Decomposability of High-Dimensional Diversity Measures: Quasi U-Statistics, Martingales and Nonstandard Asymptotics *

Aluísio Pinheiro[1], Pranab Kumar Sen[2] and Hildete Prisco Pinheiro[1]

[1] *Departamento de Estatística* UNICAMP, Brazil

[2] *Department of Biostatistics* UNC-Chapel Hill

Abstract

In complex diversity analysis, specially arising in genetics, genomics, ecology and other high-dimensional (and sometimes low sample size) data models, typically subgroup-decomposability (analogous to ANOVA decomposability) arises. In group-divergence of diversity measures in a high-dimension low sample size scenario, it is shown that Hamming distance-type statistics lead to a general class of quasi U-statistics having a martingale (array) property, providing key to the study of general (nonstandard) asymptotics. Neither the stochastic independence nor homogeneity of the marginal probability laws play a basic role. A genomic MANOVA model is presented as an illustration.

1 Introduction

For the classical analysis of variance (ANOVA) models, a decomposition of the total sum of squares into two (or more) additive components provides the basis for statistical inference. Homoscedasticity and normality of errors insure the scope for exact (finite sample)

*AMS 2000 subject classifications: Primary-62G10; secondary-62G20, 92D20. Key words and phrases: Categorical Data, Dependence, DNA, Genomics, Hamming distance, Orthogonal system, Permutation measure, Second-order asymptotics, Second-order decomposability. Acknowledgment of support: this research was funded in part by FAEP-UNICAMP (412/04), FAPESP (03/10105-2) and CNPq (474329/2004_6)

inference. In multivariate (M)ANOVA, in addition, a larger sample size (depending on the dimension) is generally needed. A greater challenge is encountered in the analysis of high-dimensional (purely) qualitative data models which abound in genomics and bioinformatics. Typically, K, the number of positions (loci), is far larger than n, the number of sequences, i.e., $K >> n$. In some cases, even n may be small. Moreover, for such qualitative categorical data models, conventional discrete multivariate analysis tools are of little use.

For qualitative data (and even for quantitative ones), diversity analysis has evolved as a viable alternative for statistical modelling and analysis. To motivate our approach, in the next section, we introduce adequate diversity measures and incorporate them in a suitable MANOVA or group-divergence models for high-dimensional data models with special emphasis on categorical ones. A general formulation of Hamming distance type functionals underlies our approach. The sample counterparts of such functionals are suitable (generalized) U-statistics [3], and the proposed subgroup-decomposability has led to a class of quasi U-statistics. Since we have in mind the scenario $K >> n$, it is quite anticipated that standard statistical theory may not be appropriate here. Of particular interest is the distribution theory of such quasi U-statistics under the hypothesis of the homogeneity of several groups, which is the central theme of the present study.

Section 3 deals with the formulation of a general class of quasi U-statistics. A martingale (array) characterization is established in Section 4. This is then incorporated in the derivation of the main results. The curse of dimensionality problem arising in a genomics setup is focused in Section 5. The Hoeffding decomposition of U-statistics [4; 9] is extended here to quasi U-statistics. This decomposition along with the martingale property provide us with the necessary tool for the study of general (nonstandard) asymptotics (when $K >> n$). The concluding section deals with some general remarks.

2 Preliminary Notions

Let $X_1, \ldots, X_n$ be n independent and identically distributed (i.i.d.) random variables (r.v.), not necessarily continuous or even quantitative, having a distribution F. Whenever the X_i are quantitative, diversity or dispersion is measured in terms of the spread of F;

common measure is the standard deviation σ_F, where $\sigma_F^2 = \mathrm{E}(X - \mathrm{E}X)^2$ may also be written as

$$\sigma_F^2 = \mathrm{E}d(X_1, X_2), \text{ where } d(x,y) = \frac{1}{2}(x-y)^2. \tag{2.1}$$

Suppose now that we have a second sample $Y_1, \ldots, Y_m$ drawn independently from a distribution G, so that $\sigma_G^2 = \mathrm{E}d(Y_1, Y_2) = \frac{1}{2}\mathrm{E}(Y_1 - Y_2)^2$. Further, we denote by $\gamma(F,G) = \mathrm{E}[d(X,Y)]$, and note that

$$\gamma(F,G) = \frac{1}{2}(\sigma_F^2 + \sigma_G^2) + \frac{1}{2}(\mathrm{E}X - \mathrm{E}Y)^2 \geq \frac{1}{2}(\sigma_F^2 + \sigma_G^2), \quad \forall F, G, \tag{2.2}$$

where the equality sign holds only when $\mathrm{E}X = \mathrm{E}Y$. This simple inequality directly extends to the multi sample case and makes no specific homoscedasticity assumption.

Consider next a multivariate extension where the $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})'$ have mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and dispersion matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively. Then, again, we have

$$\Gamma(F,G) = \frac{1}{2}\mathrm{E}\{(\mathbf{X} - \mathbf{Y})(\mathbf{X} - \mathbf{Y})'\} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)', \tag{2.3}$$

so that for every $\boldsymbol{\lambda} \in \mathbb{R}^p$,

$$\begin{aligned} \boldsymbol{\lambda}'\Gamma(F,G)\boldsymbol{\lambda} &= \frac{1}{2}\{\boldsymbol{\lambda}'(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{\lambda}\} + \frac{1}{2}\|\boldsymbol{\lambda}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|^2 \\ &\geq \boldsymbol{\lambda}'\left\{\frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\right\}\boldsymbol{\lambda}, \quad \forall F, G, \end{aligned} \tag{2.4}$$

where the equality sign holds only when $\boldsymbol{\lambda}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$. If we choose a real-valued function $\phi(\mathbf{A})$ of a positive semi-definite (p.s.d.) $\mathbf{A}$ as a norm, we like to confine ourselves to a class of $\phi(\cdot)$, such that

$$\phi(\Gamma(F,G)) \geq \frac{1}{2}\left\{\phi(\boldsymbol{\Sigma}_1) + \phi(\boldsymbol{\Sigma}_2)\right\}, \quad \forall \phi \in \Phi, \tag{2.5}$$

where the equality sign holds only when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. Since the sample counterparts of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are both U-statistics and that of $\Gamma(F,G)$ is a generalized U-statistic, (2.5) can be incorporated in a MANOVA testing problem (for the homogeneity of the mean vectors) without putting too much emphasis on the homogeneity of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

Among the possible choices of $\phi(\cdot)$, the commonly used ones are (i) the generalized variance criterion $\phi(\mathbf{A}) = |\mathbf{A}|$, the determinant of $\mathbf{A}$ (usually raised to the power $1/p$),

(ii) the trace criterion $\phi(\mathbf{A}) = trace\mathbf{A}$, and (iii) Roy's [6] largest root criterion $\phi(\mathbf{A}) = ch_{max}(\mathbf{A})$. Note that

$$ch_{max}(\mathbf{A}) \geq \frac{1}{p} trace(\mathbf{A}) \geq |\mathbf{A}|^{1/p} \geq 0, \tag{2.6}$$

where $|\mathbf{A}|$ could be equal to zero only if $\mathbf{A}$ is not of full rank. For the trace criterion, (2.4) leads to (2.5) in an additive way, and it holds in a sub-additive way for the Roy's criterion too. The classical likelihood ratio test for the equality of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is based on the homogeneity assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, so that whenever $\boldsymbol{\Sigma}$ is positive definite (p.d.)

$$\boldsymbol{\Sigma}^{-1}\Gamma(F, G) = \mathbf{I}_p + \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)', \tag{2.7}$$

and hence, the generalized variance criterion applies as well.

