CORRELATION ANALYSIS

V. Nollau
Institute of Mathematical Stochastics, Technical University of Dresden, Germany

Keywords: Random vector, multivariate normal distribution, simple correlation, partial correlation, multiple correlation, canonical correlation.

Contents

Summary

Correlation analysis is one of the most important aspects of multivariate statistical theory. Based on the different definitions of correlation coefficients (ordinary, partial, multiple and canonical), which (generally) measure the linear association between random variables or groups of random variables, a statistical analysis enables to explore the joint performance of the variables and to determine the effect of each of these variables in the presence of the others.

1. Correlation Between Two Random Variables (Simple Correlation)

Let $\mathfrak{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a 2-dimensional random vector with the expectation $\mathbb{E}(\mathfrak{X}) = \boldsymbol{\mu}$ (that means $\mathbb{E}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \boldsymbol{\mu}$) and the covariance matrix $\Gamma_{\mathfrak{X}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$.

Then the (*simple or ordinary*) *correlation coefficient* of X_1 and X_2 is defined by

$$\varrho = \varrho_{X_1, X_2} = \frac{\operatorname{cov}(X_1, X_2)}{\sqrt{\operatorname{var}(X_1) \cdot \operatorname{var}(X_2)}} \tag{1}$$

with

$$\operatorname{cov}(X_1, X_2) = \operatorname{cov}(X_2, X_1) = \mathbb{E}\big((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)\big) \tag{2}$$

and $\operatorname{var}(X_i) = \mathbb{E}\left((X_i - \mathbb{E}X_i)^2\right) > 0, \ i = 1, 2.$ (3)

This correlation coefficient is a quantitative measure for the (linear) association - called *correlation* - between the random variables X_1 and X_2 with the following properties

$-1 \le \varrho \le 1$
$\big(\varrho = 1(= -1 \text{ resp.})$ is called *positive* (*negative* resp.) *maximal correlation.*$\big)$

If and only if $|\varrho| = 1$ (maximal correlation) there exist real constants a_1, a_2, b with

$$a_1 Y_1 + a_2 Y_2 + b = 0.$$

If one relabels the random variables Y_1 and Y_2 by

$$Y_1 = aX_1 + b \quad (a > 0, \ b \text{ real})$$

and

$$Y_2 = cX_2 + d \quad (c > 0, \ d \text{ real}),$$

then the correlation coefficient between Y_1 and Y_2 is the same as the correlation coefficient between X_1 and X_2:

$$\varrho_{Y_1, Y_2} = \varrho_{X_1, X_2}.$$

(This property especially shows that the correlation coefficient is a quantitative measure for the *linear* association between two random variables.)

If a random *d*-dimensional vector $\mathfrak{X}$ has the covariance matrix

$$\Gamma_{\mathfrak{X}} = \boldsymbol{\Sigma} = (\sigma_{jk})_{\substack{j=1,\dots,d \\ k=1,\dots,d}} \tag{4}$$

with

$$\sigma_{jk} = \begin{cases} \operatorname{var}(X_j) \ (> 0) & j = k \\ \operatorname{cov}(X_j, X_k) & j \ne k, \end{cases} \tag{5}$$

then

$$\varrho_{X_j,X_k} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} = \frac{\operatorname{cov}\left(X_j, X_k\right)}{\sqrt{\operatorname{var}\left(X_j\right)\operatorname{var}\left(X_k\right)}} \tag{6}$$

is the correlation coefficient between two components of $\mathfrak{X}$, say, X_j and X_k.

Given a (mathematical) sample $\mathfrak{X}_1,\dots,\mathfrak{X}_n$ with

$$\mathfrak{X}_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}, \quad i = 1,\dots,n$$

(independent observations of $\mathfrak{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$), the correlation coefficient

$\varrho = \varrho_{X_1,X_2}$

is estimated by the (*ordinary*) *sample correlation coefficient*

$$\hat{\varrho} = \hat{\varrho}(n) = \frac{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_{1\bullet}\right)\left(X_{2i} - \bar{X}_{2\bullet}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_{1\bullet}\right)^2 \cdot \sum_{i=1}^{n}\left(X_{2i} - \bar{X}_{2\bullet}\right)^2}} \tag{7}$$

with

$$\bar{X}_{1\bullet} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} \quad \text{and} \quad \bar{X}_{2\bullet} = \frac{1}{n}\sum_{i=1}^{n} X_{2i}.$$

