Analysis of Variance based on the Hamming Distance

By Hildete Prisco Pinheiro *
Department of Statistics
State University of Campinas
Françoise Seillier-Moiseiwitsch
Department of Biostatistics
University of North Carolina at Chapel Hill
Pranab Kumar Sen
Department of Biostatistics
University of North Carolina at Chapel Hill

Abstract

The interest here is the comparison of sequences within or between groups. Sequences are considered on an individual basis, i.e., all possible pairwise comparisons within and across groups are performed. We develop a categorical analysis-of-variance framework based on Hamming distances, the proportion of positions at which two aligned sequences differ, and estimate the variability between, within and across groups. We assume that the sequences are independent, but the positions may not be. In this context U-statistics are utilized to represent the average distance between and within groups as well as the overall distance. The total sum of squares is decomposed into within-, between- and across-group sums of squares. The latter term is new: it does not appear in the classical set-up. Generalized-U-statistics theory (Puri & Sen, 1971; Lee, 1990; Sen & Singer, 1993) is used to find the asymptotic distributions of each sum of squares. Test statistics are developed to assess homogeneity among groups.

1. Introduction

The focus here lies in the comparison of sequences. The sequences are considered on an individual basis in the sence that they are compared to each other: all possible pairwise comparisons within and across groups are performed.

*This research was funded in part by CAPES, FAPESP (Brazilian Institutions), the National Science Foundation, the American Foundation for AIDS Research and the National Institutes of Health.
Key words and phrases: Analysis of Variance, Categorical Data, Hamming Distance, U-statistics

We develop an analysis-of-variance framework based on Hamming distances and estimate the variability between, within and across groups (Section 2). In the within sum of squares, we are estimating the variability among individuals within a group around the average distance within this group. In the across sum of squares, we are estimating the variability of individuals across two groups with respect to the average distance between those groups. In the between sum of squares, we estimate the variability in the group average distances around the overall distance.

Weir (1990a) describes an analysis of variance for the genetic variation in the population, in particular for the amount of observed *heterozygosity*. The variance of the estimate of the average heterozygosity is broken down to show the contribution of populations, loci and individuals by setting out the calculations in a framework similar to that of an analysis of variance. Our situation is a little different because we would like to construct a categorical analysis of variance based on Hamming distances (Seillier-Moiseiwitsch et al., 1994 and references therein), assuming that the sequences are independent, but the positions may not be. The Hamming distance is the proportion of positions at which two aligned sequences differ.

In this context U-statistics are utilized to represent the average distance between and within groups as well as the overall distance (Sections 3 and 4). The total sum of squares is decomposed into within-, between- and across-group sums of squares. The latter term is new: it does not appear in the classical set-up. Generalized-U-statistics theory (Puri & Sen, 1971; Lee, 1990; Sen & Singer, 1993) is used to find the asymptotic distributions of each sum of squares. In Section 5 test statistics are developed to assess homogeneity among groups. The power of the tests are discussed in Section 6. Finally, a data analysis is shown in Section 7.

2. The Total Sum of Squares and its decomposition

Let $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \ldots, X_{ik}^g)'$ be a random vector representing sequence i of group g. Suppose $i = 1, \ldots, N$, $k = 1, \ldots, K$ and $g = 1, \ldots, G$. So, X_{ik}^g represents either the amino acid or the nucleotide present at position k of sequence i in group g (e.g., at the nucleotide level, $x_{ik}^g \in \{\mathbf{A}, \mathbf{C}, \mathbf{T}, \mathbf{G}\}$).

Consider $\mathbf{X}_i^{g_1}$ and $\mathbf{X}_j^{g_2}$.

Definition 1
The *Hamming Distance* $D_{ij}^{(g_1 g_2)}$ is a descriptive statistic for sequence comparison defined by

$$\begin{aligned} D_{ij}^{(g_1,g_2)} &= \frac{1}{K}\sum_{k=1}^{K} I(X_{ik}^{g_1} \neq X_{jk}^{g_2}) \qquad (2.1) \\ &= \frac{1}{K} \times (\textit{number of positions where } X_i^{g_1} \textit{ and } X_j^{g_2} \textit{ differ}), \end{aligned}$$

and when $g_1 = g_2 = g$,

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

■

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,

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

Define the average distance within a group as

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

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

$$\bar{D}_\cdot^{(g_1,g_2)} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij}^{(g_1,g_2)} = \frac{1}{N^2 K}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K} I(X_{ik}^{g_1} \neq X_{jk}^{g_2})$$

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

$$\begin{aligned}
\bar{D}_\cdot &= \left[G\binom{N}{2} + N^2\binom{G}{2}\right]^{-1}\left(\sum_{g=1}^{G}\sum_{1\le i<j\le N} D_{ij}^g + \sum_{1\le g_1<g_2\le G}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij}^{(g_1,g_2)}\right) \\
&= \binom{NG}{2}^{-1}\left(\sum_{g=1}^{G}\binom{N}{2}\bar{D}_\cdot^g + \sum_{1\le g_1<g_2\le G} N^2\bar{D}_\cdot^{(g_1,g_2)}\right)
\end{aligned}$$

which is a linear combination of U-statistics.

