A double robust test of conditional independence

Xiaogang Duan[1] & Jing Qin[2]

[1]School of Mathematical Sciences, Laboratory of Mathematics and Complex Systems of Ministry of Education, Beijing Normal University Beijing 100875, P.R.China.

[2]National Institute of Allergy and Infectious Diseases, National Institute of Health, USA.

Abstract

Conditional independence assumption is popular in many fields as long as it makes sense intuitively. However, there are few methods available for testing its validity in the statistical literature, especially when the underlying variables are continuous. In this paper we propose a supremum-type statistic for testing the conditional independence assumption, based on partial sums of the product of respective residuals. The proposed test is doubly robust in the sense that it produces valid result if either one of the conditional mean models is correctly specified. To implement our test, we present a resampling procedure for calculating the p-values. Simulation studies suggest that the new method is very competitive in terms of controlling Type I error rate and power. An AIDS data set is used for illustration.

Keywords: Conditional association; Double robustness; Inflated Type I error; Resampling.

1. Introduction

We are considering the problem of testing the conditional independence between two variables Y_1 and Y_2 given a p-vector of covariates $X = (X_1, \ldots, X_p)$ based on an independent and identically sample $\{(Y_{1i}, Y_{2i}, X_i) : i = 1, \ldots, n\}$ from the population of (Y_1, Y_2, X); that is, the null hypothesis we are interested in is

$$Y_1 \perp Y_2 \mid X, \tag{1}$$

where $Y_1 \perp Y_2 \mid X$ indicates that Y_1 and Y_2 are conditionally independent given X (Dawid, 1979). Assumption (1) is widely used in many fields; see for example Dawid (1979), Rubin and Rosenbaum (1983), Rosenbaum (1984), Prentice (1989).

When X is a categorical variable, Korn (1984) and Taylor (1987) developed a weighted sum of Kendall's tau over the categorical variable X. When X is a continuous variable, Goodman (1959) and Quade (1974) considered a partial version of Kendall's tau by using the number of local concordant and disconcordant pairs of observations. In other words, they compared values of Y_1 and Y_2 only for observations where X values are close. Unfortunately, it would be difficult to find two observations which are close to each other for continuous random variables. A partial rank correlation coefficient based on comparing pairs for which the values of the conditioning variable follow each other in a numerical ordering was studied by Gripenberg (1992). However, what exactly his statistic estimates is not clear. For continuous variables, there are also several nonparametric tests for conditional independence in the literature; see Su and White (2014) and references therein. However, kernel methods have to be used, which may not be practically convenient.

The conditional independence assumption is different from the familiar unconditional independence assumption $Y_1 \perp Y_2$. For the latter, there are many statistics available for testing its validity. Popular examples include Pearson's correlation, Spearman's rho and Kendall's tau. However, for testing conditional independence, the commonly used method appears to be the test based on Pearson's partial correlation, whose population version is

$$r_{12|\mathrm{X}} = \frac{r_{12} - r_{1\mathrm{X}}^\top R_X^{-1} r_{2\mathrm{X}}}{(1 - r_{1\mathrm{X}}^\top R_X^{-1} r_{1\mathrm{X}})^{1/2}(1 - r_{2\mathrm{X}}^\top R_X^{-1} r_{2\mathrm{X}})^{1/2}},$$

where r_{12} is the usual correlation between Y_1 and Y_2, $r_{1\mathrm{X}}$ is a column vector of length p with the j'th entry being Pearson' correlation between Y_1 and X_j $(j = 1, \ldots, p)$, $r_{2\mathrm{X}}$ is defined in a similar fashion, and R_X denotes the correlation matrix of X. In fact, $r_{12|\mathrm{X}}$ equals the correlation between errors in the linear regressions $y_1 = \beta^\top x + \epsilon_1$ and $y_2 = \gamma^\top x + \epsilon_2$. Kendall's partial tau and Spearman's partial rho can be defined analogously. It is well known that, the partial-type measures are not necessarily zero even (1) holds (Korn 1984). As a result, they are not valid for measuring conditional independence without further restrictive distributional assumptions. Moreover those tests may produce inflated Type I errors.

In this paper, we propose a new index for measuring the conditional association between Y_1 and Y_2 given X. Our index requires a model assumption of the underlying variables, but only through the respective marginal regression functions $E(Y_1|X)$ and $E(Y_2|X)$. Further, when (1) is true, the new index is exactly zero if either one of the two conditional mean models is correctly specified. This provides a double protection for the use of our method. A supremum-type test is constructed for testing the validity of the conditional independence assumption, based on a sample version of the proposed population index. The test statistic equals the supremum of a partial sum process defined by the product of the respective residuals. To implement our test, we present a resampling procedure for calculating the p-values. Numerical results indicate that the new test is well behaved for finite samples. An AIDS data set is used to illustrate our method.

2. Methodology

2.1 A double robust index

We first introduce a new index for measuring the conditional association between Y_1 and Y_2 given X. A test statistic based on this index is straightforward. Let $m_1(x)$ and $m_2(x)$ be two functions of x such that both $E\{m_1^2(X)\}$ and $E\{m_2^2(X)\}$ are finite. Define

$$G = \sup_{x \in R^p} |E[\{Y_1 - m_1(X)\}\{Y_2 - m_2(X)\}I(X \le x)]| \,. \tag{2}$$

where $I(\cdot)$ is the indicator function, and the event $\{X \le x\}$ indicates that all the components of X are less than or equal to those of x. The following proposition indicates that G is a doubly robust index for measuring the conditional association between Y_1 and Y_2 given X.

Proposition 1 Assume that either $m_1(x) = E(Y_1|x)$ or $m_2(x) = E(Y_2|x)$ holds. Then $G = 0$ if and only if the conditional correlation between Y_1 and Y_2 given X is zero almost surely with respect to the distribution of X.

