

ON GENERATION OF RANDOM CORRELATION MATRICES

Hsiang-Chuan Liu

Abstract

We provide a geometric method to generate a $p \times p$ Correlation matrix ρ with (i, j) th entry:

$$\rho_{ij} = \text{Cos}\theta_i \text{Cos}\theta_j \text{Cos}[\phi_i - \phi_j] + \text{Sin}\theta_i \text{Sin}\theta_j$$

where θ_i and θ_j are independent variates from $U(0, \pi)$

$$\begin{cases} \phi_i \text{ and } \phi_j \text{ are independent variates from } U(0, 2\pi) \\ i, j = 1, 2, \dots, p \end{cases}$$

θ and ϕ are statistically independent

Introduction

The P -variate normal distribution with mean vector \underline{u} and variance-covariance matrix Σ (non-singular) is defined by the pdf:

$$f(\underline{x}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{u})\Sigma^{-1}(\underline{x} - \underline{u})\right] \dots \dots \dots (1)$$

we denote this distribution :

$$N_p(\underline{u}, \Sigma) \dots \dots \dots (2)$$

In simulation, an obvious method to generate a P vector \underline{X} with this distribution is to generate Z_1, Z_2, \dots, Z_p as independent $N_1(0, 1)$ variates forming the P vector Z having $n_p(Q, I_p)$ then form \underline{X} as

$$\underline{X} = A Z + \underline{u} \dots \dots \dots (3)$$

where A is Sed and

$$A = V^{\frac{1}{2}} \rho^{\frac{1}{2}} \dots \dots \dots (4)$$

$$\Sigma = V^{\frac{1}{2}} \rho^{\frac{1}{2}} \dots \dots \dots (5)$$

$$V^{\frac{1}{2}} = \text{diag}(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \dots, \sqrt{\sigma_{pp}}) \dots \dots \dots (6)$$

is the standard deviation matrix, ρ is the correlation matrix.

If we want first to choose some different u 's and Σ 's.

It is easy to select u and V , but somewhat difficult to select ρ , since any correlation matrix must be non-negative definite.

Here we'll propose a geometric method.

Geometric Meaning of Correlation Coefficient

In sample case :

Let $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_p, Y_p)\}$

be a random sample with size p , denote

$$\begin{aligned} \underline{X} &= (x_1, x_2, \dots, x_p) \\ &= (X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_p - \bar{X}) \\ \underline{Y} &= (y_1, y_2, \dots, y_p) = (Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_p - \bar{Y}) \end{aligned}$$

$$\text{where } \begin{cases} \bar{X} = \frac{1}{p} \sum_{i=1}^p X_i \\ \bar{Y} = \frac{1}{p} \sum_{i=1}^p Y_i \end{cases}$$

then

$$\begin{aligned} \underline{X} \cdot \underline{Y} &= \sum_{i=1}^p x_i y_i = \sum_{i=1}^p (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= (p-1)S_{xy} \dots \dots \dots (8) \end{aligned}$$

$$\begin{aligned} \|\underline{X}\| &= \sqrt{\underline{X} \cdot \underline{X}} = \left(\sum_{i=1}^p x_i^2 \right)^{\frac{1}{2}} \\ &= \left(\sum_{i=1}^p (X_i - \bar{X})^2 \right)^{\frac{1}{2}} \\ &= \sqrt{(p-1) S_x} \dots \dots \dots (9) \end{aligned}$$

$$\begin{aligned} \|\underline{Y}\| &= \left(\sum_{i=1}^p (Y_i - \bar{Y})^2 \right)^{\frac{1}{2}} \\ &= \sqrt{(p-1) S_y} \dots \dots \dots (10) \end{aligned}$$

hence

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^p (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^p (X_i - \bar{X})^2 \right) \left(\sum_{i=1}^p (Y_i - \bar{Y})^2 \right)}} = \frac{S_{xy}}{S_x S_y} \\ &= \frac{\underline{X} \cdot \underline{Y}}{\|\underline{X}\| \cdot \|\underline{Y}\|} = \text{Cos} \theta_{xy} \dots \dots \dots (11) \end{aligned}$$

In population case :

(Definition 1)

Let $g = g(X, Y)$, $h = h(X, Y)$ be two functions of random vector (X, Y)

(i) The inner product of two random functions g and h is defined as

$$g \cdot h = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)h(x, y)f(x, y)dx dy \dots\dots\dots(12)$$

(ii) The norm of random function g is defined as

$$\|g\| = \sqrt{g \cdot g} \dots\dots\dots(13)$$

here suppose

$$\begin{cases} g(x, y) = x - u_x \\ h(x, y) = y - u_y \end{cases} \dots\dots\dots(14)$$

$$\text{where } \begin{cases} u_x = \int_{-\infty}^{\infty} xf(x)dx \\ u_y = \int_{-\infty}^{\infty} yf(y)dy \end{cases}$$

