行政院國家科學委員會專題研究計畫 成果報告

轉換資料的核迴歸估計量的漸進分布
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 100-2118-M-004-006-
執 行 期 間 ：100 年 08 月 01 日至 101 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-兼任助理人員：林昱航
博士班研究生-兼任助理人員：鄭宇翔

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 101 年 07 月 31 日

中 文 摘 要 ： 本研究欲探討資料經過經驗分布轉換後，所得到的核迴歸估計量的漸進分布。 研究結果顯示, 資料經過經驗分布轉換後，所得到的核迴歸估計量的漸進分布，會和資料經過真正的分布轉換後，所得到的核迴歸估計量的聯合漸進分布相同。

中文關鍵詞： 核迴歸估計量, 經驗分布轉換, 漸進分布

英 文 摘 要 ： The proposed research is concerned with asymptotic distributions of kernel regression estimators when the data are transformed using empirical CDF ‘s. It is found that the asymptotic distribution is the same as that for the case where the data are transformed using the true CDF ‘s.

英文關鍵詞： kernel regression estimator, empirical CDF, asymptotic distribution

1 Introduction, Literature Review and Objectives

Suppose that X, Y and Z are random variables with CDF's F_X, F_Y and F_Z respectively. Suppose that λ is a smooth function defined on $[0,1]$. Consider the estimation of

$$g_1(u_0) \equiv E[g(F_X(X), F_Y(Y))|\lambda(F_Z(Z)) = u_0]$$

based on a random sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ from the distribution of (X, Y, Z). A kernel estimator based on kernel function k and bandwidth h is given by

$$\hat{g}_1(u_0) = \frac{\sum_{i=1}^n g(\hat{F}_X(X_i), \hat{F}_Y(Y_i))k([u_0 - \lambda(\hat{F}_Z(Z_i))]/h)}{\sum_{i=1}^n k([u_0 - \lambda(\hat{F}_Z(Z_i))]/h)}, \tag{1}$$

where $\hat{F}_X$, $\hat{F}_Y$ and $\hat{F}_Z$ are empirical CDF's based on $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$ and $(Z_1, \ldots, Z_n)$ respectively.

The motivation of studying the joint asymptotic distribution of kernel estimators of the form in (1) is related to the implementation of the testing procedure for conditional independence in Huang(2010)[4] based on maximal conditional correlation. The test in [4] is for testing the conditional independence of X and Y given Z based on $\{(X_i, Y_i, Z_i)\}_{i=1}^n$. To obtain the test statistic in [4], it is convenient to transform the X_i's, Y_i's and Z_i's so that they are in the range $[0,1]$. A convenient way to transform the data is to apply $\hat{F}_X$, $\hat{F}_Y$ and $\hat{F}_Z$ to X_i's, Y_i's and Z_i's respectively. In Cheng and Huang (2012) [2], the test in [4] is applied to dependent data and data are transformed using empirical CDF's.

From the simulation results in [2], it seems that the approach of transforming data using empirical CDF's works well. However, this approach is not yet theoretically justified. In the derivation of the asymptotic distribution of the test statistic in [4] or [2], it is crucial to make use of the asymptotic distribution of kernel estimators of the form in (1) with $\hat{F}_X(X_i)$, $\hat{F}_Y(Y_i)$ and $\hat{F}_Z(Z_i)$ replaced by X_i, Y_i and Z_i respectively. In order to derive the asymptotic distribution of the test statistic in [4] for the case where data are transformed using empirical CDF's, it is necessary to find the asymptotic distributions of kernel estimators of the form in (1).

