GAUSS MIXTURE VECTOR QUANTIZATION

Robert M. Gray

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

ABSTRACT

Gauss mixtures are a popular class of models in statistics and statistical signal processing because they can provide good £ts to smooth densities, because they have a rich theory, and because the can be well estimated by existing algorithms such as the EM algorithm. We here extend an information theortic extremal property for source coding from Gaussian sources to Gauss mixtures using high rate quantization theory and extend a method originally used for LPC speech vector quantization to provide a Lloyd clustering approach to the design of Gauss mixture models. The theory provides formulas relating minimum discrimination information (MDI) for model selection and the mean squared error resulting when the MDI criterion is used in an optimized robust classi£ed vector quantizer. It also provides motivation for the use of Gauss mixture models for robust compression systems for general random vectors.

1. INTRODUCTION

Gaussian models play a fundamentally important role in statistical signal processing and statistics for a variety of well known reasons, including their wealth of nice mathematical properties. As a notable example, a k-dimensional Gaussian random vector X with pdf g mean vector m, and covariance matrix K with determinant $|K|$,

$$g(x)=\frac{1}{(2\pi)^{\frac{k}{2}}|K|^{\frac{k}{2}}}\exp\left(-\frac{1}{2}(x-m)^tK^{-1}(x-m)\right)$$

has a simple Shannon differential entropy:

$$h_g=-\int dxf(x)\ln f(x)=\frac{1}{2}\ln(2\pi e)^k|K| \qquad (1)$$

Less well known outside of information theory is the fact that the Gaussian distribution plays an important extremal role in Shannon rate distortion theory. Shannon showed that the largest differential entropy for a given mean and covariance is achieved with the Gaussian density and Sakrison showed that of all distributions with a £xed mean and covariance, the Gaussian has the largest Shannon rate-distortion function [7]. Lapidoth strenthened this result to show that for iid Gaussian sources, a code designed for a Gaussian source yields the same rate and distortion on an arbitrary source with the same second order moments [5]. This characterizes the Gaussian source as a "worst case" source for data compression and provides an approach to robust compression. A problem with this approach is that it can be too conservative, designing a code for a single Gaussian model may yield a trustworthy rate-distortion tradeoff, but it may be far worse than that obtainable using a better source model. This motivates Gauss mixtures from a compression point of view: a Gauss mixture source may provide a "locally worst case" model if suitably used, yielding robust codes with better performance than a single Gaussian. This also suggests that, analgous to a Lloyd clustering design approach to LPC speech vector quantizers using a minimum discrimination information distortion measure (the Itakura-Saito distortion) [3], a speci£c Lloyd clustering approach can be used to design general Gauss mixture models based on training data, providing a possibly useful alternative to the popular EM algorithm. As shall be seen, the clustering approach does yield some theoretical results that yield a new interpretation of minimum discrimination information distortion measures.

2. QUANTIZATION

The basic quantization notation and results may be found, e.g., in [2]. We here recall several relevant de£nitions and results. Assume that X is a k-dimensional random vector with smooth probability density function (pdf) f. A Lloyd-optimal vector quantizer is summarized by

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

• 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 is $Q(x)=\beta(\alpha(x))$

• 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.

For simplicity we assume the traditional mean squared er-

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

ror distortion measure

$$d(x,y)=\sum_{n=0}^{k-1}|x_n-y_n|^2=(x-y)^t(x-y),$$

but the approach extends to more general measures (and will be explicitly considered for the minimum discrimination information (MDI) distortion measure for models or pdf's. 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 $p(y_i)$ = 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 minimum Lagrangian distortion mapping $\alpha(x)=\mathrm{argmin}_i[d(x,y_i)+\lambda r(y_i)]$ 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.

3. HIGH-RATE THEORY

Here we follow the approach of Gersho [1] (see also [2]), which is an intuitive deriviation of the results of Zador [9] using the quantizer point density ideas of Lloyd. Assume that $R(Q)$ is large and define *quantizer point density* of Q by

$$\#\{n:y_n\in S\}\approx\int_S\Lambda(x)\,dx$$

Then

$$D_f(Q)\approx b_k\int\Lambda^{-\frac{2}{k}}(x)f(x)\,dx$$

where b_k depends only on the dimension, and

$$R_f(Q)\approx\int f(x)\ln\frac{\Lambda(x)}{f(x)}\,dx$$

These approximations and Hölder's & Jensen's inequalities imply that for variable rate quantization the optimal is $\Lambda(x)=e^{R-h_f}$. The key observation is that the optimal Λ is constant, which means that for large rate the optimal quantizer is approximately a tessalating VQ such as a lattice quantizer. This implies that

