Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2015 *Paper* 113

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

Atanu Bhattacharjee Dr.*

*Malabar Cancer Centre, atanustat@gmail.com

http://biostats.bepress.com/cobra/art113

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

Atanu Bhattacharjee Dr.

Abstract

Distance Correlation is another newer choice to compute the relation between variables. However, the Bayesian counterpart of Distance Correlation is not established. In this paper, Bayesian counterpart of Distance Correlation is pro- posed. Proposed method is illustrated with Liver Chirrhosis Marker data. The relevant studies information about relation between AST and ALT is used to formulate the prior information for Bayesian computation. The Distance Correlation between AST and ALT (both are liver performance marker) is computed with 0.44. The credible interval is observed with (0.41, 0.46).Bayesian counter- part to compute Distance correlation is simple and handy.

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

Atanu Bhattacharjee

Division of Clinical Research and Biostatistics

Malabar Cancer Centre, Thalassery, Kerala-670103, India

Abstract

Distance Correlation is another newer choice to compute the relation between variables. However, the Bayesian counterpart of Distance Correlation is not established. In this paper, Bayesian counterpart of Distance Correlation is proposed. Proposed method is illustrated with Liver Chirrhosis Marker data. The relevant studies information about relation between AST and ALT is used to formulate the prior information for Bayesian computation. The Distance Correlation between AST and ALT (both are liver performance marker) is computed with 0.44. The credible interval is observed with (0.41, 0.46).Bayesian counterpart to compute Distance correlation coefficient is simple and handy.

Keywords: `ICC`, `Cannonical Correlation`, `Credible Interval`, `Distance Covariance`, `Conjugate Prior`

Introduction

The dependence between random vectors can be measured by distance correlation(DC). It is equally appropriate for equal and unequal dimensional measurement [1, 2]. The range of DC is [0, 1]. It provides platform to measure with multivariate independence. It is a generalized form of Pearson correlation.It is found consistent for all dependent alternatives through finite second

Email address: `atanustat@gmail.com` (Atanu Bhattacharjee)

moments [3].The bias outcome of DC through different dimension also been tested[3]. The unbiased t-test is found suitable to test the independence nature for distance correlation.Pearson correlation and Spearman's correlation are the most explored measurement in medical research over the last century. Both are widely explored tool to explore relation between variable. The DC is extended for high dimensional data [4]. The application of distance correlation for functional data also been extended recently through Hilber space [5].However, the limitation of DC to be applied for high dimension data also being elaborated[6]. Recently, several new tools are available to the scientific community for more complex issue through Cannonical , Rank and Renyi correlation [4]. However, all of them having some advantages and limitations[7]. The joint independences of random variable can be explored through DC [2]. It is matrix inversion free. Dependences measurement between two random variables can be observed and tested through it [8]. In experimental study, relation between two variable of interest plays always important role. Medical practice is based on known information about relation between two variables. Preventive and curative measures of medical disciplines is stand on relative relation between variables.Particularly, it is essential in experimental and medical research as tool to explore complex relation between random variables. In this article, we first elaborate the DC. Next, we discuss Bayesian approach in general and then Bayesian approach to compute Distance Correlation. In this work the default Bayesian approach is presented to compute the DC. The method is illustrated with clinical trial example. The intention of this work is to present some handy tool to the researcher involved to explored the relation between variables.

Distance Covariance and Distance Correlation

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

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

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 vector X and Y are in R^p and R^q respectively. The hypothesis is $H_0 : f_{X,Y} = f_X f_Y$ and $H_1 : f_{X,Y} \neq f_X f_Y$ The distance variance is

