Analysis of Variance for Hamming Distances Applied to Unbalanced Designs *

By Hildete Prisco Pinheiro
Department of Statistics
State University of Campinas, Brazil
Françoise Seillier-Moiseiwitsch
Department of Biostatistics
University of North Carolina at Chapel Hill, U.S.A.
Pranab Kumar Sen
Department of Biostatistics
University of North Carolina at Chapel Hill

Abstract

The interest here is the between- and within-group comparison of genomic sequences. All possible pairwise comparisons within and across groups are performed. Thus, unlike in analyses relying on measures of diversity (such as the Gini-Simpson index), sequences are considered on an individual basis. We develop a categorical analysis-of-variance framework for Hamming distances. This metric measures the proportion of positions at which two aligned sequences differ. We assume that the sequences are distantly related, but do not require that positions along the genome be independent. The total sum of squares is decomposed into within-, between- and across-group expressions. The latter term does not appear in the classical set-up. The theory of generalized U-statistics is utilized to find the asymptotic distribution of each sum of squares. Test statistics to assess homogeneity among groups are constructed.

1. Introduction

The focus of this paper lies in the comparison of genomic sequences, be they either DNA or protein sequences. These sequences are grouped in some way and the heterogeneity in these groups is quantified. The objective is to assess whether the variability is constant across groups. For instance, for sequences from the human immunodeficiency virus (HIV), a number of papers have investigated whether heterogeneity in the envelope gene is the same for all clades (see Seillier-Moiseiwitsch et al. (1994) for a review).

*This research was funded in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (202/92-02), Fundação de Amparo à Pesquisa do Estado de São Paulo (1998/12199-4), the National Science Foundation (DMS-9305588), the American Foundation for AIDS Research (70428-15-RF) and the National Institutes of Health (R29-GM49804 and P30-HD37260).

Key words and phrases: Amino Acid, Analysis of Variance, Categorical Data, Genome, Hamming Distance, Nucleotide, U-statistics, Bootstrap

Weir (1990) described an analysis of variance for the genetic variation in populations, as measured by the observed *heterozygosity*. The variance of the average heterozygosity is broken down to show the contribution of populations, loci and individuals by setting out the calculations as in an analysis of variance. Our situation is somewhat different because we would like to consider genomic regions rather than a small number of loci. This is why we selected the Hamming distance as our metric. The Hamming distance is the proportion of positions at which two aligned sequences differ. The sequences are regarded as independent but no assumption is imposed on the correlation between positions.

We develop an analysis-of-variance framework based on Hamming distances. The sequences are considered on an individual basis in the sense that they are compared to each other: all possible pairwise comparisons within and across groups are performed. We estimate the variability between, within and across groups (Section 2). In the within-group sum of squares, we are estimating the variability among sequences within a group around the average distance within this group. In the across-group sum of squares, we are estimating the variability of sequences across two groups with respect to the average distance between those groups. In the between-group sum of squares, we estimate the variability in the group average distances around the overall average. U-statistics are utilized to represent the average distance between and within groups as well as the overall distance (Sections 3, 4 and 5). The total sum of squares is decomposed into within-, between- and across-group sums of squares. The latter term does not appear in the classical set-up. The theory for generalized U-statistics (Puri & Sen, 1971; Lee, 1990; Sen & Singer, 1993) is used to find the asymptotic distributions of these sums of squares. In Section 6, test statistics are developed to assess homogeneity among groups. The power of the tests is discussed in Section 7. A data analysis is described briefly in Section 8. The paper closes with a discussion in Section 9.

2. The Total Sum of Squares and its Decomposition

Let X_{ik}^g be the label (i.e., an amino acid for protein sequences or a nucleotide $x_{ik}^g \in \{\mathbf{A}, \mathbf{C}, \mathbf{T}, \mathbf{G}\}$ for DNA sequences) present at position k ($k = 1, \ldots, K$) of sequence i ($i = 1, \ldots, N_g$) in group g ($g = 1, \ldots, G$). Then $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \ldots, X_{iK}^g)'$ is the random vector representing sequence i of group g. Consider $\mathbf{X}_i^g$ and $\mathbf{X}_j^{g'}$.

Definition
The *Hamming Distance* $D_{ij}^{(g,g')}$ between sequences i of group g and j of group g' is

$$\begin{aligned} D_{ij}^{(g,g')} &= \frac{1}{K}\sum_{k=1}^{K} I(X_{ik}^g \neq X_{jk}^{g'}) \\ &= \frac{1}{K} \times \text{number of positions at which } \mathbf{X}_i^g \text{ and } \mathbf{X}_j^{g'} \text{ differ} \end{aligned}$$

where I is the indicator function (i.e., $I(A) = 1$ when A is true, and 0 otherwise), and when $g = g'$,

$$D_{ij}^g = \frac{1}{K}\sum_{k=1}^{K} I(X_{ik}^g \neq X_{jk}^g)\,.$$

Let $\theta_k^g = P\{X_{ik}^g \neq X_{jk}^g\}$ and $\bar{\theta}_\cdot^g = \frac{1}{K}\sum_{k=1}^{K}\theta_k^g$. Then,

$$E[D_{ij}^g] = \frac{1}{K}\sum_{k=1}^{K} E[I(X_{ik}^g \neq X_{jk}^g)] = \frac{1}{K}\sum_{k=1}^{K}\theta_k^g = \bar{\theta}_\cdot^g \quad .$$

Define the average distance within a group as

$$\bar{D}_{\cdot}^{g} = \binom{N_g}{2}^{-1} \sum_{1\le i<j\le N} D_{ij}^{g} = \binom{N_g}{2}^{-1} \frac{1}{K} \sum_{1\le i<j\le N} \sum_{k=1}^{K} I(X_{ik}^{g} \neq X_{jk}^{g})$$

which is a U-statistic of degree 2 (Lee, 1990). The average distance between groups g and g' is

$$\bar{D}_{\cdot}^{(g,g')} = \frac{1}{N_g N_{g'}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} D_{ij}^{(g,g')} = \frac{1}{N_g N_{g'} K} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} \sum_{k=1}^{K} I(X_{ik}^{g} \neq X_{jk}^{g'})$$

which is a two-sample U-statistics of degree (1,1) (Hoeffding, 1948; Puri & Sen, 1971; Lee, 1990). The overall distance is

$$\begin{aligned}
\bar{D}_{\cdot} &= \left[\sum_{g=1}^{G} \binom{N_g}{2} + \sum_{1\le g<g'\le G} N_g N_{g'}\right]^{-1} \left(\sum_{g=1}^{G} \sum_{1\le i<j\le N_g} D_{ij}^{g} + \sum_{1\le g<g'\le G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} D_{ij}^{(g,g')}\right) \\
&= \binom{\sum_{g=1}^{G} N_g}{2}^{-1} \left(\sum_{g=1}^{G} \binom{N_g}{2} \bar{D}_{\cdot}^{g} + \sum_{1\le g<g'\le G} N_g N_{g'} \bar{D}_{\cdot}^{(g,g')}\right)
\end{aligned}$$

which is a linear combination of U-statistics.

The Total Sum of Squares can be decomposed as

$$TSS = \sum_{g=1}^{G} \sum_{1\le i<j\le N_g} (D_{ij}^{g} - \bar{D}_{\cdot})^2 + \sum_{1\le g<g'\le G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} (D_{ij}^{(g,g')} - \bar{D}_{\cdot})^2 \tag{2.1}$$

$$\begin{aligned}
&= \sum_{g=1}^{G} \sum_{1\le i<j\le N_g} (D_{ij}^{g} - \bar{D}_{\cdot}^{g})^2 + \sum_{g=1}^{G} \sum_{1\le i<j\le N_g} (\bar{D}_{\cdot}^{g} - \bar{D}_{\cdot})^2 \\
&\quad + \sum_{1\le g<g'\le G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} (D_{ij}^{(g,g')} - \bar{D}_{\cdot}^{(g,g')})^2 + \sum_{1\le g<g'\le G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} (\bar{D}_{\cdot}^{(g,g')} - \bar{D}_{\cdot})^2
\end{aligned} \tag{2.2}$$

$$= WSS + BSS + AWSS + ABSS$$

where WSS, BSS, $AWSS$ and $ABSS$ stand, respectively, for the within-, between-, across/within- and across/between-group sum of squares.

3. Connections Between Sums of Squares and U-statistics

There are G groups with N_g sequences each. Yet, we can disregard the group clustering and think of the sequences as a random sample of size $\sum_{g=1}^{G} N_g = M$. Then

$$\begin{aligned}
TSS &= \sum_{1\le i<j\le M} (D_{ij} - \bar{D}_{\cdot})^2 \\
&= \left(\frac{M(M-1)}{2} - 1\right) \binom{\frac{M(M-1)}{2}}{2}^{-1} \frac{1}{2} \sum_{\substack{i<j,\ i'<j' \\ i\le i' \text{ or } j\le j'}} (D_{ij} - D_{i'j'})^2
\end{aligned} \tag{3.1}$$

$$
\begin{aligned}
WSS &= \sum_{g=1}^{G} \sum_{1\le i<j\le N_g} (D_{ij}^g - \bar{D}_{\cdot}^g)^2 \\
&= \sum_{g=1}^{G} \left(\frac{N_g(N_g-1)}{2} - 1 \right) \binom{\frac{N_g(N_g-1)}{2}}{2}^{-1} \frac{1}{2} \sum_{\substack{i<j,\, i'<j' \\ i\le i' \text{ or } j\le j'}} (D_{ij}^g - D_{i'j'}^g)^2
\end{aligned} \tag{3.2}
$$

and

$$
\begin{aligned}
AWSS &= \sum_{1\le g<g'\le G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} (D_{ij}^{(g,g')} - \bar{D}_{\cdot}^{(g,g')})^2 \\
&= \sum_{1\le g<g'\le G} (N_g N_{g'} - 1) \binom{N_g N_{g'}}{2}^{-1} \frac{1}{2} \sum_{i=1}^{N_g} \sum_{j=1}^{N_{g'}} \sum_{i'=1}^{N_g} \sum_{\substack{j'=1 \\ i\le i' \text{ or } j\le j'}}^{N_{g'}} (D_{ij}^{(g,g')} - D_{i'j'}^{(g,g')})^2
\end{aligned} \tag{3.3}
$$

These sums of squares can also be expressed as linear combinations of U-statistics (Pinheiro, 1997). For instance, WSS is a linear combination of one-sample U-statistics of degrees 3 and 4, and $AWSS$ two-sample U-statistics of degrees (2,2) and (2,1).

4. Asymptotic Distributions and Decompositions of U-statistics

Let F denote the distribution function of X_i and U^m be a U-statistic of degree m, computed from a sample of size n, with kernel $\phi(X_1, \ldots, X_m)$ and $E(U^m) = \theta(F) = \theta$.

$$
U^m \equiv U(X_1, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{1\le i_1<\cdots<i_m\le n} \phi(X_{i_1}, \ldots, X_{i_m}), \quad n \ge m \tag{4.1}
$$

where $\theta(F) = E_F\{\phi(X_1, \ldots, X_m)\} = \int \ldots \int \phi(x_1, \ldots, x_m)\, dF(x_1) \ldots dF(x_m)$

Let

$$
\Psi_c(x_1, \ldots, x_c) \equiv \mathrm{E}\{\phi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\} \tag{4.2}
$$

$$
\psi_c(x_1, \ldots, x_c) \equiv \mathrm{E}\{\phi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m) - \theta\} \tag{4.3}
$$

$$
\xi_c \equiv \mathrm{E}\{\Psi_c^2(X_1, \ldots, X_c)\} - \theta^2 \quad \text{and} \quad \xi_0 \equiv 0. \tag{4.4}
$$

The function Ψ_c has the following properties (Lee, 1990, p. 11):
(i) $\Psi_c(x_1, \ldots, x_c) = \mathrm{E}\{\Psi_d(x_1, \ldots, x_c, X_{c+1}, \ldots, X_d)\}$ for $1 \le c < d \le m$,
(ii) $\mathrm{E}\{\Psi_c(x_1, \ldots, x_c)\} = \mathrm{E}\{\phi(X_1, \ldots, X_m)\}$.

