The Canadian Journal of Statistics
Vol. 38, No. 3, 2010, Pages 408–433
La revue canadienne de statistique

A distribution free test to detect general dependence between a response variable and a covariate in the presence of heteroscedastic treatment effects

Haiyan WANG[1]*, Siti TOLOS[1] and Suojin WANG[2]

[1]*Department of Statistics, Kansas State University, Manhattan, KS 66503, USA*
[2]*Department of Statistics, Texas A&M University, College Station, TX 77843, USA*

Key words and phrases: Dependency measure; hypothesis testing; k-nearest neighbours; copula.

MSC 2000: Primary 62G10; secondary 62J02.

Abstract: In this paper, we present a test of independence between the response variable, which can be discrete or continuous, and a continuous covariate after adjusting for heteroscedastic treatment effects. The method involves first augmenting each pair of the data for all treatments with a fixed number of nearest neighbours as pseudo-replicates. Then a test statistic is constructed by taking the difference of two quadratic forms. The statistic is equivalent to the average lagged correlations between the response and nearest neighbour local estimates of the conditional mean of response given the covariate for each treatment group. This approach effectively eliminates the need to estimate the nonlinear regression function. The asymptotic distribution of the proposed test statistic is obtained under the null and local alternatives. Although using a fixed number of nearest neighbours pose significant difficulty in the inference compared to that allowing the number of nearest neighbours to go to infinity, the parametric standardizing rate for our test statistics is obtained. Numerical studies show that the new test procedure has robust power to detect nonlinear dependency in the presence of outliers that might result from highly skewed distributions. *The Canadian Journal of Statistics* 38: 408–433; 2010 © 2010 Statistical Society of Canada

Résumé: Dans cet article, nous présentons un test d'indépendance entre la variable réponse qui peut être discrète ou continue, et une covariable continue après avoir pris en compte l'hétéroscédasticité des effets de traitements. La méthode commence en ajoutant à chaque paire de données de tous les traitements un nombre fixe de plus proches voisins comme pseudo-réplicats. Par la suite, un test statistique est obtenu en prenant la différence de deux formes quadratiques. Cette statistique est équivalente à la corrélation décalée moyenne entre la variable réponse et l'estimation de sa moyenne conditionnelle, étant donné les covariances pour chacun des groupes de traitements, basée sur les plus proches voisins. Cette approche élimine effectivement le besoin d'estimer une fonction de régression non linéaire. La distribution asymptotique de la statistique de test proposée est obtenue sous l'hypothèse nulle et sous des hypothèses alternatives locales. Quoique l'utilisation d'un nombre fixe de plus proches voisins pose une difficulté majeure dans l'inférence, comparativement à laisser le nombre de plus proches voisins aller à l'infini, le taux paramétrique standardisé de nos statistiques de tests est atteint. Des études numériques montrent que la nouvelle procédure de test a une puissance robuste pour détecter la dépendance non linéaire en présence de valeurs aberrantes qui peuvent résulter de distributions très asymétriques. *La revue canadienne de statistique* 38: 408–433; 2010 © 2010 Société statistique du Canada

1. INTRODUCTION

Statistical tools to detect nonlinear relationships between variables are commonly needed in various practices. Let (X_{ij}, Y_{ij}) denote the jth independent observation from a univariate covariate

* *Author to whom correspondence may be addressed.*
E-mail: hwang@ksu.edu

and response in the ith treatment, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$. The a treatments can be factor level combinations from multiple factors. The response Y_{ij} can depend on X_{ij} through its distribution or some moments. Correlation-based approaches such as Pearson's correlation, Spearman's ρ, or Kendall's τ evaluate monotone relationships between two variables without accounting for the effect of factors. The method of alternating conditional expectations (ACE, Breiman & Friedman, 1985) is an extended correlation approach that transforms both the response and covariate to achieve maximum correlation. The dependence of Y_{ij} on X_{ij} in the mean through a linear or nonlinear function is extensively studied in the literature. Examples that allow hypothesis testing include likelihood methods from linear or generalized linear models, drop test (Terpstra & Mckean, 2005), generalized additive models (GAM) using a smoother such as spline or Loess (Hastie & Tibshirani, 1990), or penalized smoothing spline (Wood, 2000, 2008).

These approaches have provided flexible tools to discover the dependence between variables. However, practical data often do not satisfy the assumptions required by these methods. For example, correlation-based approaches typically are not sensitive enough to pick up nonmonotone dependence; likelihood-based methods are restrictive to the distributional assumptions; ACE assumes that conditional on the transformed covariates, the transformed response variable follows a normal distribution with constant variance; the GAM approaches are only applicable to exponential families and outliers can seriously distort the transformations leading to inaccurate inference. In a particular example we considered (see Section 3.2), all these methods except ACE found a significant relationship between the response and covariate when an outlier (influential observation) is in the data and produced a totally opposite result when the outlier is replaced by the group median response. Our simulation study found out that the type I error rate at level 0.01 in the presence of outliers produced from mixture distributions with a lognormal component can be as high as 0.206 for the GAM methods, and 0.748 for the correlation-based approaches. Robust methods valid for distributions beyond the exponential family that are resistant to outliers while maintaining high power to detect nonlinear dependence are yet to be developed.

Inherently, whether two variables are independent or not is defined through distribution functions. One may consider using mutual information (MI) as a dependency measure (D'haeseleer et al., 1998; Butte & Kohane, 2000). The MI measures the expected (under the joint distribution) log ratio of the joint probability density function (pdf) and the product of the marginal density functions. It equals zero if and only if the variables are independent. Before the MI can be used, the joint and marginal pdfs need to be estimated from the same set of data. In addition, there is no theory available to determine the threshold of significance for the dependence. Other directions for testing mutual independence without estimating the pdfs are through combinations of asymptotically independent Cramér–von Mises statistics derived from a Möbius decomposition of the empirical copula process (Deheuvels, 1981; Genest & Rémillard, 2004, and the references therein), or based on a normalized estimated distance between the joint and the marginal characteristic functions. When there are heteroscedastic treatment effects, it is not clear how to extend these tests to testing of independence in the presence of treatment effects. In addition, when the response variable is discrete, copulas are not unique and independence between the two variables is not equivalent to the independence copula ($C(u, v) = uv$) (Genest & Nešlehová, 2007).

In this paper, we present a nonparametric test to effectively detect general dependence between two variables in the presence of heteroscedastic treatment effects. A fixed number of nearest-neighbour pseudo-replicates will be used to augment each pair of treatment level and covariate value combination. We construct a test statistic through comparing two quadratic forms, both of which are estimators of a common linear combination of the variances and conditional variances. The asymptotic results are obtained under both the null hypothesis and local alternatives. Note that the regular standardizing rate for a nonparametric test statistic is N^{α}, where $0 < \alpha < 1/2$. By using a fixed number of nearest-neighbours augmentation, the standardizing rate for our test

statistics achieves the rate for parametric analysis $\sqrt{N}$ under both the null and local alternatives. Our empirical studies show that the proposed test maintains the intended type I error control while achieving competitive or better power compared to available methods when the data have unusual observations from a skewed distribution such as a lognormal distribution.

The rest of the paper is organized as follows. Section 2 will describe how to construct the test statistic and give the asymptotic results under the null hypothesis and local alternatives. Section 3 is devoted to numerical investigations including the analysis of two real data sets and simulation studies. Technical arguments are given in the Appendix.

2. MAIN RESULTS

2.1. Construction of Test Statistics

The following notation and conditions will be used throughout this manuscript. Let (X_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, be a random sample from treatment i. Suppose $Y_{ij}|X_{ij} = x \sim F_i(y|x)$ for some unknown conditional distribution function $F_i(y|x)$. Assume that the fourth conditional central moments of Y_{ij} given $X_{ij} = x$ are uniformly bounded for all i and x. Let $f_{X,i}(x)$ and $F_{X,i}(x)$ be the marginal density and distribution functions of X_{ij}. Assume $F_{X,i}(x)$ is differentiable for all values of x. Denote $\widehat{F}_{X,i}(x) = n_i^{-1} \sum_{j=1}^{n_i} I(X_{ij} \le x)$ the empirical distribution of X_{ij}. Assume that $\min_{1\le i\le a} n_i$ and $\max_{1\le i\le a} n_i$ are of the same order. Denote $\mathbf{X} = (X_{11}, \ldots, X_{1n_1}, \ldots, X_{an_a})'$ to be the vector of all covariate values.

Independence between the two variables in all treatments corresponds to the null hypothesis:

$$H_0\text{: } F_i(y|x) \text{ does not depend on } x, \text{ for all } i, y. \tag{1}$$

As pointed out by a reviewer, when the marginal distributions of both X and Y are continuous, the null hypothesis can be stated in terms of copula. The copula of the bivariate distribution $F_{X,Y,i}$ in treatment i with continuous margins $F_{X,i}$, $F_{Y,i}$ is the unique function $C : [0, 1]^2 \longrightarrow [0, 1]$ such that $F_{X,Y,i}(x, y) = C\{F_{X,i}(x), F_{Y,i}(y)\}$ (Nelson, 2006). Hence, although the margins can vary between treatments, H_0 is equivalent to $C(u, v) = uv$. This can be tested in a classical setting (same margins) using, for example, the method of Genest & Rémillard (2004). For discrete response variables, quasi-copula has been studied by some authors (cf. Genest, Quesada Molina, Rodríguez Lallena & Sempi 1999). However, independence does not imply $C(u, v) = uv$ in discrete cases due to nonuniqueness of quasi-copula and much research remains to be done (Genest & Nešlehová, 2007).

Note that the difference between this problem and the testing of independence using a single sample from the same distribution is that the data here in different treatments could have different distributions. We would like to effectively use these data so that the samples from all treatments contribute to the power of the test and therefore reduce the sample size requirement for each treatment. To achieve this, we augment each treatment under the null hypothesis to have more observations using k-nearest neighbours. For convenience, we take k to be an odd number. Specifically, treatment i and covariate value $X_{i_1 j_1}$ define a cell indexed by (i, c), where $c = \sum_{i'=1}^{i_1} n_{i'} - n_{i_1} + j_1$. In other words, for each i, there are $N = \sum_{i_1=1}^{a} n_{i_1}$ cells as i_1 goes from 1 to a and j_1 goes from 1 to n_{i_1}. We augment each cell (i, c) using observations from treatment i as follows. The set of indices for the covariate values used in the augmented cell (i, c) is denoted by C_{ic}.

(1) For $i_1 = i$, the cell (i, c) contains (X_{ij_1}, Y_{ij_1}). In addition, we select $k - 1$ pairs of other observations in treatment i whose covariate values are among the k-closest to X_{ij_1} in rank, that is, (X_{ij}, Y_{ij}) is selected for augmentation of the cell (i, c) if and only if $n_i|\widehat{F}_{X,i}(X_{ij_1}) -$

 DOI: 10.1002/cjs

TABLE 1: Illustration of data augmentation.

Original		Treatment 1							Treatment 2							
Y		1	3	4	4	9	10	13	2	5	7	7	8	13	17	18
X		0.5	1.7	2.0	2.1	4.3	4.9	6.4	1.0	2.7	3.6	3.6	3.9	6.4	8.5	8.8
Rank (X)		1	2	3	4	5	6	7	1	2	3	4	5	6	7	8
Augmented	c	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	y_1	1	3	4	4	9	10	13	3	4	9	9	9	13	13	13
Treatment 1	y_2	3	1	3	4	4	9	10	1	4	4	4	4	10	10	10
	y_3	4	4	4	9	10	13	9	4	9	10	10	10	9	9	9
	y_1	2	5	5	5	13	13	13	2	5	7	7	8	13	17	18
Treatment 2	y_2	5	2	2	2	8	8	8	5	2	7	7	13	8	13	17
	y_3	7	7	7	7	17	17	17	7	7	5	5	7	17	18	13

The top panel contains data from two treatments. The bottom panel gives the augmented observations for cell (i, c) when $k = 3$ and i refers to treatment i, $i = 1, 2$.

$\widehat{F}_{X,i}(X_{ij})| \le (k-1)/2$ for X_{ij} whose rank is at least $(k-1)/2$ away from the rank of the boundary points.

