行政院國家科學委員會專題研究計畫 成果報告

條件相關性度量及條件獨立檢定
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2119-M-004-001-
執 行 期 間 ：95 年 10 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人：黃子銘

計畫參與人員：碩士班研究生-兼任助理：歐陽致平

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 96 年 10 月 22 日

1 Introduction

Conditional independence tests have different applications. For example, to make variable selection in a regression model, Li, Cook and Nachtsheim (2005) proposed to test the independence between the response variable and a predictor variable given other predictors, and remove the predictor variable when the test is not significant. Su and White (2005) pointed out that conditional independence tests can be used for testing Granger non-causality for two time series and choosing a proper model for a certain family of semi-parametric models.

For the case where the variables involved are discrete, there are many existing tests for conditional independence. For the case with continuous variables, there are relatively few results. Li et al (2005) proposed a test of conditional independence, which is constructed by projecting two variables to the space generated by the conditioned variable and then testing the independence between the residuals. Su and White (2005, 2006) proposed tests based on a weighted Hellinger distance between the conditional densities or based on the difference between the conditional characteristic functions.

It is desirable to construct a test of conditional independence of two random vectors X and Y given a random vector Z based on some measure of conditional association, where the measure of conditional association have the following properties:

P1 The measure can be defined for any types of random vectors, including both discrete and continuous variables.

P2 The measure is invariant when one-to-one transforms are applied to each vector.

P3 The measure is between 0 and 1, where 0 corresponds to independence and 1 corresponds to full dependence.

Part or all of Properties P1 - P3 have been considered by various authors in different contexts. Some examples are as follows.

Romanovič (1975) defined the maximum partial correlation between two σ-fields given a third σ-field. According to Romanovič's definition, the maximum partial correlation between $\sigma(X)$ and $\sigma(Y)$ given $\sigma(Z)$ is

$$\sup_{f,g} \operatorname{corr}\left(f(X,Z) - E(f(X,Z)|Z), g(Y,Z) - E(g(Y,Z)|Z)\right),$$

where $\sigma(X)$ denotes the σ-field generated by the random vector X. The maximum partial correlation between $\sigma(X)$ and $\sigma(Y)$ given $\sigma(Z)$ can serve as a measure of conditional association of X and Y given Z, and it satisfies Properties P1 - P3.

Su and White (2005) proposed a test of conditional independence which is based on a test statistic that is a weighted Hellinger distance between the

conditional density of X given Z and the conditional density of Y given Z. Such a statistic can serve as a measure of conditional association, and they chose Hellinger distance so that the test statistic has the invariant property P2.

Dauxois and Nkiet (1998) proposed to use canonical coefficients obtain in nonlinear canonical analysis (NLCA) to construct measures of association and tests of independence. The following is a straightforward extension of Dauxois and Nkiet (1998)'s definition of the canonical coefficients to the conditional case.

Definition 1. Suppose that there exist pairs of functions (f_i, g_i): $i = 0$, 1, ..., such that for each i, (f_i, g_i) is a pair of functions (f, g) that maximizes $E\left(f(X,Z)g(Y,Z)|Z\right)$ subject to $E(f^2(X,Z)|Z) = 1 = E(g^2(Y,Z)|Z)$ and

$$E(f(X,Z)f_j(X,Z)|Z) = 0 = E(g(Y,Z)g_j(Y,Z)|Z) \text{ for } j < i.$$

Define $\rho_i(X,Y|Z) = E\left(f_i(X,Z)g_i(Y,Z)|Z\right)$ for each i. The $\rho_i(X,Y|Z)$'s will be referred as canonical coefficients.

Suppose that the (f_i, g_i)'s in Definition 1 exist, then a proper combination of $\rho_i(X,Y|Z)$'s can give a measure of conditional association. Some examples of such a combination are $\rho_1(X,Y|Z)$ and $-\sum_k \log(1-\rho_k^2(X,Y|Z))$, whose unconditional counterparts are two commonly used measures of association, as mentioned in Huang, Lee and Hsiao (2006).

Among the various approaches for constructing measures of conditional association described above, the NLCA approach offers the most flexibility. *The objective* of this project is to construct a test of conditional independence based on measures of conditional association from NLCA. However, it is not clear what conditions need to be added to guarantee the existence of the (f_i, g_i)'s. In the report, an alternative definition for the canonical coefficient $\rho_1(X,Y|Z)$ is provided to avoid finding such conditions.