$f(x)$ and $f(y)$ are marginal pdf's of X and Y respectively
then

$$\begin{aligned} g \cdot h &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - u_x)(y - u_y)f(x, y)dx dy \\ &= \sigma_{xy} \dots\dots\dots(15) \end{aligned}$$

where σ_{xy} is the covariance of X and Y

$$\begin{aligned} \|g\| &= \sqrt{g \cdot g} = \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - u_x)^2 f(x, y)dx dy \right]^{\frac{1}{2}} \\ &= \left[\int_{-\infty}^{\infty} (x - u_x)^2 f(x)dx \right]^{\frac{1}{2}} \\ &= \sigma_x \dots\dots\dots(16) \end{aligned}$$

similarly we have

$$\|h\| = \sqrt{h \cdot h} = \sigma_y \dots\dots\dots(17)$$

where σ_x and σ_y are standard deviation of x and y respectively

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{g \cdot h}{\|g\| \cdot \|h\|} = \text{Cos } \theta_{gh} \dots\dots\dots(18)$$

Therefore, we know that any correlation coefficient of two random variables may be viewed as the cosine of the angle between two corresponding vectors, conversely, any cosine of the angle between two vectors can be viewed as a correlation coefficient of two corresponding random vectors.

Generating Random Correlation Matrices

In view of the above discussion, we can generate any $p \times p$ correlation matrix by considering the cosine of the angles of the sampling vectors from the unit sphere :

$$S = \{ (X_1, X_2, X_3) \mid X_1^2 + X_2^2 + X_3^2 = 1, X_1, X_2, X_3 \in \mathbb{R} \} \dots\dots\dots (19)$$

Let $\underline{X}_i = (X_{1i}, X_{2i}, X_{3i})$ be any vector belonging to S

$i = 1, 2, \dots, p$

write \underline{X}_i in spherical coordinate :

$$\exists ! \theta_i \in (0, \pi), \quad \exists ! \phi_i \in (0, 2\pi)$$

such that

$$\begin{cases} X_{1i} = \cos\theta_i \cos\phi_i \\ X_{2i} = \cos\theta_i \sin\phi_i \\ X_{3i} = \sin\theta_i \end{cases} \dots\dots\dots (20)$$

then we have

$$\begin{aligned} \rho_{ij} &= \frac{\underline{X}_i \cdot \underline{X}_j}{|\underline{X}_i| \cdot |\underline{X}_j|} = \underline{X}_i \cdot \underline{X}_j \\ &= \cos\theta_i \cos\phi_i \cos\theta_j \cos\phi_j + \cos\theta_i \sin\phi_i \cos\theta_j \sin\phi_j + \sin\theta_i \sin\theta_j \\ &= \cos\theta_i \cos\theta_j \cos(\phi_i - \phi_j) + \sin\theta_i \sin\theta_j \dots\dots\dots (21) \end{aligned}$$

Therefore the desired random correlation matrix is

$$\rho = [\rho_{ij}]_{p \times p} \dots\dots\dots (22)$$

Conclusion

In consideration of the geometric meaning of correlation coefficients, we obtain a generating method to build a $p \times p$ random correlation matrix ρ with (i,j)th

entry :

$$\rho_{ij} = \cos\theta_i \cos\theta_j \cos(\phi_i - \phi_j) + \sin\theta_i \sin\theta_j \dots\dots\dots (23)$$

- where
- θ_i and θ_j are independent variates
 - from $U(0, \pi)$
 - ϕ_i and ϕ_j are independent variates
 - from $U(0, 2\pi)$
 - $i, j = 1, 2, \dots, p$
 - θ and ϕ are statistically independent

References

Andersom, T.W.(1984) An Introduction to multivariate Statistical Analysis, 2nd ed. John Weley & Sons.

Chalmers, C.P.(1975) Generation of Correlation Matrices with a Given Eigen-Structure, JSCS, 133-139.

Kennedy, Jr. W.J. & Gentle, J.E.(1980) Statistical Computing, translated by Taipei Hua Tai Book Store(1982).

Lang, S. (1985) Linear Algebra 2nd ed. Addison Wesley Publishing Co.

Maindonald (1984) Statistical Computation translated by Taipei Hua Tai Book Store.

ABSTRACT

Several factors which control a pupil's ability to interpret and handle questions of the type mentioned above are discussed. These factors are: (a) the limitation on the capacity of the working memory, (b) the noise which swamps the signal, and (c) the language or the words used in an item. This study is based on theoretical arguments with empirical evidence recently conducted in the Centre for Science Education, University of Glasgow, Scotland. Starting from a brief overview of the information processing theory, the author explores subsequently the three mentioned factors which have an effect upon the pupils' performance in the exam situation.

Based upon the currently popular perspective of information processing theory, there are three factors which control a pupil's ability to interpret and handle questions of the type mentioned above: (a) the limitation on the capacity of the working memory, (b) the noise which swamps the signal, and (c) the language or the words used in an item. This study is based on theoretical arguments with empirical evidence recently conducted in the Centre for Science Education, University of Glasgow, Scotland. Starting from a brief overview of the information processing theory, the author explores subsequently the three mentioned factors which have an effect upon the pupils' performance in the exam situation.

Research findings were in fact in favour of the argument-based or information processing model and they call for attention not to make lists of irrelevant questions of diverse subject rather than a set of psychological features. When writing test questions, the following conditions are recommended:

1. What am I trying to ask?
2. How much information can be manipulated?
3. Does the student have a perceptible a priori intention to solve the question he is asking?
4. Is the noise justified on the objectives being tested?
5. Am I aggravating the conditions above by the words I use?