Now

$$
\mathrm{Var}(U^m) = \binom{n}{m}^{-2} \sum_{c=0}^{m} \sum^{(c)} \mathrm{Cov}\{\phi(X_{i_1}, \ldots, X_{i_m}) \phi(X_{j_1}, \ldots, X_{j_m})\}
$$

where $\sum^{(c)}$ stands for summation over all subscripts such that

$$
1 \le i_1 < i_2 < \cdots < i_m \le n, \quad 1 \le j_1 < j_2 < \cdots < j_m \le n,
$$

and exactly c equations $i_k = j_h$ are satisfied. By (4.4), each term in $\sum^{(c)}$ is equal to ξ_c. The number of terms in $\sum^{(c)}$ is

$$\frac{n(n-1)\cdots(n-2m+c+1)}{c!(m-c)!(m-c)!} = \binom{m}{c}\binom{n-m}{m-c}\binom{n}{m} \tag{4.5}$$

Since $\xi_0 = 0$,

$$\mathrm{Var}(U^m) \quad = \quad \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\xi_c \tag{4.6}$$

The inequality $0 \le \xi_c \le \frac{c}{d}\xi_d$ with $1 \le c < d \le m$ (Hoeffding, 1948) leads to

$$\frac{m^2}{n}\xi_1 \le \mathrm{Var}(U^m) \le \frac{m}{n}\xi_m$$

From (4.6) and (4.5), $n\mathrm{Var}(U^m)$ is a decreasing function of n which tends to its lower bound $m^2\xi_1$ as n increases, i.e.,

$$\mathrm{Var}(U^m) \quad = \quad \frac{m^2}{n}\xi_1 + O(n^{-2}) \tag{4.7}$$

Therefore, if $E(\phi^2) < \infty$ and $\xi_1 > 0$,

$$n^{1/2}(U^m - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, m^2\xi_1) \tag{4.8}$$

(Hoeffding, 1948).

Definition 1
$F_n(x)$ is the empirical distribution function (d.f.)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \epsilon(x - X_i) \qquad x \in I\!R^p, \qquad n \ge 1$$

with $\epsilon(u)$ being 1 if all p coordinates of u are nonnegative and 0 otherwise. ■

We may rewrite (4.1) as

$$U^m = n^{-[m]} \sum_{1 \le i_1 \ne \cdots \ne i_m \le n} \int_{I\!R^{pm}} \cdots \int \phi(x_1, \ldots, x_m) \prod_{j=1}^{m} d(\epsilon(x_j - X_{i_j})),$$

where $n^{-[m]} = (n^{[m]})^{-1} = \{n \ldots (n-m+1)\}^{-1}$.

Writing $d(\epsilon(x_j - X_{i_j})) = dF(x_j) + d[\epsilon(x_j - X_{i_j}) - F(x_j)]$, $1 \le j \le m$, we obtain

$$U^m = \theta(F) + \sum_{h=1}^{m} \binom{m}{h} U_h^m \qquad n \ge m \tag{4.9}$$

where $U_h^m = n^{-[h]} \sum_{1 \le i_1 \ne \cdots \ne i_h \le n} \int_{I\!R^{ph}} \cdots \int \Psi_h(x_1, \ldots, x_h) \prod_{j=1}^{h} d[\epsilon(x_j - X_{i_j}) - F(x_j)]$ for $1 \le h \le m$. Further, if we write

$$\begin{aligned} \Psi_h^\circ(x_1, \ldots, x_h) \quad &= \quad \Psi_h(x_1, \ldots, x_h) - \sum_{j=1}^{h} \Psi_{h-1}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_h) \\ &\quad + \cdots + (-1)^h \theta(F), \qquad \forall\ (x_1, \ldots, x_h) \in I\!R^{ph}, \end{aligned} \tag{4.10}$$

for $1 \le h \le m$, we obtain

$$U_h^m = \binom{n}{h}^{-1} \sum_{1 \le i_1 < \cdots < i_h \le n} \Psi_h^\circ(X_{i_1}, \ldots, X_{i_h}), \quad 1 \le h \le m \tag{4.11}$$

and the U_h^m are themselves U-statistics. Note that for $h = 2$, we have

$$\begin{aligned} \mathrm{E}(U_2^m) &= \mathrm{E}(\Psi_2^\circ(X_1, X_2)) = \mathrm{E}(\Psi_2(X_1, X_2)) - \mathrm{E}(\Psi_1(X_1)) \\ &- \mathrm{E}(\Psi_1(X_2)) + \theta(F) \\ &= \theta(F) - \theta(F) - \theta(F) + \theta(F) = 0 \end{aligned}$$

Let

$$\Psi_{h,h-1}^\circ(x_1, \ldots, x_{h-1}) = \mathrm{E}[\Psi_h^\circ(X_1, \ldots, X_{h-1}, X_h) \mid X_1, \ldots, X_{h-1}]$$

Then

$$\begin{aligned} \Psi_{21}^\circ(X_1) &\equiv \mathrm{E}[\Psi_2^\circ(X_1, X_2) \mid X_1] \\ &= \mathrm{E}[\Psi_2(X_1, X_2) \mid X_1] - \mathrm{E}[\Psi_1(X_1) \mid X_1] - \mathrm{E}[\Psi_1(X_2) \mid X_1] + \theta(F) \\ &= \Psi_1(X_1) - \Psi_1(X_1) - \mathrm{E}(\Psi_1(X_2)) + \theta(F) \\ &= \theta(F) - \theta(F) = 0 \end{aligned}$$

$\xi_1^\circ \equiv \mathrm{E}\left[\Psi_{21}^\circ(X_1)\right]^2 - \left(\mathrm{E}(U_2^m)\right)^2 = 0$ and by (4.6),

$$\begin{aligned} \mathrm{Var}(U_2^m) &= \frac{4(n-2)}{n(n-1)}\xi_1^\circ + \frac{2\xi_2^\circ}{n(n-1)} \\ &= \frac{4}{n}\xi_1^\circ + O(n^{-2}) = O(n^{-2}) \end{aligned} \tag{4.12}$$

Consequently, $U_2^m = O_p(n^{-1})$.

From direct computation, $\mathrm{E}(U_h^m) = 0, \ \forall\, 1 \le h \le m$ and

$$\mathrm{Var}(U_h^m) = \mathrm{E}[(U_h^m)^2] = O(n^{-h}), \qquad h = 1, 2, \ldots, m; \tag{4.13}$$

and we can write

$$U^m = \theta(F) + \frac{m}{n}\sum_{i=1}^{n}[\Psi_1(X_i) - \theta(F)] + O_p(n^{-1}) \tag{4.14}$$

Now we consider multiple-sample U-statistics.

Let $\{\mathbf{X}_i^{(j)};\ i \ge 1\},\ j = 1, \ldots, c\,(\ge 2)$ be independent sequences of independent random vectors, where $\mathbf{X}_i^{(j)}$ has a distribution function $F^{(j)}(\mathbf{x}),\ \mathbf{x} \in I\!R^p$, for $j = 1, \ldots, c$. Let $\mathbf{F} = (F^{(1)}, \ldots, F^{(c)})$ and $\phi(\mathbf{X}_i^{(j)},\ 1 \le i \le m_j,\ 1 \le j \le c)$ be a Borel-measurable *kernel of degree* $\mathbf{m} = (m_1, \ldots, m_c)$, where without loss of generality we assume that ϕ is symmetric in the $m_j(\ge 1)$ arguments of the jth set, for $j = 1, \ldots, c$. Let $m_0 = m_1 + \cdots + m_c$ and

$$\theta(\mathbf{F}) = \int_{I\!R^{m_0}} \cdots \int \phi(\mathbf{x}_i^{(j)},\ 1 \le i \le m_j,\ 1 \le j \le c) \prod_{j=1}^{c}\prod_{i=1}^{m_j} dF^{(j)}(\mathbf{x}_i^{(j)}) \tag{4.15}$$

Definition 2
For a set of samples of sizes $\mathbf{n} = (n_1, n_2, \ldots, n_c)$ with $n_j \geq m_j,\ 1 \leq j \leq c$, $\mathbf{m} = (m_1, m_2, \ldots, m_c)$ the *generalized U-statistic* for $\theta(\mathbf{F})$ is

$$U^{(\mathbf{m})} = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{(\mathbf{n})}^{*} \phi(\mathbf{X}_{\alpha}^{(j)},\ \alpha = i_{j1}, \ldots, i_{jm_j},\ 1 \leq j \leq c), \tag{4.16}$$

where the summation $\sum_{(\mathbf{n})}^{*}$ extends over all $1 \leq i_{j1} < \ldots < i_{jm_j} \leq n_j,\ 1 \leq j \leq c$. $U^{(\mathbf{m})}$ is an unbiased estimator of $\theta(\mathbf{F})$. ■

Now, for every $d_j :\ 0 \leq d_j \leq m_j,\ 1 \leq j \leq c$, let $\mathbf{d} = (d_1, \ldots, d_c)$ and

$$\Psi_{d_1 \ldots d_c}(\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{d_j}^{(j)},\ 1 \leq j \leq c) \equiv \mathrm{E}(\phi(\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{d_j}^{(j)}, \mathbf{X}_{d_j+1}^{(j)}, \ldots, \mathbf{X}_{m_j}^{(j)},\ 1 \leq j \leq c)) \tag{4.17}$$

so that $\Psi_{\mathbf{0}} = \theta(\mathbf{F})$ and $\Psi_{\mathbf{m}} = \phi$. Then

$$\xi_{\mathbf{d}}(\mathbf{F}) = \mathrm{E}\left(\Psi_{\mathbf{d}}^2(\mathbf{X}_1^{(j)}, \ldots, \mathbf{X}_{d_j}^{(j)},\ 1 \leq j \leq c)\right) - \theta^2(\mathbf{F}), \qquad \mathbf{0} \leq \mathbf{d} \leq \mathbf{m} \tag{4.18}$$

so that $\xi_{\mathbf{0}}(\mathbf{F}) = 0$. Then, for every $\mathbf{n} \geq \mathbf{m}$ (Sen, 1981),

$$\mathrm{Var}\left[U^{(\mathbf{m})}\right] \ = \ \sum_{j=1}^{c} n_j^{-1} \sigma_j^2 \left[1 + O(n_0^{-1})\right] \tag{4.19}$$

where $n_0 = \min(n_1, \ldots, n_c)$ and

$$\sigma_j^2 = m_j^2\, \xi_{\delta_{j1}, \ldots, \delta_{jc}}(\mathbf{F}) \quad j = 1, \ldots, c \tag{4.20}$$

with $\delta_{\alpha\beta} = 1$ or 0 according to whether $\alpha = \beta$ or not.

For a two-sample U-statistic of degree (m_1, m_2), the kernel is

$$\phi(X_1, \ldots, X_{m_1}; Y_1, \ldots, Y_{m_2})$$

and

$$U^{(m_1, m_2)} = \binom{n_1}{m_1}^{-1} \binom{n_2}{m_2}^{-1} \sum_{(\mathbf{n}, \mathbf{m})}^{*} \phi(X_{i_1 1}, \ldots, X_{i_1 m_1}; Y_{i_2 1}^{(2)}, \ldots, Y_{i_2 m_2})$$

where $\sum_{(\mathbf{n}, \mathbf{m})}^{*}$ extends over all $1 \leq i_{j1} < \ldots < i_{jm_j} \leq n_j,\ j = 1, 2$. $U^{(m_1, m_2)}$ is an unbiased estimator of $\theta(F^{(1)}, F^{(2)})$.