Another related problem considered is the estimation of

$$g_2(u_0) \equiv E[g(X, Y)|\lambda(F_Z(Z)) = u_0]$$

based on a random sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ from the distribution of (X, Y, Z). A kernel estimator based on kernel function k and bandwidth h is given by

$$\hat{g}_2(u_0) = \frac{\sum_{i=1}^n g(X_i, Y_i)k([u_0 - \lambda(\hat{F}_Z(Z_i))]/h)}{\sum_{i=1}^n k([u_0 - \lambda(\hat{F}_Z(Z_i))]/h)}, \tag{2}$$

The motivation for deriving the asymptotic distributions of kernel estimators of the form in (2) is to improve the performance of kernel regression estimators. In a regression model

$$Y_i = f(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3}$$

one may estimate f nonparametrically using smoothing methods such as kernel estimation or local polynomial smoothing. The kernel regression estimator for $f(x)$ is

$$\hat{E}(Y_1|X_1 = x) = \frac{\sum_{i=1}^{n} Y_i k((x - X_i)/h)}{\sum_{i=1}^{n} k((x - X_i)/h)},$$

where k is the kernel function and h is the bandwidth. Suppose that X_i is one-dimensional. If the errors ε_i's are IID, of mean zero and independent of X_i's, then it is well known (see Schuster (1972) [5] for example) that $\sqrt{nh}(\hat{E}(Y_1|X_1 = x) - E(Y_1|X_1 = x))$ converges in distribution to $N(0, v(x))$, where

$$v(x) = \frac{\int k^2(s)ds E(\varepsilon_1^2)}{f_X(x)} \tag{4}$$

and f_X is the density of X_1. The variance term in (4) indicates that the estimation error can be large when $f_X(x)$ is small. Therefore, if the distribution of X_i is highly clustered, then the performance of $\hat{E}(Y_1|X_1 = x)$ can be un satisfactory for many x's, as mentioned in Fan, Hu and Truong (1994) [3] .Such a result makes sense since there are few X_i's near x whose corresponding Y_i's can be used in estimating $E(Y_1|X_1 = x)$ using $\hat{E}(Y_1|X_1 = x)$. To overcome this problem, one may apply the empirical CDF transform to the X_i's and then use kernel estimator based on the transformed X_i's, which are roughly uniformly distributed, and then estimate f based on the estimator of $E(Y_1|F_X(X_1) = x)$.

The results of this study show that, under proper conditions, the asymptotic distributions of $\hat{g}_1(u_0)$ and $\hat{g}_2(u_0)$ are the same as those of $\tilde{g}_1(u_0)$ and $\tilde{g}_2(u_0)$, where for $i = 1, 2$, $\tilde{g}_i(u_0)$ is the same as $\hat{g}_i(u_0)$ except that the $\hat{F}_X$, $\hat{F}_Y$ and $\hat{F}_Z$ in (1) and (2) are replaced by F_X, F_Y and F_Z respectively.

2 Approach and Main Result

The derivations of the asymptotic distributions for $\hat{g}_1(u_0)$ and $\hat{g}_2(u_0)$ are similar, so only the derivation for $\hat{g}_1(u_0)$ is presented. Let

$$\tilde{m}(u_0) = \frac{1}{nh} \sum_{i=1}^{n} [g(\hat{F}_X(X_i), \hat{F}_Y(Y_i)) - g_1(u_0)] k \left(\frac{u_0 - \lambda(\hat{F}_Z(Z_i))}{h} \right)$$

and

$$\tilde{f}_\lambda(u_0) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_i))}{h}\right),$$

Then

$$\hat{g}_1(u_0) - g_1(u_0) = \frac{\tilde{m}(u_0)}{\tilde{f}_\lambda(u_0)}.$$

To approximate $\tilde{m}(u_0)$, for $i = 1, \ldots, n$, let

$$U_i = [g(F_X(X_i), F_Y(Y_i)) - g_1(u_0)]k\left(\frac{u_0 - \lambda(F_Z(Z_i))}{h}\right)$$

and let $m(u_0) = \sum_{i=1}^{n} U_i/nh$. Let $d_1(u_0) = \tilde{m}(u_0) - m(u_0)$, then it will be shown that $d_1(u_0)$ is $o_p(1/\sqrt{nh})$. To approximate $\tilde{f}_\lambda(u_0)$, let

