# Gini's contribution to Multivariate Statistical Analysis

Angela Montanari and Paola Monari

Dipartimento di Scienze Statistiche
Università di Bologna
e-mail:montanar@stat.unibo.it

**Abstract**

Corrado Gini (1884-1964) may be considered the greatest Italian statistician. We believe that his important contributions to statistics, however mainly limited to the univariate context, may be profitably employed in modern multivariate statistical methods, aimed at overcoming the curse of dimensionality by decomposing multivariate problems into a series of suitably posed univariate ones.
In this paper we critically summarize Gini's proposals and consider their impact on multivariate statistical methods both reviewing already well established applications and suggesting new perspectives.
Particular attention will be devoted to classification and regression trees, multiple linear regression, linear dimension reduction methods and transvariation based discrimination.

**Key words and phrases**: classification and regression trees, Gini's mean difference, heterogeneity, concentration ratio, linear regression, principal components, discriminant analysis, multivariate median, transvariation

**About the authors:** Angela Montanari is Full Professor of Statistics in the Faculty of Statistical Sciences at Bologna University, Bologna, Italy (e-mail:montanar@stat.unibo.it) Paola Monari is Full Professor of Statistics in the Faculty of Statistical Sciences at Bologna University, Bologna, Italy (e-mail:monari@stat.unibo.it)

## 1. Introduction

Corrado Gini may be considered the greatest Italian statistician. He intensely worked from 1905 up to the end of his life; a long period during which, modern statistics acquired its distinguishing features and reached its highest levels. He studied Law at the University of Bologna, but like the great figures of the early twentieth century, he was

an eclectic and multidisciplinary scientist whose interests ranged from statistics to biometry, from economics to sociology and demography.

During his long life he wrote 87 books and more than 800 papers and brought a lot of new ideas in many aspects of statistical theory and practice.

Gini's contribution to statistics can be divided into three periods. During the first one, until 1920, he introduced his most original and innovative developments to the theory of averages, variability and statistical relations.

Roughly up to the beginning of the second world war, Gini became more and more involved in the social and economic problems connected with his institutional activities, as a League of nations expert and advisor of the Italian Government. All his chief methodological contributions fall into the first two periods, while the critical review of the foundations of statistics, which ultimately gave Gini's method its unity, belongs to the third period.

In the last fifty years, the interest about Gini's methodological contributions has been mainly focused on the measures of concentration and on the application of income distribution modelling to the evaluation of human capital. A wide literature has flourished on these subjects (Dagum, 1987), whose origin may be traced back to Gini's papers dated 1912 and 1914. In the same time, the statistical community has progressively diminished its interest in the foundations of statistical inference and has deeply addressed multivariate inferential issues, encouraged by the always improving tools provided by automatic data processing. It has long been believed that no trace of Gini's work could be found in this wide branch of the statistical methodology, and for a long time his studies on the analysis of variability and on the problem of classification have been neglected, studies which gave rise to authentically original indices like the mean differences and the measures of transvariation.

It's true that in his vast production Gini never explicitly addressed multivariate issues, if one excludes some works on multidimensional transvariation and on the multivariate median on which we will dwell later, and it was a precise choice of interest.

He was well aware of the developments in the field that were taking place at his time. As the editor and founder of Metron he had got in touch with R.A.Fisher, also publishing his paper " On the probable error of a coefficient of correlation deduced from a small sample" which Pearson had refused for Biometrika. He debated with those two masters on the logical foundations of statistical inference and hypothesis testing, but, according to our recollections, he was not specifically concerned with multivariate

methods. He was very well acquainted with the Anglo Saxon contributions, for instance Pearson's 1901 paper "On lines and planes of closest fit to a system of points in space" which is always quoted in the statistical literature as the origin of principal components analysis. In his paper "Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali" (1921) Gini commented upon it, noticing that Pearson's solution may be really useful in the error in variables case (Fuller, 1987) when the fit has a descriptive aim, but may fail to fulfil the purpose when the aim is that of detecting the true relationship which would have emerged had the variables been measured without error. When the latter is the goal, Gini showed that Pearson's solution implicitly admits errors of equal intensity in all the variables involved and suggested a new method capable of dealing with more general error conditions, allowing for Pearson's solution as a special case. But he never went beyond considering the two variable case.

Recently larger and larger data sets, also as far as the number of observed variables is concerned, have become available thus requiring the development of new techniques also capable to face the so called curse of dimensionality. The result has been the flourishing of methods whose common feature is the reinterpretation of the solution of multivariate problems as the solution of a sequence of suitably posed univariate ones. In this new setting Gini's contribution emerges as original and fundamental (Monari and Montanari, 2003).

