KYBERNETIKA — VOLUME 22 (1986), NUMBER 4

MULTIMODAL DISCRETE KARHUNEN-LOÈVE EXPANSION

JIŘÍ GRIM

An approach is suggested which combines the idea of discrete Karhunen-Loève expansion with the centroid method of cluster analysis to optimize an approximation of multimodal data. The resulting algorithm is illustrated on the classical iris data of Fisher.

1. INTRODUCTION

The discrete Karhunen-Loève (K.-L.) expansion is well known to be useful in different areas because of its optimal properties and a simple implementation. Without affecting any aspect of the general problem we shall consider this method in the context of approximating. Particularly let $\mathbf{X}$ be a real random vector with a mean μ and a covariance matrix Σ

$$(1.1)\qquad \mathbf{X} = (X_1, X_2, \ldots, X_d)^T \in \mathbb{R}_d\,; \quad \mathsf{E}\{\mathbf{X}\} = \mu \in \mathbb{R}_d\,; \quad \mathsf{E}\{(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\} = \Sigma\,.$$

Given a vector $\mathbf{c} \in \mathbb{R}_d$ and a complete vector basis $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d\}$ of the space $\mathbb{R}_d$ we can write

$$(1.2)\qquad \mathbf{X} = \mathbf{c} + \sum_{i=1}^{d} Y_i \mathbf{u}_i\,; \quad \mathbf{c} = (c_1, c_2, \ldots, c_d)^T \in \mathbb{R}_d\,; \quad \mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{id})^T \in \mathbb{R}_d$$

where $Y_i = Y_i(\mathbf{X})$ are coordinates of $\mathbf{X}$ related to the basis U. If U is an orthonormal basis we have a simple expression for Y_i:

$$(1.3)\qquad Y_i = Y_i(\mathbf{X}) = \mathbf{u}_i^T(\mathbf{X} - \mathbf{c})\,; \quad i = 1, 2, \ldots, d\,.$$

For this reason an approximation $\mathbf{X}^+$ of the vector $\mathbf{X}$ is usually assumed in the form of truncated orthonormal expansion (1.2):

$$(1.4)\qquad \mathbf{X}^+ = \mathbf{c} + \sum_{i=1}^{d_0} Y_i(\mathbf{X})\, \mathbf{u}_i\,; \quad d_0 < d\,; \quad \mathbf{c} \in \mathbb{R}_d\,; \quad \mathbf{u}_i \in U\,.$$

One can see that, as the vectors $\mathbf{c}, \mathbf{u}_1, \ldots, \mathbf{u}_{d_0}$ are constant parameters of the mapping

(1.4), the approximating vector $\boldsymbol{X} \in \mathbb{R}_d$ is actually determined only by d_0 variables $Y_1, Y_2, \ldots, Y_{d_0}$.

To optimize the choice of the parameters $\mathbf{c}$ and U the natural mean square error criterion is generally used which can be expressed as follows (cf. (1.2)–(1.4)):

$$(1.5) \qquad \mathsf{E}\{\|\boldsymbol{X} - \boldsymbol{X}^+\|^2\} = \mathsf{E}\{\|\boldsymbol{X} - \mathbf{c}\|^2 - \sum_{i=1}^{d_0} Y_i^2\} =$$

$$= \mathsf{E}\{\|\boldsymbol{X} - \mathbf{c}\|^2\} - \sum_{i=1}^{d_0} \boldsymbol{u}_i^{\mathrm{T}}\, \mathsf{E}\{(\boldsymbol{X} - \mathbf{c})(\boldsymbol{X} - \mathbf{c})^{\mathrm{T}}\}\, \boldsymbol{u}_i\,.$$

Further, using the relations (1.1), we can write

$$(1.6) \qquad \mathsf{E}\{\|\boldsymbol{X} - \boldsymbol{X}^+\|^2\} = \mathsf{E}\{\|\boldsymbol{X} - \boldsymbol{\mu}\|^2\} + \|\boldsymbol{\mu} - \mathbf{c}\|^2 -$$

$$- \sum_{i=1}^{d_0} \boldsymbol{u}_i^{\mathrm{T}} \Sigma \boldsymbol{u}_i - \sum_{i=1}^{d_0} \boldsymbol{u}_i^{\mathrm{T}}(\boldsymbol{\mu} - \mathbf{c})(\boldsymbol{\mu} - \mathbf{c})^{\mathrm{T}} \boldsymbol{u}_i =$$

$$= \mathsf{E}\{\|\boldsymbol{X} - \boldsymbol{\mu}\|^2\} + (\boldsymbol{\mu} - \mathbf{c})^{\mathrm{T}} \Big[\sum_{i=d_0+1}^{d} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}\Big](\boldsymbol{\mu} - \mathbf{c}) - \sum_{i=1}^{d_0} \boldsymbol{u}_i^{\mathrm{T}} \Sigma \boldsymbol{u}_i\,.$$

The second term in the last expression represents a positively semi-definite quadratic form which is zero if we set $\mathbf{c} = \boldsymbol{\mu}$. We have then

$$(1.7) \qquad \mathsf{E}\{\|\boldsymbol{X} - \boldsymbol{X}^+\|^2\} = \mathsf{E}\{\boldsymbol{X} - \boldsymbol{\mu}\|^2\} - \sum_{i=1}^{d_0} \boldsymbol{u}_i^{\mathrm{T}} \Sigma \boldsymbol{u}_i\,.$$

