ELSEVIER

Chemometrics and Intelligent Laboratory Systems 40 (1998) 33–52

Chemometrics and intelligent laboratory systems

Extraction of representative subsets by potential functions method and genetic algorithms

C. Pizarro Millán [a], M. Forina [b,*], C. Casolino [b], R. Leardi [b]

[a] Departamento de Quimica, Universidad de La Rioja, c/Obispo Bustamante, 3, E-26001 Logroño, Spain
[b] Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, Università degli Studi di Genova, Via Brigata Salerno (ponte), I-16147 Genoa, Italy

Received 12 May 1997; revised 4 August 1997; accepted 27 October 1997

Abstract

Two procedures are suggested to select a representative subset from a large data set. The first is based on the use of the estimate of the multivariate probability density distribution by means of the potential functions technique. The first object selected for the subset is that for which the probability density is larger. Then, the distribution is corrected, by subtraction of the contribution of the selected object multiplied by a selection factor. The second procedure uses genetic algorithms to individuate the subset that reproduces the variance–covariance matrix with the minimum error. Both methods meet the requirement to obtain a representative subset, but the results obtained with the method based on potential functions are generally more satisfactory in the case when the original set is not a random sample from an infinite population, but is the finite population itself. Several examples show how the extraction of a representative subset from a large data set can give some advantages in the use of representation techniques (i.e., eigenvector projection, non-linear maps, Kohonen maps) and in class modelling techniques. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Object selection; Potential functions; Genetic algorithms

1. Introduction

Often in clinical, environmental, food chemistry and in quality control many samples are analysed, frequently with the objective of identification (classification and class modelling). The selection of samples to be analysed among a lot of candidate samples is generally a difficult task. The visual representation, elaboration and sometimes storage of the huge amounts of data collected on analysed samples can give some drawbacks. Visual representation, where the objective is to obtain a global information about the data set, and specially about the discriminating ability of the chemical information, can be less efficient when many hundreds of objects are represented. Moreover, it may happen that the statistical sample has not the same size for all the categories in the problem, so that the plots are unbalanced.

In these cases it is necessary to extract a subsample from the statistical sample or from a part of the statistical sample (e.g., from a category). In the case

[*] Corresponding author. Fax: +39-10-3532684; e-mail: forina@anchem.unige.it.

0169-7439/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII S0169-7439(97)00080-4

of a completely random sample this is a, generally, bad representation of the population; the common procedure of obtaining a subset by random extraction from it can therefore be considered acceptable. Frequently the original sample can not be considered as a random sample: some known or unknown factors may have selected levels, and only a small part of the dispersion in the objects is a consequence of random errors. The most important part of dispersion is the consequence of the level of the above factors. This is the case of the selection of samples for analysis among candidate samples, e.g., in a real population of a wine, the producers (described by parameters as geographical co-ordinates, exposition of the vineyard, age of vines, percentage of the cultivars in the wine, $\cdots$) are an example of a factor at some fixed levels. The dispersion in the original sample gives information about the producers. So, in the case of real samples, the subsample must represent at best the original sample; this means that the multivariate probability distribution estimated from the subsample or the parameters describing it (moments: means, variances, covariances) must be as close as possible to those estimated from the original sample. The extraction of a subset for this kind of real problems (where each chemical sample in the subset must be analysed to obtain the required chemical information) is the main objective of this work.

Two techniques have been developed here. The first one is based on the fit between distributions estimated using the potential functions method from the original sample and the subsample; the second one applies the genetic algorithms with the goal of obtaining a subset whose variance–covariance matrix is as similar as possible to the variance–covariance matrix of the original sample.

Several techniques for the selection of samples are already available, but none of them has the goal of obtaining a subset whose distribution is as close as possible to the original distribution. When the objective is to obtain a subset for calibration the selected samples in the subset must span the whole data domain. In this case algorithms as those used to obtain a D-optimal subset [1] or the Kennard–Stone algorithm [2] can be applied. When the goal is to select one object in each of the ‘interesting regions’ of the data space, techniques based on clustering [3–6] can be applied.

2. Theory

2.1. Potential functions

Potential functions (PF) were born in the early 1950s as a classification method under the name of kernel density estimators. The best-known method for PF in analytical chemistry was written by Coomans and Broeckaert [7]. PF have been also transformed to obtain a class-modelling technique [8] and they are the basis for the CLUPOT clustering procedure [9].

PF evaluate the probability density function by using the local density of the objects instead of parameters such as the mean and standard deviation. The probability function is calculated as the sum of the individual contribution of the objects in the training set.

In fixed-potential PF the shape of the individual contributions is the same for all the objects in the class. In variable-potential PF the shape of the contribution depends on the local density of the objects. In this work fixed-potential PF were used.

The individual contribution can have different shapes. The one commonly used has the form of the Gaussian function. In the univariate case it is given by the following expression:

$$\phi(x, x_i) = \frac{1}{I}\frac{1}{u\sqrt{2\pi}}\mathrm{e}^{-(1/2)[(x-x_i)/u]^2} \tag{1}$$

where $\phi(x, x_i)$ is the contribution of object i to the probability density evaluated in the point x; x_i is the value of the variable X for the object i; I is the number of objects used to obtain the estimate of the probability density function (training set); u is the smoothing parameter; it is analogous to the standard deviation in the normal distribution and determines how broad the individual contribution is.

The probability distribution function in a point x is the sum of the individual contributions

$$f(x) = \sum_{i=1}^{I} \phi(x, x_i) \tag{2}$$

The selection of the smoothing parameter is performed by a cross-optimisation procedure (e.g., leave-one-out). Let be $f(x_j)^* = \sum_{i \neq j}\phi(x_j, x_i)$ the probability density computed in x_j, when object j is left out, by the contribution of the $I-1$ objects in the training set. The first value of the smoothing parameter is the standard deviation of the variable. Then, it decreases.

The optimum smoothing parameter is that for which the product of the densities $P = \prod_{j=1}^{I} f(x_j)^*$ computed for the objects in the evaluation set is maximum.

In the case of multivariate data, the individual potential must take into account the V variables, and the probability distribution function in a point x is

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{I} \phi(\boldsymbol{x}, \boldsymbol{x}_i) \\ &= \sum_{i=1}^{I} \left(\frac{1}{I} \prod_{v=1}^{V} \frac{1}{ks_v\sqrt{2\pi}} \mathrm{e}^{-(1/2)[(x_v - x_{iv})/(ks_v)]^2} \right) \end{aligned} \tag{3}$$

where $f(\boldsymbol{x})$ is the probability density or cumulative potential in the point of co-ordinates $\boldsymbol{x}$; $\phi(\boldsymbol{x}, \boldsymbol{x}_i)$ is the individual contribution of object $\boldsymbol{x}_i$ in the point $\boldsymbol{x}$, k is the smoothing coefficient; s_v is the standard deviation of the variable v.

In Eq. (3) the individual contribution is obtained as the product of the marginal contributions of the V variables, so that it has the appearance of a multivariate normal distribution of non-correlated variables. However, in spite of the non-correlated individual contribution, the resulting density distribution obtains the correlation between variables as the result of the position of the objects in the space.

The smoothing parameter in Eq. (1) is substituted in the multivariate case of Eq. (3) by the product of the smoothing factor for the standard deviation of the variable, so that only one parameter, k, is optimised, instead of as many smoothing parameters as there are variables.

When more categories are studied, the number of objects I, the smoothing factor k, the standard deviations s_v are generally computed for each category.

2.2. Potential functions applied to object selection

The selection of a subset of M objects from a sample of $I > M$ objects is performed as follows:

(a) The cumulative potential $f(\boldsymbol{x}_j)$

$$f(\boldsymbol{x}_j) = \sum_{i=1}^{I} \phi(\boldsymbol{x}_j, \boldsymbol{x}_i) \tag{4}$$

is computed for each object j.

(b) The object with the highest cumulative potential is selected.

(c) The potential in each point is recalculated after subtraction of the individual contribution of the selected object k multiplied by a selection factor r; r usually is

$$r = I/M \tag{5}$$

i.e., the total number of objects I divided by the number of objects M to be selected. So:

$$f(\boldsymbol{x})^{\text{new}} = \sum_{i=1}^{I} \phi(\boldsymbol{x}, \boldsymbol{x}_i) - r\phi(\boldsymbol{x}, \boldsymbol{x}_k) \tag{6}$$

or generally with $L \leq M$ selected objects:

$$f(\boldsymbol{x})^{\text{new}} = \sum_{i=1}^{I} \phi(\boldsymbol{x}, \boldsymbol{x}_i) - r \sum_{l=1}^{L} \phi(\boldsymbol{x}, \boldsymbol{x}_{k(l)}) \tag{7}$$

where l is the order of selection and $k(l)$ is the index of the selected object.

