Ann Inst Stat Math (2011) 63:1165–1182
DOI 10.1007/s10463-010-0271-z

A class of asymptotically normal degenerate quasi U-statistics

Aluísio Pinheiro · Pranab Kumar Sen · Hildete P. Pinheiro

Received: 6 February 2009 / Revised: 1 September 2009 / Published online: 23 February 2010

Abstract Some quasi U-statistics, unlike other variants of U-statistics, arising in distance based tests for homogeneity of groups, have first-order stationary kernels of degree 2, and yet they enjoy asymptotic normality under suitable hypotheses of invariance. Central limit theorems for a more general class of quasi U-statistics with possibly higher order stationarity (and degree) are formulated with the aid of appropriate martingale (array) characterizations as well as permutational invariance structures.

Keywords Genomics · Hamming distance · Martingale · Orthogonal system · Permutation measure · Second-order asymptotics · Higher order decomposability

1 Introduction

In a possibly multi-dimensional setup, not necessarily confined to quantitative data models, some tests for group divergence (or homogeneity), based on appropriate

Acknowledgment of support: This research was funded in part by FAPESP (08/51097-6; 08/09286-6;09/14176-8) and CNPq (306993/2008-2; 480919/2009-7; 306240/2009-2).

A. Pinheiro (✉) · H. P. Pinheiro
Departamento de Estatística, Universidade Estadual de Campinas,
Cidade Universitária Zeferino Vaz, Cx. Postal 6065,
Campinas, SP, CEP 13083-970, Brazil
e-mail: pinheiro@ime.unicamp.br

H. P. Pinheiro
e-mail: hildete@ime.unicamp.br

P. K. Sen
Department of Biostatistics and Statistics and Operations Research,
School of Public Health, UNC at Chapel Hill, 3105E McGavran-Greenberg,
CB# 7420, Chapel Hill, NC 27599-7420, USA
e-mail: pksen@bios.unc.edu

metric or distance-norm, involve some variants of symmetric functions (Rubin and Vitale 1980; van Zwet 1984) structurally similar to the classical Hoeffding U-statistics (Hoeffding 1948a, 1961), and yet they are different from other variants of (such as incomplete, weighted, and generalized) U-statistics (Janson 1984; Major 1994; O'Neil and Redner 1993).

We term these statistics as quasi U-statistics. In particular, for qualitative data models, tests based on the Hamming distance (Pinheiro et al. 2005) give rise to kernels of degree 2, which under the hypothesis of homogeneity, are stationary of order one. A special feature of such distance-based measures is the subgroup decomposability that allows a partitioning of the pooled group measure into two (between and within) components with ordered expectations; a degeneracy occurs when the groups are all homogeneous.

Taking clue from this basic observation, it was shown (Pinheiro et al. 2005) that under the hypothesis of homogeneity whereas the classical ANOVA decomposition yields a non-normal distribution, even asymptotically, the quasi U-statistics approach can have asymptotic normality under appropriate regularity conditions (Pinheiro et al. 2009). For instance, consider G groups with n_g observations in group g, for $g = 1, \ldots, G$. For a symmetric kernel $\phi(\mathbf{a}, \mathbf{b})$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, for some $K \geq 1$, let $U_{n_g}^{(g)}$, $g = 1, \ldots, G$, and $U_{n_g, n_{g'}}^{(g,g')}$, $1 \leq g < g' \leq G$, be the natural U-statistics estimators for the within gth group and between (g, g')th groups diversity measures (Sen 1999), respectively. Moreover, the combined sample diversity measure can be written as $U_n = W_n + B_n$, representing the sample within-groups and between-groups measures. When the distributions $F_1, \ldots, F_G$ associated with the G groups are not all the same, $n^{1/2}(B_n - \mathrm{E}(B_n))$ is asymptotically normal, as expected. Under group homogeneity, $\mathrm{E}(B_n) = 0$ and these U-statistics will be stationary of order one. However, nB_n can still be asymptotically normal, under suitable conditions. We refer to Pinheiro et al. (2005) for Hamming distance based genomic studies and Sen (2006) for microarray studies based on robust U-statistics.

The proposed class of quasi U-statistics, along with the preliminary notion, are introduced in Sect. 2. The main results on this unanticipated asymptotic normality through a specific martingale array construction are presented in Sect. 3. Five versions are discussed, involving either a finite second, $(2 + \delta)$th or fourth moment on the kernel, and some growth condition on the sum of squares of the coefficients, as well as two different estimators for the kernel second moment and different stationarity orders on the kernels. Some general discussions are appended in Sect. 4.

2 Preliminary results on variants of U-statistics

Let T_n be a linear combination defined by

$$T_n = \sum_{i_1, \ldots, i_m}^{1,n} \eta_{n, i_1 \cdots i_m} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}), \tag{1}$$

such that: (i) $\eta_{n, i_1 \ldots i_m}$ are weight functions, (ii) $\sum_{i_1, \ldots, i_m}^{1,n}$ is taken on all strictly ordered permutations of $1, \ldots, n$, (iii) $\phi(\cdot, \ldots, \cdot)$ is a kernel of degree m, stationary of order

r ($1 \leq r \leq m$), for which we let $\theta = \mathrm{E}\phi(X_1, \ldots, X_m)$, and (iv) $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are i.i.d. random vectors of dimension K, not necessarily quantitative in nature (Pinheiro et al. 2005).

Some configurations of $\eta_{n,i_1 \ldots i_m}$ lead to special classes of (generalized) U-statistics, as follows. If $\eta_{n,i_1 \ldots i_m} \equiv \binom{n}{m}^{-1}$ and $r \geq 1$, T_n is a degenerate U-statistics of degree m whose projection variances are such that $0 = \sigma_1^2 = \cdots = \sigma_r^2 < \sigma_{r+1}^2$; then T_n has a degeneracy of order r and $n^{(r+1)/2}(T_n - \theta)$ converges to a (possibly) infinite linear combination of independent random variables, each distributed accordingly to a $(r+1)$-dimensional Wiener integral (Dynkin and Mandelbaum 1983).

If $K = 1$ and the $\eta_{n,i_1 \ldots i_m}$ assume 0 or 1 values only, T_n is said to be an *incomplete* U-statistic. Asymptotic distribution of T_n will be either a linear combination of independently distributed. Wiener integrals or a mixture of such a distribution with an independent normal r.v., under suitable sampling conditions (Janson 1984). For a class of so-called *conditional* U-statistics, where the weights can be decomposed as $\eta_{n,i_1 \ldots i_m} = e(i_1) \ldots e(i_m)$, being $e(\cdot)$ the marginal weight function, asymptotic normality follows from Stute (1991). Moreover, the conditional nature of the class derives from the fact that weights are defined as random functions of another set of i.i.d. r.v.'s.

