Journal of Modern Applied Statistical Methods

Volume 13 | Issue 1 Article 23

5-1-2014

Distance Correlation Coefficient: An Application with Bayesian Approach in Clinical Data Analysis

Atanu Bhattacharjee
Malabar Cancer Centre, Kerala, India, atanustat@gmail.com

Journal of Modern Applied Statistical Methods
May 2014, Vol. 13, No. 1, 354-366.

ISSN 1538 − 9472

Distance Correlation Coefficient: An Application with Bayesian Approach in Clinical Data Analysis

Atanu Bhattacharjee
Malabar Cancer Centre
Kerala, India

The distance correlation coefficient – based on the product-moment approach – is one method by which to explore the relationship between variables. The Bayesian approach is a powerful tool to determine statistical inferences with credible intervals. Prior information about the relationship between BP and Serum cholesterol was applied to formulate the distance correlation between the two variables. The conjugate prior is considered to formulate the posterior estimates of the distance correlations. The illustrated method is simple and is suitable for other experimental studies.

Keywords: Conjugate prior, credible interval, distance covariance, canonical correlation

Introduction

The correlation coefficient is a widely used tool to observe the association between two random variables in experimental research. The assessment of relation between two variables (X, Y) is a common problem. The Pearson and Spearman correlation coefficients are wonderful tools to explore the relationship between two variables. Canonical, rank and Renyi correlations are the most widely used tools to investigate the strength of relation between random vectors (Bickel & Xu, 2009). The Renyi (1959) correlation becomes zero if X and Y are independent, thus, the Renyi (1959) is examined on maximal correlation. The Pearson correlation coefficient computation is simpler than the Renyi correlation coefficient. It is well-known that Pearson's product correlation coefficient ρ becomes zero for bivariate normal independence. In the multivariate case, the diagonal matrix Σ becomes independent, but it is unable to specify dependence

Dr. Atanu Bhattacharjee is in the Division of Clinical Research and Biostatistics. Email at: atanustat@gmail.com.

for general case. It may be concluded the ρ and Σ are not able to characterize independence in general.

The joint independence of random variables can be explored through distance correlation and is measured with product moment correlation ρ. It is the measures of correlation with multivariate dependence coefficients through arbitrary random vectors. Basically, the distance correlation is a product-moment correlation and a generalized form of bivariate measures of dependency. It is a very useful and unexplored area for statistical inference.

A new type of coefficient applicable to measure the dependence between random vectors of equal or unequal distance is useful for complicated dependence structures in multivariate data. The introduction of distance correlation is well detailed (Szekely, et al., 2007) and it can be computed with a simple formula of sample size $n > 2$. It is free with matrix inversion and estimation of parameters. The distance correlation has the advantage over there. The literature on testing measures of dependence is rich (Anderson, 2003; Blomqvist, 1950; Hollander & Wolfe, 1999; Blum, et al., 1961). The Likelihood Ratio Test (LRT) and Wilks Lambda are applicable for multivariate data but fail if dimension exceeds the sample size. The proposed method distance correlation with Bayesian approach is completely new.

In another aspect, it is general practice to ignore prior information about the relation between variables and establish the new correlation. As an alternative, the Bayesian approach takes the opportunity to incorporate the prior information of the variables to establish the inference about correlation. The posterior estimate of the correlation coefficient is applied to explore the relation between maternal weight and infant birth-weight (Bashir, 1997). The Bayesian approach is an attractive method for estimating tools because it incorporates previous studies' observations into its calculation. The aim of this article is to elaborate the application of a Bayesian approach in distance correlation. The work is illustrated with the estimation of distance correlation between serum cholesterol and BP. The data are captured from two different studies detailed below.