## 2. Gini indexes for classification and regression trees

Among Gini's ideas, the best known to statisticians working in multivariate statistical analysis is what is generally called Gini index, largely employed in the context of classification tree methodology (Brieman, Friedman, Olshen and Stone, 1984).

In a J class problem, denoting by **x** the p-dimensional measurement vector corresponding to a given case and by X the measurement space defined to contain all possible measurement vectors, a classifier is a partition of X into J disjoint subsets $A_1$, …, $A_j$, …, $A_J$, $X = \bigcup_j A_j$, such that for every $x \in A_j$ the predicted class is $j$. In a binary tree structured classifier such a partition is reached by repeated splits of subsets of X into two descendant subsets beginning with X itself. According to tree terminology a generic subset of X is called a node t. The terminal subsets (called terminal nodes) form a partition of X. Each terminal subset is denoted by a class label. The partition

corresponding to the classifier is got by putting together all the terminal subsets corresponding to the same class.

A first interesting feature of the method is that the splits are formed by conditions on the coordinates of **x**, thus translating a multivariate problem into a sequence of suitably posed univariate ones. A second aspect worth underlining here, as it directly leads to Gini index, is that each split is selected so that the data in each of the descendant subsets are "purer" than the data in the parent subset.

This requires to define an impurity function which, according to Brieman *et al*., is a function $\phi$ defined on the set of all J-tuples of numbers $(p_1,..., p_J)$ satisfying $p_j \geq 0, j = 1,..., J, \sum_j p_j = 1$ with the properties

(i) $\phi$ is maximum only at the point $(\frac{1}{J}, \frac{1}{J},..., \frac{1}{J})$

(ii) $\phi$ achieves its minimum only at the points (1,0,…0); (0,1,…,0); … (0,0,…,1)

(iii) $\phi$ is a symmetric function of $(p_1,..., p_J)$.

Given an impurity function $\phi$, the impurity measure of any node *t*, may be defined as

$$i(t) = \phi(p(1 \mid t),..., p(j \mid t),...p(J \mid t)) \quad (1)$$

where $p(j \mid t),\ j = 1,..., J$, denotes the proportion of cases in node *t* belonging to class *j*.

If a split *s* of a node *t* sends a proportion $p_R$ of the data cases in *t* to $t_R$ and the proportion $p_L$ to $t_L$, the decrease in impurity due to the split may be defined as

$$\delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L) \quad (2)$$

and the split $s^*$ which maximises it is selected.

A useful expression for $i(t)$ is indeed represented by the Gini index, which is defined as

$$i(t) = \sum_{j \neq i} p(j \mid t) p(i \mid t) \quad (3)$$

or equivalently as

$$i(t) = \left( \sum_j p(j \mid t) \right)^2 - \sum_j p^2(j \mid t) = 1 - \sum_j p^2(j \mid t). \quad (4)$$

Expression (4) satisfies properties (i)-(iii) and is a concave function of class probabilities, thus guaranteeing that $\delta i(s,t) \geq 0$, for any split *s*.

As Brieman *et al*. suggest, this index has an "interesting interpretation". If one assigns an object selected at random from a node *t* to class *i* with probability $p(i \mid t)$, given that

the estimated probability the item is actually in class *j* is $p(j|t)$, the estimated probability of misclassification associated to this assignment rule is the Gini index (3).

Light and Margolin (1971) introduced a different interpretation in terms of variances: in a node *t,* assign all class *j* objects the value 1, and all other objects the value 0. Then the sample variance of these values is $p(j|t)[1-p(j|t)]$. If this is repeated for all the *J* classes and the variances summed, the result is again the Gini index. Infact

$$\sum_j p(j|t)[1-p(j|t)] = 1 - \sum_j p^2(j|t).$$

A third interpretation is the one Gini gave in his almost never quoted paper "Variabilità e Mutabilità" (1912) (reprinted in Gini, 1939), where he first introduced his index.

This original interpretation is deeply rooted in Gini's theory of variability and so a short digression is here necessary in order to view it in the proper light. In Gini's view, the goal of a variability measure differs according to the nature of the characters which are being studied. If the character keeps its intensity, but appears with different values only because of random or systematic measurement errors (the repeated measures of the same quantity, for instance) the goal of a variability measure is that of determining how much the observed quantities differ from the true one. On the contrary if the character really takes different values for different statistical units (i.e. income, weight, etc.) the goal of a variability measure is that of determining how much the observed quantities differ from each other.