2 Measures of Conditional Association

In this section, the canonical coefficients $\rho_i(X,Y|Z)$'s are defined using a new approach. In Definition 1, it is clear that $\rho_0(X,Y|Z) = 1$ with $f_0(X,Z) = 1 = g_0(Y,Z)$. Therefore, only the $\rho_i(X,Y|Z)$'s with $i \geq 1$ will be defined again, and the new definitions involve only pairs of functions in

$$S = \{(f,g) : E(f^2(X,Z)|Z) = 1 = E(g^2(Y,Z)|Z) \text{ and } E(f(X,Z)|Z) = 0 = E(g(Y,Z)|Z)\}.$$

As mentioned in Section 1, the purpose for introducing alternative definitions for the $\rho_i(X,Y|Z)$'s is to avoid dealing with the existence of the maximizers (f_i, g_i)'s. To achieve this goal, the maximums of certain conditional expectations in the original definitions of the $\rho_i(X,Y|Z)$'s will be replaced by suitable supremums. In particular, $\sup_{(f,g)\in S^*} E(f(X,Z)g(Y,Z)|Z)$ needs to be defined for $S^* \subset S$, for which the following fact is used:

Fact 1 *For $S^* \subset S$, there exists a sequence $\{(\alpha_n, \beta_n)\}$ in S^* such that*

(i) The sequence $\{E(\alpha_n(X,Z)\beta_n(Y,Z)|Z)\}$ is increasing (not necessarily strictly), and

(ii) for every $(f,g) \in S^*$, $E(f(X,Z)g(Y,Z)|Z) \leq \lim_{n\to\infty} E(\alpha_n(X,Z)\beta_n(Y,Z)|Z)$ almost surely.

Furthermore, if (i) and (ii) hold for $\{(\alpha_n(X,Z), \beta_n(Y,Z))\} = \{(\alpha_{n,1}(X,Z), \beta_{n,1}(Y,Z))\}$ or $\{(\alpha_{n,2}(X,Z), \beta_{n,2}(Y,Z))\}$, then

$$\lim_{n\to\infty} E(\alpha_{n,1}(X,Z)\beta_{n,1}(Y,Z)|Z) = \lim_{n\to\infty} E(\alpha_{n,2}(X,Z)\beta_{n,2}(Y,Z)|Z) \quad (1)$$

almost surely.

Fact 1 allows one to define $\sup_{(f,g)\in S^*} E(f(X,Z)g(Y,Z)|Z)$:

Definition 2. For $S^* \subset S$,

$$\sup_{(f,g)\in S^*} E(f(X,Z)g(Y,Z)|Z) = \lim_{n\to\infty} E(\alpha_n(X,Z)\beta_n(Y,Z)|Z),$$

where $\{(\alpha_n, \beta_n)\}$ is a sequence in S^* that satisfies (i) and (ii) in Fact 1.

Proof for Fact 1. First, note that (1) holds because for every n,

$$E(\alpha_{n,2}(X,Z)\beta_{n,2}(Y,Z)|Z) \leq \lim_{n\to\infty} E(\alpha_{n,1}(X,Z)\beta_{n,1}(Y,Z)|Z)$$

and

$$E(\alpha_{n,1}(X,Z)\beta_{n,1}(Y,Z)|Z) \leq \lim_{n\to\infty} E(\alpha_{n,2}(X,Z)\beta_{n,2}(Y,Z)|Z)$$

almost surely. It remains to find a sequence $\{(\alpha_n, \beta_n)\}$ that satisfies (i) and (ii). Let $\{(\alpha_n^*(X,Z), \beta_n^*(Y,Z)\}$ be a sequence in S^* such that $E(\alpha_n^*(X,Z)\beta_n^*(Y,Z))$ increases to $\sup_{(f,g)\in S^*} E(f(X,Z)g(Y,Z))$.

Let $(\alpha_1(X,Z), \beta_1(Y,Z)) = (\alpha_1^*(X,Z), \beta_1^*(Y,Z))$, and for $n \geq 2$, define

$$(\alpha_n(X,Z), \beta_n(Y,Z)) = \begin{cases} (\alpha_n^*(X,Z), \beta_n^*(Y,Z)) & \text{if } E(\alpha_n^*(X,Z)\beta_n^*(Y,Z)|Z) > E(\alpha_{n-1}(X,Z)\beta_{n-1}(Y,Z)|Z); \\ (\alpha_{n-1}(X,Z), \beta_{n-1}(Y,Z)) & \text{otherwise.} \end{cases}$$