Let us recall now that if the symmetrical matrix Σ is nonsingular then it has d real positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ and the corresponding eigenvectors $\mathbf{v}_i$ can be chosen to be orthonormal. Thus, introducing the matrices

$$(1.8) \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_d \end{pmatrix}; \quad (\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_d)$$

$$V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)\,; \quad \mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j = \delta_{ij}$$

we can write

$$(1.9) \qquad \Sigma = V \Lambda V^{\mathrm{T}}\,.$$

From the extremal properties of eigenvalues and eigenvectors it follows directly that the second term in (1.7) is maximized by the eigenvectors of the matrix Σ which are associated with the d_0 largest eigenvalues. The optimal parameters of the approximation formula (1.4) are therefore

$$(1.10) \qquad \mathbf{c} = \boldsymbol{\mu}_i\,; \quad \boldsymbol{u}_i = \mathbf{v}_i\,; \quad i = 1, 2, \ldots, d_0$$

and the following relations are easily verified

$$(1.11) \quad \mathsf{E}\{\boldsymbol{X}^+\} = \boldsymbol{\mu}\,; \quad \mathsf{E}\{\|\boldsymbol{X} - \boldsymbol{X}^+\|^2\} = \sum_{i=d_0+1}^{d} \lambda_i\,; \quad \mathsf{E}\{Y_i(\boldsymbol{X})\, Y_j(\boldsymbol{X})\} = \lambda_i \delta_{ij}\,.$$

Instead of exact statistical properties we are given usually only a sample S of independent observations of the approximated random vector $\boldsymbol{X}$:

$$S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \; ; \quad \mathbf{x}_n \in \mathbb{R}_d \, . \tag{1.12}$$

The above formulas are also applicable in this case if we replace the expectation operator by the corresponding sample mean. Thus the parameters of the approximation formula

$$\mathbf{x}^+ = \mathbf{x}^+(\mathbf{x}) = \mathbf{c} + \sum_{i=1}^{d_0} y_i(\mathbf{x}) \, \mathbf{v}_i \; ; \quad y_i(\mathbf{x}) = \mathbf{v}_i^{\mathrm{T}}(\mathbf{x} - \mathbf{c}) \; ; \quad \mathbf{x} \in \mathbb{R}_d \tag{1.13}$$

minimize the mean square error criterion

$$Q = \frac{1}{N} \sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{x}^+\|^2 = \frac{1}{N} \sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{c} - \sum_{i=1}^{d_0} y_i(\mathbf{x}) \, \mathbf{v}_i\|^2 \tag{1.14}$$

if

$$\mathbf{c} = \frac{1}{N} \sum_{\mathbf{x} \in S} \mathbf{x} \; ; \quad A = \frac{1}{N} \sum_{\mathbf{x} \in S} (\mathbf{x} - \mathbf{c})(\mathbf{x} - \mathbf{c})^{\mathrm{T}} \tag{1.15}$$

and $\mathbf{v}_1, \ldots, \mathbf{v}_{d_0}$ are the eigenvectors associated with the d_0 largest eigenvalues $(\lambda_1 \geqq \ldots \ldots \lambda_{d_0} \geqq \ldots)$ of the matrix A. Also, in analogy with (1.11), we can write

$$\frac{1}{N} \sum_{\mathbf{x} \in S} \mathbf{x}^+(\mathbf{x}) = \mathbf{c} \; ; \quad \frac{1}{N} \sum_{\mathbf{x} \in S} y_i(\mathbf{x}) \, y_j(\mathbf{x}) = \mathbf{v}_i^{\mathrm{T}} A \mathbf{v}_j = \lambda_i \delta_{ij} \; ; \tag{1.16}$$

$$Q = \frac{1}{N} \sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{x}^+\|^2 = \frac{1}{N} \sum_{\mathbf{x} \in S} \sum_{i=d_0+1}^{d} y_i^2(\mathbf{x}) = \sum_{i=d_0+1}^{d} \lambda_i \, . \tag{1.17}$$

Let us recall that the optimal approximation formula (1.13) is derived from the global characteristics (1.15) (cf. also (1.4) (1.10), (1.1)). If the underlying probability distribution is multimodal then the parameters $\mathbf{c}$, A in (1.15) (or $\boldsymbol{\mu}$, Σ in (1.1)) represent merely mean values of characteristics corresponding to different modes. In such a case the attainable accuracy of approximation would be probably low. For this reason it may occur useful to combine the formula (1.13) with a method of cluster analysis. First the population (or a given sample S) to be approximated is partitioned into clusters and then the approximation formula is applied independently to different clusters in a "hybrid" way.

In this paper an approach is suggested which unifies the two distinct methods into a single iterative algorithm closely related to the centroid method of cluster analysis. The partition of the approximated population into clusters is optimized simultaneously with the respective approximation formulas in order to minimize the global mean square error. In this way the approximation accuracy increases even if the population does not contain any well separated clusters.

