Journal of Classification 23:143-167 (2006)
DOI: 10.1007/s00357-006-0008-0

On a Transvariation Based Measure of Group Separability

Daniela G. Calò

University of Bologna, Italy

Abstract: In this paper, the potentialities of transvariation (Gini, 1959) in measuring the separation between two groups of multivariate observations are explored. With this aim, a modified version of Gini's notion of multidimensional transvariation is proposed. According to Gini (1959), two groups G_1 and G_2 are said to *transvary* on the k-dimensional variable $\boldsymbol{X}=(X_1,\ldots, X_h,\ldots, X_k)$ if there exists at least one pair of units, belonging to different groups, such that for $h=1, \ldots, k$ the sign of the difference between their X_h values is opposite to that of $m_{1h} - m_{2h}$, where m_{1h} and m_{2h} are the corresponding group mean values of X_h. We introduce a modification that allows us to derive a measure of group separation, which can be profitably used in discriminating between two groups. The performance of the measure is tested through simulation experiments. The results show that the proposed measure is not sensitive to distributional assumptions and highlight its robustness against outliers.

Key words: Discriminant analysis; Separability measures; Transvariation; Robustness.

I am particularly grateful to Professor Camilo Dagum, for his patience and precious comments during my work. I also thank the Reviewers for their remarks, which greatly improved the manuscript, and Professor Angela Montanari for her interesting discussions and helpful suggestions.

Author's address: Dipartimento di Scienze Statistiche, Università di Bologna, via delle Belle Arti, 41, I-40126 Bologna, Italy, fax:++39 051 2098269, email: danielagiovanna.calo @unibo.it

Dedication

I would like to dedicate this article to the memory of Camilo Dagum (Emeritus professor of the University of Ottawa, Canada, and Full professor at the Faculty of Statistical Sciences, University of Bologna, Italy) who suddenly passed away on November 5, 2005 in Ottawa. His departure is a great loss to the world's statistical and economic scientific community. He has been a source of inspiration to his students over a long career of teaching and research that spanned more than fifty years.

Camilo Dagum's versatile scientific knowledge, combined with an inquisitive mind, enabled him to carry out a remarkable scientific work in Economics, Statistics, Econometrics and Philosophy of Science. His rigorous scientific contributions have stood the passing of time, and many among them, belong to today scientific paradigm. He pursued research on functional and personal income distributions, inequality within and between income distributions, wealth distribution, personal and national human capital and poverty, producing pioneering and seminal papers in all the above topics which gave birth to new research paths.

He worked with Professor Corrado Gini at the University of Rome, Italy, on the theory of transvaration, subject to which he made significant contributions with applications in economics. In fact, the theory of transvariation was the main topic of his doctorate dissertation which led to a series of papers later published in Spanish, Italian, English, French and German.

Above all, Professor Dagum will always be remembered as a man of great honesty, unparallel humanity , and a true Gentle man.

I am personally indebted to him for his continuing support and encouragement to carry out research on the theory of transvariation.

1. Introduction

In the discriminant analysis context, an important issue that should be addressed before any classification rule is devised regards a measure of whether any rule we can construct (given a set of k variables) is likely to be effective enough for the research purposes. An indication about the best discrimination performance that can be achieved may be yielded by estimating how much separated the class conditional distributions are. According to Hand, two classes "*... are described as 'perfectly separable' or simply 'separable' if the support regions of the population distributions do not intersect. This means that, at any given point of the measurement space, objects from only one class will be observed*" (Hand 1997). In most practical

cases, however, classes (or populations) are not perfectly separable and the definition of a suitable measure of separability is required.

Several measures of distance or divergence between two distributions have been proposed in the statistical literature (see Hand (1997) or Krzanowski and Marriott (1995) for a review). They may be grouped in two main categories, corresponding to two conceptually different approaches: *probabilistic measures* and *distance-based measures*.

Probabilistic separability measures are derived as a measure of the 'distance' between the class-conditional probability density functions, f_1 and f_2: in this category one can find measures based on ideas of information theory, like Jeffrey's (1948) *divergence*, measures related to Bhattacharyya's (1943) *affinity coefficient* between f_1 and f_2 , and measures built as a decreasing function of the misclassification errors (see, for example, Lissack and Fu 1976). The use of probabilistic separability measures requires that some assumptions are made about the class probability distributions; otherwise, in a nonparametric approach, f_1 and f_2 can be approximated resorting to nonparametric density estimation (in high-dimensional settings however, the latter solution is prone to the curse of dimensionality).

Within the distance-based approach, when the observed variables are continuous, the most commonly employed measure of between-class distance in the two class case is the Mahalanobis distance

$$\Delta^2 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2),$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the population mean vectors and $\boldsymbol{\Sigma}$ is the common covariance matrix (Mahalanobis 1930, 1936). In this measure, widely used in classical multivariate statistics, the distributions are assumed to be homoscedastic and are summarized in terms of their first-order and second-order moments. Mitchell and Krzanowski (1985) have shown that the Mahalanobis distance is appropriate when the class conditional distributions are members of the elliptic class having fixed shape but varying location: in fact, in this case the two distributions are completely specified by their mean vectors and their common covariance matrix. If the two populations are normally distributed - the most important special case of elliptical distributions - with identical covariance matrix, it can be shown that the amount of overlap between them is a function of the Mahalanobis distance. However, for skewed distributions Δ^2 may not succeed in distinguishing between separated or overlapping distributions, as Figure 1 clearly shows.

An estimate of Δ^2 may be easily obtained in a plug-in fashion. It is well known that the conventional sample covariance matrix, and consequently the sample Mahalanobis distance, are highly sensitive to out-

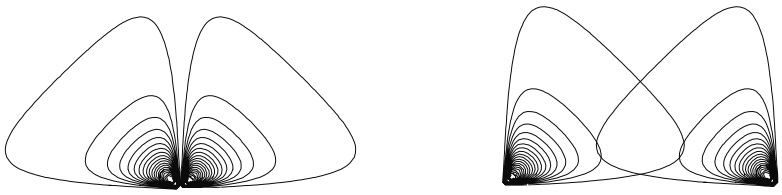

Figure 1. Two bivariate skewed distributions having the same Mahalanobis distance but a different amount of overlap.

lying observations. (A natural way to make it robust is by replacing the population mean vectors and the matrix $\mathbf{\Sigma}$ with more robust location and scale estimates). Furthermore, when data stem from heteroscedastic distributions, the use of the pooled covariance matrix in the estimated Mahalanobis distance might reduce its performance.

More generally, in the distance-based approach, distances between two classes are found starting from distances between individuals. Given a suitable measure of the distance between two k-variate observations, the quantity of interest is the average of this distance over pairs of units coming one from each class (Devijver and Kittler 1982). Among the different choices of pairwise distance measures, the Euclidean metric allows for both analytical and computational simplifications. However, the measures derived according to this approach show a behaviour analogous to that of Mahalanobis distance, as they are related to the same statistics which are involved in the sample Mahalanobis distance. Following Rao's work (Rao 1982) on diversity and dissimilarity indices, Cuadras, Fortiana and Oliva (1997) have introduced between-class distances based on dissimilarities between observations and studied their properties in nonparametric discrimination; the main benefit of this approach is that the use of dissimilarities allows for any combination of data types. For numeric variables, measuring dissimilarity through the squared Euclidean distance or through the Manhattan distance completely neglects the covariance structure, and when a weighted Euclidean distance is employed (with weight equal to the inverse covariance matrix), Mahalanobis distance is obtained again.