$$\hat{f}_\lambda(u_0) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{u_0 - \lambda(F_Z(Z_i))}{h}\right),$$

and $d_2(u_0) = \tilde{f}_\lambda(u_0) - \hat{f}_\lambda(u_0)$. It can also be shown that $d_2(u_0)$ is $o_p(1)$. Therefore,

$$\sqrt{nh}(\hat{g}_1(u_0) - g_1(u_0)) = \sqrt{nh}(m(u_0))/\hat{f}_\lambda(u_0) + o_p(1),$$

so the limiting distribution of $\sqrt{nh}(\hat{g}_1(u_0) - g_1(u_0))$ is the same as that of $\sqrt{nh}(m(u_0))/\hat{f}_\lambda(u_0)$, which is the same as $\sqrt{nh}(\hat{g}_1(u_0) - g_1(u_0))$ except that the empirical CDF's $\hat{F}_X$, $\hat{F}_Y$ and $\hat{F}_Z$ are replaced by the true CDF's F_X, F_Y and F_Z respectively.

To show that $d_1(u_0) = o_P(1/\sqrt{nh})$ and $d_2(u_0) = o_P(1)$, the following assumptions are made.

- g_1' and g_1'' are continuous and bounded.
- $\lambda^{(k)}$: $1 \leq k \leq 5$ are bounded. λ' is bounded away from zero. λ^{-1} is bounded.
- $\lambda(F_Z(Z_1))$ has a density function f_λ, and $f_\lambda^{(k)}$ are bounded for $0 \leq k \leq 3$.
- The cumulative distribution functions of (X_1, Z_1) and (Y_1, Z_1) are continuously differentiable and the partial derivatives are bounded.
- k is five-times continuously differentiable, and k is supported on $[-1, 1]$. $\int uk(u)du = 0$.
- g and all the partial derivatives of g of order 3 or lower are bounded.

- Let $f(\cdot|u)$ be the conditional probability density function of $(F_X(X_1), F_Y(Y_1))$ given $F_Z(Z_1) = u$. $f(\cdot|u)$ has a continuous and bounded partial second derivative with respect to u.
- nh^4 tends to ∞ as n tends to ∞ and $nh^5 = O(1)$.

The main result is stated in Theorem 1.

Theorem 1 *Under the above assumptions,* $d_1(u_0) = o_p(1/\sqrt{nh})$ *and* $d_2(u_0) = o_P(1)$.

3 Proof of Theorem 1

The proof of Theorem 1 consists of the mean and variance calculation of $d_1(u_0)$. In Section 3.1, it is shown that $Var(d_1(u_0)) = o(1/(nh))$; in Section 3.2, it is shown that $E(d_1(u_0)) = o(h^2)$. Therefore, $d_1(u_0) = o_p(1/\sqrt{nh})$. The proof of $d_2(u_0) = o_P(1)$ is done in Chen and Huang (2007) [1].

3.1 Variance calculation of $d_1(u_0)$

The variance of $d_1(u_0)$ is calculated below.

$$\begin{aligned} Var(d_1(u_0)) &= Var\left[\frac{1}{nh}\sum_{i=1}^{n}[g(\hat{F}_X(X_i), \hat{F}_Y(Y_i)) - g_1(u_0)]k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_i))}{h}\right) - m(u_0)\right] \\ &= \frac{V_1}{nh^2} + \frac{(n-1)V_2}{nh^2}, \end{aligned} \tag{5}$$

where V_1 is the variance of

$$[g(\hat{F}_X(X_1), \hat{F}_Y(Y_1)) - g_1(u_0)]k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_1))}{h}\right) - U_1$$

and V_2 is the covariance between

$$[g(\hat{F}_X(X_1), \hat{F}_Y(Y_1)) - g_1(u_0)]k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_1))}{h}\right) - U_1$$

and

$$[g(\hat{F}_X(X_2), \hat{F}_Y(Y_2)) - g_1(u_0)]k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_2))}{h}\right) - U_2.$$

