

The Library of the Department of Statistics  
North Carolina State University

**MODELLING VARIABILITY IN THE HIV GENOME**

by

**Hildete Prisco Pinheiro**

Department of Biostatistics  
University of North Carolina

Institute of Statistics  
Mimeo Series No. 2186T

December 1997

MODELLING VARIABILITY IN THE HIV GENOME

by

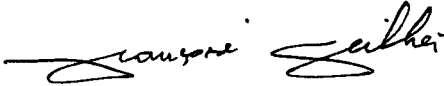
Hildete Prisco Pinheiro

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.


Chapel Hill

1997

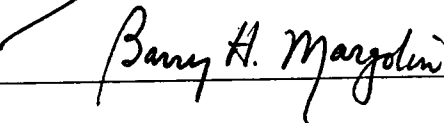
Approved by:



\_\_\_\_\_  
Advisor



\_\_\_\_\_  
Advisor



\_\_\_\_\_  
Reader

©1997

Hildete Prisco Pinheiro  
ALL RIGHTS RESERVED

## ABSTRACT

HILDETE PRISCO PINHEIRO: Modelling Variability in the HIV Genome. (Under the direction of Dr. Françoise Seillier-Moiseiwitsch and Dr. Pranab Kumar Sen.)

In modelling the mutational process in DNA sequences of the human immunodeficiency virus (HIV) one cannot assume independence among positions along the sequences. Sites are analyzed, at the nucleotide or amino-acid level, by comparing each sequence to the consensus sequence. The state at each site is regarded as a binary event (i.e., there is a mutation or not). Two families of models are considered. Each sequence can be thought of as a degenerate lattice, then autologistic models are applicable. The probability of mutation at a specific site, given all others, has an exponential form with, as predictor, a function of neighboring sites. Since there is no closed form for the likelihood function, a Markov-chain Monte-Carlo procedure is called upon. The second model is based on the Bahadur representation for the joint distribution of dichotomous responses and assumes only pairwise dependence among sites. Parameter estimation is performed via the maximum-likelihood method. An extension to three categories is suggested to evaluate the type of mutation (transition or transversion) or the absence of mutation. For the autologistic model, two dummy variables represent the three categories. The estimates of the parameters can also be obtained by Markov-chain Monte-Carlo methods.

It is also of interest to compare sequence variability between and within groups. These groups can be HIV-infected individuals from different geographical regions, different high-risk groups or different subtypes of the virus. Two analyses of variance for categorical data are proposed. One is based on an approach developed by Simpson (1949). We consider the variability in the distribution of the categorical response. Sequences are not considered on an individual basis. A test statistic is developed assuming independence among positions. The other is based on the Hamming distance, which is the proportion of positions where there is a difference between two aligned sequences. Sequences are considered on an individual basis. The interest is now in estimating the variability between, within and across groups. U-statistics represent

average distances. 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 decomposition of sum of squares. In order to find out the distributions of these sum squares, we use generalized U-statistics theory. Test statistics are constructed to test the hypothesis of homogeneity among the groups.

## Acknowledgments

I would like to thank my advisors Drs. Seillier-Moiseiwitsch and Sen for their guidance, patience, and encouragement, and all my committee members Drs. Margolin, Newman and Swanstrom for their helpful comments and suggestions. In particular, I would like to thank Dr. Seillier for her friendship and for providing me with the opportunity to work in the HIV and the Brain Tumor projects. I am grateful for the financial support provided by CAPES.

The love and encouragement exhibited by my parents, my in-laws and friends, especially Pai-Lien Chen and Hsiao-Chuan Tien for their love and support during my stressful periods. Most of all, I would like to thank my loving and understanding husband for his unconditional love, encouragement and the sharing of the family work. My especial thanks goes to my daughter Taís for providing me such wonderful moments with her smiles, first steps and all the joys of motherhood.

# Contents

<b>1</b>	<b>Introduction and Literature Review</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Biological Background . . . . .	3
1.3	Standard Statistical Procedures . . . . .	6
1.3.1	Markov Chain Models . . . . .	6
1.3.2	Genetic Variability . . . . .	8
1.4	Specific Biological Problems and Synopsis of the Solutions . . . . .	18
<b>2</b>	<b>Modelling the Mutation Process</b>	<b>20</b>
2.1	The Autologistic Model . . . . .	20
2.1.1	Estimation Procedure . . . . .	25
2.2	Model based on the Bahadur Representation . . . . .	29
2.2.1	Estimation Procedure . . . . .	31
2.3	Models for Three Categories . . . . .	31
2.3.1	Overview of the Literature . . . . .	31
2.3.2	An Alternative Extension of the Autologistic Model . . . . .	34
<b>3</b>	<b>Analyzing the Variability in DNA Sequences</b>	<b>37</b>
3.1	Variation in Categorical Data . . . . .	38

3.2	Partitioning the Measures of Diversity . . . . .	39
3.3	The Probabilistic Model . . . . .	43
3.4	Moments of Diversity Measures . . . . .	44
3.5	The Test Statistic . . . . .	48
3.6	The Power of the Test . . . . .	61
<b>4</b>	<b>Analysis of Variance based on the Hamming Distance</b>	<b>65</b>
4.1	The Total Sum of Squares and its decomposition . . . . .	66
4.2	Connections Between Sums of Squares and U-Statistics . . . . .	69
4.3	Asymptotic Distributions . . . . .	77
4.3.1	One sample U-statistics . . . . .	77
4.3.2	Multiple-Sample U-statistics . . . . .	91
4.4	Combining the U-statistics . . . . .	121
4.5	Test Statistics . . . . .	141
4.5.1	Power of the Tests . . . . .	142
<b>5</b>	<b>Numerical Studies and Data Analysis</b>	<b>147</b>
5.1	Modelling the Mutation Process . . . . .	147
5.1.1	The Autologistic Model . . . . .	147
5.1.2	Model based on the Bahadur Representation . . . . .	152
5.2	Analyzing the Variability in DNA Sequences . . . . .	153
5.2.1	Simulations . . . . .	153
5.2.2	Data Analysis . . . . .	156
<b>6</b>	<b>Conclusion and Future Research</b>	<b>158</b>
6.1	Concluding Summary . . . . .	158
6.2	Future Research . . . . .	160



6.2.1	An Application of the Autologistic Model to sequences from the <i>nef</i> gene . . . . .	160
6.2.2	Extension of the Autologistic Model to more than Two Categories	161
6.2.3	An Alternative Null Hypothesis for the Analysis of Diversity Measures . . . . .	161
6.2.4	Diversity Measures for Sequences with Dependent Positions .	162
6.2.5	Inclusion of Covariates in the Analysis of Diversity Measures .	165
6.2.6	Analysis of Variance based on the Hamming Distance when Sequences are not Independent . . . . .	165
6.2.7	Tests Based on Contrasts in the Analysis of Variance for the Hamming Distances . . . . .	166
	<b>Appendix A</b>	<b>167</b>
	<b>Appendix B</b>	<b>169</b>
	Bibliography . . . . .	178

# List of Tables

1.1	Analysis of Variance for Heterozygosity . . . . .	12
1.2	Formulae for the Number of Nucleotide Substitutions . . . . .	15
1.3	Pattern of Nucleotide Substitution . . . . .	16
3.1	Summary of the Data (one position) . . . . .	40
3.2	Contingency Table (K positions) . . . . .	42
5.1	Convergence Results . . . . .	150
5.2	Model 1 . . . . .	150
5.3	Model 2 . . . . .	151
5.4	Bahadur Representation Model (35 positions) . . . . .	152
5.5	Bahadur Representation Model (12 positions) . . . . .	153
5.6	Results of Simulations for Diversity Measures ( $p_{ck} = p_c$ ) . . . . .	154
5.7	Results of Simulations for Diversity Measures ( $p_{ck} \neq p_c$ ) . . . . .	155
5.8	Percentiles of the Bootstrap Dist. for Diversity Measures . . . . .	157
6.1	Contingency Table for Two Categories (dependent positions) . . . . .	163
6.2	Contingency Table for $C$ categories (dependent positions) . . . . .	163

# Chapter 1

## Introduction and Literature Review

### 1.1 Introduction

A brief review of molecular biology and genetic variability is presented in Section 1.2.

The primary interest is to model the mutation process exhibited in DNA sequences of the human immunodeficiency virus (HIV). The data set consists of sequences from different individuals, with  $n$  sites per sequence. We cannot assume independence among sites. Each sequence is compared to the consensus sequence at the nucleotide or amino-acid level.

Statistical methods to measure variability of DNA sequences, in general, are available and are discussed in Section 1.3.1, but one needs to be careful when modelling such data because of the known dependencies among neighboring nucleotide positions. An interesting question arises: how does one quantify the notion of *neighbor*? Tavaré & Giddings (1989) determine the distance over which there is base-frequency dependency by modelling the sequence as a Markov chain and estimating the order of the chain. For a high-order chain, the number of parameters may be too large to be reliably estimated with sequences of moderate length. Also, the dependency is located on one side only and this may not be true in reality. Positions along a sequence are

not ordered like time points. Raftery & Tavaré (1994) proposed an alternative model to reduce the number of parameters, *the mixture transition distribution model*. They apply it to repeated patterns in a DNA sequence. In these papers, the authors look at a single long sequence and try to see if there is a stochastic pattern for base-pair composition along the sequence. Dependencies on one side only are allowed. However, positions along a sequence cannot be regarded as a time-series: the sequence codes for a three-dimensional structure.

In our case we want to look at several sequences and compare them with the consensus one. We also would like to look at dependencies on both sides. Another question of interest is where are the “hot spots” in the sequence. If we have an estimate of the rate of mutation at each site along the sequences, we might be able to answer that question.

Another interest is to compare sequence variability between and within groups. Comparisons can be performed between and within individuals as well as between and within groups. These groups can be HIV-infected individuals from different geographical areas, different high-risk groups or different subtypes of the virus. Our interest would be if the variability is similar in each group.

The analysis of variance for continuous variables is a well-established procedure to compare a number of means. The experimental design guides its format (Cochran & Cox, 1957). When the variables are categorical, but ordered, rank analysis of variance can be used as a nonparametric method to test for mean differences (Daniel, 1978). In our case the variables are categorical and not ordered. Therefore, we need to find other alternative methods.

Weir (1990a) describes an analysis of variance for the genetic variation in the population, in particular for the amount of observed *heterozygosity* (Section 1.3.2). 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 Section 1.4 the biological problems and solutions given in this manuscript are discussed.

## 1.2 Biological Background

*Nucleotides* are the building blocks of genomes and each nucleotide has three components: a sugar, a phosphate and a base. The sugar may be one of two kinds: *ribose* or *deoxyribose*. In any given nucleic acid macromolecule, all the sugars are of the same kind. A nucleic acid with ribose is called *Ribonucleic Acid* or RNA, one with deoxyribose, *Deoxyrinucleic Acid* or DNA. DNA has four bases: *Adenine* (**A**), *Cytosine* (**C**), *Guanine* (**G**) and *Thymine* (**T**), where *Adenine* fits together with *Thymine* and *Guanine* with *Cytosine*. These are the so-called *base pairs*. A sequence of base pairs may be thought of as a series of "words" specifying the order of amino acids (each coded by three nucleotides) in a protein. To transform the DNA "words" into amino acids, some sophisticated molecular machinery is needed. Now, RNA comes into play. RNA also has four bases: **A**, **C**, **G** and *Uracil* (**U**) in place of **T**. *Adenine* is now complementary to *Uracil*. *Transcription* is the process by which a region of DNA is teased apart and a molecule of RNA is built along one strand by an enzyme, the *RNA Polymerase*, to begin protein synthesis. Each base of RNA is complementary to the corresponding base of DNA. The *messenger* RNA or mRNA then carries the genetic information from the DNA to the protein factory (Gonick & Wheelis, 1991).

A *retrovirus* has the ability to reverse the normal flow of genetic information from genomic DNA to mRNA (Varmus & Brown, 1989). Its genomic RNA encodes an enzyme that makes a DNA copy of its RNA and incorporates this DNA into the host genome (Gonick & Wheelis, 1991). HIV belongs to this family. Retroviruses have been subdivided into three subgroups on the basis of the in-vivo disease they produce and, more recently, on the basis of sequence homology (Teich, 1984; Varmus & Brown, 1989): *oncovirus*, *lentivirus* and *spumavirus*. HIV is classified as a lentivirus. These cause slow, chronic diseases (Cullen, 1991).

The human immunodeficiency virus evolves rapidly. HIV-1 and HIV-2 are two strains of HIV, with HIV-1 being the most common in the world and in the U.S. HIV-1 has nine genes (*gag*, *pol*, *vif*, *vpr*, *vpu*, *tat*, *rev*, *env* and *nef*), of which eight are conserved among the other primate lentiviruses. Most of the internal part of the genome is densely packed with protein-coding domains, such that some genes overlap. The DNA form of the genome is bounded by a repeated sequence, the LTR. The 5' copy of the LTR contains important transcription signals, while the 3' copy is used to encode part of the *nef* gene and the 3' end-defining polyadenylation site in the transcribed RNA (Seillier-Moiseiwitsch et al., 1994). The pattern of nucleotide variation is not constant over the whole genome. For instance, the genes encoding internal virion proteins, *gag* and *pol*, are more conserved than the gene coding for the envelope of the virus, *env*. Also, the nucleotide differences in *env* change the encoded amino acids more frequently, and therefore, the amino acid sequence exhibits more variation in *env* than in *gag* and *pol* (Coffin, 1986).

The viral envelope is the only gene product in direct contact with the host environment. It is thus not surprising that this gene has the most variable sequence (Hahn et al., 1985; Coffin, 1986; Seillier-Moiseiwitsch et al., 1994). To illustrate the diversity of HIV, *env* sequences were isolated at different times from each of a number of individuals (Hahn et al., 1986) and the greatest variation is within the so-called *hypervariable regions* V1-V5. The viruses from an individual differ from one another, but not as much as viruses from different individuals. Such high variation seems most likely due to a combination of two aspects. First, there may be a strong selection by immunological pressure for variation in these regions and, second, a lack of selection against variation may permit the presence of almost any sequence that does not interrupt translation.

In different organisms most genes do not follow a pattern in the two first codon positions, but in HIV there is a preference for **A** at the expense of **C**, in particular. In the third codon position the shift towards **A** is even greater, while in most other organisms **A** is rare in the third codon position. What is particular to HIV is that *purines* (**A** and **G**) predominate over *pyrimidines* (**C** and **T**). The overrepresentation of **A** is highest in *pol* and lowest in *env*, and may be related to the great genetic

variability of HIV. As **A** does not appear to be concentrated in the hypervariable segments of *env*, there is so far no plausible explanation for this peculiar coding strategy (Kypr & Mrázek, 1987).

The genetic variability of HIV-1 is relatively high compared to other retroviruses (Mansky & Temin, 1995). Error rates of purified HIV-1 reverse transcriptase determined with a DNA template (of the *lacZα* peptide gene) range from  $5 \times 10^{-4}$  to  $6.7 \times 10^{-4}$  (Roberts et al., 1988). To test the hypothesis that the mutation rate for HIV-1 is comparable to that of purified HIV-1 reverse transcriptase, Mansky & Temin (1995) developed a system to measure forward mutation rates with an HIV-1 vector containing the *lacZα* peptide gene as a reporter for mutations. They found that the forward mutation rate of HIV-1 in a single cycle of replication is  $3.4 \times 10^{-5}$  mutations per base pair per cycle. The in-vivo mutation rate of HIV-1 is therefore lower than the error rate of purified reverse transcriptase by a factor of 20 (Mansky & Temin, 1995). Explanations for this difference are: the association of viral or nonviral accessory proteins during reverse transcription, the influence of cellular mismatch repair mechanisms, and/or differences between the reverse transcriptase produced in vivo with that assayed in vitro.

Decomposing DNA sequences into tables of codon usage (i.e., frequency tables of codons among organisms) loses a lot of information and their interpretation can be misleading (Sharp, 1986). For example, looking at codon similarities between HIV and other viruses, one can think that the moloney murine leukemia virus is the most closely related to HIV, when in fact HIV codon usage is most similar to that of the influenza virus and cauliflower mosaic virus. It is also important to note that, unlike divergence in protein and DNA sequences, differentiation at the level of codon usage is not linear in time.

Genomic comparisons of virus isolates have shown that HIV-1 variants in Africa are both highly diverse and generally distinct from those in North America and Europe. Analysis based on *gag* or *env* indicate that African isolates are more heterogeneous than the North American/European group. For instance, McCutchan et al. (1992) compare 22 HIV-1 isolates from Zambia and 16 from North America. The Zambian isolates are most distant from a North American virus isolate, HIV<sub>MN</sub>: mean

difference from  $HIV_{MN}$  was 13.6%, with a range of 14.4-17.1%. Among the Zambian isolates the mean pairwise difference was 7.1% with a range of 4.4-9.8%. PCR and sequencing results indicate that HIV-1 isolates from Zambia are relatively homogeneous, but they are distinct from previously described HIV-1 isolates, clustering in the separate branch of the phylogenetic tree.

Sequences can be compared at either the nucleotide or amino-acid level. Nucleotide substitutions can be evaluated for mutations that cause changes in amino acids (nonsynonymous) vs. mutations that do not (silent or synonymous). Furthermore, we can have substitutions between purines only ( $A \leftrightarrow G$ ) or pyrimidines only ( $C \leftrightarrow T$ ), termed *transitions*, or we can have mutations between a purine and a pyrimidine ( $A \leftrightarrow C$ ,  $A \leftrightarrow T$ ,  $G \leftrightarrow C$ , or  $G \leftrightarrow T$ ), called *transversions*.

## 1.3 Standard Statistical Procedures

### 1.3.1 Markov Chain Models

Markov chain models have been used to analyze DNA sequences because of the format of these data. Tavaré & Giddings (1989) estimate the order of the chain.

Let  $\mathbf{X} = \{X_n, n = 1, 2, \dots\}$  be a stochastic process with  $m$  states. If these states represent nucleotides, then  $m$  is 4 ( $A = 1$ ,  $C = 2$ ,  $G = 3$  and  $T = 4$ ).  $\mathbf{X}$  is called a Markov chain of order  $k$  if

$$\begin{aligned} \Pr\{X_{n+1} = i_{n+1} \mid X_n = i_n, \dots, X_{n-k+1} = i_{n-k+1}, \dots, X_1 = i_1\} \\ = \Pr\{X_{n+1} = i_{n+1} \mid X_n = i_n, \dots, X_{n-k+1} = i_{n-k+1}\} \end{aligned}$$

for all  $n > k - 1$  and for all choices of states  $i_1, i_2, \dots, i_{n-1}$  from  $\{1, 2, \dots, m\}$ . In other words, the distribution of the next base in the sequence is determined by the  $k$  previous ones. When  $k = 0$ , the bases are independently distributed.

The transition probabilities are denoted by

$$p(i_1, i_2, \dots, i_k; i_{k+1}) \equiv \Pr\{X_{k+1} = i_{k+1} \mid X_k = i_k, \dots, X_2 = i_2, X_1 = i_1\} \quad (1.3.1)$$

The goals are to estimate the order  $k$  of the Markov chain as well as the transition



probabilities, and to test various hypotheses about the DNA sequence(s). For a single sequence, a typical hypothesis is that of independence among the bases, i.e., one tests whether  $k = 0$ .

Suppose the sequence of interest has length  $N$ . Let  $n(i_1, i_2, \dots, i_r)$  be the number of transitions  $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_r$  observed in the sequence and  $r = 1, 2, \dots$ . Since we are dealing with a multinomial distribution, it is known that the maximum likelihood estimator of  $p(i_1, i_2, \dots, i_k; i_{k+1})$  is

$$\hat{p}(i_1, i_2, \dots, i_k; i_{k+1}) = n(i_1, i_2, \dots, i_k, i_{k+1}) / n(i_1, i_2, \dots, i_k, +) \quad (1.3.2)$$

where

$$n(i_1, i_2, \dots, i_k, +) \equiv \sum_j n(i_1, i_2, \dots, i_k, j)$$

Note that for a  $k$ -th order Markov chain with  $m$  states, there are  $m^k(m-1)$  independent parameters to estimate. So, for a high-order chain, the number of parameters may be too big and a very large data set is needed. In this situation, very long sequences are required to get reliable estimates of the parameters.

To reduce the parameter space, Raftery (1985) suggests the following reparametrization for a  $k$ -th order Markov-chain model

$$p(i_1, i_2, \dots, i_k; i_{k+1}) = \sum_{j=1}^k \lambda_j q(i_j, i_{k+1}) \quad (1.3.3)$$

where  $\mathbf{Q} = \{q(i, j), 1 \leq i, j \leq m\}$  is a stochastic matrix,

$$q(i, j) \geq 0 \text{ and } \sum_{l=1}^m q(j, l) = 1 \quad j = 1, \dots, m,$$

and

$$\lambda_1 + \lambda_2 + \dots + \lambda_k = 1$$

The number of independent parameters is now reduced to  $m(m-1) + k - 1$ .

The only disadvantage with this reduced model is the estimation procedure. While in model (1.3.1) there is a simple expression for the maximum-likelihood estimate of  $p(i_1, i_2, \dots, i_k; i_{k+1})$ , given by (1.3.2), in model (1.3.3), the maximum-likelihood estimates of  $\lambda$ 's and  $q(i, j)$ 's must be found by numerically maximizing the

log-likelihood

$$L = \sum n(i_1, i_2, \dots, i_k, i_{k+1}) \ln \left\{ \sum_{j=1}^k \lambda_j q(i_j, i_{k+1}) \right\}$$

Raftery & Tavaré (1994) propose a computational algorithm for maximum-likelihood estimation of the parameters in model (1.3.3) by reducing the large number of constraints.

### 1.3.2 Genetic Variability

Weir (1990a) introduces a simple measure of genetic variation in a population: the observed *heterozygosity*. Let  $n_{luv}$  be the observed number of *heterozygotes*  $A_u A_v$ ,  $u \neq v$ , at a locus  $l$  in a sample of size  $n$ . Then the sample heterozygote frequency at locus  $l$  is

$$\tilde{H}_l = \sum_u \sum_{u \neq v} \frac{n_{luv}}{n}$$

If there are  $m$  loci, the *average heterozygosity* is

$$\tilde{H} = \frac{1}{m} \sum_{l=1}^m \tilde{H}_l$$

Since  $\tilde{H}_l$  is the sum of heterozygote counts that are multinomially distributed, each  $\tilde{H}_l$  is binomially distributed with

$$\mathbf{E}(\tilde{H}_l) = H_l \quad \text{and} \quad \text{Var}(\tilde{H}_l) = \frac{1}{n} H_l (1 - H_l)$$

where  $H_l$  is the proportion of heterozygotes at locus  $l$  in the population.  $\tilde{H}_l$  can also be written as

$$\tilde{H}_l = \frac{1}{n} \sum_{j=1}^n x_{jl}$$

where

$$x_{jl} = \begin{cases} 1 & \text{if individual is heterozygous at locus } l \\ 0 & \text{otherwise} \end{cases}$$

and for one population, we have

$$\mathbf{E}(x_{jl}) = H_l, \quad \mathbf{E}(x_{jl}^2) = H_l \quad \text{and} \quad \mathbf{E}(x_{jl} x_{j'l}) = H_l^2,$$

assuming that the individuals are independent within a sample.

Let  $\tilde{H} = \frac{1}{m} \sum_l \tilde{H}_l$  be the estimate of the average heterozygosity within a population. Then,

$$E(\tilde{H}) = \frac{1}{m} \sum_l H_l = H$$

To compute the variance of  $\tilde{H}$ , we need to take into account the covariance between heterozygosities at different loci, since they are not independent.

$$\begin{aligned} \text{Cov}(\tilde{H}_l, \tilde{H}_{l'}) &= E\left(\frac{1}{n} \sum_j x_{jl} \frac{1}{n} \sum_{j'} x_{j'l'}\right) - H_l H_{l'} \\ &= \frac{1}{n^2} E\left(\sum_j x_{jl} x_{j'l'} + \sum_j \sum_{j' \neq j} x_{jl} x_{j'l'}\right) - H_l H_{l'} \\ &= \frac{1}{n^2} [n H_{ll'} + n(n-1) H_l H_{l'}] - H_l H_{l'} \\ &= \frac{1}{n} (H_{ll'} - H_l H_{l'}) \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\tilde{H}) &= \frac{1}{m^2} \left[ \sum_l \text{Var}(\tilde{H}_l) + \sum_l \sum_{l' \neq l} \text{Cov}(\tilde{H}_l, \tilde{H}_{l'}) \right] \\ &= \frac{1}{nm^2} \left[ \sum_l H_l (1 - H_l) + \sum_l \sum_{l' \neq l} (H_{ll'} - H_l H_{l'}) \right] \end{aligned}$$

where the two-locus heterozygosity  $H_{ll'} = E(x_{jl} x_{j'l'})$  is the probability that a random individual is heterozygous at loci  $l$  and  $l'$ .

The sample variance of single-locus heterozygosities is

$$\begin{aligned} s_H^2 &= \frac{1}{m-1} \sum_l (\tilde{H}_l - \tilde{H})^2 \\ &= \frac{1}{m} \sum_l \tilde{H}_l^2 - \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} \tilde{H}_l \tilde{H}_{l'} \end{aligned}$$

with

$$\begin{aligned} E(s_H^2) &= \frac{1}{m} \sum_l \left[ H_l^2 + \frac{1}{n} H_l (1 - H_l) \right] \\ &\quad - \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} \left[ H_l H_{l'} + \frac{1}{n} (H_{ll'} - H_l H_{l'}) \right] \end{aligned}$$

Note that  $E(\tilde{H}_l^2) = H_l^2 + \text{Var}(\tilde{H}_l)$  and  $E(\tilde{H}_l \tilde{H}_{l'}) = H_l H_{l'} + \text{Cov}(\tilde{H}_l, \tilde{H}_{l'})$ .

To get the variance between populations, we need to take into account the dependence between members of the same population, caused by the founder effect. Let

$$M_l = E(x_{jl} x_{j'l}), \quad j \neq j'$$

where  $M_l$  is the probability that two individuals in the same population are heterozygous. Then,

$$\begin{aligned} E(\tilde{H}_l) &= H_l \\ \text{Var}(\tilde{H}_l) &= \frac{1}{n^2} E \left( \sum_j x_{jl}^2 + \sum_j \sum_{j' \neq j} x_{jl} x_{j'l} \right) - [E(\tilde{H}_l)]^2 \\ &= (M_l - H_l^2) + \frac{1}{n} (H_l - M_l) \end{aligned}$$

Averaging over all  $m$  loci

$$\tilde{H} = \frac{1}{m} \sum_l \tilde{H}_l = \frac{1}{nm} \sum_j \sum_l x_{jl}$$

Denote by  $M_{ll'}$  the probability that two random individuals from the same population are heterozygous, one at locus  $l$  and the other at locus  $l'$ :

$$M_{ll'} = E(x_{jl} x_{j'l'})$$

Then,

$$\begin{aligned} E(\tilde{H}) &= \frac{1}{m} \sum_l E(\tilde{H}_l) = \frac{1}{m} \sum_l H_l \\ \text{Var}(\tilde{H}) &= \frac{1}{m^2 n^2} E \left( \sum_j \sum_l x_{jl}^2 + \sum_j \sum_{j' \neq j} \sum_l x_{jl} x_{j'l} \right. \\ &\quad \left. + \sum_j \sum_l \sum_{l' \neq l} x_{jl} x_{j'l'} + \sum_j \sum_{j' \neq j} \sum_{l' \neq l} x_{jl} x_{j'l'} \right) - H^2 \\ &= \frac{1}{m^2} \left[ \sum_l (M_l - H_l^2) + \sum_l \sum_{l' \neq l} (M_{ll'} - H_l H_{l'}) \right] \\ &\quad + \frac{1}{m^2 n} \left[ \sum_l (H_l - M_l) + \sum_l \sum_{l' \neq l} (H_{ll'} - M_{ll'}) \right] \end{aligned} \quad (1.3.4)$$

The four terms in expression (1.3.4) can be rearranged to show how populations, loci, and individuals contribute to the variance of the average heterozygosity by setting out the calculations in a framework similar to that used for an analysis of variance.

Now, an index  $i$  is added to the indicator variable to denote the population being sampled, i.e.,

$$x_{ijl} = \begin{cases} 1 & \text{if individual } j \text{ from population } i \text{ is heterozygous at locus } l \\ 0 & \text{otherwise} \end{cases}$$

When  $m$  loci scored on  $n$  individuals taken from each of  $r$  populations, the indicator variables can be represented by the linear model

$$x_{ijl} = \alpha_i + \beta_{ij} + \gamma_l + (\alpha\gamma)_{il} + (\beta\gamma)_{ijl}, \quad i = 1, \dots, n, \quad j = 1, \dots, r, \quad l = 1, \dots, m$$

where  $\alpha_i$  represents the population effect,  $\beta_{ij}$  the individual-within-population effect,  $\gamma_l$  the locus effect,  $(\alpha\gamma)_{il}$  the population-by-locus interaction and  $(\beta\gamma)_{ijl}$  the "locus by individual within population" interaction.  $\gamma_l$  is a fixed effect, since the same loci are repeatedly scored (the investigator is interested in these particular loci only) and all the other effects are considered random.

$$\begin{aligned} E(\alpha_i) &= 0, \quad \text{Var}(\alpha_i) = \sigma_p^2 \\ E(\beta_{ij}) &= 0, \quad \text{Var}(\beta_{ij}) = \sigma_{i/p}^2 \\ E(\gamma_l) &= H_l \\ E((\alpha\gamma)_{il}) &= 0, \quad \text{Var}((\alpha\gamma)_{il}) = \sigma_{pl}^2, \\ E((\beta\gamma)_{ijl}) &= 0, \quad \text{Var}((\beta\gamma)_{ijl}) = \sigma_{li/p}^2 \end{aligned}$$

The analysis of variance format is shown in Table 1.1 and expectations of the terms defined in this table are

$$\begin{aligned} E(x_{ijl}^2) &= \sigma_p^2 + \sigma_{i/p}^2 + \gamma_l^2 + \sigma_{pl}^2 + \sigma_{li/p}^2 \\ E(x_{ij.}^2) &= m^2 \sigma_p^2 + m^2 \sigma_{i/p}^2 + \left( \sum_l H_l \right)^2 + m \sigma_{pl}^2 + m \sigma_{li/p}^2 \\ E(x_{i.l}^2) &= n^2 \sigma_p^2 + n \sigma_{i/p}^2 + n^2 H_l^2 + n^2 \sigma_{pl}^2 + n \sigma_{li/p}^2 \\ E(x_{i..}^2) &= n^2 m^2 \sigma_p^2 + n m^2 \sigma_{i/p}^2 + n^2 \left( \sum_l H_l \right)^2 + n^2 m \sigma_{pl}^2 + n m \sigma_{li/p}^2 \\ E(x_{.l.}^2) &= r n^2 \sigma_p^2 + r n \sigma_{i/p}^2 + r^2 n^2 H_l^2 + r n^2 \sigma_{pl}^2 + r n \sigma_{li/p}^2 \\ E(x_{...}^2) &= r n^2 m^2 \sigma_p^2 + r n m^2 \sigma_{i/p}^2 + r^2 n^2 \left( \sum_l H_l \right)^2 + r n^2 m \sigma_{pl}^2 + r n m \sigma_{li/p}^2 \end{aligned}$$

and

$$E(SS_1) = r n m \sigma_p^2 + r m \sigma_{i/p}^2 + \frac{r n}{m} \left( \sum_l H_l \right)^2 + r n \sigma_{pl}^2 + r \sigma_{li/p}^2$$

Table 1.1: Analysis of Variance for Heterozygosity

Source	d.f.	Sum of Squares	Expected Mean Square
Populations	$r - 1$	$SS_1 - C$	$\sigma_{i/p}^2 + m\sigma_{i/p}^2 + n\sigma_{pl}^2 + mn\sigma_p^2$
Individuals	$r(n - 1)$	$SS_2 - SS_1$	$\sigma_{i/p}^2 + m\sigma_{i/p}^2$
within populations			
Loci	$m - 1$	$SS_3 - C$	$\sigma_{i/p}^2 + n\sigma_{pl}^2 + L$
Loci by	$(r - 1)(m - 1)$	$SS_4 - SS_1$	$\sigma_{i/p}^2 + n\sigma_{ip}^2$
populations		$-SS_3 + C$	
Loci by	$r(n - 1)(m - 1)$	$SS_5 - SS_2$	$\sigma_{i/p}^2$
individuals		$-SS_4 + SS_1$	
within populations			
Total	$mnr - 1$	$SS_5 - C$	

$$SS_1 = \frac{1}{nm} \sum_i x_{i..}^2$$

$$SS_4 = \frac{1}{n} \sum_i \sum_l x_{i..l}^2$$

$$x_{ij.} = \sum_l x_{ijl}$$

$$x_{..l} = \sum_i \sum_j x_{ijl}$$

$$SS_2 = \frac{1}{m} \sum_i \sum_j x_{ij.}^2$$

$$SS_5 = \sum_i \sum_j \sum_l x_{ijl}^2$$

$$x_{i..} = \sum_j x_{ijl}$$

$$x_{...} = \sum_i \sum_j \sum_l x_{ijl}$$

$$SS_3 = \frac{1}{rn} \sum_l x_{..l}^2$$

$$C = \frac{1}{mnr} x_{...}^2$$

$$x_{i..} = \sum_j \sum_l x_{ijl}$$

$$\begin{aligned}
E(SS_2) &= rnm\sigma_p^2 + rnm\sigma_{i/p}^2 + \frac{rn}{m}(\sum_l H_l)^2 + rn\sigma_{pl}^2 + rn\sigma_{li/p}^2 \\
E(SS_3) &= nm\sigma_p^2 + m\sigma_{i/p}^2 + rn \sum_l H_l + nm\sigma_{pl}^2 + m\sigma_{li/p}^2 \\
E(SS_4) &= rnm\sigma_p^2 + rm\sigma_{i/p}^2 + rn \sum_l H_l + rnm\sigma_{pl}^2 + rm\sigma_{li/p}^2 \\
E(SS_5) &= rnm\sigma_p^2 + rnm\sigma_{i/p}^2 + rn \sum_l H_l + rnm\sigma_{pl}^2 + rnm\sigma_{li/p}^2 \\
E(C) &= nm\sigma_p^2 + m\sigma_{i/p}^2 + \frac{rn}{m}(\sum_l H_l)^2 + n\sigma_{pl}^2 + \sigma_{li/p}^2
\end{aligned}$$

These expressions lead to the expected mean squares in Table 1.1 and

$$L = \frac{1}{m-1} \sum_l (H_l - H)^2, \quad m > 1$$

The four variance components are:

for the population effect,

$$\sigma_p^2 = \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (M_{ll'} - H_l H_{l'}),$$

for the individual-within-population effect,

$$\sigma_{i/p}^2 = \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (H_{ll'} - M_{ll'}),$$

for the population-by-locus interaction,

$$\sigma_{pl}^2 = \frac{1}{m} \sum_l (M_l - H_l^2) - \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (M_{ll'} - H_l H_{l'})$$

and for the locus-by-individual-within-population interaction,

$$\sigma_{li/p}^2 = \frac{1}{m} \sum_l (H_l - M_l) - \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (H_{ll'} - M_{ll'}).$$

For comparative studies of DNA sequences, statistical methods for estimating the number of nucleotide substitutions are required as are models for the molecular evolution of the sequences (Gojobori et al., 1990).

Denote by  $I(t)$  the probability that two nucleotide bases at corresponding (homologous) sites at time  $t$  are identical to each other. First, let us assume that the

substitution rate is the same for all pairs of nucleotides and constant over time. The substitution process is described by a single parameter,  $\alpha$  say.

Consider the probability that two homologous nucleotide sites are identical to each other at time  $t$  and are also identical at time  $t + 1$ . In this case, there are two mutually exclusive events: one when both bases change into two other identical bases, having probability  $3\alpha^2$ , and the other when both nucleotide sites remain unchanged, having probability  $(1 - 3\alpha)^2$ . Therefore,

$$\Pr \left\{ \begin{array}{l} \text{two nucleotide bases remain identical at the site} \\ \text{at time } t + 1 \text{ when they are identical at time } t \end{array} \right\} = [(1 - 3\alpha)^2 + 3\alpha^2]I(t) \quad (1.3.5)$$

Now, consider the case where the bases are different from each other at time  $t$  and become identical at time  $t + 1$ . There are also two mutually exclusive events in this case. The first event is when a change occurs at one of the two corresponding sites but the other site remains unchanged. This occurs with probability  $2\alpha(1 - 3\alpha)$ . The second is when both nucleotide bases change into two other identical bases simultaneously. This occurs with probability  $2\alpha^2$ . Therefore,

$$\Pr \left\{ \begin{array}{l} \text{two nucleotide sites become identical at time } t + 1 \\ \text{when they are different from each other at time } t \end{array} \right\} = [2\alpha(1 - 3\alpha) + 2\alpha^2][1 - I(t)] \quad (1.3.6)$$

Thus, from (1.3.5) and (1.3.6) we obtain

$$I(t + 1) = ((1 - 3\alpha)^2 + 3\alpha)I(t) + (2\alpha(1 - 3\alpha) + 2\alpha^2)(1 - I(t)) \quad (1.3.7)$$

Satisfying the initial condition that  $I(0) = 1$ , the solution of the above equation is

$$I(t) = \frac{1}{4}[1 + 3(1 - 8\alpha + 16\alpha^2)^t] \quad (1.3.8)$$

Since  $\alpha$  is generally very small, the terms in  $\alpha^2$  in (1.3.8) are negligible and we get

$$I(t) = \frac{1}{4}[1 + 3(1 - 8\alpha)^t] \quad (1.3.9)$$

By solving a differential equation derived by substituting  $dI(t)/dt$  for  $I(t + 1) - I(t)$  in equation (1.3.7) (Nei, 1975) we get

$$I(t) = 1 - \frac{3}{4}(1 - e^{-8\alpha t}) \quad (1.3.10)$$



The mean number of nucleotide substitutions accumulated per site at time  $t$  is  $2 \times 3\alpha t$ . The factor 2 comes from the fact that the divergence of two sequences involves two evolutionary lines each having the time length  $t$ . Let  $K = 2 \times 3\alpha t$  be the evolutionary distance in terms of accumulated changes (i.e., the number of nucleotide substitutions) and  $F_D = 1 - I(t)$ . Then, from (1.3.10),

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3} F_D\right) \quad (1.3.11)$$

The standard error of  $K$  is given by

$$\sigma_K = \frac{\sqrt{\frac{1}{n} F_D (1 - F_D)}}{1 - \frac{4}{3} F_D}$$

where  $n$  is the total number of sites compared (Kimura & Ohta, 1972).

Gojobori et al. (1990) also construct two, three, four and six-parameter methods for estimating the total number of nucleotide substitutions. A summary of the formulae for  $K$  appears in Table 1.2 and the pattern of nucleotide substitutions according to each model in Table 1.3.

Table 1.2: Formulae for the Number of Nucleotide Substitutions

Model	Estimating Formula
One Parameter	$K = -\frac{3}{4} \ln(1 - F_D)$
Two Parameters	$K = -\frac{1}{2} \ln\{(1 - 2P - Q)\sqrt{(1 - 2Q)}\}$
Three Parameters	$K = -\frac{1}{4} \ln\{1 - 2P - 2\bar{Q})(1 - 2P - 2R)(1 - 2\bar{Q} - 2R)\}$
Four Parameters	$K = -\frac{1}{4} \ln \left\{ \frac{(S_{13} - \bar{Q}_1)(S_{24} - \bar{Q}_2) - [(P - R)/2]^2}{w(1-w)} \left[ 1 - \frac{P+R}{2w(1-w)} \right]^{8w(1-w)-1} \right\}$
Six Parameters	$K = -pq \ln \left( \frac{B_1}{pq} \right) - \frac{2qAqT}{p} \ln \left[ \frac{p}{3qAqT} \left( F_{12} - B_1 + \frac{3E_{12}}{B_1} \right) \right]$ $\quad - \frac{2qCqG}{q} \ln \left[ \frac{q}{3qCqG} \left( F_{34} - B_1 + \frac{3E_{34}}{B_1} \right) \right]$

For the two-parameter model,  $\alpha$  and  $\beta$  are the rates of transition and transversion, respectively. Let  $P + Q = F_D$ , where

$$P = P(t) = \frac{1}{4} - \frac{1}{2} e^{-4(\alpha+\beta)t} + \frac{1}{4} e^{-8\beta t} \quad \text{and} \quad Q = Q(t) = \frac{1}{2} - \frac{1}{2} e^{-8\beta t}.$$

$P$  and  $Q$  represent, respectively, the fractions of nucleotide sites with transition and transversion differences between the two sequences compared. Then  $k = \alpha + 2\beta$  is the

Table 1.3: Pattern of Nucleotide Substitution

Original Nucleotide	Substituted Nucleotide			
	A	T	C	G
One-parameter model				
A	-	$\alpha$	$\alpha$	$\alpha$
T	$\alpha$	-	$\alpha$	$\alpha$
C	$\alpha$	$\alpha$	-	$\alpha$
G	$\alpha$	$\alpha$	$\alpha$	-
Two-parameter model				
A	-	$\beta$	$\beta$	$\alpha$
T	$\beta$	-	$\alpha$	$\beta$
C	$\beta$	$\alpha$	-	$\beta$
G	$\alpha$	$\beta$	$\beta$	-
Three-parameter model				
A	-	$\beta$	$\gamma$	$\alpha$
T	$\beta$	-	$\alpha$	$\gamma$
C	$\gamma$	$\alpha$	-	$\beta$
G	$\alpha$	$\gamma$	$\beta$	-
Four-parameter model				
A	-	$\gamma$	$\theta\alpha$	$\alpha$
T	$\gamma$	-	$\alpha$	$\theta\alpha$
C	$\theta\beta$	$\beta$	-	$\gamma$
G	$\beta$	$\theta\beta$	$\gamma$	-
Six-parameter model				
A	-	$\alpha_1$	$\alpha$	$\alpha$
T	$\beta_1$	-	$\alpha$	$\alpha$
C	$\beta$	$\beta$	-	$\alpha_2$
G	$\beta$	$\beta$	$\beta_2$	-

number of nucleotide substitutions per site per year, and  $K = 2kt$  is the total number of nucleotide substitutions per site between the two sequences which diverged from the common ancestor  $t$  years ago.

For the three-parameter model (Kimura, 1981),  $P(t)$  is the fraction of sites (in the two sequences being compared) having **TC** or **AG** nucleotide pairs at time  $t$ ,  $\bar{Q}(t)$  the fraction of sites having **TA** or **CG** nucleotide pairs, and  $R(t)$  is the fraction of sites having **TG** or **CA** nucleotide pairs:

$$P = P(t) = [1 - e^{-4(\alpha+\beta)t} - e^{-4(\alpha+\gamma)t} + e^{-4(\beta+\gamma)t}]/4,$$

$$\bar{Q} = \bar{Q}(t) = [[1 - e^{-4(\alpha+\beta)t} + e^{-4(\alpha+\gamma)t} - e^{-4(\beta+\gamma)t}]/4,$$

$$R(t) = [1 + e^{-4(\alpha+\beta)t} - e^{-4(\alpha+\gamma)t} - e^{-4(\beta+\gamma)t}]/4,$$

and  $K = 2(\alpha + \beta + \gamma)t$  is the total number of nucleotide substitutions per site.

For the four-parameter model proposed by Takahata & Kimura (1981),  $w$  is the fraction of **A + T** in the two DNA sequences compared.  $S_{13}$  is the fraction of sites having **AA** or **TT** nucleotide pairs.  $S_{24}$  represents the fraction of sites having **CC** or **GG** nucleotide pairs.  $\bar{Q}_1$  is the fraction of sites having **AT** pairs.  $Q_1$  is the fraction of sites having **GC** pairs.  $P$  is the fraction of sites having **CT** or **AG** pairs, and  $Q$  is the fraction of sites having **GT** or **AC** pairs.

The six-parameter method is based on the model proposed by Kimura (1981) and Gojobori et al. (1982) derived its exact formulation.  $q_A$ ,  $q_T$ ,  $q_C$ , and  $q_G$  stand, respectively, for the contents of **A**, **T**, **C** and **G** in the DNA sequences compared. Also,

$$p = q_A + q_T, \quad q = q_C + q_G$$

$$B_1 = pq - (x_{AC} + x_{AG} + x_{TC} + x_{TG}),$$

$$E_{12} = (q_Aq - x_{AC} - x_{AG})(q_Tq - x_{TC} - x_{TG}),$$

$$E_{34} = (q_Cp - x_{AC} - x_{TC})(q_Gp - x_{AG} - x_{TG}),$$

$$F_{12} = x_{AA} + x_{TT} - x_{AT} - p^2 + 3q_Aq_T$$

$$\text{and } F_{34} = x_{CC} + x_{GG} - x_{CG} - q^2 + 3q_Cq_G,$$

where  $x_{ii}$  represents the fraction of sites having the same base pairs  $i$ , and  $2x_{ij}$  ( $i \neq j$ ) is the fraction of sites having different base pairs  $i$  and  $j$  ( $i, j = \text{A, C, T, G}$ ).

## 1.4 Specific Biological Problems and Synopsis of the Solutions

In modelling the mutation process exhibited in DNA sequences of the human immunodeficiency virus (HIV) one cannot assume independency among sites. Suppose the data set consists of a set of nucleotide (or amino-acid) sequences from different individuals, with  $n$  sites per sequence. Each sequence is compared to the consensus sequence at the nucleotide (or amino-acid) level. First, we consider the mutation process at each site as a binary event (i.e., there is a mutation or not).

We formulate two models (Chapter 2). In the first model (Section 2.1), each sequence is thought of as a degenerate lattice and we use results pertaining to spatial theory to build up an autologistic model, where the probability of mutation at a specific site, given all others, has an exponential form with as predictor a function of neighboring sites. Also, since we have sequences from different individuals, we have several independent lattices, not just one as is the case in image reconstruction where this theory is used. Parameter estimation is problematic: there is no closed form for the likelihood function to get maximum likelihood estimates. We therefore resort to Markov-chain Monte-Carlo procedures to estimate the parameters. In particular, we use the Metropolis algorithm, to generate random vectors, with the pseudolikelihood estimates as initial values.

The second model (Section 2.2) is based on a Bahadur representation for the joint distribution of the responses for the  $n$  dichotomous sites, assuming pairwise dependence only among sites. The probability of mutation is no longer assumed to have an exponential form. Parameter estimation is performed via the maximum-likelihood method.

Then, we extend the model to three categories denoting the type of mutation (*transition* or *transversion*) or the absence of mutation. We construct an extension of the autologistic model (Section 2.3), keeping in mind that these categories are not ordered. We define two dummy variables to represent the three categories. The model is similar to the autologistic model used in the binary case, with vectors as the response variable for each site. The estimates are again obtained via Monte-Carlo

Markov-chain simulation.

The other problem of interest is the comparison of sequences. Comparisons can be performed between and within individuals as well as between and within groups. For example, we may be interested in comparing DNA sequences from different geographical areas to see whether the variability is similar in each. We develop two approaches.

In Chapter 3 we propose a categorical analysis of variance extending the work of Simpson (1949) and Light & Margolin (1971) to consider a number of sites (because in the context of interest, a single position yields little information). The sequences are not considered on an individual basis, i.e., at each position we count the number of sequences in each category and the variability in the distribution of these counts is considered. A test statistic is developed, assuming independence among positions, to test the hypothesis that the probabilities of being at a certain category at a specific position is the same for all groups.

The analysis of variance proposed in Chapter 4 is based on Hamming distances. In comparing two aligned sequences, the Hamming distance is the proportion of positions where they differ. Our interest now is in estimating the variability between, within and across groups. We make all possible pairwise sequence comparisons within and across groups. Note that now sequences are considered on an individual basis. We use U-statistics 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 usual decomposition. In order to find the distributions of these terms, we use generalized U-statistics theory (Puri & Sen, 1971; Lee, 1990; Sen & Singer, 1993). We develop statistics to test the hypothesis of homogeneity among groups.

Numerical studies for the test statistics developed in Chapter 3 and data analysis, using the methodologies proposed in this dissertation, are shown in Chapter 5.

The conclusions and some directions for future research are discussed in Chapter 6.

# Chapter 2

## Modelling the Mutation Process

Our primary interest is in modelling whether there is a mutation or not at each site. We compare each sequence to the consensus sequence and look for differences. The response is thus binary and two alternative models are formulated: an *autologistic* model (Section 2.1), and a Bahadur representation for the joint distribution of Bernoulli trials (Section 2.2). We assume only pairwise dependence among sites. Parameter estimation is discussed in each case.

### 2.1 The Autologistic Model

The autologistic model was introduced by Besag (1972, 1974, 1975) and is widely suited to spatial binary data (Cressie, 1993). This parametric family can handle situations involving both spatial correlation and dependence on covariates. Applications include modelling the distribution of plant species in terms of climate variables like temperature and rainfall (Huffer & Wu, 1995a), the effect of soil variables on disease incidence in plants (Gumpertz & Graham, 1995) and network autocorrelation data (Smith, Calloway & Morrissey, 1995). It is important to point out that in these papers the whole data set consists of a single lattice, but in our case we have a number of independent lattices (i.e., sequences).

Let

$$y(i) = \begin{cases} 0 & \text{if there is no mutation at site } i \\ 1 & \text{if there is a mutation at site } i \end{cases}$$

Construct  $1 \times n$  vectors, where  $n$  is the total number of sites along a sequence:

$$\mathbf{y} = (y(1) \dots y(n)) \quad \text{and} \quad \mathbf{0} = (0 \ 0 \ \dots \ 0).$$

The model formulation requires that we define the concept of *neighborhood* and introduce the *positivity condition*.

**Definition 2.1**

A site  $j$  is a *neighbor* of site  $i$  if the conditional distribution of  $Y(i)$ , given all other site values, depends functionally on  $y(j)$ , for  $j \neq i$ . Also define

$$N_i = \{j : j \text{ is a neighbor of } i\} \tag{2.1.1}$$

to be the *neighborhood* of site  $i$ . ■

**Positivity Condition**

Let  $Y$  be a discrete variable associated with  $n$  sites. Define  $\zeta = \{\mathbf{y} : \Pr(\mathbf{y}) > 0\}$  and  $\zeta_i = \{y(i) : \Pr(y(i)) > 0\}$ ,  $i = 1, \dots, n$ . Then the *positivity condition* is satisfied if  $\zeta = \zeta_1 \times \dots \times \zeta_n$ . For a continuous variable, the same definition applies except that  $\Pr(\cdot)$  is replaced by  $f(\cdot)$ . ■

This condition states that the support of the joint distribution is the cartesian product of the supports for the marginal distributions. It implies that considering the elements  $\{y(i) : i, \dots, n\}$  jointly does not rule out combinations allowed in the set  $\{y(1)\} \times \{y(2)\} \times \dots \times \{y(n)\}$ . For instance, this condition is invalidated for an infectious disease model. Because of the one-way nature of infection, the following situation is not allowed:  $y(i) = 1$  when  $\{y(j) = 0, j \in N_i\}$ , i.e.,  $y(i)$  is affected when all the neighbors of  $i$  are not affected.

Without loss of generality, assume that 0 can occur at each site. Let

$$Q(\mathbf{y}) = \log\{\Pr(\mathbf{y})/\Pr(\mathbf{0})\}, \quad \mathbf{y} \in \zeta, \tag{2.1.2}$$

where  $\zeta$  is the support of the distribution of  $\mathbf{Y}$ . Then the knowledge of  $Q(\cdot)$  is equivalent to the knowledge of  $\Pr(\cdot)$ , because

$$\Pr(\mathbf{y}) = \exp(Q(\mathbf{y})) / \sum_{\mathbf{y} \in \zeta} \exp(Q(\mathbf{y}))$$

in the case of discrete  $y$ 's. A similar formula, with integrals replacing summations, applies for continuous  $y$ 's.

**Proposition 2.1** (Cressie, 1993)

The function  $Q$  satisfies the following two properties:

i.

$$\frac{\Pr(y(i) \mid \{y(j) : j \neq i\})}{\Pr(0(i) \mid \{y(j) : j \neq i\})} = \frac{\Pr(\mathbf{y})}{\Pr(\mathbf{y}_i)} = \exp(Q(\mathbf{y}) - Q(\mathbf{y}_i)) \quad (2.1.3)$$

where  $0(i)$  denotes the event  $Y(i) = 0$  and  $\mathbf{y}_i \equiv (y(1), \dots, y(i-1), 0, y(i+1), \dots, y(n))$

ii.  $Q$  can be expanded uniquely on  $\zeta$  as

$$\begin{aligned} Q(\mathbf{y}) = & \sum_{1 \leq i \leq n} y(i) G_i(y(i)) + \sum_{1 \leq i < j \leq n} y(i) y(j) G_{ij}(y(i), y(j)) \\ & + \sum_{1 \leq i < j < k \leq n} y(i) y(j) y(k) G_{ijk}(y(i), y(j), y(k)) + \dots \\ & + y(1) \dots y(n) G_{1\dots n}(y(1), \dots, y(n)), \quad \mathbf{y} \in \zeta. \end{aligned} \quad (2.1.4)$$

■

Note that although the expansion (2.1.4) is unique, the function  $\{G_{ij\dots}\}$  are not uniquely specified. By defining  $G_{ij\dots}(y(i), y(j), \dots) \equiv 0$  whenever one of the arguments is 0, uniqueness is obtained.

**Example** (Liang, Zeger & Qaqish, 1992)

The representation of  $Q(\cdot)$  in (2.1.4) has been used to represent the probability distribution for a vector  $\mathbf{x}$  of binary responses in a saturated log-linear model:

$$\Pr(\mathbf{x}) = \exp\left\{u_0 + \sum_{j=1}^n u_j x_j + \sum_{j < k} u_{jk} x_j x_k + \dots + u_{12\dots n} x_1 \dots x_n\right\} \quad (2.1.5)$$

where there are  $2^n - 1$  parameters  $u = (u_1, \dots, u_n, u_{11}, u_{12}, \dots, u_{n-1,n}, \dots, u_{12\dots n})'$ . These have straightforward interpretations in terms of conditional probabilities. For



example,

$$u_j = \text{logit}\{\Pr(x_j = 1 \mid x_k = 0, k \neq j)\}, \quad j = 1, \dots, n,$$

$$u_{jk} = \log \text{OR}(x_j, x_k \mid x_l = 0, l \neq j, k), \quad j < k = 1, \dots, n,$$

and

$$u_{123} = \log \text{OR}(x_1, x_2 \mid x_3 = 1, x_l = 0, l > 3) - \log \text{OR}(x_1, x_2 \mid x_3 = 0, x_l = 0, l > 3) \quad (2.1.6)$$

where

$$\text{OR}(v, w) = \frac{\Pr(v = w = 1) \Pr(v = w = 0)}{\Pr(v = 1, w = 0) \Pr(v = 0, w = 1)}$$

The implication of properties (i) and (ii) is that the expansion (2.1.4) for  $Q(\mathbf{y})$  is actually made up of conditional probabilities. For instance,

$$y(i) G_i(y(i)) = \log \left[ \frac{\Pr(y(i) \mid \{0(j) : j \neq i\})}{\Pr(0(i) \mid \{0(j) : j \neq i\})} \right]$$

$$y(i) y(j) G_{ij}(y(i), y(j)) = \log \left\{ \frac{\Pr(y(i) \mid y(j), \{0(l) : l \neq i, j\})}{\Pr(0(i) \mid y(j), \{0(l) : l \neq i, j\})} \times \frac{\Pr(0(i) \mid \{0(l) : l \neq i\})}{\Pr(y(i) \mid \{0(l) : l \neq i\})} \right\}$$

■

From (2.1.2), the joint probability distribution of  $\mathbf{y}$ ,  $\Pr(\mathbf{y})$ , is proportional to  $\exp(Q(\mathbf{y}))$ . Finding the proportionality constant as a function of the parameters enables us to write down the full likelihood and to obtain the maximum likelihood estimates of the parameters. Unfortunately, this is not always possible. Further, there is a powerful theorem regarding the form the function  $Q$  must take so that the conditional expressions combine consistently into a proper joint distribution. We must first define a *clique*.

### Definition 2.2

A *clique* is a set of sites that consists either of a single site or of sites that are all neighbors of each other. ■

### Theorem 2.1 (Hammersley-Clifford, 1971)

Suppose that  $\mathbf{Y}$  is distributed according to a Markov random field on  $\zeta$  that satisfies

the positivity condition. Then, the negpotential function  $Q(\cdot)$  given by (2.1.4) must satisfy the property that if sites  $i, j, \dots, n$  do not form a clique, then  $G_{ij\dots n}(\cdot) = 0$ , where cliques are defined by the neighborhood structure  $\{N_i : i = 1, \dots, n\}$ . ■

Assuming only pairwise dependence between sites,

$$Q(\mathbf{y}) = \sum_{1 \leq i \leq n} y(i) G_i(y(i)) + \sum_{1 \leq i < j \leq n} y(i) y(j) G_{ij}(y(i), y(j)) \quad (2.1.7)$$

Since  $y(i)$  is either 1 or 0, the only values for the functions  $G$  needed in (2.1.7) are  $G_i(1) \equiv \alpha_i$  and  $G_{ij}(1, 1) \equiv \gamma_{ij}$ . Thus,

$$Q(\mathbf{y}) = \sum_{i=1}^n \alpha_i y(i) + \sum_{1 \leq i < j \leq n} \gamma_{ij} y(i) y(j) = \log\{\Pr(\mathbf{y})/\Pr(\mathbf{0})\} \quad (2.1.8)$$

where  $\gamma_{ij} = 0$  unless sites  $i$  and  $j$  are neighbors. Then,

$$Q(\mathbf{y}) - Q(\mathbf{y}_i) = \alpha_i y(i) + \sum_{j=1}^n \gamma_{ij} y(i) y(j)$$

where  $\gamma_{ij} = \gamma_{ji}$  and, to maintain identifiability of the parameters,  $\gamma_{ii} = 0$ . Therefore, from (2.1.3),

$$\frac{\Pr(y(i) \mid \{y(j) : j \neq i\})}{\Pr(0(i) \mid \{y(j) : j \neq i\})} = \exp\{\alpha_i y(i) + \sum_{j=1}^n \gamma_{ij} y(i) y(j)\}$$

Because  $y(i) = 0$  or 1,

$$\Pr(0(i) \mid \{y(j) : j \neq i\}) = \frac{1}{1 + \exp(\alpha_i + \sum_{j=1}^n \gamma_{ij} y(j))}$$

and

$$P(y(i) \mid \{y(j) : j \neq i\}) = \frac{\exp(\alpha_i y(i) + \sum_{j=1}^n \gamma_{ij} y(i) y(j))}{1 + \exp(\alpha_i + \sum_{j=1}^n \gamma_{ij} y(j))} \quad (2.1.9)$$

Even with just pairwise dependence this formulation may involve too many parameters for data sets of moderate size and we need to reduce the number of parameters by imposing some constraints. For instance,

- For the  $\alpha$ 's,

1.  $\alpha_i$ 's all different;
  2.  $\alpha_i = \alpha, \forall i$ ;
  3. a different  $\alpha_i$  for each specific amino acid (or nucleotide) in the consensus sequence.
- For the  $\gamma$ 's,
    1.  $\gamma$ 's all different;
    2.  $\gamma_{ij} = 0$  if  $|i - j| > d$ , where  $d$  is some chosen threshold;
    3.  $\gamma_{ij} = \gamma \rho^{|i-j|}$ ;
    4. equal  $\gamma_{ij}$  for each pair of specific amino acids (or nucleotide) in the consensus sequence;
    5.  $\gamma_{ij} = \gamma, \forall i, j$ .

