Quantization with an Information-Theoretic Distortion Measure

Jean Cardinal

October 23, 2002

Abstract

We consider the following problem: given two nonindependent random variables, find a quantizer for the first variable such that the mutual information between the quantizer output and the second variable is maximized. This problem can be seen as a quantization problem with an information-theoretic distortion measure. It arises in numerous research areas such as classical data compression, neural information processing and cryptography. We propose a survey of these applications and analyze a general local optimization algorithm based on known methods for vector quantizer design. This algorithm is compared to agglomerative methods.

1 Introduction

Let $X_A \in \mathcal{X}_A$ and $X_B \in \mathcal{X}_B$ be two nonindependent random variables. We denote $I(X_A; X_B) = H(X_A) - H(X_A \mid X_B)$ the mutual information between X_A and X_B, where $H(X)$ is the entropy of the random variable X. The mutual information is the amount of information "shared" by the two variables. We wish to quantize X_A using a quantizer mapping $\alpha : \mathcal{X}_A \to \{1, 2, \ldots, N\}$ so that the mutual information $I(\alpha(X_A); X_B)$ is as high as possible. In other words, we want to classify the outcomes of X_A so that the classes summarize the features of X_A that are relevant to X_B. The mutual information $I(\alpha(X_A); X_B)$ is upper-bounded by $I(X_A; X_B)$. The upper bound can be achieved at the high-resolution quantization limit, that is, for an arbitrarily high number of classes. We denote $K = \alpha(X_A)$ the quantizer output.

This problem is ubiquitous in many different research areas, and our first section summarizes the occurences of this problem known to the author. Next, we show that the problem is actually a vector quantization problem with the Kullback-Leibler divergence as a distortion measure. We propose a variant of the generalized Lloyd algorithm, similar to the well-known K-means clustering algorithm. A property of locally optimal quantizers is given, leading to a concise formula for the contribution of a quantization cell in the average distortion. We also describe an entropy-constrained version of the method that maximizes the mutual information with a constraint on the output entropy of the quantizer. This happens to be the goal of a previously published

method called agglomerative information bottleneck [11]. We propose a variation in this agglomerative algorithm that takes into account the change in output entropy and compare the various methods on small and medium scale data sets. As a conclusion, we mention the use of Gaussian modeling for handling continuous random variables with Gaussian behaviours.

In general, if X_A is continuous, α should be called a quantizer, while if X_A takes value on a finite domain, it should rather be called a classifier or a clustering. Similarly, quantization cells could be called clusters and quantization index could be referred to as classes or cluster labels. In the following we decided to use the quantization terminology, keeping in mind that X_A is not necessarily continuous.

2 Applications and Previous Works

Traditional quantization aims at minimizing a distortion measure defined in the signal space, such as the mean squared error [6]. There is however already some literature on quantization for maximal mutual information. This idea has actually emerged recently in rather different contexts we now detail.

2.1 Context quantization

In a recent contribution from Wu and Chou [13], a maximal mutual information quantizer is utilized to classify context vectors in data compression applications. We wish to predict the distribution of a variable X_B given a context X_A. Since the range of X_A is potentially very large, or even continuous, there is a need in quantizing X_A to a limited number of context classes. Wu and Chou describe how to map X_A optimally onto the set $\{1, 2, \ldots, N\}$ with a quantizer α so that the conditional entropy $H(X_B \mid \alpha(X_A))$ is minimal. This is strictly equivalent to maximizing the mutual information $I(\alpha(X_A); X_B)$. In particular, they exhibit an exact polynomial algorithm for the binary case $\mathcal{X}_B = \{0, 1\}$. They also mention the Lloyd-like approach we develop here.

Note that the implementation of this idea in a practical data compression system is not straightforward, since optimal quantization, as we will see in the next section, requires the knowledge of the probability distribution of X_B given X_A, which is not available, otherwise we could use it directly to encode X_B.

