UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2015, Vol. 8, No. 1, pp. 33–44

Graphical method using neighborhoods for detecting outliers[1]

Método gráfico usando vecindades para detectar outliers

Juan Carlos Correa Morales[a]
jccorrea@unal.edu.co

Víctor Ignacio López-Ríos[b]
vilopez@unal.edu.co

Abstract

We propose a new graphical method to help us to uncover potential outliers in multivariate samples. The idea behind the method is to analyze the behavior of a growing neighborhood of each data point. This method is very robust and allows to find outliers in very complex structures.

***Keywords*:** Mahalanobi distance method, multivariate outliers, neighbors method.

Resumen

Se propone un nuevo método gráfico que ayuda a descubrir datos atípicos en muestras multivariables. La idea detrás del método es analizar el comportamiento de una vecindad creciente alrededor de cada observación en la muestra de datos. Este método es muy robusto y permite encontrar datos atípicos en estructuras muy complejas.

***Palabras clave*:** datos atípicos multivariables, método de la distancia de Mahalanobis, método de vecinos.

1 Introduction

The detection of outliers is a common task in applied statistics. Usually this is done variable by variable, but the detection of multivariate outliers is more

[1] Correa, J. C., López, V. I. (2015). Graphical method using neighborhoods for detecting outliers. *Comunicaciones en Estadística*, **8**(1), 33-44.

[a] Profesor Asociado, Escuela de Estadística,Grupo de Investigación en Estadística, Universidad Nacional de Colombia, Sede Medellín. Colombia.

[b] Profesor Asociado, Escuela de Estadística, Grupo de Investigación en Estadística, Universidad Nacional de Colombia, Sede Medellín. Colombia.

complex than in the univariate case (Campbell 1978). The difficulty of finding multivariate outliers have been addressed by several authors (Cléroux et al. 1986). Even the definition of multivariate outlier is unclear (Barnett & Lewis 1994) and Finney (2006) discussed the problems associated with this concept.

Several methods have been proposed to detect potential multivariate outliers, specially in samples that come from spherical distributions. Wilks (1963) proposed the first approximation to detect multivariate outliers. Caroni & Prescott (1992) proposed a sequential test based on Wilk's test for detection of multivariate outliers. Wang et al. (1997) proposed a test based on a modified likelihood ratio test. Peña & Prieto (2001) proposed a method for detecting multivariate outliers based on projections that optimizes a kurtosis coefficient. Different authors have proposed the use of the influence function to analyze the effect of the presence of outliers in the data (Cléroux et al. 1986, Gillespie 1993, Boente et al. 2002).

Several graphic proposals have been made for multivariate data (Everitt & Nicholls 1975), but they are not very useful to detect multivariate outliers. Rohlf (1975) generalized the gap test for detecting multivariate outliers using whether a Q-Q plot or a formal test. Other authors proposed a Q-Q plot based on Wilks's statistic (Bacon-Shone & Fung 1987). Khattree & Naik (1995) used a Q-Q plot based on a modification of the Mahalanobis distance. Hadi (1992) and Hardin & Rocke (2005) used a robustified version of the Mahalanobis distance for detecting multivariate outliers. Muruzábal & noz (1997) proposed a graphical technique called Self-Organizing Maps as a tool for detecting multivariate plots. It was based on neural concepts to detect multivariate outliers but the interpretation is no very clear. Pison & Van (2004) used plots obtained out of robustified multivariate techniques such as PCA. Hubert & Rousseeuw (2005) presented a robust algorithm for doing PCA. This alternative can be used to detect multivariate outliers.

Traditional descriptive univariate methods such as the boxplot do not detect outliers but point out candidate observations as potential outliers by observing the tails of the plot. It is a task of the analyst to determine if these points can be removed from the analysis because he/she considers them as outliers. We consider a multivariate outlier as an observation that comes from a different distribution that the one that generates the data. We propose a new graphical method that can be used for detecting potential outliers in multivariate samples coming from distributions with complex structures. This method is simple and easy to implement.