2. THE CENTROID METHOD OF CLUSTER ANALYSIS

The most simple way to approximate a random vector $\mathbf{X}$ is to use a finite number of constants according to a suitable partition of the sample space. Let $D(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_d$ be a decision function on $\mathbb{R}_d$

$$D\colon \mathbb{R}_d \to \{1, 2, \ldots, M\} \tag{2.1}$$

generating a partition of the sample space $\mathbb{R}_d$ into M subsets. Then we can write an approximation formula

$$\mathbf{x}^+ = \mathbf{c}_{D(\mathbf{x})}\,; \quad \mathbf{x} \in \mathbb{R}_d \tag{2.2}$$

where $\mathbf{c}_m \in \mathbb{R}_d$ are constant vectors used to approximate the respective subsets. Denoting $\mathcal{S}_M$ the partition of the sample S induced by the decision function D:

$$\mathcal{S}_M = \{S_1, S_2, \ldots, S_M\}\,; \quad S_m = \{\mathbf{x} \in S\colon D(\mathbf{x}) = m\}\,; \quad m = 1, 2, \ldots, M \tag{2.3}$$

we can express the mean square error of this approximation of S:

$$Q = \frac{1}{N} \sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{c}_{D(\mathbf{x})}\|^2 = \frac{1}{N} \sum_{m=1}^{M} \sum_{\mathbf{x} \in S_m} \|\mathbf{x} - \mathbf{c}_m\|^2\,. \tag{2.4}$$

It is easy to see that for a fixed partition $\mathcal{S}_M$ the function Q is minimized by the respective means of the sets S_m, i.e. by

$$\mathbf{c}_m = \frac{1}{|S_m|} \sum_{\mathbf{x} \in S_m} \mathbf{x}\,; \quad m = 1, 2, \ldots, M\,. \tag{2.5}$$

Consequently, to obtain an optimal approximation (2.2), the criterion $Q = Q(\mathcal{S}_M)$ is to be minimized by a proper choice of the partition $\mathcal{S}_M$.

The criterion (2.4) is often used also to define an optimal partition of a given set S into M clusters. For this purpose the function $Q(\mathcal{S}_M)$ is usually minimized by the following well known "centroid" – or "nearest mean" iterative algorithm:

Step 1: Given a partition $\mathcal{S}_M$ of S compute the cluster centers $\mathbf{c}_m$ by eqs.

$$\mathbf{c}_m = \frac{1}{|S_m|} \sum_{\mathbf{x} \in S_m} \mathbf{x}\,; \quad m = 1, 2, \ldots, M\,. \tag{2.6}$$

Step 2: Using the cluster centers $\mathbf{c}_m$ define a new partition $\mathcal{S}'_M = \{S'_1, S'_2, \ldots, S'_M\}$:

$$S'_m = \{\mathbf{x} \in (S - \bigcup_{k=1}^{m-1} S'_k)\colon \|\mathbf{x} - \mathbf{c}_m\|^2 \leqq \|\mathbf{x} - \mathbf{c}_j\|^2\,,\ j = 1, 2, \ldots, M\} \quad m = 1, 2, \ldots, M\,. \tag{2.7}$$

Here the partition $\mathcal{S}'_M$ in Step 2 can be equivalently specified (cf. (2.3)) by the decision function

$$D(\mathbf{x}) = \min\,\{1 \leqq m \leqq M\colon \|\mathbf{x} - \mathbf{c}_m\|^2 \leqq \|\mathbf{x} - \mathbf{c}_j\|^2;\ j = 1, 2, \ldots, M\}\,; \quad \mathbf{x} \in \mathbb{R}_d \tag{2.8}$$

The centroid method of cluster analysis and its numerous modifications have been considered by many authors (see e.g. the references in [3]). An important advantage of this algorithm is the computational simplicity and also the fact that the function $Q(\mathscr{S}_M)$ is nonincreasing at each iteration of the eqs. (2.6), (2.7). As the number of partitions of the set S is finite, the algorithm converges to a minimum of the nonnegative function $Q(\mathscr{S}_M)$ in a finite number of steps. A more detailed discussion of these questions can be found e.g. in [9]. On the other side the separation of clusters is linear and frequently only a local minimum is achieved – especially when the dimensionality is high.

The centroid algorithm can be derived (cf. [5]) as a particular case of the classification maximum likelihood method suggested by John [7]. Unfortunately this alternative technique of identifying normal mixtures was shown to produce asymptotically biased estimates of the parameters of components [1], [8]. The theoretical justification of the centroid method of cluster analysis is therefore questionable (cf. [5]). The normal mixture model identified by means of the EM algorithm (see e.g. [2], [5], [6], [12]) seems to be preferable as a method of cluster analysis as long as the data set is not extremely large.

It should be emphasized, however, that in approximation problems the very objective is to minimize the mean square error $Q(\mathscr{S}_M)$ whereas in cluster analysis the same criterion is only an intuitively chosen tool to find an optimal partition of the set S into clusters. This formulational difference has some important consequences. Thus, in the context of approximating the linear separation property is unessential as well as any comparison with the finite mixture model. Further, any local solution is equally applicable if the achieved approximation error is sufficiently small. A difficult task in cluster analysis is to determine the true number of cluster M whereas in approximation problems it may be viewed as an input parameter.

3. MULTIMODAL DISCRETE KARHUNEN-LOÈVE EXPANSION

A natural way to approximate multimodal populations is to apply the discrete K.-L. expansion to each mode. However, this approach can be viewed as a generalized centroid method and also analogously optimized.