In the new step (b) the object with the highest new potential is then selected as the second of the subset.

Steps (b) and (c) are repeated till the predetermined number M of objects has been selected.

When M objects have been selected:

$$\begin{aligned} f(\boldsymbol{x})^{\text{final}} &= \sum_{i=1}^{I} \phi(\boldsymbol{x}, \boldsymbol{x}_i) - r \sum_{l=1}^{M} \phi(\boldsymbol{x}, \boldsymbol{x}_{k(l)}) \\ &= f(\boldsymbol{x})^{\text{original}} - \frac{I}{M} \sum_{l=1}^{M} \\ &\quad \times \left(\frac{1}{I} \prod_{v=1}^{V} \frac{1}{ks_v\sqrt{2\pi}} \mathrm{e}^{-(1/2)[(x_v - x_{k(l),v})/ks_v]^2} \right) \\ &= f(\boldsymbol{x})^{\text{original}} - f(\boldsymbol{x})^{\text{selection}} \end{aligned} \tag{8}$$

This equation justifies the choice of the value of the selection factor r in Eq. (5); the residual density $f(\boldsymbol{x})^{\text{final}}$ would be null when $f(\boldsymbol{x})^{\text{original}} = f(\boldsymbol{x})^{\text{selection}}$.

However, the suitable smoothing coefficient for the subset must be evaluated after the selection. The value of k for the subset differs from that obtained with the original sample, so that $f(\boldsymbol{x})^{\text{subset}}$ is not equal to $f(\boldsymbol{x})^{\text{selection}}$.

So r has not the constraint that the sum of individual contribution of the selected objects indicated as $f(\boldsymbol{x})^{\text{selection}}$ gives a probability function with area 1. The value of r chosen for the selection step can be larger than I/M when it is advisable to favour the representation of low-density parts of the original distribution; it can be smaller than I/M when objects in low-density space are considered non-representative or outliers. A large value of r favours negative values in $f(\boldsymbol{x})^{\text{new}}$ and especially in $f(\boldsymbol{x})^{\text{final}}$, but $f(\boldsymbol{x})^{\text{subset}}$ is computed from Eq. (3) as $f(\boldsymbol{x})^{\text{original}}$, with M instead of I and with a different value of k, so that it is always non-negative.

The degree of fit between $f(\boldsymbol{x})^{\text{subset}}$ and $f(\boldsymbol{x})^{\text{original}}$ is the measure of the goodness of the selected subset. The possibility to choose a value of r different from I/M introduces some subjectivity in the evaluation of the degree of fit.

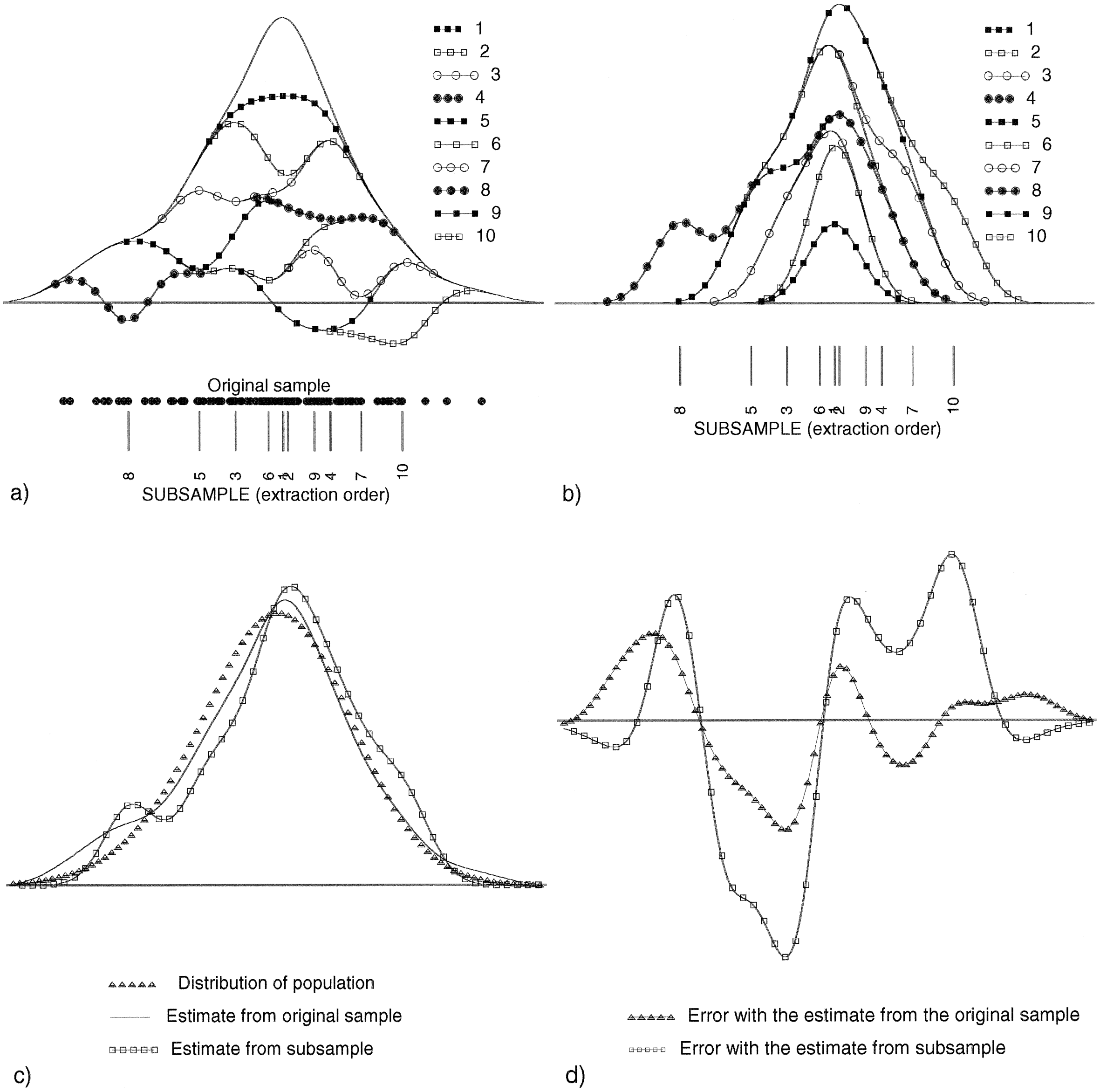


Fig. 1. Evolution with the extraction of objects for the subsample in an univariate case: (a) $f(\boldsymbol{x})^{\text{new}}$ (Eq. (6)), (b) $f(\boldsymbol{x})^{\text{selection}}$ (Eq. (8)), (c) $f(\boldsymbol{x})^{\text{true}}$, $f(\boldsymbol{x})^{\text{original}}$, $f(\boldsymbol{x})^{\text{subset}}$, (d) $f(\boldsymbol{x})^{\text{original}} - \text{f}(\boldsymbol{x})^{\text{true}}$, $f(\boldsymbol{x})^{\text{subset}} - f(\boldsymbol{x})^{\text{true}}$.

Fig. 1a shows how the density of probability function $f(\boldsymbol{x})^{\text{new}}$ is modified by selecting the objects and Fig. 1b how $f(\boldsymbol{x})^{\text{original}}$ is reconstructed as $f(\boldsymbol{x})^{\text{selection}}$ from the selected subset. Only the parts of the probability functions modified by the selection are marked in Fig. 1.

In this case a subset of 10 objects was selected from a sample of 100 objects (data set FIRST, see below) extracted from a population with probability density $f(\boldsymbol{x})^{\text{true}}$. The reconstruction has been obtained with the individual contributions of the selected objects in the original sample multiplied by $r = 10$ $(I/M = 100/10)$. Smoothing coefficient was 0.4, obtained by leave-one-out procedure.

Fig. 1c shows $f(\boldsymbol{x})^{\text{true}}$, $f(\boldsymbol{x})^{\text{original}}$ and $f(\boldsymbol{x})^{\text{final}}$; Fig. 1d shows on a magnified scale the difference between the two estimates of the probability density and $f(\boldsymbol{x})^{\text{true}}$.

2.3. Genetic algorithms

Genetic algorithms (GA) are inspired by evolution theory, according to which the evolution of a species is mainly ruled by the ‘struggle for life’. Under this principle the ‘best’ individuals (i.e., those beings whose genetic material is the best for the environmental conditions in which they live) have both the greatest probabilities of surviving and the greatest probabilities of winning the battles engaged for reproduction, thereby propagating their genome. Furthermore, when two good individuals mate, the combination of their genomes can generate offspring with even better genetic material. As a result, from this particular point of view a population evolves in such a way that its average fitness to the environment after generation $i+1$ is usually higher than after generation i.