In this paper, we propose a measure of separation between two groups of multivariate observations, which simultaneously takes into account location, variability and skewness characteristics of the class-conditional distributions, while being a moment-free statistic. It is based on Gini's notion of transvariation (Gini 1916 1959).

The use of univariate transvariation as a measure of group distance is not new. As it will be shown in the next section, univariate transvariation between two groups is related to Hand's concept of separability: in particular, univariate transvariation measures highlight different aspects of the overlap between the groups. Dagum (1980) introduced two measures related to univariate transvariation in order to evaluate the *economic distance* between two populations through a measure of inequality between the corresponding income distributions. Dagum's proposal opened a new field of research in econometrics about the measures of distance between income distributions (Shorrocks 1982): in particular, Ebert (1984) and Chakravarty and Dutta (1987) proposed to axiomatically characterize the measure, by asking it to satisfy suitable properties derived from economic theory arguments; Yitzhaki (1994) and Deutsch and Silber (1997) further explored the potentialities of transvariation in measuring the degree of overlapping between distributions (for a review, see Dagum 2005). However relevant within the domain of income inequalities, all these contributions are not suitable for our purpose as they are necessarily focused on the univariate setting, income being the only feature they deal with, whereas we want to measure multivariate separation.

In the field of discrimination and classification, Montanari (2004) has recently proposed to derive a two-group linear discriminant function as the linear combination of the observed variables along which a measure of transvariation is minimized; this solution is searched for through a projection pursuit algorithm. Several transvariation measures are employed and compared in that work, and finally a new linear discriminant function is introduced which outperforms Fisher's linear discriminant function when that is not optimal. The approach of Montanari deals again only with univariate transvariation, as it looks for suitable one-dimensional projections.

The aim of the present paper is to explore the potentialities of *multivariate* transvariation in measuring the separation between two groups. The main issue is that Gini's definition of transvariation with respect to more than one variable (which is the natural generalisation of the univariate one) may not always be interpreted as a measure of inseparability. We therefore suggest a modification of Gini's notion of multidimensional transvariation, which is derived by considering sequentially the group transvariation along any single variable, but on suitably chosen subsets of units, as illustrated in Section 3; at the end of Section 3 two illustrative examples on real data are presented. In Section 4 the performance of the proposed measure is assessed by simulation studies.

2. Univariate Transvariation

Univariate transvariation between two groups has been defined by Gini (1916) as follows:

Definition 1.

Two groups G_1 and G_2, of n_1 and n_2 units respectively, are said to *transvary* on the variable X with respect to their corresponding mean values m_{1X} and m_{2X} ($m_{1X} \neq m_{2X}$), if the sign of at least one of the differences $x_{1i}-x_{2j}$ ($i=1,\ldots, n_1$; $j=1,\ldots, n_2$) which can be defined between the X values belonging to the groups is opposite to that of $m_{1X}-m_{2X}$.

Any pair of units ($i \in G_1$, $j \in G_2$) satisfying this condition is said to transvary. The number of transvarying pairs is denoted by:

$$s_{12} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \eta\left(x_{1i}, x_{2j} \right), \tag{1}$$

where

$$\eta\left(x_{1i}, x_{2j}\right) = \begin{cases} 1 & \text{if } \left(x_{1i} - x_{2j} \right)\left(m_{1X} - m_{2X} \right) < 0 \\ 0 & \text{if } \left(x_{1i} - x_{2j} \right)\left(m_{1X} - m_{2X} \right) > 0 \\ 1/2 & \text{if } x_{1i} = x_{2j}, \end{cases}$$

since the convention is adopted of counting half the number of the pairs having identical members as transvarying pairs.

Gini (1959) introduced several measures of univariate transvariation between two groups, each highlighting different aspects of group transvariation, which have also been extended to more than two groups (see Dagum 1959, 1965). Because of its robustness properties (Montanari 2004), we focused our attention on transvariation probability.

By using the median in place of the mean value in Definition 1, Gini (1916) defines transvariation probability as the ratio of the actual number of transvarying pairs to its maximum

$$tp = s_{12}/max(s_{12}) \tag{2}$$

Thus, transvariation probability takes values in the interval [0,1] and the more the two groups overlap, the greater the values it takes. Its complement to 1 formally translates the notion of separability described by Hand (see Section 1).

Gini notes that s_{12} tends to increase, *ceteris paribus*, as the distance between the group medians decreases and that it finally reaches its maximum when the medians coincide. But it can be shown that, after performing such a shift, the value of s_{12} may not be the maximum attainable. This happens because Gini's remark is strictly appropriate if both of the variable distributions are symmetric: in this case, it can also be shown that $max(s_{12})=n_1n_2/2$. According to Gini, this value can be taken as an approximation of the unknown maximum value in more general situations. However, he himself admits that, in some cases, transvariation probability may even take values greater than 1 when this plug-in value is introduced in (2). Therefore, the maximum of s_{12} should be determined by trial and error, that is by shifting one of the groups by various amounts (for further discussions, see Gini (1959)). Finally, $max(s_{12})$ depends on the data and should be computed numerically, since no general closed form exists for it: this represents an unpleasant aspect of the definition given in (2). We were intrigued by this issue, which motivated the following remarks.

Gini introduced transvariation probability in order to evaluate the degree of uncertainty in predicting the sign of the difference between the feature values in any two units, each belonging to one of the two groups, by means of the sign of the difference between the corresponding group medians. In our opinion, it seems more coherent with Gini's aim to consider the median of the n_1n_2 differences instead of the difference between the group medians. In fact, denoting the median of $D=\{x_{1i}-x_{2j},\ i=1,\ldots,\ n_1;\ j=1,\ldots,\ n_2\}$ by δ, then when $\delta>0$ ($\delta<0$) the frequency of positive (negative) values in D is greater than 0.5. Thus, the (sign of the) statistic δ has an objective predicting power. Moreover, when $\delta=0$, negative and positive differences are equally likely, *i.e*. the relative positions of the two groups do not allow any prediction and uncertainty is maximized[1]. Thus, the situation of maximum transvariation may be obtained by shifting one of the groups so that the median of the n_1n_2 differences is equal to 0. The resulting number of transvarying pairs is always equal to $n_1n_2/2$, whatever the group distributions are.

For these reasons, we suggest the following modification of Gini's Definition 1.

---

1. The statistic δ is known in the statistical literature as the Hodges-Lehmann estimator of the location shift parameter Δ in the translation model, saying that the group parental populations are the same except one of them is shifted by the amount Δ (Hodges and Lehman 1963). Let F and G be the distribution functions corresponding to the two populations; then, the translation model is $G(t)=F(t-\Delta)$, for every t.

Definition 1*.

Two groups G_1 and G_2, of n_1 and n_2 units respectively, are said to transvary on the variable X if the sign of at least one of the $n_1 n_2$ differences $x_{1i}-x_{2j}$ which can be defined between the X values belonging to the groups is opposite to that of the *median of such differences*.

