Université d'Ottawa • University of Ottawa

PERMISSION TO REPRODUCE AND DISTRIBUTE THE THESIS

PERMISSION DE REPRODUIRE ET DE DISTRIBUER LA THÈSE

NOM DE L'AUTEUR / NAME OF AUTHOR:	Xin Gao
ADRESSE POSTALE / MAILING ADDRESS:	4545 Walkley Ave. Montreal, Quebec H4B 2K8
GRADE / DEGREE:	**ANNÉE D'OBTENTION / YEAR GRANTED**
Ph.D - Mathematics	2003
TITRE DE LA THÈSE / TITLE OF THESIS:	
Rank Test for Interaction in Two-Way Layouts with Application in Genetic Analysis	

L'auteur permet, par la présente, la consultation et le prêt de cette thèse en conformité avec les règlements établis par le bibliothécaire en chef de l'Université d'Ottawa. L'auteur autorise aussi l'Université d'Ottawa, ses successeurs et cessionnaires, à reproduire cet exemplaire par photograhie ou photocopie pour fins de prêt ou de vente au prix coûtant aux bibliothèques ou aux chercheurs qui en feront la demande.

Les droits de publication par tout autre moyen et pour vente au public demeureront la propriété de l'auteur de la thèse sous réserve des règlements de l'Université d'Ottawa en matière de publication de thèses.

The author hereby permits the consultation and the lending of this thesis pursuant to the regulations established by the Chief Librarian of the University of Ottawa. The author also authorizes the University of Ottawa, its successors and assignees, to make reproductions of this copy by photographic means or by photocopying and to lend or sell such reproductions at cost to libraries and to scholars requesting them.

The right to publish the thesis by other means and to sell it to the public is reserved to the author, subject to the regulations of the University of Ottawa governing the publication of theses.

N.B. LE MASCULIN COMPREND ÉGALEMENT LE FÉMININ

Aug 26 2003

DATE

(AUTEUR) SIGNATURE (AUTHOR)

Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

GAO, Xin

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Ph. D. (Mathematics)

GRADE - DEGREE

Mathematics

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Rank Test for Interaction in Two-Way Layouts with Application in Genetic Analysis

M. Alvo

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

P. Cabilio

A. Dabrowski

M. Saliban Barrera

M. Zarepour

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

SIGNATURE

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

RANK TESTS FOR INTERACTION IN TWO-WAY LAYOUTS WITH APPLICATION IN GENETIC ANALYSIS

By

Xin Gao

August 2003

A Thesis

submitted to the School of Graduate Studies and Research

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy in Mathematics[1]

[1]The Ph.D. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Acknowledgments

I would like to thank my advisor Dr. Alvo for his patient guidance and useful advice throughout my thesis work. I also want to express my thanks to Dr. Scott, Dr. Dabrowski, Ms. Dalrymple, Ms. Giroux and Ms. Hotte for their support and assistance to me during my studies at University of Ottawa. I want to thank my parents, my parents-in law, sisters and brothers for their love and support. Finally I want to thank my husband and my son for the inspiration they gave me. This work is supported by Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship.

Abstract

To fully dissect complex traits, it is desirable to have methods able to test gene-gene interaction for human genetic data. This thesis provides a framework to process human quantitative trait data such that the problem becomes a hypothesis test of interaction for a two-way layout design with unequal replicates. Three new nonparametric rank tests are proposed. Their limiting distributions under Pitman alternatives and asymptotic relative efficiencies are studied. The tests are extended to unbalanced designs. We also introduce the notion of composite linear rank statistics and prove asymptotic normality under mild conditions. Consistent estimators are provided for the limiting variance-covariance matrix of arbitrary linear rank statistics.

KEY WORDS: rank test, interaction, genetics, Pitman alternatives, asymptotic relative efficiency, composite linear rank statistics, consistent estimator.

Dedication

To my parents- Kecheng, Changli, Minghao and Shuiying

To my son- Xingming

To my husband- Hong

Contents

Acknowledgments ii

Abstract iii

Dedication iv

Notations 1

1 Introduction 4

2 Literature Review 7

2.1 Existing nonparametric methods for testing interaction 7

3 Composite Linear Rank Statistics 13

3.1 Asymptotic normality of simple linear rank statistics 13

3.2 Asymptotic normality of composite linear rank statistics 16

4 Tests for Balanced Designs 24

4.1 Proposed new rank statistics 24

4.2 Expectations under the null hypothesis 27

4.3 Limiting distributions under the null hypothesis 29

4.4 Limiting distributions under the alternatives and Asymptotic Relative Efficiency 42

4.5 Estimation of the covariance structures 48

4.6 Covariance structure between row rank sum and column rank sum . . 57

5 Extension to Unbalanced Designs **59**

5.1 Extension to unbalanced designs with proportional cell weights 59

5.2 Extension to unbalanced designs with arbitrary cell weights 63

6 Monte Carlo Simulation Study **68**

6.1 Monte Carlo simulation study for balanced designs 68

6.2 Monte Carlo study for unbalanced designs 73

7 Application and Discussion **90**

7.1 Application of rank tests on gene interaction when genotypes are known 90

7.2 Discussion 94

7.3 Appendix 101

List of Tables

1 Convergence of the consistent estimators 57

2 Underlying distributions for the simulated errors 69

3 Cell sizes for the simulated unbalanced design 74

4 Genotype combinations frequencies at locus A and B 93

List of Figures

1 Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the normal distribution. 75

2 Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the uniform distribution. 76

3 Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Lognormal distribution. 77

4 Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the double exponential distribution. 78

5 Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Cauchy distribution. 79

6 Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the normal distribution. 80

7 Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the uniform distribution. 81

8 Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Lognormal distribution. 82

9 Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the double exponential distribution. 83

10 Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Cauchy distribution. 84

11 The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the normal distribution. 85

12 The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the uniform distribution. 86

13 The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the Lognormal distribution. 87

14 The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the double exponential distribution. 88

15 The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the Cauchy distribution. 89

Notation

Ω: a set of independent variables $X_1, X_2, \ldots, X_N$

$\mathcal{A}$: a collection of subsets in Ω

$C_A(X_i)$: coefficient associated with set A and variable X_i

$\overline{C_A}$: the average of $C_A(X_i)$ over set A

n_A: the number of variables in set A

ϕ: score generating function with bounded second derivative

$H_A(x) = \frac{1}{n_A} \sum_{\{i|X_i \in A\}} F_i(x)$

$F_{ij}(x) = F(x - \theta - \alpha_i - \beta_j - \delta_{ij})$

$H_{i.}(x) = \frac{1}{J} \sum_{j=1}^{J} F_{ij}(x)$

$H_{.j}(x) = \frac{1}{I} \sum_{i=1}^{I} F_{ij}(x)$

$H(x) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} F_{ij}(x)$

$F_{ij;N}(x) = F(x - \theta - \alpha_i - \beta_j - \delta_{ij}/\sqrt{N})$

$F_{ij;0}(x) = F(x - \theta - \alpha_i - \beta_j)$

$H_{i;N}(x) = \frac{1}{J} \sum_{j=1}^{J} F_{ij;N}(x)$

$H_{i;0}(x) = \frac{1}{J}\sum_{j=1}^{J} F_{ij;0}(x)$

$H_{j;N}(x) = \frac{1}{I}\sum_{i=1}^{I} F_{ij;N}(x)$

$H_{j;0}(x) = \frac{1}{I}\sum_{i=1}^{I} F_{ij;0}(x)$

$H_N(x) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} F_{ij;N}(x)$

$H_0(x) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} F_{ij;0}(x)$

$u(x) = 1$ when $x \geq 0$ and $u(x) = 0$ otherwise

$L_N = \max(\max_{i=1}^{n_1+n_2} (c_i^{(1)} - \bar{c}^{(1)})^2, \max_{i=n_2+1}^{n_1+n_2+n_3} (c_i^{(2)} - \bar{c}^{(2)})^2)$

r_{ijn}: rank of X_{ijn} among the ith row

c_{ijn}: rank of X_{ijn} among the jth column.

$a_{ijn} = r_{ijn} + c_{ijn}$

$S_N(i,j) = \sum_{n=1}^{N} \alpha_{n_i.}(r_{ijn})$

$S_N = (S_N(1,1), S_N(1,2), \ldots, S_N(I,J))'$

$T_N(i,j) = \sum_{n=1}^{N} \alpha_{n_{.j}}(c_{ijn})$

$T_N = (T_N(1,1), T_N(1,2), \ldots, T_N(I,J))'$

$\mathbf{I}_I$ and $\mathbf{I}_J$: the identity matrices of dimensions I and J

$\mathbf{J}_I$ and $\mathbf{J}_J$: the matrices with all elements equal to one and dimensions I and J

$A_1 = (\frac{-1}{I})\mathbf{I}_J$, $A = \mathbf{J}_I \otimes A_1 + \mathbf{I}_I \otimes \mathbf{I}_J$

$B_1 = \mathbf{I}_J + (\frac{-1}{J})\mathbf{J}_J$, $B = \mathbf{I}_I \otimes B_1$

$\tilde{S}_N(i,j) = S_N(i,j) - \frac{1}{I}\sum_{a=1}^{I} S_N(a,j)$

$\tilde{S}_N = (\tilde{S}(1,1), \tilde{S}(1,2), \ldots, \tilde{S}(I,J))'$

$\tilde{T}_N(i,j) = T_N(i,j) - \frac{1}{J}\sum_{b=1}^{J} T_N(i,b)$

$\tilde{T}_N = (\tilde{T_N}(1,1), \tilde{T_N}(1,2), \ldots, \tilde{T}(I,J))'$

$\tilde{S}_N = AS_N,\ \tilde{T}_N = BT_N$

$\Sigma_1 = \frac{1}{N}\lim_{N\to\infty} \mathrm{var}(S_N)$

$\Sigma_2 = \frac{1}{N}\lim_{N\to\infty} \mathrm{var}(T_N)$

$\Sigma_{12} = \frac{1}{N}\lim_{N\to\infty} \mathrm{cov}(S_N, T_N)$

$\Psi_{ij}^{(i,j)} = \sum_{b\neq j} F_{ib}(X_{ij1})$

$\Psi_{ib}^{(i,j)} = -F_{ij}(X_{ib1})$ for $b \neq j$

Chapter 1

Introduction

A key problem in understanding human complex traits is the study of the interaction (genetic epistasis) among multiple genes which are responsible for the disease. Most diseases and traits do not follow a simple Mendelian inheritance pattern. For example, a recent report from the Collaborative Study on the Genetics of Asthma (CSGA) [27] has shown strong evidence of complex interaction between genes in region D11S2002 and D12S2070. Complex traits, such as asthma, are likely to be governed by two or more genes which act together to determine an individual's risk of susceptibility to disease. The big obstacle in testing gene interaction arises in the inability of conducting human crosses to suit any prespecified experimental design.

Each gene has different copies denoted by alleles, such as A and a. Genotype stands for the unordered pair of alleles, such as AA, Aa or aa. The measurement

of the quantitative trait is influenced by the genotypes of the genes as well as their interaction. The biological problem can be formulated as a statistical test of interaction in an experimental design. For a more detailed explanation of how the gene-gene interaction can be formulated as interaction terms in a linear model, readers are referred to Section 7.1. The key feature of this design is that prior to genotyping a selected individual, we have no information about the actual genotypes of the individual. Thus even the number of replicates for each genotype combination is not known in advance. There are nonparametric tests available to test for main effects in two-way layouts. However, very few nonparametric tests are available to test for interaction effect, and none are applicable to unbalanced designs. When the assumption of normality is violated, which is true for some cases of gene-gene interaction, there is a need to provide a robust, distribution-free test to deal with unbalanced designs. To address this gap, in Chapter three, we introduce composite linear rank statistics (CLRS) and investigate their asymptotic behavior. This constitutes the theoretical basis for the new rank tests. In Chapter four, we propose three new rank statistics for testing interaction which are quadratic forms of CLRS. We study the properties of the tests under the null hypothesis and under a sequence of Pitman Alternatives and obtain the asymptotic relative efficiencies of the tests. We provide as well consistent estimators for the limiting covariance matrix of the CLRS associated with the tests. In Chapter five, we extend our tests to accommodate unbalanced designs with

proportional cell weight. We introduce a new definition of weighted rank statistics and extend the tests to unbalanced designs with arbitrary cell weight. Finally, Monte Carlo simulations provide a comparison of the type I error and the power for several tests. Our tests are shown to perform well for several types of error distributions.