Distance Covariance and Distance Correlation

Distance covariance between the random variables X and Y can be defined with the marginal characteristic function $f_X(t)$ and $f_Y(s)$ by:

$$V^2(X,Y) = \left[f_{(X,Y)}(t,s) - f_x(t) f_Y(s) \right]^2. \quad (1)$$

The function $f_{(X,Y)}$ is a joint characteristic function of X and Y. The terms s and t are vectors and the product of t and s is $< t, s >$. The distance covariance measures the distance $\| f_{(X,Y)}(t,s) - f_x(t) f_Y(s) \|$ between the joint characteristic and marginal characteristics functions. The random vectors X and Y are in R^p and R^q respectively. The hypotheses are $H_0 : f_{X,Y} = f_X f_Y$ and $H_1 : f_{X,Y} \neq f_X f_Y$. The distance variance is:

$$V(X) = \left[f_{(X,X)}(t,s) - f_X(t) f_X(s) \right]. \tag{2}$$

The distance correlation between X and Y is defined with finite first moments $R(X, Y)$ by

$$R^2(X,Y) = \frac{V^2(X,Y)}{\sqrt{V^2(X) V^2(Y)}} > 0 \text{ , otherwise = 0.}$$

The distance covariance $V_n(X, Y)$ is defined with

$$V_n^2(X,Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl}. \tag{3}$$

Similarly it can be defined with

$$V_n^2(X,X) = \frac{1}{n^2}. \tag{4}$$

The parameter $a_{kl} = |X_k - Y_l|$, $\bar{a}_{k.} = \frac{1}{n} \sum_{l=1}^{n} a_{kl}$, $\bar{a}_{.l} = \frac{1}{n} \sum_{k=1}^{n} a_{kl}$ and $\bar{a}_{..} = \frac{1}{n^2} \sum_{k=1}^{n} a_{kl}$

$$A = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}. \tag{5}$$

B_{kl} is defined similarly.

Properties of Distance Correlation Coefficient

The distance correlation provides the scope to generalize the correlation between variables (X and Y) by R is defined on arbitrary dimensions $R=0$ independent of X and Y. The range of the distance correlation is $0 \le R \le 1$. R can be defined as the function of Pearson correlation coefficient ρ with $R(X,Y) < |\rho(X,Y)|$ with equality when $\rho = \pm 1$.

Importance of Distance Covariance

The random variables X and Y are expressed as $A_i = X_i + \varepsilon_i$ and $B_j = Y_j + \varepsilon_j$ respectively. The error terms ε_i and ε_j are independent with the variables X_i and Y_j. The relation between random functions A_i and B_j is irrelevant, but the relation between X_i and Y_j is important and a matter of concern. The strength of relation between X and Y can be measured through distance correlation in this scenario.

Distance Correlation in One-sided Test

The frequency approach tests the problem through $p(X)$ value of the null hypothesis H_0. By contrast, Bayesian measures through posterior probability $p(H_0 / X)$. Let the data follow a normal distribution (θ, σ^2) with null hypothesis $H_0 : \theta \le 0$ and alternate $H_1 : \theta > 0$. The frequency and robust Bayesian often coincide (Casella & Berger, 2002). Let the marginal distance correlation ρ be applied between $p(X) = 1 - \Phi(X/\sigma)$ and $p(H_0 / X)$. The distance correlation should be greater than or equal to zero. Because $p(X)$ and $p(H_0 / X)$ both decrease with respect to X.

Distance Correlation between Parameter and Unbiased Estimator

Suppose, (θ, X) are random variables with joint characteristics function $f_{(X,Y)}(t,s)$ and the marginal distribution of θ is π. The estimator of θ is $\delta(X)$ and square error loss is $r(\pi, \delta) = E\left[\delta(X) - \theta\right]^2$ and risk is $\delta_\pi(X) = E(\theta / X)$. The distance correlation between θ and $\delta(X)$ is

$$\rho\left(\theta,\delta\left(X\right)\right)=\frac{\mathrm{var}\left(\theta\right)+\mathrm{cov}\left\{\theta,b\left(\theta\right)\right\}}{\sqrt{\mathrm{var}\left(\theta\right)}\sqrt{\mathrm{var}\left\{\theta+\left(\theta\right)\right\}+\tau\left(\pi,\delta\right)-E\left\{b^{2}\left(\theta\right)\right\}}}. \quad (6)$$