Let us now turn our attention to qualitative data models. Suppose now X_i can have C qualitative (categorical) responses, labelled as $1, 2, \ldots, C$ (≥ 2), with respective probabilities $\pi_1, \pi_2, \ldots, \pi_C$. Thus, F is a multinomial distribution represented by $\boldsymbol{\pi}_F = (\pi_1^{(F)}, \pi_2^{(F)}, \ldots, \pi_C^{(F)})'$, defined on the simplex $\mathcal{S}_{C-1} = \{\mathbf{x} \in [0,1]^C : \mathbf{x}'\mathbf{1} = 1\}$. Quantitative measures of central tendency and spread (dispersion) are not meaningful in this context, yet diversity measures can be formulated in terms of the vector $\boldsymbol{\pi}_F$. [2] and [8], apparently unaware of Gini's work, proposed the measure

$$I(\boldsymbol{\pi}_F) = 1 - \boldsymbol{\pi}_F'\boldsymbol{\pi}_F = \sum_{c=1}^{C} \pi_c^{(F)}(1 - \pi_c^{(F)}). \tag{2.8}$$

This is known as the Gini-Simpson index (of bio-diversity) and has a simple interpretation if we let $d(X_1, X_2) = I(X_1 \neq X_2)$, then

$$I(\boldsymbol{\pi}_F) = \mathrm{E}_F[d(X_1, X_2)] = P\{X_1 \neq X_2\} = 1 - \boldsymbol{\pi}_F'\boldsymbol{\pi}_F. \tag{2.9}$$

In a similar manner, for the multinomial law G with a probability vector $\boldsymbol{\pi}_G$, we define the Gini-Simpson index as $I(\boldsymbol{\pi}_G)$. Let then

$$I(F, G) = P\{X \neq Y\} = 1 - P\{X = Y\} = 1 - \boldsymbol{\pi}_F'\boldsymbol{\pi}_G. \tag{2.10}$$

It is easy to verify that

$$I(F, G) \geq \frac{1}{2}\left\{I(\boldsymbol{\pi}_F) + I(\boldsymbol{\pi}_G)\right\}, \quad \forall F, G, \tag{2.11}$$

where the equality holds only when $F = G$ (or $\boldsymbol{\pi}_F = \boldsymbol{\pi}_G$), both defined on a common simplex $\mathcal{S}_{C-1}$. Thus, if one is basically interested in testing the homogeneity of $\boldsymbol{\pi}_F$ and $\boldsymbol{\pi}_G$ against such a diversity alternative, it seems desirable to incorporate the measures $I(F,G)$, $I(\boldsymbol{\pi}_F)$ and $I(\boldsymbol{\pi}_G)$ in the formulation. Sample counterparts of these measures are all (generalized) U-statistics, and that would afford the possibility of using established statistical methods for drawing statistical conclusions.

Consider now a multidimensional qualitative data model where the $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})'$ are K-vectors, each X_{ik} taking on one of the C (≥ 2) qualitative responses labelled as $1, 2, \ldots, C$. It is possible to allow the number of categories different for different k ($= 1, 2, \ldots, K$), but for simplicity of presentation, we sacrifice the generality.

Let $\mathbf{c} = (c_1, \ldots, c_k)'$ with each c_k taking on the labels $1, \ldots, C$, and let

$$\mathcal{C}_K = \{\mathbf{c} : c_k = 1, \ldots, C; k = 1, \ldots, K\} \tag{2.12}$$

so that the cardinality of $\mathcal{C}_K$ is equal to C^K. Let

$$\boldsymbol{\pi} = \{\pi(\mathbf{c}) : \mathbf{c} \in \mathcal{C}_K\}; \quad \pi(\mathbf{c}) = P\{\mathbf{X} = \mathbf{c}\}, \quad \mathbf{c} \in \mathcal{C}_K. \tag{2.13}$$

The distribution F of $\mathbf{X}$ relates to this multidimensional law with the probability set $\boldsymbol{\pi}$; we denote it $\boldsymbol{\pi}_F$. Similarly, for $\mathbf{Y}$ with a distribution G also defined on $\mathcal{C}_K$, the probability set is denoted by $\boldsymbol{\pi}_G$. We intend to compare $\boldsymbol{\pi}_F$ and $\boldsymbol{\pi}_G$ with emphasis on their diversities aspects, specially when K is very large, $K >> n$, thus creating a challenging statistical task.

In many applications there is a greater emphasis on marginal diversity measures which are to be combined into a single overall measure, so that one could incorporate the K marginal Gini-Simpson indexes in a way analogous to the trace criterion for the quantitative case. With that in mind, we introduce the Hamming distance as

$$\begin{aligned} d_H(\mathbf{X}_1, \mathbf{X}_2) &= K^{-1} \sum_{k=1}^{K} I(X_{ik} \neq X_{jk}) \\ &= K^{-1} \sum_{k=1}^{K} d(X_{ik}, X_{jk}), \end{aligned} \tag{2.14}$$

where $d(x,y) = I(x \neq y)$ is the distance function underlying the Gini-Simpson index. Let $F^{(k)}$, $k = 1, \ldots, K$, stand for the marginal distributions of the X_{ik}, $k = 1, \ldots, K$, and

let $I(F^{(k)})$ be the corresponding Gini-Simpson indexes. Then, the population Hamming distance $\mathcal{H}_F$ for F is

$$\mathcal{H}_F = E_F[d_H(\mathbf{X}_1, \mathbf{X}_2)] = \frac{1}{K}\sum_{k=1}^{K} P\{X_{1k} \neq X_{2k}\} = \frac{1}{K}\sum_{k=1}^{K} I(F^{(K)}). \tag{2.15}$$

Thus, $\mathcal{H}_F$ is a diversity measure for F based on the Hamming distance in (2.14). We define $\mathcal{H}_G$ in an analogous way. Further, let $\mathcal{H}(F,G) = E[d_H(\mathbf{X}, \mathbf{Y})]$, where $\mathbf{X}$ and $\mathbf{Y}$ come respectively from F and G; it is the Hamming distance between F and G. Using (2.11) for each marginal index, we obtain that

$$\begin{aligned} \mathcal{H}(F,G) &= \frac{1}{K}\sum_{k=1}^{K} I(F^{(k)}, G^{(k)}) \geq \frac{1}{2}\left\{\frac{1}{K}\sum_{k=1}^{K} I(F^{(k)}) + \frac{1}{K}\sum_{k=1}^{K} I(G^{(k)})\right\} \\ &= \frac{1}{2}\left\{\mathcal{H}_F + \mathcal{H}_G\right\}, \end{aligned} \tag{2.16}$$

where the equality sign holds only when $F^{(k)} \equiv G^{(k)}$, $\forall k = 1, \ldots, K$.