### 2.1.1 Estimation Procedure

Suppose the data consist of  $m$  independent sequences with  $n$  sites. In fact, we have  $y_k(i)$ , where  $i = 1, \dots, n$  and  $k = 1, \dots, m$ . Then,  $\mathbf{y}(i)$  is a  $m \times 1$  vector.

We can get approximate estimates by using the *pseudolikelihood* function  $L_P$  (Besag, 1975), i.e. the product of the conditional probabilities, which specifies the model assuming independence among sites.

$$L_P(\boldsymbol{\theta}; \mathbf{y}(1), \dots, \mathbf{y}(n)) = \prod_{i=1}^n \Pr(\mathbf{y}(i) \mid \{\mathbf{y}(j) : j \neq i\}; \boldsymbol{\theta})$$

where  $\boldsymbol{\theta} = (\alpha_1, \dots, \alpha_n, \gamma_{12}, \dots, \gamma_{n-1n})'$  is the parameter vector. Since the sequences are independent, we can write

$$\begin{aligned} L_P(\boldsymbol{\theta}; \mathbf{y}(1), \dots, \mathbf{y}(n)) &= \prod_{k=1}^m \prod_{i=1}^n \Pr(y_k(i) \mid \{y_k(j) : j \neq i\}) \\ &= \prod_{k=1}^m \prod_{i=1}^n \left\{ \frac{\exp[\alpha_i y_k(i) + \sum_{j=1}^n \gamma_{ik} y_k(i) y_k(j)]}{1 + \exp[\alpha_i + \sum_{j=1}^n \gamma_{ij} y_k(j)]} \right\} \end{aligned} \quad (2.1.10)$$

The maximum pseudolikelihood estimate (MPLE) is the value of  $\boldsymbol{\theta}$  which maximizes  $L_P$ , i.e., (2.1.10). The MPLE is consistent and asymptotically normal (Comets,

1992; Comets & Gidas, 1992). It has some disadvantages: its efficiency is unknown and expected to be inferior to that of the maximum likelihood estimate (MLE). Also, there is no direct way of obtaining standard errors for the estimates.

In order to calculate the MLE we need to proceed by Markov-chain Monte-Carlo simulation as proposed by Geyer & Thompson (1992). We can write the joint distribution of  $\mathbf{y} = (\mathbf{y}(1), \dots, \mathbf{y}(n))$  as

$$\Pr\{\mathbf{y}(1), \dots, \mathbf{y}(n); \boldsymbol{\theta}\} = C(\boldsymbol{\theta})F(\mathbf{y}(1), \dots, \mathbf{y}(n); \boldsymbol{\theta}).$$

where  $F$  is an explicitly computable function and  $C(\boldsymbol{\theta}) = \Pr\{\mathbf{0}(1), \dots, \mathbf{0}(n); \boldsymbol{\theta}\}$ . For the autologistic model,

$$F = \exp \left\{ \sum_{k=1}^m \sum_{i=1}^n \alpha_i y_k(i) + \sum_{k=1}^m \sum_{1 \leq i < j \leq n} \gamma_{ij} y_k(i) y_k(j) \right\} \quad (2.1.11)$$

Fix some specific value of  $\boldsymbol{\theta}$ ,  $\boldsymbol{\theta}_0$  say. Suppose we can generate a random vector  $(\mathbf{Y}'(1), \dots, \mathbf{Y}'(n))$  from the distribution with  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . The random variable

$$\frac{F(\mathbf{Y}'(1), \dots, \mathbf{Y}'(n); \boldsymbol{\theta})}{F(\mathbf{Y}'(1), \dots, \mathbf{Y}'(n); \boldsymbol{\theta}_0)}$$

has mean

$$\begin{aligned} & \sum_{\mathbf{y}'(1), \dots, \mathbf{y}'(n)} \frac{F(\mathbf{y}'(1), \dots, \mathbf{y}'(n); \boldsymbol{\theta})}{F(\mathbf{y}'(1), \dots, \mathbf{y}'(n); \boldsymbol{\theta}_0)} C(\boldsymbol{\theta}_0) F(\mathbf{y}'(1), \dots, \mathbf{y}'(n); \boldsymbol{\theta}_0) \\ &= C(\boldsymbol{\theta}_0) \sum_{\mathbf{y}'(1), \dots, \mathbf{y}'(n)} F(\mathbf{y}'(1), \dots, \mathbf{y}'(n); \boldsymbol{\theta}) \\ &= \frac{C(\boldsymbol{\theta}_0)}{C(\boldsymbol{\theta})}, \quad \text{since } \sum_{\mathbf{y}'(1), \dots, \mathbf{y}'(n)} C(\boldsymbol{\theta}) F(\mathbf{y}'(1), \dots, \mathbf{y}'(n); \boldsymbol{\theta}) = 1. \end{aligned}$$

Then, suppose we have a (not necessarily independent) set of random vectors  $(\mathbf{Y}^{(r)}(1), \dots, \mathbf{Y}^{(r)}(n))$ ,  $r = 1, 2, \dots, R$ , each drawn from the distribution  $C(\boldsymbol{\theta}_0)F(\cdot; \boldsymbol{\theta}_0)$  for some fixed  $\boldsymbol{\theta}_0$ . Denote the observations by  $(\mathbf{Y}(1), \dots, \mathbf{Y}(n))$ . The function

$$H(\boldsymbol{\theta}) = \frac{1}{R} \sum_{r=1}^R \frac{F(\mathbf{Y}^{(r)}(1), \dots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta})}{F(\mathbf{Y}^{(r)}(1), \dots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta}_0)} \times \frac{F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta})} \quad (2.1.12)$$

is an unbiased estimator of the likelihood ratio of  $\boldsymbol{\theta}_0$  to  $\boldsymbol{\theta}$ ,

$$\frac{C(\boldsymbol{\theta}_0) F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{C(\boldsymbol{\theta}) F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta})}$$

since

$$\begin{aligned} E[H(\boldsymbol{\theta})] &= \frac{1}{R} \frac{F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta})} \sum_{r=1}^R E \left[ \frac{F(\mathbf{Y}^{(r)}(1), \dots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta})}{F(\mathbf{Y}^{(r)}(1), \dots, \mathbf{Y}^{(r)}(n); \boldsymbol{\theta}_0)} \right] \\ &= \frac{C(\boldsymbol{\theta}_0)F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta}_0)}{C(\boldsymbol{\theta})F(\mathbf{Y}(1), \dots, \mathbf{Y}(n); \boldsymbol{\theta})} \end{aligned}$$

To obtain the MLE of  $\boldsymbol{\theta}$ , use equation (2.1.12), based on a single simulated sequence, and minimize it with respect to  $\boldsymbol{\theta}$  (Geyer & Thompson, 1992).

Here is a summary of the procedure:

1. Choose a model, as in equation (2.1.9).
2. Use the MPLE to generate an initial point estimate ( $\boldsymbol{\theta}_0$ ) for the unknown parameter vector.
3. Use Gibbs sampling or the Metropolis algorithm to generate a simulated sequence of random vectors from the distribution with parameter vector  $\boldsymbol{\theta}_0$ . Discard an initial warm-up sample, then generate random vectors  $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)} \dots$ . Select  $R$  vectors ( $R$  should be at least 1000) by sampling every  $k$  (50, say) because of the dependence between successive vectors.
4. Use equation (2.1.12) to define  $H(\boldsymbol{\theta})$ , the simulated likelihood ratio of  $\boldsymbol{\theta}_0$  to  $\boldsymbol{\theta}$ .
5. Using numerical optimization, maximize the function  $-\log H(\boldsymbol{\theta})$  to obtain the MLE  $\hat{\boldsymbol{\theta}}$ , and use the equivalent form of the observed information matrix to evaluate standard errors for these estimates.

Alternatively in (3), instead of sampling every  $k$ ,  $R$  samplers can be initiated at different values, discarding an initial warm-up at each. ■

The Gibbs Sampler algorithm (Geman & Geman, 1984; Gelfand & Smith, 1990) and the Metropolis algorithm (Metropolis et al., 1953) are special cases of the Hastings algorithm (Hastings, 1970; Peskun, 1973). An explanation of how these algorithms work and of the differences between them follows.

## Hastings' Algorithm

Suppose we want to generate a discrete or continuous random variable (scalar or vector)  $Y$  which takes values in a space  $\tau$  with density  $\pi(t)$ . We assume that  $\pi(\cdot)$  is completely specified up to a normalizing constant. We also select a Markov transition kernel  $q(s, t)$ ,  $s, t \in \tau$ , which is used to generate trial values of  $Y$ . In the discrete case, the interpretation is that if the current values of  $Y$  is  $s$ , the next trial value is  $t$  with probability  $q(s, t)$ . The choice of  $q(\cdot, \cdot)$  is almost arbitrary and the performance of the algorithm may vary according to this choice. The algorithm follows.

**Step 0:** Take an arbitrary starting value  $Y = t_0$  and set  $i = 0$ .

**Step 1:** Given  $Y = t_i = s$ , choose a new trial value according to the probability distribution  $q(s, t)$ ,  $s, t \in \tau$ , where  $\tau = \{0, 1\}$  and

$$q(s, t) = \Pr(Y_1 = t \mid Y_0 = t_0 = s)$$

is an arbitrary Markov kernel (generating an irreducible and aperiodic chain).

**Step 2:** Calculate

$$\alpha(s, t) = \min \left\{ \frac{\pi(t) q(t, s)}{\pi(s) q(s, t)}, 1 \right\},$$

where  $\pi(t)$  is a distribution specified up to a normalizing constant. Here  $\pi(\cdot) = F(\cdot; \theta_0)$ , where  $F$  is given by (2.1.11).

**Step 3:** Generate a random variable

$$U = \begin{cases} 1 & \text{with probability } \alpha(s, t) \\ 0 & \text{with probability } 1 - \alpha(s, t) \end{cases}$$

**Step 4:** If  $U = 1$ , move to  $t$ , i.e.,  $t_{i+1} = t$ . Otherwise  $t_{i+1} = t_i$ .

**Step 5:** Set  $i := i + 1$ . ■

## Metropolis' Algorithm

$q$  is taken to be a symmetric random walk, i.e.,  $q(s, t) = q(t, s)$ . Then, in step 2:

$$\alpha(s, t) = \min \left\{ \frac{\pi(t)}{\pi(s)}, 1 \right\}$$
■

## The Gibbs Sampler Algorithm

For the Gibbs Sampler, each of the updating steps corresponds to generating a new value from a particular conditional distribution, i.e., we always accept the new value. Then, in step 2:  $\alpha(s, t) = 1$ . In our specific case, we are generating  $m$  independent sequences of length  $n$  and the Gibbs sampler proceeds as follows:

**Step 0:** Take arbitrary starting values, say  $(y_1, y_2, \dots, y_n)$ . Define  $y_j^{(0)} = y_j$ ,  $j = 1, \dots, n$ . Let  $i = 0$ .

**Step 1:** Given the current values of  $y_1^{(i)}, y_2^{(i)}, \dots, y_n^{(i)}$ , generate a new pseudo-random value from the conditional distribution of  $Y_1$  given  $Y_j = y_j^{(i)}$ ,  $j = 2, \dots, n$ , and call it  $y_1^{(i+1)}$ . This probability distribution is given in equation (2.1.9), i.e.,

$$P(y^{(i)} | \{y^{(j)} : j \neq i\}) = \frac{\exp(\alpha_i y^{(i)} + \sum_{j=1}^n \gamma_{ij} y^{(i)} y^{(j)})}{1 + \exp(\alpha_i + \sum_{j=1}^n \gamma_{ij} y^{(j)})}$$

**Step 2:** Given the current values of  $y_1^{(i+1)}, y_3^{(i)}, \dots, y_n^{(i)}$ , generate a pseudo-random value from the conditional distribution of  $Y_2$  given  $Y_1 = y_1^{(i+1)}, Y_j = y_j^{(i)}$ ,  $j = 3, \dots, n$ , and call it  $y_2^{(i+1)}$ .

Continue updating one component at a time until...

**Step n:** Given the current values of  $y_1^{(i+1)}, y_2^{(i+1)}, \dots, y_{n-1}^{(i+1)}$ , generate a pseudo-random value from the conditional distribution of  $Y_n$  given  $Y_j = y_j^{(i+1)}$ ,  $j = 1, \dots, n-1$ , and call it  $y_n^{(i+1)}$ .

**Step n+1:** Set  $i := i + 1$  and return to Step 1.

A single cycle through steps 1 to n+1 completes one iteration of the algorithm. We will discard an initial warm-up sample, then generate  $R$  random vectors by sampling the chain every  $k$  cycles. ■

## 2.2 Model based on the Bahadur Representation

Let  $\mathbf{y}_k = (y_k(1) \dots y_k(n))$  be the  $1 \times n$  vector representing the binary responses for the  $n$  sites along sequence  $k$ , i.e., whether there is a mutation or not at each site along the sequence. Since each  $y_k(i)$  ( $i = 1, \dots, n$ ) assumes the value 1 or 0, the

random variable  $Y_k(i)$  is a Bernoulli trial with parameter  $\xi_i = \Pr\{Y_k(i) = 1\}$ . Then,  $E[Y_k(i)] = \xi_i$  and  $\text{Var}[Y_k(i)] = \xi_i(1 - \xi_i)$ .

Let

$$Z_k(i) = \frac{Y_k(i) - \xi_i}{\sqrt{\xi_i(1 - \xi_i)}}$$

and

$$r_{ij} = E[Z_k(i) Z_k(j)], r_{ijl} = E[Z_k(i) Z_k(j) Z_k(l)], \dots, r_{12\dots n} = E[Z_k(1) Z_k(2) \dots Z_k(n)]$$

So,  $r_{ij}$  are second-order correlations,  $r_{ijl}$  third-order correlations, and so on. Hence, there are  $\binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - n - 1$  correlation parameters.

Denote by  $\mathbf{P}_{[1]k}$  the joint distribution of  $y_k(i)$ 's when they are independent, i.e.,

$$\mathbf{P}_{[1]k}(y_k(1), \dots, y_k(n)) = \prod_{i=1}^n \xi_i^{y_k(i)} (1 - \xi_i)^{1 - y_k(i)} \quad (2.2.1)$$

Let  $\mathbf{P}(\mathbf{y}_k)$  be the distribution of  $\mathbf{Y}_k$ , where  $\mathbf{Y}_k = (Y_k(1) \dots Y_k(n))$  is a  $1 \times n$  vector denoting the response random vector for the  $k$ -th sequence.

**Proposition** (Bahadur, 1961)

For every  $\mathbf{y}_k = (y_k(1), \dots, y_k(n))$ ,

$$\mathbf{P}(\mathbf{y}_k) = \mathbf{P}_{[1]k}(\mathbf{y}_k) f(\mathbf{y}_k), \quad (2.2.2)$$

where

$$f(\mathbf{y}_k) = 1 + \sum_{i < j} r_{ij} z_k(i) z_k(j) + \dots + r_{12\dots n} z_k(1) z_k(2) \dots z_k(n) \quad (2.2.3)$$

■

If we assume pairwise dependence only, the distribution of  $\mathbf{Y}_k$  is

$$\begin{aligned} \mathbf{P}(\mathbf{y}_k) &= \mathbf{P}_{[1]k}(\mathbf{y}_k) \left[ 1 + \sum_{i < j} r_{ij} z_k(i) z_k(j) \right] \\ &= \prod_{i=1}^n \xi_i^{y_k(i)} (1 - \xi_i)^{1 - y_k(i)} \left[ 1 + \sum_{i < j} r_{ij} z_k(i) z_k(j) \right] \\ &= \prod_{i=1}^n \xi_i^{y_k(i)} (1 - \xi_i)^{1 - y_k(i)} \\ &\quad \times \left[ 1 + \sum_{i < j} r_{ij} \left( \frac{y_k(i) - \xi_i}{\sqrt{\xi_i(1 - \xi_i)}} \right) \left( \frac{y_k(j) - \xi_j}{\sqrt{\xi_j(1 - \xi_j)}} \right) \right] \end{aligned} \quad (2.2.4)$$



## 2.2.1 Estimation Procedure

Let  $\theta = (\xi_1, \dots, \xi_n, r_{12}, \dots, r_{1n}, r_{23}, \dots, r_{2n}, \dots, r_{n-1n})'$ . An estimate of  $\theta$  can be obtained by maximum likelihood. The likelihood function for sequence  $k$  is given by (2.2.4). For  $m$  independent sequences, the likelihood function is

$$\begin{aligned} L(\theta; \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m) &= \prod_{k=1}^m \prod_{i=1}^n \xi_i^{y_k(i)} (1 - \xi_i)^{1-y_k(i)} \\ &\times \left[ 1 + \sum_{i < j} r_{ij} \left( \frac{y_k(i) - \xi_i}{\sqrt{\xi_i(1 - \xi_i)}} \right) \left( \frac{y_k(j) - \xi_j}{\sqrt{\xi_j(1 - \xi_j)}} \right) \right] \end{aligned} \quad (2.2.5)$$

## 2.3 Models for Three Categories

In this section, we extend the number of categories of interest, investigating not only if there is a mutation, but also the type of mutation. The events can fall into one of three categories: transitions ( $A \leftrightarrow G$  or  $C \leftrightarrow T$ ), transversions ( $A \leftrightarrow C$ ,  $A \leftrightarrow T$ ,  $G \leftrightarrow C$ , or  $G \leftrightarrow T$ ) or no mutation. An extension of the autologistic model is proposed by creating dummy variables to represent the categories (Section 2.3.2).

### 2.3.1 Overview of the Literature

Little has been done to extend the classical autologistic model. Strauss (1977) thought of a black-and-white picture as a binary lattice and generalized his analysis for a multicolored representation. With a set of  $n$  sites, associate with site  $i$  a random variable  $y_i$  determining its color. The 'color'  $y_i = 0$  is available at each site. Therefore  $\Pr(\mathbf{0}) > 0$  and as in equation (2.1.2) we can define:

$$Q(\mathbf{y}) = \log\{\Pr(\mathbf{y})/\Pr(\mathbf{0})\}, \quad (2.3.1)$$

where  $\mathbf{y} = (y_1, \dots, y_n)$ . Following equation (2.1.4) he expands  $Q(\mathbf{y})$  as

$$Q(\mathbf{y}) = \sum_{1 \leq i \leq n} y_i G_i(y_i) + \sum_{1 \leq i < j \leq n} y_i y_j G_{ij}(y_i, y_j) + \dots + y_1 \dots y_n G_{1\dots n}(y_1, \dots, y_n) \quad (2.3.2)$$

Consider now the special case where the sites are arranged in a regular lattice, and each site has one of  $c + 1$  colors, denoted by  $y_i = 0, 1, \dots, c$ . Two assumptions are needed to obtain a workable statistical model.

(a) *Homogeneity*

If  $\mathbf{S}$  is a set of sites, then  $\Pr\{Y_i = y_i, Y_j = y_j, \dots, Y_s = y_s : (i, j, \dots, s) \in \mathbf{S}\}$  should be the same for any set of sites derived from  $\mathbf{S}$  by translation, rotation or reflexion. This condition requires that all subscripts be dropped from the functions  $G$ .

(b) *Pairwise dependence only*

With assumptions (a) and (b), we have

$$Q(\mathbf{y}) = \sum_{1 \leq i \leq n} y_i G(y_i) + \sum_{1 \leq i < j \leq n} y_i y_j G(y_i, y_j).$$

For the multicolored case, it is more convenient to write

$$u_r = r G(r) \quad \text{and} \quad v_{rs} = r s G(r, s).$$

Then

$$Q(\mathbf{y}) = \sum_{r=1}^c m_r u_r + \sum_{r=1}^c \sum_{s=1}^c n_{rs} v_{rs} \tag{2.3.3}$$

where  $m_r$  is the number of sites with color  $r$  and  $n_{rs}$  is the number of pairs of neighboring sites where one has color  $r$  and the other color  $s$ , and

$$\Pr(\mathbf{y}) \propto \exp\left\{\sum_{r=1}^c m_r u_r + \sum_{r=1}^c \sum_{s=1}^c n_{rs} v_{rs}\right\}$$

Equation (2.3.3) usually involves too many parameters, and in analyzing actual data one hopes that the following assumption is valid.

(c) *Color indifference*

Suppose that for any pair of neighboring sites  $i$  and  $j$

$$Q(0, \dots, 0, y_i, 0, \dots, 0, y_j, 0, \dots, 0) - Q(0, \dots, 0, y_j, 0, \dots, 0)$$

is independent of  $y_j$ , for  $y_j \neq y_i$ . This means that the affinity between any two colors is the same as that between any other two, which may be appropriate when the colors

are unordered. It follows from (2.3.3) that  $v_{rs}$  is constant for all  $r, s$  ( $r \neq s$ ). Since  $\sum_{r \neq s} n_{rs} + \sum_r n_{rr}$  is fixed, for a given lattice, reparametrize and take  $v_{rs} = 0$  for  $r \neq s$ . Thus, with assumptions (a)-(c), the model becomes

$$\Pr(\mathbf{y}) \propto \exp\left(\sum_{r=1}^c m_r u_r + \sum_{s=1}^c n_s v_s\right), \quad (2.3.4)$$

where the subscripts of the  $n$ 's and  $v$ 's have been dropped, and here  $n_s$  represents the number of pair of sites with the same color. It is possible to express (2.3.4) in a conditional form: for any internal site  $y_{rs}$

$$\Pr(y_{rs} = j \mid \text{other sites}) = \frac{\exp(u_j + v_j n_{jrs})}{1 + \sum_{i=1}^c \exp(u_i + v_i n_{irs})} \quad (2.3.5)$$

where  $n_{jrs}$  is the number of sites colored  $j$  with neighboring sites  $(r, s)$ .

Because of limitations in the estimation procedure, the model is simplified even further and reduces to a single parameter  $v$  or  $u$ . When conditioning on the observed  $m_r$ 's, the  $u_r$ 's become nuisance parameters and  $v_r$  is the focus of attention, where  $v_r$  measures the "clustering tendency". Then

$$\Pr(\mathbf{y}; v_1, \dots, v_c) \propto \exp\left(\sum_{s=1}^c n_s v_s\right) \quad (2.3.6)$$

Assuming that the strength of attraction is the same for all colors; i.e.,  $v_1 = v_2 = \dots = v_c = v$ , (2.3.6) becomes

$$\Pr(\mathbf{y}; v) \propto \exp(v y), \quad (2.3.7)$$

where  $y$  is the number of adjacencies of like color. When conditioning on the observed  $n_s$ 's, the  $v_s$ 's become the nuisance parameters and  $u_r$  is the "coloring tendency". Then,

$$\Pr(\mathbf{y}; u_1, \dots, u_c) \propto \exp\left(\sum_{r=1}^c m_r u_r\right) \quad (2.3.8)$$

If we assume that the "coloring tendency" is the same for all colors, i.e.,  $u_1 = u_2 = \dots = u_c = u$ , (2.3.8) becomes

$$\Pr(\mathbf{y}; u) \propto \exp(u y), \quad (2.3.9)$$

where  $y$  is the number of adjacencies with the same color.

A  $\chi^2$ -approximation to the distribution of  $\mathbf{Y}$  is obtained after calculating the first two cumulants of  $y$ . The MLE for  $v$  or  $u$  is calculated from this distribution. The resulting models are so simplistic that they are of little practical use.

## 2.3.2 An Alternative Extension of the Autologistic Model

Let  $\mathbf{y}_k(i) = (y_{1k}(i) \ y_{2k}(i))'$ , where

$$y_{1k}(i) = \begin{cases} 1 & \text{if site } i \text{ of sequence } k \text{ is in category 1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$y_{2k}(i) = \begin{cases} 1 & \text{if site } i \text{ of sequence } k \text{ is in category 2} \\ 0 & \text{otherwise} \end{cases}$$

Note that if  $y_{1k}(i) = 0$  and  $y_{2k}(i) = 0$  there is no mutation at site  $i$  of sequence  $k$ .

Thus, we have  $\mathbf{y}_k(i) = (0 \ 0)'$  if there is no mutation at site  $i$  of sequence  $k$ ,  $\mathbf{y}_k(i) = (1 \ 0)'$  if the substitution falls in category 1 and  $\mathbf{y}_k(i) = (0 \ 1)'$  if the mutation falls in category 2.

Let  $\mathbf{G}_i(\mathbf{y}_k(i)) = (G_i^{(1)}(y_{1k}(i)) \ G_i^{(2)}(y_{2k}(i)))'$ .

Then,

$$\begin{aligned} Q(\mathbf{y}) &= \sum_{i=1}^n \mathbf{y}'_k(i) \mathbf{G}_i(\mathbf{y}_k(i)) + \sum_{1 \leq i < j \leq n} y_{1k}(i) y_{1k}(j) G_{i,j}^{(1,1)}(y_{1k}(i), y_{1k}(j)) \\ &\quad + \sum_{1 \leq i < j \leq n} y_{2k}(i) y_{2k}(j) G_{i,j}^{(2,2)}(y_{2k}(i), y_{2k}(j)) \\ &\quad + \sum_{1 \leq i < j \leq n} y_{1k}(i) y_{2k}(j) G_{i,j}^{(1,2)}(y_{1k}(i), y_{2k}(j)) \\ &= \sum_{i=1}^n [y_{1k}(i) G_i^{(1)}(y_{1k}(i)) + y_{2k}(i) G_i^{(2)}(y_{2k}(i))] \\ &\quad + \sum_{1 \leq i < j \leq n} y_{1k}(i) y_{1k}(j) G_{i,j}^{(1,1)}(y_{1k}(i), y_{1k}(j)) \\ &\quad + \sum_{1 \leq i < j \leq n} y_{2k}(i) y_{2k}(j) G_{i,j}^{(2,2)}(y_{2k}(i), y_{2k}(j)) \\ &\quad + \sum_{1 \leq i < j \leq n} y_{1k}(i) y_{2k}(j) G_{i,j}^{(1,2)}(y_{1k}(i), y_{2k}(j)) \end{aligned}$$

Assuming homogeneity, all subscripts can be dropped from the functions  $G$ , which yields

$$\begin{aligned} Q(\mathbf{y}) &= \sum_{i=1}^n [y_{1k}(i) G^{(1)}(y_{1k}(i)) + y_{2k}(i) G^{(2)}(y_{2k}(i))] \\ &\quad + \sum_{1 \leq i < j \leq n} y_{1k}(i) y_{1k}(j) G^{(1,1)}(y_{1k}(i), y_{1k}(j)) \end{aligned}$$

$$\begin{aligned}
& + \sum_{1 \leq i < j \leq n} y_{2k}(i) y_{2k}(j) G^{(2,2)}(y_{2k}(i), y_{2k}(j)) \\
& + \sum_{1 \leq i < j \leq n} y_{1k}(i) y_{2k}(j) G^{(1,2)}(y_{1k}(i), y_{2k}(j)) \\
& = m_{1k} \alpha_1 + m_{2k} \alpha_2 + n_{11k} \gamma_{11} + n_{22k} \gamma_{22} + n_{12k} \gamma_{12}
\end{aligned}$$

where  $G^{(1)} = \alpha_1$ ,  $G^{(2)} = \alpha_2$ ,  $G^{(1,1)} = \gamma_{11}$ ,  $G^{(2,2)} = \gamma_{22}$ ,  $G^{(1,2)} = \gamma_{12}$ .  $m_{1k}$  and  $m_{2k}$  are the numbers of sites along sequence  $k$  in category 1 and category 2, respectively.  $n_{11k}$ ,  $n_{22k}$  and  $n_{12k}$  are the numbers of pairs of sites from sequence  $k$  with both in category 1, with both in category 2, and with one in category 1 and the other in category 2, respectively.

If we do not assume homogeneity among sites,

$$\begin{aligned}
Q(\mathbf{y}_k) - Q(\mathbf{y}_{ki}) & = y_{1k}(i) \alpha_{1i} + y_{2k}(i) \alpha_{2i} \\
& + \sum_{j=1}^n [y_{1k}(i) y_{1k}(j) \gamma_{ij}^{(1,1)} + y_{2k}(i) y_{2k}(j) \gamma_{ij}^{(2,2)} + y_{1k}(i) y_{2k}(j) \gamma_{ij}^{(1,2)}]
\end{aligned}$$

where  $G_i^{(1)} = \alpha_{1i}$ ,  $G_i^{(2)} = \alpha_{2i}$ ,  $G_{ij}^{(1,1)} = \gamma_{ij}^{(1,1)}$ ,  $G_{ij}^{(2,2)} = \gamma_{ij}^{(2,2)}$ ,  $G_{ij}^{(1,2)} = \gamma_{ij}^{(1,2)}$

Therefore,

$$\begin{aligned}
& \frac{\Pr(\mathbf{y}_k(i) \mid \{\mathbf{y}_k(j) : j \neq i\})}{\Pr(\mathbf{0}_k(i) \mid \{\mathbf{y}_k(j) : j \neq i\})} \\
& = \exp\{Q(\mathbf{y}) - Q(\mathbf{y}_i)\} \\
& = \exp\{\mathbf{y}'(i) \boldsymbol{\alpha} + \sum_{j=1}^n (y_{1k}(i) y_{1k}(j) \quad y_{2k}(i) y_{2k}(j) \quad y_{1k}(i) y_{2k}(j)) \boldsymbol{\gamma}\}
\end{aligned}$$

where  $\boldsymbol{\alpha} = (\alpha_{1i} \quad \alpha_{2i})'$ ,  $\mathbf{y}_k(i) = (y_{1k}(i) \quad y_{2k}(i))'$ ,  $\boldsymbol{\gamma} = (\gamma_{ij}^{(1,1)} \quad \gamma_{ij}^{(2,2)} \quad \gamma_{ij}^{(1,2)})'$  and  $\mathbf{0}(i) = (0 \quad 0)'$ .

So,

$$\Pr(\mathbf{y}_k(i) \mid \{\mathbf{y}_k(j) : j \neq i\}) = \frac{\exp(\mathbf{B}' \mathbf{y}_k(i))}{1 + \exp(B_1) + \exp(B_2)} \quad (2.3.10)$$

where  $\mathbf{y}_k(i) \in \{(0 \ 0)', (1 \ 0)', (0 \ 1)'\}$  and  $\mathbf{B} = (B_1 \ B_2)'$  with

$$B_1 = \alpha_{1i} + \sum_{j=1}^n y_{1k}(j) \gamma_{ij}^{(1,1)} + \sum_{j=1}^n y_{2k}(j) \gamma_{ij}^{(1,2)}$$

and

$$B_2 = \alpha_{2i} + \sum_{j=1}^n y_{2k}(j) \gamma_{ij}^{(2,2)} + \sum_{j=1}^n y_{1k}(j) \gamma_{ij}^{(1,2)}.$$

## Estimation Procedure

The estimation procedure in the three-category case is analogous to that of the classical autologistic model (Section 2.2). Only now, we have to generate vectors on the set  $\{(0\ 0)', (1\ 0)', (0\ 1)'\}$ . Recall that the vector  $(0\ 0)'$  represents no mutation,  $(1\ 0)'$  category 1 and  $(0\ 1)'$  category 2. These vectors are generated from the conditional distribution (2.3.10), with initial estimates obtained by maximization of the pseudo-likelihood function.

## Chapter 3

# Analyzing the Variability in DNA Sequences

The focus here is the comparison of sets of sequences. For example, we are interested in comparing DNA sequences from the human immunodeficiency virus (HIV) from different geographical areas to see whether the variability is similar in each. Similarly, when we study several individuals and obtain a set of sequences from each individual at different time points, our interest lies in estimating the variability between and within individuals.

Simpson (1949) proposed a measure of diversity for categorical data in terms of frequencies for each category. On the basis of a similar measure of variation, Light & Margolin (1971) developed an analysis of variance (CATANOVA), for one-way tables, suitable for categorical variables. The properties of the components of variation are investigated under a common multinomial model. This framework can be used to compare the variability of the response variable at a single position between and within groups. We consider a number of sites because, in the context of interest (the analysis of HIV-1 sequences), a single position yields little information. Components of variation are derived from the fact that the sum of squares of deviations from the mean can be expressed as a function of the squares of the pairwise differences for all possible pairs (Section 3.1). The sequences are not considered on an individual basis but only as contributing to the overall variability in the distribution of the categorical

response. We partition the measures of diversity according to the factors considered (Section 3.2) assuming independence among positions (Section 3.3). We develop a test statistic for the null hypothesis of homogeneity among groups (Sections 3.4 and 3.5) and assess its power (Section 3.6).

### 3.1 Variation in Categorical Data

For categorical data, the mean is an ill-defined concept. Therefore, measures of variation, such as the variance, which are meaningful for continuous variables, no longer apply. Gini (1912) found an alternative way of characterizing variation and developed a measure of variation for categorical data.

Let  $X_1, X_2, \dots, X_N$  denote measurements of  $N$  independent experimental units. The variance of  $X$  may be expressed as  $E\phi(X_1, X_2)$ , where  $\phi(a, b) = \frac{1}{2}(a - b)^2$  (Hoeffding, 1948). In a similar fashion, the sum of squares is

$$\begin{aligned} SS &= \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N (X_i - X_j)^2 \\ &= \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^2 = \frac{1}{N} \sum_{1 \leq i < j \leq N} d_{ij}^2 \end{aligned} \quad (3.1.1)$$

where  $\bar{X} = \sum_{i=1}^N X_i / N$  and  $d_{ij} = X_i - X_j$ .

In the present context, each  $X_i$  falls into one of  $C$  possible categories. Define  $d(X_i, X_j) = d_{ij}$  as

$$d_{ij} = \begin{cases} 1 & \text{if } X_i \text{ and } X_j \text{ belong to different categories} \\ 0 & \text{if } X_i \text{ and } X_j \text{ belong to the same category.} \end{cases} \quad (3.1.2)$$

#### Definition 3.1

The variation for categorical responses  $X_1, \dots, X_N$  is

$$D_N = \frac{1}{2N} \sum_{j=1}^N \sum_{i=1}^N d_{ij}^2 = \frac{1}{2N} \sum_{j=1}^N \sum_{i=1}^N d_{ij} \quad (3.1.3)$$

where  $d_{ij}$  is defined in (3.1.2). ■



As each response assumes one and only one of the  $C$  possible categories, the data is summarized by the vector  $\Phi = (n_1, \dots, n_C)$  where  $n_i$  is the number of responses in the  $i$ th category ( $i = 1, \dots, C$ ), so that  $\sum_{i=1}^C n_i = N$ . Then the variation in the responses is defined as

$$\begin{aligned} D_N &= \frac{1}{2N} \sum_{i \neq j} n_i n_j = \frac{1}{2N} \left[ N^2 - \sum_{i=1}^C n_i^2 \right] = \frac{N}{2} - \frac{1}{2N} \sum_{i=1}^C n_i^2 \\ &= \frac{N}{2} \left\{ 1 - \sum_{i=1}^C \left( \frac{n_i}{N} \right)^2 \right\}. \end{aligned} \quad (3.1.4)$$

If  $p_1, \dots, p_C$  stand for the probabilities of  $X$  belonging to these  $C$  categories, the Simpson's Index of ecological diversity is defined as

$$I_S(\mathbf{p}) = 1 - \mathbf{p}'\mathbf{p} = 1 - \sum_{i=1}^C p_i^2 \quad (3.1.5)$$

and its corresponding sample counterpart is

$$\hat{I}_S(\mathbf{p}) = 1 - \hat{\mathbf{p}}'\hat{\mathbf{p}} = 1 - \sum_{i=1}^C \hat{p}_i^2, \quad (3.1.6)$$

where  $\hat{p}_i = n_i/N$ ,  $i = 1, \dots, C$  relate to the sample proportion. Therefore, we have

$$D_N = \frac{N}{2} \hat{I}_S(\mathbf{p})$$

The definitions (3.1.4) and (3.1.5) are motivated by two properties.

1. The variation of  $N$  categorical responses is minimized if and only if they all belong to the same category, i.e.,  $p_i = 1$ ,  $\forall i = 1, \dots, C$ .
2. The variation of  $N$  responses is maximized when the responses are distributed among the available categories as evenly as possible, i.e.,  $p_i = 1/C$ ,  $\forall i = 1, \dots, C$ .

## 3.2 Partitioning the Measures of Diversity

Let  $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \dots, X_{iK}^g)'$  be a random vector representing sequence  $i$  of group  $g$ . Suppose  $i = 1, \dots, N$ ,  $k = 1, \dots, K$  and  $g = 1, \dots, G$ . So,  $X_{ik}^g$  represents

position  $k$  of sequence  $i$  of group  $g$ .  $X_{ik}^g$  is a categorical variable assuming  $C$  (un-ordered) categories. For instance, if comparisons are made at the nucleotide level,  $x_{ik}^g \in \{\mathbf{A}, \mathbf{C}, \mathbf{T}, \mathbf{G}\}$  and there are 4 categories.

First, assume there is only one position for each sequence. We summarize the data in Table 3.1.

Table 3.1: Summary of the Data (one position)

Sequence	Group				
	1	2	3	...	G
1	$x_1^1$	$x_1^2$	$x_1^3$	...	$x_1^G$
2	$x_2^1$	$x_2^2$	$x_2^3$	...	$x_2^G$
3	$x_3^1$	$x_3^2$	$x_3^3$	...	$x_3^G$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$x_N^1$	$x_N^2$	$x_N^3$	...	$x_N^G$

Now  $d_{ij}$  is defined as

$$d_{ij} = \begin{cases} 1 & \text{if } X_i^g \neq X_j^g \\ 0 & \text{if } X_i^g = X_j^g \end{cases}$$

The total number of responses is

$$NG = \sum_{g=1}^G n_{.g} = \sum_{c=1}^C n_{c.} = \sum_{c=1}^C \sum_{g=1}^G n_{cg}$$

where  $n_{cg}$  is the number of responses in category  $c$  for group  $g$  and  $N = n_{.g} = \sum_{c=1}^C n_{cg}$  is the number of responses for group  $g$ , which here is simply the number of sequences in each group. The Total Simpson Index (TSI) is

$$TSI = 1 - \sum_{c=1}^C \left( \frac{n_{c.}}{NG} \right)^2 \quad (3.2.1)$$

The dispersion within group  $g$  (i.e., within  $x_1^g, x_2^g, \dots, x_N^g$ ) is

$$1 - \sum_{c=1}^C \left( \frac{n_{cg}}{n_{.g}} \right)^2 \quad (3.2.2)$$

Therefore, the within-group Simpson Index (WSI) is found by averaging (3.2.2) over all  $g$ 's:

$$WSI = \frac{1}{G} \sum_{g=1}^G \left\{ 1 - \sum_{c=1}^C \left( \frac{n_{cg}}{n_{\cdot g}} \right)^2 \right\} = 1 - G \sum_{g=1}^G \sum_{c=1}^C \left( \frac{n_{cg}}{NG} \right)^2 \quad (3.2.3)$$

The between-group Simpson Index (BSI) is

$$BSI = TSI - WSI = G \sum_{g=1}^G \sum_{c=1}^C \left( \frac{n_{cg}}{NG} \right)^2 - \sum_{c=1}^C \left( \frac{n_{c\cdot}}{NG} \right)^2 \quad (3.2.4)$$

Now, assume there are  $K$  positions along each sequence. We have  $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \dots, X_{iK}^g)'$  and  $\mathbf{X}^g = (\mathbf{X}_1^g, \mathbf{X}_2^g, \dots, \mathbf{X}_N^g)'$ . The data are summarized in Table 3.2.

The total number of responses is

$$NGK = \sum_{g=1}^G n_{\cdot g} = \sum_{c=1}^C n_{c\cdot} = \sum_{k=1}^K n_{\cdot k} = \sum_{c=1}^C \sum_{g=1}^G \sum_{k=1}^K n_{cgk}$$

The interest is in assessing the homogeneity among groups: the null hypothesis is that  $p_{cgk} = p_{ck}$  where  $p_{cgk}$  is the population probability of belonging to category  $c$  in group  $g$  at position  $k$ . One could argue that this is the classical Pearson's  $\chi^2$  test, but note that the classical Pearson's  $\chi^2$  statistic for Table 3.2 is

$$\begin{aligned} \chi_P^2 &= \sum_{g=1}^G \sum_{c=1}^C \sum_{k=1}^K \frac{\left( \frac{n_{cgk}}{N} - \frac{n_{c\cdot k}}{NG} \right)^2}{n_{c\cdot k}/NG} \\ &= \sum_{g=1}^G \sum_{c=1}^C \sum_{k=1}^K \frac{G \left( n_{cgk} - \frac{n_{c\cdot k}}{G} \right)^2}{N n_{c\cdot k}} \end{aligned}$$

with  $K(G-1)(C-1)$  degrees of freedom. The limiting  $\chi^2$ -distribution is a close approximation only when the cell frequencies  $n_{cgk}$ 's are all large (at least 5). In analyzing amino-acid sequences, we know that these conditions are not met. The distribution at a single position usually exhibits a few polymorphisms with very low frequencies. Just as for Fisher's exact test, the exact null distribution is difficult to implement for small values of  $N$ , when  $G$  or  $K$  is not small. Moreover, if the number of degrees of freedom of the  $\chi^2$ -statistic is large but the noncentrality parameter is not proportionally so, the resulting test is likely to have less power than some tests

Table 3.2: Contingency Table (K positions)

Group	Position	1	2	...	C	Total
1	1	$n_{111}$	$n_{211}$	...	$n_{C11}$	$n_{.11} = N$
1	2	$n_{112}$	$n_{212}$	...	$n_{C12}$	$n_{.12} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
1	K	$n_{11K}$	$n_{21K}$	...	$n_{C1K}$	$n_{.1K} = N$
Total		$n_{11.}$	$n_{21.}$	...	$n_{C1.}$	$n_{.1.} = NK$
2	1	$n_{121}$	$n_{221}$	...	$n_{C21}$	$n_{.21} = N$
2	2	$n_{122}$	$n_{222}$	...	$n_{C22}$	$n_{.22} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
2	K	$n_{12K}$	$n_{22K}$	...	$n_{C2K}$	$n_{.2K} = N$
Total		$n_{12.}$	$n_{22.}$	...	$n_{C2.}$	$n_{.2.} = NK$
...	...	...	...	...	...	...
G	1	$n_{1G1}$	$n_{2G1}$	...	$n_{CG1}$	$n_{.G1} = N$
G	2	$n_{1G2}$	$n_{2G2}$	...	$n_{CG2}$	$n_{.G2} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
G	K	$n_{1GK}$	$n_{2GK}$	...	$n_{CGK}$	$n_{.GK} = N$
Total		$n_{1G.}$	$n_{2G.}$	...	$n_{CG.}$	$n_{.G.} = NK$
TOTAL		$n_{1..}$	$n_{2..}$	...	$n_{C..}$	$n_{...} = NGK$

directed towards specific alternatives. Note that the number of degrees of freedom here is usually large: for instance, comparing two groups of sequences 100 nucleotide long yields 300 degrees of freedom. For these reasons we use another approach to assess homogeneity among groups.

The variation within the  $g$ th group at the  $k$ th position is

$$1 - \sum_{c=1}^C \left( \frac{n_{cgk}}{n_{.gk}} \right)^2 = 1 - \sum_{c=1}^C \left( \frac{n_{cgk}}{N} \right)^2,$$

since  $n_{.gk} = N$ . The variation within the  $g$ th group is

$$1 - \sum_{c=1}^C \left( \frac{n_{cg.}}{n_{.g.}} \right)^2 = 1 - \sum_{c=1}^C \left( \frac{n_{cg.}}{NK} \right)^2$$

since  $n_{.g.} = NK$ . The measures of dispersions are

$$WSI = \frac{1}{G} \sum_{g=1}^G \left\{ 1 - \sum_{c=1}^C \left( \frac{n_{cg.}}{NK} \right)^2 \right\} = 1 - G \sum_{c=1}^C \left( \frac{n_{c.}}{NGK} \right)^2 \quad (3.2.5)$$

$$TSI = 1 - \sum_{c=1}^C \left( \frac{n_{c.}}{NGK} \right)^2, \quad (3.2.6)$$

$$\text{and } BSI = TSI - WSI = G \sum_{g=1}^G \sum_{c=1}^C \left( \frac{n_{cg.}}{NGK} \right)^2 - \sum_{c=1}^C \left( \frac{n_{c.}}{NGK} \right)^2. \quad (3.2.7)$$

### 3.3 The Probabilistic Model

Assuming that responses in different groups are independent, for each group and each position, the responses  $(n_{1gk}, n_{2gk}, \dots, n_{Cgk})$  follow a multinomial distribution:

$$\Pr\{n_{1gk}, n_{2gk}, \dots, n_{Cgk}\} = \binom{N}{n_{1gk} \dots n_{Cgk}} \prod_{c=1}^C (p_{cgk})^{n_{cgk}},$$

where  $\sum_{c=1}^C p_{cgk} = 1$ ,  $p_{cgk} > 0$ ,  $c = 1, \dots, C$ ,  $k = 1, \dots, K$  and  $g = 1, \dots, G$

$$E(n_{cgk}) = Np_{cgk} \quad \text{Var}(n_{cgk}) = Np_{cgk}(1 - p_{cgk})$$

and  $\text{Cov}(n_{c_1g_1k_1}, n_{c_2g_2k_2}) = -\delta Np_{c_1g_1k_1} p_{c_2g_2k_2}$  where

$$\delta = \begin{cases} 1 & \text{if } g_1 = g_2 \text{ and } k_1 = k_2 \\ 0 & \text{otherwise} \end{cases}$$

$n_{cgk}$  denotes number of responses in category  $c$  at position  $k$  for group  $g$  and  $p_{cgk}$  the probability of being at category  $c$  at position  $k$  for group  $g$ .

If we assume that the positions are independent, the model is

$$\begin{aligned} & \prod_{k=1}^K \Pr\{(n_{11k}, \dots, n_{C1k}, n_{12k}, \dots, n_{C2k}, \dots, n_{1Gk}, \dots, n_{CGk})\} \\ &= \prod_{g=1}^G \prod_{k=1}^K \Pr\{(n_{1gk}, n_{2gk}, \dots, n_{Cgk})\} \\ &= \prod_{g=1}^G \prod_{k=1}^K \binom{N}{n_{1gk} \dots n_{Cgk}} \prod_{c=1}^C (p_{cgk})^{n_{cgk}} \end{aligned}$$

Then  $\mathbf{V}_g \equiv (n_{1g1} \dots n_{Cg1} \ n_{1g2} \dots n_{Cg2} \dots n_{1gK} \dots n_{CgK})'$  is a  $CK \times 1$  vector

and  $\mathbf{V} \equiv (\mathbf{V}_1 \ \mathbf{V}_2 \ \dots \ \mathbf{V}_G)'$  is a  $GCK \times 1$  vector.

$$\mathbf{E}(\mathbf{V}) \equiv \boldsymbol{\mu} \equiv N\boldsymbol{\mu}_o = N(p_{111} \dots p_{C11} \dots p_{1GK} \dots p_{CGK})' \quad (3.3.1)$$

Let  $\oplus$  denote the direct-sum operation, i.e., if  $\mathbf{C} = \mathbf{A} \oplus \mathbf{B}$ ,  $\mathbf{C} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$ .

Then

$$\begin{aligned} \text{Cov}(\mathbf{V}) \equiv \boldsymbol{\Sigma} &\equiv N\boldsymbol{\Sigma}^\circ \\ &= N(\boldsymbol{\Sigma}_{11} \oplus \boldsymbol{\Sigma}_{12} \oplus \dots \oplus \boldsymbol{\Sigma}_{1K} \oplus \boldsymbol{\Sigma}_{21} \oplus \dots \oplus \boldsymbol{\Sigma}_{2K} \oplus \dots \oplus \boldsymbol{\Sigma}_{GK}) \end{aligned} \quad (3.3.2)$$

where  $\boldsymbol{\Sigma}_{gk}$  is a  $C \times C$  matrix of the form

$$\boldsymbol{\Sigma}_{gk} = \mathbf{D}_{gk} - \boldsymbol{\mu}_{ogk}\boldsymbol{\mu}'_{ogk} \quad (3.3.3)$$

with  $\mathbf{D}_{gk}$  being a  $C \times C$  diagonal matrix with elements  $p_{1gk}, \dots, p_{Cgk}$  and

$$\boldsymbol{\mu}_{ogk} = (p_{1gk} \dots p_{Cgk})'$$

### 3.4 Moments of Diversity Measures

#### Definition 3.2

Let  $\mathbf{A} = (a_{ij})$  and  $\mathbf{B} = (b_{ij})$  be  $m \times n$  and  $p \times q$  matrices, respectively. Then the *Kronecker product*

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B})$$

is an  $mp \times nq$  matrix expressible as a partitioned matrix with  $a_{ij}\mathbf{B}$  as the  $(i, j)$ th partition,  $i = 1, \dots, m$  and  $j = 1, \dots, n$ .

■

Let

$$\mathbf{T} = \frac{1}{(NGK)^2}(\mathbf{U}_{KG} \otimes \mathbf{I}_C) = \frac{1}{(NGK)^2}\mathbf{T}^\circ \quad (3.4.1)$$

where  $\mathbf{U}_{KG}$  is a  $KG \times KG$  matrix of 1's,  $\mathbf{I}_C$  is the  $C \times C$  identity matrix and  $\mathbf{T}^\circ \equiv (NGK)^2\mathbf{T}$  is a  $CKG \times CKG$  matrix, having  $KG \times KG$  partitions with each partition being  $C \times C$  identity matrix. Let  $\mathbf{M}$  be a  $G \times G$  diagonal matrix with diagonal elements  $Gn_{.g}^2 (= G(NK)^2$  here), i.e.,  $\mathbf{M} = G(NK)^2\mathbf{I}_G$ . Then  $\mathbf{M}^{-1} = \frac{1}{G(NK)^2}\mathbf{I}_G$ .

$$\begin{aligned} \mathbf{W} &\equiv [(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C] \\ \mathbf{W} &= \frac{G}{(NGK)^2}[(\mathbf{I}_G \otimes \mathbf{U}_K) \otimes \mathbf{I}_C] = \frac{1}{G(NK)^2}\mathbf{W}^\circ \end{aligned} \quad (3.4.2)$$

Then

$$TSI = 1 - \mathbf{V}'\mathbf{T}\mathbf{V} \quad (3.4.3)$$

$$WSI = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} \quad (3.4.4)$$

Therefore,

$$\begin{aligned} BSI &= TSI - WSI = -\mathbf{V}'\mathbf{T}\mathbf{V} + \mathbf{V}'\mathbf{W}\mathbf{V} = \mathbf{V}'(-\mathbf{T} + \mathbf{W})\mathbf{V} \\ &= \mathbf{V}'\mathbf{B}\mathbf{V} \end{aligned} \quad (3.4.5)$$

where

$$\begin{aligned} \mathbf{B} &= -\mathbf{T} + \mathbf{W} = \frac{-1}{(NGK)^2}(\mathbf{U}_{KG} \otimes \mathbf{I}_C) + \frac{G}{(NGK)^2}[(\mathbf{I}_G \otimes \mathbf{U}_K) \otimes \mathbf{I}_C] \\ &= \frac{G}{(NGK)^2} \left[ (\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G}\mathbf{U}_{KG} \right] \otimes \mathbf{I}_C \equiv \frac{1}{G(NK)^2}\mathbf{B}^\circ \end{aligned} \quad (3.4.6)$$

Since

$$E(\mathbf{V}'\mathbf{T}\mathbf{V}) = \text{trace}(\mathbf{T}\Sigma) + \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu} ,$$

$$E(TSI) = 1 - \text{trace}(\mathbf{T}\Sigma) - \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu}$$

$$\begin{aligned}
&= 1 - \frac{1}{(NGK)^2} \sum_{c=1}^C \sum_{g=1}^G \sum_{k=1}^K N p_{cgk} (1 - p_{cgk}) \\
&\quad - \sum_{c=1}^C \frac{1}{(NGK)^2} \left( \sum_{g=1}^G \sum_{k=1}^K N p_{cgk} \right)^2 \\
&= 1 - \frac{GK}{N(GK)^2} + \frac{1}{N(GK)^2} \sum_{c=1}^C \left[ \sum_{g=1}^G \sum_{k=1}^K p_{cgk}^2 - N p_{c\cdot}^2 \right] \\
&= 1 - \frac{1}{NGK} + \frac{1}{N(GK)^2} \sum_{c=1}^C \left[ \sum_{g=1}^G \sum_{k=1}^K p_{cgk}^2 - N p_{c\cdot}^2 \right]
\end{aligned}$$

$$\begin{aligned}
E(WSI) &= 1 - \text{trace}(\mathbf{W}\Sigma) - \boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu} \\
&= 1 - \frac{G}{(NGK)^2} \sum_{c=1}^C \sum_{g=1}^G \sum_{k=1}^K N p_{cgk} (1 - p_{cgk}) \\
&\quad - \frac{G}{(NGK)^2} \sum_{c=1}^C \sum_{g=1}^G \left( \sum_{k=1}^K N p_{cgk} \right)^2 \\
&= 1 - \frac{GK}{NGK^2} + \frac{1}{NGK^2} \sum_{c=1}^C \sum_{g=1}^G \sum_{k=1}^K p_{cgk}^2 - \frac{1}{GK^2} \sum_{c=1}^C \sum_{g=1}^G p_{cg\cdot}^2 \\
&= 1 - \frac{1}{NK} + \frac{1}{NGK^2} \sum_{c=1}^C \sum_{g=1}^G \left[ \sum_{k=1}^K p_{cgk}^2 - N p_{cg\cdot}^2 \right]
\end{aligned}$$

$$\begin{aligned}
\text{And } E(BSI) &= \frac{1}{NK} - \frac{1}{NGK} + \frac{1}{N(GK)^2} \sum_{c=1}^C \left[ \sum_{g=1}^G \sum_{k=1}^K p_{cgk}^2 - N p_{c\cdot}^2 \right] \\
&\quad - \frac{1}{NGK^2} \sum_{c=1}^C \sum_{g=1}^G \left[ \sum_{k=1}^K p_{cgk}^2 - N p_{cg\cdot}^2 \right] \\
&= \frac{G-1}{NGK} + \frac{1}{N(GK)^2} \sum_{c=1}^C \left[ \sum_{g=1}^G \sum_{k=1}^K p_{cgk}^2 - N p_{c\cdot}^2 \right] \\
&\quad - \frac{1}{NGK^2} \sum_{c=1}^C \sum_{g=1}^G \left[ \sum_{k=1}^K p_{cgk}^2 - N p_{cg\cdot}^2 \right]
\end{aligned}$$

Define the population variation within the  $g$ th group at the  $k$ th position as

$$I_S(\mathbf{p}_{gk}) = 1 - \sum_{c=1}^C p_{cgk}^2 \quad (3.4.7)$$

$H_0 : p_{cgk} = p_{ck}$  for all  $g$  implies that

$$I_S(\mathbf{p}_{1k}) = I_S(\mathbf{p}_{2k}) = \dots = I_S(\mathbf{p}_{Gk}) = I_S(\mathbf{p}_k) ,$$



i.e., within-group variation at the  $k$ th position is the same over all the groups and this implies that

$$\| \mathbf{P}_{1k} \| = \| \mathbf{P}_{2k} \| = \cdots = \| \mathbf{P}_{Gk} \| \quad (3.4.8)$$

where  $\mathbf{p}_{gk} = (p_{1gk} \ p_{2gk} \ \cdots \ p_{Cgk})'$  is a  $C \times 1$  vector representing the probabilities of belonging to categories  $c = 1, \dots, C$  in group  $g$  and position  $k$ .