Statistical Methods

The Bayes' Theorem provides prior information about the relevant parameter for a specific statistical analysis. It is helpful to test the hypothesis in presence of posterior probability of the parameter of interest. The parameter of interest $R(X,Y)$ can be computed with posterior probability through Bayes' theorem:

$$P\left(R\left(X,Y\right)/\mathit{Information}\right)=\frac{\left(P\left(\mathit{Information}/R\left(X,Y\right)\right)\right)P\left(R\left(X,Y\right)\right)}{\left(P\left(\mathit{Information}\right)\right)}. \quad (7)$$

The term P ($R(X, Y)$) is the prior probability of $R(X, Y)$ observed from the previous study. The term P (*information*/$R(X, Y)$) is the likelihood of $R(X, Y)$ that occurred in a previous study or is in data collected by an investigator. The sum of the function $1/(P$ (*Information*)) should be equal to 1 as the theory of total Bayes theorem. The relation between posterior and prior is:

$$\mathit{PosteriorProbability}\propto \mathit{LikelihoodX}\ \mathit{PriorProbability}. \quad (8)$$

The posterior density of $R(X, Y)$ is generated with

$$P\left(R\left(X,Y\right)/x,y\right)\propto P\left(R\left(X,Y\right)\right)\frac{\left(1-R\left(X,Y\right)^{2}\right)^{(n-1)/2}}{\left(1-R\left(X,Y\right)*r\right)^{n-\left(\frac{3}{2}\right)}}. \quad (9)$$

The mean and variance of X and Y are μ_1, μ_2, σ_1^2 and σ_2^2 respectively. The mean (z) is derived from

$$e^{z}=\frac{\mu_1\sigma_2}{\mu_2\sigma_1}. \quad (10)$$

The term $R\left(X,Y\right)$ is defined by $\tanh\varepsilon$ and is assumed $\varepsilon\sim N\left(z,\frac{1}{n}\right)$. The mathematical formulations are detailed in Fisher (1915). The hyperbolic

transformation plays a role in considering the conjugate prior with normal distributions. The posterior mean can be represented with

$$\mu_{Posterior} = \varepsilon^2_{Posterior}\left[\eta_{Prior}\tanh^{-1}R(X,Y)_{Prior} + \eta_{Likelihood}\tanh^{-1}R(X,Y)_{Likelihood}\right] \quad (11)$$

$$\sigma^2_{Posterior} = \frac{1}{\eta_{Prior} + \eta_{Likelihood}} \quad (12)$$

and the prior with the form

$$P(R(X,Y)) \propto \left(1 - R(X,Y)^2\right)^c. \quad (13)$$

The prior is dependent on the choice of c; $c = 0$ gives $P(R(X,Y) \propto 1$.

Illustrated Example

Among different types of risk factors hypertension and abnormalities of lipid profiles are established reasons for coronary artery disease as observed through epidemiological and genetics studies (for details see Williams, et al., 1988). Serum cholesterol is related with blood pressure (BP) values (Ferrannini, et al., 1987; Hunt, et al., 1986, Floras, et al., 1987; Simone, et al., 1992; Sung, et al., 1997). The present work is undertaken to check whether serum cholesterol is an influencing factor of BP. The BP measurement was taken in 24 hours close observation. Study 1 was conducted to observe two drug treatment effects among liver cirrhosis patients in St. Stephen hospital during 2009 to 2011. The data on serum cholesterol and BP were observed during the study of 179 patients with follow up observations. In this article, data is considered to illustrate the application of a distance correlation coefficient between serum cholesterol and BP. In Study 2, a total of 100 patients of type 2 diabetes were observed with two types of drug treatments in Madurai Menakshi Mission Hospital in 2009. The different biochemical parameters through follow up periods were observed as an effect of drug treatment. The measurements of serum cholesterol and BP were observed through the follow up periods. The work is explored on the data to illustrate the distance correlation coefficient among the patients.