(2) For $i_1 \neq i$, that is, $(X_{i_1 j_1}, Y_{i_1 j_1})$ is not in treatment i. We first find the covariate value in treatment i that is closest to $X_{i_1 j_1}$ in rank. Denote X_{ij} to be the closest. We then select additional $k-1$ pairs of observations in treatment i such that their covariate values are among the k closest to X_{ij} in ranks centred at the rank of X_{ij}. Thus, $(X_{ij'}, Y_{ij'})$ is selected to augment cell (i, c) if $n_i|\widehat{F}_{X,i}(X_{ij'}) - \widehat{F}_{X,i}(X_{ij})| \le (k-1)/2$ if the rank of X_{ij} is at least $(k-1)/2$ away from the rank of the boundary points.

We use a small data set as an example to illustrate the augmentation method. The observations for the response and covariate in two treatments with sample sizes 7 and 8, respectively, are given in the top panel of Table 1, the augmented $k = 3$ observations are listed in the bottom panel of Table 1. For ease of explanation, the pairs of response and covariate are listed in increasing order according to the covariate values within each treatment. The response variable can be discrete and hence ties are allowed. Three typical cells are $(i, c) = (1, 1)$, $(1, 3)$, and $(2, 1)$. Cells $(1, 1)$ and $(1, 3)$ are for treatment 1 and both covariate values are originally from treatment 1. The covariate value for cell $(1, 1)$ is at the boundary and that for cell $(1, 3)$ is an interior point. For cell $(1, 1)$, the covariate value is 0.5 and its rank is 1. Therefore, the three response values 1, 3, and 4 corresponding to those covariate values ranked 1, 2, and 3 in treatment 1 are used for augmentation. For cell $(1, 3)$, the rank of the covariate value 2.0 is 3. Hence the response values 3, 4, and 4 corresponding to covariate values with rank 2, 3, or 4 are used for augmenting cell $(1, 3)$. The covariate value 0.5 corresponding to cell $(2, 1)$ is from treatment 1 but the observations to be augmented need to be from treatment 2. During the augmentation, values in a new set that contains 0.5 and all covariate values in treatment 2 are ranked. The three covariate values in treatment 2 closest to 0.5 in rank are 1.0, 2.7, and 3.6. So their corresponding response values 2, 5, and 7 are used for augmenting cell $(2, 1)$.

The first part of the augmentation is similar to the near neighbour techniques used in the regression setting (cf. Cleveland, 1979; Fan, Heckman & Wand, 1995; Hastie & Tibshirani, 1996;

Li & Gong, 2008; Wang, Akritas & Keilegom, 2008). Most of such techniques require that the number of nearest neighbours goes to infinity as the sample sizes approach infinity while Wang, Akritas & Keilegom (2008) also considered fixed number of nearest neighbours for testing of constant regression function. Our augmentation uses a special weight function defined through the empirical distribution function of the covariate. The extra augmentation in the second part is aimed at capturing possible dependence of the response variable on the covariate through its interactions with the factor. In both cases, the augmented response values in cell (i, c) are denoted as $U_{ict}, t = 1, \ldots, k$. Note that under the null hypothesis, the distribution of Y_{ij} does not depend on X_{ij}. The k-nearest neighbours are selected based on the covariate values. Therefore, the augmentation simply adds more observations under the null hypothesis. However, under the alternative, the conditional distribution of Y_{ij} depends on X_{ij}. Then such an augmentation tends to add some observations that increase the between-cell variations. The difference, $B_N - W_N$, between the average between- and within-cell variations for all treatments using the augmented observations can be used as a test statistic, where B_N and W_N are defined below with $\overline{U}_{ic\cdot} = k^{-1}\sum_{t=1}^{k} U_{ict}$, $\overline{U}_{i\cdot\cdot} = N^{-1}\sum_{c=1}^{N}\overline{U}_{ic\cdot}$:

$$
\begin{aligned}
B_N &= ka^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}\left(\overline{U}_{ic\cdot} - \overline{U}_{i\cdot\cdot}\right)^2 \\
&= ka^{-1}(N-1)^{-1}\sum_{i=1}^{a}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}\left[k^{-1}\sum_{j=1}^{n_i} Y_{ij} I\left(n_i\left|\widehat{F}_{X,i}(X_{i_1 j_1}) - \widehat{F}_{X,i}(X_{ij})\right| \le \frac{k-1}{2}\right)\right. \\
&\quad \left. -(Nk)^{-1}\sum_{i_2=1}^{a}\sum_{j_2=1}^{n_{i_2}}\sum_{j=1}^{n_i} Y_{ij} I\left(n_i\left|\widehat{F}_{X,i}(X_{i_2 j_2}) - \widehat{F}_{X,i}(X_{ij})\right| \le \frac{k-1}{2}\right)\right]^2 + O_p(N^{-1}), \\
W_N &= \{Na(k-1)\}^{-1}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t=1}^{k}(U_{ict} - \overline{U}_{ic\cdot})^2 \\
&= \{Na(k-1)\}^{-1}\sum_{i=1}^{a}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}\sum_{j=1}^{n_i}\left[Y_{ij} I\left(n_i\left|\widehat{F}_{X,i}(X_{i_1 j_1}) - \widehat{F}_{X,i}(X_{ij})\right| \le \frac{k-1}{2}\right)\right. \\
&\quad \left. - k^{-1}\sum_{j_2=1}^{n_i} Y_{ij_2} I\left(n_i\left|\widehat{F}_{X,i}(X_{i_1 j_1}) - \widehat{F}_{X,i}(X_{ij_2})\right| \le \frac{k-1}{2}\right)\right]^2 + O_p(N^{-1}).
\end{aligned}
$$

The structure of the augmented data resembles high-dimensional ANOVA (HANOVA) (Wang & Akritas, 2009) in which at least one factor has a large number of levels. The difference is that the data for HANOVA are independent while the augmented observations $\{U_{ict}, c = 1, \ldots, N, t = 1, \ldots, k\}$ are not independent since the observations are repeatedly used during the augmentation. If $k \to \infty$ with N, then techniques from nonparametric smoothing such as locally weighted kernel regression can be borrowed as the k here would play a similar role as the bandwidth for kernel regression. For a fixed finite k, the intuition that this particular approach would work comes from the fact that the asymptotic results for HANOVA can be achieved not only for independent data (Wang & Akritas, 2009) but also for data with weak dependence (Wang & Akritas, 2010) and long range dependence (Wang, Higgins & Blasi, 2010). The augmented data vectors for a treatment level viewed in the order from the smallest to largest covariate values (from left to right in Table 1) are close to weak dependent processes as $N \to \infty$. The challenge is that we have correlated

weak dependent processes instead of independent ones as in Wang & Akritas (2010) and Wang, Higgins & Blasi (2010). In this paper, the inference basically relies on a combination of counting techniques, theory for spacings of order statistics, and theory for quadratic forms.

2.2. Results Under the Null Hypothesis

The standardizing rate is $\sqrt{N}$ when k is finite. To obtain the asymptotic distribution of $\sqrt{N}(B_N - W_N)$, we first find a projection of it. Note that even though U_{ict} are independent for different i, we cannot apply Hájek's projection because such a projection will not simplify our problem since a is finite. Instead, by denoting $Z_{ict} = U_{ict} - E(U_{ict}|\mathbf{X})$, we project B_N onto

$$\text{span}\{\mathbf{Z}_c, c = 1, \ldots, N\} \text{ where } \mathbf{Z}_c = (Z_{1c1}, \ldots, Z_{ack})'. \tag{2}$$

Note that $\mathbf{Z}_c$, $c = 1, \ldots, N$, are not independent. Hence this projection is not in the traditional sense. Meanwhile, we do not have to centre B_N before the projection as required in Hájek's projection. Instead, B_N and W_N have the same expectation under the null hypothesis if the cell observations are true replicates. We proceed by partitioning the quadratic form B_N into a major summation over c and another summation over c and c', $c \neq c'$, that is, under H_0,

$$B_N = P_B(\mathbf{Z}) + S_B(\mathbf{Z}), \text{ where } \mathbf{Z} = (\mathbf{Z}_1', \ldots, \mathbf{Z}_N')',$$

and

$$P_B(\mathbf{Z}) = ka^{-1}N^{-1}\sum_{i=1}^{a}\sum_{c=1}^{N}\overline{Z}_{ic\cdot}^2, \quad S_B(\mathbf{Z}) = -ka^{-1}N^{-1}(N-1)^{-1}\sum_{i=1}^{a}\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}. \tag{3}$$

Then $P_B(\mathbf{Z})$ is a projection of B_N onto the space in (2) and $B_N - W_N = (P_B(\mathbf{Z}) - W_N) + S_B(\mathbf{Z}) = T_B + S_B(\mathbf{Z})$ where

$$\begin{aligned} T_B &= \sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}^{k}\frac{Z_{ict}Z_{ict'}}{a(k-1)N} = \sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t\neq t'}^{k}\frac{(U_{ict} - E(U_{ict}|\mathbf{X}))(U_{ict'} - E(U_{ict'}|\mathbf{X}))}{a(k-1)N} \\ &= [a(k-1)N]^{-1}\sum_{i=1}^{a}\sum_{j\neq j'}^{n_i}(Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X}))\sum_{c=1}^{N}I(j \in C_{ic})I(j' \in C_{ic}) \\ &= [a(k-1)N]^{-1}\sum_{i=1}^{a}\sum_{j\neq j'}^{n_i}(Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X}))K_{ijj'}, \end{aligned} \tag{4}$$

and

$$K_{ijj'} = \sum_{c=1}^{N}I(j \in C_{ic})I(j' \in C_{ic}). \tag{5}$$

Note that the term in (4) is closely related to the expected correlation between every pair of response values with correlation induced by their dependence on $\mathbf{X}$. The $K_{ijj'}$ in (5) serves as a weight function which connects the response locally with the empirical distribution function of X_{ij}. The T_B term in (4) is more intuitive than $\sqrt{N}(B_N - W_N)$ to evaluate the effect of X_{ij} on Y_{ij}. However, T_B cannot be calculated from the sample as $E(Y_{ij'}|\mathbf{X})$ is unknown. On the other hand, $\sqrt{N}(B_N - W_N)$ can be directly obtained from the sample.

In the following lemma, we show that $\sqrt{N}S_B(\mathbf{Z})$ is asymptotically negligible. Then we will derive the asymptotic distribution of $\sqrt{N}T_B$ by showing that it satisfies the conditions for central limit theorem for clean quadratic forms by de Jong (1987).

Lemma 1 (Projection of B_N)**.** *Let $S_B(\mathbf{Z})$ be as defined in (3). If the assumptions in Section 2.1 are satisfied, then as $N \to \infty$, $\sqrt{N} S_B(\mathbf{Z}) \to 0$ in probability.*

The proofs of Lemma 1 and Theorem 1 are given in the Appendix.

Theorem 1. *Under H_0 in (1) and the assumptions given in Section 2.1,*

$$\sqrt{N}(B_N - W_N) \to N(0, \lim_{N\to\infty} \gamma_N^2),$$

provided that the limit of γ_N^2 exists, where

$$\gamma_N^2 = \sum_{i=1}^{a} \sum_{j<j'}^{n_i} \left\{ \frac{4I(j'-j \le k-1)}{Na^2(k-1)^2} \int \sigma_i^2(u) \left[B_{ijj'}^2(u) + B_{ijj'}(u) - 2I\left(j'-j \le \frac{k-1}{2}\right)\right] \right.$$

$$\left. \times \mathrm{d}F_{X,i}(u) \int \sigma_i^2(v)\, \mathrm{d}F_{X,i}(v) \right\} + O(N^{-1}),$$

with $B_{ijj'}(u) = \sum_{i_1, i_1 \ne i}^{a} \left(\frac{n_{i_1}}{n_i} d_{i_1 i}(u) + 1 \right) [k - (j'-j)] I(j'-j \le k-1)$ and $d_{i_1 i}(u) = f_{X,i_1}(u)/f_{X,i}(u)$.

