ISSN: 0975-833X

Availableonlineathttp://www.journalcra.com

International Journal of Current Research
Vol. 11, Issue, 05, pp.3632-3635, May, 2019

DOI: https://doi.org/10.24941/ijcr.35375.05.2019

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

MULTIVARIATE NORMAL SAMPLING; AN APPROACH THROUGH PRINCIPAL COMPONENT ANALYSIS

***Bushra Shamshad and Junaid Sagheer Siddiqi**

Department of Statistics, University of Karachi, Pakistan

ARTICLEINFO

Article History:
Received 20th February, 2019
Received in revised form
24th March, 2019
Accepted 10th April, 2019
Published online 30th May, 2019

Key Words:
Principal Component Analysis, Sampling, Multivariate Sampling, Random Samples, Orthogonality.

ABSTRACT

In this paper we discuss the use of Principal Component Analysis (PCA) for simulating random samples from multivariate normal distribution, using mean vector and covariance matrix. Sampling is an important aspect in the field of Statistics. We can generate random samples from various univariate distributions either discrete or continuous. We can also generate samples from bivariate distributions for that purpose there are different tables available. But, sampling in that manner is troublesome. In this article we use PCA; a multivariate technique for the purpose of sampling. Furthermore, various properties related to the multivariate normal data can be verified by simulating the samples.

Citation: Bushra Shamshad and Junaid Sagheer Siddiqi, **2019.** "Multivariate normal sampling; An approach through principal component analysis", *International Journal of Current Research*, 11, (05), 3632-3635.

INTRODUCTION

Sampling plays an important role in a large number of statistical researches. We can generate random sample from various distributions either discrete or continuous such as Uniform, Binomial, Gamma, Exponential, Normal etc. But all these distributions are univariate. We can also generate random samples from the bivariate distributions. Tables are available for this purpose. But, it is quite difficult to handle these tables for the sampling. We tried to generate random samples from multivariate normal distribution using multivariate technique PCA. Traditionally, PCA is either carried on variance-covariance matrix, or on correlation matrix (often known as the standardized variance-covariance matrix). When the variation in the variables is of main interest and one has to encounter the variability of each variable having the same units then covariance matrix is a good choice. Otherwise if the units of measurement of the individual variates differ then correlation matrix is preferred over covariance matrix. The outcome from both the matrices will give different results. If there is no correlation between the original variates then the result will be same as that of the original variable. The transformed variables (Principal components) are of the same dimension as that of the original set of variables, say, if we have q numeric variables resultant PC's are also q in number arranged in ascending order according to their accounted variation. Each transformed PC is a linear combination of the all the original variables with associated coefficient provided in the eigen vectors of either matrix (covariance/correlation). The length of eigenvectors is typically taken as one. The main property of the transformed variables is orthogonality that is PC's are uncorrelated. In general, Principal Components (y_i) are uncorrelated linear combination of the set of observed multivariate data (X_q) often maximizing the variances of the transformed Variables, Principal Component,

$$y_i = \underset{\sim}{\beta}^{(i)} \underset{\sim}{X} \quad i = 1, 2..., q \tag{1}$$

Pearson (1901) originated the concept of PCA which then later carried out by Hoteling (1933). The application of PCA is discussed by Rao (1964), Cooley and Lohnes (1971), and Gnanadesikan (1977). Excellent statistical treatment of principal components are found in Kshirsagar (1972), Morrison (1976), and Mardia, Kent and Bibby (1979). Recently, Jolliffe and Cadima (2016) provide a review on PCA and its current development. PCA have no theoretical assumptions, makes it adaptive and hence many researchers in a variety of fields are applying PCA for transformation. Orthogonal principal component's (PC's) are then used to tailor structure in the data. Kim & Kim (2012) questioned the independence assumption subject to assumption of mulivaraite normality of the variable. They proposed an alternate method named as "Independent Component Analysis", whose components are not only uncorrelated but are independent even when the normality assumption is violated. Giuliani (2017) define PCA as "hypothesis generating tool". Baytes et al.

**Corresponding author: Bushra Shamshad*
Department of Statistics, University of Karachi, Pakistan.

(2016) proposed sparse PCA which uses stochastic gradient framework to introduce sparsity to the loading vectors of PCA. Efficiency of sparse PCA was examined by using large-scale electronic medical record data. Winters et al. (2016) use the technique for environmental disaster data.