In order to cover both quantitative and qualitative data models, in a general multi-dimensional setup, we consider a set of n independent random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a distribution F, and define a parameter $\delta(F)$ as a distance functional of F. Let $\delta(F^{(k)})$ be a similar functional of the k-th marginal distribution $F^{(k)}$, $k = 1, \ldots, K$; $\boldsymbol{\delta}((\cdot)) = (\delta(F^{(1)}), \ldots, \delta(F^{(K)}))'$. Then, we assume that $\delta(F)$ is a convex combination of $\boldsymbol{\delta}((\cdot))$. For example, we may take a convex linear function:

$$\delta(F) = \boldsymbol{\lambda}'\boldsymbol{\delta}((\cdot)) \quad : \quad \boldsymbol{\lambda} \in \mathbb{R}^{+K} \text{ and } \boldsymbol{\lambda}'\boldsymbol{\lambda} = 1. \tag{2.17}$$

Whenever the elements of $\boldsymbol{\delta}((\cdot))$ satisfy (coordinatewise) an inequality similar to (2.11), (2.17) also leads to the same for $\delta(F)$. Next, we assume that the $\delta(F^{(k)})$ are estimable parameters (or regular functionals) in the sense of [3]. Keeping in mind that the $\delta(F^{(k)})$ are distance functions (that typically applies to a pair of points), we assume that there is a kernel of degree 2 and symmetric in its arguments such that

$$\delta(F^{(k)}) = \int\int \phi(x_1, x_2) dF^{(k)}(x_1) dF^{(k)}(x_2), \tag{2.18}$$

where $\phi(\cdot)$ is nonnegative. Further, we assume that

$$\delta(F^{(k)}, G^{(k)}) = \int\int \phi(x_1, x_2) dF^{(k)}(x_1) dG^{(k)}(x_2) \tag{2.19}$$

satisfy (2.11), i.e., for all F and G,

$$\delta(F^{(k)}, G^{(k)}) \geq \frac{1}{2}\left(\delta(F^{(k)}) + \delta(G^{(k)})\right), \quad \forall k = 1, \ldots, K. \tag{2.20}$$

This implies in turn that $\delta(F, G) \geq \{\delta(F) + \delta(G)\}/2, \quad \forall F, G$. Our proposed test for the homogeneity of G groups with respect to their diversity measures is based on the $\delta(F)$ and their sample counterparts which are all (generalized) U-statistics.

3 A Class of Quasi U-Statistics

Consider G (≥ 2) independent groups of samples drawn from distributions $F_1, \ldots, F_G$, of sizes $n_1, \ldots, n_G$ respectively, where all the F_g are K-dimensional and defined on a common probability space. As in (2.17)-(2.20), consider a nonnegative kernel $\phi(\mathbf{x}, \mathbf{y})$ (expressible as a convex linear compound of the componentwise kernels $\phi(x_k, y_k)$, $k = 1, \ldots, K$). Let $\mathbf{X}_{g1}, \ldots, \mathbf{X}_{gn_g}$ denote the observations (vectors) in the g-th group, $g = 1, \ldots, G$, and let

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}), \quad g = 1, \ldots, G. \tag{3.1}$$

Note that the $U_{n_g}^{(g)}$ are U-statistics [3] and are unbiased estimators of $\delta(F_g)$, $1 \leq g \leq G$. Similarly, let

$$U_{n_g, n_{g\prime}}^{(g,g\prime)} = \frac{1}{n_g n_{g\prime}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g\prime}} \phi(\mathbf{X}_{gi}, \mathbf{X}_{g\prime j}), \quad 1 \leq g < g\prime \leq G. \tag{3.2}$$

These generalized U-statistics are unbiased estimators of $\delta(F_g, F_{g\prime})$, which satisfy (2.20). For notational simplicity, we let $n = n_1 + \cdots + n_G$ and denote $U_{n_g}^{(g)}$ and $U_{n_g, n_{g\prime}}^{(g,g\prime)}$, by $U_{n,g}$ and $U_{n,gg\prime}$ respectively. Then, note that by definition, for the combined sample U_n, we have

$$U_n = \sum_{g=1}^{G} \frac{n_g(n_g - 1)}{n(n-1)} U_{n,g} + 2 \sum_{1 \leq g < g\prime \leq G} \frac{n_g n_{g\prime}}{n(n-1)} U_{n,gg\prime}, \tag{3.3}$$

which corresponds to the *within groups* and *between groups* components. However, to include the classical MANOVA as a special case, we proceed as in [7] and [5], and consider the following sub-group decomposition:

$$\begin{aligned} U_n &= \sum_{g=1}^{G} \frac{n_g}{n} U_{n,g} + \sum_{1 \leq g < g\prime \leq G} \frac{n_g n_{g\prime}}{n(n-1)} \{2U_{n,gg\prime} - U_{n,g} - U_{n,g\prime}\} \\ &= W_n + B_n, \text{ say.} \end{aligned} \tag{3.4}$$

In (3.3), the last term is nonnegative, while in (3.4), B_n could be both positive and negative. Under the hypothesis of homogeneity of the G groups, $E(B_n) = 0$, while $E(B_n) \geq 0$ under alternatives. Hence, we intend to use B_n as an appropriate test statistic. We need to standardize B_n appropriately so that it has a nondegenerate distribution.