For $K = 1$, O'Neil and Redner (1993) and Major (1994) present asymptotic results in a more general setup for the class of *weighted* U-statistics, defined by (1). The case $m = 2$ using moment matching techniques to determine the asymptotic distribution of T_n is discussed in O'Neil and Redner (1993). Under some regularity conditions on $\eta_{n,i_1 \ldots i_m}$, a non-normal limit is proven for either $r = 1$ or $r = 0$. For $r = 0$, a class of weighted U-statistics is proved to be asymptotically normal under a second set of conditions on weights. Asymptotic normality is also established for $r = 1$ and *incomplete designs*. The common idea behind all weight-designs is the orthogonality on the set of (possibly random) weights. Major (1994) points out that the aforementioned approach cannot be adapted for $m \geq 3$; Poisson approximation is used to pursue asymptotic behavior of T_n. Four main results provide asymptotic distribution of weighted U-statistics, under different weighting schemes. In three situations $r \leq m$, and in all four cases $X_1, X_2 \ldots$ is supposed to be a sequence of i.i.d. uniformly distributed r.v.'s.

A class of quasi U-statistics based on a general mth degree kernel, stationary of order r, and having the novelty that it can be applied for any i.i.d. random vectors of arbitrary (and even increasing) dimension K, is considered here. The fact that $K \geq 2$ precludes the use of Major (1994)'s results, which are based on the inversion theorem for immediate adaptation for non-uniform distributions. More generally, the proposed class is constructed in such a way that, although ϕ can be degenerate, the chosen weights lead to a contrast, i.e., such that

$$\sum_{i_1, \ldots, i_m}^{1,n} \eta_{n,i_1 \ldots i_m} = 0, \tag{2}$$

providing asymptotically normal distributions. We should notice, for instance, that Theorem 4.1 (O'Neil and Redner 1993) attains asymptotic normality for degenerate kernels under restrictions $m = 2$ and incomplete design, which are not assumed in this work. Conditions for asymptotics in the proposed setup are mild, and can be easily

expressed and interpreted in terms of ℓ_p norm, $p \geq 2$. Motivated by the commonly used Hamming distance in the genomic context (Pinheiro et al. 2000, 2005, 2009), it appears that for the quasi U-statistics, the contrast condition (2) is an essential requirement. Sans (2), for degenerate U-statistics of general order m the asymptotic distribution is non-normal (O'Neil and Redner 1993; Major 1994). As another classical example consider the one-way ANOVA test statistic. The weights do not sum zero and the asymptotic distribution is not normal.

3 Martingale representation and asymptotic normality

In this section we present a martingale representation for T_n when $m = r + 1$, leading thereby to a martingale central limit theorem for T_n. In case $r < m - 1$, it is shown that Hoeffding's decomposition of the kernel and mild assumptions on the weights can be used to ascertain that, for large n, T_n behaves like a martingale plus a stochastically negligible remainder term.

We assume that $\phi(\cdot, \ldots, \cdot)$ is a symmetric stationary kernel of order r, centered at 0, and forms an orthogonal system for which

$$\mathrm{E}[\phi(\mathbf{X}_1, \ldots, \mathbf{X}_m) \mid \mathbf{X}_1, \ldots, \mathbf{X}_j] = 0 \text{ a.e., } \forall j \leq r \tag{3}$$

and the $\mathbf{X}_i$ are i.i.d.r.v.'s with a distribution F. Further, the (nonstochastic) $\eta_{n,i_1 \ldots i_m}$, $1 \leq i_1 < \cdots < i_m \leq n$, satisfy (2) and

$$\sum_{i_1,\ldots,i_m}^{1,n} \eta^2_{n,i_1 \cdots i_m} = M_n(\nearrow \text{ in } n \geq m). \tag{4}$$

Lemma 1 *Consider T_n as in (1), with $m = r + 1$. Define*

$$Z_{nj} = \sum_{i_1,\ldots,i_{m-1}}^{1,j-1} \eta_{n,i_1 \ldots i_{m-1} j} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1}}, \mathbf{X}_j),$$

for every $j = m, \ldots, n$, and $T_{nk} = Z_{nm} + \cdots + Z_{nk}$, for $m \leq k \leq n$. Also, let $\mathcal{B}_{nk}$ be the sub-sigma fields generated by $\mathbf{X}_i$, $i \leq k$, for $m \leq k \leq n$. Then, $\{T_{nk}, \mathcal{B}_{nk} : m \leq k \leq n\}$ is a (zero mean) martingale (array), closed on the right by T_n.

Proof Since $T_n = T_{nk} + \sum_{j=k+1}^{n} Z_{nj}$, $\mathrm{E}(T_n \mid \mathcal{B}_{nk}) = T_{nk}$, *a.e.*, $\forall k : m \leq k \leq n$ if and only if $\mathrm{E}(Z_{nk+1} \mid \mathcal{B}_{nk}) = 0$ a.e., for every $k < n$.

As ϕ is stationary of order $m - 1$, for every $k = m - 1, \ldots, n - 1$, by (3),

$$\begin{aligned} &\mathrm{E}(Z_{nk+1} \mid \mathcal{B}_{nk}) \\ &= \sum_{i_1,\ldots,i_{m-1}}^{1,k+1} \eta_{n,i_1 \ldots i_{m-1} k+1} \mathrm{E}\left[\phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1}}, \mathbf{X}_{k+1}) | \mathbf{X}_1, \ldots, \mathbf{X}_k\right] = 0 \text{ a.e.} \end{aligned} \tag{5}$$

□

Let

$$\tau_2 = \mathrm{E}\phi^2(\mathbf{X}_1, \ldots, \mathbf{X}_m) > 0. \tag{6}$$

Five theorems are presented below. Theorems 1 and 2 state that T_n, conveniently normalized, will have an asymptotically normal distribution, under finite second moment kernel and an additional uniform integrability condition. Theorems 3 and 4 introduce a finite $(2+\delta)$th and fourth moment condition on ϕ, respectively, avoiding the uniform integrability assumption. Moreover, a natural estimator of τ_2 can be obtained under convenient permutation schemes on the latter version. Theorem 5 deals with the generalization of the aforementioned theorems to the case where the stationarity order $r < m-1$.

Define

$$m_{nk} = \sum_{i_1,\ldots,i_{m-1}}^{1,k} \eta^2_{n,i_1,\ldots,i_{m-1},k}, \tag{7}$$

$$\nu_{nk} = m_{nk}\tau_2, \tag{8}$$

$$\nu_n = \nu_{nm} + \cdots + \nu_{nn} = M_n\tau_2, \tag{9}$$

for $m \le k \le n$. Note that $M_n = \sum_{k=m}^n m_{nk}$. Typically $m_{nk} = O(k^{m-1})$ and that would be sufficient for (10).

