Bernoulli **17**(2), 2011, 671–686
DOI: 10.3150/10-BEJ287

Limit theorems for functions of marginal quantiles

G. JOGESH BABU[1], ZHIDONG BAI[2,*], KWOK PUI CHOI[2,**]
and VASUDEVAN MANGALAM[3]

[1]*Department of Statistics, 326 Joab L. Thomas Building, The Pennsylvania State University, University Park, PA 16802-2111, USA. E-mail: babu@stat.psu.edu*
[2]*Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546. E-mail: *stabaizd@nus.edu.sg; **stackp@nus.edu.sg*
[3]*Department of Mathematics, Universiti Brunei Darussalam, Brunei. E-mail: mangalam@fos.ubd.edu.bn*

Multivariate distributions are explored using the joint distributions of marginal sample quantiles. Limit theory for the mean of a function of order statistics is presented. The results include a multivariate central limit theorem and a strong law of large numbers. A result similar to Bahadur's representation of quantiles is established for the mean of a function of the marginal quantiles. In particular, it is shown that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(X_{n:i}^{(1)},\ldots,X_{n:i}^{(d)})-\bar{\gamma}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_{n,i}+\mathrm{o}_P(1)$$

as $n\to\infty$, where $\bar{\gamma}$ is a constant and $Z_{n,i}$ are i.i.d. random variables for each n. This leads to the central limit theorem. Weak convergence to a Gaussian process using equicontinuity of functions is indicated. The results are established under very general conditions. These conditions are shown to be satisfied in many commonly occurring situations.

Keywords: central limit theorem; Cramér–Wold device; lost association; quantiles; strong law of large numbers; weak convergence of a process

1. Introduction

Let $\{(X_i^{(1)},X_i^{(2)},\ldots,X_i^{(d)}), i=1,2,\ldots\}$ be a sequence of random vectors such that for each j $(1\le j\le d)$, $\{X_1^{(j)},X_2^{(j)},\ldots\}$ forms a sequence of independent and identically distributed (i.i.d.) random variables. For $1\le j,k\le d$, let F_j and $F_{j,k}$ denote the distributions of $X_1^{(j)}$ and $(X_1^{(j)},X_1^{(k)})$, respectively. Let $X_{n:i}^{(j)}$ denote the ith order statistic ($\frac{i}{n}$th quantile) of $\{X_1^{(j)},X_2^{(j)},\ldots,X_n^{(j)}\}$. The vector $(X_{n:i}^{(1)},\ldots,X_{n:i}^{(d)})$ corresponds to the ith marginal order statistics. In this article, we study the asymptotic behavior of the mean of a function of marginal sample quantiles:

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_{n:i}^{(1)},\ldots,X_{n:i}^{(d)})\tag{1.1}$$

as $n\to\infty$, where $\phi:\mathbb{R}^d\to\mathbb{R}$ satisfies some mild conditions.

1350-7265 © 2011 ISI/BS

Our results, Theorems 1.1 and 1.2 stated below, were motivated in part by one of the authors considering [10] the problem of estimating the parameters in a linear regression model, $Y = \alpha + \beta X + \epsilon$, when the linkage between the variables X and Y was either partially or completely lost. Were the linkage not lost, then the least-squares estimator for β would be given by $(\sum_{i=1}^n X_i Y_i - n\bar{X_n}\bar{Y_n})/\sum_{i=1}^n (X_i - \bar{X_n})^2$, where $\bar{X_n}$ and $\bar{Y_n}$ denote the sample means of $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$. When the linkage is lost, a natural candidate to estimate β is the average of this expression over all possible permutations of the Y_i's. As the term in the denominator and the second term in the numerator are permutation invariant, it remains to consider $\frac{1}{n!}\sum_{\pi \in \mathcal{S}_n} \frac{1}{n}\sum_{i=1}^n X_i Y_{\pi(i)}$. This expression is bounded above by $\frac{1}{n}\sum_{i=1}^n X_{n:i} Y_{n:i}$ and below by $\frac{1}{n}\sum_{i=1}^n X_{n:i} Y_{n:n-i+1}$, by the well-known rearrangement inequality of Hardy–Littlewood–Pólya (see [8], Chapter 10). The asymptotic behavior of the lower bound can be deduced from that of the upper bound. The upper bound, $\frac{1}{n}\sum_{i=1}^n X_{n:i} Y_{n:i}$, is a special case of (1.1). The problem of the loss of association among paired data has attracted a lot of attention in various contexts, such as the broken sample problem, file linkage problem and record linkage (see, e.g., [2,4,7]). See item (3) in Section 4 for further results and a very brief review of the literature.

We shall first introduce some notation. We shall reserve $\{U_i\}$ for a sequence of independent random variables distributed uniformly on $(0, 1)$. Let $U_{n:i}$ be the ith order statistic of of $(U_1, \dots, U_n)$. For a probability distribution function F and $0 < t < 1$, define $F^{-1}(t) = \inf\{x : F(x) \geq t\}$.

Let ϕ be a real-valued measurable function on $\mathbb{R}^d$. For $0 < x, x_1, \dots, x_d < 1$, $\mathbf{x} = (x_1, \dots, x_d)$, and $1 \leq j, k \leq d$, define

$$\psi(\mathbf{x}) := \phi(F_1^{-1}(x_1), \dots, F_d^{-1}(x_d)), \tag{1.2}$$

$$\gamma(x) := \psi(x, x, \dots, x), \tag{1.3}$$

$$\psi_j(x) := \left.\frac{\partial \psi(\mathbf{x})}{\partial x_j}\right|_{(x,\dots,x)}, \tag{1.4}$$

$$\psi_{j,k}(\mathbf{x}) := \frac{\partial^2 \psi(\mathbf{x})}{\partial x_j \, \partial x_k}, \tag{1.5}$$

$$\tilde{\psi}_{j,k}(x) := \psi_{j,k}(x, \dots, x). \tag{1.6}$$

We shall now introduce conditions on ϕ that are used in the results:

(C1) The function $\psi(u_1, \dots, u_d)$ is continuous at $u_1 = \cdots = u_d = u$, $0 < u < 1$. That is, ψ is continuous at each point on the diagonal of $(0, 1)^d$. The function ψ need not be bounded.