Another source of variation is given by mutations. They are irregular changes with a low probability of occurrence and they affect a ‘letter’ of the coding gene. They generally result in a pathological condition, but sometimes they can produce a better individual.

GA have four basic steps: (1) coding of variables; (2) initiation of population; (3) evaluation of the responses; (4) creation of population $i+1$ from population i (select-copy, cross-over, mutation). Steps 3 and 4 alternate until a termination criterion is reached; this criterion can be based on a lack of improvement in the response or simply on a maximum number of generations or on the time allowed for the elaboration.

The basic GA can nowadays be easily found in literature [10,11]. It has anyway to be said that the ‘optimal’ configuration of the algorithm changes very much according to the very specific problem to which it is applied. In Section 2.4 the peculiarities of the GA for subset selection will be described.

2.4. Genetic algorithms applied to subset selection

GA has been applied to the problem of selecting the experimental conditions to obtain a D-optimal design [12]. In this case, the candidate points can be considered as objects, and there too the problem is to select a subset of them which has to be as informative as possible from the experimental design point of view, i.e., spanning as much as possible the experimental domain. GA has also been applied to the selection of variables [13]. This problem is very similar to the selection of representative subsets; in the former, the GA works on the columns of the data matrix, while in the latter it works on the rows.

2.4.1. Coding

In a data set with K objects, each chromosome is formed by K genes coded by a single bit (0 = absent, 1 = present).

2.4.2. Response

One of the most critical steps in GA is the definition of the response to be maximised. While in some cases its identification is straightforward (e.g., when trying to find the maximum of a mathematical function), in some other cases it is not so easy.

What we want to obtain is the selection of a subset of objects being as representative as possible of the original data set.

Two similar subsets drawn from the same population have the same variance–covariance matrix (from a ‘geometrical’ point of view, one can say that the variances describe the dimension of the swarm of points, while the covariances describe its orientation). Goal of the procedure is therefore to obtain a subset whose variance–covariance matrix $(\mathbf{C}_s)$ is as similar

Table 1
Ratio (of the estimate of the standard deviation obtained from the PF-subset [10 objects] and the estimate obtained from the original set [100 objects]) and difference (between the two estimates of the mean), as a function of the selection factor r ($k=0.4$)

(A) Data set FIRST

r	$r/(I/M)$	ratio	diff.
5	0.50	0.5584	0.1452
6	0.60	0.5601	0.1762
7	0.70	0.6887	0.2140
8	0.80	0.7884	0.1329
9	0.90	0.9448	0.1348
10	1.00	0.9724	0.1893
11	1.10	1.0706	−0.1758
12	1.20	1.2088	−0.1995
13	1.30	1.1398	−0.1854
14	1.40	1.2884	−0.2289
15	1.50	1.5678	0.1439

(B) Data sets UNIMODAL, UNIMODAL1, $\cdots$

r	UNIMODAL		UNIMODAL1		UNIMODAL2		UNIMODAL3	
	ratio	diff.	ratio	diff.	ratio	diff.	ratio	diff.
5	0.617	0.140	0.636	0.150	0.739	0.010	0.429	−0.096
6	0.800	−0.100	0.744	0.127	0.740	−0.006	0.538	−0.123
7	0.761	−0.018	0.723	0.149	0.795	−0.016	0.697	−0.125
8	0.872	0.066	0.787	0.100	0.832	−0.020	0.796	−0.129
9	0.953	0.041	0.896	0.042	0.871	−0.007	0.964	−0.140
10	1.030	−0.131	0.862	−0.091	0.862	0.033	1.006	−0.102
11	1.077	−0.137	0.931	−0.075	0.924	0.066	1.157	−0.122
12	1.048	0.178	0.977	−0.025	0.976	0.098	1.224	−0.115
13	1.058	0.111	0.981	−0.049	1.367	−0.319	1.216	−0.064
14	1.100	0.091	1.536	−0.400	1.459	−0.219	1.482	−0.282
15	1.182	0.144	1.618	−0.393	1.461	−0.237	1.610	−0.214

as possible to the variance–covariance matrix of the original data set ($\mathbf{C}_x$). The similarity is computed as the inverse of the sum of the squared elements of the difference matrix **D**:

$$\mathbf{D} = \mathbf{C}_x - \mathbf{C}_\mathrm{s} \tag{9}$$

For a data set with V variables,

$$\text{similarity} = 1 \Bigg/ \sum_{i=1}^{V} \sum_{j=1}^{V} d_{ij}^2 \tag{10}$$

A penalty function related to the number of selected object has also been added, so that in case of two subsets with the same similarity the one with less objects is favoured. Since this penalty function must be very mild, not to be too much influent in the search of the solution, the decimal logarithm of the number of objects in the subset (n) has been used.

The response to be maximised by the GA is therefore:

$$\text{response} = \frac{\text{similarity}}{\log(\text{m})} \tag{11}$$

2.4.3. Initiation of population

According to the original algorithm, the value of each bit is determined by the 'toss of a coin'. Under

Table 2
Fractional factorial design (2^{4-1}) for the influential parameters in GA

Variables	−	+
No. initial obj.	10	30
No. chromosomes	20	50
%Elitism	25	50
%Mutation	1	2

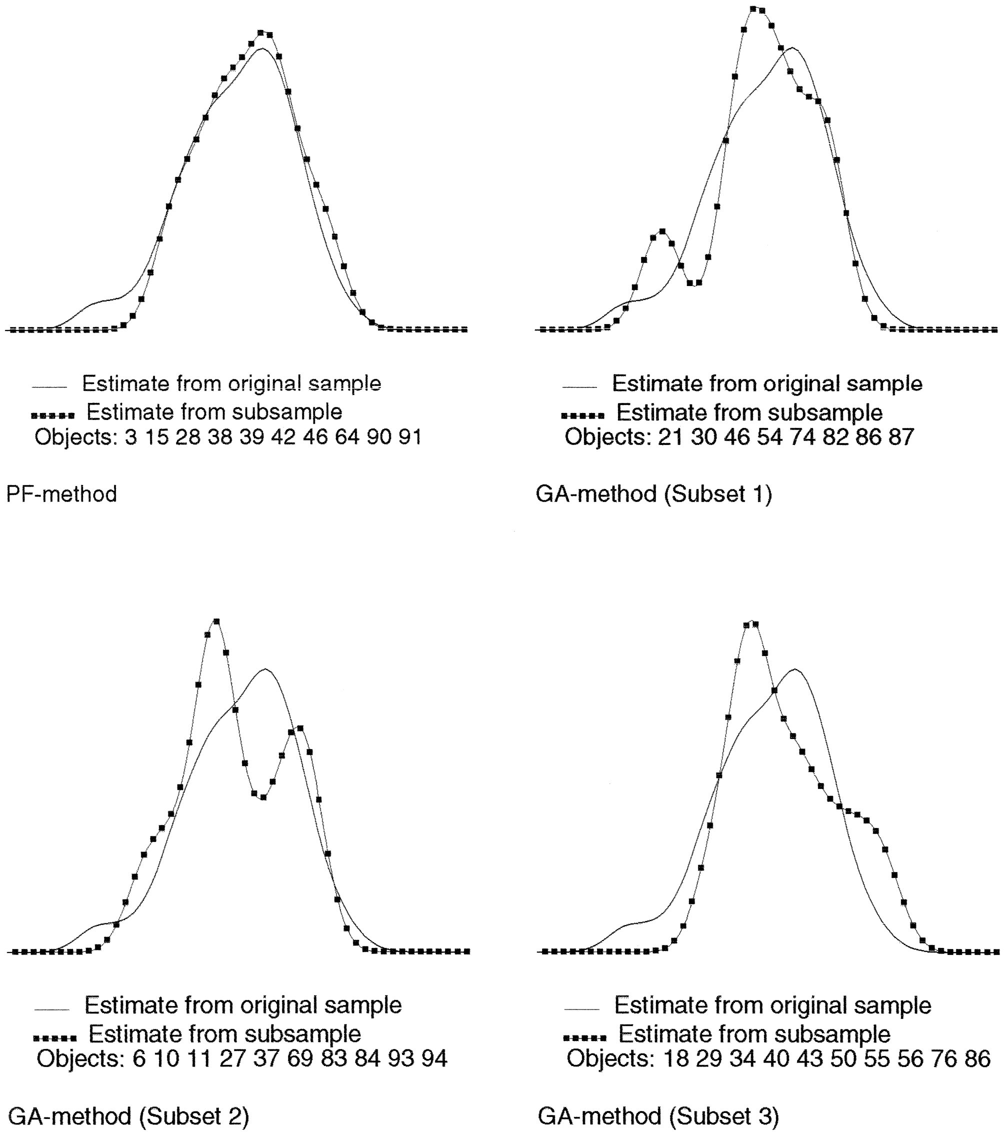


Fig. 2. Probability density function estimated from the UNIMODAL set and from the subsets ($f(x)^{\text{subset}}$) obtained by PF method and by 3 repetitions of GA method.

this hypothesis an average of 50% of the objects would be selected in every initial chromosome. This can lead to a subset with a very high number of objects.