Gini approached the latter problem by a measure, known as Gini's mean difference, which is defined, for *n* observed values of a variable *X*, as the average of all the possible differences between those values (also including the comparison of a unit with itself):

$$\Delta = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| \Big/ n^2 . \qquad (5)$$

It follows from (5) that:

$$\Delta = \frac{2\sum_{i=1}^{n}(2i-n-1)x_{(i)}}{n^2} \qquad (6)$$

(where $x_{(i)}$ denotes the *i*-th observation in the ascending ranking of the observations).

An extension of this measure to qualitative variables, such as the identifiers of the *J* classes in a *J* class classification problem, was proposed by Gini in the same paper (Gini, 1912).

Denoted by $n_1, n_2, ..., n_J$ the frequencies of each of the *J* distinct attributes of a given nominal character, such that $\sum_{j=1}^{J} n_j = n$, and setting equal to 1 the diversity between any two different attributes, the sum of the differences of a case which shows the *j*-th attribute from all the other cases will be $n_j \cdot 0 + (n - n_j) \cdot 1$, and the sum of all the possible $n^2$ differences $\sum_{j=1}^{J} n_j (n - n_j)$. Gini's mean difference will then be $\Delta = \sum_{j=1}^{J} n_j (n - n_j) \Big/ n^2$, or equivalently, after putting $p_j = n_j / n$,

$$\Delta = \sum_{j=1}^{J} p_j (1 - p_j) \tag{7}$$

where one can easily recognise equation (4) which was also derived by Brieman *et al*.

For a J class classification problem with ordered classes Brieman *et al*. suggest to resort to a different splitting criterion, they call ordered twoing, accounting for the class order. A different solution, which however requires further investigation, may be based on Gini's mean difference for ordinal characters.

Assuming equally spaced ordinal categories, Gini proposes to code the classes according to the first *J* natural numbers and then compute the average mean difference (5). This measure however doesn't fulfil all the properties advocated by Brieman *et al*. for an impurity function: it is minimum when all the units belong to the same category, but it takes its maximum value only when half the cases belong to the lowest class and half to the highest. The use of this measure as a splitting criterion requires therefore a more thorough evaluation of its properties and performances.

More promising seems the possibility of using a function based on Gini's mean difference as a splitting criterion in the regression tree context, as an alternative to ordinary least squares or least absolute deviation fitting methods.

Ordinary least squares (OLS) regression trees produce a partition of the covariate space such that, within each element of the partition, the regression function may be approximated by the mean value of the response variable *Y* corresponding to those units whose covariate values belong to that partition member.

This is obtained by choosing the total within node sum of squares as the split function:

$$\frac{1}{n} \sum_{t \in \tilde{T}} \sum_{x_j \in t} \left( y_j - \bar{y}(t) \right)^2 \tag{8}$$

(where $\tilde{T}$ is the set of terminal nodes, $y_j$ is the response value measured on the *j*-th statistical unit belonging to node *t* and $\bar{y}(t)$ is the node average response value) and by iteratively splitting nodes so as to maximise its decrease.

An alternative solution is to choose

$$\frac{1}{n}\sum_{t\in\tilde{T}}\sum_{x_j\in t}\left|y_j - M(t)\right| \tag{9}$$

(where $M(t)$ is the sample median of the $y$ values in node *t*). This leads to least absolute deviation (LAD) trees and amounts to approximate the regression surface, within each element of the partition, by the median of the response values in the node and to choose those splits which iteratively allow to minimise the sum of the absolute deviations from the node medians.

A further possibility is to choose a split function derived from Gini's mean difference:

$$\frac{1}{n^2}\sum_{t\in\tilde{T}}\sum_{\substack{x_j\in t\\ x_i\in t}}\left|y_j - y_i\right|. \tag{10}$$

As the Gini's mean difference for a variable Y may be rewritten as

$$\Delta = \frac{\sum_{i=1}^{n} d_{i,M}\left|y_i - M\right|}{\sum_{i=1}^{n} d_{i,M}} \tag{11}$$

(where *M* is the median of the *n* units and $d_{i,M}$ is the rank of the difference $|y_i - M|$ in the ascending sequence of absolute differences) that is as a weighted average of the absolute distances from the median, with weights equal to $d_{i,M}$, using it as a split function amounts to approximate once again the regression surface by the median of the response values in a given node but to chose those splits which iteratively allow to minimise the sum of the weighted absolute deviations from the node medians.

In other words, the median is still the function approximator, as in LAD trees, but is computed on possibly different sets of units. As Gini himself underlines, $\Delta$ differs from the mean absolute deviation from the median as it overweights larger deviations (see also Bowley, 1920). The standard deviation does too, but in a totally different way: while $\Delta$ gives a weight which is proportional to the rank of the deviation, the standard deviation weights each difference according to its intensity.

