10

Optimal Data Compression

M. H. van Emden
Mathematical Centre
Amsterdam

Abstract

The criteria are studied according to which it is optimal to compress data by perpendicular projection onto a subspace spanned by a set of first eigenvectors of the covariance matrix. In pattern recognition such criteria have been studied by Tou and Heydorn and by Watanabe. Several criteria have been found in a similar situation in multivariate statistics and these have been shown to be equivalent by Okamoto and Kanazawa. This paper emphasizes the equivalence of the results found in pattern recognition and in multivariate statistics. It also shows how Watanabe's approach can be extended to prove Okamoto's result and consequently also some of the better-known variational properties of the eigenvalues of a covariance matrix. In connection with a criterion in terms of entropy, a characterization of the normal distribution is given.

DATA COMPRESSION IN PATTERN CLASSIFICATION

One of the possible approaches to pattern classification proceeds in three principal steps. First, the pattern impinges on a 'retina' and the resulting (real-valued) measurements constitute a point in n-dimensional vector space. Subsequently, the 'sensory cortex' transforms this into a point in k-dimensional vector space ($k \leqslant n$) in such a way that enough information relevant to classification is retained. Data compression is regarded as the activity of the sensory cortex. Finally, in the 'motor cortex' a decision mechanism assigns the k-dimensional vector to one of the classes. This set-up is reminiscent of Rosenblatt's (1962) 'three-layer, series-coupled perceptron'.

NOTATION

Upper case letters are matrices; lower case letters are column vectors or their scalar components. An unindexed letter usually denotes the vector composed of the corresponding indexed scalars. A prime (') means transposition; it is applicable both to matrices and to vectors. Thus an inner product, sometimes

written elsewhere as (x, Wx), appears here as $x'Wx$. A diagonal matrix may be explicitly specified by $\text{diag}(\lambda_1, \ldots, \lambda_n)$, where the elements on the diagonal are $\lambda_1, \ldots, \lambda_n$. The trace (sum of the eigenvalues) and the determinant (product of the eigenvalues) of a matrix, say M, are shown as $\text{tr}(M)$ and $|M|$.

Let x be an n-dimensional random vector with mean $Ex=0$ and covariance matrix $Exx'=V(x)$. We shall suppose that x has been multiplied by a scalar in such a way that $\text{tr}(V(x))=1$. $\lambda_1(V), \lambda_2(V), \ldots, \lambda_n(V)$ are the eigenvalues of $V(x)$ in nonincreasing order and $p_1(V), p_2(V), \ldots, p_n(V)$ corresponding normalized eigenvectors. When the λs or ps are used without argument, they are understood to belong to V. When V is used without argument, it is understood to belong to x.

We may think of x as being associated with an n-dimensional probability density function or else with a sample of N vectors $s_1, \ldots, s_N$ in n-space, each of which has a weight f_i, $f_1+\ldots+f_N=1$. Such a weight may be taken to be proportional to the number of times that a pattern identical to s_i has been observed. The usual, so-called unweighted, sample is a special case of this where $f_i=1/N$ for $i=1, \ldots, N$. The weighted sample is itself a special case of a random vector x if we put $\text{Prob}(x=s_i)=f_i$ for $i=1, \ldots, N$ (Okamoto 1969). In this case we have

$$V(x)=Exx'=f_1s_1s_1'+\ldots+f_Ns_Ns_N'=SFS'$$

where $S=(s_1, \ldots, s_N)$ is an $n\times N$ matrix and $F=\text{diag}(f_1, \ldots, f_N)$.

OPTIMAL APPROXIMATION TO A RANDOM VECTOR

From now on we need to discuss only a random vector x, which is regarded as the output of the retina. The sensory cortex transforms it to a k-dimensional random vector in such a way that information relevant to classification is preserved as much as possible. Two restrictions are imposed: the transformation is to be linear and information relevant to classification is to be extracted only from the covariance matrix $V(x)$. We interpret the problem of optimal data compression as the problem of optimal approximation to a random vector by one of lower dimension.