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

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

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

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

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

Similarly it can be defined as

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

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

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

Similarly, B_{kL} 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 $R(X,Y) < |\rho(X,Y)|$ with equality when $\rho \pm 1$. The random variables X and Y are express as $A_i = X_i + \epsilon_i$ and $B_j = Y_j + \epsilon_j$ respectively. The error terms ϵ_i and ϵ_j are independent with

the variables $X_i$and Y_j . Let the relation between random functions A_i and B_j is irrelevant. But the relation between X_i and $Y_j$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 H_0. In contrast, Bayesian measures through posterior probability $p(H_0|X)$. Let the data follows normal distribution $(\theta, \sigma^2)$with null hypothesis $H_0 : \theta \leq 0$ and $H_1 : \theta > 0$. The frequency and robust Bayesian often coincide [9]. Let the marginal DC ρ is applied between $p(X) = 1-\Phi(X/\sigma)$ and $p(H_0|X)$.The DC should be greater than or equal to zero. Because $p(X)$ and $p(H_0|X)$ 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 $\delta(X)$ and square error loss is $r(\pi, \delta) = E[\delta(X) - \theta]^2$ and risk is $\delta_\pi(X) = E(\theta/X)$. The DC between θ and $\delta(X)$ is

$$\rho(\theta, \delta(X)) = \frac{var(\theta) + cov\{\theta, b(\theta)\}}{\sqrt{var(\theta)}\sqrt{var\{\theta + (\theta)\} + \tau(\pi, \delta) - E\{b^2(\theta)\}}} \tag{7}$$

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

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

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 $\frac{1}{(P(Information))}$ should be equal to 1 as the theory of total Bayes theorem. The relation between posterior and prior is

$$PosteriorProbability \propto Likelihood \times PriorProbability \tag{9}$$

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

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

Let the mean and variance of X and Y are $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ respectively. The $mean(z)$ is derived from

$$e^z = \frac{\mu_1\sigma_2^2}{\mu_2\sigma_1^2} \tag{11}$$

The term $R(X,Y)$ is defined by $tanh\epsilon$ and it is assumed $\epsilon \sim N(z, \frac{1}{n})$. The mathematical formulations are detailed in Fisher (1915). The hyperbolic transformation plays role to consider the conjugate prior with normal distributions. The posterior mean can be represented with

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

$$\sigma^2_{posterior} = \frac{1}{\eta_{prior}+\eta_{likelihood}} \tag{13}$$

The prior with the form

$$P(R(X,Y)) \propto (1-R(X,Y)^2)^c \tag{14}$$

The prior is dependent on the choice of c. The $c=0$ gives the $P(R(X,Y) \propto 1$ The specification of prior is important for testing the parameters in hypothesis H_0 and H_1. 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

$$Y = \alpha + \beta X + \epsilon \tag{15}$$

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

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

The information about β can be obtained through $\phi^{-1}(X^T X)^{-1}$. 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[10]. 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 [11] and applied for correlation coefficient [12]. The prior is like

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

$$p(\phi) \propto \frac{1}{\phi} \tag{18}$$

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

The above mentioned formula is also useful to calculate Bayes factor. The prior is applied as default prior for t-test [8]. 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 (M_1) and Model(M_0) by

$$M_1 : Y = \alpha + \beta X + \epsilon \tag{20}$$

$$M_0 : Y = \alpha + \epsilon \tag{21}$$

The model ((M_1)) states the presence of DC and absence of it by Model ((M_0)). Now, the Bayes Factor through JZS is defined [11] as,

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

$$BF_{10} = \frac{p(Y/M_1)}{p(Y/M_0)} \quad (23)$$

If the value of BF_{10} becomes more than 1, it state about presences of DC otherwise not.

Testing

Under the null hypothesis H_0, the model (M_0) is assumed and (M_1) for alternative one i.e H_1. The prior probability of null is assigned as $p(M_0)$ and alternative as $p(M_1)$. 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

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

The term$P(Y|M_1)$is the marginal likelihood of the data for alternative hypothesis. Further, the marginal likelihood is calculated as

$$p(Y|M_1) = \int_{\theta}^{\infty} p(Y|\theta, M_1)p(\theta|H_1)d\theta \quad (25)$$

Bayes Factor ([13]) is useful to compute the appearance of $P(M_1|Y)$ in comparison to $P(M_0|Y)$:

$$\frac{p(M_1|Y)}{p(M_0|Y)} = BF_{10} \times \frac{p(M_1)}{p(M_0)} \quad (26)$$

Illustrated Example

The importance of Aminotransferases to detect malfunction of live, heart, lungs, skeletal and brain [14]. The Aminotransferases can be separated into Alanine aminotransferase(ALT) and Aspartate aminotransferase(AST) [15]. Serum ALT is useful to detect various liver disease[16]. Recently, AST and ALT is found

as suitable marker for healthy Indian population for Liver Cirrhosis [17]. This study is devoted to explore the DC between AST and ALT measurements of the same individuals. The generated information between AST and ALT is used as prior information of sample size of 4917 individuals[17]. The raw data on AST and ALT of 606 individuals are detailed [18]. This work is devoted to illustrate the DC between AST and ALT observations of 606 individuals. 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 [17]. The measured distance correlation data is observed with error. Bayesian posterior estimate is computed for robust DC between ALT and AST by,