Assume that as $n \to \infty$,

$$\max_{m\le k\le n} m_{nk}/M_n \to 0, \tag{10}$$

$$Z^2_{nk}/m_{nk} \text{ are uniformly integrable.} \tag{11}$$

Theorem 1 *Let $\phi(\cdot,\cdot)$ be a degree m kernel, centered, stationary of order $m-1$, for which* **(A)** (2), (4) *and* (6)–(11) *hold.*

Then as $n \to \infty$,

$$L_n = (\nu_n)^{-1/2}T_n \xrightarrow{\mathcal{D}} N(0,1). \tag{12}$$

Proof Let

$$\begin{aligned} Z^\star_{nk} &= Z_{nk}\mathbb{I}(|Z_{nk}| < \epsilon\nu_n^{1/2}) - \mathrm{E}_{k-1}\left(Z_{nk}\mathbb{I}(|Z_{nk}| < \epsilon\nu_n^{1/2})\right) \\ &= Z_{nk}\mathbb{I}(|Z_{nk}| < \epsilon\nu_n^{1/2}) + \mathrm{E}_{k-1}\left(Z_{nk}\mathbb{I}(|Z_{nk}| > \epsilon\nu_n^{1/2})\right), \end{aligned} \tag{13}$$

as $\mathrm{E}_{k-1}Z_{nk} = 0$ a.e.. Further, note that $Z_{nk} - Z^\star_{nk} = Z_{nk}\mathbb{I}(|Z_{nk}| > \epsilon\nu_n^{1/2}) - \mathrm{E}_{k-1}\left(Z_{nk}\mathbb{I}(|Z_{nk}| > \epsilon\nu_n^{1/2})\right)$, for every $m \le k \le n$. Therefore, the martingale array

difference property extends to all $\{Z_{nk}\}$, $\{Z^{\star}_{nk}\}$ and $\{Z_{nk} - Z^{\star}_{nk}\}$. Thus,

$$\begin{aligned}
\frac{1}{v_n}\mathrm{E}\left(\sum_{k=m}^{n}\left\{Z_{nk} - Z^{\star}_{nk}\right\}\right)^2 &= \frac{1}{v_n}\sum_{k=m}^{n}\mathrm{E}\left\{Z_{nk} - Z^{\star}_{nk}\right\}^2 \\
&= \frac{1}{v_n}\sum_{k=m}^{n}\mathrm{E}\left\{Z_{nk}\mathbb{I}(|Z_{nk}| > \epsilon v_n^{1/2}) \right. \\
&\quad \left. - \mathrm{E}_{k-1}\left(Z_{nk}\mathbb{I}(|Z_{nk}| > \epsilon v_n^{1/2})\right)\right\}^2 \\
&\leq \frac{1}{v_n}\sum_{k=m}^{n}\mathrm{E}\left\{Z^2_{nk}\mathbb{I}(|Z_{nk}| > \epsilon v_n^{1/2})\right\} \\
&= \sum_{k=m}^{n}\frac{v_{nk}}{v_n}\mathrm{E}\left\{\frac{Z^2_{nk}}{v_{nk}}\mathbb{I}(|Z_{nk}| > \epsilon v_n^{1/2})\right\} = o(1), \qquad (14)
\end{aligned}$$

by (11), so it suffices to show that as $n \to \infty$, $v_n^{-1/2}\sum_{k=m}^{n} Z^{\star}_{nk} \xrightarrow{\mathcal{D}} N(0,1)$.

We let $v^{\star}_{nk} = \mathrm{E}Z^{\star 2}_{nk}$, $m \leq k \leq n$ and $v^{\star}_n = \sum_{k=m}^{n} v^{\star}_{nk}$. Then, we have

$$v^{\star}_n / v_n \uparrow 1, \quad \text{as } n \to \infty. \qquad (15)$$

So, we show that $(v^{\star}_n)^{-1}\sum_{k=m}^{n} Z^{\star 2}_{nk} \xrightarrow{p} 1$, as $n \to \infty$. Note that

$$\left\{\mathrm{E}\left[\frac{1}{v^{\star}_n}\sum_{k=m}^{n} Z^{\star 2}_{nk} - 1\right]^2\right\}^{1/2} \leq \left[(v^{\star}_n)^{-2}\mathrm{E}\left\{\sum_{k=m}^{n} Z^{\star 2}_{nk}\right\}^2\right]^{1/2} \qquad (16)$$

Since Z_{nk} is a martingale difference sequence, we can employ inequality (3.5) from Burkholder (1973) to get

$$v^{\star -2}_n \mathrm{E}\left\{\sum_{k=m}^{n} Z^{\star 2}_{nk}\right\}^2 \leq C(v^{\star}_n)^{-1},$$

for $C = 18\sqrt{2}$ so that (4) and (15) assure that $(v^{\star}_n)^{-1}\sum_{k=m}^{n} Z^{\star 2}_{nk} \xrightarrow{p} 1$.

We then apply Corollary 2.8 of McLeish (1974) to $(v^{\star}_n)^{-1/2}\sum_{k=m}^{n} Z^{\star}_{nk}$ and the proof is complete. □

Theorem 2 *Assume the regularity conditions in* Theorem 1. *Let*

$$U^{(m)}_n = \binom{n}{m}^{-1}\sum_{i_1,\ldots,i_m}^{1,n}\phi^2(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_m}). \qquad (17)$$

Then as $n \to \infty$,

$$L_n = (M_n U_n^{(m)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0, 1). \tag{18}$$

Proof Since $U_n^{(m)}$, a U-statistic and unbiased estimator of τ_2, is a reverse martingale (Sen 1981), $U_n^{(m)} \xrightarrow{a.s.} \tau_2$, as $n \to \infty$. (18) then follows from (9) and (12). □

Theorem 3 *Let* $\phi(\cdot, \cdot)$ *be a degree* m *kernel, centered, stationary of order* $m-1$ *such that*

(B.1) $\mathrm{E}\,|\phi(\mathbf{X}_1, \ldots, \mathbf{X}_m)|^{2+\delta} < \infty$, *for some positive* δ,
(B.2) (2), (4) *and* (6)–(10) *hold.*

Let $U_n^{(m)}$ *be defined by* (17). *Then as* $n \to \infty$,

$$L_n = (M_n U_n^{(m)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0, 1). \tag{19}$$

Proof Since $U_n^{(m)}$ is a nondegenerate estimator of τ_2, $U_n^{(m)} \xrightarrow{a.s.} \tau_2$, as $n \to \infty$. Moreover, the finite $(2+\delta)$th moment avoids the necessity of (11), and (19) follows in a similar manner. □

Theorem 4 addresses the situation in which the fourth moment of ϕ is finite, and utilizes a permutation-based argument. Let $b_n^{(j)} = \sum_{i_1, \ldots, i_m}^{1,n} \sum_{j_1, \ldots, j_m}^{1,n} \eta_{n, i_1 \ldots i_m} \eta_{n, j_1 \ldots j_m}$, for $j = 0, \ldots, m$. Note that $\sum_{j=0}^{m-1} b_n^{(j)} = -M_n$, and assume

$$\sum_{j=0}^{m-1} b_n^{(j)2} / n^{m+j-1} = o(M_n^2) \quad \text{as } n \to \infty. \tag{20}$$

Theorem 4 *Let* $\phi(\cdot, \cdot)$ *be a degree* m *kernel, centered, stationary of order* $m-1$ *such that*

(C.1) $\mathrm{E}\phi^4(\mathbf{X}_1, \ldots, \mathbf{X}_m) < \infty$,
(C.2) *(2),* (4), (6)–(9) *and* (20) *hold.*

Then, as $n \to \infty$,

$$L_n = (M_n U_n^{(m)})^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0, 1). \tag{21}$$