To estimate the asymptotic variance, we let $X_{i(j_*)}$ denote the order statistics of X_{ij} among covariate values in treatment i and let j_* be the ranks of X_{ij}. Then a consistent estimator of $\lim_{N\to\infty} \gamma_N^2$ is

$$\widehat{\gamma}_N^2 = \sum_{i=1}^{a} \sum_{j_*<j_*'}^{n_i} \left\{ \frac{4\, \widehat{\sigma}_i^2(X_{i(j_*)}) \widehat{\sigma}_i^2(X_{i(j_*')})}{Na^2(k-1)^2} [\widehat{B}_{ijj'}^2 + \widehat{B}_{ijj'} - 2I(j_*' - j_* \le (k-1)/2)] \right\}$$

$$\times I(j_*' - j_* \le k-1),$$

where $\widehat{B}_{ijj'} = \sum_{i_1, i_1 \ne i}^{a} (\widehat{d}_{i_1 i}(X_{i(j_*)})\, n_{i_1}/n_i + 1)[k - (j_*' - j_*)] I(j_*' - j_* \le k-1)$, and $\widehat{\sigma}_i^2(X_{ij})$ is the sample variance based on the augmented observations for the cell determined by i and X_{ij}, that is,

$$\widehat{\sigma}_i^2(X_{ij}) = \frac{1}{k-1} \left\{ \sum_{l=1}^{n_i} Y_{il}^2 I\left[\left| \widehat{F}_{X,i}(X_{il}) - \widehat{F}_{X,i}(X_{ij}) \right| \le \frac{k-1}{2n_i} \right] \right.$$

$$\left. - \frac{1}{k} \left(\sum_{l=1}^{n_i} Y_{il} I\left[\left| \widehat{F}_{X,i}(X_{il}) - \widehat{F}_{X,i}(X_{ij}) \right| \le \frac{k-1}{2n_i} \right] \right)^2 \right\}.$$

The term $k n_{i_1}/n_i \widehat{d}_{i_1 i}(X_{ij})$ is the estimate of the number of times that (X_{ij}, Y_{ij}) is selected for augmentation of the cell determined by i and $X_{i_1 j_4}$ for all $j_4 = 1, \ldots, n_{i_1}$:

$$n_{i_1}/n_i \widehat{d}_{i_1 i}(X_{ij}) = k^{-1} \sum_{j_4=1}^{n_{i_1}} I\left(|\widehat{F}_{X,i}(X_{ij}) - \widehat{F}_{X,i}(X_{i_1 j_4})| \le \frac{k-1}{2n_i} \right).$$

This is because the following equation holds when $\min_{1\le i\le a} n_i \to \infty$,

$$d_{i_1 i}(X_{ij})/n_i = k^{-1} \int I(|F_{X,i}(X_{ij} - F_{X,i}(x)| \le (k-1)/(2n_i))\, \mathrm{d}F_{X,i_1}(x) + O_p(N^{-3/2}).$$

2.3. Results Under Local Alternatives

As the response variable can be continuous or discrete, usual location alternatives are not appropriate in this setting. For example, if the observations are from a binomial or Poisson distribution, departures from the null in the mean also entail changes in the variances. Therefore, the variances of the observations under the alternatives are different from those under the null. A similar situation happens for the Gamma distribution as both the mean and variance depend on the shape and scale parameters. A change in the shape parameter would lead to a change in both the mean and the variance. Here we formulate our alternatives as departures from the null in terms of distribution functions. Consider the sequence of local alternative conditional distributions $F_{N,i}(y|x)$ that approach the marginal distribution of Y, $F_{Y,i}(y)$, in the order of $N^{-1/4}$. Define $D_{iN}(y|x) = N^{1/4}(F_{N,i}(y|x) - F_{Y,i}(y))$. Then the local alternative conditional distributions can be written as

$$F_{N,i}(y|x) = F_{Y,i}(y) + N^{-1/4} D_{iN}(y|x), \tag{6}$$

where $F_{Y,i}(y)$, $i = 1, \ldots, a$ satisfy the null hypothesis in (1) and $N^{-1/4} D_{iN}(y|x)$, $i = 1, \ldots, a$, are the deviations from the null hypothesis. This formulation allows the distribution under the alternatives to be different from that under the null.

If both X_{ij} and Y_{ij} are continuous random variables with marginal probability distributions $f_{X,i}(x)$ and $f_{Y,i}(y)$, respectively, the local alternatives can also be written in terms of generalized Farlie–Gumbel–Morgenstern copula with a sequence of dependence parameter $\theta_{i,N}$ approaching the parameter under independence at rate $N^{-1/4}$:

$$C_{i,N}(u, v) = uv[1 + \theta_{i,N} l_i(u, v)], \tag{7}$$

for some appropriate $l_i(u, v)$ whose second derivatives exist and $C_{i,N}(F_{X,i}(x), F_{X,i}(y))$ is the joint distribution of (X_{ij}, Y_{ij}). With some lengthy algebra, it can be shown that the formulation in (7) corresponds to

$$D_{iN}(y|x) = N^{1/4}\theta_{i,N} \int_{-\infty}^{y} f_{Y,i}(t) \left\{ l_i(F_{X,i}(x), F_{Y,i}(t)) + \left[F_{Y,i}(t) \frac{\partial l_i(u, v)}{\partial v} + F_{X,i}(x) \frac{\partial l_i(u, v)}{\partial u} + F_{X,i}(x) F_{Y,i}(t) \frac{\partial^2 l_i(u, v)}{\partial u \partial v} \right] \Bigg|_{u=F_{X,i}(x), v=F_{Y,i}(t)} \right\} \mathrm{d}t.$$

Example 1. The generalized Farlie–Gumbel–Morgenstern copula given in Huang & Kotz (1999) has the form $C(u, v) = uv\{1 + \theta(1 - u^p)^q(1 - v^p)^q\}$ for $0 \le u, v \le 1$, $p \ge 1$, $q \ge 1$. If the alternative hypothesis takes this form with dependence parameter $\theta_{i,N}$ for treatment i, that is,

$$C_{i,N}(u, v) = uv[1 + \theta_{i,N}(1 - u^p)^q(1 - v^p)^q], \quad p \ge 1, \ q \ge 1, \tag{8}$$

then

$$D_{iN}(y|x) = N^{1/4}\theta_{i,N}(1 - F_{X,i}^p(x))^{q-1}[1 - F_{X,i}^p(x) - pqF_{X,i}^p(x)] \times \{p^{-1} I_{F_{Y,i}^p(y)}(p^{-1}, q + 1) - qI_{F_{Y,i}^p(y)}(p^{-1} + 1, q)\},$$

where $I_t(\alpha, \beta) = \int_0^t y^{\alpha-1}(1 - y)^{\beta-1}\,\mathrm{d}y$ is the incomplete beta function for $\alpha > 0$, $\beta > 0$.

Example 2. The Clayton copula has form $C(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}$, $\theta > 0$. The alternative hypothesis from the Clayton copula with dependence parameter $\theta_{i,N}$ in treatment i is

$$C_{i,N}(u, v) = \left(u^{-\theta_{i,N}} + v^{-\theta_{i,N}} - 1\right)^{-1/\theta_{i,N}} \text{ with } \theta_{i,N} > 0. \tag{9}$$

Applying Taylor's expansion, we can write $C_{i,N}(u, v) = uv[1 + \theta_{i,N} \ln u \ln v + o(\theta_{i,N})]$. This is in the form of (7) with $l_i(u, v) = \ln u \ln v + o(1)$. This alternative hypothesis can be written in the form of (6) with

$$D_{iN}(y|x) = N^{1/4}\theta_{i,N}[\ln F_{X,i}(x) + 1]F_{Y,i}(y) \ln F_{Y,i}(y) + O(N^{1/4}\theta_{i,N}^2).$$

For generality, we state the result under the formulation in (6) since this form allows the response variable Y to be discrete while (7) only applies to continuous random variables. Let $\mathbf{H} = (H_{ict}; i = 1, \ldots, a, c = 1, \ldots, N, t = 1, \ldots, k)$ be the augmented observations under the alternatives in (6). Assume $A_i(x) = \int_{-\infty}^{\infty} y \, \mathrm{d}D_{iN}(y|x)$ and $\int_{-\infty}^{\infty} y^2 \, \mathrm{d}D_{iN}(y|x)$ are uniformly bounded for all i and x. Define

$$\eta_\phi = \lim_{N\to\infty} \frac{k}{a} \sum_{i=1}^{a} \left[\sum_{i_1=1}^{a} \frac{n_{i_1}}{N} \int_{-\infty}^{\infty} A_i^2(x) \, \mathrm{d}F_{X,i_1}(x) - \left(\sum_{i_1=1}^{a} \frac{n_{i_1}}{N} \int_{-\infty}^{\infty} A_i(x) \, \mathrm{d}F_{X,i_1}(x) \right)^2 \right].$$

Then the following result holds.

Theorem 2. *For the sequence of local alternatives $F_{N,i}(y|x)$ in (6), under the assumptions given in Section 2.1,*

$$\sqrt{N}(B_N(\mathbf{H}) - W_N(\mathbf{H})) \to N\left(ka^{-1}\eta_\phi, \lim_{N\to\infty} \gamma_{NA}^2\right),$$

provided that the limit of γ_{NA}^2 exists, where γ_{NA}^2 is defined similarly as γ_N^2 in Theorem 1 but with $\sigma_i^2(x)$ replaced by the conditional variance of Y_{ij} given $X_{ij} = x$ under the alternatives in (6).

With the local alternative hypothesis given in form (7) when the margins are absolutely continuous, the η_ϕ can be expressed in terms of $l_i(u, v)$ after writing

$$\begin{aligned}\int_{-\infty}^{\infty} A_i(x) \, \mathrm{d}F_{X,i_1}(x) = N^{1/4}\theta_{i,N} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \Bigg\{ & l_i(F_{X,i}(x), F_{Y,i}(y)) \\ & + \left[F_{Y,i}(y) \frac{\partial l_i(u, v)}{\partial v} + F_{X,i}(x) \frac{\partial l_i(u, v)}{\partial u} \right. \\ & \left. + F_{X,i}(x)F_{Y,i}(y) \frac{\partial^2 l_i(u, v)}{\partial u \partial v} \right]\Bigg|_{u=F_{X,i}(x), v=F_{Y,i}(y)} \Bigg\} \mathrm{d}F_{Y,i}(y) \, \mathrm{d}F_{X,i_1}(x),\end{aligned}$$

and

$$\begin{aligned}\int_{-\infty}^{\infty} A_i^2(x) \, \mathrm{d}F_{X,i_1}(x) = N^{1/2}\theta_{i,N}^2 \int_{-\infty}^{\infty} \Bigg[\int_{-\infty}^{\infty} y \Bigg\{ & l_i(F_{X,i}(x), F_{Y,i}(y)) + \left[F_{Y,i}(y) \frac{\partial l_i(u, v)}{\partial v} \right. \\ & \left. + F_{X,i}(x) \frac{\partial l_i(u, v)}{\partial u} + F_{X,i}(x)F_{Y,i}(y) \frac{\partial^2 l_i(u, v)}{\partial u \partial v} \right]\Bigg|_{u=F_{X,i}(x), v=F_{Y,i}(y)} \Bigg\} \mathrm{d}F_{Y,i}(y) \Bigg]^2 \mathrm{d}F_{X,i_1}(x).\end{aligned}$$

Example 1 continued. For the special copula given in (8), let the dependence parameter in treatment i be $\theta_{i,N} = \delta_i / N^{1/4}$ and

$$W_i(x) = (1 - F_{X,i}^p(x))^{q-1}[1 - F_{X,i}^p(x) - pqF_{X,i}^p(x)],$$

$$M_i = \int_{-\infty}^{\infty} \{y(1 - F_{Y,i}^p(y))^q - pqyF_{Y,i}^p(y)(1 - F_{Y,i}^p(y))^{q-1}\} \, \mathrm{d}F_{Y,i}(y).$$

Then the η_ϕ in Theorem 2 can be written as

$$\eta_{\phi,\mathrm{GFGM}} = \lim_{N\to\infty} \frac{k}{a} \sum_{i=1}^{a} \delta_i^2 M_i^2 \left\{ \sum_{i_1=1}^{a} \frac{n_{i_1}}{N} \int_{-\infty}^{\infty} W_i^2(x) \, \mathrm{d}F_{X,i_1}(x) - \left[\sum_{i_1=1}^{a} \frac{n_{i_1}}{N} \int_{-\infty}^{\infty} W_i(x) \, \mathrm{d}F_{X,i_1}(x) \right]^2 \right\}.$$