If $\mathfrak{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is normally distributed with the covariance matrix

$$\Gamma_{\mathfrak{X}} = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

then the density of $\mathfrak{X}$:

$$f_{\mathfrak{X}} : f_{\mathfrak{X}}(\boldsymbol{x}) = \frac{1}{2\pi}\left(\det\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left((\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)} \quad \left(\boldsymbol{x} = \left(x_1, x_2\right) \text{with} -\infty < x_1, x_2 < \infty\right) \tag{8}$$

has the following form

$$f_{\mathfrak{X}}(x_1,x_2)=\frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}}\mathrm{e}^{-\frac{\sigma_{22}(x_1-\mu_1)^2-2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)+\sigma_{11}(x_2-\mu_2)^2}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}} \quad (9)$$

or

$$f_{\mathfrak{X}}(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}}\mathrm{e}^{-\frac{1}{2(1-\varrho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\varrho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)} \quad (10)$$

with

$$\begin{aligned}\mu_1&=\mathbb{E}(X_1)\\ \mu_2&=\mathbb{E}(X_2)\end{aligned} \quad (11)$$

$$\begin{aligned}\sigma_{11}&=\sigma_1^2=\mathrm{var}(X_1)>0\\ \sigma_{12}&=\sigma_{21}=\sigma_1\sigma_2\varrho=\mathrm{cov}(X_1,X_2)\\ \sigma_{22}&=\sigma_2^2=\mathrm{var}(X_2)>0.\end{aligned} \quad (12)$$

In this case

$$\hat{\mu}_1=\bar{X}_{1\bullet}, \quad (13)$$

$$\hat{\mu}_2=\bar{X}_{2\bullet}, \quad (14)$$

$$\hat{\sigma}_{11}=\frac{1}{n}\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_{1\bullet}\right)^2, \quad (15)$$

$$\hat{\sigma}_{22}=\frac{1}{n}\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_{2\bullet}\right)^2, \quad (16)$$

and $\hat{\sigma}_{12}=\frac{1}{n}\sum_{i=1}^{n}\left(X_{1i}-\bar{X}_{1\bullet}\right)\left(X_{2i}-\bar{X}_{2\bullet}\right)$ (17)

are the so-called *maximum likelihood estimators* of $\mu_1,\mu_2,\sigma_{11},\sigma_{22}$, and σ_{12} resp. (compare *Statistical Inference*), that means

$$L\left(\mathfrak{X}_1,...,\mathfrak{X}_n;\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}_{11},\hat{\sigma}_{22},\hat{\sigma}_{12}\right) = \max_{(\mu_1,\mu_2,\sigma_{11},\sigma_{22},\sigma_{12})\in\mathbb{R}\times\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+\times\mathbb{R}} L\left(\mathfrak{X}_1,...,\mathfrak{X}_n;\mu_1,\mu_2,\sigma_{11},\sigma_{22},\sigma_{12}\right) \quad (18)$$

with the *likelihood function L:*

$$L\left(\mathfrak{X}_1,...,\mathfrak{X}_n;\mu_1,\mu_2,\sigma_{11},\sigma_{22},\sigma_{12}\right) = \prod_{i=1}^{n} f_{\mathfrak{X}}\left(X_{1i},X_{2i}\right)$$

$$=\frac{1}{\left(2\pi\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\right)^n}\prod_{i=1}^{n}\mathrm{e}^{-\frac{\sigma_{22}\left(X_{1i}-\mu_1\right)^2-2\sigma_{12}\left(X_{1i}-\mu_1\right)\left(X_{2i}-\mu_2\right)+\sigma_{11}\left(X_{2i}-\mu_2\right)^2}{2\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}} \quad (19)$$

$$=\frac{1}{\left(2\pi\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\right)^n}\mathrm{e}^{-\frac{1}{2\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}\sum_{i=1}^{n}\left(\sigma_{22}\left(X_{1i}-\mu_1\right)^2-2\sigma_{12}\left(X_{1i}-\mu_1\right)\left(X_{2i}-\mu_2\right)+\sigma_{11}\left(X_{2i}-\mu_2\right)^2\right)}$$

($L: L\left(\boldsymbol{x}_1,...,\boldsymbol{x}_n;\mu_1,\mu_2,\sigma_{11},\sigma_{22},\sigma_{12}\right)$, $\boldsymbol{x}_1,...,\boldsymbol{x}_n\in\mathbb{R}^2$, is the density function of the $2n$-dimensional random vector $\mathfrak{X}=\left(\mathfrak{X}_1...\mathfrak{X}_n\right)^T$).