Proof (17) leads to $U_n^{(m)} = \mathrm{E}_{\mathcal{P}_n}[\phi^2(\mathbf{X}_1, \ldots, \mathbf{X}_m)]$, and let

$$\begin{aligned} U_n^{(2m-j)} &= \mathrm{E}_{\mathcal{P}_n}[\phi(\mathbf{X}_1, \ldots, \mathbf{X}_m)\phi(\mathbf{X}_1, \ldots, \mathbf{X}_j, \mathbf{X}_{m+1}, \ldots, \mathbf{X}_{2m-j})] \\ &= \frac{1}{n(n-1)\cdots(n-2m+j)} \sum_{i_1, i_{2m-j}}^{1,n} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}) \\ &\quad \times \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_j}, \mathbf{X}_{i_{m+1}}, \ldots, \mathbf{X}_{i_{2m-j}}); \ j = 0, \ldots, m; \ n \geq 2m. \end{aligned} \tag{22}$$

Note that $\mathcal{P}_n$ is a conditional (given the collection of n observations) probability measure. As noted in Theorem 3.3, $U_n^{(m)} \xrightarrow{a.s.} \tau_2$ as $n \to \infty$. $U_n^{(2m-j)}$ is a degenerate U-statistic of stationary order $m + j - 1$, for $j = 1, \ldots, m$. Then

$$\begin{aligned} \mathrm{E}_{\mathcal{P}_n}(T_n) &= \sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m} \mathrm{E}_{\mathcal{P}_n}(\phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m})) \\ &= \left(\sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m} \right) U_n^{(m)} \\ &= 0 \text{ a.e., by (2).} \end{aligned} \tag{23}$$

This also implies that $\mathrm{E}(T_n) = \mathrm{E}(\mathrm{E}_{\mathcal{P}_n}(T_n)) = 0$, $\mathrm{Var}_{\mathcal{P}_n}(T_n) = \mathrm{E}_{\mathcal{P}_n}(T_n^2)$, and

$$\begin{aligned} \mathrm{Var}_{\mathcal{P}_n}(T_n) &= \mathrm{E}_{\mathcal{P}_n} \left(\sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}) \right)^2 \\ &= \left(\sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m}^2 \right) U_n^{(m)} \\ &\quad + \sum_{j=1}^{m} \left(\sum_{i_1,\ldots,i_{m+j}}^{1,n} \eta_{n,i_1\ldots i_m} \eta_{n,i_1\ldots i_{m-j} i_{m+1}\ldots i_{m+j}} \right) U_n^{(2m-j)}. \end{aligned} \tag{24}$$

Let $\mathbf{i} = \{i_1, \ldots, i_m\}$ and $\mathbf{j} = \{j_1, \ldots, j_m\}$. Let also c be the cardinality of common indexes on $\mathbf{i}$ and $\mathbf{j}$. For $n \geq 2m$, $c = 0, \ldots, m$, and (24) can be written as

$$\begin{aligned} \mathrm{Var}_{\mathcal{P}_n}(T_n) &= \sum_{c=0}^{m} \left\{ \sum_{i_1,\ldots,i_m}^{1,n} \sum_{j_1,\ldots,j_m}^{1,n} \eta_{n,i_1\ldots i_m} \eta_{n,j_1\ldots j_m} \right\} U_n^{(2m-c)} \\ &= M_n U_n^{(m)} + \sum_{c=1}^{m} \left\{ \sum_{i_1,\ldots,i_m}^{1,n} \sum_{j_1,\ldots,j_m}^{1,n} \eta_{n,i_1\ldots i_m} \eta_{n,j_1\ldots j_m} \right\} U_n^{(2m-c)} \\ &= M_n U_n^{(m)} + \sum_{c=1}^{m-1} b_n^{(c)} U_n^{(2m-c)}. \end{aligned} \tag{25}$$

Further,

$$\begin{aligned} \mathrm{E} \left(\sum_{j=1}^{m-1} b_n^{(j)} U_n^{(2m-j)} \right)^2 &= \sum_{j=1}^{m-1} b_n^{(j)2} \mathrm{E} \left[U_n^{(2m-j)} \right]^2 \\ &= \sum_{j=1}^{m-1} b_n^{(j)2} O(n^{-(m+j-1)}). \end{aligned} \tag{26}$$

Therefore, the second term on the RHS of (25) will be $o_p(M_n)$ as long as (20) holds, which in turn means that

$$\frac{(\mathrm{Var}_{\mathcal{P}_n}(T_n) - M_n U_n^{(m)})^2}{M_n^2} = o_p(1), \tag{27}$$

i.e.,

$$U_n^{(m)} M_n / \mathrm{E}(T_n^2) \xrightarrow{p} 1, \text{ as } n \to \infty. \tag{28}$$

Led by the martingale array representation of T_n, we let

$$v_{nk} = \mathrm{E}(Z_{nk}^2 \mid \mathcal{B}_{nk-1}), \ m \le k \le n \quad \text{and} \quad V_n = \sum_{k=m}^{n} v_{nk}. \tag{29}$$

Then, by the martingale property (Lemma 1), for every $n \ge m$,

$$\mathrm{E}(V_n) = \sum_{k=m}^{n} \mathrm{E}(Z_{nk}^2) = \mathrm{E}\left(\sum_{k=m}^{n} Z_{nk}\right)^2 = \mathrm{E}(T_n^2). \tag{30}$$

Further, note that for every $k \le n$,

$$\begin{aligned} v_{nk} = & \sum_{i_1 \ldots, i_{m-1}}^{1,k} \eta_{n,i_1 \cdots i_{m-1} k}^2 \varphi_{m-1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1}}) \\ & + \sum_{j=1}^{m-1} \sum_{i_1, \ldots, i_{m+j-1}}^{1,k} \eta_{n,i_1 \ldots i_{m-1} k} \eta_{n,i_1 \ldots i_{m+j-1} i_m \ldots i_{m+j-1} k} \\ & \times \varphi_{m+j-1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m+j-1}}) \end{aligned} \tag{31}$$

where

$$\begin{aligned} \varphi_{m-1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1}}) &= \mathrm{E}[\phi^2(\mathbf{X}_i, \ldots, \mathbf{X}_{i_{m-1}} \mathbf{X}_k) \mid \mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1}}] \\ &\quad (\to \mathrm{E}\varphi_{m-1}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1}}) = \tau_2); \end{aligned} \tag{32}$$

$$\begin{aligned} \varphi_{m-1+j}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1+j}}) = \mathrm{E}[&\phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1+j}}, \mathbf{X}_k) \\ &\times \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1+j}}, \mathbf{X}_{i_m}, \ldots, \mathbf{X}_{i_{2m-2+j}}, \mathbf{X}_k) \\ &\mid \mathbf{X}_{i_1}, \ldots \mathbf{X}_{2m-2+j}], \end{aligned} \tag{33}$$

for $i_1, \ldots, i_{2m-2+j} < k$ so that $\mathrm{E}\varphi_{m-1+j}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{m-1+j}}) = 0$. Therefore, $\mathrm{E}v_{nk} = \tau_2 \sum_{i_1, \ldots, i_{m-1}}^{1,k} \eta_{n,i_1 \cdots i_{m-1} k}^2$, $\forall\ k \le m$, and hence, $\mathrm{E}V_n = \mathrm{E}T_n^2$, as expected from Lemma 1. Further,

$$
\begin{aligned}
V_n/\mathrm{E}T_n^2 &= \sum_{k=m}^{n} \sum_{i_1,\ldots,i_{m-1}}^{1,k} \eta^2_{n,i_1\ldots i_{m-1}k}\varphi_{m-1}(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_{m-1}})/\{M_n\tau_2\} \\
&\quad + \sum_{k=m}^{n}\sum_{j=1}^{m-1} \sum_{i_1,\ldots,i_{m-1+j}}^{1,k} \eta_{n,i_1\ldots i_{m-1}k}\eta_{n,i_1\ldots i_{m-j}i_m\ldots i_{m-1+j}k} \\
&\quad \times \varphi_{m-1+j}(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_{m-1+j}})/\{M_n\tau_2\} \\
&= A_n + B_n, \quad say. \qquad (34)
\end{aligned}
$$