If one is interested only in the hypothesis stated in (3.4.8), the hypothesis of homogeneity among the groups ( $p_{cgk} = p_{ck}$ ) is not necessarily true. Here, we consider  $H_0 : p_{cgk} = p_{ck}$ .

$$\begin{aligned} E_0(TSI) &= 1 - \frac{1}{NGK} + \frac{1}{N(GK)^2} \sum_{c=1}^C \left[ G \sum_{k=1}^K p_{ck}^2 - NG^2 p_c^2 \right] \\ &= 1 - \frac{1}{NGK} + \frac{1}{NGK^2} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - NG p_c^2 \right] \end{aligned} \quad (3.4.9)$$

$$E_0(WSI) = 1 - \frac{1}{NK} + \frac{1}{NK^2} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - N p_c^2 \right] \quad (3.4.10)$$

$$\begin{aligned} E_0(BSI) &= \frac{G-1}{NGK} + \frac{1}{N(GK)^2} \sum_{c=1}^C \left[ G \sum_{k=1}^K p_{ck}^2 - NG^2 p_c^2 \right] \\ &\quad - \frac{G}{NGK^2} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - N p_c^2 \right] \\ &= \frac{G-1}{NGK} + \frac{1}{NGK^2} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 - \frac{1}{NK^2} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \\ &= \frac{G-1}{NGK} \left[ 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right] \end{aligned} \quad (3.4.11)$$

Since  $\mathbf{V}$  follows a multinomial distribution, from (3.3.1) and (3.3.2), asymptotically,

$$\frac{\mathbf{V}}{\sqrt{N}} \xrightarrow{d} N(\sqrt{N} \boldsymbol{\mu}_0, \boldsymbol{\Sigma}^\circ) \quad (3.4.12)$$

where  $\boldsymbol{\Sigma}^\circ = \boldsymbol{\Sigma}_1 \oplus \boldsymbol{\Sigma}_2 \oplus \cdots \oplus \boldsymbol{\Sigma}_G$ ,  $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_{g1} \oplus \boldsymbol{\Sigma}_{g2} \oplus \cdots \oplus \boldsymbol{\Sigma}_{gK}$ ,  $g = 1, \dots, G$  and  $\boldsymbol{\Sigma}_{gk}$  is given by (3.3.3). Under  $H_0$ , for any  $g = 1, \dots, G$ ,

$$\boldsymbol{\Sigma}_{gk} = \boldsymbol{\Sigma}_{0k} \quad \text{and} \quad \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_0^* = \boldsymbol{\Sigma}_{01} \oplus \boldsymbol{\Sigma}_{02} \oplus \cdots \oplus \boldsymbol{\Sigma}_{0K} \quad (3.4.13)$$

where  $\boldsymbol{\Sigma}_{0k}$  is a  $C \times C$  matrix

$$\boldsymbol{\Sigma}_{0k} = \mathbf{D}_k - \boldsymbol{\mu}_{0k} \boldsymbol{\mu}'_{0k} \quad (3.4.14)$$

with  $\mathbf{D}_k$  being a  $C \times C$  diagonal matrix with elements  $p_{1k}, \dots, p_{Ck}$  and  $\boldsymbol{\mu}_{ok} = (p_{1k} \dots p_{Ck})'$ .

Therefore, under  $H_0$ ,

$$\boldsymbol{\Sigma} = N\boldsymbol{\Sigma}^\circ = N\boldsymbol{\Sigma}_0^\circ = N(\mathbf{I}_G \otimes \boldsymbol{\Sigma}_0^*) \quad (3.4.15)$$

Now,

$$\begin{aligned} \text{i. Cov}(BSI, WSI) &= \text{Cov}(BSI, TSI - BSI) \\ &= \text{Cov}(BSI, TSI) - \text{Var}(BSI) \end{aligned} \quad (3.4.16)$$

$$\begin{aligned} \text{ii. Cov}(TSI, WSI) &= \text{Cov}(TSI, TSI - BSI) \\ &= \text{Var}(TSI) - \text{Cov}(TSI, BSI) \end{aligned} \quad (3.4.17)$$

### 3.5 The Test Statistic

Note that  $BSI$  can be written as

$$BSI = \mathbf{V}'\mathbf{B}\mathbf{V} = \sum_{g=1}^G \sum_{c=1}^C \left[ \frac{n_{cg}}{NK\sqrt{G}} - \frac{n_{c..}NK\sqrt{G}}{(NGK)^2} \right]^2 \quad (3.5.1)$$

Let  $\theta_{cgk} = n_{cgk} - Np_{ck} = n_{cgk} - E_0(n_{cgk})$ . Then

$$\theta_{c..} = \sum_{g=1}^G \sum_{k=1}^K \theta_{cgk} = n_{c..} - NG \sum_{k=1}^K p_{ck}$$

Also, under  $H_0$ ,  $\boldsymbol{\theta} = (\theta_{111} \dots \theta_{CG1} \dots \theta_{CGK})'$  is asymptotically

$$\frac{\boldsymbol{\theta}}{\sqrt{N}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_0^\circ) \quad (3.5.2)$$

where  $\boldsymbol{\Sigma}_0^\circ$  is as in (3.4.15). So,

$$\begin{aligned} BSI &= \mathbf{V}'\mathbf{B}\mathbf{V} \\ &= \sum_{g=1}^G \sum_{c=1}^C \left[ \frac{\theta_{cg} + Np_{c.}}{NK\sqrt{G}} - \frac{(\theta_{c..} + NGp_{c.})}{(NGK)^2} NK\sqrt{G} \right]^2 \\ &= \left[ \frac{\theta_{cg}}{NK\sqrt{G}} - \frac{\theta_{c..}}{(NGK)^2} NK\sqrt{G} \right]^2 = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} = \frac{1}{G(NK)^2} \boldsymbol{\theta}'\mathbf{B}^\circ\boldsymbol{\theta} \end{aligned}$$

Hence, under  $H_0$   $BSI$  is asymptotically

$$BSI \sim \sum_{i=1}^{CGK} \lambda_i (\chi_1^2)_i, \quad (3.5.3)$$

where  $(\chi_1^2)_i$ 's are independent  $\chi^2$ -random variables with 1 degree of freedom and  $\{\lambda_i, i = 1, \dots, CGK\}$  is the set of characteristic roots of

$$NB\Sigma_0^\circ = \frac{1}{NGK^2} \mathbf{B}^\circ \Sigma_0^\circ = \frac{1}{NGK^2} \left[ \left( (\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G} \mathbf{U}_{KG} \right) \otimes \mathbf{I}_C \right] (\mathbf{I}_G \otimes \Sigma_0^*)$$

by (3.4.6) and (3.4.15).

Looking at  $\Sigma_{0k}$ , it is easy to see that its rank is at most  $(C-1)$ , because of the restriction that  $\sum_{c=1}^C p_{ck} = 1$ . In fact, the rank of each  $\Sigma_{0k}$  is  $(C-1)$ , since any of its  $(C-1)$  columns are linearly independent. Therefore, the rank of  $\Sigma_0^*$  is  $K(C-1)$ . Further,  $\frac{1}{NGK^2} \mathbf{B}^\circ \Sigma_0^\circ$  is a  $G \times G$  partition matrix with elements the  $\Sigma_{0k}$ 's matrices premultiplied by some constants ( $G-1$  or  $-1$ ).

Note that for any  $k$

$$\begin{aligned} & \overbrace{(G-1)\Sigma_{0k} + (G-1)\Sigma_{0k} + \dots + (G-1)\Sigma_{0k}}^{(G-1) \text{ terms}} - (G-1)(G-1)\Sigma_{0k} \\ &= (G-1)^2 \Sigma_{0k} - (G-1)^2 \Sigma_{0k} = \mathbf{0} \end{aligned}$$

and so

$$\begin{aligned} & (G-1)(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^* - \overbrace{[(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^* + \dots + (\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^*]}^{(G-1) \text{ terms}} \\ &= (G-1)(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^* - (G-1)(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^* = \mathbf{0} \end{aligned}$$

In order to get the characteristic roots of  $\frac{1}{NGK^2} \mathbf{B}^\circ \Sigma_0^\circ$  we need to solve the equation

$$\left| \frac{1}{NGK^2} \mathbf{B}^\circ \Sigma_0^\circ - \lambda \mathbf{I}_{CKG} \right| = 0 \quad (3.5.4)$$

But  $CK$  of the characteristic roots are zero and the others are the characteristic roots of  $G(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^*$ , with multiplicity  $(G-1)$ . Thus,

$$BSI \sim \frac{1}{NGK^2} \sum_{i=1}^{KC} \lambda_i (\chi_{(G-1)}^2)_i \quad (3.5.5)$$

where  $\{\lambda_i : 1, \dots, KC\}$  is the set of characteristic roots of  $(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^*$ . Since  $\{p_{ck}, c = 1, \dots, C\}$  is unknown and need to be estimated, so are  $\{\lambda_{ik}, i = 1, \dots, C-1\}$ . Determining the characteristic roots of a multinomial covariance matrix is not straightforward. Roy et al. (1960) studied this problem without actually presenting the closed-form expression for the roots. The characteristic equation for each  $k$  is

$$\left\{ 1 - \sum_{c=1}^C \left( \frac{p_{ck}^2}{p_{ck} - \lambda} \right) \right\} \prod_{c=1}^C (p_{ck} - \lambda) = 0 \quad (3.5.6)$$

It is easy to see that  $\lambda = 0$  is a root, but identifying the other roots must proceed numerically.

Now,

$$TSI = 1 - \mathbf{V}'\mathbf{T}\mathbf{V}$$

Since  $N\mathbf{T}\Sigma_0^\circ$  is not idempotent, the distribution of  $\mathbf{V}'\mathbf{T}\mathbf{V}$  is not  $\chi_{(\text{rank}(\mathbf{T}), \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu})}^2$ . Under  $H_0$ , however,

$$\begin{aligned} \mathbf{V}'\mathbf{T}\mathbf{V} &= \frac{1}{(NGK)^2} \mathbf{V}'\mathbf{T}^\circ\mathbf{V} = \frac{1}{(NGK)^2} \sum_{c=1}^C n_{c..}^2 \\ &= \frac{1}{(NGK)^2} \sum_{c=1}^C [\theta_{c..} + NGp_c.]^2 \\ &= \frac{1}{(NGK)^2} \sum_{c=1}^C (\theta_{c..}^2 + (NG)^2 p_c^2 + 2\theta_{c..} NGp_c.) \\ &= \frac{1}{(NGK)^2} \sum_{c=1}^C \theta_{c..}^2 + \frac{1}{K^2} \sum_{c=1}^C p_c^2 + \frac{2}{NGK^2} \sum_{c=1}^C \theta_{c..} p_c. \\ &= \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} + \frac{1}{K^2} \sum_{c=1}^C p_c^2 + \mathbf{A}'\boldsymbol{\theta} \\ &= \frac{1}{(NGK)^2} \boldsymbol{\theta}'\mathbf{T}^\circ\boldsymbol{\theta} + \frac{1}{K^2} \sum_{c=1}^C p_c^2 + \mathbf{A}'\boldsymbol{\theta} \end{aligned} \quad (3.5.7)$$

where  $\mathbf{A} = (\mathbf{A}^* \mathbf{A}^* \dots \mathbf{A}^*)'$  is a  $CGK \times 1$  vector and  $\mathbf{A}^*$  is a  $1 \times CK$  vector of the form

$$\mathbf{A}^* = \frac{2}{NGK^2} (p_1 \dots p_C \cdot p_1 \dots p_C \dots p_1 \dots p_C.)$$

### Lemma 3.1

$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$  and  $\mathbf{A}'\boldsymbol{\theta}$  are not independent.

**Proof:**

$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} = \frac{1}{(NGK)^2} \boldsymbol{\theta}'\mathbf{T}^\circ\boldsymbol{\theta}$  and  $\mathbf{A}'\boldsymbol{\theta}$  would be independent if and only if

$$\mathbf{A}'N\Sigma_0^\circ \frac{1}{(NGK)^2} \mathbf{T}^\circ = \mathbf{0} \quad (\text{Searle, 1971}).$$

$$\frac{1}{N(GK)^2} \mathbf{A}'\Sigma_0^\circ \mathbf{T}^\circ$$

$$\begin{aligned}
&= \frac{1}{N(GK)^2} (\mathbf{A}^* \mathbf{A}^* \dots \mathbf{A}^*) (\mathbf{I}_G \otimes \boldsymbol{\Sigma}_0^*) (\mathbf{U}_{KG} \otimes \mathbf{I}_C) \\
&= \frac{G}{N(GK)^2} (\mathbf{A}^* \boldsymbol{\Sigma}_0^* (\mathbf{U}_K \otimes \mathbf{I}_C) \mathbf{A}^* \boldsymbol{\Sigma}_0^* (\mathbf{U}_K \otimes \mathbf{I}_C) \dots \mathbf{A}^* \boldsymbol{\Sigma}_0^* (\mathbf{U}_K \otimes \mathbf{I}_C))
\end{aligned}$$

Let  $\mathbf{a} = (p_1 \dots p_C)'$ . Recall  $\boldsymbol{\Sigma}_0^* = \boldsymbol{\Sigma}_{01} \oplus \dots \oplus \boldsymbol{\Sigma}_{0K}$

$$\mathbf{A}^* \boldsymbol{\Sigma}_0^* (\mathbf{U}_K \otimes \mathbf{I}_C) = \frac{2}{NGK^2} (\mathbf{a}' \boldsymbol{\Sigma}_{01} \mathbf{a}' \boldsymbol{\Sigma}_{02} \dots \mathbf{a}' \boldsymbol{\Sigma}_{0K}) (\mathbf{U}_K \otimes \mathbf{I}_C)$$

For each  $k$ , from (3.4.14)

$$\mathbf{a}' \boldsymbol{\Sigma}_{0k} = [p_{1k}(p_{1\cdot} - \sum_{c=1}^C p_c p_{ck}) \ p_{2k}(p_{2\cdot} - \sum_{c=1}^C p_c p_{ck}) \ \dots \ p_{Ck}(p_{C\cdot} - \sum_{c=1}^C p_c p_{ck})]$$

and the first element of the vector  $\mathbf{A}^* \boldsymbol{\Sigma}_0^* (\mathbf{U}_K \otimes \mathbf{I}_C)$  is

$$\begin{aligned}
&\frac{2}{NGK^2} \left( p_{11}(p_{1\cdot} - \sum_{c=1}^C p_c p_{c1}) + p_{12}(p_{1\cdot} - \sum_{c=1}^C p_c p_{c2}) + \dots + p_{1K}(p_{1\cdot} - \sum_{c=1}^C p_c p_{cK}) \right) \\
&= \frac{2}{NGK^2} \left( \sum_{k=1}^K p_{1k}(p_{1\cdot} - \sum_{c=1}^C p_c p_{ck}) \right) \\
&= \frac{2}{NGK^2} \left( p_{1\cdot}^2 - \sum_{c=1}^C p_c \sum_{k=1}^K p_{1k} p_{ck} \right) \neq 0
\end{aligned}$$

Hence,  $\boldsymbol{\theta}' \mathbf{T} \boldsymbol{\theta}$  and  $\mathbf{A}' \boldsymbol{\theta}$  are not independent. ■

Now,

$$\boldsymbol{\theta}' \mathbf{T} \boldsymbol{\theta} \sim \sum_{i=1}^{KCG} \lambda_i (\chi_1^2)_i, \quad \mathbf{A}' \boldsymbol{\theta} \sim N(\mathbf{0}, N \mathbf{A}' \boldsymbol{\Sigma}_0^* \mathbf{A})$$

and

$$\mathbf{V}' \mathbf{T} \mathbf{V} \sim \sum_{i=1}^{KCG} \lambda_i (\chi_1^2)_i + N(\mathbf{0}, N \mathbf{A}' \boldsymbol{\Sigma}_0^* \mathbf{A}) + \delta_1$$

where  $\{\lambda_i, i = 1, \dots, KCG\}$  is the set of characteristic roots of

$$N \mathbf{T} \boldsymbol{\Sigma}_0^* = \frac{1}{N(GK)^2} \mathbf{T}^* \boldsymbol{\Sigma}_0^*$$
 and

$$\delta_1 = \frac{1}{K^2} \sum_{c=1}^C p_c^2 = \boldsymbol{\mu}' \mathbf{T} \boldsymbol{\mu} \text{ under } H_0 \tag{3.5.8}$$

Recall that  $\mathbf{T}^* = (\mathbf{U}_{KG} \otimes \mathbf{I}_C)$ . Thus, the characteristic roots of  $\mathbf{T}^* \boldsymbol{\Sigma}_0^*$  are the characteristic roots of  $(\mathbf{U}_K \otimes \mathbf{I}_C) \boldsymbol{\Sigma}_0^*$  with multiplicity  $G$ . Therefore,

$$\mathbf{V}' \mathbf{T} \mathbf{V} = \frac{1}{(NGK)^2} \mathbf{V}' \mathbf{T}^* \mathbf{V} \sim \frac{1}{N(GK)^2} \sum_{i=1}^{KCG} \lambda_i (\chi_G^2)_i + N(\mathbf{0}, N \mathbf{A}' \boldsymbol{\Sigma}_0^* \mathbf{A}) + \delta_1 \tag{3.5.9}$$

where  $\{\lambda_i : i = 1, \dots, KC\}$  is the set of characteristic roots of  $(\mathbf{U}_K \otimes \mathbf{I}_C)\boldsymbol{\Sigma}_0^*$ .

An alternative of the distribution of  $\mathbf{V}'\mathbf{T}\mathbf{V}$  goes along the following lines. Let  $\mathbf{R}$  be a  $CGK \times CGK$  matrix such that  $\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}' = \mathbf{I}_{CGK}$  and

$$\mathbf{Y} = \mathbf{R}\mathbf{V} \Rightarrow \mathbf{V} = \mathbf{R}^{-1}\mathbf{Y}$$

Then,

$$\mathbf{Y} \sim N(\mathbf{R}\boldsymbol{\mu}, \mathbf{I}_{CGK})$$

and

$$\mathbf{V}'\mathbf{T}\mathbf{V} = \mathbf{Y}'(\mathbf{R}^{-1})'\mathbf{T}\mathbf{R}^{-1}\mathbf{Y} \equiv \mathbf{Y}'\mathbf{C}\mathbf{Y}$$

Let  $\mathbf{P}$  be an orthogonal matrix such that  $(\mathbf{P}^{-1})'\mathbf{C}\mathbf{P}^{-1}$  is a diagonal matrix and

$$\mathbf{Y}^* \equiv \mathbf{P}\mathbf{Y} \Rightarrow \mathbf{Y} = \mathbf{P}^{-1}\mathbf{Y}^* = \mathbf{P}'\mathbf{Y}^*$$

Then,

$$\mathbf{Y}^* \sim N(\mathbf{P}\mathbf{R}\boldsymbol{\mu}, \mathbf{I}_{CGK})$$

$$\mathbf{V}'\mathbf{T}\mathbf{V} = \mathbf{Y}'\mathbf{C}\mathbf{Y} = (\mathbf{Y}^*)'\mathbf{P}\mathbf{C}\mathbf{P}'\mathbf{Y}^* = (\mathbf{Y}^*)'\mathbf{C}^*\mathbf{Y}^*$$

where  $\mathbf{C}^* \equiv \mathbf{P}(\mathbf{R}^{-1})'\mathbf{T}\mathbf{R}^{-1}\mathbf{P}'$ .

Hence,

$$\mathbf{V}'\mathbf{T}\mathbf{V} = (\mathbf{Y}^*)'\mathbf{C}^*\mathbf{Y}^* \sim \sum_{i=1}^{CGK} c_i^* (\chi_1^2(\delta_i)) \quad \text{with } \delta_i \equiv c_i^*(\mu_i^*)^2 \quad (3.5.10)$$

where  $c_i^*$ 's are the diagonal elements of  $\mathbf{C}^*$  and  $\mu_i^*$  is the  $i$ th row of the vector  $\boldsymbol{\mu}^* = \mathbf{P}\mathbf{R}\boldsymbol{\mu}$ .

As for

$$WSI = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} ,$$

under  $H_0$

$$\mathbf{V}'\mathbf{W}\mathbf{V} = \frac{1}{G(NK)^2} \mathbf{V}'\mathbf{W}^*\mathbf{V} = \frac{1}{G(NK)^2} \sum_{g=1}^G \sum_{c=1}^C n_{cg}^2 \quad \text{from (3.2.5)}$$

$$\begin{aligned}
&= \frac{1}{G(NK)^2} \sum_{g=1}^G \sum_{c=1}^C (\theta_{cg} + Np_c)^2 \\
&= \frac{1}{G(NK)^2} \sum_{g=1}^G \sum_{c=1}^C \theta_{cg}^2 + \frac{2}{NGK^2} \sum_{c=1}^C \theta_{c \cdot} p_c + \frac{1}{GK^2} \sum_{g=1}^G \sum_{c=1}^C p_c^2 \\
&= \boldsymbol{\theta}' \mathbf{W} \boldsymbol{\theta} + \mathbf{A}' \boldsymbol{\theta} + \frac{1}{K^2} \sum_{c=1}^C p_c^2 \\
&= \boldsymbol{\theta}' \mathbf{W} \boldsymbol{\theta} + \mathbf{A}' \boldsymbol{\theta} + \delta_1 \quad \text{from (3.5.8)} \tag{3.5.11}
\end{aligned}$$

Again,

$$\mathbf{V}' \mathbf{W} \mathbf{V} = \frac{1}{G(NK)^2} \mathbf{V}' \mathbf{W}^\circ \mathbf{V} \sim \sum_{i=1}^{CGK} \lambda_i (\chi_1^2)_i + N(\mathbf{0}, NA' \Sigma_0^\circ \mathbf{A}) + \delta_1$$

where  $\{\lambda_i, i = 1, \dots, CGK\}$  is the set of characteristic roots of

$$N \mathbf{W} \Sigma_0^\circ = \frac{1}{NGK^2} \mathbf{W}^\circ \Sigma_0^\circ.$$

Recall that  $\mathbf{W}^\circ = [(\mathbf{I}_G \otimes \mathbf{U}_K) \otimes \mathbf{I}_C]$ . Hence, the characteristic roots of  $\mathbf{W}^\circ \Sigma_0^\circ$  have multiplicity  $G$ . Then,

$$\mathbf{V}' \mathbf{W} \mathbf{V} = \frac{1}{G(NK)^2} \mathbf{V}' \mathbf{W}^\circ \mathbf{V} \sim \frac{1}{NGK^2} \sum_{i=1}^{KC} \lambda_i (\chi_G^2)_i + N(\mathbf{0}, NA' \Sigma_0^\circ \mathbf{A}) + \delta_1 \tag{3.5.12}$$

where  $\{\lambda_i : i = 1, \dots, KC\}$  is the set of characteristic roots of  $(\mathbf{U}_K \otimes \mathbf{I}_C) \Sigma_0^\circ$  and  $\delta_1$  is given by (3.5.8).

Alternatively, similarly to the derivation of (3.5.10)

$$\mathbf{V}' \mathbf{W} \mathbf{V} = (\mathbf{X}^*)' \mathbf{D}^* \mathbf{X}^* \sim \sum_{i=1}^{CGK} d_i^* (\chi_1^2(\delta_i)) \quad \text{with } \delta_i = d_i^* (\nu_i^*)^2 \tag{3.5.13}$$

where  $d_i^*$ 's are the diagonal elements of  $\mathbf{D}^* = \mathbf{P}_2 (\mathbf{R}_2^{-1})' \mathbf{W} \mathbf{R}_2^{-1} \mathbf{P}_2'$  and  $\nu_i^*$  is the  $i$ th row of the vector  $\boldsymbol{\nu}^* = \mathbf{P}_2 \mathbf{R}_2 \boldsymbol{\mu}$ . Note that

$$\mathbf{X}^* \sim N(\boldsymbol{\nu}^*, \mathbf{I}_{CGK})$$

where  $\mathbf{R}_2$  is a  $CGK \times CGK$  matrix such that  $\mathbf{R}_2 \Sigma \mathbf{R}_2' = \mathbf{I}_{CGK}$  and  $\mathbf{P}_2$  is an orthogonal matrix such that  $(\mathbf{P}_2^{-1})' (\mathbf{R}_2^{-1})' \mathbf{W} \mathbf{R}_2^{-1} \mathbf{P}_2^{-1}$  is a diagonal matrix.

As the above distributional results for the sums of squares do not readily yield the first moments, we calculate them directly.

Let

$$\begin{aligned}\theta_1 &\equiv E_0(BSI) = \frac{G-1}{NGK} \left[ 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right] \\ \theta_2 &\equiv E_0(TSI) = 1 - \frac{1}{NGK} + \frac{1}{NGK^2} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - NG p_c^2 \right] \\ \theta_3 &\equiv E_0(WSI) = 1 - \frac{1}{NK} + \frac{1}{NK^2} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - N p_c^2 \right]\end{aligned}$$

from (3.4.11), (3.4.9) and (3.4.10). The variances are calculated using the following theorem.

**Theorem 3.1** (Searle, 1971)

When  $\mathbf{X}$  is  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , the  $r$ th cumulant of  $\mathbf{X}'\mathbf{A}\mathbf{X}$  is

$$K_r(\mathbf{X}'\mathbf{A}\mathbf{X}) = 2^{r-1}(r-1)! [\text{tr}(\mathbf{A}\boldsymbol{\Sigma})^r + r\boldsymbol{\mu}'\mathbf{A}(\boldsymbol{\Sigma}\mathbf{A})^{r-1}\boldsymbol{\mu}]$$

■

Since  $\frac{\mathbf{V}}{\sqrt{N}} \sim N(\sqrt{N}\boldsymbol{\mu}_\circ, \boldsymbol{\Sigma}^\circ)$  and  $\frac{\boldsymbol{\theta}}{\sqrt{N}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^\circ)$ ,

$$\text{Var}(BSI) = \text{Var}(\mathbf{V}'\mathbf{B}\mathbf{V}) = \text{Var}(\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}) = 2 \text{trace}(\mathbf{B}N\boldsymbol{\Sigma}^\circ)^2 \quad (3.5.14)$$

$$\text{Var}(TSI) = \text{Var}(\mathbf{V}'\mathbf{T}\mathbf{V}) = 2 \text{trace}(\mathbf{T}N\boldsymbol{\Sigma}^\circ)^2 + 4N\boldsymbol{\mu}'_\circ\mathbf{T}N\boldsymbol{\Sigma}^\circ\mathbf{T}N\boldsymbol{\mu}_\circ \quad (3.5.15)$$

$$\text{Var}(WSI) = \text{Var}(\mathbf{V}'\mathbf{W}\mathbf{V}) = 2 \text{trace}(\mathbf{W}N\boldsymbol{\Sigma}^\circ)^2 + 4N\boldsymbol{\mu}'_\circ\mathbf{W}N\boldsymbol{\Sigma}^\circ\mathbf{W}N\boldsymbol{\mu}_\circ \quad (3.5.16)$$

and under  $H_0$

$$\text{Var}_0(BSI) = 2 \text{trace} \left( \frac{1}{NGK^2} \mathbf{B}^\circ \boldsymbol{\Sigma}_0^\circ \right)^2 = \frac{2}{(NGK^2)^2} \text{trace}(\mathbf{B}^\circ \boldsymbol{\Sigma}_0^\circ)^2 \quad (3.5.17)$$

$$\begin{aligned}\text{Var}_0(TSI) &= 2 \text{trace} \left( \frac{1}{N(GK)^2} \mathbf{T}^\circ \boldsymbol{\Sigma}_0^\circ \right)^2 + \frac{4}{N(GK)^4} \boldsymbol{\mu}'_\circ \mathbf{T}^\circ \boldsymbol{\Sigma}_0^\circ \mathbf{T}^\circ \boldsymbol{\mu}_\circ \\ &= \frac{2}{N^2(GK)^4} \text{trace}(\mathbf{T}^\circ \boldsymbol{\Sigma}_0^\circ)^2 + \frac{4}{N(GK)^4} \boldsymbol{\mu}'_\circ \mathbf{T}^\circ \boldsymbol{\Sigma}_0^\circ \mathbf{T}^\circ \boldsymbol{\mu}_\circ \quad (3.5.18)\end{aligned}$$

$$\text{Var}_0(WSI) = \frac{2}{(NG)^2 K^4} \text{trace}(\mathbf{W}^\circ \boldsymbol{\Sigma}_0^\circ)^2 + \frac{4}{NG^2 K^4} \boldsymbol{\mu}'_\circ \mathbf{W}^\circ \boldsymbol{\Sigma}_0^\circ \mathbf{W}^\circ \boldsymbol{\mu}_\circ \quad (3.5.19)$$

Let

$$T_{N,1} \equiv BSI - \theta_1, \quad T_{N,2} \equiv TSI - \theta_2, \quad T_{N,3} \equiv WSI - \theta_3$$



Note that

$$(i) \sum_{i=1}^{KCG} \lambda_i (\chi_G^2)_i = O_p(N^{-1})$$

since  $\{\lambda_i : i = 1, \dots, KC\}$  is the set of characteristic roots of  $\frac{1}{N(GK)^2}(\mathbf{U}_K \otimes \mathbf{I}_C)\boldsymbol{\Sigma}_0^*$  and  $\boldsymbol{\Sigma}_0^* = O(1)$ .

$$(ii) \mathbf{A}'\boldsymbol{\theta} = O_p(N^{-1/2}) \text{ since } \mathbf{A}'\boldsymbol{\theta} \sim N(\mathbf{0}, N\mathbf{A}'\boldsymbol{\Sigma}_0^*\mathbf{A}) \text{ and } \mathbf{A} = O(N^{-1})$$

$$(iii) \delta_1 = \frac{1}{K^2} \sum_{c=1}^C p_c^2 = O(1) .$$

Then,

$$\begin{aligned} T_{N,2} &= TSI - \theta_2 = 1 - \mathbf{V}'\mathbf{T}\mathbf{V} - \theta_2 \\ &= 1 - \left( O_p(N^{-1}) + O_p(N^{-1/2}) + \frac{1}{K^2} \sum_{c=1}^C p_c^2 \right) \\ &\quad - \left( 1 + O(N^{-1}) - \frac{1}{K^2} \sum_{c=1}^C p_c^2 \right) \\ &= O_p(N^{-1/2}) \end{aligned}$$

Similarly,

$$\begin{aligned} T_{N,3} &= WSI - \theta_3 = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} - \theta_3 \\ &= 1 - \left( O_p(N^{-1}) + O_p(N^{-1/2}) + \frac{1}{K^2} \sum_{c=1}^C p_c^2 \right) \\ &\quad - \left( 1 + O(N^{-1}) - \frac{1}{K^2} \sum_{c=1}^C p_c^2 \right) \\ &= O_p(N^{-1/2}) \end{aligned}$$

and

$$BSI = \mathbf{V}'\mathbf{B}\mathbf{V} = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} = O_p(N^{-1})$$

Then

$$\begin{aligned} F_1 &\equiv N \left( \frac{BSI}{WSI} \right) = N \left( \frac{BSI}{T_{N,3} + \theta_3} \right) \\ &= N \left( \frac{BSI}{\theta_3} \right) \left[ 1 + \frac{T_{N,3}}{\theta_3} \right]^{-1} \\ &= N \left( \frac{BSI}{\theta_3} \right) + O_p(N^{-1/2}) \\ &= N \left( \frac{BSI}{\theta_3^0} \right) + O_p(N^{-1/2}) \end{aligned}$$

since  $T_{N,3} = O_p(N^{-1/2})$ ,  $\frac{N(BSI)T_{N,3}}{\theta_3^2} = O_p(N^{-1/2})$ ,  $N(BSI) = O_p(1)$  and

$$\begin{aligned}\theta_3 &= 1 - \frac{1}{NK} \left( 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right) - \frac{1}{K^2} \sum_{c=1}^C p_c^2 \\ &= 1 - \frac{1}{K^2} \sum_{c=1}^C p_c^2 + O(N^{-1}) \\ &= \theta_3^o + O(N^{-1})\end{aligned}$$

By (3.5.3), asymptotically

$$F_1 = N \frac{BSI}{\theta_3^o} \sim \frac{1}{\theta_3^o} \sum_{i=1}^{CGK} \lambda_i (\chi_1^2)_i \quad (3.5.20)$$

where  $\{\lambda_i : 1, \dots, CGK\}$  is the set of characteristic roots of  $\frac{1}{NGK^2} \mathbf{B}^o \Sigma_0^*$ .

Under  $H_0$ , asymptotically

$$\begin{aligned}E_0(F_1) &= N \left[ \frac{E_0(BSI)}{\theta_3^o} \right] = \frac{N\theta_1}{\theta_3^o} \\ &= \frac{(G-1)}{GK\theta_3^o} \left[ 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right]\end{aligned}$$

$$\text{Var}_0(F_1) = N^2 \left[ \frac{\text{Var}_0(BSI)}{(\theta_3^o)^2} \right] = \frac{N^2 2 \text{trace}(\mathbf{B}^o \Sigma_0^o)^2}{N^2 (GK^2 \theta_3^o)^2} = \frac{2 \text{trace}(\mathbf{B}^o \Sigma_0^o)^2}{(GK^2 \theta_3^o)^2}$$

Since  $p_{ck}$ 's are unknown, one can only get estimates for the  $\lambda_i$ 's, i.e., the characteristic roots of  $\Sigma_0^o$ . To derive the distribution of  $F_1$ , estimate  $\{p_{ck}\}$ , then get the characteristic roots of  $\Sigma_0^o$  based on those estimates.

Alternatively, since each of the elements on the R.H.S. of (3.5.20) are i.i.d.  $\chi_1^2$ 's, if  $\frac{\max(\lambda_i)}{\sqrt{\sum_{i=1}^{CGK} \lambda_i^2}} \rightarrow 0$ , we can apply the C.L.T. for  $CGK$  large,

$$K^2 \left( F_1 - N \frac{\theta_1}{\theta_3^o} \right) \sim N(0, \sigma^2) \quad (3.5.21)$$

where  $\sigma^2 = \frac{2 \text{trace}(\mathbf{B}^o \Sigma_0^o)^2}{(G\theta_3^o)^2}$ .

Using a similar approach, Light & Margolin (1971) developed an analysis of variance for categorical data (CATANOVA) and these two approaches are equivalent. For a one-way table, the Total Sum of Squares (TSS) is

$$TSS = \frac{NG}{2} - \frac{1}{2NG} \sum_{c=1}^C n_c^2 = \frac{NG}{2} TSI \quad (3.5.22)$$

The variation within group  $g$  (i.e., within  $x_1^g, x_2^g, \dots, x_N^g$ ) is

$$\frac{n_{\cdot g}}{2} - \frac{1}{2n_{\cdot g}} \sum_{c=1}^C n_{cg}^2 \quad (3.5.23)$$

Therefore, the within-group sum of squares (WSS) is found by summing (3.5.23) over  $g$ :

$$WSS = \sum_{g=1}^G \left( \frac{n_{\cdot g}}{2} - \frac{1}{2n_{\cdot g}} \sum_{c=1}^C n_{cg}^2 \right) = \frac{GN}{2} - \frac{1}{2N} \sum_{g=1}^G \sum_{c=1}^C n_{cg}^2 = \frac{NG}{2} WSI \quad (3.5.24)$$

The between sum of squares (BSS) is

$$\begin{aligned} BSS &= TSS - WSS = \frac{NG}{2} - \frac{1}{2NG} \sum_{c=1}^C n_c^2 - \frac{GN}{2} + \frac{1}{2N} \sum_{g=1}^G \sum_{c=1}^C n_{cg}^2 \\ &= \frac{1}{2NG} \left\{ G \left( \sum_{c=1}^C \sum_{g=1}^G n_{cg}^2 \right) - \sum_{c=1}^C n_c^2 \right\} = \frac{NG}{2} BSI \end{aligned} \quad (3.5.25)$$

Extending these for several positions, the variation within the  $g$ -th group at the  $k$ -th position is

$$\frac{n_{\cdot gk}}{2} - \frac{1}{2n_{\cdot gk}} \sum_{c=1}^C n_{cgk}^2 = \frac{N}{2} - \frac{1}{2N} \sum_{c=1}^C n_{cgk}^2,$$

since  $n_{\cdot gk} = N$ . The variation within the  $g$ -th group is

$$\frac{n_{\cdot g}}{2} - \frac{1}{2n_{\cdot g}} \sum_{c=1}^C n_{cg}^2 = \frac{NK}{2} - \frac{1}{2NK} \sum_{c=1}^C n_{cg}^2,$$

since  $n_{\cdot g} = NK$ . The sum of squares are

$$\begin{aligned} WSS &= \sum_{g=1}^G \left( \frac{NK}{2} - \frac{1}{2NK} \sum_{c=1}^C n_{cg}^2 \right) = \frac{NGK}{2} - \frac{1}{2NK} \sum_{g=1}^G \sum_{c=1}^C n_{cg}^2 \quad (3.5.26) \\ &= \frac{NGK}{2} WSI, \end{aligned}$$

$$TSS = \frac{NGK}{2} - \frac{1}{2NGK} \sum_{c=1}^C n_{c\cdot}^2 = \frac{NGK}{2} TSI, \quad (3.5.27)$$

$$\begin{aligned} \text{and } BSS &= TSS - WSS = \frac{1}{2NK} \sum_{g=1}^G \sum_{c=1}^C n_{cg}^2 - \frac{1}{2NGK} \sum_{c=1}^C n_{c\cdot}^2 \\ &= \frac{1}{2NGK} \left( G \sum_{g=1}^G \sum_{c=1}^C n_{cg}^2 - \sum_{c=1}^C n_{c\cdot}^2 \right) = \frac{NGK}{2} BSI. \end{aligned} \quad (3.5.28)$$

In this setup a test statistic could be

$$F_1^* = \frac{BSS/(G-1)}{WSS/(NGK-G)} = \frac{BSI/(G-1)}{WSI/(NGK-1)} = \frac{(NGK-G)}{N(G-1)} F_1 \quad (3.5.29)$$

Let

$$\begin{aligned} \theta_1^* &\equiv E_0(BSS) = \frac{NGK}{2} E_0(BSI) = \frac{G-1}{2} \left[ 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right] \\ \theta_2^* &\equiv E_0(TSS) = \frac{NGK}{2} E_0(TSI) = \frac{NGK-1}{2} + \frac{1}{2K} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - NG p_c^2 \right] \\ \theta_3^* &\equiv E_0(WSS) = \frac{NGK}{2} E_0(WSI) = \frac{NGK-G}{2} + \frac{G}{2K} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - N p_c^2 \right] \end{aligned}$$

from (3.4.11), (3.4.9) and (3.4.10).

Under  $H_0$

$$\text{Var}_0(BSS) = 2 \text{trace} \left( \frac{1}{2K} \mathbf{B}^\circ \Sigma_0^\circ \right)^2 = \frac{1}{2K^2} \text{trace}(\mathbf{B}^\circ \Sigma_0^\circ)^2 \quad (3.5.30)$$

$$\text{Var}_0(TSS) = \frac{1}{2(GK)^2} \text{trace}(\mathbf{T}^\circ \Sigma_0^\circ)^2 + \frac{N}{(GK)^2} \boldsymbol{\mu}'_0 \mathbf{T}^\circ \Sigma_0^\circ \mathbf{T}^\circ \boldsymbol{\mu}_0 \quad (3.5.31)$$

$$\text{Var}_0(WSS) = \frac{1}{2K^2} \text{trace}(\mathbf{W}^\circ \Sigma_0^\circ)^2 + \frac{N}{K^2} \boldsymbol{\mu}'_0 \mathbf{W}^\circ \Sigma_0^\circ \mathbf{W}^\circ \boldsymbol{\mu}_0 \quad (3.5.32)$$

Let

$$S_1 \equiv BSS - \theta_1^*, \quad S_{N,2} \equiv TSS - \theta_2^*, \quad S_{N,3} \equiv WSS - \theta_3^*$$

and

$$a_1 \equiv \frac{\theta_1^*}{(G-1)} = \frac{1}{2} \left[ 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right] \quad (3.5.33)$$

$$\begin{aligned} a_2 &\equiv \frac{\theta_2^*}{(NGK-1)} = \frac{1}{2} - \frac{1}{2K^2} \sum_{c=1}^C \left( \sum_{k=1}^K p_{ck} \right)^2 + O(N^{-1}) \\ &= \frac{1}{2} - \frac{1}{2K^2} \sum_{c=1}^C p_c^2 + O(N^{-1}) \\ &= a_2^* + O(N^{-1}) \end{aligned} \quad (3.5.34)$$

$$\begin{aligned} a_3 &\equiv \frac{\theta_3^*}{(NGK-G)} = \frac{1}{2} - \frac{1}{2K^2} \sum_{c=1}^C \left( \sum_{k=1}^K p_{ck} \right)^2 + O(N^{-1}) \\ &= \frac{1}{2} - \frac{1}{2K^2} \sum_{c=1}^C p_c^2 + O(N^{-1}) \\ &= a_2^* + O(N^{-1}) \end{aligned} \quad (3.5.35)$$

Now

$$\begin{aligned}
\frac{S_{N,2}}{(NGK-1)} &= \frac{TSS - \theta_2^*}{(NGK-1)} = \frac{\frac{NGK}{2}(1 - \mathbf{V}'\mathbf{TV}) - \theta_2^*}{(NGK-1)} \\
&= \frac{1}{2} - \left( O_p(N^{-1}) + O_p(N^{-1/2}) + \frac{1}{2K^2} \sum_{c=1}^C p_c^2 \right) \\
&\quad - \left( \frac{1}{2} - \frac{1}{2K^2} \sum_{c=1}^C p_c^2 + O(N^{-1}) \right) \\
&= O_p(N^{-1/2})
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{S_{N,3}}{(NGK-G)} &= \frac{WSS - \theta_3^*}{(NGK-G)} = \frac{\frac{NGK}{2}(1 - \mathbf{V}'\mathbf{WV}) - \theta_3^*}{(NGK-G)} \\
&= \frac{1}{2} - \left( O_p(N^{-1}) + O_p(N^{-1/2}) + \frac{1}{2K^2} \sum_{c=1}^C p_c^2 \right) \\
&\quad - \left( \frac{1}{2} - \frac{1}{2K^2} \sum_{c=1}^C p_c^2 + O(N^{-1}) \right) \\
&= O_p(N^{-1/2})
\end{aligned}$$

Then

$$\begin{aligned}
F_1^* &\equiv \frac{BSS/(G-1)}{WSS/(NGK-G)} = \frac{\frac{\theta_1^*}{(G-1)} + \frac{S_1}{(G-1)}}{\frac{\theta_3^*}{(NGK-G)} + \frac{S_{N,3}}{(NGK-G)}} \\
&= \frac{a_1 + S_1/(G-1)}{a_3 + \frac{S_{N,3}}{(NGK-G)}} \\
&= \frac{a_1 + S_1/(G-1)}{a_3} \left[ 1 + \frac{S_{N,3}}{a_3(NGK-G)} \right]^{-1} \\
&= \frac{a_1}{a_3} + \frac{S_1}{(G-1)a_3} + O_p(N^{-1/2}) \\
&= \frac{BSS}{a_3(G-1)} + O_p(N^{-1/2}) \\
&= \frac{BSS}{a_2^*(G-1)} + O_p(N^{-1/2})
\end{aligned}$$

since  $\frac{S_{N,3}}{(NGK-G)} = O_p(N^{-1/2})$ ,  $\frac{a_1 S_{N,3}}{a_3^2(NGK-G)} = O_p(N^{-1/2})$  and  $a_3 = a_2^* + O(N^{-1})$ .

By (3.5.3), asymptotically

$$F_1^* = \frac{BSS}{a_2^*(G-1)} \sim \frac{1}{a_2^*(G-1)} \sum_{i=1}^{CGK} \lambda_i (\chi_i^2), \quad (3.5.36)$$

where  $\{\lambda_i : 1, \dots, CGK\}$  is the set of characteristic roots of  $\frac{1}{2K} \mathbf{B}^\circ \Sigma_0^*$ .

Under  $H_0$ , asymptotically

$$\begin{aligned} E_0(F_1^*) &= \frac{a_1}{a_2^*} = \frac{E(BSS)}{a_2^*(G-1)} = \frac{\theta_1}{a_2^*(G-1)} \\ &= \frac{1}{2a_2^*} \left[ 1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \right] \\ \text{Var}_0(F_1^*) &= \frac{\text{Var}_0(BSS)}{(a_2^*(G-1))^2} = \frac{\text{trace}(\mathbf{B}^\circ \Sigma_0^*)^2}{2(Ka_2^*(G-1))^2} \end{aligned}$$

Applying the C.L.T. for  $CGK$  large,

$$K \left( F_1^* - \frac{a_1}{a_2^*} \right) \sim N(0, \sigma_\star^2) \quad (3.5.37)$$

where  $\sigma_\star^2 = \frac{\text{trace}(\mathbf{B}^\circ \Sigma_0^*)^2}{2(a_2^*(G-1))^2}$ .

Note that  $a_2^* = \frac{1}{2} \theta_3^2$  and that for large  $N$ , we can say that  $F_1^* = \frac{GK}{(G-1)} F_1$ . Therefore,

it can be shown that  $\sigma_\star^2 = \frac{G^2}{(G-1)^2} \sigma^2$ .

Note that, as  $N \rightarrow \infty$ ,  $\frac{a_1}{a_3} \rightarrow \frac{a_1}{a_2^*}$  and

$$\begin{aligned} \frac{a_1}{a_2^*} &= \frac{1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2}{1 - \frac{1}{K^2} \sum_{c=1}^C \left( \sum_{k=1}^K p_{ck} \right)^2} \\ &= \frac{1 - \frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2}{1 - \sum_{c=1}^C \left( \sum_{k=1}^K \frac{p_{ck}}{K} \right)^2} \\ &= \frac{1 - \sum_{c=1}^C \left( \frac{1}{K} \sum_{k=1}^K (p_{ck} - \bar{p}_c)^2 + \bar{p}_c^2 \right)}{1 - \sum_{c=1}^C \bar{p}_c^2}, \quad \text{where } \bar{p}_c = \sum_{k=1}^K \frac{p_{ck}}{K} \\ &= 1 - \frac{\frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K (p_{ck} - \bar{p}_c)^2}{1 - \sum_{c=1}^C \bar{p}_c^2} \end{aligned}$$

Note that  $\frac{a_1}{a_2^*}$  is bounded by 0 and 1, since  $\frac{1}{K} \sum_{c=1}^C \sum_{k=1}^K p_{ck}^2 \geq \sum_{c=1}^C \left( \frac{1}{K} \sum_{k=1}^K p_{ck} \right)^2$ .

Also, if there is homogeneity among positions within a group,  $p_{c1} = p_{c2} = \dots = p_{cK} = \bar{p}_c$  and  $\frac{a_1}{a_2^*} = 1$ .

Note that

$$\begin{aligned}
\frac{a_3}{a_2} &= \frac{1 + \frac{1}{K(NK-1)} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - N \left( \sum_{k=1}^K p_{ck} \right)^2 \right]}{1 + \frac{1}{K(NGK-1)} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - NG \left( \sum_{k=1}^K p_{ck} \right)^2 \right]} \\
&= 1 + \frac{\frac{N(G-1)}{(NK-1)(NGK-1)} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - \frac{1}{K} \left( \sum_{k=1}^K p_{ck} \right)^2 \right]}{1 + \frac{1}{K(NGK-1)} \sum_{c=1}^C \left[ \sum_{k=1}^K p_{ck}^2 - NG \left( \sum_{k=1}^K p_{ck} \right)^2 \right]} \\
&= 1 + \frac{\frac{N(G-1)}{(NK-1)(NGK-1)} \sum_{c=1}^C \sum_{k=1}^K (p_{ck} - \bar{p}_c)^2}{1 + \frac{1}{K(NGK-1)} \sum_{c=1}^C \sum_{k=1}^K (p_{ck} - \bar{p}_c)^2 - \sum_{c=1}^C \bar{p}_c^2}
\end{aligned}$$

Again, if there is homogeneity among positions within a group,  $p_{c1} = p_{c2} = \dots = p_{cK} = \bar{p}_c$  and  $\frac{a_3}{a_2} = 1$ .

### 3.6 The Power of the Test

Let us consider an alternative hypothesis, i.e.,

$$p_{cgk} = \frac{1}{\sqrt{N}} \gamma_{cgk} + p_{ck} .$$

Thus,  $\gamma_{cgk} = 0$  yields the null hypothesis  $H_0 : p_{cgk} = p_{ck}$ . The interest here is in the case where  $\gamma_{cgk} \neq 0$ . Then,

$$\theta_{cgk} = n_{cgk} - N \left( \frac{1}{\sqrt{N}} \gamma_{cgk} + p_{ck} \right)$$

$$\theta_{cg.} = n_{cg.} - \sqrt{N} \gamma_{cg.} - N p_{c.} , \quad \theta_{c..} = n_{c..} - \sqrt{N} \gamma_{c..} - NG p_{c.}$$

and

$$\begin{aligned}
BSI &= \mathbf{V}'\mathbf{B}\mathbf{V} = \sum_{g=1}^G \sum_{c=1}^C \left[ \frac{n_{cg.}}{NG\sqrt{G}} - \frac{n_{c..}NK\sqrt{G}}{(NGK)^2} \right]^2 \\
&= \sum_{g=1}^G \sum_{c=1}^C \left[ \frac{\theta_{cg.} + Np_{c.} + \sqrt{N}\gamma_{cg.}}{NK\sqrt{G}} - \frac{(\theta_{c..} + NGp_{c.} + \sqrt{N}\gamma_{c..})NK\sqrt{G}}{(NGK)^2} \right]^2 \\
&= \sum_{g=1}^G \sum_{c=1}^C \left[ \frac{\theta_{cg.}}{NK\sqrt{G}} - \frac{\theta_{c..}\sqrt{G}}{NKG^2} + \frac{\gamma_{cg.}}{K\sqrt{NG}} - \frac{\gamma_{c..}\sqrt{G}}{KG^2\sqrt{N}} \right]^2 \\
&= \sum_{g=1}^G \sum_{c=1}^C \left( \frac{\theta_{cg.}}{NK\sqrt{G}} - \frac{\theta_{c..}NK\sqrt{G}}{(NGK)^2} \right)^2
\end{aligned}$$

$$\begin{aligned}
& + 2 \sum_{g=1}^G \sum_{c=1}^C \left( \frac{\theta_{cg}}{NK\sqrt{G}} - \frac{\theta_{c..}NK\sqrt{G}}{(NGK)^2} \right) \left( \frac{\gamma_{cg}}{K\sqrt{NG}} - \frac{\gamma_{c..}\sqrt{G}}{KG^2\sqrt{N}} \right) \\
& + \sum_{g=1}^G \sum_{c=1}^C \left( \frac{\gamma_{cg}}{K\sqrt{NG}} - \frac{\gamma_{c..}\sqrt{G}}{KG^2\sqrt{N}} \right)^2 \\
& = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + (\mathbf{A}_1 - \mathbf{A}_2)'\boldsymbol{\theta} + N^2(\mathbf{A}_1 - \mathbf{A}_2)'(\mathbf{A}_1 - \mathbf{A}_2)
\end{aligned}$$

where  $\mathbf{B}$  is as in (3.4.6),  $\boldsymbol{\theta} = (\theta_{111} \dots \theta_{c_1 g_1} \dots \theta_{CGK})'$ ,  $\mathbf{A}_1$  and  $\mathbf{A}_2$  are  $CGK \times 1$  vectors of the form

$$\begin{aligned}
\mathbf{A}_1 & \equiv \frac{2}{K^2GN^{3/2}}(\mathbf{A}_{11}^* \dots \mathbf{A}_{1G}^*) \text{ with} \\
\mathbf{A}_{1g}^* & = (\gamma_{1g} \dots \gamma_{Cg} \gamma_{1g} \dots \gamma_{Cg} \dots \gamma_{1g} \dots \gamma_{Cg}) \text{ for each } g = 1, \dots, G \\
\mathbf{A}_2 & \equiv \frac{2}{(KG)^2N^{3/2}}(\mathbf{A}_2^* \dots \mathbf{A}_2^*) \text{ with} \\
\mathbf{A}_2^* & = (\gamma_{1..} \dots \gamma_{C..} \gamma_{1..} \dots \gamma_{C..} \dots \gamma_{1..} \dots \gamma_{C..})
\end{aligned}$$

Recall that  $\frac{\boldsymbol{\theta}}{\sqrt{N}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^\circ)$  and  $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} \sim \sum_{i=1}^{CGK} \lambda_i (\chi_{1i}^2)$ , with  $\lambda_i$ 's being the characteristic roots of  $\frac{1}{NGK^2}\mathbf{B}^\circ\boldsymbol{\Sigma}^\circ$ , where

$$\boldsymbol{\Sigma}^\circ = (\boldsymbol{\Sigma}_{11} \oplus \boldsymbol{\Sigma}_{12} \oplus \dots \oplus \boldsymbol{\Sigma}_{1K} \oplus \boldsymbol{\Sigma}_{21} \oplus \dots \oplus \boldsymbol{\Sigma}_{GK}) \quad (3.6.1)$$

and  $\boldsymbol{\Sigma}_{gk}$  is as in (3.3.3). Now, let  $\mathbf{A}_1^\circ = N^{3/2}\mathbf{A}_1$  and  $\mathbf{A}_2^\circ = N^{3/2}\mathbf{A}_2$ . Then

$$N(\mathbf{A}_1 - \mathbf{A}_2)'\boldsymbol{\theta} \sim N(\mathbf{0}, (\mathbf{A}_1^\circ - \mathbf{A}_2^\circ)'\boldsymbol{\Sigma}^\circ(\mathbf{A}_1^\circ - \mathbf{A}_2^\circ))$$

### Lemma 3.2

$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$  and  $\mathbf{A}_1'\boldsymbol{\theta}$  are not independent.

**Proof:**

$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$  and  $\mathbf{A}_1'\boldsymbol{\theta}$  are independent if and only if  $\mathbf{A}_1'N\boldsymbol{\Sigma}^\circ\mathbf{B} = \mathbf{0}$ .

$$\begin{aligned}
\mathbf{A}_1'N\boldsymbol{\Sigma}^\circ\mathbf{B} & = \mathbf{A}_1' \frac{1}{NGK^2} \boldsymbol{\Sigma}^\circ \mathbf{B}^\circ \\
& = \frac{2}{K^2GN^{3/2}}(\mathbf{A}_{11}^* \dots \mathbf{A}_{1G}^*) \frac{1}{NGK^2} \left[ \left( (\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G} \mathbf{U}_{KG} \right) \otimes \mathbf{I}_C \right] \\
& = \frac{2}{K^3G^2N^{5/2}}(\mathbf{A}_{11}^*\boldsymbol{\Sigma}_1 \dots \mathbf{A}_{1G}^*\boldsymbol{\Sigma}_G) \left[ \left( (\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G} \mathbf{U}_{KG} \right) \otimes \mathbf{I}_C \right]
\end{aligned}$$

Let  $\mathbf{a}_{1g}^* = (\gamma_{1g} \dots \gamma_{Cg})$  be a  $1 \times C$  vector. Then write  $\mathbf{A}_{1g}^* = (\mathbf{a}_{1g}^* \dots \mathbf{a}_{1g}^*)$  and  $\mathbf{A}_{1g}^*\boldsymbol{\Sigma}_g$  can be written as

$$\mathbf{A}_{1g}^*\boldsymbol{\Sigma}_g = (\mathbf{a}_{1g}\boldsymbol{\Sigma}_{g1} \mathbf{a}_{1g}\boldsymbol{\Sigma}_{g2} \dots \mathbf{a}_{1g}\boldsymbol{\Sigma}_{gK}) \text{ for each } g = 1, \dots, G$$



Now,

$$\mathbf{a}_{1g}\boldsymbol{\Sigma}_{1k} = \left[ p_{1gk}(\gamma_{1g} - \sum_{c=1}^C \gamma_{cg} \cdot p_{c1k}) \cdots p_{Cgk}(\gamma_{Cg} - \sum_{c=1}^C \gamma_{cg} \cdot p_{c1k}) \right]$$

for each  $g = 1, \dots, G$  and  $k = 1, \dots, K$ . The first element of  $\mathbf{A}'_1 N \boldsymbol{\Sigma} \mathbf{A}$  is

$$\begin{aligned} & \frac{2}{K^3 G^2 N^{5/2}} \left( \frac{G-1}{G} \sum_{k=1}^K p_{11k}(\gamma_{11} - \sum_{c=1}^C \gamma_{c1} \cdot p_{c1k}) - \frac{1}{G} \sum_{g=2}^G \sum_{k=1}^K p_{1gk}(\gamma_{1g} - \sum_{c=1}^C \gamma_{cg} \cdot p_{c1k}) \right) \\ &= \frac{2}{K^3 G^2 N^{5/2}} \left( \sum_{k=1}^K p_{11k}(\gamma_{11} - \sum_{c=1}^C \gamma_{c1} \cdot p_{c1k}) - \frac{1}{G} \sum_{g=1}^G \sum_{k=1}^K p_{1gk}(\gamma_{1g} - \sum_{c=1}^C \gamma_{cg} \cdot p_{c1k}) \right) \neq 0 \end{aligned}$$

■

Since  $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$  and  $\mathbf{A}'_1\boldsymbol{\theta}$  are not independent,  $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$  and  $(\mathbf{A}_1 - \mathbf{A}_2)'\boldsymbol{\theta}$  are also not independent. So, the distribution of  $\mathbf{V}'\mathbf{B}\mathbf{V}$  is not the convolution of a linear combination of  $\chi^2$ -random variables and a normal distribution.

Let  $\mathbf{A} = \mathbf{A}_1 - \mathbf{A}_2$ , then

$$\begin{aligned} \mathbf{V}'\mathbf{B}\mathbf{V} &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}'\boldsymbol{\theta} + N^2\mathbf{A}'\mathbf{A} \\ &= (\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A})'(\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}) - \frac{1}{4}\mathbf{A}'\mathbf{B}^{-1}\mathbf{A} + N^2\mathbf{A}'\mathbf{A} \\ &= (\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A})'(\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}) + \frac{1}{4}\mathbf{A}'(4N^2\mathbf{I} - \mathbf{B}^{-1})\mathbf{A} \end{aligned}$$

and let  $\mathbf{X} = (\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A})$ ,  $\boldsymbol{\Gamma} = \frac{1}{GNK^2}(\mathbf{B}^\circ)^{1/2}\boldsymbol{\Sigma}^\circ[(\mathbf{B}^\circ)^{1/2}]'$  and  $\boldsymbol{\mu}_B = \frac{1}{2}\mathbf{B}^{1/2}\mathbf{A}$ .

Then,

$$\sqrt{N}\mathbf{X} \sim N(\sqrt{N}\boldsymbol{\mu}_B, N\boldsymbol{\Gamma})$$

Let  $\mathbf{P}$  be an orthogonal matrix (i.e.,  $\mathbf{P}'\mathbf{P} = \mathbf{I}$ ) such that  $\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}' = \boldsymbol{\Lambda}$ , where  $\boldsymbol{\Lambda}$  is a diagonal matrix, and

$$\mathbf{Y} = \mathbf{P}\mathbf{X} \Rightarrow \mathbf{X} = \mathbf{P}'\mathbf{Y}$$

Then,

$$\mathbf{Y} \sim N(\mathbf{P}\boldsymbol{\mu}_B, \boldsymbol{\Lambda})$$

and

$$\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{P}\mathbf{P}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} \sim \sum_{i=1}^G \lambda_i (\chi_1^2(\delta_i)), \quad (3.6.2)$$

where  $\lambda_i$ 's are the diagonal elements of  $\mathbf{\Lambda}$ ,  $\delta_i = \frac{a_i^2}{\lambda_i}$ , with  $a_i$  being the  $i$ th row of the vector  $\frac{1}{2}\mathbf{P}\mathbf{B}^{-1/2}\mathbf{A}$ , which are linear combination of the  $\gamma_{cgk}$ 's.