2.2 Test statistic

To test the conditional independence assumption, we propose to use a sample version of the index G defined above. For this purpose, we need two working models for both $E(Y_1|x)$ and $E(Y_2|x)$. We assume that the conditional mean of Y_1 and Y_2 depends on X respectively by

$$E(Y_1|x) = \mu_1(z_1^\top \beta), \tag{3}$$

$$E(Y_2|x) = \mu_2(z_2^\top \gamma), \tag{4}$$

where μ_1 and μ_2 are known functions, $z_1 = z_1(x)$ and $z_2 = z_2(x)$ are fixed vector-valued transformations of x, β and γ are vectors of unknown regression parameters with respective dimension q_1 and q_2. Our formulation of the conditional mean model is very general and includes generalized linear model and polynomial regression as its special cases. For example, it accommodates the case that $z_1 = (1, x)^\top$ and $z_2 = (1, x, x^2)^\top$ for a scalar covariate x. We further assume that $\mathrm{var}(Y_1|x) = \xi_1(\mu_1)$ and $\mathrm{var}(Y_2|x) = \xi_2(\mu_2)$, where ξ_1 and ξ_2 are known functions.

Let $\hat{\beta}$ be the quasilikelihood estimator of β (Wedderburn, 1974), which is a solution to the estimating equation $\sum_{i=1}^n U_{1i}(\beta) = \sum_{i=1}^n \dot{\mu}_1(Z_{1i}^\top\beta)\xi_{1i}^{-1}(\beta)Z_{1i}\{Y_{1i} - \mu_1(Z_{1i}^\top\beta)\} = 0$, where $U_{1i}(\beta) = \dot{\mu}_1(Z_{1i}^\top\beta)\xi_{1i}^{-1}(\beta)Z_{1i}\{Y_{1i} - \mu_1(Z_{1i}^\top\beta)\}$ and $\xi_{1i}(\beta) = \xi_1\{\mu_1(Z_{1i}^\top\beta)\}$. Correspondingly, $\hat{\gamma}$ is a solution to $\sum_{i=1}^n U_{2i}(\gamma) = \sum_{i=1}^n \dot{\mu}_2(Z_{2i}^\top\gamma)\xi_{2i}^{-1}(\gamma)Z_{2i}\{Y_{2i} - \mu_2(Z_{2i}^\top\gamma)\} = 0$, with a similar definition of $U_{2i}(\gamma)$ and $\xi_{2i}(\gamma)$. Here and throughout, $\dot{a}(t)$ denotes the first order derivative for a real-valued function $a(t)$. Let β^* and γ^* be the respective probability limits of $\hat{\beta}$ and $\hat{\gamma}$. Under some mild regularity conditions (see for example page of Tsiatis, 2006), $n^{1/2}(\hat{\beta} - \beta^*)$ is asymptotically equivalent to $A_1^{-1}(\beta^*)S_{1n}(\beta^*)$ and $n^{1/2}(\hat{\gamma} - \gamma^*)$ is asymptotically equivalent to $A_2^{-1}(\gamma^*)S_{2n}(\gamma^*)$, where $A_1 = -\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \dot{U}_{1i}(\beta)$, $A_2 = -\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \dot{U}_{2i}(\gamma)$, $S_{1n}(\beta^*) = n^{-1/2}\sum_{i=1}^n U_{1i}(\beta^*)$ and $S_{2n}(\gamma^*) = n^{-1/2}\sum_{i=1}^n U_{2i}(\gamma^*)$.

Consider the following statistic,

$$W_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_{1i} - \mu_1(Z_{1i}^\top\hat{\beta})\}\{Y_{2i} - \mu_2(Z_{2i}^\top\hat{\gamma})\}I(X_i \le t),$$

where $t = (t_1, \cdots, t_p)^\top \in R^p$, $I(\cdot)$ is the indicator function, and the event $\{x \le t\}$ indicates that all the components of x are less than or equal to those of t. Consequently, $W_n(t)$ is a multiparameter stochastic process. The proposed test statistic is

$$G_n = \sup_{t\in R^p} |W_n(t)|\,. \tag{5}$$

By a similar argument to Su and Wei (1991), we can show that G_n converges to G in probability as n goes to infinity.

To approximate the large sample null distribution of G_n, we will adopt a resampling technique similar to Su and Wei (1991) and Lin, Wei and Ying (2002). To this end, we need to find an asymptotically equivalent representation of $W_n(t)$ for each fixed t. Under some mild regularity conditions (see, for example, van der vaart 2000, chapter 5), $W_n(t)$ is asymptotically equivalent to $W_{0n}(t) + W_{1n}(t) + W_{2n}(t)$, where $W_{0n}(t) = n^{-1/2}\sum_{i=1}^n e_{1i}(\beta^*)e_{2i}(\gamma^*)I(X_i \le t)$,

$W_{1n}(t) = \eta_{1n}(t;\beta^*,\gamma^*)A_2^{-1}(\gamma^*)S_{2n}(\gamma^*)$ and $W_{2n}(t) = \eta_{2n}(t;\beta^*,\gamma^*)A_1^{-1}(\beta^*)S_{1n}(\beta^*)$, with

$$\eta_{1n}(t;\beta^*,\gamma^*) = -\frac{1}{n}\sum_{i=1}^{n} e_{1i}(\beta^*)\dot{\mu}_2(Z_{2i}^\top\gamma^*)Z_{2i}^\top I(X_i \le t),$$

$$\eta_{2n}(t;\beta^*,\gamma^*) = -\frac{1}{n}\sum_{i=1}^{n} e_{2i}(\gamma^*)\dot{\mu}_1(Z_{1i}^\top\beta^*)Z_{1i}^\top I(X_i \le t),$$