An initial probability of selection has thus been added so that the average number of objects present in the initial population can be selected.

The maximum number of objects present in each combination is asked for. This constraint will be retained also in the subsequent steps. If the number of objects in a chromosome is higher than this number its response will be zero.

2.4.4. Select-copy

The probability for a chromosome of being selected and copied into the following population is:

$$\text{prob}(i) = \text{response}(i)/\text{sum}(\text{responses}) \qquad (12)$$

2.4.5. Cross-over

The uniform cross-over has been used (for each gene a random number determines if it will undergo cross-over).

2.4.6. Elitism

To avoid the loss of highly informative solutions the e bests chromosomes of each generation are passed unchanged to the following generation. This means that at the end of the run the e best chromosomes are the e best solutions found till then.

2.4.7. Control of replicates

The presence of 'twins' is not allowed; if two identical chromosomes are present in the same population, one gene of one of them is randomly selected and swapped (from 0 to 1, or vice versa).

2.4.8. Stop condition

The GA stops after a predetermined number of generations took place without any 'new entry' among the e elitist chromosomes.

2.4.9. 'Heats' and 'final'

To better exploit the results obtained in several runs a new run is performed in which the initial population is composed by the best chromosomes found in each of the previous runs. This final run can be seen a refinement of the already good solutions found. In it, exploitation prevails on exploration. With the objects corresponding to the best chromosome $f(\boldsymbol{x})^{\text{final}}$ is computed by the use of the potential functions.

2.5. Software

The programs for both procedures have been written in Matlab 4.2c.1 (The Math Works, Natick, MA).

3. Data

The two selection techniques were applied to several real and simulated data sets; in this paper the results obtained on some simulated data sets and on three real data sets are reported.

3.1. Simulated data set FIRST

This set has been used for Fig. 1. 100 objects were extracted from an univariate distribution obtained by the sum of two normal distributions with a small difference (0.1) between locations (a1) and standard deviations 1.3 and 1.7.

3.2. Simulated data set UNIMODAL

The data set UNIMODAL consists of 100 objects described by one variable distributed as $N(0, 1)$. Some results (Table 1) obtained with other data sets of 100 objects extracted from the same population are also reported. These data sets are referred as UNIMODAL1, UNIMODAL2, $\cdots$.

3.3. Simulated data set BIMODAL

The set BIMODAL has been generated with two unimodal distributions similar to the previous one. The number of objects is 165.

3.4. Data set IRIS

It is the well known data set used by Fisher in his fundamental paper introducing linear discriminant analysis. 150 objects, 50 in each of the 3 categories, are described by 4 variables.

3.5. Data set WINES

The data set WINES consists of 178 samples of three types of Italian wine; Barolo (59 objects), Gri-

Table 3
Means and standard deviations for data sets UNIMODAL and BIMODAL and for the subsets

Set	Mean	Standard deviation
Unimodal original	0.0551	0.9365
Unimodal PF	0.2437	0.8388
Unimodal GA subset 1	−0.0053	0.9368
Unimodal GA subset 2	−0.0409	0.9367
Unimodal GA subset 3	0.3490	0.9365
Bimodal original	1.2328	1.7350
Bimodal PF	1.2518	1.9495
Bimodal GA subset 1	1.6525	1.7350
Bimodal GA subset 2	1.9192	1.7346
Bimodal GA subset 3	1.3371	1.7189

gnolino (71) and Barbera (48). Every sample is described by 27 chemical variables [14].

3.6. Data set GRIGNOLINO

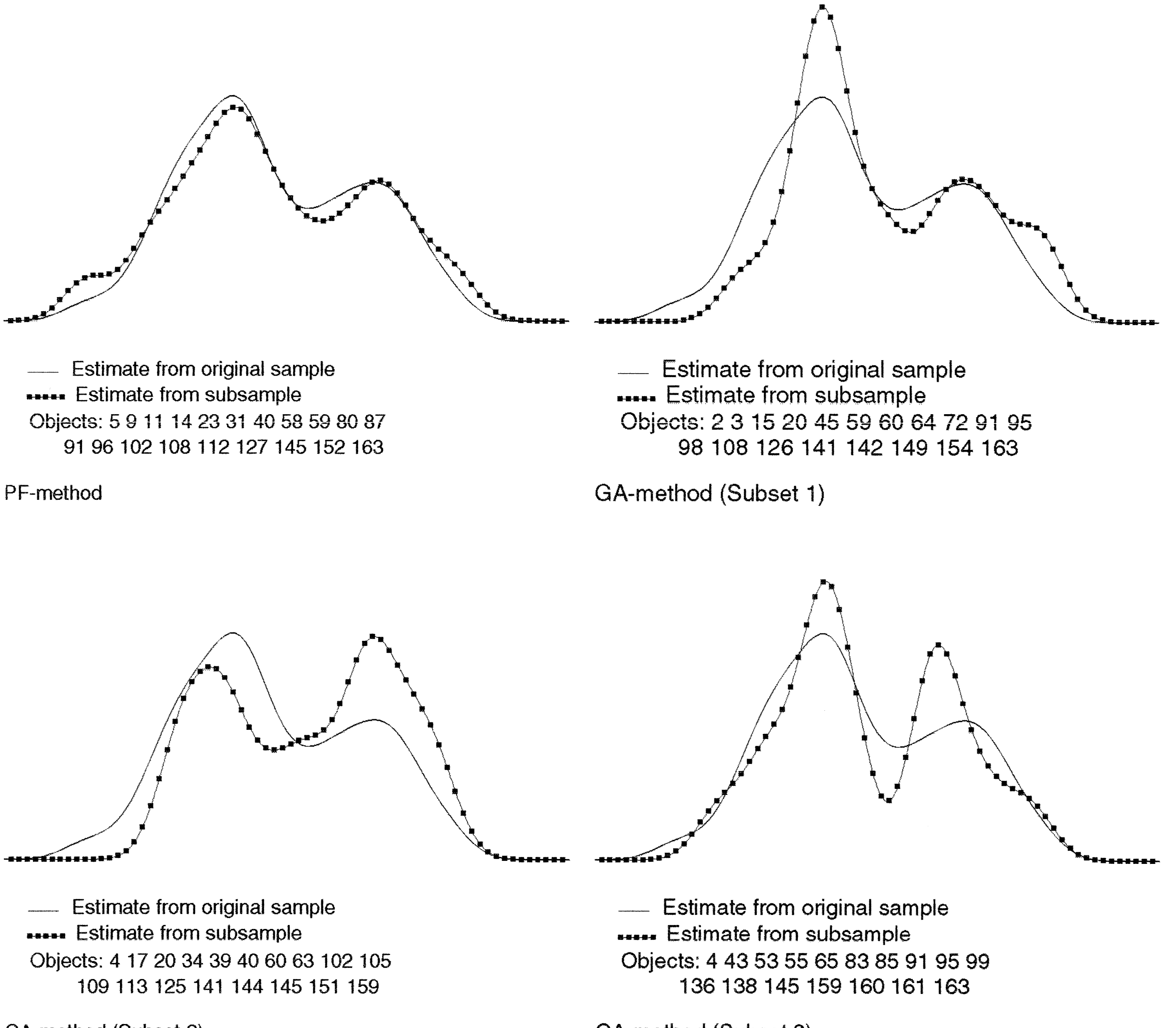
It is class 2 of data set WINES.

3.7. Data set ITAOIL

The data set ITAOIL consists of 547 samples of olive oil, from different Italian regions (1: North-Apulia; 2: Calabria; 3: South-Apulia; 4: Sicily; 5: Inland-Sardinia; 6: Coast-Sardinia; 7: East-Liguria; 8: West-Liguria; 9: Umbria). Every sample is described by 8 fatty acids. The number of samples of each category is given in Table 5.

3.8. Data set APULIA

It is class 3 of data set ITAOIL.



PF-method

GA-method (Subset 1)

GA-method (Subset 2)

GA-method (Subset 3)

Fig. 3. Probability density function of the BIMODAL set and of the subsets ($f(x)^{\text{subset}}$) obtained by PF method and by 3 repetitions of GA method.