The use of the split function based on Gini's mean difference may then lead to regression trees whose properties are intermediate between those of OLS and LAD trees, but much work has still to be done.

It may be interesting to note that , after putting $p(t) = n(t)/n$ (where $n(t)$ is the number of units belonging to node *t*), the function which is minimised at each step in the growing of OLS trees can be interpreted as the weighted average of within node variances ($s^2(t)$), $\sum_{t\in\tilde{T}} s^2(t)p(t)$, the one which is minimised in LAD trees as the weighted average of the within node sum of absolute deviations from the median ($\bar{d}(t)$) $\sum_{t\in\tilde{T}} \bar{d}(t)p(t)$, the one which would be minimised in what we might call MD (mean difference) trees is the weighted average of within node Gini's mean difference ($\Delta(t)$) $\sum_{t\in\tilde{T}} \Delta(t)p^2(t)$.

Before concluding this section on classification and regression trees a further index due to Gini (or better its complement to 1) is worth remembering as it may be interpreted as an impurity measure in the sense of Brieman *et al*. for it satisfies properties (i)-(iii). It is Gini's concentration ratio, introduced in 1914 in order to provide a suitable measure for income concentration and widely known in the statistical community involved in the study of income distributions.

Denoting by $p_{(1)}, p_{(2)},..., p_{(J)}$ the class frequencies arranged in ascending order, and putting $l(j) = j/J$ and $q(j) = p_{(1)} + p_{(2)} + ... + p_{(j)}$, Gini's ratio is

$$R = \frac{\sum_{j=1}^{J-1}[l(j) - q(j)]}{\sum_{j=1}^{J-1} l(j)} \tag{12}$$

and its complement

$$R' = 1 - R = \frac{2}{J-1}\sum_{j=1}^{J-1}(J-j)p_{(j)} \tag{13}$$

The properties of this index as an impurity measure for classification trees have so far not been investigated (see Brizzi (2002) for a thorough study of the descriptive properties of this and of other heterogeneity measures). Gini's ratio has on the contrary already provided good results when used as an alternative to the OLS criterion for the construction of regression trees aimed at modelling income inequalities (Costa, Galimberti, Montanari 2005)

## 3. Gini's mean difference for regression and linear dimension reduction

As already mentioned, Gini's mean difference can be defined in a variety of ways, each of which provides some insight into it. Stuart (1954) has derived the following expression:

$$\Delta = 4\,cov\left(X, F(X)\right) = \Delta_{xx} = E_{X_1} E_{X_2} \left|X_1 - X_2\right| \tag{14}$$

(where $F(X)$ denotes the cumulative distribution function of the random variable *X*). Hence $\Delta$ is interpreted as a function of the covariance between a variate and its rank and its empirical estimate $\tilde{\Delta}_{xx}$ is obtained by (5).

Expression (14) suggests a possible way of defining the analogous of the classical covariance and correlation based on the concept of mean difference. What may be called Gini covariance (Taguchi, 1981; Schechtman and Yitzhaki, 1987) or co-difference is then defined as

$$4\,\mathrm{cov}\left(Y, F(X)\right) = \Delta_{XY} = E_{(X_1,Y_1)} E_{(X_2,Y_2)} \left\{\left[\mathrm{sgn}(X_1 - X_2)\right](Y_1 - Y_2)\right\} \tag{15}$$

or as

$$4\,\mathrm{cov}\left(X, F(Y)\right) = \Delta_{YX} = E_{(X_1,Y_1)} E_{(X_2,Y_2)} \left\{\left[\mathrm{sgn}(Y_1 - Y_2)\right](X_1 - X_2)\right\}. \tag{16}$$

$(X_1, Y_1)$ and $(X_2, Y_2)$ are mutually independent pairs of random variables with the same joint density function and

$$sgn(x) = \begin{cases} -1 \text{ for } x < 0 \\ 0 \text{ for } x = 0 \\ 1 \text{ for } x > 0 \end{cases}$$

according to which variable is ranked.

$\Delta_{XY}$ can be empirically estimated as

$$\tilde{\Delta}_{XY} = \sum_i \sum_j \left\{\left[\mathrm{sgn}\left(x_i - x_j\right)\right]\left(y_i - y_j\right)\right\} \Big/ n^2 \tag{17}$$

whose computational version is

$$\tilde{\Delta}_{XY} = \frac{2\sum_{i=1}^{n} (2i - n - 1) y_{r(x_i)}}{n^2} \tag{18}$$

(where $y_{r(x_i)}$ is the observation on Y that corresponds to the *i*-th lowest value of X).