Furthermore, it holds

$$\mathbb{E}\hat{\boldsymbol{\mu}}=\mathbb{E}\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad (20)$$

$$\boldsymbol{\Gamma}_{\hat{\boldsymbol{\mu}}}=\begin{pmatrix}\mathbb{E}\left(\hat{\mu}_1-\mu_1\right)^2 & \mathbb{E}\left(\hat{\mu}_1-\mu_1\right)\left(\hat{\mu}_2-\mu_2\right)\\ \mathbb{E}\left(\hat{\mu}_1-\mu_1\right)\left(\hat{\mu}_2-\mu_2\right) & \mathbb{E}\left(\hat{\mu}_2-\mu_2\right)^2\end{pmatrix}$$

$$=\begin{pmatrix}\operatorname{var}\left(\hat{\mu}_1\right) & \operatorname{cov}\left(\hat{\mu}_1,\hat{\mu}_2\right)\\ \operatorname{cov}\left(\hat{\mu}_1,\hat{\mu}_2\right) & \operatorname{var}\left(\hat{\mu}_2\right)\end{pmatrix} \quad (21)$$

$$=\frac{1}{n}\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix},$$

and the *sample covariance matrix*

$$\hat{\mathbf{\Gamma}}=\begin{pmatrix}\hat{\sigma}_{11} & \hat{\sigma}_{12}\\ \hat{\sigma}_{12} & \hat{\sigma}_{22}\end{pmatrix} \tag{22}$$

has the (probability) density

$$f_{\hat{\mathbf{\Gamma}}}(s_{11},s_{12},s_{22})=f_{\begin{pmatrix}\hat{\sigma}_{11}\\ \hat{\sigma}_{12}\\ \hat{\sigma}_{22}\end{pmatrix}}(s_{11},s_{12},s_{22})$$

$$=\begin{cases}\dfrac{n^{n-1}}{4\pi\Gamma(n-1)}\cdot\dfrac{\left(s_{11}s_{22}-s_{12}^2\right)^{\frac{n-4}{2}}}{\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)^{\frac{n-1}{2}}}\cdot e^{-\frac{n\left(\sigma_{22}s_{11}-2\sigma_{12}s_{12}+\sigma_{11}s_{22}\right)}{2\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)}} & \text{if } s_{11}>0,\ s_{22}>0 \text{ and } s_{12}^2<s_{11}s_{22}\\ 0 & \text{elsewhere}\end{cases} \tag{23}$$

with the Gamma-function $\Gamma:\Gamma(p)=\int\limits_0^\infty t^{p-1}\mathrm{e}^{-t}\mathrm{d}t\,(p>0)$.

This implies the (probability) density $f_{\hat{\varrho}}$ of the sample correlation coefficient

$$f_{\hat{\varrho}}(r)=\begin{cases}\dfrac{n-2}{\pi}\left(1-\varrho^2\right)^{\frac{n-1}{2}}\left(1-r^2\right)^{\frac{n-4}{2}}\displaystyle\int\limits_0^1\frac{x^{n-2}\mathrm{d}x}{(1-\varrho rx)^{n-1}\sqrt{1-x^2}} & \text{if } |r|\le 1\\ 0 & \text{elsewhere}\end{cases} \tag{24}$$

and the sample function (statistic)

$$T=\sqrt{n-2}\,\frac{\hat{\varrho}}{\sqrt{1-\hat{\varrho}^2}} \tag{25}$$

is t-distributed with $n-2$ degrees of freedom.

Thus to test the hypothesis $H_0:\varrho=0$ (versus the alternative $H_A:\varrho\neq 0$) one uses the statistic (25).