and $e_{1i}(\beta^*) = Y_{1i} - \mu_1(Z_{1i}^\top\beta^*)$ and $e_{2i}(\gamma^*) = Y_{2i} - \mu_2(Z_{2i}^\top\gamma^*)$ for $i = 1,\cdots,n$. Following Su and Wei (1991) and Lin, Wei and Ying (2002), the asymptotic null distribution of $W_n(t)$ can be approximated by that of $\widehat{W}_n(t) = \widehat{W}_{0n}(t) + \widehat{W}_{1n}(t) + \widehat{W}_{2n}(t)$. Here,

$$\widehat{W}_{0n}(t) = n^{-1/2}\sum_{i=1}^{n} V_i\ e_{1i}(\hat{\beta})e_{2i}(\hat{\gamma})I(X_i \le t),$$

$$\widehat{W}_{1n}(t) = \eta_{1n}(t;\hat{\beta},\hat{\gamma})A_{2n}^{-1}(\hat{\gamma})\hat{S}_{2n}(\hat{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n} e_{1i}(\hat{\beta})\dot{\mu}_2(Z_{2i}^\top\hat{\gamma})\{Z_{2i}^\top A_{2n}^{-1}(\hat{\gamma})\hat{S}_{2n}(\hat{\gamma})\}I(X_i \le t),$$

$$\widehat{W}_{2n}(t) = \eta_{2n}(t;\hat{\beta},\hat{\gamma})A_{1n}^{-1}(\hat{\beta})\hat{S}_{1n}(\hat{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} e_{2i}(\hat{\gamma})\dot{\mu}_1(Z_{1i}^\top\hat{\beta})\{Z_{1i}^\top A_{1n}^{-1}(\hat{\beta})\hat{S}_{1n}(\hat{\beta})\}I(X_i \le t),$$

where $A_{1n}(\hat{\beta}) = -n^{-1}\sum_{i=1}^{n}\dot{U}_{1i}(\hat{\beta})$, $A_{2n}(\hat{\gamma}) = -n^{-1}\sum_{i=1}^{n}\dot{U}_{2i}(\hat{\gamma})$, $\hat{S}_{1n}(\hat{\beta}) = n^{-1/2}\sum_{i=1}^{n} V_iU_{1i}(\hat{\beta})$, $\hat{S}_{2n}(\hat{\gamma}) = n^{-1/2}\sum_{i=1}^{n} V_iU_{2i}(\hat{\gamma})$; and $\{V_1,\cdots,V_n\}$ are random sample from standard normal distribution and are independent of $\{(Y_{11},Y_{21},X_1),\cdots,(Y_{1n},Y_{2n},X_n)\}$. It is worth noting that $\widehat{W}_{1n}(t)$ and $\widehat{W}_{2n}(t)$ depend on the resampling through $\hat{S}_{1n}(\hat{\beta})$ and $\hat{S}_{2n}(\hat{\gamma})$. Further, the above formulation facilitates numerical computation. Suppose that g_n is the observed value of G_n. We can compute the p-value of the test statistic by repeatedly generating random samples from $N(0,1)$.

3. Simulation studies

In this section, we carried out simulations to investigate the finite sample performance of new test statistic and to compare it, in terms of Type I error rate and power, with Pearson's partial correlation, Kendall's partial tau and Spearman's partial rho, as described in the introductory section. All the simulations are conducted using the R language (R Core Team 2013).

We considered two model setups. For all our simulations, we used 5000 replications to generate the desired results. In each replication, we computed the p-values using the resampling method described in the last section, with resampling size equal to 1000. For the partial procedures, the p-values were computed using R's package ppcor (Kim 2012).

For the first model, the true data generating process is as follows. We first generated a scalar X from $U(-1,1)$; given $X = x$, Y_1 and Y_2 were generated respectively through $Y_1 = \beta_0 + \beta_1 x + \beta_2 x^2 + \sigma_1|x|^{\lambda/2}\epsilon_1 + \eta e$ and $Y_2 = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \sigma_2|x|^{\lambda/2}\epsilon_2 + \eta e$. Here, $\beta_0 = \beta_1 = \beta_2 = 1$ and $\gamma_0 = \gamma_1 = \gamma_2 = 1$ are true values of the regression parameters, (σ_1,σ_2,η) is a vector of scale parameters, $(\epsilon_1,\epsilon_2,e) \overset{\text{iid}}{\sim} N(0,1)$, and λ is a type of tuning parameter which controls the

heteroscedasticity of the respective models. For this model, the conditional correlation between Y_1 and Y_2 given $X = x$ is $\eta^2\{(\sigma_1^2|x|^\lambda + \eta^2)(\sigma_2^2|x|^\lambda + \eta^2)\}^{-1/2}$. We considered four parameter combinations of $(\lambda, \sigma_1, \sigma_2)$, which are $(0, 1, 1)$, $(0, 0.5, 1)$, $(2, 1, 1)$ and $(2, 0.5, 1)$ respectively.

To implement our test procedure, it is required to assume a model for both $E(Y_1|x)$ and $E(Y_2|x)$. We consider two types of working model specification. The proposed test statistics associated are denoted as G_n^1 and G_n^2 respectively. For the first specification, the assumed model for $E(Y_1|x)$ is $\beta_0 + \beta_1 x + \beta_2 x^2$ and so is correctly specified; but the assumed model for $E(Y_2|x)$ is incorrectly specified as $\gamma_0 + \gamma_1 x$. For the second specification, both mean models are correctly specified. For both specifications, the conditional variances are assumed to be constant. As a result, the variance functions are correctly specified when $\lambda = 0$ but incorrectly specified when $\lambda \neq 0$ in the true data generating process.

Table 1 summarizes the empirical sizes for the different procedures we investigated. In the table, the column labeled G_n^1 corresponds to results of proposed test statistic with the first type model specification, that is, with correct model for $E(Y_1|x)$ but incorrect model for $E(Y_2|x)$; the column labeled G_n^2 corresponds to results with the second type of model specification, that is, correct models for both $E(Y_1|x)$ and $E(Y_2|x)$. The last three columns of the table are results of the three partial test statistics.

