Suboptimality of the Karhunen-Loève Transform for Fixed-Rate Transform Coding

Kenneth Zeger

University of California, San Diego, Department of ECE
La Jolla, CA 92093-0407 USA

***Abstract*- An open problem in source coding theory has been whether the Karhunen-Loève transform (KLT) is optimal for a system that orthogonally transforms a vector source, scalar quantizes the components of the transformed vector using optimal bit allocation, and then inverse transforms the vector. Huang and Schultheiss proved in 1963 that for a Gaussian source the KLT is mean squared optimal in the limit of high quantizer resolution. It is often assumed and stated in the literature that the KLT is also optimal in general for non-Gaussian sources. We disprove such assertions by demonstrating that the KLT is not optimal for certain nearly bimodal Gaussian and uniform sources. In addition, we show the unusual result that for vector sources with independent identically distributed Laplacian components, the distortion resulting from scalar quantizing the components can be reduced by including an orthogonal transform that adds inter-component dependency.**

I. INTRODUCTION

The Karhunen-Loève transform (KLT) plays a fundamental role in a variety of disciplines, including statistical pattern matching, filtering, estimation theory, and source coding. In many of these applications, the KLT is known to be "optimal" in various senses. In certain applications, however, there is widespread belief that the KLT is optimal for general conditions even though proofs only exist under very restrictive assumptions. We demonstrate here that, in fact, some of this accepted "folklore" of the KLT optimality is wrong. We specifically focus on the role of the KLT in source coding.

The main application of the KLT in source coding is in scalar quantized transform coding. In such a system, an input vector is linearly transformed into another vector of the same dimension, whose components are scalar quantized. The resulting vector is then linearly transformed to get an estimate of the original input vector. The goal is to find the pair of linear transforms and the allocation of a fixed bit budget among the various scalar quantizers that minimize the end-to-end mean squared error. This system is in general suboptimal to vector quantization but has served as an important theoretical model in order to gain understanding of optimal quantization. Its scalar quantizers are easy to implement but the transform is signal dependent which can be expensive in practice. It was shown in [5] that if the vector source is Gaussian and the bit budget is asymptotically large, then the Karhunen-Loève transform and its inverse are an optimal pair of transforms. The KLT decorrelates the input vector components, thus making them independent in the Gaussian case. The proof in [5] relies on the source being Gaussian and no subsequent extensions to non-Gaussian sources have been published in the nearly four decades that followed. Recently, Goyal, Zhuang, and Vetterli (Proof 1 of Theorem 6 in [3, p. 1628]) and Telatar (Proof 2 of Theorem 6 in [3, p. 1629]) improved the result by showing the KLT is optimal for Gaussian inputs without making any high resolution assumptions. [1]

Despite the lack of rigorous generalizations of the KLT optimality for transform coding, many people today assert that the KLT is optimal in a more general sense, if not in complete generality for all possible sources. In fact, numerous textbooks and other scholarly publications routinely quote the KLT as being optimal with respect to the transform coding / bit allocation problem, despite the lack of proof.

In the present paper, we demonstrate a class of sources for which the KLT is not optimal in the transform coding / bit allocation sense. In fact we show that the KLT makes things substantially worse for this class of sources. To the best of our knowledge, no such assertion has previously appeared in the literature for fixed-rate transform coders. [2]

The research was supported in part by the National Science Foundation. The author's email is zeger@ucsd.edu

[1] A second application of the KLT in transform coding has been called *truncation coding* or *zonal sampling* [6] in which a certain fixed number of the components of the transformed source vector are quantized using zero bits, and the remaining components are quantized with infinite accuracy. This is a special (suboptimal) case of the bit allocation transform coding system previously described. In this case it has been shown that the KLT minimizes the mean squared error over all possible choices of orthogonal transforms. Some authors confuse this form of optimality (which holds for all stationary sources) with optimality in the bit allocation version of transform coding described previously.

[2] In general there may be multiple KLTs for a given source. Feng and Effros [1] have recently shown for variable rate transform coding systems

In light of this result, one might conjecture that a weaker theorem holds, namely that if the KLT produces independent vector components then it is optimal. Such a conjecture would directly generalize the Huang-Schultheiss result, since the KLT produces independent components for vector Gaussian sources. However, we also demonstrate the somewhat surprising fact that such a conjecture is false, by exhibiting a source (i.e. the Laplacian) for which quantizing independent components is not optimal.

