Generalized Distance Measures for Asymmetric Multivariate Distributions

Marco Riani and Sergio Zani
Istituto di Statistica, Facoltà di Economia, Univ. of Parma, Italy

Abstract:[1] In this paper we suggest a non parametric generalization of the Mahalanobis distance which enables to take into account the differing spread of the data in the different directions. The output is an easy to handle metric which can be conveniently used both in an exploratory stage of the analysis for the detection of multivariate outliers and successively as a tool for non parametric discriminant analysis, multidimensional scaling and cluster analysis. In addition, the use of this metric can provide information about multivariate transformations and multiple outliers.

Keywords: Mahalanobis distance, equidistance contours, convex hull peeling, B-spline smoothing.

1 Introduction

The distance concept plays an important role in many topics of multivariate analysis. One of the most widely used distance measures is the one of Mahalanobis (d_M) (e.g. Dasgupta, 1993). This distance is appropriate for use in sample spaces where there exist differential variances and correlations between variables. In this metric, contours of equi-distance in the p-dimensional space are p-dimensional hyperellipsoids. Consequently, when we use d_M we implicitly assume that the spread of the data in the different directions is symmetric. Therefore, in presence of highly asymmetric data the use of this distance seems questionable. More generally, in a preliminary stage of the analysis it seems preferable to use a metric which does not assume an underlying distribution. Finally, even if the ellipticity hypothesis is satisfied, in order to construct d_M we still have to face the problem of estimating the means of the variables and the covariance matrix. The breakdown point of d_M is 0 and the presence of multiple outliers can cause masking and swamping problems (e.g. Barnett and Lewis, 1994).

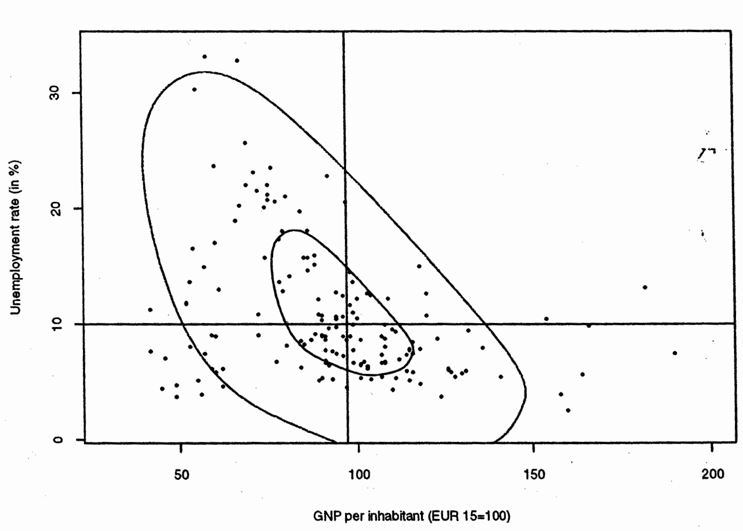
The purpose of this paper is to suggest a simple generalization of d_M which: (a) is non parametric, (b) is robust to the presence of atypical observations, (c) keeps into account the differing spread of the data in the different directions, (d) reduces to the usual d_M when the ellipticity hypothesis is satisfied and the data are not contaminated.

[1] The program for computing the non parametric distance suggested in this paper is available upon request. We can be contacted at `statdue@ipruniv.cce.unipr.it` or `zani@ipruniv.cce.unipr.it`. The authors wish to thank Aldo Corbellini for programming help.

2 Description of the method

First let us consider the case in which the number of variables (p) is equal to 2. Our approach starts defining a non parametric bivariate central region and a robust centroid. As pointed out by Riani *et al.* (1998), a natural and completely non parametric way for defining a central region in $\mathbb{R}^2$ is through the use of the so called convex hull peeling. Successive convex hulls are peeled until the first one is obtained which includes not more than 50% of the data (and so asymptotically half of the data). This 50%-hull is smoothed using a B-spline, constructed from cubic polynomial pieces, which uses the vertices of the 50%-hull to provide information about the location of the knots. (From now on this spline will be called 50%-spline). Zani *et al.* (1998) discuss several choices of a robust centroid. In this work we use the intersection of the two least squares lines built with the observations which lie inside or at the boundary of the 50%-spline. As an illustration of the suggested approach let us consider the data referred to 160 European regions reported in Figure 1. On the x-axis we have the Index Numbers of GNP per inhabitant, PPS (EUR 15=100). On the y-axis we have the unemployment rate in % (Source: Eurostat, REGIO, 1996). Figure 1 also reports the 50%-spline. As emerges clearly from the plot, the spread of the data in the differing directions is different, therefore the traditional approach based on d_M does not seem to be appropriate. Two straight lines in correspondence of the bivariate centroid have also been drawn. Zani *et al.*, (1998) show that in order to obtain an outer contour which under the assumption of bivariate normality leaves outside a percentage of observations close to 1%, we must multiply the distance from the