(C2) There exist K and $c_0 > 0$ such that

$$|\psi(x_1, \dots, x_d)| \leq K\left(1 + \sum_{j=1}^d |\gamma(x_j)|\right) \qquad \text{for } (x_1, \dots, x_d) \in (0, c_0)^d \cup (1 - c_0, 1)^d.$$

(C3) Let $\mu_{n:i} = i/(n+1)$. For $1 \leq j, k \leq d$,

$$\frac{1}{n}\sum_{i=1}^n \bigl(\mu_{n:i}(1 - \mu_{n:i})\bigr)^{3/2} (\psi_j(\mu_{n:i}))^2 \longrightarrow \int_0^1 \bigl(x(1-x)\bigr)^{3/2} (\psi_j(x))^2 \, \mathrm{d}x < \infty$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\big(\mu_{n:i}(1-\mu_{n:i})\big)^{3/2}|\tilde{\psi}_{j,k}(\mu_{n:i})| \longrightarrow \int_0^1 \big(x(1-x)\big)^{3/2}|\tilde{\psi}_{j,k}(x)|\,\mathrm{d}x < \infty.$$

(C4) For all large m, there exist $K = K(m) \geq 1$ and $\delta > 0$ such that

$$|\psi(\mathbf{y}) - \psi(\mathbf{x}) - \langle \mathbf{y}-\mathbf{x}, \nabla\psi(\mathbf{x})\rangle| \leq K \sum_{j,k=1}^{d} |(y_j - x)(y_k - x)|\big(1 + |\psi_{j,k}(\mathbf{x})|\big),$$

whenever $\|\mathbf{y}-\mathbf{x}\|_{\ell_1} < \delta$ and $\min_{1\leq j\leq d} y_j(1-y_j) > x(1-x)/m$, where $\mathbf{x} = (x, \ldots, x)$, $\mathbf{y} = (y_1, \ldots, y_d) \in (0,1)^d$. Here, $\|\mathbf{y}\|_{\ell_1} := |y_1| + \cdots + |y_d|$ denotes the ℓ_1-norm of $\mathbf{y}$ and $\nabla\psi(\mathbf{x})$ denotes the gradient of ψ.

Condition (C3) holds if the functions $(x(1-x))^{3/2}(\psi_j(x))^2$ and $(x(1-x))^{3/2}|\tilde{\psi}_{j,k}(x)|$ are Riemann integrable over (0, 1) and satisfy K-pseudo convexity for $1 \leq j,k \leq d$. A function g is said to be K-pseudo convex if $g(\lambda x + (1-\lambda)y) \leq K(\lambda g(x) + (1-\lambda)g(y))$.

To state the main results, recall the definition of γ in (1.3).

Theorem 1.1. *Let* $\{(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}), i = 1,2,\ldots\}$ *be a sequence of random vectors such that for each* j $(1 \leq j \leq d)$, $\{X_1^{(j)}, X_2^{(j)}, \ldots\}$ *forms a sequence of i.i.d. random variables. Suppose* ϕ *satisfies conditions* (C1)–(C2), F_j *is continuous for* $1 \leq j \leq d$ *and* γ *is Riemann integrable. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\phi\big(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\big) \xrightarrow{a.s.} \bar{\gamma}$$

as $n \to \infty$, *where* $\bar{\gamma} = \int_0^1 \gamma(y)\,\mathrm{d}y$.

Note that we need only the independence of the jth marginal random variables, for each j. The result does not depend on the joint distribution of $(X_1^{(1)}, \ldots, X_1^{(d)})$.

Theorem 1.2. *Let* $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(d)})$ *be i.i.d. random vectors. Suppose* ϕ *satisfies conditions* (C1)–(C4), F_j *is continuous for* $1 \leq j \leq d$ *and* γ *is Riemann integrable. Then,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi\big(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\big) - \sqrt{n}\bar{\gamma} = \frac{1}{\sqrt{n}}\sum_{\ell=1}^{n} Z_{n,\ell} + \mathrm{o}_P(1), \tag{1.7}$$

where $Z_{n,\ell} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} W_{j,\ell}(i/n)\psi_j(i/(n+1))$, $W_{j,\ell}(x) = I(U_\ell^{(j)} \leq x) - x$ *for* $1 \leq \ell \leq n$ *and* $\bar{\gamma}$ *is defined as in Theorem* 1.1. *Further, as* $n \to \infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi\big(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\big) - \sqrt{n}\bar{\gamma} \xrightarrow{dist} N(0,\sigma^2), \tag{1.8}$$

where $G_{j,k}(x,y) = F_{j,k}(F_j^{-1}(x), F_k^{-1}(y))$ *and*

$$\sigma^2 = \lim_{n\to\infty} \operatorname{Var}(Z_{n,1}) = 2\sum_{j=1}^{d} \int_0^1 \int_0^y x(1-y)\psi_j(x)\psi_j(y)\,\mathrm{d}x\,\mathrm{d}y$$

$$+ 2 \sum_{1\le j<k\le d} \int_0^1 \int_0^1 \bigl(G_{j,k}(x,y) - xy\bigr)\psi_j(x)\psi_k(y)\,\mathrm{d}x\,\mathrm{d}y.$$

This theorem can be extended to m functions $\phi_1, \ldots, \phi_m$ simultaneously using the Cramér–Wold device (see [3]), as in the corollary below. Let $\psi_j(x;r)$ denote the partial derivative of $\phi_r(F_1^{-1}(x_1), \ldots, F_d^{-1}(x_d))$ with respect to x_j evaluated at $x_1 = \cdots = x_d = x$.

Corollary 1.1. *Let* $\phi_1, \ldots, \phi_m$ *satisfy conditions* (C1)–(C4). *For* $1 \le r \le m$, *if we define* $T_n(\phi_r) = \sum_{i=1}^n \phi_r(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)})$ *and* $\bar\gamma_r = E\phi_r(F_1^{-1}(U), F_2^{-1}(U), \ldots, F_d^{-1}(U)$, *then*

$$\frac{1}{\sqrt{n}}(T_n(\phi_1), \ldots, T_n(\phi_m)) - \sqrt{n}(\bar\gamma_1, \ldots, \bar\gamma_m) \xrightarrow{dist} N(0, \Sigma) \qquad \text{as } n \to \infty,$$

where the (r,s)*th element* $\sigma_{r,s}$ *of* Σ, *is given by*

$$\sum_{j=1}^{d} \int_0^1 \int_0^y x(1-y)\bigl(\psi_j(x;r)\psi_j(y;s) + \psi_j(x;s)\psi_j(y;r)\bigr)\,\mathrm{d}x\,\mathrm{d}y$$

$$+ \sum_{1\le j<k\le d} \int_0^1 \int_0^1 \bigl(G_{j,k}(x,y) - xy\bigr)\bigl(\psi_j(x;r)\psi_k(y;s) + \psi_j(x;s)\psi_k(y;r)\bigr)\,\mathrm{d}x\,\mathrm{d}y.$$

Proof. Use the Cramér–Wold device and Theorem 1.2. In computing $\sigma_{r,s}$, we used

$$2\sigma_{r,s} = \lim_{n\to\infty} \bigl(\operatorname{Var}(Z_{n,1,r} + Z_{n,1,s}) - \operatorname{Var}(Z_{n,1,r}) - \operatorname{Var}(Z_{n,1,s})\bigr),$$

where $Z_{n,1,r} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d W_{j,1}(i/n)\psi_j(i/(n+1); r)$. □

Our results can be adapted to provide a suitable test statistic for testing equality of marginal distributions against various alternative hypotheses using suitable choices for ϕ.