Chapter 2

Literature Review

2.1 Existing nonparametric methods for testing interaction

In this thesis, we will be concerned with testing for interaction in a 2-way layout design specified by the semiparametric linear model

$$X_{ijn} = \theta + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijn}, \quad i = 1, \ldots, I, j = 1, \ldots, J, n = 1, \ldots, N, \qquad (2.1.1)$$

where the errors ϵ_{ijn} are iid random variables with unspecified absolute continuous cdf F, θ is a common mean, α_i represents a treatment effect, β_j a block effect and γ_{ij} an interaction effect.

In practice, especially in biological settings, there exist situations where the errors are not normally distributed. Biological readings such as blood pressure and

cholesterol level are limited to certain ranges. Other example includes penetrance of developing certain disease, which is not normally distributed [11].

When the normal assumption is violated, there are rank tests available to test for interaction.

Aligned tests advocated by Mehra, Sen and Mansouri et al [17], [21], [15] require preprocessing of the data by subtracting the Lehmann estimates of the row effect and column effect. In a two-way layout, the observations x_{ijk} are aligned by subtracting the Lehmann estimates of the mean row effect and mean column effect. The Lehmann estimate of the main effect is based on the median of all the averages of pairwise values [12]. After alignment, the observations are ranked together and the classical analysis of variance is applied to the overall rankings to detect the interaction effect. Aligned tests still have limitations. First they are not invariant under monotone increasing transformations of the data. Secondly, aligned tests cannot be applied to unbalanced designs, which limits their use in genetic data.

Rank transform method is another popular but controversial test for interaction. All the observations are first ranked together. Then the usual method of analysis of variance is applied to the overall rankings. The method is appealing in its simplicity. Moreover it can be implemented using existing commercial software. However, as first pointed out by Fligner [5] and later on proved by Thompson [25], the asymptotic

distribution of the rank transform test for interaction in a balanced two-way layout is $\chi^2_{(IJ-I-J+1)}/(IJ-I-J+1)$ if and only if there are exactly two levels of each main effect, i.e. $I = J = 2$. This limits the use of the rank transform method to designs with factor levels not exceeding 2. In other cases, the expectation of the test statistic is divergent even under the null hypothesis. This conclusion also agrees with the simulation results performed by Blair, Sawilowsky and Higgins [3] which detect unacceptably large type I errors. Since the method has been endorsed by the SAS manual and widely used by practitioners in all applied sciences, there is a great need to emphasize its shortcomings in most cases.

There are several tests proposed by Lemmer and Stoker [14], Crouse [4], and de Kroon and van der Lann [9] that require the absence of one of the main effects. Patil and Hoel [18], Marden and Muyot [16] proposed nonparametric tests for their own definitions of interaction based on paired comparisons. For a $R \times C$ design, denote $\mu_{ii'jj'} = P(X_{ij'} \leq X_{ij}) - P(X_{i'j'} \leq X_{i'j})$, where $i \neq i'$ and $j \neq j'$. They introduced a measure of interaction of the factor A and the factor B by $\delta^2_{AB} = \mu'\mu$, where $\mu = \{\mu_{ii'jj'}, 1 \leq i < i' \leq R, 1 \leq j < j' \leq C\}'$. This definition of interaction involves a large number of parameters, one parameter for each pair of cell combination, resulting in a test statistic with a large number of degrees of freedom $f = RC(R-1)(C-1)/4$. Another drawback of this method is that utilizing pairwise comparisons leads to potential power loss compared to methods based on overall rankings or within block

rankings because of sample size considerations. As indicated by their simulation studies, their method requires moderate large cell sizes ($n > 30$) for satisfactory performance.

Testing hypotheses for unbalanced designs is a challenging question and remains largely unsolved in the literature. The lack of efficient methodologies to cope with unbalanced designs complicates the application of rank methods to practical settings. Bhapkar and Gore [2] proposed a nonparametric test based on U-statistics to test interaction for orthogonal design, i.e. $n_{ij} = n_{i.}n_{.j}/N$, where $n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$ and $N = \sum_i \sum_j n_{ij}$.

In contrast to the semiparametric linear model specified by equation 2.1.1, Akritas and Arnold et al [1] considered a novel nonparametric model which only specifies that observations in different cells have different distribution functions. They proposed the use of nonparametric hypotheses instead of parametric hypotheses for this nonparametric model. Let $H_{i.} = \frac{1}{J}\sum_{j=1}^{J} F_{ij}$, $H_{.j} = \frac{1}{I}\sum_{i=1}^{I} F_{ij}$ and $H = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} F_{ij}$. Let A be the row-factor and B be the column-factor. The hypothesis for main effects, and interaction effects given by them are as follows:

$$H_0(A) : A_i = H_{i.} - H = 0, \quad \forall i = 1, \ldots, a.$$

$$H_0(AB) : C_{ij} = F_{ij} - H_{i.} - H_{.j} + H = 0, \quad \forall i = 1, \ldots, a, j = 1, \ldots, b.$$

$$H_0(A|B) : A_i + C_{ij} = F_{ij} - H_{.j} = 0, \quad \forall i = 1, \ldots, a, \quad j = 1, \ldots, b,$$

where $H_0(A)$ stands for no main factor A effect, $H_0(AB)$ stands for no interaction effect, and $H_0(A|B)$ tests for no simple factor A effect, which means the factor A has no effect through either the main effects or interaction. Rank tests were developed to test for these nonparametric hypotheses. Since the linear model is more commonly used and the main and interaction effects can be more easily explained, this thesis is focused on the semiparametric linear model.

Except for the approach of Akritas et al [1] which is based on a different model, all the approaches reviewed above are restricted to special designs. For instance, rank transform method is only valid for 2×2 designs. Aligned tests are restricted to balanced designs. The tests due to Lemmer & Stoker [14], Crouse [4], and de Kroon & van der Lann [9] require one of the main effects to be absent. The test due to Bhapkar and Gore [2] is dealt with orthogonal designs. In contrast to this list of existing methods, the new rank tests proposed in this thesis are applicable to a very wide range of designs, including balanced designs and unbalanced designs with arbitrary cell replicates, designs with or without the presence of main effects.

Another novelty of our approach is the way of pooling ranking informations from overlapped subsets, for instance, combining row ranks and column ranks together. In contrast, traditional rank tests are limited to use only overall ranking or within block rankings, which utilize ranking information from one set or non-overlapped subsets. Thus our method provides a new way to gather ranking information in the practice

of nonparametric testing.

Most important of all, this thesis develops a new concept of composite linear rank statistics (CLRS). All the ranking methods in literature rely on the theoretical work regarding the asymptotic normality of simple linear rank statistics. This thesis establishes the asymptotic normality of sums of correlated linear rank statistics, thus providing a very general extension to the Hájek theorem [8].

The combination of the concept of CLRS and the estimation method of its limiting variance and covariance structure serves as a new tool to construct rank tests for different hypotheses adaptively and mechanically. A direct application of this tool is to construct a weighted rank to solve the hypothesis testing in unbalanced designs. There is a great potential to employ CLRS to solve other nonparametric hypotheses testing problems.

Chapter 3

Composite Linear Rank Statistics

3.1 Asymptotic normality of simple linear rank statistics

In Hájek's paper [8], the asymptotic normality of simple linear rank statistics was established under very general conditions. Let $X_1, \ldots, X_N$ be independent random variables with continuous distribution functions $F_1, \ldots, F_N$ respectively. Let R_i be the rank of the X_i among $X_1, \ldots, X_N$. Using Hájek's notation, let c_i, $i = 1, \ldots, N$ be regression constants and let $\alpha_N(x)$ be generated by a real valued function $\phi(x)$ having a bounded second derivative either as

$$\alpha_N(i) = \phi(\frac{i}{N+1})$$

or

$$\alpha_N(i) = E\phi(U_N^{(i)}).$$

Here $U_N^{(i)}$ stands for the ith order statistic from a uniform distribution on $(0,1)$. A simple linear rank statistic takes the form of

$$S = \sum_{i=1}^{N} c_i \alpha_N(R_i).$$

Let

$$\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i,$$

$$\bar{\phi} = \int_0^1 \phi(x)dx,$$

$$H(x) = \frac{1}{N}\sum_{i=1}^{N} F_i(x),$$

and

$$\mu = \sum_{i=1}^{N} c_i \int \phi(H(x))dF_i(x).$$

Define $u(x) = 1$, when $x \geq 0$ and $u(x) = 0$, when $x < 0$.

Theorem 3.1.1. *(Hájek [8]) Let*

$L_i(x) = \frac{1}{N}\sum_{j=1}^{N}(c_j - c_i)\int[u(y-x) - F_i(y)]\phi'(H(y))d\dot{F}_j(y)$

and

$$\sigma^2 = \sum_{i=1}^{N} \text{var}(L_i(X_i)).$$

If for every $\epsilon > 0$ *there exists* K_ϵ *such that*

$$\text{var}(S) > K_\epsilon max_{1\le i\le N}(c_i - \bar{c})^2,$$

then

$$\max_{-\infty<x<\infty} |P(S - ES < x(\text{var}\, S)^{\frac{1}{2}}) - \Phi(x)| < \epsilon,$$

with $\Phi(x)$ *denoting the cdf of standard normal distribution. The conclusion still holds if* $\text{var}(S)$ *is replaced by* σ^2. *If* $\sum_i^n c_i^2$ *is bounded by a multiple of* $\sum_i^n (c_i - \bar{c})^2$, ES *can be replaced by* μ *in the conclusion.*

Comment: According to the proof of the theorem, K_ϵ has an explicit expression and it increases as ϵ decreases. The $\text{var}(S)$ is unknown and σ^2 has a closed form expression instead. In practice, we can use σ^2 to estimate $\text{var}(S)$.

Theorem 3.1.2. *(Hájek [8]) Let* $Z_i = L_i(X_i)$. *Assume the score generating function* ϕ *has a bounded second derivative. Consider the statistic* $S = \sum_{i=1}^N c_i \alpha_N(R_i)$. *Then there exists a constant* $M = M(\phi)$ *independent of* N, *such that*

$$E(S - ES - \sum_{i=1}^N Z_i)^2 \le MN^{-1} \sum_{i=1}^N (c_i - \bar{c})^2$$

and

$$(ES - \mu)^2 \le MN^{-1} \sum_{i=1}^N c_i^2.$$

3.2 Asymptotic normality of composite linear rank statistics

First we study the asymptotic behavior of sums of correlated simple linear rank statistics. Each simple linear rank statistic is defined on a different subset and all those subsets are not necessarily disjoint. The following theorem proves the asymptotic normality of sums of correlated linear rank statistics under mild conditions.

Let $X_1, ...X_{n_1}, X_{n_1+1}, ...X_{n_1+n_2}, X_{n_1+n_2+1}, ...X_{n_1+n_2+n_3}$ be independent random variables and let X_i follow distribution F_i.

Let $S_1 = \sum_{i=1}^{n_1+n_2} c_i^{(1)} \alpha(R_i^{(1)})$ be a linear rank statistic where $R_i^{(1)}$ is the rank of X_i among the set $\{X_1, ...X_{n_1}, X_{n_1+1}, ...X_{n_1+n_2}\}$ and $c_i^{(1)}$ is the regression coefficient associated with $R_i^{(1)}$.

Let $S_2 = \sum_{i=n_1+1}^{n_1+n_2+n_3} c_i^{(2)} \alpha(R_i^{(2)})$ be a linear rank statistic, where $R_i^{(2)}$ is the rank of X_i among the set $\{X_{n_1+1}, \cdots, X_{n_1+n_2}, X_{n_1+n_2+1}, \cdots, X_{n_1+n_2+n_3}\}$ and $c_i^{(2)}$ is the regression coefficient associated with $R_i^{(2)}$.

Let $\bar{c}^{(1)} = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} c_i^{(1)}$ and $\bar{c}^{(2)} = \frac{1}{n_2+n_3} \sum_{i=n_1+1}^{n_1+n_2+n_3} c_i^{(2)}$.