Then $\{(\alpha_n(X,Z), \beta_n(Y,Z))\}$ is a sequence in S^* that satisfies (i). To see that $\{(\alpha_n(X,Z), \beta_n(Y,Z))\}$ also satisfies (ii), for (α, β) in S^*, Define

$$(\alpha_n^{**}, \beta_n^{**}) = \begin{cases} (\alpha, \beta) & \text{if } E(\alpha(X,Z)\beta(Y,Z)|Z) > \lim_{n\to\infty} E(\alpha_n(X,Z)\beta_n(Y,Z)|Z); \\ (\alpha_n, \beta_n) & \text{otherwise.} \end{cases}$$

Then

$$E(\alpha_n^{**}(X,Z)\beta_n^{**}(Y,Z)|Z) = \max\{E(\alpha(X,Z)\beta(Y,Z)|Z), E(\alpha_n(X,Z)\beta_n(Y,Z)|Z)\}$$

and

$$E(\alpha_n^{**}(X,Z)\beta_n^{**}(Y,Z)) = \sup_{(f,g)\in S^*} E(f(X,Z)g(Y,Z)) = E(\alpha_n(X,Z)\beta_n(Y,Z)),$$

so $E(\alpha_n^{**}(X,Z)\beta_n^{**}(Y,Z)|Z) = E(\alpha_n(X,Z)\beta_n(Y,Z)|Z)$ almost surely and (ii) holds. The proof of Fact 1 is complete.

With Definition 2, $\rho_1(X,Y|Z)$ can be re-defined as follows:

Definition 3. $\rho_1(X,Y|Z) = \sup_{(f,g)\in S} E(f(X,Z)g(Y,Z)|Z)$.

Note that if the maximizer (f_1, g_1) in Definition 1 exists, it is clear that $\rho_1(X,Y|Z) = E(f_1(X,Z)g_1(Y,Z)|Z)$ using Definition 3. Therefore, the definition for $\rho_1(X,Y|Z)$ in Definition 3 can be viewed as a generalized version of that in Definition 1.

It might be possible to define the $\rho_k(X,Y|Z)$'s for $k \geq 2$ without assuming the existence of the the (f_i, g_i)'s in Definition 1. However, the definition is currently under construction and is not reported here.

Below are some remarks for the $\rho_k(X,Y|Z)$'s.

1. $\rho_k(X,Y|Z)$'s satisfy Properties P1 and P2 and are between 0 and 1. $\rho_1(X,Y|Z)$ satisfies Property P3. That is, when X and Y are conditionally independent given Z, $\rho_1(X,Y|Z) = 0$. When X is a function of Y and Z or Y is a function of X and Z, $\rho_1(X,Y|Z) = 1$.

2. When Z is a constant vector, $\rho_k(X,Y|Z)$'s are the canonical coefficients in Dauxois and Nkiet (1998).

3. It is stated in Dauxois and Nkiet (1998) that when the joint distribution of X and Y is bivariate normal $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, the first canonical coefficient $\rho_1(X,Y) = |\rho|$. This result implies that, when the joint distribution for X, Y and Z is multivariate normal and X and Y are both univariate,

$$\begin{aligned} \rho_1(X,Y|Z) &= \left| \frac{E(X - E(X|Z))(Y - E(Y|Z))|Z)}{(E(X - E(X|Z))^2|Z))^{1/2} (E(Y - E(Y|Z))^2|Z))^{1/2}} \right| \\ &= \left| \frac{E(X - E(X|Z))(Y - E(Y|Z)))}{(E(X - E(X|Z))^2))^{1/2} (E(Y - E(Y|Z))^2))^{1/2}} \right|, \end{aligned}$$

which also equals the absolute value of the usual partial correlation coefficient.

3 A Test of Conditional Independence

Testing conditional independence is equivalent to testing if $E(\rho_1(X,Y|Z)) = 0$. In this section, an estimator for $E(\rho_1(X,Y|Z))$ is proposed, and its asymptotic distribution is derived to give a test of conditional independence.

3.1 Estimation of $E(\rho_1(X,Y|Z))$

An estimator for $E(\rho_1(X,Y|Z))$ can be constructed using basis approximation. First, suppose that there exist basis functions $\{\phi_{p,i} : 1 \le i \le p, p \ge 1\}$, $\{\psi_{q,j} : 1 \le j \le q, q \ge 1\}$ and $\{\theta_{r,k} : 1 \le k \le r, k \ge 1\}$ such that for $\{p_n\}$ and q_n with $\lim_{n\to\infty} p_n = \infty$ and $\lim_{n\to\infty} q_n = \infty$,