The Total Sum of Squares can be decomposed as

$$\begin{aligned}
TSS &= \sum_{g=1}^{G}\sum_{1\le i<j\le N}(D_{ij}^g - \bar{D}_\cdot)^2 + \sum_{1\le g_1<g_2\le G}\sum_{i=1}^{N}\sum_{j=1}^{N}(D_{ij}^{(g_1,g_2)} - \bar{D}_\cdot)^2 \qquad (2.2) \\
&= \sum_{g=1}^{G}\sum_{1\le i<j\le N}(D_{ij}^g - \bar{D}_\cdot^g)^2 + \sum_{g=1}^{G}\sum_{1\le i<j\le N}(\bar{D}_\cdot^g - \bar{D}_\cdot)^2 \\
&+ \sum_{1\le g_1<g_2\le G}\sum_{i=1}^{N}\sum_{j=1}^{N}(D_{ij}^{(g_1,g_2)} - \bar{D}_\cdot^{(g_1,g_2)})^2 + \sum_{1\le g_1<g_2\le G}\sum_{i=1}^{N}\sum_{j=1}^{N}(\bar{D}_\cdot^{(g_1,g_2)} - \bar{D}_\cdot)^2 \\
&= WSS + BSS + AWSS + ABSS
\end{aligned}$$

where WSS stands for Within Sum of Squares, BSS for Between Sum of Squares, $AWSS$ for Across Within Sum of Squares and $ABSS$ for Across Between Sum of Squares

3.Connections Between Sums of Squares and U-statistics Since we have G groups of N sequences, we can disregard the group clustering and think of the sequences as a random sample of size NG. Then

$$\begin{aligned} TSS &= \sum_{1\le i<j\le NG} (D_{ij} - \bar{D}.)^2 \\ &= \left(\frac{NG(NG-1)}{2} - 1\right) \binom{\frac{NG(NG-1)}{2}}{2}^{-1} \sum_{\substack{i<j,\ i'<j' \\ i\le i' \text{ or } j\le j'}} \frac{(D_{ij} - D_{i'j'})^2}{2} \end{aligned} \tag{3.3}$$

$$\begin{aligned} WSS &= \sum_{g=1}^{G} \sum_{1\le i<j\le N} (D^g_{ij} - \bar{D}^g_{\cdot})^2 \\ &= \left(\frac{N(N-1)}{2} - 1\right) \binom{\frac{N(N-1)}{2}}{2}^{-1} \sum_{g=1}^{G} \sum_{\substack{i<j,\ i'<j' \\ i\le i' \text{ or } j\le j'}} \frac{(D^g_{ij} - D^g_{i'j'})^2}{2} \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} AWSS &= \sum_{1\le g_1<g_2\le G} \sum_{i=1}^{N}\sum_{j=1}^{N} (D^{(g_1,g_2)}_{ij} - \bar{D}^{(g_1,g_2)}_{\cdot})^2 \\ &= (N^2-1)\binom{N^2}{2}^{-1} \sum_{1\le g_1<g_2\le G} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{i'=1}^{N} \sum_{\substack{j'=1 \\ i\le i' \text{ or } j\le j'}}^{N} \frac{(D^{(g_1,g_2)}_{ij} - D^{(g_1,g_2)}_{i'j'})^2}{2} \end{aligned} \tag{3.5}$$

The above 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 $AWSS$ is a linear combination of two-sample U-statistics of degrees (2,2) and (2,1).

4. Asymptotic Distributions and decompositions of U-statistics

Let U_n be a U-statistic of degree m with kernel $\phi(X_1, \ldots, X_m)$ and $E(U_n) = \theta(F) = \theta$.

$$U_n = U(X_1, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{1\le i_1<\cdots<i_m\le n} \phi(X_{i_1}, \ldots, X_{i_m}), \quad n \ge m \tag{4.6}$$

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

$$\Psi_c(x_1, \ldots, x_c) \equiv \mathrm{E}\{\phi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)\} \tag{4.7}$$

$$\psi_c(x_1, \ldots, x_c) \equiv \mathrm{E}\{\phi(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m) - \theta\}, \tag{4.8}$$

$$\xi_c \equiv \mathrm{E}\{\psi_c^2(X_1, \ldots, X_c)\} = \mathrm{E}\{\Psi_c^2(X_1, \ldots, X_c)\} - \theta^2 \quad \text{and} \quad \xi_0 \equiv 0. \tag{4.9}$$

Theorem 1
The function Ψ_c defined in (4.7) has the properties
(i) $\Psi_c(x_1, \ldots, x_c) = \mathrm{E}\{\Psi_d(x_1, \ldots, x_c, X_{c+1}, \ldots, X_d)\}$ for $1 \leq c < d \leq m$,
(ii) $\mathrm{E}\{\Psi_c(x_1, \ldots, x_c)\} = \mathrm{E}\{\phi(X_1, \ldots, X_m)\}$. ■

The proof appears in Lee (1990, p. 11).

By (4.6) and (4.8),

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

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

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

and exactly c equations $i_k = j_h$ are satisfied. By (4.9), each term in $\sum^{(c)}$ is equal to ξ_c. The number of terms in $\sum^{(c)}$ is

$$\frac{n(n-1)\cdots(n-2m+c+1)}{c!(m-c)!(m-c)!} = \binom{m}{c}\binom{n-m}{m-c}\binom{n}{m} \tag{4.10}$$

Since $\xi_0 = 0$,

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\xi_c \tag{4.11}$$

Hoeffding (1948) obtained the inequality: $0 \leq \xi_c \leq \frac{c}{d}\xi_d \quad 1 \leq c < d \leq m$, which leads to

$$\frac{m^2}{n}\xi_1 \leq \mathrm{Var}(U_n) \leq \frac{m}{n}\xi_m$$

Now, from (4.11) and (4.10)

$$\mathrm{Var}(U_n) = \frac{m^2}{n}\left(\frac{n-m}{n-1}\right)\cdots\left(\frac{n-2m+2}{n-m+1}\right)\xi_1 + \cdots + \frac{m!}{n(n-1)\ldots(n-m+1)}\xi_m$$

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

$$\mathrm{Var}(U_n) = \frac{m^2}{n}\xi_1 + O(n^{-2}) \tag{4.12}$$

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

$$n^{1/2}(U_n - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, m^2\xi_1), \quad \text{(Hoeffding, 1948)} \tag{4.13}$$