To compute V_1 and V_2, we first introduce some notation. For $i = 1, 2$, let $\Delta_X(X_i) = \hat{F}_X(X_i) - F_X(X_i)$, $\Delta_Y(Y_i) = \hat{F}_Y(Y_i) - F_Y(Y_i)$ and $\Delta_Z(Z_i) = \hat{F}_Z(Z_i) - F_Z(Z_i)$. Let

$$g_5(x) = k\left(\frac{u_0 - \lambda(x)}{h}\right)$$

and for $i = 1$, 2 and $\ell \geq 0$, let

$$a_\ell(Z_i) = g_5^{(\ell)}(F_Z(Z_i))(\Delta_Z(Z_i))^\ell \tag{6}$$

$$b_0(X_1, Y_1, Z_1) = g(F_X(X_1), F_Y(Y_1)) - g_1(\lambda(F_Z(Z_1))),$$

$$b_1(Z_1) = g_1(\lambda(F_Z(Z_1))) - g_1(u_0)$$

$$b_2(X_1, Y_1) = g_x(F_X(X_1), F_Y(Y_1))(\hat{F}_X(X_1) - F_X(X_1)),$$

$$b_3(X_1, Y_1) = g_y(F_X(X_1), F_Y(Y_1))(\hat{F}_Y(Y_1) - F_Y(Y_1)),$$

$$b_4(X_1, Y_1) = \frac{g_{xx}(F_X(X_1), F_Y(Y_1))(\hat{F}_X(X_1) - F_X(X_1))^2}{2},$$

$$b_5(X_1, Y_1) = g_{xy}(F_X(X_1), F_Y(Y_1))(\hat{F}_X(X_1) - F_X(X_1))(\hat{F}_Y(Y_1) - F_Y(Y_1)),$$

and

$$b_6(X_1, Y_1) = \frac{g_{yy}(F_X(X_1), F_Y(Y_1))(\hat{F}_Y(Y_1) - F_Y(Y_1))^2}{2},$$

where $g_{xx}(x, y) = \dfrac{\partial^2}{\partial x^2} g(x, y)$, $g_{xy}(x, y) = \dfrac{\partial^2}{\partial x \partial y} g(x, y)$ and $g_{yy}(x, y) = \dfrac{\partial^2}{\partial y^2} g(x, y)$, then we have (i) and (ii):

(i) V_2 is the covariance between

$$\left(\sum_{i=0}^{4} a_i(Z_1)\right)\left(b_0(X_1, Y_1, Z_1) + b_1(Z_1) + \sum_{i=2}^{6} b_i(X_1, Y_1)\right) - U_1$$

and

$$\left(\sum_{i=0}^{4} a_i(Z_2)\right)\left(b_0(X_2, Y_2, Z_2) + b_1(Z_2) + \sum_{i=2}^{6} b_i(X_2, Y_2)\right) - U_2$$

plus a term of order $o(h/n)$.

(ii) $V_1 = o(h)$.

Below we will explain why (i) and (ii) hold. Note that

$$E\left|\prod_{i=1}^{2}\left[\Delta_X(X_i)^{a_i}\Delta_Y(Y_i)^{b_i}\left(\frac{\Delta_Z(Z_i)}{h}\right)^{c_i}\right]\right| = O\left(\frac{1}{n^{s/2}h^{c_1+c_2}}\right), \tag{7}$$

where $s = \sum_{i=1}^{2}(a_i + b_i + c_i)$. When $c_1 + c_2 \geq 5$, (7) implies that

$$\frac{1}{h^2}E\left|\prod_{i=1}^{2}\left[\Delta_X(X_i)^{a_i}\Delta_Y(Y_i)^{b_i}\left(\frac{\Delta_Z(Z_i)}{h}\right)^{c_i}\right]\right| = \left(\frac{1}{nh}\right)O\left(\frac{1}{nh^4}\right)^{(s-2)/2} = o\left(\frac{1}{nh}\right).$$

