1

Gauss Mixture Quantization: Clustering Gauss Mixtures

Robert M. Gray¹

Abstract

Gauss mixtures are a popular class of models in statistics and statistical signal processing because Gauss mixtures can provide good fits to smooth densities, because they have a rich theory, because they can yield good results in applications such as classification and image segmentation, and because they can be well estimated by existing algorithms such as the EM algorithm. We here use high rate quantization theory to develop a variation of an information theoretic extremal property for Gaussian sources and its extension to Gauss mixtures. This extends a method originally used for LPC speech vector quantization to provide a clustering approach to the design of Gauss mixture models. The theory provides formulas relating minimum discrimination information (MDI) selection of Gaussian components of a Gauss mixture and the mean squared error resulting when the MDI criterion is used in an optimized robust classified vector quantizer. It also provides motivation for the use of Gauss mixture models for robust compression systems for random vectors with estimated second order moments but unknown distributions.

1.1 Introduction

Gauss mixtures have played an important role in modeling random processes for purposes of both theory and design. They have been used in statistics and statistical signal processing for many decades, but they have recently enjoyed renewed popularity in a variety of problems. Simply stated, a Gauss mixture probability density function (pdf) is a weighted sum of a collection of distinct Gaussian pdf's. A single Gaussian pdf describing a

¹Information Systems Laboratory, Department of Electrical Engineering, Stanford University, rmgray@stanford.edu. This work was partially supported by the National Science Foundation under Grant Number 0073050 and by the Hewlett Packard Corporation.

k -dimensional random variable X has the form

$$g(x) = \mathcal{N}(x, \mu, K) = \frac{1}{(2\pi)^{\frac{k}{2}} |K|^{\frac{k}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^t K^{-1} (x - \mu) \right),$$

where $K = E[(X - EX)(X - EX)^t]$ is the $k \times k$ covariance matrix, $|K|$ the determinant of k , and $\mu = EX$ the mean vector. We assume that the covariance is nonsingular and all moments are finite. A Gauss mixture density has the form

$$\bar{g}(x) = \sum_i w_i g_i(x) = \sum_i w_i \mathcal{N}(x, \mu_i, K_i),$$

where

$$\sum_i w_i = 1$$

and we assume that there are only a finite number of components.

There are two distinct ways of interpreting the physical meaning of a mixture density which lead to different methods of estimating mixture densities based on observed data. One view is to consider the weighted sum directly as a single pdf with a particular structure and try to match such a pdf directly to an observed sequence of sample vectors produced by the random phenomenon to be modeled. For example, the EM algorithm takes the mixture model literally and tries to match a weighted sum of distributions to observed data, generally attempting to approximate a maximum likelihood estimation assuming the components of the mixture to be Gaussian. Relative entropy matching techniques also usually deal directly with the weighted sum, yielding complicated formulas including sums of Gaussians within logarithms [26]. An alternative viewpoint, which we adopt, is to instead view the mixture as a “doubly stochastic” phenomenon. First nature selects which component of the mixture is to be used using the probability mass function p . Once the component, say i , is chosen, the actual random variable is then generated using the pdf g_i . The interpretation using the doubly stochastic idea is that one is in fact observing a sample vector produced by one of the Gaussian components of the mixture, but one does not know *which one*. If one is viewing a sequence of vectors from such a source instead of a single vector (as is usually the case), the underlying model is that of a composite or “switched” process, that is, the observed vectors are produced by a random slowly varying switch that successively connects distinct information sources to the output. The marginal pdf (for an individual vector) for such a source is a mixture pdf. Composite source models have been successful in the analysis of a variety of signal processing and information theory problems and are explicit or implicit in the most popular speech modeling algorithms.

The primary issue at hand is how to design a Gauss mixture considered as collection of Gaussian components together with a probability mass func-

tion providing their prior probabilities in the context of certain specific applications to be described. The answer provided is to use a clustering algorithm that attempts to match a collection of individual Gaussian components to a set of learning or training data. The underlying prior will follow from the relative frequencies at which these components are selected by the algorithm. The number of components will follow as a natural byproduct of the clustering algorithm.

Gauss mixture models played a fundamental, although originally implicit, role in the development in the 1970s of linear predictive coded speech (LPC) modeling, which yielded the first truly low bit rate speech compression systems with reasonable quality and which still dominate speech coding and recognition applications. LPC was originally introduced as the partial correlation method (PARCOR) based on an approximately maximum likelihood fit of Gaussian components to nonoverlapping segments of speech by Itakura and Saito [18]. It 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 or Gaussian process with the property that at any given instant one is observing a single Gaussian component of the mixture.

The most popular means of fitting a Gauss mixture model to data is the EM algorithm, but similar ends can be achieved using Lloyd (or Steinhaus or k -means or principal points) clustering techniques [21] (see, e.g., [11] for a technical and historical survey) with suitable distortion measures between observed data and proposed models, a method described for low rate speech coding using the Itakura-Saito distortion in [16]. 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 (Kullback-Leibler numbers, cross entropies) [19]. Lloyd clustering techniques have several potential advantages over the EM algorithm. The first is the use of minimum distortion rules for selecting the “best” Gaussian component of a mixture to fit the current observed data, a form of nearest neighbor selection of a model for an observation. A second advantage is the existence of explicit formulas describing centroids with respect to 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. A final advantage is that the number of components in a Gauss mixture model is automatically determined by the clustering algorithm. Entropy constrained Lloyd clustering automatically eliminates unused clusters so as to improve the tradeoff between distortion and rate. When the algorithm converges, the number of cluster centers – which is also the number of components in the mixture model – must be at least a locally optimum choice in terms of balancing rate and distortion.

One of the key properties of the Gaussian model is its role as a “worst case” in compression/source coding problems, a characterization most developed by Sakrison [24] and Lapidoth [20]. We describe a recent extension of this property to Gauss mixture models and explore applications to situations in which successive vectors produced by an information source are classified as belonging to one of a collection of Gaussian models which together form a Gauss mixture.

1.2 Preliminaries

The reader is referred to [11] for a summary, history, and extensive bibliography of quantization, including its intimate connections with clustering. We here recall several relevant definitions and results.