We may rewrite (4.6) as

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

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

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

$$U_n = \theta(F) + \sum_{h=1}^{m} \binom{m}{h} U_{n,h} \qquad n \ge m \tag{4.14}$$

where

$$U_{n,h} = n^{-[h]} \sum_{1 \le i_1 \ne \cdots \ne i_h \le n} \int_{R^{ph}} \cdots \int \Psi_h(x_1, \ldots, x_h) \prod_{j=1}^{h} d[c(x_j - X_{i_j}) - F(x_j)]$$

for $1 \le h \le m$. Further, if we write

$$\begin{aligned} \Psi_h^\circ(x_1, \ldots, x_h) \quad &= \quad \Psi_h(x_1, \ldots, x_h) - \sum_{j=1}^{h} \Psi_{h-1}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_h) \\ &\quad + \cdots + (-1)^h \theta(F), \quad \forall\ (x_1, \ldots, x_h) \in R^{ph}, \end{aligned} \tag{4.15}$$

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

$$U_{n,h} = \binom{n}{h}^{-1} \sum_{1 \le i_1 < \cdots < i_h \le n} \Psi_h^\circ(X_{i_1}, \ldots, X_{i_h}), \qquad 1 \le h \le m \tag{4.16}$$

and the $U_{n,h}$ are themselves U-statistics. From direct computation, $\mathrm{E}(U_{n,h}) = 0,\ \forall\ 1 \le h \le m$ and

$$\mathrm{Var}(U_{n,h}) = \mathrm{E}(U_{n,h}^2) = O(n^{-h}), \qquad h = 1, 2, \ldots, m; \tag{4.17}$$

and we can write

$$U_n = \theta(F) + \frac{m}{n} \sum_{i=1}^{n} [\Psi_1(X_i) - \theta(F)] + O_p(n^{-1}) \tag{4.18}$$

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 $F^{(j)}(\mathbf{x}),\ \mathbf{x} \in R^p$, 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

$$\theta(\mathbf{F}) = \int_{R^{m_0}} \cdots \int \phi(\mathbf{x}_i^{(j)},\ 1 \le i \le m_j,\ 1 \le j \le c) \prod_{j=1}^{c} \prod_{i=1}^{m_j} dF^{(j)}(\mathbf{x}_i^{(j)}) \tag{4.19}$$

Definition 2

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

$$U(\mathbf{n}) = \prod_{j=1}^{c} \binom{n_j}{m_j}^{-1} \sum_{(\mathbf{n})}^{*} \phi(\mathbf{X}_\alpha^{(j)},\ \alpha = i_{j1}, \ldots, i_{jm_j},\ 1 \leq j \leq c), \tag{4.20}$$

where the summation $\sum_{(\mathbf{n})}^{*}$ extends over all $1 \leq i_{j1} < \ldots < i_{jm_j} \leq n_j,\ 1 \leq j \leq c$. $U(\mathbf{n})$ 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

$$\Psi_{d_1 \ldots d_c}(\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{d_j}^{(j)},\ 1 \leq j \leq c) \equiv \mathrm{E}(\phi(\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_{d_j}^{(j)}, \mathbf{X}_{d_j+1}^{(j)}, \ldots, \mathbf{X}_{m_j}^{(j)},\ 1 \leq j \leq c)) \tag{4.21}$$

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

$$\xi_{\mathbf{d}}(\mathbf{F}) = \mathrm{E}\left(\Psi_{\mathbf{d}}^2(\mathbf{X}_1^{(j)}, \ldots, \mathbf{X}_{d_j}^{(j)},\ 1 \leq j \leq c)\right) - \theta^2(\mathbf{F}), \qquad \mathbf{0} \leq \mathbf{d} \leq \mathbf{m} \tag{4.22}$$

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

$$\mathrm{Var}\,[U(\mathbf{n})] \;=\; \sum_{j=1}^{c} n_j^{-1} \sigma_j^2\, [1 + O(n_0^{-1})] \tag{4.23}$$

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

$$\sigma_j^2 = m_j^2\, \xi_{\delta_{j1}, \ldots, \delta_{jc}}(\mathbf{F}) \quad j = 1, \ldots, c \tag{4.24}$$

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

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

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

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

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

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

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

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

4. Combining the U-statistics We can write

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

where

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

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

are one-sample U-statistics of degree 3 and

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

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

are one-sample U-statistics of degree 4. The expected value of WSS is

$$E(WSS) = \sum_{g=1}^{G}(N-2)\left\{\mu_{g1} + \frac{(N-3)}{4}\mu_{g2}\right\} \ .$$

Under H_0, there is homogeneity among groups, i.e., for any g, $\theta^g_k = \theta_k$ and $\theta^g_{k_1k_2} = \theta_{k_1k_2}$, thus

$$E_0(WSS) = G(N-2)\left\{\mu_1 + \frac{(N-3)}{4}\mu_2\right\}$$

where

$$\mu_1 = \frac{2}{K^2}\left[\sum_{k=1}^{K}\theta_k + \sum_{k_1\neq k_2}\theta_{k_1k_2} - \sum_{k=1}^{K}\theta_k(i,j;i,j') - \sum_{k_1\neq k_2}\theta_{k_1k_2}(i,j;i,j'))\right] \quad (4.28)$$

and

$$\mu_2 = \frac{2}{K^2}\left\{\sum_{k=1}^{K}\theta_k(1-\theta_k) + \sum_{k_1\neq k_2}(\theta_{k_1k_2} - \theta_{k_1}\theta_{k_2})\right\} \tag{4.29}$$