Standard normal transformation through PC

For the purpose of sampling, the property of Orthogonality of principal component is very useful. This procedure of multivariate sampling becomes quite easy when using this method. As, in univariate sampling we assume that population parameters are known to generate a random sample. Similarly, in multivariate sampling we consider a mean vector and a covariance matrix.

In this section, we discussed the method and show that how a normally distributed random vector, whose probability distribution is multivariate normal with mean vector μ (of order m×1) and variance covariance matrix Σ (of order m×m), can be obtained from m independent univariate normal distribution, using the transformation taken from Guttman, (1982).

$$Z = P'(y- \mu) \tag{2}$$

Where,

P: matrix if eigenvector of a variance covariance matrix.
Y: normally distributed random vector.
μ: mean vector of the random vector y.

The m random vector Y is said to be multi-normally distributed (provided in Equation 3), with probability density function $P_y(y_1,y_2,.....,y_m)$ is of the form

$$g_y(\mathbf{y}) = g_y(y_1,y_2,.....,y_m)$$

$$\frac{\left|\Sigma^{-1}\right|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}}\exp[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})]$$

For $(y_1, y_{2...}, y_m) \in R^m$, the matrix $\sum^{-1}=(\sigma^{ij})$ is an m×m positive definite matrix of constant, with $\sigma^{ij}=\sigma^{ji}$, and $\mu = (\mu_1, \mu_{2...}\mu_m)'$ is such that $-\infty < \mu_j < \infty, j = 1,2,3...,m$.
Now consider the transformation

$$z = P'(y - \mu)$$

The jacobian of transformation is

$$\left|J\right| = \frac{\sqrt{\left|\mathbf{P}'\Sigma^{-1}\mathbf{P}\right|}}{\sqrt{\left|\Sigma^{-1}\right|}} \tag{4}$$

Since $\sum^{-1}$ is positive definite symmetric, there exist an orthogonal matrix P such that, $P'\sum^{-1}P = D$ or $\sum^{-1} = PDP'$, where D is a diagonal matrix:

$$D = \begin{pmatrix} \lambda_1 & 0 & . & . & 0 \\ 0 & \lambda_2 & . & . & 0 \\ 0 & 0 & . & . & 0 \\ 0 & 0 & . & . & 0 \\ 0 & 0 & . & . & \lambda_m \end{pmatrix} \quad ,\lambda_j>0$$

With λ_ja characteristic root of $\sum^{-1}$, $\left|D\right| = \left|\sum^{-1}\right| = \prod_{j=1}^{m}\lambda_j$, since P is orthogonal. The density function of $z = P'(y-\mu)$ is obtained as

$$g(z) = g(z_1,z_2,\ldots,z_n) =$$

$$\frac{\left|\Sigma^{-1}\right|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}}\exp[-\frac{1}{2}\mathbf{z}'(\mathbf{P}'\Sigma^{-1}\mathbf{P})\mathbf{z}].\frac{\left|\sqrt{\mathbf{P}'\Sigma^{-1}\mathbf{P}}\right|}{\sqrt{\left|\Sigma^{-1}\right|}} \tag{5}$$

$$= \frac{\sqrt{|\mathbf{D}|}}{(2\pi)^{\frac{m}{2}}}\exp[-\frac{1}{2}\mathbf{z}'\mathbf{D}\mathbf{z}]$$

Since D is diagonal so that $z'Dz = \sum_{j=1}^{m}\lambda_j z_j^2$. Thus,

$$g_z(z) = \prod_{j=1}^{m}\frac{\sqrt{\lambda_j}}{(2\pi)^{\frac{m}{2}}}\exp[-\frac{\lambda_j z_j^2}{2}] \tag{6}$$

That is to say, $g_z(z)$ in Equation (6) is the product of marginal densities of random variables z_j's, are distributed independently as univariate normal with mean 0 and variance $1/\lambda_j$.