1.2.1 Information Sources

An information source is a stationary random process $\{X(n)\}$ where each $X(n)$ is a random vector taking values in $\mathbb{R}^k$ according to a generic probability distribution P_X corresponding to a generic random vector X described by a smooth probability density function (pdf) f . Smoothness of the densities is required for the approximations we use which are based on Gersho’s conjecture.

We focus on the generic random vector $X = (X_0, X_1, \dots, X_{k-1})$ with the understanding that for any information theoretic problem in fact there will be a sequence of random vectors with the same distribution. For most information theory applications we do not need to assume that the successive vectors are independent, but this is often done because most coding structures code each vector in a manner independent of past and future, so that the average behavior is the same whether or not the vectors are independent. The vectors can be thought of rectangular blocks of pixels in an image, of transformed versions of those pixels using a Fourier or wavelet transform, or as features extracted from an image.

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 integer-indexed vector. In the latter case the covariance matrix is easily described in vector notation as, but in the former case it is often more convenient to deal directly with the covariance function $\{K(i, j); i, j \in \mathcal{Z}_N^2\}$. 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 Toeplitz. We will focus on the case where X is simply considered as a vector without regard to its origins, e.g., to how a raster

image is scanned to convert it into a vector. It is worth pointing out that no matter how a raster X is “vectorized” into an integer-indexed vector, the resulting determinant will be unchanged since any vector versions of X will be related to others by a permutation matrix.

One of the useful properties of a Gaussian distribution is that it has a simple Shannon differential entropy:

$$h(g) = - \int dx f(x) \ln f(x) = \frac{1}{2} \ln(2\pi e)^k |K|. \quad (1.1)$$

Less well known outside of information theory is the fact that the Gaussian distribution plays an important extremal role in Shannon rate distortion theory, one version of which will be described in more detail shortly. Sakrison showed that of all distributions with a fixed mean and covariance, the Gaussian has the largest Shannon rate-distortion function [24], that is, the Gaussian requires the largest rate to meet a given fidelity requirement and hence has the worst performance over all sources with the same covariance. Lapidoth strengthened 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 [20]. 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 yield far worse performance 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” or “conditionally worst case” model if suitably used, yielding robust codes with better performance than a single Gaussian would provide.

1.2.2 Quantization

A vector quantizer Q (or Shannon lossy compression code or source code subject to a fidelity criterion) can be described by the following mappings and sets:

- an encoder $\alpha : \Omega \rightarrow \mathcal{I}$, where $\mathcal{I}$ is a countable index set, and an associated measurable partition $\mathcal{S} = \{S_i; i \in \mathcal{I}\}$ such that $\alpha(x) = i$ if $x \in S_i$,
- a decoder $\beta : \mathcal{I} \rightarrow \Omega$, and an associated reproduction codebook $\mathcal{C} = \{\beta(i); i \in \mathcal{I}\}$,
- an index coder $\psi : \mathcal{I} \rightarrow \{0, \dots, D-1\}^*$, the space of all finite-length D -ary strings, and the associated length $L : \mathcal{I} \rightarrow \{1, 2, \dots\}$ defined

by $L(i) = \text{length}(\psi(i))$. ψ is assumed to be uniquely decodable (a lossless or noiseless code)².

- The overall quantizer is

$$Q(x) = \beta(\alpha(x)). \quad (1.2)$$

Without loss of generality we assume that the codevectors $\beta(i)$; $i \in \mathcal{I}$ are all distinct. Since the index coder must be uniquely decodable, the Kraft inequality (e.g., [6]) requires that the codelengths $L(i)$ satisfy

$$\sum_i D^{-L(i)} \leq 1. \quad (1.3)$$

It is convenient to measure lengths in a normalized fashion and hence we define the length function of the code in nats as $\ell(i) = L(i) \ln D$ so that Kraft's inequality becomes

$$\sum_i e^{-\ell(i)} \leq 1. \quad (1.4)$$

A set of codelengths $\ell(i)$ is said to be *admissible* if (1.4) holds.

As in Cover and Thomas [6], it is convenient to remove the restriction of integer D -ary codelengths and hence we define any collection of nonnegative real numbers $\ell(i)$; $i \in \mathcal{I}$ to be an *admissible length function* if it satisfies (1.4). The primary reason for dropping the constraint is to provide a useful tool for proving results, but the general definition can be interpreted as an approximation since if $\ell(i)$ is an admissible length function, then for a code alphabet of size D the actual integer codelengths $L(i) = \lceil \frac{\ell(i)}{\ln D} \rceil$ will satisfy the Kraft inequality. Abbreviating $P_f(S_i)$ to p_i the average length (in nats) will satisfy

$$\sum_i p_i \ell(i) \leq (\ln D) \sum_i p_i L(i) < \sum_i p_i \ell(i) + \ln D.$$

If this is normalized by $1/k$, then the actual average length can be made arbitrarily close to the average length function by choosing a sufficiently large dimension. Let $\mathcal{A}$ denote the collection of all admissible length functions ℓ .

With a slight abuse of notation we will use the symbol Q to denote both the composite of encoder α and decoder β as in (1.2) and the complete quantizer comprising the triple (α, β, ℓ) . The meaning should be clear from context.

The instantaneous rate of a quantizer is defined by $r(\alpha(x)) = \ell(\alpha(x))$. The average rate is

$$R_f(Q) = R_f(\alpha, \ell) = Er(\alpha(X)) = \sum_i p_i \ell(i). \quad (1.5)$$

²This is stronger than invertibility since it requires that sequences be invertible.

Given a quantizer Q , the Shannon entropy of the quantizer is defined in the usual fashion by

$$H_f(Q) = - \sum_i p_i \ln p_i$$

and we assume that $p_i > 0$ for all i .