In statistics an equivalent problem has been studied by Pearson (1901). Since Hotelling's (1933) work on it, the method of approximating a random vector by its perpendicular projection onto a subspace spanned by a set of k first eigenvectors of the covariance matrix has become widely known as the 'method of principal components'. The optimality criteria used by Pearson and Hotelling are different, and Rao (1965) has introduced yet another one leading to the same approximation. Okamoto and Kanazawa (1968) and Okamoto (1969) investigated the relation between these criteria. In the latter paper a theorem is presented that indicates a whole class of criteria that lead to the same approximation and of which the earlier are special cases.

In pattern recognition, the same problem of approximation to a random vector has been encountered, but different names were used: 'feature

selection' or 'data compression'. Possibly as a result of this, the problem was solved anew (Watanabe 1965, Tou and Heydorn 1967). One of the results of Tou and Heydorn is a direct consequence of the properties of the principal components approximation. Watanabe uses a criterion that leads to the same approximation but is more powerful in the sense that it simultaneously characterizes all solutions for $k=1, \ldots, n$. Again, this may be shown to be a consequence of Okamoto's theorem. In this paper we have chosen the opposite direction and we shall show that Watanabe's approach may be extended to yield Okamoto's theorem as a consequence and thereby also some of the more widely known extremal properties of the eigenvalues of a covariance matrix, such as Fischer's max-min theorem.

WATANABE'S CRITERION

Let U be a square matrix whose columns are an orthonormal set $u_1, \ldots, u_n$, that is, $U'U=I$, the identity matrix. Then we have:

$$x=Ix=U'Ux=UU'x=u_1u_1'x+ \ldots +u_nu_n'x.$$

Here the scalar random variables $u_1'x, \ldots, u_n'x$ are the components of x with respect to the basis $u_1, \ldots, u_n$. Because of the invariance of the trace under a similarity transformation, we have:

$$\mathrm{tr}(V(U'x))=\mathrm{tr}(U'VU)=\mathrm{tr}(U^{-1}VU)=\mathrm{tr}(V)=1.$$

This implies that, whatever orthonormal base we choose, the variances of the components add up to one.

Watanabe (1965) chose as approximation to x its perpendicular projection onto a subspace spanned by k vectors of the basis. He selected the most 'significant' k vectors, where significance of a basis vector was interpreted to be the variance of the corresponding component (a component with small variance gives little information about the difference between various occurrences of x, which is what we are interested in). Therefore, the subspace to be chosen is spanned by the basis vectors corresponding to the components that have the largest variances. The total amount of variance collected in this way is the greater the more unequal the sum of variances is distributed over the basis vectors. The obvious way to express equality of a partition of 1 into n nonnegative numbers $\rho_1, \ldots, \rho_n$ is an entropy-like function $H(\rho)=-\phi(\rho_1) \ldots -\phi(\rho_n)$ where ϕ is a continuous and convex function.

Watanabe's results may be summarized as follows. A random vector x is approximated by its perpendicular projection onto the subspace spanned by the basis vectors $u_1, \ldots, u_k$ for which the corresponding components have the largest variances $\rho_1, \ldots, \rho_k$. The approximation is considered optimal for $k=1, \ldots, n$ if the basis is chosen such that $H(\rho)$ is minimum.

Theorem 1

The minimum is attained if and only if $u_1=p_1, \ldots, u_n=p_n$.

Proof. Let P be the matrix of which the columns are $p_1, \ldots, p_n$. U is an arbitrary orthogonal matrix.

$$V(U'x)=EU'xx'U=U'VU=U'P\Lambda P'U=Q\Lambda Q',$$

where $\Lambda=\text{diag}\,(\lambda_1,\ldots,\lambda_n)$ and where $Q=U'P$. Q is the product of orthogonal matrices and is therefore itself orthogonal. $V(U'x)=Q\Lambda Q'$ implies that

$$\rho_i=\lambda_1 q_{i1}^2+\ldots+\lambda_n q_{in}^2 \quad \text{or} \quad \rho_i=\lambda_1 r_{i1}+\ldots+\lambda_n r_{in} \tag{1}$$