Note that

$$\theta_k = \mathrm{P}(X_{ik}\neq X_{jk}) = \sum_{c=0}^{C-1} p_k(c)[1-p_k(c)] \tag{4.30}$$

$$\begin{aligned}\theta_{k_1k_2} &= \mathrm{P}(X_{ik_1}\neq X_{jk_1}; X_{ik_2}\neq X_{jk_2})\\ &= \sum_{c_1,c_2=0}^{C-1} p_{k_1k_2}(c_1,c_2)\left[\sum_{\substack{c_3=0\\ c_3\neq c_1}}^{C-1}\sum_{\substack{c_4=0\\ c_4\neq c_2}}^{C-1} p_{k_1k_2}(c_3,c_4)\right]\end{aligned} \tag{4.31}$$

Decomposing WSS, under H_0,

$$\begin{aligned}WSS &= G(N-2)\left(\mu_1 + \frac{(N-3)}{4}\mu_2\right)\\ &+ (N-2)\frac{3}{N}G\sum_{i=1}^{N}[\Psi_{(1)1}(\mathbf{X}_i)-\mu_1] + O_p(1)\\ &+ \frac{(N-2)(N-3)}{N}G\sum_{i=1}^{N}[\Psi_{(2)1}(\mathbf{X}_i)-\mu_2] + O_p(N)\end{aligned} \tag{4.32}$$

and the associated mean square expression is

$$\begin{aligned}WMS &\equiv \frac{WSS}{G\binom{N}{2}} = \frac{2WSS}{GN(N-1)}\\ &= \frac{(N-2)(N-3)}{N(N-1)2}\left\{\mu_2 + \frac{4}{N}\sum_{i=1}^{N}[\Psi_{(2)1}(\mathbf{X}_i)-\mu_2]\right\} + O_p(N^{-1})\end{aligned} \tag{4.33}$$

with

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

and

$$\mathrm{Var}_0(WMS) = \frac{4(N-2)^2(N-3)^2}{GN^2(N-1)^2}\frac{\xi_1^{(2)}}{N} + O(N^{-2})$$

For $AWSS$,

$$\begin{aligned}AWSS &= \sum_{1\leq g_1<g_2\leq G}\left[\frac{(N-1)^2}{4}\left(\mathbf{U}_{4,1}^{(2,2)} + \mathbf{U}_{4,2}^{(2,2)}\right)\right.\\ &\left. + \frac{(N-1)}{2}\left(\mathbf{U}_{5,2}^{(2,1)} + \mathbf{U}_{5,1}^{(1,2)}\right)\right]\end{aligned}$$

where

$$\mathbf{U}_{\mathbf{4},\mathbf{1}}^{(\mathbf{2},\mathbf{2})} = \left[\binom{N}{2}\binom{N}{2}\right]^{-1} \sum_{i\neq i'} \sum_{\substack{j\neq j'\\ j\neq i'}} (D_{ij}^{(g_1,g_2)} - D_{i'j'}^{(g_1,g_2)})^2 \quad \text{and}$$

$$\mathbf{U}_{\mathbf{4},\mathbf{2}}^{(\mathbf{2},\mathbf{2})} = \left[\binom{N}{2}\binom{N}{2}\right]^{-1} \sum_{\substack{i\neq j\\ i\neq i'}} \sum_{j\neq j'} (D_{ij}^{(g_1,g_2)} - D_{i'j'}^{(g_1,g_2)})^2$$

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

$$\mathbf{U}_{\mathbf{5},\mathbf{1}}^{(\mathbf{1},\mathbf{2})} = \left[\binom{N}{1}\binom{N}{2}\right]^{-1} \sum_{\substack{i=1\\ i\neq j'}}^{N} \sum_{\substack{1\leq j,\, j'\leq N\\ j\neq j'}} (D_{ij}^{(g_1,g_2)} - D_{ij'}^{(g_1,g_2)})^2 \quad \text{and}$$

$$\mathbf{U}_{\mathbf{5},\mathbf{2}}^{(\mathbf{2},\mathbf{1})} = \left[\binom{N}{2}\binom{N}{1}\right]^{-1} \sum_{\substack{j=1\\ j\neq i'}}^{N} \sum_{i\neq i'} (D_{ij}^{(g_1,g_2)} - D_{i'j}^{(g_1,g_2)})^2$$

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

$$\mathrm{E}(AWSS) = (N-1) \sum_{1\leq g_1<g_2\leq G} \left(\frac{(N-1)}{2}\mu_{(g_1,g_2)4} + \mu_{(g_1,g_2)5} \right)$$

and under H_0

$$\mathrm{E}_0(AWSS) = \frac{G(G-1)(N-1)}{2} \left(\frac{(N-1)}{2}\mu_4 + \mu_5 \right)$$

where $\mu_4 = \mu_2$ is given by (4.29) and $\mu_5 = \mu_1$ is given by (4.28).

$AWSS$ can be decomposed as

$$\begin{aligned} AWSS \;=\; & \sum_{1\leq g_1<g_2\leq G} \Bigg[\frac{(N-1)^2}{2} \Bigg(\mu_{(g_1,g_2)4} + \frac{2}{N}\sum_{i=1}^{N} (\Psi_{(4)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1,g_2)4}) \\ & + \frac{2}{N}\sum_{j=1}^{N} (\Psi_{(4)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1,g_2)4}) + O_p(N^{-1}) \Bigg) \\ & + \frac{(N-1)}{2} \Bigg(2\mu_{(g_1,g_2)5} + \frac{1}{N}\sum_{i=1}^{N} (\Psi_{(5)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1,g_2)5}) \\ & + \frac{2}{N}\sum_{j=1}^{N} (\Psi_{(5)01}(\mathbf{X}_j^{g_1}) - \mu_{(g_1,g_2)5}) + \frac{2}{N}\sum_{i=1}^{N} (\Psi_{(5)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1,g_2)5}) \\ & + \frac{1}{N}\sum_{j=1}^{N} (\Psi_{(5)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1,g_2)5}) + O_p(N^{-1}) \Bigg) \Bigg] \end{aligned} \tag{4.34}$$