Gini correlation is therefore

$$C(X,Y) = \frac{\Delta_{XY}}{\Delta_{YY}} \text{ and } C(Y,X) = \frac{\Delta_{YX}}{\Delta_{XX}}.$$

The first striking feature of these quantities is that unlike ordinary covariance and Pearson's correlation, neither Gini covariance nor Gini correlation are symmetric.

However Schechtman and Yitzhaki have proved that

1) $-1 \le C(X,Y) \le 1$ for all X,Y
2) if X and Y are independent random variables $C(X,Y) = C(Y,X) = 0$
3) if X and Y are exchangeable random variables, then $C(X,Y) = C(Y,X)$
4) if (X,Y) has a bivariate normal distribution with correlation $\rho$, then $C(X,Y) = C(Y,X) = \rho$

In order to maintain symmetry in all cases Taguchi (1981) suggests the squared Gini correlation coefficient

$$\frac{\left|cov(X,F(Y))cov(Y,F(X))\right|}{cov(X,F(X))cov(Y,F(Y))} \tag{19}$$

which may be proved to vary between 0 and 1.

The correlation measure that varies between –1 and 1 is then given by the square root of (19) with the sign determined by

$$sgn\left(\frac{cov(X,F(Y))}{cov(X,F(X))} + \frac{cov(Y,F(X))}{cov(Y,F(Y))}\right). \tag{20}$$

These concepts have been widely employed in order to develop what have been called Gini multiple linear regression and Gini principal components analysis.

The first problem, multiple linear regression, has been approached from various perspectives, all based on the idea of Gini's mean difference.

Olkin and Yitzhaki (1992) for instance suggest to estimate the parameters of the linear multiple model

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$ where $E(\varepsilon) = 0$

by minimising

$\tilde{\Delta}_{ee} = \mathrm{cov}(e, R(e)) = \sum_i e_i R(e_i)$ where *e* represents the sample residual and *R(e)* is the rank of the estimated error term.

The method doesn't allow to estimate $\alpha$, which is therefore estimated resorting to the OLS intercept estimator, based on the previously obtained regression coefficients.

Minimisation of Gini's mean difference of the sample error term yields the first order conditions

$$\mathrm{cov}(X_k, R(e)) = 0 \quad k = 1,..., p \tag{21}$$

which are analogous to the normal equations in OLS regression estimation.

Since the error *e* is a function of all the regression weights the solution of (21) yields the partial effect of $X_k$ on Y , which is similar to that obtained from OLS.

Observing that $R(e)$ is strongly dependent on how the residuals have been estimated and that the rank of the true population residual corresponding to a given covariate vector is unknown, Podder (2002) suggests a more general framework within which a different version of Gini regression can be derived.

His proposal is to obtain an estimator for the vector parameter $\beta$ by optimising a weighted sum of the residuals, which is a translation invariant measure of dispersion:

$$\sum_{i=1}^{n} w_i e_i \tag{22}$$

under the constraint that $\sum_{i=1}^{n} w_i = 0$.

The weights $w_i = 2(2i-n-1)/n^2$ which appear in the computational expression of Gini's mean difference satisfy this constraint. With such weights one can easily verify that $\alpha$ cannot be directly estimated and Podder too suggests to estimate it by the OLS intercept estimator.

The open problem is then to identify the best variable to be ranked, that is the suitable variable to which the rank *i* in Gini's weight refers.

While Olkin and Yitzhaki suggest to rank the residuals themselves, Podder proposes to weight each residual by a function of the rank of the corresponding $X_k$ value. As minimising (22) amounts to make it vanish (see Podder for a detailed proof), by equating $\sum_{i=1}^{n} w_i e_i = \sum_{i=1}^{n} w_i \left(y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip}\right)$ to 0 with respect to each of the regressors (i.e. by choosing the weights according to the rank of each of the regressors) one obtains a set of $p$ normal equations which suggest that the residuals of a linear regression model estimated according to Podder's suggestion satisfy the condition $\operatorname{cov}\left(R(X_k), e\right) = 0 \quad \forall k$. Denoting the estimated covariance between the rank of the *k*-th regressor and the *j*-th regressor as $\tilde{\Delta}_{kj}$ and the estimated covariance between the rank of the *k*-th regressor and the dependent variable as $\tilde{\Delta}_{ky}$, the *k*-th normal equation is

$$\tilde{\beta}_1\tilde{\Delta}_{k1} + \tilde{\beta}_2\tilde{\Delta}_{k2} + \ldots + \tilde{\beta}_p\tilde{\Delta}_{kp} = \tilde{\Delta}_{ky} \qquad (k = 1,\ldots p)$$