2.2 Information bottleneck method

In the so-called *information bottleneck method* from Tishby, Pereira and Bialek [12], a stochastic map of the form $p(k \mid x_A)$ plays the role of the quantizer α, and minimizes $I(K; X_B)$ subject to a constraint on the value of $I(K; X_A)$. Tishby et al. describe an algorithm for computing these maps based on classical developments in rate-distortion theory and similar to the Blahut-Arimoto algorithm [3]. In a subsequent development of the method [11], they describe a greedy heuristic algorithm for designing "hard clusters", i.e. a deterministic quantizer, minimizing the same criteria. The algorithm

is based on iterative merging of quantization cells, choosing at each step the merging that leads to the lowest decrease in mutual information. It is shown to perform efficiently on document clustering tasks. Note that in that case the constraint on $I(K; X_A)$ reduces to a constraint on $H(K)$, since $H(K \mid X_A) = 0$.

2.3 Neural computation

Let us also notice several recent contributions in the neural computation community based on the same ideas, such as [10, 4].

In [10], a soft clustering method is presented that minimizes the Kullback-Leibler divergence between estimated distributions of auxiliary data, what we call X_B, conditioned on primary data, here called X_A. As shown in the paper and in the next section, this turns out to be equivalent to maximizing the mutual information between the categories and the auxiliary data X_B. They propose a Hebb-type online learning algorithm and relate this method to competitive learning and the information bottleneck method. The algorithm is illustrated on a gene expression data clustering task. In [4], a maximum mutual information method is presented to discover relevant features in stimuli of the cricket cercal sensory system using a joint relationship between stimuli and neural responses. The approach is strongly similar to the information bottleneck method and makes use of a maximum entropy approach borrowed from rate-distortion theory. It is also the only paper, apart from [13], that uses the word "quantization". In both cases, no entropy constraint was used, and the Lloyd algorithm is not mentioned.

2.4 Cryptography

Suppose that two parties, say Alice and Bob, have access to nonindependent random variables X_A and X_B, respectively. They wish to agree on a secret digital key extracted from these data. A simple way to proceed is as follows: Alice constructs a key $K = \alpha(X_A)$, and sends a message to Bob, on a public channel, so that he can reconstruct K using this information and his own data X_B. The information leaked to a potential eavesdropper is equal to the size of Alice's message, which is lower bounded by $H(K \mid X_B)$. We wish to maximize the quantity of information shared by Alice and Bob, taking into account the lower bound on the leaked information. This amounts to maximizing $H(K) - H(K \mid X_B) = I(K; X_B)$.

This idea has recently been applied to a key distillation procedure using continuous quantum states [1].

3 Algorithm

We propose a method that follows the developments provided in [13] and inspired from the Lloyd optimality conditions for vector quantizers [6]. We assume that K belongs to the set $\{1, 2, \ldots, N\}$.

In order to avoid distinguishing continuous and discrete random variables, we use the following synthetic notation:

$$\langle f, g \rangle = \begin{cases} \int f(x)g(x)dx & \text{in the continuous case,} \\ \sum_x f(x)g(x) & \text{in the discrete case.} \end{cases}$$

We also denote by P_X the probability distribution of the random variable X.

α is a solution of

$$\begin{aligned}
& \arg\max_{\alpha} I(K; X_B) && (1)\\
= \; & \arg\max_{\alpha} H(X_B) - H(X_B \mid K) && (2)\\
= \; & \arg\min_{\alpha} H(X_B \mid K) && (3)\\
= \; & \arg\min_{\alpha} H(X_B \mid K) - H(X_B \mid X_A) && (4)\\
= \; & \arg\min_{\alpha} \; -\langle P_{X_A}, \langle P_{X_B|X_A}, \log P_{X_B|K} \rangle \rangle + \langle P_{X_A}, \langle P_{X_B|X_A}, \log P_{X_B|X_A} \rangle \rangle && (5)\\
= \; & \arg\min_{\alpha} \langle P_{X_A}, \langle P_{X_B|X_A}, \log \frac{P_{X_B|X_A}}{P_{X_B|K}} \rangle \rangle && (6)\\
= \; & \arg\min_{\alpha} E_{X_A}[D(P_{X_B|X_A} \parallel P_{X_B|K})]. && (7)
\end{aligned}$$

The function $D(p \parallel q) = \langle p, \log \frac{p}{q} \rangle$ is called the *Kullback-Leibler (K-L) divergence* or the *relative entropy* of p with respect to q [3].