4. Results and discussion

4.1. Influence of the method parameters

In the PF-method the influential parameters are the smoothing coefficient, k, and the selection factor, r. A systematic study was made on the univariate data sets FIRST, UNIMODAL, UNIMODAL1, $\cdots$, by studying the mean and the standard deviation of the subset as a function of k and r. Both parameters have no systematic effect on the estimate of the mean, whose values are irregularly distributed around the value of the original set. The effect of the smoothing coefficient k on the estimate of the standard deviation is irregular. In contrast, the parameter r has, as expected, a strong influence on the standard deviation. Small values of r (below the value suggested by Eq. (5)) favour the selection of objects near to the maximum of the probability distribution, so that the standard deviation of the subset is smaller than that of the original data set (see Table 1); large values of r favour the selection of some objects corresponding to low values of the probability density, so that the standard deviation of the subset is large. There is no interaction effect between the two parameters.

The results described below were obtained with the value of the smoothing coefficient k obtained by the usual leave-one-out procedure and the value I/M of Eq. (5) for the selection factor r.

In the GA-method four parameters (population size, number of elitist chromosomes, probability of initial selection and probability of mutation) have to be defined. It is therefore important to know if and how they affect the final results. To study this problem a fractional factorial design (2^{4-1}) has been performed (see Table 2) with the data set IRIS. The elaboration was stopped when 100 consecutive chromosomes were evaluated without a 'new entry' among the elitist chromosomes (due to the different population size, a stop criterion based on the number of generations would have favoured the experiments with the largest population size). In this case none of the parameters is significant. This is very important, since it means that the structure of the GA is rather robust and therefore the proposed parameters can be used without having to perform a set-up for each data set.

4.2. Results on simulated and real data sets

The parameters utilised for the GA in all the examples were: 20 chromosomes, 10 elitist chromosomes, probability mutation 1%; the stop criterion was 3 generations without modification in the elitist chromosomes and 5 'heats'; the top 4 chromosomes of each heat passed to the final run, in which the stop criterion was 6 generations without modification. The results reported in the examples are those corresponding to the best chromosome. In some cases (sets

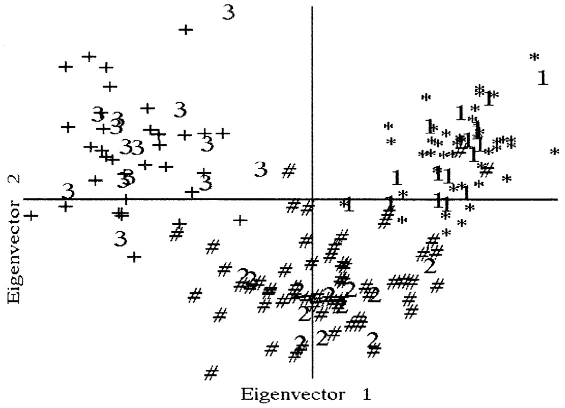


Fig. 4. Data set WINES (1, 2, 3: selected samples represented with their category index; $*$, #, $+$: unselected samples) in the space of first two eigenvectors of the original data. PF method.

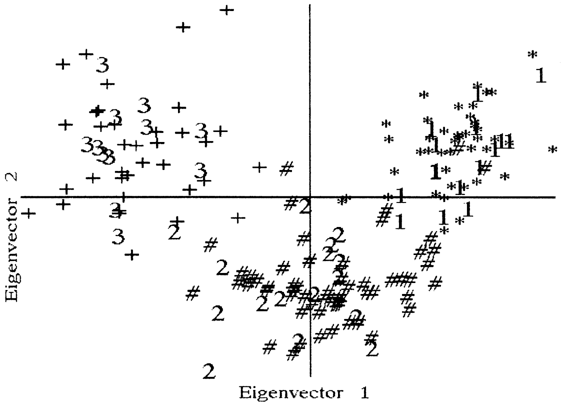


Fig. 5. Data set WINES (1, 2, 3: selected samples represented with their category index; $*$, #, $+$: unselected samples) in the space of first two eigenvectors of the original data. GA method.

UNIMODAL and BIMODAL) three results are reported, corresponding to the best chromosome obtained in repeated GA procedures, with different starting randomisation.

4.2.1. Data set UNIMODAL

The representative subset was composed by 10% (as maximum) of the original objects. In Fig. 2, the probability density functions of the original data set and of the selected subsets are shown. In Table 3 the mean and the standard deviation of the original data set and of the subsets are reported.

The subset obtained by PF is the one better reconstructing the original distribution, in spite of the noticeable difference in the variance, due to the fact that the subset has no objects representative of the more negative values on the left of the distribution. GA-method produces subsets with an excellent reproduction of the variance, but with a form of the estimated probability distribution very different from that obtained with the original data set. The three subsets obtained from GA have about the same variance and different mean; the fit with the original distribution does not depend on the reproduction of the mean: Fig. 2 shows that the GA subset 1 with the best reproduction of the mean is that with the worse fit with the original distribution.

4.2.2. Data set BIMODAL

The representative subset was composed by 20 objects (as maximum). Fig. 3 shows the original and reconstructed probability density functions. In Table 3 the mean and the standard deviation of the original data set and of the subsets are reported.

Also in this case PF-method produces a subset with a bad reproduction of the variance, in this case with a positive error. The mean is reproduced fairly well. The probability density distribution estimated from the subset fits well that estimated from the original data set.

GA-method subsets reproduce generally very well the variance; they are able to detect the bimodal character of the distribution and the approximate location of the probability maxima, but the shape of the distribution is rather different from the original one.

4.2.3. Data set WINES

The number of the samples in the subsets was 44 (25%), and the selection was made on the whole data set, without taking into account the subdivision in categories. Figs. 4 and 5 show the complete data set and the subsets selected by the two methods in the space of the two first principal components of the original data set. Table 4 shows the details of the principal component analysis of the original data set and of the two subsets. There is no significant difference in the eigenvalues of the two subsets. In contrast, the within-category dispersions are very different, especially in the case of category 2 (Grignolino) where the square root of the determinant of the variance–covariance matrix, as used in the expression of

Table 4
Results with data set WINES

	Original	Potential Functions	Genetic Algorithm
Objects in Class 1 Barolo	59	18	16
Objects in Class 2 Grignolino	71	12	16
Objects in Class 3 Barbera	48	14	12
Principal component analysis — autoscaled variables, cumulate % explained variance			
With 1 component	25.30	28.16	28.93
With 2 components	41.05	47.00	46.90
With 3 components	50.41	57.65	58.75
With 4 components	57.74	64.29	65.67
With 5 components	62.90	69.96	70.44
Determinant of covariance matrix (5 components)			
Class 1	0.628	0.536	0.189
Class 2	25.742	3.139	31.150
Class 3	2.603	2.995	0.881

the multivariate normal probability distribution, corresponding to the standard deviation in the univariate case, is 5.1 for the original data, 1.8 for the PF subset, 5.6 for the GA subset. In the case of the other two classes PF subset has dispersion characteristics very similar to those of the original subset.

The structure of the original data set can be defined by means of the 351 correlation coefficients between the variables. For each correlation coefficient the 95% confidence interval was computed.

In the case of PF subset, 3 correlation coefficients are very different from the original correlation coefficients (difference from the nearest limit of the confidence interval greater than 0.2); 29 correlation coefficients are rather different from the original correlation coefficients (difference from the nearest limit of the confidence interval between 0.05 and 0.2); 47 correlation coefficients are a little different (less than 0.05); the other 272 correlation coefficients are within the confidence interval.

In the case of GA subset, 3 correlation coefficients are very different, 38 rather different, 51 a little different, 259 within the confidence interval. So, it seems that PF subset preserves better the original structure. However, when looking to the single categories, the opposite result can be obtained. In the case of the category Grignolino, where PF subset retains only 12 objects, 4 less than the Grignolino objects in the GA subset, the PF subset shows 36 very different, 77 rather different, 25 a little different correlation coefficients, with only 213 correlation coefficients within the 95% confidence interval around the original correlation coefficients. In the case of GA subset, the number of heavy errors is one half, 18; the number of medium errors is the same, 77; the number of light errors is 33, and 223 correlation coefficients are within the confidence interval. These results confirm that in the case of categorised data the selection made on the whole data set can produce worse results in one or more categories.

Figs. 6–8 show the model of class 2 (data set GRIGNOLINO), computed by means of potential functions used as a class-modelling technique [8] by using as modelling variables the scores in the space of the two principal components of the original data of the modelled class. The class model has a boundary which corresponds to a critical value of the probability density, $f(\mathbf{x})^{\text{Class Boundary}}$, and the iso-lines in

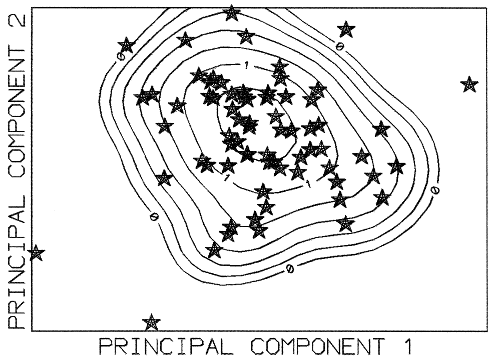


Fig. 6. Data set GRIGNOLINO. Class model obtained with the original data set. Iso lines represent constant values of the function $\log_{10} f(\mathbf{x})/f(\mathbf{x})^{\text{class boundary}}$.