Particularly, let $\mathscr{S}_M = \{S_1, \ldots, S_M\}$ be a partition of S generated by a decision function $D(\mathbf{x})$. Applying the formula (1.13) to each set $S_m \in \mathscr{S}_M$ we obtain the following approximation:

$$(3.1) \qquad \mathbf{x}^+ = \sum_{m=1}^{M} \delta(m, D(\mathbf{x})) \left[\mathbf{c}_m + \sum_{i=1}^{d_0} y_{mi}(\mathbf{x}) \, \mathbf{v}_{mi}\right] ; \quad y_{mi}(\mathbf{x}) = (\mathbf{x} - \mathbf{c}_m)^{\mathrm{T}} \, \mathbf{v}_{mi} \, .$$

Let us recall (cf. Sec. 1) that the parameters $\mathbf{c}_m, \mathbf{v}_{m1}, \ldots, \mathbf{v}_{md_0}$ minimize the mean square error on the respective subsets S_m. Thus, to obtain optimal parameters in the formula (3.1), the global mean square error given by the equation

$$(3.2)\quad Q = Q(\mathcal{S}_M) = \frac{1}{N} \sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{x}^+(\mathbf{x})\|^2 = \frac{1}{N} \sum_{m=1}^{M} \sum_{\mathbf{x} \in S_m} \|\mathbf{x} - \mathbf{c}_m - \sum_{i=1}^{d_0} y_{mi}(\mathbf{x})\, \mathbf{v}_{mi}\|^2$$

is to be minimized as a function of the partition $\mathcal{S}_M$.

For this purpose we can modify the centroid algorithm as follows:

Step 1: Given a partition $\mathcal{S}_M$ compute the centers

$$(3.3)\qquad \mathbf{c}_m = \frac{1}{|S_m|} \sum_{\mathbf{x} \in S_m} \mathbf{x}\,; \quad m = 1, 2, \ldots, M$$

and construct the bases $V_m = \{\mathbf{v}_{m1}, \mathbf{v}_{m2}, \ldots, \mathbf{v}_{md_0}\}$, $m = 1, 2, \ldots, M$ by choosing the eigenvectors associated with the d_0 largest eigenvalues of the respective covariance matrices A_m:

$$(3.4)\qquad A_m = \frac{1}{|S_m|} \sum_{\mathbf{x} \in S_m} (\mathbf{x} - \mathbf{c}_m)(\mathbf{x} - \mathbf{c}_m)^{\mathrm{T}}\,; \quad m = 1, 2, \ldots, M\,.$$

Step 2: Using the centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M$ and the bases $V_1, V_2, \ldots, V_M$ compute the quantities

$$(3.5)\qquad y_{mi}(\mathbf{x}) = (\mathbf{x} - \mathbf{c}_m)^{\mathrm{T}}\, \mathbf{v}_{mi}\,; \quad \mathbf{x} \in S\,; \quad m = 1, 2, \ldots, M\,; \quad i = d_0 + 1, \ldots, d$$

and define a new partition $\mathcal{S}'_M$ of S by

$$(3.6)\qquad S'_m = \{\mathbf{x} \in (S \setminus \bigcup_{j=1}^{m-1} S'_j)\colon \sum_{i=d_0+1}^{d} y^2_{mi}(\mathbf{x}) \leqq \sum_{i=d_0+1}^{d} y^2_{ji}(\mathbf{x}),\ j = 1, 2, \ldots, M\}$$

$$m = 1, 2, \ldots, M\,.$$

Let us note that the partition $\mathcal{S}'_M$ (cf. (3.6)) may be specified by a related decision function $D(\mathbf{x})$, (cf. (2.3)):

$$(3.7)\quad D(\mathbf{x}) = \min\{1 \leqq m \leqq M\colon \sum_{i=d_0+1}^{d} y^2_{mi}(\mathbf{x}) \leqq \sum_{i=d_0+1}^{d} y^2_{ji}(\mathbf{x}),\ j = 1, 2, \ldots, M\}\,;$$

$$\mathbf{x} \in \mathbb{R}_d$$

or equivalently by the function

$$(3.8)$$

$$D(\mathbf{x}) = \min\{1 \leqq m \leqq M\colon \|\mathbf{x} - \mathbf{c}_m\|^2 - \sum_{i=1}^{d_0} y^2_{mi}(\mathbf{x}) \leqq \|\mathbf{x} - \mathbf{c}_j\|^2 - \sum_{i=1}^{d_0} y^2_{ji}(\mathbf{x}),$$

$$j = 1, \ldots, M\}\,; \quad \mathbf{x} \in \mathbb{R}_d$$

which is advantageous for d_0 small. In some cases it could be also of interest that, applying the formula (1.17) to each subset S_m, we can write

$$(3.9)\qquad Q(\mathcal{S}_M) = \sum_{m=1}^{M} \frac{|S_m|}{N} \left[\frac{1}{|S_m|} \sum_{\mathbf{x} \in S_m} \|\mathbf{x} - \mathbf{x}^+\|^2\right] = \sum_{m=1}^{M} \frac{|S_m|}{N} \sum_{i=d_0+1}^{d} \lambda_{mi}$$

where λ_{mi}, $i = 1, 2, \ldots, d$ are the eigenvalues of the covariance matrix A_m.

One can easily verify that the criterion (3.2) is nonincreasing at each iteration of eqs. (3.3)–(3.6): In the first step the approximation error is minimized independently for each subset S_m by means of the optimal parameters $\mathbf{c}_m$, V_m (cf. Sec. 1). In the second step the approximation error decreases for each reclassified vector and remains unchanged otherwise. Thus, as the number of partitions of the set S is finite, the algorithm converges to a possibly local minimum of the function $Q(\mathscr{S}_M)$ in a finite number of iterations. Let us remark also that, by increasing the number of subsets M, the attainable approximation arror is nonincreasing.