For any admissible length function ℓ the divergence inequality [6] implies that

$$\begin{aligned} R_f(Q) - H_f(Q) &= \sum_i p_i \ln \frac{p_i}{e^{-\ell(i)}} \\ &\geq \sum_i p_i \ln \frac{p_i}{e^{-\ell(i)} / \sum_j e^{-\ell(j)}} \\ &\geq 0 \end{aligned}$$

with equality if and only if $\ell(i) = -\ln p_i$. Thus in particular

$$H_f(Q) = \inf_{\ell \in \mathcal{A}} R_f(\alpha, \ell). \quad (1.6)$$

We assume a distortion measure $d(x, \hat{x}) \geq 0$ and measure performance by average distortion

$$\begin{aligned} D_f(Q) &= D_f(\alpha, \beta) \\ &= Ed(X, Q(X)) \\ &= Ed(X, \beta(\alpha(X))). \end{aligned}$$

For simplicity we assume squared error distortion with average

$$d(x, \hat{x}) = \|x - \hat{x}\|^2 = \sum_{i=0}^{k-1} |x_i - \hat{x}_i|^2,$$

where $x = (x_0, \dots, x_{k-1})$. 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 optimal performance is the minimum distortion achievable for a given rate $R \geq 0$, the *operational distortion-rate function*

$$\delta_f(R) = \inf_{Q: R_f(Q) \leq R} D_f(Q) \quad (1.7)$$

$$= \inf_{\alpha, \beta, \ell: R_f(\alpha, \ell) \leq R} D_f(\alpha, \beta). \quad (1.8)$$

In order to describe necessary conditions for optimality of a code and provide a clustering algorithm for the design of codes having these properties, it is convenient to use a Lagrangian formulation of variable rate vector quantization [4]. Define for each value of a Lagrangian multiplier $\lambda > 0$ a Lagrangian distortion

$$\begin{aligned} \rho_\lambda(x, i) &= d(x, \beta(i)) + \lambda r(\alpha(i)) \\ &= d(x, \beta(i)) + \lambda \ell(i) \end{aligned}$$

a corresponding performance

$$\begin{aligned}\rho(f, \lambda, Q) &= Ed(X, Q(X)) + \lambda E\ell(\alpha(X)) \\ &= D_f(Q) + \lambda R_f(Q),\end{aligned}$$

and an optimal performance

$$\rho(f, \lambda) = \inf_Q \rho(f, \lambda, Q)$$

where the infimum is over all quantizers $Q = (\alpha, \beta, \ell)$, $\ell \in \mathcal{A}$. The Lagrangian formulation yields necessary conditions for optimality that generalize Lloyd's original formulation [21] of optimal scalar quantization. (See [11] for simple proofs and a history of these conditions and their intimate connections with k -means and other similar clustering techniques.) Intuitively the conditions simply capsulize the fact that each of the three components of the quantizer be optimal for the other two. In particular, for a given decoder β and length function ℓ , the optimal encoder is

$$\alpha(x) = \operatorname{argmin}_i (d(x, \beta(i)) + \lambda \ell(i))$$

(ties are broken arbitrarily). The optimal decoder for a given encoder and length function is the usual Lloyd centroid:

$$\beta(i) = \operatorname{argmin}_y E(d(X, y) | \alpha(X) = i),$$

the conditional expectation in the case of squared error distortion. The optimal length function for the given encoder and decoder is, as we have seen, the negative log probabilities of the encoder output:

$$\ell(i) = -\ln p_i.$$

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.

1.3 High-rate quantization theory

High-rate quantization theory treats the asymptotic behavior of the rate-distortion tradeoff as the rate becomes large and the distortion becomes small. Here we follow the approach to high rate analysis developed by Gersho [7] (see also [11]), which is an intuitive derivation of the rigorous results of Zador [28] (see also [13]) using the quantizer point density ideas of Lloyd.

Define the volume $V(S)$ of a set S as the integral over S with respect to Lebesgue measure, $V(S) = \int_S dx$. Assume that there are asymptotically many quantization levels with vanishingly small cell volume and that the

density f is sufficiently well behaved. Assume also that $y_i \in S_i$, where as before $S_i = \{x : \alpha(x) = i\}$. In fact this is optimal for most common distortion measures. Under these assumptions, $f(x) \approx f(y_i)$; $x \in S_i$. From the mean value theorem of calculus

$$P_f(S_i) = \int_{S_i} f(x) dx \approx V(S_i) f(y_i)$$

and hence

$$f(y_i) \approx \frac{P_f(S_i)}{V(S_i)},$$

which implies that

$$D_f(Q) \approx \frac{1}{k} \sum_i P_f(S_i) \int_{S_i} \frac{\|x - y_i\|^2}{V(S_i)} dx.$$

For each i , let y_i be the Euclidean centroid of S_i and hence

$$\int_{S_i} \frac{\|x - y_i\|^2}{V(S_i)} dx$$

is the moment of inertia of the region S_i about its centroid.

It is convenient to use normalized moments of inertia so that they are invariant to scale: For any measurable set S define

$$M(S) = \frac{1}{kV(S)^{2/k}} \int_S \frac{\|x - y(S)\|^2}{V(S)} dx$$

where $y(S)$ denotes the centroid of S . If $c > 0$ and $cS = \{cx : x \in S\}$, then $M(S) = M(cS)$ and M depends only on shape and not upon scale. Thus

$$\begin{aligned} D &\approx \sum_i P_f(S_i) M(S_i) V(S_i)^{2/k} \\ &= \sum_i f(y_i) M(S_i) V(S_i)^{1+2/k} \end{aligned}$$

To be more precise, high rate quantization theory considers the behavior of a sequence of codes $\mathcal{C}_n$ or corresponding quantizers Q_n with increasing rate and decreasing distortion. In order to guarantee that distortion is decreasing, the reproduction vectors must become increasingly dense in the support set of the underlying pdf. We assume that as $n \rightarrow \infty$, the reproduction vectors $\mathcal{C}_n = \{y_{n,i}; i = 1, \dots, N\}$ have a smooth point density $\Lambda(x)$ in the sense that $\Lambda(x) \geq 0$ and

$$\# \text{ reproduction vectors of } \mathcal{C}_n \text{ in a set } S \rightarrow \int_S \Lambda(x) dx; \text{ all } S.$$

Note that a point density is like a mass density rather than a probability density since its integral need not be 1 (or even finite).

