Distance Correlation Measures Applied to Analyze Relation between Variables in Liver Cirrhosis Marker Data

Santam Chakraborty[1], Atanu Bhattacharjee[2]

[1] *Department of Radiation Oncology, Tata Memorial Hospital, Mumbai, India*

[2] *Division of Clinical Research and Biostatistics, Malabar Cancer Centre, Thalassery, India Malabar Cancer Centre, Thalassery, Kerala-670103, India*

[*] **Corresponding author:** Atanu Bhattacharjee, Division of Clinical Research and Biostatistics, Malabar Cancer Centre, Thalassery, India Malabar Cancer Centre, Thalassery, Kerala-670103, India; E-mail: atanustat@gmail.com

Abstract

Distance correlation (DC) is a new choice to compute the relation between variables. However, the Bayesian counterpart of Distance Correlation is not well established. In this paper, a Bayesian counterpart of Distance Correlation is proposed. The proposed method is illustrated with Liver Cirrhosis Marker data. Previously published data on the relation between aspartate transaminase (AST) and alanine transaminase (ALT) is used to formulate the prior information for Bayesian computation. The computed DC using the proposed method between AST and ALT (both of which are markers of liver function) is 0.44. The credible interval is ranges 0.41 to 0.46. Bayesian counterpart proposed herein to compute DC coefficient is simple and handy.

Keywords: ICC, Cannonical correlation, Credible interval, Distance covariance, Conjugate prior

Introduction

The statistical dependence between two random vectors (irrespective of the measurement dimension) can be measured by distance correlation (DC).[1,2] DC ranges between 0−1 with 0 indicating that the vectors are completely in- dependant statistically. As a generalized form of Pearson correlation it provides a method to measure multivariate independence. Szekely et al have shown that it is consistent for all dependent alternatives through finite second moments.[3] The bias outcome of DC through different dimensions are also tested.[3] The unbiased T test is considered suitable for testing the independence of variables using distance correlation.

The use of DC has been extended for high dimensional data.[4] The application of DC for functional data has also been extended recently through Hilbert space.[5] Recently, several new tools are available to the scientific community for more complex issue through Cannonical (consideration of linear combinations between variables through maximum correlation with each other), Rank and Renyi correlation (through observing the cosine angle between the linear subspaces of mean zero square integral real-valued random variables from individual random variable.[4] However, all of them having some advantages and limitations.[6] The joint independences of random variable can be explored through DC.[2] It is a matrix inversion free approach. Dependences measurement between two random variables can be observed and tested through matrix inversion free approach.[7] Experimental and observational studies in clinical medicine usually rely on

exploring the relation between two variables of interest (for example understanding how high blood pressure and increased total cholesterol in serum are related with each other to predict the risk of myocardial infarction). The objective of this present study is to demonstrate the use of a Bayesian approach to DC and formulate a methodology for calculation. The method is then illustrated with clinical trial example.

Distance Covariance and Distance Correlation

Distance covariance between the random variables X and Y is defined with marginal characteristic function of fY (t) and fY (s) by,

$$V^2 \ (X,Y) = [f_{(X,Y)} \ (t,s) - f_X(t) f_Y \ (s)]^2 \rightarrow (1)$$

The function f (X, Y) is joint characteristics function of X and Y. The terms s and t are the vectors and the product of t and s is < t, s > .The distance covariance measures the distance f(X, Y) (t, s)-f(X) f(Y) (s) between the joint characteristic function and marginal characteristics function. The random vec-tor X and Y are in Rp and Rq respectively. The hypothesis is H0: fX,Y = fX fY and H1 : fX,Y ≠ fX fY The distance variance is

$$V(X) = [f_{x,y}(t,s) \ - f_x(t) \ f_x \ (s)] \rightarrow (2)$$

DC between X and Y is defined with finite first moments R(X, Y) by

$$R^2(X, Y) = \frac{V^2(X,Y)}{\sqrt{V^2(X) \, V^2 \ (Y)}} > 0 \rightarrow (3)$$

The distance covariance Vn (X, Y) is defined with

$$V_n^2(X,Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A \ B_{kl} \rightarrow (4)$$

Similarly it can be defined as:

$$V_n^2(X,X) = \frac{1}{n^2} \rightarrow (5)$$

The parameters are $a_{kl} = \ X_l - Y_l$,, $\bar{a_k} = \frac{1}{n}\Sigma_{k=1}^{n} a_{kl}$ and $\bar{a}_{..} = \frac{1}{n^2}\Sigma_{k=1}^{n} a_{kl}$

$$A = a_{kl} - \bar{a}_{.l} + \bar{a}_{..} \rightarrow (6)$$

Similarly, BkL is defined.