Remark 1.1. Since the finite-dimensional distributions converge to multivariate normal distributions, the weak convergence to a Gaussian process indexed by $t \in T$ (T being an interval of $\mathbb{R}$) can be established under a condition such as equicontinuity of $\{\phi_t : t \in T\}$.

Remark 1.2. In Theorem 1.1, we just require i.i.d. for each component. No further assumptions are made on how the components are related. We need a stronger assumption in Theorem 1.2, namely, that the rows are i.i.d. random vectors. Interestingly, the variance of the limiting normal only depends on the 2-dimensional marginal distributions.

Remark 1.3. Conditions (C1) and (C2) are, in general, easy to verify. Condition (C3) is used to control the behavior of the function ψ around the neighborhood of $(0,\dots,0)$ and $(1,\dots,1)$ in $(0,1)^d$. For example, if we suppose that $X_1^{(j)}$ is uniformly distributed over $(0,1)$ for $j=1,2$ and $\phi(x,y):=((x+y)/2)^{-\alpha}(1-(x+y)/2)^{-\alpha}$, then (C3) holds if $0<\alpha<1/4$. However, the first limit in (C3) fails if $\alpha\ge 1/4$ and the second limit in (C3) fails if $\alpha\ge 1/2$.

Remark 1.4. By a compactness argument, condition (C1) is shown to be equivalent to

(C1′) For any $c\in(0,\frac{1}{2})$, $\lim_{\delta\to 0}\omega(c,\delta)=0$, where

$$\begin{aligned}\omega(c,\delta):=\sup\{|\psi(x_1,\dots,x_d)-\gamma(y)|:|x_i-y|<\delta,c<y,\\ x_i<1-c,1\le i\le d\}.\end{aligned}\tag{1.9}$$

Proofs of Theorems 1.1 and 1.2 are given in Sections 2 and 3, respectively. The results are illustrated by means of examples and counterexamples in the last section.

2. Proof of Theorem 1.1

The main idea of the proof of Theorem 1.1 comes from the observation that

$$\frac{1}{n}\sum_{i=1}^{n}\phi\bigl(X_{n:i}^{(1)},\dots,X_{n:i}^{(d)}\bigr)=\frac{1}{n}\sum_{i=1}^{n}\psi\bigl(U_{n:i}^{(1)},\dots,U_{n:i}^{(d)}\bigr)\approx\int_0^1\psi(u,\dots,u)\,\mathrm{d}u.$$

The cases where i is close to 1 or n need to be carefully analyzed as ψ could be unbounded near 0 and 1.

Proof of Theorem 1.1. Let $U_i^{(j)}=F_j(X_i^{(j)})$ for $1\le i\le n, 1\le j\le d$. Therefore, $\{U_1^{(j)},U_2^{(j)},\dots\}$ forms a sequence of i.i.d. uniformly distributed random variables and $F_j^{-1}(U_i^{(j)})=X_i^{(j)}$ with probability 1. Recall that $U_{n:i}^{(j)}$ denotes the ith order statistic of $U_1^{(j)},\dots,U_n^{(j)}$. We write $\mu_{n:i}=EU_{n:i}^{(j)}=i/(n+1)$. Recall, also, that $\psi(x_1,\dots,x_d)=\phi(F_1^{-1}(x_1),\dots,F_d^{-1}(x_d))$ and that $\gamma(x)=\psi(x,\dots,x)$. For any $\epsilon\in(0,c_0)$,

$$\frac{1}{n}\sum_{i=1}^{n}\phi\bigl(X_{n:i}^{(1)},\dots,X_{n:i}^{(d)}\bigr)=\frac{1}{n}\sum_{i=1}^{n}\psi\bigl(U_{n:i}^{(1)},\dots,U_{n:i}^{(d)}\bigr)=\Gamma_n+R_{n,1}+R_{n,2}+R_{n,3}\tag{2.1}$$

almost surely, where

$$\Gamma_n=\frac{1}{n}\sum_{i=1}^{n}\gamma(\mu_{n:i}),$$

$$R_{n,1}=\frac{1}{n}\sum_{1\le i<\epsilon n}\bigl(\psi\bigl(U_{n:i}^{(1)},\dots,U_{n:i}^{(d)}\bigr)-\gamma(\mu_{n:i})\bigr),$$

$$R_{n,2} = \frac{1}{n} \sum_{\epsilon n \le i \le (1-\epsilon)n} \big(\psi\big(U_{n:i}^{(1)}, \ldots, U_{n:i}^{(d)}\big) - \gamma(\mu_{n:i})\big),$$

$$R_{n,3} = \frac{1}{n} \sum_{(1-\epsilon)n < i \le n} \big(\psi\big(U_{n:i}^{(1)}, \ldots, U_{n:i}^{(d)}\big) - \gamma(\mu_{n:i})\big).$$

Since γ is Riemann integrable, the Riemann sum

$$\Gamma_n \to \int_0^1 \gamma(y)\,\mathrm{d}y = E(\phi(F_1^{-1}(U), \ldots, F_d^{-1}(U))) \qquad \text{as } n \to \infty.$$

Thus, it remains to show that $R_{n,i} \longrightarrow^{\text{a.s.}} 0$ as $n \to \infty$ for $i = 1, 2$ and 3.

For $1 \le j \le d$, by then Glivenko–Cantelli lemma, $\sup_{x\in(0,1)} |\hat{F}_{n;j}(x) - x| \longrightarrow^{\text{a.s.}} 0$ as $n \to \infty$, where $\hat{F}_{n;j}$ is the empirical distribution function of $\{U_i^{(j)} : 1 \le i \le n\}$. For $1 \le i \le n$, $1 \le j \le d$, we have

$$\big|U_{n:i}^{(j)} - \mu_{n:i}\big| \le \left|U_{n:i}^{(j)} - \frac{i}{n}\right| + \frac{1}{n} = \big|U_{n:i}^{(j)} - \hat{F}_{n;j}\big(U_{n:i}^{(j)}\big)\big| + \frac{1}{n} \le \frac{1}{n} + \sup_{x\in(0,1)} |x - \hat{F}_{n;j}(x)|.$$

Hence, it follows that as $n \to \infty$,

$$\delta_n := \max\big\{\big|U_{n:i}^{(j)} - \mu_{n:i}\big|, 1 \le i \le n, 1 \le j \le d\big\} \xrightarrow{\text{a.s.}} 0. \tag{2.2}$$