Define

$$\begin{aligned} &\theta(F_{n_1}^{(1)}, F_{n_2}^{(2)}) \\ &\quad = \ n_1^{-m_1} n_2^{-m_2} \times \sum_{i_{11}=1}^{n_1} \cdots \sum_{i_{1m_1}=1}^{n_1} \sum_{i_{21}=1}^{n_2} \cdots \sum_{i_{2m_2}=1}^{n_2} \phi(X_{i_1 1}, \ldots, X_{i_1 m_1}, Y_{i_2 1}, \ldots, Y_{i_2 m_2}) \end{aligned}$$

Then,

$$| U^{(m_1,m_2)} - \theta(F_{n_1}^{(1)}, F_{n_2}^{(2)}) | = O_p(n_0^{-1}); \quad n_0 = \min(n_1, n_2) \tag{4.21}$$

provided the variance of $U^{(m_1,m_2)}$ exists.

$$\begin{aligned}
&\Psi_{d_1 d_2}(x_1, \ldots, x_{d_1}, y_1, \ldots, y_{d_2}) \\
&= E\{\phi(x_1, \ldots, x_{d_1}, X_{d_1+1}, \ldots, X_{m_1}; y_1, \ldots, y_{d_2}, Y_{d_2+1}, \ldots, Y_{m_2})\} ,
\end{aligned}$$

$$\xi_{d_1 d_2} = E\{\Psi_{d_1 d_2}^2(X_1, \ldots, X_{d_1}, Y_1, \ldots, Y_{d_2})\} - \theta^2(F^{(1)}, F^{(2)}) \tag{4.22}$$

for $d_1 = 0, \ldots, m_1$, $d_2 = 0, \ldots, m_2$. ($\xi_{00} \equiv 0$). Then,

$$(\gamma_{n_1,n_2})^{-1}(U^{(m_1,m_2)} - \theta(F^{(1)}, F^{(2)})) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1), \tag{4.23}$$

where

$$\gamma_{n_1,n_2}^2 = \left(\frac{m_1^2}{n_1}\right)\xi_{10} + \left(\frac{m_2^2}{n_2}\right)\xi_{01}. \tag{4.24}$$

The decomposition for $U^{(\mathbf{m})}$ can be developed similarly to the one-sample U-statistic. For a two-sample U-statistic of degree (m_1, m_2), we have

$$\begin{aligned}
U^{(m_1,m_2)} &= \theta(\mathbf{F}) + \frac{m_1}{n_1}\sum_{i=1}^{n_1}[\Psi_{10}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2}\sum_{i=1}^{n_2}[\Psi_{01}(Y_i) - \theta(\mathbf{F})] \\
&+ O_p(n_0^{-1})
\end{aligned} \tag{4.25}$$

where $n_0 = min(n_1, n_2)$.

The above expression can be generalized for multiple-sample U-statistics. For instance, the decomposition for a three-sample and four-sample U-statistics are as follows

$$\begin{aligned}
U^{(m_1,m_2,m_3)} &= \theta(\mathbf{F}) + \frac{m_1}{n_1}\sum_{i=1}^{n_1}[\Psi_{100}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2}\sum_{i=1}^{n_2}[\Psi_{010}(Y_i) - \theta(\mathbf{F})] \\
&+ \frac{m_3}{n_3}\sum_{i=1}^{n_3}[\Psi_{001}(Z_i) - \theta(\mathbf{F})] + O_p(n_0^{-1})
\end{aligned} \tag{4.26}$$

where $n_0 = min(n_1, n_2, n_3)$ and

$$\begin{aligned}
U^{(m_1,m_2,m_3,m_4)} &= \theta(\mathbf{F}) + \frac{m_1}{n_1}\sum_{i=1}^{n_1}[\Psi_{1000}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2}\sum_{i=1}^{n_2}[\Psi_{0100}(Y_i) - \theta(\mathbf{F})] \\
&+ \frac{m_3}{n_3}\sum_{i=1}^{n_3}[\Psi_{0010}(Z_i) - \theta(\mathbf{F})] + \frac{m_4}{n_4}\sum_{i=1}^{n_4}[\Psi_{0001}(W_i) - \theta(\mathbf{F})] \\
&+ O_p(n_0^{-1})
\end{aligned} \tag{4.27}$$

where $n_0 = min(n_1, n_2, n_3, n_4)$.

5. Combining the U-statistics

We can write

$$\begin{aligned}
WSS &= \sum_{g=1}^{G}\frac{(N_g - 2)}{3}[\mathbf{U_{1,1}^{(3)}} + \mathbf{U_{1,2}^{(3)}} + \mathbf{U_{1,3}^{(3)}}] \\
&+ \sum_{g=1}^{G}\frac{(N_g - 2)(N_g - 3)}{12}[\mathbf{U_{2,1}^{(4)}} + \mathbf{U_{2,2}^{(4)}} + \mathbf{U_{2,3}^{(4)}}]
\end{aligned}$$

where

$$\mathbf{U}^{(\mathbf{3})}_{\mathbf{1,1}} = \binom{N_g}{3}^{-1} \sum_{i<j<j'} (D^g_{ij} - D^g_{ij'})^2 \,, \qquad \mathbf{U}^{(\mathbf{3})}_{\mathbf{1,2}} = \binom{N_g}{3}^{-1} \sum_{i<i'<j} (D^g_{ij} - D^g_{i'j})^2 \quad \text{and}$$

$$\mathbf{U}^{(\mathbf{3})}_{\mathbf{1,3}} = \binom{N_g}{3}^{-1} \sum_{i<j<j'} (D^g_{ij} - D^g_{jj'})^2$$

are one-sample U-statistics of degree 3 and

$$\mathbf{U}^{(\mathbf{4})}_{\mathbf{2,1}} = \binom{N_g}{4}^{-1} \sum_{i<j<i'<j'} (D^g_{ij} - D^g_{i'j'})^2 \,, \quad \mathbf{U}^{(\mathbf{4})}_{\mathbf{2,2}} = \binom{N_g}{4}^{-1} \sum_{i<i'<j<j'} (D^g_{ij} - D^g_{i'j'})^2 \quad \text{and}$$

$$\mathbf{U}^{(\mathbf{4})}_{\mathbf{2,3}} = \binom{N_g}{4} \sum_{i<i'<j'<j} (D^g_{ij} - D^g_{i'j'})^2$$

are one-sample U-statistics of degree 4.

Note that $U^{(m)}_{j,k}$ represents a U-statistic of degree m with expected value μ_j, i.e., $E(U^{(m)}_{j,k}) = \mu_j$ for any k. The expected value of WSS is

$$E(WSS) \;=\; \sum_{g=1}^{G} (N_g - 2) \left\{ \mu_{1g} + \frac{(N_g - 3)}{4} \mu_{2g} \right\} \;.$$

where $\mu_{1g} = E(\mathbf{U}^{(\mathbf{3})}_{\mathbf{1,1}}) = E(\mathbf{U}^{(\mathbf{3})}_{\mathbf{1,2}}) = E(\mathbf{U}^{(\mathbf{3})}_{\mathbf{1,3}}) = E(D^g_{ij} - D^g_{ij'})^2$ and
$\mu_{2g} = E(\mathbf{U}^{(\mathbf{4})}_{\mathbf{2,1}}) = E(\mathbf{U}^{(\mathbf{4})}_{\mathbf{2,2}}) = E(\mathbf{U}^{(\mathbf{4})}_{\mathbf{2,3}}) = E[(D^g_{ij} - D^g_{i'j'})^2]$.
Under H_0, there is homogeneity among groups, i.e., for any g, $\theta^g_k = \theta_k$ and
$\theta^g_{k_1k_2} = \theta_{k_1k_2}$. Therefore, $\mu_{1g} = \mu_1$ and $\mu_{2g} = \mu_2$, thus

$$E_0(WSS) \;=\; \sum_{g=1}^{G} (N_g - 2) \left\{ \mu_1 + \frac{(N_g - 3)}{4} \mu_2 \right\}$$

where

$$\mu_{1g} = \mu_1 = \frac{2}{K^2} \left[\sum_{k=1}^{K} \theta_k + \sum_{k_1 \neq k_2} \theta_{k_1k_2} - \sum_{k=1}^{K} \theta_k(i,j;i,j') - \sum_{k_1 \neq k_2} \theta_{k_1k_2}(i,j;i,j') \right] \tag{5.1}$$

and

$$\mu_{2g} = \mu_2 = \frac{2}{K^2} \left\{ \sum_{k=1}^{K} \theta_k(1-\theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1k_2} - \theta_{k_1}\theta_{k_2}) \right\} \tag{5.2}$$

with

$$\theta_k = \theta_k(i,j) = \mathrm{P}(X_{ik} \neq X_{jk}) = \sum_{c=0}^{C-1} p_k(c)[1 - p_k(c)] \tag{5.3}$$

$$\begin{aligned} \theta_{k_1k_2} \;&=\; \theta_{k_1,k_2}(i,j) = \mathrm{P}(X_{ik_1} \neq X_{jk_1}; X_{ik_2} \neq X_{jk_2}) \\ &=\; \sum_{c_1,c_2=0}^{C-1} p_{k_1k_2}(c_1,c_2) \left[\sum_{\substack{c_3=0 \\ c_3 \neq c_1}}^{C-1} \sum_{\substack{c_4=0 \\ c_4 \neq c_2}}^{C-1} p_{k_1k_2}(c_3,c_4) \right] \end{aligned} \tag{5.4}$$

$$\theta_k^g(i,j;i,j') = P(X_{ik}^g \neq X_{jk}^g,\, X_{ik}^g \neq X_{j'k}^g) = \sum_{c=0}^{C-1} p_k^g(c)[1-p_k^g(c)]^2, \tag{5.5}$$

$$\begin{aligned}\theta_{k_1k_2}^g(i,j;i,j') &= P(X_{ik_1}^g \neq X_{jk_1}^g,\, X_{ik_2}^g \neq X_{j'k_2}^g) \\ &= \sum_{c_1,\,c_2=0}^{C-1} p_{k_1k_2}^g(c_1,c_2)[1-p_{k_1}^g(c_1)][1-p_{k_2}^g(c_2)],\end{aligned} \tag{5.6}$$

$$p_k(c) = P(X_{ik}^g = c) \quad \text{and} \quad p_{k_1k_2}^g(c_1,c_2) = P(X_{ik_1}^g = c_1, X_{ik_2}^g = c_2) \tag{5.7}$$