If we consider a tiny cell S containing a only a single reproduction vector, then from the mean value theorem

$$1 \approx \int_S \Lambda(x) dx \approx \Lambda(x)V(S); \quad \text{all } x \in S$$

and hence since $y_{n,i} \in S_{n,i}$,

$$V(S_{n,i}) \approx \frac{1}{\Lambda(y_{n,i})}.$$

Define for the sequence of codebooks $\mathcal{C}_n$; $n = 1, 2, \dots$ the function $B_i^{(n)}(x)$ as the cell of the codebook $\mathcal{C}_n$ which contains x . Assume that as $n \rightarrow \infty$ and $B_i^{(n)}(x) \rightarrow x$, then

$$M(B_i^{(n)}(x)) \rightarrow m(x),$$

the so-called *inertial profile* of the sequence of codebooks [22]. This function is assumed to be smooth and occasionally it can be evaluated and usually bound above and below. We then have for large n that

$$\begin{aligned} D_f(Q_n) &\approx \sum_i f(y_{n,i}) M(S_{n,i}) V(S_{n,i})^{1+2/k} \\ &\approx \sum_i f(y_{n,i}) m(y_{n,i}) \frac{V(S_{n,i})}{\Lambda(y_{n,i})^{2/k}} \\ &\approx \int f(x) \frac{m(x)}{\Lambda(x)^{2/k}} dx = E\left[\frac{m(X)}{\Lambda(X)^{2/k}}\right] \end{aligned}$$

Gersho's conjecture [7] can be described as follows [11]. If $f(x)$ is smooth and the allowed rate $R_n = R_f(Q_n)$ of (1.5) is large, then, regardless of Λ , the optimal quantizer Q_n has cells $S_{n,i}$ that are (approximately) scaled, rotated, and translated copies of S^* , the convex polytope that tessellates $\mathcal{R}^k$ with minimum normalized moments of inertia $M(S)$, i.e.,

$$m^*(x) = c_k = \min_S M(S) = M(S^*).$$

In this case the average distortion becomes

$$D_f(Q_n) \approx c_k E_f\left(\left(\frac{1}{\Lambda(X)}\right)^{2/k}\right). \quad (1.9)$$

The minimum average rate of the quantizer given the encoder and decoder is given by the entropy of the quantizer:

$$E_f \ell(Q_n(X)) = H_f(Q_n) = - \sum_i P_f(S_{n,i}) \ln P_f(S_{n,i}).$$

and the approximations then imply that

$$P_f(S_{n,i}) \approx f(y_{n,i}) / (\Lambda(y_{n,i})) = f(y_{n,i}) V(S_{n,i})$$

so that

$$\begin{aligned}
E_f \ell(\alpha(X)) &= H_f(Q_n) \\
&\approx - \sum_i \frac{f(y_{n,i})}{\Lambda(y_{n,i})} \ln \frac{f(y_{n,i})}{\Lambda(y_{n,i})} \\
&= - \sum_i V(S_{n,i}) f(y_{n,i}) \ln f(y_{n,i}) + \sum_i V(S_{n,i}) f(y_{n,i}) \ln(\Lambda(y_{n,i})) \\
&\approx - \int dy f(y) \ln f(y) + \int dy f(y) \ln(\Lambda(y)) \\
&\approx h(f) - E(\ln \frac{1}{\Lambda(X)}), \tag{1.10}
\end{aligned}$$

where $h(f)$ is the differential entropy. Thus for large N

$$\begin{aligned}
H_f(Q_n) &\approx h(f) - E_f(\ln \frac{1}{\Lambda(X)}) \\
&= h(f) - \frac{k}{2} E(\ln(\frac{1}{\Lambda(X)})^{2/k}). \tag{1.11}
\end{aligned}$$

From Jensen's inequality,

$$\begin{aligned}
H_f(Q_n) &\geq h(f) - \frac{k}{2} \ln E_f((\frac{1}{\Lambda(X)})^{2/k}) \\
&\approx h(f) - \frac{k}{2} \ln c_k D_f(Q_n) \tag{1.12}
\end{aligned}$$

with equality iff $\Lambda(X)$ is a constant, in which case from (1.9) and (1.12) the constant is given by

$$\Lambda(x) = (\frac{c_k}{D_f(Q_n)})^{k/2} = e^{(H_f(Q_n) - h(f))}. \tag{1.13}$$

Thus for large n using the optimal point density

$$D_f(Q_n) \approx c_k e^{\frac{2}{k}(h(f) - H_f(Q_n))} \tag{1.14}$$

The optimal performance is then

$$\delta_f(R) = \inf_{Q: H_f(Q) \leq R} D_f(Q) \approx c_k e^{\frac{2}{k}(h(f) - R)}. \tag{1.15}$$

This approximate argument produces results that agree with the rigorous version [28, 13]. In particular, under suitable technical conditions the asymptotic performance is characterized by

$$\lim_{R \rightarrow \infty} 2^{\frac{2}{k}R} \delta_f(R) = b_k 2^{\frac{2}{k}h(f)} \tag{1.16}$$

where b_k is Zador's constant, which depends only on k and not f . If Gersho's conjecture is true, this identifies b_k with c_k .

The Lagrangian form follows a similar argument. Define

$$\rho(f, \lambda) = \inf_{\alpha, \beta, \ell} (E_f \|X - \beta(\alpha(X))\|^2 + \lambda E_f \ell(\alpha(X)))$$

$$= \inf_{\alpha, \beta} (E_f \|X - \beta(\alpha(X))\|^2 + \lambda H_f(Q_n))$$

and consider the behavior as $\lambda \rightarrow 0$. Since the limit corresponds to placing less cost on the entropy and more on the distortion, one would expect the entropy to grow and the distortion to shrink, which will happen under the same asymptotics as in the non-Lagrangian argument, that is, the rate will grow while the distortion will shrink so that the high rate approximations previously used will still hold, yielding for small λ the approximations

$$\begin{aligned} & \frac{D_f(Q)}{\lambda} + H_f(Q) + \frac{k}{2} \ln \lambda \\ & \approx \frac{c_k E_f((\frac{1}{\Lambda(X)})^{2/k})}{\lambda} + h(f) - \frac{k}{2} \ln E_f(\ln(\frac{1}{\lambda \Lambda(X)})^{2/k}) \\ & = \frac{k}{2} \left(E_f\left(\left(\frac{2c_k}{k\lambda\Lambda(X)}\right)^{2/k}\right) - \ln E_f\left(\ln\left(\frac{2c_k}{k\lambda\Lambda(X)^{2/k}}\right) - 1\right) \right) + h(f) + \frac{k}{2} \ln \frac{2ec_k}{k} \\ & \geq h(f) + \frac{k}{2} \ln \frac{2ec_k}{k} \end{aligned}$$