2 The method

Let us asume that $x_1, x_2, \ldots, x_n$ is a sample drawn from a multivariate population. We compute the matrix with the interpoint distances. We can use, for example, the Euclidean distance:

$$D(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2} \tag{1}$$

For each data point, x_i, $i = 1, 2, \ldots, n$, we compute a nearest neighbor function that can be expressed as:

$$N_{x_i}(d) = \frac{\sum_{j=1}^{n} I_{V(x_i,d)}(x_j)}{n} \tag{2}$$

where I_A is an indicator function of A and $V(x_i, d)$ denotes a neighborhood of ratio d with center x_i, so:

$$V(x_i, d) = \{x : D(x, x_i) \leq d\},$$

d varies in the interval $(0, \infty)$. Intuitively we can think of an outlier as a data point that does not have close neighbors. But how close is close? We propose to observe a growing neighborhood around each data point. If this neighborhood is large enough and still empty this is an indication that the point is a candidate to be consider as an outlier. In this spirit we plot N_{x_i} vs. d. All curves are plotted on the same graphic. We can identify outliers by looking those functions that behave different from the main body of curves.

This method will be illustrated in the following section.

We may use other distance functions. One possible choice is the squared Mahalanobis distance:

$$D(x_i, x_j; \Sigma, \mu) = (x_i - \mu)^T \Sigma^{-1} (x_j - \mu) \tag{3}$$

where Σ and μ are replaced by unbiased estimators. It is possible to use robust estimators of them.
We compare our graphical proposal method with Mahalanobis distance method, which consists in identifying potential outlier if the Mahalanobis distance is greater than a quantile of the chi-square distribution, so x_i $i = 1, 2, \ldots, N$ is a potential outlier if

$$D(x_i; \Sigma, \mu) := D(x_i, x_i; \Sigma, \mu) > \chi^2_{p,1-\alpha/2},$$

when $\alpha = 0.05$, the cut-off value for the robust Mahalanobis distance the value $\chi^2_{p,0.975}$ is suggested, which is the 97.5% quantile of the chi-square distribution with p degrees of freedom (Rousseeuw & Van Driessen 1999).

Seber (1984) points out that the Mahalanobis distance could reduce the clarity of the clusters and one outlier could be thought as a cluster of a single point. This is illustrated below in those cases in which we have distributions with a hole and the outlier is placed in the center of the hole. This situation could appear when we are dealing with mixture distributions.

3 Examples

3.1 Artificial data

To illustrate this method let us assume we have a sample from an elliptical distribution with a hole in the middle. First, we generate a bivariate sample of mixture of normal variates. Figure 1 a. shows the sample and one artificial outlier is placed in the center. Figure 1 b. shows the N nearest-neighbor functions (Proportion of neighbors of a point in a given distance vs. distance). It is easy to identify the curves associated to the potential outliers. In this case it is possible to see three curves that correspond to the two outliers detected by Mahalanobis distance, observations labeled as 33 and 116 (*Mah. Outliers*) and the artificial outlier (*Art. Outlier*). We can see that the graphical proposal does not produce the false signals with those points that the Mahalanobis distance chooses as a potential outliers and we know that they are not outliers. However the artificial outlier is detected by the proposal method and it is masked by Mahalanobis distance method.