Let  $c$  be a constant  $\frac{1}{4}\mathbf{A}'(4N^2\mathbf{I} - \mathbf{B}^{-1})\mathbf{A}$ . Then,

$$\begin{aligned}\Pr(F_1 \geq u) &= \Pr\left(\frac{N(\mathbf{X}'\mathbf{X} + c)}{\theta_3^o} \geq u\right) \\ &= \Pr(N\mathbf{X}'\mathbf{X} \geq \theta_3^o u - Nc)\end{aligned}\tag{3.6.3}$$

Note that  $Nc = O(1)$  since  $\mathbf{B}^{-1} = O(N^2)$  and  $\mathbf{A}'\mathbf{A} = O(N^{-3})$ . As the noncentrality parameter  $\delta_i$  increases, the distribution of each of the noncentral  $\chi^2$ -random variables shifts to the right, therefore the probability in (3.6.3) goes to 1 and the power of the test converges to 1.

## Chapter 4

# Analysis of Variance based on the Hamming Distance

The interest here still lies in the comparison of sequences. Now they are considered on an individual basis in that they are compared to each other: all possible pairwise comparisons within and across groups are performed. We develop an analysis-of-variance framework for Hamming distances and estimate the variability between, within and across groups. 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. 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.

## 4.1 The Total Sum of Squares and its decomposition

Let  $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \dots, X_{ik}^g)'$  be a random vector representing sequence  $i$  of group  $g$ . Suppose  $i = 1, \dots, N$ ,  $k = 1, \dots, K$  and  $g = 1, \dots, G$ . 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 4.1

The *Hamming Distance*  $D_{ij}^{(g_1, g_2)}$  is a descriptive statistic for sequence comparison:

$$\begin{aligned} D_{ij}^{(g_1, g_2)} &= \frac{1}{K} \sum_{k=1}^K I(X_{ik}^{g_1} \neq X_{jk}^{g_2}) \\ &= \frac{1}{K} \times (\text{number of positions where } X_i^{g_1} \text{ and } X_j^{g_2} \text{ differ}), \end{aligned} \quad (4.1.1)$$

and when  $g_1 = g_2 = g$ ,

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

### Definition 4.2

Let  $X_1, X_2, \dots$  be independent observations with distribution  $F$ . They may be vec-

tors, but for simplicity we confine our attention to the scalar case. Consider a *parameter function*  $\theta = \theta(F)$  for which there is an unbiased estimator, i.e.,  $\theta(F)$  may be represented as

$$\theta(F) = E_F\{\phi(X_1, \dots, X_m)\} = \int \dots \int \phi(x_1, \dots, x_m) dF(x_1) \dots dF(x_m), \quad (4.1.2)$$

for some function  $\phi = \phi(x_1, \dots, x_m)$ , called a *kernel*. Without loss of generality, assume that  $\phi$  is *symmetric*. For, if not, it may be replaced by the *symmetric kernel*

$$\frac{1}{m!} \sum_p \phi(x_{i_1}, \dots, x_{i_m}), \quad (4.1.3)$$

where  $\sum_p$  denotes the summation over the  $m!$  permutations  $(i_1, \dots, i_m)$  of  $(1, \dots, m)$ . ■

### Definition 4.3

For any kernel  $\phi$ , the corresponding *U-statistic* for estimation of  $\theta$  on the basis of a sample  $X_1, \dots, X_n$  of size  $n \geq m$  is obtained by averaging the kernel  $\phi$  symmetrically over the observations:

$$U_n = U(X_1, \dots, X_n) = \frac{1}{\binom{n}{m}} \sum_c \phi(X_{i_1}, \dots, X_{i_m}), \quad (4.1.4)$$

where  $\sum_c$  denotes the summation over the  $\binom{n}{m}$  combinations of  $m$  distinct elements  $\{i_1, \dots, i_m\}$  from  $\{1, \dots, n\}$ . This U-statistic is said to be of *degree*  $m$ . Clearly,  $U_n$  is an *unbiased* estimate of  $\theta$ . ■

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

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

Define the average distance within a group as

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

which is a U-statistic of degree 2. The average distance between two groups is

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

which, as we will see later, is a two-sample U-statistics of degree (1,1). The overall distance is

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

which is a linear combination of U-statistics.

The Total Sum of Squares

$$TSS = \sum_{g=1}^G \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)^2 + \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^{(g_1, g_2)} - \bar{D}^{(g_1, g_2)})^2 \quad (4.1.5)$$

can be decomposed as follows

$$\begin{aligned} &\sum_{g=1}^G \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)^2 + \sum_{g=1}^G \sum_{1 \leq i < j \leq N} (\bar{D}^g - \bar{D})^2 \\ &+ 2 \sum_{g=1}^G \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)(\bar{D}^g - \bar{D}) \\ &+ \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^{(g_1, g_2)} - \bar{D}^{(g_1, g_2)})^2 + \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (\bar{D}^{(g_1, g_2)} - \bar{D})^2 \\ &+ 2 \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^{(g_1, g_2)} - \bar{D}^{(g_1, g_2)})(\bar{D}^{(g_1, g_2)} - \bar{D}). \end{aligned}$$

Since

$$\sum_{g=1}^G \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)(\bar{D}^g - \bar{D}) = 0$$

and

$$\sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^{(g_1, g_2)} - \bar{D}^{(g_1, g_2)})(\bar{D}^{(g_1, g_2)} - \bar{D}) = 0,$$

$$TSS = \sum_{g=1}^G \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)^2 + \sum_{g=1}^G \sum_{1 \leq i < j \leq N} (\bar{D}^g - \bar{D})^2$$

$$\begin{aligned}
& + \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^{(g_1, g_2)} - \bar{D}^{(g_1, g_2)})^2 + \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (\bar{D}^{(g_1, g_2)} - \bar{D})^2 \\
& = \text{Within Sum of Squares (WSS)} + \text{Between Sum of Squares (BSS)} \\
& + \text{Across Within Sum of Squares (AWSS)} \\
& + \text{Across Between Sum of Squares (ABSS)}
\end{aligned}$$

## 4.2 Connections Between Sums of Squares and U-Statistics

The extension of U-statistics to the case of several samples is straightforward.

### Definition 4.4

Consider  $k$  independent collections of independent observations  $\{X_1^{(1)}, X_2^{(1)}, \dots\}, \dots, \{X_1^{(k)}, X_2^{(k)}, \dots\}$  taken from distributions  $F^{(1)}, \dots, F^{(k)}$ , respectively.

Let  $\theta = \theta(F^{(1)}, \dots, F^{(k)})$  denote a parametric function for which there is an unbiased estimator, i.e.,

$$\theta = E\{\phi(X_1^{(1)}, \dots, X_{m_1}^{(1)}, \dots; X_1^{(k)}, \dots, X_{m_k}^{(k)})\},$$

where  $\phi$  is assumed, without loss of generality, to be symmetric within each of its  $k$  blocks of arguments. Corresponding to the kernel  $\phi$  and assuming  $n_1 \geq m_1, \dots, n_k \geq m_k$ , the  $k$ -sample U-statistic of degree  $(m_1, m_2, \dots, m_k)$  for estimation of  $\theta$  is

$$U_n = \frac{1}{\prod_{j=1}^k \binom{n_j}{m_j}} \sum_c \phi(X_{i_{11}}^{(1)}, \dots, X_{i_{1m_1}}^{(1)}; \dots; X_{i_{k1}}^{(k)}, \dots, X_{i_{km_k}}^{(k)}) \quad (4.2.1)$$

Here  $\{i_{j1}, \dots, i_{jm_j}\}$  denotes a set of  $m_j$  distinct elements of the set  $\{1, 2, \dots, n_j\}$ ,  $1 \leq j \leq k$ , and  $\sum_c$  denotes the summation over all such combinations. ■

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

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

$$\begin{aligned}
&= \frac{2}{NG(NG-1)} \left\{ \sum_{g=1}^G \left[ \sum_{1 \leq i < j < j' \leq N} (D_{ij}^g - D_{ij'}^g)^2 \right. \right. \\
&+ \sum_{1 \leq i < i' < j \leq N} (D_{ij}^g - D_{i'j}^g)^2 + \sum_{1 \leq i < j < j' \leq N} (D_{ij}^g - D_{jj'}^g)^2 \\
&+ \sum_{1 \leq i < j < i' < j' \leq N} (D_{ij}^g - D_{i'j'}^g)^2 + \sum_{1 \leq i < i' < j < j' \leq N} (D_{ij}^g - D_{i'j'}^g)^2 \\
&+ \left. \sum_{1 \leq i < i' < j' < j \leq N} (D_{ij}^g - D_{i'j'}^g)^2 \right] \\
&+ \sum_{1 \leq g_1 < g_2 \leq G} \left[ \sum_{1 \leq i < j \leq N} \sum_{1 \leq i' < j' \leq N} (D_{ij}^{g_1} - D_{i'j'}^{g_2})^2 \right] \\
&+ \sum_{1 \leq g_1 < g_2 \leq G} \left[ \sum_{1 \leq i < i' \leq N} (D_{ii}^{(g_1, g_2)} - D_{i'i'}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, j' \leq N \\ i \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 \right. \\
&+ \sum_{\substack{1 \leq i, i' \leq N \\ i \neq i'}} (D_{ii}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_2)})^2 + \sum_{1 \leq i < j \leq N} (D_{ij}^{(g_1, g_2)} - D_{ji}^{(g_1, g_2)})^2 \\
&+ \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{jj'}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 \\
&+ \sum_{\substack{1 \leq i, j, i' \leq N \\ i \neq j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, i', j' \leq N \\ i \neq i' \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 \\
&+ \left. \sum_{\substack{1 \leq i, j, i', j' \leq N \\ i \neq j \neq i' \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 \right] \\
&+ \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \left[ \sum_{\substack{1 \leq i, j' \leq N \\ i \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{j'j'}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, j' \leq N \\ i \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2 \right. \\
&+ \sum_{i=1}^N (D_{ii}^{(g_1, g_2)} - D_{ii}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, j \leq N \\ i \neq j}} (D_{ij}^{(g_1, g_2)} - D_{ij}^{(g_1, g_3)})^2 \\
&+ \sum_{\substack{1 \leq i, j \leq N \\ i \neq j}} (D_{ij}^{(g_1, g_2)} - D_{ii}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, j \leq N \\ i \neq j}} (D_{ij}^{(g_1, g_2)} - D_{jj}^{(g_1, g_3)})^2 \\
&+ \sum_{\substack{1 \leq i, i' \leq N \\ i \neq i'}} (D_{ii}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, j \leq N \\ i \neq j}} (D_{ij}^{(g_1, g_2)} - D_{ji}^{(g_1, g_3)})^2 \\
&+ \sum_{\substack{1 \leq i, j, i' \leq N \\ i \neq j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, j, i' \leq N \\ i \neq j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_3)})^2 \\
&+ \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{jj'}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2 \\
&+ \left. \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2 \right]
\end{aligned}$$



$$\begin{aligned}
& + \sum_{\substack{1 \leq i, j, i' \leq N \\ i \neq j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 + \sum_{\substack{1 \leq i, i', j' \leq N \\ i \neq i' \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 \\
& + \sum_{\substack{1 \leq i, j, i', j' \leq N \\ i \neq j \neq i' \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 \Big] \\
& + \sum_{\substack{1 \leq \theta_1 \neq \theta_2 \neq \theta_3 \neq \theta_4 \leq G \\ \theta_1 < \theta_2, \theta_3 < \theta_4}} \left[ \sum_{i=1}^N \sum_{j=1}^N \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_3, g_4)})^2 \right] \\
& + \sum_{\substack{1 \leq \theta_1, \theta_2 \leq G \\ \theta_1 \neq \theta_2}} \left[ \sum_{1 \leq i < j \leq N} (D_{ij}^{g_1} - D_{ii}^{(g_1, g_2)})^2 + \sum_{1 \leq i < j \leq N} (D_{ij}^{g_1} - D_{jj}^{(g_1, g_2)})^2 \right. \\
& + \sum_{\substack{1 \leq i, j, i' \leq N \\ i' \neq j \neq i', i < j}} (D_{ij}^{g_1} - D_{i'i}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, j, j' \leq N \\ j' \neq i \neq j \neq j', i < j}} (D_{ij}^{g_1} - D_{j'j'}^{(g_1, g_2)})^2 \\
& + \sum_{1 \leq i < j \leq N} (D_{ij}^{g_1} - D_{ji}^{(g_1, g_2)})^2 + \sum_{1 \leq i < j \leq N} (D_{ij}^{g_1} - D_{ij}^{(g_1, g_2)})^2 \\
& + \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j', i < j}} (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, j, j' \leq N \\ i \neq j \neq j', i < j}} (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \\
& + \sum_{\substack{1 \leq i, j, i' \leq N \\ i \neq j \neq i', i < j}} (D_{ij}^{g_1} - D_{i'j}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, j, i', j' \leq N \\ i \neq j \neq i' \neq j', i < j}} (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \Big] \\
& + \sum_{\substack{1 \leq \theta_1 \neq \theta_2 \neq \theta_3 \leq G \\ \theta_2 < \theta_3}} \left[ \sum_{1 \leq i < j \leq N} \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_2, g_3)})^2 \right] \Big\} \tag{4.2.2}
\end{aligned}$$

Note that the total number of terms in (4.1.5) is

$$\begin{aligned}
\binom{\frac{NG(NG-1)}{2}}{2} &= \frac{1}{8} NG(NG-1)[NG(NG-1) - 2] \\
&= \frac{NG}{8} [N^3 G^3 - 2N^2 G^2 - NG + 2]
\end{aligned}$$

Separating the different U-statistics and adding up the number of terms we get

$$\begin{aligned}
& G \binom{\frac{N(N-1)}{2}}{2} + \frac{G(G-1)}{2} \binom{N}{2} + \frac{G(G-1)}{2} \binom{N^2}{2} \\
& + \frac{G(G-1)(G-2)}{2} N^4 + \frac{G(G-1)(G-2)(G-3)}{8} N^4 \\
& + G(G-1) \left[ \binom{N}{2} N^2 \right] + \frac{G(G-1)(G-2)}{2} \left[ \binom{N}{2} N^2 \right] \\
& = \frac{NG}{8} [N^3 G^3 - 2N^2 G^2 - NG + 2]
\end{aligned}$$

We now characterize each term in the above decomposition (4.2.2):

- one-sample U-statistics of degree 3:

$$U_{1,1}^{(3)} = \binom{N}{3}^{-1} \sum_{i < j < j'} (D_{ij}^g - D_{ij'}^g)^2, \quad U_{1,2}^{(3)} = \binom{N}{3}^{-1} \sum_{i < i' < j} (D_{ij}^g - D_{i'j}^g)^2 \quad \text{and}$$

$$U_{1,3}^{(3)} = \binom{N}{3}^{-1} \sum_{i < j < j'} (D_{ij}^g - D_{jj'}^g)^2$$

- one-sample U-statistics of degree 4:

$$U_{2,1}^{(4)} = \binom{N}{4}^{-1} \sum_{i < j < i' < j'} (D_{ij}^g - D_{i'j'}^g)^2, \quad U_{2,2}^{(4)} = \binom{N}{4}^{-1} \sum_{i < i' < j < j'} (D_{ij}^g - D_{i'j'}^g)^2 \quad \text{and}$$

$$U_{2,3}^{(4)} = \binom{N}{4}^{-1} \sum_{i < i' < j' < j} (D_{ij}^g - D_{i'j'}^g)^2$$

- two-sample U-statistics of degree (2,2):

$$U_{3}^{(2,2)} = \left[ \binom{N}{2} \binom{N}{2} \right]^{-1} \sum_{i < j} \sum_{i' < j'} (D_{ij}^{g_1} - D_{i'j'}^{g_2})^2$$

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

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

since

$$\sum_{\substack{i \neq i' \\ j \neq j'}} \sum_{\substack{j \neq j' \\ j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 = \sum_{i < i'} (D_{ii}^{(g_1, g_2)} - D_{i'i'}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, j', i', \leq N \\ i \neq i' \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2$$

$$+ \sum_{\substack{i \neq j \neq i' \neq j' \\ j < i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2$$

has  $\binom{N}{2} + N(N-1)(N-2) + \frac{\binom{N}{1}\binom{N-1}\binom{N-3}}{2} = \binom{N}{2} \binom{N}{2}$  terms and

$$\sum_{\substack{i \neq j \\ i \neq i'}} \sum_{\substack{j \neq j' \\ j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 = \sum_{i < j} (D_{ij}^{(g_1, g_2)} - D_{ji}^{(g_1, g_2)})^2 + \sum_{i \neq j \neq j'} (D_{ij}^{(g_1, g_2)} - D_{jj'}^{(g_1, g_2)})^2$$

$$+ \sum_{\substack{i \neq j \neq i' \neq j' \\ j > i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2$$

has  $\binom{N}{2} + N(N-1)(N-2) + \frac{\binom{N}{1}\binom{N-1}{2}\binom{N-3}{1}}{2} = \binom{N}{2}\binom{N}{2}$  terms.

• two-sample U-statistic of degree (1,2):

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

since

$$\begin{aligned} \sum_{\substack{i=1 \\ i \neq j'}}^N \sum_{\substack{1 \leq j, j' \leq N \\ j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 &= \sum_{\substack{1 \leq i, j \leq N \\ i \neq j}} (D_{ii}^{(g_1, g_2)} - D_{ij}^{(g_1, g_2)})^2 \\ &+ \sum_{i \neq j \neq j'} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 \end{aligned}$$

has  $N(N-1) + \frac{\binom{N}{1}\binom{N-1}{1}\binom{N-2}{1}}{2} = \binom{N}{1}\binom{N}{2}$  terms.

• two-sample U-statistics of degree (2,1):

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

$$U_{6,1}^{(2,1)} = \left[ \binom{N}{2} \binom{N}{1} \right]^{-1} \sum_{1 \leq i < j \leq N} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \quad \text{and}$$

$$U_{6,2}^{(2,1)} = \left[ \binom{N}{2} \binom{N}{1} \right]^{-1} \sum_{1 \leq i < j \leq N} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2$$

since

$$\sum_{\substack{j=1 \\ j \neq i'}}^N \sum_{i \neq i'} (D_{ij}^{(g_1, g_2)} - D_{i'j}^{(g_1, g_2)})^2 = \sum_{i \neq i'} (D_{ii}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_2)})^2 + \sum_{i \neq j \neq i'} (D_{ij}^{(g_1, g_2)} - D_{i'j}^{(g_1, g_2)})^2$$

has  $N(N-1) + \frac{N(N-1)(N-2)}{2} = \binom{N}{2}\binom{N}{1}$  terms,

$$\begin{aligned} \sum_{1 \leq i < j \leq N} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 &= \sum_{i < j} (D_{ij}^{g_1} - D_{ii}^{(g_1, g_2)})^2 + \sum_{i < j} (D_{ij}^{g_1} - D_{ij}^{(g_1, g_2)})^2 \\ &+ \sum_{i < j} \sum_{\substack{j' \\ j' \neq j \neq i}} (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \end{aligned}$$

has  $\binom{N}{2} + \binom{N}{2} + \binom{N}{2}\binom{N-2}{1} = \binom{N}{2}\binom{N}{1}$  terms, and

$$\begin{aligned} \sum_{1 \leq i < j \leq N} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 &= \sum_{i < j} (D_{ij}^{g_1} - D_{jj}^{(g_1, g_2)})^2 \\ &+ \sum_{i < j} (D_{ij}^{g_1} - D_{ji}^{(g_1, g_2)})^2 + \sum_{i < j} \sum_{\substack{1 \leq j' \leq N \\ i \neq j \neq j'}} (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 \end{aligned}$$

has  $\binom{N}{2} + \binom{N}{2} + \binom{N}{2} \binom{N-2}{1} = \binom{N}{2} \binom{N}{1}$  terms.

• two-sample U-statistics of degree (3,1):

$$U_{7,1}^{(3,1)} = \left[ \binom{N}{3} \binom{N}{1} \right]^{-1} \sum_{i < j < i'} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2,$$

$$U_{7,2}^{(3,1)} = \left[ \binom{N}{3} \binom{N}{1} \right]^{-1} \sum_{i' < i < j} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \quad \text{and}$$

$$U_{7,3}^{(3,1)} = \left[ \binom{N}{3} \binom{N}{1} \right]^{-1} \sum_{i < i' < j} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2$$

since

$$\begin{aligned} \sum_{\substack{i \neq j \neq i' \\ i < j}} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 &= \sum_{\substack{i \neq j \neq j' \\ i < j}} (D_{ij}^{g_1} - D_{j'j'}^{(g_1, g_2)})^2 + \sum_{\substack{i \neq j \neq i' \\ i < j}} (D_{ij}^{g_1} - D_{i'j}^{(g_1, g_2)})^2 \\ &\quad + \sum_{\substack{i \neq j \neq i' \\ i < j}} (D_{ij}^{g_1} - D_{i'i}^{(g_1, g_2)})^2 + \sum_{\substack{i \neq j \neq i' \neq j' \\ i < j}} (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \end{aligned}$$

has  $\binom{N}{1} \binom{N}{2} \binom{N-2}{1} = \binom{N}{2} \binom{N-2}{1} + \binom{N}{2} \binom{N-2}{1} + \binom{N}{2} \binom{N-2}{1} + \binom{N}{2} \binom{N-2}{1} \binom{N-3}{1} = 3 \binom{N}{3} \binom{N}{1}$  terms, and

$$\begin{aligned} \sum_{\substack{i \neq j \neq i' \\ i < j}} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 &= \sum_{i < j < i'} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 + \sum_{i' < i < j} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \\ &\quad + \sum_{i < i' < j} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \quad \text{has } 3 \binom{N}{3} \binom{N}{1} \text{ terms.} \end{aligned}$$

• three-sample U-statistics of degree (2,1,1):

$$U_8^{(2,1,1)} = \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i < j} \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_2, g_3)})^2$$

$$U_{9,1}^{(2,1,1)} = \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i \neq i'} \sum_{\substack{j=1 \\ j \neq i'}}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2$$

$$U_{9,2}^{(2,1,1)} = \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i \neq j} \sum_{\substack{i'=1 \\ i' \neq i}}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2$$

since

$$\sum_{\substack{i \neq i' \\ j \neq i'}} \sum_{j=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2$$

$$\begin{aligned}
&= \sum_{i \neq i'} (D_{ii}^{(g_1, g_2)} - D_{i'i'}^{(g_1, g_3)})^2 + \sum_{i \neq i'} (D_{ii}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_3)})^2 \\
&\quad + \sum_{i \neq i' \neq j'} (D_{ii}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 + \sum_{\substack{i \neq j \neq i' \\ j < i'}} (D_{ij}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_3)})^2 \\
&\quad + \sum_{i \neq j \neq i'} (D_{ij}^{(g_1, g_2)} - D_{i'i'}^{(g_1, g_3)})^2 + \sum_{\substack{i \neq j \neq i' \neq j' \\ j < i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2
\end{aligned}$$

has  $N(N-1) + N(N-1) + N(N-1)(N-2) + \frac{N(N-1)(N-2)}{2} + N(N-1)(N-2) + \frac{N(N-1)(N-2)(N-3)}{2} = \binom{N}{2} \binom{N}{1} \binom{N}{1}$  terms, and

$$\begin{aligned}
&\sum_{i \neq j} \sum_{\substack{i'=1 \\ i \neq i'}}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 \\
&= \sum_{i \neq j} (D_{ij}^{(g_1, g_2)} - D_{ji}^{(g_1, g_3)})^2 + \sum_{i \neq j} (D_{ij}^{(g_1, g_2)} - D_{jj}^{(g_1, g_3)})^2 \\
&\quad + \sum_{i \neq j \neq j'} (D_{ij}^{(g_1, g_2)} - D_{jj'}^{(g_1, g_3)})^2 + \sum_{i \neq j \neq i'} (D_{ij}^{(g_1, g_2)} - D_{i'j}^{(g_1, g_3)})^2 \\
&\quad + \sum_{\substack{i \neq j \neq i' \\ j > i'}} (D_{ij}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_3)})^2 + \sum_{\substack{i \neq j \neq i' \neq j' \\ j > i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2
\end{aligned}$$

has  $N(N-1) + N(N-1) + N(N-1)(N-2) + N(N-1)(N-2) + \frac{N(N-1)(N-2)}{2} + \frac{N(N-1)(N-2)(N-3)}{2} = \binom{N}{2} \binom{N}{1} \binom{N}{1}$  terms.

• three-sample U-statistic of degree (1,1,1):

$$\mathbf{U}_{10}^{(1,1,1)} = \left[ \binom{N}{1} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i=1}^N \sum_{j=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2, \text{ since}$$

$$\begin{aligned}
&\sum_{i=1}^N \sum_{j=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2 \\
&= \sum_{i \neq j} (D_{ij}^{(g_1, g_2)} - D_{ij}^{(g_1, g_3)})^2 + \sum_{i=1}^N (D_{ii}^{(g_1, g_2)} - D_{ii}^{(g_1, g_3)})^2 \\
&\quad + \sum_{i \neq j'} (D_{ii}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2 + \sum_{i \neq j} (D_{ij}^{(g_1, g_2)} - D_{ii}^{(g_1, g_3)})^2 \\
&\quad + \sum_{i \neq j \neq j'} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2
\end{aligned}$$

has  $N(N-1) + N + N(N-1) + N(N-1) + N(N-1)(N-2) = \binom{N}{1} \binom{N}{1} \binom{N}{1}$  terms.

• four-sample U-statistic of degree (1,1,1,1):

$$U_{11}^{(1,1,1,1)} = N^{-4} \sum_{i=1}^N \sum_{j=1}^N \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_3, g_4)})^2$$

Further,

$$WSS = \sum_{g=1}^G \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)^2$$

can be shown to be a U-statistic.

For each  $g$ , we can write

$$\begin{aligned} WSS_g &= \sum_{1 \leq i < j \leq N} (D_{ij}^g - \bar{D}^g)^2 \\ &= \left( \frac{N(N-1)}{2} - 1 \right) \left( \frac{N(N-1)}{2} \right)^{-1} \sum_{\substack{i < j, i' < j' \\ i \leq i' \text{ or } j \leq j'}} \frac{(D_{ij}^g - D_{i'j'}^g)^2}{2} \\ &= \left( \frac{N(N-1)}{2} - 1 \right) \left( \frac{N(N-1)}{2} \right)^{-1} \frac{1}{2} \left[ \sum_{1 \leq i < j < j' \leq N} (D_{ij}^g - D_{ij'}^g)^2 \right. \\ &\quad + \sum_{1 \leq i < i' < j \leq N} (D_{ij}^g - D_{i'j}^g)^2 + \sum_{1 \leq i < j < j' \leq N} (D_{ij}^g - D_{jj'}^g)^2 \\ &\quad + \sum_{1 \leq i < j < i' < j' \leq N} (D_{ij}^g - D_{i'j'}^g)^2 + \sum_{1 \leq i < i' < j < j' \leq N} (D_{ij}^g - D_{i'j'}^g)^2 \\ &\quad \left. + \sum_{1 \leq i < i' < j' < j \leq N} (D_{ij}^g - D_{i'j'}^g)^2 \right] \end{aligned}$$

Therefore,

$$\begin{aligned} WSS &= \left( \frac{N(N-1)}{2} - 1 \right) \left( \frac{N(N-1)}{2} \right)^{-1} \sum_{g=1}^G \sum_{\substack{i < j, i' < j' \\ i \leq i' \text{ or } j \leq j'}} \frac{(D_{ij}^g - D_{i'j'}^g)^2}{2} \\ &= \frac{4}{N(N-1)} \sum_{g=1}^G \frac{1}{2} \left[ \sum_{1 \leq i < j < j' \leq N} (D_{ij}^g - D_{ij'}^g)^2 \right. \\ &\quad + \sum_{1 \leq i < i' < j \leq N} (D_{ij}^g - D_{i'j}^g)^2 + \sum_{1 \leq i < j < j' \leq N} (D_{ij}^g - D_{jj'}^g)^2 \\ &\quad + \sum_{1 \leq i < j < i' < j' \leq N} (D_{ij}^g - D_{i'j'}^g)^2 + \sum_{1 \leq i < i' < j < j' \leq N} (D_{ij}^g - D_{i'j'}^g)^2 \\ &\quad \left. + \sum_{1 \leq i < i' < j' < j \leq N} (D_{ij}^g - D_{i'j'}^g)^2 \right] \\ &= \frac{2}{N(N-1)} \sum_{g=1}^G \left[ \binom{N}{3} (U_{1,1}^{(3)} + U_{1,2}^{(3)} + U_{1,3}^{(3)}) + \binom{N}{4} (U_{2,1}^{(4)} + U_{2,2}^{(4)} + U_{2,3}^{(4)}) \right] \end{aligned}$$

Now,

$$\begin{aligned}
AWSS &= \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^{(g_1, g_2)} - \bar{D}^{(g_1, g_2)})^2 \\
&= (N^2 - 1) \binom{N^2}{2}^{-1} \sum_{1 \leq g_1 < g_2 \leq G} \sum_{i=1}^N \sum_{j=1}^N \sum_{i'=1}^N \sum_{\substack{j'=1 \\ i \leq i' \text{ or } j \leq j'}}^N \frac{(D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2}{2} \\
&= \frac{2}{N^2} \sum_{1 \leq g_1 < g_2 \leq G} \frac{1}{2} \left[ \sum_{1 \leq i < i' \leq N} (D_{ii}^{(g_1, g_2)} - D_{i'i'}^{(g_1, g_2)})^2 \right. \\
&\quad + \sum_{\substack{1 \leq i, j' \leq N \\ i \neq j'}} (D_{ii}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 + \sum_{\substack{1 \leq i, i' \leq N \\ i \neq i'}} (D_{ii}^{(g_1, g_2)} - D_{i'i}^{(g_1, g_2)})^2 \\
&\quad + \sum_{1 \leq i < j \leq N} (D_{ij}^{(g_1, g_2)} - D_{ji}^{(g_1, g_2)})^2 + \sum_{i \neq j \neq j'} (D_{ij}^{(g_1, g_2)} - D_{jj'}^{(g_1, g_2)})^2 \\
&\quad + \sum_{i \neq j \neq j'} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 + \sum_{i \neq j \neq i'} (D_{ij}^{(g_1, g_2)} - D_{i'j}^{(g_1, g_2)})^2 \\
&\quad \left. + \sum_{i \neq i' \neq j'} (D_{ii}^{(g_1, g_2)} - D_{i'i'}^{(g_1, g_2)})^2 + \sum_{i \neq j \neq i' \neq j'} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 \right] \\
&= \frac{1}{N^2} \sum_{1 \leq g_1 < g_2 \leq G} \left[ \binom{N}{2} \binom{N}{2} (\mathbf{U}_{4,1}^{(2,2)} + \mathbf{U}_{4,2}^{(2,2)}) \right. \\
&\quad \left. + \binom{N}{2} \binom{N}{1} (\mathbf{U}_{5,2}^{(2,1)} + \mathbf{U}_{5,1}^{(1,2)}) \right]
\end{aligned}$$

## 4.3 Asymptotic Distributions

### 4.3.1 One sample U-statistics

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

$$\begin{aligned}
U_n &= U(X_1, \dots, X_n) = n^{-[m]} \sum_{1 \leq i_1 \neq \dots \neq i_m \leq n} \phi(X_{i_1}, \dots, X_{i_m}) \\
&= \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} \phi(X_{i_1}, \dots, X_{i_m}), \quad n \geq m \tag{4.3.1}
\end{aligned}$$

where  $n^{-[m]} = (n^{[m]})^{-1} = \{n \dots (n - m + 1)\}^{-1}$ . From (4.1.2) we know that

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

**Definition 4.5** von Mises (1947)

Define a von Mises' *differentiable statistical function* as

$$\begin{aligned}\theta(F_n) &= \int \cdots \int \phi(x_1, \dots, x_m) dF_n(x_1) \cdots dF_n(x_m) \\ &= n^{-m} \sum_{i_1=1}^n \cdots \sum_{i_m=1}^n \phi(X_{i_1}, \dots, X_{i_m}) \quad n \geq 1\end{aligned}\tag{4.3.2}$$

where  $F_n(x)$  is the empirical d.f.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n c(x - X_i) \quad x \in R^p, \quad n \geq 1$$

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

Let

$$\Psi_c(x_1, \dots, x_c) \equiv E\{\phi(x_1, \dots, x_c, X_{c+1}, \dots, X_m)\}\tag{4.3.3}$$

$$\psi_c(x_1, \dots, x_c) \equiv E\{\phi(x_1, \dots, x_c, X_{c+1}, \dots, X_m) - \theta\},\tag{4.3.4}$$

$$\xi_c \equiv E\{\psi_c^2(X_1, \dots, X_c)\} = E\{\Psi_c^2(X_1, \dots, X_c)\} - \theta^2 \quad \text{and} \quad \xi_0 \equiv 0.\tag{4.3.5}$$

**Theorem 4.1**

The function  $\Psi_c$  defined in (4.3.3) has the properties

- (i)  $\Psi_c(x_1, \dots, x_c) = E\{\Psi_d(x_1, \dots, x_c, X_{c+1}, \dots, X_d)\}$  for  $1 \leq c < d \leq m$ ,
- (ii)  $E\{\Psi_c(x_1, \dots, x_c)\} = E\{\phi(X_1, \dots, X_m)\}$ . ■

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

By (4.3.1) and (4.3.4),

$$\begin{aligned}\text{Var}(U_n) &= \binom{n}{m}^{-2} \text{Var}\left\{ \sum_{1 \leq i_1 < \cdots < i_m \leq n} \phi(X_{i_1}, \dots, X_{i_m}) \right\} \\ &= \binom{n}{m}^{-2} \sum_{c=0}^m \sum^{(c)} \text{Cov}\{\phi(X_{i_1}, \dots, X_{i_m}) \phi(X_{j_1}, \dots, X_{j_m})\}\end{aligned}$$

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 (4.3.5), each term in  $\sum^{(c)}$  is equal to  $\xi_c$ . The number of terms in  $\sum^{(c)}$  is

$$\frac{n(n-1)\cdots(n-2m+c+1)}{c!(m-c)!(m-c)!} = \binom{m}{c} \binom{n-m}{m-c} \binom{n}{m} \quad (4.3.6)$$

Since  $\xi_0 = 0$ ,

$$\begin{aligned} \text{Var}(U_n) &= \binom{n}{m}^{-2} \sum_{c=0}^m \binom{n}{m} \binom{m}{c} \binom{n-m}{m-c} \xi_c \\ &= \binom{n}{m}^{-1} \sum_{c=1}^m \binom{m}{c} \binom{n-m}{m-c} \xi_c \end{aligned} \quad (4.3.7)$$

Hoeffding (1948) obtained the following inequality:

$$0 \leq \xi_c \leq \frac{c}{d} \xi_d \quad 1 \leq c < d \leq m \quad (4.3.8)$$

which leads to

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

Now, from (4.3.7) and (4.3.6)

$$\begin{aligned} \text{Var}(U_n) &= \binom{n}{m}^{-1} \left\{ \frac{m(n-m)!}{(m-1)!(n-2m+1)!} \xi_1 + \cdots + \xi_m \right\} \\ &= \frac{m(m-1)\cdots 1}{n(n-1)\cdots(n-m+1)} \\ &\quad \times \left\{ \frac{m(n-m)(n-m-1)\cdots(n-m-(m-1)+1)}{(m-1)!} \xi_1 + \cdots + \xi_m \right\} \\ &= \frac{m^2(n-m)(n-m-1)\cdots(n-2m+1+1)}{n(n-1)\cdots(n-m+1)} \xi_1 + \cdots \\ &\quad + \frac{m(m-1)\cdots 1}{n(n-1)\cdots(n-m+1)} \xi_m \\ &= \frac{m^2}{n} \left( \frac{n-m}{n-1} \right) \cdots \left( \frac{n-2m+2}{n-m+1} \right) \xi_1 + \cdots \\ &\quad + \frac{m!}{n(n-1)\cdots(n-m+1)} \xi_m \end{aligned}$$

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

$$\lim_{n \rightarrow \infty} n\text{Var}(U_n) = m^2\xi_1$$

or

$$\text{Var}(U_n) = m^2 n^{-1} \xi_1 + O(n^{-2}) \quad (4.3.9)$$

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

$$n^{1/2}(U_n - \theta) \xrightarrow{d} N(0, m^2 \xi_1), \quad (\text{Hoeffding, 1948}) \quad (4.3.10)$$

In order to compute the covariance of two U-statistics, say  $U_n^{(\gamma)}$  and  $U_n^{(\delta)}$  (U-statistics of degrees  $m_{(\gamma)}$  and  $m_{(\delta)}$  respectively), we need to take into account the common elements in the two U-statistics. Let

$$E(U_n^{(\gamma)}) = E\{\phi_{(\gamma)}(X_1, \dots, X_{m_{(\gamma)}})\} = \theta_{(\gamma)},$$

$$E(U_n^{(\delta)}) = E\{\phi_{(\delta)}(X_1, \dots, X_{m_{(\delta)}})\} = \theta_{(\delta)},$$

$$\begin{aligned} \psi_{(\gamma)c}(x_1, \dots, x_c) &= E\{\phi_{(\gamma)}(x_1, \dots, x_c, X_{c+1}, \dots, X_{m_{(\gamma)}}) - \theta_{(\gamma)}\} \\ &= E\{\phi_{(\gamma)}(x_1, \dots, x_c, X_{c+1}, \dots, X_{m_{(\gamma)}})\} - \theta_{(\gamma)} \end{aligned}$$

$$\Psi_{(\gamma)c}(x_1, \dots, x_c) = E\{\phi_{(\gamma)}(x_1, \dots, x_c, X_{c+1}, \dots, X_{m_{(\gamma)}})\}$$

and

$$\begin{aligned} \xi_c^{(\gamma, \delta)} &= E\{\psi_{(\gamma)c}(X_1, \dots, X_c) \psi_{(\delta)c}(X_1, \dots, X_c)\} \\ &= E\{\Psi_{(\gamma)c}(X_1, \dots, X_c) \psi_{(\delta)c}(X_1, \dots, X_c)\} - \theta_{(\gamma)} \theta_{(\delta)} \end{aligned}$$

where  $c = 1, \dots, m_{(\gamma)}$  and  $\gamma, \delta = 1, \dots, q$ .

If  $\gamma = \delta$ , we have

$$\xi_c^{(\gamma)} = \xi_c^{(\gamma, \gamma)} = E\{\psi_{(\gamma)c}^2(X_1, \dots, X_c)\}$$

which is the term needed for the variance of a U-statistic.

Let

$$\sigma(U_n^{(\gamma)}, U_n^{(\delta)}) = E\{(U_n^{(\gamma)} - \theta_{(\gamma)})(U_n^{(\delta)} - \theta_{(\delta)})\}$$

be the covariance of  $U_n^{(\gamma)}$  and  $U_n^{(\delta)}$ .

If  $m_{(\gamma)} \leq m_{(\delta)}$ ,

$$\sigma(U_n^{(\gamma)}, U_n^{(\delta)}) = \binom{n}{m_{(\gamma)}}^{-1} \sum_{c=1}^{m_{(\gamma)}} \binom{m_{(\delta)}}{c} \binom{n-m_{(\delta)}}{m_{(\delta)}-c} \xi_c^{(\gamma, \delta)}. \quad (4.3.11)$$

From (4.3.11)

$$\lim_{n \rightarrow \infty} n\sigma(U_n^{(\gamma)}, U_n^{(\delta)}) = m_{(\gamma)}m_{(\delta)}\xi_1^{(\gamma, \delta)} \quad (4.3.12)$$

Therefore, by (4.3.12) it is sufficient to compute  $\xi_1^{(\gamma, \delta)}$ , i.e., it is enough to consider only the case of one element in common between the kernels.

To obtain a decomposition of  $\theta(F_n)$  and  $U_n$ , for every  $1 \leq h \leq m$ , let

$$V_{n,h} \equiv \int_{R^{ph}} \cdots \int \Psi_h(x_1, \dots, x_h) \prod_{j=1}^h d[F_n(x_j) - F(x_j)] \quad x \in R^{ph} \quad (4.3.13)$$

with each  $x$  being a  $p$  dimensional vector and since there are  $h$  of them, the integral is in  $R^{ph}$ . Then,

$$V_{n,1} = n^{-1} \sum_{i=1}^n [\Psi_1(X_i) - \theta(F)]$$

Writing  $dF_n(x_i) = dF(x_i) + d[F_n(x_i) - F(x_i)]$ ,  $i = 1, \dots, m$ , we

$$\theta(F_n) = \theta(F) + \sum_{h=1}^m \binom{m}{h} V_{n,h}, \quad n \geq 1 \quad (4.3.14)$$

Similarly, we may rewrite (4.3.1) as

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

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

$$U_n = \theta(F) + \sum_{h=1}^m \binom{m}{h} U_{n,h} \quad n \geq m \quad (4.3.15)$$

where

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

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

$$\begin{aligned} \Psi_h^0(x_1, \dots, x_h) &= \Psi_h(x_1, \dots, x_h) - \sum_{j=1}^h \Psi_{h-1}(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_h) \\ &\quad + \cdots + (-1)^h \theta(F), \quad \forall (x_1, \dots, x_h) \in R^{ph}, \end{aligned} \quad (4.3.16)$$

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

$$U_{n,h} = \binom{n}{h}^{-1} \sum_{1 \leq i_1 < \dots < i_h \leq n} \Psi_h^\circ(X_{i_1}, \dots, X_{i_h}), \quad 1 \leq h \leq m \quad (4.3.17)$$

So, the  $U_{n,h}$  are themselves U-statistics. Note that for  $h = 2$ , we have

$$\begin{aligned} \mathbf{E}(U_{n,2}) &= \mathbf{E}(\Psi_2^\circ(X_1, X_2)) = \mathbf{E}(\Psi_2(X_1, X_2)) - \mathbf{E}(\Psi_1(X_1)) \\ &\quad - \mathbf{E}(\Psi_1(X_2)) + \theta(F) \\ &= \theta(F) - \theta(F) - \theta(F) + \theta(F) = 0 \end{aligned}$$

Let

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

Then

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

$$\xi_1^\circ \equiv \mathbf{E}[\Psi_{21}^\circ(X_1)]^2 - (\mathbf{E}(U_{n,2}))^2 = 0$$

and by (4.3.7),

$$\begin{aligned} \text{Var}(U_{n,2}) &= \frac{4(n-2)}{n(n-1)} \xi_1^\circ + \frac{2\xi_2^\circ}{n(n-1)} \\ &= \frac{4}{n} \xi_1^\circ + O(n^{-2}) \\ &= O(n^{-2}) \end{aligned} \quad (4.3.18)$$

Consequently,  $U_{n,2} = O_p(n^{-1})$ .

For  $h = 3$  we have

$$\begin{aligned} \mathbf{E}(U_{n,3}) &= \mathbf{E}(\Psi_3^\circ(X_1, X_2, X_3)) \\ &= \mathbf{E}(\Psi_3(X_1, X_2, X_3)) - \mathbf{E}(\Psi_2(X_1, X_2)) - \mathbf{E}(\Psi_2(X_1, X_3)) \\ &\quad - \mathbf{E}(\Psi_2(X_2, X_3)) + \mathbf{E}(\Psi_1(X_1)) + \mathbf{E}(\Psi_1(X_2)) + \mathbf{E}(\Psi_1(X_3)) - \theta(F) \\ &= 4\theta(F) - 4\theta(F) \\ &= 0 \end{aligned}$$

$$\begin{aligned}
\Psi_{31}^{\circ}(X_1) &= \mathbb{E}[\Psi_3^{\circ}(X_1, X_2, X_3) \mid X_1] \\
&= \mathbb{E}[\Psi_3(X_1, X_2, X_3) \mid X_1] - \mathbb{E}[\Psi_2(X_1, X_2) \mid X_1] - \mathbb{E}[\Psi_2(X_1, X_3) \mid X_1] \\
&\quad - \mathbb{E}[\Psi_2(X_2, X_3) \mid X_1] + \mathbb{E}[\Psi_1(X_1) \mid X_1] + \mathbb{E}[\Psi_1(X_2) \mid X_1] \\
&\quad + \mathbb{E}[\Psi_1(X_3) \mid X_1] - \theta(F) \\
&= \Psi_1(X_1) - \Psi_1(X_1) - \Psi_1(X_1) - \theta(F) + \Psi_1(X_1) + \theta(F) + \theta(F) - \theta(F) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\Psi_{32}^{\circ}(X_1, X_2) &= \mathbb{E}[\Psi_3^{\circ}(X_1, X_2, X_3) \mid X_1, X_2] \\
&= \mathbb{E}[\Psi_3(X_1, X_2, X_3) \mid X_1, X_2] - \mathbb{E}[\Psi_2(X_1, X_2) \mid X_1, X_2] \\
&\quad - \mathbb{E}[\Psi_2(X_1, X_3) \mid X_1, X_2] \mathbb{E}[\Psi_2(X_2, X_3) \mid X_1, X_2] + \mathbb{E}[\Psi_1(X_1) \mid X_1, X_2] \\
&\quad + \mathbb{E}[\Psi_1(X_2) \mid X_1, X_2] + \mathbb{E}[\Psi_1(X_3) \mid X_1, X_2] - \theta(F) \\
&= \Psi_2(X_1, X_2) - \Psi_2(X_1, X_2) - \Psi_1(X_1) - \Psi_1(X_2) + \Psi_1(X_1) + \Psi_1(X_2) \\
&\quad + \theta(F) - \theta(F) \\
&= 0
\end{aligned}$$

Then,

$$\begin{aligned}
\xi_1^{\circ} &= \mathbb{E}[\Psi_{31}^{\circ}(X_1)]^2 - (\mathbb{E}(U_{n,3}))^2 = 0 \\
\xi_2^{\circ} &= \mathbb{E}[\Psi_{32}^{\circ}(X_1, X_2)]^2 - (\mathbb{E}(U_{n,3}))^2 = 0
\end{aligned}$$

By (4.3.7),

$$\begin{aligned}
\text{Var}(U_{n,3}) &= \frac{6}{n(n-1)(n-2)} \left[ 3 \binom{n-3}{2} \xi_1^{\circ} + \binom{3}{2} \binom{n-3}{1} \xi_2^{\circ} + \xi_3^{\circ} \right] \\
&= \frac{1}{n(n-1)(n-2)} (9(n-3)(n-4)\xi_1^{\circ} + 18(n-3)\xi_2^{\circ} + 6\xi_3^{\circ}) \\
&= O(n^{-3}) \tag{4.3.19}
\end{aligned}$$

From direct computation,  $\mathbb{E}(U_{n,h}) = 0$ ,  $\forall 1 \leq h \leq m$  and

$$\text{Var}(U_{n,h}) = \mathbb{E}(U_{n,h}^2) = O(n^{-h}), \quad h = 1, 2, \dots, m; \tag{4.3.20}$$

Since  $U_{n,1} = V_{n,1}$ , we can write

$$U_n = \theta(F) + \frac{m}{n} \sum_{i=1}^n [\Psi_1(X_i) - \theta(F)] + O_p(n^{-1}) \tag{4.3.21}$$

**Theorem 4.2**

For  $c = 1, 2, \dots, h - 1$ , and  $h = 1, 2, \dots, m$ ,  $\Psi_{hc}^\circ(x_1, \dots, x_c) = 0$

**Proof:** Lee (1990)

Using the integral representation

$$\begin{aligned}
 & \Psi_h^\circ(x_1, \dots, x_h) \\
 &= \int \cdots \int \phi(u_1, \dots, u_m) \prod_{j=1}^h [d(c(u_j - X_{i_j})) - dF(u_j)] \prod_{j=h+1}^m dF(u_j) \quad (4.3.22) \\
 &= \int \cdots \int \phi(u_1, \dots, x_h, \dots, u_m) \prod_{j=1}^{h-1} [d(c(u_j - X_{i_j})) - dF(u_j)] \prod_{j=h+1}^m dF(u_j) \\
 &\quad - \int \cdots \int \phi(u_1, \dots, u_m) \prod_{j=1}^{h-1} [d(c(u_j - X_{i_j})) - dF(u_j)] \prod_{j=h}^m dF(u_j) \\
 &= \int \cdots \int \phi(u_1, \dots, u_m) \prod_{j=1}^{h-1} [d(c(u_j - X_{i_j})) - dF(u_j)] \prod_{j=h+1}^m dF(u_j) \\
 &\quad - \Psi_{h-1}^\circ(x_1, \dots, x_{h-1})
 \end{aligned}$$

Integrating both sides with respect to  $u_h$  with measure  $dF(u_h)$  yields

$$\begin{aligned}
 \mathbb{E}(\Psi_h^\circ(x_1, \dots, x_{h-1}, X_h)) &= \Psi_{h-1}^\circ(x_1, \dots, x_{h-1}) - \Psi_{h-1}^\circ(x_1, \dots, x_{h-1}) \\
 &= 0
 \end{aligned}$$

and so  $\Psi_{h,h-1}^\circ(x_1, \dots, x_{h-1}) = \mathbb{E}\{\Psi_h^\circ(x_1, \dots, x_{h-1}, X_h)\} = 0$  from (4.3.3).

It follows from Theorem 4.1 (i) that

$$\Psi_{hc}^\circ(x_1, \dots, x_c) = \mathbb{E}\{\Psi_{h,h-1}^\circ(x_1, \dots, x_c, X_{c+1}, \dots, X_{h-1})\} = 0$$

■

**Theorem 4.3**

(i) Let  $h < h'$  and let  $S_1$  and  $S_2$  be a  $h$ -subset of  $\{1, 2, \dots, n\}$  and a  $h'$ -subset of  $\{1, 2, \dots, n\}$  respectively. Then

$$\text{Cov}(\Psi_h^\circ(S_1), \Psi_{h'}^\circ(S_2)) = 0 \quad \text{and} \quad \text{Cov}(U_{n,h}, U_{n,h'}) = 0$$

(ii) Let  $S_1$  and  $S_2$  be two distinct  $h$ -subsets of  $\{1, 2, \dots, n\}$ . Then

$$\text{Cov}(\Psi_h^\circ(S_1), \Psi_h^\circ(S_2)) = 0$$

**Proof:** Lee (1990)

(i) Since  $E(U_{n,h}) = E(\Psi_h^\circ(X_1, \dots, X_h)) = 0$ , we only need to prove that

$$E(\Psi_h^\circ(S_1), \Psi_{h'}^\circ(S_2)) = 0$$

Since  $h < h'$ , there is an element in  $S_2$  that is not in  $S_1$ , therefore there is a r.v.,  $X_{h'}$  say, that appears in  $\Psi_{h'}^\circ(S_2)$  but not in  $\Psi_h^\circ(S_1)$ . Thus we can write

$$\begin{aligned} E\{\Psi_h^\circ(S_1)\Psi_{h'}^\circ(S_2)\} &= E\{E(\Psi_h^\circ(S_1)\Psi_{h'}^\circ(S_2) \mid X_{h'})\} \\ &= E\{\Psi_h^\circ(S_1)E(\Psi_{h'}^\circ(S_2) \mid X_{h'})\} \end{aligned}$$

since  $\Psi_h^\circ(S_1)$  and  $X_{h'}$  are independent. But  $E(\Psi_{h'}^\circ(S_2) \mid X_{h'}) = \Psi_{h'}^\circ(X_{h'}) = 0$  by Theorem 4.2 and so  $E\{\Psi_h^\circ(S_1)\Psi_{h'}^\circ(S_2)\} = 0$ , proving the first part.

$$\text{Cov}(U_{n,h}, U_{n,h'}) = \binom{n}{h}^{-1} \binom{n}{h'}^{-1} \sum_{(n,h)} \sum_{(n,h')} \text{Cov}(\Psi_h^\circ(S_1), \Psi_{h'}^\circ(S_2)) = 0$$

where the sum  $\sum_{(n,h)}$  is taken over all subsets  $1 \leq i_1 < \dots < i_h \leq n$  of  $\{1, 2, \dots, n\}$ .