In order to solve this system of *p* equations one has to resort to Gini's difference-codifference matrix of order *p*

$$\mathbf{D}_{xx} = \begin{bmatrix} \widetilde{\Delta}_{11} & \widetilde{\Delta}_{12} & \dots & \widetilde{\Delta}_{1p} \\ \widetilde{\Delta}_{21} & \widetilde{\Delta}_{22} & \dots & \widetilde{\Delta}_{2p} \\ \dots & \dots & \dots & \dots \\ \widetilde{\Delta}_{p1} & \widetilde{\Delta}_{p2} & \dots & \widetilde{\Delta}_{pp} \end{bmatrix} \tag{23}$$

which we assume to be full rank. Similarly we can define the codifference vector

$$\mathbf{D}_{xy} = \begin{bmatrix} \widetilde{\Delta}_{1y} \\ \widetilde{\Delta}_{2y} \\ \dots \\ \widetilde{\Delta}_{py} \end{bmatrix}. \tag{24}$$

The matrix form of the set of normal equations is then $\mathbf{D}_{xx}\widetilde{\beta} = \mathbf{D}_{xy}$ and its solution is $\widetilde{\beta} = \mathbf{D}_{xx}^{-1}\mathbf{D}_{xy}$ whose structure closely resembles the one obtained by OLS.

Podder derives some properties of his regression estimator, he proves it is unbiased, consistent, asymptotically normal and provides an expression for its variance. He also suggests it is more robust than the OLS estimator against outlying observations, but much work in this direction has still to be done and a thorough study of the performances of this estimator in presence of data pathologies (not only outliers or leverage points, but also multicollinearity) is necessary.

Based on $\mathbf{D}_{xx}$ transpose (it has already been stressed that $\mathbf{D}_{xx}$ is not symmetric) Baccini and de Falguerolles (1993) propose to derive a Gini analogue of principal components. Just as the coefficients of Hotelling's principal components can be derived after a singular value decomposition of the covariance matrix of the observed variables, the coefficients of Gini principal components are derived from the singular value decomposition of the codifference matrix $\mathbf{D}'_{xx}$.

However admissible from a computational point of view, the interpretation of the solution is not completely satisfactory, as it doesn't clearly show which optimality properties Gini components enjoy. $\mathbf{D}_{xx}$ cannot be viewed as an estimate of the population variance covariance matrix: Gini himself showed that variance and mean difference highlight different aspects of variability and answer different questions. Furthermore while classical principal components are the maximum variance linear combinations of the observed variables, it cannot be proved that the left singular vectors of $\mathbf{D}'_{xx}$ identify the directions along which Gini's mean difference is maximum.

A more interesting solution may perhaps be derived by casting the problem within the so called projection pursuit framework (Huber, 1985) and directly looking for the linear combination of the observed variables which optimises Gini's mean difference. The

search can be repeated by restricting it each time to the orthogonal complement of the directions identified in the preceding steps.

Liner combinations of the observed variables have found a wide use also in the context of discriminant analysis. The first idea goes back to Fisher (1936) who suggested to search directly the linear combination of the *p* measured characteristics which maximises group separation, defined as the ratio of "between" to "within" group variance under the condition of homoscedasticity. Apparently, it does not require any distributional assumption, but normality or at least symmetry is actually implicitly assumed. Furthermore it is well known that Fisher's function is not robust against outlying observations and against violations of normality and homoscedasticity.

In order to maintain the ease of interpretation of Fisher's function while avoiding the normality and heteroscedasticity assumption Posse (1992) proposed a projection pursuit version of linear discriminant analysis based on the search of the linear combination showing the minimum total probability of misclassification.

It obviously gives the best error rates as far as the classification of the units belonging to the training sample is concerned, but for small or moderate sample sizes nothing guarantees its good performances on new cases whose group membership is to be determined, in other words, it may be derailed by a sort of overfitting effect. A different promising solution may be obtained by looking for the linear combination which optimises group separation in terms of Gini's transvariaton (Montanari and Calò 1998; Montanari 2004).

According to Gini (1916), two groups $g_1$ and $g_2$ are said to transvariate on a variable $X$, with respect to their corresponding mean values $m_{x1}$ and $m_{x2}$ if the sign of some of the $n_1 n_2$ differences $x_{i1} - x_{j2}$ $(i = 1,2,...,n_1 \quad j = 1,2,...n_2)$ which can be defined between the $x$ values belonging to the two groups is opposite to that of $m_{x1} - m_{x2}$. Any difference satisfying this condition is called "a transvariation" and $\left|x_{i1} - x_{j2}\right|$ is its intensity. (It's worth noting that Wilcoxon- Mann Whitney two sample test is based on the same idea).

In order to measure the transvariation between two groups Gini first introduced, among others, the concepts of transvariation probability, transvariation intensity and transvariation area.