Example 2 continued. For the alternative hypothesis given by the Clayton copula in (9), let

$$W_{i,C}(x) = \ln F_{X,i}(x) + 1, \qquad M_{i,C} = \int_{-\infty}^{\infty} y \left\{ 1 + \ln[F_{Y,i}(y)] \right\} \, \mathrm{d}F_{Y,i}(y).$$

Then the η_ϕ in Theorem 2 can be written in the form below if $\delta_i = \theta_{i,N} N^{1/4}$:

$$\eta_{\phi,C} = \lim_{N\to\infty} \frac{k}{a} \sum_{i=1}^{a} \delta_i^2 M_{i,C}^2 \left\{ \sum_{i_1=1}^{a} \frac{n_{i_1}}{N} \int_{-\infty}^{\infty} W_{i,C}^2(x) \, \mathrm{d}F_{X,i_1}(x) - \left[\sum_{i_1=1}^{a} \frac{n_{i_1}}{N} \int_{-\infty}^{\infty} W_{i,C}(x) \, \mathrm{d}F_{X,i_1}(x) \right]^2 \right\}.$$

3. NUMERICAL RESULTS

The following tests will be considered for comparison with the proposed test (pNP) in this section: the score test from GAM using spline (GAM Spline) or Loess smoothing (GAM Loess) with quasilikelihood, the drop test, the likelihood ratio test from GAM using penalized splines (GAM Pspline), the likelihood ratio test from linear models (LRT), the test of association based on Pearson's correlation, Spearman's ρ, and Kendall's τ. Additional comparisons with the test of Genest & Rémillard (2004) for copulas are given in the last subsection. All the computation is carried out in R 2.8.1. Package *gam* is used for GAM spline or Loess smoothing; package *mgcv* is used for GAM Pspline; package *acepack* is used for ACE test. Command *cor.test* is used for the three correlation-based tests. The drop test is obtained from http://www.stat.wmich.edu/mckean/HMC/Rcode/ww.r. Except for the proposed test and three correlation-based tests, the significance of dependence on the covariate for the rest of the tests is obtained through comparing the log-likelihood or residual deviance from two models using an F test (see Chapter 12 of Faraway 2006), one with the covariate, treatment, and their interaction

effects, and one with only the treatment effect. Comparison with ACE is only given in Section 3.3 and is removed from other comparisons because this test consistently produces highly inflated type I error rates.

For the proposed test, we conducted simulations with the number of nearest neighbours $k = 3, 5, 7$ when $n_i = 30$ and 50 for a few data generation settings (linear alternative, quadratic alternative, binary data with log-odds to be the cosine function of the covariate). We observed a slight reduction in the type I error and slight increase in the power as k increases. However, the difference is too small to discriminate among the different k values. To be concise in presenting the simulation work and data analysis, we only provide the results for $k = 3$.

3.1. Analysis of Ozone Concentration Data—Detection of Nonlinear Dependence

The ozone data in the R faraway package contains daily measurements of ozone concentrations (O3) and eight meteorological variables in the Los Angeles basin for 330 days of 1976. We consider the relationship of ozone concentration with two other variables, day of the year (doy) and wind speed, for illustration. Wind has only 11 integer values. We split it into four intervals representing wind level to be low for values 0, 1, 2, medium for values 3, 4, 5, medium high for values 6, 7, 8, and high for values 9, 10, 11. The scatter plot of the data in Figure 1 suggests that the variable doy is related to the O3 in a quadratic relationship. However, this relationship is not evident due to large variations of O3. The variation of O3 is low at small or large values of doy and increases as O3 value approaches the peak concentration. A similar variation pattern is observed for O3 versus wind. This suggests strong heteroscedasticity for wind levels and that the conditional variance of O3 given doy changes with doy. Regression-based methods typically only evaluate if the mean regression function depends on the covariate regardless of whether the conditional variance depends on the covariate or not. In this example, even if the quadratic relationship of O3 on doy can be attributed to its dependence on the wind level, the dependence of O3 on doy through variances is still apparent. When applying all the tests mentioned in the beginning of this section, a significant doy effect on O3 is detected by the proposed test (P-value = 0), GAM with spline ($p = 9.6 \times 10^{-34}$), GAM with Loess smoothing ($p = 1.9 \times 10^{-33}$), GAM with penalized spline ($p = 6.9 \times 10^{-36}$). None of the other tests are significant (P-values are 0.390 for the drop test, 0.214 for the likelihood ratio test, 0.186 for Kendall's correlation test, 0.335 for Spearman's correlation test, and 0.220 for Pearson's correlation test). This is reasonable because this group of

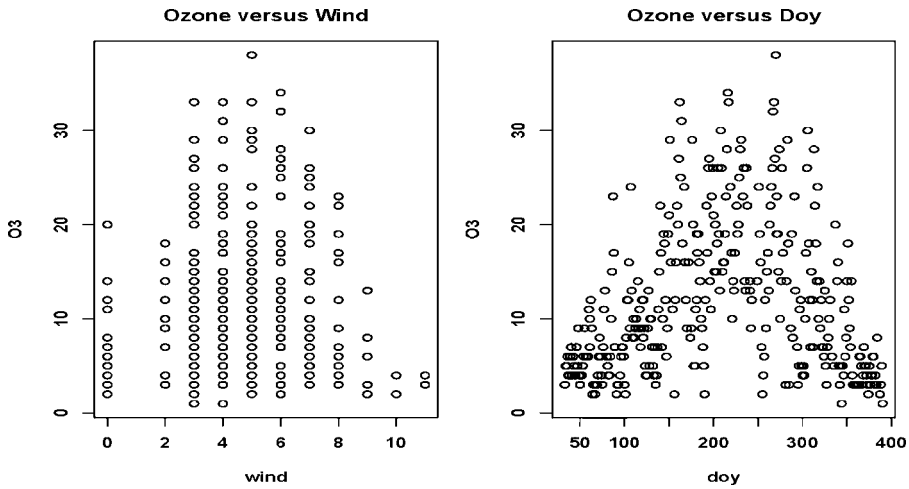

FIGURE 1: Scatter plot of zone versus wind or doy.

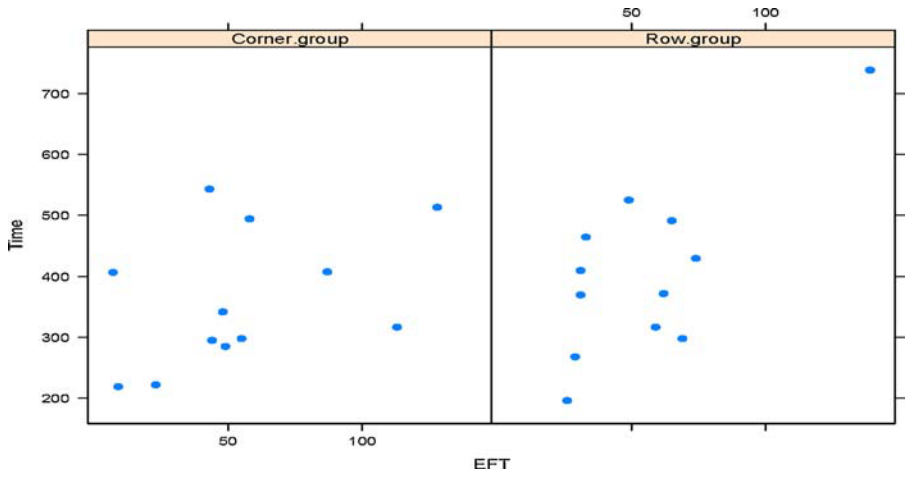

FIGURE 2: Scatter plot of time versus EFT for each instruction group. [Color figure can be viewed in the online issue, which is available at http://www.interscience.wiley.com.]

tests only access monotone relationships. Type I error analysis for these tests will be investigated in Section 3.3.

3.2. Application to EFT Study—Resistance to Outliers

In this subsection, we apply the proposed test to a data set in Aitkin et al. (1989, p. 70) that contains a sample of 24 children randomly selected from fifth-grade children attending a state primary school in a Sydney suburb. Each child was assigned to one of two experimental groups given different instructions: Corner group and Row group. The total time in seconds to conduct a test of Wechsler Intelligence Scale for Children (WISC) was recorded for each child. Each child was also tested for "field dependence" using Embedded Figures Test (EFT). The objective of the study was to evaluate if the time to complete a WISC test was affected by field dependence. Figure 2 displays time versus EFT for each group. The observation (139, 739) at the upper right corner has a large influence on linear or nonlinear regression fit.

Five different linear models were considered in Aitkin et al. (1989, pp. 83–104) with extensive discussions. They advised the readers to be cautious with small sample sizes since some of the fitted models produced conflicting interpretations. We applied all the tests considered in this section to this data set. The *P*-value for each test is given in the top row of Table 2. The proposed test is the only one that yielded nonsignificant result. All other tests are significant at 0.05 level although some are not significant at 0.01 level. The second row in Table 2 gives the *P*-values of the tests when the outlier (139, 739) is replaced by the median time in the row group. With this single change, the proposed test produced consistent results but all the other tests have a dramatic change in their *P*-values yielding nonsignificant results at 0.05 level.

TABLE 2: *P*-values for test of no association before and after the outlier is replaced by the median time in the row group.

	pNP	GAM Spline	GAM Loess	GAM Pspline	Drop Test	LRT	Kendall	Spearman	Pearson
Original data	0.729	0.033	0.041	0.013	0.035	0.002	0.017	0.021	0.006
Outlier replaced	0.369	0.385	0.347	0.145	0.305	0.306	0.059	0.067	0.178

The ground truth of whether TIME is associated with EFT or not is not known, but Aitkin et al. (1989, p. 76) did give a comment: "It is worth stressing that none of the models is a true representation of the population. If we could take a complete census of fifth-grade children in the school, and administer the EFT and WISC tests to all of them, we would find that the mean completion time for children with each EFT score in each experimental group did not lie on a straight line." We reiterate the advice of Aitkin et al. (1989) that the sample size in the data set is very small. Application to such a data set using an asymptotic test procedure needs a careful interpretation of the test result. In the next subsection, we will explore the empirical performance of these tests via simulation studies.

3.3. Simulation Studies

In this subsection, we report simulation studies conducted to investigate the type I error and power performance for the tests applied in Sections 3.1 and 3.2. The type I error estimates are obtained for data having various probabilities of containing outliers. The power is presented for one setting.

For one group, the covariate values X_{1j} were independently generated from Unif(7, 128) (7 and 128 are the minimum and maximum values of EFT in the corner group), and the response values Y_{1j} were independently generated from Unif(219, 543) (219 and 543 are the minimum and maximum values of time in the corner group). The response and covariate for the other group were generated from a mixture of a Beta and a lognormal distribution as follows:

$$\begin{cases} (r_2 - r_1)Z_{2j} \text{ where } Z_{2j} \sim \text{Beta}(1.2, 3) & \text{with probability } p_0 \\ 10Q_{2j} \text{ where } Q_{2j} \sim \text{lognormal}(1.2, 2) & \text{with probability } 1 - p_0, \end{cases} \tag{10}$$

where r_1 and r_2 are the lower and upper bounds of the observed real data. That is, $r_1 = 26$, $r_2 = 74$ for EFT were used to generate X_{2j}, and $r_1 = 196$, $r_2 = 525$ for time were used to generate Y_{2j}.

The type I error estimates at level 0.01 based on 2,000 runs for different values of p_0 and n_i are given in Table 3. It can be seen that the proposed test tends to be conservative under H_0 and is the only test that has an acceptable type I error rate under all mixing proportions. Smaller p_0 corresponds to a bigger mixture percentage for the lognormal observations which leads to a higher chance of outliers. The type I error rates for the GAM tests increase as the chance of having outliers increases. The drop test has a similar pattern as the GAM Loess test even though it has less type I errors. An opposite pattern is observed for the three correlation-based tests. The type I error for the LRT test is inflated but does not change as dramatically as the other available tests. The ACE test has consistently high type I errors (at least 0.22) for all cases. Therefore, we eliminate ACE from further comparisons.