Then, the number of transvarying pairs according to Definition 1* is

$$s_{12}^{*} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \eta\left(x_{1i}, x_{2j} \right) \tag{3}$$

with

$$\eta\left(x_{1i}, x_{2j}\right) = \begin{cases} 1 & \text{if } \left(x_{1i} - x_{2j} \right) \; median\left\{ x_{1i} - x_{2j} \right\} < 0 \\ 0 & \text{if } \left(x_{1i} - x_{2j} \right) \; median\left\{ x_{1i} - x_{2j} \right\} > 0 \\ 1/2 & \text{if } x_{1i} = x_{2j} \; . \end{cases}$$

As a result, transvariation probability may be defined as

$$tp^{*} = s_{12}^{*} / max(s_{12}^{*}) = 2\, s_{12}^{*} / (n_1 n_2) \tag{4}$$

since $max(s_{12}^{*}) = n_1 n_2 / 2$.

It is easy to see that the quantity s_{12}^{*}, defined in (3), is related to the Mann-Whitney-Wilcoxon U statistic: $s_{12}^{*} = \min(U, \; n_1 n_2 - U)$. Thus, transvariation probability in (4) is but twice an estimate of the minimum between $p = P({}_1X < {}_2X)$ and its complement to 1, where ${}_1X$ (${}_2X$) is a random member from the parental population of group G_1 (G_2). This probability is an interesting parameter in the two-sample problem, particularly in applied reasearch, due to its natural sense and interpretability. Many efforts have been devoted in the statistical literature to attach confidence bounds to it (see Hollander and Wolfe (1999) for a review). Recently, Fligner and Policello (1981) have proposed a modification of the Mann-Whitney-Wilcoxon statistic which can be used to test the null hypothesis H_0: $p = 1/2$, *i.e.* the hypothesis of maximum transvariation probability.

3. Multivariate Transvariation

Gini's extension of the notion of transvariation between two groups in the multivariate context requires that at least one pair of units, each taken from one of the groups, simultaneously transvaries on each variable according to Definition 1. More precisely, multivariate transvariation is defined as follows (Gini and Livada 1943; Dagum 1971):

Definition 2.

Two groups G_1 and G_2, of n_1 and n_2 units respectively, are said to transvary on the k-dimensional variable $\boldsymbol{X}$ with respect to their corresponding mean vectors $\boldsymbol{m}_{1X}$ and $\boldsymbol{m}_{2X}$, if there exists at least one pair $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2j})$, where $i \in G_1$ and $j \in G_2$, such that for h=1,…, k the sign of the h-th entry in vector $\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}$ is opposite to that of the h-th entry in vector $\boldsymbol{m}_{1X} - \boldsymbol{m}_{2X}$ (this entry not being null).

Any pair of units ($i \in G_1$, $j \in G_2$) satisfying this condition is said to jointly transvary. By using the marginal median vector in place of the mean vector in Definition 2, Gini and Livada define multivariate transvariation probability as the ratio of the number of such jointly transvarying pairs to its maximum. No closed form for the maximum is given: it occurs when one of the groups is shifted so that the group marginal median vectors coincide.

Several nonparametric (rank-based) multivariate techniques apply univariate nonparametric methods to analyze the multivariate observations componentwise (see, for example, Hettmansperger (1984)). However, they often have difficulties in cases of dependence between component variates. This is true for multivariate transvariation as well, since it can no longer yield an indication of group overlapping, as opposite to the univariate case. In fact, the number of pairs satisfying Definition 2 may be greater than 0 even if the groups are completely separated in the multidimensional space. Figure 2 shows an illustrative example. The groups are taken from two bivariate normal distributions. As it can be easily seen, the mean values of both the variables are greater in G_1 than in G_2. Taken the pair (i, j) as an example (marked with a cross in the figure), both the variables assume in unit $i \in G_1$ smaller values than those they take in unit $j \in G_2$. In other words, this pair of units satisfies Gini's definition of bidimensional transvariation on the observed variables. Therefore, even if the groups are perfectly separable in R^2 (since their convex hulls do not intersect), they do transvary according to Definition 2.

It is worth noting that the natural generalization of the univariate definition to the multidimensional context would be appropriate provided that the class conditional densities of the variable $\boldsymbol{X}$ are somehow estimated. In this case, the unknown probability of observing a jointly transvarying pair after choosing at random one member from each population could be estimated by deriving the density of the vector variable $\boldsymbol{X}_1 - \boldsymbol{X}_2$ (where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote the values $\boldsymbol{X}$ assumes in the two classes) and integrating it on the right hyper-quadrant.

Trying to overcome the difficulty illustrated in Figure 2 amounts to defining a transvariation-based measure of group separation in the multi-

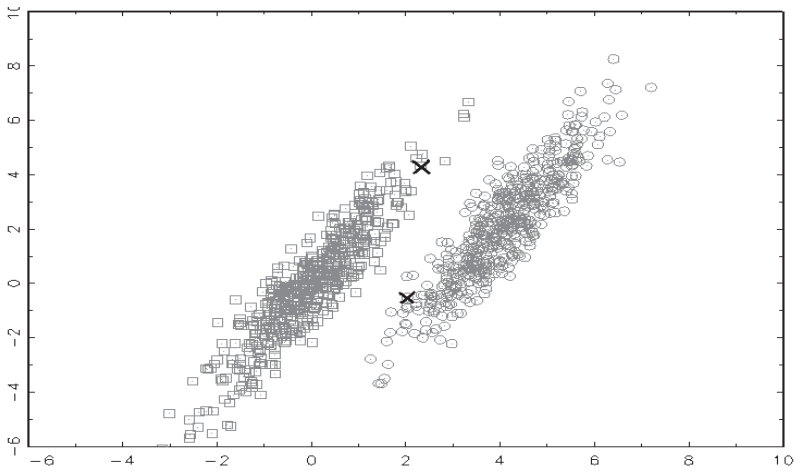

Figure 2. The symbols denote group membership: the dot for group G_1 and the square for group G_2. Even if the groups are perfectly separable in R^2, there are some pairs of units, belonging to different groups, which transvary according to Definition 2 (one of them is marked with a cross).

variate space. Due to the reasons illustrated in Section 2, in the following we will refer to the notion of univariate transvariation given in Definition 1*.

A good measure of group separation should be able to detect and explore the region where the groups overlap, or where their boundary lies (provided it exists): an indication about the amount of overlapping could be derived from the number of units lying in this region.

When only one variable, say X, is observed, this region may be identified in transvariation terms: more precisely, it corresponds to the X range in the set $G_1' \cup G_2'$, where G_1' (and G_2', respectively) is the set of the units in G_1 (G_2) which transvary on X with at least one unit in G_2 (G_1). The elements of $G_1' \cup G_2'$ are the members of the transvarying pairs: in this sense the subsets G_1' and G_2' are important, as far as separability is concerned, since variable X does not succeed in discriminating between them in transvariation terms.

In the multivariate case, each unit is observed with respect to several variables at the same time. The joint characteristics of the data set may be studied also by conditioning with respect to a subset of the variables. When dealing with joint transvariation, this means that when one more variable is observed besides X, attention should be restricted to the previously described subsets G_1' and G_2'. When turning back to the modification we wish to introduce in Definition 2, this implies that the median of the differences (see

Definition 1*) between the current variable values in the two groups should not be computed on the whole set of $n_1 n_2$ pairs: this would mean ignoring any information provided by the variables already considered. Its computation should be restricted to the subsets G_1' and G_2' of the elements which transvary on the variables so far examined.