Decomposing WSS, under H_0,

$$\begin{aligned}WSS &= \sum_{g=1}^{G}(N_g-2)\left(\mu_1 + \frac{(N_g-3)}{4}\mu_2\right) \\ &+ \sum_{g=1}^{g}(N_g-2)\frac{3}{N_g}\sum_{i=1}^{N_g}[\Psi_{(1)1}(\mathbf{X}_i)-\mu_1] + O_p(1) \\ &+ \sum_{g=1}^{G}\frac{(N_g-2)(N_g-3)}{N_g}\sum_{i=1}^{N_g}[\Psi_{(2)1}(\mathbf{X}_i)-\mu_2] + O_p(N_g)\end{aligned} \tag{5.8}$$

and the associated mean square expression is

$$\begin{aligned}WMS &\equiv \frac{WSS}{\sum_{g=1}^{G}\binom{N_g}{2}} = \frac{2WSS}{\sum_{g=1}^{G}N_g(N_g-1)} \\ &= \frac{2}{\sum_{g=1}^{G}N_g(N_g-1)}\left\{\sum_{g=1}^{G}(N_g-2)\left(\mu_1 + \frac{(N_g-3)}{4}\mu_2\right)\right. \\ &+ \sum_{g=1}^{g}(N_g-2)\frac{3}{N_g}\sum_{i=1}^{N_g}[\Psi_{(1)1}(\mathbf{X}_i)-\mu_1] + O_p(1) \\ &+ \left.\sum_{g=1}^{G}\frac{(N_g-2)(N_g-3)}{N_g}\sum_{i=1}^{N_g}[\Psi_{(2)1}(\mathbf{X}_i)-\mu_2] + O_p(N_g)\right\} \\ &= \frac{\sum_{g=1}^{G}(N_g-2)(N_g-3)}{2\sum_{g=1}^{G}N_g(N_g-1)}\left\{\mu_2 + \frac{4}{N_g}\sum_{i=1}^{N_g}(\Psi_{(2)1}(\mathbf{X}_i)-\mu_2)\right\} + O_p(N_0^{-1})\end{aligned} \tag{5.9}$$

with

$$\mathrm{E}_0(WMS) = \frac{\mu_2}{2} + O(N_g^{-1})$$

and

$$\begin{aligned}\mathrm{Var}_0(WMS) &= \frac{4\sum_{g=1}^{G}[N_g^3 - 10N_g^2 + 37N_g - 60 + 36/N_g]}{[\sum_{g=1}^{G}N_g(N_g-1)]^2}\xi_1^{(2)} + O(N_g^{-2}) \\ &= \frac{4\sum_{g=1}^{G}N_g^3}{[\sum_{g=1}^{G}N_g(N_g-1)]^2}\xi_1^{(2)} + O(N_g^{-2})\end{aligned} \tag{5.10}$$

For $AWSS$,

$$\begin{aligned} AWSS &= \sum_{1\le g<g'\le G}\Bigg[\frac{(N_g-1)(N_{g'}-1)}{4}\left(\mathbf{U}_{4,1}^{(2,2)}+\mathbf{U}_{4,2}^{(2,2)}\right) \\ &+ \frac{(N_g-1)}{2}\mathbf{U}_{5,2}^{(2,1)}+\frac{(N_{g'}-1)}{2}\mathbf{U}_{5,1}^{(1,2)}\Bigg] \end{aligned}$$

where

$$\mathbf{U}_{\mathbf{4,1}}^{(\mathbf{2,2})} = \left[\binom{N_g}{2}\binom{N_{g'}}{2}\right]^{-1}\sum_{i\neq i'}\sum_{\substack{j\neq j'\\ j\neq i'}}(D_{ij}^{(g,g')}-D_{i'j'}^{(g,g')})^2 \text{ and}$$

$$\mathbf{U}_{\mathbf{4,2}}^{(\mathbf{2,2})} = \left[\binom{N_g}{2}\binom{N_{g'}}{2}\right]^{-1}\sum_{\substack{i\neq j\\ i\neq i'}}\sum_{j\neq j'}(D_{ij}^{(g,g')}-D_{i'j'}^{(g,g')})^2$$

are two-sample U-statistics of degree (2,2) and

$$\mathbf{U}_{\mathbf{5,1}}^{(\mathbf{1,2})} = \left[\binom{N_g}{1}\binom{N_{g'}}{2}\right]^{-1}\sum_{\substack{i=1\\ i\neq j'}}^{N_g}\sum_{\substack{1\le j,\, j'\le N_{g'}\\ j\neq j'}}(D_{ij}^{(g,g')}-D_{ij'}^{(g,g')})^2 \quad \text{and}$$

$$\mathbf{U}_{\mathbf{5,2}}^{(\mathbf{2,1})} = \left[\binom{N_g}{2}\binom{N_{g'}}{1}\right]^{-1}\sum_{\substack{j=1\\ j\neq i'}}^{N_{g'}}\sum_{\substack{1\le i,i'\le N_g\\ i\neq i'}}(D_{ij}^{(g,g')}-D_{i'j}^{(g,g')})^2$$

are two sample U-statistics of degree (1,2) and (2,1), respectively.

$$\mathrm{E}(AWSS) = \sum_{1\le g<g'\le G}\left[\frac{(N_g-1)(N_{g'}-1)}{2}\mu_{4(g,g')}+\frac{(N_g-1)}{2}\mu_{5(g,g')}+\frac{(N_{g'}-1)}{2}\mu_{5(g,g')}\right]$$

with $\mu_{4(g,g')}=E(\mathbf{U}_{4,1}^{(2,2)})=E(\mathbf{U}_{4,2}^{(2,2)})$ and $\mu_{5(g,g')}=E(\mathbf{U}_{5,1}^{(1,2)})=E(\mathbf{U}_{5,2}^{(2,1)})$.

Under H_0, $\mu_{4(g,g')}=\mu_4$ and $\mu_{5(g,g')}=\mu_5$, therefore

$$\mathrm{E}_0(AWSS) = \frac{1}{2}\sum_{1\le g<g'\le G}[(N_g-1)(N_{g'}-1)\mu_4+(N_g+N_{g'}-2)\mu_5]$$

where $\mu_4=\mu_2$ is given by (5.2) and $\mu_5=\mu_1$ is given by (5.1).

$AWSS$ can be decomposed as

$$\begin{aligned} AWSS &= \sum_{1\le g<g'\le G}\Bigg[\frac{(N_g-1)(N_{g'}-1)}{2}\Bigg(\mu_{4(g,g')}+\frac{2}{N_g}\sum_{i=1}^{N_g}(\Psi_{(4)10}(\mathbf{X}_i^g)-\mu_{4(g,g')}) \\ &+ \frac{2}{N_{g'}}\sum_{j=1}^{N_{g'}}(\Psi_{(4)01}(\mathbf{X}_j^{g'})-\mu_{(g,g')4})+O_p(N_0^{-1})\Bigg) \\ &+ \frac{(N_g-1)}{2}\Bigg(\mu_{5(g,g')}+\frac{2}{N_g}\sum_{i=1}^{N_g}(\Psi_{(5)10}(\mathbf{X}_i^g)-\mu_{5(g,g')}) \\ &+ \frac{1}{N_{g'}}\sum_{j=1}^{N_{g'}}(\Psi_{(5)01}(\mathbf{X}_j^{g'})-\mu_{5(g,g')})\Bigg) \end{aligned}$$

$$+ \quad \frac{(N_{g'}-1)}{2}\left(\mu_{5(g,g')} + \frac{1}{N_g}\sum_{i=1}^{N_g}(\Psi_{(5)10}(\mathbf{X}_i^g) - \mu_{5(g,g')}) \right.$$

$$\left. + \quad \frac{2}{N_{g'}}\sum_{j=1}^{N_{g'}}(\Psi_{(5)01}(\mathbf{X}_j^{g'}) - \mu_{5(g,g')}) + O_p(N_0^{-1})\right)\Bigg] \tag{5.11}$$

with $N_0 = \min_{1\le g\le G}(N_g)$.

The associated mean-square expression is

$$\begin{aligned} AWMS &= \frac{AWSS}{\sum_{g<g'} N_g N_{g'}} \\ &= \frac{1}{2}\sum_{1\le g<g'\le G} \frac{(N_g-1)(N_{g'}-1)}{\left[\sum_{g<g'} N_g N_{g'}\right]}\Bigg[\mu_{4(g,g')} + \frac{2}{N_g}\sum_{i=1}^{N_g}(\Psi_{(4)10}(\mathbf{X}_i^g) - \mu_{4(g,g')}) \\ &+ \frac{2}{N_{g'}}\sum_{j=1}^{N_{g'}}(\Psi_{(4)01}(\mathbf{X}_j^{g'}) - \mu_{4(g,g')})\Bigg] + O_p(N_0^{-1}) \end{aligned} \tag{5.12}$$

$$\mathrm{E}_0(AWMS) = \frac{\mu_4}{2} + O(N_0^{-1})$$

Note that $\mathrm{E}_0(AWMS) = \mathrm{E}_0(WMS)$, since under H_0, $\mu_4 = \mu_2$.

$$\begin{aligned} \mathrm{Var}_0(AWMS) &= \frac{1}{(\sum_{g<g'} N_g N_{g'})^2}\left\{\sum_{g<g'}\frac{(N_g-1)^2(N_{g'}-1)^2}{4}\left(\frac{4}{N_g}\xi_{10}^{(4)} + \frac{4}{N_{g'}}\xi_{01}^{(4)}\right)\right\} + O(N_0^{-2}) \\ &= \frac{1}{(\sum_{g<g'} N_g N_{g'})^2}\left\{\sum_{g<g'}(N_g-1)^2(N_{g'}-1)^2\left(\frac{\xi_{10}^{(4)}}{N_g} + \frac{\xi_{01}^{(4)}}{N_{g'}}\right)\right\} + O(N_0^{-2}) \end{aligned} \tag{5.13}$$

Now

$$BSS = \sum_{g=1}^{G}\frac{N_g(N_g-1)}{2}(\bar{D}_\cdot^g - \bar{D}_\cdot)^2 = \frac{1}{2}\mathbf{D}_1'\mathbf{D}_1$$

where $\mathbf{D}_1$ is the $G\times 1$ vector

$$\mathbf{D}_1 = \left(\sqrt{N_1(N_1-1)}(\bar{D}_\cdot^1 - \bar{D}_\cdot) \ \ldots \ \sqrt{N_G(N_G-1)}(\bar{D}_\cdot^G - \bar{D}_\cdot)\right)'$$

Note that

$$\mathrm{E}(\bar{D}_\cdot^g) = \frac{1}{\binom{N_g}{2}}\sum_{1\le i<j\le N_g}\mathrm{E}(D_{ij}^g) = \frac{1}{K}\sum_{k=1}^{K}\theta_k^g = \bar{\theta}_\cdot^g$$

$$\mathrm{E}(\bar{D}_\cdot^{(g,g')}) = \frac{1}{K}\sum_{k=1}^{K}\theta_k^{(g,g')} = \bar{\theta}_\cdot^{(g,g')}$$

and

$$\mathrm{E}(\bar{D}_\cdot) = \binom{M}{2}^{-1} \left[\sum_{g=1}^{G} \frac{N_g(N_g-1)}{2} \bar{\theta}_\cdot^g + \sum_{g<g'} N_g N_{g'} \bar{\theta}_\cdot^{(g,g')} \right]$$

where $M = \sum_{g=1}^{G} N_g$.