From the previous developments, we see that a realization x_A of the continuous value X_A should be mapped by α to the index $\alpha(x_A)$ such that

$$\alpha(x_A) = \arg\min_{k=1}^{N} D(P_{X_B|X_A=x_A} \parallel P_{X_B|K=k}), \tag{8}$$

that is, to the index k whose associated distribution $P_{X_B|K=k}$ is the nearest neighbor of $P_{X_B|X_A=x_A}$ in terms of the K-L divergence. This is equivalent to the first Lloyd's optimality condition in classical vector quantization.

The nearest neighbor condition in Eqn. (8), however, is tail-biting: the mapping α is defined through the distributions $P_{X_B|K}$, which in turn depend on α. This observation suggests an algorithm in which the mapping and the conditional distributions are updated alternately. Let us define $\{f_k\}_{k=1}^N$ the *codebook* of probability distributions for X_B and the *quantization cells* $\mathcal{Q}_k = \{x_A \in \mathcal{X}_A \mid \alpha(x_A) = k\}$, i.e. the subsets of $\mathcal{X}_A$ whose elements are mapped to the same quantization index k. The quantizer α is completely defined by the partition $\{\mathcal{Q}_k\}_{k=1}^N$. Algorithm 1 is applied, starting with any initial quantizer α. A suitable tie-breaking rule is used in the update step for $\mathcal{Q}_k$.

While this algorithm is an adaptation of the well-known generalized Lloyd algorithm, we can consider that the agglomerative information bottleneck technique [11] is an adaptation of the Pairwise Nearest Neighbor algorithm [5] for vector quantizer design.

The local optimization algorithm can be implemented in practice using a training set $\mathcal{T}$ of outcomes of X_A and applying the nearest neighbor rule (8) for each element of the set. The algorithm becomes as described in Algorithm 2.

Algorithm 1 A general alternate optimization algorithm

repeat
  for $k = 1, 2, \ldots, N$ **do**
    $f_k \leftarrow E[P_{X_B|X_A} \mid X_A \in \mathcal{Q}_k]$
  end for
  for $k = 1, 2, \ldots, N$ **do**
    $\mathcal{Q}_k \leftarrow \{x_A \in \mathcal{X}_A \mid \forall j \neq k \; D(P_{X_B|X_A=x_A} \parallel f_j) > D(P_{X_B|X_A=x_A} \parallel f_k)\}$
  end for
until variation of $E_{X_A}[D(P_{X_B|X_A} \parallel P_{X_B|K})]$ becomes negligible

Algorithm 2 A practical alternate optimization algorithm

repeat
  for $k = 1, 2, \ldots, N$ **do**
    $\mathcal{T}_k \leftarrow \{x_A \in \mathcal{T} \mid \alpha(x_A) = k\}$
    $f_k \leftarrow (\sum_{x_A \in \mathcal{T}_k} P_{X_B|X_A=x_A})/|\mathcal{T}_k|$
  end for
  for each $x_A \in \mathcal{T}$ **do**
    $\alpha(x_A) \leftarrow \arg\min_{k=1}^{N} D(P_{X_B|X_A=x_A} \parallel f_k)$
  end for
until variation of $\sum_{x_A \in \mathcal{T}} D(P_{X_B|X_A=x_A} \parallel f_{\alpha(x_A)})/|\mathcal{T}|$ becomes negligible

Note that when X_B is a continuous random variable, evaluating the K-L divergence may be a difficult task, and it is probably simpler to quantize X_B with a high resolution quantizer, so that probability distributions are vectors and integrals reduce to discrete sums. For this reason and for clarity, we will henceforth assume that the set $\mathcal{X}_B$ is finite.

4 Properties

Quantization cells $\mathcal{Q}_k$ have no special structure. It is not necessary, in particular, that values of X_A that are close to each other lead to similar distributions for X_B. On