II. MAIN RESULTS

Let X be a k-dimensional zero mean random vector (i.e. the *source*) with real components $X_1, \ldots, X_k$, and correlation matrix $\Phi_X = E[XX^t]$. Let T be a $k \times k$ orthogonal matrix (i.e. the *transform*) with real elements and let $Y = TX$ be a transformed random vector with components $Y_1, \ldots, Y_k$. A *scalar quantizer* with resolution r bits is a mapping $Q : \mathbb{R} \to \mathbb{R}$ whose range (called a *codebook*) has cardinality 2^r. Let $Q_1, \ldots, Q_k$ be scalar quantizers with corresponding resolutions $r_1, \ldots, r_k$. For any k-dimensional vector $x = (x_1, \ldots, x_k)^t$, let $Q(x) = (Q_1(x_1), \ldots, Q_k(x_k))^t$ be a vector of scalar quantized components of x. We restrict attention to orthogonal transforms since it suffices to demonstrate the nonoptimality of the KLT over this restricted class.

The *mean squared error* at resolution r of this transform coded scalar quantization system is:

$$d_T(r) = \min_{Q:\sum_{i=1}^k r_i = kr} E[\|X - T^t Q(TX)\|^2].$$

The minimization above is taken over all scalar quantizers $Q_1 \ldots Q_k$ and all bit allocations $(r_1, \ldots, r_k)$ whose average resolution is r. When T is the identity matrix I we omit the subscript and simply write $d(r)$, which is the mean squared error corresponding to the classical scalar bit allocation problem. The optimality of a transform is considered in an asymptotic sense (e.g. [2]). The *coding gain* obtained by using transform T instead of transform U is defined as

$$G_{T,U} = \lim_{r\to\infty} d_U(r)/d_T(r).$$

An orthogonal transform T is said to be *optimal* with respect to the source X if $G_{T,U} \geq 1$ for all orthogonal transforms U. All of the results to follow refer to optimality in this fixed-rate transform coding sense.

The *Karhunen-Loève transform* is the linear map given by a $k \times k$ orthogonal matrix M^t such that $M^t \Phi_X M$ is a diagonal matrix. The matrix M decorrelates the random vector X since

$$\Phi_{M^t X} = E[(M^t X)(M^t X)^t] = M^t \Phi_X M.$$

that a bad KLT for a given source can be arbitrarily worse than the best KLT for the source.

Lemma 1 *(Huang and Schultheiss, 1963 in [5]): If a source is Gaussian then the Karhunen-Loève transform is optimal.*

It has been conjectured that Lemma 1 holds for non-Gaussian sources as well. Our main result is Theorem 1 below which shows that this conjecture is false.

Theorem 1 *There exist sources for which the Karhunen-Loève transform is not optimal.*

Theorem 2 *There exist sources for which the Karhunen-Loève transform produces independent components and yet is not optimal.*

It is clear that Theorem 1 follows from Theorem 2 but we include both theorems for the following reason. The proof of Theorem 1 demonstrates that transforming a decorrelated source with dependent components to a correlated source can improve the overall performance. The proof of Theorem 2 demonstrates that transforming a source with independent components to a decorrelated source with dependent components can improve the overall performance. Also, the coding gain achieved in the proof of Theorem 1 ranges from about 5.6 dB in one case and is about 3.4 dB in a second case, whereas the coding gain achieved in the proof of Theorem 1 is only about 1.3 dB. Thus both theorems are interesting, each in their own right. The corollary below follows from Theorem 2.

Corollary 1 *For an i.i.d. Laplacian source $\{X_i\}$ it is asymptotically better to scalar quantize successive sums and differences than to scalar quantize the source directly. That is, for all i,*

$$\lim_{r\to\infty} 2^{2r} \left\{ E\left[\left(\frac{Q(X_i + X_{i+1}) + Q(X_i - X_{i+1})}{2} - X_i\right)^2\right] + E\left[\left(\frac{Q(X_i + X_{i+1}) - Q(X_i - X_{i+1})}{2} - X_{i+1}\right)^2\right]\right\}$$
$$< \lim_{r\to\infty} 2^{2r} \cdot \left(2E[(\hat{Q}(X_i) - X_i)^2]\right)$$

where $\hat{Q}$ and Q are optimal rate r scalar quantizers for X_i and $X_i + X_{i+1}$, respectively.

The following lemma is a specialization to two dimensions of a high resolution bit allocation result of Huang and Schultheiss [5] (see also [2, pp. 228-232]).