For power comparisons, we consider departures from the null hypothesis in a quadratic relationship for the variables in one group: $X_{1j} \sim \text{Unif}(7, 128)$, $Y_{1j} = \tau(X_{1j} - E(X_{1j}))^2 + \epsilon_{1j}$, where $\epsilon_{1j} \sim \text{Unif}(-5, 15)$. For the other group, X_{2j} were generated from the mixture distribution in (10) with $p_0 = 0.1$ and Y_{2j} were independently generated from the mixture distribution in (10) with $p_0 = 0.6$ and were independently from X_{2j}.

The proportions of rejections at level 0.01 when $n_i = 12$ are presented in Figure 3 as τ increases from 0 to 2.5. The plot is busy for smaller values of τ so these values are presented also in Table 4. The power estimates were also obtained for some additional values of τ between 2.5 and 10. But the power stays at the plateau so they are not presented. The value $\tau = 0$ corresponds to the null hypothesis. The GAM Loess and GAM Pspline have similar power to the proposed test but they have inflated type I error rates. The GAM spline has lower power than the other two GAM tests. The three correlation-based tests have inflated type I errors under H_0 due to outliers and the proportion of rejections reduces to the true level as τ increases. This is because X_{ij} and Y_{ij} are uncorrelated although Y_{1j} is not independent of X_{1j} and the signal-to-noise ratio increases

TABLE 3: Proportion of rejections under H_0 at level 0.01.

	Mixture Proportion		Estimated Type I Error at 0.01 Level									
n_i	p_{0X}	p_{0Y}	pNP	ACE	GAM Spline	GAM Loess	GAM Pspline	Drop Test	LRT	Kendall	Spearman	Pearson
	0.1	0.6	0.012	0.307	0.139	0.110	0.183	0.129	0.077	0.041	0.049	0.080
	0.1	0.1	0.013	0.351	0.121	0.097	0.149	0.056	0.074	0.037	0.037	0.048
12	0.2	0.2	0.008	0.340	0.094	0.084	0.134	0.050	0.061	0.071	0.075	0.050
	0.4	0.4	0.006	0.309	0.085	0.070	0.104	0.047	0.064	0.134	0.142	0.075
	0.5	0.5	0.007	0.289	0.073	0.071	0.088	0.035	0.054	0.191	0.213	0.118
	0.6	0.6	0.010	0.282	0.044	0.039	0.066	0.034	0.043	0.256	0.292	0.153
	0.1	0.6	0.013	0.330	0.101	0.151	0.202	0.147	0.056	0.068	0.071	0.072
	0.1	0.1	0.005	0.366	0.134	0.090	0.163	0.058	0.060	0.069	0.069	0.043
20	0.2	0.2	0.005	0.342	0.130	0.081	0.156	0.052	0.058	0.119	0.128	0.048
	0.4	0.4	0.004	0.311	0.092	0.065	0.126	0.048	0.055	0.236	0.246	0.061
	0.5	0.5	0.008	0.298	0.093	0.068	0.096	0.037	0.045	0.374	0.403	0.090
	0.6	0.6	0.009	0.284	0.070	0.059	0.072	0.034	0.037	0.508	0.547	0.134
	0.1	0.6	0.006	0.270	0.082	0.150	0.206	0.176	0.047	0.117	0.120	0.070
	0.1	0.1	0.008	0.336	0.143	0.082	0.185	0.063	0.048	0.094	0.095	0.040
30	0.2	0.2	0.005	0.306	0.146	0.072	0.172	0.054	0.048	0.159	0.169	0.040
	0.4	0.4	0.004	0.260	0.111	0.062	0.126	0.046	0.049	0.416	0.426	0.050
	0.5	0.5	0.004	0.251	0.101	0.066	0.107	0.044	0.042	0.560	0.575	0.063
	0.6	0.6	0.005	0.228	0.077	0.052	0.084	0.036	0.042	0.717	0.749	0.114

DOI: 10.1002/cjs

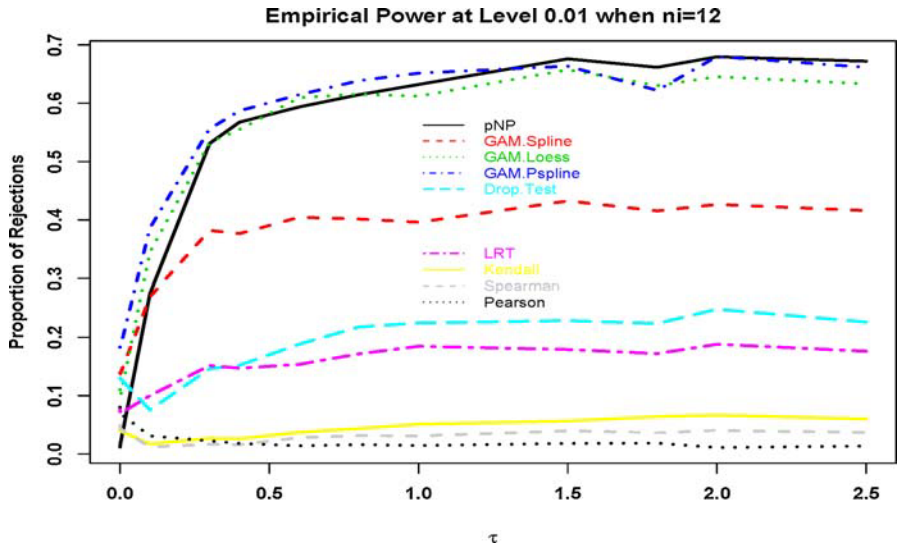

FIGURE 3: Empirical power at level 0.01 based on 2, 000 runs when the data were from a mixture of beta and lognormal distribution with $n_i = 12$. [Color figure can be viewed in the online issue, which is available at http://www.interscience.wiley.com.]

as τ increases. The power of the drop test and LRT test lies in between the GAM Spline and the three correlation-based tests. For $n_i = 20$ or 30, the proportions of rejections for all tests are also reported in Table 4. In this simulation setting, the proposed test outperforms all other tests in terms of both the estimated type I error and power. We remark that the GAM tests were developed for the exponential family and the mixture component log-normal distribution is not a member of the exponential family. This explains the observed lower power for the GAM tests.

In summary, our simulation studies suggest that the proposed test not only offers reliable type I error estimates for our simulated data in the presence of outliers which lead to inflated type I error estimates for the GAM and other tests, but also maintains a high power to detect nonlinear dependence.

3.4. Performance on Copulas

Upon the request of a reviewer, we also considered numerical comparisons for data generated with copulas. Dependence pattern from the following three copulas were considered: the Frank copula that has symmetric dependence in both tails, the Clayton copula that has stronger dependence in the left tail than in the right tail, and the Gumbel copula that has strong right tail dependence and weak left tail dependence.

We simulated 30 pairs of uniform random variables for each treatment from the Gumbel, Frank, and Clayton copulas using the *archmCopula* command in R package *copula*. Denote the dependence parameter for treatment i as θ_i, $i = 1, 2$. The data under the null hypothesis of independence were generated with $\theta_1 = \theta_2 = 1$ for Gumbel copula, $\theta_1 = \theta_2 = 0$ for the Frank copula. The margins of the Clayton copula become independent as the dependence parameter approaches zero. We take $\theta_1 = \theta_2 = 10^{-8}$. The data under the alternatives were generated with $\theta_2 = 2$ and θ_1 given in Table 5. As the dependence parameter moves away from its value under the null, the dependence increases.

For linear dependence, the uniform variables were converted to standard normal variables directly via the standard normal quantile function,

$$X_{ij} = \Phi^{-1}(u_{ij}), \qquad Y_{ij} = \Phi^{-1}(v_{ij}) \quad \text{for } i = 1, 2. \tag{11}$$

TABLE 4: Empirical power at level 0.01 based on 2, 000 runs.

		Estimated Power at 0.01 Level								
n_i	τ	pNP	GAM Loess	GAM Spline	GAM Pspline	Drop Test	LRT	Kendall	Spearman	Pearson
12	0	0.012	0.110	0.139	0.183	0.129	0.077	0.041	0.049	0.080
	0.1	0.274	0.345	0.269	0.386	0.075	0.100	0.017	0.012	0.031
	0.3	0.531	0.532	0.382	0.556	0.146	0.151	0.027	0.016	0.022
	0.4	0.568	0.555	0.376	0.587	0.152	0.147	0.026	0.016	0.018
20	0	0.013	0.101	0.151	0.202	0.147	0.056	0.068	0.071	0.072
	0.1	0.485	0.423	0.273	0.435	0.085	0.080	0.024	0.021	0.031
	0.2	0.692	0.599	0.336	0.556	0.133	0.106	0.029	0.021	0.019
	0.3	0.775	0.651	0.364	0.608	0.160	0.116	0.032	0.025	0.012
	0.5	0.853	0.768	0.430	0.666	0.215	0.117	0.041	0.024	0.008
	1.0	0.927	0.831	0.440	0.736	0.248	0.128	0.064	0.042	0.010
	1.5	0.959	0.869	0.460	0.730	0.246	0.121	0.074	0.053	0.007
	2.0	0.964	0.875	0.469	0.730	0.245	0.126	0.102	0.074	0.007
	2.5	0.976	0.883	0.478	0.751	0.260	0.126	0.084	0.058	0.006
	3.0	0.973	0.883	0.455	0.740	0.254	0.122	0.073	0.048	0.004
30	0	0.006	0.082	0.150	0.206	0.176	0.047	0.117	0.120	0.070
	0.1	0.438	0.458	0.263	0.453	0.098	0.061	0.022	0.016	0.047
	0.2	0.682	0.666	0.340	0.536	0.127	0.077	0.032	0.020	0.027
	0.3	0.778	0.750	0.372	0.580	0.176	0.086	0.040	0.026	0.018
	0.5	0.879	0.830	0.415	0.643	0.232	0.102	0.056	0.034	0.005
	1.0	0.955	0.895	0.446	0.688	0.270	0.105	0.090	0.064	0.007
	1.5	0.972	0.909	0.452	0.701	0.283	0.110	0.093	0.066	0.004
	2.0	0.989	0.935	0.459	0.708	0.275	0.104	0.107	0.076	0.002
	2.5	0.991	0.936	0.457	0.704	0.299	0.123	0.116	0.084	0.002
	3.0	0.989	0.936	0.464	0.694	0.275	0.103	0.138	0.088	0.002

For quadratic dependence, the uniform variables were first transformed to a quadratic form before further conversion via the standard normal quantile function,

$$X_{ij} = \Phi^{-1}(u_{ij}), \qquad Y_{ij} = \Phi^{-1}(2(v_{ij} - 0.5)^2 + 0.25) \quad \text{for } i = 1, 2. \tag{12}$$

The proportions of rejections for the proposed test (pNP) and the test by Genest & Rémillard (2004) (GR) are reported in Table 5. The Genest & Rémillard (2004) test was carried out with the *indepTest* command in the *copula* package.

Note that the dependence between X_{ij} and Y_{ij} were specified through the copula. For data obtained through transformation in (11), the GR test has better power than the proposed test. The transformations in (11) and (12) do not remove the dependence between X_{ij} and Y_{ij}. However, it can be seen from Table 5 that the quadratic transformation in (12) makes it harder to detect the

TABLE 5: Empirical power at level 0.01 for the Genest & Rémillard (2004) test (GR) and the proposed test (pNP) based on 1, 000 runs when $n_i = 30$.