(ii) If  $S_1 \cap S_2$  is empty, the result follows by independence. Otherwise, suppose that  $S_2 - S_1 = \{i_1, \dots, i_c\}$  with  $c < h$ . (If  $S_1 \subseteq S_2$  then consider  $S_1 - S_2$ ). Then

$$\begin{aligned} E(\Psi_h^\circ(S_1)\Psi_h^\circ(S_2)) &= E\{E(\Psi_h^\circ(S_1)\Psi_h^\circ(S_2) \mid X_{i_1}, \dots, X_{i_c})\} \\ &= E\{\Psi_h^\circ(S_1)E(\Psi_h^\circ(S_2) \mid X_{i_1}, \dots, X_{i_c})\} \\ &= 0 \end{aligned}$$

since  $\Psi_{hc}^\circ(x_1, \dots, x_c) = 0$  by Theorem 4.2. ■

For the one-sample U-statistics of degree 3,

$$U_{1,1}^{(\mathbf{s})} = \binom{N}{3}^{-1} \sum_{i < j < j'} (D_{ij}^g - D_{ij'}^g)^2, \quad U_{1,2}^{(\mathbf{s})} = \binom{N}{3}^{-1} \sum_{i < i' < j} (D_{ij}^g - D_{i'j}^g)^2$$

$$\text{and } U_{1,3}^{(\mathbf{s})} = \binom{N}{3}^{-1} \sum_{i < j < j'} (D_{ij}^g - D_{jj'}^g)^2,$$

$$\begin{aligned} \mu_{g1} = E(U_{1,1}^{(\mathbf{s})}) &= E(U_{1,2}^{(\mathbf{s})}) = E(U_{1,3}^{(\mathbf{s})}) = E(D_{ij}^g - D_{ij'}^g)^2 \\ &= \frac{2}{K^2} \left[ \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g - \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) \right] \end{aligned}$$

$$\begin{aligned}
& - \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \Big] \\
& = \frac{2}{K^2} \left[ \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g - \sum_{k=1}^K \theta_k^g(i, j; i, j') - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g(i, j; i, j') \right]
\end{aligned}$$

where, for any  $g$ ,

$$\theta_k^g(i, j) = \theta_k^g = P(X_{ik}^g \neq X_{jk}^g) = \sum_{c=0}^{C-1} p_k^g(c)[1 - p_k^g(c)], \quad (4.3.23)$$

$$\begin{aligned}
\theta_{k_1 k_2}^g(i, j) & = \theta_{k_1 k_2}^g = P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \\
& = \sum_{c_1, c_2=0}^{C-1} p_{k_1 k_2}^g(c_1, c_2) \left[ \sum_{\substack{c_3=0 \\ c_3 \neq c_1}}^{C-1} \sum_{\substack{c_4=0 \\ c_4 \neq c_2}}^{C-1} p_{k_1 k_2}^g(c_3, c_4) \right], \quad (4.3.24)
\end{aligned}$$

$$\theta_k^g(i, j; i, j') = P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) = \sum_{c=0}^{C-1} p_k^g(c)[1 - p_k^g(c)]^2, \quad (4.3.25)$$

$$\begin{aligned}
\theta_{k_1 k_2}^g(i, j; i, j') & = P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \\
& = \sum_{c_1, c_2=0}^{C-1} p_{k_1 k_2}^g(c_1, c_2)[1 - p_{k_1}^g(c_1)][1 - p_{k_2}^g(c_2)], \quad (4.3.26)
\end{aligned}$$

$$p_k(c) = P(X_{ik}^g = c) \quad \text{and} \quad p_{k_1 k_2}^g(c_1, c_2) = P(X_{ik_1}^g = c_1, X_{ik_2}^g = c_2) \quad (4.3.27)$$

If the null hypothesis of interest is that there is no between-group difference in Hamming distances, i.e.,  $H_0: \theta_k^1 = \dots = \theta_k^G = \theta_k$  and  $\theta_{k_1 k_2}^1 = \dots = \theta_{k_1 k_2}^G = \theta_{k_1 k_2}$ , then

$$\mu_{g1} = \mu_1 = \frac{2}{K^2} \left[ \sum_{k=1}^K \theta_k + \sum_{k_1 \neq k_2} \theta_{k_1 k_2} - \sum_{k=1}^K \theta_k(i, j; i, j') - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}(i, j; i, j') \right]$$

Note that  $\phi_{1,1}(\mathbf{X}_i^g, \mathbf{X}_j^g, \mathbf{X}_{j'}^g) = (D_{ij}^g - D_{ij'}^g)^2$  is not a symmetric kernel. Therefore, we need to replace it by the symmetric kernel

$$\phi'_{1,1}(\mathbf{X}_i^g, \mathbf{X}_j^g, \mathbf{X}_{j'}^g) = \frac{1}{3} \left[ (D_{ij}^g - D_{ij'}^g)^2 + (D_{jj'}^g - D_{ji}^g)^2 + (D_{j'i}^g - D_{j'j}^g)^2 \right].$$

Since the sequences are i.i.d.,

$$E[\phi_{1,1}(\mathbf{X}_i^g, \mathbf{X}_j^g, \mathbf{X}_{j'}^g)] = E[\phi'_{1,1}(\mathbf{X}_i^g, \mathbf{X}_j^g, \mathbf{X}_{j'}^g)] = \mu_{g1}$$



$$\begin{aligned}
\psi_{(1,1)1}(\mathbf{x}_i^g) &\equiv E[\phi'_{1,1}(\mathbf{x}_i^g, \mathbf{X}_j^g, \mathbf{X}_{j'}^g) - \mu_{g1}] \\
&= \frac{2}{3K^2} \left\{ 2 \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) + 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right. \\
&\quad - \sum_{k=1}^K \left( P(X_{jk}^g \neq x_{ik}^g) \right)^2 - \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\
&\quad - 2 \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \\
&\quad - 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) - 2 \sum_{k=1}^K \theta_k^g \\
&\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g + 3 \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) \\
&\quad \left. + 3 \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right\}
\end{aligned}$$

By (4.3.21),

$$U_{1,k}^{(3)} = \mu_{g1} + \frac{3}{N} \sum_{i=1}^N \left( \Psi_{(1,k)1}(\mathbf{X}_i^g) - \mu_{g1} \right) + O_p(N^{-1}) \quad \text{for } k = 1, 2, 3.$$

Note that  $E(U_{1,k}^{(3)}) = 0$  and  $\text{Var}(U_{1,k}^{(3)}) = \frac{9}{N} \xi_1^{(1)} + O_p(N^{-2})$

### Example: Two categories

Let, for any  $g$ ,

$$p_k^g(0) = P(X_{jk}^g = 0) \quad , \quad p_k^g(1) = P(X_{jk}^g = 1)$$

$$p_{k_1 k_2}^g(1, 1) = P(X_{jk_1}^g = 1, X_{jk_2}^g = 1) \quad , \quad p_{k_1 k_2}^g(0, 1) = P(X_{jk_1}^g = 0, X_{jk_2}^g = 1)$$

$$p_{k_1 k_2}^g(1, 0) = P(X_{jk_1}^g = 1, X_{jk_2}^g = 0) \quad , \quad p_{k_1 k_2}^g(0, 0) = P(X_{jk_1}^g = 0, X_{jk_2}^g = 0).$$

Note that  $p_k^g(0) + p_k^g(1) = 1$  and  $p_{k_1 k_2}^g(1, 1) + p_{k_1 k_2}^g(0, 1) + p_{k_1 k_2}^g(1, 0) + p_{k_1 k_2}^g(0, 0) = 1$

Therefore, there are 4 parameters for each group. ■

Now,

$$\begin{aligned}
\psi_{(1,1)1}^2(\mathbf{x}_i^g) &= \frac{4}{9K^4} \left\{ 4 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right)^2 + 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right)^2 \right. \\
&\quad \left. + 4 \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right. \\
&\quad \left. + 4 \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right. \\
&\quad \left. + 4 \sum_{k=1}^K \theta_k^g + 4 \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right\}
\end{aligned}$$

$$\begin{aligned}
& + \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right)^2 + \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right)^2 \\
& + 4 \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right)^2 + 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right)^2 \\
& + \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right. \\
& + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \left. \right]^2 \\
& + 8 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \\
& - 4 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right) \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right) \\
& - 4 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right) \\
& - 8 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right) \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right) \\
& - 8 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \\
& + 4 \left( \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g) \right) \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right. \\
& + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq x_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \left. \right] \\
& - 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right) \\
& - 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right) \\
& - 8 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right) \\
& - 8 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \\
& + 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right.
\end{aligned}$$

$$\begin{aligned}
& + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \Bigg] \\
& + 2 \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right) \\
& + 4 \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right) \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right) \\
& + 4 \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \\
& - 2 \left( \sum_{k=1}^K (P(X_{jk}^g \neq x_{ik}^g))^2 \right) \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right. \\
& \left. + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \right] \\
& + 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right) \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right) \\
& + 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \\
& - 2 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \right) \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right. \\
& \left. + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \right] \\
& + 8 \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right) \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \\
& - 4 \left( \sum_{k=1}^K P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) \right) \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right. \\
& \left. + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \right] \\
& - 4 \left( \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \right) \left[ -2 \left( \sum_{k=1}^K \theta_k^g + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^g \right) \right. \\
& \left. + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \right] \Bigg] \\
& + 3 \left( \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) + \sum_{k_1 \neq k_2} P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) \right) \Bigg] \Bigg]
\end{aligned}$$

and

$$\xi_1^{(1)} = E[\psi_{(1,1)1}^2(\mathbf{X}_i^g)]$$

For the 1-sample U-statistics of degree 4,

$$U_{2,1}^{(4)} = \binom{N}{4}^{-1} \sum_{i < j < i' < j'} (D_{ij}^g - D_{i'j'}^g)^2, \quad U_{2,2}^{(4)} = \binom{N}{4}^{-1} \sum_{i < i' < j < j'} (D_{ij}^g - D_{i'j'}^g)^2$$

$$\text{and } U_{2,3}^{(4)} = \binom{N}{4}^{-1} \sum_{i < i' < j' < j} (D_{ij}^g - D_{i'j'}^g)^2,$$

$$\begin{aligned} \mu_{g2} &= E(U_{2,1}^{(4)}) = E(U_{2,2}^{(4)}) = E(U_{2,3}^{(4)}) = E[(D_{ij}^g - D_{i'j'}^g)^2] \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k^g (1 - \theta_k^g) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^g - \theta_{k_1}^g \theta_{k_2}^g) \right\} \end{aligned} \quad (4.3.28)$$

and under  $H_0$ ,

$$\mu_{g2} = \mu_2 = \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k (1 - \theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right\}$$

$$\phi_{2,1}(\mathbf{X}_i^g, \mathbf{X}_j^g, \mathbf{X}_{i'}^g, \mathbf{X}_{j'}^g) = (D_{ij}^g - D_{i'j'}^g)^2 = \left[ \sum_{k=1}^K \frac{I(X_{ik}^g \neq X_{jk}^g)}{K} - \sum_{k=1}^K \frac{I(X_{i'k}^g \neq X_{j'k}^g)}{K} \right]^2$$

$$\begin{aligned} \psi_{(2,1)1}(\mathbf{x}_i^g) &\equiv E[\phi_{2,1}(\mathbf{x}_i^g, \mathbf{X}_j^g, \mathbf{X}_{i'}^g, \mathbf{X}_{j'}^g) - \mu_{g2}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^g) P(X_{jk}^g \neq x_{ik}^g) + \sum_{k=1}^K \theta_k^g (2\theta_k^g - 1) \right. \\ &\quad + \sum_{k_1 \neq k_2} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^g \theta_{k_2}^g - \theta_{k_1 k_2}^g) \\ &\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^g P(X_{jk_1}^g \neq x_{ik_1}^g) \right\} \end{aligned} \quad (4.3.29)$$

which becomes under  $H_0$

$$\begin{aligned} \psi_{(2,1)1}(\mathbf{x}_i) &\equiv \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k) P(X_{jk} \neq x_{ik}) + \sum_{k=1}^K \theta_k (2\theta_k - 1) \right. \\ &\quad + \sum_{k_1 \neq k_2} P(X_{jk_1} \neq x_{ik_1}, X_{jk_2} \neq x_{ik_2}) + \sum_{k_1 \neq k_2} (2\theta_{k_1} \theta_{k_2} - \theta_{k_1 k_2}) \\ &\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2} P(X_{jk_1} \neq x_{ik_1}) \right\} \end{aligned}$$

$$\xi_1^{(2)} = E[\psi_{(2,1)1}^2(\mathbf{X}_i^g)]$$

By (4.3.21),

$$U_{2,k}^{(4)} = \mu_{g2} + \frac{4}{N} \sum_{i=1}^N \left( \Psi_{(2,k)1}(\mathbf{X}_i^g) - \mu_{g2} \right) + O_p(N^{-1}) \quad \text{for } k = 1, 2, 3.$$

Note that  $E(U_{2,k}^4) = \mu_{g2}$  and  $\text{Var}(U_{2,k}) = \frac{16}{N} \xi_1^{(2)} + O_p(N^{-2})$ .

### 4.3.2 Multiple-Sample U-statistics

Let  $\{\mathbf{X}_i^{(j)}; i \geq 1\}$ ,  $j = 1, \dots, c (\geq 2)$  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  $j = 1, \dots, c$ . Let  $\mathbf{F} = (F^{(1)}, \dots, F^{(c)})$  and  $\phi(\mathbf{X}_i^{(j)}, 1 \leq i \leq m_j, 1 \leq j \leq c)$  be a Borel-measurable *kernel of degree*  $\mathbf{m} = (m_1, \dots, m_c)$ , where without loss of generality we assume that  $\phi$  is symmetric in the  $m_j (\geq 1)$  arguments of the  $j$ th set, for  $j = 1, \dots, c$ . Let  $m_0 = m_1 + \dots + m_c$  and

$$\theta(\mathbf{F}) = \int_{R^{m_0}} \dots \int \phi(\mathbf{x}_i^{(j)}, 1 \leq i \leq m_j, 1 \leq j \leq c) \prod_{j=1}^c \prod_{i=1}^{m_j} dF^{(j)}(\mathbf{x}_i^{(j)}) \quad (4.3.30)$$

#### Definition 4.6

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

$$U(\mathbf{n}) = \prod_{j=1}^c \binom{n_j}{m_j}^{-1} \sum_{(\mathbf{n})}^* \phi(\mathbf{X}_\alpha^{(j)}, \alpha = i_{j1}, \dots, i_{jm_j}, 1 \leq j \leq c), \quad (4.3.31)$$

where the summation  $\sum_{(\mathbf{n})}^*$  extends over all  $1 \leq i_{j1} < \dots < i_{jm_j} \leq n_j$ ,  $1 \leq j \leq c$ .  $U(\mathbf{n})$  is an unbiased estimator of  $\theta(\mathbf{F})$ . The generalized von Mises' functional is

$$\theta(\mathbf{F}(\cdot, \mathbf{n})) = \left( \prod_{j=1}^c n_j^{-m_j} \right) \prod_{j=1}^c \left\{ \sum_{i_{j1}}^{n_j} \dots \sum_{i_{jm_j}}^{n_j} \phi(\mathbf{X}_\alpha^{(j)}, \alpha = i_{j1}, \dots, i_{jm_j}, 1 \leq j \leq c) \right\} \quad (4.3.32)$$

■

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

$$\Psi_{d_1 \dots d_c}(\mathbf{x}_1^{(j)}, \dots, \mathbf{x}_{d_j}^{(j)}, 1 \leq j \leq c) \equiv E(\phi(\mathbf{x}_1^{(j)}, \dots, \mathbf{x}_{d_j}^{(j)}, \mathbf{X}_{d_j+1}^{(j)}, \dots, \mathbf{X}_{m_j}^{(j)}, 1 \leq j \leq c)) \quad (4.3.33)$$

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

$$\xi_{\mathbf{d}}(\mathbf{F}) = E \left( \Psi_{\mathbf{d}}^2(\mathbf{X}_1^{(j)}, \dots, \mathbf{X}_{d_j}^{(j)}, 1 \leq j \leq c) \right) - \theta^2(\mathbf{F}), \quad \mathbf{0} \leq \mathbf{d} \leq \mathbf{m} \quad (4.3.34)$$

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

$$\begin{aligned} \text{Var}[U(\mathbf{n})] &= \left[ \prod_{j=1}^c \binom{n_j}{m_j} \right]^{-1} \left\{ \sum_{d_1=0}^{m_1} \dots \sum_{d_c=0}^{m_c} \prod_{j=1}^c \binom{m_j}{d_j} \binom{n_j - m_j}{m_j - d_j} \xi_{d_1, \dots, d_c}(\mathbf{F}) \right\} \\ &= \sum_{j=1}^c n_j^{-1} \sigma_j^2 [1 + O(n_0^{-1})] \end{aligned} \quad (4.3.35)$$

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

$$\sigma_j^2 = m_j^2 \xi_{\delta_{j1}, \dots, \delta_{jc}}(\mathbf{F}) \quad j = 1, \dots, c \quad (4.3.36)$$

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

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

and

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

where  $\sum_{(\mathbf{n}, \mathbf{m})}^*$  extends over all  $1 \leq i_{j1} < \dots < i_{j m_j} \leq n_j$ ,  $j = 1, 2$ .  $U(n_1, n_2)$  is an unbiased estimator of  $\theta(F^{(1)}, F^{(2)})$ .

Define

$$\begin{aligned} \theta(F_{n_1}^{(1)}, F_{n_2}^{(2)}) \\ = n_1^{-m_1} n_2^{-m_2} \times \sum_{i_1 1=1}^{n_1} \dots \sum_{i_1 m_1=1}^{n_1} \sum_{i_2 1=1}^{n_2} \dots \sum_{i_2 m_2=1}^{n_2} \phi(X_{i_1 1}, \dots, X_{i_1 m_1}, Y_{i_2 1}, \dots, Y_{i_2 m_2}) \end{aligned}$$

Then,

$$|U_{n_1, n_2} - \theta(F_{n_1}^{(1)}, F_{n_2}^{(2)})| = O_p(n_0^{-1}); \quad n_0 = \min(n_1, n_2) \quad (4.3.37)$$

provided the variance of  $U_{n_1, n_2}$  exists.

$$\begin{aligned} \psi_{d_1 d_2}(x_1, \dots, x_{d_1}, y_1, \dots, y_{d_2}) \\ = E\{\phi(x_1, \dots, x_{d_1}, X_{d_1+1}, \dots, X_{m_1}; y_1, \dots, y_{d_2}, Y_{d_2+1}, \dots, Y_{m_2}) - \theta(F^{(1)}, F^{(2)})\}, \end{aligned}$$

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

$$\begin{aligned} \xi_{d_1 d_2} &= E\{\psi_{d_1 d_2}^2(X_1, \dots, X_{d_1}, Y_1, \dots, Y_{d_2})\} \\ &= E\{\Psi_{d_1 d_2}^2(X_1, \dots, X_{d_1}, Y_1, \dots, Y_{d_2})\} - \theta^2(F^{(1)}, F^{(2)}) \end{aligned} \quad (4.3.38)$$

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

$$(\gamma_{n_1, n_2})^{-1}(U(n_1, n_2) - \theta(F^{(1)}, F^{(2)})) \xrightarrow{d} N(0, 1), \quad (4.3.39)$$

where

$$\gamma_{n_1, n_2}^2 = \left(\frac{m_1^2}{n_1}\right) \xi_{10} + \left(\frac{m_2^2}{n_2}\right) \xi_{01}. \quad (4.3.40)$$

The decomposition for  $U(\mathbf{n})$  can be developed similarly to the one-sample U-statistic. For a two-sample U-statistic, let

$$\begin{aligned} & \Psi_{h_1, h_2}^\circ(x_1, \dots, x_{h_1}; y_1, \dots, y_{h_2}) \\ &= \int \cdots \int \phi(u_1, \dots, u_{m_1}; v_1, \dots, v_{m_2}) \prod_{j=1}^{h_1} [d(c(u_j - X_{i_j})) - dF^1(u_j)] \\ & \quad \times \prod_{j=h_1+1}^{m_1} dF^1(u_j) \prod_{l=1}^{h_2} [d(c(v_l - Y_{i_l})) - dF^2(v_l)] \prod_{l=h_2+1}^{m_2} dF^2(v_l) \end{aligned}$$

then it follows as in the one-sample case that

$$\begin{aligned} & \phi(x_1, \dots, x_{m_1}; y_1, \dots, y_{m_2}) \\ &= \sum_{h_1=0}^{m_1} \sum_{h_2=0}^{m_2} \sum_{m_1, h_1} \sum_{m_2, h_2} \Psi_{h_1, h_2}^\circ(x_{j_1}, \dots, x_{j_{h_1}}; y_{l_1}, \dots, y_{l_{h_2}}) \end{aligned} \quad (4.3.41)$$

where  $\sum_{m_1, h_1}$  is taken over all subsets  $1 \leq j_1 < \cdots < j_{h_1} \leq m_1$  of  $\{1, 2, \dots, m_1\}$ .

**Theorem 4.4** (Lee, 1990)

The two-sample U-statistic admits the representation

$$U_{n_1, n_2} = \sum_{h_1=0}^{m_1} \sum_{h_2=0}^{m_2} \binom{m_1}{h_1} \binom{m_2}{h_2} U_{n_1, n_2}^{(h_1, h_2)} \quad (4.3.42)$$

where  $U_{n_1, n_2}^{(h_1, h_2)}$  is the generalized U-statistic based on the kernel  $\Psi_{h_1, h_2}^\circ$  and is given by

$$U_{n_1, n_2}^{(h_1, h_2)} = \binom{n_1}{h_1}^{-1} \binom{n_2}{h_2}^{-1} \sum_{(n_1, h_1)} \sum_{(n_2, h_2)} \Psi_{h_1, h_2}^\circ(X_{j_1}, \dots, X_{j_{h_1}}; Y_{l_1}, \dots, Y_{l_{h_2}}).$$

The functions  $\Psi_{h_1, h_2}^\circ$  implicitly defined by (4.3.41) satisfy

- (i)  $\mathbf{E}\{\Psi_{h_1, h_2}^\circ(X_1, \dots, X_{h_1}; Y_1, \dots, Y_{h_2})\} = 0$
- (ii)  $\text{Cov}(\Psi_{h_1, h_2}^\circ(S_1, S_2), \Psi_{h'_1, h'_2}^\circ(S'_1, S'_2)) = 0$  for all integers  $h_1, h_2, h'_1, h'_2$  and sets  $S_1, S_2, S'_1, S'_2$  unless  $h_1 = h'_1, h_2 = h'_2$  and  $S_1 = S'_1$  and  $S_2 = S'_2$ . ■

The generalized U-statistics  $U_{n_1, n_2}^{(h_1, h_2)}$  are thus all uncorrelated. Their variances are given by

$$\text{Var}(U_{n_1, n_2}^{(h_1, h_2)}) = \binom{n_1}{h_1}^{-1} \binom{n_2}{h_2}^{-1} \delta_{h_1, h_2}^2 \quad (4.3.43)$$

where  $\delta_{h_1, h_2}^2 = \text{Var}(\Psi_{h_1, h_2}^\circ(X_1, \dots, X_{h_1}; Y_1, \dots, Y_{h_2}))$ .

From (4.3.42) we can write

$$U(n_1, n_2) - \theta(\mathbf{F}) = U_1(n_1) + U_2(n_2) + U^*(n_1, n_2)$$

where

$$\begin{aligned} U_1(n_1) &= \binom{n_1}{m_1}^{-1} \sum_{(n_1, m_1)} \Psi_{m_1 0}^\circ(X_{i_1}, \dots, X_{i_{m_1}}) - \theta(\mathbf{F}); \\ U_2(n_2) &= \binom{n_2}{m_2}^{-1} \sum_{(n_2, m_2)} \Psi_{0 m_2}^\circ(Y_{i_1}, \dots, Y_{i_{m_2}}) - \theta(\mathbf{F}); \\ U^*(n_1, n_2) &= \sum_{h_1=1}^{m_1} \sum_{h_2=1}^{m_2} \binom{m_1}{h_1} \binom{m_2}{h_2} U_{n_1, n_2}^{(h_1, h_2)} \\ &= U(n_1, n_2) - \theta(\mathbf{F}) - U_1(n_1) - U_2(n_2), \end{aligned}$$

and  $U^*(n_1, n_2)$  is also a generalized U-statistic for which

$$\begin{aligned} &\text{Var}(U^*(n_1, n_2)) \\ &= \mathbf{E}[U^*(n_1, n_2)]^2 \\ &= \mathbf{E}[U(n_1, n_2) - \theta(\mathbf{F}) - U_1(n_1) - U_2(n_2)]^2 \\ &= \text{Var}(U(n_1, n_2)) + \mathbf{E}(U_1^2(n_1)) + \mathbf{E}(U_2^2(n_2)) \\ &\quad - 2\mathbf{E}[(U(n_1, n_2) - \theta(\mathbf{F}))U_1(n_1)] - 2\mathbf{E}[(U(n_1, n_2) - \theta(\mathbf{F}))U_2(n_2)] \\ &\quad + 2\mathbf{E}(U_1(n_1)U_2(n_2)) \\ &= \text{Var}(U(n_1, n_2)) + \mathbf{E}(U_1^2(n_1)) + \mathbf{E}(U_2^2(n_2)) \\ &\quad - 2\mathbf{E}[U_1^2(n_1) + U_1(n_1)U^*(n_1, n_2)] - 2\mathbf{E}[U_2^2(n_2) + U_2(n_2)U^*(n_1, n_2)] \end{aligned}$$



$$\begin{aligned}
&= \text{Var}(U(n_1, n_2)) - E(U_1^2(n_1)) - E(U_2^2(n_2)) \\
&= \sum_{h_1=1}^{m_1} \sum_{h_2=1}^{m_2} \binom{m_1}{h_1} \binom{m_2}{h_2} \binom{n_1 - m_1}{m_1 - h_1} \binom{n_2 - m_2}{m_2 - h_2} \binom{n_1}{m_1}^{-1} \binom{n_2}{m_2}^{-1} \xi_{h_1 h_2}(\mathbf{F}) \\
&= \frac{m_1^2 m_2^2}{n_1 n_2} \xi_{11}(\mathbf{F}) + O(n_0^{-3})
\end{aligned}$$

since the generalized U-statistics  $U_{n_1, n_2}^{(h_1, h_2)}$  are all uncorrelated (Theorem 4.4).

If  $\text{Var}(U^*(n_1, n_2)) = O(n_0^{-2})$ , then  $U^*(n_1, n_2) = O_p(n_0^{-1})$ . Also,  $U_1(n_1) - \theta(\mathbf{F})$  and  $U_2(n_2) - \theta(\mathbf{F})$  are one-sample U-statistics. Therefore,  $U(n_1, n_2)$  can be written as

$$\begin{aligned}
U(n_1, n_2) &= \theta(\mathbf{F}) + \frac{m_1}{n_1} \sum_{i=1}^{n_1} [\Psi_{10}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2} \sum_{i=1}^{n_2} [\Psi_{01}(Y_i) - \theta(\mathbf{F})] \\
&\quad + O_p(n_0^{-1})
\end{aligned} \tag{4.3.44}$$

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

$$\begin{aligned}
U(n_1, n_2, n_3) &= \theta(\mathbf{F}) + \frac{m_1}{n_1} \sum_{i=1}^{n_1} [\Psi_{100}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2} \sum_{i=1}^{n_2} [\Psi_{010}(Y_i) - \theta(\mathbf{F})] \\
&\quad + \frac{m_3}{n_3} \sum_{i=1}^{n_3} [\Psi_{001}(Z_i) - \theta(\mathbf{F})] + O_p(n_0^{-1})
\end{aligned} \tag{4.3.45}$$

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

$$\begin{aligned}
U(n_1, n_2, n_3, n_4) &= \theta(\mathbf{F}) + \frac{m_1}{n_1} \sum_{i=1}^{n_1} [\Psi_{1000}(X_i) - \theta(\mathbf{F})] + \frac{m_2}{n_2} \sum_{i=1}^{n_2} [\Psi_{0100}(Y_i) - \theta(\mathbf{F})] \\
&\quad + \frac{m_3}{n_3} \sum_{i=1}^{n_3} [\Psi_{0010}(Z_i) - \theta(\mathbf{F})] + \frac{m_4}{n_4} \sum_{i=1}^{n_4} [\Psi_{0001}(W_i) - \theta(\mathbf{F})] \\
&\quad + O_p(n_0^{-1})
\end{aligned} \tag{4.3.46}$$

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

For the two-sample U-statistic of degree (2,2),

$$\mathbf{U}_s^{(2,2)} = \binom{N}{2}^{-1} \binom{N}{2}^{-1} \sum_{1 \leq i < j \leq N} \sum_{1 \leq i' < j' \leq N} (D_{ij}^{g_1} - D_{i'j'}^{g_2})^2, \tag{4.3.47}$$

$$E(D_{ij}^{g_1} - D_{i'j'}^{g_2})^2 = E[(D_{ij}^{g_1})^2 + (D_{i'j'}^{g_2})^2 - 2D_{ij}^{g_1} D_{i'j'}^{g_2}].$$

$$\begin{aligned}
E(D_{ij}^g)^2 &= E \left[ \frac{1}{K} \sum_{k=1}^K I(X_{ik}^g \neq X_{jk}^g) \right]^2 \\
&= \frac{1}{K^2} E \left[ \sum_{k=1}^K I^2(X_{ik}^g \neq X_{jk}^g) + 2 \sum_{k_1 < k_2} I(X_{ik_1}^g \neq X_{jk_1}^g) I(X_{ik_2}^g \neq X_{jk_2}^g) \right] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(X_{ik}^g \neq X_{jk}^g) + 2 \sum_{k_1 < k_2} P(X_{ik_1}^g \neq X_{jk_1}^g; X_{ik_2}^g \neq X_{jk_2}^g) \right\} \\
&= \frac{1}{K^2} \sum_{k=1}^K \theta_k^g + \frac{2}{K^2} \sum_{k_1 < k_2} \theta_{k_1 k_2}^g
\end{aligned}$$

And, for any  $g_1$  and  $g_2$ ,

$$\begin{aligned}
E(D_{ij}^{g_1}; D_{i'j'}^{g_2}) &= E \left\{ \frac{1}{K^2} \left[ \sum_{k=1}^K I(X_{ik}^{g_1} \neq X_{jk}^{g_1}) \right] \left[ \sum_{k=1}^K I(X_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \right] \right\} \\
&= \frac{1}{K^2} \left\{ E \left[ \sum_{k=1}^K I(X_{ik}^{g_1} \neq X_{jk}^{g_1}) I(X_{i'k}^{g_2} \neq X_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} I(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) I(X_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right] \right\} \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_1}, X_{i'k}^{g_2} \neq X_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, X_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right\} \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_1}) P(X_{i'k}^{g_2} \neq X_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) P(X_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right\} \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K \theta_k^{g_1} \theta_k^{g_2} + \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} \theta_{k_2}^{g_2} \right\} = \frac{1}{K^2} \left\{ \sum_{k_1=1}^K \sum_{k_2=1}^K \theta_{k_1}^{g_1} \theta_{k_2}^{g_2} \right\}
\end{aligned}$$

since the sequences in groups  $g_1$  and  $g_2$  are independent.

Therefore,

$$\begin{aligned}
E(\mathbf{U}_s^{(2,2)}) &\equiv \mu_{(g_1, g_2)3} \equiv \theta(F^{(1)}, F^{(2)}) \equiv E(D_{ij}^{g_1} - D_{i'j'}^{g_2})^2 \\
&= \frac{1}{K^2} \sum_{k=1}^K \theta_k^{g_1} + \frac{2}{K^2} \sum_{k_1 < k_2} \theta_{k_1 k_2}^{g_1} + \frac{1}{K^2} \sum_{k=1}^K \theta_k^{g_2} + \frac{2}{K^2} \sum_{k_1 < k_2} \theta_{k_1 k_2}^{g_2} \\
&\quad - \frac{2}{K^2} \left( \sum_{k=1}^K \theta_k^{g_1} \theta_k^{g_2} + \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} \theta_{k_2}^{g_2} \right) \\
&= \frac{1}{K^2} \left[ \sum_{k=1}^K (\theta_k^{g_1} + \theta_k^{g_2} - 2\theta_k^{g_1} \theta_k^{g_2}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{g_1} + \theta_{k_1 k_2}^{g_2} - 2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2}) \right]
\end{aligned} \tag{4.3.48}$$

Under  $H_0$ :  $\theta_k^1 = \dots = \theta_k^G = \theta_k$  and  $\theta_{k_1 k_2}^1 = \dots = \theta_{k_1 k_2}^G = \theta_{k_1 k_2}$  Then,

$$\mu_{(g_1, g_2)3} = \mu_3 = \frac{2}{K^2} \left( \sum_{k=1}^K \theta_k (1 - \theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right) \quad (4.3.49)$$

$$\begin{aligned} & \psi_{(3)10}(\mathbf{x}_i^{g_1}) \\ &= E\{\phi_3(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_1}; \mathbf{X}_{i'}^{g_2}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)3}\} \\ &= E\left\{ \left[ \frac{1}{K} \sum_{k=1}^K I(x_{ik}^{g_1} \neq X_{jk}^{g_1}) - \frac{1}{K} \sum_{k=1}^K I(X_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \right]^2 - \mu_{(g_1, g_2)3} \right\} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) + 2 \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) + \sum_{k=1}^K \theta_k^{g_2} \right. \\ &\quad \left. + 2 \sum_{k_1 < k_2} \theta_{k_1 k_2}^{g_2} - 2 \sum_{k=1}^K \theta_k^{g_2} P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right\} - \mu_{(g_1, g_2)3} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) + 2 \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right. \\ &\quad \left. - 2 \sum_{k=1}^K \theta_k^{g_2} P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) - \sum_{k=1}^K \theta_k^{g_1} \right. \\ &\quad \left. - 2 \sum_{k_1 < k_2} \theta_{k_1 k_2}^{g_1} + 2 \sum_{k=1}^K \theta_k^{g_1} \theta_k^{g_2} + 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} \theta_{k_2}^{g_2} \right\} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) + 2 \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right. \\ &\quad \left. + \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right. \\ &\quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right\} \end{aligned}$$

Under  $H_0$ ,

$$\begin{aligned} & \psi_{(3)10}(\mathbf{x}_i) \\ &= E\{\phi_3(\mathbf{x}_i, \mathbf{X}_j; \mathbf{X}_{i'}, \mathbf{X}_{j'}) - \mu_3\} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K \theta_k (2\theta_k - 1) - \sum_{k=1}^K (2\theta_k - 1) P(x_{ik} \neq X_{jk}) \right. \\ &\quad \left. + 2 \sum_{k_1 < k_2} P(x_{ik_1} \neq X_{jk_1}, x_{ik_2} \neq X_{jk_2}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2} P(x_{ik_1} \neq X_{jk_1}) \right\} \end{aligned}$$

$$+ \sum_{k_1 \neq k_2} (2\theta_{k_1} \theta_{k_2} - \theta_{k_1 k_2}) \Big\}$$

Similarly,

$$\begin{aligned} & \psi_{(3)01}(\mathbf{x}_{i'}^{g_2}) \\ &= E\{\phi_3(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{x}_{i'}^{g_2}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)3}\} \\ &= E\left\{\left[\frac{1}{K} \sum_{k=1}^K I(X_{ik}^{g_1} \neq X_{jk}^{g_1}) - \frac{1}{K} \sum_{k=1}^K I(x_{i'k}^{g_2} \neq X_{j'k}^{g_2})\right]^2 - \mu_{(g_1, g_2)3}\right\} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K \theta_k^{g_1} + 2 \sum_{k_1 < k_2} \theta_{k_1 k_2}^{g_1} + \sum_{k=1}^K P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \right. \\ & \quad + 2 \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}, x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) - 2 \sum_{k=1}^K \theta_k^{g_1} P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \\ & \quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \right\} - \mu_{(g_1, g_2)3} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) + 2 \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right. \\ & \quad - 2 \sum_{k=1}^K \theta_k^{g_1} P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) - \sum_{k=1}^K \theta_k^{g_2} \\ & \quad \left. - 2 \sum_{k_1 < k_2} \theta_{k_1 k_2}^{g_2} + 2 \sum_{k=1}^K \theta_k^{g_1} \theta_k^{g_2} + 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} \theta_{k_2}^{g_2} \right\} \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) + 2 \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}, x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right. \\ & \quad + \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \\ & \quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right\} \end{aligned}$$

Under  $H_0$ ,  $\psi_{(3)10}(\mathbf{x}_i) = \psi_{(3)01}(\mathbf{x}_{i'})$ . Then,

$$\begin{aligned} & \psi_{(3)10}^2(\mathbf{x}_i^{g_1}) \\ &= \frac{1}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right)^2 \\ & \quad + \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right)^2 \end{aligned}$$

$$\begin{aligned}
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right) \\
& + \frac{2}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_1}^{g_1} \neq X_{jk_2}^{g_1}) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right) \\
& - \frac{8}{K^4} \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right) \\
& + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right) \\
& + \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& = \frac{1}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right)^2 \\
& + \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right) \\
& + \frac{2}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right)
\end{aligned}$$

$$\begin{aligned}
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& + \frac{4}{K^4} \left[ \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_1}^{g_1} \neq X_{jk_2}^{g_1}) (1 - 2\theta_{k_1}^{g_2}) P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \right. \\
& + \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_1}^{g_1} \neq X_{jk_2}^{g_1}) (1 - 2\theta_{k_2}^{g_2}) P(x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \\
& + \left. \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_1}^{g_1} \neq X_{jk_2}^{g_1}) (1 - 2\theta_{k_3}^{g_2}) P(x_{ik_3}^{g_1} \neq X_{jk_3}^{g_1}) \right] \\
& - \frac{8}{K^4} \left[ \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \theta_{k_1}^{g_2} P(x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right. \\
& + \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \theta_{k_1}^{g_2} P(x_{ik_3}^{g_1} \neq X_{jk_3}^{g_1}) \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \theta_{k_3}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \theta_{k_2}^{g_2} P(x_{ik_3}^{g_1} \neq X_{jk_3}^{g_1}) \\
& + \left. \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \theta_{k_3}^{g_2} P(x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right] \\
& + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}; x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& - \frac{4}{K^4} \left[ \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_2}) P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \theta_{k_1}^{g_2} P(x_{ik_2}^{g_1} \neq X_{jk_2}^{g_1}) \right. \\
& + \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_2}) P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \theta_{k_2}^{g_2} P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \\
& + \left. \sum_{k_1 \neq k_2 \neq k_3} (1 - 2\theta_{k_1}^{g_2}) P(x_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}) \theta_{k_2}^{g_2} P(x_{ik_3}^{g_1} \neq X_{jk_3}^{g_1}) \right] \\
& + \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(x_{ik}^{g_1} \neq X_{jk}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right)
\end{aligned}$$

$$- \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{i k_1}^{g_1} \neq X_{j k_1}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right)$$

Similarly,

$$\begin{aligned} & \psi_{(3)01}^2(\mathbf{x}_{i'}^{g_2}) \\ &= \frac{1}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}; x_{i' k_2}^{g_2} \neq X_{j' k_2}^{g_2}) \right)^2 \\ &+ \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i' k}^{g_2} \neq X_{j' k}^{g_2}) \right)^2 \\ &+ \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right)^2 \\ &+ \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 < k_2} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}; x_{i' k_2}^{g_2} \neq X_{j' k_2}^{g_2}) \right) \\ &+ \frac{2}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i' k}^{g_2} \neq X_{j' k}^{g_2}) \right) \\ &- \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}) \right) \\ &- \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\ &+ \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}; x_{i' k_2}^{g_2} \neq X_{j' k_2}^{g_2}) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i' k}^{g_2} \neq X_{j' k}^{g_2}) \right) \\ &- \frac{8}{K^4} \left( \sum_{k_1 < k_2} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}; x_{i' k_2}^{g_2} \neq X_{j' k_2}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}) \right) \\ &+ \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}; x_{i' k_2}^{g_2} \neq X_{j' k_2}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\ &- \frac{4}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i' k}^{g_2} \neq X_{j' k}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}) \right) \\ &+ \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i' k}^{g_2} \neq X_{j' k}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\ &- \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\ &= \frac{1}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{i' k_1}^{g_2} \neq X_{j' k_1}^{g_2}; x_{i' k_2}^{g_2} \neq X_{j' k_2}^{g_2}) \right)^2 \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \right)^2 + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right) \\
& + \frac{2}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\
& + \frac{4}{K^4} \left[ \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_1}^{g_2} \neq X_{j'k_2}^{g_2}) (1 - 2\theta_{k_1}^{g_1}) P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \right. \\
& + \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_1}^{g_2} \neq X_{j'k_2}^{g_2}) (1 - 2\theta_{k_2}^{g_1}) P(x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \\
& + \left. \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_1}^{g_2} \neq X_{j'k_3}^{g_2}) (1 - 2\theta_{k_3}^{g_1}) P(x_{i'k_3}^{g_2} \neq X_{j'k_3}^{g_2}) \right] \\
& - \frac{8}{K^4} \left[ \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \theta_{k_1}^{g_1} P(x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right. \\
& + \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \theta_{k_1}^{g_1} P(x_{i'k_3}^{g_2} \neq X_{j'k_3}^{g_2}) \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \theta_{k_3}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \theta_{k_2}^{g_1} P(x_{i'k_3}^{g_2} \neq X_{j'k_3}^{g_2}) \\
& + \left. \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \theta_{k_3}^{g_1} P(x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right] \\
& + \frac{4}{K^4} \left( \sum_{k_1 < k_2} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}; x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right)
\end{aligned}$$



$$\begin{aligned}
& - \frac{4}{K^4} \left[ \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_1}) P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \theta_{k_1}^{g_1} P(x_{i'k_2}^{g_2} \neq X_{j'k_2}^{g_2}) \right. \\
& + \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_1}) P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \\
& \left. + \sum_{k_1 \neq k_2 \neq k_3} (1 - 2\theta_{k_1}^{g_1}) P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \theta_{k_2}^{g_1} P(x_{i'k_3}^{g_2} \neq X_{j'k_3}^{g_2}) \right] \\
& + \frac{1}{K^4} \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(x_{i'k}^{g_2} \neq X_{j'k}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\
& - \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} P(x_{i'k_1}^{g_2} \neq X_{j'k_1}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right)
\end{aligned}$$

$$\begin{aligned}
\xi_{10}^{(3)} & = E\{\psi_{10}^2(\mathbf{X}_i^{g_1})\} \\
& = \frac{1}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right)^2 \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 < k_2} \left[ \sum_{u, v=0}^1 (p_{k_1 k_2}^{g_1}(u, v))^2 p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \right. \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_1}(u, v) p_{k_1 k_3}^{g_1}(u, t) p_{k_1 k_2 k_3}^{g_1}(1-u, 1-v, 1-t) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_3}^{g_1}(u, t) p_{k_2 k_3}^{g_1}(v, t) p_{k_1 k_2 k_3}^{g_1}(1-u, 1-v, 1-t) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_1}(u, v) p_{k_2 k_3}^{g_1}(v, t) p_{k_1 k_2 k_3}^{g_1}(1-u, 1-v, 1-t) \right] \\
& + 2 \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \left[ \sum_{u, v, t, z=0}^1 p_{k_1 k_3}^{g_1}(u, t) p_{k_2 k_4}^{g_1}(v, z) \right. \\
& \quad \left. \times p_{k_1 k_2 k_3 k_4}^{g_1}(1-u, 1-v, 1-t, 1-z) \right] \left. \right\} \\
& + \frac{1}{K^4} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{g_2})^2 \left[ \sum_{u=0}^1 (1 - p_k^{g_1}(u))^2 p_k^{g_1}(u) \right] \right. \\
& + 2 \sum_{k_1 < k_2} (1 - 2\theta_{k_1}^{g_2})(1 - 2\theta_{k_2}^{g_2}) \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u)) (1 - p_{k_2}^{g_1}(v)) p_{k_1 k_2}^{g_1}(u, v) \right] \left. \right\} \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 \neq k_2} (\theta_{k_2}^{g_2})^2 \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_1}(u))^2 p_{k_1}^{g_1}(u) \right] \right. \\
& + 2 \sum_{k=1}^K \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} (\theta_{k_1}^{g_2})^2 \left[ \sum_{u, v=0}^1 (1 - p_{k_2}^{g_1}(u)) (1 - p_{k_3}^{g_1}(v)) p_{k_2 k_3}^{g_1}(u, v) \right] \left. \right\}
\end{aligned}$$

$$\begin{aligned}
& + \sum_{k=1}^K \sum_{k_1 \neq k_2 \neq k_3} \theta_{k_2}^{g_2} \theta_{k_3}^{g_2} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_1}(u))^2 p_{k_1}^{g_1}(u) \right] \\
& + 2 \sum_{k_1 < k_2} \theta_{k_1}^{g_2} \theta_{k_2}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u)) (1 - p_{k_2}^{g_1}(v)) p_{k_1 k_2}^{g_1}(u, v) \right] \\
& + 2 \sum_{k=1}^K \sum_{k_1 \neq k_2 \neq k_3} \theta_{k_1}^{g_2} \theta_{k_2}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u)) (1 - p_{k_3}^{g_1}(v)) p_{k_1 k_3}^{g_1}(u, v) \right] \\
& + 2 \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \theta_{k_1}^{g_2} \theta_{k_3}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_2}^{g_1}(u)) (1 - p_{k_4}^{g_1}(v)) p_{k_2 k_4}^{g_1}(u, v) \right] \Big\} \\
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 < k_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \right) \\
& + \frac{2}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) \left[ \sum_{u=0}^1 (1 - p_k^{g_1}(u)) p_k^{g_1}(u) \right] \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} \left[ \sum_{u=0}^1 (1 - p_k^{g_1}(u)) p_k^{g_1}(u) \right] \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_1} (2\theta_k^{g_2} - 1) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 < k_2} (1 - 2\theta_{k_1}^{g_2}) \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_1}^{g_1}(1-u)) p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \right. \\
& + \sum_{k_1 < k_2} (1 - 2\theta_{k_2}^{g_2}) \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_2}^{g_1}(1-v)) p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} (1 - 2\theta_{k_3}^{g_2}) \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_3}^{g_1}(t)) p_{k_1 k_2 k_3}^{g_1}(1-u, 1-v, t) \right] \Big\} \\
& - \frac{8}{K^4} \left\{ \sum_{k_1 < k_2} \theta_{k_1}^{g_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_2}^{g_1}(1-v)) p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \right. \\
& + \sum_{k_1 < k_2} \theta_{k_2}^{g_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_1}^{g_1}(1-u)) p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_1}^{g_2} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_3}^{g_1}(t)) p_{k_1 k_2 k_3}^{g_1}(1-u, 1-v, t) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_3}^{g_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_1}^{g_1}(1-u)) p_{k_1 k_2}^{g_1}(1-u, 1-v) \right] \Big\}
\end{aligned}$$

$$\begin{aligned}
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_2}^{g_2} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_3}^{g_1}(t)) p_{k_1 k_2 k_3}^{g_1}(1 - u, 1 - v, t) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_3}^{g_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) (1 - p_{k_2}^{g_1}(1 - v)) p_{k_1 k_2}^{g_1}(1 - u, 1 - v) \right] \Big\} \\
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \left( \sum_{k_1 < k_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) p_{k_1 k_2}^{g_1}(1 - u, 1 - v) \right] \right) \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_2}) \theta_{k_1}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u)) (1 - p_{k_2}^{g_1}(v)) p_{k_1 k_2}^{g_1}(u, v) \right] \right. \\
& + \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_2}) \theta_{k_2}^{g_2} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_1}(u))^2 p_{k_1}^{g_1}(u) \right] \\
& + \left. \sum_{k_1 \neq k_2 \neq k_3} (1 - 2\theta_{k_1}^{g_2}) \theta_{k_2}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u)) (1 - p_{k_3}^{g_1}(v)) p_{k_1 k_3}^{g_1}(u, v) \right] \right\} \\
& + \frac{1}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_2}) \left[ \sum_{u=0}^1 (1 - p_k^{g_1}(u)) p_k^{g_1}(u) \right] \right) \\
& - \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_1}) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_1}(u)) p_{k_1}^{g_1}(u) \right] \right)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\xi_{01}^{(3)} & = E\{\psi_{01}^2(\mathbf{X}_i^{g_2})\} \\
& = \frac{1}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right)^2 \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 < k_2} \left[ \sum_{u, v=0}^1 (p_{k_1 k_2}^{g_2}(u, v))^2 p_{k_1 k_2}^{g_2}(1 - u, 1 - v) \right] \right. \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_2}(u, v) p_{k_1 k_3}^{g_2}(u, t) p_{k_1 k_2 k_3}^{g_2}(1 - u, 1 - v, 1 - t) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_3}^{g_2}(u, t) p_{k_2 k_3}^{g_2}(v, t) p_{k_1 k_2 k_3}^{g_2}(1 - u, 1 - v, 1 - t) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_2}(u, v) p_{k_2 k_3}^{g_2}(v, t) p_{k_1 k_2 k_3}^{g_2}(1 - u, 1 - v, 1 - t) \right] \\
& + 2 \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \left[ \sum_{u, v, t, z=0}^1 p_{k_1 k_3}^{g_2}(u, t) p_{k_2 k_4}^{g_2}(v, z) \right. \\
& \left. \times p_{k_1 k_2 k_3 k_4}^{g_2}(1 - u, 1 - v, 1 - t, 1 - z) \right] \Big\}
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{K^4} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{g_1})^2 \left[ \sum_{u=0}^1 (1 - p_k^{g_2}(u))^2 p_k^{g_2}(u) \right] \right. \\
& + 2 \sum_{k_1 < k_2} (1 - 2\theta_{k_1}^{g_1})(1 - 2\theta_{k_2}^{g_1}) \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_2}(u)) (1 - p_{k_2}^{g_2}(v)) p_{k_1 k_2}^{g_2}(u, v) \right] \Big\} \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 \neq k_2} (\theta_{k_2}^{g_1})^2 \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_2}(u))^2 p_{k_1}^{g_2}(u) \right] \right. \\
& + 2 \sum_{k=1}^K \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} (\theta_{k_1}^{g_1})^2 \left[ \sum_{u, v=0}^1 (1 - p_{k_2}^{g_2}(u)) (1 - p_{k_3}^{g_2}(v)) p_{k_2 k_3}^{g_2}(u, v) \right] \\
& + \sum_{k=1}^K \sum_{k_1 \neq k_2 \neq k_3} \theta_{k_2}^{g_1} \theta_{k_3}^{g_1} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_2}(u))^2 p_{k_1}^{g_2}(u) \right] \\
& + 2 \sum_{k_1 < k_2} \theta_{k_1}^{g_1} \theta_{k_2}^{g_1} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_2}(u)) (1 - p_{k_2}^{g_2}(v)) p_{k_1 k_2}^{g_2}(u, v) \right] \\
& + 2 \sum_{k=1}^K \sum_{k_1 \neq k_2 \neq k_3} \theta_{k_1}^{g_1} \theta_{k_2}^{g_1} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_2}(u)) (1 - p_{k_3}^{g_2}(v)) p_{k_1 k_3}^{g_2}(u, v) \right] \\
& + 2 \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \theta_{k_1}^{g_1} \theta_{k_3}^{g_1} \left[ \sum_{u, v=0}^1 (1 - p_{k_2}^{g_2}(u)) (1 - p_{k_4}^{g_2}(v)) p_{k_2 k_4}^{g_2}(u, v) \right] \Big\} \\
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right)^2 \\
& + \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 < k_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_2}(u, v) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \right) \\
& + \frac{2}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) \left[ \sum_{u=0}^1 (1 - p_k^{g_2}(u)) p_k^{g_2}(u) \right] \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} \left[ \sum_{u=0}^1 (1 - p_k^{g_2}(u)) p_k^{g_2}(u) \right] \right) \\
& - \frac{4}{K^4} \left( \sum_{k=1}^K \theta_k^{g_2} (2\theta_k^{g_1} - 1) \right) \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 < k_2} (1 - 2\theta_{k_1}^{g_1}) \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_2}(u, v) (1 - p_{k_1}^{g_2}(1-u)) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \right. \\
& + \sum_{k_1 < k_2} (1 - 2\theta_{k_2}^{g_1}) \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_2}(u, v) (1 - p_{k_2}^{g_2}(1-v)) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} (1 - 2\theta_{k_3}^{g_1}) \left[ \sum_{u, v, t=0}^1 p_{k_1 k_2}^{g_2}(u, v) (1 - p_{k_3}^{g_2}(t)) p_{k_1 k_2 k_3}^{g_2}(1-u, 1-v, t) \right] \Big\}
\end{aligned}$$

$$\begin{aligned}
& - \frac{8}{K^4} \left\{ \sum_{k_1 < k_2} \theta_{k_1}^{g_1} \left[ \sum_{u,v=0}^1 p_{k_1 k_2}^{g_2}(u,v) (1 - p_{k_2}^{g_2}(1-v)) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \right. \\
& + \sum_{k_1 < k_2} \theta_{k_2}^{g_1} \left[ \sum_{u,v=0}^1 p_{k_1 k_2}^{g_2}(u,v) (1 - p_{k_1}^{g_2}(1-u)) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_1}^{g_1} \left[ \sum_{u,v,t=0}^1 p_{k_1 k_2}^{g_2}(u,v) (1 - p_{k_3}^{g_2}(t)) p_{k_1 k_2 k_3}^{g_2}(1-u, 1-v, t) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_3}^{g_1} \left[ \sum_{u,v=0}^1 p_{k_1 k_2}^{g_2}(u,v) (1 - p_{k_1}^{g_2}(1-u)) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_2}^{g_1} \left[ \sum_{u,v,t=0}^1 p_{k_1 k_2}^{g_2}(u,v) (1 - p_{k_3}^{g_2}(t)) p_{k_1 k_2 k_3}^{g_2}(1-u, 1-v, t) \right] \\
& + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2}} \theta_{k_3}^{g_1} \left[ \sum_{u,v=0}^1 p_{k_1 k_2}^{g_2}(u,v) (1 - p_{k_2}^{g_2}(1-v)) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \left. \right\} \\
& + \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \left( \sum_{k_1 < k_2} \left[ \sum_{u,v=0}^1 p_{k_1 k_2}^{g_2}(u,v) p_{k_1 k_2}^{g_2}(1-u, 1-v) \right] \right) \\
& + \frac{4}{K^4} \left\{ \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_1}) \theta_{k_1}^{g_1} \left[ \sum_{u,v=0}^1 (1 - p_{k_1}^{g_2}(u)) (1 - p_{k_2}^{g_2}(v)) p_{k_1 k_2}^{g_2}(u,v) \right] \right. \\
& + \sum_{k_1 \neq k_2} (1 - 2\theta_{k_1}^{g_1}) \theta_{k_2}^{g_1} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_2}(u))^2 p_{k_1}^{g_2}(u) \right] \\
& + \left. \sum_{k_1 \neq k_2 \neq k_3} (1 - 2\theta_{k_1}^{g_1}) \theta_{k_2}^{g_1} \left[ \sum_{u,v=0}^1 (1 - p_{k_1}^{g_2}(u)) (1 - p_{k_3}^{g_2}(v)) p_{k_1 k_3}^{g_2}(u,v) \right] \right\} \\
& + \frac{1}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \left( \sum_{k=1}^K (1 - 2\theta_k^{g_1}) \left[ \sum_{u=0}^1 (1 - p_k^{g_2}(u)) p_k^{g_2}(u) \right] \right) \\
& - \frac{4}{K^4} \left( \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{g_2} - \theta_{k_1 k_2}^{g_2}) \right) \left( \sum_{k_1 \neq k_2} \theta_{k_2}^{g_1} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_2}(u)) p_{k_1}^{g_2}(u) \right] \right)
\end{aligned}$$

$\mathbf{U}_3^{(2,2)}$  (4.3.47) can be decomposed as

$$\begin{aligned}
\mathbf{U}_3^{(2,2)} &= \mu_{(g_1, g_2)3} + \frac{2}{N} \sum_{i=1}^N [\Psi_{(3)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)3}] \\
&+ \frac{2}{N} \sum_{i'=1}^N [\Psi_{(3)01}(\mathbf{X}_{i'}^{g_2}) - \mu_{(g_1, g_2)3}] + O_p(N^{-1}) \quad (4.3.50)
\end{aligned}$$

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

$$U_{4,1}^{(2,2)} = \left[ \binom{N}{2} \binom{N}{2} \right]^{-1} \sum_{\substack{i \neq i' \\ j \neq j'}} \sum_{\substack{j \neq j' \\ j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 \quad (4.3.51)$$

$$U_{4,2}^{(2,2)} = \left[ \binom{N}{2} \binom{N}{2} \right]^{-1} \sum_{\substack{i \neq i' \\ i \neq j}} \sum_{\substack{j \neq j' \\ j \neq i'}} (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 \quad (4.3.52)$$

Let

$$\begin{aligned} \mu_{(g_1, g_2)4} &= E(D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_2)})^2 = E[(D_{ij}^{(g_1, g_2)})^2 + (D_{i'j'}^{(g_1, g_2)})^2 - 2D_{ij}^{(g_1, g_2)} D_{i'j'}^{(g_1, g_2)}] \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k^{(g_1, g_2)} (1 - \theta_k^{(g_1, g_2)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} - \theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_2)}) \right\} \end{aligned}$$

where

$$\theta_k^{(g_1, g_2)} = P(X_{ik}^{g_1} \neq X_{jk}^{g_2}) = \sum_{u=0}^1 p_k^{g_1}(u) p_k^{g_2}(1-u)$$

and

$$\theta_{k_1 k_2}^{(g_1, g_2)} = P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{jk_2}^{g_2}) = \sum_{u,v=0}^1 p_{k_1 k_2}^{g_1}(u, v) p_{k_1 k_2}^{g_2}(1-u, 1-v)$$

Under  $H_0$ :  $\theta_k^{(g_1, g_2)} = \theta_k^g = \theta_k$  and  $\theta_{k_1 k_2}^{(g_1, g_2)} = \theta_{k_1 k_2}^g = \theta_{k_1} \theta_{k_2}$ .

$$\mu_{(g_1, g_2)4} = \mu_4 = \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k (1 - \theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right\} \quad (4.3.53)$$

$\psi_{(4)10}(\mathbf{x}_i^{g_1})$

$$\begin{aligned} &= E(\phi_4(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)4}) \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_2)}) P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_2} \neq x_{ik_2}^{g_1}) \right. \\ &\quad + \sum_{k=1}^K \theta_k^{(g_1, g_2)} (2\theta_k^{(g_1, g_2)} - 1) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_2)} - \theta_{k_1 k_2}^{(g_1, g_2)}) \\ &\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_1, g_2)} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}) \right\} \end{aligned}$$

Under  $H_0$ ,

$$\begin{aligned}
\psi_{(4)10}(\mathbf{x}_i) &= E(\phi_4(\mathbf{x}_i, \mathbf{X}_j, \mathbf{X}_{i'}, \mathbf{X}_{j'}) - \mu_4) \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k) P(X_{jk} \neq x_{ik}) + \sum_{k_1 \neq k_2} P(X_{jk_1} \neq x_{ik_1}, X_{jk_2} \neq x_{ik_2}) \right. \\
&\quad + \sum_{k=1}^K \theta_k (2\theta_k - 1) + \sum_{k_1 \neq k_2} (2\theta_{k_1} \theta_{k_2} - \theta_{k_1 k_2}) \\
&\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2} P(X_{jk_1} \neq x_{ik_1}) \right\}
\end{aligned}$$

$$\begin{aligned}
\psi_{(4)01}(\mathbf{x}_j^{g_2}) &= E(\phi_4(\mathbf{X}_i^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)4}) \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_2)}) P(X_{ik}^{g_1} \neq x_{jk}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq x_{jk_2}^{g_2}) \right. \\
&\quad + \sum_{k=1}^K \theta_k^{(g_1, g_2)} (2\theta_k^{(g_1, g_2)} - 1) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_2)} - \theta_{k_1 k_2}^{(g_1, g_2)}) \\
&\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_1, g_2)} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}) \right\}
\end{aligned}$$

Again, under  $H_0$ ,  $\psi_{(4)01}(\mathbf{x}_j) = \psi_{(4)10}(\mathbf{x}_i)$  and

$$\xi_{10}^{(4)} = \xi_{01}^{(4)} = E[\psi_{(4)10}^2(\mathbf{X}_i)]$$

The decomposition of  $U_{4,k}^{(2,2)}$ ,  $k = 1, 2$ , is then

$$\begin{aligned}
U_{4,k}^{(2,2)} &= \mu_{(g_1, g_2)4} + \frac{2}{N} \sum_{i=1}^N [\Psi_{(4)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)4}] \\
&\quad + \frac{2}{N} \sum_{j=1}^N [\Psi_{(4)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)4}] + O_p(N^{-1}) \tag{4.3.54}
\end{aligned}$$

For the two-sample U-statistic of degree (1,2),

$$U_{51}^{(1,2)} = \left[ \binom{N}{1} \binom{N}{2} \right]^{-1} \sum_{\substack{i=1 \\ i \neq j'}}^N \sum_{\substack{1 \leq j, j' \leq N \\ j \neq j'}} (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 \tag{4.3.55}$$

Let

$$\mu_{(g_1, g_2)5} = E(D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_2)})^2 = E[(D_{ij}^{(g_1, g_2)})^2 + (D_{ij'}^{(g_1, g_2)})^2 - 2D_{ij}^{(g_1, g_2)} D_{ij'}^{(g_1, g_2)}]$$

$$= \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k^{(g_1, g_2)} + \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_2)} - \sum_{k=1}^K \theta_k^{(g_1, g_2; g_1, g_2)}(i, j; i, j') \right. \\ \left. - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_2; g_1, g_2)}(i, j; i, j') \right\}$$

where

$$\theta_k^{(g_1, g_2; g_1, g_2)}(i, j; i, j') = P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) = \sum_{u=0}^1 p_k^{g_1}(u)[1 - p_k^{g_2}(u)]^2$$

$$\theta_{k_1 k_2}^{(g_1, g_2; g_1, g_2)}(i, j; i, j') = P(X_{ik_1}^{g_1} \neq X_{jk_2}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_1}^{g_2}) \\ = \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v)[1 - p_{k_1}^{g_2}(u)][1 - p_{k_2}^{g_2}(v)]$$

and under  $H_0$ ,

$$\mu_{(g_1, g_2)5} = \mu_5 \\ = \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k + \sum_{k_1 \neq k_2} \theta_{k_1 k_2} - \sum_{k=1}^K \theta_k(i, j; i, j') - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}(i, j; i, j') \right\}$$

Note that under  $H_0$ ,  $\mu_1 = \mu_5$ .

$$\psi_{(5)10}(\mathbf{x}_i^{g_1}) = E[\phi_5(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)5}] \\ = \frac{2}{K^2} \left\{ \sum_{k=1}^K P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_2} \neq x_{ik_2}^{g_1}) \right. \\ \left. - \sum_{k=1}^K P(X_{jk}^{g_2} \neq x_{ik}^{g_1})P(X_{j'k}^{g_2} \neq x_{ik}^{g_1}) - \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1})P(X_{j'k_2}^{g_2} \neq x_{ik_2}^{g_1}) \right. \\ \left. + \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_2}) \right. \\ \left. - \sum_{k=1}^K \theta_k^{(g_1, g_2)} - \theta_{k_1 k_2}^{(g_1, g_2)} \right\}$$

$$\psi_{(5)01}(\mathbf{x}_j^{g_2}) = E[\phi_5(\mathbf{X}_i^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)5}] \\ = \frac{1}{K^2} \left\{ \sum_{k=1}^K P(X_{ik}^{g_1} \neq x_{jk}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq x_{jk_2}^{g_2}) \right. \\ \left. - 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq x_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) - 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_2}) \right\}$$



$$\begin{aligned}
& + \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_2}) \\
& - \left. \sum_{k=1}^K \theta_k^{(g_1, g_2)} - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_2)} \right\}
\end{aligned}$$

For the two-sample U-statistics of degree (2,1)

$$\mathbf{U}_{5,2}^{(2,1)} = \left[ \binom{N}{2} \binom{N}{1} \right]^{-1} \sum_{\substack{j=1 \\ j \neq i'}}^N \sum_{i \neq i'} (D_{ij}^{(g_1, g_2)} - D_{i'j}^{(g_1, g_2)})^2, \quad (4.3.56)$$

$$E(\mathbf{U}_{5,2}^{(2,1)}) = E(\mathbf{U}_{5,1}^{(1,2)}) = \mu_{(g_1, g_2)5}$$

and the decompositions for  $\mathbf{U}_{5,1}^{(1,2)}$  and  $\mathbf{U}_{5,2}^{(2,1)}$  are

$$\begin{aligned}
\mathbf{U}_{5,1}^{(1,2)} &= \mu_{(g_1, g_2)5} + \frac{1}{N} \sum_{i=1}^N [\Psi_{(5)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)5}] \\
&+ \frac{2}{N} \sum_{j=1}^N [\Psi_{(5)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)5}] + O_p(N^{-1}) \quad (4.3.57)
\end{aligned}$$

$$\begin{aligned}
\mathbf{U}_{5,2}^{(2,1)} &= \mu_{(g_1, g_2)5} + \frac{2}{N} \sum_{i=1}^N [\Psi_{(5)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)5}] \\
&+ \frac{1}{N} [\Psi_{(5)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)5}] + O_p(N^{-1}) \quad (4.3.58)
\end{aligned}$$