Lemma 2 *Let $Z = (Z_1, Z_2)$ be a two-dimensional random vector with marginal probability density functions f_{Z_1} and f_{Z_2}, respectively. Let Q_1 and Q_2 be scalar quantizers with resolutions r_1 and r_2, respectively. Then, the minimum*

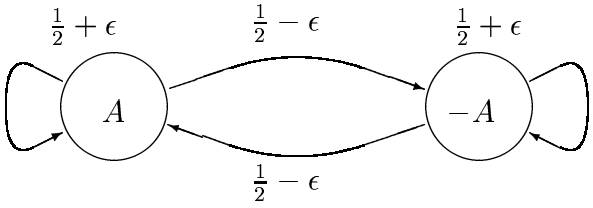

Fig. 1. Stationary Markov process S_n.

mean squared error of independently scalar quantizing the components of Z is

$$\lim_{r\to\infty} d(r)\cdot 2^{2r} = \frac{1}{6}\sqrt{\|f_{Z_1}\|_{1/3}\cdot\|f_{Z_2}\|_{1/3}}$$

where $\|f\|_{1/3} = \left(\int_{-\infty}^{\infty} f^{1/3}(u)du\right)^3$ *and* $d(r) = \min_{Q_1,Q_2:r_1+r_2\le 2r} E[(Z_1-Q_1(Z_1))^2] + E[(Z_2-Q_2(Z_2))^2]$.

III. PROOF OF THEOREM 1: NONOPTIMAL KLT WITH CORRELATED TRANSFORM COEFFICIENTS

Let W_n be an i.i.d. real random sequence with $E[W_n] = 0$, $E[W_n^2] = \sigma^2$, and symmetric probability density function f_W. Let S_n be the stationary 2-state Markov process shown in Figure 1 with $A > 0$ and $\epsilon > 0$.

That is, $S_n = A$ in one state and $S_n = -A$ in the other state and the value of ϵ determines the tendency of the process to remain in its current state. Furthermore, assume the processes W_n and S_n are independent of each other. Let $X_n = W_n + S_n$ be a scalar source and define the 2-dimensional random vector $X = (X_n, X_{n+1})^t$. Then $E[X] = (0,0)^t$ and the correlation matrix is (independent of n)

$$\Phi_X = E[XX^t] = \begin{bmatrix} \sigma^2+A^2 & 2\epsilon A^2 \\ 2\epsilon A^2 & \sigma^2+A^2 \end{bmatrix}.$$

The matrix Φ_X is diagonalized as $\Phi_X = M\Lambda M^t$, where

$$\Lambda = \begin{bmatrix} \sigma^2+A^2(1-2\epsilon) & 0 \\ 0 & \sigma^2+A^2(1+2\epsilon) \end{bmatrix}$$

is the diagonal matrix of eigenvalues of Φ_X and

$$M = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad (1)$$

is an orthogonal matrix whose columns are the corresponding eigenvectors of Φ_X. The vector $Y = M^tX$ is the Karhunen-Loève transform of the vector X, and M^t simply rotates X through an angle of 45° counterclockwise. We will show that optimally scalar quantizing (in the high resolution limit) the components of the correleted random vector X produces a smaller mean squared error than optimally scalar quantizing the components of the decorrelated vector Y. (Note that Y has dependent components).

Since the Markov process S_n is stationary, we assume without loss of generality that $X = (X_1, X_2)^t$. Notice that by symmetry the scalar components of X are identically distributed with probability density function:

$$f_{X_1}(u) = f_{X_2}(u) = \frac{1}{2}f_W(u+A) + \frac{1}{2}f_W(u-A). \quad (2)$$

Suppose $f_W(u) = 0$ for all $u \notin [-A/2, A/2]$. Then for all u, either $f_W(u+A) = 0$ or $f_W(u-A) = 0$, and thus

$$\begin{aligned}\|f_{X_1}\|_{1/3} &= \left(\int_{-\infty}^{\infty}\left[\frac{1}{2}f_W(u+A)+\frac{1}{2}f_W(u-A)\right]^{1/3}du\right)^3 \\ &= \frac{1}{2}\left(\int_{-\infty}^{\infty}\left[f_W^{1/3}(u+A)+f_W^{1/3}(u-A)\right]du\right)^3 \\ &= 4\cdot\|f_W\|_{1/3}.\end{aligned} \quad (3)$$