Columns G_n^1 and G_n^2 of table 1 suggest that the empirical size of the proposed procedure is quite close to its nominal counterparts as long as one of the working models for the conditional mean $E(Y_1|x)$ and $E(Y_2|x)$ is correctly specified. This is true even when we incorrectly model the conditional variance functions for both $\text{var}(Y_1|x)$ and $\text{var}(Y_2|x)$, although the proposed test performs a little bit worse for this situation than it were when $\lambda = 0$. For example, the first two blocks of table 1 corresponding to $\lambda = 0$, seems better than the last two blocks which correspond to results when we use constant variance function for estimating β and γ while actually they are linear functions of $|x|$. On the contrary, it is seen from the last three columns of table 1 that Pearson's partial correlation, Kendall's partial tau and Spearman's partial rho all produce inflated Type I errors. The discrepancy between the true and stated significance levels is particularly serious either when the constant error variance in the true data generating process turns small with $\lambda = 0$ but σ_1 reducing from 1 to 0.5, or the true conditional variance of Y_1 and Y_2 depends on the conditional variable x, corresponding to $\lambda = 2$. For example, at the commonly referred stated significance level 0.05, the false positive rate of the test based on Pearson's partial correlation is bigger than 0.54 when $\lambda = 2$. In other words, it nearly falsely rejects half the null when actually it is true.

To reinforce previous findings, we display in Figure 1 the histograms of the empirical p values for the different procedures. It is seen that the empirical distribution of the p values associated with the proposed test is quite close to distribution of a $U(0, 1)$ variable under all four parameter combinations we examined, while those for partial tests are seriously right skewed. To be more intuitive, we also display in table 2 the basic mean and standard error summaries of the empirical p-values of the different procedures when the null hypothesis is true. The empirical numerical characteristics of the proposed test are consistent with the population mean and standard deviation of a $U(0, 1)$ variable, although the consistency is less satisfactory when the true model is heteroscedastic.

Due to the substantial difference of the actual Type I errors, it is less appropriate to compare the power directly of the different procedures. Instead, we compare them in terms of the adjusted power,

which is computed as follows. For a specified procedure with test statistic T say, we first generated 5000 samples independently under the null hypothesis. Based on the ith sample, we computed the test statistic $T_i^{(0)}$. We worked out the $(1-\alpha)$'th quantile of the sample $\{T_1^{(0)}, \cdots, T_{5000}^{(0)}\}$, and denoted as $\hat{C}_\alpha$. We then generated another 5000 independent samples under a specified alternative, and computed the associated test statistic $T_i^{(1)}$ for $i = 1, \cdots, 5000$. The resulting adjusted power is $\frac{1}{5000}\sum_{i=1}^{5000} I(T_i^{(1)} \geq \hat{C}_\alpha)$, where $I(A)$ is an indicator function taking value 1 if A is true and 0 otherwise. The adjusted power curves of the different procedures are shown in figure 2. It appears that the proposed test is very competitive compared with the partial based tests, as long as one conditional mean models is correctly specified. Also, it seems that the gain from additionally assuming a correct conditional mean model is very limited for the specific true model we considered.

Our second model setup is a partial repetition of a model considered in Korn (1984). Specifically, we first generated a scalar covariate X from $U(0,1)$; given $X = x$, Y_1 and Y_2 were generated through the same way as our first model setup except changing the conditional mean structure as $2 - 2I(x < 1/2) - 2x + 4xI(x < 1/2)$. In particular, the current model with $\lambda = 0$ and $\eta = 0$ reduces to the model considered by Korn (1984), who has designed the original model to examine the possible ranges of the limiting values of the three partial association indices under conditional independence assumption. The corresponding results for this model are displayed in table 3, 4 and figure 4 and 3; they are quite consistent with results from the first model, and so summary of them is omitted here.

4. Application to an AIDS data

We now apply the proposed method to data from the AIDS Clinical Trials Group protocol 175 (ACTG175). ACTG175 was a randomized clinical trial to compare monotherapy with zidovudine or didanosine with combined therapy with zidovudine and didanosine or zidovudine and zalcitabine in adults infected with the human immunodeficiency virus type I whose CD4 T cell counts were between 200 and 500 per cubic millimeter; see Hammer et al. (1996) for details.

It is of interest here to test whether Y_1, the CD4 count at 96±5 weeks, is independent of Y_2, the CD4 count at 20±5 weeks, given X_1, the baseline CD4 count, for the different subpopulation defined by gender of patients and arms of treatment a patient belonging to. In other words we are interested in testing how does a patient' conditions of the baseline CD4s determine his/her future response. Totally, there are 1336 subjects with complete observations on (Y_1, Y_2, X_1).

We applied the proposed method to each of the eight samples defined by gender and arms of treatment. All the regression models involved were assumed to be linear in X_1, and the regression parameters were estimated by the least square methods. We computed the p-values of the proposed test by the resampling method introduced in section 2. The resampling size was 1000.

Table 5 summarizes the results. The p-values of the table indicate that, there is a strong evidence for the existence of association between CD4 count at 96±5 weeks and CD4 count at 20±5 weeks conditional on the CD4 count at baseline, whatever the gender of a patient and whatever treatment group a patient belongs to. Although it is relatively less stronger for female patients than for male patients, this is very likely due to the lower sample size of the different female subgroups. In fact, if we combined the different treatment groups of female patients with a total sample size 218, the

resulting p-value is less than 10^{-4}, implying an association between Y_1 and Y_2 given X_1 for female patients.