The other two-sample U-statistics of degree (2,1) are

$$\mathbf{U}_{6,1}^{(2,1)} = \left[ \binom{N}{2} \binom{N}{1} \right]^{-1} \sum_{1 \leq i < j \leq N} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \quad \text{and} \quad (4.3.59)$$

$$\mathbf{U}_{6,2}^{(2,1)} = \left[ \binom{N}{2} \binom{N}{1} \right]^{-1} \sum_{1 \leq i < j \leq N} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 \quad (4.3.60)$$

Now

$$\begin{aligned}
\mu_{(g_1, g_2)6} &= E(D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 = E(D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{g_1} + \theta_k^{(g_1, g_2)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{g_1} + \theta_{k_1 k_2}^{(g_1, g_2)}) \right. \\
&\quad \left. - 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_1}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) - 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_2}) \right\} \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{g_1} + \theta_k^{(g_1, g_2)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{g_1} + \theta_{k_1 k_2}^{(g_1, g_2)}) \right. \\
&\quad \left. - 2 \sum_{k=1}^K \theta_k^{(g_1, g_1; g_1, g_2)}(i, j; i, j') - 2 \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_1; g_1, g_2)}(i, j; i, j') \right\}
\end{aligned}$$

and under  $H_0$ ,

$$\begin{aligned}\mu_{(g_1, g_2)6} &= \mu_6 = \mu_1 = \mu_5 \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k + \sum_{k_1 \neq k_2} \theta_{k_1 k_2} - \sum_{k=1}^K \theta_k(i, j; i, j') - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}(i, j; i, j') \right\}\end{aligned}$$

Since the kernels

$$\phi_{61}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{j'}^{g_2}) = (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \quad \text{and} \quad \phi_{62}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{j'}^{g_2}) = (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2$$

are not symmetric, we need to work with the symmetric kernels

$$\begin{aligned}\phi'_{61} &= \frac{1}{2} \left\{ (D_{ij}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 + (D_{ji}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 \right\} \quad \text{and} \\ \phi'_{62} &= \frac{1}{2} \left\{ (D_{ij}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 + (D_{ji}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \right\} \quad \text{with}\end{aligned}$$

$$\begin{aligned}\psi_{(6)10}(\mathbf{x}_i^{g_1}) &\equiv \mathbb{E}[\phi'_{61}(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)6}] \\ &= \frac{1}{2K^2} \left\{ 2 \sum_{k=1}^K P(X_{jk}^{g_1} \neq x_{ik}^{g_1}) + 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) \right. \\ &\quad + \sum_{k=1}^K P(X_{j'k}^{g_2} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{j'k_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{j'k_2}^{g_2} \neq x_{ik_2}^{g_1}) \\ &\quad - 2 \sum_{k=1}^K P(X_{jk}^{g_1} \neq x_{ik}^{g_1}, X_{j'k}^{g_2} \neq x_{ik}^{g_1}) - 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{j'k_2}^{g_2} \neq x_{ik_2}^{g_1}) \\ &\quad - 2 \sum_{k=1}^K P(X_{jk}^{g_1} \neq x_{ik}^{g_1}, X_{jk}^{g_1} \neq X_{j'k}^{g_2}) - 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq X_{j'k_2}^{g_2}) \\ &\quad + 4 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_1}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) + 4 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_2}) \\ &\quad \left. - \sum_{k=1}^K (2\theta_k^{g_1} + \theta_k^{(g_1, g_2)}) - \sum_{k_1 \neq k_2} (2\theta_{k_1 k_2}^{g_1} + \theta_{k_1 k_2}^{(g_1, g_2)}) \right\}\end{aligned}$$

and

$$\begin{aligned}\psi_{(6)01}(\mathbf{x}_{j'}^{g_2}) &\equiv \mathbb{E}[\phi'_{61}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{x}_{j'}^{g_2}) - \mu_{(g_1, g_2)6}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(X_{ik}^{g_1} \neq x_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{j'k_1}^{g_2}, X_{ik_2}^{g_1} \neq x_{j'k_2}^{g_2}) \right. \\ &\quad \left. - 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_1}, X_{ik}^{g_1} \neq x_{j'k}^{g_2}) - 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, X_{ik_2}^{g_1} \neq x_{j'k_2}^{g_2}) \right\}\end{aligned}$$

$$\begin{aligned}
& + 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_1}, X_{ik}^{g_1} \neq X_{j'k}^{g_2}) + 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_1}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_2}) \\
& - \left. \sum_{k=1}^K \theta_k^{(g_1, g_2)} - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_2)} \right\}
\end{aligned}$$

$U_{6,k}^{(2,1)}$ , for  $k = 1, 2$ , can be decomposed as

$$\begin{aligned}
U_{6,k}^{(2,1)} &= \mu_{(g_1, g_2)6} + \frac{2}{N} \sum_{i=1}^N [\Psi_{(6)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)6}] \\
&+ \frac{1}{N} \sum_{j'=1}^N [\Psi_{(6)01}(\mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)6}] + O_p(n^{-1}) . \tag{4.3.61}
\end{aligned}$$

For the two-sample U-statistics of degree (3,1)

$$U_{7,1}^{(3,1)} = \left[ \binom{N}{3} \binom{N}{1} \right]^{-1} \sum_{i < j < i'} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2, \tag{4.3.62}$$

$$U_{7,2}^{(3,1)} = \left[ \binom{N}{3} \binom{N}{1} \right]^{-1} \sum_{i' < i < j} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \text{ and} \tag{4.3.63}$$

$$U_{7,3}^{(3,1)} = \left[ \binom{N}{3} \binom{N}{1} \right]^{-1} \sum_{i < i' < j} \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2, \tag{4.3.64}$$

$$\begin{aligned}
\mu_{(g_1, g_2)7} &= E(D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K [\theta_k^{g_1} (1 - 2\theta_k^{(g_1, g_2)}) + \theta_k^{(g_1, g_2)}] + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{g_1} + \theta_{k_1 k_2}^{(g_1, g_2)} - 2\theta_{k_1}^{g_1} \theta_{k_2}^{(g_1, g_2)}) \right\}
\end{aligned}$$

which becomes under  $H_0$

$$\begin{aligned}
\mu_{(g_1, g_2)7} &= \mu_7 = \mu_2 = \mu_3 = \mu_4 \\
&= \frac{2}{K^2} \left\{ \sum_{k=1}^K [\theta_k (1 - \theta_k)] + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right\}
\end{aligned}$$

Since the kernels

$$\phi_{71}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) = (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2, \quad i < j < i',$$

$$\phi_{72}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) = (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2, \quad i' < i < j \text{ and}$$

$$\phi_{73}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) = (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2, \quad i < i' < j$$

are not symmetric, we switch to the symmetric ones, i.e.,

$$\phi'_{71}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2})$$

$$\begin{aligned}
&= \frac{1}{3} \left\{ (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 + (D_{ii'}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 + (D_{ji'}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \right\}, \quad i < j < i', \\
\phi'_{72}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) \\
&= \frac{1}{3} \left\{ (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 + (D_{ii'}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 + (D_{ji'}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \right\}, \quad i' < i < j \text{ and} \\
\phi'_{73}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) \\
&= \frac{1}{3} \left\{ (D_{ij}^{g_1} - D_{i'j'}^{(g_1, g_2)})^2 + (D_{ii'}^{g_1} - D_{jj'}^{(g_1, g_2)})^2 + (D_{ji'}^{g_1} - D_{ij'}^{(g_1, g_2)})^2 \right\}, \quad i < i' < j
\end{aligned}$$

with

$$\begin{aligned}
\psi_{(7)10}(\mathbf{x}_i^{g_1}) &\equiv \mathbb{E}[\phi'_{71}(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)7}] \\
&= \frac{1}{3K^2} \left\{ \sum_{k=1}^K (2 - 4\theta_k^{(g_1, g_2)}) P(X_{jk}^{g_1} \neq x_{ik}^{g_1}) + 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) \right. \\
&\quad - \sum_{k=1}^K (\theta_k^{(g_1, g_2)} + 2\theta_k^{g_1} - 6\theta_k^{g_1} \theta_k^{(g_1, g_2)}) - \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} + 2\theta_{k_1 k_2}^{g_1} - 6\theta_{k_1}^{g_1} \theta_{k_2}^{(g_1, g_2)}) \\
&\quad - 4 \sum_{k_1 \neq k_2} (\theta_{k_2}^{(g_1, g_2)} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}) + \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(X_{j'k}^{g_2} \neq x_{ik}^{g_1})) \\
&\quad \left. + \sum_{k_1 \neq k_2} P(X_{j'k_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{j'k_2}^{g_2} \neq x_{ik_2}^{g_1}) - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} P(X_{j'k_2}^{g_2} \neq x_{ik_2}^{g_1}) \right\}
\end{aligned}$$

and

$$\begin{aligned}
\psi_{(7)01}(\mathbf{x}_{j'}^{g_2}) &\equiv \mathbb{E}[\phi'_{71}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{x}_{j'}^{g_2}) - \mu_{(g_1, g_2)7}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(X_{i'k}^{g_1} \neq x_{j'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{i'k_1}^{g_1} \neq x_{j'k_1}^{g_2}, X_{i'k_2}^{g_1} \neq x_{j'k_2}^{g_2}) \right. \\
&\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} P(X_{i'k_2}^{g_1} \neq x_{j'k_2}^{g_2}) - \sum_{k=1}^K \theta_k^{(g_1, g_2)} (1 - 2\theta_k^{g_1}) \\
&\quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{(g_1, g_2)} - \theta_{k_1 k_2}^{(g_1, g_2)}) \right\}
\end{aligned}$$

Note that under  $H_0$ ,  $\psi_{(7)10}(\mathbf{x}_i) = \psi_{(7)01}(\mathbf{x}_{j'})$  and  $\xi_{10}^{(7)} = \xi_{01}^{(7)} \mathbb{E}(\psi_{(7)10}^2(\mathbf{X}_i))$ . The decomposition of  $\mathbf{U}_{7,k}^{(s,1)}$ , for  $k = 1, 2, 3$  is

$$\begin{aligned}
\mathbf{U}_{7,k}^{(s,1)} &= \mu_{(g_1, g_2)7} + \frac{3}{N} \sum_{i=1}^N [\Psi_{(7)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)7}] \\
&\quad + \frac{1}{N} \sum_{j'=1}^N [\Psi_{(7)01}(\mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)7}] + O_p(N^{-1}) \tag{4.3.65}
\end{aligned}$$

The three-sample U-statistics of degree (2,1,1) is

$$U_8^{(2,1,1)} = \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i < j} \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{g_1} - D_{i'j'}^{(g_2, g_3)})^2 \quad (4.3.66)$$

with

$$\begin{aligned} \mu_{(g_1 g_2 g_3)8} &= E(D_{ij}^{g_1} - D_{i'j'}^{(g_2, g_3)})^2 \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{g_1} + \theta_k^{(g_2, g_3)} - 2\theta_k^{g_1} \theta_k^{(g_2, g_3)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{g_1} + \theta_{k_1 k_2}^{(g_2, g_3)} - 2\theta_{k_1}^{g_1} \theta_{k_2}^{(g_2, g_3)}) \right\} \end{aligned}$$

which becomes under  $H_0$

$$\begin{aligned} \mu_{(g_1 g_2 g_3)8} &= \mu_8 = \mu_2 = \mu_3 = \mu_4 = \mu_7 \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K [\theta_k(1 - \theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2})] \right\}. \end{aligned}$$

Also,

$$\begin{aligned} \psi_{(8)100}(\mathbf{x}_i^{g_1}) &\equiv E[\phi_8(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_2}, \mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)8}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_2, g_3)}) P(X_{jk}^{g_1} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) \right. \\ &\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_2, g_3)} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}) - \sum_{k=1}^K \theta_k^{g_1} (1 - 2\theta_k^{(g_2, g_3)}) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{(g_2, g_3)} - \theta_{k_1 k_2}^{g_1}) \right\} \end{aligned}$$

$$\begin{aligned} \psi_{(8)010}(\mathbf{x}_i^{g_2}) &\equiv E[\phi_8(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{x}_i^{g_2}, \mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)8}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(X_{j'k}^{g_3} \neq x_{i'k}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{j'k_1}^{g_3} \neq x_{i'k_1}^{g_2}, X_{j'k_2}^{g_3} \neq x_{i'k_2}^{g_2}) \right. \\ &\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} P(X_{j'k_2}^{g_3} \neq x_{i'k_2}^{g_2}) - \sum_{k=1}^K \theta_k^{(g_2, g_3)} (1 - 2\theta_k^{g_1}) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{(g_2, g_3)} - \theta_{k_1 k_2}^{(g_2, g_3)}) \right\} \end{aligned}$$

$$\begin{aligned} \psi_{(8)001}(\mathbf{x}_{j'}^{g_3}) &\equiv E[\phi_8(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_1}, \mathbf{X}_{i'}^{g_2}, \mathbf{x}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)8}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{g_1}) P(X_{i'k}^{g_2} \neq x_{j'k}^{g_3}) + \sum_{k_1 \neq k_2} P(X_{i'k_1}^{g_2} \neq x_{j'k_1}^{g_3}, X_{i'k_2}^{g_2} \neq x_{j'k_2}^{g_3}) \right. \\ &\quad \left. - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{g_1} P(X_{i'k_2}^{g_2} \neq x_{j'k_2}^{g_3}) - \sum_{k=1}^K \theta_k^{(g_2, g_3)} (1 - 2\theta_k^{g_1}) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{g_1} \theta_{k_2}^{(g_2, g_3)} - \theta_{k_1 k_2}^{(g_2, g_3)}) \right\} \end{aligned}$$

Under  $H_0$ ,  $\psi_{(8)100}^2(\mathbf{x}_i) = \psi_{(8)010}^2(\mathbf{x}_i) = \psi_{(8)001}^2(\mathbf{x}_{j'})$  and

$$\xi_{100}^{(8)} = \xi_{010}^{(8)} = \xi_{001}^{(8)} = E(\psi_{(8)100}^2(\mathbf{X}_i))$$

The decomposition for  $U_8^{(2,1,1)}$  is

$$\begin{aligned} U_8^{(2,1,1)} &= \mu_{(g_1 g_2 g_3)8} + \frac{2}{N} \sum_{i=1}^N [\Psi_{(8)100}(\mathbf{X}_i^{g_1}) - \mu_{(g_1 g_2 g_3)8}] \\ &+ \frac{1}{N} \sum_{i'=1}^N [\Psi_{(8)010}(\mathbf{X}_{i'}^{g_2}) - \mu_{(g_1 g_2 g_3)8}] + \frac{1}{N} \sum_{j'=1}^N [\Psi_{(8)001}(\mathbf{X}_{j'}^{g_3}) - \mu_{(g_1 g_2 g_3)8}] \\ &+ O_p(N^{-1}) \end{aligned} \quad (4.3.67)$$

The other three-sample U-statistics of degree (2,1,1) are

$$U_{9,1}^{(2,1,1)} = \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i \neq i'} \sum_{j=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 \quad (4.3.68)$$

$$U_{9,2}^{(2,1,1)} = \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i \neq j} \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 \quad (4.3.69)$$

with

$$\begin{aligned} \mu_{(g_1 g_2 g_3)9} &= E(D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{(g_1, g_2)} + \theta_k^{(g_1, g_3)} - 2\theta_k^{(g_1, g_2)} \theta_k^{(g_1, g_3)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} \right. \\ &\quad \left. + \theta_{k_1 k_2}^{(g_1, g_3)} - 2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_3)}) \right\} \end{aligned}$$

which becomes under  $H_0$

$$\begin{aligned} \mu_{(g_1 g_2 g_3)9} &= \mu_9 = \mu_2 = \mu_3 = \mu_4 = \mu_7 = \mu_8 \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K [\theta_k(1 - \theta_k)] + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right\} \end{aligned}$$

Since

$$\phi_{91}(\mathbf{X}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) = (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2, \quad i \neq i', j \neq j' \text{ and}$$

$$\phi_{92}(\mathbf{X}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) = (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2, \quad i \neq j, i \neq i'$$

are not symmetric, we consider the symmetric kernels

$$\phi'_{91}(\mathbf{X}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) = \frac{1}{2} [(D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 + (D_{i'j}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2],$$

$i \neq i', j \neq j'$  and

$$\phi'_{92}(\mathbf{X}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) = \frac{1}{2} [(D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_1, g_3)})^2 + (D_{i'j}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2],$$

$i \neq j, i \neq i'$  with

$$\psi_{(9)100}(\mathbf{x}_i^{g_1}) \equiv E[\phi'_{91}(\mathbf{x}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)9}]$$

$$\begin{aligned}
&= \frac{1}{2K^2} \left\{ \sum_{k=1}^K P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_2} \neq x_{ik_2}^{g_1}) \right. \\
&\quad - 2 \sum_{k=1}^K \theta_k^{(g_1, g_3)} P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_1, g_3)} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}) \\
&\quad + \sum_{k=1}^K P(X_{j'k}^{g_3} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{j'k_1}^{g_3} \neq x_{ik_1}^{g_1}, X_{j'k_2}^{g_3} \neq x_{ik_2}^{g_1}) \\
&\quad - 2 \sum_{k=1}^K \theta_k^{(g_1, g_2)} P(X_{j'k}^{g_3} \neq x_{ik}^{g_1}) - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{(g_1, g_2)} P(X_{j'k_2}^{g_3} \neq x_{ik_2}^{g_1}) \\
&\quad + \sum_{k=1}^K (4\theta_k^{(g_1, g_2)} \theta_k^{(g_1, g_3)} - \theta_k^{(g_1, g_3)} - \theta_k^{(g_1, g_2)}) \\
&\quad \left. + \sum_{k_1 \neq k_2} (4\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_3)} - 2\theta_{k_1 k_2}^{(g_1, g_3)} - 2\theta_{k_1 k_2}^{(g_1, g_2)}) \right\}
\end{aligned}$$

$$\begin{aligned}
\psi_{(9)010}(\mathbf{x}_j^{g_2}) &\equiv E[\phi'_{91}(\mathbf{X}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)9}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_3)}) P(X_{ik}^{g_1} \neq x_{jk}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq x_{jk_2}^{g_2}) \right. \\
&\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_1, g_3)} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}) - \sum_{k=1}^K (\theta_k^{(g_1, g_2)} - 2\theta_k^{(g_1, g_2)} \theta_k^{(g_1, g_3)}) \\
&\quad \left. - \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} - 2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_3)}) \right\}
\end{aligned}$$

$$\begin{aligned}
\psi_{(9)001}(\mathbf{x}_{j'}^{g_3}) &\equiv E[\phi'_{91}(\mathbf{X}_i^{g_1}, \mathbf{X}_{i'}^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{x}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)9}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_2)}) P(X_{i'k}^{g_1} \neq x_{j'k}^{g_3}) + \sum_{k_1 \neq k_2} P(X_{i'k_1}^{g_1} \neq x_{j'k_1}^{g_3}, X_{i'k_2}^{g_1} \neq x_{j'k_2}^{g_3}) \right. \\
&\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{(g_1, g_2)} P(X_{i'k_2}^{g_1} \neq x_{j'k_2}^{g_3}) - \sum_{k=1}^K (\theta_k^{(g_1, g_3)} - 2\theta_k^{(g_1, g_2)} \theta_k^{(g_1, g_3)}) \\
&\quad \left. - \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_3)} - 2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_1, g_3)}) \right\}
\end{aligned}$$

Under  $H_0$ ,  $\psi_{(9)100}(\mathbf{x}_i) = \psi_{(9)010}(\mathbf{x}_j) = \psi_{(9)001}(\mathbf{x}_{j'})$  and

$$\xi_{100}^{(9)} = \xi_{010}^{(9)} = \xi_{001}^{(9)} = E(\psi_{(9)100}^2(\mathbf{X}_i))$$

The decomposition for  $\mathbf{U}_{9,k}^{(2,1,1)}$  for  $k = 1, 2$  is

$$\mathbf{U}_{9,k}^{(2,1,1)} = \mu_{(g_1, g_2, g_3)9} + \frac{2}{N} \sum_{i=1}^N [\Psi_{(9)100}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2, g_3)9}]$$

$$\begin{aligned}
& + \frac{1}{N} \sum_{j=1}^N [\Psi_{(9)010}(\mathbf{X}_j^{g_2}) - \mu_{(g_1 g_2 g_3)9}] + \frac{1}{N} \sum_{j'=1}^N [\Psi_{(9)001}(\mathbf{X}_{j'}^{g_3}) - \mu_{(g_1 g_2 g_3)9}] \\
& + O_p(N^{-1})
\end{aligned} \tag{4.3.70}$$

For the three-sample U-statistic of degree (1,1,1)

$$\mathbf{U}_{10}^{(1,1,1)} = \left[ \binom{N}{1} \binom{N}{1} \binom{N}{1} \right]^{-1} \sum_{i=1}^N \sum_{j=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2, \text{ since} \tag{4.3.71}$$

$$\begin{aligned}
\mu_{(g_1 g_2 g_3)10} & = E(D_{ij}^{(g_1, g_2)} - D_{ij'}^{(g_1, g_3)})^2 \\
& = \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{(g_1, g_2)} + \theta_k^{(g_1, g_3)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} + \theta_{k_1 k_2}^{(g_1, g_3)}) \right. \\
& \quad - 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_3}) \\
& \quad \left. - 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_3}) \right\} \\
& = \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{(g_1, g_2)} + \theta_k^{(g_1, g_3)}) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} + \theta_{k_1 k_2}^{(g_1, g_3)}) \right. \\
& \quad \left. - 2 \sum_{k=1}^K \theta_k^{(g_1, g_2; g_1, g_3)}(i, j; i, j') - 2 \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_2; g_1, g_3)}(i, j; i, j') \right\}
\end{aligned}$$

where

$$\theta_k^{(g_1, g_2; g_1, g_3)}(i, j; i, j') = P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_3}) = \sum_{u=0}^1 p_k^{g_1}(u) p_k^{g_2}(1-u) p_k^{g_3}(1-u)$$

$$\begin{aligned}
\theta_{k_1 k_2}^{(g_1, g_2; g_1, g_3)}(i, j; i, j') & = P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_3}) \\
& = \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) p_{k_1}^{g_2}(1-u) p_{k_2}^{g_3}(1-v)
\end{aligned}$$

then under  $H_0$

$$\begin{aligned}
\mu_{(g_1 g_2 g_3)10} & = \mu_{10} = \mu_1 = \mu_5 = \mu_6 \\
& = \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k + \sum_{k_1 \neq k_2} \theta_{k_1 k_2} - \sum_{k=1}^K \theta_k(i, j; i, j') - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}(i, j; i, j') \right\}.
\end{aligned}$$

Also,

$$\psi_{(10)100}(\mathbf{x}_i^{g_1}) \equiv E[\phi_{10}(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)10}]$$



$$\begin{aligned}
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_2} \neq x_{ik_2}^{g_1}) \right. \\
&+ \sum_{k=1}^K P(X_{j'k}^{g_3} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{j'k_1}^{g_3} \neq x_{ik_1}^{g_1}, X_{j'k_2}^{g_3} \neq x_{ik_2}^{g_1}) \\
&- 2 \sum_{k=1}^K P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) P(X_{j'k}^{g_3} \neq x_{ik}^{g_1}) - 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}) P(X_{j'k_2}^{g_3} \neq x_{ik_2}^{g_1}) \\
&+ 2 \sum_{k=1}^K P(X_{jk}^{g_2} \neq X_{ik}^{g_1}, X_{j'k}^{g_3} \neq x_{ik}^{g_1}) + 2 \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq X_{ik_1}^{g_1}, X_{j'k_2}^{g_3} \neq X_{ik_2}^{g_1}) \\
&\left. - \sum_{k=1}^K (\theta_k^{(g_1, g_2)} + \theta_k^{(g_1, g_3)}) - \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} + \theta_{k_1 k_2}^{(g_1, g_3)}) \right\}
\end{aligned}$$

$$\begin{aligned}
\psi_{(10)010}(\mathbf{x}_j^{g_2}) &\equiv E[\phi_{10}(\mathbf{X}_i^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)10}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - \theta_k^{(g_1, g_3)}) P(X_{ik}^{g_1} \neq x_{jk}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq x_{jk_2}^{g_2}) \right. \\
&- 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}) \theta_{k_2}^{(g_1, g_3)} + 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_3}) \\
&\left. + 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_3}) - \sum_{k=1}^K \theta_k^{(g_1, g_2)} - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_2)} \right\}
\end{aligned}$$

and

$$\begin{aligned}
\psi_{(10)001}(\mathbf{x}_{j'}^{g_3}) &\equiv E[\phi_{10}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{x}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)10}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - \theta_k^{(g_1, g_2)}) P(X_{ik}^{g_1} \neq x_{j'k}^{g_3}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{j'k_1}^{g_3}, X_{ik_2}^{g_1} \neq x_{j'k_2}^{g_3}) \right. \\
&- 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{j'k_1}^{g_3}) \theta_{k_2}^{(g_1, g_2)} + 2 \sum_{k=1}^K P(X_{ik}^{g_1} \neq X_{jk}^{g_2}, X_{ik}^{g_1} \neq X_{j'k}^{g_3}) \\
&\left. + 2 \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq X_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq X_{j'k_2}^{g_3}) - \sum_{k=1}^K \theta_k^{(g_1, g_3)} - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}^{(g_1, g_3)} \right\}
\end{aligned}$$

The decomposition for  $\mathbf{U}_{10}^{(1,1,1)}$  is

$$\begin{aligned}
\mathbf{U}_{10}^{(1,1,1)} &= \mu_{(g_1 g_2 g_3)10} + \frac{1}{N} \sum_{i=1}^N [\Psi_{(10)100}(\mathbf{X}_i^{g_1}) - \mu_{(g_1 g_2 g_3)10}] \\
&+ \frac{1}{N} \sum_{j=1}^N [\Psi_{(10)010}(\mathbf{X}_j^{g_2}) - \mu_{(g_1 g_2 g_3)10}] + \frac{1}{N} \sum_{j'=1}^N [\Psi_{(10)001}(\mathbf{X}_{j'}^{g_3}) - \mu_{(g_1 g_2 g_3)10}] \\
&+ O_p(N^{-1}) \tag{4.3.72}
\end{aligned}$$

For the four-sample U-statistic of degree (1,1,1,1)

$$U_{11}^{(1,1,1,1)} = N^{-4} \sum_{i=1}^N \sum_{j=1}^N \sum_{i'=1}^N \sum_{j'=1}^N (D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_3, g_4)})^2, \quad (4.3.73)$$

$$\begin{aligned} \mu_{(g_1 g_2 g_3 g_4)11} &= E(D_{ij}^{(g_1, g_2)} - D_{i'j'}^{(g_3, g_4)})^2 \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (\theta_k^{(g_1, g_2)} + \theta_k^{(g_3, g_4)} - 2\theta_k^{(g_1, g_2)} \theta_k^{(g_3, g_4)}) \right. \\ &\quad \left. + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2}^{(g_1, g_2)} + \theta_{k_1 k_2}^{(g_3, g_4)} - 2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_3, g_4)}) \right\} \end{aligned}$$

and under  $H_0$

$$\begin{aligned} \mu_{(g_1 g_2 g_3 g_4)11} &= \mu_{11} = \mu_2 = \mu_3 = \mu_4 = \mu_7 = \mu_8 = \mu_9 \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K [\theta_k(1 - \theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2})] \right\}. \end{aligned}$$

Now

$$\begin{aligned} \psi_{(11)1000}(\mathbf{x}_i^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{x}_{i'}^{g_3}, \mathbf{x}_{j'}^{g_4}) &\equiv E[\phi_{11}(\mathbf{x}_i^{g_1}, \mathbf{x}_j^{g_2}, \mathbf{x}_{i'}^{g_3}, \mathbf{x}_{j'}^{g_4}) - \mu_{(g_1, g_2, g_3, g_4)11}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_2)}) P(X_{jk}^{g_2} \neq x_{ik}^{g_1}) + \sum_{k_1 \neq k_2} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_2} \neq x_{ik_2}^{g_1}) \right. \\ &\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_3, g_4)} P(X_{jk_1}^{g_2} \neq x_{ik_1}^{g_1}) + \sum_{k=1}^K \theta_k^{(g_1, g_2)} (2\theta_k^{(g_3, g_4)} - 1) \\ &\quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_3, g_4)} - \theta_{k_1 k_2}^{(g_1, g_2)}) \right\} \end{aligned}$$

$$\begin{aligned} \psi_{(11)0100}(\mathbf{x}_i^{g_2}, \mathbf{x}_j^{g_1}, \mathbf{x}_{i'}^{g_3}, \mathbf{x}_{j'}^{g_4}) &\equiv E[\phi_{11}(\mathbf{x}_i^{g_2}, \mathbf{x}_j^{g_1}, \mathbf{x}_{i'}^{g_3}, \mathbf{x}_{j'}^{g_4}) - \mu_{(g_1, g_2, g_3, g_4)11}] \\ &= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_3, g_4)}) P(X_{ik}^{g_1} \neq x_{jk}^{g_2}) + \sum_{k_1 \neq k_2} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}, X_{ik_2}^{g_1} \neq x_{jk_2}^{g_2}) \right. \\ &\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_2}^{(g_3, g_4)} P(X_{ik_1}^{g_1} \neq x_{jk_1}^{g_2}) + \sum_{k=1}^K \theta_k^{(g_1, g_2)} (2\theta_k^{(g_3, g_4)} - 1) \\ &\quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_3, g_4)} - \theta_{k_1 k_2}^{(g_1, g_2)}) \right\} \end{aligned}$$

$$\psi_{(11)0010}(\mathbf{x}_i^{g_3}, \mathbf{x}_j^{g_4}, \mathbf{x}_{i'}^{g_1}, \mathbf{x}_{j'}^{g_2}) \equiv E[\phi_{11}(\mathbf{x}_i^{g_3}, \mathbf{x}_j^{g_4}, \mathbf{x}_{i'}^{g_1}, \mathbf{x}_{j'}^{g_2}) - \mu_{(g_1, g_2, g_3, g_4)11}]$$

$$\begin{aligned}
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_2)}) P(X_{j'k}^{g_4} \neq x_{i'k}^{g_3}) + \sum_{k_1 \neq k_2} P(X_{j'k_1}^{g_4} \neq x_{i'k_1}^{g_3}, X_{j'k_2}^{g_4} \neq x_{i'k_2}^{g_3}) \right. \\
&\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{(g_1, g_2)} P(X_{j'k_2}^{g_4} \neq x_{i'k_2}^{g_3}) + \sum_{k=1}^K \theta_k^{(g_3, g_4)} (2\theta_k^{(g_1, g_2)} - 1) \\
&\quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_3, g_4)} - \theta_{k_1 k_2}^{(g_3, g_4)}) \right\}
\end{aligned}$$

$$\begin{aligned}
\psi_{(11)0001}(\mathbf{x}_{j'}^{g_4}) &\equiv E[\phi_{11}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_2}, \mathbf{X}_{i'}^{g_3}, \mathbf{x}_{j'}^{g_4}) - \mu_{(g_1, g_2, g_3, g_4)11}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^{(g_1, g_2)}) P(X_{i'k}^{g_3} \neq x_{j'k}^{g_4}) + \sum_{k_1 \neq k_2} P(X_{i'k_1}^{g_3} \neq x_{j'k_1}^{g_4}, X_{i'k_2}^{g_3} \neq x_{j'k_2}^{g_4}) \right. \\
&\quad - 2 \sum_{k_1 \neq k_2} \theta_{k_1}^{(g_1, g_2)} P(X_{i'k_2}^{g_3} \neq x_{j'k_2}^{g_4}) + \sum_{k=1}^K \theta_k^{(g_3, g_4)} (2\theta_k^{(g_1, g_2)} - 1) \\
&\quad \left. + \sum_{k_1 \neq k_2} (2\theta_{k_1}^{(g_1, g_2)} \theta_{k_2}^{(g_3, g_4)} - \theta_{k_1 k_2}^{(g_3, g_4)}) \right\}
\end{aligned}$$

Under  $H_0$ ,  $\psi_{(11)1000}(\mathbf{x}_i^{g_1}) = \psi_{(11)0100}(\mathbf{x}_j^{g_2}) = \psi_{(11)0010}(\mathbf{x}_{i'}^{g_3}) = \psi_{(11)0001}(\mathbf{x}_{j'}^{g_4})$  and

$$\xi_{1000}^{(11)} = \xi_{0100}^{(11)} = \xi_{0010}^{(11)} = \xi_{0001}^{(11)} = E[\psi_{(11)1000}^2(\mathbf{X}_i)]$$

The decomposition for  $\mathbf{U}_{11}^{(1,1,1,1)}$  is then

$$\begin{aligned}
\mathbf{U}_{11}^{(1,1,1,1)} &= \mu_{(g_1, g_2, g_3, g_4)11} + \frac{1}{N} \sum_{i=1}^N [\Psi_{(11)1000}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2, g_3, g_4)11}] \\
&\quad + \frac{1}{N} \sum_{j=1}^N [\Psi_{(11)0100}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2, g_3, g_4)11}] \\
&\quad + \frac{1}{N} \sum_{i'=1}^N [\Psi_{(11)0010}(\mathbf{X}_{i'}^{g_3}) - \mu_{(g_1, g_2, g_3, g_4)11}] \\
&\quad + \frac{1}{N} \sum_{j'=1}^N [\Psi_{(11)0001}(\mathbf{X}_{j'}^{g_4}) - \mu_{(g_1, g_2, g_3, g_4)11}] + O_p(N^{-1}) \quad (4.3.74)
\end{aligned}$$

## 4.4 Combining the U-statistics

We know that

$$WSS = \frac{2}{N(N-1)} \sum_{g=1}^G \left[ \binom{N}{3} (\mathbf{U}_{1,1}^{(g)} + \mathbf{U}_{1,2}^{(g)} + \mathbf{U}_{1,3}^{(g)}) + \binom{N}{4} (\mathbf{U}_{2,1}^{(g)} + \mathbf{U}_{2,2}^{(g)} + \mathbf{U}_{2,3}^{(g)}) \right]$$

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

with expected value

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

Under  $H_0$ , there is homogeneity among groups, i.e., for any  $g$ ,  $\theta_k^g = \theta_k$  and  $\theta_{k_1 k_2}^g = \theta_{k_1 k_2}$ , thus

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

where

$$\mu_1 = \frac{2}{K^2} \left[ \sum_{k=1}^K \theta_k + \sum_{k_1 \neq k_2} \theta_{k_1 k_2} - \sum_{k=1}^K \theta_k(i, j; i, j') - \sum_{k_1 \neq k_2} \theta_{k_1 k_2}(i, j; i, j') \right] \quad (4.4.1)$$

and

$$\mu_2 = \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k(1 - \theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right\} \quad (4.4.2)$$

Note that

$$\theta_k = P(X_{ik} \neq X_{jk}) = \sum_{c=0}^{C-1} p_k(c)[1 - p_k(c)] \quad (4.4.3)$$

$$\begin{aligned}
\theta_{k_1 k_2} &= P(X_{ik_1} \neq X_{jk_1}; X_{ik_2} \neq X_{jk_2}) \\
&= \sum_{c_1, c_2=0}^{C-1} p_{k_1 k_2}(c_1, c_2) \left[ \sum_{\substack{c_3=0 \\ c_3 \neq c_1}}^{C-1} \sum_{\substack{c_4=0 \\ c_4 \neq c_2}}^{C-1} p_{k_1 k_2}(c_3, c_4) \right]
\end{aligned} \quad (4.4.4)$$

Decomposing  $WSS$ ,

$$\begin{aligned}
WSS &= \frac{(N-2)}{3} \sum_{g=1}^G 3 \left\{ \mu_{g1} + \frac{3}{N} \sum_{i=1}^N [\Psi_{(1)1}(\mathbf{X}_i^g) - \mu_{g1}] + O_p(N^{-1}) \right\} \\
&+ \frac{(N-2)(N-3)}{12} \sum_{g=1}^G 3 \left\{ \mu_{g2} + \frac{4}{N} \sum_{i=1}^N [\Psi_{(2)1}(\mathbf{X}_i^g) - \mu_{g2}] \right. \\
&\left. + O_p(N^{-1}) \right\}
\end{aligned} \quad (4.4.5)$$

and under  $H_0$ ,

$$\begin{aligned}
WSS &= G(N-2) \left( \mu_1 + \frac{(N-3)}{4} \mu_2 \right) \\
&\quad + (N-2) \frac{3}{N} G \sum_{i=1}^N [\Psi_{(1)1}(\mathbf{X}_i) - \mu_1] + O_p(1) \\
&\quad + \frac{(N-2)(N-3)}{N} G \sum_{i=1}^N [\Psi_{(2)1}(\mathbf{X}_i) - \mu_2] + O_p(N) \tag{4.4.6}
\end{aligned}$$

and the associated mean square expression is

$$\begin{aligned}
WMS &\equiv \frac{WSS}{G \binom{N}{2}} = \frac{2WSS}{GN(N-1)} \\
&= \frac{(N-2)(N-3)}{N(N-1)2} \left\{ \mu_2 + \frac{4}{N} \sum_{i=1}^N [\Psi_{(2)1}(\mathbf{X}_i) - \mu_2] \right\} + O_p(N^{-1}) \tag{4.4.7}
\end{aligned}$$

with

$$\begin{aligned}
E_0(WMS) &= \frac{(N-2)(N-3)}{2N(N-1)} \mu_2 + O(N^{-1}) \\
&= \frac{\mu_2}{2} + O(N^{-1})
\end{aligned}$$

Let  $\mathbf{U}_{wss} = (\mathbf{U}_{1,1}^{(s)}, \mathbf{U}_{1,2}^{(s)}, \mathbf{U}_{1,3}^{(s)}, \mathbf{U}_{2,1}^{(4)}, \mathbf{U}_{2,2}^{(4)}, \mathbf{U}_{2,3}^{(4)})'$  and  $\Sigma_{wss}$  be the variance-covariance matrix of  $\mathbf{U}_{wss}$ . In order to get the variance of  $WSS$ , we need to know the elements of  $\Sigma_{wss}$ . Since  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$  are i.i.d.,  $\text{Cov}(\mathbf{U}_{1,k}^{(s)}, \mathbf{U}_{1,l}^{(s)})$  and  $\text{Cov}(\mathbf{U}_{2,k}^{(4)}, \mathbf{U}_{2,l}^{(4)})$  (for  $1 \leq k < l \leq 3$ ) are nothing but  $\text{Var}(\mathbf{U}_{1,k}^{(s)})$  and  $\text{Var}(\mathbf{U}_{2,k}^{(4)})$ , respectively. To illustrate this result, compute  $\text{Cov}(\mathbf{U}_{2,1}^{(4)}, \mathbf{U}_{2,2}^{(4)})$ . From (4.3.29)

$$\begin{aligned}
\psi_{(2,1)1}(\mathbf{x}_1^g) &= E[\phi_{2,1}(\mathbf{x}_1^g, \mathbf{X}_2^g, \mathbf{X}_3^g, \mathbf{X}_4^g) - \mu_{g2}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^g) P(X_{2k}^g \neq x_{1k}^g) + \sum_{k=1}^K \theta_k^g (2\theta_k^g - 1) \right. \\
&\quad + \sum_{k_1 \neq k_2} P(X_{2k_1}^g \neq x_{1k_1}^g, X_{2k_2}^g \neq x_{1k_2}^g) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^g \theta_{k_2}^g - \theta_{k_1 k_2}^g) \\
&\quad \left. + 2 \sum_{k_1 \neq k_2} \theta_{k_2}^g P(X_{2k_1}^g \neq x_{1k_1}^g) \right\}
\end{aligned}$$

and

$$\begin{aligned}
\psi_{(2,2)1}(\mathbf{x}_1^g) &= E[\phi_{2,2}(\mathbf{x}_1^g, \mathbf{X}_6^g, \mathbf{X}_5^g, \mathbf{X}_7^g) - \mu_{g2}] \\
&= \frac{1}{K^2} \left\{ \sum_{k=1}^K (1 - 2\theta_k^g) P(X_{6k}^g \neq x_{1k}^g) + \sum_{k=1}^K \theta_k^g (2\theta_k^g - 1) \right.
\end{aligned}$$

$$\begin{aligned}
& + \sum_{k_1 \neq k_2} P(X_{6k_1}^g \neq x_{1k_1}^g, X_{6k_2}^g \neq x_{1k_2}^g) + \sum_{k_1 \neq k_2} (2\theta_{k_1}^g \theta_{k_2}^g - \theta_{k_1 k_2}^g) \\
& + 2 \sum_{k_1 \neq k_2} \theta_{k_2}^g P(X_{6k_1}^g \neq x_{1k_1}^g) \}
\end{aligned}$$

Hence,  $\psi_{(2,1)1}(\mathbf{x}_1^g) = \psi_{(2,2)1}(\mathbf{x}_1^g)$  and  $\psi_{(2,1)1}(\mathbf{x}_1^g)\psi_{(2,2)1}(\mathbf{x}_1^g) = \psi_{(2,1)1}^2(\mathbf{x}_1^g)$ .

Now, let

$$\begin{aligned}
\psi_{(1)1}(\mathbf{x}_i^g) &= \psi_{(1,1)1}(\mathbf{x}_i^g) = \psi_{(1,2)1}(\mathbf{x}_i^g) = \psi_{(1,3)1}(\mathbf{x}_i^g), \\
\psi_{(2)1}(\mathbf{x}_i^g) &= \psi_{(2,1)1}(\mathbf{x}_i^g) = \psi_{(2,2)1}(\mathbf{x}_i^g) = \psi_{(2,3)1}(\mathbf{x}_i^g), \\
\xi_1^{(1)} &= E(\psi_{(1)1}^2(\mathbf{X}_i^g)), \quad \xi_1^{(2)} = E(\psi_{(2)1}^2(\mathbf{X}_i^g)) \text{ and } \xi_1^{(1,2)} = E(\psi_{(1)1}(\mathbf{X}_i^g)\psi_{(2)1}(\mathbf{X}_i^g))
\end{aligned}$$

Therefore, the variance-covariance matrix of

$\mathbf{U}_{wss} = (\mathbf{U}_{1,1}^{(s)}, \mathbf{U}_{1,2}^{(s)}, \mathbf{U}_{1,3}^{(s)}, \mathbf{U}_{2,1}^{(4)}, \mathbf{U}_{2,2}^{(4)}, \mathbf{U}_{2,3}^{(4)})'$  is

$$\boldsymbol{\Sigma}_{wss} = \begin{pmatrix} 9\xi_1^{(1)} & 12\xi_1^{(1,2)} \\ 12\xi_1^{(1,2)} & 16\xi_1^{(2)} \end{pmatrix} \otimes \mathbf{J}_3$$

where  $\mathbf{J}_3$  is a  $3 \times 3$  matrix of 1's. So,

$\text{Var}(WSS)$

$$\begin{aligned}
&= \sum_{g=1}^G \left\{ \frac{(N-2)^2}{9} \text{Var}(\mathbf{U}_{1,1}^{(s)} + \mathbf{U}_{1,2}^{(s)} + \mathbf{U}_{1,3}^{(s)}) \right. \\
&+ \frac{(N-2)^2(N-3)^2}{144} \text{Var}(\mathbf{U}_{2,1}^{(4)} + \mathbf{U}_{2,2}^{(4)} + \mathbf{U}_{2,3}^{(4)}) \\
&+ \left. 2 \frac{(N-2)^2(N-3)}{36} \text{Cov}[\mathbf{U}_{1,1}^{(s)} + \mathbf{U}_{1,2}^{(s)} + \mathbf{U}_{1,3}^{(s)}, \mathbf{U}_{2,1}^{(4)} + \mathbf{U}_{2,2}^{(4)} + \mathbf{U}_{2,3}^{(4)}] \right\} \\
&= \sum_{g=1}^G \left\{ \frac{(N-2)^2}{9} \left( 3 \times \frac{9}{N} \xi_1^{(1)} + 2 \left( 3 \times \frac{9}{N} \xi_1^{(1)} \right) \right) \right. \\
&+ \frac{(N-2)^2(N-3)^2}{144} \left( 3 \times \frac{16}{N} \xi_1^{(2)} + 2 \left( 3 \times \frac{16}{N} \xi_1^{(2)} \right) \right) \\
&+ \left. 2 \frac{(N-2)^2(N-3)}{36} \left( 9 \times \frac{12}{N} \xi_1^{(1,2)} \right) \right\} \\
&= \frac{(N-2)^2}{N} \sum_{g=1}^G \left\{ 9\xi_1^{(1)} + (N-3)^2 \xi_1^{(2)} + 6(N-3) \xi_1^{(1,2)} \right\}
\end{aligned}$$

$$\begin{aligned}
\text{Var}(WMS) &= \frac{4}{G^2 N^2 (N-1)^2} \text{Var}(WSS) \\
&= \frac{4(N-2)^2(N-3)^2}{GN^2(N-1)^2} \frac{\xi_1^{(2)}}{N} + O(N^{-2})
\end{aligned}$$

For  $AWSS$ ,

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

and under  $H_0$

$$\begin{aligned}
E_0(AWSS) &= \frac{G(G-1)}{2} \left[ \frac{(N-1)^2}{2} \mu_4 + (N-1) \mu_5 \right] \\
&= \frac{G(G-1)(N-1)}{2} \left( \frac{(N-1)}{2} \mu_4 + \mu_5 \right)
\end{aligned}$$

where  $\mu_4 = \mu_2$  is given by (4.4.2) and  $\mu_5 = \mu_1$  is given by (4.4.1).

$AWSS$  can be decomposed as

$$\begin{aligned}
AWSS &= \sum_{1 \leq g_1 < g_2 \leq G} \left[ \frac{(N-1)^2}{2} \left( \mu_{(g_1, g_2)4} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(4)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)4}) \right. \right. \\
&\quad \left. \left. + \frac{2}{N} \sum_{j=1}^N (\Psi_{(4)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)4}) + O_p(N^{-1}) \right) \right. \\
&\quad \left. + \frac{(N-1)}{2} \left( 2\mu_{(g_1, g_2)5} + \frac{1}{N} \sum_{i=1}^N (\Psi_{(5)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)5}) \right. \right. \\
&\quad \left. \left. + \frac{2}{N} \sum_{j=1}^N (\Psi_{(5)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)5}) + \frac{2}{N} \sum_{i=1}^N (\Psi_{(5)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)5}) \right. \right. \\
&\quad \left. \left. + \frac{1}{N} \sum_{j=1}^N (\Psi_{(5)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)5}) + O_p(N^{-1}) \right) \right] \quad (4.4.8)
\end{aligned}$$

The associated mean-square expression is

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

$$\begin{aligned}
&= \frac{(N-1)^2}{N^2 G(G-1)} \sum_{1 \leq g_1 < g_2 \leq G} \left[ \mu_{(g_1, g_2)4} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(4)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)4}) \right. \\
&\quad \left. + \frac{2}{N} \sum_{j=1}^N (\Psi_{(4)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)4}) \right] + O_p(N^{-1}) \tag{4.4.9}
\end{aligned}$$

$$\begin{aligned}
E_0(AWMS) &= \frac{2}{N^2 G(G-1)} E_0(AWSS) \\
&= \frac{(N-1)}{N^2} \left( \frac{(N-1)}{2} \mu_4 + \mu_5 \right) \\
&= \frac{(N-1)^2}{2N^2} \mu_4 + O(N^{-1}) \\
&= \frac{\mu_4}{2} + O(N^{-1})
\end{aligned}$$

Note that  $E_0(AWMS) = E_0(WMS)$  since under  $H_0$ ,  $\mu_4 = \mu_2$

$$\begin{aligned}
\text{Var}(AWSS) &= \frac{(N-1)^2}{4} \sum_{1 \leq g_1 < g_2 \leq G} \left\{ \frac{(N-1)^2}{4} \text{Var}(\mathbf{U}_{4,1}^{(2,2)} + \mathbf{U}_{4,2}^{(2,2)}) \right. \\
&\quad \left. + \text{Var}(\mathbf{U}_{5,1}^{(1,2)} + \mathbf{U}_{5,2}^{(2,1)}) \right. \\
&\quad \left. + (N-1) \text{Cov}[\mathbf{U}_{4,1}^{(2,2)} + \mathbf{U}_{4,2}^{(2,2)}, \mathbf{U}_{5,1}^{(1,2)} + \mathbf{U}_{5,2}^{(2,1)}] \right\} \\
&= \frac{(N-1)^2}{4} \sum_{1 \leq g_1 < g_2 \leq G} \left\{ \frac{(N-1)^2}{4} \left( 4 \times \left( \frac{4}{N} \xi_{10}^{(4)} + \frac{4}{N} \xi_{01}^{(4)} \right) \right) \right. \\
&\quad \left. + \frac{1}{N} \xi_{10}^{(5,1)} + \frac{4}{N} \xi_{01}^{(5,1)} + \frac{4}{N} \xi_{10}^{(5,2)} + \frac{1}{N} \xi_{01}^{(5,2)} \right. \\
&\quad \left. + 2 \times \left( \frac{2}{N} \xi_{10}^{(5,1),(5,2)} + \frac{2}{N} \xi_{01}^{(5,1),(5,2)} \right) \right. \\
&\quad \left. + (N-1) \left( 2 \times \left( \frac{2}{N} \xi_{10}^{(4),(5,1)} + \frac{4}{N} \xi_{01}^{(4),(5,1)} \right) \right) \right. \\
&\quad \left. + 2 \times \left( \frac{4}{N} \xi_{10}^{(4),(5,2)} + \frac{2}{N} \xi_{01}^{(4),(5,2)} \right) \right\}
\end{aligned}$$

$$\begin{aligned}
\text{Var}(AWMS) &= \frac{4}{N^4 G^2 (G-1)^2} \text{Var}(AWSS) \\
&= \frac{(N-1)^4}{2N^4 G(G-1)} \left( \frac{4}{N} \xi_{10}^{(4)} + \frac{4}{N} \xi_{01}^{(4)} \right) + O(N^{-2})
\end{aligned}$$

For  $TSS$ ,

$$TSS = \frac{1}{NG(NG-1)} \left\{ \sum_{g=1}^G \left[ \binom{N}{3} (U_{1,1}^{(3)} + U_{1,2}^{(3)} + U_{1,3}^{(3)}) \right] \right\}$$



$$\begin{aligned}
& + \binom{N}{4} (U_{2,1}^{(4)} + U_{2,2}^{(4)} + U_{2,3}^{(4)}) + \sum_{1 \leq g_1 < g_2 \leq G} \left[ \binom{N}{2} \binom{N}{2} U_3^{(2,2)} \right] \\
& + \sum_{1 \leq g_1 < g_2 \leq G} \left[ \binom{N}{2} \binom{N}{2} (U_{4,1}^{(2,2)} + U_{4,2}^{(2,2)}) + \binom{N}{1} \binom{N}{2} (U_{5,1}^{(1,2)} + U_{5,2}^{(2,1)}) \right] \\
& + \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \left[ \binom{N}{1} \binom{N}{1} \binom{N}{1} U_{10}^{(1,1,1)} + \binom{N}{2} \binom{N}{1} \binom{N}{1} (U_{9,1}^{(2,1,1)} + U_{9,2}^{(2,1,1)}) \right] \\
& + \sum_{\substack{1 \leq g_1 \neq g_2 \neq g_3 \neq g_4 \leq G \\ g_1 < g_2, g_3 < g_4}} \left[ \binom{N}{1} \binom{N}{1} \binom{N}{1} \binom{N}{1} U_{11}^{(1,1,1,1)} \right] \\
& + \sum_{\substack{1 \leq g_1, g_2 \leq G \\ g_1 \neq g_2}} \left[ \binom{N}{2} \binom{N}{1} (U_{6,1}^{(2,1)} + U_{6,2}^{(2,1)}) \right] \\
& + \binom{N}{3} \binom{N}{1} (U_{7,1}^{(3,1)} + U_{7,2}^{(3,1)} + U_{7,3}^{(3,1)}) \\
& + \sum_{\substack{1 \leq g_1 \neq g_2 \neq g_3 \leq G \\ g_2 < g_3}} \left[ \binom{N}{2} \binom{N}{1} \binom{N}{1} U_8^{(2,1,1)} \right] \}
\end{aligned}$$

$$\begin{aligned}
E(TSS) &= \frac{1}{NG(NG-1)} \left\{ \frac{N(N-1)(N-2)}{2} \sum_{g=1}^G \mu_{g1} \right. \\
& + \frac{N(N-1)(N-2)(N-3)}{8} \sum_{g=1}^G \mu_{g2} \\
& + \frac{N^2(N-1)^2}{4} \sum_{1 \leq g_1 < g_2 \leq G} [\mu_{(g_1, g_2)3} + 2\mu_{(g_1, g_2)4}] \\
& + N^2(N-1) \sum_{1 \leq g_1 < g_2 \leq G} \mu_{(g_1, g_2)5} + N^3 \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \mu_{(g_1, g_2, g_3)10} \\
& + N^3(N-1) \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \mu_{(g_1, g_2, g_3)9} + N^4 \sum_{\substack{1 \leq g_1, g_2, g_3, g_4 \leq G \\ g_1 \neq g_2 \neq g_3 \neq g_4}} \mu_{(g_1, g_2, g_3, g_4)11} \\
& + N^2(N-1) \sum_{\substack{1 \leq g_1, g_2 \leq G \\ g_1 \neq g_2}} \mu_{(g_1, g_2)6} + \frac{N^2(N-1)(N-2)}{2} \sum_{\substack{1 \leq g_1, g_2 \leq G \\ g_1 \neq g_2}} \mu_{(g_1, g_2)7} \\
& \left. + \frac{N^3(N-1)}{2} \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \mu_{(g_1, g_2, g_3)8} \right\}
\end{aligned}$$

and under  $H_0$ ,

$$E_0(TSS) = \frac{1}{NG-1} \left\{ \frac{(N-1)(N-2)}{2} \mu_1 + \frac{(N-1)(N-2)(N-3)}{8} \mu_2 \right\}$$

$$\begin{aligned}
& + \frac{(G-1)}{8} N(N-1)^2 \mu_3 + \frac{(G-1)}{4} N(N-1)^2 \mu_4 \\
& + \frac{(G-1)}{2} N(N-1) \mu_5 + \frac{(G-1)(G-2)}{2} N^2 \mu_{10} \\
& + \frac{(G-1)(G-2)}{2} N^2 (N-1) \mu_9 + \frac{(G-1)(G-2)(G-3)}{8} N^3 \mu_{11} \\
& + (G-1) N(N-1) \mu_6 + (G-1) \frac{N(N-1)(N-2)}{2} \mu_7 \\
& + (G-1)(G-2) \frac{N^2(N-1)}{4} \mu_8 \}
\end{aligned}$$

$TSS$  can be decomposed as

$$\begin{aligned}
TSS &= \frac{1}{NG(NG-1)} \left\{ \frac{N(N-1)(N-2)}{2} \sum_{g=1}^G [\mu_{g1} + \frac{3}{N} \sum_{i=1}^N (\Psi_{(1)1}(\mathbf{X}_i^g) - \mu_{g1}) \right. \\
& + O_p(N^{-1})] \\
& + \frac{N(N-1)(N-2)(N-3)}{8} \sum_{g=1}^G \left[ \mu_{g2} + \frac{4}{N} \sum_{i=1}^N (\Psi_{(2)1}(\mathbf{X}_i^g) - \mu_{g2}) + O_p(N^{-1}) \right] \\
& + \frac{N^2(N-1)^2}{4} \sum_{1 \leq g_1 < g_2 \leq G} \left[ \mu_{(g_1, g_2)3} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(3)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)3}) \right. \\
& + \left. \frac{2}{N} \sum_{i'=1}^N (\Psi_{(3)01}(\mathbf{X}_{i'}^{g_2}) - \mu_{(g_1, g_2)3}) + O_p(N^{-1}) \right] \\
& + \frac{N^2(N-1)^2}{2} \sum_{1 \leq g_1 < g_2 \leq G} \left[ \mu_{(g_1, g_2)4} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(4)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)4}) \right. \\
& + \left. \frac{2}{N} \sum_{j=1}^N (\Psi_{(4)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)4}) + O_p(N^{-1}) \right] \\
& + N^2(N-1) \sum_{1 \leq g_1 < g_2 \leq G} \left[ \mu_{(g_1, g_2)5} + \frac{1}{N} \sum_{i=1}^N (\Psi_{(5,1)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)5}) \right. \\
& + \frac{2}{N} \sum_{j=1}^N (\Psi_{(5,1)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)5}) + \frac{2}{N} \sum_{i=1}^N (\Psi_{(5,2)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)5}) \\
& + \left. \frac{1}{N} \sum_{j=1}^N (\Psi_{(5,2)01}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2)5}) + O_p(N^{-1}) \right] \\
& + N^3 \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \left[ \mu_{(g_1, g_2, g_3)10} + \frac{1}{N} \sum_{i=1}^N (\Psi_{(10)100}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2, g_3)10}) \right. \\
& + \left. \frac{1}{N} \sum_{j=1}^N (\Psi_{(10)010}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2, g_3)10}) \right]
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{N} \sum_{j'=1}^N (\Psi_{(10)001}(\mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)10}) + O_p(N^{-1}) \Big] \\
& + N^3(N-1) \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \left[ \mu_{(g_1, g_2, g_3)9} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(9)100}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2, g_3)9}) \right. \\
& + \frac{1}{N} \sum_{j=1}^N (\Psi_{(9)010}(\mathbf{X}_j^{g_2}) - \mu_{(g_1, g_2, g_3)9}) \\
& + \left. \frac{1}{N} \sum_{j'=1}^N (\Psi_{(9)001}(\mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)9}) + O_p(N^{-1}) \right] \\
& + N^4 \sum_{\substack{1 \leq g_1, g_2, g_3, g_4 \leq G \\ g_1 \neq g_2 \neq g_3 \neq g_4}} \left[ \mu_{(g_1, g_2, g_3, g_4)11} + \frac{1}{N} \sum_{i=1}^N (\Psi_{(11)1000}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2, g_3, g_4)11}) \right. \\
& + \frac{1}{N} \sum_{i'=1}^N (\Psi_{(11)0100}(\mathbf{X}_{i'}^{g_2}) - \mu_{(g_1, g_2, g_3, g_4)11}) \\
& + \frac{1}{N} \sum_{j=1}^N (\Psi_{(11)0010}(\mathbf{X}_j^{g_3}) - \mu_{(g_1, g_2, g_3, g_4)11}) \\
& + \left. \frac{1}{N} \sum_{j'=1}^N (\Psi_{(11)0001}(\mathbf{X}_{j'}^{g_4}) - \mu_{(g_1, g_2, g_3, g_4)11}) + O_p(N^{-1}) \right] \\
& + N^2(N-1) \sum_{\substack{1 \leq g_1, g_2 \leq G \\ g_1 \neq g_2}} \left[ \mu_{(g_1, g_2)6} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(6)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)6}) \right. \\
& + \left. \frac{1}{N} \sum_{j'=1}^N (\Psi_{(6)01}(\mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)6}) + O_p(N^{-1}) \right] \\
& + \frac{N^2(N-1)(N-2)}{2} \sum_{\substack{1 \leq g_1, g_2 \leq G \\ g_1 \neq g_2}} \left[ \mu_{(g_1, g_2)7} + \frac{3}{N} \sum_{i=1}^N (\Psi_{(7)10}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2)7}) \right. \\
& + \left. \frac{1}{N} \sum_{j'=1}^N (\Psi_{(7)01}(\mathbf{X}_{j'}^{g_2}) - \mu_{(g_1, g_2)7}) + O_p(N^{-1}) \right] \\
& + \frac{N^3(N-1)}{2} \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \left[ \mu_{(g_1, g_2, g_3)8} + \frac{2}{N} \sum_{i=1}^N (\Psi_{(8)100}(\mathbf{X}_i^{g_1}) - \mu_{(g_1, g_2, g_3)8}) \right. \\
& + \frac{1}{N} \sum_{i'=1}^N (\Psi_{(8)010}(\mathbf{X}_{i'}^{g_2}) - \mu_{(g_1, g_2, g_3)8}) \\
& + \left. \frac{1}{N} \sum_{j'=1}^N (\Psi_{(8)001}(\mathbf{X}_{j'}^{g_3}) - \mu_{(g_1, g_2, g_3)8}) + O_p(N^{-1}) \right] \Big\}
\end{aligned}$$

