ”Query by image and video content: the QBIC system”m” *IEEE Computer*, Vol. 28, pp. 23–32, September 1995.