This proposal entails a sequential procedure, whose results depend on the variable ordering. It seems reasonable to take the variables in increasing order of (univariate) transvariation so as to use, at each step, as much information as possible: in fact, the more the two groups transvary, the less confident we are in the sign of the median of the differences.

Therefore, the suggested modification of Gini's Definition 2 gives rise to the following algorithm for evaluating the k-dimensional transvariation. (Here the previous notation $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2j})$ is simplified to $(\boldsymbol{x}_i, \boldsymbol{x}_j)$):

- $s \leftarrow 1$
- Transvariation probability (see tp in (3)) between G_1 and G_2 on each variable is computed
- The variable X_s corresponding to the minimum tp value is considered
- The set of the pairs (i, j) such that
  $(x_{si} - x_{sj}) \cdot \ median\{(x_{si} - x_{sj}), i \in G_1, j \in G_2\} < 0$
  is recorded to determine G_1' and G_2'
- Do while s less than k
  - Transvariation probability, tp, between G_1' and G_2' on each of the remaining $k-s$ variables is computed
  - $s \leftarrow s+1$
  - The variable X_s corresponding to the minimum tp value is considered
  - The current set of the pairs (i, j) is updated by selecting, among its elements, the pairs for which
    $(x_{si} - x_{sj}) \cdot \ median\{(x_{si} - x_{sj}), i \in G_1', j \in G_2'\} < 0$
  - Subsets G_1' and G_2' are updated
- End

The resulting set of selected pairs are the jointly transvarying pairs according to the modified definition. Their number will be denoted by ${}_k s^*_{12}$.

As an example, let us apply the proposed sequential definition on the data illustrated in Figure 2, where Gini's multivariate transvariation failed in measuring group separation. The first variable to be considered, X_1, is the one whose values are reported in abscissa, since it yields the minimum transvariation probability value. The median of the differences $x_{1i} - x_{1j}$ ($i \in G_1$, $j \in G_2$) is positive; G_1' and G_2' are determined. At step 2 the "restricted" median of the differences between the values of the remaining

variable, $x_{2i}-x_{2j}$ ($i \in G_1'$, $j \in G_2'$), is computed. It is negative. Then, the current set of transvarying pairs (i, j), $i \in G_1$, $j \in G_2$ are those for which $x_{1i}<x_{1j}$ and $x_{2i}>x_{2j}$. As it can be easily seen, there are no units in G_1 whose X_1 value is smaller, and whose X_2 value is greater, than that of any unit in G_2 ($_2 s_{12}^* = 0$). Therefore, the groups G_1 and G_2 do not transvary on the observed two-dimensional vector variable.

Multivariate transvariation probability may be defined again as the ratio of $_k s_{12}^*$ to its maximum

$$_k tp = {}_k s_{12}^* / \max({}_k s_{12}^*). \tag{5}$$

Coherently with Definition 1*, the maximum is obtained as the number of transvarying pairs after shifting one group so that for each variable the median of the differences is equal to 0. Finally, we propose a measure of group separability by taking the complement to 1 of multivariate transvariation probability.

Simulation studies have shown that in the normal case the modified transvariation probability tends, for larger data sets, to the transvariation probability obtained by explicitly using the normality assumption (Gini and Livada 1943), which is a function of mean vectors and of variance-covariance matrices just like the Mahalanobis distance. Thus, in the normal case the modified transvariation probability and the Mahalanobis distance yield equivalent information as far as separability is concerned.

A numerical example

The proposed procedure has been applied to Fisher's iris data, after selecting observations from the two species Versicolor and Virginica; the results are illustrated in Figures 3-5. The data set consists of 100 units, 50 from each of the two groups, which have been observed with respect to the following variables:

X_1=sepal length
X_2=sepal width
X_3=petal length
X_4=petal width

At the beginning, the entire data set is considered: let's assume G_1=Versicolor, G_2=Virginica. The variable showing the lowest transvariation probability is X_3, and the median of the 50×50 differences between the X_3 values in G_1 and those in G_2 is negative.

Then, by identifying the members of the pairs which transvary along X_3 (marked with the cross in Figure 3, whichever their group membership is) subsets G_1' and G_2' are determined. Transvariation probability between these

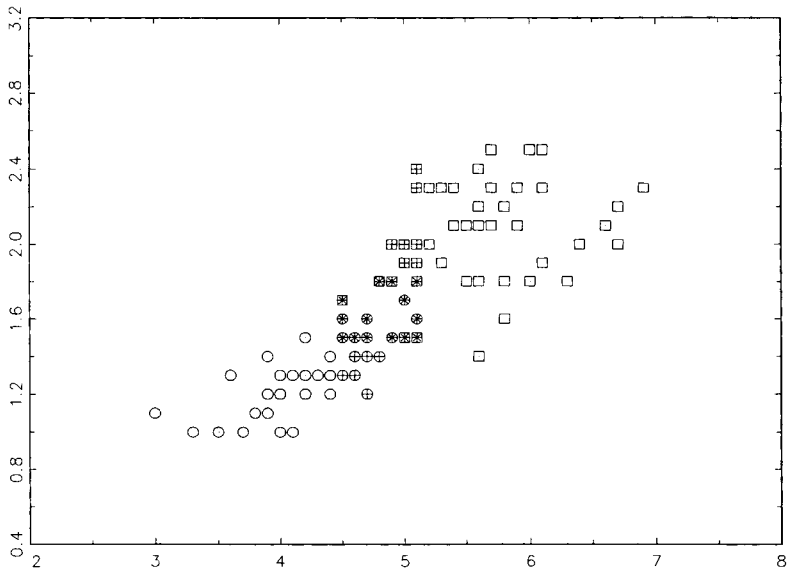

Figure 3. The square and the dot denote group membership (the dot corresponds to the species Versicolor and the square to the species Virginica). The values of X_3 are reported on the x-axis and those of X_4 on the y-axis. The symbol + denotes the members of pairs transvarying with respect to X_3 (*i.e.* the elements of the subsets G_1' and G_2' at the first step of the procedure). Among them, the members of pairs which transvary with respect to X_4 (i.e. the elements of the subsets G_1' and G_2' at the second step of the procedure) are additionally marked with the symbol $\times$.

subsets of observations is computed with respect to each of the remaining variables. Variable X_4 yields the minimum transvariation probability value; the sign of the median of the differences between the X_4 values belonging to G_1' and to G_2' is negative.

Among the elements of G_1' and G_2', the members of the pairs which transvary with respect to X_4 are selected (denoted with the additional symbol $\times$ in Figure 3) and thus subsets G_1' and G_2' are updated. Transvariation probability between these current subsets of observations is computed with respect to each of the remaining variables (X_1 and X_2), and variable X_2 is selected, whose corresponding median of the differences is positive.

Subsets G_1' and G_2' are updated by selecting, among the elements on the current subsets, the members of the pairs transvarying along X_2 (denoted with the cross + in Figure 4). The median of the differences between the values of the last variable, X_1, in these new subsets is positive.

At the end of the procedure, subsets G_1' and G_2' are determined by selecting, among the elements on the current subsets, the members of the pairs which transvary with respect to X_1 (marked with the cross + in Figure 5): more specifically, the units in G_1' correspond to specimens 17, 19, 23, 29, 34, 35 in the Versicolor group and those in G_2' correspond to specimens 20, 24, 27, 28, 34, 39, 50 in the Virginica one. Each element of G_1' (G_2') is the first (second) member of at least one transvarying pair along X_h, for h=1,…,4.