Figure 1: Bivariate boxplot of Unemployment rate in % versus Index Numbers of GNP per inhabitant for 160 European regions (50% and 99% contours).



50%-spline to th
the outermost c
the bivariate cor
outermost conto
from the center.
the differing spr
us consider sepa
distance betwee

Distance from t

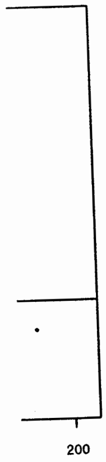
In our metric t
distance from th
generality we ca
lies on the 50%
on the former u

Figure 2: Exa
for 4 of the Eu

straight line
A' be the in
in this case i
at a distanc
of every poi
intersection
Figure 2 u
(E) and Un
point B wo
the ratio (
differing sp

es (p) is equal
central region
a natural and
$I\!R^2$ is through
ulls are peeled
)% of the data
oothed using a
the vertices of
knots. (From
(1998) discuss
ie intersection
hich lie inside
the suggested
ons reported in
oer inhabitant,

ndex Numbers
% contours).



ient rate in %
0%-spline. As
ring directions
es not seem to
ariate centroid
obtain an outer
aves outside a
tance from the

50%-spline to the robust center by 1.68. Using this coefficient we obtained the outermost contour reported in Figure 1. The units which lie outside the bivariate contour can be interpreted as atypical. The 50%-spline and the outermost contour can be interpreted as non parametric equidistance contours from the center. In this way we can take into account non parametrically the differing spread of the data. This leads us to define a new metric. Let us consider separately the distance of one point from the centroid and the distance between two generic points.

Distance from the centroid in $I\!R^2$

In our metric the observations which lie on the 50%-spline have the same distance from the centroid. In order to define a measure unit, without loss of generality we can set equal to 1 the distance from the center of a point which lies on the 50%-spline. The distance of every other observation can be based on the former unit of measure. For example, in Figure 2 let us consider the straight line $\overline{OA}$ passing through the robust centroid and point A. Let point A' be the intersection of segment $\overline{OA}$ with the 50%-spline. The distance $\overline{OA}$ in this case is 2.55 times the distance $\overline{OA'}$. Therefore in our metric point A lies at a distance 2.55 from the origin. In general: the distance from the centroid of every point K depends on the ratio between $\overline{OK}$ and $\overline{OH}$ where H is the intersection of the straight line passing through $\overline{OK}$ with the 50%-spline. In Figure 2 units A e B respectively correspond to the regions Extremadura (E) and Uusimaa (FIN). Note that using the standardized Euclidean distance point B would be much closer to the center than point A. Using our metric the ratio $\overline{OB}/\overline{OB'}$ is 2.50. This implies that if we take into account the differing spread of the data in the different directions these two regions have

Figure 2: Example of computation of the distance from the robust centroid for 4 of the European regions reported in Figure 1 (Standardized variables).