If $F_1, \ldots, F_G$ are not all the same, $E(B_n) > 0$. Further, B_n, being a linear combination of generalized U-statistics, is attracted by the central limit theorem, and hence $n^{1/2}(B_n - E(B_n))$ will be asymptotically normal. The situation will be somewhat different when $F_1 \equiv \cdots \equiv F_G$ (i.e, the G groups are homogeneous). Our main interest centers on this nonstandard situation. First, $EB_n = 0$, and B_n can thereby assume both negative and positive values. Secondly, these generalized U-statistics are then stationary of order one (in the sense of [3]) for which generally non-normal asymptotic distributions prevail. To explore this situation fully, we note that when the G groups are homogeneous, $\mathbf{X}_{gi}, \quad 1 \leq i \leq n_g, \quad 1 \leq g \leq G$ are i.i.d.r.v.'s with a common distribution F. We let

$$\phi_1(\mathbf{x}) = E[\phi(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1 = \mathbf{x}]; \quad \phi_0 = E[\phi(\mathbf{X}_1, \mathbf{X}_2)]. \tag{3.5}$$

Then, we write

$$\begin{aligned} \phi(\mathbf{X}_1, \mathbf{X}_2) &= \phi_0 + \{\phi_1(\mathbf{X}_1) - \phi_0\} + \{\phi_1(\mathbf{X}_2) - \phi_0\} \\ &\quad + \{\phi(\mathbf{X}_1, \mathbf{X}_2)\} - \phi_1(\mathbf{X}_1) - \phi_1(\mathbf{X}_2) + \phi_0\} \\ &= \phi_0 + \psi_1(\mathbf{X}_1) + \psi_1(\mathbf{X}_2) + \psi_2(\mathbf{X}_1, \mathbf{X}_2), \end{aligned} \tag{3.6}$$

which is the Hoeffding decomposition of U-statistics [9]. The nice properties are (i) $\psi_1(\mathbf{X}_i)$ are i.i.d r.v.'s centered at 0, (ii) $\psi_2(\mathbf{X}_i, \mathbf{X}_j)$ are orthogonal and also centered at 0, so that $\mathrm{E}[\psi_2(\mathbf{X}_1, \mathbf{X}_2)\psi_2(\mathbf{X}_3, \mathbf{X}_4)] = 0$ and

$$\mathrm{E}[\psi_2(\mathbf{X}_1, \mathbf{X}_2)\psi_2(\mathbf{X}_1, \mathbf{X}_3)] = 0, \tag{3.7}$$

and (iii) relabelling the $\mathbf{X}_{gi}$, $1 \leq i \leq n_g$, $1 \leq g \leq G$ simply as $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with the convention that the first n_1 indexes relate to group 1, the next n_2 to group 2, ..., the last n_G to group G, we may rewrite B_n equivalently as

$$\binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \eta_{nij} \psi_2(\mathbf{X}_i, \mathbf{X}_j), \tag{3.8}$$

where

$$\eta_{nij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ come from different groups,} \\ -\frac{(n-n_g)}{(n_g-1)}, & \text{if } i \text{ and } j \text{ are both from group } g,\ 1 \le g \le G. \end{cases} \tag{3.9}$$

Thus,

$$\begin{aligned} &\sum_{1\le i<j\le n} \eta_{nij} = 0, \\ &\sum_{1\le i<j\le n} \eta_{nij}^2 = \binom{n}{2}(G-1)\left\{1 + \tfrac{1}{n}\textstyle\sum_{g=1}^{G} \dfrac{n-n_g}{(n_g-1)(G-1)}\right\}. \end{aligned} \tag{3.10}$$

Motivated by (3.8)-(3.10), in the next section, we proceed to study a general class of quasi U-statistics and their asymptotic properties.

4 Martingale Property

We consider a general statistic

$$T_n = \sum_{1\le i<j\le n} \eta_{nij}\phi(\mathbf{X}_i, \mathbf{X}_j), \quad n \ge 4, \tag{4.1}$$

where $\phi(x, y)$ is a first-order stationary kernel, centered at 0, and forms an orthogonal system for which

$$\mathrm{E}[\phi(\mathbf{X}_1, \mathbf{X}_2) \mid \mathbf{X}_1] = \phi_1(\mathbf{X}_1) = 0 \ a.e., \tag{4.2}$$

$$\mathrm{E}[\phi(\mathbf{X}_1, \mathbf{X}_2)\phi(\mathbf{X}_1, \mathbf{X}_3)] = 0,\ \mathrm{E}\phi^2(\mathbf{X}_1, \mathbf{X}_2) < \infty, \tag{4.3}$$

and the $\mathbf{X}_i$ are i.i.d.r.v.'s with a distribution F. Further, the (nonstochastic) $\eta_{nij}, 1 \le i < j \le n$, satisfy

$$\sum_{1\le i<j\le n} \eta_{nij} = 0, \quad \sum_{1\le i<j\le n} \eta_{nij}^2 = M_n(\nearrow \text{ in } n \ge 2). \tag{4.4}$$

Let $Z_{nj} = \sum_{i=1}^{j-1} \eta_{nij}\phi(\mathbf{X}_i, \mathbf{X}_j),\ j = 2, \ldots, n$ and let

$$T_{nk} = Z_{n2} + \ldots + Z_{nk},\ 2 \le k \le n;\ \ T_n = T_{nn}. \tag{4.5}$$

Further, let $\mathcal{B}_{nk} = \mathcal{B}(\mathbf{X}_i, i \le k)$ be a nondecreasing (in j) sub-sigma field generalized by the $\mathbf{X}_i,\ i \le k$, for $2 \le k \le n$.

Theorem 4.1 *For first-order stationary kernel, under (4.1)-(4.3),* $\{T_{nk}, \mathcal{B}_{nk} : 2 \le k \le n\}$ *is a (zero mean) martingale (array), closed on the right by* T_n.

Proof

We need to show that

$$\mathrm{E}(T_n \mid \mathcal{B}_{nk}) = T_{nk}, \ a.e., \ \forall 2 \le k \le n. \tag{4.6}$$

Since $T_n = T_{nk} + \sum_{j=k+1}^{n} Z_{nj}$, for $k < n$, it suffices to show that

$$\mathrm{E}(Z_{nk+1} \mid \mathcal{B}_{nk}) = 0 \ a.e., \text{ for every } k < n. \tag{4.7}$$

Towards this note that

$$\begin{aligned} \mathrm{E}(Z_{nk+1} \mid \mathcal{B}_{nk}) &= \sum_{i=1}^{k} \eta_{nik+1} E(\phi(\mathbf{X}_i, \mathbf{X}_{k+1}) \mid \mathcal{B}_{nk}) \\ &= \sum_{i=1}^{k} \eta_{nik+1} \phi_1(\mathbf{X}_i) = 0 \ \ a.e., \end{aligned} \tag{4.8}$$

by (4.2), for every $k = 1, \ldots, n-1$. □

Having the martingale (array) characterization, we are naturally tempted to incorporate suitable martingale array central limit theorems for our study of asymptotics for T_n in (4.1). In this context, we also note that the $\mathbf{X}_i$ are i.i.d.r.v.'s, so that their joint distribution remains invariant under any permutation of the indices $1, 2, \ldots, n$ (among themselves); this (discrete) uniform probability measure on $n!$ equally likely permutations is denoted by $\mathcal{P}_n$. Also, let