The martingale representation $A_n - 1 = a_{nm-1} + \cdots + a_{nn-1}$, where

$$
a_{nk} = \sum_{i_1,\ldots,i_{m-2}}^{1,k-1} \eta^2_{n,i_1,\ldots,i_{m-2},k}[\varphi_{m-1}(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_k}) - \tau_2]/\{M_n\tau_2\},
$$

implies that $A_n \overset{a.s.}{\to} 1$ as $n \to \infty$. Also, $\mathrm{E}B_n = 0$, and, given that the φ_{m-1+j} are orthogonal,

$$
\begin{aligned}
\mathrm{E}B_n^2 &= \sum_{k=m}^{n}\sum_{j=1}^{m-1}\left[\sum_{i_1,\ldots,i_{m-1+j}}^{1,k} \eta^2_{n,i_1\ldots i_{m-1}k}\eta^2_{n,i_1\ldots i_{m-j}i_m\ldots i_{m-1+j}k}\right] \\
&\quad \times \mathrm{E}\varphi^2_{m-1+j}(\mathbf{X}_1,\ldots,\mathbf{X}_{m-1+j})/\left\{M_n^2\tau_2^2\right\}. \qquad (35)
\end{aligned}
$$

Note that $\mathrm{E}\varphi^2_m(\mathbf{X}_1,\ldots,\mathbf{X}_m) \le \mathrm{E}\phi^4(\mathbf{X}_1,\ldots,\mathbf{X}_m) < \infty$ and all other φ_{m-1+j}'s are degenerate U-statistics. By (20), we have $\mathrm{E}B_n^2 \to 0$ as $n \to \infty$ so that $B_n = o_p(1)$. Thus,

$$
V_n/\mathrm{E}(T_n^2) \overset{p}{\to} 1, \quad \text{as } n \to \infty. \qquad (36)
$$

By virtue of (28) and (36), we are in a position to use the martingale (array) central limit theorem (Dvoretzky 1972) to establish (21), and it suffices to verify the Lindeberg condition: $\forall\ \epsilon > 0$, as $n \to \infty$, $\sum_{k=2}^{n} \mathrm{E}(Z_{nk}^2 I(|Z_{nk}| > \epsilon\sqrt{\mathrm{E}(T_n^2)}))/\mathrm{E}(T_n^2) \to 0$. Since the Z_{nk} have finite moments at least up to the order 4, instead of the Lindeberg condition, we may as well use (the more restrictive) Liapounoff condition, and the proof of this follows along the lines of (30)–(36). □

Although in many important applications, $r = m-1$, in some instances degeneracy can be of order $r < m-1$. If that is the case, we can write

$$
T_n = \sum_{k=0}^{m} T_n^{(k)} \qquad (37)
$$

$$
T_n^{\star} = \sum_{k=0}^{r+1} T_n^{(k)}, \qquad (38)
$$

for

$$\phi_c(x_1,\ldots,x_c) = \mathrm{E}\{\phi(X_1,\ldots,X_m)|X_1 = x_1,\ldots,X_c = x_c\} \quad (39)$$
$$c = 1,\ldots,m$$
$$\phi_c^0(x_1,\ldots,x_c) = \phi_c(x_1,\ldots,x_c) - \theta \quad c = 0,\ldots,m \quad (40)$$
$$\psi_1(x_1) = \phi_c^0(x_1) \quad (41)$$
$$\psi_2(x_1,x_2) = \phi_2^0(x_1,x_2) - \phi_0^1(x_1) - \phi_0^1(x_2) + \phi_0^0 \quad (42)$$
$$\psi_3(x_1,x_2,x_3) = \phi_3^0(x_1,x_2,x_3) - \phi_2^0(x_1,x_2) - \phi_2^0(x_1,x_3) - \phi_2^0(x_2,x_3) + \phi_0^1(x_1) + \phi_0^1(x_2) + \phi_0^1(x_3) - \phi_0^0 \quad (43)$$
$$\ldots$$
$$\psi_m(x_1,\ldots,x_m) = \phi_m^0(x_1,\ldots,x_m) - \sum \phi_{m-1}^0(x_1,\ldots,x_{m-1}) + \cdots + (-1)^m\phi_0^0 \quad (44)$$

and

$$T_n^{(k)} = \sum_{i_1,\ldots,i_k}^{1,n} \eta_{n,i_1\cdots i_k}\psi_k(X_{i_1},\ldots,X_{i_k}). \quad (45)$$

Using Hoeffding's projection technique on $\phi(\cdot,\cdots,\cdot)$ and the fact that $\eta_{n,i_1\ldots i_m}$ are centered in zero, one is able to write

$$T_n = \sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m} \sum_{j=1}^{k}\sum_{o}\psi_j(X_{v_1},\ldots,X_{v_j}), \quad (46)$$

where $\sum_o$ is taken for all subsets $\{v_1,\ldots,v_j\}$ of j elements out of $\{i_1,\ldots,i_m\}$.

We can then write

$$T_n = \sum_{j=1}^{n}\sum_{v_1,\ldots,v_j}^{1,n}\left(\sum_{o*}\eta_{n,i_1\ldots i_m}\right)\psi_j(X_{v_1},\ldots,X_{v_j}), \quad (47)$$

where $\sum_{o*}$ is taken for all sets $\{i_1,\ldots,i_m\}$ of which $\{v_1,\ldots,v_j\}$ is a subset, and $\mathrm{E}\psi_j^2(X_1,\ldots,X_j) = \sigma_{\psi,j}^2$, $0 = \sigma_{\psi,1}^2 = \ldots = \sigma_{\psi,j}^r < \sigma_{\psi,j}^{r+1}$. Moreover, $T_n - T_n^\star = \sum_{j=r+2}^m \sum_{v_1,\ldots,v_j}^{1,n}\left(\sum_{o*}\eta_{n,i_1\ldots i_m}\right)\psi_j(X_{v_1},\ldots,X_{v_j})$, so that

$$\begin{aligned}\mathrm{E}(T_n - T_n^\star)^2 &= \sum_{j=r+2}^{m}\sum_{v_1,\ldots,v_j}^{1,n}\left(\sum_{o*}\eta_{n,i_1\ldots i_m}\right)^2\mathrm{E}\psi_j^2(X_{v_1},\ldots,X_{v_j})\\ &\le C\sum_{j=r+2}^{m} n^{-j}\binom{m}{r+1}\sum_{v_1,\ldots,v_{r+1}}^{1,n}\eta_{n,v_1,\ldots,v_j}^{\star 2}\\ &= O(n^{-(r+2)}M_{n,r+1}), \end{aligned} \quad (48)$$

for a constant $C > 0$, where the $\eta^{\star}_{n,v_1,\ldots,v_j}$ represent the weights relative to the kernel $\psi_{r+1}(\cdot,\ldots,\cdot)$, stationary of order r and $M_{n,r+1} = \sum^{1,n}_{v_1,\ldots,v_{r+1}} \eta^{\star 2}_{n,v_1,\ldots,v_j}$, similarly to (4), and $\psi_j(\cdot,\ldots,\cdot)$ is a stationary kernel of order j, for $j = r+2,\ldots,m$.