As it appears the classification rule (3.7) was introduced by Watanabe [10] for use in pattern recognition. To characterize this rule let us note that for any vector $\mathbf{x}$ from a subspace spanned by a basis V_m, i.e. for

$$\mathbf{x} = \mathbf{c}_m + \sum_{i=1}^{d_0} \eta_i \mathbf{v}_{mi} \; ; \quad \eta_i \in \mathbb{R}_1 \tag{3.10}$$

it holds

$$\sum_{i=d_0+1}^{d} y_{mi}^2(\mathbf{x}) = \sum_{i=d_0+1}^{d} [(\mathbf{x} - \mathbf{c}_m)^{\mathrm{T}} \mathbf{v}_{mi}]^2 = 0 \, . \tag{3.11}$$

Thus any vector $\mathbf{x}$ satisfying eq. (3.10) will be assigned to the mth class (except for possible ties) even if the distance $\|\mathbf{x} - \mathbf{c}_m\|^2$ is large. The classification rule (3.7) is therefore suitable for separation of classes of elongated form lying in different subspaces. From the point of view of approximation, this question is not of qualitative importance as it may influence only the resulting accuracy.

4. NUMERICAL EXAMPLE

To compare the considered multimodal modifications of the discrete K.-L. expansion we chose the classical iris data of Fisher [4] (see also [11]). The sample includes 150 observations of three different species of iris (iris setosa: $S_1^* = \{\mathbf{x}_1, \ldots \ldots, \mathbf{x}_{50}\}$, iris versicolor: $S_2^* = \{\mathbf{x}_{51}, \ldots, \mathbf{x}_{100}\}$, iris virginica: $S_3^* = \{\mathbf{x}_{101}, \ldots, \mathbf{x}_{150}\}$) with each observation consisting of four measurements: $\mathbf{x} = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$ (x_1 – sepal length, x_2 – sepal width, x_3 – petal length, x_4 – petal width).

First we applied the centroid algorithm (2.6)–(2.7) to comupte an optimal partition of the iris data into three clusters. The best solution we obtained is defined by the centers

$$\begin{aligned} \mathbf{c}_1 &= (5{\cdot}006,\ 3{\cdot}428,\ 1{\cdot}462,\ 0{\cdot}246)^{\mathrm{T}} \, , \\ \mathbf{c}_2 &= (5{\cdot}884,\ 2{\cdot}741,\ 4{\cdot}388,\ 1{\cdot}434)^{\mathrm{T}} \, , \\ \mathbf{c}_3 &= (6{\cdot}854,\ 3{\cdot}077,\ 5{\cdot}715,\ 2{\cdot}054)^{\mathrm{T}} \, , \end{aligned} \tag{4.1}$$

and yields the partition $\mathscr{S}_3$, (cf. (2.7)) which is not identical with the original one:

$$\mathscr{S}_3 = \{S_1, S_2, S_3\} \; ; \quad S_1 = S_1^* \; ; \quad |S_2| = 61 \; ; \quad |S_3| = 39 \; ; \tag{4.2}$$

$$|S_2 \cap S_2^*| = 47 \; ; \quad |S_2 \cap S_3^*| = 14 \; ; \quad |S_3 \cap S_2^*| = 3 \; ; \quad |S_3 \cap S_3^*| = 36 \, .$$

The corresponding value of the mean squares error criterion is

$$Q = \frac{1}{150} \sum_{m=1}^{3} \sum_{\mathbf{x} \in S_m} \|\mathbf{x} - \mathbf{c}_m\|^2 = 0{\cdot}526 \,. \tag{4.3}$$

Several other local minima we found were characterized by significantly higher values of the criterion (4.3).

Accordingly to the "hybrid" approach the standard K.-L. expansion was applied independently to the cluster S_1, S_2 and S_3. First the sample covariance matrix was computed for each of the clusters and then the respective eigenvectors and eigenvalues:

$$\begin{array}{lllll}
\mathbf{v}_{11} = (\;\; 0{\cdot}669, & 0{\cdot}734, & 0{\cdot}096, & 0{\cdot}063)^{\mathrm{T}}, & \lambda_{11} = 0{\cdot}232\,, \\
\mathbf{v}_{12} = (\;\; 0{\cdot}598, & -0{\cdot}621, & 0{\cdot}490, & 0{\cdot}131)^{\mathrm{T}}, & \lambda_{12} = 0{\cdot}036\,, \\
\mathbf{v}_{13} = (-0{\cdot}440, & 0{\cdot}275, & 0{\cdot}832, & 0{\cdot}195)^{\mathrm{T}}, & \lambda_{13} = 0{\cdot}026\,, \\
\mathbf{v}_{14} = (-0{\cdot}036, & -0{\cdot}020, & -0{\cdot}240, & 0{\cdot}970)^{\mathrm{T}}, & \lambda_{14} = 0{\cdot}009\,,
\end{array} \tag{4.4}$$

$$\begin{array}{lllll}
\mathbf{v}_{21} = (\;\; 0{\cdot}531, & 0{\cdot}231, & 0{\cdot}739, & 0{\cdot}344)^{\mathrm{T}}, & \lambda_{21} = 0{\cdot}423\,, \\
\mathbf{v}_{22} = (\;\; 0{\cdot}767, & 0{\cdot}203, & -0{\cdot}403, & -0{\cdot}457)^{\mathrm{T}}, & \lambda_{22} = 0{\cdot}124\,, \\
\mathbf{v}_{23} = (-0{\cdot}253, & 0{\cdot}937, & -0{\cdot}184, & 0{\cdot}154)^{\mathrm{T}}, & \lambda_{23} = 0{\cdot}064\,, \\
\mathbf{v}_{24} = (\;\; 0{\cdot}256, & -0{\cdot}163, & -0{\cdot}509, & 0{\cdot}806)^{\mathrm{T}}, & \lambda_{24} = 0{\cdot}017\,,
\end{array} \tag{4.5}$$