$$\delta_f(R)\approx b_k e^{\frac{2}{k}h_f}e^{-2\frac{R}{k}}.$$

For a Gaussian pdf,

$$\delta_g(R)\approx b_k e^{\frac{2}{k}h_f}e^{-2\frac{R}{k}}=b_k(2\pi e)e^{-\frac{2}{k}R}|K|^{\frac{1}{k}}$$

Combining this fact with the extremal property of the Gaussian pdf for differential entropy immediately provides a high-rate quantization variation on Sakrison's result:

$$\sup_{f:E_f[(X-EX)(X-EX)^t]=K}\delta_f(R)=\delta_g(R)$$

This property suggests a further extension. It is often the case that full knowledge of the covariance of a random vector is lacking, for example one might only trustworthy estimates of covariance values for small lags, as in the case of low order correlations in LPC speech modeling. If the supremum above is instead taken over all f with only the partial information, then the worst case will be achieved by the Gaussian pdf with the covariance consistent with the partial information and having the largest determinant. This is the famous "maximum determinant" or MAXDET problem [8] Given index set $\overline{\mathcal{N}}$ and partial covariance $\Sigma_{\overline{\mathcal{N}}}=\{\Sigma_{i,j};\ (i,j)\in\overline{\mathcal{N}}$, find $\max_{K:K_{\overline{\mathcal{N}}}=\Sigma_{\overline{\mathcal{N}}}}|K|$. The K achieving the maximum (if it exists) is MAXDET extension of $\Sigma_{\overline{\mathcal{N}}}$.

4. QUANTIZER MISMATCH

Suppose now that Q is optimized for a Gaussian g, but applied to f. Then using the $\log p(\alpha(X)=i)$ approximation to the rate computed using the design Gaussian pdf, taking expectations with respect to f yields

$$\begin{aligned}D_f(Q)&\approx b_k\int\Lambda^{-\frac{2}{k}}(x)f(x)\,dx\\&=b_k e^{-\frac{2}{k}(R-h_g)}\approx D_g(Q)\\R_f(Q)&\approx\int f(x)\ln\frac{\Lambda(x)}{g(x)}\,dx\\&=R-\int f(x)\ln g(x)-h_g\end{aligned}$$

which for Gaussian g yields

$$\begin{aligned}R_f(Q)-R&\approx\frac{1}{2}\mathrm{Tr}(K_g^{-1}K_f)\\&+\frac{1}{2}(m_g-m_f)^tK_g^{-1}(m_g-m_f)-\frac{k}{2}\end{aligned}$$

If $m_f=m_g$ and$K_f=K_g$, then $D_f(Q)\approx D_g(Q)$ and $R_f(Q)\approx R$, reminiscent of Lapidoth's fixed rate result for iid Gaussian processes: the performance predicted for the Gaussian is actually achieved for the nonGaussian.

5. GAUSS MIXTURE VQ

Suppose that X has mixture pdf f generated by classifying X into classes $l = 1, 2, \ldots$. For the moment the classifier is arbitrary, but later its design will be considered. Let $L = L(X)$ denote the integer-valued class. Then $\{f_{X|L}(x|l), p_l = \Pr(L = l)\}$ is a mixture model for f. A separate VQ can then be designed for each class, yielding a classified VQ structure. For each class l one can estimate a conditional mean m_l and covariance K_l, possibly only partially. The worst case source for quantizing this source is then the MAXDET Gaussian. Design an optimal code Q_l for each Gaussian component g_l, yielding a (D_l, R_l) distortion/rate pair with performance that can be approximated using the high rate formulas. This yields a two-step classified VQ: First classify X into Gaussian model g_l described by (m_l, K_l), then quantize using optimal quantizer Q_l for g_l.