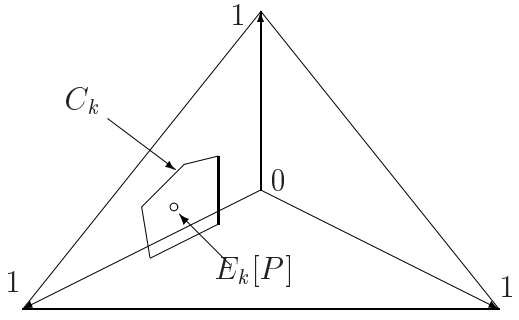


Figure 1: A cell C_k on the probability simplex for $|\mathcal{X}_B| = 3$

the other hand, there exist quantization cells $\mathcal{C}_k$ on the probability simplex, the set of vectors of size $|\mathcal{X}_B|$ with positive components summing to one. These cells contain all probability mass functions for X_B corresponding to a given quantization index k: $\mathcal{C}_k = \{P_{X_B|X_A=x_A} \mid \alpha(x_A) = k\}$. An illustration is given on Fig. 1. These cells are connected and bounded by $(|\mathcal{X}_B| - 2)$-dimensional hyperplanes. We first show that the optimal value of f_k within a cell is the average probability mass function in that cell. In other words, vector quantizers minimizing the K-L divergence obey the centroid rule. In the following, $g(.)$ is the probability density function of the distribution $P = (P_1, P_2, \ldots, P_{|\mathcal{X}_B|})$ of X_B within the cell $\mathcal{C}_k$ and the expectation $E_k[.]$ denotes the expected value within that same cell. Hence $E_k[P]$ is the $|\mathcal{X}_B|$-dimensional vector $(\int_{C_k} P_1 g(P)dP, \int_{C_k} P_2 g(P)dP, \ldots, \int_{C_k} P_{|\mathcal{X}_B|} g(P)dP)$. Note that if X_A is a discrete random variable, the probability density function $g(.)$ is actually a probability mass function. The following discussion is without loss of generality.

We wish to show that $f_k = E_k[P]$ is the solution of

$$\min \int_{C_k} \langle P, \log \frac{P}{f_k} \rangle g(P)dP \tag{9}$$

subject to $\langle 1, f_k \rangle = 1$. This is easily achieved by writing the Lagrangian cost

$$J = \lambda \langle 1, f_k \rangle + \int_{C_k} \langle P, \log \frac{P}{f_k} \rangle g(P)dP. \tag{10}$$

Taking the derivative for each component j leads to

$$\frac{\delta J}{\delta f_{k,j}} = \lambda - \frac{1}{f_{k,j}} \int_{C_k} P_j g(P)dP = 0 \tag{11}$$

by identification, we find $\lambda = 1$ and $f_{k,j} = \int_{C_k} P_j g(P)dP = E_k[P_j]$, hence $f_k = E_k[P]$.

This centroid rule is important, because it proves that the alternate optimization algorithm converges: each of the two steps minimizes the K-L divergence, and since this quantity is always positive, the algorithm must converge to a quantizer that is locally optimal with respect to both the nearest neighbor and the centroid rule.

Let us now compute the exact average K-L divergence D_k within the cell $\mathcal{C}_k$:

$$\begin{aligned} D_k &= \int_{C_k} \langle P, \log \frac{P}{f_k} \rangle g(P)dP & (12) \\ &= \int_{C_k} \left(\langle P, \log P \rangle - \langle P, \log f_k \rangle \right) g(P)dP & (13) \\ &= -E_k[H(P)] - \langle E_k[P], \log f_k \rangle & (14) \end{aligned}$$

but since $f_k = E_k[P]$ we obtain

$$D_k = H(E_k[P]) - E_k[H(P)]. \tag{15}$$

When $g(.)$ is actually a probability mass function, this expression is known as the generalized Jensen-Shannon divergence [7]. We conclude that minimization of the average K-L divergence or the average Jensen-Shannon divergence within a cluster are similar problems.

5 Rate-Constrained Quantization

Just as in standard vector quantization [2], we can define a rate-constrained algorithm that does not restrict the range of K but rather put a constraint on the rate of α. The rate is defined differently according to the target application.

In the information bottleneck method, a constraint is put on $I(K;X_A)$ which we have shown to reduce to $H(K)$ when α is deterministic. We thus obtain a classical entropy-constrained vector quantizer. In the cryptography application, the rate is defined as the minimal quantity of information needed by Bob to reconstruct K, hence $H(K \mid X_B)$.