The Figure 2 a., shows a sample from a star-shape distribution which was defined as a mixture of two bivariate normal distributions,

$$(x_1, x_2)^T \sim 0.5N_2(0, 0, \rho_1 = -0.99) + 0.5N_2(0, 0, \rho_2 = 0.3),$$

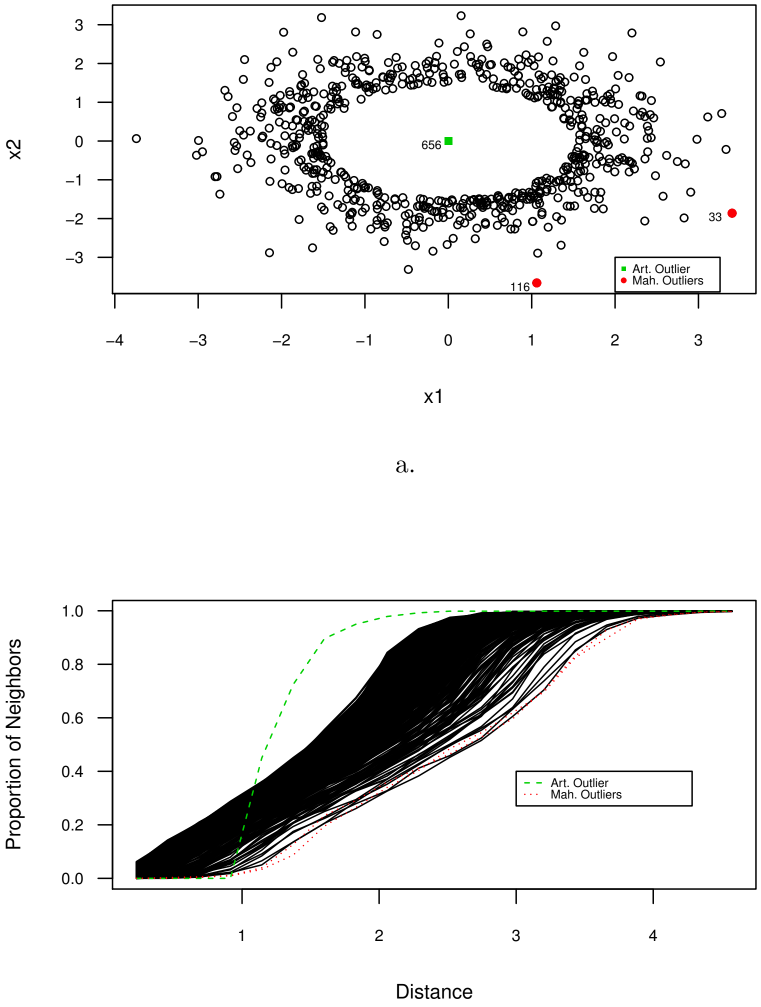
with a hole at the center. Then we put one artificial outlier in the middle. The Mahalanobis distance detected seven potential outliers, observations labeled by 44, 51, 61, 115, 132, 170 y 224. Figure 2 b., shows the plot of the N functions. It is possible to identify the curve associated to artificial outlier. There are three curves separated from the main body of curves, one of them corresponds to artificial outlier and the others two were detected by Mahalanobis as potential outliers. We observe that only two out of seven potential outliers detected by Mahalanobis distance were indentified by our proposal method. These outliers identified by the Mahalanobis distance method are all not real outliers, but the artificial outlier was not detected by Mahalanobis distance method.

The Figure 3 a., shows a sample from a star-shape distribution which was defined as a mixture of four bivariate normal distributions,

$$\begin{aligned}(x_1, x_2)^T \sim \quad & 0.25N_2(0, 0, \rho_1 = -0.99) + 0.25N_2(0, 0, \rho_2 = 0.5) \\ & +0.25N_2(0, 0, \rho_2 = -0.5) + 0.25N_2(0, 0, \rho_2 = 0.99),\end{aligned}$$

with a hole at the center. Then we put three artificial outliers in the middle. The Mahalanobis distance detected only one potential outlier, that does not correspond to any of the three real outliers. The proposed method uncovered the three real outliers (see Figure 3 b.)