$$\lim_{n\to\infty} \inf_{p\le p_n, r\le r_n, \alpha_{p,r,i,k}} E\left(\alpha(X,Z) - \sum_{1\le i\le p, 1\le k\le r} \alpha_{p,r,i,k}\phi_{p,i}(X)\theta_{r,k}(Z)\right)^2 = 0 \tag{2}$$

and

$$\lim_{n\to\infty} \inf_{q\le q_n, r\le r_n, \beta_{q,r,j,k}} E\left(\beta(Y,Z) - \sum_{1\le j\le q, 1\le k\le r} \beta_{q,r,j,k}\psi_{q,j}(Y)\theta_{r,k}(Z)\right)^2 = 0 \tag{3}$$

for any $\alpha(X,Z)$ and $\beta(Y,Z)$ with finite second moments. Furthermore, it is assumed that for each (p,q), there exist coefficients $\alpha_{p,i}$'s and $\beta_{q,j}$'s such that

$$1 = \sum_{1\le i\le p} \alpha_{p,i}\phi_{p,i}(X) \text{ and } 1 = \sum_{1\le j\le q} \beta_{q,j}\psi_{q,j}(Y).$$

Then $\rho_1(X,Y|Z)$ can be approximated by $\sup_{(f,g)\in S_{p_n,q_n}} E(f(X,Z)g(Y,Z)|Z)$, where $S_{p_n,q_n} = \{(f,g) \in S : f(X,Z) = \sum_{i=1}^{p_n} \alpha_i(Z)\phi_{p_n,i}(X)$ and $g(Y,Z) = \sum_{j=1}^{q_n} \beta_j(Z)\psi_{q_n,j}(Y)\}$. Denote the supremum by $\rho_{p_n,q_n}(Z)$. Then specific approximation result is stated as follows.

Fact 2 *Suppose that* $\lim_{n\to\infty} p_n = \infty$ *and* $\lim_{n\to\infty} q_n = \infty$, *then* $\lim_{n\to\infty} E(|\rho_1(X,Y|Z) - \rho_{p_n,q_n}(Z)|) = 0$.

The proof of Fact 2 follows from the approximation properties (2) and (3):

$$\begin{aligned}\rho_1(X,Y|Z) &\approx Ef(X,Z)g(Y,Z)|Z = \text{ corr } (f(X,Z)g(Y,Z)|Z) \\ &\approx \quad \text{corr } (f^*(X,Z), g^*(Y,Z)|Z) \le \rho_{p_n,q_n}(Z)\end{aligned}$$

for some $(f,g) \in S$ and some $(f^*,g^*) \in S_{p_n,q_n}$. Based on Fact 2, it is reasonable to estimate $\rho_1(X,Y|Z)$ using an estimator for $\rho_{p_n,q_n}(Z)$. To make such estimation possible, it is assumed that for each (p,q,i,j), $E(\phi_{p,i}(X)\psi_{q,j}(Y)|Z)$

is equal to some continuous function of Z almost surely. Then an estimator for $\rho_{p_n,q_n}(z)$ and $\rho_1(X,Y|Z=z)$ is

$$\max_{\{\alpha_i\}_{i=1}^{p_n},\{\beta_j\}_{j=1}^{q_n}} \sum_{i,j} \alpha_i\beta_j \left(\hat{E}(\phi_i(X)\psi_j(Y)|Z=z) - \hat{E}(\phi_i(X)|Z=z)\hat{E}(\psi_j(Y)|Z=z)\right),$$

where the maximum is taken over all $\{\alpha_i\}_{i=1}^{p_n}$ and $\{\beta_j\}_{j=1}^{q_n}$ such that

$$\sum_{1\le i\le p_n, 1\le j\le q_n} \alpha_i\alpha_j \left(\hat{E}(\phi_i(X)\phi_j(X)|Z=z) - \hat{E}(\phi_i(X)|Z=z)\hat{E}(\phi_j(X)|Z=z)\right) = 1$$

and

$$\sum_{1\le i\le p_n, 1\le j\le q_n} \beta_i\beta_j \left(\hat{E}[\psi_i(Y)\psi_j(Y)|Z=z] - \hat{E}(\psi_i(Y)|Z=z)\hat{E}(\psi_j(Y)|Z=z)\right) = 1.$$

Here $\hat{E}[g(X,Y)|Z=z] = \sum_{i=1}^{N_n} g(X_i,Y_i)k_h(z-Z_i)/\sum_{i=1}^{N_n} k_h(z-Z_i)$, where $N_n \to \infty$ as $n\to\infty$, for $z=(z(1),\ldots,z(d))$, $k_h(z) = \prod_{j=1}^{d} h^{-1}k_0(z(j)/h)$ and k_0 is a symmetric probability density function on R. Denote the above estimator for $\rho_1(X,Y|Z=z)$ by $\hat{\rho}(z)$, then $N_{\rho,n}^{-1}\sum_{i=N_n+1}^{N_n+N_{\rho,n}} \hat{\rho}(Z_i)$ is an estimator for $E(\rho_1(X,Y|Z))$, where $N_{\rho,n}\to\infty$ as $n\to\infty$.