Therefore,

$$\nu_1 \equiv \mathrm{E}(\bar{D}_\cdot^g - \bar{D}_\cdot) = \bar{\theta}_\cdot^g - \binom{M}{2}^{-1} \left[\sum_{g=1}^{G} \frac{N_g(N_g-1)}{2} \bar{\theta}_\cdot^g + \sum_{g<g'} N_g N_{g'} \bar{\theta}_\cdot^{(g,g')} \right] \tag{5.14}$$

Since $\bar{D}_\cdot^g$ is a U-statistic of degree 2,

$$\mathrm{Var}(\bar{D}_\cdot^g) = \frac{4}{N_g} \xi_1^{(12)} + O(N_g^{-2})$$

where $\xi_1^{(12)} \equiv \mathrm{E}[\psi_{(12)1}^2(\mathbf{X}_i^g)]$, and since $\bar{D}_\cdot^{(g,g')}$ is a two-sample U-statistic of degree $(1,1)$,

$$\mathrm{Var}(\bar{D}_\cdot^{(g,g')}) = \frac{1}{N_g} \xi_{10}^{(13)} + \frac{1}{N_{g'}} \xi_{01}^{(13)} + O(N_0^{-2}) \tag{5.15}$$

where $\xi_{10}^{(13)} \equiv \mathrm{E}[\psi_{(13)10}^2(\mathbf{X}_i^g)]$ and $\xi_{01}^{(13)} \equiv \mathrm{E}[\psi_{(13)01}^2(\mathbf{X}_j^{g'})]$. Under H_0,

$$\begin{aligned} \psi_{(12)1}^2(\mathbf{x}_i) &= \frac{1}{K^2} \sum_{k=1}^{K} \mathrm{P}^2(X_{jk} \neq x_{ik}) + \left(\bar{\theta}_\cdot\right)^2 - \frac{2}{K} \bar{\theta}_\cdot \sum_{k=1}^{K} \mathrm{P}(X_{jk} \neq x_{ik}) \\ &+ \frac{1}{K^2} \sum_{k_1 \neq k_2} \mathrm{P}(X_{jk_1} \neq x_{ik_1}; X_{jk_2} \neq x_{ik_2}) \end{aligned}$$

We are assuming that under H_0 there is homogeneity across or within groups, i.e., $\theta_k^1 = \theta_k^2 = \cdots = \theta_k^G = \theta_k$ and $\theta_k^{(g,g')} = \theta_k^g = \theta_k$. Therefore, under H_0,

$$\sqrt{N_g} \left(\bar{D}_\cdot^g - \bar{\theta}_\cdot \right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, 4\xi_1^{(12)}) \tag{5.16}$$

and

$$\gamma_{13}^{-1} \left(\bar{D}_\cdot^{(g,g')} - \bar{\theta}_\cdot \right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \tag{5.17}$$

where $\gamma_{13}^2 = \frac{1}{N_g} \xi_{10}^{(13)} + \frac{1}{N_{g'}} \xi_{01}^{(13)} = \left(\frac{1}{N_g} + \frac{1}{N_{g'}} \right) \xi_1^{(12)}$ by (5.15) and under H_0.

If $\bar{D}_\cdot$ is a linear combination of normal variables, then $\bar{D}_\cdot$ also follows a normal distribution.

$$\bar{D}_\cdot = \binom{M}{2}^{-1} \left[\sum_{g=1}^{G} \frac{N_g(N_g-1)}{2} \bar{D}_\cdot^g + \sum_{1 \leq g < g' \leq G} N_g N_{g'} \bar{D}_\cdot^{(g,g')} \right]$$

where $M = \sum_{g=1}^{G} N_g$. Under H_0,

$$\eta_1 \equiv \mathrm{E}_0(\bar{D}_\cdot) = \binom{M}{2}^{-1} \left[\sum_{g=1}^{G} \binom{N_g}{2} + \sum_{1 \leq g < g' \leq G} N_g N_{g'} \right] \bar{\theta}_\cdot = \bar{\theta}_\cdot$$

$$
\begin{aligned}
\sigma_1^2 &\equiv \mathrm{Var}_0(\bar{D}_\cdot) \\
&= \frac{4}{M^2(M-1)^2}\left\{\sum_{g=1}^{G} N_g(N_g-1)^2\xi_1^{(12)} + \sum_{1\le g<g'\le G} N_g^2 N_{g'}^2\left(\frac{1}{N_g}\xi_{10}^{(13)} + \frac{1}{N_{g'}}\xi_{01}^{(13)}\right)\right. \\
&+ \left. 2\sum_{\substack{1\le g,g',g^\dagger\le G \\ g\ne g'\ne g^\dagger}} N_g N_{g'} N_{g^\dagger}\xi_{10}^{(13,1;13,2)} + 2\sum_{g=1}^{G}\sum_{g'=g+1}^{G} N_g(N_g-1)N_{g'}\xi_1^{(12,13)}\right\}
\end{aligned}
$$

where $\xi_{10}^{(13,1;13,2)} = \mathrm{E}\{\psi_{(13,1)10}(\mathbf{X}_i^g)\psi_{(13,2)10}(\mathbf{X}_i^g)\}$ and
$\psi_{(13,2)10}(\mathbf{x}_i^g) = \mathrm{E}[\phi_{13,2}(\mathbf{x}_i^g, \mathbf{X}_j^{g^\dagger}) - \bar{\theta}^{(g,g^\dagger)}]$. Under H_0,
$\psi_{(12)1}(\mathbf{X}_i) = \psi_{(13)10}(\mathbf{X}_i) = \psi_{(13)01}(\mathbf{X}_j) = \psi_{(13,1)10}(\mathbf{X}_i) = \psi_{(13,2)10}(\mathbf{X}_i)$. Therefore,
$\xi_1^{(12)} = \xi_{10}^{(13)} = \xi_{01}^{(13)} = \xi_{10}^{(13,1;13,2)} = \xi_1^{(12,13)}$ and

$$
\begin{aligned}
\sigma_1^2 &= \frac{4\xi_1^{(12)}}{M^2(M-1)^2}\left\{\sum_{g=1}^{G} N_g(N_g-1)^2 + \sum_{1\le g<g'\le G} N_g N_{g'}\left(N_g + N_{g'}\right)\right. \\
&+ \left. 2\sum_{\substack{1\le g,g',g^\dagger\le G \\ g\ne g'\ne g^\dagger}} N_g N_{g'} N_{g^\dagger} + 2\sum_{g=1}^{G}\sum_{g'=g+1}^{G} N_g(N_g-1)N_{g'}\right\} \qquad (5.18)
\end{aligned}
$$

Hence, under H_0,

$$
\sigma_1^{-1}\left(\bar{D}_\cdot - \bar{\theta}_\cdot\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)
$$

Now, under H_0,

$$
\nu_1 = \mathrm{E}_0(\bar{D}_\cdot^g - \bar{D}_\cdot) = \bar{\theta}_\cdot - \bar{\theta}_\cdot = 0 \qquad (5.19)
$$

and

$$
\begin{aligned}
\tau_{1g}^2 &\equiv \mathrm{Var}_0(\bar{D}_\cdot^g - \bar{D}_\cdot) \\
&= \mathrm{Var}_0(\bar{D}_\cdot^g) + \mathrm{Var}(\bar{D}_\cdot) - 2\mathrm{Cov}_0(\bar{D}_\cdot^g, \bar{D}_\cdot) \\
&= \left[\frac{4}{N_g} - \frac{8(N_g-1)}{M(M-1)}\right]\xi_1^{(12)} + \sigma_1^2 - \frac{8(M-N_g)}{M(M-1)}\xi_1^{(12,13)} \qquad (5.20)
\end{aligned}
$$

where $\xi_1^{(12,13)} \equiv \mathrm{E}\{\psi_{(12)1}(\mathbf{X}_i^g)\psi_{(13)10}(\mathbf{X}_i^g)\} = \xi_1^{(12)}$, since $\psi_{(12)1}(\mathbf{X}_i) = \psi_{(13)10}(\mathbf{X}_i)$ under H_0.

Then,

$$
\begin{aligned}
\tau_{1g}^2 &= 4\xi_1^{(12)}\left\{\frac{1}{N_g} + \frac{\sum_{g=1}^{G} N_g(N_g-1)^2}{M^2(M-1)^2} + \sum_{g<g'}\frac{N_g N_{g'}(N_{g'}+N_g)}{M^2(M-1)^2} + 2\sum_{g\ne g'\ne g^\dagger}\frac{N_g N_{g'} N_{g^\dagger}}{M^2(M-1)^2}\right. \\
&+ \left. 2\sum_{g=1}^{G}\sum_{g'=g+1}^{G}\frac{N_g(N_g-1)N_{g'}}{M^2(M-1)^2} - 2\frac{(N_g-1)}{M(M-1)} - 2\frac{(M-N_g)}{M(M-1)}\right\} \qquad (5.21)
\end{aligned}
$$

So,

$$
\tau_{1g}^{-1}(\bar{D}_\cdot^g - \bar{D}_\cdot) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)
$$

Since BSS is a quadratic form of normal random variables,

$$BSS = \frac{1}{2}\mathbf{D}_1'\mathbf{D}_1 \sim \frac{1}{2}\sum_{g=1}^{G}\lambda_g\left(\chi_1^2\right)_g$$

which is a linear combination of χ_1^2 random variables, where λ_g's are the characteristic roots of $\mathrm{Var}(\mathbf{D}_1) = \mathbf{\Sigma}_1$. Note that the diagonal elements of $\mathbf{\Sigma}_1$ are $N_g(N_g-1)\tau_{1g}^2$ and the off-diagonal elements, under H_0, are

$$\begin{aligned}\sqrt{N_g(N_g-1)N_{g'}(N_{g'}-1)}\tau_{1gg'} &= \sqrt{N_g(N_g-1)N_{g'}(N_{g'}-1)}\mathrm{Cov}_0(\bar{D}_{\cdot}^{g}-\bar{D}_{\cdot},\bar{D}_{\cdot}^{g'}-\bar{D}_{\cdot})\\ &= \sqrt{N_g(N_g-1)N_{g'}(N_{g'}-1)}\left(-\frac{4\xi_1^{(12)}}{M(M-1)}\left[N_g-1+M-N_g+N_{g'}-1+M-N_{g'}\right]+\sigma_1^2\right)\\ &= \sqrt{N_g(N_g-1)N_{g'}(N_{g'}-1)}\left(-\frac{8\xi_1^{(12)}}{M}+\sigma_1^2\right) \qquad (5.22)\end{aligned}$$

where σ_1^2 is given by (5.18).

Now,

$$\mathrm{E}_0(BSS) = \frac{1}{2}\,\mathrm{trace}(\mathbf{\Sigma}_1) = \frac{1}{2}\sum_{g=1}^{G}N_g(N_g-1)\tau_{1g}^2$$

and

$$\mathrm{Var}_0(BSS) = \frac{1}{4}\,\mathrm{trace}(\mathbf{\Sigma}_1)^2$$

Let

$$BMS = \frac{BSS}{\sum_{g=1}^{G}\binom{N_g}{2}} \quad = \quad \frac{\mathbf{D}_1'\mathbf{D}_1}{\sum_{g=1}^{G}N_g(N_g-1)}$$

Then

$$\mathrm{E}_0(BMS) = \frac{\mathrm{trace}(\mathbf{\Sigma}_1)}{\sum_{g=1}^{G}N_g(N_g-1)} = \frac{\sum_{g=1}^{G}N_g(N_g-1)\tau_{1g}^2}{\sum_{g=1}^{G}N_g(N_g-1)}$$

and

$$\mathrm{Var}_0(BMS) = \frac{1}{\left[\sum_{g=1}^{G}\binom{N_g}{2}\right]^2}\mathrm{Var}_0(BSS) = \frac{\mathrm{trace}(\mathbf{\Sigma}_1)^2}{[\sum_{g=1}^{G}N_g(N_g-1)]^2}$$

For $ABSS$ we have,

$$ABSS = \sum_{1\le g<g'\le G}\sum_{i=1}^{N_g}\sum_{j=1}^{N_{g'}}(\bar{D}_{\cdot}^{(g,g')}-\bar{D}_{\cdot})^2 = \mathbf{D}_2'\mathbf{D}_2$$

where $\mathbf{D}_2 = [\sqrt{N_1N_2}(\bar{D}_{\cdot}^{(1,2)}-\bar{D}_{\cdot}), \sqrt{N_1N_3}(\bar{D}_{\cdot}^{(1,3)}-\bar{D}_{\cdot}), \ldots, \sqrt{N_{G-1}N_G}(\bar{D}_{\cdot}^{(G-1,G)}-\bar{D}_{\cdot})]'$ is a $\frac{G(G-1)}{2}\times 1$ vector.