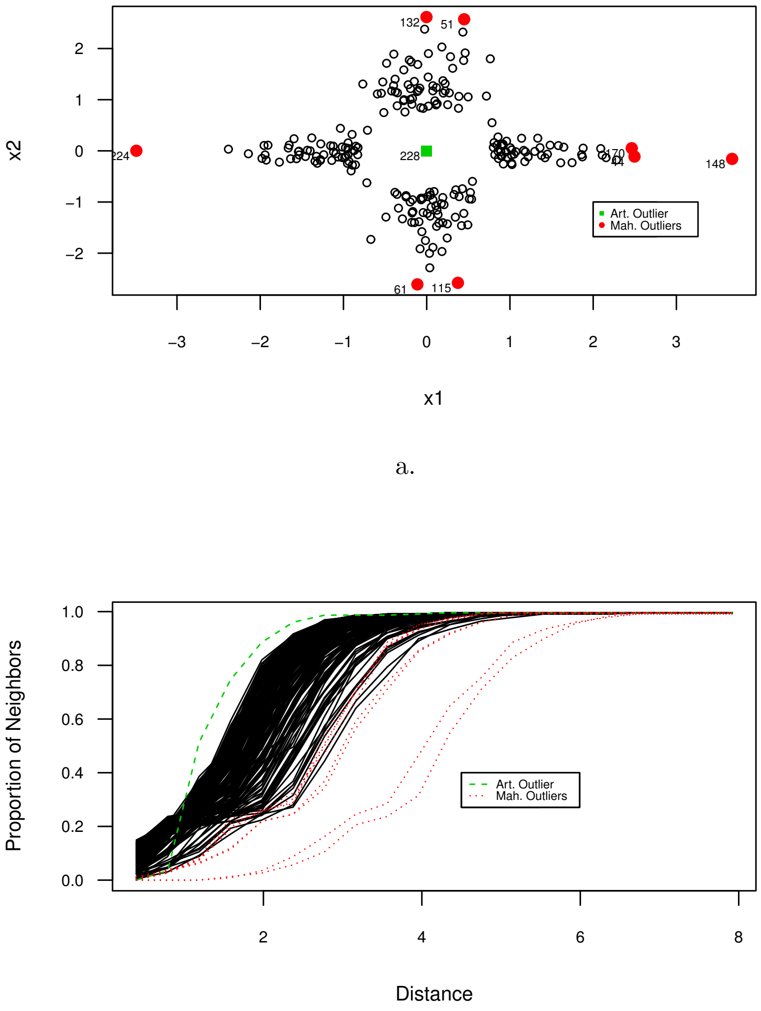
We also generate samples from multivariate normals in several dimensions and we observe similar behaviour of the proposal method. It always uncover the artificial outliers but Mahalanobis distance method fails to detect them.



a.

b.

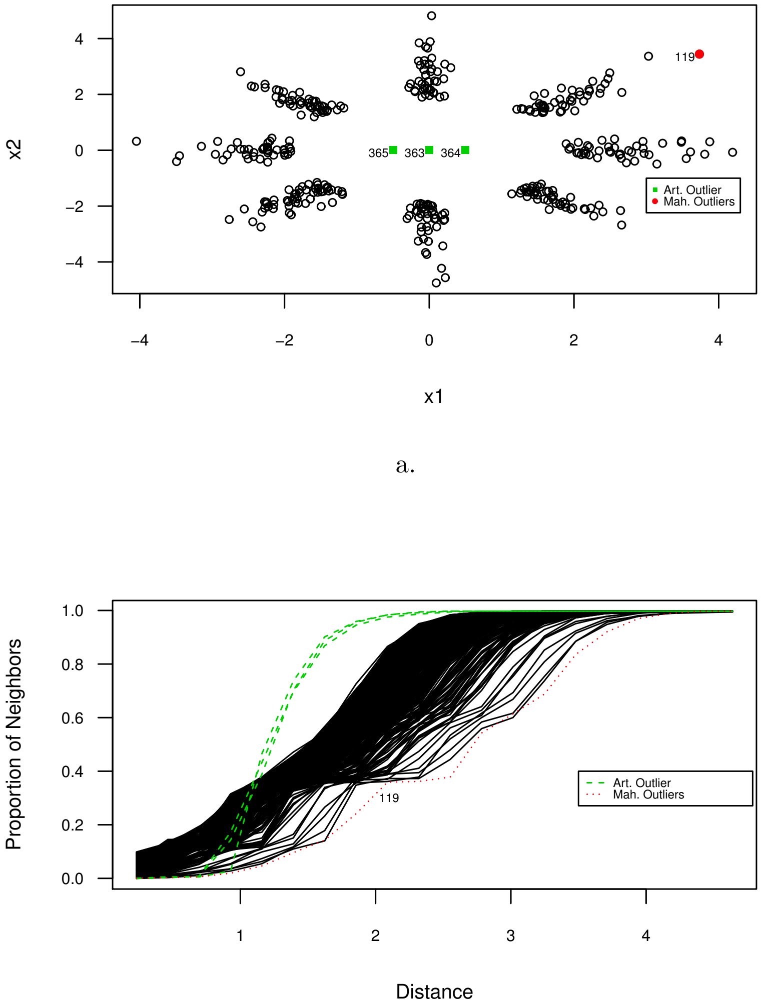
Figure 1: *a. Sample from bivariate normal with a hole and one artificial outlier b. Proportions of neighbors of a point from star-shape distribution with a hole and three outliers vs. distance. Source: own elaboration.*