Frequentist Test for Distance Correlation

To check the distance correlation between BP and serum cholesterol, the variables were defined. For every i[th] person the BP is denoted with x_i and serum cholesterol by y_i. In Study 1, the null hypothesis to test the distance correlation coefficient is assumed as zero, i.e. $R(X, Y) = 0$. The dcor.ttest(x,y) function in the *energy* package of *R* i386 3.0.1 is applied to test the null hypothesis. Results show that the distance correlation rejects the null hypothesis that $p = 0.01$. Consequently, researchers may feel that it is possible to reject the null hypothesis of no correlation between BP and serum cholesterol.

Bayesian test for Distance Correlation

The measure of evidence of

$$p\left(H_0 / x\right)$$

is the probability of

H_0 is true with $X = x$ is

$$P\left(H_0 / x\right) = P\left(\theta \le 0 / x\right) = \frac{\int_0^{\infty} f\left(x-\theta\right)\pi\left(\theta\right)}{\int_{-\infty}^{\infty} f\left(x-\theta\right)\pi\left(\theta\right)}. \tag{14}$$

In both studies, it was assumed that the BP and serum cholesterol are correlated to each of the others. The relation with distance covariance was examined using a Bayesian approach. The relation between BP and serum cholesterol during surgery in patients was observed from anesthesia data. The estimated distance correlation between serum cholesterol and BP may be measured with error. The error arises due to the presence of small sample size. The fluctuation of observed correlation in different studies may be due to different sample sizes. Using several studies, a meta-analysis can be conducted to estimate the real correlation between BP and cholesterol. However, if lacking several studies, the Bayesian posterior estimate is applied to estimate the robust distance correlation between BP and serum cholesterol.

$$\sigma^2_{posterior} = \frac{1}{\eta_{Prior} + \eta_{Likelihood}} = \frac{1}{179 + 100} = 0.035 \tag{15}$$

$$\mu_{Posterior} = 0.035(179 \tanh^{-1} 0.58 + 100 \tanh^{-1} 0.43) \tag{16}$$

$$\mu_{posterior} = 0.0035\left(179 * 0.67 + 100 * 0.47\right) \tag{17}$$

The confidence interval is

$$\mu_{Posterior} \pm 1.96\sqrt{\left(\sigma^2_{post}\right)} = 0.58 \pm 1.96(0.0035)^{\frac{1}{2}} \tag{18}$$

i.e. (0.69, 0.47). This shows the posterior estimate of distance correlation $R(X, Y)$ is 0.58 with credible interval (0.69, 0.47). The estimate observed at the 95 confidence interval is 0.23 (0.28, 0.18). The values can be compared and a conclusion can be drawn. The simple approach for distance correlation can be extended to other experimental research.

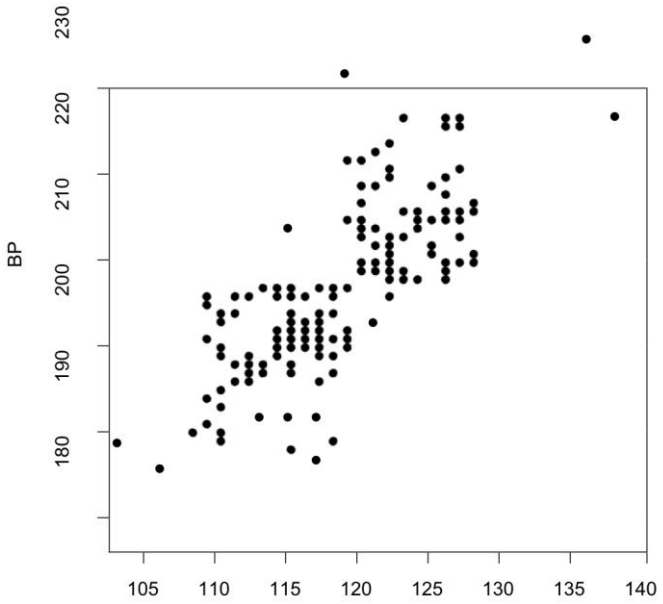

Figure 1. Relationship between BP and serum cholesterol. A positive correlation suggests that serum cholesterol is the influencing factor for high BP.