Let

$$\nu_2 \equiv \mathrm{E}(\bar{D}_{\cdot}^{(g,g')}-\bar{D}_{\cdot}) = \bar{\theta}_{\cdot}^{(g,g')} - \binom{M}{2}^{-1}\left[\sum_{g=1}^{G}\frac{N_g(N_g-1)}{2}\bar{\theta}_{\cdot}^{g} + \sum_{g<g'}N_gN_{g'}\bar{\theta}_{\cdot}^{(g,g')}\right]$$

Under H_0,

$$\nu_2 = \mathrm{E}_0(\bar{D}_{\cdot}^{(g,g')} - \bar{D}_{\cdot}) = \bar{\theta}_{\cdot} - \bar{\theta}_{\cdot} = 0 \tag{5.23}$$

and

$$\begin{aligned}
\tau^2_{2(g,g')} &\equiv \mathrm{Var}(\bar{D}_{\cdot}^{(g,g')} - \bar{D}_{\cdot}) \\
&= \mathrm{Var}(\bar{D}_{\cdot}^{(g,g')}) + \mathrm{Var}(\bar{D}_{\cdot}) - 2\mathrm{Cov}(\bar{D}_{\cdot}^{(g,g')}, \bar{D}_{\cdot}) \\
&= \frac{1}{N_g}\xi_{10}^{(13)} + \frac{1}{N_{g'}}\xi_{01}^{(13)} + \sigma_1^2 - \frac{4}{M(M-1)}\Big[(N_g-1)\xi_1^{(12,13)} + (N_{g'}-1)\xi_1^{(12,13)} + N_{g'}\xi_{01}^{(13)} \\
&+ 2(M-N_g-N_{g'})\xi_{10}^{(13,1;13,2)}\Big]
\end{aligned} \tag{5.24}$$

Note that under H_0 there is homogeneity among groups,

$$\Psi_{(13)10}(\mathbf{x}_i) = \Psi_{(13)01}(\mathbf{x}_j) = \Psi_{(13,1)10}(\mathbf{x}_i) = \Psi_{(13,2)10}(\mathbf{x}_i) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{P}(X_{ik} \neq x_{jk})$$

since the sequences are i.i.d.

Therefore, $\Psi_{(13,1)10}(\mathbf{x}_i)\Psi_{(13,2)10}(\mathbf{x}_i) = \Psi^2_{(13)10}(\mathbf{x}_i)$ and

$$\xi_{10}^{(13,1;13,2)} = \xi_{10}^{(13)} = \xi_{01}^{(13)} = \xi_1^{(12)} = \xi_1^{(12,13)}$$

So, under H_0,

$$\tau^2_{2(g,g')} = \left[\frac{(N_g+N_{g'})}{N_gN_{g'}} - \frac{8}{M}\right]\xi_1^{(12)} + \sigma_1^2 \tag{5.25}$$

As in BSS,

$$ABSS \sim \sum_{i=1}^{G(G-1)/2} \lambda_i \left(\chi^2_1\right)_i$$

where λ_i's are the characteristic roots of $\mathbf{\Sigma}_2 = \mathrm{Var}(\mathbf{D}_2)$. The diagonal elements of $\mathbf{\Sigma}_2$ are $N_gN_{g'}\tau^2_{2(g,g')}$ and, if all groups are different, the off-diagonal elements are

$$\sqrt{N_gN_{g'}N_{g^\dagger}N_{g^\ddagger}}\mathrm{Cov}(\bar{D}_{\cdot}^{(g,g')} - \bar{D}_{\cdot}, \bar{D}_{\cdot}^{(g^\dagger,g^\ddagger)} - \bar{D}_{\cdot}) = \sqrt{N_gN_{g'}N_{g^\dagger}N_{g^\ddagger}}\left(-\frac{8}{M}\xi_1^{(12)} + \sigma_1^2\right) \tag{5.26}$$

and if there is one group in common, i.e., $g = g^\dagger$ or $g' = g^\dagger$,

$$N_g\sqrt{N_{g'}N_{g^\dagger}}\mathrm{Cov}(\bar{D}_{\cdot}^{(g,g')} - \bar{D}_{\cdot}, \bar{D}_{\cdot}^{(g,g^\dagger)} - \bar{D}_{\cdot}) = N_g\sqrt{N_{g'}N_{g^\dagger}}\left(\frac{\xi_1^{(12)}}{N_g} - \frac{8\xi_1^{(12)}}{M} + \sigma_1^2\right) \tag{5.27}$$

Now

$$\mathrm{E}_0(ABSS) = \mathrm{trace}(\mathbf{\Sigma}_2) = \sum_{g<g'} N_gN_{g'}\tau^2_{2(g,g')}$$

$$\mathrm{Var}_0(ABSS) = \mathrm{trace}(\mathbf{\Sigma}_2)^2$$

The corresponding mean-square term is defined as

$$ABMS = \frac{ABSS}{\sum_{g<g'} N_g N_{g'}} = \frac{\mathbf{D}_2'\mathbf{D}_2}{\sum_{g<g'} N_g N_{g'}}$$

Then

$$\mathrm{E}_0(ABMS) = \frac{\mathrm{trace}(\mathbf{\Sigma}_2)}{\sum_{g<g'} N_g N_{g'}} = \frac{\sum_{g<g'} N_g N_{g'} \tau^2_{2(g,g')}}{\sum_{g<g'} N_g N_{g'}}$$

$$\mathrm{Var}_0(ABMS) = \frac{\mathrm{trace}(\mathbf{\Sigma}_2)^2}{[\sum_{g<g'} N_g N_{g'}]^2}$$

6. Test Statistics

One alternative is to compare WMS with $AWMS$. Let $T_1 = \frac{WMS}{AWMS}$. Under H_0,

$$\frac{WMS}{AWMS} = \frac{\dfrac{\sum_{g=1}^{G}(N_g-2)(N_g-3)}{2\sum_{g=1}^{G} N_g(N_g-1)} \left\{ \mu_2 + \dfrac{4}{N_g}\sum_{i=1}^{N_g}(\Psi_{(2)1}(\mathbf{X}_i) - \mu_2) \right\} + O_p(N_0^{-1})}{\dfrac{\sum_{g<g'}(N_g-1)(N_{g'}-1)}{2\sum_{g<g'} N_g N_{g'}} \left\{ \mu_2 + \dfrac{2}{N_g}\sum_{i=1}^{N_g}(\Psi_{(2)1}(\mathbf{X}_i) - \mu_2) + \dfrac{2}{N_{g'}}\sum_{i=1}^{N_{g'}}(\Psi_{(2)1}(\mathbf{X}_i) - \mu_2) \right\} + O_p(N_0^{-1})}$$

Provided that $N_0 = \min_{1\le g\le G}(N_g)$ and $N_g = O(N_0),\ \forall\, g = 1, \ldots G$, we have that $\frac{WMS}{AWMS} \xrightarrow{p} 1$ as $N_0 \to \infty$, i.e, asymptotically the distribution of $\frac{WMS}{AWMS}$ is degenerate.

By (5.21) and (5.22) we have that $\mathbf{\Sigma}_1 = O(N_0)$ and by (5.25), (5.26) and (5.27), $\mathbf{\Sigma}_2 = O(N_0)$).

$$BMS = \frac{BSS}{\sum_{g=1}^{G}\binom{N_g}{2}} \sim \frac{\sum_{g=1}^{G}\lambda_{1g}\left(\chi_1^2\right)_g}{\sum_{g=1}^{G} N_g(N_g-1)}$$

where λ_{1g}'s are the characteristic roots of $\mathrm{Var}(\mathbf{D}_1) = \mathbf{\Sigma}_1$.

$$ABMS = \frac{ABSS}{\sum_{g<g'} N_g N_{g'}} \sim \frac{1}{\sum_{g<g'} N_g N_{g'}} \sum_{i=1}^{G(G-1)/2} \lambda_{2i}\left(\chi_1^2\right)_i$$

where λ_{2i}'s are the characteristic roots of $\mathrm{Var}(\mathbf{D}_2) = \mathbf{\Sigma}_2$.

Also, under H_0, by theoretical results pertaining to U-statistics

$$\sqrt{N_g}(WMS - \mu_2/2) \to \mathrm{N}\left(0, 4\xi_1^{(2)}\right)$$

and

$$\sqrt{N_0}(AWMS - \mu_2/2) \to \mathrm{N}\left(0, \xi_1^{(2)}\right) \quad .$$

Thus,

$$BMS = O_p(N_0^{-1}) \quad \text{and} \quad ABMS = O_p(N_0^{-1})$$

while

$$WMS = O_p(N_g^{-1/2}) \quad \text{and} \quad AWMS = O_p(N_0^{-1/2})$$

Define

$$T_{N,2} \equiv N_0 \left(\frac{BMS}{WMS}\right) \quad \text{and} \quad T_{N,3} \equiv N_0 \left(\frac{ABMS}{AWMS}\right) \quad .$$

Since, BMS and $ABMS$ are the dominating terms in $T_{N,2}$ and $T_{N,3}$, respectively, we can write

$$T_{N,2} = \frac{2N_0(BMS)}{\mu_2} + O_p(N_0^{-1/2})$$

and

$$T_{N,3} = \frac{2N_0(ABMS)}{\mu_2} + O_p(N_0^{-1/2})$$

Therefore,

$$T_{N,2} \sim \frac{2N_0}{\mu_2} \frac{\sum_{g=1}^{G} \lambda_{1g} \left(\chi_1^2\right)_g}{\sum_{g=1}^{G} N_g(N_g - 1)}$$

and

$$T_{N,3} \sim \frac{2N_0}{\mu_2 \sum_{g<g'} N_g N_{g'}} \sum_{i=1}^{G(G-1)/2} \lambda_{2i} \left(\chi_1^2\right)_i$$

Because the elements of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are unknown, the characteristic roots of these matrices are also unknown. Therefore, the above distributions do not have a closed analytic form and we call upon resampling methods, such as the bootstrap, to generate the reference distribution for the test statistic.

7. Power of the Tests

Lemma 1
Let $\mathbf{T}_n$ be a vector of random variables that can be expressed as

$$\mathbf{T}_n = \boldsymbol{\nu} + \frac{1}{\sqrt{n}}\mathbf{U}_n + \mathbf{R}_n$$

where $\mathbf{R}_n = O_p(n^{-1})$.
If $Q(\mathbf{T}) = \mathbf{T}'\mathbf{A}\mathbf{T}$ is a quadratic form on $\mathbf{T}$. Then,

$$\begin{aligned} Q(\mathbf{T}) &= \mathbf{T}'\mathbf{A}\mathbf{T} = \{\boldsymbol{\nu} + \frac{1}{\sqrt{n}}\mathbf{U}_n + \mathbf{R}_n\}'\mathbf{A}\{\boldsymbol{\nu} + \frac{1}{\sqrt{n}}\mathbf{U}_n + \mathbf{R}_n\} \\ &= Q(\boldsymbol{\nu}) + \frac{2}{\sqrt{n}}\boldsymbol{\nu}'\mathbf{A}\mathbf{U}_n + \frac{1}{n}Q(\mathbf{U}_n) + 2\boldsymbol{\nu}'\mathbf{A}\mathbf{R}_n + O_p(n^{-3/2}) \end{aligned}$$

If $\boldsymbol{\nu} = \mathbf{0}$ then $Q(\mathbf{T}) = \frac{1}{n}Q(\mathbf{U}_n) + O_p(n^{-3/2})$. ■

In our case, $\mathbf{T} = \mathbf{D}_1$ and the quadratic form is $Q(\mathbf{D}_1) = \mathbf{D}_1'\mathbf{D}_1$. Note that we can write,

$$\begin{aligned}\mathbf{D}_1'\mathbf{D}_1 &= \sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot)^2 = \sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1)^2 \\ &+ 2\,\nu_1 \sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1) + \sum_{g=1}^{G} N_g(N_g-1)\,\nu_1^2\end{aligned}$$