Let

$$L_i(X_i) = \frac{1}{n_1+n_2} \sum_{j=1}^{n_1+n_2} (c_j^{(1)} - c_i^{(1)}) \int [u(y - X_i) - F_i(y)] \phi'(H_1(y)) dF_j(y)$$

$$\sigma_1^2 = \sum_{i=1}^{n_1+n_2} \text{var}[L_i(X_i)]$$

$$H_1(x) = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} F_i(x)$$

$$\mu_1 = \sum_{i=1}^{n_1+n_2} c_i^{(1)} \int \phi(H_1(x))dF_i(x).$$

Also let

$$L_i^*(X_i) = \frac{1}{n_2+n_3}\sum_{j=n_1+1}^{n_1+n_2+n_3} (c_j^{(2)} - c_i^{(2)}) \int [u(y-X_i) - F_i(y)]\phi'(H_2(y))dF_j(y)$$

$$\sigma_2^2 = \sum_{i=n_1+1}^{n_1+n_2+n_3} \text{var}[L_i^*(X_i)]$$

$$H_2(x) = \frac{1}{n_2+n_3}\sum_{i=n_1+1}^{n_1+n_2+n_3} F_i(x)$$

$$\mu_2 = \sum_{i=n_1+1}^{n_1+n_2+n_3} c_i^{(2)} \int \phi(H_2(x))dF_i(x).$$

Let $Z_i = L_i(X_i)$ and $Z_i^* = L_i^*(X_i)$ and define

$$W_i = \begin{cases} Z_i, i = 1, \cdots, n_1 \\ Z_i + Z_i^*, i = n_1+1, \cdots, n_1+n_2 \\ Z_i^*, i = n_1+n_2+1, \cdots, n_1+n_2+n_3 \end{cases} \tag{3.2.1}$$

The W_i's are independent of each other, with $E(W_i) = 0$.

Let $\sigma_N^2 = \sum_{i=1}^{n_1+n_2+n_3} \text{var}(W_i)$, and $\mu = \mu_1 + \mu_2$.

Theorem 3.2.1. *Let S_1 and S_2 be defined as above. Also let $L_N = \max(\max_{i=1}^{n_1+n_2} (c_i^{(1)} - \bar{c}^{(1)})^2, \max_{i=n_2+1}^{n_1+n_2+n_3}(c_i^{(2)} - \bar{c}^{(2)})^2)$. If the following condition holds:*

$$\lim_{\min(n_1+n_2, n_2+n_3)\to\infty} \frac{L_N}{\sigma_N^2} = 0. \tag{3.2.2}$$

then $\frac{S_1+S_2-E(S_1+S_2)}{\sigma} \longrightarrow N(0,1)$. The conclusion still holds if σ_N is replaced by $(\mathrm{var}(S_1+S_2))^{\frac{1}{2}}$. If $\max(\max_{i=1}^{n_1+n_2}(c_i^{(1)})^2, \max_{i=n_2+1}^{n_1+n_2+n_3}(c_i^{(2)})^2) \leq KL_N$ for some constant K, $E(S_1+S_2)$ can be replaced by μ.

Proof. Since, S_1 and S_2 are dependent, normality does not follow using classical results. Let $N = n_1 + n_2 + n_3$. Using Theorem 3.1.2, we can find constants $M_1 = M_1(\phi)$, $M_2 = M_2(\phi)$, such that

$$E(S_1 - ES_1 - \sum_{i=1}^{n_1+n_2} Z_i)^2 \leq \frac{M_1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} (c_i^{(1)} - \bar{c}^{(1)})^2,$$

and

$$E(S_2 - ES_2 - \sum_{i=n_1+1}^{n_1+n_2+n_3} Z_i^*)^2 \leq \frac{M_2}{n_2+n_3} \sum_{i=n_1+1}^{n_1+n_2+n_3} (c_i^{(2)} - \bar{c}^{(2)})^2.$$

Then

$$\begin{aligned}
&E(S_1+S_2 - E(S_1+S_2) - \sum_{i=1}^{n_1+n_2} Z_i - \sum_{i=n_1+1}^{n_1+n_2+n_3} Z_i^*)^2 \\
&\leq 2E(S_1 - ES_1 - \sum_{i=1}^{n_1+n_2} Z_i)^2 + 2E(S_2 - ES_2 - \sum_{i=n_1}^{n_1+n_2+n_3} Z_i^*)^2 \\
&\leq \frac{2M_1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} (c_i - \bar{c}^{(1)})^2 + \frac{2M_2}{n_2+n_3} \sum_{i=n_1+1}^{n_1+n_2+n_3} (c_i - \bar{c}^{(2)})^2 \\
&\leq 2(M_1+M_2)L_N.
\end{aligned} \tag{3.2.3}$$

Because $|Z_i| \leq 2\max_{i=1}^{n_1+n_2}|c_i-\bar{c}^{(1)}|\sup[\phi'(t)]$, $|Z_i^*| \leq 2\max_{i=n_1+1}^{n_1+n_2+n_3}|c_i-\bar{c}^{(2)}|\sup[\phi'(t)]$, then $|W_i| \leq 4\sqrt{L_N}\sup[\phi'(t)]$.

Let $\epsilon > 0$. In view of (3.2.2) and since W_i is bounded, for any $\delta > 0$, there exist a $N_\delta > 0$ such that for all $N > N_\delta$, $W_i < \delta\sigma_N$. Therefore, the Lindeberg condition holds:

$$\lim_{N\to\infty} \sigma^{-2} \sum_{i=1}^{n_1+n_2+n_3} \int_{|x|>\delta\sigma_N} x^2 dP(W_i \leq x) = 0$$

Thus by the Lindeberg theorem,

$$\sup_x |P(\sum_{i=1}^{n_1+n_2+n_3} W_i < x\sigma_N) - \Phi(x)| < \frac{1}{4}\epsilon, \tag{3.2.4}$$

where $\Phi(x)$ is the cdf of a standard normal distribution.

In view of the uniform continuity of $\Phi(x)$, there exists a $\beta > 0$, such that

$$|\Phi(x+\beta) - \Phi(x)| < \frac{1}{4}\epsilon, -\infty < x < \infty \tag{3.2.5}$$

Combining (3.2.4) and (3.2.5), we have $\sup_x |P(\sum_{i=1}^{n_1+n_2+n_3} W_i < x\sigma + \beta\sigma_N) - \Phi(x)| < \frac{1}{2}\epsilon$. Now

$$\begin{aligned}
&P(S_1+S_2-E(S_1+S_2) < x\sigma_N) \\
\leq &P(\sum_{i=1}^{n_1+n_2+n_3} W_i < x\sigma_N + \beta\sigma_N) + P(|S_1+S_2-E(S_1+S_2) - \sum_{i=1}^{n_1+n_2+n_3} W_i| > \beta\sigma_N) \\
\leq &\Phi(x) + \frac{1}{2}\epsilon + E(S_1+S_2-E(S_1+S_2) - \sum_{i=1}^{n_1+n_2+n_3} W_i)^2/(\beta^2\sigma_N^2) \\
\leq &\Phi(x) + \frac{1}{2}\epsilon + \frac{2(M_1+M_2)L_N}{\beta^2\sigma_N^2}
\end{aligned} \tag{3.2.6}$$

Using (3.2.2), we have

$$\beta^2\sigma_N^2 \geq 4\epsilon^{-1}(M_1+M_2)L_N$$

Consequently, $P(S_1+S_2-E(S_1+S_2)<x\sigma_N)\leq\Phi(x)+\epsilon$.

Similarly for the other direction, we have

$$\begin{aligned}
&P(S_1+S_2-E(S_1+S_2)<x\sigma_N)\\
\geq&P(\sum_{i=1}^{n_1+n_2+n_3}W_i<x\sigma_N-\beta\sigma_N)-P(|S_1+S_2-E(S_1+S_2)-\sum_{i=1}^{n_1+n_2+n_3}W_i|>\beta\sigma_N)\\
\geq&\Phi(x)-\frac{1}{2}\epsilon-E(S_1+S_2-E(S_1+S_2)-\sum_{i=1}^{n_1+n_2+n_3}W_i)^2/(\beta^2\sigma_N^2)\\
\geq&\Phi(x)-\epsilon
\end{aligned}\tag{3.2.7}$$

Thus, $\sup_x|P(S_1+S_2-E(S_1+S_2)<x\sigma_N)-\Phi(x)|<\epsilon$.

Note that $EW_i=0$. By Cauchy-Schwartz inequality,

$$\begin{aligned}
&(\sigma_N-\mathrm{var}(S_1+S_2)^{\frac{1}{2}})^2\\
=&\sigma_N^2+\mathrm{var}(S_1+S_2)-2\sigma\,\mathrm{var}(S_1+S_2)^{\frac{1}{2}}\\
\leq&E(\sum_{i=1}^{N}W_i^2)+E(S_1+S_2-E(S_1+S_2))^2-2E[(\sum_{i=1}^{N}W_i)(S_1+S_2-E(S_1+S_2)]\\
=&E[S_1+S_2-E(S_1+S_2)-\sum_{i=1}^{N}W_i]^2\\
\leq&2(M_1+M_2)L_N
\end{aligned}\tag{3.2.8}$$

Thus $|\sigma_N - (\mathrm{var}(S_1+S_2))^{\frac{1}{2}}| \le \sqrt{2(M_1+M_2)L_N}$. Since $\lim_N \sqrt{L_N}/\sigma_N = 0$ by the condition 3.2.2, we have $\lim_N \sigma_N/(\mathrm{var}(S_1+S_2)))^{\frac{1}{2}} = 1$. Thus σ_N can be replaced by $(\mathrm{var}(S_1+S_2))^{\frac{1}{2}}$ in the conclusion.

From Theorem 3.1.2, we have

$$(ES_1 - \mu_1)^2 \le M_1 \max_{i=1}^{n_1+n_2} (c_i^{(1)})^2$$

and

$$(ES_2 - \mu_2)^2 \le M_2 \max_{i=n_1}^{n_1+n_2+n_3} (c_i^{(2)})^2.$$

If

$$\max(\max_{i=1}^{n_1+n_2} (c_i^{(1)})^2, \max_{i=n_2+1}^{n_1+n_2+n_3} (c_i^{(2)})^2) \le KL_N$$

for some constant K, we have

$$(E(S_1+S_2) - \mu)^2 \le K(M_1+M_2)L_N.$$

Combining with (3.2.3),

$$E(S_1 + S_2 - \mu - \sum_{i=1}^{n_1+n_2+n_3} W_i)^2 \le (K+2)(M_1+M_2)L_N.$$

Thus $E(S_1+S_2)$ can be replaced by μ in the conclusion if the specified condition holds.

□

As an extension of the theorem above, we consider a more general correlation structure. Let Ω be a set consisting of N random variables $X_1, X_2, \ldots, X_N$. Let $\mathcal{A}$

be a collection of subsets in Ω, not necessarily disjoint. Assume the cardinality of $\mathcal{A}$ denoted by $\#\mathcal{A}$ is fixed. Let $R_A(X_i)$ be the rank of X_i among set A, $C_A(X_i)$ is the constant coefficient associated with X_i. Let $\overline{C_A}$ be the average of $C_A(X_i)$ over set A. Let n_A be the cardinality of set A. Also define $H_A(x) = \frac{1}{n_A}\sum_{\{i|X_i\in A\}} F_i(x)$ and $L_A(X_i) = \frac{1}{n_A}\sum_{X_j\in A}(C_A(X_j)-C_A(X_i)\int[u(y-X_i)-F_i(y)]\phi'(H_A(y))dF_j(y)$, if $X_i \in A$. $L_A(X_i) = 0$, if $X_i \notin A$. Let W_i denote $\sum_{A\in\mathcal{A}} L_A(X_i)$. Let $\mu_A = \sum_{X_i\in A} C_A(X_i)\int \phi(H_A(x))dF_i(x)$ and $\mu = \sum_{A\in\mathcal{A}}\mu_A$.

Theorem 3.2.2. *Let the composite linear rank statistic*

$$S = \sum_{A\in\mathcal{A}} S_A = \sum_{A\in\mathcal{A}}\sum_{X_i\in A} C_A(X_i)\alpha_{n_A}(R_A(X_i)).$$

Let

$$\sigma_N^2 = \sum_{i=1}^{N}\mathrm{var}(W_i).$$

If the following condition holds:

$$\lim_{\min_{A\in\mathcal{A}}(n_A)\to\infty} \frac{\sup_{A\in\mathcal{A}}\sup_{X_i\in A}(C_A(X_i)-\overline{C_A})^2}{\sigma_N^2} = 0,$$

then

$$\frac{S-ES}{\sigma_N} \longrightarrow N(0,1).$$

The conclusion still holds if σ_N is replaced by $\mathrm{var}(S)^{\frac{1}{2}}$.

If $\sup_{A\in\mathcal{A}}\sup_{X_i\in A}(C_A(X_i))^2 \le K\sup_{A\in\mathcal{A}}\sup_{X_i\in A}(C_A(X_i)-\overline{C_A})^2$ for some constant K, ES can be replaced by μ in the conclusion.