The estimator $\hat{\rho}(z)$ can be obtained using SVD (single value decomposition), which makes it easy to do the computation.

3.2 Asymptotic distribution of the estimator and test of conditional independence

To build a test for conditional independence based on the estimator $\hat{\rho}$, it is necessary to derive the asymptotic distribution of the estimator under the conditional independence hypothesis. An asymptotic property of the estimator $\hat{\rho}(z)$ is given in the following theorem.

Theorem 1 *Suppose that $p_n = p$ and $q_n = q$ do not depend on n, then $\sqrt{N_n h_n^d}(\hat{\rho}(z) - \rho_{p,q}(z))$ converges in distribution as $n\to\infty$. Furthermore, if X and Y are conditionally independent given Z, then $N_n h_n^d c_K \hat{\rho}^2(z)$ converges in distribution to the maximum eigenvalue of CC^T as $n\to\infty$, where $c_K = f(z)/\int k_0^2(s)ds$, f is the pdf of Z and C is an $p\times q$ matrix with 0's in the first row and first column and other elements are IID $N(0,1)$.*

The proof of Theorem 1 requires only minor modification of Lemma 7.1 in Dauxois and Nkiet (1998) and is left out. Also, the asymptotic joint distribution of the estimators of certain conditional expectations is needed, which is taken directly from the lecture notes by James Powell titled "Notes on Nonparametric Regression Estimation", which is available at

`http://emlab.berkeley.edu/users/powell/e241a_sp07/nrnotes.pdf`

The asymptotic distribution of the test statistic $N_{\rho,n}^{-1}\sum_{i=N_n+1}^{N_n+N_{\rho,n}}\hat{\rho}(Z_i)$ is normal, with mean and variance equal to the mean and variance of $\hat{\rho}(Z_1)$, which can be approximated using estimators from Bootstrap or the asymptotic distribution of $\hat{\rho}(z)$. However, if p_n and q_n both tend to ∞, then the distribution of the maximum eigenvalue of the random matrix CC^T does not converge. Therefore, a more general version of Theorem 1 needs to be derived in order to understand the behavior of $\hat{\rho}(z)$ when both p_n and q_n tend to ∞.

References

[1] Baba, K.,Shibata, R. and Sibuya, M. (2004). *Australian and New Zealand Journal of Statistics* **46** 657–664.

[2] Daudin, J. J. (1980). *Biometrika* **67** 581–590.

[3] Dauxois, J. and Nkiet, G. M. (1998). *Annals of Statistics* **26** 1254–1278.

[4] Delgado, M. A. and González-Manteiga, W. (2001). *Annals of Statistics* 1469–1507.

[5] Dossou-Gbete, S. and Pousse, A. (1991). *Statistics* **22** 479–491.

[6] Huang, S.-Y.,Lee, M.-H. and Hsiao, C. K. (2006). *Draft*

[7] Johnstone, I. (2001). *Annals of Statistics* **29** 295–327.

[8] Lawrance, A. J. (1976). *The American Statistician* **30** 146–149.

[9] Li, L.,Cook, R. D. and Nachtsheim, C. J. (2005). *Journal of the Royal Statistical Society. Series B* **67** 285–299.

[10] Linton, O. and Gozalo, P. (1997). *Discussion paper, Cowles Foundation for Research in Economics, Yale University*

[11] Muirhead, R. J. and Waternaux, C. M. (1980). *Biometrika* **67** 31–43.

[12] Romanovič, V. A. (1975). *Izvestiya Vysshikh Uchebnykh Zavedenii̯ Matematika* **10** 94–96.

[13] Roussas, G. G. and Tran, L. T. (1992). *Annals of Statistics* **20** 98–120.

[14] Schuster, E. F. (1972). *Annals of Mathematical Statistics* **43** 84–88.

[15] Su, L. and White, H. (2005). *Submitted to Econometric Theory.*

[16] Su, L. and White, H. (2006). *Submitted to Journal of Econometrics*

[17] Waternaux, C. M. (1976). *Biometrika* **63** 639–645.