These two entropy constraints can be easily integrated in the previous algorithm. For instance if the rate is defined as $H(K)$ then the objective function we wish to minimize is

$$E_{X_A}[D(P_{X_B|X_A} \parallel P_{X_B|K})] + \lambda H(K). \tag{16}$$

λ is a positive Lagrangian multiplier that controls the tradeoff between the K-L divergence and the entropy of the quantizer output. This translates straightforwardly to the following modified nearest neighbor rule for encoding a realization x_A of X_A:

$$\alpha(x_A) = \arg\min_k \left(D(P_{X_B|X_A=x_A} \parallel P_{X_B|K=k}) - \lambda \log P[K=k] \right). \tag{17}$$

6 Agglomerative Algorithm

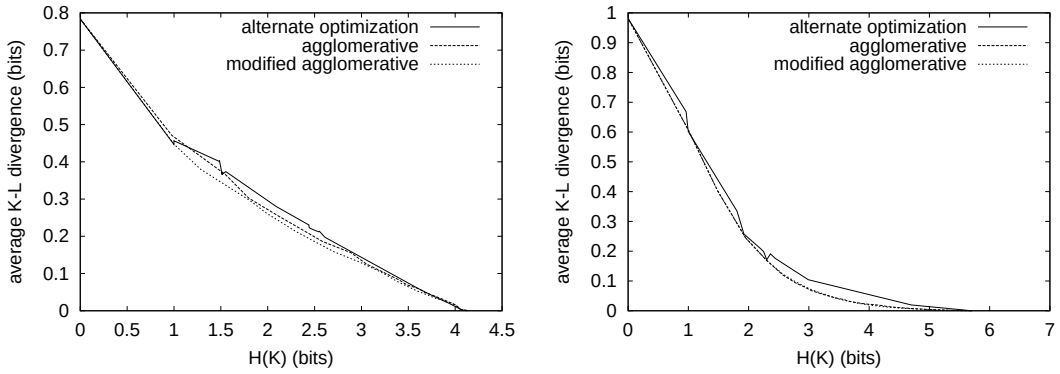
We already mentioned the agglomerative information bottleneck method, the only previously published method that deals with the exact same problem. In this method, quantization cells are iteratively merged until the mutual information reaches a given threshold. At each step, the pair of cells that leads to the minimum decrease in mutual information – or, equivalently, the minimum increase in K-L divergence – is merged. A property of the divergence measure makes the computation of the K-L divergence variation easy. We denote ΔD_{ij} the variation of the average K-L divergence when the quantization cells $\mathcal{Q}_i$ and $\mathcal{Q}_j$ (or $\mathcal{C}_i$ and $\mathcal{C}_j$) are merged, and $D_{i\cup j}$ the distortion in the merged cell. We also define

$$\pi_i = \frac{P[K=i]}{P[K=i]+P[K=j]}, \tag{18}$$

and $\pi_j = 1 - \pi_i$. Then

$$\begin{aligned}
& \frac{\Delta D_{ij}}{P[K=i]+P[K=j]} && (19)\\
= \;& D_{i\cup j} - \pi_i D_i - \pi_j D_j && (20)\\
= \;& (H(E_{i\cup j}[P]) - E_{i\cup j}[H(P)]) - \pi_i (H(E_i[P]) - E_i[H(P)]) \\
& -\pi_j (H(E_j[P]) - E_j[H(P)]) && (21)\\
= \;& H(\pi_i E_i[P] + \pi_j E_j[P]) - \pi_i H(E_i[P]) - \pi_j H(E_j[P]), && (22)
\end{aligned}$$

which is, again, a Jensen-Shannon divergence between $E_i[P]$ and $E_j[P]$.



(a) digrams (b) clusters

Figure 2: Performance comparisons

We propose a variant of this algorithm that takes into account the variation $\Delta H(K)$ of the output entropy. Instead of choosing the merge that minimizes the average K-L divergence increase, we choose the merge that minimizes the following *marginal return*:

$$\frac{-\Delta D_{ij}}{\Delta H(K)} = \frac{H(\pi_i E_i[P] + \pi_j E_j[P]) - \pi_i H(E_i[P]) - \pi_j H(E_j[P])}{\log(P[K=i] + P[K=j]) - \pi_i \log P[K=i] - \pi_j \log P[K=j]}, \quad (23)$$

i.e. that selects the most profitable merge in terms of the rate-divergence slope.

7 Experimental Comparison

We tested the following three algorithms

1. entropy-constrained alternate optimization, as presented in section 5,
2. agglomerative information bottleneck as presented in [11],
3. modified agglomerative information bottleneck as presented in the previous section,

on two simple data sets. In the first set, 'digrams', X_A is defined as a letter in $\{a, b, \ldots, z\}$ taken in an english text, and X_B is the letter that immediatly follows it. In the second set, 'clusters', the joint distribution of X_A, X_B is defined as a Gaussian mixture made up of four two-dimensional Gaussian pdf of variances 1/16 whose centers are chosen randomly in $[0, 1]^2$. In this set, X_B has been quantized to 30 different values, and we defined 200 training vectors. Results for the two sets are shown on Fig. 2. For readability, we have linked successive solution pairs.