$$\begin{aligned} U_n^{(2)} &= \mathrm{E}_{\mathcal{P}_n}[\phi(\mathbf{X}_1, \mathbf{X}_2)] \\ &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \phi(\mathbf{X}_i, \mathbf{X}_j); \end{aligned} \tag{4.9}$$

$$\begin{aligned} U_n^{(3)} &= \mathrm{E}_{\mathcal{P}_n}[\phi(\mathbf{X}_1, \mathbf{X}_2)\phi(\mathbf{X}_1, \mathbf{X}_3)] \\ &= \frac{1}{n(n-1)(n-2)} \sum_{1 \le i \ne j \ne l \le n} \phi(\mathbf{X}_i, \mathbf{X}_j)\phi(\mathbf{X}_i, \mathbf{X}_l); \end{aligned} \tag{4.10}$$

$$\begin{aligned} U_n^{(2,2)} &= \mathrm{E}_{\mathcal{P}_n}[\phi^2(\mathbf{X}_1, \mathbf{X}_2)] \\ &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \phi^2(\mathbf{X}_i, \mathbf{X}_j). \end{aligned} \tag{4.11}$$

Note that $\mathcal{P}_n$ is a conditional (given the collection of n observations) probability measure, and $U_n^{(2)}$, $U_n^{(3)}$ and $U_n^{(2,2)}$ are all suitable U-statistics [3]. Then

$$\begin{aligned} \mathrm{E}_{\mathcal{P}_n}(T_n) &= \sum_{1\le i<j\le n} \eta_{nij} E_{\mathcal{P}_n}(\phi(\mathbf{X}_i,\mathbf{X}_j)) \\ &= \left(\sum_{1\le i<j\le n} \eta_{nij}\right) U_n^{(2)} \\ &= 0 \ \ a.e., \text{ by (4.2).} \end{aligned} \tag{4.12}$$

This also implies that $E(T_n) = E(E_{\mathcal{P}_n}(T_n)) = 0$. In the same way, defining

$$\begin{aligned} U_n^{(4)} &= \mathrm{E}_{\mathcal{P}_n}\{\phi(\mathbf{X}_1,\mathbf{X}_2)\phi(\mathbf{X}_3,\mathbf{X}_4)\} \\ &= \{n^{[4]}\}^{-1} \sum_{1\le i\ne j\ne r\ne s\le n} \phi(\mathbf{X}_i,\mathbf{X}_j)\phi(\mathbf{X}_r,\mathbf{X}_s), \end{aligned} \tag{4.13}$$

we obtain that

$$\begin{aligned} V_{\mathcal{P}_n}(T_n) &= \mathrm{E}_{\mathcal{P}_n}(T_n^2) = \sum_{1\le i<j\le n} \eta_{nij}^2 U_n^{(2,2)} + \\ &+ \sum\nolimits_1^* \eta_{nij}\eta_{nrs} U_n^{(3)} + \sum\nolimits_2^* \eta_{nij}\eta_{nrs} U_n^{(4)}, \end{aligned} \tag{4.14}$$

where the summation $\sum_1^*$ extends over $n^{[3]}$ terms: $(i,j),(r,s)$ for which exactly one index is common between (i,j) and (r,s), and $\sum_2^*$ over the remaining $\binom{n}{2}\binom{n-2}{2}$ terms for which i,j,r,s are all distinct. As $\sum_{1\le i<j\le n}\eta_{nij} = 0$, we have

$$\sum\nolimits_1^* \eta_{nij}\eta_{nrs} = -\sum_{1\le i<j\le n} \eta_{nij}^2 - \sum\nolimits_2^* \eta_{nij}\eta_{nrs}, \tag{4.15}$$

and further $\sum_2^* \eta_{nij}\eta_{nrs} = O(M_n^{3/2})$ if not $O(M_n)$. Thus,

$$\mathrm{E}_{\mathcal{P}_n}(T_n^2) = \left(U_n^{(2,2)} - U_n^{(3)}\right) \sum_{1\le i<j\le n} \eta_{nij}^2 + \left(U_n^{(4)} - U_n^{(3)}\right) \sum\nolimits_2^* \eta_{nij}\eta_{nrs}. \tag{4.16}$$

Now, under the condition $\tau_2 = \mathrm{E}\phi^2(\mathbf{X}_i,\mathbf{X}_j) < \infty$, $\mathrm{E}U_n^{(3)} = 0 = \mathrm{E}U_n^{(4)}$, and

$$U_n^{(2,2)}/\tau_2 \to 1 \quad \text{a.s./}L_1 - norm, \quad \text{as } n\to\infty, \tag{4.17}$$

while $U_n^{(3)}$ and $U_n^{(4)}$ are stationary of order 2 and 3 respectively, so that if $\mathrm{E}\phi^4(\mathbf{X}_1,\mathbf{X}_2) < \infty$ then letting (WLOG) $M_n = O(n^2)$.

$$\mathrm{E}(U_n^{(3)})^2/\tau_2^2 = O(n^{-3}); \quad \mathrm{E}(U_n^{(4)})^2/\tau_2^2 = O(n^{-4}). \tag{4.18}$$

Therefore, $|U_n^{(4)} - U_n^{(3)}|/\tau_2 = O_p(n^{-3/2})$, so that writing $V_n^* = U_n^{(2,2)} - U_n^{(3)}$, we have

$$V_n^* M_n / \mathrm{E}(T_n^2) \xrightarrow{p} 1, \text{ as } n \to \infty. \tag{4.19}$$

Motivated by these findings, we intend to study the asymptotics for T_n. In addition to (4.4) , we assume that as $n \to \infty$,

$$\sum_{1 \le i \ne j < k \le n} \eta_{nik}^2 \eta_{njk}^2 \, / M_n^2 \to 0 \, , \tag{4.20}$$

$$\sum_{1 \le i < j \le n} \eta_{nij}^4 \, / M_n^2 \to 0 \, . \tag{4.21}$$

(Whenever $max\{|\eta_{nij}| : 1 \le i < j \le n\} = o(\sqrt{M_n})$, both (4.20) and (4.21) hold).

