A Robust Minimax Variance Estimator of a Correlation Coefficient

Georgiy Shevlyakov
State Technical University, Department of Mathematics
Polytechnichskaya, 29, St.Petersburg, 195251, Russia
shev@stat.hop.stu.neva.ru

Consider robust minimax estimation of the correlation coefficient ρ of the following class of bivariate distribution densities (the parameters of location and scale are assumed known)

$$f(x,y|\rho) = \frac{1}{\beta_u(\rho)}\, g\left(\frac{u}{\beta_u(\rho)}\right) \frac{1}{\beta_v(\rho)}\, g\left(\frac{v}{\beta_v(\rho)}\right),\ u = \frac{x}{\sqrt{2}} + \frac{y}{\sqrt{2}},\ v = \frac{x}{\sqrt{2}} - \frac{y}{\sqrt{2}}, \tag{1}$$

where $g(\cdot)$ is an univariate symmetric density belonging to a certain class $\mathcal{G}$.

For instance, if $\beta_u(\rho) = \sqrt{1+\rho}$, $\beta_v(\rho) = \sqrt{1-\rho}$ and $g(z)$ is the standard normal density, then formula (1) yields the standard bivariate normal density: $f(x,y|\rho) = N(x,y|0,0,1,1,\rho)$. Using other forms of univariate distribution densities, say the Laplace or Cauchy, one can construct the bivariate analogs of the corresponding univariate distributions.

The principal idea of introducing class (1) means that for any random pair (X,Y) the transformation $U = X+Y,\ V = X-Y$ gives uncorrelated (U,V), and the estimation of their scale solves the problem of estimation of correlation between (X,Y).

The correlation coefficient ρ of class (1) depends on the scale parameters β_u and β_v as follows: $\rho = (\beta_u^2 - \beta_v^2)/(\beta_u^2 + \beta_v^2)$.

Given $(x_i, y_i)_1^n$, consider the following estimation procedure:

• evaluate the M-estimates of scale $\widehat{\beta}_u$ and $\widehat{\beta}_v$ as the solutions of the following equations

$$\sum \chi(u_i/\widehat{\beta}_u) = 0, \qquad \sum \chi(u_i/\widehat{\beta}_v) = 0, \tag{2}$$

where $\chi(\cdot)$ is a score function and

$$u_i = x_i/\sqrt{2} + y_i/\sqrt{2}, \quad v_i = x_i/\sqrt{2} - y_i/\sqrt{2}, \qquad i = 1,\ldots,n;$$

• evaluate the estimate of ρ of the form

$$\widehat{\rho} = (\widehat{\beta}_u^2 - \widehat{\beta}_v^2)/(\widehat{\beta}_u^2 + \widehat{\beta}_v^2).$$

Thus for class (1), the problem of robust estimation of a correlation coefficient is reduced to the problem of robust estimation of scale of the principal variables u and v.

Theorem *Under general conditions of regularity put on distribution densities and score functions [1] providing consistency and asymptotical normality of the M-estimators of scale (2), the estimator $\widehat{\rho}$ is consistent and asymptotically normal with the following variance*

$$Var\,(\widehat{\rho}) = \frac{2(1-\rho^2)^2}{n}\, V(\chi, g), \quad V(\chi, g) = \frac{\int \chi^2(z) g(z)\, dz}{\left(\int z\chi'(z) g(z)\, dz\right)^2}. \tag{3}$$

The second formula in (3) is the asymptotic variance of M-estimator of scale [1], therefore we obtain a wide spectrum of robust minimax estimators of a correlation coefficient, since there exist robust minimax estimators of scale for various classes of distributions $\mathcal{G}$.

Corollary *In the class of ε-contaminated normal densities*

$$f(x,y) \geq (1-\varepsilon)\ N(x,y|0,0,1,1,\rho), \quad 0 \leq \varepsilon < 1,$$

the minimax robust estimator of ρ is given by

$$\widehat{\rho} = \left(\sum_{r_1+1}^{n-r_2} u_{(i)}^2 - \sum_{r_1+1}^{n-r_2} v_{(i)}^2 \right) \Bigg/ \left(\sum_{r_1+1}^{n-r_2} u_{(i)}^2 + \sum_{r_1+1}^{n-r_2} v_{(i)}^2 \right),$$

where the numbers r_1 and r_2 of the trimmed smallest and greatest order statistics $u_{(i)}$ and $v_{(i)}$ depend on the value of the contamination parameter ε: $r_1 = r_1(\varepsilon),\ r_2 = r_2(\varepsilon)$.

In the limiting case as $\varepsilon \to 1$, $r_1, r_2 \to [n/2]$, the estimators of scale $\widehat{\beta}_u$ and $\widehat{\beta}_v$ tend to the median absolute deviations $med|u|$ and $med|v|$, respectively, and hence $\widehat{\rho}$ tends to the median correlation coefficient

$$\widehat{\rho}_{med} = (med^2|u| - med^2|v|)/(med^2|u| + med^2|v|).$$

The median correlation coefficient also possesses the highest qualitative robustness properties [2]: its breakdown point equals 1/2.

Thus, the median correlation coefficient can be regarded as a correlative analog of the sample median and median absolute deviation – the wellknown robust estimators of location and scale possessing both quantitative minimax and highest qualitative robustness properties.

REFERENCES

Huber, P.J. (1981). *Robust Statistics. Wiley, New York.*

Shevlyakov, G.L. (1997). On Robust Estimation of a Correlation Coefficient. *J. Math. Sc.*, **83**, 90-94.

RESUME

Synthèse des estimateurs robustes minimax (selon Huber) du coefficient de corrélation dans la classe des répartitions normales contaminées avec les cas particuliers suivants: le coefficient de corrélation de l'échantillon et le coefficient de corrélation médiane.