The associated mean-square expression is

$$AWMS = \frac{AWSS}{\binom{G}{2}N^2} = \frac{2AWSS}{N^2G(G-1)}$$

$$= \frac{(N-1)^2}{N^2 G(G-1)} \sum_{1 \le g_1 < g_2 \le G} \left[\mu_{(g_1,g_2)4} + \frac{2}{N} \sum_{i=1}^{N} (\Psi_{(4)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1,g_2)4}) \right.$$

$$\left. + \quad \frac{2}{N} \sum_{j=1}^{N} (\Psi_{(4)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1,g_2)4}) \right] + O_p(N^{-1}) \tag{4.35}$$

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

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

$$\mathrm{Var}_0(AWMS) \;=\; \frac{(N-1)^4}{2N^4 G(G-1)} \left(\frac{4}{N} \xi_{10}^{(4)} + \frac{4}{N} \xi_{01}^{(4)} \right) + O(N^{-2})$$

Now

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

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

$$\mathbf{D}_1 = (\bar{D}_{\cdot}^{1} - \bar{D}_{\cdot} \; \ldots \; \bar{D}_{\cdot}^{G} - \bar{D}_{\cdot})'$$

Note that

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

$$\mathrm{E}(\bar{D}_{\cdot}^{(g_1,g_2)}) = \frac{1}{K} \sum_{k=1}^{K} \theta_k^{(g_1,g_2)} = \bar{\theta}_{\cdot}^{(g_1,g_2)}$$

and

$$\mathrm{E}(\bar{D}_{\cdot}) \;=\; \frac{(N-1)}{G(NG-1)} \sum_{g=1}^{G} \bar{\theta}_{\cdot}^{g} + \frac{2N}{G(NG-1)} \sum_{1 \le g_1 < g_2 \le G} \bar{\theta}_{\cdot}^{(g_1,g_2)}$$

Therefore,

$$\nu_1 \equiv \mathrm{E}(\bar{D}_{\cdot}^{g} - \bar{D}_{\cdot}) = \bar{\theta}_{\cdot}^{g} - \frac{(N-1)}{G(NG-1)} \sum_{g=1}^{G} \bar{\theta}_{\cdot}^{g} + \frac{2N}{G(NG-1)} \sum_{1 \le g_1 < g_2 \le G} \bar{\theta}_{\cdot}^{(g_1,g_2)}$$

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

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

where $\xi_1^{(12)} \equiv \mathrm{E}[\psi_{(12)1}^2(\mathbf{X}_i^g)]$, and since $\bar{D}_{\cdot}^{(g_1,g_2)}$ is a two-sample U-statistic of degree $(1,1)$,

$$\mathrm{Var}(\bar{D}_{\cdot}^{(g_1,g_2)}) = \frac{1}{N} \xi_{10}^{(13)} + \frac{1}{N} \xi_{01}^{(13)} + O(N^{-2}) \tag{4.36}$$

where $\xi_{10}^{(13)} \equiv \mathrm{E}[\psi^2_{(13)10}(\mathbf{X}_i^{g_1})]$ and $\xi_{01}^{(13)} \equiv \mathrm{E}[\psi^2_{(13)01}(\mathbf{X}_j^{g_2})]$. Under H_0,

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

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 $\theta_k^{(g_1,g_2)} = \theta_k^g = \theta_k$. Therefore, under H_0,

$$\sqrt{N}\left(\bar{D}^g_. - \bar{\theta}.\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, 4\xi_1^{(12)}) \tag{4.37}$$

and

$$\gamma_{13}^{-1}\left(\bar{D}^{(g_1,g_2)}_. - \bar{\theta}.\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \tag{4.38}$$

where $\gamma_{13}^2 = \frac{1}{N}\xi_{10}^{(13)} + \frac{1}{N}\xi_{01}^{(13)} = \frac{2}{N}\xi_1^{(12)}$ by (4.36).

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

$$\bar{D}. = \frac{(N-1)}{G(NG-1)}\sum_{g=1}^{G}\bar{D}^g_. + \frac{2N}{G(NG-1)}\sum_{1\leq g_1<g_2\leq G}\bar{D}^{(g_1,g_2)}_.$$

Under H_0,

$$\eta_1 \equiv \mathrm{E}_0(\bar{D}.) = \frac{(N-1)\bar{\theta}. + N(G-1)\bar{\theta}.}{(NG-1)} = \bar{\theta}.$$

$$\begin{aligned}\sigma_1^2 &\equiv \mathrm{Var}_0(\bar{D}.) \\ &+ \frac{4N^2}{G^2(NG-1)^2} \\ &= \frac{(N-1)^2}{G(NG-1)^2}\frac{4}{N}\xi_1^{(12)} + \frac{2N^2(G-1)}{G(NG-1)^2}\left[\left(\frac{1}{N}(\xi_{10}^{(13)} + \xi_{01}^{(13)})\right)\right. \\ &+ \left. 2(G-2)\frac{1}{N}\xi_{10}^{(13,1;13,2)}\right] + \frac{2N(N-1)}{G^2(NG-1)^2}G(G-1)\frac{2}{N}\xi_1^{(12,13)}\end{aligned}$$

where $\xi_{10}^{(13,1;13,2)} = \mathrm{E}\{\psi_{(13,1)10}(\mathbf{X}_i^{g_1})\psi_{(13,2)10}(\mathbf{X}_i^{g_1})\}$ and $\psi_{(13,2)10}(\mathbf{x}_i^{g_1}) = \mathrm{E}[\phi_{13,2}(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_3}) - \bar{\theta}^{(g_1,g_3)}]$. 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, $\xi_1^{(12)} = \xi_{10}^{(13)} = \xi_{01}^{(31)} = \xi_{10}^{(13,1;13,2)} = \xi_1^{(12,13)}$ and

$$\sigma_1^2 = [(N-1)^2 + N(G-1)(NG-1)]\frac{4\xi_1^{(12)}}{NG(NG-1)^2} \tag{4.39}$$

