Generalized Distance Measures for Asymmetric Multivariate Distributions

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

Abstract: 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.

The program for computing the non parametric distance suggested in this paper is available upon request. We can be contacted at statdu@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 \( z \)-axis we have the Index Numbers of GNP per inhabitant, 50%-spline to the outermost contour and the bivariate centroid from the center of the differing spread. We consider separating the distance between 50% and the outermost contour, from the center of the differing spread. We consider separating the distance between from the center of the differing spread.

Distance from the center of the differing spread.

We consider separating the distance between.

Distance from the center of the differing spread.

We consider separating the distance between.

Distance from the center of the differing spread.

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

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

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

Figure 2: Example for 4 of the European regions.

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 straight line \( A' \) be the in this case i at a distance of every point intersection. Figure 2: \( (E) \) and \( U \) point \( B \) with the ratio \( \gamma \) differing \( \sigma \).
As \((p)\) is equal central region a natural and \(R^2\) is through
ulls are peeled 10\% of the data weled using a the vertices of
knots. (From (1998) discuss the intersection which lie inside
the suggested ms reported in per inhabitant,

index Numbers % contours).

Distance from the centroid in \(R^2\)
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

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

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 \(R^2\)
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
(F) and Uusimaa (FIN). Note that using the standardized Euclidean distance
distance point \(B\) would be much closer to the center than point \(A\). Using our metric
the ratio \(\overline{OB}/\overline{OF}\) 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 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 practice. 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.

**Remark 2:** In our metric this new space approximates the hypothesis is true 50%-spline tends to become an ellipse (Atkinson and Riani, 1997). In addition, convex hull peeling is invariant under linear transformations of the data.

**Remark 3:** It simply need an dispersion or splines or 75% adapt more this results in Figure 3 show parametrical this new space approximate this new space.

| Table 1: Squared Distances |

In our metric this new space approximates the hypothesis is true 50%-spline tends to become an ellipse (Atkinson and Riani, 1997). In addition, convex hull peeling is invariant under linear transformations of the data.

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

**Extensions**

In presence because we purpose of this results in Figure 3 show parametrical this new space equal to 3.3 with radius referring to moved (non directions). The sum of $d_M$ plus a for simplic |
ce using the equidistant.

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.

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

<table>
<thead>
<tr>
<th>Region</th>
<th>$d_{RZ}^2$</th>
<th>$d_{M}^2$</th>
<th>$d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Extremadura</td>
<td>6.528</td>
<td>11.070</td>
<td>13.041</td>
</tr>
<tr>
<td>B: Uusimaa</td>
<td>6.275</td>
<td>2.064</td>
<td>1.331</td>
</tr>
<tr>
<td>C: Ceuta y Melilla</td>
<td>11.056</td>
<td>14.139</td>
<td>14.961</td>
</tr>
<tr>
<td>D: Bremen</td>
<td>11.338</td>
<td>5.573</td>
<td>4.935</td>
</tr>
</tbody>
</table>

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

Distance between two points in $\mathbb{R}^2$

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. Similarly, units A and B lie approximately on the circumference whose radius is 2.5.

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 $\lfloor p/2 \rfloor$ 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 \( x = (x, y, z)' \) from 0 can be decomposed as:

\[
x'\Sigma^{-1}x = \frac{x^2 + y^2 - 2p_{xy}xy}{1 - p_{xy}^2} + \left( z\sqrt{1 - p_{xz}^2} - zp_{xz,y}\sqrt{1 - p_{yz}^2} - yp_{yz,z}\sqrt{1 - p_{xz}^2} \right)^2
\]

(1)

where \( \Sigma \) is the correlation matrix and \( p_{j,l,k} \) 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 \( z(x, y) \):

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

(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 different directions.

References


Abstract:

regression
scaling mo
Greatest M
their inclus
between th
(ALSCAL)

Keywords
Disparities

1. Disp:
Conv

Given a \( \delta \)
The order
"(i,j) is le

\[ \delta(i, j) \leq \]

Given the nonmetric \( X_1, X_2, \ldots \) distances
much as
if the dis
of dissim

if \( \delta(i, j) \leq \)