Distance based correlation

Enyukov I.

StatPoint, Moscow

Distance based correlation is a nonparametric measure of the relation between a dependent variable y and a set of predicting variables (X). The measure is designed as follows.

Let $d_y^2(A,B) = (y(A) - y(B))^2$ be the squared distance between the values of the variable y for a pair of objects A and B, where $y(A)$ is the value of the variable y for the object A. And let $D(A,B)$ be a distance measure (for example, euclidean) between objects A and B in the predicting variables space.

Let $\bar{d}_y^2$ be the mean value of the paired squared distances by the variable y and $\bar{d}_y^2(\varepsilon)$ be the *conditional mean value* of the paired squared distance by the variable y for the object pairs *A* and *B* that satisfy the condition $D(A,B) < \varepsilon$:

$$\bar{d}_y^2(\varepsilon) = \{ Ed_y^2(A,B) / D(A,B) < \varepsilon \}$$

Then define the *distance based correlation coefficient* with *a level* ε as

$$R(\varepsilon) = 1 - \bar{d}_y^2(\varepsilon) / \bar{d}_y^2$$

When $\varepsilon \to \infty$ then $R(\varepsilon) \to 0$. If the variable y is independent of the predicting variables *X* the coefficient $R(\varepsilon) = 0$ for any value $\varepsilon \geq 0$.

Now we can define the *distance based correlation coefficient* as follows:

$$R_{dist}(y,X) = \lim_{\varepsilon \to 0} R(\varepsilon)$$

Example. Let y and *X* be normally distributed together. Then $R_{dist}(y,X) = R_{y,X}^2$, where $R_{y,X}^2$ is the square of the multiple correlation coefficient between the variable y, on the one hand, and the set of the variables *X*, on the other hand.

Estimating the distance based correlation. Supposing there is a sample of size n of $(p+1)-$variate observations (y_i, X_i), where X_i are p-variate vectors and $i = 1,...,n$.

Define a sequence of thresholds $\varepsilon_1 < \varepsilon_2 < ... < \varepsilon_k$. Then we have the sequence of estimates

$$\hat{R}(\varepsilon_i) = 1 - \hat{d}_y^2(\varepsilon_i) / \hat{d}_y^2,$$

where

$\hat{d}_y^2$ - estimated mean squared distance by variable y between all pairs of observations from the sample;

$\hat{d}_y^2(\varepsilon_i)$ - estimated conditional mean value of squared distance by variable y between the pairs of observations X_m, X_l that satisfy the condition $D(X_m, X_l) < \varepsilon_i, m < l$.

Now let n_i be the number of pairs X_m, X_l that satisfy the condition $D(X_m, X_l) < \varepsilon_i, m < l$. The problem is the following. If the value ε_i is small the number n_i can also be small and the value $\hat{R}(\varepsilon_i)$ has valuable variance. On the other hand, if the value ε_i is large enough, the estimate $\hat{R}(\varepsilon_i)$ is biased in the direction of decreasing. In particular, the value of $\hat{R}(\varepsilon_i)$ can be a negative one.

To solve the problem, we use the result showing that, if y is independent of X and y is normally distributed, the value $f_i = n(n-c_i)(1/(1-\hat{R}(\varepsilon_i)) - c_i/n)$ is F-distributed (approximately) with $n-c_i$ and c_i degrees of freedom, where $c_i \approx n_i^{1/2}$.

Using the distribution we can get for any $\hat{R}(\varepsilon_i)$ its *P*-value

$$P_i = 1 - F(n-c_i, c_i, \hat{R}(\varepsilon_i)),$$

that is the probability to obtain at the least the observed value $\hat{R}(\varepsilon_i)$ if the theoretical value $R(\varepsilon_i)$ is 0.

Then we use $\hat{R}(\varepsilon_i)$ with the minimal *P*-value (if the value is smaller than a threshold, for example 0.05) as the estimate of the distance based correlation coefficient. If the minimal *P*-value exceeds the threshold we suggest that $R_{dist}(y, X) = 0$.

Suppressing the influence of outliers. Because $\overline{d}_y^2$ is in fact the doubled variance of the variable y the estimate $\hat{d}_y^2$ is rather sensitive to the presence of outliers. In the presence of outliers the value of $\hat{d}_y^2$ increases as a rule. It can force the estimate of distance based correlation be too optimistic. To avoid it we use some robust estimator of the variance of variable y.

Applications. Distance correlation can be used together with regression analysis procedures in multivariate analysis. It can also be used as a test of predictability of time series and as the base of "threshold" forecasting. For these aims the statistics is implemented in DataScope-system for interactive graphical data analysis and TimeLand- the system for time series analysis and forecasting.

Relations with other statistics. The distance based correlation coefficient is related for example, with such statistics as the correlation integral (Grassberger, Procaccia), BDS-statistics (Brock, Hsiesh, LeBaron). For instance, the ratio $\ln(2n_i/n(n-1))/\ln(\varepsilon_i)$ gives an estimate of the "real" dimension of the data.

References

Grassberger P., Procaccia I. (1983). Measuring the Strangeness of Strange Attractors, *Physica* 9D, 189-208.

Brock W., Hsiech D., W., LeBaron B. (1991). Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence. MIT Press