Hence, under H_0,

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

Now

$$\nu_1 = \mathrm{E}_0(\bar{D}^g_{\cdot} - \bar{D}_{\cdot}) = \bar{\theta}_{\cdot} - \bar{\theta}_{\cdot} = 0 \tag{4.40}$$

and

$$\begin{aligned}\tau_1^2 &\equiv \mathrm{Var}_0(\bar{D}^g_{\cdot} - \bar{D}_{\cdot}) \\ &= \left[1 - 2\frac{(N-1)}{G(NG-1)}\right]\frac{4}{N}\xi_1^{(12)} + \sigma_1^2 - \frac{4N(G-1)}{G(NG-1)}\frac{2}{N}\xi_1^{(12,13)}\end{aligned} \tag{4.41}$$

where $\xi_1^{(12,13)} \equiv \mathrm{E}\{\psi_{(12)1}(\mathbf{X}_i^{g_1})\psi_{(13)10}(\mathbf{X}_i^{g_1})\} = \xi_1^{(12)}$, since $\psi_{(12)1}(\mathbf{X}_i) = \psi_{(13)10}(\mathbf{X}_i)$ under H_0.

Then,

$$\tau_1^2 = \{(N-1)^2 + (NG-1)[N(G-1) + (NG-1)(G-2)]\}\frac{4\xi_1^{(12)}}{NG(NG-1)^2} \tag{4.42}$$

So,

$$\tau_1^{-1}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot}) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$$

Since BSS is a quadratic form of normal random variables,

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

which is a linear combination of χ_1^2 random variables, where λ_g's are the characteristic roots of $\mathrm{Var}(\mathbf{D}_1) = \boldsymbol{\Sigma}_1$. Note that the diagonal elements of $\boldsymbol{\Sigma}_1$ are τ_1^2 and the off-diagonal elements, under H_0, are

$$\begin{aligned}&\mathrm{Cov}_0(\bar{D}^{g_1}_{\cdot} - \bar{D}_{\cdot}, \bar{D}^{g_2}_{\cdot} - \bar{D}_{\cdot}) \\ &\quad= \left[\frac{(N-1)^2 - (NG-1)(NG+N-2)}{(NG-1)}\right]\frac{4\xi_1^{(12)}}{NG(NG-1)} < 0\end{aligned}$$

since $(NG-1)(NG+N-2) > (N-1)^2$.

Now,

$$\mathrm{E}_0(BSS) = \frac{N(N-1)}{2}\,\mathrm{trace}(\boldsymbol{\Sigma}_1) = \frac{N(N-1)}{2}\,G\tau_1^2$$

and

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

Let

$$BMS = \frac{BSS}{G\binom{N}{2}} = \frac{1}{G}\,\mathbf{D}_1'\mathbf{D}_1$$

Then

$$\mathrm{E}_0(BMS) = \frac{1}{G}\,\mathrm{E}_0(BSS) = \tau_1^2$$

and

$$\mathrm{Var}_0(BMS) = \frac{1}{G^2}\mathrm{Var}_0(BSS) = \frac{1}{G^2}\,\mathrm{trace}(\boldsymbol{\Sigma}_1)^2$$

For $ABSS$ we have,

$$ABSS = \sum_{1\le g_1<g_2\le G}\sum_{i=1}^{N}\sum_{j=1}^{N}(\bar{D}_{\cdot}^{(g_1,g_2)} - \bar{D}_{\cdot})^2 = N^2\mathbf{D}_2\mathbf{D}_2$$

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

Let

$$\nu_2 \equiv \mathrm{E}(\bar{D}_{\cdot}^{(g_1,g_2)} - \bar{D}_{\cdot}) = \bar{\theta}_{\cdot}^{(g_1,g_2)} - \frac{(N-1)}{G(NG-1)}\sum_{g=1}^{G}\bar{\theta}_{\cdot}^{g} - \frac{2N}{G(NG-1)}\sum_{1\le g_1<g_2\le G}\bar{\theta}_{\cdot}^{(g_1,g_2)}$$

Under H_0,

$$\nu_2 = \mathrm{E}_0(\bar{D}_{\cdot}^{(g_1,g_2)} - \bar{D}_{\cdot}) = \bar{\theta}_{\cdot} - \bar{\theta}_{\cdot} = 0 \tag{4.43}$$

and

$$\begin{aligned}
\tau_2^2 &\equiv \mathrm{Var}(\bar{D}_{\cdot}^{(g_1,g_2)} - \bar{D}_{\cdot})\\
&= \mathrm{Var}(\bar{D}_{\cdot}^{(g_1,g_2)}) + \mathrm{Var}(\bar{D}_{\cdot}) - 2\mathrm{Cov}(\bar{D}_{\cdot}^{(g_1,g_2)}, \bar{D}_{\cdot})\\
&= \frac{1}{N}\left(\xi_{10}^{(13)} + \xi_{01}^{(13)}\right) + \sigma_1^2 - \frac{4(N-1)}{G(NG-1)}\frac{2}{N}\xi_1^{(12,13)}\\
&- \frac{4N}{G(NG-1)}\left[\frac{1}{N}(\xi_{10}^{(13)} + \xi_{01}^{(13)}) + 2(G-2)\frac{1}{N}\xi_{10}^{(13,1;13,2)}\right]
\end{aligned} \tag{4.44}$$