It is clear that z_j; j = 1,2,…,m, follow normal distribution with zero mean vector and with variance $V(z_j) = \frac{1}{\lambda_j}$, on its diagonal, such that each z_j's are independent of each other. That is,

$$Z = P'(y- \mu) \sim N(\underline{0}, D^{-1}) \tag{3}$$

Let consider another transformation $U = \mathbf{D}^{\frac{1}{2}}Z$; such that

$$E(U) = \mathbf{D}^{\frac{1}{2}}E(Z) = 0 \tag{7}$$

and

$$\text{cov}(U) = \mathbf{D}^{\frac{1}{2}}E(ZZ')\mathbf{D}^{\frac{1}{2}} = I \tag{8}$$

Then, $U \sim N(0, I)$.

Since $U = \mathbf{D}^{\frac{1}{2}}Z = \mathbf{D}^{\frac{1}{2}}\{P'(y- \mu)\}$ we get y as;

$$y = P\mathbf{D}^{-\frac{1}{2}}U + \mu \tag{9}$$

Thus, the expected value of the variable y is

$$E(y) = E(P\mathbf{D}^{-\frac{1}{2}}U + \mu) = \mu \tag{10}$$

and the variance of y is

$PD^{-1/2} D^{-1/2}P$
$=PD^{-1} P$ (11)

The result stated in Equation (10) and (11) can be explained as if a random vector variable y has probability density function, then the constants μ of the quadratic form in the exponent are the mean value of y, and the inverse of the matrix of the quadratic form is the dispersion matrix of y.

Steps of sampling

- Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_p)$ and $\sum = (\sigma_{ij})_{m \times m}$ be the known mean vector and variance covariance matrix, respectively.
- Carry out eigen analysis of the given variance covariance matrix Σ. Let λ_1, λ_2,…λ_m are m characteristic roots of the variance covariance matrix and their corresponding eigenvectors are $[e_1, e_2,\ldots,e_m]$ where
- Construct a diagonal matrix $\mathbf{D}^{\frac{1}{2}}$, whose diagonal entries are the reciprocal of square root of the eigenvalues.
- Generate a sample matrix from standard normal distribution of order n×m.
- Then using y = P $\mathbf{D}^{\frac{1}{2}}$ U + μ transform it into y, the resultant data set obtained is a multivariate sample, whose estimated mean vector and estimated variance covariance matrix is approximately same as the population mean vector and covariance matrix.

R codes for Multivariate Normal Sampling Using PCA (through Covariance Matrix)

```
> x=matrix(c(4, 0, 0, 0, 9, 0, 0, 0,
1),nrow=3,ncol=3,byrow=TRUE)
>eigen=eigen(x)
>eigen

eigen() decomposition
$`values`
[1] 9 4 1

$vectors
[,1] [,2] [,3]
[1,]    0    1    0
[2,]    1    0    0
[3,]    0    0    1

>eigen_values=eigen$values
>eigen_vector=eigen$vectors
>eigen_vector
[,1] [,2] [,3]
[1,]    0    1    0
[2,]    1    0    0
[3,]    0    0    1
>sqrt_values=sqrt (eigen_values)
> D=diag (sqrt_values, nrow=3, ncol=3, names=T)
> D
[,1] [,2] [,3]
[1,]    3    0    0
[2,]    0    2    0
[3,]    0    0    1
> P=eigen_vector
> P
[,1] [,2] [,3]
[1,]    0    1    0
[2,]    1    0    0
[3,]    0    0    1
> U=matrix (NA, 3, 10000)
>for (i in 1:3){
+   U[i,]=rnorm(10000,0,1)}
> UD=D%*%U
> y=t(P)%*%UD
>cov(t(y))


 [,1]       [,2]       [,3]
[1,] 3.99531331 0.03118828 0.02612224
[2,] 0.03118828 9.08823815 0.01811803
[3,] 0.02612224 0.01811803 0.99334097
```