Proof. Using the projection method in each subset A respectively, then add up the projections together from all the subsets. In this way, construct independent variables W_i as functions of X_i

$$W_i = \sum_{A \in \mathcal{A}} L_A(X_i).$$

As in the proof of theorem (3.2.1), we can show $E(S - E(S) - \sum W_i)^2 \leq \#\mathcal{A} \sum_{A \in \mathcal{A}} E(S_A - E(S_A) - \sum L_A(X_i))^2$, thus $E(S - E(S) - \sum W_i)^2 \to 0$ as $n \to \infty$. This demonstrates that the difference between the centered S and sums of the Ws converges to 0 in L^2 norm. Since W_i's are bounded and $\sigma^2 \to \infty$, the Lindeberg theorem can be applied to $W_1, \ldots, W_N$. Combining the results above, we can prove the asymptotic normality of S following the same arguments in the proof of Theorem 3.2.1.

□

Comment: Based on Cramer-Wold device, the theorem above can be restated as follows: If the condition in Theorem 3.2.2 holds, the vector of simple linear rank statistics, $(S_A, \quad A \in \mathcal{A})$ has asymptotic multivariate normal distribution.

Chapter 4

Tests for Balanced Designs

4.1 Proposed new rank statistics

First consider the 2-way layout design with equal replications per cell.

$$X_{ijn} = \theta + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijn}. \quad i = 1, \ldots, I,$$
$$j = 1, \ldots, J,$$
$$n = 1, \ldots, N,$$

where ϵ_{ijn} are iid random variables with absolute continuous cdf F. Let $F_{ij}(x) = F(x - \alpha_i - \beta_j - \gamma_{ij})$ be the distribution function for X_{ijn}. The usual hypothesis to be tested is: $H_0 : \gamma_{ij} = 0, \forall$ i and j, vs $H_1 : \gamma_{ij} \neq 0$, for some i and j.

Define r_{ijn} to be the rank of X_{ijn} among the i^{th} row and c_{ijn} to be the rank of X_{ijn} among the j^{th} column.

Let $n_{i.} = \sum_{j=1}^{J} n_{ij}$ and $n_{.j} = \sum_{i=1}^{I} n_{ij}$. Define $S_N(i,j) = \sum_{n=1}^{N} \alpha_{n_{i.}}(r_{ijn})$, and $S_N = (S_N(1,1), S_N(1,2), \ldots, S_N(I,J))'$, the vector of row sums of length IJ. Similarly let $T_N(i,j) = \sum_{n=1}^{N} \alpha_{n_{.j}}(c_{ijn})$, and $T_N = (T_N(1,1), T_N(1,2), \ldots, T_N(I,J))'$, the vector of column sums of length IJ.

Let $\Sigma_1 = \lim_{N\to\infty} \frac{1}{N} \operatorname{var}(S_N)$, $\Sigma_2 = \lim_{N\to\infty} \frac{1}{N} \operatorname{var}(T_N)$ and $\Sigma_{12} = \lim_{N\to\infty} \frac{1}{N} \operatorname{cov}(S_N, T_N)$. Let $\hat{\Sigma_1}$, $\hat{\Sigma_2}$, $\hat{\Sigma_{12}}$ be corresponding consistent estimators.

Define $\tilde{S}_N(i,j) = S_N(i,j) - \frac{1}{I} \sum_{a=1}^{I} S_N(a,j)$, and

$$\tilde{S}_N = (\tilde{S}(1,1), \tilde{S}(1,2), \ldots, \tilde{S}(I,J))'.$$

Similarly define $\tilde{T}_N(i,j) = T_N(i,j) - \frac{1}{J} \sum_{b=1}^{J} T_N(i,b)$, and

$$\tilde{T}_N = (\tilde{T_N}(1,1), \tilde{T_N}(1,2), \ldots, \tilde{T}(I,J))'.$$

Let $\mathbf{I}_I$ and $\mathbf{I}_J$ be the identity matrices of dimension I and J respectively. Let $\mathbf{J}_I$ and $\mathbf{J}_J$ be matrices with all elements equal to one and dimensions equal to I and J respectively. Set $A_1 = (\frac{-1}{I})\mathbf{I}_J$, $A = \mathbf{J}_I \otimes A_1 + \mathbf{I}_I \otimes \mathbf{I}_J$.

Set $B_1 = \mathbf{I}_J + (\frac{-1}{J})\mathbf{J}_J$, $B = \mathbf{I}_I \otimes B_1$. Then we can show that $\tilde{S}_N = AS_N$, $\tilde{T}_N = BT_N$.

Our proposed test statistics are:

$$W_1 = \frac{1}{N}(AS_N)'(A\hat{\Sigma_1}A')^{-1}(AS_N) \tag{4.1.1}$$

$$W_2 = \frac{1}{N}(BT_N)(B\hat{\Sigma_2}B')^{-1}(BT_N) \tag{4.1.2}$$

$$W_3 = \frac{1}{N}(AS_N + BT_N)'(A\hat{\Sigma_1}A' + A\hat{\Sigma_{12}}B' + B\hat{\Sigma_2}B')^{-1}(AS_N + BT_N) \tag{4.1.3}$$

Since the covariance matrices are singular, generalized inverses are used in the test statistics. It is worthy to note that W_1, W_2 and W_3 have an interesting attribute. Let $a_{ijn} = r_{ijn} + c_{ijn}$,

let

$$a_{ij.} = \sum_{n=1}^{N} a_{ijn},$$

$$a_{i..} = \sum_{j=1}^{J} \sum_{n=1}^{N} a_{ijn},$$

$$a_{.j.} = \sum_{i=1}^{I} \sum_{n=1}^{N} a_{ijn},$$

$$a_{...} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{n=1}^{N} a_{ijn}.$$

Similar definitions apply to $r_{ij.}, r_{i..}, r_{.j.}, r_{...}$, and $c_{ij.}, c_{i..}, c_{.j.}, c_{...}$

Lemma 4.1.1. *If $\alpha(.)$ is the Wilcoxon score, the (i,j) element of $AS_N + BT_N$ is equal to $a_{ij.} - \frac{1}{I}a_{.j.} - \frac{1}{J}a_{i..} + \frac{1}{IJ}a_{...}$. Analogous representations exist for W_1 and W_2. The (i,j) element of AS_N is equal to $r_{ij.} - \frac{1}{I}r_{.j.} - \frac{1}{J}r_{i..} + \frac{1}{IJ}r_{...}$. The (i,j) element of BT_N is equal to $c_{ij.} - \frac{1}{I}c_{.j.} - \frac{1}{J}c_{i..} + \frac{1}{IJ}c_{...}$.*

Proof. Note that

$$\begin{aligned}
&a_{ij.} - \frac{1}{I}a_{.j.} - \frac{1}{J}a_{i..} + \frac{1}{IJ}a_{...} \\
=&r_{ij.} - \frac{1}{I}r_{.j.} - \frac{1}{J}r_{i..} + \frac{1}{IJ}r_{...} + c_{ij.} - \frac{1}{I}c_{.j.} - \frac{1}{J}c_{i..} + \frac{1}{IJ}c_{...} \\
=&r_{ij.} - \frac{1}{I}r_{.j.} - \frac{1}{J}\frac{NJ(NJ+1)}{2} + \frac{INJ(NJ+1)}{2IJ} + c_{ij.} - \frac{1}{J}c_{.j.} \\
&- \frac{1}{I}\frac{NI(NI+1)}{2} + \frac{JNI(NI+1)}{2IJ} \\
=&r_{ij.} - \frac{1}{I}r_{.j.} + c_{ij.} - \frac{1}{J}r_{i..}
\end{aligned} \tag{4.1.4}$$

where $r_{ij.} - \frac{1}{I}r_{.j.} + c_{ij.} - \frac{1}{J}r_{i..}$ is the (i,j) element of $AS_N + BT_N$. □

The representation appearing in lemma 4.1.1 bears a strong resemblance to the form of the parametric test of interaction. Specifically, the parametric test statistic is given by

$$F = \frac{(N-1)IJ\sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij.} - \frac{1}{I}X_{.j.} - \frac{1}{J}X_{i..} + \frac{1}{IJ}X_{...})^2}{(I-1)(J-1)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{N}(X_{ijn} - \frac{1}{N}X_{ij.})^2},$$

where the X_{ijn} terms in the numerator play the role of a_{ijn}.

4.2 Expectations under the null hypothesis

In order to have a central χ^2 distribution for the proposed test statistics, the vectors of the row rank sums and the column rank sums need to have zero expectation under the null hypothesis of no interaction.

Theorem 4.2.1. *Under the null hypothesis H_0 of no interaction, $E[\tilde{S_N}] = 0$ and $E[\tilde{T_N}] = 0$ for any continuous score function ϕ.*

Proof. Recall that $\tilde{S_N} = AS_N$, and $\tilde{T_N} = BT_N$, We first prove the result when ϕ is the identity function.

$$\begin{aligned} E(\alpha_{n_{i.}}(r_{ijn})) &= E(\phi(\frac{1}{n_{i.}+1}\sum_{a=1}^{J}\sum_{k=1}^{N} I(X_{iak} \le X_{ijn}))), \text{where } I \text{ is the indicator function.} \\ &= \frac{1}{n_{i.}+1}\sum_{a=1}^{J}\sum_{k=1}^{N} EI(X_{iak} \le X_{ijn}) \\ &= \frac{1}{n_{i.}+1}\sum_{a=1}^{J}\sum_{k=1}^{N} P(X_{iak} \le X_{ijn}) \\ &= \frac{1}{n_{i.}+1}\sum_{a=1}^{J} N\int F_{ia}dF_{ij} \end{aligned} \tag{4.2.1}$$

Under $H_0, F_{ia} = F(x-\alpha_i-\beta_a), F_{ij} = F(x-\alpha_i-\beta_j)$, we have

$$\begin{aligned} \int F_{ia}dF_{ij} &= \int F(x-\alpha_i-\beta_a)dF(x-\alpha_i-\beta_j) \\ &= \int F(z-\beta_a)dF(z-\beta_j) \end{aligned} \tag{4.2.2}$$

It follows that neither (4.2.2) nor (4.2.1) depends on i, since the $n_{i.}$s are the same for different i in balanced designs. Hence

$$\begin{aligned} E(\tilde{S_N}(i,j)) &= E(\sum_{n=1}^{N}\alpha_{n_{i.}}(R_{ijn}) - \frac{1}{I}\sum_{b=1}^{I}\sum_{n=1}^{N}\alpha_{n_{b.}}(R_{bjn})) \\ &= 0 \end{aligned} \tag{4.2.3}$$

Now let ϕ be the polynomial of degree l given by $\phi(x) = x^l$.

We expand $E(\phi(\frac{1}{n_{i.}+1}\sum_{a=1}^{J}\sum_{k=1}^{N} I(X_{iak} \le X_{ijn})))$.

Let Q be a subset of the product space $(1,2,\ldots,J) \times (1,2,\ldots,N)$ and d be an arbitrary constant. Note that $I^m(x) = I(x)$ for any integer m. By the conditioning argument on the X_{ijn}'s, each term in the expansion is of the form

$$
\begin{aligned}
& E[d \prod_{(a,k)\in Q} I(X_{iak} \leq X_{ijn}) \\
=& dP(\cap_{(a,k)\in Q}(X_{iak} \leq X_{ijn})) \\
=& d \int \prod_{(a,k)\in Q} F_{ia} dF_{ij},
\end{aligned}
\tag{4.2.4}
$$

where both d and the integral are independent of i. Hence we have shown that each term in the expansion of $E(\phi(\sum\sum I(X_{iak} \leq X_{ijn})/(n_{i.}+1)))$ is independent of i, and hence the result holds for any polynomial. For any arbitrary continuous function ϕ defined on [0,1], the Weierstrass theorem guarantees the existence of a sequence of polynomial $\phi_m(x)$ such that ϕ_m converges uniformly to ϕ over [0,1]. Hence $\lim_{m\to\infty} E(\phi_m(\frac{R_{ijn}}{n_{i.}+1})) = E(\phi(\frac{R_{ijn}}{n_{i.}+1}))$. Since $E(\phi_m(\frac{R_{ijn}}{n_{i.}+1}))$ is independent of i, it follows that the limit is also independent of i. Similarly we can prove $E[\tilde{T_N}] = E[BT_N] = 0$ under the null hypothesis. □

4.3 Limiting distributions under the null hypothesis

The goal for this section is to derive the limiting distributions of the proposed test statistics under the null hypothesis. In order to achieve this goal, expressions for the limiting covariance matrices need to be derived. Secondly, the asymptotic multivariate normality of the vector of row rank sums and column rank sums has to be established.