with equality iff

$$\Lambda(x) = \left(\frac{2c_k}{k\lambda}\right)^{k/2}, \text{ all } x, \quad (1.17)$$

a constant point density function. The rigorous version of this result is proved in [13] and shows that under suitable conditions (including the assumptions that P_f is absolutely continuous with respect to Lebesgue measure, $h(f)$ is finite, and $H_f(Q)$ is finite when Q corresponds to a partition of Euclidean space into cubes)

$$\lim_{\lambda \rightarrow 0} \left(\frac{\rho(f, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right) = \frac{k}{2} \ln \frac{2eb_k}{k} + h(f) \quad (1.18)$$

where

$$\frac{k}{2} \ln \frac{2eb_k}{k} = \theta_k \triangleq \inf_{\lambda > 0} \left(\frac{\rho(u_1, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right), \quad (1.19)$$

where u_1 is the uniform probability density on a k -dimensional unit cube.

A consequence of the result is that given a sequence λ_n decreasing to 0, there is a sequence of quantizers Q_n for which

$$\lim_{n \rightarrow \infty} \left(\frac{D_f(Q_n)}{\lambda} + E_f \ell_n(Q_n) + \frac{k}{2} \ln \lambda_n \right) = h(f) + \frac{k}{2} \ln \frac{2ec_k}{k}, \quad (1.20)$$

where ℓ is the optimal length function for the encoder/decoder (α_n, β_n) . Such a sequence of quantizers is said to be *asymptotically optimal* for the density f .

1.4 The Gaussian Case

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:

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

The Lagrangian formulation yields the same conclusion since from (1.20) the asymptotic performance depends on the source only through the differential entropy: the worst possible Lagrangian performance for a high rate quantizer applied to a source with known second order moments is achieved by the Gaussian source.

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 have 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 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 [25] Given an index set $\overline{N}$ and a partial covariance $\Sigma_{\overline{N}} = \{\Sigma_{i,j}; (i,j) \in \overline{N}\}$, find $\max_{K: K_{\overline{N}} = \Sigma_{\overline{N}}} |K|$. The K achieving the maximum (if it exists) is the MAXDET extension of $\Sigma_{\overline{N}}$.

1.5 Quantizer Mismatch

Suppose now that Q is optimized for a Gaussian g , but applied to another source f . To quantify the performance change resulting from such a mismatch, we define the *relative entropy* (or Kullback-Leibler number or Kullback-Leibler I-divergence or directed divergence or discrimination). The reader is referred to [19, 23, 2, 3, 9] for thorough treatments of relative entropy and its properties. Given two pdf's f and g with the property that the induced measures P_f and P_g are such that P_f is absolutely continuous with respect to P_g , then following Csiszar's notation [2, 3] define

$$I(f||g) = \int dx f(x) \ln \frac{f(x)}{g(x)}.$$

Consider the asymptotic approximations for a sequence of quantizers Q_n and assume that n is large. The distortion derivation mimics that of (1.9)

and the point density is the optimal point density for g as given by (1.13) with g in place of f :

$$D_f(Q_n) \approx c_k E_f \left(\left(\frac{1}{\Lambda(X)} \right)^{2/k} \right) \approx D_g(Q_n). \quad (1.21)$$

The average rate is found analogous to (1.10), but now the pdf being averaged over f differs from that used for the optimal length function, g :

$$\begin{aligned} E_f \ell(\alpha(X)) &\approx - \sum_i \frac{f(y_{n,i})}{\Lambda(y_{n,i})} \ln \frac{g(y_{n,i})}{\Lambda(y_{n,i})} \\ &= - \sum_i V(S_{n,i}) f(y_{n,i}) \ln g(y_{n,i}) + \sum_i V(S_{n,i}) f(y_{n,i}) \ln(\Lambda(y_{n,i})) \\ &\approx - \int dy f(y) \ln g(y) + \int dy f(y) \ln(\Lambda(y)) \\ &\approx H_g(Q_n) - h(g) - \int dy f(y) \ln g(y) \end{aligned} \quad (1.22)$$

When g is Gaussian this becomes

$$\begin{aligned} R_f(Q_n) - R &\approx \frac{1}{2} \text{Trace}(K_g^{-1} K_f) \\ &\quad + \frac{1}{2} (\mu_g - \mu_f)^t K_g^{-1} (\mu_g - \mu_f) - \frac{k}{2}. \end{aligned}$$

If $\mu_f = \mu_g$ and $K_f = K_g$, then $D_f(Q_n) \approx D_g(Q_n)$ and $R_f(Q_n) \approx R$, reminiscent of Lapidoth's fixed rate result for iid Gaussian processes: the performance predicted for the Gaussian is actually achieved for the non-Gaussian. The loss in performance here is that in theory one could do better if the code had been designed for f and not for g .

The mismatch result can be translated into the Lagrangian formulation, where the result provides particular insight for the applications to come. The basic high rate result of (1.20) implies that given a Gaussian source described by a pdf g there exists a sequence of Lagrange multipliers λ_n decreasing to 0 and a corresponding asymptotically optimal sequence of codes $(\alpha_n, \beta_n, \ell_n)$ such that

$$\lim_{n \rightarrow \infty} \left(\frac{D_g(\alpha_n, \beta_n)}{\lambda_n} + E_g \ell(\alpha_n) + \frac{k}{2} \ln \lambda_n \right) = \frac{k}{2} \ln \frac{2e}{k} b_k + h(g), \quad (1.23)$$

the optimal value for g . Using the distortion approximation of (1.21) and the rate approximation of (1.22) with the quantizer point density function Λ_n defined by (1.17) we have for high rate that

$$\begin{aligned} &\frac{D_f(Q_n)}{\lambda_n} + E_f(\ell_n) + \frac{k}{2} \ln \lambda_n \\ &\approx \frac{k}{2} \left(\frac{E_f \left((c_k \frac{1}{\Lambda(X)})^{2/k} \right)}{\lambda_n} \ln E_f \left(\ln \left(\frac{2ec_k}{k\lambda\Lambda(X)^{2/k}} \right) \right) - \int dy f(y) \ln g(y) \right) \end{aligned}$$

$$= - \int dy f(y) \ln g(y) + \frac{k}{2} \ln\left(\frac{2ec_k}{k}\right) \quad (1.24)$$