On the average the total information rate for the lth component is $R_l - \ln p_l$, the number of nats needed to specify quantizer used plus the encoder output for that quantizer. The overall average distortion is then $D(Q) = \sum_l D_l p_l$ and the overall average rate $R(Q) = \sum_l R_l p_l + H(p)$. From high rate theory $D_l \approx \delta(R_l) \approx b_k e^{\frac{2}{k}(h_l - R_l)}$ whence $D(Q) = b_k \sum_l e^{\frac{2}{k}(h_l - R_l)} p_l$. The optimal rate allocation $\{R_l\}$ minimizing $b_k \sum_l e^{\frac{2}{k}(h_l - R_l)} p_l$ subject to $\sum_l R_l p_l + H(p) \leq R$ is readily solved by Lagrangian methods or directly using convexity arguments: $R_l = h_l + R - H(p) - \overline{h}$ where

$$\overline{h} = \sum_l h_l p_l = h(X|L) \ \ , \ \ H(p) = H(L)$$

The Lagrangian multiplier for the modified distortion for each quantizer is the same: $\lambda_l = (2b_k/k)e^{-\frac{2}{k}(R - H(p) - \overline{h})}$.

With this assignment it turns out that the optimum quantizer point density for all the quantizers is the same, $\Lambda(x) = e^{R - \overline{h} - H(p)}$, and that the conditional average distortion for each component is the same, $D_l = b_k e^{-\frac{2}{k}(R - \overline{h} - H(p))}$ so that $D = b_k e^{-\frac{2}{k}(R - \overline{h} - H(p))}$. Plugging in for the Gaussian case

$$D_{\mathrm{MSE}} = b_k (2\pi e) e^{\frac{2}{k}(\frac{1}{2} \sum_l p_l \ln |K_l| + H(p) - R)} \tag{2}$$

From the robustness property, *this formula also gives the performance for nonGaussian source classified into a mixture with* $\{(m_l, K_l, p_l)\}$*!*

6. MINIMUM DISCRIMINATION INFORMATION CLASSIFICATION

Given the classified quantizer structure, the question now arises how to select a good set of Gaussian models (m_l, K_l) and how to classify an observed input X into one of the models. The traditional solution for designing Gauss mixture models is the expectation-maximization (EM) algorithm, but we here adopt a VQ/clustering approach similar to that used for designing LPC models in very low rate speech coding [3].

Suppose that $\overline{\mathcal{N}}$ is the collection of indices for which we trust covariance estimates $K(i, j)$. The allowed models are maximum (differential) entropy pdf's for partial covariances, i.e., each model will be specified by $K_{\overline{\mathcal{N}}} = \{K(i, j);\ (i, j) \in \overline{\mathcal{N}}\}$ as the Gaussian pdf with K given by the MAXDET extension of $K_{\overline{\mathcal{N}}}$.

To match the input X to a Gaussian model g_l specified by (m_l, K_l) assume for the moment that X yields a pdf estimate $\hat{f}$ and measure the distortion or "distance" from the input to g_l by the discrimination information (relative entropy, Kullback-Leibler information [4])

$$H(\hat{f}||g_l) = \int \hat{f}(x) \ln \frac{\hat{f}(x)}{g_l(x)}\, dx.$$

To form $\hat{f}$ from X, assume partial second order information: a constant mean estimate $\hat{m}$, e.g., a sample average, and covariance values for some index set $\overline{\mathcal{N}}$, e.g., sample averages for small lags. Effectively we are assuming a stationary random field. Choose $\hat{f}$ as the density consistent with these second order moments 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. If g is assumed to be Gaussian, than $\hat{f}$ will also be Gaussian and

$$\begin{aligned} H(\hat{f}||g_l) \quad &= \quad \frac{1}{2}\Bigg(\log \frac{|K_l|}{|\hat{K}|} + \mathrm{Tr}(\hat{K}K_l^{-1}) \\ &\qquad + (\hat{m} - m_l)^t K_l^{-1} (\hat{m} - m_l) - k\Bigg) \end{aligned}$$

Properties of maximum entropy models imply that the trace term depends only on $\hat{K}_{\overline{\mathcal{N}}}$ (e.g., [6]) imply that the relative entropy is minimized over all $\hat{f}$ by maximizing $|\hat{K}|$, i.e., by choosing $\hat{f}$ as the *maximum entropy estimate* given the estimated partial covariance, yielding

$$\begin{aligned} d_{\mathrm{MDI}}(X, g_l) \quad &= \quad \frac{1}{2}\Bigg(\log \frac{|K_l|}{|\hat{K}_X|} + \mathrm{Tr}(\hat{K}_X K_l^{-1}) \\ &\qquad + (\hat{m}_X - m_l)^t K_l^{-1} (\hat{m}_X - m_l) - k\Bigg) \end{aligned}$$

where $\hat{K}_X = \mathrm{argmax}_{K : K_{\overline{\mathcal{N}}} = \Sigma_{\overline{\mathcal{N}}}} |K|$ is the MAXDET extension of the partial covariance based on the input X.