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

$$\mu_{Posterior} = 0.00018(4917 tanh^{-1} + 606 tanh^{-1} 0.80) \tag{28}$$

$$\mu_{posterior} = 0.44 \tag{29}$$

The confidence interval is

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

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 BF_{10} in equation (23). The BF_{10} is calculated with 8.3. It is the evidence in favor of M_1 in comparison to model M_0. The presence of DC is tested through g prior.

Discussion

Recently, the testing process to check the presences of DC has been attempted. 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 [7]. 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 Biochemcal 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 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.

Acknowledgement

I thank to Dr.Vijay M Patil, Department of Medical Oncology,Tata Memorial Hospital and Dr. Tapesh Bhattacharyya , Malabar Cancer Centre for their constructive suggestions for this manuscript.

Competing Interest

None Declared

References

[1] G. J. Székely, M. L. Rizzo, et al., Brownian distance covariance, The annals of applied statistics 3 (4) (2009) 1236–1265.

[2] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al., Measuring and testing dependence by correlation of distances, The Annals of Statistics 35 (6) (2007) 2769–2794.

[3] G. J. Székely, M. L. Rizzo, The distance correlation t-test of independence in high dimension, Journal of Multivariate Analysis 117 (2013) 193–213.

[4] A. Gretton, K. Fukumizu, B. K. Sriperumbudur, Discussion of: Brownian distance covariance, The annals of applied statistics (2009) 1285–1294.

[5] R. Lyons, et al., Distance covariance in metric spaces, The Annals of Probability 41 (5) (2013) 3284–3305.

[6] J. R. Schott, Testing for complete independence in high dimensions, Biometrika 92 (4) (2005) 951–956.

[7] A. Bhattacharjee, Distance correlation coefficient: An application with bayesian approach in clinical data analysis, Journal of Modern Applied Statistical Methods 13 (1) (2014) 23.

[8] J. Blum, J. Kiefer, M. Rosenblatt, Distribution free tests of independence based on the sample distribution function, The annals of mathematical statistics (1961) 485–498.

[9] R. A. Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, Biometrika (1915) 507–521.

[10] E. George, D. P. Foster, Calibration and empirical bayes variable selection, Biometrika 87 (4) (2000) 731–747.

[11] F. Liang, R. Paulo, G. Molina, M. A. Clyde, J. O. Berger, Mixtures of g priors for bayesian variable selection, Journal of the American Statistical Association 103 (481).

[12] T. W. Anderson, An introduction to multivariate statistical analysis.

[13] J. O. Berger, L. R. Pericchi, The intrinsic bayes factor for model selection and prediction, Journal of the American Statistical Association 91 (433) (1996) 109–122.

[14] S. Reitman, S. Frankel, A colorimetric method for the determination of serum glutamic oxalacetic and glutamic pyruvic transaminases., American journal of clinical pathology 28 (1) (1957) 56–63.

[15] F. Wróblewski, The clinical significance of transaminase activities of serum, The American journal of medicine 27 (6) (1959) 911–923.

[16] D. Prati, E. Taioli, A. Zanella, E. Della Torre, S. Butelli, E. Del Vecchio, L. Vianello, F. Zanuso, F. Mozzi, S. Milani, et al., Updated definitions of healthy ranges for serum alanine aminotransferase levels, Annals of internal medicine 137 (1) (2002) 1–10.

[17] S. Kumar, A. Amarapurkar, D. Amarapurkar, Serum aminotransferase levels in healthy population from western india, The Indian journal of medical research 138 (6) (2013) 894.

[18] H. Southworth, J. E. Heffernan, Extreme value modelling of laboratory safety data from clinical studies, Pharmaceutical statistics 11 (5) (2012) 361–366.