Thus applying the asymptotically optimal codes designed for g to the source f results in

$$\begin{aligned} & \lim_{n \rightarrow \infty} \left(\frac{D_f(Q_n)}{\lambda_n} + E_f \ell(\alpha_n) + \frac{k}{2} \ln \lambda_n \right) \\ &= - \int dy f(y) \ln g(y) + \frac{k}{2} \ln\left(\frac{2ec_k}{k}\right) \\ &= h(f) + I(f||g) + \frac{k}{2} \ln\left(\frac{2ec_k}{k}\right). \end{aligned}$$

Thus

$$\begin{aligned} & \lim_{n \rightarrow \infty} \left(\frac{D_f(Q_n)}{\lambda_n} + E_f \ell(\alpha_n) + \frac{k}{2} \ln \lambda_n \right) \\ &= \lim_{\lambda \rightarrow \infty} \inf_{\alpha, \beta, \ell} \left(\frac{D_f(\alpha, \beta)}{\lambda} + E_f \ell(\alpha) + \frac{k}{2} \ln \lambda \right) + I(f||g). \quad (1.25) \end{aligned}$$

In words, applying a sequence of codes which are asymptotically optimal for g to a mismatched source f yields a loss of performance from the optimal asymptotic performance for f of $I(f||g)$. Thus $I(f||g)$ can be viewed as a measure of the mismatch of two sources with respect to quantization or lossy compression performance. This is analogous to the similar role of discrete relative entropy for quantifying the mismatch in lossless coding. This fact will be used shortly to provide a distortion measure or cost function for clustering models or density functions. The reader is referred to [14] for a more thorough and precise development of the mismatch result.

1.6 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 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 in [16]. The method is simply an extension of the speech case to multiple dimensional sources such as images. As earlier discussed, we fit the mixture to the data by trying to map individual Gaussian components to each observed vector using a measure of the distortion or badness of approximation. This allows the use of Lloyd clustering ideas instead of the usual EM fitting. A philosophical difference is that we here

admit the Gaussian models being fitted are just that, models. We are not claiming the actual data is itself Gaussian and we admit we do not know the true underlying densities, hence we cannot claim to using an ML method. Instead, as in speech, we argue that local Gaussian behavior provides a local worst case fit, and a distortion measure will be used to quantify the quality of that fit. In the case of speech, using Gaussian models to synthesize reproduced speech or to recognize words or phrases in speech can yield synthetic speech that sounds good and the best known recognizers. As a result, we can admit that in fact actual speech is highly nonGaussian, yet provide good signal processing by fitting Gaussian components to chunks of speech, providing overall an implicit Gauss mixture model.

Since each Gaussian model is described by its mean and covariance matrix, say (μ_l, K_l) for the l th 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 $\hat{\mu}_x$ and $\hat{K}_x$. This effectively assumes that it is the second order characteristics which are important. Typically $\hat{\mu}_x$ is assumed to be a constant vector, as in the speech case where it is assumed to be 0. When a 0 mean assumption is not appropriate (as in untransformed image pixels), the mean can be estimated by a sample average. The covariance might be estimated, for example, by assuming that the underlying process is locally spatially stationary and computing scalar sample averages

$$\hat{K}_{x,m}(n) = \frac{\sum_{i,j:|i-j|=n} (x_i - m)(x_j - m)}{N(n)}; \quad n \in \mathcal{Z}_N^2 \quad (1.26)$$

where, e.g., one might choose $N(n) = \#\{i, j : |i-j| = n\}$. Choosing $m = \hat{\mu}_x$ 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 pdf’s with the known second order moments). This is an example of the MAXDET algorithm [25].

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 $I(\hat{f}||g_l) = \int \hat{f} \ln \hat{f}/g_l$. 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[19, 16] This follows

in exactly the same manner as the proof that the Gaussian pdf yields the maximum differential entropy for a given covariance as proved, e.g., in [6]. Thus

$$\begin{aligned} d_{MDI}(x, (\mu_l, K_l)) &= I(\hat{f}||g_l) \\ &= \frac{1}{2} \left[\ln \frac{|K_l|}{|\hat{K}_x|} + \text{Tr}(\hat{K}_x K_l^{-1}) + (\hat{\mu}_x - \mu_l)^t K_l^{-1} (\hat{\mu}_x - \mu_l) - k \right]. \end{aligned}$$

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

$$\begin{aligned} d_{MDI}(x, (\mu_l, K_l)) &= \frac{1}{2} \left(\ln \frac{|K_l|}{|\hat{K}|} + \sum_{i,j \in \mathcal{Z}_N^2} \hat{K}_x(i,j) K_l^{-1}(i,j) + \right. \\ &\quad \left. \sum_{i,j \in \mathcal{Z}_N^2} (\hat{\mu}_x(i) - \mu_l(i)) (\hat{\mu}_x(j) - \mu_l(j)) K_l^{-1}(i,j) - k \right) \\ &= \frac{1}{2} \left(\ln \frac{|K_l|}{|\hat{K}_x|} + \sum_{i,j \in \mathcal{Z}_N^2} K_l^{-1}(i,j) [\hat{K}_x(i,j) + (\hat{\mu}_x(i) - \mu_l(i)) (\hat{\mu}_x(j) - \mu_l(j))] - k \right) \\ &= \frac{1}{2} \left[\ln \frac{|K_l|}{|\hat{K}_x|} + \sum_{i,j \in \mathcal{Z}_N^2} K_l^{-1}(i,j) \hat{K}_{x,\mu_l}(i,j) - k \right] \end{aligned}$$

Itakura and Saito [18] originally derived their “error matching measure” by an approximate maximum likelihood argument. A similar informal argument can be used here to argue that minimizing the above MDI distortion measure is approximately equivalent to a maximum likelihood selection of which Gaussian component is in effect based on observed data.