where ρ_i is the ith diagonal element of $V(U'x)$ and $r_{ik}=q_{ik}^2$ are the nonnegative elements of a matrix R. Because of the orthogonality of Q, both row and column sums of R equal 1. We can now write (1) as $\rho=R\lambda$. The effect of multiplying by R may be called a linear averaging transformation: each element of ρ is a weighted mean of the elements of λ. We shall say that λ *majorizes* ρ when, supposing the ρs and λs to be arranged in nonincreasing order, the following relations hold:

$$\begin{aligned}&\lambda_1+\ldots+\lambda_k\geqslant\rho_1+\ldots+\rho_k \quad \text{for } k=1,\ldots,n-1 \text{ and}\\ &\lambda_1+\ldots+\lambda_n=\rho_1+\ldots+\rho_n.\end{aligned} \tag{2}$$

Now, a necessary and sufficient condition for λ to majorize ρ is that $\rho=R\lambda$ for some linear averaging transformation R (Hardy, Littlewood, and Pólya 1934, theorem 46).

Let ϕ be some continuous and convex function, then according to Jensen's inequality:

$$\phi(\rho_i)=\phi\left(\sum_{j=1}^{n} r_{ij}\lambda_j\right)\leqslant\sum_{j=1}^{n} r_{ij}\phi(\lambda_j).$$

Equality occurs only if either for some $j=j_1$, $r_{ij}=1$ and, consequently, $r_{ij}=0$ for $j\neq j_1$, or if $\lambda_1=\ldots=\lambda_n$.

$$H(\rho)=\sum_{i=1}^{n}-\phi(\rho_i)\geqslant\sum_{i=1}^{n}\sum_{j=1}^{n}-r_{ij}\phi(\lambda_j)=\sum_{j=1}^{n}-\phi(\lambda_j)=H(\lambda)$$

Equality occurs only if there is equality for each pair of terms. Suppose $u_1=p_1,\ldots,u_n=p_n$, then $\rho_1=\lambda_1,\ldots,\rho_n=\lambda_n$ and $H(\rho)$ attains the minimum $H(\lambda)$. Suppose $H(\rho)=H(\lambda)$. Either we have $\lambda_1=\ldots=\lambda_n$ and any orthonormal basis is a set of eigenvectors. Or we have $\rho_1=\lambda_1,\ldots,\rho_n=\lambda_n$ and this also implies $u_1=p_1,\ldots,u_n=p_n$. This completes the proof of Theorem 1.

Corollary

$$\begin{aligned}&\max u'Vu=\lambda_1 \text{ under condition that } u'u=1 \text{ and}\\ &\min u'Vu=\lambda_n \text{ under condition that } u'u=1.\end{aligned} \tag{3}$$

Proof. $\rho_i=u_i'Vu_i$ and λ majorizes ρ for any choice of an orthonormal set $u_1,\ldots,u_n$. The conditions (2) imply that $\lambda_1\geqslant\rho_1$ and $\lambda_n\leqslant\rho_n$.

Let us denote in the following by U_k an $n\times k$ matrix whose columns are $u_1,\ldots,u_k$, by P_k an $n\times k$ matrix whose columns are $p_1,\ldots,p_k$, which is a set of first k eigenvectors of V. Q_k is a square orthogonal matrix of order k.

Theorem 2 (Okamoto 1969)

For a fixed $k=1,\ldots,n$, all eigenvalues of $U_k'VU_k$ are maximized at $\lambda_1(V),\ldots,\lambda_k(V)$ by the choice $U_k=P_kQ_k'$. The subspace spanned by the columns of U_k is unique if and only if $\lambda_k(V)>\lambda_{k+1}(V)$.

Proof. Let us proceed by induction on k. Corollary (3) shows the theorem to be true for $k=1$. Suppose it holds for $k=i-1$. We would now like to choose the columns of U_i in such a way that $\lambda_1(U_i'VU_i), \ldots, \lambda_i(U_i'VU_i)$ are maximized. Suppose that $\lambda_j(U_i'VU_i)>\lambda_j(V)$ for some $j=1, \ldots, i-1$. If we leave out the last column of U_i we would have a contradiction to the supposition that the theorem holds for $k=i-1$. We have, therefore, $u_1=p_1, \ldots, u_{i-1}=p_{i-1}$. The vector u_i must be orthogonal to $u_1, \ldots, u_{i-1}$ and of length 1. This implies that $u_i=\alpha_i p_i+ \ldots +\alpha_n p_n$ where $u_i'u_i=1$ implies $\alpha_i^2+ \ldots +\alpha_n^2=1$.