$$\begin{array}{lllll}
\mathbf{v}_{31} = (\;\; 0{\cdot}674, & 0{\cdot}078, & 0{\cdot}727, & 0{\cdot}105)^{\mathrm{T}}, & \lambda_{31} = 0{\cdot}413\,, \\
\mathbf{v}_{32} = (-0{\cdot}384, & 0{\cdot}472, & 0{\cdot}195, & 0{\cdot}769)^{\mathrm{T}}, & \lambda_{32} = 0{\cdot}113\,, \\
\mathbf{v}_{33} = (\;\; 0{\cdot}504, & 0{\cdot}678, & -0{\cdot}535, & -0{\cdot}029)^{\mathrm{T}}, & \lambda_{33} = 0{\cdot}091\,, \\
\mathbf{v}_{34} = (\;\; 0{\cdot}380, & -0{\cdot}559, & -0{\cdot}384, & 0{\cdot}629)^{\mathrm{T}}, & \lambda_{34} = 0{\cdot}035\,.
\end{array} \tag{4.6}$$

The approximation based on the centers (4.1) and the respective first two eigenvectors of (4.4)–(4.6), ($d_0 = 2$, cf. (3.1)) yield the mean square error

$$Q = \frac{1}{150} \sum_{m=1}^{3} \sum_{\mathbf{x} \in S_m} \|\mathbf{x} - \mathbf{c}_m - y_{m1}(\mathbf{x})\, \mathbf{v}_{m1} - y_{m2}(\mathbf{x})\, \mathbf{v}_{m2}\|^2 = 0{\cdot}077 \tag{4.7}$$

Obviously the result (4.7) may be influenced by the chosen method of cluster analysis. Thus, to obtain another independent solution, the method of mixtures was applied. Particularly, the normal mixture with three components

$$f(\mathbf{x}) = \sum_{m=1}^{3} \frac{w_m}{\sqrt{((2\pi)^4 \det A_m)}} \exp\left\{-\tfrac{1}{2}(\mathbf{x} - \mathbf{c}_m)^{\mathrm{T}} A_m^{-1}(\mathbf{x} - \mathbf{c}_m)\right\}, \quad \mathbf{x} \in \mathbb{R}_4 \tag{4.8}$$

was fitted to the sample. Using the EM algorithm we obtained the following m.-l. estimates of the parameter w_m, $\mathbf{c}_m$ and A_m:

(4.9)

$$\begin{array}{ll}
w_1 = {\cdot}333, & \mathbf{c}_1 = (5{\cdot}006,\ 3{\cdot}428,\ 1{\cdot}462,\ 0{\cdot}246)^{\mathrm{T}}, \\
w_2 = {\cdot}299, & \mathbf{c}_2 = (5{\cdot}915,\ 2{\cdot}778,\ 4{\cdot}202,\ 1{\cdot}297)^{\mathrm{T}}, \\
w_3 = {\cdot}368, & \mathbf{c}_3 = (6{\cdot}644,\ 2{\cdot}949,\ 5{\cdot}480,\ 1{\cdot}985)^{\mathrm{T}},
\end{array}
\qquad
A_1 = \begin{pmatrix} {\cdot}122 & {\cdot}097 & {\cdot}016 & {\cdot}010 \\ {\cdot}097 & {\cdot}141 & {\cdot}011 & {\cdot}009 \\ {\cdot}016 & {\cdot}011 & {\cdot}030 & {\cdot}006 \\ {\cdot}010 & {\cdot}009 & {\cdot}006 & {\cdot}011 \end{pmatrix};$$

$$A_2 = \begin{pmatrix} \cdot 275 & \cdot 097 & \cdot 185 & \cdot 054 \\ \cdot 097 & \cdot 093 & \cdot 091 & \cdot 043 \\ \cdot 185 & \cdot 091 & \cdot 201 & \cdot 061 \\ \cdot 054 & \cdot 043 & \cdot 061 & \cdot 032 \end{pmatrix} ; \quad A_3 = \begin{pmatrix} \cdot 378 & \cdot 092 & \cdot 303 & \cdot 061 \\ \cdot 092 & \cdot 110 & \cdot 084 & \cdot 056 \\ \cdot 303 & \cdot 084 & \cdot 328 & \cdot 074 \\ \cdot 061 & \cdot 056 & \cdot 074 & \cdot 086 \end{pmatrix} \tag{4.10}$$

and further, by means of the Bayes decision rule, the corresponding partition of S:

$$\mathscr{S}'_3 = \{S'_1, S'_2, S'_3\} ; \quad S'_1 = S^*_1 ; \quad |S'_2| = 45 ; \quad |S'_3| = 55 ; \tag{4.11}$$
$$|S'_2 \cap S^*_2| = 45 ; \quad |S'_2 \cap S^*_3| = 0 ; \quad |S'_3 \cap S^*_2| = 5 ; \quad |S'_3 \cap S^*_3| = 50 .$$