1.7 MDI Centroids

As in the analogous speech case [16], the MDI 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 μ_l and K_l must minimize the conditional expected distortion.

$$\begin{aligned} E[d_{MDI}(X, g_l) \mid \alpha(X) = l] &= \frac{1}{2} E \left[\ln \frac{|K_l|}{|\hat{K}_X|} + \text{Tr}(\hat{K}_X K_l^{-1}) \right. \\ &\quad \left. + (\hat{\mu}_X - \mu_l)^t K_l^{-1} (\hat{\mu}_X - \mu_l) - k \mid \alpha(X) = l \right] \end{aligned}$$

where $\hat{\mu}_X$ and $\hat{K}_X$ are the mean and the covariance estimates for observation X . The mean centroids are given by $\mu_l = E[\hat{\mu}_X \mid \alpha(X) = l]$ 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 μ_l we need K_l to minimize

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

where $\bar{K}_l = E[(X - \mu_l)(X - \mu_l)^t]$ since the first three terms are just the relative entropy between two Gaussian distributions with the given covariances and 0 means, and this is nonnegative from the divergence inequality (see, e.g., [6]). Equality holds if $K_l = \bar{K}_l$.

1.8 MDI Clustering

Application of the Lloyd algorithm to the MDI distortion measures yields a Gaussian model VQ, a mapping of input vectors X (e.g., image blocks) into a Gaussian model chosen from a codebook of such models. The collection of models together with the probability of their occurrence when the quantizer is used on the training data provides a complete Gauss mixture model. Under reasonably general conditions, the Lloyd algorithm converges. If the algorithm converges to a stationary point, the centroid formulas yield the resulting MDI distortion in terms of the model covariances of the codebook and their probabilities of occurrence, 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: $d_{\text{ECMDI}}(x, g_l) = d_{\text{MDI}}(x, g_l) - \mu \ln w_l$. Applying the MDI centroid expression provides a simple formula for the average ECMDI distortion:

$$D_{\text{ECMDI}} = \frac{1}{2} \sum_l w_l \ln |K_l| - E[\ln |\hat{K}_X|] + \mu H(p) \quad (1.27)$$

In Lloyd clustering with variable rate codes, standard practice is to begin with a large number of codewords. As the algorithm runs it will eliminate unused codewords, so that the total is monotonically decreasing. If the algorithm is initiated with too few codewords, then the number will usually not reduce with successive iterations and the usual practice is to restart the algorithm with a larger number. This practice is reinforced by theoretical results to the effect that under suitable conditions the optimal codes will have a finite number of reproduction vectors. “Suitable conditions” in-

clude the case of squared error distortion and pdf's with fine support and Gaussian pdf's [5]

1.9 Gauss Mixture VQ

We now consider an application where the MDI clustering approach to Gauss mixture modeling arises naturally. Suppose that X has mixture pdf f generated by classifying X into classes $l = 1, 2, \dots$. For the moment the classifier is arbitrary, but obviously the clustering algorithm of the previous section could be used. Let $L = L(X)$ denote the integer-valued class. Then $\{f_{X|L}(x|l), w_l = \Pr(L = l)\}$ is a mixture model for f . A separate VQ can then be designed for each class, i.e., for each Gaussian source with a specified mean and covariance, yielding a classified VQ structure. For each class l one can estimate a conditional mean μ_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 (μ_l, K_l) , then quantize using optimal quantizer Q_l for g_l .

The idea for classified vector quantization dates back to switched quantizers and was extended to vector quantization for image coding by Gersho and Ramamurthi [8] in 1982. The approach has been highly successful for image coding applications by adapting the quantization or compression algorithms to local statistical behavior within an image. For a modern survey of such methods incorporating wavelet transforms and scalar quantizers, the reader is referred to Yoo, Ortega, and Yu [27]. Our treatment here, however, emphasizes arbitrary vector sources and vector quantizers and the theoretical performance when the bit rate is large.

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

$$\bar{h} = \sum_l h_l w_l = h(X|L) \quad , \quad H(p) = H(L)$$

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

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

$$D_{\text{MSE}} = b_k (2\pi e) e^{\frac{2}{k}(\frac{1}{2} \sum_l w_l \ln |K_l| + H(p) - R)} \quad (1.28)$$

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

If one designs a classified VQ using the MDI as described earlier, 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 (assuming high rate and optimal bit allocation across the classes) is

$$D_{\text{MSE}} = b_k (2\pi e) e^{\frac{2}{k}(D_{\text{ECMDI}} + E[\ln |\hat{K}_X|])} \quad (1.29)$$

where b_k is a constant depending only on the dimension and not on the underlying pdf's and the MDI Lagrangian is chosen as $\mu = 1$. 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. The relation shows that designing a Gauss mixture model using the MDI clustering algorithm described is equivalent to designing a robust classified vector quantizer with the overall goal of minimizing the reproduction mean squared error. The rightmost term above does not depend on the code, only on the estimation technique used to estimate the covariance.

Preliminary simulation results on various image databases may be found in [1, 15] and work is proceeding on applications to image classification and segmentation.

1.10 Parting Thoughts

The approach here described provides a hindsight interpretation as to why modern code excited linear prediction (CELP) techniques work as well as they do. They can be viewed as fitting a reproduction to an observed waveform by first selecting a Gaussian AR model (by LPC analysis and simple quantization), then populating a custom codebook for this model by driving the implied inverse filter by a codebook of vectors randomly generated from a memoryless Gaussian process, and then finding the closest resulting vector in the resulting output codebook to the original input vector. High rate quantization theory implies that driving the AR filters with independent inputs produces an approximately optimal codebook for

the Gaussian AR source and hence a robust codebook for all sources with approximately the same second order moments. The design of the AR filters is a form of MDI clustering, and we have seen that this indeed results in an approximately optimal classified quantizer.

The ideas here suggest that similar methods will work for image coding and image segmentation, and simulations are underway to study these possibilities. Eventually it would be of interest to compare the MDI clustering approach to the traditional EM algorithm approach to the design of Gauss mixtures. An application where the two techniques could be compared in terms of overall performance is that of a vector quantization based system since Hedelin and Skoglund [17] used EM to design a vector quantizer based on Gauss mixture models. For the applications described here, however, it is the MDI approach that is the most natural: if a classifier or coder is to select a model using the minimum MDI distortion rule, then it is natural to design the codebook of models so as to minimize the average MDI distortion, i.e., to use a clustering algorithm.