Recall the definition of $\omega(c, \delta)$ in (1.9). Since $U_{n:i}^{(j)} \in (\mu_{n:i} - \delta_n, \mu_{n:i} + \delta_n)$ for $1 \le j \le d$ and for each integer i in the interval $[n\epsilon, n(1-\epsilon)]$, we have $|\psi(U_{n:i}^{(1)}, \ldots, U_{n:i}^{(d)}) - \gamma(\mu_{n:i})| \le \omega(\epsilon, \delta_n)$, provided $\delta_n < \epsilon/2$. Hence, if $\delta_n < \epsilon/2$, by (2.2) and (C1′) (which is equivalent to (C1) by Remark 1.4 in Section 1), we have

$$|R_{n,2}| \le \frac{1}{n} \sum_{\epsilon n \le i \le (1-\epsilon)n} \big|\psi\big(U_{n:i}^{(1)}, \ldots, U_{n:i}^{(d)}\big) - \gamma(\mu_{n:i})\big| \le \omega(\epsilon, \delta_n) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$. By (C2),

$$|R_{n,1}| \le K \sum_{j=1}^{d} R_{n,1,j} + \frac{1}{n} \sum_{1 \le i < \epsilon n} |\gamma(\mu_{n:i})| + K\epsilon,$$

where $R_{n,1,j} = n^{-1} \sum_{1\le i<\epsilon n} |\gamma(U_{n:i}^{(j)})|$ for $1 \le j \le d$. Clearly, if $U_{n:(\epsilon n)+1}^{(j)} \le 2\epsilon$, then

$$R_{n,1,j} \le n^{-1} \sum_{1 \le i \le n} \big|\gamma\big(U_i^{(j)}\big)\big| I\big(U_i^{(j)} \le 2\epsilon\big).$$

Note that, with probability 1, $U^{(j)}_{n:(\epsilon n)+1} \le 2\epsilon$ for all large n and the right-hand side of the above inequality goes to $\int_0^{2\epsilon} |\gamma(y)|\,\mathrm{d}y$ a.s. as $n \to \infty$. Hence,

$$\limsup_{n\to\infty} |R_{n,1}| \le (Kd+1)\left(\epsilon + \int_0^{2\epsilon} |\gamma(y)|\,\mathrm{d}y\right) \qquad \text{a.s.}$$

As $|\gamma|$ is integrable, letting ϵ tend to zero, we conclude that $R_{n,1} \xrightarrow{\text{a.s.}} 0$. A similar argument will show that $R_{n,3} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. This completes the proof of Theorem 1.1. □

3. Proof of Theorem 1.2

As in the proof of Theorem 1.1, we introduce $U_i^{(j)} = F_j(X_i^{(j)})$ for $1 \le i \le n, 1 \le j \le d$. It follows that $(U_i^{(1)}, \ldots, U_i^{(d)})$, $1 \le i \le n$, are i.i.d. random vectors. For $1 \le j, k \le d$, note that $G_{j,k}$ is the joint distribution of $(U_1^{(j)}, U_1^{(k)})$. In particular, $G_{j,j}(x, y) = \min\{x, y\}$, for $1 \le j \le d$. Using the notation introduced in Section 1, we outline some key approximations used in the proof of Theorem 1.2. In particular, (1.7) follows from

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\bigl(U^{(1)}_{n:i}, \ldots, U^{(d)}_{n:i}\bigr) - \sqrt{n}\bar{\gamma} \\
&\quad\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bigl(\psi\bigl(U^{(1)}_{n:i}, \ldots, U^{(d)}_{n:i}\bigr) - \gamma(\mu_{n,i})\bigr) \\
&\quad\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{d} \bigl(U^{(j)}_{n:i} - \mu_{n,i}\bigr)\psi_j(\mu_{n:i}) \\
&\quad\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{\ell=1}^{n} \bigl(I\bigl(U^{(j)}_{\ell} - i/n\bigr)\bigr)\psi_j(\mu_{n:i}).
\end{aligned}$$

The proof of the first approximation, which is about $\sqrt{n}$ times the difference between the Riemann sum and the integral $\bar{\gamma}$, is non-trivial and is handled in Lemma 3.3. We use Bahadur's representation of quantiles in the last approximation. We start with some technical lemmas, the first of which is well known (see [6], page 36).

Lemma 3.1. *Suppose that $U_{n:1} \le \cdots \le U_{n:n}$ denote the order statistics of n independent random variables that are uniformly distributed over* $(0, 1)$. *Then, for $1 \le i \le n$,*

$$\operatorname{Var}(U_{n:i}) = \frac{\mu_{n:i}(1-\mu_{n:i})}{(n+2)} \le \frac{1}{n}.$$

Lemma 3.2. *Under condition* (C3), *the limiting variance σ^2 is well defined.*

Proof. It suffices to show that for $1 \le j, k \le d$,

$$\beta_1 := \iint_{0<x<y<1} |G_{j,k}(x, y) - xy||\psi_j(x)\psi_k(y)| \, \mathrm{d}x \, \mathrm{d}y < \infty, \tag{3.1}$$

$$\beta_2 := \iint_{0<y<x<1} |G_{j,k}(x, y) - xy||\psi_j(x)\psi_k(y)| \, \mathrm{d}x \, \mathrm{d}y < \infty. \tag{3.2}$$

To prove (3.1), we introduce $W_j(x) := I(U_1^{(j)} \le x) - x$. Here, $W_j(x)$ has mean 0 and variance $x(1-x)$. Furthermore, $EW_j(x)W_k(y) = G_{j,k}(x, y) - xy$. Thus, $EW_j(x)W_j(y) = x(1-y)$ when $x < y$. By the Cauchy–Schwarz inequality, β_1^2 is bounded above by

$$\begin{aligned}
&\biggl(E\int_0^1\int_0^y \bigl(x/(1-y)\bigr)^{1/4}|\psi_j(x)||W_j(x)|\bigl((1-y)/x\bigr)^{1/4}|\psi_k(y)||W_k(y)| \, \mathrm{d}x \, \mathrm{d}y\biggr)^2 \\
&\quad\le E\int_0^1\int_0^y \bigl(x/(1-y)\bigr)^{1/2}(\psi_j(x))^2(W_j(x))^2 \, \mathrm{d}x \, \mathrm{d}y \\
&\qquad\qquad \times E\int_0^1\int_0^y \bigl((1-y)/x\bigr)^{1/2}(\psi_k(y))^2(W_k(y))^2 \, \mathrm{d}x \, \mathrm{d}y \\
&\quad= \int_0^1\int_0^y x^{3/2}(1-x)(1-y)^{-1/2}(\psi_j(x))^2 \, \mathrm{d}x \, \mathrm{d}y \\
&\qquad\qquad \times \int_0^1\int_0^y x^{-1/2}y(1-y)^{3/2}(\psi_k(y))^2 \, \mathrm{d}x \, \mathrm{d}y \\
&\quad= 4\int_0^1 x^{3/2}(1-x)^{3/2}(\psi_j(x))^2 \, \mathrm{d}x \\
&\qquad \times \int_0^1 y^{3/2}(1-y)^{3/2}(\psi_k(y))^2 \, \mathrm{d}x < \infty.
\end{aligned}$$

Similarly, we can prove (3.2). This completes the proof of Lemma 3.2. □

Lemma 3.3. *Let $\phi : (0, 1)^d \to \mathbb{R}$ satisfy condition* (C3). *Suppose that the function γ associated with ϕ and defined in* (1.3) *is Riemann integrable. We then have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma(\mu_{n:i}) - \sqrt{n}\int_0^1 \gamma(x) \, \mathrm{d}x \to 0$$

as $n \to \infty$.