Properties

The DC provides the scope to generalize the correlation between variables (X and Y) by R. It is defined on arbitrary dimensions R=0 for independent of X and Y. The range of DC is 0 < R < 1.The R can be defined as the function of Pearson correlation coefficient ρ with R(X,Y) < |ρ(X,Y)| with equality when ρ ± 1. The random variables X and Y are expressed as Ai=Xi+∈i and Bj=Yj+∈j respectively. The error terms ∈i and ∈j are independent with the variables Xi and Yj. Let the relation between random functions Ai and Bj is irrelevant. But the relation between Xi and Yj is importance and matter of concerned. The strength of relation between X and Y can be measured through DC in this scenario.

In One-sided Test

The frequency approach test the problem through p(X) value of the null hypothesis H0. In contrast, Bayesian measures through posterior probability p (H0 |X). Let the data follows normal distribution (θ, σ2) with null hypothesis H0: θ ≤ 0 and H1: θ>0. The frequency and robust Bayesian often coincide.[8] Let the marginal DC ρ is applied between p(X)=1−Φ(X/σ) and p (H0 |X).The DC should be greater than or equal to zero. Because p(X) and p (H0 |X) both are decreasing with respect to X.

Parameter and Unbiased Estimator

Suppose, (θ, X) are the random variables with joint characteristics function f(X, Y) (t, s) and marginal distribution of θ is π. The estimator of θ is δ(X) and square error loss is r(π, δ) = E[δ(X) − θ]2 and risk is δπ(X) = E(θ/X).The DC between θ and δ(X) is

$$\rho(\theta,\delta(X)) = \frac{\text{var}(\theta) + cov\{\theta b(\theta)\}}{\sqrt{\text{var}(\theta)}\sqrt{\text{var}\{\theta + (\theta)\}} + \tau(\Pi,\delta) - E\left\{b^2(\theta)\right\}} \rightarrow (7)$$

Method

The Bayes' Theorem provides the prior information about the relevant parameter for the 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

$$P(R(X,Y)/Information) = \frac{P(Information/R(X,Y))P(R(X,Y))}{P(Information)} \rightarrow (8)$$

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 likelihood of R(X,Y) occurred in the previous study or data collected by the investigator. The sum of the function 1 should be equal to 1 as the theory of total Bayes theorem. The relation between posterior and prior is

$$PosteriorProbability \propto Likelihood \times PriorProbability \rightarrow (9)$$

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

$$P(R(X,Y)/x,y) \propto P(R(X,Y))\frac{(1-R(X,Y)^2)^{(n-1)/2}}{(1-R(X,Y)\times r)^{n-\frac{3}{2}}} \rightarrow (10)$$

Let the mean and variance of X and Y are μ1, μ2, σ21, σ22 respectively. The mean (z) is derived from

$$e^z = \frac{\mu_1\sigma_2^2}{\mu_2\sigma_1^2} \rightarrow (11)$$

The term R(X, Y) is defined by tanhϵ and it is assumed ϵ N (z, 1).The mathematical formulations are detailed in Fisher (1915). The hyperbolic trans- formation plays role to consider the conjugate prior with normal distributions.

The posterior mean can be represented with

$$\mu_{\text{posterior}} = \varepsilon^2_{\text{posterior}}[\eta_{prior}\tanh^{-1}R(x,y)_{prior} + \eta_{likelihood}\tanh^{-1}R(x,y)_{likelihood}] \rightarrow (12)$$

$$\sigma^2_{\text{posterior}} = \frac{1}{\eta_{prior} + \eta_{likelihood}} \rightarrow (13)$$

The prior with the form

$$P(R(X, Y))\alpha(1 - R(X, Y)^2)^c \rightarrow (14)$$

The prior is dependent on the choice of c. The c=0 gives the P (R(X,Y) 1 the specification of prior is important for testing the parameters in hypothesis H0 and H1. The main focus of research in Bayesian approach is the specification of prior. The prior specification is carried out through regression modeling. Let the response of interest (Y), covariates (X) ,error(ϵ) and intercept (α) are in regression line through $Y = \alpha + \beta X + \varepsilon \rightarrow (15)$

Zellner (1986) has introduced the g prior for the above mentioned β coefficient. However, it is the extension of Jeffrey's prior on the error precision ϕ with uniform prior of interest α by

$$p(\beta \mid \phi, g, X) = N(0, \frac{g}{\phi}(X^T X)^{-1}), p(\phi, \alpha)\alpha\frac{1}{\phi} \rightarrow (16)$$