a.

b.

Figure 2: *a. Mixture of bivariate normals:* $0.5N_2(0,0,\rho_1 = -0.99) + 0.5N_2(0,0,\rho_2 = 0.3)$ *with a hole. b. Proportions of neighbors of a point from a mixture of normal variates:* $0.5N_2(0,0,\rho_1 = -0.99) + 0.5N_2(0,0,\rho_2 = 0.3)$ *vs. distance. Source: own elaboration.*



a.

b.

Figure 3: *a. Sample from star-shape distribution with a hole and three outliers b. Proportions of neighbors of a point from star-shape distribution with a hole and three outliers vs. distance. Source: own elaboration.*

3.2 Data from the olympics

Dawkins (1989) presents the National records at various track races from 100 meters to the marathon for men and women. We use the women data presented by him. It contains information about 55 countries and seven track races. Figure 4 shows 55 nearest-neighbor functions (Proportion of neighbors of a point in a given distance vs. distance). We can see clearly that there are two outliers, Wsamoa and Mauritania.

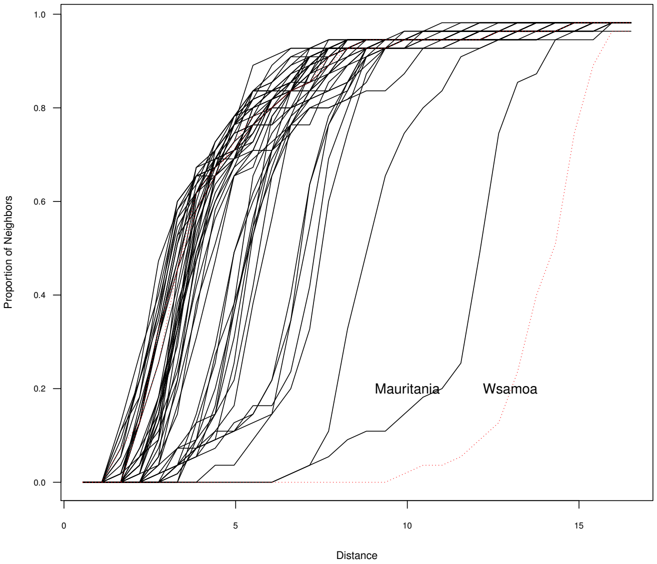


Figure 4: *Proportions of neighbors vs. distance. Source: own elaboration.*

4 Conclusions

We have proposed a graphical method that allows to visualize the neighborhood of a multivariate point. This permits us to identify potential outlier points if we see that the neighborhood of a point is almost empty within a reasonable distance. This method can be used to uncover clusters of data points too. This proposal can be easily implemented in the standard statistical software and it is useful to use with other methods as Mahalanobis distance.

Recibido: 6 de noviembre de 2014
Aceptado: 5 de marzo de 2015

References

Bacon-Shone, J. & Fung, W. K. (1987), 'A new graphical method for detecting single and multiple outliers in univariate and multivariate data', *Applied Statistics* **36**(2), 153–162.

Barnett, V. & Lewis, T. (1994), *Outliers in Statistical Data*, 3era edn, John Wiley & Sons Ltd.: Chichester.

Boente, G., Pires, A., Rodrigues, I. M. & Campbell, N. A. (2002), 'Influence functions and outlier detection under the common principal components model: A robust approach', *Biometrika* **89**(4), 861–875.

Campbell, N. A. (1978), 'The influence function as an aid in outlier detection in discriminant analysis', *Applied Statistics* **27**(3), 251–258.

Caroni, C. & Prescott, P. (1992), 'Sequential application of wilks's multivariate outlier test', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **41**(2), 355–364.

Cléroux, R., Helbling, J. M. & Ranger, N. (1986), 'Some methods of detecting multivariate outliers', *Computational Statistics Quaterly* **3**, 177–195.

Dawkins, B. (1989), 'Mutivariate analysis of national track records', *The American Statistician* **43**(2), 110–115.

Everitt, B. S. & Nicholls, P. (1975), 'Visual techniques for representing multivariate data', *Journal of the Royal Statistical Society. Series D (The Statistician)* **24**(1), 37–49.

Finney, D. J. (2006), 'Calibration guidelines challenge outlier practices', *The American Statistician* **60**(4), 309–314.

Gillespie, E. S. (1993), 'An application of multivariate outlier detection in assessing family characteristics for bank advertisements', *Journal of the Royal Statistical Society. Series D* **42**(3), 231–235.

Hadi, A. S. (1992), 'Identifying multiple outliers in multivariate data', *Journal of the Royal Statistical Society. Series B (Methodological)* **54**(3), 761–771.

Hardin, J. & Rocke, D. M. (2005), 'The distribution of robust distances', *Journal of Computational and Graphical Statistics* **14**(4), 928–946.

Hubert, M. & Rousseeuw, P. J. (2005), 'Robpca: A new approach to robust principal component analysis', *Technometrics* **47**(1), 64–79.

Khattree, R. & Naik, D. N. (1995), *Applied Multivariate Statistics with SAS Software*, SAS Institute Inc: Cary NC.

Muruzábal & noz, M. (1997), 'On the visualization of outliers via self-organizing maps', *Journal of Computational and Graphical Statistics* **6**(4), 355–382.

Peña, D. & Prieto, F. J. (2001), ‘Multivariate outlier detection and robust covariance matrix estimation’, *Technometrics* **43**(3), 286–300.

Pison, G. & Van, S. (2004), ‘Diagnostic plots for robust multivariate methods’, *Journal of Computational and Graphical Statistics* **13**(2), 310–329.

Rohlf, F. J. (1975), ‘Generalization of the gap test for the detection of multivariate outliers’, *Biometrics* **31**, 93–101.

Rousseeuw, P. J. & Van Driessen, K. (1999), ‘A fast algorithm for the minimum covariance determinant estimator’, *Technometrics* **41**(3), 212–223.

Seber, G. A. F. (1984), *Multivariate Observations*, John Wiley & Sons Inc.: New York.

Wang, S., Woodward, W. A., Gray, H. L., Wiechecki, S. & Sain, S. R. (1997), ‘A new test for outlier detection from a multivariate mixture distribution’, *Journal of Computational and Graphical Statistics* **6**(3), 285–299.

Wilks, S. S. (1963), ‘Multivariate statistical outliers’, *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)* **25**(4), 407–426.

A Computational codes

```
library(robust)
proporcion.vecinos<-function(Matriz.distancias, d)
    {proporcion <- (apply(ifelse(Matriz.distancias < d, 1, 0), 1, "sum")
     - 1)/nrow(Matriz.distancias)
     proporcion
     }
matriz.distancias<-function(X)
     {n <- nrow(X)
       covr<-covRob(X)$cov
       invS<-solve(covr)
  distancia <- matrix(rep(0, n * n), ncol = n)
  for(i in 1:(n - 1))
          {for(j in 2:n)
            {x1<-matrix(X[i,],ncol=1)
             x2<-matrix(X[j,],ncol=1)
       distancia[i, j] <- sqrt(t(x1-x2)%*%invS%*%(x1-x2))
  distancia[j, i] <- distancia[i, j]
  }
      }
      distancia
       }
grafique.outlier<-function(X,DD,outM,outA)
      {distancias <- matriz.distancias(X)
```

```
   dist.maxima <- max(distancias)
   radios <- (dist.maxima * (1:DD))/DD
    temp <- matrix(rep(0, nrow(distancias)), nrow = 1)
    for(i in 1:DD)
         {d       <- radios[i]
     temp  <- rbind(temp, proporcion.vecinos(distancias, d))
      }
       temp      <- temp[2:nrow(temp),  ]
            numcol    <- ncol(temp)
            r         <- outM+outA
            tipo.linea<-c(rep(1,numcol-r),rep(2,outA),rep(3,outM))
          matplot(radios,temp,type='l',col=c(rep('black',numcol-r),
          rep(3,outA),rep(2,outM)),
        lty=tipo.linea, las=1,xlab= "Distance",
         ylab= "Proportion of Neighbors",
        cex.lab=0.6,cex.axis=0.5 )
            }
#Example 1
datos<-matrix(rnorm(100),ncol=5)
X<-datos
library(MASS)
met.mahalanobis<-function(X)
  {
  cov.datos<-cov.mcd(X)$cov
  tmp<-X %*%solve(cov.datos)
  dist.mahalanobis<-X%*%t(tmp)
  d.mahalanobis<-diag(dist.mahalanobis)
  d.mahalanobis
  }
example1<-read.table(file.choose(),header=F)[,2:3]
plot(example1, las=1,xlab= "x1",ylab= "x2", cex.lab=0.6,cex.axis=0.5)
mah<-met.mahalanobis(as.matrix(example1))
cuantil<-qchisq(0.975,2)
plot(1:nrow(example1),mah,type='n',xlab='Observation number',
     ylab='MCD Mahalanobis distances')
points(1:nrow(example1),mah,pch='x')
indices<-which(mah>cuantil)
text(indices-6,mah[indices],indices,cex=0.5,col='red')
outA<-1
outM<-length(indices)
r<-outM+outA
nr<-nrow(example1)
example11<-example1[-indices,]
example11<-rbind(example11,example1[indices,])
plot(example1, las=1,xlab= "x1",ylab= "x2", cex.lab=0.6,cex.axis=0.5,cex=0.7)
points(example1[indices,],col=2:(li+1),pch=19,cex=0.7)
```

```
text(example1[indices,1]-0.2,example1[indices,2]-.1,indices, cex=0.4,col=’black’)
text(example1[(nr-outA+1):nr,1]-0.2,example1[(nr-outA+1):nr,2]-.1,
    (nr-outA+1):nr, cex=0.4,col=’black’)
points(example11[(nr-r+1):nr,],col=c(3:(outA+3-1),rep(2,outM)),
     pch=c(rep(15,outA),rep(19,outM)))
legend(2,-1,pch=c(15,19),col=c(3,2), legend=c("Art. Outlier", "Mah. Outliers"),
     cex=0.4,text.width=1)
grafique.outlier(as.matrix(example11),20,outM,outA)
legend(4.5,0.4,lty=c(2, 3), legend=c("Art. Outlier","Mah. Outliers"),col=c(3,2),
     cex=0.4,text.width=1)
```