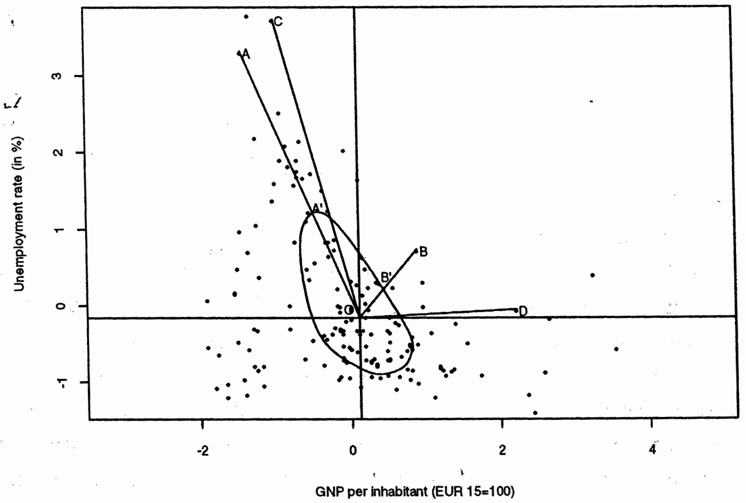
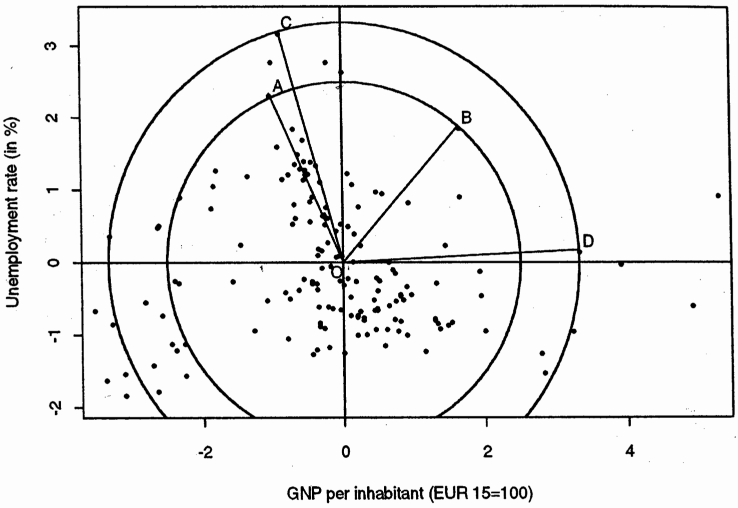


Figure 3: Plot of the points of Figure 2 in the transformed space using the suggested metric (The two superimposed circumferences define equidistant contours from the robust centroid).



approximately the same distance. Following similar arguments we can claim that point C (Ceuta y Melilla, E) has a distance from the centroid (3.32) which is exactly equal to that of point D (Bremen, D).

In Table 1 we compare for units A,B, C and D our non parametric distance (d_{RZ}) from the centroid with the Euclidean one (d) (using standardized variables) and the one of Mahalanobis (d_M). This table shows that for units A and C the values of d_{RZ} and d_M are smaller than those produced by the Euclidean distance. The opposite happens for units B and D. In fact d_{RZ} and d_M correctly keep into account the negative correlation ($r = -0.359$) between the two variables. The values of our metric compared to d_M are smaller for units A and C and greater for the two remaining units. Our metric is based on non parametric contours and enables to reduce (increase) the distance for the units located in the directions where the spread of the data is high (low). The assumption of d_M of symmetric spread usually is not met in practise. For example, in the data reported in Figure 1 it is clear that the spread of the data in the South-West direction is much more evident than that in the North-East direction. The ellipses (here not reported for lack of space) which represent the equidistance contours in the Mahalanobis metric, treat in the same way the units located South-West with those located North-East. This explains why for example, the value of d_{RZ} for unit D is much bigger than that reported by d_M.

Remark 1: If the underlying hypothesis of elliptic distribution is true the 50%-spline tends to become an ellipse (Atkinson and Riani, 1997). In addition, convex hull peeling is invariant under linear transformations of the data.

Table 1: Squa
metrics.

Therefore the
hypothesis is
Remark 2: In
ply need an
dispersion or
Remark 3: It
spline or 75%
adapt more
this results in
Figure 3 sho
parametrical
this new spa
on the robu
approximate
Distance bet
In our metri
into circles.
equal to 3.3
with radius
referring to
moved (non
directions.
two points i
Extensions
In presence
because we
purpose of
to the d_M
marginal b
variables.
the sum of
d_M plus a
for simplic

Table 1: Squared distances from the centroid of four regions using different metrics.

	d^2_{RZ}	d^2_M	d^2
A: Extremadura	6.528	11.070	13.041
B: Uusimaa	6.275	2.064	1.331
C: Ceuta y Melilla	11.050	14.139	14.961
D: Bremen	11.038	5.573	4.935

Therefore the suggested metric simply reduces to the d_M when the ellipticity hypothesis is satisfied and the data are not contaminated.

Remark 2: In order to calculate the former distances, in our metric we simply need an estimate of the centroid but we do not have to estimate either dispersion or correlation parameters.