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

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

So, under H_0,

$$\tau_2^2 = \{2(N-1)^2 + (NG-1)[2N(G-1) + (NG-1)(G-4)]\}\,\frac{2\xi_1^{(12)}}{NG(NG-1)^2} \tag{4.45}$$

As in BSS,

$$ABSS \sim N^2 \sum_{i=1}^{G(G-1)/2} \lambda_i \left(\chi_1^2\right)_i$$

where λ_i's are the characteristic roots of $\boldsymbol{\Sigma}_2 = \text{Var}(\mathbf{D}_2)$. The diagonal elements of $\boldsymbol{\Sigma}_2$ are τ_2^2 and, if all groups are different, the off-diagonal elements are

$$\begin{aligned}
&\text{Cov}(\bar{D}_{\cdot}^{(g_1,g_2)} - \bar{D}_{\cdot}, \bar{D}_{\cdot}^{(g_3,g_4)} - \bar{D}_{\cdot}) \\
&\quad = [(N-1)^2 - (NG-1)(NG+N-2)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} < 0
\end{aligned}$$

and if $g_1 = g_2$ or $g_1 = g_3$ or $g_2 = g_3$,

$$\begin{aligned}
&\text{Cov}(\bar{D}_{\cdot}^{(g_1,g_2)} - \bar{D}_{\cdot}, \bar{D}_{\cdot}^{(g_1,g_3)} - \bar{D}_{\cdot}) \\
&\quad = \{4(N-1)^2 + (NG-1)[4N(G-1) + (G-8)(NG-1)]\} \frac{\xi_1^{(12)}}{NG(NG-1)^2} \ .
\end{aligned}$$

Now

$$\text{E}_0(ABSS) = N^2 \text{trace}(\boldsymbol{\Sigma}_2) = N^2 \frac{G(G-1)}{2} \tau_2^2$$

$$\text{Var}_0(ABSS) = N^4 \text{trace}(\boldsymbol{\Sigma}_2)^2$$

The corresponding mean-square term is defined as

$$ABMS = \frac{ABSS}{N^2 \binom{G}{2}} = \frac{2}{G(G-1)} \mathbf{D}_2' \mathbf{D}_2$$

Then

$$\text{E}_0(ABMS) = \frac{2}{G(G-1)} \text{trace}(\boldsymbol{\Sigma}_2) = \tau_2^2$$

$$\text{Var}_0(ABMS) = \frac{4}{G^2(G-1)^2} \text{trace}(\boldsymbol{\Sigma}_2)^2$$

5. Test Statistics

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

$$\frac{WMS}{AWMS} = \frac{\frac{(N-2)(N-3)}{2N(N-1)} \left\{\mu_2 + \frac{4}{N} \sum_{i=1}^N (\Psi_{(2)1}(\mathbf{X}_i) - \mu_2)\right\} + O_p(N^{-1})}{\frac{(N-1)^2}{2N^2} \left\{\mu_2 \frac{4}{N} \sum_{i=1}^N (\Psi_{(2)1}(\mathbf{X}_i) - \mu_2)\right\} + O_p(N^{-1})}$$

But, $\frac{WMS}{AWMS} \xrightarrow{p} 1$ as $N \to \infty$, i.e, asymptotically the distribution of $\frac{WMS}{AWMS}$ is degenerate.

Let $\boldsymbol{\Sigma}_1 = \frac{1}{N}\boldsymbol{\Sigma}_1^\star$ and $\boldsymbol{\Sigma}_2 = \frac{1}{N}\boldsymbol{\Sigma}_2^\star$. Under H_0,

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

$$ABMS = \frac{ABSS}{N^2\binom{G}{2}} \sim \frac{2}{NG(G-1)} \sum_{i=1}^{G(G-1)/2} \lambda_{2i}^\star \left(\chi_1^2\right)_i$$

where $\lambda_{1g}^\star$'s and $\lambda_{2i}^\star$'s are the characteristic roots of $\boldsymbol{\Sigma}_1^\star$ and $\boldsymbol{\Sigma}_2^\star$, respectively. Also, under H_0, by theoretical results pertaining to U-statistics

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

and

$$\sqrt{N}(AWMS - \mu_2/2) \to \mathrm{N}\left(0, \frac{4}{G(G-1)}\xi_1^{(2)}\right) \quad .$$

Thus,

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

while

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

Define

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

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

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

and

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

Therefore,

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

and

$$T_{N,3} \sim \frac{4}{G(G-1)\mu_2} \sum_{i=1}^{G(G-1)/2} \lambda_{2i}^\star \left(\chi_1^2\right)_i$$

Because the elements of $\boldsymbol{\Sigma}_1^\star$ and $\boldsymbol{\Sigma}_2^\star$ 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.

6. 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 $Q(\mathbf{T}) = \dfrac{1}{n}Q(\mathbf{U}_n) + O_p(n^{-3/2})$.

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,

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

Let $\mathbf{V}_N = \mathbf{D}_1 - \boldsymbol{\nu}_1$, where $\boldsymbol{\nu}_1$ is a vector $G \times 1$ with elements ν_1. Then, $\mathrm{E}(\mathbf{V}_N) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{V}_N) = \boldsymbol{\Sigma}_1 = \dfrac{1}{N}\boldsymbol{\Sigma}_1^\star = O(N^{-1})$. 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 $\sqrt{N}\mathbf{V}_N \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_1^\star)$,

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

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

$$2\sqrt{N}\boldsymbol{\nu}_1'\mathbf{V}_N \sim \mathrm{N}\left(\mathbf{0}, 4\boldsymbol{\nu}_1'\boldsymbol{\Sigma}_1^\star\boldsymbol{\nu}_1\right)$$

Now,

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

$$\left(\frac{T_{N,2} - 2N\nu_1^2/\mu_2}{4\sqrt{N}\nu_1/(G\mu_2)}\right) = \frac{N\sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1)^2}{2\sqrt{N}\nu_1} + \sqrt{N}\sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1) + O_p(N^{-1})$$