But what really matters in multivariate transvariation is the number of the pairs which simultaneously transvary along each variable, ${}_4s_{12}^*$ (note that each member of such pairs is forced to belong, respectively, to the subsets G_1' and G_2' described just above). In this example only one pair among the 50×50 pairs ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$), $i \in G_1$ and $j \in G_2$, was found to satisfy this condition, that is $x_{1i} \leq x_{1j}$, $x_{2i} \leq x_{2j}$, $x_{3i} \geq x_{3j}$, $x_{4i} \leq x_{4j}$. This pair is composed by specimen 34 in the Versicolor group, $\boldsymbol{x}_i$ =(6.0, 2.7, 5.1, 1.6), and specimen 34 in the Virginica one, $\boldsymbol{x}_j$ =(6.3, 2.8, 5.1, 1.5). Multivariate transvariation probability between the two species Versicolor and Virginica with respect to the observed variables results to be equal to 0.0048, indicating that the groups are almost perfectly separable.

An additional application

We present an application of the proposed separability measure on real data in a variable selection context. The data set is taken from Reaven and Miller (1979) and concerns an investigation on diabetes. It lists the values of the following 5 variables for 145 non-obese adult human subjects: relative weight, fasting plasma glucose, glucose area, insulin area and steady state plasma glucose. It also indicates if the subject suffers from chemical diabetes, from overt diabetes or is normal.

Imagine that we want to select a subset of the 5 above mentioned variables for discriminative purposes. For example, we could be interested in identifying the most informative pair of variables in discriminating between chemical and overt diabetic classes. For this purpose, the graphical inspection of the scatter-plot matrix of the predictors can be useful. In addition, for each of the

$$\binom{5}{2}$$

different pairs of variables, both Mahalanobis distance and the proposed separability measure have been computed. Our aim is to compare the rank-

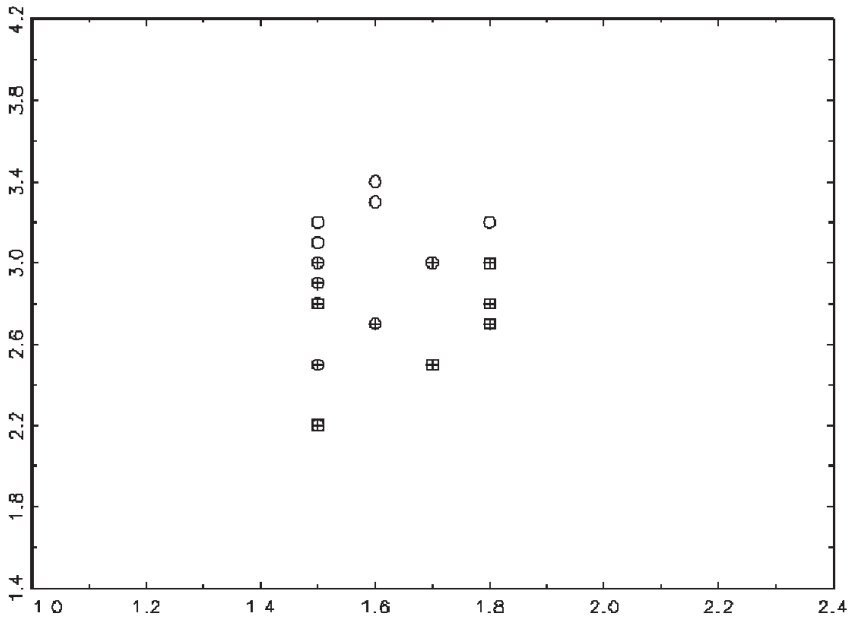

Figure 4. The square and the dot denote group membership (the dot corresponds to the species Versicolor and the square to the species Virginica). The elements of the subsets G_1' and G_2' at the second step of the procedure are plotted. The values of X_4 are on the x-axis and those of X_2 on the y-axis. The symbol + denotes the members of the pairs transvarying with respect to X_2 (*i.e.* the elements of the subsets G_1' and G_2' at the third step of the procedure).

ings of the pairs (in terms of group separation) yielded by these measures and, if the rankings are different, to see which of the two measures is in accord with the scatter-plot indications.

The pair of variables yielding the maximum value of Mahalanobis distance is 'glucose area – insulin area'. However, as the scatter-plot in the left panel of Figure 6 clearly shows, there is some overlapping between the groups with respect to these variables. On the contrary, the first 4 places of the ranking yielded by the transvariation-based measure are taken by the 4 pairs having the variable 'plasma' as a member (with values of modified bivariate transvariation probability ranging from 0 to 0.018). It suggests that the variable 'plasma' is fundamental in the separation between the groups and that it is almost sufficient, as clearly indicated by the scatter-plots (the right panel of Figure 6 reports, as an example, the plot of the pair 'plasma – relative weight').

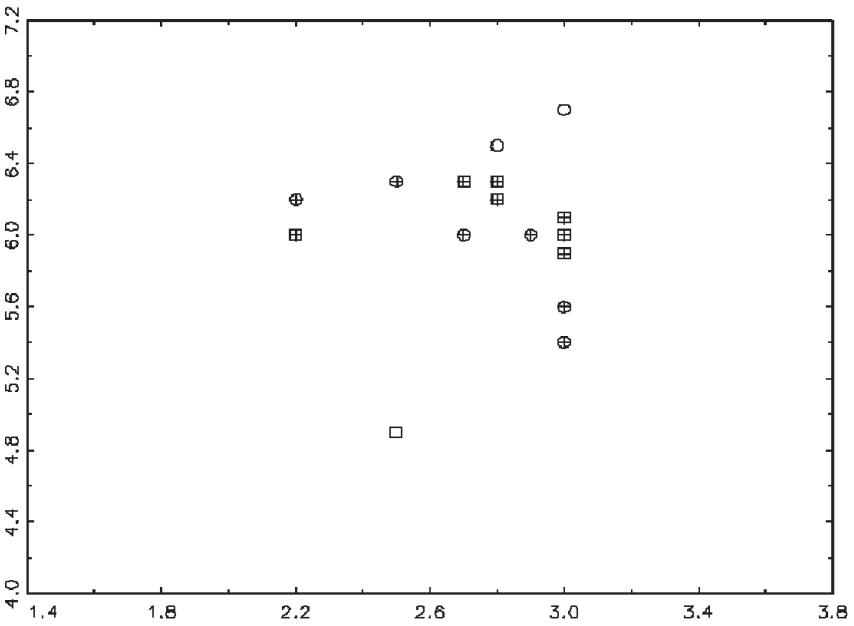

Figure 5. The square and the dot denote group membership (the dot corresponds to the species Versicolor and the square to the species Virginica). The elements of the subsets G_1' and G_2' at the third step of the procedure are plotted. The values of X_2 are on the x-axis and those of X_1 on the y-axis. The symbol + denotes the members of the pairs transvarying with respect to X_1 (*i.e.* the elements of the subsets G_1' and G_2' at the fourth and last step of the procedure).