Again, according to the hybrid approach, we computed the sample means and sample covariance matrices for each of the clusters S'_1, S'_2, S'_3 (they are generally different from (4.9), (4.10)) and then the respective eigenvectors and eigenvalues:

$$\begin{aligned} \mathbf{c}_1 &= (5{\cdot}006,\ 3{\cdot}428,\ 1{\cdot}462,\ 0{\cdot}246)^{\mathrm{T}} , \\ \mathbf{c}_2 &= (5{\cdot}904,\ 2{\cdot}776,\ 4{\cdot}193,\ 1{\cdot}293)^{\mathrm{T}} , \\ \mathbf{c}_3 &= (6{\cdot}554,\ 2{\cdot}951,\ 5{\cdot}489,\ 1{\cdot}989)^{\mathrm{T}} , \end{aligned} \tag{4.12}$$

$$\begin{aligned} \mathbf{v}_{11} &= (\ 0{\cdot}669,\ \ 0{\cdot}734,\ \ 0{\cdot}096,\ \ 0{\cdot}064)^{\mathrm{T}} , & \lambda_{11} &= 0{\cdot}232 , \\ \mathbf{v}_{12} &= (\ 0{\cdot}598,\ -0{\cdot}621,\ \ 0{\cdot}490,\ \ 0{\cdot}131)^{\mathrm{T}} , & \lambda_{12} &= 0{\cdot}036 , \\ \mathbf{v}_{13} &= (-0{\cdot}440,\ \ 0{\cdot}275,\ \ 0{\cdot}832,\ \ 0{\cdot}195)^{\mathrm{T}} , & \lambda_{13} &= 0{\cdot}026 , \\ \mathbf{v}_{14} &= (-0{\cdot}036,\ -0{\cdot}020,\ -0{\cdot}240,\ \ 0{\cdot}970)^{\mathrm{T}} , & \lambda_{14} &= 0{\cdot}009 , \end{aligned} \tag{4.13}$$

$$\begin{aligned} \mathbf{v}_{21} &= (\ 0{\cdot}712,\ \ 0{\cdot}335,\ \ 0{\cdot}587,\ \ 0{\cdot}189)^{\mathrm{T}} , & \lambda_{21} &= 0{\cdot}476 , \\ \mathbf{v}_{22} &= (-0{\cdot}654,\ \ 0{\cdot}579,\ \ 0{\cdot}356,\ \ 0{\cdot}332)^{\mathrm{T}} , & \lambda_{22} &= 0{\cdot}065 , \\ \mathbf{v}_{23} &= (-0{\cdot}246,\ -0{\cdot}665,\ \ 0{\cdot}701,\ -0{\cdot}075)^{\mathrm{T}} , & \lambda_{23} &= 0{\cdot}038 , \\ \mathbf{v}_{24} &= (\ 0{\cdot}070,\ -0{\cdot}332,\ -0{\cdot}192,\ \ 0{\cdot}921)^{\mathrm{T}} , & \lambda_{24} &= 0{\cdot}007 , \end{aligned} \tag{4.14}$$

$$\begin{aligned} \mathbf{v}_{31} &= (\ 0{\cdot}714,\ \ 0{\cdot}220,\ \ 0{\cdot}645,\ \ 0{\cdot}159)^{\mathrm{T}} , & \lambda_{31} &= 0{\cdot}683 , \\ \mathbf{v}_{32} &= (-0{\cdot}292,\ \ 0{\cdot}723,\ -0{\cdot}076,\ \ 0{\cdot}622)^{\mathrm{T}} , & \lambda_{32} &= 0{\cdot}111 , \\ \mathbf{v}_{33} &= (-0{\cdot}547,\ -0{\cdot}404,\ \ 0{\cdot}671,\ \ 0{\cdot}295)^{\mathrm{T}} , & \lambda_{33} &= 0{\cdot}057 , \\ \mathbf{v}_{34} &= (\ 0{\cdot}324,\ -0{\cdot}516,\ -0{\cdot}357,\ \ 0{\cdot}708)^{\mathrm{T}} , & \lambda_{34} &= 0{\cdot}034 . \end{aligned} \tag{4.15}$$

The approximation based on the first two eigenvectors yields the mean square error

$$Q(\mathscr{S}'_3) = 0{\cdot}059 . \tag{4.16}$$

Finally the generalized centroid algorithm (3.3)−(3.6) was applied to compute an optimal multimodal discrete K.-L. expansion. The best solution we obtained is defined by the centers

$$\begin{aligned} \mathbf{c}_1 &= (5{\cdot}006,\ 3{\cdot}428,\ 1{\cdot}462,\ 0{\cdot}246)^{\mathrm{T}} , \\ \mathbf{c}_2 &= (6{\cdot}157,\ 2{\cdot}870,\ 4{\cdot}768,\ 1{\cdot}549)^{\mathrm{T}} , \\ \mathbf{c}_3 &= (6{\cdot}381,\ 2{\cdot}874,\ 5{\cdot}062,\ 1{\cdot}819)^{\mu} , \end{aligned} \tag{4.17}$$

and vectors

$$\begin{array}{lll}
 & \mathbf{v}_{11} = (\;\; 0{\cdot}669, \;\; 0{\cdot}734, \;\; 0{\cdot}096, \;\; 0{\cdot}064)^{\mathrm{T}}, & (\lambda_{11} = 0{\cdot}232) \\
(4.18) & \mathbf{v}_{12} = (\;\; 0{\cdot}598, \; -0{\cdot}621, \;\; 0{\cdot}490, \;\; 0{\cdot}131)^{\mathrm{T}}, & (\lambda_{12} = 0{\cdot}036) \\
 & \mathbf{v}_{13} = (-0{\cdot}440, \;\; 0{\cdot}275, \;\; 0{\cdot}832, \;\; 0{\cdot}195)^{\mathrm{T}}, & (\lambda_{13} = 0{\cdot}026) \\
 & \mathbf{v}_{14} = (-0{\cdot}036, \; -0{\cdot}020, \; -0{\cdot}240, \;\; 0{\cdot}970)^{\mathrm{T}}, & (\lambda_{14} = 0{\cdot}009)
\end{array}$$