$$U_i'VU_i=U_i'P\Lambda P'U_i=\text{diag}\,(\lambda_1, \ldots, \lambda_{i-1}, \alpha_i^2\lambda_i+ \ldots +\alpha_n^2\lambda_n).$$

Therefore, $\lambda_j(U_iVU_i)=\lambda_j$ for $j=1, \ldots, i-1$ and $\lambda_i(U_i'VU_i)=\alpha_i^2\lambda_i+ \ldots +\alpha_n^2\lambda_n$. This last eigenvalue is maximized for $\alpha_i=1, \alpha_{i+1}= \ldots =\alpha_n=0$, which implies that $u_i=p_i$.

FISCHER'S MAX-MIN THEOREM

The fact that Okamoto's theorem can be proved quite simply makes one wonder whether a more widely known theorem like Fischer's max-min theorem may be proved by means of Okamoto's theorem (theorem 2) in a simple way. That this is indeed the case may be seen in this section.

Theorem 3 ('Fischer's max-min theorem')
$\lambda_k(V)=\max \min\ c'Vc$, where minimization is over c satisfying $c'c=1$, $c'c_{k+1}=0, \ldots, c'c_n=0$, where $c_{k+1}, \ldots, c_n$ are independent and maximization is over $c_{k+1}, \ldots, c_n$. The maximum occurs for $c_{k+1}, \ldots, c_n$ spanned by a set of $n-k$ last eigenvectors of V.

Proof. Let U be an orthogonal matrix whose columns $u_1, \ldots, u_n$ are such that $u_{k+1}, \ldots, u_n$ are spanned by $c_{k+1}, \ldots, c_n$ and $u_k=c$. By corollary (3), the minimum of $c'Vc$ is μ_k, a smallest eigenvalue of $U_k'VU_k$, when we keep $c_{k+1}, \ldots, c_n$ fixed. By theorem 2, the maximum of μ_k is λ_k and this is achieved for $u_1, \ldots, u_k$, the columns of U_k, spanned by a set $p_1, \ldots, p_k$ of first eigenvectors of V.

CLOSEST FIT OF LINES AND PLANES

Let D be a linear transformation that maps x onto a vector in the subspace spanned by $u_1, \ldots, u_k$ and let $D_1=U_kU_k'$ be the perpendicular projection of x onto this subspace. Suppose we want to find D and $u_1, \ldots, u_k$ such that the sum of the variances of the components of the error vector, $E(x-Dx)'(x-Dx)$, is minimum. We have, by Pythagoras' theorem:

$$(x-Dx)'(x-Dx)=(x-D_1x)'(x-D_1x)+(D_1x-Dx)'(D_1x-Dx)$$

which shows that, for fixed $u_1, \ldots, u_k$, D must be chosen as $D=D_1=U_kU_k'$. We now have to find the $u_1, \ldots, u_k$ for which the minimum occurs.

$$1=\text{tr}(V(x))=\text{tr}(Exx')=Ex'x=E((x-D_1x)'(x-D_1x)+(D_1x)'(D_1x))=$$
$$E(x-D_1x)'(x-D_1x)+E(D_1x)'(D_1x)=\text{tr}(V(x-D_1x))+\text{tr}(V(D_1(x))).$$
$$\text{tr}(V(x-D_1x))=1-\text{tr}(V(D_1x))=1-\text{tr}(ED_1xx'D_1')$$
$$=1-\text{tr}(U_kU_k'VU_kU_k')=1-\text{tr}(U_k'VU_k).$$

Apparently, the $u_1, \ldots, u_k$ we are looking for are those that maximize $\mathrm{tr}(U_k' V U_k)$. Theorem 2 states that the choice $u_1 = p_1, \ldots, u_k = p_k$ maximizes each eigenvalue of $U_k' V U_k$ and therefore also maximizes their sum.