The sum $S_N(i,j)$ can be written as a linear rank statistic: $S_N(i,j) = \sum_{b=1}^{J} \sum_{n=1}^{N} d_{ibn}$ $\alpha_{n_{i.}}(r_{ibn})$, where d_{ibn} are regression constants defined as:

$$d_{ibn} = \begin{cases} 1 & \text{if } b = j, \\ 0 & \text{if } b \neq j. \end{cases}$$

Note that d_{ibn} is not a function of n. Before proceeding to the next lemma, we show that by using integration by parts, the following type of expression which frequently occurs in the later sections can be simplified, see equation (2.26) in [8].

$$\begin{aligned} &\int [u(y-x) - F_i(y)]\phi'(H_{i.}(y))dF_j(y) \\ &= \int_{-\infty}^{x} -F_i(y)\phi'(H(y))dF_j(y) + \int_{x}^{\infty} (1 - F_i(y))\phi'(H(y))dF_j(y) \\ &= \int_{-\infty}^{\infty} -F_i(y)\phi'(H(y))dF_j(y) + \int_{x}^{\infty} \phi'(H(y))dF_j(y) \\ &= \text{const} + \int_{x}^{\infty} \phi'(H(y))dF_j(y) \end{aligned} \tag{4.3.1}$$

The following lemma gives the closed expressions for the limiting variance and covariance of the row rank sums.

Lemma 4.3.1. *Let $S_N(i,j) = \sum_{n=1}^{N} r_{ijn}$ denote the row rank sum in the (i,j)th cell in a balanced two-way layout. Assume the continuity of F and the existence of a bounded second derivative of the score generating function ϕ. Also let $H_{i.}$ denote*

$1/J\sum_{j=1}^{J} F_{ij}$. *Then*

$$\begin{aligned}\sigma^2(i,j) &= \lim_{N\to\infty}\frac{1}{N}\operatorname{var}(S_N(i,j)) \\ &= \operatorname{var}(\phi(H_{i.}(X_{ij1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x))\end{aligned} \tag{4.3.2}$$

$$\begin{aligned}\sigma^2(i,j,i,j') &= \lim_{N\to\infty}\frac{1}{N}\operatorname{cov}(S_N(i,j),S_N(i,j')) \\ &= \operatorname{cov}(\phi(H_{i.}(X_{ij1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x), \\ &\quad \phi(H_{i.}(X_{ij'1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij'}(x))\end{aligned} \tag{4.3.3}$$

Proof. Using the projection method as in the proof of theorem 3.2.2, construct random variables W_{ibn} :

$$\begin{aligned}W_{ibn} &= \frac{1}{n_{i.}}\sum_{a=1}^{J}\sum_{k=1}^{N}(d_{iak}-d_{ibn})\int[u(x-X_{ibn})-F_{ib}(x)]\phi'(H_{i.}(x))dF_{ia}(x) \\ &= \frac{1}{J}\sum_{a=1}^{J}(d_{ia1}-d_{ibn})\int[u(x-X_{ibn})-F_{ib}(x)]\phi'(H_{i.}(x)dF_{ia}(x)\end{aligned} \tag{4.3.4}$$

When $b=j$,

$$W_{ijn} = -\frac{1}{J}\sum_{a\neq j}\int[u(x-X_{ijn})-F_{ij}(x)]\phi'(H_{i.}(x)dF_{ia}(x).$$

By (4.3.1),

$$\begin{aligned}\mathrm{var}(W_{ijn}) &= \mathrm{var}(-\frac{1}{J}\sum_{a\neq j}\int_{X_{ijn}}^{\infty}\phi'(H_{i.}(x))dF_{ia}(x)) \\ &= \mathrm{var}(-\frac{1}{J}\sum_{a=1}^{J}\int_{X_{ijn}}^{\infty}\phi'(H_{i.}(x))dF_{ia}(x)+\frac{1}{J}\int_{X_{ijn}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x)) \\ &= \mathrm{var}(-\int_{X_{ijn}}^{\infty}\phi'(H_{i.}(x))dH_{i.}(x)+\frac{1}{J}\int_{X_{ijn}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x)) \\ &= \mathrm{var}(\phi'(H_{i.}(X_{ijn}))+\frac{1}{J}\int_{X_{ijn}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x))\end{aligned}$$

When $b \neq j$,

$$W_{ibn} = \frac{1}{J}\int[u(x-X_{ibn})-F_{ib}(x)]\phi'(H_{i.}(x))dF_{ij}(x).$$

By (4.3.1),

$$\mathrm{var}(W_{ibn}) = \mathrm{var}(\frac{1}{J}\int_{X_{ibn}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x))$$

According to theorem 3.2.2,

$$\begin{aligned}\sigma^2(i,j) &= \lim_{N\to\infty}\frac{1}{N}\sum_{b=1}^{J}\sum_{n=1}^{N}\mathrm{var}(W_{ibn}) \\ &= \mathrm{var}(\phi(H_{i.}(X_{ij1}))+\frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij;N}(x))\end{aligned}$$

Following the same simplification, we can obtain the formula for the limiting covariance. □

Comment: When ϕ is the identity function, then the formula can be simplified to

$$\sigma^2(i,j) = \mathrm{var}(H_{i.}(X_{ij1}) - \frac{1}{J}\sum_{b=1}^{J}F_{ij}(X_{ib1})) \tag{4.3.5}$$

Analogous results hold for $T_N(i,j)$.

Lemma 4.3.2. *Let $f(x)$ denote the density function of $F(x)$. If the score generating function ϕ is a strictly monotone function and has a bounded second derivative, and $f(x)$ has only finite many points of discontinuity, then $\sigma^2(i,j) > 0$.*

Proof. According to lemma 4.3.1, we have

$$\begin{aligned}\sigma^2(i,j) = \operatorname{var}\Big(\phi(H_{i.}(X_{ij1})) + \frac{1}{J}\int_{X_{ij1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x)\Big) \\ + \frac{1}{J^2}\sum_{b\neq j}\operatorname{var}\Big(\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x)\Big).\end{aligned} \tag{4.3.6}$$

If $\sigma^2(i,j) = 0$, it follows that

$$\operatorname{var}(\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x)) = 0, \forall b \neq j. \tag{4.3.7}$$

$$\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x) = c_b, \quad \text{a.s., } c_b \text{ is a constant.} \tag{4.3.8}$$

Since f has finitely many points of discontinuity, there exist (a,b) in R, such that $\forall$ X_{ib1} in (a,b), $f_{ij}(X_{ib1}) \neq 0$. Let μ be the measure induced by the distribution function F_{ib}. Therefore, there exist $R \in \Omega$ $(X_{ib1} : \Omega \to \mathcal{R})$ and $\mu(R) = 0$, such that $\forall w \in B = (X_{ib1}^{-1}((a,b)) \cap R^c)$, $\int_{X_{ib1}}^{\infty}\phi'(H_{i.}(x))dF_{ij}(x) = c_b$. Let $G(x) = \int_x^{\infty}\phi'(H_{i.}(t))dF_{ij}(t)$. Then for any $x_1 \in B$ and $x_2 \in B$, $G(x_1) - G(x_2) = \int_{x_1}^{x_2}\phi'(H_{i.}(t))\ dF_{ij}(t) \neq 0$, since the integrand is strictly greater than zero or less than zero in set B. This contradicts the fact that $G(x)$ is a constant in set B. □

Note: The condition imposed in the lemma above is satisfied by many commonly used distributions, such as the normal, the double exponential, etc. Let $\mu_N(i,j) = N\int \phi(H_{i.}(x))dF_{ij}(x)$, and $\nu_N(i,j) = N\int \phi(H_{.j}(x))dF_{ij}(x)$.

Theorem 4.3.3. *Under the conditions stated in Lemma* 4.3.2,

$$\frac{1}{\sqrt{N}}\frac{S_N(i,j) - E[S_N(i,j)]}{\sigma(i,j)} \longrightarrow N(0,1),$$

as $N \to \infty$. *Moreover,* $E[S_N(i,j)]$ *can be replaced by* $\mu_N(i,j)$ *in the conclusion.*

Proof. In view of the lemma above, $\sigma^2(i,j) > 0$ensures $\lim_{N\to\infty} \text{var}(S_N(i,j)) \to \infty$. $S_N(i,j)$ is a composite linear rank statistic defined on set $\{X_{i11},\ldots,X_{iJN}\}$ with regression constants $d_{ian} = 1$, if $a = j$, and $d_{ian} = 0$, if $a \neq j$.

Set $\overline{d_{i..}} = \frac{1}{JN}\sum_{a=1}^{J}\sum_{n=1}^{N} d_{ian} = \frac{1}{J}$.

Thus, $\max_{a=1}^{J}(d_{ian} - \overline{d_{i..}})^2 = \frac{(J-1)^2}{J^2}$ which is uniformly bounded as $N \to \infty$. Thus according to theorem 3.2.2, $S_N(i,j)$ is asymptotically normal. Since $\max_{a=1}^{J} d_{ian}^2 = 1$ is bounded by a multiple of $\max_{a=1}^{J}(d_{ian} - \overline{d_{i..}})^2$, $E[S_N(i,j)]$ can be replaced by $\mu_N(i,j)$ in the conclusion. Similarly we can prove the asymptotic normality result for $T_N(i,j)$. Also $E[T_N(i,j)]$ can be replaced by $\nu_N(i,j)$ in the conclusion. □

Recall the expressions for the limiting variance and covariance in Lemma 4.3.1. Let vector $l = (l(1,1), l(1,2),\ldots,l(I,J))$, where

$$l(i,j) = \phi(H_{i.}(X_{ij1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty} \phi'(H_{i.}(x))dF_{ij}(x).$$

Let Σ_1 be a IJ by IJ matrix with $\Sigma_1((i-1)J+j,(i-1)J+j') = \text{cov}(l(i,j),l(i,j'))$ and $\Sigma_1((i-1)J+j,(i'-1)J+j') = 0$ for $i \neq i'$. Let vector $p = (p(1,1),p(1,2),\ldots,p(I,J))$, where

$$p(i,j) = \phi(H_{\cdot j}(X_{ij1})) + \frac{1}{I}\sum_{a=1}^{I}\int_{X_{aj1}}^{\infty}\phi'(H_{\cdot j}(x))dF_{ij}(x).$$

Let Σ_2 be a IJ by IJ matrix with $\Sigma_2((i-1)J+j,(i'-1)J+j) = \text{cov}(p(i',j),p(i',j))$ and $\Sigma_2((i-1)J+j,(i'-1)J+j') = 0$ for $j \neq j'$. Let $\mu_N = (\mu_N(i,j))$, $i = 1,\ldots,I$ and $j = 1,\ldots,J$. Let $\nu_N = (\nu_N(i,j))$, $i = 1,\ldots,I$ and $j = 1,\ldots,J$.

Theorem 4.3.4. *Under the conditions stated in Lemma* 4.3.2,

$$\frac{1}{\sqrt{N}}(S_N - E(S_N)) \longrightarrow N_{IJ}(0,\Sigma_1),$$

$$\frac{1}{\sqrt{N}}(T_N - E(T_N)) \longrightarrow N_{IJ}(0,\Sigma_2),$$

as $N \to \infty$. *Moreover* $E(S_N)$ *and* $E(T_N)$ *can be replaced by* μ_N *and* ν_N.

Proof. In line with theorem 3.2.2, set $\Omega = \{X_{111},\ldots,X_{IJN}\}$.