In addition, when $s-(c_1+c_2)\geq 3$, (7) implies that

$$\frac{1}{h^2}E\left|\prod_{i=1}^{2}\left[\Delta_X(X_i)^{a_i}\Delta_Y(Y_i)^{b_i}\left(\frac{\Delta_Z(Z_i)}{h}\right)^{c_i}\right]\right| = \left(\frac{1}{nh}\right)O\left(\frac{1}{nh^2}\right)^{(s-2)/2} = o\left(\frac{1}{nh}\right).$$

Thus we do not have to consider terms involving $a_\ell(Z_i)$ for $\ell\geq 5$ or $\Delta_X(X_i)^a$ or $\Delta_Y(Y_i)^a$ for $a\geq 3$ in the calculation of V_2 (up to $o(h/n)$). From the above discussion, (i) holds.

To see that (ii) holds, note that when $c_1+c_2\geq 1$, (7) implies that

$$\frac{1}{h}E\left|\prod_{i=1}^{2}\left[\Delta_X(X_i)^{a_i}\Delta_Y(Y_i)^{b_i}\left(\frac{\Delta_Z(Z_i)}{h}\right)^{c_i}\right]\right| = O\left(\frac{1}{nh^4}\right)^{s/2} = o(1),$$

and that when $s-(c_1+c_2)\geq 1$, (7) implies that

$$\frac{1}{h}E\left|\prod_{i=1}^{2}\left[\Delta_X(X_i)^{a_i}\Delta_Y(Y_i)^{b_i}\left(\frac{\Delta_Z(Z_i)}{h}\right)^{c_i}\right]\right| = O\left(\frac{1}{nh^2}\right)^{s/2} = o(1).$$

Therefore, we do not have to consider terms involving $\Delta_X(X_i)^a$, $\Delta_Y(Y_i)^a$, or $\Delta_Z(Z_i)^a$ for $a\geq 1$ in the calculation of V_1 (up to the order $o(h)$), so (ii) holds.

Below we will calculate the covariance between

$$\left(\sum_{i=0}^{4}a_i(Z_1)\right)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{6}b_i(X_1,Y_1)\right)-U_1$$

and

$$\left(\sum_{i=0}^{4}a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{6}b_i(X_2,Y_2)\right)-U_2$$

to show that $V_2=o(h/n)$.

We first compute the covariance between

$$a_4(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{6}b_i(X_1,Y_1)\right)$$

and

$$\left(\sum_{i=0}^{4}a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{6}b_i(X_2,Y_2)\right)-U_2.$$

From (7), the above covariance is the covariance between

$$a_4(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)\right)$$

and

$$a_0(Z_2)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)\right)-U_2=0$$

plus a term of order $o(h/n)$, so V_2 is the covariance between

$$\left(\sum_{i=0}^{3}a_i(Z_1)\right)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{6}b_i(X_1,Y_1)\right)-U_1$$

and

$$\left(\sum_{i=0}^{3}a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{6}b_i(X_2,Y_2)\right)-U_2$$

plus a term of order $o(h/n)$.

Next we compute the covariance between

$$a_3(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{6}b_i(X_1,Y_1)\right)$$

and

$$\left(\sum_{i=0}^{3}a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{6}b_i(X_2,Y_2)\right)-U_2,$$

which is

$$Cov\left[a_3(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)\right),a_0(Z_2)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)\right)-U_2\right]$$
$$+O\left(\frac{1}{n^2h^3}\right)+o(h/n)=o(h/n).$$

Therefore, V_2 is the covariance between

$$\left(\sum_{i=0}^{2}a_i(Z_1)\right)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{6}b_i(X_1,Y_1)\right)-U_1$$

and

$$\left(\sum_{i=0}^{2}a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{6}b_i(X_2,Y_2)\right)-U_2$$

plus a term of order $o(h/n)$.