Pearson (1901) considered a set of $N \geqslant n$ points in n-space and sought a k-dimensional subspace that gives closest fit to this set, that is, a k-dimensional subspace such that the sum of squares of the perpendicularly-projecting lines from each of the points onto this subspace is a minimum. He concluded that the subspace sought is the one spanned by a set of first k eigenvectors of the covariance matrix of the set of points. Again, Tou and Heydorn (1967) derived this result in the one of their approaches to feature selection that they called 'estimation optimality'.

SCATTER AND ENTROPY

Hotelling (1933) considered the problem of approximating x by $y = U_k' x$ in such a way that $|V(y)|$, the scatter of y, is as large as possible (again under the constraint $U_k' U_k = I_k$). The solution is identical to the one given by theorem 2, because the determinant is the product of eigenvalues, which are all nonnegative.

This is closely related to a result about the entropy of a normal distribution derived in Tou and Heydorn (1967). Let y be any normally-distributed k-dimensional random vector. Its density is given by:

$$g(y) = |V(y)|^{-\frac{1}{2}} \cdot (2\pi)^{-\frac{1}{2}k} \cdot \exp\left(-\tfrac{1}{2} tr((V(y))^{-1} yy')\right).$$

It may be verified that the entropy

$$\begin{aligned} H(y) &= \int \cdots \int -g(y_1, \ldots, y_k) \log\left(g(y_1, \ldots, y_k)\right) dy_1 \ldots dy_k \\ &= \tfrac{1}{2} k \log(2\pi) + \tfrac{1}{2} \log|V(y)| + \tfrac{1}{2} k. \end{aligned} \tag{4}$$

If x is normally distributed, any perpendicular projection $y = U_k' x$ is also normally distributed. Theorem 2 shows, with eq. (4), that, for fixed $k = 1, \ldots, n$, if we want to choose U_k such that $H(U_k' x)$ is maximum, this is achieved by taking $U_k = P_k Q_k$.

A GENERAL CRITERION FOR OPTIMAL APPROXIMATION

The previous two sections were concerned with special optimality criteria. Here it will be shown that theorem 2 allows a general formulation of optimal approximation to a random vector. We assume x to be approximated by its perpendicular projection onto a subspace of k dimensions. This subspace is spanned by the columns $u_1, \ldots, u_k$ of a matrix U_k ($U_k' U_k = 1$). The projection, which is a k-dimensional random vector, is an optimal approximation if it is, in some suitable sense, as large as possible. This may be interpreted as making the covariance matrix $V(U_k' x)$ as large as possible, and the interpretation of this has been given by Okamoto and Kanazawa (1968) (the account we give in this section is a slightly modified version of theirs in order to avoid a difficulty with the determinant function).

Let f be a real-valued function with nonnegative definite matrices of order n as argument satisfying the following conditions:

$f(V) \leqslant f(V+W)$ for any nonnegative definite W and

$f(V)=f(P'VP)$ for any orthogonal P.

These conditions are satisfied if and only if f is identical to some function $g(\lambda_{i_1}(V), \ldots, \lambda_{i_j}(V))$ of a subset of the eigenvalues of V that is monotone nondecreasing in each of its arguments. The criteria of the previous two sections, namely $\operatorname{tr}(U_k'VU_k)$ and $|U_k'VU_k|$ are special cases of this. Rao (1965) used the norm of $U_k'VU_k$ as criterion, which is also a special case.

A CRITERION IN TERMS OF ENTROPY

According to the previous criteria, we were concerned with maximizing the approximating vector. Equivalently, we may minimize the error vector. The result is the same, but it is worth while to state it from this point of view because of the relation to yet another approach to the problem.

Suppose that $U_{n-k}=(u_1, \ldots, u_{n-k})$ and we shall consider the perpendicular projection of x onto the subspace spanned by the columns of U_{n-k} as the error vector. Its covariance matrix $V(U_{n-k}'x)=U_{n-k}'V(x)U_{n-k}$. The criterion which we consider now is the entropy H of the error vector, where $H(f)=\int -f(z)\log(f(z))\,dz$, where f is the probability density function of $U_{n-k}'x$ and integration is over the subspace spanned by the columns of U_{n-k}. The problem is to choose U_{n-k} such that the entropy of the error vector $U_{n-k}'x$ is minimum. However, the entropy depends on the probability density function f. Good (1965, 1968) argued that in many situations it makes sense to estimate probabilities in such a way that entropy is maximized under known constraints. He advocated the principle of 'minimaxing entropy': maximize entropy to find a probability distribution and, when planning an experiment, which is analogous to our choice of U_{n-k}, minimize the expected maximum entropy. The minimax characterization of principal components to be given below is reminiscent of this. In our case, the constraint is that the distribution must have the same covariance matrix as the given error vector.