5. Concluding remarks

We have proposed a double robust procedure for testing the conditional independence assumption. Simulation studies suggest that the new test performs well for finite samples with a variety of data generating processes. Although we formulate the problem with a scalar Y_1 and Y_2, the proposed method is clearly applicable when Y_1 and Y_2 are both random vectors with fixed dimension q_1 and q_2, respectively. Under these circumstances, the proposed test statistic is $G_n^0 = \max_{\{k=1,\cdots,q_1,l=1,\cdots,q_2\}} G_{n,kl}$, where

$$G_{n,kl} = \sup_{x\in R^p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_{1i} - \mu_1(Z_{1i}^\top \hat{\beta})\}_{(k)} \{Y_{2i} - \mu_2(Z_{2i}^\top \hat{\gamma})\}_{(l)} I(X_i \le t) \right|,$$

where $\hat{\beta}$ and $\hat{\gamma}$ are GEE estimators of the β and γ, and $a_{(k)}$ denotes the kth component of a vector a.

One potential issue about the proposed test is that it possibly converges to zero even when the null hypothesis is not true; for example, when Y_1 and Y_2 given X is not conditionally independent but conditionally uncorrelated everywhere. In other words, the proposed test lose power under this circumstance. A possible solution is to improve the index G in (2) to be $G_0 = \sup_{y_1,y_2} G(y_1, y_2)$, where

$$G(y_1, y_2) = \sup_x \left| E[\{I(Y_1 \le y_1) - m_{1y_1}(X)\}\{I(Y_2 \le y_2) - m_{2y_2}(X)\} I(X \le x)] \right|.$$

Then it can be shown that $G_0 = 0$ is equivalent to the $Y_1 \perp Y_2 \mid X$, as long as one of $m_{1y_1}(x) = F_1(y_1|x)$ or $m_{2y_2}(x) = F_2(y_1|x)$ is correctly specified. Then one can test assumption (1) based on a sample version of G_0.

Anothe issue is that the proposed test becomes computationally infeasible when the dimension of X is relatively high, since the number of function evaluations for computing G_n is an exponential function of p. This is a common problem for high dimensional data analysis. To relieve this disaster, a possible solution is to improve our test statistic G_n based on the technique of principle component analysis. In particular, let Σ_n be the sample variance-covariance matrix of $X_1, \ldots, X_n$, and let $\nu_1, \ldots, \nu_p$ be the eigenvectors of Σ_n with corresponding eigenvalues $\tau_1, \ldots, \tau_p$. For this situation, we test (1) based on the following test statistic

$$G_n^* = \sup_{t\in R^p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_{1i} - \mu_1(Z_{1i}^\top \hat{\beta})\}\{Y_{2i} - \mu_2(Z_{2i}^\top \hat{\gamma})\} I(X_i^* \le t) \right|,$$

where $X_i^* = (X_{1i}^*, \ldots, X_{qi}^*)^\top$ is the first q principle components of X_i with $X_{ji}^* = \nu_j^\top X_i$ for $j = 1, \ldots, q$; q is usually small compared to p. However, G_n^* may be powerful only in limited cases. It is very interesting in future work to develop a powerful test procedure for testing assumption (1) without too much dependence on the underlying variable dimension.

References

[1] Dawid, A. P. (1979), Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1-31.

[2] Gripenberg, G. (1992). Confidence intervals for partial rank correlations. *Journal of the American Statistical Association*, **87**, 546-551.

[3] Goodman, L. A. (1959). Partial tests for partial tau. *Biometrika*, **46**, 425-432.

[4] Hammer S.M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, **335**, 1081-1090.

[5] Lin, D.Y., Wei, L.J. and Ying, Z. (2002). Model-Checking Techniques Based on Cumulative Residuals. *Biometrics*, **58**, 1-12.

[6] Kendall, M. G. (1942). Partial rank correlation. *Biometrika*, **32**, 277-283.

[7] Kim, S. (2012). ppcor: Partial and Semi-partial (Part) correlation. R package version 1.0. http://CRAN.R-project.org/package=ppcor.

[8] Korn, E. L. (1984). The ranges of limiting values of some partial correlations under conditional independence. *The American Statistician*, **38**, 61-62.

[9] Korn, E. L. (1984). Kendall's tau with a blocking variable. *Biometrics*, **40**, 209-214.

[10] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[11] Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, **8**, 431-440.

[12] Quade, D. (1974). Nonparametric partial correlation. in H. M. Blalock, Jr. (ed.), *Measurement in the Social Science*, 369-398. Chicago, Aldine.

[13] Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, **49**, 425-435.

[14] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.

[15] Su, J.Q. and Wei, L.J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, **86**, 420-426.

[16] Su, L. and White, H. (2014). Testing Conditional Independence via Empirical Likelihood. Forthcoming.

[17] Taylor, J. M. G. (1987). Kendall's and Spearman's correlation coefficient in the presence of a blocking variable. *Biometrics*, **43**, 409-415.

[18] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. Springer-Verlag, New York.

[19] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

[20] Wedderburn, R.W.M. (1974).Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

Table 1: Empirical sizes of different procedures for testing conditional independence based on data generated from model 1. The sample size is $n = 100$; the replication times is 5000; the resampling size is 1000.

Nominal level	G_n^1	G_n^2	Pearson	Kendall	Spearman
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (1, 1)$			
0.01	0.0084	0.0084	0.0408	0.0356	0.0896
0.05	0.0508	0.0514	0.1242	0.1190	0.2508
0.10	0.1098	0.1086	0.2068	0.1902	0.3780
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
0.01	0.0116	0.0116	0.1448	0.0892	0.2422
0.05	0.0504	0.0506	0.3186	0.2264	0.5000
0.10	0.1048	0.1034	0.4254	0.3272	0.6400
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (1, 1)$			
0.01	0.0086	0.0072	0.3826	0.3496	0.5332
0.05	0.0554	0.0564	0.5682	0.5422	0.7446
0.10	0.1158	0.1156	0.6614	0.6426	0.8378
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
0.01	0.0074	0.0062	0.7650	0.6078	0.7876
0.05	0.0474	0.0496	0.8824	0.7664	0.9272
0.10	0.1130	0.1128	0.9256	0.8368	0.9624