Define the Total Mean Square ( $TMS$ ) as

$$TMS = \frac{TSS}{\binom{NG}{2}} = \frac{2TSS}{NG(NG-1)}$$

$$\begin{aligned} E_0(TMS) &= \frac{2}{NG(NG-1)} E_0(TSS) \\ &= \frac{(N-1)(N-2)(N-3)}{4NG(NG-1)^2} \mu_2 + \frac{(G-1)N(N-1)^2}{4NG(NG-1)^2} \mu_3 \\ &\quad + \frac{(G-1)N(N-1)^2}{2NG(NG-1)^2} \mu_4 + \frac{(G-1)(G-2)N^2(N-1)}{NG(NG-1)^2} \mu_9 \\ &\quad + \frac{(G-1)(G-2)(G-3)N^3}{4NG(NG-1)^2} \mu_{11} + \frac{(G-1)N(N-1)(N-2)}{NG(NG-1)^2} \mu_7 \\ &\quad + \frac{(G-1)(G-2)N^2(N-1)}{2NG(NG-1)^2} \mu_8 \end{aligned}$$

But, under  $H_0$ ,

$$\begin{aligned} \mu_2 &= \mu_3 = \mu_4 = \mu_7 = \mu_8 = \mu_9 = \mu_{11} \\ &= \frac{2}{K^2} \left\{ \sum_{k=1}^K \theta_k(1-\theta_k) + \sum_{k_1 \neq k_2} (\theta_{k_1 k_2} - \theta_{k_1} \theta_{k_2}) \right\} \end{aligned}$$

Therefore,

$$\begin{aligned} E_0(TMS) &= \frac{1}{NG(NG-1)^2} \mu_2 \left[ \frac{(N-1)(N-2)(N-3)}{4} \right. \\ &\quad + \frac{(G-1)N(N-1)^2}{4} + \frac{(G-1)N(N-1)^2}{2} \\ &\quad + (G-1)(G-2)N^2(N-1) + \frac{(G-1)(G-2)(G-3)N^3}{4} \\ &\quad \left. + (G-1)N(N-1)(N-2) + \frac{(G-1)(G-2)N^2(N-1)}{2} \right] \\ &= \frac{\mu_2}{4G^3} [1 + (G-1)(7 + (G-2)(G+3))] + O(N^{-1}) \end{aligned}$$

Now

$$\begin{aligned} BSS &= \frac{N(N-1)}{2} \sum_{g=1}^G (\bar{D}^g - \bar{D}.)^2 \\ &= \frac{N(N-1)}{2} \mathbf{D}'_1 \mathbf{D}_1 \end{aligned}$$

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

$$\mathbf{D}_1 = (\bar{D}^1 - \bar{D}., \dots, \bar{D}^G - \bar{D}.)'$$

Note that

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

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

and

$$\begin{aligned} E(\bar{D}) &= \frac{(N-1)}{G(NG-1)} \sum_{g=1}^G E(\bar{D}^g) + \frac{2N}{G(NG-1)} \sum_{1 \leq g_1 < g_2 \leq G} E(\bar{D}^{(g_1, g_2)}) \\ &= \frac{(N-1)}{G(NG-1)} \sum_{g=1}^G \bar{\theta}^g + \frac{2N}{G(NG-1)} \sum_{1 \leq g_1 < g_2 \leq G} \bar{\theta}^{(g_1, g_2)} \end{aligned}$$

Therefore,

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

Let  $\mathbf{U}_{12}^{(2)} = \bar{D}^g$  and  $\mathbf{U}_{13}^{(1,1)} = \bar{D}^{(g_1, g_2)}$ . Then,

$$\phi_{12}(\mathbf{X}_i^g, \mathbf{X}_j^g) = D_{ij}^g = \frac{1}{K} \sum_{k=1}^K I(X_{ik}^g \neq X_{jk}^g)$$

$$\phi_{13}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_2}) = D_{ij}^{(g_1, g_2)} = \frac{1}{K} \sum_{k=1}^K I(X_{ik}^{g_1} \neq X_{jk}^{g_2})$$

$$\Psi_{(12)1}(\mathbf{x}_i^g) = E[\phi_{12}(\mathbf{X}_i^g, \mathbf{X}_j^g) \mid \mathbf{X}_i^g] = \frac{1}{K} \sum_{k=1}^K P(X_{jk}^g \neq x_{ik}^g)$$

$$\Psi_{(13)10}(\mathbf{x}_i^{g_1}) = E[\phi_{13}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_2}) \mid \mathbf{X}_i^{g_1}] = \frac{1}{K} \sum_{k=1}^K P(X_{jk}^{g_2} \neq x_{ik}^{g_1})$$

$$\Psi_{(13)01}(\mathbf{x}_j^{g_2}) = E[\phi_{13}(\mathbf{X}_i^{g_1}, \mathbf{X}_j^{g_2}) \mid \mathbf{X}_j^{g_2}] = \frac{1}{K} \sum_{k=1}^K P(X_{ik}^{g_1} \neq x_{jk}^{g_2})$$

Under  $H_0$ ,

$$\begin{aligned} \Psi_{(12)1}(\mathbf{X}_i) &= \Psi_{(13)10}(\mathbf{X}_i) = \Psi_{(13)01}(\mathbf{X}_j) \\ &= \frac{1}{K} \sum_{k=1}^K P(X_{jk} \neq x_{ik}) \end{aligned} \tag{4.4.10}$$

since the sequences  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$  are i.i.d. and  $P(X_{jk} \neq x_{ik}) = P(X_{ik} \neq x_{jk})$ .

$$\psi_{(12)1}(\mathbf{x}_i) = \Psi_{(12)1}(\mathbf{x}_i) - \bar{\theta}.$$

and under  $H_0$

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

Example: Two categories

$$\begin{aligned} P(X_{jk} \neq x_{ik}) &= P(X_{jk} \neq 0)I(x_{ik} = 0) + P(X_{jk} \neq 1)I(x_{ik} = 1) \\ &= p_k(1)I(x_{ik} = 0) + (1 - p_k(1))I(x_{ik} = 1) \end{aligned}$$

$$\begin{aligned} &P(X_{jk_1} \neq x_{ik_1}; X_{jk_2} \neq x_{ik_2}) \\ &= P(X_{jk_1} \neq 0, x_{ik_2} \neq 0)I(x_{ik_1} = 0, x_{ik_2} = 0) \\ &\quad + P(X_{jk_1} \neq 0, x_{ik_2} \neq 1)I(x_{ik_1} = 0, x_{ik_2} = 1) \\ &\quad + P(X_{jk_1} \neq 1, x_{ik_2} \neq 0)I(x_{ik_1} = 1, x_{ik_2} = 0) \\ &\quad + P(X_{jk_1} \neq 1, x_{ik_2} \neq 1)I(x_{ik_1} = 1, x_{ik_2} = 1) \\ &= p_{k_1 k_2}(1, 1)I(x_{ik_1} = 0, x_{ik_2} = 0) + p_{k_1 k_2}(1, 0)I(x_{ik_1} = 0, x_{ik_2} = 1) \\ &\quad + p_{k_1 k_2}(0, 1)I(x_{ik_1} = 1, x_{ik_2} = 0) + p_{k_1 k_2}(0, 0)I(x_{ik_1} = 1, x_{ik_2} = 1) \end{aligned}$$

$$\begin{aligned} \psi_{(12)1}^2(\mathbf{x}_i) &= \frac{1}{K^2} \sum_{k=1}^K [(p_k(1))^2 I(X_{ik} = 0) + (1 - p_k(1))^2 I(X_{ik} = 1)] \\ &\quad + (\bar{\theta})^2 - \frac{2}{K} \bar{\theta} \cdot \sum_{k=1}^K [p_k(1)I(X_{ik} = 0) + (1 - p_k(1))I(X_{ik} = 1)] \\ &\quad + \frac{1}{K^2} \sum_{k_1 \neq k_2} [p_{k_1 k_2}(1, 1)I(X_{ik_1} = 0, X_{ik_2} = 0) \\ &\quad + p_{k_1 k_2}(1, 0)I(X_{ik_1} = 0, X_{ik_2} = 1) + p_{k_1 k_2}(0, 1)I(X_{ik_1} = 1, X_{ik_2} = 0) \\ &\quad + p_{k_1 k_2}(0, 0)I(X_{ik_1} = 1, X_{ik_2} = 1)] \end{aligned}$$

and

$$\xi_1^{(12)} = E[\psi_{(12)1}^2(\mathbf{X}_i)]$$

$$\begin{aligned}
&= \frac{1}{K^2} \sum_{k=1}^K \left[ (p_k(1))^2 p_k(0) + (1 - p_k(1))^2 p_k(1) \right] + (\bar{\theta})^2 \\
&\quad - \frac{2}{K} \bar{\theta} \sum_{k=1}^K [p_k(1) p_k(0) + (1 - p_k(1)) p_k(1)] \\
&\quad + \frac{1}{K^2} \sum_{k_1 \neq k_2} [p_{k_1 k_2}(1, 1) p_{k_1 k_2}(0, 0) + p_{k_1 k_2}(1, 0) p_{k_1 k_2}(0, 1) \\
&\quad + p_{k_1 k_2}(0, 1) p_{k_1 k_2}(1, 0) + p_{k_1 k_2}(0, 0) p_{k_1 k_2}(1, 1)] \\
&= \frac{1}{K^2} \sum_{k=1}^K \left[ (p_k(1))^2 p_k(0) + (1 - p_k(1))^2 p_k(1) \right] + (\bar{\theta})^2 \\
&\quad - \frac{2}{K} \bar{\theta} \sum_{k=1}^K [p_k(1) p_k(0) + (1 - p_k(1)) p_k(1)] \\
&\quad + \frac{2}{K^2} \sum_{k_1 \neq k_2} [p_{k_1 k_2}(1, 1) p_{k_1 k_2}(0, 0) + p_{k_1 k_2}(1, 0) p_{k_1 k_2}(0, 1)]
\end{aligned}$$

■

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

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

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

$$\text{Var}(\bar{D}^{(g_1, g_2)}) = \frac{1}{N} \xi_{10}^{(13)} + \frac{1}{N} \xi_{01}^{(13)} + O(N^{-2}) \quad (4.4.11)$$

where  $\xi_{10}^{(13)} \equiv \text{E}[\psi_{(13)10}^2(\mathbf{X}_i^{g_1})]$  and  $\xi_{01}^{(13)} \equiv \text{E}[\psi_{(13)01}^2(\mathbf{X}_j^{g_2})]$ .

We are assuming that under  $H_0$  there is homogeneity across or within groups, i.e.,  $\theta_k^1 = \theta_k^2 = \dots = \theta_k^G = \theta_k$  and  $\theta_k^{(g_1, g_2)} = \theta_k^g = \theta_k$ . Therefore, under  $H_0$ ,

$$\sqrt{N} (\bar{D}^g - \bar{\theta}) \xrightarrow{d} \text{N}(0, 4\xi_1^{(12)}) \quad (4.4.12)$$

and

$$\gamma_{13}^{-1} (\bar{D}^{(g_1, g_2)} - \bar{\theta}) \xrightarrow{d} \text{N}(0, 1) \quad (4.4.13)$$

where  $\gamma_{13}^2 = \frac{1}{N} \xi_{10}^{(13)} + \frac{1}{N} \xi_{01}^{(13)} = \frac{2}{N} \xi_1^{(12)}$  by (4.4.11) and (4.4.10).

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

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

$$\begin{aligned}
&= \frac{(N-1)}{G(NG-1)} \sum_{g=1}^G \left[ \bar{\theta}^g + \frac{2}{N} \sum_{i=1}^N (\Psi_{(12)1}(\mathbf{X}_i^g) - \bar{\theta}^g) \right] + O_p(N^{-1}) \\
&+ \frac{2N}{G(NG-1)} \sum_{1 \leq g_1 < g_2 \leq G} \left[ \bar{\theta}^{(g_1, g_2)} + \frac{1}{N} \sum_{i=1}^N (\Psi_{(13)10}(\mathbf{X}_i^{g_1}) - \bar{\theta}^{(g_1, g_2)}) \right. \\
&+ \left. \frac{1}{N} \sum_{i=1}^N (\Psi_{(13)01}(\mathbf{X}_i^{g_2}) - \bar{\theta}^{(g_1, g_2)}) \right] + O_p(N^{-1})
\end{aligned}$$

Under  $H_0$ ,

$$\eta_1 \equiv E_0(\bar{D}.) = \frac{(N-1)\bar{\theta} + N(G-1)\bar{\theta}}{(NG-1)} = \bar{\theta}.$$

$$\begin{aligned}
\sigma_1^2 &\equiv \text{Var}_0(\bar{D}.) \\
&= \frac{(N-1)^2}{G^2(NG-1)^2} \sum_{g=1}^G \text{Var}_0(\bar{D}^g) \\
&+ \frac{4N^2}{G^2(NG-1)^2} \\
&\times \left[ \sum_{1 \leq g_1 < g_2 \leq G} \text{Var}_0(\bar{D}^{(g_1, g_2)}) + 2 \sum_{\substack{1 \leq g_1, g_2, g_3 \leq G \\ g_1 \neq g_2 \neq g_3}} \text{Cov}_0(\bar{D}^{(g_1, g_2)}, \bar{D}^{(g_1, g_3)}) \right] \\
&+ 2 \frac{N(N-1)}{G^2(G-1)^2} \text{Cov}_0 \left( \sum_{g=1}^G \bar{D}^g, \sum_{1 \leq g_1 < g_2 \leq G} \bar{D}^{(g_1, g_2)} \right) \\
&= \frac{(N-1)^2}{G^2(NG-1)^2} G \left( \frac{4}{N} \xi_1^{(12)} \right) \\
&+ \frac{4N^2}{G^2(NG-1)^2} \left[ \frac{G(G-1)}{2} \left( \frac{1}{N} (\xi_{10}^{(13)} + \xi_{01}^{(13)}) \right) \right. \\
&+ \left. 2 \frac{G(G-1)(G-2)}{2} \left( \frac{1}{N} \xi_{10}^{(13,1;13,2)} \right) \right] \\
&+ \frac{2N(N-1)}{G^2(NG-1)^2} \sum_{g_1=1}^G \sum_{g_2 \neq g_1} \text{Cov}_0(\bar{D}^{g_1}, \bar{D}^{(g_1, g_2)}) \\
&= \frac{(N-1)^2}{G(NG-1)^2} \frac{4}{N} \xi_1^{(12)} + \frac{2N^2(G-1)}{G(NG-1)^2} \left[ \left( \frac{1}{N} (\xi_{10}^{(13)} + \xi_{01}^{(13)}) \right) \right. \\
&+ \left. 2(G-2) \frac{1}{N} \xi_{10}^{(13,1;13,2)} \right] + \frac{2N(N-1)}{G^2(NG-1)^2} G(G-1) \frac{2}{N} \xi_1^{(12,13)}
\end{aligned}$$

where  $\xi_{10}^{(13,1;13,2)} = E\{\psi_{(13,1)10}(\mathbf{X}_i^{g_1}) \psi_{(13,2)10}(\mathbf{X}_i^{g_1})\}$  and

$\psi_{(13,2)10}(\mathbf{x}_i^{g_1}) = E[\phi_{13,2}(\mathbf{x}_i^{g_1}, \mathbf{X}_j^{g_3}) - \bar{\theta}^{(g_1, g_3)}]$ . 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,



$$\xi_1^{(12)} = \xi_{10}^{(13)} = \xi_{01}^{(31)} = \xi_{10}^{(13,1;13,2)} = \xi_1^{(12,13)} \text{ and}$$

$$\begin{aligned} \sigma_1^2 &= \frac{(N-1)^2}{G(NG-1)^2} \frac{4}{N} \xi_1^{(12)} + \frac{2N^2(G-1)}{G(NG-1)^2} \left[ \left( \frac{2}{N} \xi_1^{(12)} \right) + 2(G-2) \frac{1}{N} \xi_1^{(12)} \right] \\ &\quad + \frac{2N(N-1)}{G^2(NG-1)^2} G(G-1) \frac{2}{N} \xi_1^{(12)} \\ &= [(N-1)^2 + N^2(G-1)^2 + N(N-1)(G-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\ &= [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \end{aligned} \quad (4.4.14)$$

Hence, under  $H_0$ ,

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

Now

$$\nu_1 = E_0(\bar{D}^g - \bar{D} \cdot) = \bar{\theta} \cdot - \bar{\theta} \cdot = 0 \quad (4.4.15)$$

and

$$\begin{aligned} \tau_1^2 &\equiv \text{Var}_0(\bar{D}^g - \bar{D} \cdot) \\ &= \text{Var}_0(\bar{D}^g) + \text{Var}(\bar{D} \cdot) - 2\text{Cov}_0(\bar{D}^g, \bar{D} \cdot) \\ &= \frac{4}{N} \xi_1^{(12)} + \sigma_1^2 - 2\text{Cov}_0 \left( \bar{D}^g, \frac{(N-1)}{G(NG-1)} \sum_{g=1}^G \bar{D}^g \right) \\ &\quad - 2\text{Cov}_0 \left( \bar{D}^g, \frac{2N}{G(NG-1)} \sum_{\substack{1 \leq g_1 < g_2 \leq G \\ g_1 \neq g_2}} \bar{D}^{(g_1, g_2)} \right) \\ &= \frac{4}{N} \xi_1^{(12)} + \sigma_1^2 - 2 \frac{(N-1)}{G(NG-1)} \text{Var}_0(\bar{D}^g) \\ &\quad - \frac{4N}{G(NG-1)} \sum_{\substack{g_2=1 \\ g_1 \neq g_2}}^G \text{Cov}_0(\bar{D}^{g_1}, \bar{D}^{(g_1, g_2)}) \\ &= \frac{4}{N} \xi_1^{(12)} + \sigma_1^2 - 2 \frac{(N-1)}{G(NG-1)} \frac{4}{N} \xi_1^{(12)} - \frac{4N}{G(NG-1)} \sum_{\substack{g_2=1 \\ g_1 \neq g_2}}^G \frac{2}{N} \xi_1^{(12,13)} \\ &= \left[ 1 - 2 \frac{(N-1)}{G(NG-1)} \right] \frac{4}{N} \xi_1^{(12)} + \sigma_1^2 - \frac{4N(G-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12,13)} \end{aligned} \quad (4.4.16)$$

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

Then,

$$\begin{aligned}
\tau_1^2 &= \left[ 1 - 2 \frac{(N-1 + N(G-1))}{G(NG-1)} \right] \frac{4}{N} \xi_1^{(12)} \\
&\quad + [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\
&= [(G-2)(NG-1)^2 + (N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\
&= \{(N-1)^2 + (NG-1)[N(G-1) + (NG-1)(G-2)]\} \frac{4\xi_1^{(12)}}{NG(NG-1)^2}
\end{aligned} \tag{4.4.17}$$

So,

$$\tau_1^{-1}(\bar{D}^g - \bar{D}.) \xrightarrow{d} N(0, 1)$$

Therefore, as in Chapter 3,

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

where  $\lambda_g$ 's are the characteristic roots of  $\text{Var}(\mathbf{D}_1) = \Sigma_1$ . Note that the diagonal elements of  $\Sigma_1$  are  $\tau_1^2$  and the off-diagonal elements are

$$\begin{aligned}
&\text{Cov}(\bar{D}^{g_1} - \bar{D}., \bar{D}^{g_2} - \bar{D}.) \\
&= \text{Cov}(\bar{D}^{g_1}, \bar{D}^{g_2}) - \text{Cov}(\bar{D}^{g_1}, \bar{D}.) - \text{Cov}(\bar{D}^{g_2}, \bar{D}.) + \text{Var}(\bar{D}.) \\
&= -2 \text{Cov}(\bar{D}^g, \bar{D}.) + \text{Var}(\bar{D}.) \\
&= -2 \frac{(N-1)}{G(NG-1)} \frac{4}{N} \xi_1^{(12)} - \frac{4N(G-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12,13)} + \sigma^2
\end{aligned}$$

which becomes under  $H_0$

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

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

Now,

$$E_0(BSS) = \frac{N(N-1)}{2} \text{trace}(\Sigma_1) = \frac{N(N-1)}{2} G\tau_1^2$$

and

$$\text{Var}_0(BSS) = \frac{N^2(N-1)^2}{4} \text{trace}(\Sigma_1)^2$$

Let

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

Then

$$E_0(BMS) = \frac{1}{G} E_0(BSS) = \tau_1^2$$

and

$$\text{Var}_0(BMS) = \frac{1}{G^2} \text{Var}_0(BSS) = \frac{1}{G^2} \text{trace}(\Sigma_1)^2$$

For *ABSS* we have,

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

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

Let

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

Under  $H_0$ ,

$$\nu_2 = E_0(\bar{D}^{(g_1, g_2)} - \bar{D}.) = \bar{\theta} - \bar{\theta} = 0 \quad (4.4.18)$$

and

$$\begin{aligned} \tau_2^2 &\equiv \text{Var}(\bar{D}^{(g_1, g_2)} - \bar{D}.) \\ &= \text{Var}(\bar{D}^{(g_1, g_2)}) + \text{Var}(\bar{D}.) - 2\text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}.) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N} \left( \xi_{10}^{(13)} + \xi_{01}^{(13)} \right) + \sigma_1^2 - 2\text{Cov} \left( \bar{D}^{(g_1, g_2)}, \frac{(N-1)}{G(NG-1)} \sum_{g=1}^G \bar{D}^g \right) \\
&\quad - 2\text{Cov} \left( \bar{D}^{(g_1, g_2)}, \frac{2N}{G(NG-1)} \sum_{1 \leq g_1 < g_2 \leq G} \bar{D}^{(g_1, g_2)} \right) \\
&= \frac{1}{N} \left( \xi_{10}^{(13)} + \xi_{01}^{(13)} \right) + \sigma_1^2 - \frac{2(N-1)}{G(NG-1)} \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}^{g_1}) \\
&\quad - \frac{2(N-1)}{G(NG-1)} \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}^{g_2}) \\
&\quad - \frac{4N}{G(NG-1)} \text{Cov}(\bar{D}^{(g_1, g_2)}, \sum_{1 \leq g_1 < g_2 \leq G} \bar{D}^{(g_1, g_2)}) \\
&= \frac{1}{N} \left( \xi_{10}^{(13)} + \xi_{01}^{(13)} \right) + \sigma_1^2 - \frac{2(N-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12,13)} \\
&\quad - \frac{2(N-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12,13)} \\
&\quad - \frac{4N}{G(NG-1)} \left[ \text{Var}(\bar{D}^{(g_1, g_2)}) + \sum_{\substack{g_3=1 \\ g_3 \neq g_1 \neq g_2}}^G \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}^{(g_1, g_3)}) \right. \\
&\quad \left. + \sum_{\substack{g_3=1 \\ g_3 \neq g_1 \neq g_2}}^G \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}^{(g_2, g_3)}) \right] \\
&= \frac{1}{N} \left( \xi_{10}^{(13)} + \xi_{01}^{(13)} \right) + \sigma_1^2 - \frac{4(N-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12,13)} \\
&\quad - \frac{4N}{G(NG-1)} \left[ \frac{1}{N} \left( \xi_{10}^{(13)} + \xi_{01}^{(13)} \right) + 2(G-2) \frac{1}{N} \xi_{10}^{(13,1;13,2)} \right] \tag{4.4.19}
\end{aligned}$$

Note that

$$\begin{aligned}
\Psi_{(13,1)10}(\mathbf{x}_i^{g_1}) &= \frac{1}{K} \sum_{k=1}^K \text{P}(X_{jk}^{g_2} \neq x_{ik}^{g_1}) \\
\Psi_{(13,2)10}(\mathbf{x}_i^{g_1}) &= \frac{1}{K} \sum_{k=1}^K \text{P}(X_{jk}^{g_3} \neq x_{ik}^{g_1}) \\
\Psi_{(13)10}(\mathbf{x}_i^{g_1}) &= \frac{1}{K} \sum_{k=1}^K \text{P}(X_{jk}^{g_2} \neq x_{ik}^{g_1}) \\
\Psi_{(13)01}(\mathbf{x}_j^{g_2}) &= \frac{1}{K} \sum_{k=1}^K \text{P}(X_{ik}^{g_1} \neq x_{jk}^{g_2})
\end{aligned}$$

and under  $H_0$  there is homogeneity among groups,

$$\Psi_{(13)10}(\mathbf{x}_i) = \Psi_{(13)01}(\mathbf{x}_j) = \Psi_{(13,1)10}(\mathbf{x}_i) = \Psi_{(13,2)10}(\mathbf{x}_i)$$

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  $H_0$ ,

$$\begin{aligned} \tau_2^2 &= \frac{2}{N}\xi_1^{(12)} + [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\ &\quad - \frac{4(N-1)}{G(NG-1)} \frac{2}{N}\xi_1^{(12)} - \frac{4N}{G(NG-1)}(G-1) \frac{2}{N}\xi_1^{(12)} \\ &= [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} + \frac{2(G-4)\xi_1^{(12)}}{NG} \\ &= \{2(N-1)^2 + (NG-1)[2N(G-1) + (NG-1)(G-4)]\} \frac{2\xi_1^{(12)}}{NG(NG-1)^2} \end{aligned} \quad (4.4.20)$$

As in *BSS*,

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

where  $\lambda_i$ 's are the characteristic roots of  $\Sigma_2 = \text{Var}(\mathbf{D}_2)$ . The diagonal elements of  $\Sigma_2$  are  $\tau_2^2$  and, if all groups are different, the off-diagonal elements are

$$\begin{aligned} &\text{Cov}(\bar{D}^{(g_1, g_2)} - \bar{D}, \bar{D}^{(g_3, g_4)} - \bar{D}) \\ &= \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}^{(g_3, g_4)}) - \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}) \\ &\quad - \text{Cov}(\bar{D}^{(g_3, g_4)}, \bar{D}) + \text{Var}(\bar{D}) \\ &= -2\text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}) + \text{Var}(\bar{D}) \\ &= \frac{-4(N-1)}{G(NG-1)} \frac{2}{N}\xi_1^{(12,13)} - \frac{4N}{G(NG-1)} \left[ \frac{1}{N}(\xi_{10}^{(13)} + \xi_{01}^{(13)}) \right. \\ &\quad \left. + 2(G-2) \frac{1}{N}\xi_{10}^{(13,1;13,2)} \right] + \sigma_1^2 \\ &= \frac{-4(N-1)}{G(NG-1)} \frac{2}{N}\xi_1^{(12)} - \frac{4N}{G(NG-1)}(G-1) \frac{2}{N}\xi_1^{(12)} \\ &\quad + [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\ &= -\frac{8\xi_1^{(12)}}{NG} + \left[ \frac{(N-1)^2 + N(G-1)(NG-1)}{(NG-1)} \right] \frac{4\xi_1^{(12)}}{NG(NG-1)} \\ &= [(N-1)^2 - (NG-1)(NG+N-2)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} < 0 \end{aligned}$$

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

$$\begin{aligned}
& \text{Cov}(\bar{D}^{(g_1, g_2)} - \bar{D}., \bar{D}^{(g_1, g_3)} - \bar{D}.) \\
&= \text{Cov}(\bar{D}^{(g_1, g_2)} - \bar{D}., \bar{D}^{(g_2, g_3)} - \bar{D}.) \\
&= \text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}^{(g_1, g_3)}) - 2\text{Cov}(\bar{D}^{(g_1, g_2)}, \bar{D}.) + \text{Var}(\bar{D}.) \\
&= \frac{1}{N} \xi_{10}^{(13,1;13,2)} - \frac{4(N-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12,13)} - \frac{4N}{G(NG-1)} \left[ \frac{1}{N} (\xi_{10}^{(13)} + \xi_{01}^{(13)}) \right. \\
&\quad \left. + 2(G-2) \frac{1}{N} \xi_{10}^{(13,1;13,2)} \right] + \sigma_1^2 \\
&= \frac{1}{N} \xi_1^{(12)} - \frac{4(N-1)}{G(NG-1)} \frac{2}{N} \xi_1^{(12)} - \frac{4N}{G(NG-1)} (G-1) \frac{2}{N} \xi_1^{(12)} \\
&\quad + [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\
&= (G-8) \frac{\xi_1^{(12)}}{NG} \\
&\quad + [(N-1)^2 + N(G-1)(NG-1)] \frac{4\xi_1^{(12)}}{NG(NG-1)^2} \\
&= \{4(N-1)^2 + (NG-1)[4N(G-1) + (G-8)(NG-1)]\} \frac{\xi_1^{(12)}}{NG(NG-1)^2} .
\end{aligned}$$

Now

$$E_0(ABSS) = N^2 \text{trace}(\Sigma_2) = N^2 \frac{G(G-1)}{2} \tau_2^2$$

$$\text{Var}_0(ABSS) = N^4 \text{trace}(\Sigma_2)^2$$

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

$$E_0(ABMS) = \frac{2}{G(G-1)} \text{trace}(\Sigma_2) = \tau_2^2$$

$$\text{Var}_0(ABMS) = \frac{4}{G^2(G-1)^2} \text{trace}(\Sigma_2)^2$$

## 4.5 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  $N \rightarrow \infty$ , i.e, asymptotically the distribution of  $\frac{WMS}{AWMS}$  is degenerate.

Let  $\Sigma_1 = \frac{1}{N}\Sigma_1^*$  and  $\Sigma_2 = \frac{1}{N}\Sigma_2^*$ . Under  $H_0$ ,

$$BSS \sim \frac{(N-1)}{2} \sum_{g=1}^G \lambda_{1g}^* (\chi_1^2)_g$$

where  $\lambda_{1g}^*$ 's are the characteristic roots of  $\Sigma_1^*$ .

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

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

where  $\lambda_{2i}^*$ 's are the characteristic roots of  $\Sigma_2^*$ .

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

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

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

and

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

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

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

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

$$\begin{aligned} T_{N,2} &= N \left[ \frac{BMS}{WMS - \mu_2/2 + \mu_2/2} \right] = N \left( \frac{BMS}{\mu_2/2} \right) \left[ 1 + \frac{(WMS - \mu_2/2)}{\mu_2/2} \right]^{-1} \\ &= \frac{2N(BMS)}{\mu_2} + O_p(N^{-1/2}) \end{aligned}$$

and

$$\begin{aligned} T_{N,3} &= N \left[ \frac{ABMS}{AWMS - \mu_2/2 + \mu_2/2} \right] = N \left( \frac{ABMS}{\mu_2/2} \right) \left[ 1 + \frac{(AWMS - \mu_2/2)}{\mu_2/2} \right]^{-1} \\ &= \frac{2N(ABMS)}{\mu_2} + O_p(N^{-1/2}) \end{aligned}$$

Therefore,

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

and

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

Because the elements of  $\Sigma_1^*$  and  $\Sigma_2^*$  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.

### 4.5.1 Power of the Tests

#### Lemma 4.1

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,

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

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,

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

Let  $\mathbf{V}_N = \mathbf{D}_1 - \boldsymbol{\nu}_1$ , where  $\boldsymbol{\nu}_1$  is a vector  $G \times 1$  with elements  $\nu_1$ . Then,  $E(\mathbf{V}_N) = \mathbf{0}$  and  $\text{Var}(\mathbf{V}_N) = \boldsymbol{\Sigma}_1 = \frac{1}{N}\boldsymbol{\Sigma}_1^* = O(N^{-1})$ . 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  $\sqrt{N}\mathbf{V}_N \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1^*)$ ,

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

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

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

Now,

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

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

Note that

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

and

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

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

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

Thus,

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

i.e., this test is consistent.

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

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

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

Note that

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

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

$$T_{N,2} = \frac{2N}{G\mu_2} \mathbf{V}'_N \mathbf{V}_N + \frac{4\sqrt{N}}{G\mu_2} (\gamma_1^*)' \mathbf{V}_N + \frac{2(\gamma_1^*)^2}{\mu_2} + O_p(N^{-1/2})$$

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

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

Recall that

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

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,

$$(\boldsymbol{\gamma}_1^*)' \boldsymbol{\Sigma}_1 = \gamma_1^* [\tau_1^2 + (G-1)\tau_{12} \cdots \tau_1^2 + (G-1)\tau_{12}]$$

and

$$\begin{aligned} \tau_1^2 + (G-1)\tau_{12} &= 0 \\ \Leftrightarrow G(N-1)^2 + (NG-1)[N(G-1) + (NG-1)(G-2)] \\ &\quad - (G-1)(NG+N-2) = 0 \\ \Leftrightarrow G(N-1)^2 + (NG-1)[(NG-1)(G-2) - (NG-2)(G-1)] &= 0 \\ \Leftrightarrow G(N-1)^2 + (NG-1)^2(G-2) - (NG-1)(NG-2)(G-1) &= 0 \\ \Leftrightarrow N^2G + 2NG - 2NG^2 - 2N^2G^2 + 3NG^2 + N^2G^2 - 3NG &= 0 \\ \Leftrightarrow N-1 + G - NG = 0 \Leftrightarrow N(1-G) + G - 1 = 0 \\ \Leftrightarrow N = 1 \end{aligned}$$

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

Now, write

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

and

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

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

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

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  $\mathbf{P}\boldsymbol{\Sigma}_1^*\mathbf{P}' = \boldsymbol{\Lambda}$ , 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,

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

Hence,

$$N\mathbf{D}'_1\mathbf{D}_1 = \mathbf{Y}'\mathbf{Y} \sim \sum_{i=1}^G \lambda_i (\chi_1^2(\delta_i)) \quad (4.5.1)$$

where  $\delta_i = \frac{(\nu_{1i}^*)^2}{\lambda_i}$ ,  $\lambda_i$ 's are the diagonal elements of the diagonal matrix  $\boldsymbol{\Lambda}$  and  $\nu_{1i}^*$  is the  $i$ th row of the vector  $\boldsymbol{\nu}_1^* = \mathbf{P}\boldsymbol{\gamma}_1^*$ . By (4.5.1),

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

Since we have a linear combination of non-central chi-square random variables, when  $\nu_1 = \frac{\boldsymbol{\gamma}_1^*}{\sqrt{N}}$ ,

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

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  $T_{N,3}$ .

# Chapter 5

## Numerical Studies and Data Analysis

### 5.1 Modelling the Mutation Process

As an example, we consider a set of HIV-1 sequences from LaRosa et al. (1990) and Myers et al. (1992) which span the V3 loop of the envelope gene. The data consist of 87 sequences with 35 amino acids each. They are all from the B clade (North America, Western Europe, Brazil and Thailand). Within the V3 loop are found determinants for T-cell-adaptation and macrophage tropism (the wild-type phenotype). These strains (T-cell adapted and macrophage tropic) are subject to different selection pressures and therefore linkage patterns within the V3 loop are most likely to differ as well.

#### 5.1.1 The Autologistic Model

Recall from Chapter 2 that  $y(i) = 0$  or  $1$ ,

$$\Pr(0(i) \mid \{y(j) : j \in N_i\}) = \frac{1}{1 + \exp\left(\alpha_i + \sum_{j \in N_i} \gamma_{ij} y(j)\right)}$$

and

$$P(y(i) | \{y(j) : j \in N_i\}) = \frac{\exp\left(\alpha_i y(i) + \sum_{j \in N_i} \gamma_{ij} y(i) y(j)\right)}{1 + \exp\left(\alpha_i + \sum_{j \in N_i} \gamma_{ij} y(j)\right)} \quad (5.1.1)$$

where  $N_i = \{j : j \text{ is a neighbor of } i\}$  is the neighborhood of site  $i$ .

Since the data set is of moderate size and this model involves too many parameters, we reduce the number of parameters by imposing some restrictions.

### Model 1

- The neighborhood is defined as the immediate positions, i.e, the neighborhood of site  $i$  is  $i - 1$  and  $i + 1$ . At the two extremities of the sequences, the neighborhood is only the next one for the left extremity and only the previous one for the right extremity, i.e., the neighborhood of site 1 is site 2 and the neighborhood of site  $n$  is  $n - 1$ .

- $\alpha_i = \alpha, \forall i$
- $\gamma_{ij} = \gamma, \forall i, j$ .

The model is then,

$$P(y(i) | \{y(j) : 0 < |i - j| \leq 1\}) = \frac{\exp\left(\alpha y(i) + \sum_{j \in N_i} \gamma y(i) y(j)\right)}{1 + \exp\left(\alpha + \sum_{j \in N_i} \gamma y(j)\right)} \quad (5.1.2)$$

The maximum pseudo-likelihood estimates (MPLE) and the simulated maximum likelihood estimates (SMLE) are shown in Table 5.2. We used the Metropolis algorithm to obtain the MCMC maximum-likelihood estimates: 1000 samplers were generated with different initial values, discarding an initial warm-up of 500 iterations in each.

To verify if we reached convergence after a burn-in of 500, we considered the first 100 samplers and compared at each of the 33 positions the empirical probabilities with the probabilities under model (5.1.2).

For each position we computed the statistic  $Z$

$$Z_i = \frac{\hat{p}_i - p_i}{\sqrt{p_i(1-p_i)/m}}, \quad i = 1, \dots, 33$$

where,

$$p_i = P(y(i) | \{y(j) : 0 < |i-j| \leq 1\}) = \frac{\exp\left(\tilde{\alpha} y(i) + \sum_{j \in N_i} \tilde{\gamma} y(i) y(j)\right)}{1 + \exp\left(\tilde{\alpha} + \sum_{j \in N_i} \tilde{\gamma} y(j)\right)},$$

$\tilde{\alpha}$  and  $\tilde{\gamma}$  are the MPLE of  $\alpha$  and  $\gamma$ , respectively, and  $\hat{p}_i$  is the empirical distribution computed from the first 100 samplers.

For each position we compute 4 probabilities:

$$P(y(i) = 0 | y(i-1) = 0, y(i+1) = 0) \quad , \quad P(y(i) = 0 | y(i-1) = 0, y(i+1) = 1)$$

$$P(y(i) = 0 | y(i-1) = 1, y(i+1) = 0) \quad \text{and} \quad P(y(i) = 0 | y(i-1) = 1, y(i+1) = 1)$$

For the empirical distribution we have,

$$\hat{p}_{i1} = \frac{(\#y(i) = 0, y(i-1) = 0, y(i+1) = 0)}{(\#y(i-1) = 0, y(i+1) = 0)},$$

$$\hat{p}_{i2} = \frac{(\#y(i) = 0, y(i-1) = 0, y(i+1) = 1)}{(\#y(i-1) = 0, y(i+1) = 1)},$$

$$\hat{p}_{i3} = \frac{(\#y(i) = 0, y(i-1) = 1, y(i+1) = 0)}{(\#y(i-1) = 1, y(i+1) = 0)} \quad \text{and}$$

$$\hat{p}_{i4} = \frac{(\#y(i) = 0, y(i-1) = 1, y(i+1) = 1)}{(\#y(i-1) = 1, y(i+1) = 1)}$$

We only looked at 33 positions, because the first and last positions remain constant. So, we then have  $4 \times 33$  statistics to compute and the results in Table 5.1 show that only 6 out of 132 (4%) of those statistics have a significant result. Hence, if we adopt a level of significance of 5%, a burn-in of 500 is satisfactory.

For this model,

$$\log \left\{ \frac{P(y(i) = 1 | y(i-1) = 0, y(i+1) = 0)}{1 - P(y(i) = 1 | y(i-1) = 0, y(i+1) = 0)} \right\} = \alpha$$

Hence,  $\alpha$  is the log-odds of mutation at a certain site when there is no mutation in its neighborhood. If  $\alpha$  is negative, the probability of mutation at site  $i$  is less than

Table 5.1: Convergence Results

Statistic	Positions											
	2	3	4	5	6	7	8	9	10	11	12	
$Z_{i1}$	1.41	0.06	0.88	0.30	1.89	-1.44	0.05	0.98	-0.42	-1.15	1.65	
$Z_{i2}$	-0.36	-0.87	1.62	-1.29	0.82	-0.38	-1.65	1.25	0.59	-1.03	-0.27	
$Z_{i3}$	0.65	-0.83	-0.62	-0.35	1.37	-1.19	0.03	-0.63	0.08	0.78	-0.82	
$Z_{i4}$	-0.80	-1.47	-0.55	2.62*	0.54	-0.88	0.07	-0.50	1.34	-0.78	0.77	
Statistic	13	14	15	16	17	18	19	20	21	22	23	
$Z_{i1}$	0.65	-0.27	0.71	0.14	2.45*	-0.89	0.83	-0.83	1.06	-0.43	0.55	
$Z_{i2}$	-0.89	0.44	-0.10	-1.59	1.03	-0.25	0.12	0.12	-0.32	0.09	-1.54	
$Z_{i3}$	0.56	-1.77	0.61	-0.50	0.06	0.47	-1.13	0.83	0.47	-0.80	-0.33	
$Z_{i4}$	-0.95	-0.67	0.04	0.26	0.54	-2.34*	0.96	-0.64	-0.40	-1.57	-0.40	
Statistic	24	25	26	27	28	29	30	31	32	33	34	
$Z_{i1}$	0.92	-2.01*	-0.55	1.01	-0.17	0.89	0.10	-1.55	-0.17	-2.26*	-0.69	
$Z_{i2}$	1.52	-0.22	-0.57	1.32	0.09	-0.10	0.99	0.79	0.82	-1.72	0.88	
$Z_{i3}$	-0.74	0.21	0.47	0.37	0.07	0.85	0.20	0.27	-1.17	1.27	0.02	
$Z_{i4}$	-0.28	-0.76	-1.19	-0.24	-0.25	0.27	-1.02	-1.20	3.91*	1.22	-1.43	

\*  $|Z_i| > 1.96$ .

Table 5.2: Model 1

	MPLE	SMLE
$\alpha$	-1.7117	-1.7117
$\gamma$	0.7813	0.8204



the probability of no mutation at that site, given no mutation at its neighborhood. Since we are assuming that  $\alpha_i = \alpha$ , the probability of no mutation is 5.53 ( $\exp(1.71)$ ) times higher than the probability of mutation given that there is no mutation in the neighborhood. When there is mutation in the neighborhood of site  $i$ , the log odds of mutation is a function of both  $\alpha$  and  $\gamma$ . So, when there is mutation in the neighborhood of site  $i$ , if  $\gamma$  is large and positive the log-odds of mutation at this site increases, and if  $\gamma$  is negative the log-odds of mutation at this site decreases.

## Model 2

- The neighborhood is defined as in model 1.
- $\alpha_i = \alpha, \forall i$
- $\gamma_{ij} = \gamma \rho^{|i-j|}$ .

The model is then,

$$P(y(i) | \{y(j) : 0 < |i - j| \leq 1\}) = \frac{\exp(\alpha y(i) + \sum_{j=1}^n \gamma \rho^{|i-j|} y(i) y(j))}{1 + \exp(\alpha + \sum_{j=1}^n \gamma \rho^{|i-j|} y(j))} \quad (5.1.3)$$

Table 5.3: Model 2

	MPLE	SMLE
$\alpha$	-0.0582	-0.0582
$\gamma$	1.5715	1.5715
$\rho$	0.4679	0.4679

There is no difference between the pseudolikelihood estimates and the simulated maximum-likelihood estimates, indicating that the dependency among neighboring positions is weak. Again we are assuming that  $\alpha_i = \alpha$  and the probability of no mutation is 1.05 ( $\exp(0.06)$ ) times higher than the probability of mutation given that there is no mutation in the neighborhood. When there is mutation in the neighborhood of site  $i$ , the log odds of mutation is a function of  $\alpha$ ,  $\gamma$ ,  $\rho$  and the distance between sites  $i$  and  $j$  (since  $\rho$  has power  $|i - j|$ ). So, when there is mutation in the neighborhood of site  $i$ , the log-odds of mutation at this site increases, since  $\gamma$  and  $\rho$  are positive.

### 5.1.2 Model based on the Bahadur Representation

From Section 2.3,  $\mathbf{y}_k = (y_k(1), \dots, y_k(n))$  is a  $n \times 1$  vector representing the binary responses for the  $n$  sites along sequence  $k$ , i.e, whether there is a mutation from the consensus or not at each site along the sequence. Here, we consider  $k = 87$  sequences and  $n = 35$  sites.

#### Probabilistic model

$$\begin{aligned} \mathbf{P}(\mathbf{y}_k) &= \prod_{i=1}^n \xi_i^{y_k(i)} (1 - \xi_i)^{1-y_k(i)} \\ &\times \left[ 1 + \sum_{i < j} r_{ij} \left( \frac{y_k(i) - \xi_i}{\sqrt{\xi_i(1 - \xi_i)}} \right) \left( \frac{y_k(j) - \xi_j}{\sqrt{\xi_j(1 - \xi_j)}} \right) \right] \end{aligned} \quad (5.1.4)$$

The number of parameters to be estimated in the above model is too big ( $35 + \binom{35}{2}$ ) compared with the number of observations (87) in the data set. Therefore, we cannot work with this general model. To get reliable estimates we need to reduce the parameter space. First, let  $r_{ij} = 0$  if  $|i - j| \geq 1$  and  $\xi_i = \xi, \forall i$ . Then, for illustration purposes, let us discard all positions with less than 20% of mutation and still consider  $r_{ij} = 0$  if  $|i - j| \geq 1$ . In this case we have 12 positions and 23 (12+11) parameters. The 12 positions we are now dealing with are: 9, 10, 11, 13, 14, 19, 20, 22, 23, 24, 25 and 32. The results are shown in Tables 5.4 and 5.5.

Table 5.4: Bahadur Representation Model (35 positions)

Parameter	$\xi$	$r_{1,2}$	$r_{2,3}$	$r_{3,4}$	$r_{4,5}$	$r_{5,6}$	$r_{6,7}$	$r_{7,8}$	$r_{8,9}$
MLE	0.24	0.57	0.47	0.75	0.48	0.65	0.32	0.40	-0.02
Parameter	$r_{9,10}$	$r_{10,11}$	$r_{11,12}$	$r_{12,13}$	$r_{13,14}$	$r_{14,15}$	$r_{15,16}$	$r_{16,17}$	$r_{17,18}$
MLE	0.11	0.14	-0.11	-0.05	0.25	0.01	0.53	0.70	0.46
Parameter	$r_{18,19}$	$r_{19,20}$	$r_{20,21}$	$r_{21,22}$	$r_{22,23}$	$r_{23,24}$	$r_{24,25}$	$r_{25,26}$	$r_{26,27}$
MLE	-0.07	0.36	-0.21	0.13	0.11	0.31	0.26	-0.20	0.22
Parameter	$r_{27,28}$	$r_{28,29}$	$r_{29,30}$	$r_{30,31}$	$r_{31,32}$	$r_{32,33}$	$r_{33,34}$	$r_{34,35}$	
MLE	-0.01	0.49	0.00	0.49	-0.11	-0.26	0.42	0.47	

Comparing the results of Tables 5.4 and 5.5, we see that the assumption of equal probability of mutation for all positions (model 1) is not tenable. For instance,

Table 5.5: Bahadur Representation Model (12 positions)

Parameter	$\xi_9$	$\xi_{10}$	$\xi_{11}$	$\xi_{13}$	$\xi_{14}$	$\xi_{19}$	$\xi_{20}$	$\xi_{22}$	$\xi_{23}$
MLE	0.30	0.35	0.53	0.73	0.18	0.27	0.55	0.40	0.26
Parameter	$\xi_{24}$	$\xi_{25}$	$\xi_{32}$	$r_{9,10}$	$r_{10,11}$	$r_{11,13}$	$r_{13,14}$	$r_{14,19}$	$r_{19,20}$
MLE	0.58	0.73	0.30	-0.15	0.16	-0.16	0.77	0.49	0.30
Parameter	$r_{20,22}$	$r_{22,23}$	$r_{23,24}$	$r_{24,25}$	$r_{25,32}$				
MLE	-0.22	0.09	0.31	-0.27	0.14				

in Table 5.5, the estimated probability of mutation at position 14 is 0.18 while at position 13 it is 0.73, despite the fact that  $r_{13,14}$  is 0.77, indicating high correlation between these adjacent sites. A negative estimate of  $r_{ij}$  means that the mutation rate at position  $i$  is lower than at position  $j$ .

## 5.2 Analyzing the Variability in DNA Sequences

### 5.2.1 Simulations

In order to look at the behavior of the asymptotic distributions, we generated  $N$  sequences, with  $K$  positions each, in  $G$  groups and computed the test statistic  $F_1^*$  (3.5.29) and the standardized  $F_1^*$  (i.e.,  $K \left( F_1^* - \frac{a_1}{a_2^*} \right) / \sigma_*$ ). We performed 500 simulations (i.e., the above procedure was repeated 500 times). In Table 5.6 the data were generated using the same probability across all positions, i.e.,  $p_{ck} = p_c$ , while for Table 5.7, different probabilities were used across positions. The tables show the number below and above some quantiles of the standard normal distribution.

The number of groups seems to make a big difference on the asymptotic results. When we increase the number of groups to 10, the results are much better (see the lines for  $N = 100, K = 10, G = 10$  and  $N = 200, K = 10, G = 10$ ). The number of sequences also plays an important role, but we need to keep in mind that the ratio of the number of sequences ( $N$ ) to the number of positions ( $K$ ) should not be small (ideally it should be at least 5). Therefore, for nucleotide sequences we need

Table 5.6: Results of Simulations for Diversity Measures ( $p_{ck} = p_c$ )

$N$	$K$	$G$	Percentiles of the Std. Normal Dist.								
			1	2.5	5	10	90	95	97.5	99	
20	10	2	0*	0**	0**	0**	52	36*	27**	16**	
20	10	5	0*	1**	2**	20**	48	30	21*	16**	
20	50	2	0*	0**	0**	0**	59	49**	38**	27**	
50	10	2	0*	0**	0**	0**	67*	43**	28**	19**	
50	10	5	0*	0**	4**	33*	51	34	23**	13**	
50	50	2	0*	0**	0**	0**	59	41**	32**	27**	
100	10	2	0*	0**	0**	0**	52	38*	30**	24**	
100	10	5	0*	1**	5**	26**	63	38*	26**	11*	
100	10	10	0*	4*	11*	40	42	30	14	10*	
100	50	2	0*	0**	0**	0**	49	36*	21*	16**	
100	50	5	0*	0**	3**	23**	56	32	18	12**	
200	10	5	0*	0**	9**	30*	48	32	24**	18**	
200	10	10	1	3*	14*	40	52	31	17	7	

\* between 2SD and 3SD or  $-2SD$  and  $-3SD$ .

\*\* greater than 3SD or smaller than  $-3SD$ .

Table 5.7: Results of Simulations for Diversity Measures ( $p_{ck} \neq p_c$ )

$N$	$K$	$G$	Percentiles of the Std. Normal Dist.								
			1	2.5	5	10	90	95	97.5	99	
20	10	2	0*	0**	0**	0**	59	36*	23**	16**	
20	10	5	0*	0**	4**	24**	54	37*	22*	13**	
20	50	2	0*	0**	0**	0**	63	39*	29**	20**	
50	10	2	0*	0**	0**	0**	60	36*	28**	19**	
50	10	5	0*	1**	6**	43	51	33	17	10	
50	50	2	0*	0**	0**	0**	53	37*	24**	14**	
100	10	2	0*	0**	0**	0**	53	42**	31**	21**	
100	10	5	0*	1**	10**	27**	41	28	19	12**	
100	10	10	1	3	13*	49	53	30	18	7	
100	50	2	0*	0**	0**	0**	46	28	22*	17**	
100	50	5	0*	1**	9**	28**	57	35*	22*	10*	
200	10	5	0*	1**	8**	34*	54	30	18	12**	
200	10	10	1	8	17	40	54	30	17	9	

\* between 2SD and 3SD or  $-2SD$  and  $-3SD$ .

\*\* greater than 3SD or smaller than  $-3SD$ .

at least five times more sequences than the number of positions in order to apply the asymptotic results. The scenario where the probabilities are different across positions yields equally good results.

## 5.2.2 Data Analysis

The data set consists of two groups (subtype B and not B) with 46 sequences each. The nucleotide sequences are all from different individuals and span the protease region. There are therefore four categories. After aligning the sequences and discarding the positions with no change, we end up with 155 positions.

Looking at the simulation results we see that our data set is not large enough for the asymptotic results to apply. We therefore rely on resampling techniques, such as the bootstrap. Here is a summary of the procedure:

1. Estimate  $p_{ck}$  from the data, i.e.,  $\hat{p}_{ck} = \frac{n_{c1k} + n_{c2k}}{2N}$  and compute the statistic  $F_1$ .
2. Generate  $N = 46$  sequences, with  $K = 155$  positions each, in each of the  $G = 2$  groups, using  $\hat{p}_{ck}$ .
3. Recompute the test statistic  $F_1$  from the generated data and store it.
4. Repeat steps 2 and 3 1,000 times.

The p-value is then  $\frac{\#F_1's \geq F_1^{obs}}{1000}$ .

The results are

$$WSI = 0.7005 \quad TSI = 0.7007 \quad BSI = 0.0002 \quad \text{and} \quad F_1^{obs} = 0.011$$

The percentiles of the bootstrap distribution are given in Table 5.8 and the observed p-value is less than 1/1001. This means that relative to the within-clade variation, there is significant variability between the two clades.

Table 5.8: Percentiles of the Bootstrap Dist. for Diversity Measures

90%	95%	97.5%	99%	99.5%	99.9%
0.0015	0.0020	0.0023	0.0027	0.0030	0.0040

# Chapter 6

## Conclusion and Future Research

### 6.1 Concluding Summary

The autologistic model formulated in Section 2.1 is able to handle situations involving spatial binary data and dependence on covariates. One problem is that the estimation procedure is not straightforward and computer-intensive MCMC procedures are needed. Also, the general model formulation involves too many parameters and when applied to data sets of moderate size, we need to reduce the number of parameters by imposing restrictions. For the data set we use, the number of sequences is small compared to the number of positions, and we end up with a very simple model because of the restrictions we impose on the parameter space. For instance, we assume equal mutation rate and equal correlation structure over all positions, which may not be true in reality.

The model based on the Bahadur representation (Section 2.2) can also handle dependent binary data, but the inclusion of dependent covariates is not as easy as in the autologistic model. An advantage of this model is that the likelihood function has a closed form and the parameter estimates are obtained by the maximum-likelihood approach using numerical optimization methods, such as the Newton-Raphson procedure. Again, we have to reduce the parameter space when applying this model to our data set, because the sample size (the number of independent sequences) is small compared to the number of positions. Ideally, the ratio between the number



of sequences and the number of positions should be at least 5. When we discard all positions with frequency of mutation less than 20%, we see that the mutation rate varies a lot from one position to the other, confirming the fact that the assumption of equal mutation rate over all positions does not hold. For the extension of the autologistic model to three categories (Section 2.3) we also need large data sets, since the general formulation of this model involves even more parameters than the one for a binary response.

When using the diversity measures (Chapter 3) to analyze the variability in DNA sequences, we assume independence among positions and we only consider sequences from independent individuals. The power of the test developed in Chapter 3 is evaluated and we conclude that the test is consistent, i.e., as the sample size increases, the power of the test goes to 1. Since we know that positions in DNA sequences may not be independent, more work needs to be done relaxing this assumption and this is discussed in Section 6.2.4. The simulations (Section 5.2.1) show that the asymptotic test statistic (3.5.21), developed in Chapter 3, behaves better when the number of groups is large (at least 10) and the number of sequences ( $N$ ) is large compared to the number of positions ( $K$ ). The ratio between  $N$  and  $K$  should be at least 10. Also, when the data are generated using different probabilities across positions the asymptotic results are as good as those obtained when generated using equal probabilities. For small sample sizes we can use resampling techniques, such as the bootstrap, to generate the reference distribution and see where the observed test statistic falls. The data set used for illustration consists of two groups (subtype B and not B) with 46 sequences in each. Sequences are originated from different individuals and the sequences span the protease region. After aligning the sequences, we discard the positions with no change and we end up with 155 positions. The results show that relative to the within-clade variation, there is significant variability between the two clades.