Proof. As $\gamma'(x) = \psi_1(x) + \cdots + \psi_d(x)$, condition (C3) implies that $(x(1-x))^{3/2}(\gamma'(x))^2$ is Riemann integrable. We have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma(\mu_{n:i}) - \sqrt{n}\int_0^1 \gamma(x)\,\mathrm{d}x \\
&\quad = \sqrt{n}\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n}\bigl(\gamma(\mu_{n:i}) - \gamma(x)\bigr)\,\mathrm{d}x \\
&\quad = \sqrt{n}\sum_{i=1}^{n}\biggl(\int_{(i-1)/n}^{\mu_{n:i}}\int_x^{\mu_{n:i}}\gamma'(y)\,\mathrm{d}y\,\mathrm{d}x - \int_{\mu_{n:i}}^{i/n}\int_x^{\mu_{n:i}}\gamma'(y)\,\mathrm{d}y\,\mathrm{d}x\biggr) \\
&\quad = \sqrt{n}\int_0^1 g_n(y)\gamma'(y)\,\mathrm{d}y,
\end{aligned}$$

where

$$g_n(y) = \begin{cases} y - (i-1)/n, & \text{if } (i-1)/n \le y < i/(n+1), 1 \le i \le n, \\ y - i/n, & \text{if } i/(n+1) \le y < i/n, 1 \le i \le n. \end{cases}$$

Note that

$$|g_n(y)| \le \begin{cases} y, & \text{if } 0 < y \le 1/n, \\ 1/n, & \text{if } 1/n < y < 1 - 1/n, \\ 1 - y, & \text{if } 1 - 1/n \le y < 1. \end{cases}$$

Therefore,

$$\begin{aligned}
&\biggl(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma(\mu_{n:i}) - \sqrt{n}\int_0^1 \gamma(x)\,\mathrm{d}x\biggr)^2 \\
&\quad = n\biggl(\int_0^1 g_n(y)\gamma'(y)\,\mathrm{d}y\biggr)^2 \\
&\quad \le n\int_0^1 (g_n(y))^2\bigl(y(1-y)\bigr)^{-3/2}\,\mathrm{d}y\int_0^1 \bigl(y(1-y)\bigr)^{3/2}(\gamma'(y))^2\,\mathrm{d}y.
\end{aligned}$$

Since the second term above is finite by (C3), Lemma 3.3 will follow if we can show that the first term goes to 0 as $n \to \infty$. Note that

$$\begin{aligned}
&n\int_0^1 (g_n(y))^2\bigl(y(1-y)\bigr)^{-3/2}\,\mathrm{d}y \\
&\quad \le 2^{3/2}n\biggl(\int_0^{1/n}\sqrt{y}\,\mathrm{d}y + \int_{1-1/n}^1 \sqrt{1-y}\,\mathrm{d}y\biggr) + \frac{1}{n}\int_{1/n}^{1-1/n} y^{-3/2}(1-y)^{-3/2}\,\mathrm{d}y \\
&\quad \le \frac{8\sqrt{2}}{3\sqrt{n}} + (1-n^{-1})^{-3/4}n^{-1/4}\int_0^1 y^{-3/4}(1-y)^{-3/4}\,\mathrm{d}y \to 0.
\end{aligned}$$

□

Lemma 3.4. *Let $U_{n:i}$ denote the ith order statistic of an i.i.d. sample of size n from the uniform distribution over $(0,1)$. Define $\mathcal{A}_{m,n} = \bigcap_{1\le i\le n}\{U_{n:i}(1-U_{n:i}) > \mu_{n:i}(1-\mu_{n:i})/m\}$. We then have $\lim_{m\to\infty}\sup_{n\ge 1} P(\mathcal{A}_{m,n}) = 1$.*

Proof. By symmetry considerations, we only need to prove

$$\lim_{m\to\infty}\sup_{n\ge 1} P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i}(1-U_{n:i}) > \mu_{n:i}(1-\mu_{n:i})/m\}\biggr) = 1. \tag{3.3}$$

For any $\varepsilon > 0$, we can choose n_0 such that for all $n > n_0$, $P(U_{n:((n+1)/2)} \ge 2/3) < \varepsilon/2$ and

$$\begin{aligned}
&P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i}(1-U_{n:i}) > \mu_{n:i}(1-\mu_{n:i})/m\}\biggr)\\
&\quad\ge P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i} > 3\mu_{n:i}/m\}\biggr) - P\bigl(\bigl\{U_{n:((n+1)/2)} \ge 2/3\bigr\}\bigr)\\
&\quad\ge P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i} > 3\mu_{n:i}/m\}\biggr) - \varepsilon/2.
\end{aligned}$$

Obviously, we can find a constant m_0 such that for all $m > m_0$,

$$\sup_{1\le n\le n_0} P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i}(1-U_{n:i}) > \mu_{n:i}(1-\mu_{n:i})/m\}\biggr) > 1-\varepsilon.$$

If we can choose a constant m_1 such that for all $m > m_1$,

$$\sup_{n>n_0} P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i} > 3\mu_{n:i}/m\}\biggr) \ge 1-\varepsilon/2,$$

then, for all $m > \max(m_0, m_1)$,

$$\sup_{n\ge 1} P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i}(1-U_{n:i}) > \mu_{n:i}(1-\mu_{n:i})/m\}\biggr) > 1-\varepsilon.$$

Therefore, the proof of Lemma 3.4 reduces to establishing that

$$\lim_{m\to\infty}\sup_{n>1} P\biggl(\bigcap_{1\le i\le n/2}\{U_{n:i} > \mu_{n:i}/m\}\biggr) = 1. \tag{3.4}$$

Recall the representation formula for the order statistics from a sequence of uniform random variables, $U_{n:i} \overset{\text{dist}}{=} S_i/S_{n+1}$, where $e_1,\dots,e_{n+1}$ are i.i.d. exponentially distributed random variables with $E(e_i) = 1$ and $S_i = e_1 + \cdots + e_i$. If

$$M = \inf_{1\le i\le n<\infty} \frac{S_i/i}{S_{n+1}/(n+1)} > \frac{1}{m},$$

then, for all $1 \le i \le n/2$, we have $\frac{S_i/i}{S_{n+1}/(n+1)} > 1/m$. This, in turn, implies that, as $m \to \infty$,

$$\lim_{m\to\infty} \sup_{n\ge 1} P\biggl(\bigcap_{1\le i\le n/2} \biggl\{ \frac{S_i/i}{S_{n+1}/(n+1)} > \frac{1}{m} \biggr\} \biggr) \ge \lim_{m\to\infty} P(M > 1/m) = P(M > 0).$$