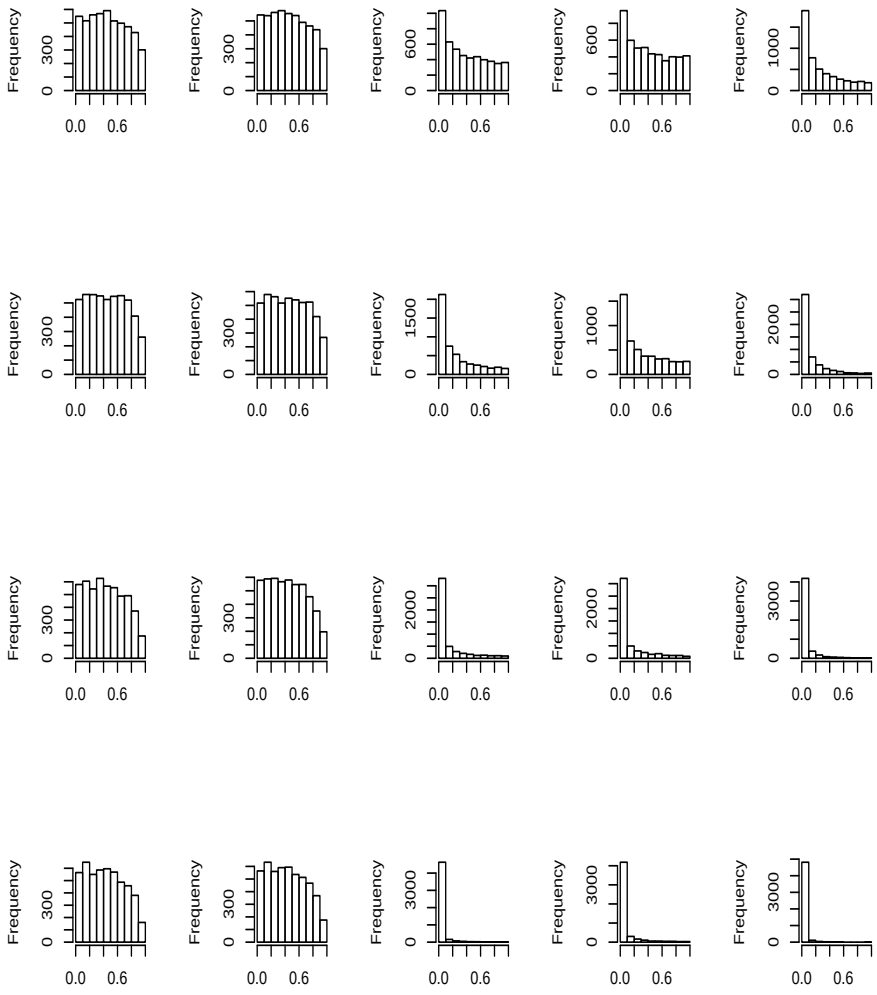

Figure 1: Histograms of the empirical p values for the different testing procedures when data are generated through model 1. The rows correspond in turn to values of the parameter combination $(\lambda, \sigma_1, \sigma_2)$: $(0, 1, 1)$, $(0, 0.5, 1)$, $(2, 1, 1)$, $(2, 0.5, 1)$; the columns correspond in turn to testing procedures: proposed test with correct μ_1 but incorrect μ_2, proposed test with both correct μ_1 and μ_2, Pearson' partial correlation, Kendall's partial tau, Spearman's partial rho.