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.

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

where $\hat{m}_X$ and $\hat{K}_X$ are the mean and the covariance estimates for observation X. Some matrix algebra leads to the conclusion that the centroids are $m_l = E[\hat{m}_X|\alpha(X)=l]$ regardless of K_l and $K_l = \overline{K}_l \equiv E[\hat{K}_X|\alpha(X)=l]$.

Application of the Lloyd algorithm to the MDI distortion yields a model VQ, a mapping of input vectors X (e.g., image blocks) into a model. The determinant maximization can be done by MAXDET, although highly structured problems usually have significantly faster algorithms, e.g., Levinson's algorithm in the speech example. Since we are considering variable rate systems, it is natural to consider an entropy constrained VQ for the models as well: $d_{\mathrm{ECMDI}}(x,g_l) = d_{\mathrm{MDI}}(x,g_l) - \lambda \ln p_l$. This model VQ using the MDI distortion provides a classifier for use with the classified robust VQ compression scheme.

The MDI centroid formula provides a simple formula for the average ECMDI distortion:

$$D_{\mathrm{ECMDI}} = \frac{1}{2}\sum_l p_l \ln|K_l| - E[\ln|\hat{K}_X|] + \lambda H(p) \quad (3)$$

In the special case where the MDI Lagrangian $\lambda = 1$, then from (2)

$$D_{\mathrm{MSE}} = b_k(2\pi e)e^{\frac{2}{k}(D_{\mathrm{ECMDI}}+E[\ln|\hat{K}_X|])} \quad (4)$$

which relates the MSE in the resulting classified VQ to the ECMDI distortion used to design the classifier, providing a new interpretation of MDI classification as the classifier which minimizes the overall MSE when used in a classified compression system.

The minimum average Lagrangian distortion for rate $R > 0$ must be smaller than that for rate $R = 0$ (since the constraint set is larger), $D_0 = \frac{1}{2}\ln\overline{K} - E[\ln|\hat{K}_X|]$, where $\overline{K} = E[\hat{K}_X]$. Thus

$$\frac{1}{2}\sum_l p_l \ln|K_l| + \lambda H(p) \leq \frac{1}{2}\ln\overline{K} \quad (5)$$

If again the MDI $\lambda = 1$, this result confirms the intuition that D_{MSE} using a Gauss mixture VQ is smaller than the distortion resulting from a single Gauss model, i.e., a Gauss mixture worst case is better than a single Gaussian worst case.

As a final interpretation of the ECMDI classifier, suppose that given an input vector X we can form an estimate of the underlying pdf, $\hat{f}$, with covariance $\hat{K}$ and mean $\hat{m}$. An alternative classification rule would be to choose the model $g_l \leftrightarrow Q_l$ which yields the smallest average $D_{\hat{f}}(Q_l) + \lambda R_f(Q_l)$. From the quantizer mismatch results with nominal rate R_l: $D_{\hat{f}}(Q_l) = b_k e^{-\frac{2}{k}(R_l-h_l)} = b_k e^{-\frac{2}{k}(R-H(p)-\overline{h})}$, which does not depend on l and hence does not effect the choice. Some algebra then yields

$$\begin{aligned}R_{\hat{f}}(Q_l) &= \frac{1}{2}\ln(2\pi e)^k + R - H(p) - \overline{h} + \frac{1}{2}\ln|\hat{K}|\\&\quad + d_{\mathrm{ECMDI}}((\hat{m},\hat{K}),(m_l,K_l))\end{aligned}$$

so that selection of l to minimize the expected Lagrangian MSE/rate distortion is equivalent to the ECMDI selection of l.

7. REFERENCES

[1] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory.*, vol. 25, pp. 373–380, July 1979.

[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] H. Lev-Ari, S.R. Parker, and T. Kailath, "Multidimensional maximum-entropy covariance extension," *IEEE Transactions on Information Theory*, Vol. 35, pp. 497–508, May 1989.

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

[8] 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.

[9] P. L. Zador, "Topics in the asymptotic quantization of continuous random variables," Bell Laboratories Technical Memorandum, 1966.