Theorem 5 *Let ϕ be a kernel of degree m, stationary of order $r < m-1$. Suppose that one out of the following set of conditions,* **(A)**, **(B1)–(B2)**, *or* **(C1)–(C2)**, *holds. Moreover, let*

$$M_{n,r} = o(n^{r+1}),\ r = 2,\ldots,m-1. \tag{49}$$

Then,

$$L_n = (Var(T_n))^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1) \quad as\ n \to \infty. \tag{50}$$

Proof By (48), if (49) holds, $T_n - T^{\star}_n \xrightarrow{p} 0$, as $n \to \infty$. Therefore T_n has the same limiting distribution of $T^{\star}_n$. Under (49), and one of the set of conditions **(A)**, **(B1)–(B2)**, or **(C1)–(C2)**, $T^{\star}_n$ will be asymptotically normal. □

Remarks (i) We should note that, contrary to the moments matching procedure (O'Neil and Redner 1993), which cannot be generalized to $m \geq 3$, m has no significant bearing on the proposed martingale representation. Moreover, one should also notice that although the number of different weight values (where grouping is desirable) does not change with the sample size, membership to an arbitrary group is not ordered in n and weights η's can, in this class, vary with n.

(ii) For the case $m = 2, r = 1$, M_n will be typically $O(n^2)$; in the case of genomics, the η's will be bounded (Pinheiro et al. 2005).

With the current data acquisition capabilities, the problem of high dimensional data sets and appropriate statistical estimation and hypothesis testing is of crucial importance. We present two results for large K and categorical random vectors. The basic assumption needed for that is mixing along the random vectors. We prove those results for random weights η and kernels of order $r = m-1$. The stated proofs can be easily adapted for deterministic weights and $r < m-1$ as well.

Theorem 6 *Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be a sequence of i.i.d. $K \times 1$ categorical random vectors. Let $\phi(\cdot,\ldots,\cdot)$ be a kernel of degree m such that*

$$\phi(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_m}) = \frac{1}{K}\sum_{l=1}^{K} \phi^{\star}(X_{i_1 l},\ldots,X_{i_m l}), \tag{51}$$

for some kernel, stationary of order $m-1$, $\phi^{\star}(\cdot,\ldots,\cdot)$. Let T_n be defined by (1). Assume that one out of the following set of conditions:

(a) (6)–(9), (11), $m_{nk} = o_p(M_n)$ *as $n \to \infty$ hold*;
(b) (6)–(9), $m_{nk} = o_p(M_n)$ *as $n \to \infty$ and* **(B.1)** *hold* ;
(c) (6), $\sum_{j=0}^{m} b_n^{(j)2}/n^{m+j-1} = o_p(M_n^2)$ *as $n \to \infty$ and* **(C.1)** *hold.*

Suppose that $\{\eta_{ni_1\ldots i_m},\ 1 \le i_1 < \cdots < i_m \le n\ ,\ n \ge m\}$ *is a triangular array of random variables independent of* $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\ ,\ n \ge m\}$, *and*

$$\sum_{1\le i_1<\cdots<i_m\le n} \eta^2_{ni_1\ldots i_m} - M_n = o_p(M_n) \quad as\ n \to \infty. \tag{52}$$

Suppose also that

$$\sum_{1\le l<q\le K} \mathrm{E}\left[\phi^\star(X_{i_1 l}, \ldots, X_{i_m l})\phi^\star(X_{i_1 q}, \ldots, X_{i_m q})\right] = O(K) \quad as\ K \to \infty. \tag{53}$$

Then

$$(K M_n U_n^{(m)})^{-1/2} T_n \overset{\mathcal{D}}{\to} N(0,1) \quad as\ n \to \infty\ and\ K \to \infty, \tag{54}$$

where $U_n^{(m)}$ *is defined by* (17).

Proof First consider the array $\{\eta_{n,i_1\ldots i_m} : 1 \le i_1 < \cdots < i_m \le n\},\ n \ge m$ to be deterministic, such that $\sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m} = 0$ and $\sum_{i_1,\ldots,i_m}^{1,n} \eta^2_{n,i_1\ldots i_m} = \binom{n}{m}$. We take K a positive integer and WLOG $M_n = \binom{n}{m}$.

By Theorem 1, the martingale characterization of T_n is achieved and, by Theorem 2, (54) follows.

Next, consider the case of stochastic $\{\eta_{n,i_1\ldots i_m}, 1 \le i_1 < \cdots < i_m \le n\}$ and choose one set of conditions **(a)**, **(b)** or **(c)**. Since the $\eta_{n,i_1\ldots i_m}$ are independent of the $\mathbf{X}_i,\ i \le n$, the permutation law $\mathcal{P}_n$ remains intact conditionally on $\{\eta_{n,i_1\ldots i_m}, 1 \le i_1 < \cdots < i_m \le n\}$. Let then $\bar\eta_n = \binom{n}{m}^{-1} \sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m}$, so that letting $\eta^\circ_{n,i_1\ldots i_m} = \eta_{n,i_1\ldots i_m} - \bar\eta_n,\ 1 \le i_1 < \ldots < i_0 \le n$, we have $\sum_{i_1,\ldots,i_m}^{1,n} \eta^\circ_{n,i_1\ldots i_m j} = 0$. Then, we can write

$$T_n = \sum_{i_1,\ldots,i_m}^{1,n} \eta^\circ_{n,i_1\ldots i_m} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}) + \binom{n}{m} \bar\eta_n W_n^{(m)}, \tag{55}$$

where $W_n^{(m)} = \binom{n}{m}^{-1} \sum_{i,\ldots,i_m}^{1,n} \phi(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}) \to 0$ a.s./L_2-norm, and $W_n^{(m)} = O_p(n^{-m/2})$. As such, if with $n \to \infty$,