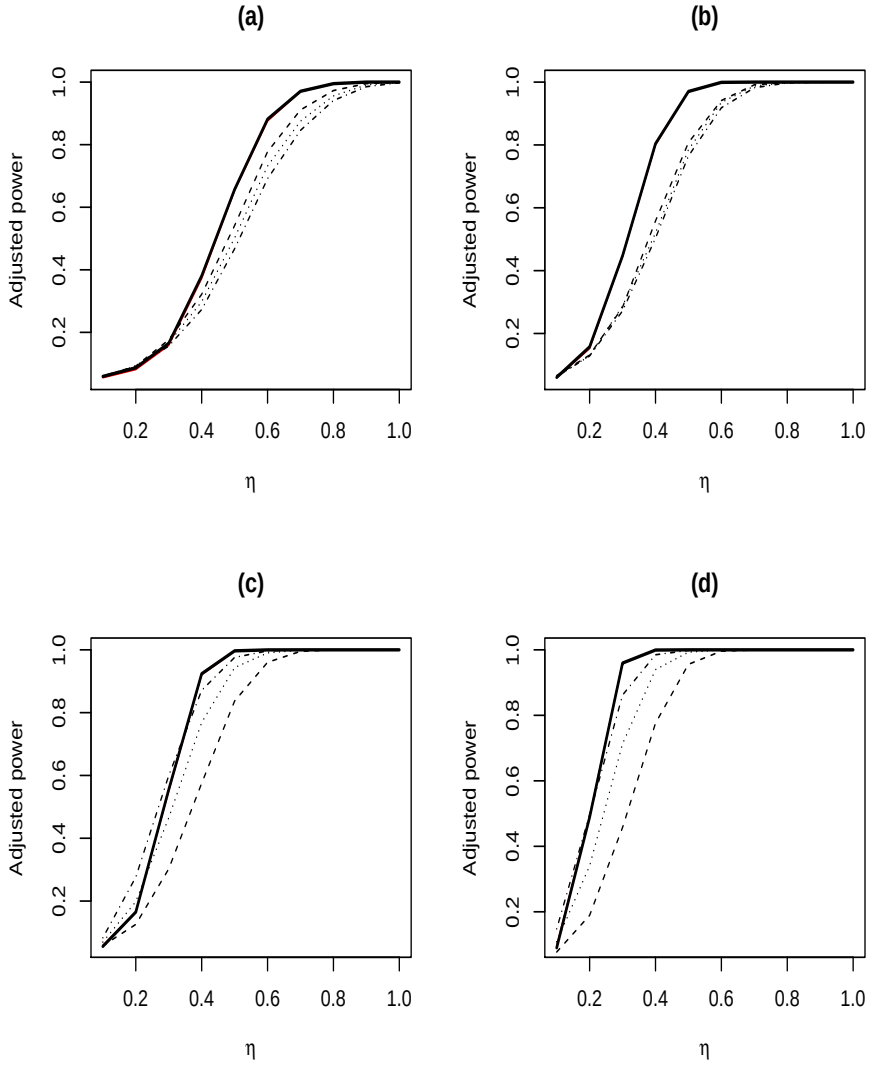

Figure 2: Adjusted powers of the different testing procedures for model 1: red solid line for G_n^1, black solid thick line for G_n^2, dashed line for Pearson, dotted line for Kendall, dotdash line for Spearman; (a) $\lambda = 0, \sigma_1 = 1, \sigma_2 = 1$, (b) $\lambda = 0, \sigma_1 = 0.5, \sigma_2 = 1$, (c) $\lambda = 2, \sigma_1 = 1, \sigma_2 = 1$, (d) $\lambda = 2, \sigma_1 = 0.5, \sigma_2 = 1$.

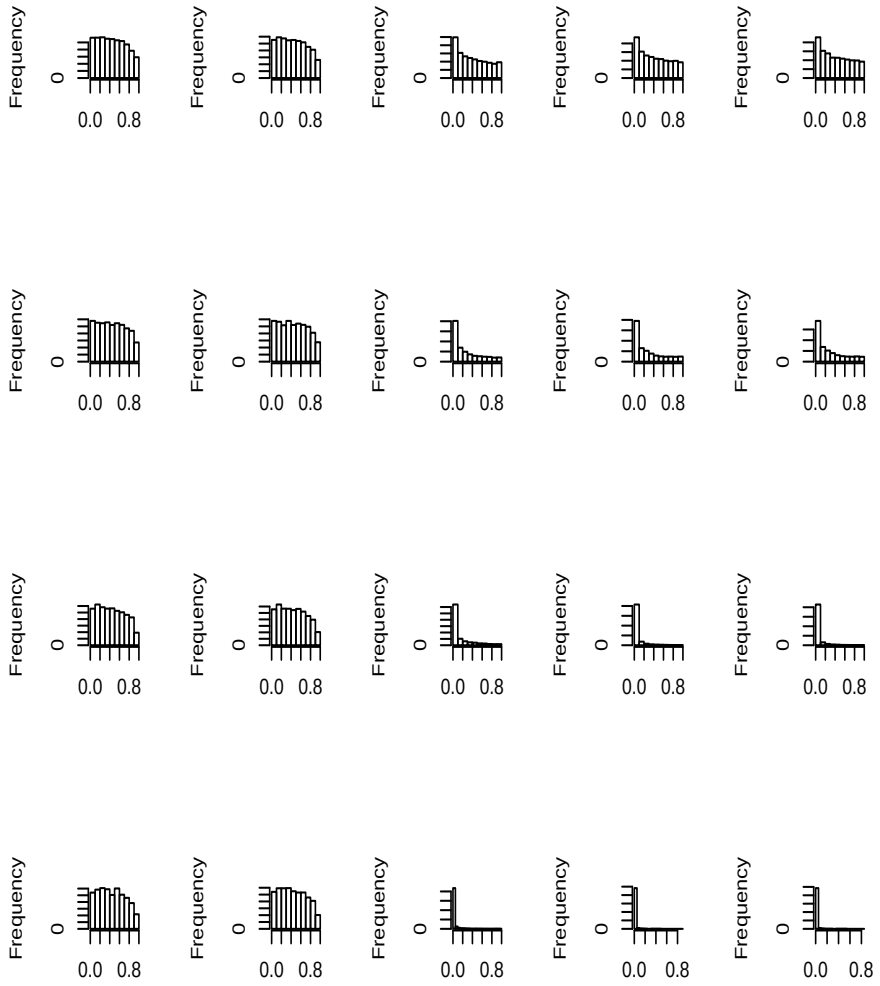

Figure 3: Histograms of the empirical p values for the different testing procedures when data are generated through model 2. The rows correspond in turn to values of the parameter combination $(\lambda, \sigma_1, \sigma_2)$: $(0, 1, 1)$, $(0, 0.5, 1)$, $(2, 1, 1)$, $(2, 0.5, 1)$; the columns correspond in turn to testing procedures: proposed test with correct μ_1 but incorrect μ_2, proposed test with both correct μ_1 and μ_2, Pearson' partial correlation, Kendall's partial tau, Spearman's partial rho.