4. A Simulation Study

The proposed separability measure has been tested on several simulated data sets and its performance compared with that of the sample Mahalanobis distance, as a commonly used measure of group separation, and of the Matusita distance (Matusita 1956), as a prototype of probabilistic separability measures:

$$D_M\left(\hat{f}_1, \hat{f}_2\right) = \left[\int \left[\sqrt{\hat{f}_1(\boldsymbol{x})} - \sqrt{\hat{f}_2(\boldsymbol{x})}\right]^2 d\boldsymbol{x}\right]^{\frac{1}{2}},$$

where $\hat{f}_1$ and $\hat{f}_2$ denote the probability density estimates obtained by the kernel method (Silverman 1986; Wand and Jones 1995) and numerical integration is applied.

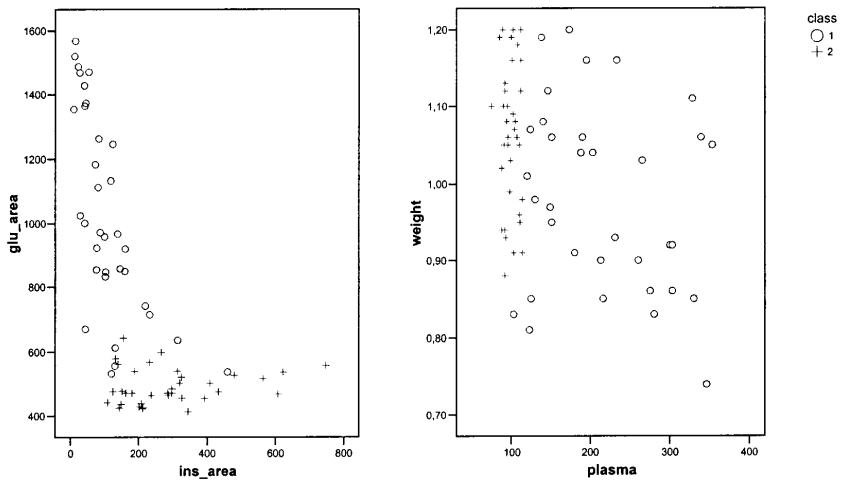

Figure 6. Two bivariate scatter-plots of the diabetes data by Reaven and Miller (1979). Codes 1 and 2 stand for "chemical diabetic" and "overt diabetic", respectively.

The sequential procedure described in Section 3, the separability measures described above and the instructions for generating the simulated samples have been implemented in GAUSS.

The distributional situations considered in this Monte Carlo study are reported in Table 1: all of them are three-variate. For each situation, 100 data sets of size 100 (50 units from each population) were generated. Each data set was projected onto the planes obtained by rotating the (x_1, x_2) plane about x_1 axis through the projection matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \end{bmatrix}$$

where the rotation angle $\theta = \pi/180, 2\pi/180, \ldots, \pi$.

Group separation has been evaluated on each plane by means of Mahalanobis distance, of Matusita distance and of the modified multivariate transvariation probability, and finally the value θ_l corresponding to the maximum group separation in the l-th data set according to each of the three measures has been recorded. The distributions of θ_l over the 100 simulated data sets for each of the three measures have been derived and summar-

Table 1. Distributional Situations

	Parental Distributions	
	Π_1	Π_2
Situation 1	Normal $\boldsymbol{\mu}_1=(0,1,1)$ $\boldsymbol{\Sigma}_1 = \mathbf{V} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -0.5 \\ 0 & -0.5 & 1 \end{bmatrix}$	Normal $\boldsymbol{\mu}_2=(0,0,0)$ $\boldsymbol{\Sigma}_2=\mathbf{V}$
Situation 2	Log-normal $\mathbf{t}_1=(0,0,1)$ $\boldsymbol{\mu}_1=(0,0,0)$ $\boldsymbol{\Sigma}_1=\mathbf{I_3}$	Log-normal $\mathbf{t}_2=(0,0,0)$ $\boldsymbol{\mu}_2=\boldsymbol{\mu}_1$ $\boldsymbol{\Sigma}_2=\mathbf{I_3}$
Situation 3	Log-normal $\mathbf{t}_1=(0,0,0)$ $\boldsymbol{\mu}_1=(0,0,1)$ $\boldsymbol{\Sigma}_1=\mathbf{I_3}$	Log-normal $\mathbf{t}_2=\mathbf{t}_1$ $\boldsymbol{\mu}_2=(0,0,0)$ $\boldsymbol{\Sigma}_2=\mathbf{I_3}$

	Contaminating Distributions	
	Π_1'	Π_2'
Situation 1a	Normal $\boldsymbol{\mu}_1=(0,1,1)$ $\boldsymbol{\Sigma}_1=144*\mathbf{V}$	Normal $\boldsymbol{\mu}_2=(0,0,0)$ $\boldsymbol{\Sigma}_2=144*\mathbf{V}$
Situation 1b	Normal $\boldsymbol{\mu}_1=(0,0,0)$ $\boldsymbol{\Sigma}_1=144*\mathbf{V}$	Normal $\boldsymbol{\mu}_2=(0,1,1)$ $\boldsymbol{\Sigma}_2=144*\mathbf{V}$

ized by means of the following descriptive measures (Upton and Fingleton 1985):

the mean direction $\overline{\theta}$, defined as the solution of the system

$$\begin{cases} \overline{\theta} = \arccos\left(\overline{C} \Big/ \sqrt{\overline{C}^2 + \overline{S}^2} \right) \\ \overline{\theta} = \arcsin\left(\overline{S} \Big/ \sqrt{\overline{C}^2 + \overline{S}^2} \right) , \end{cases}$$

where

$$\bar{C} = \frac{1}{L}\sum_{l=1}^{L}\cos(\theta_l), \qquad \bar{S} = \frac{1}{L}\sum_{l=1}^{L}\sin(\theta_l),$$

and the measure of dispersion

$$D = \frac{1}{L}\sum_{l=1}^{L}\left[1-\cos(\theta_l - \bar{\theta})\right].$$

The results are shown in Table 2.

The parameters of the normal homoscedastic distributions of Situation 1 have been chosen so that the plane maximizing population separability corresponds to 45 degrees. In situations 1a and 1b these distributions are contaminated by 15% of the normal distributions listed in the last part of Table 1: in the former situation the outlying distributions are contaminating only the scale, their marginal standard deviation being equal to 12. In addition, mean vectors are exchanged in the latter situation.

The simulation results show that in the normal homoscedastic case the measures we compared yield the same information as far as separability is concerned. As far as the comparison between the modified transvariation probability and the Mahalanobis distance is concerned, this result is coherent with what has been anticipated in Section 3 and is illustrated in Figure 7.

In the normal homoscedastic case, the Mahalanobis distance attains its maximum exactly where the transvariation-based measure is minimized. Obviously, the latter performs worse in terms of precision, since it does not rely on any distributional assumption.

The good performance of D_M is probably due to the solution we adopted to the problem of automatically specifying the bandwidth parameters in the formulation of the multivariate kernel density estimator: since the issue of data driven bandwidth selection in multivariate kernel density estimation has not been resolved so far, we resorted to a "rule of thumb" (see, for example, Scott, 1996, p.152) derived under the assumption that the unknown density is a *k*-variate normal one, which is just the case of Situation 1. In its general formulation, the rule requires that the observed variable variances (or the whole covariance matrix) are estimated; if the sample variance is used (as we did), D_M inherits the sensitiveness of this estimator to outlying observations, as well as the Mahalanobis distance. This is clearly shown in Table 1 where, in Situation 1a and Situation 1b, both the separability measures prove to be derailed by outlying observations in detecting the best plane. The measure based on the modified transvariation probability seems to be affected by outlying data to a lesser extent. This is

Table 2. Summary results for the angle θ corresponding to maximum group separation according to several separability measures: mean directions (in degrees) and dispersion measures (in brackets) for 100 replications in each distributional situation.