Theorem 4 (Shannon 1948)
Of all distributions having a given covariance matrix, the normal distribution with that covariance matrix has maximum entropy.

Using the results of section 7, we arrive at the following characterization of the principal components solution:

$$\min \max H(f(U_{n-k}'x)) =$$
$$\tfrac{1}{2}(n-k)\log(2\pi)+\tfrac{1}{2}\log(\lambda_{k+1}\times \ldots \times \lambda_n)+\tfrac{1}{2}(n-k),$$

where maximization is over all distributions having the given covariance matrix and minimization is over $n\times(n-k)$ matrices U_{n-k}. The maximum occurs for the normal distribution and the minimum occurs for $U_{n-k}=(p_{n-k+1}, \ldots, p_n)$, a set of $n-k$ last eigenvectors of V.

A MAXIMUM ENTROPY CHARACTERIZATION OF THE NORMAL DISTRIBUTION

Shannon's theorem (theorem 4) is not quite satisfactory, because the entropy of an n-dimensional normal distribution with covariance matrix V turns out to be:

$$H=\tfrac{1}{2}n\log(2\pi)+\tfrac{1}{2}\log(|V|)+\tfrac{1}{2}n.$$

Apparently, the entropy is not completely specified by the individual covariances v_{ij}, but only by $|V|$ and Shannon's condition can be relaxed to stating this determinant. But then there is no unique distribution for which the maximum of entropy is achieved. In this section we are concerned with a less stringent constraint that leads to a uniquely determined maximizing distribution.

Theorem 5

Let W be a positive definite real symmetric matrix of order n. Of all density functions f with zero average, of which the covariance matrix V satisfies

$$\operatorname{tr}(VW)\leqslant n, \tag{5}$$

$$f(x)=|W|^{\frac{1}{2}}(2\pi)^{-\frac{1}{2}n}\exp(-\tfrac{1}{2}x'Wx) \tag{6}$$

has maximum entropy, which is

$$H(f)=\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}\log(|W|)+\tfrac{1}{2}n. \tag{7}$$

Furthermore, any distribution satisfying eq. (5) and not identical to eq. (6) has an entropy less than eq. (7).

Proof. Besides the entropy $H(f)=\int -f(x)\log(f(x))\,dx$ of the density function f, we shall also consider its energy $U(f)=\int \tfrac{1}{2}f(x)x'Wx\,dx$. We first determine the density f that maximizes H under the constraints $\int f(x)\,dx=1$ and $U(f)=\tfrac{1}{2}n$. This is equivalent to the maximization without constraints of:

$$H+\lambda(\tfrac{1}{2}n-U)+\mu(\int f(x)\,dx-1),$$

where λ and μ are Lagrange multipliers.

$$\begin{aligned}
&H+\lambda(\tfrac{1}{2}n-U)+\mu(\int f(x)\,dx-1)\\
&=\int f(x)\log(1/f(x))\,dx-\int f(x)\tfrac{1}{2}\lambda x'Wx\,dx+\int \mu f(x)\,dx+\tfrac{1}{2}n\lambda-\mu\\
&=\int f(x)\log\left(\frac{\exp(\mu)\exp(-\tfrac{1}{2}\lambda x'Wx)}{f(x)}\right)dx+\tfrac{1}{2}\lambda n-\mu\\
&\leqslant\int f(x)\left(\frac{\exp(\mu)\exp(-\tfrac{1}{2}\lambda x'Wx)}{f(x)}-1\right)dx+\tfrac{1}{2}\lambda n-\mu\\
&=\exp(\mu)\int \exp(-\tfrac{1}{2}\lambda x'Wx)\,dx-1+\tfrac{1}{2}\lambda n-\mu.
\end{aligned}$$

The maximum occurs if and only if in each point x we have

$$f(x)=\exp(\mu)\exp(-\tfrac{1}{2}\lambda x'Wx).$$

The choice $\lambda=1$, $\mu=\tfrac{1}{2}\log|W|-\tfrac{1}{2}n\log(2\pi)$ satisfies the constraints.