Next we compute the covariance between

$$\left(\sum_{i=0}^{2}a_i(Z_1)\right)\left(\sum_{i=4}^{6}b_i(X_1,Y_1)\right)$$

and

$$\left(\sum_{i=0}^{2} a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{6} b_i(X_2,Y_2)\right)-U_2,$$

which is

$$\begin{aligned} & Cov\left[a_0(Z_1)\left(\sum_{i=4}^{6} b_i(X_1,Y_1)\right), a_0(Z_2)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)\right)-U_2\right]+o(h/n) \\ = \;& 0+o(h/n)=o(h/n), \end{aligned}$$

so V_2 is the covariance between

$$\left(\sum_{i=0}^{2} a_i(Z_1)\right)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{3} b_i(X_1,Y_1)\right)-U_1$$

and

$$\left(\sum_{i=0}^{2} a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{3} b_i(X_2,Y_2)\right)-U_2$$

plus a term of order $o(h/n)$.

Next we compute the covariance between

$$a_2(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{3} b_i(X_1,Y_1)\right)$$

and

$$\left(\sum_{i=0}^{2} a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{3} b_i(X_2,Y_2)\right)-U_2,$$

which is

$$\begin{aligned} & Cov\left[a_2(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)\right), a_0(Z_2)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)\right)-U_2\right]+o(h/n) \\ = \;& 0+o(h/n)=o(h/n). \end{aligned}$$

Therefore, V_2 is the covariance between

$$\left(\sum_{i=0}^{1} a_i(Z_1)\right)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)+\sum_{i=2}^{3} b_i(X_1,Y_1)\right)-U_1$$

and

$$\left(\sum_{i=0}^{1} a_i(Z_2)\right)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)+\sum_{i=2}^{3} b_i(X_2,Y_2)\right)-U_2$$

plus a term of order $o(h/n)$, which is the sum of

$$a_{112} = Cov\left[a_1(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)\right), a_1(Z_2)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)\right)\right],$$

$$a_{113} = 2Cov\left[a_1(Z_1)\left(b_0(X_1,Y_1,Z_1)+b_1(Z_1)\right), a_0(Z_2)\left(\sum_{i=2}^{3} b_i(X_2,Y_2)\right)\right],$$

$$a_{114} = Cov\left[a_0(Z_1)\left(\sum_{i=2}^{3} b_i(X_1,Y_1)\right), a_0(Z_2)\left(\sum_{i=2}^{3} b_i(X_2,Y_2)\right)\right],$$

and

$$a_{102} = 2Cov\left[a_0(Z_1)\left(\sum_{i=2}^{3} b_i(X_1,Y_1)\right), a_0(Z_2)\left(b_0(X_2,Y_2,Z_2)+b_1(Z_2)\right)\right]$$

plus a term of order $o(h/n)$.

To compute a_{112}, let

$$a_{11}(z) = -k'\left(\frac{u_0-\lambda(F_Z(z))}{h}\right)\lambda'(F_Z(z)) \tag{8}$$

and

$$g_{10}(c) = E\left[a_{11}(Z_1)b_1(Z_1)[\min(c, F_Z(Z_1)) - cF_Z(Z_1)]\right],$$

then

$$a_{112} = \frac{1}{nh}E\left[a_{11}(Z_2)b_1(Z_2)g_{10}(F_Z(Z_2))\right] + O\left(\frac{1}{n^2h^2}\right).$$

Express

$$\begin{aligned} g_{10}(c) &= -h\int_0^c k\left(\frac{u_0-\lambda(u)}{h}\right)\left[(g_1(\lambda(u))-g_1(u_0))(u-cu)\right]'du \\ &\quad -h\int_c^1 k\left(\frac{u_0-\lambda(u)}{h}\right)\left[(g_1(\lambda(u))-g_1(u_0))(c-cu)\right]'du, \end{aligned}$$

then $g_{10}(c)/(h^2)$ is bounded by a constant that does not depend on c. Since $E|a_{11}(Z_2)| = O(h)$, $a_{112} = O\left(\frac{h^2}{n}\right) + o(h/n) = o(h/n)$.