On the theoretical side, the development of this paper rests on Gersho's conjecture, which is widely believed but yet unproved. The basic high rate quantization results have been rigorously proved by completely different techniques which do not involve the concept of a quantizer point density function [13]. Similar techniques have been applied to recently provide a rigorous proof of the mismatch result, but this result has not yet been validated by by scrutiny of colleagues and reviewers. A preprint of the proof of the basic mismatch result and its applications to developing the heuristic results developed here may be found in [14].

We close with a few thoughts on further possible extensions. If in addition to estimating w_l of mixture components from relative frequencies in order to construct Gauss mixture models, one could also estimate conditional probabilities for the index of a Gauss component given the previous index, providing a hidden Markov model (HMM) without resort to EM or forward-backward algorithms. The basic results described here all require an assumption of high rate, so a natural question is whether the approximations remain valid for moderate or low rates. There is anecdotal evidence to the effect that as with many high rate results, these results often hold approximately in low rate situations.

The use of clustered models suggests an approach to some simple image segmentation problems that mimics a simple and powerful early approach to simple speech recognition, the recognition of isolated utterances such as the alphabet and numbers. Given a labeled training set, design a separate quantizer for each class using a classified VQ, i.e., each class has a Gauss mixture model associated with it, not just a single Gaussian component. Classify new data by encoding it using all of the quantizers in parallel and select the one with the smallest overall distortion. This might provide reasonable performance, for example, in texture recognition applications.

This technique is currently being tested for content-addressable browsing and for image segmentation applications.

Acknowledgements

The author would like to acknowledge the helpful comments and suggestions on the manuscript by Ms. Maya Gupta and by an anonymous referee.

References

- [1] *Robust Image Compression using Gauss Mixture Models*, PhD Dissertation, Department of Electrical Engineering, Stanford University, June 2001.
- [2] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observations,” *Studia Scientiarum Mathematicarum Hungarica*, Vol. 2, pp. 299–318,, 1967.
- [3] I. Csiszár, “On an extremum problem of information theory,” *Studia Scientiarum Mathematicarum Hungarica*, pp. 57–70, 1974.
- [4] P.A. Chou, T. Lookabaugh, and R.M. Gray, “Entropy-constrained vector quantization,” *IEEE Trans. Acoust. Speech and Signal Proc.*, Vol. 37, pp. 31–42, January 1989.
- [5] Philip A. Chou and B. Betts, “When optimal entropy-constrained quantizers have only a finite number of codewords,” IEEE International Symposium on Information Theory. Full paper available at <http://research.microsoft.com/pachou/publications.htm>.
- [6] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.
- [7] A. Gersho, “Asymptotically optimal block quantization,” *IEEE Trans. Inform. Theory.*, vol. 25, pp. 373–380, July 1979.
- [8] A. Gersho and B. Ramamurthi, “Image coding using vector quantization,” *Proc. Intl. Conf. on Acoust. Speech, and Signal Processing*, vol. 1, pp. 428–431, Paris, April 1982.
- [9] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, 1990.
- [10] R.M. Gray, “Gauss Mixture Vector Quantization,” *Proceedings ICASSP 2001*.
- [11] R.M. Gray and D.L. Neuhoff, “Quantization,” *IEEE Transactions on Information Theory*, Vol. 44, pp. 2325–2384, October 1998.
- [12] R.M. Gray and J. Li, “On Zador’s Entropy-Constrained Quantization Theorem,” *Proceedings Data Compression Conference 2001*, IEEE Computer Science Press, Los Alimitos, pp. 3–12, March 2001.
- [13] R.M. Gray, T. Linder, and J. Li, “A Lagrangian Formulation of Zador’s Entropy-Constrained Quantization Theorem,” *IEEE Transactions on Information Theory*, pp. 695–707, Vol. 48, Number 3, March 2002.
- [14] R.M. Gray and T. Linder, “Mismatch in high rate entropy constrained vector quantization,” preprint available at <http://ee.stanford.edu/~gray/mismatch.pdf>.
- [15] R.M. Gray, J.C. Young, and A. K. Ayer, “Minimum discrimination information clustering: modeling and quantization with Gauss mixtures,” *Proceedings 2001 IEEE International Conference on Image Processing*, Thessaloniki, Greece, October 2001.

- [16] 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.
- [17] P. Hedelin and J. Skoglund, "Vector Quantization based on Gaussian mixture models," *IEEE Transactions on Speech and Audio Processing*, Vol.8, no.4, p.385–401, July 2000.
- [18] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Proc. 6th Int'l Congress of Acoustics*, pp. C-17–C-20, Tokyo, Japan, August 1968.
- [19] S. Kullback. *Information Theory and Statistics*, Dover, New York, 1968. (Reprint of 1959 edition published by Wiley.)
- [20] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.
- [21] S. P. Lloyd. Least squares quantization in PCM. Unpublished Bell Laboratories Technical Note. Portions presented at the Institute of Mathematical Statistics Meeting Atlantic City New Jersey September 1957. Published in special issue on quantization, *IEEE Trans. Inform. Theory*, vol. 28, pp. 129–137, Mar. 1982.
- [22] S. Na and D. L. Neuhoff, "Bennett's integral for vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, pp. 886–900, July 1995.
- [23] M. S. Pinsker, *Information and information stability of random variables and processes*, Holden Day, San Francisco, 1964. (Translated by A. Feinstein from the Russian edition published in 1960 by Izd. Akad. Nauk. SSSR.)
- [24] D. J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. 21, pp. 301–309, May 1975.
- [25] 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.
- [26] Yuhong Yang and Andrew Barron "An Asymptotic Property of Model Selection Criteria," *IEEE Transaction on Information Theory*, vol. 44, pp. 95–116, 1998.
- [27] Y. Yoo, A. Ortega, and B. Yu, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Transactions on Image Processing*, Vol. 8, pp. 1702–1715, December 1999.
- [28] P. L. Zador, "Topics in the asymptotic quantization of continuous random variables," Bell Laboratories Technical Memorandum, 1966.