Then we have

$$f(x)=|W|^{\frac{1}{2}}(2\pi)^{-\frac{1}{2}n}\exp(-\tfrac{1}{2}x'Wx) \text{ and} \tag{8}$$

$$H(f)=\tfrac{1}{2}\lambda n-\mu=\tfrac{1}{2}n+\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}\log(|W|). \tag{9}$$

This derivation can be applied directly to the distribution of the velocity components of a molecule of an ideal gas to yield Maxwell's distribution. In that case W would be the identity matrix, but the full generality of W may well be useful to find the distribution in cases where the quadratic form for the energy is more complicated.

Note that $U(f)=\int f(x)\tfrac{1}{2}x'Wx\,dx=\tfrac{1}{2}\mathrm{tr}(VW)$, so that we found a maximum for the entropy under the condition that $\mathrm{tr}(VW)=n$. The same maximum would be found under the condition $\mathrm{tr}(VW)\leqslant n$, for suppose for the moment that inequality holds. The geometric-arithmetic mean inequality implies that

$$|VW|^{1/n}\leqslant \mathrm{tr}(VW)/n \tag{10}$$

so, in that case, we would have $|V|<|W|^{-1}$. But the maximization could be carried out with $W_1=V^{-1}$ and we would find an entropy

$$H=\tfrac{1}{2}n+\tfrac{1}{2}n\log(2\pi)-\tfrac{1}{2}\log(|W_1|).$$

Therefore, any distribution, whether normal or not, for which $|V|<|W|^{-1}$, has an entropy smaller than eq. (9).

The maximizing distribution must therefore have $|V|\geqslant|W|^{-1}$. Inequality is impossible because of eq. (10). The only remaining case we have to investigate is that of a distribution different from eq. (8), but with $|V|=|W|^{-1}$ and also normal, achieves the same maximum entropy. Then we have:

$$1=|V|\,|W|=|VW|=|VW|^{1/n}\leqslant \mathrm{tr}(VW)/n=1,$$

where the last equality is the constraint. We must therefore have equality in eq. (10), which implies $V=W^{-1}$. This proves that eq. (8) uniquely maximizes H.

Acknowledgement

The research reported here was carried out at the Mathematical Centre, Amsterdam and entirely supported by the Hugo de Vries Laboratory for Systematic Botany in the University of Amsterdam as part of its program to develop mathematical methods for plant-ecological data.

REFERENCES

Good, I.J. (1965) *The estimation of probabilities.* Cambridge, Mass.: MIT Press.

Good, I.J. (1968) Some statistical methods in machine-intelligence research. *Virginia J. of Sci.*, **19**, 101–10.

Hardy, G.H., Littlewood, J.E. & Pólya, G. (1934) *Inequalities.* Cambridge: Cambridge University Press.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *J. educ. Psych.*, **26**, 417–41, 498–520.

Okamoto, M. & Kanazawa, M. (1968) Minimization of eigenvalues of a matrix and optimality of principal components. *Ann. Math. Stat.*, **39**, 859–63.

Okamoto, M. (1969) Optimality of principal components. *Multivariate Analysis II*, (ed. Krishnaiah, P.R.). New York: Academic Press.

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, **2**, 559–72.

Rao, C.R. (1956) *Linear statistical inference and its applications*. New York: Wiley.
Rosenblatt, F. (1962) *Principles of neurodynamics*. Washington, D.C.: Spartan Books.
Tou, J. & Heydorn, R.P. (1967) Some approaches to optimum feature extraction. *Computer and information sciences II* (ed. Tou, J.). New York: Academic Press.
Watanabe, S. (1965) Karhunen-Loève expansion and factor analysis. *Proc. fourth Prague conference on information theory*, 635–59.