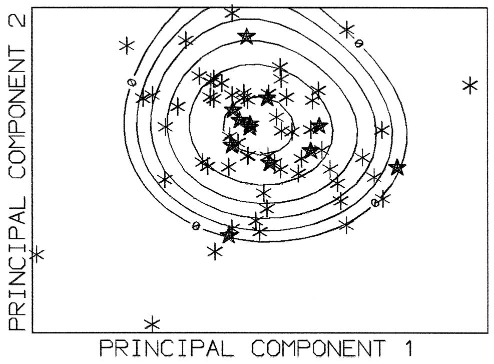


Fig. 7. Data set GRIGNOLINO. Class model obtained with the subset from PF-method.

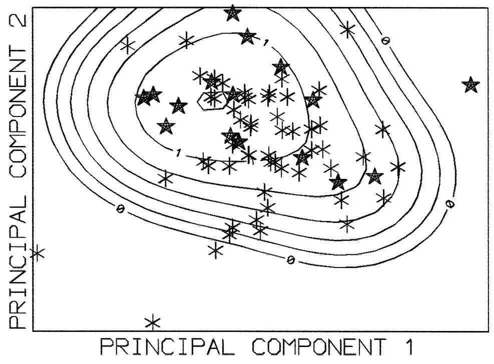


Fig. 8. Data set GRIGNOLINO. Class model obtained with the subset from GA-method.

Figs. 6–8 correspond to constant values of the function

$$\log_{10}\left(\frac{f(\boldsymbol{x})}{f(\boldsymbol{x})^{\text{class boundary}}}\right) \qquad (13)$$

PF method selects too many objects in the central part of the class space, and the class model becomes too small; GA method selects too many objects far from the class centroid; so, due also to a larger value of the smoothing coefficient computed (by means of cross validation) the class model is too large, especially in the upper part, outside the figure.

In spite of the poor representation of the single class obtained with the selection performed on the whole data set, the model obtained with the PF subset seems better than the model obtained with the GA subset.

The selection was repeated on the single class GRIGNOLINO. GA method was used three times, so that three subsets were obtained. Because of the constraint of maximum 25% of selected objects in the subset, the PF subset had 18 objects, as two GA subsets, whereas the other GA subset had 17 objects.

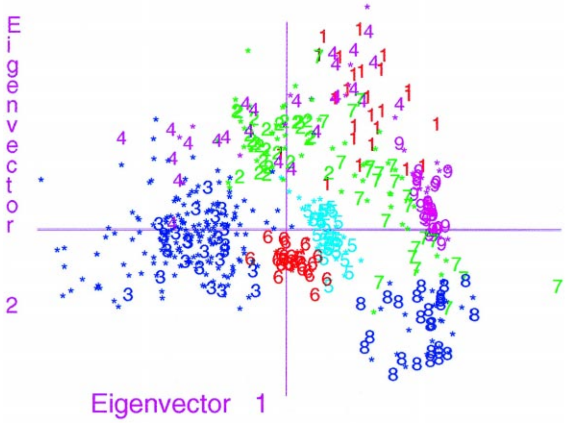
Potential function class models were obtained with the scores of the two first principal components of the class. PF-subset model is just slightly larger than that in Fig. 7, so that the performance is very similar to that of the model computed with the 71 original data. One of the GA-subset models can be considered satisfactory. The other two GA-subset models are a bad representation of the original class model. Taking into account also the result in Fig. 8, three GA-subset models out of four retain one or two objects outside the class boundary, so surely not representative of the typical objects of the class.

4.2.4. Data set ITAOIL

In this case the subsets were obtained by working on each category separately, with the objective to obtain a set with balanced number of objects in each category. In the case of the category with the smallest number of objects (North-Apulia) no selection was made.

The subsets contain 25 objects (maximum in the case of GA method) for each class. Figs. 9 and 10 show the selected objects in the space of the two first principal components of the whole data set. Figs. 11–13 show the score plots of the two first principal components for the original data set and that of the two selections. In Table 5 some details are reported, i.e., the variance explained by the first components, and the measure of the within-class dispersion obtained by the determinant of the class covariance matrix in the space of the first five components.

The principal components of the data set with 25 (maximum) objects in each categories show approxi-



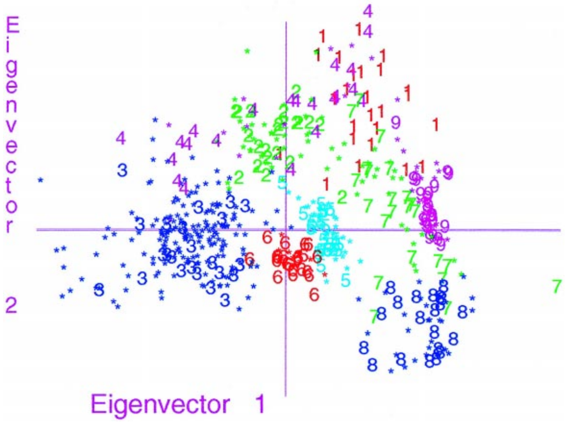


Fig. 9. Data set ITAOIL (1, 2, ..., 9: objects selected by PF method represented by their category index; ∗: unselected samples) in the space of the two first components of the original data set.

Fig. 10. Data set ITAOIL (1, 2, ..., 9: objects selected by GA method represented by their category index; ∗: unselected samples) in the space of the two first components of the original data set.

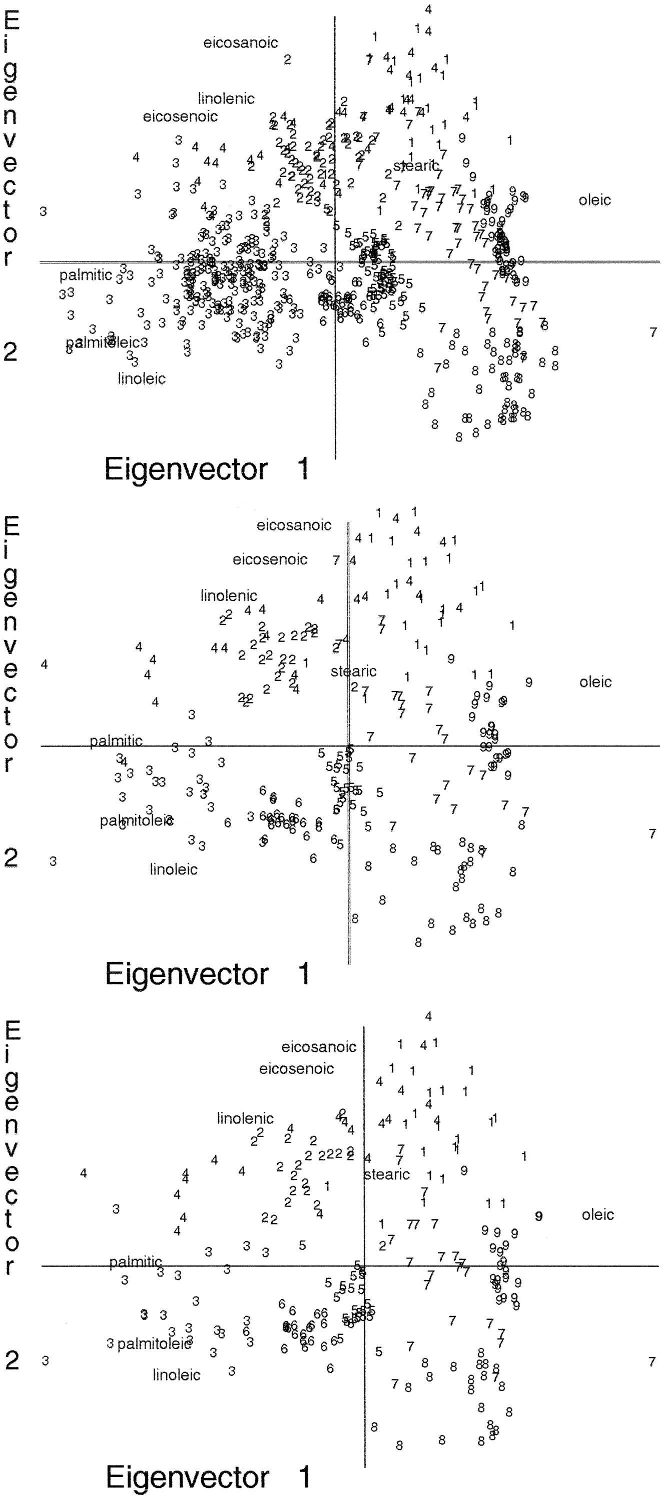

Eigenvector 2
eicosanoic
linolenic
eicosenoic
stearic
oleic
palmitic
palmitoleic
linoleic
Eigenvector 1
Eigenvector 2
eicosanoic
eicosenoic
linolenic
stearic
oleic
palmitic
palmitoleic
linoleic
Eigenvector 1
Eigenvector 2
eicosanoic
eicosenoic
linolenic
stearic
oleic
palmitic
palmitoleic
linoleic
Eigenvector 1