The analysis of variance based on Hamming distances (Chapter 4) has the advantage of considering the sequences on an individual basis, since we make all pairwise comparisons within and across groups. The sequences are assumed to be independent in each group, but we should also think about developing tests when the sequences

cannot be considered as independent. For instance, if the interest is in comparisons of sequences within and between individuals, the sequences within the same individual cannot be considered as independent. This is discussed in Section 6.2.6. We decompose the total sum of squares into within-, between- and across-group sums of squares, with the latter term being new: it does not appear in the usual decomposition. The assumption of independence among positions is relaxed and test statistics are developed based on U-statistics theory. We found that these tests are consistent and their power goes to 1 as the sample size increases. The distributions of these test statistics do not have a closed analytical form, and therefore, we need to call upon resampling techniques, such as the bootstrap, to perform the test.

## 6.2 Future Research

### 6.2.1 An Application of the Autologistic Model to sequences from the *nef* gene

In the application of the autologistic model of Section 2.1 no covariates are considered. Now, we would like to include covariates and define neighborhoods in a biologically meaningful manner. In particular, the three-dimensional molecular structure of the *nef* gene is known and it is possible to define neighborhoods according to molecular distances. Having the spatial coordinates for each sequence position, we can compute genetic distances. At the amino-acid level, possible covariates are size (e.g., small or large) and group (hydrophobic or hydrophylic). The goal is to estimate the probability of mutation taking into account the neighborhood of the site and the most significant covariates. A mutation occurs when there is a change in amino acid with respect to the consensus sequence.

## 6.2.2 Extension of the Autologistic Model to more than Two Categories

The theoretic aspects of an extension of the autologistic model for three categories are developed in Section 2.3, but no application is shown. Now, we will extend this model for any  $C$  ( $C > 2$ ) categories and apply it to some data. Markov-chain Monte-Carlo procedures will estimate the model parameters.

## 6.2.3 An Alternative Null Hypothesis for the Analysis of Diversity Measures

The null hypothesis of interest in Chapter 3 is that there is homogeneity among the groups, i.e., the category probability at a certain position is the same over all groups. Now, we would like to test a less restrictive hypothesis. Recall from Chapter 3 that the population variation within the  $g$ th group at the  $k$ th position is

$$I_S(\mathbf{p}_{gk}) = 1 - \sum_{c=1}^C p_{cgk}^2 \quad (6.2.1)$$

If the null hypothesis of interest is just

$$H_0 : I_S(\mathbf{p}_{1k}) = I_S(\mathbf{p}_{2k}) = \cdots = I_S(\mathbf{p}_{Gk}) ,$$

i.e., the within-group variation at the  $k$ th position is the same over all groups, it implies that

$$\| \mathbf{p}_{1k} \| = \| \mathbf{p}_{2k} \| = \cdots = \| \mathbf{p}_{Gk} \| \quad (6.2.2)$$

where  $\mathbf{p}_{gk} = (p_{1gk} \ p_{2gk} \ \dots \ p_{Cgk})'$  is a  $C \times 1$  vector representing the probabilities of belonging to categories  $c = 1, \dots, C$  in group  $g$  and position  $k$ . Note that if this hypothesis is true, the hypothesis of Chapter 3 ( $H_0 : p_{cgk} = p_{ck}$ ) is not necessarily true.

The distribution of the test statistic under this less restrictive null hypothesis need to be derived and we need to be aware of the structure of the covariance matrix in this situation. The latter is now a little more complicated since we cannot assume that individual probabilities are the same over all groups.

## 6.2.4 Diversity Measures for Sequences with Dependent Positions

The test statistic developed on Chapter 3 assumes independence among positions along the sequences. A simple scenario for dependence among positions is depicted by a first-order Markov chain.

First, assume that there are only 2 categories, i.e.,  $C = 2$ . For a first-order Markov chain

$$\begin{aligned} \Pr\{X_{iK}^g = x_{iK}^g \mid X_{i1}^g = x_{i1}^g, X_{i2}^g = x_{i2}^g, \dots, X_{iK-1}^g = x_{iK-1}^g\} \\ = \Pr\{X_{iK}^g = x_{iK}^g \mid X_{iK-1}^g = x_{iK-1}^g\} \end{aligned}$$

Let

$$\begin{aligned} p_k^g(2 \mid 1) \\ = \Pr\{\text{being in category 2 at position } k \text{ for group } g \mid \text{at position } k-1 \text{ it is at category 1}\} \end{aligned}$$

Then,

$$p_k^g(1 \mid 1) = 1 - p_k^g(2 \mid 1) \text{ and } p_k^g(1 \mid 2) = 1 - p_k^g(2 \mid 2)$$

Let  $n_{gk} \equiv n_{gk}(1)$  denote the number of responses in category 1 at position  $k$  for group  $g$ ,  $N$  the total number of responses at position  $k$ ,  $N - n_{gk} \equiv n_{gk}(2)$  the number of responses in category 2 at position  $k$  for group  $g$ ,  $n_k^g(a \mid b)$  the number of responses in category  $a$  at position  $k$  for group  $g$  given that at position  $k-1$  it is in category  $b$ , with  $a, b = 1, 2$ .

The contingency table for two categories is shown in Table 6.1 and for  $C$  categories in Table 6.2.

For each group's responses  $(n_{g1}, n_{g2}, \dots, n_{gK})$ ,

$$\begin{aligned} \Pr\{(n_{g1}, n_2^g(1 \mid 1), n_2^g(1 \mid 2), \dots, n_K^g(1 \mid 1), n_K^g(1 \mid 2))\} \\ = \binom{N}{n_{g1}} p_1^g(1)^{n_{g1}} (1 - p_1^g(1))^{(N - n_{g1})} \\ \times \binom{n_{g1}}{n_2^g(1 \mid 1)} p_2^g(1 \mid 1)^{n_2^g(1 \mid 1)} (1 - p_2^g(1 \mid 1))^{(n_{g1} - n_2^g(1 \mid 1))} \end{aligned}$$

Table 6.1: Contingency Table for Two Categories (dependent positions)

		<u>Position <math>k</math></u>			
		Category			
		1	2	Total	
Posit. $k - 1$	Category	1	$n_k^g(1   1)$	$n_k^g(2   1)$	$n_{g,k-1}(1)$
		2	$n_k^g(1   2)$	$n_k^g(2   2)$	$n_{g,k-1}(2)$
Total		$n_{g,k}(1)$	$n_{g,k}(2)$	$n_{\cdot g} = N$	

$$\begin{aligned}
 & \times \binom{N - n_{g1}}{n_2^g(1 | 2)} p_2^g(1 | 2)^{n_2^g(1|2)} (1 - p_2^g(1 | 2))^{(N - n_{g1} - n_2^g(1|2))} \\
 & \dots \times \binom{n_{gK-1}}{n_K^g(1 | 1)} p_K^g(1 | 1)^{n_K^g(1|1)} (1 - p_K^g(1 | 1))^{(n_{gK-1} - n_K^g(1|1))} \\
 & \times \binom{N - n_{gK-1}}{n_K^g(1 | 2)} p_K^g(1 | 2)^{n_K^g(1|2)} (1 - p_K^g(1 | 2))^{(N - n_{gK-1} - n_K^g(1|2))}
 \end{aligned}$$

Table 6.2: Contingency Table for  $C$  categories (dependent positions)

		<u>Position <math>k</math></u>					
		Category					
		1	2	...	C	Total	
Position $k - 1$	Categ.	1	$n_k^g(1   1)$	$n_k^g(2   1)$	...	$n_k^g(C   1)$	$n_{gk-1}(1)$
		2	$n_k^g(1   2)$	$n_k^g(2   2)$	...	$n_k^g(C   2)$	$n_{gk-1}(2)$
		$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
		C	$n_k^g(1   C)$	$n_k^g(2   C)$	...	$n_k^g(C   C)$	$n_{gk-1}(C)$
Total		$n_{gk}(1)$	$n_{gk}(2)$	...	$n_{gk}(C)$	$n_{\cdot g} = N$	

Since the groups are independent, we can take the product over all groups. The probabilistic model is

$$\begin{aligned}
 & \prod_{g=1}^G \Pr\{(n_{g1}, n_2^g(1 | 1), n_2^g(1 | 2), \dots, n_K^g(1 | 1), n_K^g(1 | 2))\} \\
 & = \prod_{g=1}^G \binom{N}{n_{g1}} p_1^g(1)^{n_{g1}} (1 - p_1^g(1))^{(N - n_{g1})}
 \end{aligned}$$

$$\begin{aligned}
& \times \binom{n_{g1}}{n_2^g(1|1)} p_2^g(1|1)^{n_2^g(1|1)} (1 - p_2^g(1|1))^{(n_{g1} - n_2^g(1|1))} \\
& \times \binom{N - n_{g1}}{n_2^g(1|2)} p_2^g(1|2)^{n_2^g(1|2)} (1 - p_2^g(1|2))^{(N - n_{g1} - n_2^g(1|2))} \\
& \dots \times \binom{n_{gK-1}}{n_K^g(1|1)} p_K^g(1|1)^{n_K^g(1|1)} (1 - p_K^g(1|1))^{(n_{gK-1} - n_K^g(1|1))} \\
& \times \binom{N - n_{g,K-1}}{n_K^g(1|2)} p_K^g(1|2)^{n_K^g(1|2)} (1 - p_K^g(1|2))^{(N - n_{gK-1} - n_K^g(1|2))}
\end{aligned}$$

Let

$\mathbf{V}_1 \equiv (n_{11}(1) \ n_{11}(2) \ \dots \ n_{G1}(1) \ n_{G1}(K))'$  be a  $2G \times 1$  vector,

$\mathbf{V}_k \equiv (n_k^1(1|1) \ n_k^1(2|1) \ n_k^1(1|2) \ n_k^1(2|2) \ \dots \ n_k^G(1|2) \ n_k^G(2|2))'$ , be a  $4G \times 1$  vector for  $k = 2, \dots, K$  and  $\mathbf{V} = (\mathbf{V}_1 \ \mathbf{V}_2 \ \dots \ \mathbf{V}_K)'$  be a  $2G(2K - 1) \times 1$  vector.

$$\boldsymbol{\mu}_1 \equiv \mathbf{E}(\mathbf{V}_1) = (Np_1^1(1) \ Np_1^1(2) \ \dots \ Np_1^G(1) \ Np_1^G(2))' \quad (6.2.3)$$

$$\boldsymbol{\mu}_k \equiv \mathbf{E}(\mathbf{V}_k) = (n_{1k-1}(1)p_k^1(1|1) \ \dots \ n_{1k-1}(2)p_k^1(2|2) \ \dots \ n_{Gk-1}(2)p_k^G(2|2))' \quad (6.2.4)$$

for  $k = 2, \dots, K$

$$\boldsymbol{\mu} \equiv \mathbf{E}(\mathbf{V}) = (\boldsymbol{\mu}_1 \ \boldsymbol{\mu}_2 \ \dots \ \boldsymbol{\mu}_K)' \quad (6.2.5)$$

$\boldsymbol{\Sigma} \equiv \text{Cov}(\mathbf{V})$

$$= (\boldsymbol{\Sigma}_{11} \oplus \boldsymbol{\Sigma}_{21} \oplus \dots \oplus \boldsymbol{\Sigma}_{G1} \oplus \boldsymbol{\Sigma}_{12|1} \oplus \boldsymbol{\Sigma}_{12|2} \oplus \dots \oplus \boldsymbol{\Sigma}_{G2|1} \oplus \boldsymbol{\Sigma}_{G2|2} \oplus \dots \oplus \boldsymbol{\Sigma}_{GK|1} \oplus \boldsymbol{\Sigma}_{GK|2})' \quad (6.2.6)$$

where  $\boldsymbol{\Sigma}_{g1}$  is a  $2 \times 2$  matrix

$$\boldsymbol{\Sigma}_{g1} = N \begin{pmatrix} p_1^g(1)(1 - p_1^g(1)) & -p_1^g(1)p_2^g(1) \\ -p_1^g(1)p_2^g(1) & p_2^g(1)(1 - p_2^g(1)) \end{pmatrix}$$

and  $\boldsymbol{\Sigma}_{gk|c}$ , for  $c = 1, 2$ , is a  $2 \times 2$  matrix

$$\boldsymbol{\Sigma}_{gk|c} = n_{gk-1}(c) \begin{pmatrix} p_k^g(1|c)(1 - p_k^g(1|c)) & -p_k^g(1|c)p_k^g(2|c) \\ -p_k^g(1|c)p_k^g(2|c) & p_k^g(2|c)(1 - p_k^g(2|c)) \end{pmatrix}$$

Therefore,  $\boldsymbol{\Sigma}$  is a square matrix of dimension  $2G(2K - 1)$ .

Note that for  $C$  categories  $\mathbf{V}$  is a  $CG(C(K - 1) + 1) \times 1$  vector. The probabilistic model for  $C$  categories is

$$\prod_{g=1}^G \Pr\{(n_{g1}(1), \dots, n_{g1}(C), \dots, n_K^g(1|C), \dots, n_K^g(C|C))\}$$

$$\begin{aligned}
&= \prod_{g=1}^G \binom{N}{n_{g1}(1) \dots n_{g1}(C)} \prod_{c=1}^C [p_1^g(c)]^{n_{g1}(c)} \\
&\quad \times \prod_{k=2}^K \prod_j^C \binom{n_{gk-1}(j)}{n_k^g(1|j) \dots n_k^g(C|j)} \prod_{c=1}^C [p_k^g(c|j)]^{n_k^g(c|j)}
\end{aligned}$$

So, we need to find the distribution of the sums of squares under this model and construct a test statistic.

More general dependency will be considered: e.g.,  $k$ th order Markov chains and autologistic models.

### 6.2.5 Inclusion of Covariates in the Analysis of Diversity Measures

If we want to include covariates in the set up of the analysis of diversity measures and these covariates are categorical, we can treat them as additional factors in the analysis of variance. If the covariates are continuous, the appropriate model for this situation would be a log-linear model, and it does not seem to be possible to extend this model for the case of non-independent positions. However, most of the time the variables of interest are categorical. Even in the cases of variables like viral load or CD4 count, they are usually categorized.

### 6.2.6 Analysis of Variance based on the Hamming Distance when Sequences are not Independent

The theory of generalized U-statistics for independent random vectors (i.e, sequences) was used to get the distribution of sums of squares and develop test statistics in Chapter 4. In the case of dependence among sequences, the theory of U-statistics for non-independent random vectors may be called upon to find the distribution of sums of squares.

### **6.2.7 Tests Based on Contrasts in the Analysis of Variance for the Hamming Distances**

If the test of homogeneity among groups turns out significant, we would like to identify the most heterogeneous groups. So, we need to construct tests to compare all possible pairs of groups and to order them according to their degree of variability.



# Appendix A

$$\left( \sum_{k=1}^K c_k \right) \left( \sum_{k_1 \neq k_2} a_{k_1} b_{k_2} \right) = \sum_{k_1 \neq k_2 \neq k_3} c_{k_1} a_{k_2} b_{k_3} + \sum_{k_1 \neq k_2} c_{k_1} a_{k_1} b_{k_2} + \sum_{k_1 \neq k_2} c_{k_1} b_{k_1} a_{k_2} \quad (\text{A.1})$$

$$\left( \sum_{k=1}^K c_k \right) \left( \sum_{k_1 \neq k_2} a_{k_1} a_{k_2} \right) = \sum_{k_1 \neq k_2 \neq k_3} c_{k_1} a_{k_2} a_{k_3} + 2 \sum_{k_1 \neq k_2} c_{k_1} a_{k_1} a_{k_2} \quad (\text{A.2})$$

$$\left( \sum_{k=1}^K a_k b_k \right) \left( \sum_{k_1 < k_2} c_{k_1} c_{k_2} \right) = \sum_{k_1 \neq k_2} a_{k_1} b_{k_1} c_{k_1} c_{k_2} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} a_{k_1} b_{k_1} c_{k_2} c_{k_3} \quad (\text{A.3})$$

$$\left( \sum_{k=1}^K a_k \right) \left( \sum_{k=1}^K b_k \right) = \sum_{k=1}^K a_k b_k + \sum_{k_1 \neq k_2} a_{k_1} b_{k_2} \quad (\text{A.4})$$

$$\left( \sum_{k=1}^K a_k \right) \left( \sum_{k=1}^K b_k c_k \right) = \sum_{k=1}^K a_k b_k c_k + \sum_{k_1 \neq k_2} a_{k_1} b_{k_2} c_{k_2} \quad (\text{A.5})$$

$$\left( \sum_{k=1}^K a_k \right)^2 = \sum_{k=1}^K a_k^2 + \sum_{k_1 \neq k_2} a_{k_1} a_{k_2} \quad (\text{A.6})$$

$$\left( \sum_{k=1}^K a_k \right) \left( \sum_{k_1 < k_2} b_{k_1} c_{k_2} \right) = \sum_{k_1 < k_2} a_{k_1} b_{k_1} c_{k_2} + \sum_{k_2 < k_1} a_{k_1} b_{k_2} c_{k_1} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} a_{k_1} b_{k_2} c_{k_3} \quad (\text{A.7})$$

$$\begin{aligned} \left( \sum_{k_1 \neq k_2} a_{k_1} b_{k_2} \right)^2 &= \sum_{k_1 \neq k_2} (a_{k_1} b_{k_2})^2 + 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} a_{k_1} b_{k_2} a_{k_1} b_{k_3} \\ &+ 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} a_{k_2} b_{k_1} a_{k_3} b_{k_1} + 2 \sum_{k_1 < k_2} a_{k_1} b_{k_2} a_{k_2} b_{k_1} \\ &+ 2 \sum_{k_1 \neq k_2 \neq k_3} a_{k_1} b_{k_2} a_{k_3} b_{k_1} + \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} a_{k_1} b_{k_2} a_{k_3} b_{k_4} \end{aligned} \quad (\text{A.8})$$

$$\begin{aligned}
\left( \sum_{k_1 \neq k_2} a_{k_1} b_{k_2} \right) \left( \sum_{k_1 < k_2} c_{k_1} c_{k_2} \right) &= \sum_{k_1 \neq k_2} a_{k_1} b_{k_2} c_{k_1} c_{k_2} + \sum_{k_1 \neq k_2 \neq k_3} a_{k_1} b_{k_2} c_{k_1} c_{k_3} \\
&+ \sum_{k_1 \neq k_2 \neq k_3} a_{k_1} b_{k_2} c_{k_2} c_{k_3} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_3 < k_4}} a_{k_1} b_{k_2} c_{k_3} c_{k_4}
\end{aligned} \tag{A.9}$$

$$\begin{aligned}
\left( \sum_{k_1 < k_2} a_{k_1} b_{k_2} \right) \left( \sum_{k_1 < k_2} c_{k_1} d_{k_2} \right) &= \sum_{k_1 < k_2} a_{k_1} b_{k_2} c_{k_1} d_{k_2} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2, k_1 < k_3}} a_{k_1} b_{k_2} c_{k_1} d_{k_3} \\
&+ \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2, k_3 < k_1}} a_{k_1} b_{k_2} c_{k_3} d_{k_1} + \sum_{k_1 < k_2 < k_3} a_{k_1} b_{k_2} c_{k_2} d_{k_3} \\
&+ \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_3, k_2 < k_3}} a_{k_1} b_{k_3} c_{k_2} d_{k_3} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_1 < k_2, k_3 < k_4}} a_{k_1} b_{k_2} c_{k_3} d_{k_4}
\end{aligned} \tag{A.10}$$

$$\begin{aligned}
\left( \sum_{k_1 < k_2} a_{k_1} a_{k_2} \right) \left( \sum_{k_1 < k_2} c_{k_1} c_{k_2} \right) &= \sum_{k_1 < k_2} a_{k_1} a_{k_2} c_{k_1} c_{k_2} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2, k_1 < k_3}} a_{k_1} a_{k_2} c_{k_1} c_{k_3} \\
&+ \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_2, k_3 < k_1}} a_{k_1} a_{k_2} c_{k_3} c_{k_1} + \sum_{k_1 < k_2 < k_3} a_{k_1} a_{k_2} c_{k_2} c_{k_3} \\
&+ \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_1 < k_3, k_2 < k_3}} a_{k_1} a_{k_3} c_{k_2} c_{k_3} + \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_1 < k_2, k_3 < k_4}} a_{k_1} a_{k_2} c_{k_3} c_{k_4}
\end{aligned} \tag{A.11}$$

$$\begin{aligned}
\left( \sum_{k_1 < k_2} a_{k_1} b_{k_2} \right)^2 &= \sum_{k_1 < k_2} (a_{k_1} b_{k_2})^2 + 2 \sum_{k_1 < k_2 < k_3} a_{k_1} b_{k_2} a_{k_1} b_{k_3} + 2 \sum_{k_1 < k_2 < k_3} a_{k_1} b_{k_3} a_{k_2} b_{k_3} \\
&+ \sum_{k_1 < k_2 < k_3} a_{k_1} b_{k_2} a_{k_2} b_{k_3} + 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_1 < k_3, k_2 < k_4}} a_{k_1} b_{k_3} a_{k_2} b_{k_4}
\end{aligned} \tag{A.12}$$

# Appendix B

## Example: Two categories

Let, for any  $g$ ,

$$\begin{aligned} p_k^g(0) &= P(X_{jk}^g = 0) \quad , \quad p_k^g(1) = P(X_{jk}^g = 1) \\ p_{k_1 k_2}^g(1, 1) &= P(X_{jk_1}^g = 1, X_{jk_2}^g = 1) \quad , \quad p_{k_1 k_2}^g(0, 1) = P(X_{jk_1}^g = 0, X_{jk_2}^g = 1) \\ p_{k_1 k_2}^g(1, 0) &= P(X_{jk_1}^g = 1, X_{jk_2}^g = 0) \quad , \quad p_{k_1 k_2}^g(0, 0) = P(X_{jk_1}^g = 0, X_{jk_2}^g = 0). \end{aligned}$$

Note that  $p_k^g(0) + p_k^g(1) = 1$  and  $p_{k_1 k_2}^g(1, 1) + p_{k_1 k_2}^g(0, 1) + p_{k_1 k_2}^g(1, 0) + p_{k_1 k_2}^g(0, 0) = 1$

Also, let

$$\theta_k^g = P(X_{ik}^g \neq X_{jk}^g) = \sum_{u=0}^1 p_k^g(u)[1 - p_k^g(u)] \quad (\text{B.1})$$

$$\theta_{k_1 k_2}^g = P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{jk_2}^g) = \sum_{u, v=0}^1 p_{k_1 k_2}^g(u, v)p_{k_1 k_2}^g(1 - u, 1 - v) \quad (\text{B.2})$$

$$P(X_{ik}^g \neq X_{jk}^g, X_{ik}^g \neq X_{j'k}^g) = \sum_{u=0}^1 p_k^g(u)[1 - p_k^g(u)]^2 \quad (\text{B.3})$$

$$P(X_{ik_1}^g \neq X_{jk_1}^g, X_{ik_2}^g \neq X_{j'k_2}^g) = \sum_{u, v=0}^1 p_{k_1 k_2}^g(u, v)[1 - p_{k_1}^g(u)][1 - p_{k_2}^g(v)] \quad (\text{B.4})$$

$$P(X_{jk}^g \neq x_{ik}^g) = \sum_{u=0}^1 [1 - p_k^g(u)]I(x_{ik}^g = u) \quad (\text{B.5})$$

$$(P(X_{jk}^g \neq x_{ik}^g))^2 = \sum_{u=0}^1 [1 - p_k^g(u)]^2 I(x_{ik}^g = u) \quad (\text{B.6})$$

$$P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) = \sum_{u=0}^1 p_k^g(u)[1 - p_k^g(u)]I(x_{ik}^g = u) \quad (\text{B.7})$$

$$P(X_{jk}^g \neq x_{ik}^g)P(X_{jk}^g \neq X_{j'k}^g, X_{jk}^g \neq x_{ik}^g) = \sum_{u=0}^1 [1 - p_k^g(u)]^2 p_k^g(u)I(x_{ik}^g = u) \quad (\text{B.8})$$

$$\begin{aligned} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\ = \sum_{u,v=0}^1 [1 - p_{k_1}^g(u)][1 - p_{k_2}^g(v)]I(x_{ik_1}^g = u, x_{ik_2}^g = v) \end{aligned} \quad (\text{B.9})$$

$$\begin{aligned} (P(X_{jk_1}^g \neq x_{ik_1}^g))^2 P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\ = (P(X_{jk_1}^g \neq x_{ik_1}^g))^2 P(X_{jk_2}^g \neq x_{ik_2}^g, X_{j'k_1}^g \neq x_{ik_1}^g) \\ = \sum_{u,v=0}^1 [1 - p_{k_1}^g(u)]^3 [1 - p_{k_2}^g(v)]I(x_{ik_1}^g = u, x_{ik_2}^g = v) \end{aligned} \quad (\text{B.10})$$

$$\begin{aligned} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_2}^g \neq x_{ik_2}^g, X_{j'k_3}^g \neq x_{ik_3}^g) \\ = \sum_{u,v,z=0}^1 p_{k_1}^g(u)[1 - p_{k_1}^g(u)][1 - p_{k_2}^g(v)][1 - p_{k_3}^g(z)]I(x_{ik_1}^g = u, x_{ik_2}^g = v, x_{ik_3}^g = z) \end{aligned} \quad (\text{B.11})$$

$$\begin{aligned} P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\ = P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_2}^g \neq x_{ik_2}^g, X_{j'k_1}^g \neq x_{ik_1}^g) \\ = \sum_{u,v=0}^1 p_{k_1}^g(u)[1 - p_{k_1}^g(u)]^2 [1 - p_{k_2}^g(v)]I(x_{ik_1}^g = u, x_{ik_2}^g = v) \end{aligned} \quad (\text{B.12})$$

$$\begin{aligned} P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g)P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_3}^g \neq x_{ik_3}^g) \\ = \sum_{u,v,z=0}^1 [1 - p_{k_1}^g(u)]^2 [1 - p_{k_2}^g(v)][1 - p_{k_3}^g(z)]I(x_{ik_1}^g = u, x_{ik_2}^g = v, x_{ik_3}^g = z) \end{aligned} \quad (\text{B.13})$$

$$P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) = \sum_{u,v=0}^1 p_{k_1 k_2}^g(u, v)I(x_{ik_1}^g = 1 - u, x_{ik_2}^g = 1 - v) \quad (\text{B.14})$$

$$P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) = \sum_{u,v=0}^1 p_{k_1}^g(1 - u)p_{k_1 k_2}^g(u, v)I(x_{ik_2}^g = 1 - v) \quad (\text{B.15})$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g)P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v=0}^1 p_{k_1 k_2}^g(1-u, 1-v)[1-p_{k_1}^g(u)][1-p_{k_2}^g(v)]I(x_{ik_1}^g = u, x_{ik_2}^g = v) \quad (\text{B.16})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_3}^g \neq x_{ik_3}^g)P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1 k_3}^g(1-u, 1-v)[1-p_{k_1}^g(u)][1-p_{k_2}^g(z)]I(x_{ik_1}^g = u, x_{ik_2}^g = z, x_{ik_3}^g = v) \\
& \hspace{15em} (\text{B.17})
\end{aligned}$$

$$\begin{aligned}
& (P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g))^2 = \sum_{u,v=0}^1 [p_{k_1}^g(1-u)]^2 [p_{k_1 k_2}^g(u, v)]^2 I(x_{ik_2}^g = 1-v) \\
& + \sum_{u,v=0}^1 p_{k_1}^g(1-u)p_{k_1}^g(u)p_{k_1 k_2}^g(u, v)p_{k_1 k_2}^g(1-u, v)I(x_{ik_2}^g = 1-v) \quad (\text{B.18})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g)P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v,z=0}^1 [1-p_{k_1}^g(u)][1-p_{k_2}^g(v)]p_{k_1}^g(1-z)p_{k_1 k_2}(z, 1-v)I(x_{ik_1}^g = u, x_{ik_2}^g = v) \\
& \hspace{15em} (\text{B.19})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{j'k_3}^g \neq x_{ik_3}^g)P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v,z,l=0}^1 [1-p_{k_1}^g(u)][1-p_{k_3}^g(v)]p_{k_1}^g(1-z)p_{k_1 k_2}(z, 1-l)I(x_{ik_1}^g = u, x_{ik_2}^g = l, x_{ik_3}^g = v) \\
& \hspace{15em} (\text{B.20})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_2}^g \neq x_{ik_2}^g, X_{j'k_3}^g \neq x_{ik_3}^g)P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v,z=0}^1 [1-p_{k_2}^g(u)][1-p_{k_3}^g(v)]p_{k_1}^g(1-z)p_{k_1 k_2}(z, 1-u)I(x_{ik_2}^g = u, x_{ik_3}^g = v) \\
& \hspace{15em} (\text{B.21})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_3}^g \neq x_{ik_3}^g, X_{j'k_4}^g \neq x_{ik_4}^g)P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v,z,l=0}^1 [1-p_{k_3}^g(u)][1-p_{k_4}^g(v)]p_{k_1}^g(1-z)p_{k_1 k_2}(z, 1-l)I(x_{ik_2}^g = l, x_{ik_3}^g = u, x_{ik_4}^g = v) \\
& \hspace{15em} (\text{B.22})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_2}^g \neq X_{j'k_2}^g, X_{jk_3}^g \neq x_{ik_3}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1}^g(u)[1-p_{k_1}^g(u)]p_{k_2}^g(1-z)p_{k_2k_3}^g(z,v)I(x_{ik_1}^g=u, x_{ik_3}^g=1-v) \quad (B.23)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1}^g(u)[1-p_{k_1}^g(u)]p_{k_1}^g(1-z)p_{k_1k_2}^g(z,v)I(x_{ik_1}^g=u, x_{ik_2}^g=1-v) \quad (B.24)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_2}^g \neq X_{j'k_2}^g, X_{jk_1}^g \neq x_{ik_1}^g) \\
&= \sum_{u,v=0}^1 p_{k_1}^g(u)[1-p_{k_1}^g(u)]p_{k_2}^g(1-v)p_{k_1k_2}^g(1-u,v)I(x_{ik_1}^g=u) \quad (B.25)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_2}^g \neq x_{ik_2}^g)P(X_{jk_1}^g \neq X_{j'k_1}^g, X_{jk_3}^g \neq x_{ik_3}^g) \\
&= \sum_{u,v,z,l=0}^1 p_{k_1}^g(1-u)p_{k_1k_2}^g(u,v)p_{k_1}^g(1-z)p_{k_1k_3}^g(z,l)I(x_{ik_2}^g=1-v, x_{ik_3}^g=1-l) \quad (B.26)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_2}^g \neq X_{j'k_2}^g, X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_3}^g \neq X_{j'k_3}^g, X_{jk_1}^g \neq x_{ik_1}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_2}^g(1-u)p_{k_1k_2}^g(v,u)p_{k_3}^g(1-z)p_{k_1k_3}^g(v,z)I(x_{ik_1}^g=1-v) \quad (B.27)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g)P(X_{j'k_2}^g \neq x_{ik_2}^g) = \sum_{u,v=0}^1 [1-p_{k_1}^g(u)][1-p_{k_2}^g(v)]I(x_{ik_1}^g=u, x_{ik_2}^g=v) \quad (B.28)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g)P(X_{j'k_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v=0}^1 [1-p_{k_1}^g(1-u)]p_{k_1k_2}^g(u,v)I(x_{ik_1}^g=1-u, x_{ik_2}^g=1-v) \quad (B.29)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g)P(X_{jk_2}^g \neq x_{ik_2}^g, X_{j'k_3}^g \neq x_{ik_3}^g) \\
&= \sum_{u,v,z=0}^1 [1-p_{k_1}^g(u)][1-p_{k_2}^g(v)][1-p_{k_3}^g(z)]I(x_{ik_1}^g=u, x_{ik_2}^g=v, x_{ik_3}^g=z) \quad (B.30)
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g)P(X_{j'k_2}^g \neq x_{ik_2}^g, X_{j'k_3}^g \neq x_{ik_3}^g) \\
&= \sum_{u,v,z=0}^1 [1 - p_{k_1}^g(u)]p_{k_2k_3}^g(v, z)I(x_{ik_1}^g = u, x_{ik_2}^g = 1 - v, x_{ik_3}^g = 1 - z) \quad (\text{B.31})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g)P(X_{j'k_1}^g \neq x_{ik_1}^g, X_{j'k_3}^g \neq x_{ik_3}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1k_2}^g(u, v)p_{k_1k_3}^g(u, z)I(x_{ik_1}^g = 1 - u, x_{ik_2}^g = 1 - v, x_{ik_3}^g = 1 - z) \quad (\text{B.32})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g)P(X_{j'k_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v=0}^1 [p_{k_1k_2}^g(u, v)]^2 I(x_{ik_1}^g = 1 - u, x_{ik_2}^g = 1 - v) \quad (\text{B.33})
\end{aligned}$$

$$\begin{aligned}
& P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g)P(X_{j'k_3}^g \neq x_{ik_3}^g, X_{j'k_4}^g \neq x_{ik_4}^g) \\
&= \sum_{u,v,z,l=0}^1 p_{k_1k_2}^g(u, v)p_{k_3k_4}^g(z, l)I(x_{ik_1}^g = 1 - u, x_{ik_2}^g = 1 - v, x_{ik_3}^g = 1 - z, x_{ik_4}^g = 1 - l) \quad (\text{B.34})
\end{aligned}$$

$$\begin{aligned}
& (P(X_{jk_1}^g \neq x_{ik_1}^g))^2 P(X_{j'k_2}^g \neq x_{ik_2}^g, X_{j'k_3}^g \neq x_{ik_3}^g) \\
&= \sum_{u,v,z=0}^1 [1 - p_{k_1}^g(u)]^2 p_{k_2k_3}^g(v, z)I(x_{ik_1}^g = u, x_{ik_2}^g = 1 - v, x_{ik_3}^g = 1 - z) \quad (\text{B.35})
\end{aligned}$$

$$\begin{aligned}
& (P(X_{jk_1}^g \neq x_{ik_1}^g))^2 P(X_{j'k_1}^g \neq x_{ik_1}^g, X_{j'k_2}^g \neq x_{ik_2}^g) \\
&= \sum_{u,v=0}^1 [1 - p_{k_1}^g(1 - u)]^2 p_{k_1k_2}^g(u, v)I(x_{ik_1}^g = 1 - u, x_{ik_2}^g = 1 - v) \quad (\text{B.36})
\end{aligned}$$

$$(P(X_{jk_1}^g \neq x_{ik_1}^g))^2 P(X_{j'k_1}^g \neq x_{ik_1}^g) = \sum_{u=0}^1 [1 - p_{k_1}^g(u)]^3 I(x_{ik_1}^g = u) \quad (\text{B.37})$$

$$(P(X_{jk_1}^g \neq x_{ik_1}^g, X_{jk_2}^g \neq x_{ik_2}^g))^2 = \sum_{u,v=0}^1 (p_{k_1k_2}^g(u, v))^2 I(x_{ik_1}^g = 1 - u, x_{ik_2}^g = 1 - v) \quad (\text{B.38})$$

$$\begin{aligned}
& P(X_{j_1k}^g \neq x_{ik}^g)P(X_{j_2k}^g \neq x_{ik}^g, X_{j_3k}^g \neq x_{ik}^g) \\
&= \sum_{u=0}^1 p_k^g(u)[1 - p_k^g(u)]^2 I(x_{ik}^g = u) \quad (\text{B.39})
\end{aligned}$$

$$\begin{aligned}
& P(X_{j_1 k_1}^g \neq x_{i k_1}^g)P(X_{j_2 k_2}^g \neq X_{j_3 k_2}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g) \\
&= \sum_{u,v=0}^1 [1 - p_{k_1}^g(u)]p_{k_2}^g(v)[1 - p_{k_2}^g(v)]I(x_{i k_1}^g = u, x_{i k_2}^g = v)
\end{aligned} \tag{B.40}$$

$$\begin{aligned}
& P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_1}^g \neq x_{i k_1}^g)P(X_{j_1 k_1}^g \neq x_{i k_1}^g, X_{j_1 k_2}^g \neq x_{i k_2}^g) \\
&= \sum_{u,v=0}^1 p_{k_1}^g(1-u)[1 - p_{k_1}^g(1-u)]p_{k_1 k_2}^g(u,v)I(x_{i k_1}^g = 1-u, x_{i k_2}^g = 1-v)
\end{aligned} \tag{B.41}$$

$$\begin{aligned}
& P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_1}^g \neq x_{i k_1}^g)P(X_{j_1 k_2}^g \neq x_{i k_2}^g, X_{j_1 k_3}^g \neq x_{i k_3}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1}^g(u)[1 - p_{k_1}^g(u)]p_{k_2 k_3}^g(v,z)I(x_{i k_1}^g = u, x_{i k_2}^g = 1-v, x_{i k_3}^g = 1-z)
\end{aligned} \tag{B.42}$$

$$\begin{aligned}
& P(X_{j_1 k_1}^g \neq x_{i k_1}^g)P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g) \\
&= \sum_{u,v,z=0}^1 [1 - p_{k_1}^g(z)]p_{k_1}^g(1-u)p_{k_1 k_2}^g(u,v)I(x_{i k_1}^g = z, x_{i k_2}^g = 1-v)
\end{aligned} \tag{B.43}$$

$$\begin{aligned}
& P(X_{j_1 k_1}^g \neq x_{i k_1}^g)P(X_{j_2 k_2}^g \neq X_{j_3 k_2}^g, X_{j_2 k_3}^g \neq x_{i k_3}^g) \\
&= \sum_{u,v,z=0}^1 [1 - p_{k_1}^g(z)]p_{k_2}^g(1-u)p_{k_2 k_3}^g(u,v)I(x_{i k_1}^g = z, x_{i k_3}^g = 1-v)
\end{aligned} \tag{B.44}$$

$$\begin{aligned}
& P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g)P(X_{j_1 k_1}^g \neq x_{i k_1}^g, X_{j_1 k_2}^g \neq x_{i k_2}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1}^g(1-z)p_{k_1 k_2}^g(z,v)p_{k_1 k_2}^g(u,v)I(x_{i k_1}^g = 1-u, x_{i k_2}^g = 1-v)
\end{aligned} \tag{B.45}$$

$$\begin{aligned}
& P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g)P(X_{j_1 k_1}^g \neq x_{i k_1}^g, X_{j_1 k_3}^g \neq x_{i k_3}^g) \\
&= \sum_{u,v,z,l=0}^1 p_{k_1}^g(1-u)p_{k_1 k_2}^g(u,v)p_{k_1 k_3}^g(z,l)I(x_{i k_1}^g = 1-z, x_{i k_2}^g = 1-v, x_{i k_3}^g = 1-l)
\end{aligned} \tag{B.46}$$

$$\begin{aligned}
& P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g)P(X_{j_1 k_2}^g \neq x_{i k_2}^g, X_{j_1 k_3}^g \neq x_{i k_3}^g) \\
&= \sum_{u,v,z=0}^1 p_{k_1}^g(1-z)p_{k_1 k_2}^g(z,u)p_{k_2 k_3}^g(u,v)I(x_{i k_2}^g = 1-u, x_{i k_3}^g = 1-v)
\end{aligned} \tag{B.47}$$



$$\begin{aligned}
& P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g) P(X_{j_1 k_3}^g \neq x_{i k_3}^g, X_{j_1 k_4}^g \neq x_{i k_4}^g) \\
&= \sum_{u,v,z,l=0}^1 p_{k_1}^g(1-u) p_{k_1 k_2}^g(u,v) p_{k_3 k_4}^g(z,l) I(x_{i k_2}^g = 1-v, x_{i k_3}^g = 1-z, x_{i k_4}^g = 1-l)
\end{aligned} \tag{B.48}$$

$$\begin{aligned}
& P(X_{j_1 k_2}^g \neq x_{i k_2}^g) P(X_{j_2 k_1}^g \neq X_{j_3 k_1}^g, X_{j_2 k_2}^g \neq x_{i k_2}^g) \\
&= \sum_{u,v=0}^1 [1 - p_{k_2}^g(1-v)] p_{k_1}^g(1-u) p_{k_1 k_2}^g(u,v) I(x_{i k_2}^g = 1-v)
\end{aligned} \tag{B.49}$$

Thus

$$\begin{aligned}
& \left[ \sum_{k=1}^K (1 - 2\theta_k^{g_2}) P(X_{jk}^{g_1} \neq x_{ik}^{g_1}) \right]^2 \\
&= \sum_{k=1}^K (1 - 2\theta_k^{g_2})^2 \left( P(X_{jk}^{g_1} \neq x_{ik}^{g_1}) \right)^2 \\
&\quad + 2 \sum_{k_1 < k_2} (1 - 2\theta_{k_1}^{g_2}) P(X_{j k_1}^{g_1} \neq x_{i k_1}^{g_1}) (1 - 2\theta_{k_2}^{g_2}) P(X_{j k_2}^{g_1} \neq x_{i k_2}^{g_1}) \\
&= \sum_{k=1}^K \sum_{u=0}^1 (1 - 2\theta_k^{g_2})^2 [1 - p_k^{g_1}(u)]^2 I(x_{ik}^{g_1} = u) \\
&\quad + 2 \sum_{k_1 < k_2} \sum_{u,v=0}^1 (1 - 2\theta_{k_1}^{g_2}) (1 - 2\theta_{k_2}^{g_2}) [1 - p_{k_1}^{g_1}(u)] [1 - p_{k_2}^{g_1}(v)] I(x_{i k_1}^{g_1} = u, x_{i k_2}^{g_1} = v)
\end{aligned} \tag{B.50}$$

$$\begin{aligned}
& \left[ \sum_{k_1 \neq k_2} \theta_{k_2}^{g_2} P(X_{j k_1}^{g_1} \neq x_{i k_1}^{g_1}) \right]^2 \\
&= \sum_{k_1 \neq k_2} \left( \theta_{k_2}^{g_2} \right)^2 \left( P(X_{j k_1}^{g_1} \neq x_{i k_1}^{g_1}) \right)^2 \\
&\quad + 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} \left( \theta_{k_1}^{g_2} \right)^2 P(X_{j k_2}^{g_1} \neq x_{i k_2}^{g_1}) P(X_{j k_3}^{g_1} \neq x_{i k_3}^{g_1}) \\
&\quad + 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} \left( P(X_{j k_1}^{g_1} \neq x_{i k_1}^{g_1}) \right)^2 \theta_{k_2}^{g_2} \theta_{k_3}^{g_2} \\
&\quad + 2 \sum_{k_1 < k_2} \theta_{k_1}^{g_2} \theta_{k_2}^{g_2} P(X_{j k_1}^{g_1} \neq x_{i k_1}^{g_1}) P(X_{j k_2}^{g_1} \neq x_{i k_2}^{g_1}) \\
&\quad + 2 \sum_{k_1 \neq k_2 \neq k_3} \theta_{k_1}^{g_2} P(X_{j k_1}^{g_1} \neq x_{i k_1}^{g_1}) \theta_{k_2}^{g_2} P(X_{j k_3}^{g_1} \neq x_{i k_3}^{g_1}) \\
&\quad + \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \theta_{k_1}^{g_2} P(X_{j k_2}^{g_1} \neq x_{i k_2}^{g_1}) \theta_{k_3}^{g_2} P(X_{j k_4}^{g_1} \neq x_{i k_4}^{g_1})
\end{aligned}$$

$$\begin{aligned}
&= \sum_{k_1 \neq k_2} \left( \theta_{k_2}^{g_2} \right)^2 \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_1}(u))^2 I(x_{ik_1}^{g_1} = u) \right] \\
&+ 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} \left( \theta_{k_1}^{g_2} \right)^2 \left[ \sum_{u, v=0}^1 (1 - p_{k_2}^{g_1}(u))(1 - p_{k_3}^{g_1}(v)) I(x_{ik_2}^{g_1} = u, x_{ik_3}^{g_1} = v) \right] \\
&+ 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \\ k_2 < k_3}} \theta_{k_2}^{g_2} \theta_{k_3}^{g_2} \left[ \sum_{u=0}^1 (1 - p_{k_1}^{g_1}(u))^2 I(x_{ik_1}^{g_1} = u) \right] \\
&+ 2 \sum_{k_1 < k_2} \theta_{k_1}^{g_2} \theta_{k_2}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u))(1 - p_{k_2}^{g_1}(v)) I(x_{ik_1}^{g_1} = u, x_{ik_2}^{g_1} = v) \right] \\
&+ 2 \sum_{k_1 \neq k_2 \neq k_3} \theta_{k_1}^{g_2} \theta_{k_2}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_1}^{g_1}(u))(1 - p_{k_3}^{g_1}(v)) I(x_{ik_1}^{g_1} = u, x_{ik_3}^{g_1} = v) \right] \\
&+ 2 \sum_{k_1 \neq k_2 \neq k_3 \neq k_4} \theta_{k_1}^{g_2} \theta_{k_3}^{g_2} \left[ \sum_{u, v=0}^1 (1 - p_{k_2}^{g_1}(u))(1 - p_{k_4}^{g_1}(v)) I(x_{ik_2}^{g_1} = u, x_{ik_4}^{g_1} = v) \right]
\end{aligned} \tag{B.51}$$

$$\begin{aligned}
&\left[ \sum_{k_1 < k_2} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) \right]^2 \\
&= \sum_{k_1 < k_2} \left( P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) \right)^2 \\
&+ 2 \sum_{k_1 < k_2 < k_3} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_3}^{g_1} \neq x_{ik_3}^{g_1}) \\
&+ 2 \sum_{k_1 < k_2 < k_3} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_3}^{g_1} \neq x_{ik_3}^{g_1}) P(X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}, X_{jk_3}^{g_1} \neq x_{ik_3}^{g_1}) \\
&+ 2 \sum_{k_1 < k_2 < k_3} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}) P(X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}, X_{jk_3}^{g_1} \neq x_{ik_3}^{g_1}) \\
&+ 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_1 < k_3, k_2 < k_4}} P(X_{jk_1}^{g_1} \neq x_{ik_1}^{g_1}, X_{jk_3}^{g_1} \neq x_{ik_3}^{g_1}) P(X_{jk_2}^{g_1} \neq x_{ik_2}^{g_1}, X_{jk_4}^{g_1} \neq x_{ik_4}^{g_1}) \\
&= \sum_{k_1 < k_2} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) I(x_{ik_1}^{g_1} = 1 - u, x_{ik_2}^{g_1} = 1 - v) \right]^2 \\
&+ 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v=0}^1 p_{k_1 k_2}^{g_1}(u, v) I(x_{ik_1}^{g_1} = 1 - u, x_{ik_2}^{g_1} = 1 - v) \right] \\
&\quad \times \left[ \sum_{u, v=0}^1 p_{k_1 k_3}^{g_1}(u, v) I(x_{ik_1}^{g_1} = 1 - u, x_{ik_3}^{g_1} = 1 - v) \right] \\
&+ 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u, v=0}^1 p_{k_1 k_3}^{g_1}(u, v) I(x_{ik_1}^{g_1} = 1 - u, x_{ik_3}^{g_1} = 1 - v) \right]
\end{aligned}$$

$$\begin{aligned}
& \times \left[ \sum_{u,v=0}^1 p_{k_2 k_3}^{g_1}(u,v) I(x_{ik_2}^{g_1} = 1-u, x_{ik_3}^{g_1} = 1-v) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u,v=0}^1 p_{k_1 k_2}^{g_1}(u,v) I(x_{ik_1}^{g_1} = 1-u, x_{ik_2}^{g_1} = 1-v) \right] \\
& \times \left[ \sum_{u,v=0}^1 p_{k_2 k_3}^{g_1}(u,v) I(x_{ik_2}^{g_1} = 1-u, x_{ik_3}^{g_1} = 1-v) \right] \\
& + 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_1 < k_3, k_2 < k_4}} \left[ \sum_{u,v=0}^1 p_{k_1 k_3}^{g_1}(u,v) I(x_{ik_1}^{g_1} = 1-u, x_{ik_3}^{g_1} = 1-v) \right] \\
& \times \left[ \sum_{u,v=0}^1 p_{k_2 k_4}^{g_1}(u,v) I(x_{ik_2}^{g_1} = 1-u, x_{ik_4}^{g_1} = 1-v) \right] \\
= & \sum_{k_1 < k_2} \sum_{u,v=0}^1 \left( p_{k_1 k_2}^{g_1}(u,v) \right)^2 I(x_{ik_1}^{g_1} = 1-u, x_{ik_2}^{g_1} = 1-v) \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u,v,z=0}^1 p_{k_1 k_2}^{g_1}(u,v) p_{k_1 k_3}^{g_1}(u,z) \right. \\
& \quad \left. \times I(x_{ik_1}^{g_1} = 1-u, x_{ik_2}^{g_1} = 1-v, x_{ik_3}^{g_1} = 1-z) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u,v,z=0}^1 p_{k_1 k_3}^{g_1}(u,v) p_{k_2 k_3}^{g_1}(z,v) \right. \\
& \quad \left. \times I(x_{ik_1}^{g_1} = 1-u, x_{ik_2}^{g_1} = 1-z, x_{ik_3}^{g_1} = 1-v) \right] \\
& + 2 \sum_{k_1 < k_2 < k_3} \left[ \sum_{u,v,z=0}^1 p_{k_1 k_2}^{g_1}(u,v) p_{k_2 k_3}^{g_1}(v,z) \right. \\
& \quad \left. \times I(x_{ik_1}^{g_1} = 1-u, x_{ik_2}^{g_1} = 1-v, x_{ik_3}^{g_1} = 1-z) \right] \\
& + 2 \sum_{\substack{k_1 \neq k_2 \neq k_3 \neq k_4 \\ k_1 < k_3, k_2 < k_4}} \left[ \sum_{u,v,z,l=0}^1 p_{k_1 k_3}^{g_1}(u,v) p_{k_2 k_4}^{g_1}(z,l) \right. \\
& \quad \left. \times I(x_{ik_1}^{g_1} = 1-u, x_{ik_2}^{g_1} = 1-z, x_{ik_3}^{g_1} = 1-v, x_{ik_4}^{g_1} = 1-l) \right]
\end{aligned}$$

(B.52)

# Bibliography

- Amfoh, K. K., Shaw, R. F. and Bonney, G. E. (1994). The use of logistic models for the analysis of codon frequencies of DNA sequences in terms of explanatory variables. *Biometrics* **50**, 1054–1063.
- 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, 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.
- Bahadur, R. R. (1961). A representation of the joint distribution of responses to  $n$  dichotomous items. In *Studies in Item Analysis and Prediction*, H. Solomon (ed.). Stanford University Press. pp. 158–176.
- Berger, J. O. (1980). *Statistical Decision Theory and Bayesian Analysis*. second edn. Springer-Verlag.
- Besag, J. (1972). Nearest-neighbor systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B* **34**, 75–83.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 192–236.
- Besag, J. (1975). Statistical of non-lattice data. *The Statistician* **24**, 179–195.
- Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B* pp. 259–302.
- Besag, J. (1989). Digital image processing towards Bayesian image analysis. *Journal of Applied Statistics* **16**, 395–407.
- Bose, R. C., Chakravarti, I. M., Mahalanobis, P. C., Rao, C. R. and Smith, K. J. C. (1970). A representation of the joint distribution of responses to  $n$  dichotomous items. In *Essays in Probability and Statistics*, R. C. Bose (ed.). The University of North Carolina Press. pp. 111–132.

- Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models*. Marcel Dekker, INC.
- Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**, 167–174.
- Chakraborty, R. and Rao, C. R. (1991). Measurement of genetic variation for evolutionary studies. In *Handbook of Statistics 8*, R. Rao, C R & Chakraborty (ed.). North-Holland. pp. 271–316.
- Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes - Models and Applications*. Pion.
- Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. John Wiley & Sons, Inc.
- Coffin, J. M. (1986). Genetic variation in AIDS viruses. *Cell* **46**, 1–4.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. second edn. John Wiley & Sons, INC.
- Cullen, B. R. (1991). Human immunodeficiency virus as a prototypic complex retrovirus. *Journal of Virology* **65**, 1053–1056.
- Daniel, W. W. (1978). *Applied Nonparametric Statistics*. Houghton Mifflin Company.
- Davidson, R. R. and Bradley, R. A. (1969). Multivariate paired comparisons: The extension of a univariate model and associated estimation and test procedures. *Biometrika* **56**, 81–95.
- Dawid, A. P. (1983). Introduction to Bayesian statistics. In *Encyclopedia of Statistics 4*, S. Kotz and N. L. Johnson (eds). John Wiley & Sons. pp. 89–105.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
- Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972–985.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B* **54**, 657–699.
- Gini, C. W. (1912). Variabilita e Mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari* **3**(2), 3–159.

- Gojobori, T., Ishii, K. and Nei, M. (1982). Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *Journal of Molecular Evolution* **18**, 414–423.
- Gojobori, T., Moriyana, E. N. and Kimura, M. (1990). Statistical methods for estimating sequence divergence. *Methods in Enzimology* **183**, 531–550.
- Gonick, L. and Wheelis, M. (1991). *The Cartoon Guide to Genetics*. updated edn. Harper Perennial.
- Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*. McGraw-Hill, New York.
- Hahn, B. H., Gonda, M. A., Shaw, G. M., Popovic, M. and Hoxie, J. A. (1985). Genomic diversity of the acquired immune deficiency syndrome virus HTLV-III: Different viruses exhibit greatest divergence in their envelope genes. *Proc. Natl. Acad. Sci. USA* **82**, 4813–4817.
- Hahn, B. H., Shaw, G. M., Taylor, M. E., Redfield, R. R. and Markham, P. D. (1986). Genetic variation in HTLV-III/LAV over time in patients with AIDS. *Science* **232**, 1548–1553.
- Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript, Oxford University.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chain and their applications. *Biometrika* **57**, 97–109.
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. of Math. Stat.* **19**, 293–325.
- Huffer, F. W. and Wu, H. (1995a). Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Not published yet.
- Huffer, F. W. and Wu, H. (1995b). Variable selection in auto-logistic models. Not published yet.
- Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**, 454–458.
- Kimura, M. and Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* **2**, 87–90.
- Kullback, S. (1983). Kullback information. In *Encyclopedia of Statistics* **4**, S. Kotz and N. L. Johnson (eds). John Wiley and Sons. pp. 421–425.
- Kypr, J. and Mrázek, J. (1987). Unusual codon usage of HIV. *Nature* **327**, 20.
- Lee, A. J. (1990). *U-Statistics - Theory and Practice*. Marcel Dekker, Inc.

- Liang, K., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54**, 3–40.
- 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, 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.
- Mansky, L. M. and Temin, H. M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology* **69**, 5087–5094.
- McCutchan, F. E., Ungar, B. L. P., Hegerich, P., Roberts, C. R., Fowler, A. K., Hira, S. K., Perine, P. L. and Burke, D. S. (1992). Genetic analysis of HIV-1 isolates from zambia and an expanded phylogenetic tree for HIV-1. *Journal of Acquired Immune Deficiency Syndromes* **5**, 441–449.
- Mises, R. V. (1947). On the asymptotic distribution of the differentiable statistical functions. *Annals of Mathematical Statistics* **18**, 309–348.
- Nei, M. (1975). *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam.
- Parzen, E. (1962). *Stochastic Processes*. Holden-Day, INC.
- Puri and Sen, P. K. (1985). *Nonparametric Multivariate Analysis*.
- Raftery, A. E. (1985). A model for high-order markov chains. *Journal of the Royal Statistical Society, Ser. B* **47**, 528–539.
- Raftery, A. E. and Tavaré, S. (1994). Estimation and modeling repeated patterns in high order Markov chains with the mixture transition distribution model. *Applied Statistics* **43**, 179–199.
- Roberts, J. D., K, B. and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science* **242**, 1171–1173.
- Roy, S. N., Greenberg, B. G. and Sarhan, A. E. (1960). Evaluation of determinants, characteristic equations, and their roots for a class of patterned matrices. *Journal of the Royal Statistical Society, Ser. B* **22**(2), 348–359.
- Searle, S. R. (1971). *Linear Models*. John Wiley & Sons.
- Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley & Sons.
- 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, P. K. (1967). On some multisample permutation tests based on a class of  $u$ -statistics. *American Statistical Association Journal* pp. 1201–1213.
- Sen, P. K. (1981). *Invariance Principles and Statistical Inference*. John Wiley & Sons.
- Sen, P. K. (1995). Paired comparisons for multiple characteristics: An anocova approach. In *Statistical Theory and Practice: Papers in Honor of H. A. David*, H. N. Nagaraja, P. K. Sen and D. F. Morrison (eds). Springer-Verlag, New York. pp. 247–264.
- Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall.
- Sharp, P. M. (1986). What can AIDS virus codon usage tell us. *Nature* **324**, 114.
- Simpson, E. H. (1949). The Measurement of Diversity. *Nature* **163**, 688.
- Smith, R. L. (1994). A tutorial on Markov chain Monte Carlo. Unpublished Manuscript.
- Smith, R. L., Calloway, M. O. and Morrissey, J. P. (1994). Network autocorrelation with a binary dependent variable: A method and an application. Not yet published.
- Smith, T. F., Srinivasan, A., Schochetman, G., Marcus, M. and Myers, G. (1988). The phylogenetic history of immunodeficiency viruses. *Nature* **333**, 573–575.
- Strauss, D. J. (1977). Clustering on coloured lattices. *Journal of Applied Probability* **14**, 135–143.
- Takahata, N. and Kimura, M. (1981). *Genetics* **98**, 641.
- Tanner, M. A. (1993). *Tools for Statistical Inference*. second edn. Springer-Verlag.
- Tavaré, S. and Giddings, B. W. (1989). Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences*, M. S. Waterman (ed.). CRC Press. pp. 117–132.
- Teich, N. (1984). Taxonomy of retroviruses. In *RNA tumor viruses vol. 2*, R. Weiss, N. Teich, H. Varmus and J. Coffin (eds). Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.. pp. 25–207.
- Varmus, H. and Brown, P. (1989). Retroviruses. In *Mobile DNA*, D. E. Berg and M. M. Howe (eds). American Society for Microbiology. pp. 53–108.
- Weir, B. S. (1990a). *Genetic Data Analysis*. Sinauer Associates.
- Weir, B. S. (1990b). Variation in sequence distances. *Molecular Evolution* pp. 281–288.
- Wu, H. and Huffer, F. W. (1995). Modeling the distribution of plant species using the autologistic regression model. Not published yet.