Theorem 4.2 *If $\phi(\cdot,\cdot)$ is centered, stationary of order 1, $\mathrm{E}\phi^4(\mathbf{X}_1, \mathbf{X}_2) < \infty$, and if (4.2)-(4.4) and (4.20)-(4.21) hold, then as $n \to \infty$,*

$$L_n = (M_n V_n^*)^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1). \tag{4.22}$$

Proof: Led by Theorem 4.1, we let

$$v_{nk} = \mathrm{E}(Z_{nk}^2 \mid \mathcal{B}_{nk-1}), \; 2 \le k \le n \text{ and } V_n = \sum_{k=2}^{n} v_{nk}. \tag{4.23}$$

Then, by the martingale property (Theorem 4.1), for every $n \ge 2$,

$$\mathrm{E}(V_n) = \sum_{k=2}^{n} \mathrm{E}(Z_{nk}^2) = \mathrm{E}\left(\sum_{k=2}^{n} Z_{nk} \right)^2 = \mathrm{E}(T_n^2). \tag{4.24}$$

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

$$v_{nk} = \sum_{1 \le i < k} \eta_{nik}^2 \psi_2(\mathbf{X}_i) + \sum_{1 \le i \ne j < k} \eta_{nik} \eta_{njk} \psi_3(\mathbf{X}_i, \mathbf{X}_j), \tag{4.25}$$

where

$$\psi_2(\mathbf{X}_i) = \mathrm{E}[\phi^2(\mathbf{X}_i, \mathbf{X}_k) \mid \mathbf{X}_i] \; (\Rightarrow \mathrm{E}\psi_2(\mathbf{X}_i) = \tau_2); \tag{4.26}$$

$$\psi_3(\mathbf{X}_i, \mathbf{X}_j) = \mathrm{E}[\phi(\mathbf{X}_i, \mathbf{X}_k)\phi(\mathbf{X}_j, \mathbf{X}_k) \mid \mathbf{X}_i, \mathbf{X}_j] \;\; (i \ne j), \tag{4.27}$$

so that $\mathrm{E}(\mathbf{X}_i, \mathbf{X}_j) = \mathrm{E}U_n^{(3)} = 0,\ \forall\ i \neq j$. Therefore, $\mathrm{E}v_{nk} = \tau_2 \sum_{1\leq i<k} \eta_{nik}^2,\ \forall\ k \leq 2$, and hence, $\mathrm{E}V_n = \mathrm{E}T_n^2$, as expected from Theorem 4.1. Further,

$$\begin{aligned} V_n/\mathrm{E}T_n^2 &= \sum_{1\leq i<j\leq n} \eta_{nij}^2 \psi_2(\mathbf{X}_i)/\{M_n\tau_2\} \\ &+ \sum_{1\leq i\neq j<k\leq n} \eta_{nik}\eta_{njk}\psi_3(\mathbf{X}_i, \mathbf{X}_j)/\{M_n\tau_2\} \\ &= A_n + B_n, \quad say. \end{aligned} \tag{4.28}$$

By (4.21) and (4.26), $\mathrm{E}A_n = 1$ and $\mathrm{Var}(A_n) \to 0$ as $n \to \infty$, so that $A_n \xrightarrow{P} 1$ as $n \to \infty$. Also, $\mathrm{E}B_n = 0$, and

$$\mathrm{E}B_n^2 = \left(\mathrm{E}[\psi_3^2(\mathbf{X}_1, \mathbf{X}_2)]/\tau_2^2\right) \left(\sum_{1\leq i\neq j<k\leq n} \eta_{nik}^2 \eta_{njk}^2 \right) /M_n^2. \tag{4.29}$$

Thus, noting that $\mathrm{E}\psi_3^2(\mathbf{X}_1, \mathbf{X}_2) \leq \mathrm{E}\phi^4(\mathbf{X}_1, \mathbf{X}_2) < \infty$, by (4.20), we have $\mathrm{E}B_n^2 \to 0$ as $n \to \infty$ so that $B_n = o_p(1)$. Thus,

$$V_n/\mathrm{E}(T_n^2) \xrightarrow{p} 1, \text{ as } n \to \infty. \tag{4.30}$$

To establish (4.22), by virtue of (4.19) and (4.30), we are in a position to use the martingale (array) central limit theorem [1], and it suffices to verify the Lindeberg condition: $\forall\ \epsilon > 0$, as $n \to \infty$,

$$\sum_{k=2}^{n} \mathrm{E}(Z_{nk}^2 I(|Z_{nk}| > \epsilon\sqrt{\mathrm{E}(T_n^2)}))/\mathrm{E}(T_n^2) \to 0. \tag{4.31}$$

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 $\sum_{k=2}^{n} \mathrm{E}(Z_{nk}^4)/[\mathrm{E}(T_n^2)]^2 \to 0$ as $n \to \infty$, and the proof of this follows along the lines of (4.24)-(4.30). □

Remark: In the specific context of subgroup decomposability, treated in Section 3, the $|\eta_{nij}|$ are all bounded (and $O(1)$), and hence, (4.4) , (4.20) and (4.21) all hold, without requiring any further condition. As for the moment condition on $\phi(\cdot, \cdot)$, we may note that $\phi(\cdot, \cdot)$ is a function of a vector kernel depending on the K coordinates of the $\mathbf{X}_i$. This requires some delicate appraisal, and in the next section, we shall examine the case of Hamming distance based measures allowing inter-positions dependence to a certain extent.

5 The Genomics Case

Genomic sequences differ from the usual statistical data for two main reasons. First, because of their functional characteristics, inter-positions independence is optimistic at best. Sites relate by their molecular use and intricate (deterministic and stochastic) associations should be expected. Another source of complexity added to the analysis arises from the fact that it is not unusual to have a sample of n sequences of inidvidual lengths K, where $K >> n$.

In this section, we show that under mixing conditions one still has CLT for T_n. Theorem 5.1 shows that T_n is asymptotically normal when both K and n are large. If $K >> n$ (and n is even possibly bounded), one can not apply directly the martingale CLT but one still has a CLT for T_n, as long as some very mild moment and mixing conditions hold.

Theorem 5.1 *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. $K \times 1$ categorical random vectors. Let $\phi(\cdot,\cdot)$ be a first order U-statistics kernel such that*

$$\phi(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{K}\sum_{l=1}^{K} \phi^\star(\mathbf{X}_{il}, \mathbf{X}_{jl}), \tag{5.1}$$

for some first order U-statistics kernel $\phi^\star(\cdot,\cdot)$. Let T_n be defined by (4.1), and assume that all conditions in Theorem 4.1 hold.

Suppose that $\{\eta_{nij},\ 1 \le i < j \le n\ ,\ n \ge 1\}$ is a triangular array of random variables independent of $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\ ,\ n \ge 1\}$, and

$$\sum_{1 \le i < j \le n} \eta_{nij}^2 - M_n = o_p(M_n) \text{ as } n \to \infty. \tag{5.2}$$

Suppose also that

$$\sum_{1 \le l < m \le K} \mathrm{E}\left[\phi^\star(\mathbf{X}_{li}, \mathbf{X}_{lj})\phi^\star(\mathbf{X}_{mi}, \mathbf{X}_{mj})\right] = O(K) \text{ as } K \to \infty. \tag{5.3}$$

Then

$$(M_n V_n^\star)^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1) \quad \text{as } n \to \infty \text{ and } K \to \infty, \tag{5.4}$$

where $V_n^\star = U_n^{(2,2)} - U_n^{(3)}$, and $U_n^{(2,2)}$, $U_n^{(3)}$ are defined respectively by (4.9) and (4.10).