Let $R_i = \{X_{i11},\ldots,X_{iJN}\}$, and $\mathcal{A} = \{R_1,\ldots,R_I\}$. For any vector $\lambda = (\lambda_{ij})$ with $\lambda'\Sigma_1\lambda \geq 0$, we have that $\lambda' S_N$ is a composite linear rank statistic with $C_{R_i}(X_{ijn}) = \lambda_{ij}$. Then $\overline{\lambda}_{i\cdot} = \frac{1}{J}\sum_{a=1}^{J}\lambda_{ia} = \overline{C}_{R_i}$. To verify the condition of theorem 3.2.2, we have $\sup_{R_i}\sup_j(C_{R_i}(X_{ijn}) - \overline{C}_{R_i})^2 = \sup_i\sup_j(\lambda_{ij} - \overline{\lambda}_{i\cdot})^2$ which is uniformly bounded by

a constant. According to Theorem 3.2.2,

$$\begin{aligned}\sigma^2 &= \sum_{X_{ijn}\in\Omega} \mathrm{var}(\sum_{A\in\mathcal{A}}(L_A(X_{ijn})) \\ &= \sum_{X_{ijn}\in\Omega} \mathrm{var}(L_{R_i}(X_{ijn})),\end{aligned} \tag{4.3.9}$$

where

$$L_{R_i}(X_{ijn}) = \frac{1}{JN}\sum_{a=1}^{J}\sum_{k=1}^{N}(\lambda_{ia}-\lambda_{ij})\int[u(y-X_{ijn})-F_{ij}(y)]\phi'(H_{i.}(y))dF_{ia}(y).$$

From Lemma 4.3.1, we have $\lim_{N\to\infty}\frac{1}{N}\mathrm{var}(S_N) = \Sigma_1$. Thus

$$\lim_{N\to\infty}\frac{1}{N}\sigma^2 = \lim_{N\to\infty}\frac{1}{N}\mathrm{var}(\lambda' S_N) = \lambda'\Sigma_1\lambda \geq 0.$$

When the limit is equal to zero, the composite linear rank statistic is a constant, i.e. singular normal. When the limit is greater than zero, we have the assurance that $\sigma^2 \to \infty$, as $N \to \infty$.

Thus

$$\lim_{N\to\infty}\frac{\sup_i \sup_j(\lambda_{ij}-\overline{\lambda}_{i.})^2}{\sigma^2} = 0.$$

According to theorem 3.2.2,

$$\frac{1}{\sqrt{N}}(\lambda' S_N - E(\lambda' S_N)) \to N(0, \lambda'\Sigma_1\lambda).$$

Thus the multivariate normality of S_N follows. Similarly we can prove the multivariate normality of T_N. To show that $E(S_N)$ and $E(T_N)$ can be replaced by μ_N and ν_N, we

also have $\sup_{R_i} \sup_j (C_{R_i}(X_{ijn}))^2 = \sup_i \sup_j (\lambda_{ij})^2$ which is bounded by a multiple of $\sup_{R_i} \sup_j (C_{R_i}(X_{ijn}) - \overline{C}_{R_i})^2$.

□

Let Σ_{12} be the covariance matrix of vector l and p defined above.

Let

$$\Sigma = \begin{pmatrix} \Sigma_1, \Sigma_{12} \\ \Sigma_{12}, \Sigma_2 \end{pmatrix}.$$

Theorem 4.3.5. *Under the conditions stated in Lemma* 4.3.2,

$$\frac{1}{\sqrt{N}} \begin{pmatrix} S_N - E(S_N) \\ T_N - E(T_N) \end{pmatrix} \to N_{2IJ}(0, \Sigma), \tag{4.3.10}$$

as $N \to \infty$. *Moreover,* $E(S_N)$ *and* $E(T_N)$ *can be replaced by* μ_N *and* ν_N *in the conclusion.*

Proof. Set $\Omega = \{X_{111}, \ldots, X_{IJN}\}$. Let $R_i = \{X_{i11}, \ldots, X_{iJN}\}$, $C_i = \{X_{1j1}, \ldots, X_{IjN}\}$, and $\mathcal{A} = \{R_1, \ldots, R_I, C_1, \ldots, C_J\}$. For any vectors $\lambda = (\lambda_{ij})$ and $\beta = (\beta_{ij})$, we have $\lambda' S_N + \beta' T_N$, which is a composite linear rank statistic with $C_{R_i}(X_{ijn}) = \lambda_{ij}$, and $C_{C_j}(X_{ijn}) = \beta_{ij}$. Then $\overline{\lambda}_{i\cdot} = \frac{1}{J}\sum_{a=1}^{J} \lambda_{ia} = \overline{C}_{R_i}$, and $\overline{\beta}_{\cdot j} = \frac{1}{I}\sum_{b=1}^{I} \lambda_{bj} = \overline{C}_{C_j}$. To verify the condition of theorem 3.2.2, we have $\sup_{A\in\mathcal{A}} \sup_{X_{ijn}\in A} (C_A(X_{ijn}) - \overline{C}_A)^2 = \sup_i \sup_j \sup((\lambda_{ij} - \overline{\lambda}_{i\cdot})^2, (\beta_{ij} - \overline{\beta}_{\cdot j})^2)$ is uniformly bounded by a constant. Also we have $\sup_{A\in\mathcal{A}} \sup_{X_{ijn}\in A} (C_A(X_{ijn}))^2 = \sup_i \sup_j \sup((\lambda_{ij})^2, (\beta_{ij})^2)$ is bounded by a multiple of $\sup_{A\in\mathcal{A}} \sup_{X_{ijn}\in A} (C_A(X_{ijn}) - \overline{C}_A)^2$.

According to theorem 3.2.2,

$$\begin{aligned}\sigma^2 &= \sum_{X_{ijn}\in\Omega} \mathrm{var}(\sum_{A\in\mathcal{A}}(L_A(X_{ijn})) \\ &= \sum_{X_{ijn}\in\Omega} \mathrm{var}(L_{R_i}(X_{ijn}) + L_{C_j}(X_{ijn})),\end{aligned} \tag{4.3.11}$$

where

$$L_{R_i}(X_{ijn}) = \frac{1}{JN}\sum_{a=1}^{J}\sum_{k=1}^{N}(\lambda_{ia}-\lambda_{ij})\int[u(y-X_{ijn})-F_{ij}(y)]\phi'(H_{i.}(y))dF_{ia}(y),$$

$$L_{C_j}(X_{ijn}) = \frac{1}{IN}\sum_{b=1}^{J}\sum_{k=1}^{N}(\beta_{bj}-\beta_{ij})\int[u(y-X_{ijn})-F_{ij}(y)]\phi'(H_{.j}(y))dF_{bj}(y).$$

It follows that

$$\lim_{N\to\infty}\frac{1}{N}\sigma^2 = (\lambda',\beta')\begin{pmatrix}\Sigma_1, \Sigma_{12} \\ \Sigma_{12}, \Sigma_2\end{pmatrix}\begin{pmatrix}\lambda \\ \beta\end{pmatrix} \geq 0$$

When the limit is equal to zero, the composite linear rank statistic is a constant, i.e. singular normal. When the limit is greater than zero, we have the assurance that $\sigma^2 \to \infty$, as $N \to \infty$.

Thus

$$\lim_{N\to\infty}\frac{\sup_i \sup_j \sup((\lambda_{ij}-\overline{\lambda}_{i.})^2,(\beta_{ij}-\overline{\beta}_{.j})^2)}{\sigma^2} = 0.$$

Thus the multivariate normality of (S_N, T_N) follows. □

Theorem 4.3.6. *Let $\alpha(.)$ be the Wilcoxon score. Let f_{ij} be the probability density function of random variable X_{ijn}. Assume there exists a common support (a,b), such that f_{ij} is strictly positive for $1 \leq i \leq I, 1 \leq j \leq J$. Then under the conditions of*

Lemma 4.3.2 *and under the null hypothesis of no interaction, each of* W_1, W_2, W_3 *follows a central chisquare distribution with* $(I-1)(J-1)$ *degrees of freedom asymptotically.*

Before proceeding to the proof of Theorem 4.3.6, we prove two lemmas.

Lemma 4.3.7. *Let* v *be an* N*-dimensional random vector with covariance matrix* Σ*. The necessary and sufficient condition for the rank of* Σ *to be* $N-r$ *is that there exist exactly* r *linearly independent nonzero vectors* λ_i *with* $1 \leq i \leq r$*, such that* $\lambda_i' v = c$*, a.s., where* c *denotes a constant.*

Proof. If the rank of Σ is $N-r$, then the dimension of the null space is r. If λ_i is a basis vector in the null space of Σ, $\Sigma\lambda_i = 0$ and we have $\text{cov}(\lambda_i' v) = \lambda_i' \Sigma \lambda_i = 0$. This leads to $\lambda_i' v = c$, *a.s.*.

If there exist exactly r linearly independent nonzero vectors λ_i with $1 \leq i \leq r$, such that $\lambda_i' v = c$, *a.s.*, we have $\text{cov}(\lambda' v) = \lambda_i' \Sigma \lambda_i = 0$. Since Σ is a positive semi definite matrix, we may write $0 = \lambda_i' \Sigma \lambda_i = \lambda_i' (\Sigma^{\frac{1}{2}})(\Sigma^{\frac{1}{2}})\lambda_i$. Hence $\lambda_i' \Sigma^{\frac{1}{2}} = 0$. Since the null space of $\Sigma^{\frac{1}{2}}$ is the same as the null space of Σ, then the rank of Σ is $N-r$. □

Lemma 4.3.8. *Let* X_i *be independent random variables with distribution function* F_i*,* $1 \leq i \leq N$*. Assume* F_i *is absolutely continuous with probability density function* f_i*. Assume there exists a common support* (a, b)*, such that* f_i *is strictly positive for* $1 \leq i \leq N$*. Let* $r = (R(X_1), \ldots, R(X_N))$ *be the vector of ranks. If there exists a*

nonzero vector $\lambda = (\lambda_1, \ldots, \lambda_N)$ *such that* $\lambda' r = c$ *almost surely, then* $\lambda_i = \lambda_j$ *for all* $i \neq j$.

Proof. We first take $r_1 = (1, 2, 3, \ldots, N)$. We have $P(r_1) = P(X_1 < X_2 < \ldots, < X_N) > P(a < X_1 < X_2 < \ldots, < X_N < b) > 0$, since f_i is strictly positive on (a, b) for $1 \leq i \leq N$. Because $\lambda' r_1 = c$ almost surely and $P(r_1) > 0$, we have $\lambda' r_1 = \lambda_1 + 2\lambda_2 + 3\lambda_3 + \cdots + N\lambda_N = c$.

Next we take $r_2 = (2, 1, 3, \ldots, N)$. Similarly we have $P(r_2) > 0$. Thus we have $\lambda' r_2 = 2\lambda_1 + \lambda_2 + 3\lambda_3 + \cdots + N\lambda_N = c$. Combining the result above, we have $\lambda_1 = \lambda_2$. Repeating the same argument, we can prove $\lambda_i = \lambda_j$ for all $i \neq j$. □

The following is the proof for Theorem 4.3.6.

Proof. According to theorem 4.2.1, $E(AS_N) = 0$ and $E(BT_N) = 0$ under the null hypothesis of no interaction. Theorem 4.3.5 has established the asymptotic joint multivariate normality of S_N and T_N. Let $\hat{\Sigma_1}$, $\hat{\Sigma_2}$ and $\hat{\Sigma_{12}}$ be consistent estimates of Σ_1, Σ_2 and Σ_{12}. It follows that W_1, W_2, W_3 each follows a central chisquare distribution asymptotically.

Next we need to determine the degrees of freedom for each statistic. The degrees of freedom of the statistic W_1 is equal to the rank of the matrix $A\Sigma_1 A'$. Since $\tilde{S}_N = AS_N$, we have $\text{cov}(\tilde{S}_N) = A\Sigma_1 A'$. According to Lemma 4.3.7, to determine the rank of $A\Sigma_1 A'$, it is equivalent to finding the number of linearly independent constraints for $\tilde{S}_N$. There exist at least $I + J - 1$ linearly independent constraints.

For $j = 1, \ldots, J$, define vector $w_j = (w_{ib}, 1 \leq i \leq I, 1 \leq b \leq J | w_{ib} = 1, \text{for} \quad b = j; w_{ib} = 0, \quad o.w.)$. Thus $w_j' \tilde{S}_N = \sum_{i=1}^{I} \tilde{S}_N(i,j) = \sum_{i=1}^{I} (S_N(i,j) - \frac{1}{I} \sum_{a=1}^{I} S_N(a,j)) = \sum_{i=1}^{I} S_N(i,j) - \sum_{a=1}^{I} S_N(a,j) = 0$.

For $i = 1, \ldots, I$, define vector $z_i = (z_{aj}, 1 \leq a \leq I, 1 \leq j \leq J | z_{aj} = 1, \text{for} \quad a = i; z_{aj} = 0, \quad o.w.)$. Thus $z_i' \tilde{S}_N = \sum_{j=1}^{J} \tilde{S}_N(i,j) = \sum_{j=1}^{J} (S_N(i,j) - \frac{1}{I} \sum_{a=1}^{I} S_N(a,j)) = \sum_{j=1}^{J} S_N(i,j) - \frac{1}{I} \sum_{a=1}^{I} (\sum_{j=1}^{J} S_N(a,j)) = \frac{JN(JN+1)}{2} - \frac{JN(JN+1)}{2} = 0$. Note that there is one redundancy that the last constraint z_I can be written as a linear combination of the other w_js and z_is. Hence the total number of the constraints is $I + J - 1$.