Note that

$$\frac{N\sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1)^2}{2\sqrt{N}\nu_1} = O_p(N^{-1/2}), \text{ since } N\sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1)^2 = O_p(1)$$

and

$$\sqrt{N}\sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1) = O_p(1), \text{ since } \sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1) = O_p(N^{-1/2})$$

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

$$\left(\frac{T_{N,2} - 2N\nu_1^2/\mu_2}{4\sqrt{N}\nu_1/(G\mu_2)}\right) = \sqrt{N}\sum_{g=1}^{G}(\bar{D}^g_{\cdot} - \bar{D}_{\cdot} - \nu_1) + O_p(N^{-1/2})$$

Thus,

$$\mathrm{P}(T_{N,2} > \nu_1) = \mathrm{P}\left(Z > G\,\frac{(\mu_2 - 2N\nu_1)}{4\sqrt{N}}\right) \to 1, \quad \text{as } N \to \infty,$$

i.e., this test is consistent.

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

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

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

Note that

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

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

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

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

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

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} \tau_1^2 & \tau_{12} & \dots & \tau_{12} \\ \tau_{12} & \tau_1^2 & \dots & \tau_{12} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{12} & \tau_{12} & \dots & \tau_1^2 \end{pmatrix}$$

where

$$\tau_1^2 = \{(N-1)^2 + (NG-1)[N(G-1) + (NG-1)(G-2)]\} \frac{4\xi_1^{(12)}}{NG(NG-1)^2}$$

and

$$\tau_{12} = \{(N-1)^2 - (NG-1)(NG+N-2)\} \frac{4\xi_1^{(12)}}{NG(NG-1)^2}$$

Then,

$$\left(\boldsymbol{\gamma}_1^{\star}\right)'\boldsymbol{\Sigma}_1 = \gamma_1^{\star}\left[\tau_1^2 + (G-1)\tau_{12} \,\dots\, \tau_1^2 + (G-1)\tau_{12}\right]$$

and

$$\begin{aligned} &\tau_1^2 + (G-1)\tau_{12} = 0 \\ &\quad \Leftrightarrow G(N-1)^2 + (NG-1)[N(G-1) + (NG-1)(G-2) \\ &\quad\quad - (G-1)(NG+N-2)] = 0 \\ &\quad \Leftrightarrow N = 1 \end{aligned}$$

So, $\mathbf{V}_N'\mathbf{V}_N$ and $\left(\boldsymbol{\gamma}_1^{\star}\right)'\mathbf{V}_N$ are independent if and only if $N = 1$, which is not the case here.

Now, write

$$\frac{2}{G\mu_2}\left[N\mathbf{V}_N'\mathbf{V}_N + 2\sqrt{N}\left(\boldsymbol{\gamma}_1^{\star}\right)'\mathbf{V}_N\right] = \frac{2}{G\mu_2}\left[(\sqrt{N}\mathbf{V}_N + \boldsymbol{\gamma}_1^{\star})'(\sqrt{N}\mathbf{V}_N + \gamma_1^{\star}) - \left(\boldsymbol{\gamma}_1^{\star}\right)'\boldsymbol{\gamma}_1^{\star}\right]$$

and

$$\begin{aligned} T_{N,2} &= \frac{2N}{\mu_2}(BMS) = \frac{2N}{G\mu_2}\mathbf{D}_1'\mathbf{D}_1 \\ &= \frac{2}{G\mu_2}(\sqrt{N}\mathbf{V}_N + \boldsymbol{\gamma}_1^{\star})'(\sqrt{N}\mathbf{V}_N + \gamma_1^{\star}) + O_p(N^{-1/2}) \end{aligned}$$

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

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

The distribution of $\sqrt{N}\mathbf{D}_1'\mathbf{D}_1$ 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

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

Then,

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

Hence,

$$N\mathbf{D}_1'\mathbf{D}_1 = \mathbf{Y}'\mathbf{Y} \sim \sum_{i=1}^{G} \lambda_i \left(\chi_1^2(\delta_i)\right) \tag{6.46}$$

where $\delta_i = \dfrac{(\nu_{1i}^\star)^2}{\lambda_i}$, 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 (6.46),

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

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

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

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 $T_{N,3}$.

7. Data analysis

The data set consists of two groups of HIV infected individuals (subtype B and not B) with 46 sequences (individuals) each. The nucleotide sequences are all from epidemiologically independent individuals, which means that among the individuals in our sample, one did not infect the other, i.e., they were not sharing the same siringe, they were not partners. Since the sequences are in the nucleotide level, there are therefore four categories. After aligning the sequences and discarding the positions with no change, we end up with 155 positions.

Since the elements of $\boldsymbol{\Sigma}_1^\star$ and $\boldsymbol{\Sigma}_2^\star$ are unknown, the characteristic roots of these matrices are also unknown and the distributions of the test statsitics 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 46 sequences to each 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. Reapeat 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

$$T_{N2}obs = 20,07 \qquad T_{N3}obs = 4,17$$

For $R = 10,000$, the percentiles of the bootstrap distribution are given in Table 1 and the observed p-value for T_{N2} and T_{N3} are less than 1/10001. Therefore, we can say that relative to the within-clade variation, there is significant variability between the two clades and similarly, relative to the across-within-clade, there is significant variability acroos-between the two clades.

Table 1: Percentiles of the Bootstrap Distribution

Statistic	1%	5%	95%	99%
T_{N2}	0.0002	0.0007	2.6799	4.1866
T_{N3}	0.0000	0.0000	0.0132	0.0520

References

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

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

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

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

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

[6] R J Light and B H Margolin. 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, 1974.

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

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

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

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

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

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

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

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

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

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

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

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