Table 5
Results with data set ITAOIL

	Original	Potential functions	Genetic algorithm
Objects in class 1	25	25	25
Objects in class 2	56	25	22
Objects in class 3	206	25	23
Objects in class 4	36	25	22
Objects in class 5	65	25	25
Objects in class 6	33	25	22
Objects in class 7	50	25	23
Objects in class 8	50	25	21
Objects in class 9	51	25	23
Principal component analysis — autoscaled variables, cumulate % explained variance			
With 1 component	46.75	39.02	39.99
With 2 components	68.71	65.57	66.91
With 3 components	81.38	78.54	80.44
With 4 components	91.32	89.90	90.87
With 5 components	95.40	95.11	95.00
Determinant of covariance matrix (5 components) (multiplied by 10000)			
Class 1	3.189	3.253	2.841
Class 2	6.746	8.062	1.715
Class 3	33.383	1.982	8.716
Class 4	194.473	299.730	108.053
Class 5	0.00405	0.00230	0.00149
Class 6	0.00162	0.00370	0.00080
Class 7	3.993	4.610	1.347
Class 8	1.011	1.528	0.294
Class 9	0.00052	0.00046	0.00017

mately the same structure as those of the original data (Figs. 11–13), at least on the significant components (2, according with double-cross validation), with a very small difference in the loadings.

Table 6 shows the correlation coefficients between the 8 variables for the data set APULIA, the lower and upper limits of the 95% confidence intervals, and the correlation coefficients in the PF and GA subsets. In the case of PF subset there are 1 heavy error, 8 medium errors, 3 light errors; the other 16 correlation coefficients are within the confidence intervals. In the case of GA subset there are 2 heavy errors, 9 medium errors, 8 light errors, and only 9 correlation coefficients are within the confidence interval. So, again the PF method seems to preserve better the structure of the original space of the information.

Figs. 14–16 show the class model of South-Apulia (data set APULIA) computed by means of potential functions used as a class-modelling technique [8] by using the scores in the space of the two principal components of the original data as variables.

The samples selected by PF are more concentrated near the barycentre of the category (where the potential value is higher), and the direction of the cluster in the upper left part of the plot is ignored. The

Fig. 11. Data set ITAOIL. Biplot of original data.

Fig. 12. Data set ITAOIL. Biplot of the subset selected by PF method.

Fig. 13. Data set ITAOIL. Biplot of the subset selected by GA method.

Table 6
Correlation coefficients in data set ITAOIL — class 3

(a) Correlation coefficients in the original data set

1.000	0.722	0.065	−0.812	0.229	0.025	0.198	0.154
	1.000	0.046	−0.695	0.253	−0.049	0.021	0.077
		1.000	−0.207	−0.031	0.091	0.151	0.215
			1.000	−0.704	−0.110	−0.202	−0.109
				1.000	0.071	−0.056	−0.193
					1.000	0.431	0.300
						1.000	0.510
							1.000

(b) Lower limit of 95% confidence intervals

1.000	0.649	−0.073	−0.855	0.095	−0.113	0.062	0.017
	1.000	−0.092	−0.760	0.120	−0.185	−0.117	−0.061
		1.000	−0.335	−0.168	−0.047	0.014	0.080
			1.000	−0.767	−0.244	−0.330	−0.243
				1.000	−0.067	−0.192	−0.322
					1.000	0.312	0.169
						1.000	0.400
							1.000

(c) Upper limit of 95% confidence intervals

1.000	0.782	0.201	−0.760	0.355	0.162	0.327	0.286
	1.000	0.183	−0.616	0.377	0.089	0.158	0.212
		1.000	−0.071	0.107	0.225	0.283	0.342
			1.000	−0.627	0.028	−0.066	0.029
				1.000	0.207	0.082	−0.057
					1.000	0.536	0.420
						1.000	0.605
							1.000

(d) Correlation coefficients in the subset from potential functions

1.000	0.728	0.395	−0.769	0.009	0.022	0.128	0.373
	1.000	0.304	−0.694	0.102	−0.050	0.067	0.125
		1.000	−0.508	−0.015	0.144	0.264	0.383
			1.000	−0.567	0.129	−0.103	−0.133
				1.000	−0.359	−0.228	−0.477
					1.000	0.382	0.348
						1.000	0.561
							1.000

(e) Deviations from the correlation coefficients of the original data set (—: within the 95% confidence interval; +: outside the confidence interval, less than 0.05; + +: outside the confidence interval, more than 0.05, less than 0.20; + + +: outside the confidence interval, more than 0.20)

—	—	+ +	—	+ +	—	—	+ +
	—	+ +	—	+	—	—	—
		—	+ +	—	—	—	+
			—	+ +	+ +	—	—
				—	+ + +	+	+ +
					—	—	—
						—	—
							—

Table 6 (continued)

(f) Correlation coefficients in the subset from genetic algorithms

1.000	0.688	0.504	−0.752	0.007	0.205	0.089	0.101
	1.000	0.410	−0.835	0.463	0.036	−0.294	−0.261
		1.000	−0.488	−0.068	−0.067	0.053	0.392
			1.000	−0.624	−0.389	−0.025	−0.028
				1.000	0.330	−0.213	−0.285
					1.000	0.569	0.279
						1.000	0.667
							1.000

(g) Deviations from the correlation coefficients of the original data set (—: within the 95% confidence interval; +: outside the confidence interval, less than 0.05; + +: outside the confidence interval, more than 0.05, less than 0.20; + + +: outside the confidence interval, more than 0.20)

—	—	+ + +	+	+ +	+	—	—
	—	+ + +	+ +	+ +	—	+ +	+ +
		—	+ +	—	+	—	+
			—	+	+ +	+	—
				—	+ +	+	—
					—	+	—
						—	+ +
							—

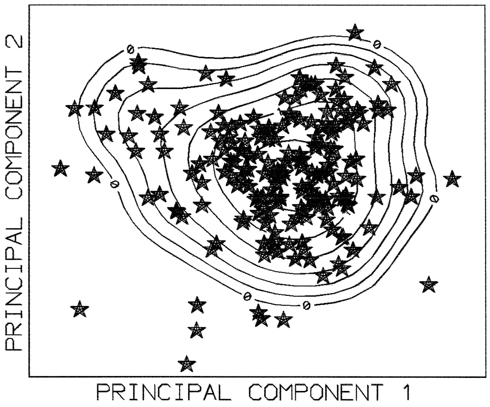
samples selected by GA are more spread throughout the space of the category, with a variance–covariance matrix more similar to the original one. When used for modelling purposes, the subsample obtained by PF, being more compact, would have a higher specificity (less objects of the other categories accepted), while the subsample obtained by GA would have a higher sensitivity (less objects of the same category rejected) and a lower specificity.

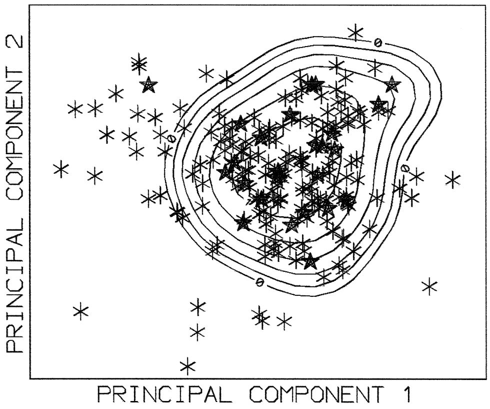
Data set APULIA is characterised by a nucleus of about 150 objects with similar characteristics. The other objects are far from the class centroid and rather dissimilar. In these cases when potential functions are used for classification purposes or to obtain a class model it is generally useful to use the variable-potential techniques, where the individual contribution of an object of the training set to the probability density function depends on the local density of the objects, measured by the distance between the object and the nearest objects. The application of variable-potential PF to the subset selection is under study.

Data set ITAOIL has also been used to study as the extraction of a representative subset from an unbalanced original set can change the performances of other techniques.