Discussion

The risk of coronary artery disease is depends on different risk factors (Levy et al., 1988; Assmann, et al., 1988). Different risk factors can be classified into global and individual risk factors; however, global risk factors are more important than individual risk factors for cardiovascular disease (Ferrara, 2002). Distance correlation is applied to explore the relation between BP and cholesterol. Distance correlation is applicable to all types of metric spaces. The idea is to apply the distance correlation by replace the Euclidean distance into metric distance. Distance covariance is also applicable to test the linear model $Y = X\beta + \varepsilon$; where (x, ε) are i.i.d. Distance covariance is defined on arbitrary dimension and it can be extended for multivariate responses. It is simple like the Pearson product moment covariance. If X and Y are sample from different metric space, then distance covariance can be measured. The Pearson correlation is the best choice to explore the relation between variables. However, it is not feasible to apply it to non-normal data. Distance covariance can also be applied to non-normal data.

The correlation coefficient for the sample average was examined with uniform prior by Daniels (1999). Extensions of the work were carried with shrinkage priors by Daniels Kass (2001). The measurements of correlation tested through logarithmic transformation of the eigenvalue (Leonard et al., 1992). Barnard et al. (2000) proposed a normal prior for a transformation of correlation coefficients. Wong, et al., (2003) give a prior probability model for graphical models and partial correlations through the sparseness of the precision matrix. Gabor et al., (2007) discussed the advantage of distance correlation over Pearson correlation: It is the generalized form of the Pearson correlation in two ways (1) its ability to measure the linear relation with consideration of all types of dependency, and (2) exposure to measure the dependency through random vectors in the arbitrary dimension. Tracz, et al., (1992) showed that the distance correlation is more suitable as a dependent index than the product moment correlation coefficient.

The Affine invariance property is important for the transformation of data in statistical inference. The Affine invariance with a group is detailed by Eaton (1989) and Giri (1996). Gabor (2007) proved distance correlation is free from Affine invariance. Correlation analysis is strong a filler to draw statistical inference in any medical research. Distance correlation is another useful tool to explore the relation between variables. Distance correlation with confidence interval is a statistical tool to sketch the inference about the relation between variables. In this study, the confidence interval between serum cholesterol and BP

was observed. The Bayesian approach was applied with credible intervals and observed with less interval estimates. Small sample sizes tend to be a problem in clinical trials due to cost and time. The Bayesian approach gives a practical concession. It is also useful choice to deal with random measurement error in weight gain relation. It is simple and accurate. The approach is helpful to explore the relation between variables more intuitively.

Conclusion

Distance correlation with a Bayesian approach is not the only choice of correlation analysis, but can be considered in many cases as an alternative of Pearson's covariance. An example with clinical trials illustrated where distance correlation can give more information not captured by traditional correlation analysis. In exploratory analysis with small sample size data, the Bayesian distance correlation is an alternative choice for the low dimensional marginal distribution of two variables. The Bayesian distance correlation can be useful to test the linear relation between variables and it can be a first choice to explore the relation between variables to made decisions about specific tools for further data analysis. Distance correlation having high value of one (or near to one) shows a strong relation between variables. The Bayesian approach is suitable tool for calculating distance correlation coefficient among variables. The work can be extended to explore the relation between bivariate observations in different experimental research. Like the correlation coefficient, distance correlation can be applied to understand the relation between variables by clinician. It can serve clinicians to know the real strength of variables and, as a result, interpretation of the results in the real life practice. In any experimental research relations between variables is unavoidable. Distance correlation can be considered as easily interpretable tool to discover the relations.

Acknowledgement

We would like to thank Dr. Shuarav in St. Stephen Hospital and Mr. Rakesh in Menakshi Mission Hospital for permission to apply data with initial review to initiate the work of this paper.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd Ed.). New York, Wiley.