The information about β can be obtained through ϕ−1 (X T X)−1. Further, specified value of g gives the exposure about observed data. The specified value of g=1 says no influences of observed data. Whereas, g=5 gives 15 weight as the observed data. The selection of value of g is very important.[9] It is considered as g=n. n is the sample size. Discussed to consider g=k. (k is the number of parameters). There are several literatures about selection of g prior. The work is contributed with Jeffrey's-Zellner-sion (JZS) prior for g-value. It was represented by Liang and his colleagues and applied for correlation coefficient.[10,11] The prior is like

$$p(\beta \mid \phi, g, X) = \int N(0, \frac{g}{\phi}(X^T X)^{-1})p(g)dy \rightarrow (17)$$

$$p(\phi)\alpha\frac{1}{\phi} \rightarrow (18)$$

$$p(g)=\frac{(\frac{n}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}g - \frac{3}{2} - \frac{n}{2g} \rightarrow (19)$$

The above mentioned formula is also useful to calculate Bayes factor. The prior is applied as default prior for t-test.[7] The Bayesian factor is applied through JZS for DC in regression line. The regression coefficient β is allowed to the application JZS prior. Our goal is to compute DC, Intercept (α), regression coefficient (β) and error term (ϵ) s detailed in equation (1). Let the equation (1) further separated into Model (M1) and Model (M0) by

$$M_1: Y = \alpha + \beta X + \varepsilon \rightarrow (20)$$

$$M_0: Y = \alpha + \varepsilon \rightarrow (21)$$

The model ((M1)) states the presence of DC and absence of it by Model ((M0)).

Now, the Bayes Factor through JZS is defined [10] as,

$$BF_{10} = \frac{(\frac{n}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}x\int_0^{\infty}(1+g)^{((n-2)/2}x[1-r^2)g] - \frac{(n-1)}{2}g - \frac{3}{2_e} - (\frac{n}{2g})dg \rightarrow (22)$$

$$BF_{10} = \frac{p(\frac{Y}{M_1})}{p(\frac{Y}{M0})} \rightarrow (23)$$

If the value of BF10 becomes more than 1, it state about presences of DC otherwise not.

Testing

Under the null hypothesis H0, the model (M0) is assumed and (M1) for alternative one i.e., H1. The prior probability of null is assigned as p (M0) and alternative as p (M1). Thereafter, Baye's theorem is applied on the observed data to compute posterior probability of the hypothesis. The appearance of posterior probability of alternative Hypothesis is computed as

$$p(\frac{Y}{M_1}) = \frac{p(\frac{Y}{M_1})p(M_1)}{p(\frac{Y}{M_1})p(M_1) + p(\frac{Y}{M_0})p(M_0)} \rightarrow (24)$$

The term P (Y |M1) is the marginal likelihood of the data for alternative hypothesis. Further, the marginal likelihood is calculated as

$$p(\frac{M_1}{Y}) = \int_{\theta}^{\infty} p(Y \left| \theta, M_1 \right.) p(\theta \left| H_1 \right.) d\theta \rightarrow (25)$$

Bayes Factor is useful to compute the appearance of P (M1|Y) in comparison to P (M0 | Y)12:

$$\frac{p(\frac{Y}{M_1})}{p(\frac{Y}{M_0})} = BF_{10} x \frac{p(M_1)}{p(M_0)} \rightarrow (26)$$

Illustrated Example

Aminotransferases are serum enzymes which are used to detect malfunction of liver, heart, lung, skeletal muscles and brain.[13] Among the aminotransferases, alanine and aspartate aminotransferase (ALT and AST respectively) are routinely measured to assess liver function.[14] Kumar et al have recently published the normal range of serum AST and ALT in over 5000 Indian blood donors and have proposed normal limits for healthy population.[15] In this ex- ample we illustrate the use of DC between AST and ALT measurements in the same individuals. The generated information between AST and ALT is used as prior information of sample size of 4917 individuals.[15] The raw data on AST and ALT of 606 individuals are detailed.[16] In both the above mentioned study, the relation between Serum alanine aminotransferase (ALT) and serum aminotransferase (AST) are observed. The relations between variables are explored through distance covariance with Bayesian approach. The first relation between ALT and AST is observed.[15] The measured distance correlation data is observed with error. Bayesian posterior estimate is computed for robust

DC between ALT and AST by,

$$\sigma^2_{\text{posterior}} = \frac{1}{\eta_{\text{prior}} + \eta_{\text{likelihood}}} = \frac{1}{4917 + 606} = 0.00018 \rightarrow (27)$$

$$\mu_{\text{posterior}} = 0.00018(4917\tanh^{-1} + 606\tanh^{-1}0.80) \rightarrow (28)$$

$$\mu_{\text{posterior}} = 0.44 \rightarrow (29)$$

The confidence interval is

$$\mu_{\text{posterior}} \pm 1.96\sqrt{(\sigma^2_{post})} = 0.44 \pm 1.96(0.00018)^{\frac{1}{2}} \rightarrow (30)$$

i.e. (0.41, 0.46). It shows the posterior estimates of DC i.e., R(X,Y) is 0.44 with credible interval (0.41, 0.46). This simple approach for DC can be extended in other experimental research. The posterior computed mean is 0.44 and sample size 606. The values are applied to obtain the BF10 in equation (23). The BF10 is calculated with 8.3. It is the evidence in favour of M1 in comparison to model M0. The presence of DC is tested through g prior.