$$\begin{array}{lll}
 & \mathbf{v}_{21} = (\;\; 0{\cdot}558, \;\; 0{\cdot}255, \;\; 0{\cdot}735, \;\; 0{\cdot}288)^{\mathrm{T}}, & (\lambda_{21} = 0{\cdot}943) \\
(4.19) & \mathbf{v}_{22} = (\;\; 0{\cdot}716, \;\; 0{\cdot}256, \; -0{\cdot}447, \; -0{\cdot}471)^{\mathrm{T}}, & (\lambda_{22} = 0{\cdot}185) \\
 & \mathbf{v}_{23} = (-0{\cdot}073, \;\; 0{\cdot}646, \; -0{\cdot}417, \;\; 0{\cdot}636)^{\mathrm{T}}, & (\lambda_{23} = 0{\cdot}031) \\
 & \mathbf{v}_{24} = (\;\; 0{\cdot}414, \; -0{\cdot}673, \; -0{\cdot}292, \;\; 0{\cdot}539)^{\mathrm{T}}, & (\lambda_{24} = 0{\cdot}018)
\end{array}$$

$$\begin{array}{lll}
 & \mathbf{v}_{31} = (\;\; 0{\cdot}561, \;\; 0{\cdot}144, \;\; 0{\cdot}749, \;\; 0{\cdot}321)^{\mathrm{T}}, & (\lambda_{31} = 1{\cdot}340) \\
(4.20) & \mathbf{v}_{32} = (-0{\cdot}347, \;\; 0{\cdot}794, \; -0{\cdot}102, \;\; 0{\cdot}489)^{\mathrm{T}}, & (\lambda_{32} = 0{\cdot}144) \\
 & \mathbf{v}_{33} = (\;\; 0{\cdot}750, \;\; 0{\cdot}225, \; -0{\cdot}621, \;\; 0{\cdot}036)^{\mathrm{T}}, & (\lambda_{33} = 0{\cdot}033) \\
 & \mathbf{v}_{34} = (-0{\cdot}046, \; -0{\cdot}546, \; -0{\cdot}207, \;\; 0{\cdot}810)^{\mathrm{T}}, & (\lambda_{34} = 0{\cdot}022)
\end{array}$$

The corresponding partition $\mathscr{S}_3''$ induced by the parameters (4.17)–(4.20), (cf. (3.6)) has the properties

$$\begin{array}{ll}
(4.21) & \mathscr{S}_3'' = \{S_1'', S_2'', S_3''\}\;; \quad S_1'' = S_1^*\;; \quad |S_2''| = 53\;; \quad |S_3''| = 47\;; \\
 & |S_2'' \cap S_2^*| = 33\;; \quad |S_2'' \cap S_3^*| = 20\;; \quad |S_3'' \cap S_2^*| = 14\;; \quad |S_3'' \cap S_3^*| = 30\,.
\end{array}$$

The approximation based on the first two eigenvectors ($d_0 = 2$) of (4.18)–(4.20) yields the mean square error

(4.22) $$Q(\mathscr{S}_3'') = 0{\cdot}047\,.$$

It is easy to see that the approximation error (4.22) is always less than (or equal to) that obtained by a hybrid approach (cf. (4.7), (4.16)) but the actual difference is data dependent for obvious reasons.

5. CONCLUSION

The use of the suggested multimodal discrete K.-L. expansion may be expected to be efficient especially when the underlying probability distribution really shows well separated modes. However in other problems the number of components M and the number of vectors d_0 may be connected with some technical limitations. In such a case the two numbers may be viewed as input parameters to be chosen as large possible. More generally, if only the total number of vectors is bounded, e.g. because of a limited storage capacity available, we could improve the approximation accuracy by allowing the number of vectors d_0 to be different in different subsets. In view of the formula (3.9) the optimal vectors can be determined in this case by ordering the quantities $|S_m|\,\lambda_{mi}$.

(Received April 23, 1985.)

REFERENCES

[1] P. Bryant and J. A. Williamson: Asymptotic behaviour of classification maximum likelihood estimates. Biometrika *65* (1978), 2, 273—281.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. *B 39* (1977), 1—38.

[3] R. O. Duda and P. E. Hart: Pattern Classification and Scene Analysis. John Wiley, New York—London 1973.

[4] R. A. Fisher: The use of multiple measurements in taxonomic problems. Ann. Eugenics *7* (1936), p. 179.

[5] J. Grim: Metody shlukové analýzy a jejich využití při zpětnovazebním řízení velkých systémů (Methods of Cluster Analysis and their Application for Feedback Control of Large Systems). Ph. D. Dissertation, Institute of Information Theory and Automation, Prague 1979.

[6] J. Grim: On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions. Kybernetika *18* (1982), 3, 173—190.

[7] S. John: On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics *12* (1970), 3, 553—563.

[8] F. H. C. Marriott: Separating mixtures of normal distributions. Biometrics *31* (1975), 767 to 769.

[9] S. Z. Selim and M. A. Ismail: K-means-type algorithms: A general convergence theorem and characterization of local optimality. IEEE Trans. Pattern Analysis Machine Intelligence *PAMI-6* (1984), 1, 81—86.

[10] S. Watanabe: Karhunen-Loève expansion and factor analysis. In: Trans. Fourth Prague Conf. on Information Theory, Academia, Prague 1967, 635—660.

[11] S. Wold: Pattern recognition by means of disjoint principal components models. Pattern Recognition *8* (1976), 3, 127—139.

[12] J. H. Wolfe: Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research *5* (1970), 329—350.

Ing. Jiří Grim, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia