A Robust Alternatives to the Covariance

Ortega, J. Fco.

[1] Facultad de CC. Económicas y Empresariales de Albacete, Plaza de la Universidad, 1. 02071. Albacete (Spain). JuanFco.Ortega@uclm.es

Keywords: Outliers. Trimming. Robustness.

1 Introduction

The covariance and its non-dimensional associated measure, the correlation coefficient, are built by using the sample means and its estructure. Thus, considering $z = \{z_1, z_2, \ldots, z_n\}$ a bidimensional sample, where $z_i = (x_i, y_i)$ with $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$, then the covariance and the correlation coefficient of x and y are $Cov(x, y)$ and $r(x, y)$ where:

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \qquad r(x,y) = \frac{Cov(x,y)}{S_x S_y} \tag{1}$$

with $\overline{x}$, $\overline{y}$, S_x and S_y the sample means and standard deviations of x and y. We know that these measures are very sensitive to the presence of outliers.

There are location and scale estimators which are more resistant in the presence of outliers, for example the ones defined by means of the trimming. So, given $x = \{x_1, x_2, \ldots, x_n\}$, the location estimators and scale estimators family (Ortega, 2004) by means of trimming are the $\alpha\_Trimmed$ and the $+\sqrt{\alpha\beta\_Trimmed}$, which are defined in:

$$\alpha\_Trim(x) = \frac{1}{n-2a}\sum_{i=a+1}^{n-a} x_{[i]} \tag{2}$$

and using:

$$\alpha\beta\_Trim(x) = C(\alpha,\beta)\ \ \beta\_Trim(\{(x_i - \alpha\_Trim(x))^2\}_i) \tag{3}$$

where: $\alpha, \beta \in [0, 50)$, with $a = Int(\alpha n/100)$ and $b = Int(\beta n/100)$; $x_{[i]}$ is the observation in the ith position of a sequence ranging from the smallest to the largest of the elements of x; $50\_Trim(x)$ is the median; and $C(\alpha, \beta)$ is a consistency coefficient.

In this paper, we proposed a alternatives measures like the ones defined in (1), using the elements (2) and (3). From those measures, we will be demonstrate their main analytic properties and their good behavior in the presence of outliers.

2 Main contributions and results

Given $z = \{z_1, z_2, \ldots, z_n\}$, like in Section 1, we define the concepts:

1. $\alpha\beta\_CoTrim(x,y) = C'(\alpha,\beta)\ \ \beta\_Trim(\{(x_i - \alpha\_Trim(x))(y_i - \alpha\_Trim(y))\}_i)$
2. $r_{\alpha\beta}(x,y) = \frac{\alpha\beta\_CoTrim(x,y)}{+\sqrt{\alpha\beta\_Trim(x)\ \ \alpha\beta\_Trim(y)}}$

where $\alpha, \beta \in [0, 50]$, $C'(\alpha, \beta)$ is a consistency coefficient and for $r_{\alpha\beta}$ is true that $\alpha\beta\_Trim(x)$ and that $\alpha\beta\_Trim(y)$ are different from 0.

For these definitions, it is true that:

1. $\alpha\beta\_CoTrim(x,y) = \alpha\beta\_CoTrim(y,x)$.
2. $\alpha\beta\_CoTrim(x,x) = \alpha\beta\_Trim(x) \geq 0$.
3. $\alpha\beta\_CoTrim(a+bx, a'+b'y) = b\ b'\ \alpha\beta\_CoTrim(x,y)$.
4. Under Normality (using simulations), the $\alpha\beta\_CoTrimmed$ are *consistent* in mean square for the covariance, and with *efficiency* between 37% (for $\beta = 50$) and 100%.
5. $r_{\alpha\beta}(x,y) = r_{\alpha\beta}(y,x)$.
6. If $b, b' \neq 0$ then $\mid r_{\alpha\beta}(a+bx, a'+b'y) \mid = \mid r_{\alpha\beta}(x,y) \mid$.
7. If there is an exact lineal relation between x e y then $\mid r_{\alpha\beta}(x,y) \mid = 1$.
8. If X and Y are random variables with symmetrical distributions, then:

$$\text{If } X \text{ and } Y \text{ are independent then } r_{\alpha\beta}(X,Y) = 0 = \alpha\beta\_CoTrim(X,Y)$$

On the other hand, $r_{\alpha\beta}(x,y)$ gives information about the lineal fit between x and y, like $r(x,y)$ but more resistant in the presence of outliers. So, the behavior of these measures in robustness is studied by means of the transformations of the *breakdown point* property in Donoho and Huber (1983). We define this property for a correlation coefficient R by $z = \{z_1, z_2, \ldots, z_n\}$, termed $\epsilon_n^*(R,z)$, as follows:

$$\epsilon_n^*(R,z) = Min_m\{\frac{m}{n}\ /\ Inf_{zc_m}\{\mid R(zc_m) \mid\} = 0\}$$

where zc_m is a contamination of z in m observations. And, the *asymptotic breakdown point*, termed $\epsilon^*(R)$, as follows :

$$\epsilon^*(R) = \lim_{n\to\infty} \epsilon_n^*(R,z).$$

Given these definitions, it is true that:

$$\epsilon_n^*(r_{\alpha\beta}, z) = Min\{\frac{a+1}{n}, \frac{b+1}{n}\} \qquad \epsilon^*(r_{\alpha\beta}) = Min\{\frac{\alpha}{100}, \frac{\beta}{100}\}$$

where $\alpha, \beta \in [0, 50]$, with $a = Int(\alpha n/100)$ and $b = Int(\beta n/100)$.

References

Y. Dodge & J. Jurecková (2000). Adaptive Regression. *Ed. Springer-Verlag N.Y.*

D.L. Donoho & P.J. Huber (1983). The notion of breakdown point. *A Festschrift for Erich Lehmann.* Eds. Bickel, P., Doksum, K. y Hodges, J.L.Jr.

D.C. Hoaglin *et al.* (2000). Understanding Robust and Exploratory Data Analysis. *Ed. Wiley and Sons.*

J.Fco. Ortega (2004). A family of scale estimators by means of trimming. *Theory and Applications of Recent Robust Methods*, pp. 259-269. Birkhäuser Verlag Basel, Switzerland.

P.J. Rousseeuw & C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, Vol.88. No.424.

P.J. Rousseeuw & A.M. Leroy (1987). Robust regression and outliers detection. *Ed. Wiley and Sons.*