Fig. 17 shows the 7 × 7 two-dimensional Kohonen maps [15] obtained with the original set and with the PF-method subset. Maps in Fig. 17 report the capital letter corresponding to the category index. Each unit is characterised by the letter of the category with the majority within the objects mapped in the unit. Blank units have not a winner class. Because of the very large number of objects in category 3, South-Apulia, the number of units assigned to it (*C*) is very high (15 out of 49). The minimum number of units, 1, is assigned to class F (Coast Sardinia). With the subset with 25 objects in each class the number of units assigned to the nine categories ranges from 2 to 9.

Subset extraction was applied also to non-linear mapping (NLM) and to simplified non-linear mapping [16] (SNLM). The direct representation by NLM in a two-dimensional representation plane of data set ITAOIL requires the optimisation of 1091 parameters (2 co-ordinates for each object, less 3 co-ordinates fixed to avoid rotation and translation); the NLM co-ordinates must be selected so that the 149 331 distances in the representation plane are as close as possible to the corresponding distances in the space of the original (autoscaled or range-scaled)





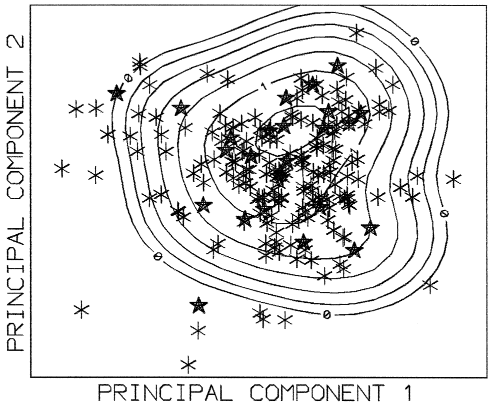


Fig. 14. Data set APULIA. Class model obtained with the original data set. Iso lines represent constant values of the function $\log_{10} f(\boldsymbol{x}) / f(\boldsymbol{x})^{\text{class boundary}}$

Fig. 15. Data set APULIA. Class model obtained with the subset from PF-method.

Fig. 16. Data set APULIA. Class model obtained with the subset from GA-method.

H	H	I	I	I	A	A
E	G	G	I	I		D
E		G	G	G	B	B
E	E			B	B	B
F		C	C	C	B	B
C	C	C	C	C	C	D
C	C	C	C	C	C	D

H	H	G	I	I		E
H	G	G	G	I		E
A	A	A	G	G	E	F
A	A	A		G		F
A	A	A	B	B		
D	D	B		C	C	C
B	B	D	D	C	C	C

Fig. 17. Data set APULIA. Kohonen maps obtained with the original data set (left) and with the subset from PF method (right).

variables. When the number of the objects is very large results of NLM are not satisfactory. So, from the set ITAOIL a subset was selected by means of the PF-method, with 3 objects in each category. On the 27 objects of the subset NLM has only 51 co-ordinates to optimise; repeated trials with different starting co-ordinates are performed, and the final result, shown in Fig. 18, refers to the trial with the minimum representation error. Class 4, Sicily, is characterised by a large dispersion of its 3 samples, corresponding to the recognised high dispersion of the class. Then, SNLM was performed; here each object in the set ITAOIL is placed in the plot so that the distance from the objects and the 27 objects selected for NLM is as close as possible to the corresponding distance in the space of the original information. Only 2 co-ordinates are optimised for each object. The final plot is shown in Fig. 19, with the dispersion polygons for the 9 categories.

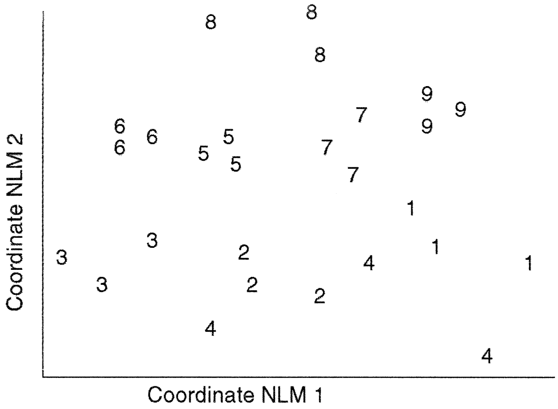


Fig. 18. Data set ITAOIL. Representation on non-linear-mapping co-ordinates of a subset with 3 representative objects for each class, selected by PF-method.

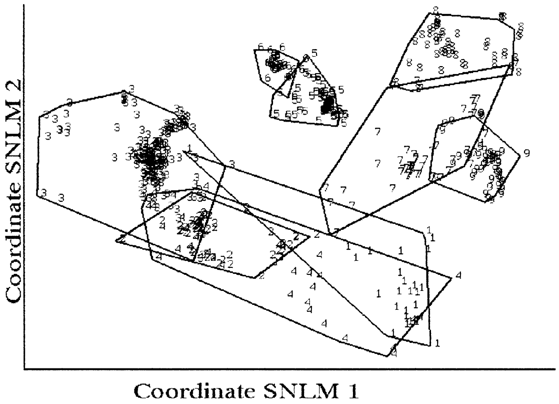


Fig. 19. Data set ITAOIL. Representation of the whole data set on the co-ordinates of simplified NLM, using as reference objects those selected for Fig. 18. Each class shown by means of his dispersion polygon.

The advantage of using a small representative subset is that a class with a multimodal distribution or with different clusters of objects can not be well represented only by its centroid, as was done previously [16] when the 9 centroids were used in the NLM step.

5. Conclusions

The extraction of a representative subset from a large data set has the obvious advantage of a reduced experimental cost in the case of real problems as that presented in the introduction; moreover it can give some advantages in the use of representation and class modelling techniques. The two methods presented here meet the requirement to obtain a representative subset. The reproduction of the dispersion characteristics, as evaluated from the fit to the probability function estimated from the original data set, shows that, in the case of simulated data, with very regular probability distribution, PF-method performs better than GA-method. When used on real data, with complex distributions, PF method selects generally a too high percentage of objects in the centre of a class, so that the class model is too small; on the contrary GA method selects too many objects near to the class boundary, so that the class model is too large. According to the purposes for which a class model is computed one of the two methods can be more appropriate. PF-method can be, in principle, forced to select more objects far from the class centroid, by increasing the selection factor r; however this procedure must be used cautiously, and always the class model obtained with the subset must be compared with that obtained with the original subset.

Both methods can be improved. PF could use variable potentials, with possible advantages when the objects of the original sample are distributed with very different density in the space of variables. GA can be improved by introducing a variable power in the equation of the similarity. These possibilities will be tested in future.

References

[1] R. Carlson, Design and Optimization in Organic Synthesis, Elsevier, Amsterdam, 1992.

[2] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.

[3] C.E. Miller, The use of chemometrics techniques in process analytical method development and operation, Chemom. Intell. Lab. Syst. 30 (1995) 11–22.

[4] T. Isaksson, T. Naes, Selection of samples for calibration in near-infrared spectroscopy. Part II: selection based on spectral measurements, Appl. Spectrosc. 44 (1990) 1152–1158.

[5] L. Kaufman, P.J. Rouseeuw, Finding Groups in Data, Wiley, New York, 1990.

[6] P.K. Hopke, L. Kaufman, The use of sampling to cluster large data sets, Chemom. Intell. Lab. Syst. 8 (1990) 195–204.

[7] D. Coomans, I. Broeckaert, Potential pattern recognition in chemical and medical decision making, Research Studies Press, Letchworth, 1986.

[8] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, J. Chemom. 5 (1991) 435–453.

[9] D.L. Massart, L. Kaufman, The interpretation of analytical chemical data by the use of cluster analysis, Wiley, New York, 1983.

[10] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms. part 1. concepts, properties and context (Tutorial), Chemom. Intell. Lab. Syst. 19 (1993) 1–33.

[11] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms. Part 2. representation, configuration and hibridization (tutorial), Chemom. Intell. Lab. Syst. 25 (1993) 99–145.

[12] A. Broudiscou, R. Leardi, R. Phan-Tan-Luu, Genetic algorithms as a tool for selection of D-optimal design, Chemom. Intell. Lab. Syst. 35 (1996) 105–116.

[13] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemom. 6 (1992) 267–281.

[14] M. Forina, C. Armanino, M. Castino, M. Ubigli, Multivariate

data analysis as a discriminating method of the origin of wines, Vitis 25 (1986) 189–201.

[15] W.J. Melssen, J.R.M. Smits, L.M.C. Buydens, G. Kateman, Using artificial neural networks for solving chemical problems — part II. Kohonen self-organizing feature maps and Hopfield networks, Chemom. Intell. Lab. Syst. 23 (1994) 267–291.

[16] M. Forina, C. Armanino, S. Lanteri, C. Calcagno, Simplified nonlinear mapping of analytical data, Ann. Chim. (Rome) 72 (1983) 641–657.