Remark 3: It is possible to use a different spline contour (for example a 60%-spline or 75%-spline). With this choice the shape of the inner region may adapt more to the spread of the units lying far from the centroid. However this results in a decrease of robustness.

Figure 3 shows the 160 European regions of Figure 1 after eliminating non parametrically the differing spred of the data in the various directions. In this new space both unit C and D lie on the same circumference centered on the robust centroid whose radius is 3.32. Similarly, units A and B lie approximately on the circumference whose radius is 2.5.

Distance between two points in $I\!R^2$

In our metric the splines which define the equidistant contours are transformed into circles. This implies, for example, that the units which have a distance equal to 3.32 from the robust center after the transformation lie on a circle with radius 3.32. In the transformed space (Figure 3) Euclidean distances referring to different directions are directly comparable because we have removed (non parametrically) the different spread of the data in the various directions. Therefore our metric is equal to the Euclidean distance between two points in the transformed space.

Extensions to p-dimensional data

In presence of p-dimensional data the situation becomes more complicated because we cannot rely on the graphical representation anymore. With the purpose of defining a distance measure which, with well behaved data, reduces to the d_M we have to consider pairs of variables such that the sum of the marginal bivariate d_M is equal to the overall d_M using the whole set of p-variables. It is possible to show that the p-variate d_M can be expressed as the sum of $p/2$ bivariate d_M if p is even and as the sum of $[p/2]$ bivariate d_M plus a univariate d_M if p is odd. For example if $p = 3$, (considering for simplicity the variables in standardized form) using the matrix inversion lemma it is possible to prove that the squared Mahalanobis distance of the trivariate vector $\mathbf{x} = (x, y, z)'$ from $\mathbf{0}$ can be decomposed as:

$$\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} = \frac{x^2 + y^2 - 2\rho_{xy}xy}{1 - \rho_{xy}^2} + \frac{\left[z\sqrt{1 - \rho_{xy}^2} - x\rho_{xz.y}\sqrt{1 - \rho_{yz}^2} - y\rho_{yz.x}\sqrt{1 - \rho_{xz}^2}\right]^2}{|\mathbf{\Sigma}|} \tag{1}$$

where $\mathbf{\Sigma}$ is the correlation matrix and $\rho_{jl.k}$ denotes the partial correlation coefficient between variables j and l given variable k.

The first term on the right hand side of equation (1) is nothing but the squared marginal d_M between the first two variables. The second term can be shown to be the squared d_M from 0 of the univariate variable $z|(x, y)$:

$$z|(x, y) = z - \frac{[(\rho_{xz} - \rho_{xy}\rho_{yz})x + (\rho_{yz} - \rho_{xy}\rho_{xz})y]}{1 - \rho_{xy}^2} \tag{2}$$

In the case of a univariate random variable the central region is given by the interquartile range and the robust center is the median. Similarly to the bivariate case, we can set equal to 1 the distance between the first (third) quartile and the median and we can use this length for defining the other distances. With 3 variables, therefore, we initially have to compute our distance using the first two variables, then we have to consider the transformed variable $z|(x, y)$. If the ellipticity hypothesis is satisfied and the data are not contaminated the simple sum of these two distances reduces to the global d_M. With elliptic distributions it is immaterial the order in which the variables are considered. Finally, in presence of asymmetric data we end up with a metric which adapts non parametrically to the various spread of the data in the different directions.

References

Atkinson, A. C., and Riani, M. (1997), Bivariate Boxplots, Multiple Outliers, Multivariate Transformations and Discriminant Analysis: the 1997 Hunter Lecture, *Environmetrics*, 8, 583-602.

Barnett, V. and Lewis, T. (1994), *Outliers in Statistical Data*, Wiley, N. Y.

Dasgupta, S. (1993), "The Evolution of the D^2-Statistic of Mahalanobis", *Sankhyā*, special volume dedicated to the memory of P.C. Mahalanobis, 442-459.

Riani, M., Zani, S. and Corbellini, A. (1998), Robust Bivariate Boxplots and Visualization of Multivariate Data, in: I. Balderjahn, R. Mathar and M. Schader (Eds.), *Classification, Data Analysis and Data Highways*, Springer Verlag, Berlin, 93-100.

Zani, S., Riani, M. and Corbellini, A. (1998), Robust Bivariate Boxplots and Multiple Outlier Detection, forthcoming in *Computational Statistics and Data Analysis*.