The problem is somewhat difficult if one wishes to test the hypothesis $H_0:\varrho=\varrho_0,\varrho_0\left(|\varrho_0|<1\right)$ is specified, versus the alternative (hypothesis)

$H_A : \varrho \neq \varrho_0$ (That means, the correlation coefficient is assumed equal to a given value ϱ_0 .). In this case R.A. Fisher (1921) (cf. Nollau, V. and Srivastava, M.S. and Carter, E.M.) suggested a transformation (Fisher's Z-transformation, c.f. Eq. (74)):

$$Z = \frac{1}{2}\ln\frac{1+\hat{\varrho}}{1-\hat{\varrho}} \tag{26}$$

with $\mathbb{E}Z = \frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)}$ and $\operatorname{var}(Z) = \frac{1}{n-3}$. (27)

With $\zeta = \frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)} (-1 < \varrho < 1)$ Fisher's *Z*-transformation has asymptotically a normal distribution $\mathrm{N}\left(\zeta, \frac{1}{n-3}\right)$, if the sample size n tends to infinity. Hence, under the hypothesis $H_0 : \varrho = \varrho_0$ the test statistic

$$(Z - \zeta_0)\sqrt{n-3} \tag{28}$$

with

$$Z = \frac{1}{2}\ln\frac{1+\hat{\varrho}(n)}{1-\hat{\varrho}(n)}, \quad \hat{\varrho}(n) = \hat{\varrho} \quad \text{(cf.Eq.(7))}, \tag{29}$$

and

$$\zeta_0 = \frac{1}{2}\ln\frac{1+\varrho_0}{1-\varrho_0} + \frac{\varrho_0}{2(n-1)} \tag{30}$$

is asymptotically standardized normally distributed.

The asymptotic distribution of *Z* also implies that an asymptotic confidence interval for ϱ is

$$\mathrm{P}\left(\tanh\left(\frac{Z - z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \right) < \varrho < \tanh\left(\frac{Z + z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \right) \right) = 1 - \alpha \tag{31}$$

for a given confidence level $1-\alpha (0 < \alpha < 1)$.

Moreover, an asymptotic test for comparing the correlation coefficients ϱ_1 and ϱ_2 of two normally distributed random vectors $\mathfrak{X}$ and $\mathcal{Y}$ can also be constructed by Fisher's transformation:

Let

$$\mathfrak{X}_1 = \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix},\ \mathfrak{X}_2 = \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix},\ldots,\mathfrak{X}_{n_1} = \begin{pmatrix} X_{1n_1} \\ X_{1n_1} \end{pmatrix} \qquad (n_1 \geq 4)$$

and (32)

$$\mathcal{Y}_1 = \begin{pmatrix} Y_{11} \\ Y_{21} \end{pmatrix},\ \mathcal{Y}_2 = \begin{pmatrix} Y_{12} \\ Y_{22} \end{pmatrix},\ldots,\ \mathcal{Y}_{n_2} = \begin{pmatrix} Y_{1n_2} \\ Y_{1n_2} \end{pmatrix} \qquad (n_2 \geq 4)$$

independent random samples from two two-dimensional normal populations $\mathrm{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathrm{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with the expectation vectors

$$\mathbb{E}\mathfrak{X}_i = \boldsymbol{\mu}_1 \quad (i = 1,\ldots,n_1)$$
$$\mathbb{E}\mathcal{Y}_i = \boldsymbol{\mu}_2 \quad (i = 1,\ldots,n_2),$$

the covariance matrices

$$\Gamma_{\mathfrak{X}_i} = \boldsymbol{\Sigma}_1 = \begin{pmatrix} \sigma_{11}^2 & \varrho_1\sigma_{11}\sigma_{12} \\ \varrho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{pmatrix} \quad (i = 1,\ldots,n_1)$$
$$\Gamma_{\mathcal{Y}_i} = \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_{21}^2 & \varrho_2\sigma_{21}\sigma_{22} \\ \varrho_2\sigma_{21}\sigma_{22} & \sigma_{22}^2 \end{pmatrix} \quad (i = 1,\ldots,n_2)$$

and the correlation coefficients

$$\varrho_1 = \varrho_{X_{1i},X_{2i}} \quad (i = 1,\ldots,n_1)$$
$$\varrho_2 = \varrho_{Y_{1i},Y_{2i}} \quad (i = 1,\ldots,n_2).$$