Transvariation probability is defined as the ratio of the number of transvariations (assuming the median as mean value) to its maximum. It takes value in the interval [0,1] and the more the two groups overlap, the greater its values. Its complement to 1

formally translates the notion of separability due to Hand (1997):"Two classes are said to be 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".

Denoted by $f_k(x)\ \ k=1,2$ the probability density function of X in $\Pi_k$, the parent population of group $g_k$, the transvariation area is

$$\int_{-\infty}^{+\infty}\Psi(x)dx \text{ where } \Psi(x)=\min\left(f_1(x),f_2(x)\right). \tag{25}$$

When the transvariation probability is zero, the two groups do not overlap and therefore the transvariation area is also zero, but the inverse is not always true. This means that the two measures usually highlight different aspects of group transvariation.

Transvariation intensity (defined with respect to the arithmetic mean) is equal to the ratio of the sum of transvariation intensities to its maximum.

The above description may have shown that transvariation measures can be profitably used to discriminate between two groups. A linear discriminant function can then be derived as the linear combination which minimises transvariation probability, intensity or area. For normal data the three solutions coincide and also coincide with Fisher linear discriminant function.

A closer look at the formal expression of transvariation area shows that it is but twice the total probability of misclassification, therefore minimising transvariation area leads to the solution obtained by Posse. Based on the study of the statistical properties of transvariation measures and on simulations Montanari (2004) shows that the linear disciminant function obtained by optimising transvariation probability generally gives the best results as it is robust against violations of normality and homoscedasticity assumptions and against the presence of outliers.

## 4. New perspectives: multivariate transvariation and multivariate median

As already mentioned in the introduction, the only explicit Gini's contribution to multivariate analysis is confined to multivariate transvariation (Gini and Livada, 1959) and to the concept of spatial median (Gini and Galvani, 1929).

In most of the cases he developed the methods for the two variable case because "it is the most frequent and most interesting for practical applications" (Gini 1959) and

mentioned the possibility of extending them to more tan two variables jointly observed on a given set of statistical units.

It is also worth noting that in his works on the above mentioned topics he had a co-author. A further proof of his limited interest in multivariate topics.

Extending his notion of transvariation between two groups to the multivariate context he required that at least one pair of units simultaneously transvariates on each variable. A more formal definition may be given as:

two groups $g_1$ and $g_2$ of $n_1$ and $n_2$ units respectively, are said to transvariate on a *p*-dimensional variable **X,** with respect to their corresponding mean vectors $\mathbf{m}_{1X}$ and $\mathbf{m}_{2X}$, if there exists at least one pair $(\mathbf{x}_i, \mathbf{x}_j)$, with $\mathbf{x}_i \in g_1$ and $\mathbf{x}_j \in g_2$, such that for any variable $X_k\,(k = 1,..., p)$ the difference $x_{ki} - x_{kj}$ is opposite to that of $m_{1X_k} - m_{2X_k}$.

Transvariation may still be measured by transvariation probability, again defined as the ratio of the number of transvariating pairs (with respect to the marginal medians) to its maximum. As opposite to the univariate case, however, transvariation probability can no longer be interpreted as a measure of group separability as it may be greater than 0 even if the groups are completely separate in the multidimensional space. On the contrary, Gini's "transvariation space", that is the multidimensional transvariation area

$$\int_{-\infty}^{\infty} \Psi(x_1, x_2, ..., x_p)\, dx_1 dx_2 ... dx_p \tag{26}$$

where

$$\Psi(x_1, x_2, ..., x_p) = \min\left(f_1(x_1, x_2, ..., x_p), f_2(x_1, x_2, ..., x_p)\right) \tag{27}$$

with $f_1$ and $f_2$ the multivariate group densities, still maintains its meaning and its equivalence to the total probability of misclassification in the equal prior case.

Calò (2004) has introduced a modified version of Gini's multidimensional transvariation probability which can be interpreted as a true measure of group separation in the multivariate space and has suggested a stepwise variable selection method in discriminant analysis which is based on this new measure.

In the statistical literature on robust location measures Gini's spatial median is quite well known, even if he dedicated to it very little space. He just defined it and proposed an application to the study of population distribution.

Given *n* points lying in $R^p$, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, the spatial median is the *p*-vector $\mathbf{M}$ which minimises the Euclidean distance of the points from it $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{M}\|$. When *p=1* the definition yields the standard univariate median.

Small (1990) presents a thorough survey of multidimensional medians, also highlighting the properties of Gini's spatial one and relating it to other possible definitions of medians in the multivariate space.

A reinterpretation of Gini's contribution on this and many other issues in multivariate analysis can surely open new fields of research, towards more general and flexible solutions.