	Distributional Situations				
	Situation 1	Situation 1a	Situation 1b	Situation 2	Situation 3
MAHALANOBIS DISTANCE	45.15 (0.003)	46.30 (0.093)	50.65 (0.182)	89.98 (0.123)	93.146 (0.137)
MATUSITA DISTANCE	45.06 (0.003)	52.16 (0.127)	50.49 (0.105)	93.38 (0.100)	90.51 (0.085)
TRANSVARIATION BASED MEASURE	45.19 (0.008)	44.53 (0.026)	44.88 (0.028)	91.05 (0.012)	89.68 (0.061)

due to the fact that in the definition of transvariation probability what matters is only if a pair of units is a transvarying pair or not, whereas the "amount" of such a transvariation is not taken into account.

The robustness of the transvariation-based measure against outlying observations has been confirmed in another simulation experiment, where two normal heteroscedastic populations were considered, and outliers contaminating scale and affecting both location and scale estimates were introduced, as in Situation 1a and Situation 1b, respectively.

Furthermore, in order to evaluate robustness against skewness, two lognormal distributions are considered in Situation 2. Both of them correspond to a unit normal distribution, but their threshold parameters, $\mathbf{t}_1$ and $\mathbf{t}_2$, are different, so that the plane maximizing population separability corresponds to a rotation angle of 90 degrees. The results reveal that transvariation probability yields an accurate estimate of this angle, displaying the most stable behaviour throughout.

In the situations so far examined the two population distributions are the same apart from a location shift. Situation 3 is different. Two log-normal distributions are considered, corresponding to two normal distributions having unit variance-covariance matrices but different mean vectors. Their maximally separated bivariate marginals correspond to a rotation angle of 90 degrees. The transvariation-based measure and the other nonparametric measures lead to close results and outperform Mahalanobis distance, which is severely affected by deviations from normality.

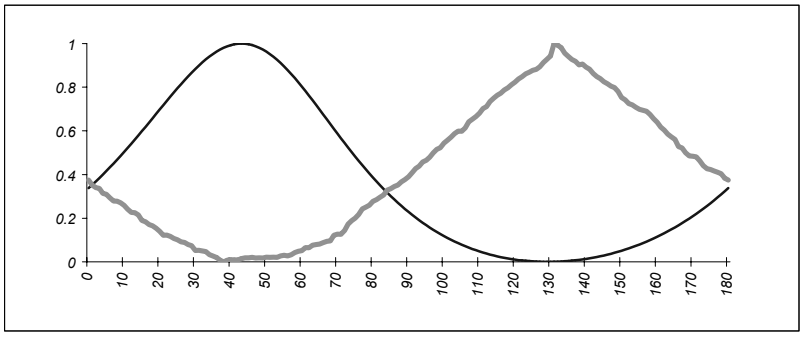

Figure 7. Values of the Mahalanobis distance (black line) and of the modified transvariation probability (gray line) as a function of the rotation angle for two samples, of 100 units each, taken from the normal homoscedastic populations of Situation1 (see Table 1). The values of both the measures have been normalized, so as to ease the graphical comparison of the two curves.

The simulation results seem therefore to suggest that the proposed definition of multivariate transvariation probability may be useful in measuring group separability and may be competitive, if not preferable, with respect to the alternatives we have considered.

At the end of this section, we present another example, where some of the measures we dealt with are compared (see Figure 8): the sample Mahalanobis distance, the sample version of Gini's multidimensional transvariation probability (Definition 2) and its version for normal data, and finally the proposed multidimensional transvariation probability. Analogously to the example illustrated in Figure 7, one data set of size 200 was generated by taking 100 units from each of two normal homoscedastic distributions with covariance matrix $\mathbf{\Sigma}$={1 0 0.8,0 1 0,0.8 0 1} and mean vectors $\boldsymbol{\mu}_1$={3,0,1} and $\boldsymbol{\mu}_2$={0,0,0}, respectively. It was projected onto the planes obtained by rotating the (x_1, x_2) plane on x_1 axis at steps π/180 wide. The values of the above-mentioned measures are illustrated in Figure 8 as a function of the rotation angle, ranging from 0 to π.

5. Concluding Remarks and Open Issues

In this paper a modified version of Gini's notion of multidimensional transvariation is proposed, which can be a useful tool for measuring the separation between two groups of multivariate observations. The measure of

group separation we present is easy to be interpreted, since it is derived as a normalized index, and it is shown to be distribution-free and robust against outlying observations.

In two-group discriminant analysis it could be used as a descriptive measure for determining how good a subset of variables is; it could be employed in feature extraction as well, as a measure to be minimized when the best discriminant low dimensional subspace is searched for (in a projection pursuit perspective). In both cases, its use in discrimination is preliminary with respect to the problem of assigning new cases to one of the groups, which has not been addressed in this paper. As a matter of fact, a coherent transvariation-based allocation rule can be devised, which however does not seem to inherit the property of being robust against outliers.

About the proposed separability measure, the following two remarks are worth noting. Firstly, both Gini's original definition of multidimensional transvariation and the one proposed are based on hyper-rectangular axis oriented sets. Thus, the number of jointly transvarying pairs is not invariant under affine transformations in both the formulations. However, it is invariant under scale transformations and location shifts. Moreover, the proposed measure may be prone to the curse of dimensionality since as the sequential procedure designed for its numerical evaluation considers further variables, statistics are computed on smaller and smaller data sets, attention being focused on subsets G_1' and G_2'.

The measures of separability can be used in variable selection, in order to evaluate the effectiveness of different sets of variables in discriminating between the classes. Discriminant rules with too many variables may be difficult to interpret, and the predictive performance of a sample classification rule tends to be affected adversely by the inclusion of irrelevant or redundant variables (McLachlan 1992); moreover, when many variables are included, the overall error rate will start to increase, due to the problem of dimensionality versus sample size, known as the *peaking phenomenon*.