Since $S_n/n \longrightarrow^{\text{a.s.}} 1$ as $n \to \infty$, we have $P(M > 0) = 1$. This implies (3.4) and hence Lemma 3.4 follows. □

Proof of Theorem 1.2. We write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\bigl(U_{n:i}^{(1)}, \dots, U_{n:i}^{(d)}\bigr) - \sqrt{n}\bar{\gamma} = I_n + \epsilon_n = S_{n,1} + S_{n,2} + \epsilon_n,$$

where

$$\begin{aligned} I_n &= n^{-1/2} \sum_{i=1}^{n} \bigl(\psi\bigl(U_{n:i}^{(1)}, \dots, U_{n:i}^{(d)}\bigr) - \gamma(\mu_{n:i})\bigr), \\ S_{n,1} &= n^{-1/2} \sum_{j=1}^{d} \sum_{i=1}^{n} \bigl(U_{n:i}^{(j)} - \mu_{n:i}\bigr)\psi_j(\mu_{n:i}), \\ S_{n,2} &= I_n - S_{n,1}, \\ \epsilon_n &= n^{-1/2} \sum_{i=1}^{n} \gamma(\mu_{n:i}) - \sqrt{n} \int_0^1 \gamma(x)\,\mathrm{d}x. \end{aligned}$$

By Lemma 3.3, $\epsilon_n \to 0$ as $n \to \infty$. We shall now show that $S_{n,2} \to^P 0$ as $n \to \infty$.

Since $\max\{|U_{n:i}^{(j)} - \mu_{n:i}| : 1 \le i \le n, 1 \le j \le d\} \longrightarrow^{\text{a.s.}} 0$, by (C4) we have

$$|S_{n,2}| I_{\mathcal{A}_{m,n}} \le \frac{K(m)}{\sqrt{n}} \sum_{j,k=1}^{d} \sum_{i=1}^{n} \bigl|\bigl(U_{n:i}^{(j)} - \mu_{n:i}\bigr)\bigl(U_{n:i}^{(k)} - \mu_{n:i}\bigr)\bigr| \bigl(1 + |\tilde{\psi}_{j,k}(\mu_{n:i})|\bigr).$$

By condition (C3), Lemma 3.1 and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\bigl|\bigl(U_{n:i}^{(j)} - \mu_{n:i}\bigr)\bigl(U_{n:i}^{(k)} - \mu_{n:i}\bigr)\bigr| \bigl(1 + |\tilde{\psi}_{j,k}(\mu_{n:i})|\bigr) \\ &\quad \le \frac{1}{n^{3/2}} \sum_{i=1}^{n} \mu_{n:i}(1 - \mu_{n:i})|\tilde{\psi}_{j,k}(\mu_{n:i})| + \frac{1}{\sqrt{n}} := J_1 + J_2 + J_3 + \frac{1}{\sqrt{n}}, \end{aligned}$$

where $J_1 = n^{-3/2}\sum_{1\le i\le\sqrt{n}}\mu_{n:i}(1-\mu_{n:i})|\tilde{\psi}_{j,k}(\mu_{n:i})|$ and J_2, J_3 are similarly defined over $\sqrt{n} < i < n-\sqrt{n}$ and $n-\sqrt{n}\le i\le n$, respectively. We have

$$\begin{aligned} J_1 &\le \frac{2}{n}\sum_{1\le i\le\sqrt{n}}\bigl(\mu_{n:i}(1-\mu_{n:i})\bigr)^{3/2}|\tilde{\psi}_{j,k}(\mu_{n:i})| \\ &\sim 2\int_0^{1/\sqrt{n}}\bigl(x(1-x)\bigr)^{3/2}|\tilde{\psi}_{j,k}(x)|\,\mathrm{d}x \to 0 \end{aligned}$$

as $n\to\infty$. Similarly, $J_3\to 0$ as $n\to\infty$. Also, as $n\to\infty$,

$$J_2 \le \frac{1}{n^{5/4}}\sum_{\sqrt{n}<i<n-\sqrt{n}}\bigl(\mu_{n:i}(1-\mu_{n:i})\bigr)^{3/2}|\tilde{\psi}_{j,k}(\mu_{n:i})| \to 0.$$

That is, we have shown that as $n\to\infty$, for any given large m, $S_{n,2}I_{\mathcal{A}_{m,n}} \to^P 0$. We can now choose a sequence of $m = m_n\to\infty$ such that $S_{n,2}I_{\mathcal{A}_{m,n}}\to^P 0$ as $n\to\infty$. By Lemma 3.4, $I_{\mathcal{A}^c_{m,n}}\to^P 0$ and hence $S_{n,2}I_{\mathcal{A}^c_{m,n}}\to^P 0$ as $m\to\infty$. Therefore, $S_{n,2}\to^P 0$.

Define $W_{j,\ell}(x) = I(U_\ell^{(j)}\le x) - x$ for $1\le j\le d$ and $1\le\ell\le n$. Observe that $W_{j,1}$ is W_j defined in the proof of Lemma 3.2 and that $\hat{F}^{-1}_{n;j}(\frac{i}{n}) = U^{(j)}_{n:i}$. By Bahadur's representation of quantiles (see, e.g., [1] or [9]),

$$\sup_{0<t<1}|\hat{F}_{n;j}(t) - t + \hat{F}^{-1}_{n;j}(t) - t| = \mathrm{O}(n^{-3/4}\log n) \qquad \text{a.s. for } 1\le j\le d.$$

Hence,

$$S_{n,1} = \frac{1}{\sqrt{n}}\sum_{\ell=1}^{n} Z_{n,\ell} + \mathrm{o}(1) \qquad \text{a.s.},$$

where, for each n,

$$Z_{n,\ell} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} W_{j,\ell}(i/n)\psi_j(\mu_{n:i})$$

are i.i.d. random variables with mean zero and

$$\begin{aligned} \operatorname{Var}(Z_{n,1}) &= \sum_{j,k=1}^{d}\frac{1}{n^2}\sum_{h,i=1}^{n}\operatorname{Cov}\bigl(W_j(h/n), W_k(i/n)\bigr)\psi_j(\mu_{n:h})\psi_k(\mu_{n:i}) \\ &= \sum_{j,k=1}^{d}\frac{1}{n^2}\sum_{h,i=1}^{n}\bigl(G_{j,k}(h/n,i/n) - hin^{-2}\bigr)\psi_j(\mu_{n:h})\psi_k(\mu_{n:i}) \\ &\to \sum_{j,k=1}^{d}\int_0^1\int_0^1\bigl(G_{j,k}(x,y) - xy\bigr)\psi_j(x)\psi_k(y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