(a) $\bar\eta_n \overset{p}{\to} 0$, and
(b) $\sum_{i_1,\ldots,i_m}^{1,n} {\eta^\circ_{n,i_1\ldots i_m}}^2 / \binom{n}{m} \overset{p}{\to} 1$,

then letting $T_n^\circ = \sum_{i,\ldots,i_m}^{1,n} \eta^\circ_{n,i_1,\ldots,i_m} \phi(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_m})$, we have

$$\binom{n}{m}^{-1/2} T_n = \binom{n}{m}^{-1/2} T_n^\circ + \binom{n}{m}^{1/2} \bar{\eta}_n W_n^{(m)} \sim \binom{n}{m}^{-1/2} T_n^\circ,$$

so that by the Slutsky's Theorem:

$$T_n / \sqrt{\binom{n}{m} U_n^{(m)}} \xrightarrow{\mathcal{D}} N(0,1).$$

Suppose now that K is large. If the weights are non-stochastic, following (51), and recalling that $0 < \tau_2 = \mathrm{E}\phi^2(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_m}) < \infty$,

$$K\tau_2 = \bar{\tau}_2 + \frac{2}{K} \sum_{1\le l<q\le K} \eta^2_{n,1\ldots m} \mathrm{E}\left[\phi^\star(X_{1l},\ldots,X_{ml})\phi^\star(X_{1q},\ldots,X_{mq})\right], \quad (56)$$

where $\bar{\tau}_2 = (1/K)\sum_{l=1}^K \eta^2_{n,1\ldots m} \mathrm{E}\phi^{\star 2}(X_{1l},\ldots,X_{ml})$. By (53), one has a finite limit for (56) when $K \to \infty$. Let $\tau_0 = \lim_{K\to\infty} K\tau_2$. Then,

$$K U_n^{(m)} \xrightarrow{p} \tau_0 \quad \text{as } n \to \infty \text{ and } K \to \infty$$

and (54) follows by Dvoretzky (1972). So that, we have as $n \to \infty$,

$$T_n / \sqrt{\binom{n}{m} \tau_0} \xrightarrow{\mathcal{D}} N(0,1).$$

If we then consider random coefficients $\{\eta_{n,i_1\ldots i_m};\ 1 \le i_1 < \cdots < i_m = n; n \ge m\}$, (54) will follow as well, because of (52) and (55). □

Theorem 7 *Let T_n be defined as in* Theorem 6. *Suppose that* (53) *holds. Then,*

$$T_n / \sqrt{\mathrm{Var}(T_n)} \xrightarrow{\mathcal{D}} N(0,1),$$

as $K \to \infty$ (either if $n \to \infty$, $n/K \to 0$, as $K \to \infty$ or if n is bounded).

Proof We apply Theorem 2.1 (Withers 1981). Let $S_n = T_n/\sqrt{M_n} = K^{-1}\sum_{k=1}^K t_{nk}/\sqrt{M_n} = K^{-1}\sum_{k=1}^K x_{nk}$, where $t_{nk} = \sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\ldots i_m}\phi^*(X_{i_1k},\ldots,X_{i_mk})$.

Since $\phi^*(\cdot,\ldots,\cdot)$ is bounded (as a function of categorical values), take, for every $k \ge 1$, $|\phi(X_{i_1k},\ldots,X_{i_mk})| \le M$ w.p.1. Then, $|x_{nk}| \le M$ w.p.1. and $\|\sum_{j=a+1}^{a+b} x_{nk}\|_{2+\epsilon} \le bM$. Hence, the rate of growth of the partial sums $(2+\epsilon)$-norm is guaranteed.

The mixing condition (53) ensures the l-mixing (Yoshihara 1993). Moreover, (53) also implies that $Var(S_n) = O(K) \to \infty$ as $K \to \infty$ and that the covariances are absolutely summable. Therefore, the CLT holds for T_n at a rate $O(\sqrt{K})$ if n is bounded or $O(n\sqrt{K})$ if both $K \to \infty$ and $n \to \infty$. □

We list below three examples. In two of them the asymptotic behavior can be derived directly from that of T_n defined by (1). The third example, quantitative data ANOVA, illustrates the importance of kernel degeneracy.

Example 1 [Hoeffding (1948b)] Let $\mathbf{Z}_1, \mathbf{Z}_2 \ldots$ i.i.d. bivariate random vectors with d.f. F and $\mathbf{Z}_i = (X_i, Y_i)$, $i \geq 1$. Consider

$$\Delta = \Delta(F) = \int D^2(x, y) \mathrm{d}f(x, y),$$

where $D(x, y) = F(x, y) - F(x, \infty)F(\infty, y)$. Then the components of $\mathbf{X}_1$ are independent if and only if $D(x, y) \equiv 0$ (Hoeffding 1948b). The kernel for the U-statistic that unbiasedly estimates Δ is given by

$$\phi(x_1, y_1; \ldots; x_5, y_5) = \frac{1}{4}\psi(x_1, x_2, x_3)\psi(x_1, x_4, x_5)\psi(y_1, y_2, y_3)\psi(y_1, y_4, y_5),$$

for

$$\begin{aligned} C(u) &= \mathbb{I}(u \geq 0), \\ \psi(x_1, x_2, x_3) &= C(x_1 - x_2) - C(x_1 - x_3). \end{aligned}$$

The null hypothesis of independence can be written as $H_0 : \Delta \equiv 0$ and the test statistic is given by

$$D_n = n^{-[5]} \sum_{*} \phi(X_{i_1}, Y_{i_1}; \ldots; X_{i_5}, Y_{i_5}),$$

where $n^{-[j]} = n \times \cdots \times (n - j + 1)$ and $\sum_*$ is taken for all $\{i_1, \ldots, i_5\} \subset \{1, \ldots, n\}$. The asymptotic variance for the first order term of Hoeffding's projection, ξ_1, is zero under H_0. Therefore, the asymptotic distribution of $\sqrt{n}D_n$ will not be normal under the null hypothesis (Hoeffding 1948b).

In the *multisample* case, however, we define D_n as a pooled sample U-statistic. Then, we can decompose $D_n = D_n(W) + D_n(B)$, where $D_n(W)$ and $D_n(B)$ are the within-groups and between-groups statistics, respectively. $D_n(B)$ is a member of the class of quasi U-statistics and, therefore, $M_n^{-1/2} D_n(B)$ will be asymptotically normal under the null hypothesis of independence. We should note that we can perform such a test for an actual multisample, in which case we will be testing componentwise independence for all sub-samples, or for a one-sample case in which we divide the sample in pseudo sub-samples. A question of statistical interest is the sampling scheme which maximizes power in either case.

Example 2 (*Quasi V-statistics of degree* 2) Note that

$$V_n = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi(\mathbf{X}_i, \mathbf{X}_j). \tag{57}$$

$T_{U,n}$ and $T_{V,n}$ are respectively defined as

$$T_{U,n} = \sum_{1 \le i < j \le n} \eta_{n,ij} \phi(\mathbf{X}_i, \mathbf{X}_j) \tag{58}$$

and

$$T_{V,n} = \sum_{i=1}^{n} \sum_{j=1}^{n} \eta_{n,ij} \phi(\mathbf{X}_i, \mathbf{X}_j), \tag{59}$$

where $\phi(\cdot, \cdot)$ is a stationary kernel of order 1.