When the primary aim of the analysis is discrimination, the relative importance of a subset of variables should be assessed in terms of the separation they provide among the groups. In the sequential procedure described in Section 3, at each step the individual contribution of each variable – allowing for those previously selected – to the separation of the populations is assessed. This suggests that a forward variable selection algorithm could be derived. We are at present evaluating the performance of this proposal as an alternative distribution-free solution to variable selection

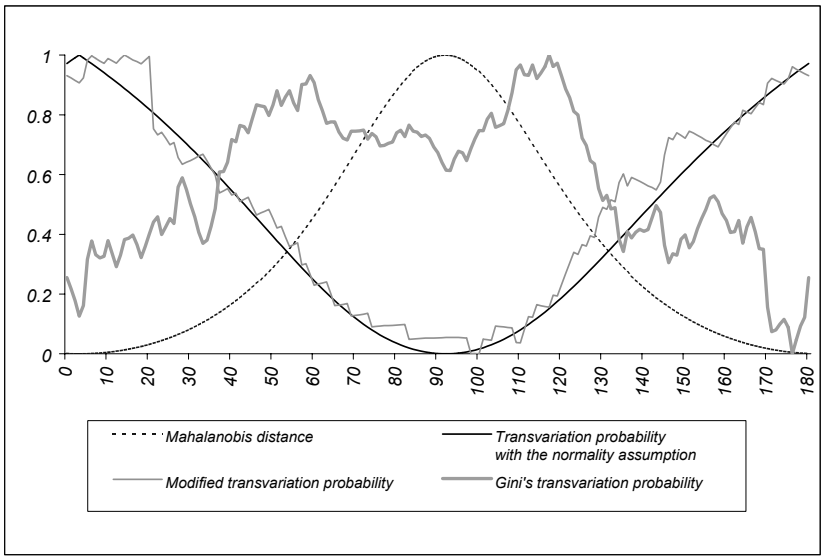

Figure 8. Values of the four measures as a function of the rotation angle for two samples, of 100 units each, taken from two normal homoscedastic populations with covariance matrix $\boldsymbol{\Sigma}$ ={1 0 0.8,0 1 0,0.8 0 1} and mean vectors $\boldsymbol{\mu}_1$={3,0,1} and $\boldsymbol{\mu}_2$={0,0,0}, respectively. The values of the measures have been normalized, so as to ease the graphical comparison of the two curves.

in discriminant analysis[2], especially when compared with the rank transformation, which is undoubtedly very simple to use and computationally less expensive (see Conover and Iman 1980).

References

BHATTACHARYYA, A. (1943), "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions", *Bulletin of the Calcutta Mathematical Society*, *35*, 99-109.

CHAKRAVARTY, S.R., and DUTTA, B. (1987), "A Note on Measures of Distance Between Income Distributions", *Journal of Economic Theory*, *41*, 185-188.

2. The classical stepwise discriminant analysis procedure is based on the criterion of no additional information: a partial (conditional) *F*-test (Rao 1973) is used to assess the individual contribution (to the separation of the two populations) of each of the remaining variables in the presence of the already selected ones. This method is based on Mahalanobis distance as a measure of separability and relies on the assumption of a homoscedastic normal model for the *k*-dimensional random feature vector in each population.

CONOVER, W.J., and IMAN, R.L. (1980), "The Rank Transformation as a Method of Discrimination with Some Examples", *Communications in Statistics-Theory and Methods*, *A9*, 465-487.

CUADRAS, C.M., FORTIANA, J., and OLIVA, F. (1997), "The Proximity of an Individual to a Population with Applications to Discriminant Analysis", *Journal of Classification, 14*, 117-136.

DAGUM, C. (1959), "Transvariazione fra più di due distribuzioni," in C. Gini (1959), 608-647.

DAGUM, C. (1965), "Probabilité et intensité de la transvariation dans l'espace á n dimensions," *Economie Appliqueé*, *18*, 507-538.

DAGUM, C. (1971), "Multivariate Transvariation Theory among Several Distributions and its Economic Applications", in *Studi di probabilità, statistica e ricerca operativa in onore di Giuseppe Pompilj*, Gubbio (Italy): Oderisi.

DAGUM, C. (1980), "Inequality Measures Between Income Distributions with Applications", *Econometrica*, *4*, 1791-1803.

DAGUM, C. (2005), "Inequality Decomposition, Directional Economic Distance, Metric Distance and Gini Dissimilarity between Income Distributions", in *Proceedings of the International Conference in Memory of Two Eminent Social Scientists: C. Gini and M. O. Lorenz* (to appear).

DEUTSCH, J., and SILBER, J. (1997), "Gini's 'Transvariazione' and the Measurement of Distance Between Distributions", *Empirical Economics*, *22*, 547-554.

DEVIJVER, P.A., and KITTLER, J. (1982), *Pattern Recognition: A Statistical Approach*, Englewood Cliffs: Prentice-Hall.

EBERT, U. (1984), "Measures of Distance Between Income Distributions", *Journal of Economic Theory*, *32*, 266-274.

FLIGNER, M.A., and POLICELLO, G.E. (1981), "Robust Rank Procedures for the Behrens-Fisher Problem", *Journal of the American Statistical Association*, *76*, 162-168.

GINI, C. (1916), "Il concetto di transvariazione e le sue prime applicazioni", *Giornale degli Economisti e Rivista di Statistica*; in C. Gini (1959), 1-55.

GINI, C. (1959), *Transvariazione*, Roma: Libreria Goliardica.

GINI, C., and LIVADA, G. (1943), "Transvariazione a più dimensioni", in *Atti della VI Riunione della Società Italiana di Statistica*; in Gini C. (1959), 216-253.

HAND, D.J. (1997), *Construction and Assessment of Classification Rules*, Chichester: Wiley.

HETTMANSPERGER, T. (1984), *Statistical Inference Based on Ranks*, New York: Wiley.

HODGES, J.L. Jr, and LEHMANN, E.L. (1963), "Estimates of Location Based on Rank Tests", *Annals of Mathematical Statistics*, *34*, 598-611.

HOLLANDER, M., and WOLFE, D.A. (1999), *Nonparametric Statistical Methods*, New York: Wiley Series in Probability and Statistics.

JEFFREYS, H. (1948), *Theory of Probability* (2nd ed.), Oxford: Clarendon Press.

KRZANOWSKI, W.J., and MARRIOTT, F.H.C. (1995), *Multivariate Analysis*, *Part 2*: *Classification, Covariance Structures and Repeated Measurements*, New York: Wiley.

LISSACK, T., and FU, K.S. (1976), "Error Estimation in Pattern Recognition Via L^{o}-Distance Between Posterior Density Functions", *IEEE Transactions on Information Theory*, *22*, 34-45.

MAHALANOBIS, P.C. (1930), "On Tests and Measures of Group Divergence", *Journal and Proceedings of the Asiatic Society of Bengal*, *26*, 541-588.

MAHALANOBIS, P.C. (1936), "On the Generalized Distance in Statistics", *Proceedings of the National Institute of Science India*, *2*, 49-55.

MATUSITA, K. (1956), "Decision Rule, Based on Distance, for the Classification Problem", *Annals of the Institute of Statistical Mathematics*, *8*, 67-77.

MCLACHLAN, G.J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

MITCHELL, A.F.S., and KRZANOWSKI, W.J. (1985), "The Mahalanobis Distance and Elliptical Disributions", *Biometrika*, *72*, 464-467.

MONTANARI, A. (2004), "Linear Discriminant Analysis and Transvariation", *Journal of Classification*, *21*, 71-88.

RAO, C.R. (1973), *Linear Statistical Inference and Its Applications*, New York: Wiley.

RAO, C.R. (1982), "Diversity and Dissimilarity Coefficients: A Unified Approach", *Theoretical Population Biology*, *21*, 24-43.

REAVEN, G.M., and MILLER, R.G. (1979), "An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis", *Diabetologia*, *16*, 17-24.

SCOTT , D. (1992), *Multivariate Density Estimation*, New York: Wiley.

SILVERMAN, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

UPTON, G., and FINGLETON, B. (1989), *Spatial Data Analysis by Example*, New York: Wiley.

WAND, M.P., and JONES, M.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

YITZHAKI, S. (1994), "Economic Distance and Overlapping of Distributions", *Journal of Econometrics*, *61*, 147-159.