Let $\alpha = \frac{1}{4}-\frac{\epsilon}{2}$ and $\beta = \frac{1}{4}+\frac{\epsilon}{2}$. The joint density of (X_1, X_2) is:

$$\begin{aligned}&f_{X_1,X_2}(u,v) = \\ &\quad \beta(f_W(u+A)f_W(v+A) + f_W(u-A)f_W(v-A)) \\ &\quad +\alpha(f_W(u+A)f_W(v-A) + f_W(u-A)f_W(v+A)).\end{aligned}$$

Since $Y_1 = \frac{1}{\sqrt{2}}(X_1-X_2)$, $Y_2 = \frac{1}{\sqrt{2}}(X_1+X_2)$, and since f_W is symmetric, the marginal densities of Y are:

$$\begin{aligned}f_{Y_1}(u) &= \sqrt{2}\cdot\Big[\alpha(f_W*f_W)\left(u\sqrt{2}+2A\right) \\ &\quad +\alpha(f_W*f_W)\left(u\sqrt{2}-2A\right) + 2\beta(f_W*f_W)\left(u\sqrt{2}\right)\Big] \\ f_{Y_2}(u) &= \sqrt{2}\cdot\Big[\beta(f_W*f_W)\left(u\sqrt{2}+2A\right) \\ &\quad +\beta(f_W*f_W)\left(u\sqrt{2}-2A\right) + 2\alpha(f_W*f_W)\left(u\sqrt{2}\right)\Big]\end{aligned}$$

where the convolution of f_W with itself is

$$(f_W*f_W)(u) = \int_{-\infty}^{\infty} f_W(t)f_W(u-t)dt.$$

From the marginal densities we obtain

$$\begin{aligned}\|f_{Y_1}\|_{1/3} &= \frac{1}{2}(2^{1/3}\beta^{1/3}+2\alpha^{1/3})^3\cdot\|f_W*f_W\|_{1/3} \\ \|f_{Y_2}\|_{1/3} &= \frac{1}{2}(2^{1/3}\alpha^{1/3}+2\beta^{1/3})^3\cdot\|f_W*f_W\|_{1/3}.\end{aligned}$$

By Lemma 2, the coding gain obtained by quantizing the correlated scalar components instead of the uncorrelated components is:

$$G_{I,M^t} = \lim_{r\to\infty}\frac{d_{M^t}(r)}{d(r)} = \frac{\sqrt{\|f_{Y_1}\|_{1/3}\cdot\|f_{Y_2}\|_{1/3}}}{\sqrt{\|f_{X_1}\|_{1/3}\cdot\|f_{X_2}\|_{1/3}}} = \frac{\|f_W*f_W\|_{1/3}}{8\|f_W\|_{1/3}}\cdot\left[(2^{1/3}\beta^{1/3}+2\alpha^{1/3})(2^{1/3}\alpha^{1/3}+2\beta^{1/3})\right]^{3/2}.$$

A. Special case: W is uniform

If the random variable W is uniform on $[-B, B]$ then

$$\|f_W * f_W\|_{1/3} \quad = \quad 4B^2$$

and therefore for all $\epsilon \in (0, 1/2]$,

$$G_{I,M^t} = \frac{27}{64} \cdot \left[(2^{1/3}\beta^{1/3} + 2\alpha^{1/3})(2^{1/3}\alpha^{1/3} + 2\beta^{1/3}) \right]^{3/2}.$$

In the limit as $\epsilon \to 0$ the coding gain is about 5.63 dB.

Figures 2 and 3 show the two-dimensional probability density functions of X and Y and also the equivalent vector quantizer codebooks, induced by optimally scalar quantizing the individual components of the random vectors. The equivalent vector quantizer codebooks are the Cartesian products of the scalar quantizer codebooks for the two components of the source vector. It can be seen that the quantizer associated with the decorrelated random vector Y is very inefficient and leads to higher mean squared error.

B. Special case: W is Gaussian

Suppose the random variable W is Gaussian with

$$f_W(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/(2\sigma^2)}$$

and $\sigma << A$. The marginal densities of X are still given by (2). For small values of σ, Equation (3) is an approximation which becomes accurate as $\sigma \to 0$, giving

$$\lim_{\sigma \to 0} \frac{\|f_{X_1}\|_{1/3}}{\sigma^2} = \lim_{\sigma \to 0} \frac{\|f_{X_2}\|_{1/3}}{\sigma^2} = 24\sqrt{3} \cdot \pi$$

and the marginal densities of Y were given earlier. Also,

$$\begin{aligned} (f_W * f_W)(u) \quad &= \quad \frac{1}{2\sigma\sqrt{\pi}} \cdot e^{-u^2/(4\sigma^2)} \\ \|f_W * f_W\|_{1/3} \quad &= \quad 12\sqrt{3} \cdot \pi\sigma^2. \end{aligned}$$