We observe that the three algorithms have similar performance in general, and that the modified version of the agglomerative algorithm performs always at least as

well as the two others. The alternate optimization algorithm does not always yield a solution close to the two others curves, but the lower convex hulls of the three curves are in most cases indistinguishable.

8 Gaussian Modeling

Let us assume that X_B is a continuous random variable and that the conditional distributions $P_{X_B|X_A}$ are Gaussian, or in some sense close to Gaussian. Then a reasonable approximation of the K-L divergence $D(P_{X_B|X_A=x_A} \parallel f_k)$ that we wish to minimize can be obtained by modeling f_k by a Gaussian pdf $\tilde{f}_k$ with the same mean and variance.

The error due to this approximation can be computed as follows:

$$\begin{aligned} D(P \parallel f_k) &= \langle P, \log \frac{P}{f_k} \rangle & (24) \\ &= \langle P, \log \frac{P}{\tilde{f}_k} \frac{\tilde{f}_k}{f_k} \rangle & (25) \\ &= D(P \parallel \tilde{f}_k) + \langle P, \log \frac{\tilde{f}_k}{f_k} \rangle & (26) \end{aligned}$$

The additional term $\langle P, \log \frac{\tilde{f}_k}{f_k} \rangle$, the "distance" between f_k and its approximation, averaged with respect to P, should be minimized. It can be computed to give an indication of the goodness of the approximation.

Let f_1 and f_2 be two Gaussian pdf with respective means μ_1 and μ_2 and standard deviations σ_1 and σ_2. It is straightforward to show that

$$D(f_1 \parallel f_2) = \ln \frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} \text{ nats.} \quad (27)$$

If $X_B|X_A$ is multivariate Gaussian, a similar approximation of the K-L divergence can be obtained by modeling f_k as a multivariate Gaussian with estimated covariance matrices. Simple formulas for the K-L divergence are still applicable. Even finer approximations have been recently proposed by Lin, Saito and Levine in [8], using higher order statistics. All these ideas straightforwardly apply to our method. Properties of the Voronoi diagram induced by the K-L divergence in a Gaussian parametric space are described in [9].

9 Conclusion

This optimization problem is ubiquitous in pattern classification and compression. It has already been studied before, but we proposed to unify different existing views through a vector quantization approach. We introduced the entropy constraint explicitly in both the Lloyd-like and the agglomerative algorithm and validated the idea on two simple experiments. We also mentioned the use of Gaussian modeling that could lead to further developments in suitable applications.

References

[1] N. J. Cerf, M. Lévy, and G. Van Assche. Quantum distribution of gaussian keys using squeezed states. *Phys. Rev. A*, 63:052311, May 2001.

[2] P. A. Chou, T. Lookabaugh, and R. M. Gray. Entropy-constrained vector quantization. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(1):31, 1989.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

[4] A. Dimitrov, J. Miller, T. Gedeon, Z. Aldworth, and A. Parker. Analysis of neural coding using quantization with an information-based distortion measure. submitted to Network: Comput. Neural Syst., 2001.

[5] W. H. Equitz. A new vector quantization clustering algorithm. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(10):1568–1575, October 1989.

[6] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inform. Theory*, 44, 1998.

[7] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151, January 1991.

[8] J.-J. Lin, N. Saito, and R. A. Levine. Edgeworth approximations of the kullback-leibler distance towards problems in image analysis. submitted for publication, 2001.

[9] K. Onishi and H. Imai. Voronoi diagram in statistical parametric space by Kullback-Leibler divergence. In *Proceedings of the 13th ACM Symposium on Computational Geometry*, pages 463–465, 1997.

[10] J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[11] N. Slonim and N. Tishby. Agglomerative information bottleneck. In *Proc. of NIPS-12*, pages 617–623. MIT Press, 2000.

[12] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[13] X. Wu, P. A. Chou, and X. Xue. Minimum conditional entropy context quantization. In *Proc. Int. Symposium on Information Theory*, 2000.