To show that there are only $I + J - 1$ linearly independent constraints, we first assume that there exists a vector $(\lambda_{ij}, 1 \leq I, 1 \leq J)$, which is not a linear combination of the w_js and z_is, such that $\lambda' \tilde{S}_N = \sum_{i,j} \lambda_{ij} \tilde{S}_N(i,j) = c$ almost surely. After reformulating the equality in terms of $S_N(i,j)$, we have $\sum_{i,j} (\lambda_{ij} - \overline{\lambda}_{.j}) S_N(i,j) = c$ almost surely. Let λ_{ij}^* denote $\lambda_{ij} - \overline{\lambda}_{.j}$. Since $(S_N(i,j), 1 \leq j \leq J)$ is independent of $(S_N(i',j), 1 \leq j \leq J)$, we have $\sum_j \lambda_{ij}^* S_N(i,j) = c$ almost surely for each i. Since for each i, $(S_N(i,j), 1 \leq j \leq J)$ is a vector of row ranks of observations $X_{i11}, \ldots, X_{iJN}$, then according to Lemma 4.3.8, in order to have the sum equal to a constant almost surely, $(\lambda_{ij}^*, 1 \leq j \leq J)$ has to take the form $(a_i, a_i, \ldots, a_i)$ for some constant a_i. Thus we have $\lambda_{ij}^* = a_i$. Since λ_{ij}^* denotes $\lambda_{ij} - \overline{\lambda}_{.j}$, we have $\lambda_{ij} = a_i + \overline{\lambda}_{.j}$, for all i, j. Thus we can formulate the vector λ in terms of linear combinations of the constraints already listed above. We have $\lambda = \sum_i a_i z_i + \sum_j \overline{\lambda}_{.j} w_j$, which contradicts the fact that λ is

linearly independent of z_i and w_j.

Thus the null set of $A\Sigma_1 A'$ is spanned by these $I + J - 1$ linear combinations. Therefore, $A\Sigma_1 A'$ has rank $IJ - (I + J - 1) = (I - 1)(J - 1)$. Similarly, we can demonstrate that $B\Sigma_2 B'$ has rank $(I - 1)(J - 1)$. Since $AS_N + BT_N = \tilde{S}_N + \tilde{T}_N$, combining the constraints on $\tilde{S}_N$ and $\tilde{T}_N$, we also have $I + J - 1$ linearly independent constraints on $AS_N + BT_N$. Thus the rank of $A\Sigma_1 A' + B\Sigma_2 B' + 2A\Sigma_{12} B'$ is $(I-1)(J-1)$.

□

We observe that our test statistics are invariant with respect to the choices of the general inverses. Also note that regardless of the existence of the main effects, our method is valid for the testing of interaction effect.

4.4 Limiting distributions under the alternatives and Asymptotic Relative Efficiency

In this section, we derive asymptotic distributions for the W statistics under Pitman alternatives. Let $\gamma = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{IJ})$ be a vector of $I \times J$ elements. Define the sequence of Pitman alternatives under investigation by $F_{ij;N}(x) = F(x - \theta - \alpha_i - \beta_j - \gamma_{ij}/\sqrt{N})$. Let the cdf of X_{ijn} under H_0 be $F_{ij;0}(x) = F(x - \theta - \alpha_i - \beta_j)$. Define $H_{i;N}(x) = \frac{1}{J}\sum_{j=1}^{J} F_{ij;N}(x)$, and $H_{i;0}(x) = \frac{1}{J}\sum_{j=1}^{J} F_{ij;0}(x)$. Also define $H_{j;N}(x) = \frac{1}{I}\sum_{i=1}^{I} F_{ij;N}(x)$, and $H_{j;0}(x) = \frac{1}{I}\sum_{i=1}^{I} F_{ij;0}(x)$.

Define the vector l_N of IJ elements of $l_N(i,j)$, where

$$l_N(i,j) = \phi(H_{i;N}(X_{ij1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty}\phi'(H_{i;N}(x))dF_{ij;N}(x).$$

Also define the vector p_N of IJ elements of $p_N(i,j)$, where

$$p_N(i,j) = \phi(H_{j;N}(X_{ij1})) + \frac{1}{I}\sum_{a=1}^{I}\int_{X_{aj1}}^{\infty}\phi'(H_{j;N}(x))dF_{ij;N}(x).$$

Set $\Sigma_{1,\gamma} = \lim_{N\to\infty}\mathrm{var}(l_N)$, $\Sigma_{2,\gamma} = \lim_{N\to\infty}\mathrm{var}(p_N)$, $\Sigma_{12,\gamma} = \lim_{N\to\infty}\mathrm{cov}(l_N, p_N)$, where the limit is taken under the Pitman alternatives.

Set vector l of elements $l(i,j)$,

$$l(i,j) = \phi(H_{i;0}(X_{ij1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty}\phi'(H_{i;0}(x))dF_{ij;0}(x).$$

Also set vector p of elements of $p(i,j)$,

$$p(i,j) = \phi(H_{j;0}(X_{ij1})) + \frac{1}{I}\sum_{a=1}^{I}\int_{X_{aj1}}^{\infty}\phi'(H_{j;0}(x))dF_{ij;0}(x).$$

According to the generalized dominated convergence theorem, $\Sigma_{1,\gamma} = \mathrm{var}(l) = \Sigma_1$, $\Sigma_{2,\gamma} = \mathrm{var}(p) = \Sigma_2$, $\Sigma_{12,\gamma} = \mathrm{cov}(l,p) = \Sigma_{12}$.

Let $\gamma_{iN} = (\gamma_{i1}/\sqrt{N}, \gamma_{i2}/\sqrt{N}, \ldots, \gamma_{iJ}/\sqrt{N})$ and $\gamma_{jN} = (\gamma_{1j}/\sqrt{N}, \gamma_{2j}/\sqrt{N}, \ldots, \gamma_{Ij}/\sqrt{N})$. Define $\mu_N(i,j;\gamma_{iN}) = N\int\phi(H_{i;N}(x))\,dF_{ij;N}(x)$ and $\mu_N(i,j;0) = N\int\phi(H_{i;0}(x))\,dF_{ij;0}(x)$. Also define $\nu_N(i,j;\gamma_{jN}) = N\int\phi(H_{j;N}(x))\,dF_{ij;N}(x)$ and $\nu_N(i,j;0) = N\int\phi(H_{j;0}(x))\,dF_{ij;0}(x)$

Let $F_i(x)$ denote $F(x-\theta-\alpha_i)$.

$$\begin{aligned}\mu_N(i,j;\gamma_{iN}) &= N\int \phi(H_{i;N}(x))dF_{ij;N}(x)\\ &= N\int \phi(\frac{1}{J}\sum_{b=1}^{J} F_{ib;N}(x))dF_{ij;N}(x)\\ &= N\int \phi(\frac{1}{J}\sum_{b=1}^{J} F_i(x-\beta_b-\frac{\gamma_{ib}}{\sqrt{N}}))dF_i(x-\beta_j-\frac{\gamma_{ij}}{\sqrt{N}})\\ &= N\int \phi(\frac{1}{J}\sum_{b=1}^{J} F_i(\mu+\beta_j-\beta_b+\frac{\gamma_{ij}}{\sqrt{N}}-\frac{\gamma_{ib}}{\sqrt{N}}))dF_i(\mu)\end{aligned}$$

When $a \neq j$,

$$\begin{aligned}\frac{\partial\mu_N(i,j;\gamma_{iN})}{\partial\gamma_{ia}} =& -\sqrt{N}\int \phi'(\frac{1}{J}\sum_{b=1}^{J} F_i(\mu+\beta_j-\beta_b+\frac{\gamma_{ij}}{\sqrt{N}}-\frac{\gamma_{ib}}{\sqrt{N}}))\\ &\frac{1}{J}F_i'(\mu+\beta_j-\beta_a+\frac{\gamma_{ij}}{\sqrt{N}}-\frac{\gamma_{ia}}{\sqrt{N}})dF_i(\mu).\end{aligned}$$

Hence,

$$\begin{aligned}G_{ij;a} &\equiv \lim_{N\to\infty}\frac{1}{\sqrt{N}}\frac{\partial\mu_N(i,j;\gamma_{iN})}{\partial\gamma_{ia}}\\ &= -\frac{1}{J}\int \phi'(\frac{1}{J}\sum_{b=1}^{J} F_i(\mu+\beta_j-\beta_b))F_i'(\mu+\beta_j-\beta_a)dF_i(\mu).\end{aligned} \tag{4.4.1}$$

$$\begin{aligned}\frac{\partial\mu_N(i,j;\gamma_{iN})}{\partial\gamma_{ij}} =&\sqrt{N}\int \phi'(\frac{1}{J}\sum_{b=1}^{J} F_i(\mu+\beta_j-\beta_b+\frac{\gamma_{ij}}{\sqrt{N}}-\frac{\gamma_{ib}}{\sqrt{N}}))\\ &\frac{1}{J}\sum_{b\neq j} F_i'(\mu+\beta_j-\beta_b+\frac{\gamma_{ij}}{\sqrt{N}}-\frac{\gamma_{ib}}{\sqrt{N}})dF_i(\mu)\end{aligned}$$

$$\begin{aligned}G_{ij;j} &\equiv \lim_{N\to\infty}\frac{1}{\sqrt{N}}\frac{\partial\mu_N(i,j;\gamma_{iN})}{\partial\gamma_{ij}}\\ &= \frac{1}{J}\int \phi'(\frac{1}{J}\sum_{b=1}^{J} F_i(\mu+\beta_j-\beta_b))\sum_{b\neq j} F_i'(\mu+\beta_j-\beta_b)dF_i(\mu).\end{aligned} \tag{4.4.2}$$

Using the multivariate Mean Value Theorem and adopting Thompson's approach [22], we have

$$\lim_{N\to\infty}\frac{1}{\sqrt{N}}(\mu_N(i,j;\gamma_{iN})-\mu_N(i,j;0))=\sum_{b=1}^{J}\gamma_{ib}\lim_{N\to\infty}\frac{1}{\sqrt{N}}\frac{\partial\mu_N(i,j;\gamma_{iN})}{\partial\gamma_{ib}}|_{\psi\gamma_{iN}} \quad (4.4.3)$$

where $0\le\psi\le 1$.

Let D be the IJ vector with element $D_{(i-1)J+j}=\sum_{b=1}^{J}\gamma_{ib}G_{ij;b}$. Let

$$\Delta_1=(AD)'(A\Sigma_1A')^{-1}(AD).$$

For the column method, let $F_j(x)$ denote $F(x-\theta-\beta_j)$. Set

$$\begin{aligned}\nu_N(i,j;\gamma_{iN}) &= N\int\phi(H_{j;N}(x))dF_{ij;N}(x)\\ &= N\int\phi(\frac{1}{I}\sum_{b=1}^{I}F_j(\mu+\alpha_i-\alpha_b+\frac{\gamma_{ij}}{\sqrt{N}}-\frac{\gamma_{bj}}{\sqrt{N}}))dF_j(\mu)\end{aligned}$$

When $a\neq i$,

$$\begin{aligned}K_{ij;a}&\equiv\lim_{N\to\infty}\frac{1}{\sqrt{N}}\frac{\partial\nu_N(i,j)}{\partial\gamma_{ia}}\\ &=-\frac{1}{I}\int\phi'(\frac{1}{I}\sum_{b=1}^{I}F_j(\mu+\alpha_i-\alpha_b))F_j'(\mu+\alpha_i-\alpha_a)dF_j(\mu).\end{aligned} \quad (4.4.4)$$

$$\begin{aligned}K_{ij;j}&\equiv\lim_{N\to\infty}\frac{1}{\sqrt{N}}\frac{\partial\nu_N(i,j)}{\partial\gamma_{ij}}\\ &=\frac{1}{I}\int\phi'(\frac{1}{I}\sum_{b=1}^{I}F_j(\mu+\alpha_i-\alpha_b)\sum_{b\neq i}F_j'(\mu+\alpha_i-\alpha_b)dF_j(\mu)\end{aligned} \quad (4.4.5)$$