Discussion

Recently, the testing process to check the presences of DC has been at- tempted. The t-test is found suitable to test the presence of DC. The relevant factors are proposed to perform it.[3] The evaluation of direct relation between two variables is important. Pearson and Spearman correlations are commonly applied tools to explore relation between variables. The strength of relation between variable can be classified by Cannonical, Rank and Renyi Correlation.[4] The widely explored correlation tool-Pearson correlation fails in multivariate data set. It becomes zero for independent bivariate normal distribution. But it failed to specify multivariate dependence in general. The limitation can be overcome by joint independence of the random variable through DC. The DC is product-moment correlation and generalized form of bivariate measures of dependency. It is very much useful and unexplored area for statistical inference. The idea of this work is to establish the application of new types of correlation tools for measurement of dependence between variables. It is more applicable for complicated multivariate data. The detailed application DC is recently established.[2] There are several advantages for application of DC over simple. The Bayesian application on DC computation has been elaborated.[6] But, the application of g-prior of DC testing is completely new. It is general tendency to avoid the prior information about the relation between variable. The Bayesian gives the scope to consider the prior information of the relation between variables to explore the strength of current relation between variables. The application of Bayesian to compute DC is illustrated and Hypothesis test statistics through Bayes Factor is detailed on Biochemical marker for liver performance. The work is illustrated with the estimation of DC between AST and ALT. It is dedicated for Bayesian test to compute DC. The simple method proposed can be used by researchers exploring the use of DC in their research work. This work is not an attempt to develop a new statistical model. But it is an effort to explore the application of Bayesian approach to compute DC. The application is illustrated with biomarker of liver cirrhosis observed through clinical trial data analysis. Bayesian can be useful to get prominent evidence for test statistics on relation between variables. Bayes factor is useful for computation of DC. It is useful to figure out the strength of hypothesis. It can be considered as easily interpretable tool to discover the relations. This illustrated tool can be widely accepted for future research to explore relation between variables.

References

1. Székely GJ., Rizzo ML. Brownian Distance Covariance. The Annals of Applied Statistics 2009; 3: 1236-1265.
2. Sz´ekely GJ., Rizzo ML., Bakirov NK. Measuring and Testing Dependence by Correlation of Distances. The Annals of Statistics 2007; 35: 2769-2794.
3. Sz´ekely GJ., Rizzo ML. The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis 2013; 117: 193-213.
4. Gretton A., Fukumizu K., Sriperumbudur BK. Discussion of: Brownian distance covariance. The annals of applied statistics 2009; 1285-1294.
5. Lyons R. Distance covariance in metric spaces. Ann Probab 2013; 41: 3284-3305.

6. Bhattacharjee A. Distance correlation coefficient: An application with bayesian approach in clinical data analysis., Journal of Modern Applied Statistical Methods 2014; 13: 23.
7. Blum JR., Kiefer J., Rosenblatt M. Distribution free tests of independence based on the sample distribution function. The annals of mathematical statistics 1961; 32: 485-498.
8. Casella G., Berger RL. Statistical inference. Duxbury Pacific Grove., CA 2002.
9. George E., Foster DP. Calibration and empirical bayes variable selection. Biometrika 2000; 87: 731-747.
10. Liang F., Paulo R., Molina G., Clyde CA., Berger JO. Mixtures of g Priors for Bayesian Variable Selection. Journal of the American Statistical Association 2008; 103: 410-423.
11. Anderson TW. An introduction to multivariate statistical analysis. New York: John Wiley and Sons, 1958; 374.
12. Berger JO., Pericchi LR. The intrinsic bayes factor for model selection and prediction., Journal of the American Statistical Association 1996; 91: 109-122.
13. REITMAN S., FRANKEL S. A colorimetric method for the determination of serum glutamic oxalacetic and glutamic pyruvic transaminases. Am J Clin Pathol 1957; 28: 56-63.
14. WROBLEWSKI F. The clinical significance of transaminase activities of serum. Am J Med 1959; 27: 911-923.
15. Kumar S., Amarapurkar A., Amarapurkar D. Serum aminotransferase levels in healthy population from western India. Indian J Med Res 2013; 138: 894-899.
16. Southworth H., Heffernan JE. Extreme value modelling of laboratory safety data from clinical studies. Pharmaceutical Statistics 2012; 11: 361-366.