Let $\mathbf{V}_N = \mathbf{D}_1 - \boldsymbol{\nu}_1$, where $\boldsymbol{\nu}_1$ is a vector $G \times 1$ with elements $\sqrt{N_g(N_g-1)}\nu_1$, $g = 1, \ldots, G$ and ν_1 is given by (5.14). Then, $\mathrm{E}(\mathbf{V}_N) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{V}_N) = \mathrm{Var}(\mathbf{D}_1) = \boldsymbol{\Sigma}_1 = N_0\boldsymbol{\Sigma}_1^\star = O(N_0)$. Therefore,

$$Q(\mathbf{D}_1) = \mathbf{D}_1'\mathbf{D}_1 = \mathbf{V}_N'\mathbf{V}_N + 2\boldsymbol{\nu}_1'\mathbf{V}_N + \boldsymbol{\nu}_1'\boldsymbol{\nu}_1$$

Since $\frac{\mathbf{V}_N}{\sqrt{N_0}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_1^\star)$,

$$\frac{\mathbf{V}_N'\mathbf{V}_N}{N_0} \sim \sum_{g=1}^{G} \lambda_g^\star \left(\chi_1^2\right)_g$$

where $\lambda_g^\star$ are the characteristic roots of $\boldsymbol{\Sigma}_1^\star$. Also,

$$\frac{2\boldsymbol{\nu}_1'\mathbf{V}_N}{\sqrt{N_0}} \sim \mathrm{N}\left(\mathbf{0}, 4\boldsymbol{\nu}_1'\boldsymbol{\Sigma}_1^\star\boldsymbol{\nu}_1\right) \quad \text{or} \quad \frac{2\boldsymbol{\nu}_1'\mathbf{V}_N}{N_0^{3/2}} \sim \mathrm{N}\left(\mathbf{0}, 4\nu_1^2\boldsymbol{\Sigma}_1^\star\right)$$

With the above results, we can see that $\mathbf{V}_N = O_p(N_0^{1/2})$, $\mathbf{V}_N'\mathbf{V}_N = O_p(N_0)$ and $\boldsymbol{\nu}_1'\mathbf{V}_N = O_p(N_0^{3/2})$.

Now,

$$T_{N,2} = \frac{2N_0}{\mu_2 \sum_{g=1}^{G} N_g(N_g-1)}\mathbf{V}_N'\mathbf{V}_N + \frac{4N_0^{3/2}\boldsymbol{\nu}_1'}{\mu_2 \sum_{g=1}^{G} N_g(N_g-1)}\left(\frac{\mathbf{V}_N}{\sqrt{N_0}}\right) + \frac{2N_0}{\mu_2}\nu_1^2 + O_p(N_0^{-1/2})$$

$$\begin{aligned}\left(\frac{T_{N,2} - 2N_0\nu_1^2/\mu_2}{4N_0^{5/2}\nu_1/[\mu_2 \sum_{g=1}^{G} N_g(N_g-1)]}\right) &= \frac{\sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1)^2}{2N_0^{3/2}\nu_1} \\ &+ \frac{1}{N_0^{3/2}} \sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1) + O_p(N_0^{-1})\end{aligned}$$

Note that

$$\frac{\sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1)^2}{2N_0^{3/2}\nu_1} = O_p(N_0^{-1/2}),$$

since $\dfrac{\sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1)^2}{N_0} = \dfrac{\mathbf{V}_N'\mathbf{V}_N}{N_0} = O_p(1)$

and

$$\frac{1}{N_0^{3/2}} \sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1) = O_p(1),$$

since

$$\frac{\nu_1}{N_0^{3/2}}\sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1) = \frac{\boldsymbol{\nu}_1'\mathbf{V}_N}{N_0^{3/2}} = O_p(1)$$

So, for a fixed $\nu_1 \neq 0$, as $N_0 \to \infty$,

$$\left(\frac{T_{N,2} - 2N_0\nu_1^2/\mu_2}{4N_0^{5/2}\nu_1/[\mu_2\sum_{g=1}^{G} N_g(N_g-1)]}\right) = \frac{1}{N_0^{3/2}}\sum_{g=1}^{G} N_g(N_g-1)(\bar{D}_\cdot^g - \bar{D}_\cdot - \nu_1) + O_p(N_0^{-1/2})$$

Thus,

$$\mathrm{P}(T_{N,2} > \nu_1) = \mathrm{P}\left(Z > \sum_{g=1}^{G} N_g(N_g-1)\frac{(\mu_2 - 2N_0\nu_1)}{4N_0^{5/2}}\right) \to 1, \quad \text{as } N_0 \to \infty,$$

i.e., this test is consistent.

Now, consider a local alternative hypothesis. Let $\nu_1 = \frac{1}{\sqrt{N_0}}\gamma_1^\star$, where $\gamma_1^\star$ is a constant. Then,

$$\begin{aligned} T_{N,2} &= \frac{2N_0}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}\mathbf{V}_N'\mathbf{V}_N \\ &+ \frac{4\gamma_1^\star}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}\left[\sqrt{N_0}\sum_{g=1}^{G} N_g(N_g-1)\left(\bar{D}_\cdot^g - \bar{D}_\cdot - \frac{1}{\sqrt{N_0}}\gamma_1^\star\right)\right] \\ &+ \frac{2}{\mu_2}\left(\gamma_1^\star\right)^2 + O_p(N_0^{-1/2}) \end{aligned}$$

$$\begin{aligned} \left(\frac{T_{N,2} - 2\left(\gamma_1^\star\right)^2/\mu_2}{4\gamma_1^\star N_0^2/\left(\mu_2\sum_{g=1}^{G} N_g(N_g-1)\right)}\right) &= \frac{\sum_{g=1}^{G} N_g(N_g-1)\left(\bar{D}_\cdot^g - \bar{D}_\cdot - \frac{1}{\sqrt{N_0}}\gamma_1^\star\right)^2}{2N_0\gamma_1^\star} \\ &+ N_0^{-3/2}\sum_{g=1}^{G} N_g(N_g-1)\left(\bar{D}_\cdot^g - \bar{D}_\cdot - \frac{1}{\sqrt{N_0}}\gamma_1^\star\right) + O_p(N_0^{-1/2}) \end{aligned}$$

Note that

$$\frac{\sum_{g=1}^{G} N_g(N_g-1)\left(\bar{D}_\cdot^g - \bar{D}_\cdot - \frac{1}{\sqrt{N_0}}\gamma_1^\star\right)^2}{2N_0\gamma_1^\star} = O_p(1) \quad \text{and} \quad \frac{\sum_{g=1}^{G} N_g(N_g-1)\left(\bar{D}_\cdot^g - \bar{D}_\cdot - \frac{1}{\sqrt{N}}\gamma_1^\star\right)}{N_0^{3/2}} = O_p(1)$$

Therefore, $T_{N,2}$ no longer follows a Normal distribution as $N_0 \to \infty$. It is a convolution of a linear combination of chi-square random variables and a normal random variable:

$$T_{N,2} = \frac{2N_0}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}\mathbf{V}_N'\mathbf{V}_N + \frac{4N_0}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}\left(\boldsymbol{\gamma}_1^\star\right)'\mathbf{V}_N + \frac{2\left(\gamma_1^\star\right)^2}{\mu_2} + O_p(N_0^{-1/2}),$$

where $\boldsymbol{\gamma}_1^\star = (\sqrt{N_1(N_1-1)}\frac{\gamma_1^\star}{\sqrt{N_0}}, \ldots, \sqrt{N_G(N_G-1)}\frac{\gamma_1^\star}{\sqrt{N_0}})'$.

$$T_{N,2} \sim \frac{2N_0^2}{\mu_2 \sum_{g=1}^{G} N_g(N_g-1)} \sum_{g=1}^{G} \lambda_{1g}^\star \left(\chi_1^2\right)_g + \mathrm{N}\left(\mathbf{0}, \frac{16N_0^3}{\mu_2^2\left[\sum_{g=1}^{G} N_g(N_g-1)\right]^2} \left(\boldsymbol{\gamma}_1^\star\right)' \boldsymbol{\Sigma}_1^\star \boldsymbol{\gamma}_1^\star\right) + \frac{2\left(\gamma_1^\star\right)^2}{\mu_2}$$

where $\lambda_{1g}^\star$ are the characteristic roots of $\boldsymbol{\Sigma}_1^\star$.

Now, let us find out whether $\mathbf{V}_N'\mathbf{V}_N$ and $\left(\boldsymbol{\gamma}_1^\star\right)'\mathbf{V}_N$ are independent. $\mathbf{V}_N'\mathbf{V}_N$ and $\left(\boldsymbol{\gamma}_1^\star\right)'\mathbf{V}_N$ are independent if and only if $\left(\boldsymbol{\gamma}_1^\star\right)'\boldsymbol{\Sigma}_1 = \mathbf{0}$ (Searle, 1971).

Recall that

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} N_1(N_1-1)\tau_{11}^2 & \sqrt{N_1(N_1-1)N_2(N_2-1)}\tau_{112} & \ldots & \sqrt{N_1(N_1-1)N_G(N_G-1)}\tau_{11G} \\ \sqrt{N_1(N_1-1)N_2(N_2-1)}\tau_{112} & N_2(N_2-1)\tau_{12}^2 & \ldots & \sqrt{N_2(N_2-1)N_G(N_G-1)}\tau_{12G} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{N_1(N_1-1)N_G(N_G-1)}\tau_{11G} & \sqrt{N_2(N_2-1)N_G(N_G-1)}\tau_{12G} & \ldots & N_G(N_G-1)\tau_{1G}^2 \end{pmatrix}$$

where

$$\begin{aligned} \tau_{1g}^2 &= 4\xi_1^{(12)} \left\{ \frac{1}{N_g} + \frac{\sum_{g=1}^{G} N_g(N_g-1)^2}{M^2(M-1)^2} + \sum_{g<g'} \frac{N_g N_{g'}(N_{g'}+N_g)}{M^2(M-1)^2} + 2 \sum_{g\neq g'\neq g^\dagger} \frac{N_g N_{g'} N_{g^\dagger}}{M^2(M-1)^2} \right. \\ &+ \left. 2\sum_{g=1}^{G} \sum_{g'=g+1}^{G} \frac{N_g(N_g-1)N_{g'}}{M^2(M-1)^2} - 2\frac{(N_g-1)}{M(M-1)} - 2\frac{(M-N_g)}{M(M-1)} \right\} \end{aligned}$$

and

$$\begin{aligned} \tau_{1gg'} &= \left(\frac{4\xi_1^{(12)}}{M}\right) \left[-2 + \frac{1}{M(M-1)^2} \left(\sum_{g=1}^{G} N_g(N_g-1)^2 + \sum_{1\leq g<g'\leq G} N_g N_{g'}\left(N_g + N_{g'}\right) \right.\right. \\ &+ \left.\left. 2 \sum_{\substack{1\leq g,g',g^\dagger \leq G \\ g\neq g'\neq g^\dagger}} N_g N_{g'} N_{g^\dagger} + 2\sum_{g=1}^{G} \sum_{g'=g+1}^{G} N_g(N_g-1)N_{g'} \right) \right] \end{aligned}$$

Then, the first element of $\left(\boldsymbol{\gamma}_1^\star\right)'\boldsymbol{\Sigma}_1$ is

$$\frac{\gamma_1^\star}{\sqrt{N_0}} \sqrt{N_1(N_1-1)}\left[N_1(N_1-1)\tau_{11}^2 + \sum_{g=2}^{G} N_g(N_g-1)\tau_{11g}\right]$$

and

$$\begin{aligned} &\frac{\gamma_1^\star}{\sqrt{N_0}} \sqrt{N_1(N_1-1)}\left[N_1(N_1-1)\tau_{11}^2 + \sum_{g=2}^{G} N_g(N_g-1)\tau_{11g}\right] = 0 \\ &\quad \Leftrightarrow N_1 = 1 \text{ or all } N_g's = 1 \text{ or } (\tau_{11}^2 = 0 \text{ and all } \tau_{11g}) = 0 \end{aligned}$$

So, $\mathbf{V}_N'\mathbf{V}_N$ and $\left(\boldsymbol{\gamma}_1^\star\right)'\mathbf{V}_N$ are independent if and only if $N_1 = 1$ or all N_g's $= 1$ or ($\tau_{11}^2 = 0$ and all τ_{11g}'s) $= 0$, which is not the case here.