Proof

For the time being, we take the array $\{\eta_{n_{ij}} : 1 \leq i < j \leq n\}$, $n \geq 2$, such that $\sum_{1\leq i<j\leq n} \eta_{nij} = 0$ and $\sum_{1\leq i<j\leq n} \eta_{nij}^2 = \binom{n}{2}$. We also take $K \in \mathbb{N}$.

By Theorem 4.1, the martingale characterization of T_n is achieved and, by Theorem 4.2, (5.4) follows.

Next, consider the case of stochastic $\{\eta_{nij}, 1 \leq i < j \leq n\}$. As in the proof of Theorem 4.1, we take (WLOG) $M_n = \binom{n}{2}$. We assume that the $\eta_{n_{ij}}$ are independent of the $\mathbf{X}_i$, $i \leq n$, so that conditionally on $\eta_n = (\eta_{nij},\ 1 \leq i < j \leq n)$, the permutation law $\mathcal{P}_n$ remains intact. Let then $\bar{\eta}_n = \binom{n}{2}^{-1}\{\sum_{1\leq i<j\leq n} \eta_{n_{ij}}\}$, so that letting $\eta^\circ_{nij} = \eta_{nij} - \bar{\eta}_n$, $1 \leq i < j \leq n$, we have $\sum_{1\leq i<j\leq n} \eta^\circ_{n_{ij}} = 0$. Then, we can write

$$T_n = \sum_{1\leq i<j\leq n} \eta^\circ_{n_{ij}} \phi(\mathbf{X}_i, \mathbf{X}_j) + \binom{n}{2} \bar{\eta}_n U_n^{(2)}, \tag{5.5}$$

where $U_n^{(2)} = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} \phi(\mathbf{X}_i, \mathbf{X}_j) \to 0$ a.s./L_2-norm, and $U_n^{(2)} = O_p(n^{-1})$. As such, if with $n \to \infty$,

(a) $\bar{\eta}_n \xrightarrow{p} 0$, and

(b) $\sum_{1\leq i<j\leq n} {\eta^\circ_{n_{ij}}}^2 / \binom{n}{2} \xrightarrow{p} 1$,

then letting $T_n^\circ = \sum_{1\leq i<j\leq n} \eta^\circ_{n_{ij}} \phi(\mathbf{X}_i, \mathbf{X}_j)$, we have

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

so that by the Slutsky's Theorem:

$$(M_n V_n^\star)^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1).$$

Suppose now that K varies. If the weights are non-stochastic, following (5.1), and recalling that $0 < \tau_2 = E\phi^2(\mathbf{X}_i, \mathbf{X}_j) < \infty$,

$$K\tau_2 = \bar{\tau}_2 + \frac{2}{K} \sum_{1\leq l<m\leq K} \eta_{n12}^2 E\left[\phi^\star(\mathbf{X}_{1l}, \mathbf{X}_{2l})\phi^\star(\mathbf{X}_{1m}, \mathbf{X}_{2m})\right], \tag{5.6}$$

where $\bar{\tau}_2 = (1/K)\sum_{l=1}^K \eta_{n12}^2 E\left[\phi^{\star 2}(\mathbf{X}_{1l}, \mathbf{X}_{2l})\right]$. By (5.3), one has a finite limit for (5.6) when $K \to \infty$. Let $\tau_0 = \lim_{K\to\infty} K\tau_2$. Then,

$$KU_n^{(4)} \xrightarrow{p} \tau_0 \text{ as } n \to \infty \text{ and } K \to \infty$$

and (5.4) follows by [1]. So that, we have as $n \to \infty$,

$$(M_n V_N^\star)^{-1/2} T_n \xrightarrow{\mathcal{D}} N(0,1).$$

If we then consider random coefficients $\{\eta_{nij}; i,j = 1,2,\ldots,n; n \geq 2\}$, (5.4) will follow as well, because of (5.2) and (5.5). □

Theorem 5.2 *Let T_n be defined as in Theorem 5.1. Suppose that (5.3) holds. Then,*

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

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 from [10]. Notice that T_n can be written as $T_n = K^{-1} \sum_{k=1}^K t_{nk}$, where $t_{nk} = \sum_{1 \leq i < j \leq n} \eta_{nij} \phi^*(\mathbf{X}_{ik}, \mathbf{X}_{jk})$. We can take $S_n = K^{-1} \sum_{k=1}^K x_{nk}$, where $x_{nk} = t_{nk} / \sqrt{\binom{n}{2}}$.

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

The mixing condition (5.3) ensures the l-mixing [11]. Moreover, (5.3) 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 $n \to \infty$. □

Using the notation from Section 3, Hoeffiding's decomposition assures that $\phi(\cdot,\cdot)$ can be written as

$$\begin{aligned} \phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}) &= \theta_{gg} + \psi_{1g}(\mathbf{X}_{gi}) + \psi_{1g}(\mathbf{X}_{gj}) + \psi_{2gg}(\mathbf{X}_{gi}, \mathbf{X}_{gj}) \\ \phi(\mathbf{X}_{gi}, \mathbf{X}_{g'j}) &= \theta_{gg'} + \psi_{1g'}(\mathbf{X}_{gi}) + \psi_{1g}(\mathbf{X}_{g'j}) + \psi_{2gg'}(\mathbf{X}_{gi}, \mathbf{X}_{g'j}), \end{aligned}$$

for $g, g' = 1, \ldots, G$, where all the ψ's are centered.

Note that $\theta_{gg} = 0$ and $\psi_{1g}(\mathbf{X}_{gi}) = \psi_{1g}(\mathbf{X}_{gj}) = 0$ a.e. . Therefore one can write

$$\begin{aligned} T_n &= \sum_{g=1}^G \sum_{i \in \mathcal{I}_g, j \in \mathcal{I}_g} \eta_{nij} \psi_{2gg}(\mathbf{X}_{gi}, \mathbf{X}_{gj}) + \sum_{g<g'} \sum_{i \in \mathcal{I}_g, j \in \mathcal{I}_{g'}} \eta_{nij} \psi_{2gg'}(\mathbf{X}_{gi}, \mathbf{X}_{gj}) + \\ &\sum_{g<g'} \sum_{i \in \mathcal{I}_g, j \in \mathcal{I}_{g'}} \eta_{nij} \theta_{(gg')} + \sum_{g<g'} \sum_{i \in \mathcal{I}_g, j \in \mathcal{I}_{g'}} \eta_{nij} \left[\psi_{1g'}(\mathbf{X}_{gi}) + \psi_{1g}(\mathbf{X}_{gj}) \right]. \end{aligned} \tag{5.7}$$

Theorems 4.1, 4.2, 5.1 and 5.2 provide the asymptotic behavior of T_n for several setups, including $n \to \infty$ and/or $K \to \infty$, deterministic or random coefficients η_{nij}. These results however assume a sample drawn from a single situation. For hypothesis testing, one needs to deal with multipopulation samples. Theorem 5.3 summarizes the results for Pitman alternatives.