Under the hypothesis $H_0 : \varrho_1 = \varrho_2$ ("The correlation coefficients of both the populations are equal.") the (test) statistic

$$T = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \qquad (33)$$

with

$$Z_1 = \frac{1}{2}\ln\frac{1+\hat{\varrho}_1}{1-\hat{\varrho}_1} \quad \text{and} \quad Z_2 = \frac{1}{2}\ln\frac{1+\hat{\varrho}_2}{1-\hat{\varrho}_2}, \tag{34}$$

$$\hat{\varrho}_1 = \frac{\sum_{i=1}^{n_1}\left(X_{1i}-\bar{X}_{1\bullet}\right)\left(X_{2i}-\bar{X}_{2\bullet}\right)}{\sqrt{\sum_{i=1}^{n_1}\left(X_{1i}-\bar{X}_{1\bullet}\right)^2 \cdot \sum_{i=1}^{n_1}\left(X_{2i}-\bar{X}_{2\bullet}\right)^2}} \tag{35}$$

$$\text{and} \quad \hat{\varrho}_2 = \frac{\sum_{i=1}^{n_2}\left(Y_{1i}-\bar{Y}_{1\bullet}\right)\left(Y_{2i}-\bar{Y}_{2\bullet}\right)}{\sqrt{\sum_{i=1}^{n_2}\left(Y_{1i}-\bar{Y}_{1\bullet}\right)^2 \cdot \sum_{i=1}^{n_2}\left(Y_{2i}-\bar{Y}_{2\bullet}\right)^2}}, \tag{36}$$

$$\bar{X}_{j\bullet} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{ji} \quad (j=1,2) \quad \text{and} \quad \bar{Y}_{j\bullet} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{ji}$$

is asymptotically standardized normally distributed.

Thus the hypothesis is to reject, if for a realization t of T based on concrete samples (cf.Eq. (32)) holds $|t| > z_{1-\frac{\alpha}{2}}$ with respect to a given significance level $1-\alpha\,(0<\alpha<1)$.

-
-
-

Bibliography

Johnson N.L. and Kotz S. (1970), (1972). *Distribution in Statistics* (*continuous univariate distributions-1, 2, continuous multivariate distributions*). New York: John Wiley & Sons. [This is a very important standard work for statistical research and applications for three decades].

Müller P.H.(ed.) (1981). *Lexikon der Stochastik*. 5. Auflage. Berlin: Akademie-Verlag. [This is a dictionary for all fields of stochastics with a comprehensive description of correlation analysis].

Muirhead R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley & Sons. [This book presents all aspects of modern multivariate statistics, especially correlation theory including canonical correlation].

Nollau V. (1979). *Statistische Analysen*. 2. Auflage. Basel und Stuttgart: Birkhäuser.[An important chapter of this book presents simple, partial and multiple correlation analysis].

Röhr: M. (1987). *Kanonische Korrelationsanalyse*. Berlin: Akademie-Verlag. [This book presents a very comprehensive study about canonical correlation analysis with many applications].

Seber G.A.F.(1984). *Multivariate Observations*. New York: John Wiley & Sons. [This monograph deals with multivariate distributions, inference for the multivariate normal distribution, dimensional reductions and discriminant analysis, cluster analysis and MANOVA (multivariate analysis of variance and covariance)].

Srivastava M.S. and Carter E.M. (1983). *An Introduction to Applied Multivariate Statistics.* New York, Amsterdam, Oxford: North Holland. [This is a textbook with the main topics: multivariate techniques as ANOVA, multivariate regression, discrimination and correlation].

Biographical sketch

V. Nollau was born in 1941 and studied mathematics and theoretical physics at the Technical University of Dresden (Germany). He graduated in 1964, obtaining doctorate in 1966 and 1971 (Dr. habil.). From 1969 he was assistant professor at TU Dresden. His main research topics were operator theory, stochastic processes and random search. In 1972 he made the first contributions to stochastic optimization and decision processes theory. Since 1990 the author is professor for stochastic analysis and control. He wrote several text works including "*Statistische Analysen*" (Linear Models in Statistics). The author is dean of the faculty of mathematics in Dresden.