Likewise, the expressions for $\|f_{Y_1}\|_{1/3}$ and $\|f_{Y_2}\|_{1/3}$ derived for sources with bounded support become accurate for a Gaussian as $\sigma \to 0$. Therefore, the limiting coding gain G_{I,M^t} obtained by quantizing the correlated scalar components instead of the uncorrelated components is:

$$\lim_{\sigma \to 0} G_{I,M^t} = \frac{1}{4} \cdot \left[(2^{1/3}\beta^{1/3} + 2\alpha^{1/3})(2^{1/3}\alpha^{1/3} + 2\beta^{1/3}) \right]^{3/2}.$$

Therefore $G_{I,M^t} > 1$ whenever $\epsilon < 0.485$ for some $\sigma > 0$. For $e \approx 0$, the coding gain obtained by not decorrelating the source X is about 3.36 dB.

IV. PROOF OF THEOREM 2: QUANTIZING INDEPENDENT TRANSFORM COEFFICIENTS CAN BE SUBOPTIMAL

In this section we demonstrate that it is not always optimal to quantize independent components. That is, the mean squared error can sometimes be reduced by orthoginally transforming a vector of independent random variables prior to quantization. The example we give is the Laplacian source in two dimensions, in which case the KLT in not unique. We show that a rotation by $45°$ is an orthogonal transform that yields dependent but uncorrelated components and results in a strictly smaller mean-squared error than simply using the identity transform. Let

$$f_{X_1}(u) = f_{X_2}(u) = \frac{1}{\sigma\sqrt{2}} \cdot e^{-|u|\sqrt{2}/\sigma}$$

and let the transform M be the same as in (1). Then

$$\|f_{X_1}\|_{1/3} = \|f_{X_2}\|_{1/3} = 54\sigma^2.$$

The joint density of (Y_1, Y_2) is

$$f_{Y_1,Y_1}(u, v) \quad = \quad \frac{1}{2\sigma^2} \cdot e^{-\frac{1}{\sigma}(|v+u|+|v-u|)}$$

and the marginal densities are

$$f_{Y_1}(u) = f_{Y_2}(u) = \frac{1}{\sigma^2} \cdot e^{-2|u|/\sigma} \left(|u| + \frac{\sigma}{2} \right).$$

Therefore,

$$\|f_{Y_1}\|_{1/3} = \|f_{Y_2}\|_{1/3} \quad = \quad \frac{81e\sigma^2}{2} \cdot \Gamma^3(4/3, 1/3)$$

where the "incomplete Gamma function" is defined as (see [4, pp. 317 (eq. 3.381.3)]): $\Gamma(u, v) = \int_v^\infty t^{u-1}e^{-t}dt$. Therefore the coding gain obtained by quantizing Y instead of the independent component source X is

$$G_{I,M^t} = \frac{4}{3e \cdot \Gamma^3(4/3, 1/3)}$$

Using the identity (see [4, p. 942 (eq. 8.356.2)]) $\Gamma(u+1, v) = u\Gamma(u, v) + v^u e^{-v}$ we have

$$\Gamma(4/3, 1/3) = \frac{1}{3} \cdot \Gamma(1/3, 1/3) + (3e)^{-1/3}.$$

Using the identity (see [4, p. 941 (eq. 8.354.2)])

$$\Gamma(u, v) = \Gamma(u) - \sum_{n=0}^{\infty} \frac{(-1)^n v^{u+n}}{n!(u+n)}$$

where the usual Gamma function is $\Gamma(u) = \int_0^\infty e^{-t}t^{u-1}dt$ we have

$$\begin{aligned} &\Gamma(1/3) - \Gamma(1/3, 1/3) \\ &\quad = 3^{-1/3} \cdot \left[3 - \frac{1}{4} + \frac{1}{42} - \frac{1}{540} + \frac{1}{8424} - \cdots \right] \approx 1.922. \end{aligned}$$

Also, using the identity (see [4, pp. 937 (eq. 8.327)])

$$\Gamma(u) = u^{u-\frac{1}{2}} e^{-u} \sqrt{2\pi} \left[1 + \frac{1}{12u} + \frac{1}{288u^2} - + O(u^{-3}) \right]$$

gives

$$\Gamma(1/3) = 3^{1/6}e^{-1/3}\sqrt{2\pi}\cdot\left[1+\frac{1}{4}+\frac{9}{288}-\cdots\right] \approx 2.567$$

and thus

$$\Gamma(4/3,1/3) \approx \frac{1}{3}\cdot(2.567-1.922)+(3e)^{-1/3} \approx 0.712.$$