Next we compute a_{113}. Let $c_{1,3}$, $c_{2,3}$ be the cumulative distribution functions of $(F_X(X_1), F_Z(Z_1))$, $(F_Y(Y_1), F_Z(Z_1))$ respectively,

$$g_8(c_0) = E\left[a_{11}(Z_1)b_1(Z_1)[c_{1,3}(c_0, F_Z(Z_1)) - c_0F_Z(Z_1)]\right],$$

and

$$g_9(c_0) = E\left[a_{11}(Z_1)b_1(Z_1)[c_{2,3}(c_0, F_Z(Z_1)) - c_0F_Z(Z_1)]\right],$$

where a_{11} is defined in (8), then

$$\begin{aligned} a_{113} &= \frac{2}{nh} E\left[a_0(Z_2) g_x(F_X(X_2), F_Y(Y_2)) g_8(F_X(X_2))\right] \\ &\quad \frac{2}{nh} E\left[a_0(Z_2) g_y(F_X(X_2), F_Y(Y_2)) g_9(F_Y(Y_2))\right] + O\left(\frac{1}{n^2 h}\right), \end{aligned}$$

and

$$\begin{aligned} g_8(c_0) &= h \int_0^1 [c_{13}(c_0, u) - c_0 u][g_1(\lambda(u)) - g_1(u_0)] dk \left(\frac{u_0 - \lambda(u)}{h}\right) \\ &= -h \int_0^1 k\left(\frac{u_0 - \lambda(u)}{h}\right) [c_{13}(c_0, u) - c_0 u]'[g_1(\lambda(u)) - g_1(u_0)] du \\ &\quad -h \int_0^1 k\left(\frac{u_0 - \lambda(u)}{h}\right) [c_{13}(c_0, u) - c_0 u][g_1(\lambda(u)) - g_1(u_0)]' du, \end{aligned}$$

so $g_8(c_0)/(h^2)$ is bounded by a constant that does not depend on c_0. Similarly, $g_9(c_0)/(h^2)$ is bounded by a constant that does not depend on c_0. Therefore,

$$a_{113} = \frac{1}{nh} O(h^3) + O\left(\frac{1}{n^2 h}\right) = o(h/n).$$

$$a_{114} = O\left(\frac{1}{n}\right) E a_0(Z_1) E a_0(Z_2) + O\left(\frac{1}{n^2}\right) E a_0(Z_1) E a_0(Z_2) = O\left(\frac{h^2}{n}\right).$$

$$a_{102} = O\left(\frac{1}{n}\right) E a_0(Z_1) E a_0(Z_2) = O\left(\frac{h^2}{n}\right) = o(h/n).$$

In summary, from the above calculation, $V_2 = o(h/n)$ and $V_1 = o(h)$. From (5), $Var(d_1(u_0)) = o(1/(nh))$.

3.2 Expectation calculation of $d_1(u_0)$

The expectation of $d_1(u_0) = \tilde{m}(u_0) - m(u_0)$ is calculated below. Since