Gumbel					Clayton					Frank				
	Quadratic		Linear			Quadratic		Linear			Quadratic		Linear	
θ_1	GR	pNP	GR	pNP	θ_1	GR	pNP	GR	pNP	θ_1	GR	pNP	GR	pNP
H_0	0.011	0.007	0.008	0.003	H_0^*	0.010	0.010	0.009	0.012	H_0	0.008	0.009	0.009	0.006
1.0	0.020	0.054	0.493	0.574	1	0.121	0.111	0.977	0.879	5	0.014	0.033	0.883	0.656
1.5	0.052	0.113	0.975	0.853	1.2	0.136	0.133	0.995	0.928	6	0.023	0.053	0.945	0.785
2.0	0.102	0.228	1.000	0.969	1.5	0.168	0.167	0.998	0.949	8	0.021	0.126	0.981	0.908
2.5	0.156	0.378	1.000	0.992	2	0.230	0.225	0.997	0.975	12	0.046	0.382	0.998	0.975
3.0	0.192	0.555	1.000	0.998	2.5	0.208	0.298	1.000	0.993	15	0.049	0.537	0.996	0.983
3.5	0.257	0.677	1.000	0.998	3	0.243	0.342	1.000	0.998	20	0.103	0.729	0.999	0.988
4.0	0.347	0.781	1.000	1.000	3.5	0.246	0.405	1.000	0.998	30	0.145	0.893	0.999	0.996
5.0	0.441	0.871			4	0.289	0.470	1.000	1.000	40	0.164	0.920	1.000	0.993
6.0	0.522	0.954			5	0.346	0.608			60	0.201	0.940	1.000	0.994
8.0	0.571	0.964			6	0.366	0.687			150	0.180	0.949	1.000	0.997
9.0	0.605	0.972			8	0.414	0.809							
10.0	0.596	0.971			12	0.505	0.928							
12.0	0.664	0.982			15	0.584	0.955							
15.0	0.678	0.992			60	0.661	0.991							

H_0 refers to $\theta_1 = \theta_2 = 1$ for the Gumbel copula, $\theta_1 = \theta_2 = 0$ for the Frank copula. H_0^* refers to $\theta_1 = \theta_2 = 10^{-8}$ for the Clayton copula as an approximation to H_0. For all data under the alternatives, $\theta_2 = 2$.

DOI: 10.1002/cjs

dependence between X_{ij} and Y_{ij}. This is particularly true for the GR test. For the data converted by formula in (12), the proposed test has much better power than the GR test.

4. SUMMARY

In this paper, we developed a nonparametric test of independence between two variables after adjusting for heteroscedastic treatment effects. The response variable can be discrete or continuous but the covariate is required to be absolutely continuous. A test statistic based on augmented observations using a fixed number of nearest neighbours is constructed. Counting techniques and spacings for order statistics are used in conjunction with the asymptotic theory for clean quadratic forms to obtain the asymptotic distribution of the test statistic under both the null hypothesis and local alternatives. As the inference is distribution free, it can be applied to a wide range of data including nonexponential families that have a high chance of unusual observations. Our numerical studies confirm that the proposed method offers a powerful alternative approach to existing methods in detecting general dependence between two variables in the presence of treatment effects.

APPENDIX

Lemma 2. *Under the conditions of Theorem 1, as* $N \to \infty$, $var(\sqrt{N}T_B) - \gamma_N^2 \to 0$.

The proof of this lemma is given after the proof of the theorems.

Sketch Proof of Theorem 1. By Lemma 1, $\sqrt{N}(B_N - W_N)$ has the same asymptotic distribution as $\sqrt{N}T_B$. The asymptotic variance of this statistic is obtained in Lemma 2.

Here we only need to show the asymptotic normality for the test statistic. Let $t^{(2)}_{iji'j'} = (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{i'j'} - E(Y_{i'j'}|\mathbf{X}))K_{ijj'}$, where $K_{ijj'}$ is defined in (5), and write

$$\sqrt{N}T_B = \frac{\sqrt{N}}{Na(k-1)} \sum_{i,i',j,j'} t^{(2)}_{iji'j'} I(i = i')I(j \neq j') = \sum_{1 \le l_1 \le N} \sum_{1 \le l_2 \le N} V_{l_1 l_2},$$

where $l_1 = l(i, j)$ and $l_2 = l(i, j')$ are defined through a one to one index mapping function

$$l(i, j) = \begin{cases} j & \text{for } i = 1 \\ \sum_{i_2=1}^{i-1} n_{i_2} + j & \text{for } i > 1, \end{cases} \tag{13}$$

and

$$V_{l_1 l_2} = \begin{cases} \frac{\sqrt{N}}{Na(k-1)}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))\, K_{l_1 l_2} & \text{for } i = i' \text{ and } j \neq j' \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Here $K_{l_1 l_2}$ is the same as $K_{ijj'}$ but using index l_1, l_2:

$$K_{l_1 l_2} = \begin{cases} \sum_{i_1}^{a} \sum_{j_1}^{n_{i_1}} I(l_1 \in C_{iX_{i_1 j_1}}) I(l_2 \in C_{iX_{i_1 j_1}}) & \text{for } i = 1 \\ \sum_{i_1}^{a} \sum_{j_1}^{n_{i_1}} I(\sum_{i_2=1}^{i-1} n_{i_2} + l_1 \in C_{iX_{i_1 j_1}}) I(\sum_{i_2=1}^{i-1} n_{i_2} + l_2 \in C_{iX_{i_1 j_1}}) & \text{for } i > 1. \end{cases}$$

Note that $V_{l_1 l_2} = V_{l_2 l_1}$. Therefore, $\sqrt{N}T_B = 2\sum_{1 \le l_1 < l_2 \le N} V_{l_1 l_2}$ is a clean quadratic form as in de Jong (1987). In order to show that $\text{var}(\sqrt{N}T_B)^{-1/2}\sqrt{N}T_B \xrightarrow{\mathcal{L}} N(0, 1)$, we will show that Proposition 3.2 in de Jong (1987) can be applied, that is, we will show that G_1, G_2, and G_3

(defined below) are of smaller order than that of $[\mathrm{var}(\sqrt{N}T_B)]^4 = O(1)$. Let $l_3 = l(i, j_3)$, and $l_4 = l(i, j_4)$.

Define

$G_1 = \sum_{1\le l_1<l_2\le N} E(V_{l_1l_2}^4)$, $G_2 = \sum_{1\le l_1<l_2<l_3\le N}\{E(V_{l_1l_2}^2 V_{l_1l_3}^2) + E(V_{l_2l_1}^2 V_{l_2l_3}^2) + E(V_{l_3l_1}^2 V_{l_3l_2}^2)\}$, and $G_3 = \sum_{1\le l_1<l_2<l_3<l_4\le N}\{E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) + E(V_{l_1l_2}V_{l_1l_4}V_{l_3l_2}V_{l_3l_4}) + E(V_{l_1l_3}V_{l_1l_4}V_{l_2l_3}V_{l_2l_4})\}$.

First, we show that $G_1 = o(1)$. It suffices to consider only the case that $V_{l_1l_2} \neq 0$. When the response has finite conditional fourth moment, there exists some finite $M_0 > 0$, such that

$$\begin{aligned} E(V_{l_1l_2}^4 I(V_{l_1l_2}\neq 0)) &= \frac{16}{N^2a^4(k-1)^4}E\{E[((Y_{l_1}-E(Y_{l_1}|\mathbf{X}))(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))|\mathbf{X})(K_{l_1l_2})]^4\} \\ &= \frac{16}{N^2a^4(k-1)^4}E\{E[((Y_{l_1}-E(Y_{l_1}|\mathbf{X}))^4 E(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))^4|\mathbf{X})K_{l_1l_2}^4]\} \\ &\le \frac{M_0}{N^2a^4(k-1)^4}E(K_{l_1l_2}^4). \end{aligned}$$

Thus, we have

$E(K_{l_1l_2}^4) = E(K_{ijj'}^4) = E\{E[\sum_{c=1}^N I(j\in C_{ic})I(j'\in C_{ic})]^4|X_{ij}, X_{ij'}\} = E(D_1 + D_2 + D_3 + D_4)$, where $D_1 = E(\sum_{c=1}^N I(j\in C_{ic})I(j'\in C_{ic})|X_{ij}, X_{ij'})$,

$D_2 = E[\sum_{c_1\neq c_2} I(j\in C_{ic_1})I(j'\in C_{ic_1})I(j\in C_{ic_2})I(j'\in C_{ic_2})|X_{ij}, X_{ij'}]I(c_1\neq c_2)$,

$D_3 = E\{E[\sum_{c_1\neq c_2\neq c_3} I(j\in C_{ic_1})I(j'\in C_{ic_1})I(j\in C_{ic_2})I(j'\in C_{ic_2})I(j\in C_{ic_3})I(j'\in C_{ic_3})|X_{ij}, X_{ij'}]\}$,

$D_4 = E\{E[\sum_{c_1}\sum_{c_2}\sum_{c_3}\sum_{c_4} I(j\in C_{ic_1})I(j'\in C_{ic_1})I(j\in C_{ic_2})I(j'\in C_{ic_2})I(j\in C_{ic_3})I(j'\in C_{ic_3})]$
$I(j\in C_{ic_4})I(j'\in C_{ic_4})|X_{ij}, X_{ij'}\}I(c_1\neq c_2\neq c_3\neq c_4)$.

It can be shown that the D_m, $m = 1, 2, 3, 4$, are of $O_p(1)$ and thus $E(K_{l_1l_2}^4) = O(1)$. In fact, $K_{ijj'}$ are bounded counts, so that

$$D_1 = E(K_{ijj'}|X_{ij}, X_{ij'}) = O_p(1)I(j'_* - j_* \le (k-1)). \tag{15}$$

The result in (15) can be obtained from (22). Next we have $D_2 \le E^2(K_{ijj'}|X_{ij},\ X_{ij'}) = O(1)I(j'_* - j_* \le (k-1))$. Similarly $D_3 \le E^3(K_{ijj'}\,|\,X_{ij},\ X_{ij'}) = O(1)I(j'_* - j_* \le (k-1))$. Lastly, $D_4 \le E^4(K_{ijj'}|X_{ij}, X_{ij'}) = O(1)I(j'_* - j_* \le (k-1))$. Therefore, $E(K_{l_1l_2})^4 = O(1)I(l_{2*} - l_{1*} \le (k-1))$, and $E(V_{l_1l_2}^4) = O(N^{-2})I(l_{2*} - l_{1*} \le k-1)$, where $l_{1*} = l(i, j*)$, $l_{2*} = l(i, j'*)$. Thus, $G_1 = O(N^{-1}) = o(1)$.

Next, we want to show that the order of G_2 is $o(1)$ when $l_1 < l_2 < l_3$, that is, $i = i'$, $j < j'$, and $j < j_3$. We first show that $E(K_{l_1l_2}^2 K_{l_1l_3}^2)$ is bounded and $E(V_{l_1l_2}^2 V_{l_1l_3}^2)$ is of order $O(N^{-2})$. By Equation (14),

$$\begin{aligned} E(V_{l_1l_2}^2 V_{l_1l_3}^2) &= E\{E[16(N^2a^4k^4)^{-1}(Y_{l_1}-E(Y_{l_1}|\mathbf{X}))^2(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))^2 K_{l_1l_2}^2 \\ &\quad\times(Y_{l_1}-E(Y_{l_1}|\mathbf{X}))^2(Y_{l_3}-E(Y_{l_3}|\mathbf{X}))^2 K_{l_1l_3}^2|\mathbf{X}]\} \\ &= E\{16(N^2a^4k^4)^{-1}E[(Y_{l_1}-E(Y_{l_1}|\mathbf{X}))^4|\mathbf{X}]E[(Y_{l_2}-E(Y_{l_2}|\mathbf{X}))^2|\mathbf{X}]K_{l_1l_2}^2 \\ &\quad E[(Y_{l_3}-E(Y_{l_3}|\mathbf{X}))^2|\mathbf{X}]K_{l_1l_3}^2\} \\ &\quad\le M_2N^{-2}a^{-4}k^{-4}E(K_{l_1l_2}^2K_{l_1l_3}^2) \text{ for some finite } M_2 > 0. \end{aligned}$$

 DOI: 10.1002/cjs

Applying the Cauchy–Schwartz inequality, we obtain

$$E(K_{l_1l_2}^2 K_{l_1l_3}^2) \le [E(K_{l_1l_2}^4)E(K_{l_1l_3}^4)]^{1/2} = O(1)I(l_2^* - l_1^* \le k-1)I(l_3^* - l_1^* \le k-1). \quad (16)$$

The last equation in (16) follows from the previous result that $E(K_{l_1l_2}^4) = O(1)$. Therefore,

$$E(V_{l_1l_2}^2 V_{l_1l_3}^2) = O(N^{-2})I(l_2^* - l_1^* \le k-1)I(l_3^* - l_1^* \le k-1).$$

It can be shown similarly that the order for $E(V_{l_2l_1}^2 V_{l_2l_3}^2)$ and $E(V_{l_3l_1}^2 V_{l_3l_2}^2)$ is $O(N^{-2})$. Therefore, $G_2 = O(N^{-1})I(l_2^* - l_1^* \le k-1)I(l_3^* - l_1^* \le k-1) = o(1)$.