The gain is then

$$G_{I,M^t} \approx \frac{4}{3e\cdot(0.712)^3} \approx 1.33 \text{ dB}.$$

Thus, since $G_{I,M^t} > 1$, the distortion in quantizing the independent Laplacian random variables X_1 and X_2 is greater than the distortion of quantizing the dependent random variables Y_1 and Y_2.

V. CONCLUSION

A family of two-dimensional sources has been demonstrated for which the Karhunen-Loève transform is suboptimal in a scalar quantized transform coding system. This settles an open question in lossy source coding theory and corrects some frequent misinterpretation of the claim of the KLT being "optimal". One particular vector source in the family used in the proof was constructed by blocking together two successive observations of a nearly bimodal symmetric scalar source, with memory containing a slight preference to repeat the same mode previously observed. The scalar source chosen was uniform but it was also shown to hold for a Gaussian provided its variance was small enough.

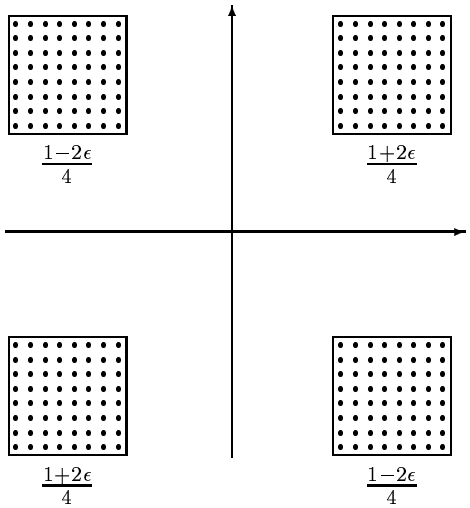

Fig. 2. Two-dimensional probability density function f_X of the correlated source X with 16-point scalar quantizers in each dimension. The density is uniform on the four squared but with different heights, as indicated. Equivalent 256-point two-dimensional vector quantizer codevectors are shown.

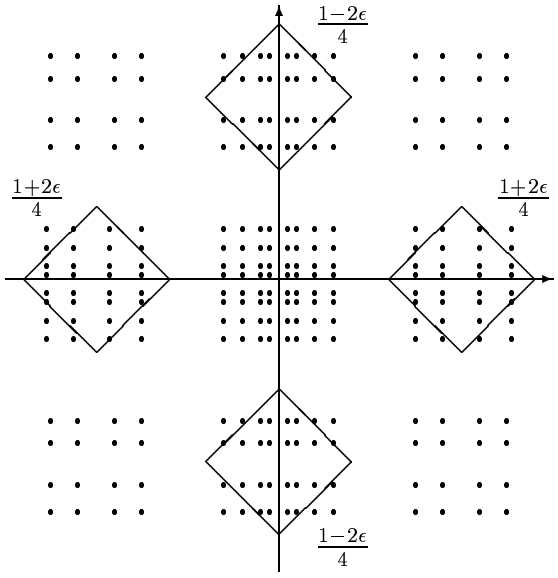

Fig. 3. Two-dimensional probability density function f_Y of the uncorrelated (i.e. Karhunen-Loève transformed) source $Y = M^tX$ with 16-point scalar quantizers in each dimension. The density f_Y is the same as that of f_X shown in Figure 2 but rotated 45° counterclockwise about the origin. Equivalent 256-point two-dimensional vector quantizer codevectors are shown.

VI. REFERENCES

[1] H. Feng and M. Effros, *A KLT can be arbitrarily worse than the optimal transform for transform coding*, Conference on Information Sciences and Systems (CISS), Princeton, NJ, March 2002 (to appear).

[2] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, Kluwer, Boston, 1992.

[3] V. K. Goyal, J. Zhuang, and M. Vetterli, "Transform coding with backward adaptive updates," *IEEE Transacations on Information Theory*, vol. 46, no. 4, pp. 1623-1633.

[4] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic Press, New York, 1980.

[5] J.-Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Transacations on Communications*, vol. 11, pp. 289-296, September 1963.

[6] N. S. Jayant and P. Noll, *Digital coding of waveforms*, Prentice-Hall, New Jersey, 1984.