Defining

$$Z_{V,nj} = 2 \sum_{i=1}^{j-1} \eta_{n,ij} \phi(\mathbf{X}_i, \mathbf{X}_j) + \eta_{n,jj} \phi(\mathbf{X}_j, \mathbf{X}_j), \tag{60}$$

$j = 1, \ldots, n$, we can write

$$\begin{aligned} T_{V,nk} &= Z_{V,n1} + \cdots + Z_{V,nk} \quad 1 \le k \le n, \\ T_{V,n} &= T_{V,nn}. \\ T_{V,n} - 2T_{U,n} &= \sum_{j=1}^{n} \eta_{n,jj} \phi(\mathbf{X}_j, \mathbf{X}_j). \end{aligned} \tag{61}$$

Thus, whenever $\sum_{j=1}^{n} \eta_{n,jj} \phi(\mathbf{X}_j, \mathbf{X}_j)/V_n \xrightarrow{p} c$ (possibly 0), asymptotic normality with the same rate holds, as long as $\sum_{i=1}^{n} \eta_{n,ii}^2 = O(M_n)$. Hence, through the asymptotic joint normality of $T_{u,n}$ and $T_{V,n} - 2T_{U,n}$, $T_{V,n}$ will also be asymptotically normal. If $\sum_{i=1}^{n} \eta_{n,ii}^2 = o(M_n)$, $T_{V,n} - 2T_{U,n}$ will be asymptotic negligible and no extra asymptotic variance term will be added. In that way, asymptotic results for $T_{V,n}$ are equivalent to the asymptotic results for $T_{U,n}$ up to a bias term. First-order stationary and generalized quasi U-statistics can also be dealt with similarly, via Hoeffding's decomposition of their respective kernels.

We present below a counterexample for Theorems 1–5, which illustrates the importance of degeneracy for the asymptotic normality.

Example 3 (*Quantitative ANOVA*) Suppose $m = 2$ and $\phi(x, y) = (x - y)^2/2$. We use the same groups and weights as in Pinheiro et al. (2005), i.e.,

$$\eta_{n,ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to different groups;} \\ -\frac{n-n_g}{n_g-1} & \text{if } i \text{ and } j \text{ both belong to group } g, \end{cases} \tag{62}$$

where G groups are defined with n_g units in each, $n = n_1 + \cdots + n_G$. We define the gth intra-group measures as $U_{gg} = S_g^2$ and the (g, g')th cross-groups measures as

$$U_{gg'} = \frac{1}{2}\left\{(\bar{X}_g - \bar{X}_{g'})^2 + U_{gg} - U_{g'g'} - \frac{1}{n_g}U_{gg} - \frac{1}{n_{g'}}U_{g'g'}\right\}, \tag{63}$$

for $g, g' = 1, \ldots, G$, $g \neq g'$. We have therefore

$$nT_n = \frac{n}{n-1}\sum_{1\leq g<g'\leq G}\frac{n_g n_{g'}}{n}(\bar{X}_g - \bar{X}_{g'})^2 - \sum_{1\leq g<g'\leq G}\frac{n_g + n_{g'}}{n-1}\theta + o_p(1), \tag{64}$$

so that nT_n will be asymptotically equivalent in distribution to a linear combination of chi-square r.v.'s each with one degree of freedom, which we write as $\sum_i \lambda_{ni}(Z_i^2 - 1)$, $\lambda_{ni} \geq 0$.

Example 3 reflects the fact that, since the kernel $\phi(x, y) = (x - y)^2/2$ is not degenerate, its first order projection variance still plays the main role in asymptotics, and results from Hoeffding (1948a) and O'Neil and Redner (1993) can therefore be applied successfully. If $K \geq 2$ but still fixed, a similar argument will lead to a χ^2 with K degrees of freedom even under mild dependency conditions (within each vector structure). If all the λ_{ni} are small, then the CLT applies to nT_n.

4 Conclusion

We present a class of quasi U-statistics admitting a martingale representation. Moreover, their contrast-like weights provides these statistics with asymptotic normality albeit their kernel's degeneracy. These results readdress some results for weighted U-statistics in the literature. Three basic features of our proposal are attractive. The asymptotic normality for degenerate kernels represents a very important and useful tool for methodological work. Moreover, the martingale representation can be implemented for any combination of kernel degree, order of degeneracy and vector dimension. The case of nonstochastic n and η's has been treated here. If n is itself stochastic, denoted by N, assuming positive integer values, the weights $\eta_{N,i_1,\ldots,i_m}$ may also become stochastic. As long as $N/\mathrm{E}N \xrightarrow{p} c > 0$ and the η_N's are independent of the $\mathbf{X}_i$, the asymptotic results pertain to this stochastic environment. Martingale limit theorems for stochastic sample sizes are applicable in this case.

Acknowledgments We would like to thank the reviewers as well as the associate editor for their most thoughtful comments and suggestions.

References

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *Annals of Probability, 1*(1), 19–42.

Dvoretzky, A. (1972). Central limit theorem for dependent random variables. In L. LeCam, et al. (Eds.), *Proceedings of the sixth Berkeley symposium of mathematical statistics and probability*, Vol. 2, pp. 513–555, Los Angeles: University of California Press.

Dynkin, E. B., Mandelbaum, A. (1983). Symmetric statistics, poisson point processes and multiple Wiener integrals. *Annals of Statistics, 11*(3), 739–745.

Hoeffding, W. (1948a). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics, 19*(3), 293–325.

Hoeffding, W. (1948b). A non-parametric test of independence. *Annals of Mathematical Statistics, 19*(4), 546–557.

Hoeffding, W. (1961). The strong law of large numbers for U-statistics. In Institute of Statistics Mimeo Series No. 302. Chapel Hill: University of North Carolina.

Janson, S. (1984). The asymptotic distribution of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 66*(4), 495–505.

Major, P. (1994). Asymptotic distributions for weighted U-statistics. *Annals of Probability, 22*(3), 1514–1535.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability, 2*(4), 620–628.

O'Neil, K. A., Redner, R. A. (1993). Asymptotic distributions of weighted U-statistics of degree 2. *Annals of Probability, 21*(2), 1159–1169.

Pinheiro, A., Sen, P. K., Pinheiro, H. P. (2009). Decomposability of high-dimensional diversity measures: quasi U-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis, 100*, 1645–1656.

Pinheiro, H. P., Seillier-Moiseiwitsch, F., Sen, P. K., Eron, J. (2000). Genomic sequence analysis and quasi-multivariate catanova. In P. K. Sen, C. R. Rao, (Eds.), *Handbook of statistics, Vol. 18: Bioenvironmental and public health statistics* (pp. 713–746). Amsterdam: Elsevier.

Pinheiro, H. P., Pinheiro, A., Sen, P. K. (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference, 130*(1/2), 325–339.

Rubin, H., Vitale, R. A. (1980). Asymptotic distribution of symmetric statistics. *Annals of Statistics, 8*(1), 165–170.

Sen, P. K. (1981). *Sequential nonparametrics: Invariance principles and statistical inference*. New York: Wiley.

Sen, P. K. (1999). Utility-oriented Simpson-type indexes and inequality measures. *Calcutta Statistical Association Bulletin, 49*(193–194), 1–22.

Sen, P. K. (2006). Robust statistical inference for high-dimensional data models with application to genomics. *Austrian Journal of Statistics, 35*(2/3), 197–214.

Stute, W. (1991). Conditional U-statistics. *Annals of Probability, 19*(2), 812–823.

van Zwet, W. R. (1984). A Berry-Esseen bound for symmetric statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 66*(3), 425–440.

Withers, C. S. (1981). Central limit theorems for dependent variables. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57*(4), 509–534.

Yoshihara, K. (1993). *Asymptotic statistics based on weakly dependent data: weakly dependent stochastics sequences and their applications*, 2. Sanseido, Tokyo.