Recall that $G_{j,k}$ is the joint distribution of $(U_1^{(j)}, U_1^{(k)})$ and that $G_{j,j}(x,y) = \min(x,y)$. To establish the convergence above, fix j,k and split the second sum above into cases according to whether h, or i, is: less than ϵn; between ϵn and $(1-\epsilon)n$; greater than $(1-\epsilon)n$. For example, when we sum over $\epsilon n \le h, i \le (1-\epsilon)n$, then it converges to $\int_\epsilon^{1-\epsilon}\int_\epsilon^{1-\epsilon} H(x,y)\,\mathrm{d}x\,\mathrm{d}y$, where $H(x,y) = (G_{j,k}(x,y) - xy)\psi_j(x)\psi_k(y)$. The sum over $1 \le h < \epsilon n$ and $\epsilon n \le i \le (1-\epsilon)n$ can be shown to converge to $\int_0^\epsilon \int_\epsilon^{1-\epsilon} H(x,y)\,\mathrm{d}x\,\mathrm{d}y$, which, from the method of proof of Lemma 3.2 and condition (C3), can be shown to converge to 0 as $\epsilon \to 0$. Similar convergences hold for other ranges of h and i.

It is now easy to see that the limit above can be written in the form of σ^2 as stated in Theorem 1.2. Note that $|Z_{n,1}| \le \sum_{j=1}^d \frac{1}{n}\sum_{i=1}^n |\psi_j(\mu_{n:i})|$. If $(1/\sqrt{n})\frac{1}{n}\sum_{i=1}^n |\psi_j(\mu_{n:i})| \to 0$ for $j = 1,2,\dots,d$, then the Lindeberg–Lévy condition holds. To see this, note that

$$\left(\frac{1}{n}\sum_{i=1}^n |\psi_j(\mu_{n:i})|\right)^2 \le \frac{1}{n}\sum_{i=1}^n \bigl(\mu_{n:i}(1-\mu_{n:i})\bigr)^{-3/2} \frac{1}{n}\sum_{i=1}^n \bigl(\mu_{n:i}(1-\mu_{n:i})\bigr)^{3/2} (\psi_j(\mu_{n:i}))^2.$$

By (C3), it is enough to establish that $I_n = \frac{1}{n^2}\sum_{i=1}^n (\frac{i}{n+1}(1-\frac{i}{n+1}))^{-3/2} \to 0$. Since

$$I_n \le \frac{4(n+1)^{3/2}}{n^2}\left(\sum_{1\le i\le (n+1)/2} i^{-3/2} + \sum_{(n+1)/2 \le i \le n} (n+1-i)^{-3/2}\right) \to 0,$$

we have, by the Lindeberg–Lévy central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{\ell=1}^n Z_{n,\ell} \xrightarrow{\text{dist}} N(0,\sigma^2).$$

Hence, $S_{n,1} \to^{\text{dist}} N(0,\sigma^2)$, which completes the proof of Theorem 1.2. □

4. Examples and counterexamples

We give some examples to show our results and counterexamples to illustrate that conditions (C1) and (C2) are necessary for Theorem 1.1 to hold.

(1) Let Z be a random variable with a continuous distribution function F. Let $g_j, 1 \le j \le d$, be continuous monotonically increasing functions. For each $1 \le j \le d$, suppose $X_1^{(j)}, X_2^{(j)}, \dots$ are independent random variables having the same distribution as $g_j(Z)$. Applying Theorem 1.1 and assuming necessary integrability conditions, we get, after changing the variable $y = F(x)$,

$$\frac{1}{n}\sum_{i=1}^n \phi\bigl(X_{n:i}^{(1)}, \dots, X_{n:i}^{(d)}\bigr) \xrightarrow{\text{a.s.}} E\phi(g_1(Z),\dots,g_d(Z)) \qquad \text{as } n \to \infty.$$

(2) Let $(X_1^{(1)}, \dots, X_1^{(d)})$, $(X_2^{(1)}, \dots, X_2^{(d)}), \dots$ be independent random vectors having the same distribution as $(U_1, \dots, U_d)$, where the U_j's are uniformly distributed over $(0, 1)$. Let $F_{j,k}$ be the joint distribution of U_j and U_k. Suppose $\phi\colon (0,1)^d \to \mathbb{R}$ is defined by $\phi(x_1, \dots, x_d) = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$, where $\alpha_j \geq 1$. Let $M = \alpha_1 + \cdots + \alpha_d$. Then $\psi = \phi$, $\gamma(x) = x^M$ and $\psi_j(x) = \alpha_j x^{M-1}$ for $1 \leq j \leq d$. We have

$$\frac{1}{n} \sum_{i=1}^{n} \bigl(X_{n:i}^{(1)}\bigr)^{\alpha_1} \cdots \bigl(X_{n:i}^{(d)}\bigr)^{\alpha_d} \xrightarrow{\text{a.s.}} \frac{1}{M+1}$$

and

$$n^{-1/2} \left(\sum_{i=1}^{n} \bigl(X_{n:i}^{(1)}\bigr)^{\alpha_1} \cdots \bigl(X_{n:i}^{(d)}\bigr)^{\alpha_d} - \frac{n}{M+1} \right) \xrightarrow{\text{dist.}} N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{2}{M^2} \sum_{1 \leq j < k \leq d} \alpha_j \alpha_k \operatorname{Cov}(U_j^M, U_k^M) + \frac{1}{(M+1)^2 (2M+1)} \sum_{j=1}^{d} \alpha_j^2.$$

(3) The study of the statistical properties when there is a loss of association among paired data has attracted a lot of attention in various contexts, such as the broken sample problem, file linkage problem and record linkage. For example, DeGroot and Goel initiated the investigation of estimating the correlation coefficient of a bivariate normal distribution based on a broken random sample in [7]. Copas and Hilton proposed statistical models to measure the evidence that a pair of records relates to the same individuals in [4]. Chan and Loh considered an approximation of the likelihood computation for large broken sample in [5]. Bai and Hsing, in [2], proved that there does not exist any consistent discrimination rule for the correlation coefficient, ρ, between X and Y when the paired sample is broken, that is, the association between X and Y is lost. When pairing is lost, the X's and Y's behave as if they were independent as far as first order asymptotics, such as the law of large numbers (see Theorem 1.1), are concerned.