**References**


BACCINI, A. and DE FALGUEROLLES A.(1993). Analysing multivariate income data with a PCA based on Gini's mean difference, *Bull. of ISI*, 59-60.

BOWLEY, A.L. (1920). *Elements of statistics*, Kin, London.

BRIEMAN, L., FRIEDMAN J.H., OLSHEN, R.A. and STONE C.J. (1984). *Classification and regression trees,* Wadsworth International Group, Belmont, California.

BRIZZI M. (2002). Alcune considerazioni sugli indici di eterogeneità normalizzati e sugli indici di connessione da essi derivabili, working paper.

CALÒ D.G. (2004). On a transvariation based measure of group separability, submitted for publication in *Journal of Classification.*

COSTA M,.GALIMBERTI G.,MONTANARI A.. (2005). *Binary segmentation methods based on Gini index: a new approach to the measurement of poverty* submitted for publication

DAGUM C.(1987). Gini Ratio, *The New Palgrave: a Dictionary of Economics*, vol. II, 529-532.

DAGUM C.(1997). Decomposition and Interpretation of Gini and the Generalized Entropy Inequality Measures, *Proceedings of the American Statistical Association*, *Business and Economic Statistics Section*, 200-205.

FISHER R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample, *Metron*, **1**, 3-32.

FISHER R.A (1936). The use of multiple measurements in taxonomic problems, *The annals of eugenics*, **7**,179-188.

FULLER W.A. (1987) *Measurement error models*, Wiley, New York .

GINI C. (1912). Variabilità e mutabilità, Studi economico-giuridici pubblicati per cura della Facoltà di Giurisprudenza della Regia Università di Cagliari, Anno III, parte 2, also reproduced in C. Gini (1939) op. cit.

GINI C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri, Atti del R. Istituto Veneto di Scienze, Lettere e Arti, LXXIII, parte II, 1203-1248.

GINI C. (1916). Il concetto di transvariazione e le sue prime applicazioni, *Giornale degli economisti* and *Rivista di statistica.*

GINI C. (1921). Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali, *Metron*, **1**, 63-82.

GINI C. (1939). *Memorie di metodologia statistica* , Vol. I, Variabilità e concentrazione, Giuffrè, Milano.

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

GINI C. and GALVANI L. (1929). Di talune estensioni dei concetti di media ai caratteri qualitativi, *Metron*, **8,** 3-209**.**

GINI C. and LIVADA G. (1959). Transvariazione a più dimensioni in C.Gini (1959).

GINI C. and LIVADA G. (1959). Nuovi contributi alla teoria della transvariazione in C. Gini (1959).

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

HUBER P.J. (1985). Projection pursuit (with discussion), the Annals of statistics, 13, 435-475.

LIGHT R.J., and MARGOLIN B.H. (1971). An analysis of variance for categorical data, *J. Amer. Statisti. Assoc.*, 66, 534-544.

MONARI P. and MONTANARI A. (2003) Corrado Gini and multivariate statistical analysis: the (so far) missing link, in *Between Data Science and Applied Data Analysis* (Eds. M. Schader, W. Gaul e M. Vichi) Springer-Verlag Berlin, Heidelberg, 321-328.

MONTANARI A. (2004). Linear discriminant analysis and transvariation, *Journal of Classification*, 21, 71-88.

MONTANARI A. and CALÒ D. G. (1998). Two Group Linear Discrimination Based on Transvariation Measures, in *Advances in Data Science and Classification* (Eds. A. Rizzi, M. Vichi and H.H. Bock) Springer-Verlag Berlin, Heidelberg, 97-104.

OLKIN, I. and YITZHAKI S. (1992). Gini regression analysis, *International statistical review*, **60**, 185-196.

PEARSON K. (1901). On lines and planes of closest fit to a system of points in space, *Philosophical Magazine*, **2**, 559-572.

PODDER N. (2002). The theory of multivariate Gini regression and its applications, working paper.

POSSE C. (1992). Projection pursuit discriminant analysis for two groups, *Communications in statistics, Theory and methods*, **21**, 1-19.

SCHECHTMAN, E. and YITZAKI S. (1987). A measure of association based on Gini's mean difference, *Communication in Statistics, Theory and Methods*, **16**, 207-231.

SMALL C.G. (1990). A survey of multidimensional medians, *International statistical review*, **58**, 263-277.

STUART A. (1954) The correlation between variate values and ranks in sample from continuous distribution, *British Journal of Statistical Psychology*, **7**, 37-44.

TAGUCHI T. (1981). On a multiple Gini's coefficient and some concentrative regressions, *Metron*, **39**, 69-98.