Assmann G. & Schulte H. (1988). The Prospective Cardiovascular Münster (PROCAM) study: prevalence of hyperlipid aemia in persons with hypertension and/or diabetes 343 mellitus and the relationship to coronary heart disease. *American Heart Journal, 116*(6 Pt. 2): 1713-1724.

Barnard, J., McCulloch, R. & Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica, 10*: 1281-311.

Bashir S. A. & Duffy S. W. (1997). The Correction of Risk Estimates for Measurement Error. *Annals of Epidemiology, 7*: 154-164.

Bickel, P .J & Ying, X. (2009). Discussion of Brownian Distance Covariance. *The Annals of Applied Statistics, 3*(4): 1266-1269.

Blomqvist, N. (1950). On a measure of dependence between two random variables. The Annals of Mathematical Statistics, 21: 593-600.

Blum , J. R., Kiefer, J. & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics, 32*(2): 485-498.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Thomson Learning.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics, 27*: 567-78.

Daniels, M. J. & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics, 57*(4): 1174-84.

Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Hayward, CA , IMS.

Ferrannini, E., Buzzigoli, G., Bonadonna, R., Giorico, M. A., Oleggini, M., Gradizdei, L., Pedrinelli, R., Brandi, L. & Bevilacgua, S. (1987). Insulin resistance in essential hypertension. *The New England Journal of Medicine, 317*: 350-357.

Ferrara L. A., Guida, L., Iannuzzi, R., Celentano, A. & Lionello, F. (2002). Serum cholesterol affects blood pressure Regulation. *Journal of Human Hypertension, 16*: 337-343.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 19*(4): 507-521.

Floras, J. S., Hassan, M. O., Jones, J. V., & Sleight, P. (1987). Pressor responses to laboratory stresses and daytime blood pressure variability. *American Journal of Hypertension, 5*(6): 715-719.

Giri, N. C. (1996). *Group Invariance in Statistical Inference*. Edge, NJ, World River Scientific.

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric Statistical Methods* (2nd Ed.). New York, Wiley.

Hunt, S. C., Williams, R. R., Smith, J. B., & Ash, K. O. (1986). Associations of three erythrocytes cation transport system with plasma lipids in Utah subjects. *American Journal of Hypertension, 8*: 30-36.

Leonard, T. & Hsu, J. S. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics , 20*: 1669-96.

Levy, D., Anderson, K. M., Savage, D. D., Kannel, W. B., Christiansen, J. C. & Castelli, W. P. (1988). Echocardiographically detected left ventricular hypertrophy: prevalence and risk factors. The Framingham Heart Study. *Annals of Internal Medicine, 108*(1): 7-13.

Renyi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica, 10*: 441-451.

de Simone, G., Daniels, S. R., Devereux, R. B., Meyer, R. A., Roman, M. J., de Divitiis, O., & Alderman, M. H. (1992). Left ventricular mass and body size in normotensive children and adults: assessment of allometric relations and impact of overweight. *Journal of the American College of Cardiology, 20*: 1251-1260. doi: 10.1016/0735-1097(92)90385-Z

Sung, B. H., Izzo, J. L., & Wilson, M. F. (1997). Effects of cholesterol reduction on BP response to mental stress in patients with high cholesterol. *American Journal of Hypertension, 10*: 592-599.

Szekely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by Correlation of distances. *The Annals of Statistics, 35*(6): 2769-2794.

Tracz, S. M., Elomore, P. B., & Pohlmann, J. T. (1992). Correlation meta-analysis: Independent and no independent cases. E*ducational and Psychological Measurement, 52*: 879-888.

Williams, R. R., Hunt, S. C., Hopkins, P. N., Stults, B. M., Wu, L. L., Hasstedt, S. J., Barlow, G. K., Stephenson, S. H., Lalouel, J. M., & Kuida, H. (1988). Familial dyslipidemic hypertension. Evidence from 58 Utah families for a syndrome present in approximately 12% of patients with essential hypertension. *The Journal of the American Medical Association, 259*(24): 3579-3586. doi: 10.1001/jama.259.24.3579

Wong, F., Carter, C. K. & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika, 90*: 809-830.