Example 1. This example shows that condition (C1) is necessary for Theorem 1.1 to hold. Let

$$\phi(x, y) = \begin{cases} 1, & \text{if } 0 < x = y < 1, \\ 0, & \text{if } 0 < x \neq y < 1. \end{cases}$$

Let $\{(X_i, Y_i)\colon 1 \leq i \leq n\}$ be a sequence of i.i.d. random vectors. We further suppose that X_i and Y_i are independent and uniformly distributed over $(0, 1)$. Since ϕ is bounded, (C2) holds, whereas (C1) does not hold. We further note that $P(X_{n:i} \neq Y_{n:i}) = 1$ for $1 \leq i \leq n$. Hence, $\sum_{i=1}^{n} \phi(X_{n:i}, Y_{n:i}) = 0$, but $\int_0^1 \phi(x, x)\, \mathrm{d}x = 1$.

Example 2. This example shows that condition (C2) is necessary for Theorem 1.1 to hold. Let $\tilde{S}_0 = (0, 1)^2$ and, for $m \geq 1$, define $\tilde{S}_m = (\frac{m}{m+1}, 1)^2$, $S_m = \tilde{S}_{m-1} \setminus \tilde{S}_m$. Let L_m be the union of three line segments:

$$L_m = \left\{\left(\frac{m}{m+1}, y\right) : \frac{m}{m+1} \le y \le 1\right\} \cup \left\{\left(x, \frac{m}{m+1}\right) : \frac{m}{m+1} \le x \le 1\right\}$$
$$\cup \left\{(x, x) : \frac{m-1}{m} \le x \le \frac{m}{m+1}\right\}.$$

Let C_m be the region inside S_m which is distance ϵ_m within L_m, where ϵ_m is chosen so that the area of C_m is m^{-8}. Write $A_m = S_m \setminus C_m$. Let ϕ be a continuous on $(0, 1)^2$ satisfying $\phi = 1$ on the diagonal, $\phi = m^3$ on A_m and $1 \le \phi \le m^3$ on C_m.

Let $\{U_i, V_j : 1 \le i, j \le n\}$ be independent and uniformly distributed on $(0, 1)$. Define $W_n = (U_{n:n}, V_{n:n})$ and $a_n = \lceil n^{1/2} \rceil$. Observe that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \phi(U_{n:i}, V_{i:n}) &\ge \frac{1}{n}\phi(W_n) \ge \frac{1}{n}\sum_{m\ge\sqrt{n}} m^3 I(W_n \in A_m)\\ &\ge \sqrt{n}\sum_{m\ge\sqrt{n}} \big(I(W_n \in S_m) - I(W_n \in C_m)\big) \qquad (4.1)\\ &\ge \sqrt{n}\left(I(W_n \in \tilde{S}_{a_n}) - I\left(W_n \in \bigcup_{m\ge\sqrt{n}} C_m\right)\right).\end{aligned}$$

We now claim that

$$I(W_n \in \tilde{S}_{a_n}) \longrightarrow 1 \qquad \text{a.s.,} \tag{4.2}$$

$$I\left(W_n \in \bigcup_{m\ge\sqrt{n}} C_m\right) \longrightarrow 0 \qquad \text{a.s.} \tag{4.3}$$

as $n \to \infty$. To prove (4.2), observe that

$$P(W_n \notin \tilde{S}_{a_n}) \le 2P\big(U_{n:n} \le a_n/(1 + a_n)\big) = 2\left(1 - \frac{1}{1 + a_n}\right)^n \approx 2\mathrm{e}^{-\sqrt{n}}.$$

This yields

$$\sum_{n=1}^{\infty} P(W_n \notin \tilde{S}_{a_n}) < \infty,$$

which, by the Borel–Cantelli lemma, implies that $I(W_n \notin \tilde{S}_{a_n}$ i.o.$) = 0$, proving (4.2). To prove (4.3), it suffices to show that

$$\sum_{n=1}^{\infty} P\left(W_n \in \bigcup_{m\ge\sqrt{n}} C_m\right) < \infty. \tag{4.4}$$

We again consider the nth term in the series in (4.4):

$$
\begin{aligned}
P\Bigg(W_n \in \bigcup_{m\ge\sqrt{n}} C_m\Bigg) &\le \sum_{m\ge\sqrt{n}} P(W_n \in C_m) \\
&\le \sum_{m\ge\sqrt{n}} P\big((U_i, V_j) \in C_m \text{ for some } 1 \le i, j \le n\big) \\
&\le n^2 \sum_{m\ge\sqrt{n}} P\big((U_1, V_1) \in C_m\big) = n^2 \sum_{m\ge\sqrt{n}} m^{-8} \le Cn^{-3/2}
\end{aligned}
$$

and hence the infinite series in (4.4) is finite. This completes the proof of (4.3). Thus, by (4.1), $\frac{1}{n}\sum_{i=1}^{n} \phi(U_{n:i}, V_{i:n})$ diverges. Furthermore, it is easy to see that condition (C2) does not hold. If (C2) were satisfied, that would imply boundedness of γ over $(1 - c_0, 1)^2$, which is not the case. This completes the construction of the counterexample.

Acknowledgements

The authors wish to thank the Editor, an Associate Editor and the referees for useful suggestions which improved the paper. They acknowledge helpful discussions with Dr. Alex Cook. The research of G.J. Babu was supported in part by NSF Grants AST-0707833 and AST-0434234. The research of Z.D. Bai was supported in part by the National University of Singapore ARF Grant R-155-000-079-112. The research of K.P. Choi is supported in part by the National University of Singapore ARF Grants R-155-000-051-112 and R-155-000-102-112.

References

[1] Babu, G.J. and Rao, C.R. (1988). Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population. *J. Multivariate Anal.* **27** 15–23. MR0971169

[2] Bai, Z.D. and Hsing, T. (2005). The broken sample problem. *Probab. Theory Related Fields* **131** 528–552. MR2147220

[3] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. *Wiley Series in Probability and Statistics: Probability and Statistics*. New York: Wiley. MR1700749

[4] Copas, J.B. and Hilton, F.J. (1990). Record linkage: Statistical models for matching computer records. *J. R. Statist. Soc. A* **153** 287–320.

[5] Chan, H.P. and Loh, W.L. (2001). A file linkage problem of DeGroot and Goel revisited. *Statist. Sinica* **11** 1031–1045. MR1867330

[6] David, H.A. (1981). *Order Statistics*. New York: Wiley. MR0286226

[7] DeGroot, M.H. and Goel, P.K. (1980). Estimation of the correlation coefficient from a broken sample. *Ann. Statist.* **8** 264–278. MR0560728

[8] Hardy, G.H., Littlewood, J.E. and Pólya, G. (1952). *Inequalities*. Cambridge: Cambridge Univ. Press.

[9] Kiefer, J. (1970). Deviations between the sample quantile process and the sample d.f. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)* 299–319. London: Cambridge Univ. Press. MR0277071

[10] Mangalam, V. (2010). Regression under lost association. To appear.

Received December 2009 and revised May 2010