Next, we show that $G_3 = o(1)$ for the case $i = i'$, $j < j' < j_3 < j_4$, that is, $l_1 < l_2 < l_3 < l_4$. This involves first showing that $E(K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3})$ is bounded and $E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) = O(N^{-2})$. Consider

$$\begin{aligned}
&E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) \\
&\quad = E\{E[N^{-2}a^{-4}k^{-4}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))K_{l_1l_2}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))(Y_{l_3} - E(Y_{l_3}|\mathbf{X})) \\
&\qquad K_{l_1l_3}(Y_{l_4} - E(Y_{l_4}|\mathbf{X}))(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))K_{l_4l_2}(Y_{l_4} - E(Y_{l_4}|\mathbf{X}))(Y_{l_3} - E(Y_{l_3}|\mathbf{X}))K_{l_4l_3}|\mathbf{X}]\} \\
&\quad = E\{E[N^{-2}a^{-4}k^{-4}(Y_{l_1} - E(Y_{l_1}|\mathbf{X}))^2(Y_{l_2} - E(Y_{l_2}|\mathbf{X}))^2(Y_{l_3} - E(Y_{l_3}|\mathbf{X}))^2(Y_{l_4} - E(Y_{l_4}|\mathbf{X}))^2] \\
&\qquad K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3}\} \le M_3 N^{-2}a^{-4}k^{-4}E(K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3}).
\end{aligned}$$

This leads to

$$\begin{aligned}
&E(K_{l_1l_2}K_{l_1l_3}K_{l_4l_2}K_{l_4l_3}) \le [E(K_{l_1l_2}K_{l_1l_3})^2 E(K_{l_4l_2}K_{l_4l_3})^2]^{1/2} \\
&\quad = O(1)I(l_2^* - l_1^* \le k-1)I(l_3^* - l_1^* \le k-1)I(l_2^* - l_4^* \le k-1)I(l_3^* - l_4^* \le k-1).
\end{aligned}$$

It can be shown similarly that $E(K_{l_1l_2}K_{l_1l_4}K_{l_3l_2}K_{l_3l_4})$ and $E(K_{l_1l_3}K_{l_1l_4}K_{l_2l_3}K_{l_2l_4})$ are also of $O(1)$. Therefore, $E(V_{l_1l_2}V_{l_1l_3}V_{l_4l_2}V_{l_4l_3}) = O(N^{-2})I(l_2^* - l_1^* \le k-1)I(l_3^* - l_1^* \le k-1)I(l_2^* - l_4^* \le k-1)I(l_3^* - l_4^* \le k-1)$. Thus $G_3 = O(N^{-1}) = o(1)$. ■

Sketch Proof of Theorem 2. Let $Z_{ict}^{(a)} = H_{ict} - E(H_{ict}|\mathbf{X})$. Then

$$\begin{aligned}
B_N(\mathbf{H}) &= ka^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}\left(\overline{H}_{ic\cdot} - \overline{H}_{i\cdot\cdot}\right)^2 \\
&= ka^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}\{E(\overline{H}_{ic\cdot}|\mathbf{X}) + [\overline{H}_{ic\cdot} - E(\overline{H}_{ic\cdot}|\mathbf{X})] - E(\overline{H}_{i\cdot\cdot}|\mathbf{X}) \\
&\quad - [\overline{H}_{i\cdot\cdot} - E(\overline{H}_{i\cdot\cdot}|\mathbf{X})]\}^2\} \\
&= ka^{-1}(N-1)^{-1}\left\{\sum_{c=1}^{N}\sum_{i=1}^{a}[\overline{Z}_{ic\cdot}^{(a)} - \overline{Z}_{i\cdot\cdot}^{(a)}]^2 + \sum_{c=1}^{N}\sum_{i=1}^{a}[E(\overline{H}_{ic\cdot}|\mathbf{X}) - E(\overline{H}_{i\cdot\cdot}|\mathbf{X})]^2\right. \\
&\quad \left. + 2\sum_{c=1}^{N}\sum_{i=1}^{a}[E(\overline{H}_{ic\cdot}|\mathbf{X}) - E(\overline{H}_{i\cdot\cdot}|\mathbf{X})][\overline{Z}_{ic\cdot}^{(a)} - \overline{Z}_{i\cdot\cdot}^{(a)}]\right\} \\
&= B_N(\mathbf{Z}^{(a)}) + ka^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}[E(\overline{H}_{ic\cdot}|\mathbf{X}) - E(\overline{H}_{i\cdot\cdot}|\mathbf{X})]^2 + O_p(N^{-3/4}).
\end{aligned}$$

Similarly, $W_N(\mathbf{H}) = W_N(\mathbf{Z}^{(a)}) + \{Na(k-1)\}^{-1}\sum_{i=1}^{a}\sum_{c=1}^{N}\sum_{t=1}^{k}(E(H_{ict}|\mathbf{X}) - E(\overline{H}_{ic\cdot}|\mathbf{X}))^2$. Note that

$$E(H_{ict}|\mathbf{X}) - E(\overline{H}_{ic\cdot}|\mathbf{X}) = N^{-1/4}\left[I(j \in C_{ic})A_i(X_{ij}) - k^{-1}\sum_{j'=1}^{n_i} I(j' \in C_{ic})A_i(X_{ij'})\right]$$
$$= o_p(N^{-1/4}), \tag{17}$$

where the last equality is a result of the mid-value theorem or first-order Taylor expansion of $A_i(X_{ij'})$ at X_{ij} combined with the property of order statistics, and the assumption of uniformly bounded condition for $A_i(x)$ and $A_i^{(2)}(x)$. The result of Theorem 1 can be applied to $\sqrt{N}(B_N(\mathbf{Z}^{(a)}) - W_N(\mathbf{Z}^{(a)}))$. Therefore, to show Theorem 2, it remains to show that $kN^{1/2}a^{-1}(N-1)^{-1}\sum_{c=1}^{N}\sum_{i=1}^{a}[E(\overline{H}_{ic\cdot}|\mathbf{X}) - E(\overline{H}_{i\cdot\cdot}|\mathbf{X})]^2 \xrightarrow{p} k\eta_\phi/a$. Similar to the argument in (17), we write

$$\frac{k\sqrt{N}}{a(N-1)}\sum_{i=1}^{a}\sum_{c=1}^{N}\left[E(\overline{H}_{ic\cdot}|\mathbf{X}) - E(\overline{H}_{i\cdot\cdot}|\mathbf{X})\right]^2$$
$$= \frac{k}{a(N-1)}\sum_{i=1}^{a}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}\left\{k^{-1}\sum_{j'=1}^{n_i}A_i(X_{ij'})\left[I\left(n_i\left|\widehat{F}_{X,i}(X_{ij'}) - \widehat{F}_{X,i}(X_{i_1j_1})\right| \le \frac{k-1}{2}\right)\right.\right.$$
$$\left.\left. - N^{-1}\sum_{i_2=1}^{a}\sum_{j_2=1}^{n_{i_2}} I\left(n_i\left|\widehat{F}_{X,i}(X_{ij'}) - \widehat{F}_{X,i}(X_{i_2j_2})\right| \le \frac{k-1}{2}\right)\right]\right\}^2$$
$$= \frac{k}{a(N-1)}\sum_{i=1}^{a}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}A_i^2(X_{i_1j_1}) - \frac{k}{a}\sum_{i=1}^{a}\left(N^{-1}\sum_{i_1=1}^{a}\sum_{j_1=1}^{n_{i_1}}A_i(X_{i_1j_1})\right)^2 + o_p(1)$$
$$= \frac{k}{a}\sum_{i=1}^{a}\left[\sum_{i_1=1}^{a}\frac{n_{i_1}}{N}\int A_i^2(x)\,\mathrm{d}F_{X,i_1}(x) - \left(\sum_{i_1=1}^{a}\frac{n_{i_1}}{N}\int A_i(x)\,\mathrm{d}F_{X,i_1}(x)\right)^2\right] + o_p(1).$$

Thus the proof is completed. ■

Proof of Lemma 1. It is sufficient to show that $E(\sqrt{N}S_B(\mathbf{Z})) \to 0$ and $\mathrm{var}(\sqrt{N}S_B(\mathbf{Z})) \to 0$. Note that $E(S_B(\mathbf{Z})) = -k\sum_{i=1}^{a}\sum_{c\neq c'}^{N}E\{E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X})\}/[aN(N-1)]$. Since $E(Y_{ij}^2|X_{ij})$ is uniformly bounded for all i, j. So there exists some finite $M_1 > 0$, $|E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X})| \le k^{-2}\sum_{t=1}^{k}\sum_{t'=1}^{k}|E(Z_{ict}Z_{ic't'}|\mathbf{X})| \le k^{-2}\sum_{t=1}^{k}\sum_{t'=1}^{k}[E(Z_{ict}^2|\mathbf{X})E(Z_{ict'}^2|\mathbf{X})]^{1/2} \le M_1$. When the observations in cell (i, c) and cell (i, c') do not have overlap, $E(Z_{ic\cdot}Z_{ic'\cdot}) = E(Z_{ic\cdot})E(Z_{ic'\cdot}) = 0$, which gives the following result:

$$\sum_{i=1}^{a}\sum_{c\neq c'}^{N}E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X}) = O_p\left(\sum_{i=1}^{a}\sum_{c\neq c'}^{N}I(|c'-c| \le k)E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}|\mathbf{X})\right) = O_p(N),$$

which then implies that $E(N^{-1/2}S_B(\mathbf{Z})) = O(N^{-1/2}) \to 0$ as $N \to \infty$.

Next we will show that $\text{var}(N^{-1/2}S_B(\mathbf{Z}))$ goes to 0 as $N \to \infty$. Since $E(\sqrt{N}S_B(\mathbf{Z}))$ goes to 0, it remains to show that $E(\sum_{c\neq c'}^{N} \overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot})^2/N^3 \to 0$. Observe that

$$
\begin{aligned}
E\left(\sum_{c\neq c'}^{N} \overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right)^2 &\leq \sum_{c\neq c'}^{N}\sum_{c_1\neq c_1'}^{N} |E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{ic_1\cdot}\overline{Z}_{ic_1'\cdot})| \\
&\leq \sum_{c,c',c_1,c_1'}^{N} |E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{ic_1\cdot}\overline{Z}_{ic_1'\cdot})|(2I_1(c,c',c_1,c_1') \\
&\quad +3I_2(c,c',c_1,c_1') + 3I_3(c,c',c_1,c_1') + 4I_4(c,c',c_1,c_1')\big|,
\end{aligned}
$$

where $I_1(\cdot)$ is the indicator function for cases that either c, c', c_1, c_1' fall into three nonoverlapping cells where two nonoverlapping cells contain one of the c's and one of the cells contains two members of c, c', c_1, c_1'; $I_2(\cdot)$ is the indicator function for cases that c, c', c_1, c_1' are evenly divided into two nonoverlapping cells; $I_3(\cdot)$ is the indicator function for cases that c, c', c_1, c_1' are in two nonoverlapping cells, such that one cell contains three of the c's and the other contains one of the c's. Finally $I_4(\cdot)$ is the indicator function for cases that c, c', c_1, c_1' are all in the same cell. Note that $E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{ic_1\cdot}\overline{Z}_{ic_1'\cdot})I_1(c,c',c_1,c_1') = 0$ and $E(\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\overline{Z}_{ic_1\cdot}\overline{Z}_{ic_1'\cdot})I_2(c,c',c_1,c_1') = 0$ since the observations in nonoverlapping cells are independent. Therefore,

$$
\begin{aligned}
\text{var}(\sqrt{N}S_B(\mathbf{Z})) &= \frac{k^2}{a^2N(N-1)^2}E\left(\sum_{c\neq c'}^{N}\overline{Z}_{ic\cdot}\overline{Z}_{ic'\cdot}\right)^2 \\
&\leq \frac{k^2}{a^2N(N-1)^2}\{O[N^2] + O[N^2]\} = O(N^{-1}),
\end{aligned}
$$

and the proof is completed. ■

Proof of Lemma 2. Write $\text{var}(\sqrt{N}T_B) = E(\text{var}(\sqrt{N}T_B|\mathbf{X})) + \text{var}(\sqrt{N}E(T_B|\mathbf{X}))$. We will show that $\text{var}(\sqrt{N}E(T_B|\mathbf{X})) = 0$ and $E(\text{var}(\sqrt{N}T_B|\mathbf{X})) - \gamma_N^2 \to 0$.