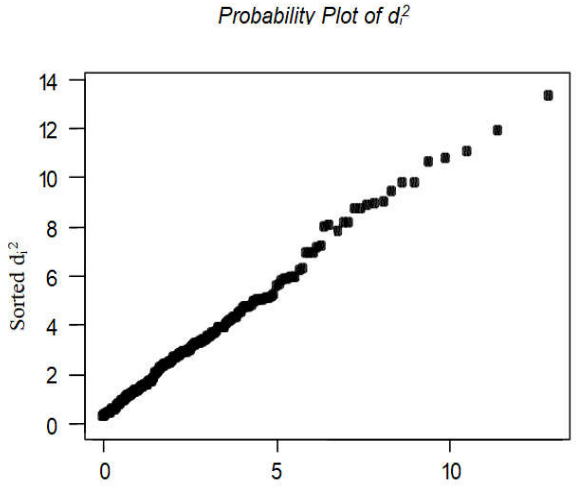
Assesment of the normality

It is prudent to check that the new transform data follows the normal distribution. A popular approach is to be the probability plots of d_i^2 described by Healy (1968).
Where,

$$d_i^2 = (y_i - \boldsymbol{\mu})' \Sigma^{-1} (y_i - \boldsymbol{\mu}) \quad (12)$$

And, $y_i = (y_{i1}, y_{i2}, \ldots, y_{im})$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_p)$ and $\sum = (\sigma_{ij})_{m \times m}$

If the data have a multivariate normal distribution then, these plot will be linear; any systematic departure from linearity signifies departure from multivariate normality of the data.



The normal probability plot forms a straight line, emphasizing that the sample generated from multivariate normal using PCA, is normally distributed.

Conclusion

From the above analysis we can see clearly that it is possible to generate sample from the multivariate normal distribution, using the Orthogonality property of principal components or more formally principal axes. The independence of eigen vectors work miraculously well. One point should be realize carefully that, in order to get good sample the size of the sample, generated from the standard normal distribution should be as large as possible. Since, we know that as the sample size increases, our estimate approaches to the population parameter.

REFERENCES

Anderson, T.W. 1984. An Introduction to Multivariate Statistical Analysis. (2nd edition). New York: John Wiley.

Barlett, M.S. 1950. Tests of significance in factor analysis. *Brit. J .Psych.,* (stat. sec.), 3, 77-85.

Baytas, I. M., Lin, K., Wang, F., Jain, A. K. and Zhou, J. 2016. Stochastic convex sparse principal component analysis. *EURASIP Journal on bioinformatics and systems biology,* 2016(1), 15. doi:10.1186/s13637-016-0045-x.

Girshick, M. A. 1939a. On the Sampling Theory of Roots of Determinantal Equations. *Ann. Math. Stat.,* 10, 203-244.

Giuliani, A. 2017. The application of principal component analysis to drug discovery and biomedical data. Drug Discovery Today, 22(7), 1069-1076.

Gnanadesikan, R. 1977. Methods for Statistical Data Analysis of Multivariate Observations. New York, Wiley.

Guttman, I. 1982. Linear Models: An Introduction. New York: John Wiley, pp. 62-66.

Hotelling, H. 1933. Analysis of a Complex of Statistical Variables into Principal Components. *J. Of Educational Psychology,* 24, 417-441, 498-520.

Johnson, R.A. 1988. Applied Multivariate Statistical Analysis. (2nd edition), Prentice-Hall, New Jersey. pp. 340-366.

Jolliffe Ian, T. and Cadima, J. 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: *Mathematical, Physical and Engineering Sciences,* 374(2065), 20150202.

Kim, D. and Kim, S.-K. 2012. Comparing patterns of component loadings: Principal Component Analysis (PCA) versus Independent Component Analysis (ICA) in analyzing multivariate non-normal data. *Behavior Research Methods,* 44(4), 1239-1243.

Krzanowski, W.J. and Marriott, F. H. C. 1995. Multivariate Analysis Part II, Classification, Covariance Structures and Repeated Measurements. London: Arnold, pp. 138-143.

Mardia, K.V., Kent, J.T. and Bibby, J.M. 1979. Multivariate Analysis. Inc. New York, Academic Press.

Morrison, D.F. 1976. Multivariate Statistical Methods. 2nd Edition, Tokyo: McGraw Hill Kogahusha.

Pearson, K. 1901. On Lines and Planes of Closet Fit to System of Points in Space. Philosophical magazine (sixth series), 2, 559-572.

Rao, C.R. 1964. The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhya, A, 329-358.

Winters, C. A., Moore, C. F., Kuntz, S. W., Weinert, C., Hernandez, T. and Black, B. 2016. Principal components analysis to identify influences on research communication and engagement during an environmental disaster. BMJ Open, 6(8), e012106.

****