Now, write

$$\left[\mathbf{V}_N'\mathbf{V}_N + 2\left(\boldsymbol{\gamma}_1^\star\right)'\mathbf{V}_N + \left(\boldsymbol{\gamma}_1^\star\right)'\boldsymbol{\gamma}_1^\star\right] = \left[(\mathbf{V}_N + \boldsymbol{\gamma}_1^\star)'(\mathbf{V}_N + \boldsymbol{\gamma}_1^\star)\right]$$

and

$$\begin{aligned} T_{N,2} &= \frac{2N_0}{\mu_2}(BMS) + O_p(N_0^{-1/2}) = \frac{2N_0}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}\mathbf{D}_1'\mathbf{D}_1 + O_p(N_0^{-1/2}) \\ &= \frac{2N_0}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}(\mathbf{V}_N + \boldsymbol{\gamma}_1^\star)'(\mathbf{V}_N + \boldsymbol{\gamma}_1^\star) + O_p(N_0^{-1/2}) \end{aligned}$$

Note that $(\mathbf{V}_N + \boldsymbol{\gamma}_1^\star) \sim \mathrm{N}(\boldsymbol{\gamma}_1^\star, \boldsymbol{\Sigma}_1)$, $\left(\frac{\mathbf{V}_N}{\sqrt{N_0}} + \boldsymbol{\gamma}_1^\star\right) \sim \mathrm{N}(\boldsymbol{\gamma}_1^\star, \boldsymbol{\Sigma}_1^\star)$ and

$$\mathbf{D}_1 \sim \mathrm{N}(\boldsymbol{\nu}_1, \boldsymbol{\Sigma}_1) \quad \text{or} \quad \frac{\mathbf{D}_1}{\sqrt{N_0}} \sim \mathrm{N}(\boldsymbol{\gamma}_1^\star, \boldsymbol{\Sigma}_1^\star).$$

The distribution of $\frac{\mathbf{D}_1'\mathbf{D}_1}{N_0}$ can also be derived the following way.

Let $\mathbf{P}$ be a $G \times G$ orthogonal matrix (i.e., $\mathbf{P}'\mathbf{P} = \mathbf{I}$) such that $\mathbf{P}\boldsymbol{\Sigma}_1^\star\mathbf{P}' = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix, and

$$\mathbf{Y} = \mathbf{P}\frac{\mathbf{D}_1}{\sqrt{N_0}} \Rightarrow \frac{\mathbf{D}_1}{\sqrt{N_0}} = \mathbf{P}'\mathbf{Y}$$

Then,

$$\mathbf{Y} \sim \mathrm{N}(\mathbf{P}\boldsymbol{\gamma}_1^\star, \boldsymbol{\Lambda}) \quad \text{and} \quad \frac{\mathbf{D}_1'\mathbf{D}_1}{N_0} = \mathbf{Y}'\mathbf{P}\mathbf{P}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y},$$

Hence,

$$\frac{\mathbf{D}_1'\mathbf{D}_1}{N_0} = \mathbf{Y}'\mathbf{Y} \sim \sum_{i=1}^{G} \lambda_i \left(\chi_1^2(\delta_i)\right) \tag{7.1}$$

where $\delta_i = \frac{(\nu_{1i}^\star)^2}{\lambda_i}$, λ_i's are the diagonal elements of the diagonal matrix $\boldsymbol{\Lambda}$ and $\nu_{1i}^\star$ is the ith row of the vector $\boldsymbol{\nu}_1^\star = \mathbf{P}\boldsymbol{\gamma}_1^\star$. By (7.1),

$$T_{N,2} = \frac{2N_0}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)}\mathbf{D}_1'\mathbf{D}_1 \sim \frac{2N_0^2}{\mu_2\sum_{g=1}^{G} N_g(N_g-1)} \sum_{i=1}^{G} \lambda_i \left(\chi_1^2(\delta_i)\right)_i$$

Since we have a linear combination of non-central chi-square random variables, when $\nu_1 = \frac{\gamma_1^\star}{\sqrt{N_0}}$,

$$\mathrm{P}(T_{N,2} > \nu_1) \to 1 \text{ as } N_0 \to \infty$$

As the distribution of $T_{N,3}$ is similar to the distribution of $T_{N,2}$, the above results about consistency and power of the test apply to $T_{N,3}$.

8. Data analysis

The data set consists of three groups of HIV infected individuals. The interest is to compare the *env* gene V3 loops from B clade macrophage-tropic, B clade t-cell adapted and clade C sequences. There is a hypothesis which says that clade C is like B clade macrophage-tropic sequences. For this data set there are 356 independent sequences from B clade macrophage-tropic, 140 B clade t-cell adapted sequences and 151 clade C sequences. The sequences are all at the amino acid level with 35 positions long and they can be downloaded from the Los Alamos repository at the address http://hiv-web.lanl.gov

Since the elements of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are unknown, the characteristic roots of these matrices are also unknown and the distributions of the test statistics do not have a closed analytic form. In view of this, we call upon resampling techniques, such as the bootstrap. Here is a summary of the procedure:

1. Compute the statistics T_{N2} and T_{N3} from the data set.
2. Sample 356 sequences for the B clade macrophage-tropic group, 140 sequences for B clade t-cell adapted group and 151 for clade C group with replacement from the pooled sample, i.e., the combined groups.
3. Recompute the test statistics T_{N2} and T_{N3} from this sample and store it.
4. Repeat steps 2 and 3 R times (R should be at least 1,000).

The p-values for the tests are then $\dfrac{\#T'_{N2}s \geq T_{N2}obs}{R}$ and $\dfrac{\#T'_{N3}s \geq T_{N3}obs}{R}$.

The results are

$$T_{N2}obs = 5.235 \qquad T_{N3}obs = 0.4173$$

For $R = 10,000$, the percentiles of the bootstrap distribution are given in Table 1.

Table 1: Percentiles of the Bootstrap Distribution

Statistic	1%	5%	95%	99%
T_{N2}	1.275	1.290	1.544	1.574
T_{N3}	0.399	0.404	0.457	0.459

When comparing the three groups, the observed p-value for T_{N2} is less than 1/10001 and for T_{N3} is 0.7586. Therefore, we can say that relative to the within-clade variation, there is significant variability between the three clades, but relative to the across-within-clade, there is no significant variability across-between the three clades.

Since we found a significant variability among the three clades, we now need to compare the variability of the groups two by two. Using the same procedure described above, we get:

For the comparison of B clade macrophage-tropic group and B clade t-cell adapted group,

$$T_{N2}obs = 5.7975 \qquad T_{N3}obs = 2.9928$$

For $R = 20,000$, the percentiles of the bootstrap distribution are given in Table 2.

Table 2: Percentiles of the Bootstrap Distribution

Statistic	1%	5%	95%	99%
T_{N2}	1.5524	1.5954	1.7682	1.8006
T_{N3}	0.7929	0.8028	0.8463	0.8546

When comparing these two groups, the observed p-value for T_{N2} and T_{N3} are less than 1/20001. Therefore, we can say that relative to the within-clade variation, there is significant variability between the B clade macrophage-tropic group and the B clade t-cell adapted group and relative to the across-within-clade, there is significant variability across-between the two clades.

For the comparison of B clade macrophage-tropic group and clade C group,

$$T_{N2}obs = 4.6724 \qquad T_{N3}obs = 2.3057$$

For $R = 20,000$, the percentiles of the bootstrap distribution are given in Table 3.

Table 3: Percentiles of the Bootstrap Distribution

Statistic	1%	5%	95%	99%
T_{N2}	1.4181	1.4331	1.4989	1.5130
T_{N3}	0.8964	0.9075	0.9579	0.9682

When comparing these two groups, the observed p-value for T_{N2} and for T_{N3} are less than 1/20001. Therefore, we can say that relative to the within-clade variation, there is significant variability between the B clade macrophage-tropic group and the clade C group. Also, relative to the across-within-clade, there is significant variability across-between the two clades.

Comparing B clade t-cell adapted group and clade C group,

$$T_{N2}obs = 5.0621 \qquad T_{N3}obs = 1.1312$$

For $R = 20,000$, the percentiles of the bootstrap distribution are given in Table 4.

Table 4: Percentiles of the Bootstrap Distribution

Statistic	1%	5%	95%	99%
T_{N2}	3.2125	3.4611	4.9459	5.3518
T_{N3}	2.0066	2.1183	2.6653	2.7810

When comparing these two groups, the observed p-value for T_{N2} is 0.0335 and for T_{N3} is 1. Therefore, we can say that relative to the within-clade variation, there is significant variability between the B clade t-cell adapted group and the clade C group, but relative to the across-within-clade, there is no significant variability across-between the two clades.

References

[Anderson and Landis, 1980] Anderson, R. J. and Landis, J. R. (1980). CATANOVA for multidimensional contingency tables: Nominal-scale response. *Communications in Statistics - Theory and Methods*, A9(11):1191–1206.

[Anderson and Landis, 1982] Anderson, R. J. and Landis, J. R. (1982). CATANOVA for multidimensional contingency tables: Ordinal-scale response. *Communications in Statistics - Theory and Methods*, 11(3):257–270.

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

[Lee, 1990] Lee, A. J. (1990). *U-Statisitcs - Theory and Practice*. Marcel Dekker, Inc.

[Light and Margolin, 1971] Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66:534–544.

[Light and Margolin, 1974] Light, R. J. and Margolin, B. H. (1974). An analysis of variance for categorical data II: Small sample comparisons with chi square and other competitors. *Journal of the American Statistical Association*, 69:755–764.

[Mises, 1947] Mises, R. V. (1947). On the asymptotic distribution of the differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348.

[Pinheiro, 1997] Pinheiro, H. P. (1997). *Modelling Variability in the HIV Genome*. PhD thesis, University of North Carolina. Mimeo Series No. 2186T.

[Pinheiro et al., 1999] Pinheiro, H. P., Seillier-Moiseiwitsch, F., and Sen, P. K. (1999). Multivariate CATANOVA and applications to DNA sequences in categorical data. Research Report 12/99, Universidade Estadual de Campinas.

[Puri and Sen, 1971] Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[Searle, 1971] Searle, S. R. (1971). *Linear Models*. John Wiley & Sons.

[Searle, 1982] Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley & Sons.

[Seillier-Moiseiwitsch et al., 1994] Seillier-Moiseiwitsch, F., Margolin, B. H., and Swanstrom, R. (1994). Genetic variability of human immunodeficiency virus: Statistical and biological issues. *Annual Review of Genetics*, 28:559–596.

[Sen, 1967] Sen, P. K. (1967). On some multisample permutation tests based on a class of u-statistics. *American Statistical Association Journal*, pages 1201–1213.

[Sen, 1981] Sen, P. K. (1981). *Invariance Principles and Statistical Inference*. John Wiley & Sons.

[Sen, 1995] Sen, P. K. (1995). Paired comparisons for multiple characteristics: An anocova approach. In Nagaraja, H. N., Sen, P. K., and Morrison, D. F., editors, *Statistical Theory and Practice: Papers in Honor of H. A. David*, pages 247–264. Springer-Verlag, New York.

[Sen and Singer, 1993] Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall.

[Weir, 1990] Weir, B. S. (1990). *Genetic Data Analysis*. Sinauer Associates.