$$\begin{aligned} E[\tilde{m}(u_0)] &= \frac{1}{nh} E\left[\sum_{i=1}^n [g(\hat{F}_X(X_i), \hat{F}_Y(Y_i)) - g_1(u_0)] k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_i))}{h}\right)\right] \\ &= \frac{1}{h} E\left[[g(\hat{F}_X(X_1), \hat{F}_Y(Y_1)) - g_1(u_0)] k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_1))}{h}\right)\right] \\ &= \frac{1}{h} E\left[k\left(\frac{u_0 - \lambda(\hat{F}_Z(Z_1))}{h}\right)\left(b_0(X_1, Y_1, Z_1) + b_1(Z_1) + \sum_{i=2}^3 b_i(X_1, Y_1)\right)\right] + O\left(\frac{1}{nh}\right) \\ &= \frac{1}{h} E\left[\left(\sum_{i=0}^4 a_i(Z_1)\right)\left(b_0(X_1, Y_1, Z_1) + b_1(Z_1) + \sum_{i=2}^3 b_i(X_1, Y_1)\right)\right] + o(h^2) \\ &= \frac{1}{h} E\left[\left(\sum_{i=0}^1 a_i(Z_1)\right)(b_0(X_1, Y_1, Z_1) + b_1(Z_1))\right] + o(h^2) \end{aligned}$$

and

$$E[m(u_0)] = \frac{1}{h} E\left[a_0(Z_1)\left(b_0(X_1, Y_1, Z_1) + b_1(Z_1)\right)\right],$$

$$\begin{aligned}
E[d_1(u_0)] &= \frac{1}{h} E\left[a_1(Z_1)\left(b_0(X_1, Y_1, Z_1) + b_1(Z_1)\right)\right] + o(h^2) \\
&= \frac{1}{h} E\left[-k'\left(\frac{u_0 - \lambda(F_Z(Z_1))}{h}\right)\lambda'(F_Z(Z_1))b_1(Z_1)\left(\frac{1 - F_Z(Z_1)}{n}\right)\right] + o(h^2) \\
&= \frac{1}{n}\int_0^1 (1-u)[g_1(\lambda(u)) - g_1(u_0)]dk\left(\frac{u_0 - \lambda(u)}{h}\right) + o(h^2) \\
&= \frac{1}{n}\int_0^1 k\left(\frac{u_0 - \lambda(u)}{h}\right)((1-u)[g_1(\lambda(u)) - g_1(u_0)])'\, du = o(h^2).
\end{aligned}$$

4 Conclusion

From Theorem 1, the difference between the usual kernel estimator and the kernel estimator based on empirical CDF transformed data is negligible under the conditions stated. It is clear that both estimators have the same asymptotic distribution.

References

[1] S. X. Chen and T. M. Huang. Nonparametric estimation of copula functions for dependence modelling. *The Canadian Journal of Statistics*, 35(2):265–282, 2007.

[2] Yu-Hsiang Cheng and Tzee-Ming Huang. A conditional independence test for dependent data based on maximal conditional correlation. *Journal of Multivariate Analysis*, 107:210–226, 2012.

[3] Jianqing Fan, Tien-Chung Hu, and Young K. Truong. Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, 21(4):433–446, 1994.

[4] Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *Annals of Statistics*, 38(4):2047–2091, 2010.

[5] Eugene F. Schuster. Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *The Annals of Mathematical Statistics*, 43:84–88, 1972.

國科會補助計畫衍生研發成果推廣資料表

日期:2012/07/31

國科會補助計畫	計畫名稱: 轉換資料的核迴歸估計量的漸進分布
	計畫主持人: 黃子銘
	計畫編號: 100-2118-M-004-006- 學門領域: 統計推論

無研發成果推廣資料

100 年度專題研究計畫研究成果彙整表

0-2118-M-004-006-

該迴歸估計量的漸進分布

備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
研究成果已寫成技術報告一篇, 網址為 http://stat.nccu.edu.tw/download.php?filename=1026_3aca95b4.pdf&dir=writing&title=--%E9%99%84%E4%BB

名稱或內容性質簡述

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 □撰寫中 ■無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

使用核迴歸估計量時，若是能使用進行經驗分布轉換後的資料，則有一些便利性．此研究結果說明了使用此種轉換的資料和使用真正分布轉換後的資料，所得到核迴歸估計量的漸進分布是相同的．這為使用經驗分布轉換提供了理論上的依據.