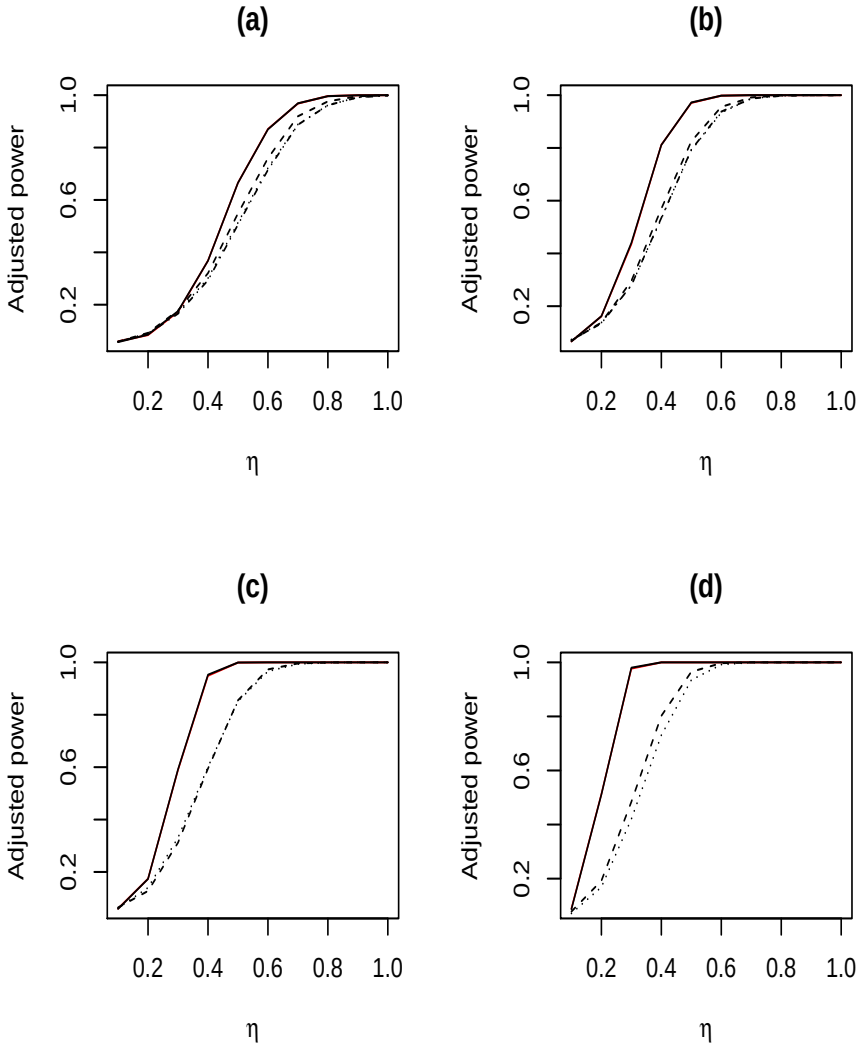

Figure 4: Adjusted powers of the different testing procedures for model 2: red solid line for G_n^1, black solid thick line for G_n^2, dashed line for Pearson, dotted line for Kendall, dotdash line for Spearman; (a) $\lambda = 0, \sigma_1 = 1, \sigma_2 = 1$, (b) $\lambda = 0, \sigma_1 = 0.5, \sigma_2 = 1$, (c) $\lambda = 2, \sigma_1 = 1, \sigma_2 = 1$, (d) $\lambda = 2, \sigma_1 = 0.5, \sigma_2 = 1$.

Table 2: Mean and standard errors of the empirical p values based on data generated from model 1; corresponding values of $U(0,1)$ are 0.5 and 0.2887, respectively. The sample size is $n = 100$; the replication times is 5000; the resampling size is 1000.

	G_n^1	G_n^2	Pearson	Kendall	Spearman
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (1, 1)$			
mean	0.4644	0.4635	0.4084	0.4254	0.2801
se	0.2732	0.2740	0.3007	0.3031	0.2791
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
mean	0.4635	0.4635	0.2516	0.3272	0.1332
se	0.2709	0.2712	0.2729	0.2982	0.1923
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (1, 1)$			
mean	0.4404	0.4412	0.1512	0.1590	0.0606
se	0.2636	0.2635	0.2397	0.2408	0.1314
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
mean	0.4368	0.4387	0.0302	0.0666	0.0153
se	0.2619	0.2624	0.1022	0.1579	0.0505

Table 3: Empirical sizes of different procedures for testing conditional independence based on data generated from model 2. The sample size is $n = 100$; the replication times is 5000; the resampling size is 1000.

Nominal level	G_n^1	G_n^2	Pearson	Kendall	Spearman
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (1, 1)$			
0.01	0.0096	0.0102	0.0396	0.0362	0.0306
0.05	0.0506	0.0548	0.1254	0.1176	0.1096
0.10	0.1132	0.1106	0.1988	0.1884	0.1828
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
0.01	0.0106	0.0118	0.1150	0.1164	0.1008
0.05	0.0568	0.0588	0.2848	0.2762	0.2610
0.10	0.1156	0.1156	0.4034	0.3904	0.3800
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (1, 1)$			
0.01	0.0076	0.0086	0.3420	0.5332	0.5764
0.05	0.0526	0.0498	0.5394	0.7498	0.7820
0.10	0.1118	0.1112	0.6398	0.8306	0.8580
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
0.01	0.0060	0.0062	0.7352	0.8982	0.9118
0.05	0.0454	0.0476	0.8718	0.9660	0.9712
0.10	0.1080	0.1080	0.9222	0.9830	0.9860

Table 4: Mean and standard errors of the empirical p values based on data generated from model 2; corresponding values of $U(0,1)$ are 0.5 and 0.2887, respectively. The sample size is $n = 100$; the replication times is 5000; the resampling size is 1000.

	G_n^1	G_n^2	Pearson	Kendall	Spearman
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (1, 1)$			
mean	0.4566	0.4538	0.4140	0.4216	0.4247
se	0.2734	0.2722	0.3008	0.2997	0.2989
		$\lambda = 0, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
mean	0.4601	0.4592	0.2787	0.2908	0.2923
se	0.2743	0.2737	0.2871	0.2940	0.2917
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (1, 1)$			
mean	0.4449	0.4454	0.1557	0.0651	0.0558
se	0.2674	0.2665	0.2337	0.1456	0.1343
		$\lambda = 2, \quad (\sigma_1, \sigma_2) = (0.5, 1)$			
mean	0.4480	0.4472	0.0312	0.0072	0.0061
se	0.2651	0.2656	0.0987	0.0362	0.0324

Table 5: The p-values of the proposed test when applied to different samples defined by gender and arms of treatment of patients. For treatment arms, 0=zidovudine, 1=zidovudine and didanosine, 2=zidovudine and zalcitabine, 3=didanosine.

gender	arms	number of patients	p-values
m	0	263	<0.000
m	1	276	<0.000
m	2	285	<0.000
m	3	300	<0.000
f	0	58	0.005
f	1	57	0.002
f	2	52	0.004
f	3	51	0.009