Theorem 5.3 *Let $\phi(\cdot,\cdot)$ be centered, stationary of order 1, $E\phi^4(\mathbf{X}_i,\mathbf{X}_j) < \infty$. Suppose $n = n_1 + \ldots + n_G$, for which n_i is the size of the sample drawn from the i-th distribution, $i = 1,\ldots,G$. Let $E\phi(\mathbf{X}_i,\mathbf{X}_j) = \theta_{rs}$, where $\mathbf{X}_i$ and $\mathbf{X}_j$ are drawn from the r-th and s-th populations, respectively. Consider the hypotheses $H_0 : \theta_{rs} = 0\ \forall r,s$ and $H_1 : \exists r \neq s$ s.t. $\theta_{rs} \neq 0$.*

Suppose the following moment conditions whithin each distribution, as follows,

$$\mathrm{E}[\phi(\mathbf{X}_i,\mathbf{X}_j) \mid \mathbf{X}_i] = 0 \ a.e., \tag{5.8}$$

$$\mathrm{E}[\phi(\mathbf{X}_i,\mathbf{X}_j)\phi(\mathbf{X}_i,\mathbf{X}_k)] = 0,\ \mathrm{E}\phi^2(\mathbf{X}_i,\mathbf{X}_j) < \infty, \tag{5.9}$$

for all $i \neq j \neq k$.

Suppose also that one of the following set of conditions holds:

*(**A.1**) - (4.4), (4.20) and (4.21) hold and $n \to \infty$.*

*(**B.1**) - $\{\eta_{nij},\ 1 \leq i < j \leq n,\ n \geq 1\}$ is a triangular array independent of $\mathbf{X}_1,\ldots,\mathbf{X}_n$ and $n, K \to \infty$.*

*(**B.2**) - (5.3) holds*

*(**C.1**) - $\{\eta_{nij},\ 1 \leq i < j \leq n,\ n \geq 1\}$ is a triangular array independent of $\mathbf{X}_1,\ldots,\mathbf{X}_n$.*

*(**C.2**) - $K \to \infty$ (either if $n \to \infty$, $n/K \to 0$, as $K \to \infty$ or if n is bounded).*

Finally, suppose a sequence of hypotheses $\{H_n\}$ contiguous to H_0. Then,

$$P_{H_n}\left(\frac{T_n}{\sqrt{M_n V_n^*}} > q_\alpha\right) \to 1 - \Phi\left(q_\alpha - \Delta\right), \tag{5.10}$$

where $E_{H_n}T_n - \sqrt{M_n\tau_0}\Delta \to 0$.

Proof

We will proceed WNLOG under conditions from Theorem 5.1 and $M_n = \binom{n}{2}$.

From (5.7),

$$T_n - ET_n = T_{n1} + T_{n2},$$

where the $\psi_{1g}, \psi_{1g'}$ etc. appear in T_{n1} in a linear (contrast) form, while T_{n2} is expressible as

$$\sum_{g,g'=1}^{G} \sum_{\{i,j\}} c_{gg'ij} \psi_{2gg'}(\mathbf{X}_{gi}, \mathbf{X}_{g'j})$$

where the $c_{gg'ij}$ are nonstochastic. Note that under $\{H_n\}$,

$$E\left[\psi_{2gg'}(\mathbf{X}_{gi}, \mathbf{X}_{g'j}) - \psi_{2gg}(\mathbf{X}_{gi}, \mathbf{X}_{g'j})\right]^2 \to 0, \text{ as } n \to 0.$$

Also,

$$E\left[\psi_{1g'}(\mathbf{X}_{gi}) - \psi_{1g}(\mathbf{X}_{gi})\right]^2 \to 0, \text{ as } n \to 0.$$

Therefore, noting that ET_n has the right order under H_n to match the variance factor, T_{n1} can be approximated up to the principal term by T_{n1}^0 in the null case, and similarly T_{n2}, we obtain the desired result. □

Remark: For the Hamming distance case, $\phi(X_{gi}, X_{g'j})$ and $\psi_{1g'}(X_{gi})$ are related by (3.6). For alternatives such that $\sqrt{K}\theta_{gg'}/\sqrt{n^2\tau_0} = O(1)$, $\forall g, g' = 1, \ldots, G$, $\psi_{1g'}(X_{gi})$ are functionals of the empirical distribution functions which differ at most by $O((n\sqrt{K})^{-1})$. Hence in those cases the contiguity condition in Theorem 5.3 can be relaxed to a weaker condition on $\theta_{gg'}$.

6 Discussion

A general class of quasi U-statistics is presented. We have proven that this class has a martingale array property and derive the asymptotic normality under a general set of conditions, for either large sample and/or high-dimensional setups.

The asymptotic normality results are valid for a large class of kernels. In particular, ANOVA-type decomposition tests in genomic studies as well as other diversity analysis can be theoretically justified by these results. One such case of special interest is that of Hamming distance genomic decomposition tests [7; 5].

References

[1] Dvoretzky , A. (1972). Central limit theorem for dependent random variables. *Proceedings Sixth Berkeley Symposium Math. Statist. Prob.* (Ed. L.LeCam et al.) Los Angeles: University of California Press, Vol.2 513–555.

[2] Gini, C. W. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari* **3***(2)* 3–159

[3] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

[4] Hoeffding, W. (1961). The strong law of large numbers for U-statistics. *Institute of Statistics Mimeo Series No 302, University of North Carolina.*

[5] Pinheiro, H.P., Pinheiro, A. and Sen, P.K. (2005). Comparison of genomic sequences using the Hamming Distance. *J. Statist. Plann. Inference* **130**(1-2) 325–339.

[6] Roy, S. N. (1953). One heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24**(3) 220–238.

[7] Sen, P. K. (1999). Utility-oriented Simpson-type indexes and inequality measures. *Calcutta Statist. Assoc. Bull.* **49** 1–22.

[8] Simpson, E. H. (1949). The measurement of diversity. *Nature* **163** 688.

[9] van Zwet, W. R. (1984). A Berry-Esseen bound for symmetric statistics. *Z. Wahrsch. und Verw. Gebiete* **66** 425–440.

[10] Withers, C. S. (1981). Central Limit Theorems for Dependent Variables. I. *Z. Wahrsch. und Verw. Gebiete* **57**(4) 509–534.

[11] Yoshihara, K. (1993). Asymptotic statistics based on weakly dependent data. *Weakly Dependent Stochastics Sequences and Their Applications, 2.* Sanseido, Tokyo.