It is clear that $\text{var}(\sqrt{N}E(T_B|\mathbf{X})) = 0$ since by the definition of T_B in (4),

$$
E(\sqrt{N}T_B|\mathbf{X}) = E\left(\frac{N^{-1/2}}{a(k-1)}\sum_{i=1}^{a}\sum_{j\neq j'}(Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X}))|\mathbf{X}\right)K_{ijj'} = 0 \quad a.s.
$$

Next, we will show that $E(\text{var}(\sqrt{N}T_B|\mathbf{X})) - \gamma_N^2 \to 0$. Let $t_{ijj'} = (Y_{ij} - E(Y_{ij}|\mathbf{X}))(Y_{ij'} - E(Y_{ij'}|\mathbf{X}))K_{ijj'}$. Then

$$
\begin{aligned}
Na^2(k-1)^2E(\text{var}(\sqrt{N}T_B|\mathbf{X})) &= E\left[\text{var}\left(\sum_{i=1}^{a}\sum_{j\neq j'}t_{ijj'}|\mathbf{X}\right)\right] = 2E\left(\sum_{i=1}^{a}\sum_{j\neq j'}E(t_{ijj'}^2|\mathbf{X})\right) \\
&= 2\sum_{i=1}^{a}\sum_{j\neq j'}E[\sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})K_{ijj'}^2].
\end{aligned}
$$

Let $X_{i(j_*)}$ be the order statistic for X_{ij} within group i so that j_* is the rank of X_{ij} among $\{X_{ij_1}, j_1 = 1, \ldots, n_i\}$. Then

$$E(\mathrm{var}(\sqrt{N}T_B|\mathbf{X})) = \frac{4}{Na^2(k-1)^2} E\left\{\sum_{i=1}^{a}\sum_{j<j'}\sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})E[K_{ijj'}^2|X_{ij}, X_{ij'}, j_*, j'_*]\right\}$$

$$= \frac{4}{Na^2(k-1)^2} E\left\{\sum_{i=1}^{a}\sum_{j<j'}\sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})[E^2(K_{ijj'}|X_{ij}, X_{ij'}, j_*, j'_*) + \mathrm{var}(K_{ijj'}|X_{ij}, X_{ij'}, j_*, j'_*)]\right\}. \quad (18)$$

The conditional expectation can be obtained by considering whether a covariate value X_c is in group i or not. For $X_c \in$ group i_1, we denote it as $X_{i_1 j_1}$. Then, if $i_1 \neq i$,

$$\Lambda_{ijj'i_1} = E(j \in C_{ic_1}, j' \in C_{ic_1})|X_{ij}, X_{ij'}, j_*, j'_*)$$

$$= P(X_{ij} \in C_{ic_1}, X_{ij'} \in C_{ic_1}|X_{ij}, X_{ij'}, j_*, j'_*) = \int_{X_{ij}-L_{ij}}^{X_{ij}+D_{ij}} f_{X,i_1}(x)\,\mathrm{d}x I(j'_* - j_* \leq k-1),$$

where D_{ij} is the upper $k/2$ spacing and L_{ij} is the lower $(k/2 - (j'_* - j_*))$ spacing from X_{ij}. Without loss of generality, assume that $j_* < j'_*$. Applying Taylor's expansion twice, we can write

$$\Lambda_{ijj'i_1} = \{f_{X,i_1}(X_{ij})/f_{X,i}(X_{ij}) \cdot [F_{X,i}(X_{ij} + D_{ij}) - F_{X,i}(X_{ij} - L_{ij})] + O_p(N^{-2})\} I(j'_* - j_* \leq k-1).$$

From the properties of spacings in Pyke (1965), we have

$$E(F_{X,i}(X_{ij} + D_{ij}) - F_{X,i}(X_{ij} - L_{ij})|X_{ij}, X_{ij'}, j_*, j'_*) = [k - (j'_* - j_*)]/(n_i + 1) \cdot I(j'_* - j_* \leq k-1).$$

Therefore, for $X_c \in$ group $i_1 \neq i$,

$$E(\Lambda_{ijj'i_1}|X_{ij}, X_{ij'}, j_*, j'_*) = \{f_{X,i_1}(X_{ij})/f_{X,i}(X_{ij})[k - j'_* + j_*]/(n_i + 1) + O_p(N^{-2})\} I(j'_* - j_* \leq k-1). \quad (19)$$

If $i_1 = i$ and $X_{ij_1} \neq X_{ij}$ and $X_{ij_1} \neq X_{ij'}$, detailed inspection yields that

$$E(\Lambda_{ijj'i}|X_{ij}, X_{ij'}, j_*, j'_*) = \left\{\left[k - (j'_* - j_*) - 2I(j'_* - j_* \leq (k-1)/2)\right]/(n_i + 1) + O_p(N^{-2})\right\} I(j'_* - j_* \leq k-1); \quad (20)$$

if $i_1 = i$ and $X_{ij_1} = X_{ij}$ (or symmetrically $X_{ij_1} = X_{ij'}$), then

$$\Lambda_{ijj'i} = I(j'_* \in C_{iX_{i(j_*)}}) = I\left(j'_* - j_* \leq (k-1)/2\right). \quad (21)$$

Collecting terms from (19), (20), and (21), with $B_{ijj'}$ defined in Theorem 1 we have

$$E\left(K_{ijj'}|X_{ij}, X_{ij'}, j_*, j'_*\right) = B_{ijj'} + I(j'_* - j_* \leq k-1)O_p(N^{-2}). \quad (22)$$

Now consider the conditional variance. Note that when $X_c \in \{X_{ij}, X_{ij'}\}$, the term in $K_{ijj'}$ is a constant. Therefore,

$$\begin{aligned}
\operatorname{var}(K_{ijj'}|X_{ij}, X_{ij'}, j_*, j'_*) &= \operatorname{var}\left(\sum_{c=1}^{N} I(j \in C_{ic})I(j' \in C_{ic})I(X_c \notin \{X_{ij}, X_{ij'}\})|X_{ij}, X_{ij'}, j_*, j'_*\right) \\
&= \sum_{c_1=1}^{N}\sum_{c_2=1}^{N}\{E[I(j \in C_{ic_1})I(j' \in C_{ic_1})I(j \in C_{ic_2}) \\
&\quad I(j' \in C_{ic_2})|X_{ij}, X_{ij'}, j_*, j'_*] \\
&\quad -E[I(j \in C_{ic_1})I(j' \in C_{ic_1})|X_{ij}, X_{ij'}, j_*, j'_*]E[I(j \in C_{ic_2}) \\
&\quad I(j' \in C_{ic_2})|X_{ij}, X_{ij'}, j_*, j'_*]\} \\
&\quad \times I(X_{c_1} \notin \{X_{ij}, X_{ij'}\})I(X_{c_2} \notin \{X_{ij}, X_{ij'}\}) \\
&= \sum_{c=1}^{N} E[I(j \in C_{ic})I(j' \in C_{ic})I(X_c \notin \{X_{ij}, X_{ij'}\})|X_{ij}, X_{ij'}, j_*, j'_*] \\
&\quad -\sum_{c=1}^{N}[E(I(j \in C_{ic})I(j' \in C_{ic})|X_{ij}, X_{ij'}, j_*, j'_*)]^2 \\
&\quad \times I(X_c \notin \{X_{ij}, X_{ij'}\}),
\end{aligned}$$

where the last equality is due to the fact that the indicator functions involving c_1 and c_2 are conditionally independent when $c_1 \neq c_2$ and neither c_1, c_2 is X_{ij} or $X_{ij'}$. Plugging (19)–(22) into the right-hand side of the equation above, we obtain

$$\begin{aligned}
&\operatorname{var}\left(K_{ijj'}|X_{ij}, X_{ij'}, j_*, j'_*\right) \\
&= \left[\left(\sum_{i_1, i_1 \neq i}^{a} \frac{n_{i_1}}{n_i} d_{i_1 i}(X_{ij}) + 1\right)[k - j'_* + j_*] - 2I(j'_* - j_* \leq (k-1)/2) + O_p(N^{-1})\right] \\
&\times I(j'_* - j_* \leq k - 1). \qquad (23)
\end{aligned}$$

Putting (22) and (23) into (18), we have

$$\begin{aligned}
E(\operatorname{var}(\sqrt{N}T_B|\mathbf{X})) = \sum_{i=1}^{a} E\Bigg\{ \sum_{j<j'}^{n_i} \frac{4\sigma_i^2(X_{ij})\sigma_i^2(X_{ij'})}{Na^2(k-1)^2}\left[B_{ijj'}^2 + B_{ijj'} - 2I\left(j'_* - j_* \leq \frac{k-1}{2}\right)\right] \\
\times I(j'_* - j_* \leq k-1)\Bigg\} + O(N^{-1}) \to \lim_{N\to\infty} \gamma_N^2,
\end{aligned}$$

completing the proof. ■

ACKNOWLEDGMENTS

This work was partially supported by an Ecological Genomics Seed Grant at Kansas State University. We are grateful to the anonymous referees whose insightful and constructive comments have led to significant improvement of this manuscript.

BIBLIOGRAPHY

M. Aitkin, D. Anderson, B. Francis & J. Hinde (1989). "*Statistical Modelling in GLIM*," Oxford University Press, New York.

L. Breiman & J. H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80, 580–598.

A. Butte & I. S. Kohane (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 418–429.

W. Cleveland (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74, 829–836.

P. de Jong (1987). A central limit theorem for generalized quadratic forms. *Probability Theory* 75, 261–277.

P. Deheuvels (1981). An asymptotic decomposition for multivariate distribution free test of independence. *Journal of Multivariate Analysis* 11, 102–113.

P. D'haeseleer, X. Wen, S. Fuhrman & R. Somogyi (1998). Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. *Proceedings of the Second International Workshop on Information Processing in Cell and Tissues*, 203–212.

J. Fan, N. E. Heckman & M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* 90, 141–150.

J. Faraway (2006). "*Extending the Linear Model with R*," Chapman and Hall/CRC, Boca Raton, FL.

C. Genest, J. J. Quesada Molina, J. A. Rodríguez Lallena & C. Sempi (1999). A characterization of quasi-copulas. *Journal of Multivariate Analysis* 69, 193–205.

C. Genest & J. Nešlehová (2007). A primer on copulas for count data. *The ASTIN Bulletin* 37, 475–515.

C. Genest & B. Rémillard (2004). Tests of independence and randomness based on the empirical copula process. *Test* 13, 335–369.

T. Hastie & R. Tibshirani (1990). "*Generalized Additive Models*," Chapman and Hall, London.

T. Hastie & R. Tibshirani (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18, 607–616.

J. Huang & S. Kotz (1999). Modifications of the Farlie-Gumbel-Morgenstern distributions. A tough hill to climb. *Metrika* 49, 135–145.

R. Li & G. Gong (2008). K-nearest-neighbour non-parametric estimation of regression functions in the presence of irrelevant variables. *Econometrics Journal* 11, 396–408.

R. Nelson (2006). "*An Introduction to Copulas*," Springer, New York.

R. Pyke (1965). Spacing (with discussion). *Journal of the Royal Statistical Society Series B* 27, 395–449.

J. T. Terpstra & J. Mckean (2005). Rank-based analyses of linear models using R. *Journal of Statistical Software* 14, http://www.jstatsoft.org/, 1–26.

H. Wang & M.G. Akritas (2009). Rank tests in heteroscedastic multi-way HANOVA. *Journal of Nonparametric Statistics* 21, 663–681.

H. Wang & M. G. Akritas (2010). Inference from heteroscedastic functional data. *Journal of Nonparametric Statistics* 22, 149–168.

H. Wang, J. Higgins & D. Blasi (2010). Distribution-free tests for no effect of treatment in heteroscedastic functional data under both weak and long range dependence. *Statistics and Probability Letters* 80, 390–402.

L. Wang, M. G. Akritas & I. V. Keilegom (2008). An ANOVA-type nonparametric diagnostic test for heteroscedastic regression models. *Journal of Nonparametric Statistics* 20, 365–382.

S. Wood (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society Series B* 62, 413–428.

S. Wood (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society Series B* 70, 495–518.

Received 5 May 2009
Accepted 10 February 2010