Let E be the IJ vector with element $E_{(i-1)J+j}=\sum_{b=1}^{I}\gamma_{ib}K_{ij;b}$. Let

$$\Delta_2=(BE)'(B\Sigma_2B')^{-1}(BE),$$

and

$$\Delta_3 = (AD + BE)'(A\Sigma_1 A' + B\Sigma_2 B' + A\Sigma_{12} B')^{-1}(AD + BE).$$

Theorem 4.4.1. *Let $\alpha(.)$ be the Wilcoxon score. Let f_{ij} be the probability density function of random variable X_{ijn}. Assume there exists a common support (a, b), such that f_{ij} is strictly positive for $1 \leq i \leq I, 1 \leq j \leq J$. Then under the conditions of Lemma 4.3.2 and under the sequence of Pitman alternatives $(\gamma_{ij;N})$ defined above,*

$$W_1 = \frac{1}{N}(AS_N)'(A\hat{\Sigma}_1 A')^{-1}(AS_N) \longrightarrow \chi^2_{(I-1)(J-1)}(\Delta_1).$$

$$W_2 = \frac{1}{N}(BT_N)'(B\hat{\Sigma}_2 B')^{-1}(BT_N) \longrightarrow \chi^2_{(I-1)(J-1)}(\Delta_2).$$

$$W_3 = \frac{1}{N}(AS_N + BT_N)'(A\hat{\Sigma}_1 A' + B\hat{\Sigma}_2 B' + A\hat{\Sigma}_{12} B')^{-1}(AS_N + BT_N)$$
$$\longrightarrow \chi^2_{(I-1)(J-1)}(\Delta_3).$$

Proof. First we can prove that $A\mu_N(i, j; 0) = 0$, following the same argument as the proof of theorem 4.2.1. Let $\mu_N(\gamma_N)$ stand for the vector $(\mu_N(i, j; \gamma_{iN}))$, then we can rewrite W_1 as follows:

$$W_1 = \frac{1}{N}(A(S_N - \mu_N(\gamma_N) + \mu_N(\gamma_N) - \mu_N(0)))'(A\Sigma_1 A')^{-1}A(S_N - \mu_N(\gamma_N) + \mu_N(\gamma_N) - \mu_N(0))$$

In view of Theorem 4.3.4, $\frac{1}{\sqrt{N}}A(S_N - \mu_N(\gamma_N)) \longrightarrow N(0, A\Sigma_1 A')$, and the rank of $A\Sigma_1 A'$ is $(I-1)(J-1)$. According to Rao and Mitra [20], we have $W_1 \longrightarrow \chi^2_{(I-1)(J-1)}(\Delta_1)$. Similarly we can prove the results for W_2 and W_3. □

The asymptotic relative efficiency of the W's relative to the corresponding parametric F statistic are the ratios of the noncentrality parameters (Puri and Sen, [19]).

When the row and column effects are both absent, ϕ is the identity function, and $\sum_j \gamma_{ij} = 0$ and $\sum_i \gamma_{ij} = 0$, we have the following simplified expressions:

$$D_{ij} = E_{ij} = \gamma_{ij} \int (F'(\mu))^2 d\mu \tag{4.4.6}$$

Assume no main effects, i.e. $F_{ij;0} = F$ for all i, j, and let ϕ be the identity function. Let U_{ijn} denote $F(X_{ijn})$, where U_{ijn}s are iid random variables following the uniform distribution $U(0,1)$. Thus, we can simplify the expression for $l(i,j)$,

$$\begin{aligned} l(i,j) &= +\phi(H_{i;0}(X_{ij1})) + \frac{1}{J}\sum_{b=1}^{J}\int_{X_{ib1}}^{\infty} \phi'(H_{i;0}(x))dF_{ij;0}(x) \\ &= -\frac{1}{J}\sum_{b\neq J} U_{ib1} + \frac{J-1}{J}U_{ij1} \end{aligned}$$

Also we can simplify $p(i,j)$,

$$\begin{aligned} p(i,j) &= +\phi(H_{j;0}(X_{ij1})) + \frac{1}{I}\sum_{a=1}^{I}\int_{X_{aj1}}^{\infty} \phi'(H_{j;0}(x))dF_{ij;0}(x) \\ &= -\frac{1}{I}\sum_{b\neq i} U_{bj1} + \frac{I-1}{I}U_{ij1} \end{aligned}$$

Thus we have $\mathrm{var}(l(i,j)) = \frac{1}{12}\frac{J-1}{J}$ and $\mathrm{cov}(l(i,j), l(i,j')) = -\frac{1}{12}\frac{1}{J}$. Similarly $\mathrm{var}(p(i,j)) = \frac{1}{12}\frac{I-1}{I}$ and $\mathrm{cov}(p(i,j), p(i',j)) = -\frac{1}{12}\frac{1}{I}$. We also have $\mathrm{cov}(l(i,j), p(i,j)) = \frac{1}{12}\frac{(I-1)(J-1)}{IJ}$, $\mathrm{cov}(l(i,j), p(i',j)) = -\frac{1}{12}\frac{(J-1)}{IJ}$, $\mathrm{cov}(l(i,j), p(i,j')) = -\frac{1}{12}\frac{(I-1)}{IJ}$ and cov

$(l(i,j),p(i',j')) = \frac{1}{12}\frac{1}{IJ}$. Therefore we can explicitly calculate the entries for Σ_1, Σ_2, Σ_{12}. For the classical F statistic, $\Delta_F = \sum_i \sum_j \gamma_{ij}^2/v$, where $v = \text{var}(\epsilon)$ stands for the common variance of the i.i.d errors.

The simplified expressions for the ARE's are as follows:

$$ARE(W_1,F) = v^2(\int [F'(x)]^2 dx)^2 \frac{\gamma' A'(A\Sigma_1 A')^{-1} A\gamma}{\gamma'\gamma},$$

$$ARE(W_2,F) = v^2(\int [F'(x)]^2 dx)^2 \frac{\gamma' B'(B\Sigma_2 B')^{-1} B\gamma}{\gamma'\gamma},$$

$$ARE(W_3,F) = v^2(\int [F'(x)]^2 dx)^2 \frac{\gamma'(A+B)'(A\Sigma_1 A' + B\Sigma_2 B' + A\Sigma_{12}B')^{-1}(A+B)\gamma}{\gamma'\gamma}.$$

It is interesting to note that $12v^2(\int [F'(x)]^2 dx)^2$ is the well-known ARE of the Kruskal-Wallis test over the F test, which is equal to $\frac{3}{\pi}$ in the standard normal case.

4.5 Estimation of the covariance structures

This section is devoted to developing consistent estimators for the covariance structures of the composite linear rank statistics we employ in the tests. For arbitrary linear rank statistics, the same method can be generalized to estimate the corresponding limiting variance and covariance matrix.

Let $W_{ibn}^{(i,j)}$, $b = 1,\ldots,J$, $n = 1,\ldots,N$ be the independent random variables constructed by the projection method on the linear rank statistic $S_N(i,j)$ (Theorem 3.2.2).

When the scoring function ϕ is the identity function, the formula for the limiting variance can be written as

$$\begin{aligned}\sigma^2(i,j) &= \lim_{N\longrightarrow\infty}\frac{1}{N}\operatorname{var}(S_N(i,j)) \\ &= \lim_{N\longrightarrow\infty}\frac{1}{N}\operatorname{var}(\sum_{b=1}^{J}\sum_{n=1}^{N}W_{ibn}^{(i,j)}) \\ &= \lim_{N\longrightarrow\infty}(\sum_{b=1}^{J}\operatorname{var}W_{ib1}^{(i,j)}) \\ &= \frac{1}{J^2}\{\operatorname{var}(\sum_{b\neq j}F_{ib}(X_{ij1})) + \sum_{b\neq j}\operatorname{var}(-F_{ij}(X_{ib1}))\}\end{aligned} \tag{4.5.1}$$

Let $\Psi_{ij}^{(i,j)} = \sum_{b\neq j}F_{ib}(X_{ij1})$ and $\Psi_{ib}^{(i,j)} = -F_{ij}(X_{ib1})$ for $b \neq j$. Then due to the independence of X_{ib1}, for $b = 1,\ldots,J$,

$$\sigma^2(i,j) = \frac{1}{J^2}\sum_{b=1}^{J}\operatorname{var}(\Psi_{ib}^{(i,j)}) \tag{4.5.2}$$

Define

$$C_{ijn}^{(i,j)} = \frac{1}{N}\sum_{b\neq j}\sum_{n'=1}^{N}u(X_{ijn} - X_{ibn'})$$

and for $b \neq j$,

$$C_{ibn}^{(i,j)} = \frac{1}{N}\sum_{n'=1}^{N}u(X_{ibn} - X_{ijn'}).$$

We propose to use $\hat{\sigma}^2(i,j) = \frac{1}{J^2}\sum_{b=1}^{J}\frac{1}{N}\sum_{n=1}^{N}(C_{ibn}^{(i,j)} - \overline{C_{ib.}^{(i,j)}})^2$ as the consistent estimator for the limiting variance of $\frac{1}{N}S_N(i,j)$.

We also define

$$\sigma^2(i,j,i',j') = \lim_{N\to\infty}\frac{1}{N}\operatorname{cov}(S_N(i,j), S_N(i',j')).$$

When $i \neq i'$, $\sigma^2(i,j,i',j') = 0$. We only need to consider the case when $i = i'$. Let $Z_{ibn}^{(i,j')}$, $C_{ibn}^{(i,j')}$, $\overline{C_{ib.}^{(i,j')}}$, $\Psi_{ib}^{(i,j')}$ for $b = 1, \ldots, J$ and $n = 1, \ldots, N$ be defined with regard to $S_N(i,j')$.

$$\begin{aligned}\sigma^2(i,j,i,j') &= \lim_{N\to\infty} \frac{1}{N} \operatorname{cov}(\sum_{b=1}^{J}\sum_{n=1}^{N} W_{ibn}^{(i,j)}, W_{ibn}^{(i,j')}) \\ &= \sum_{b=1}^{J} \operatorname{cov}(W_{ib1}^{(i,j)}, W_{ib1}^{(i,j')}) \\ &= \sum_{b=1}^{J} \operatorname{cov}(\Psi_{ib}^{(i,j)}, \Psi_{ib}^{(i,j')}).\end{aligned}$$

We propose to use the following estimator for the limiting covariance.

$$\hat{\sigma}^2(i,j,i,j') = \frac{1}{J^2}\sum_{b=1}^{J}\frac{1}{N}\sum_{n=1}^{N}(C_{ibn}^{(i,j)} - \overline{C_{ib.}^{(i,j)}})(C_{ibn}^{(i,j')} - \overline{C_{ib.}^{(i,j')}}).$$

Later in this section, let C_{ibn}, Ψ_{ib}, C_{ibn}^*, Ψ_{ib}^* stand for $C_{ibn}^{(i,j)}$, $\Psi_{ib}^{(i,j)}$, $C_{ibn}^{(i,j')}$, $\Psi_{ib}^{(i,j')}$ respectively for simplicity in notation.

Lemma 4.5.1. *Let the random variables C_{ijn} and Ψ_{ij} be defined as above. Assuming the continuity of the cumulative distribution function F, it follows that*

$$E(|C_{ijn} - \Psi_{ij}|) = O(N^{-1})$$

$$E(|C_{ijn}^2 - \Psi_{ij}^2|) = O(N^{-1})$$

Proof. Let $X_1, \ldots, X_N$ be independent random variables with absolute continuous distribution functions $F_1, \ldots, F_N$ respectively. Let $R(X_i)$ be the rank of X_i among

the set of $X_1, \ldots, X_N$ and $H(X_i)$ be the average distribution function $\frac{1}{N}\sum_{b=1}^{N} F_b(X_i)$. According to theorem 4.2 of Hájek [8],

$$E|\frac{R(X_i)}{N+1} - H(X_i)| = O(N^{-1}). \tag{4.5.3}$$

In order to prove the convergence rate of $E|C_{ijn} - \Psi_{ij}|$, we need to link C_{ijn} with the rank of X_{ijn} among a certain set and link Ψ_{ij} with the average distribution function in the corresponding set.

Consider the set $A = \{X_{ijn}\} \cup \{\cup_{b\neq j}\{X_{ib1}, \ldots, X_{ibn}\}\}$. Let $R_A(X_i)$ denote the rank of X_{ijn} in set A and H_A denote the average distribution function among this set, and we get:

$$\begin{aligned}
&E|C_{ijn} - \sum_{b\neq j} F_{ib}| \\
&= E|\frac{\sum_{b\neq j}\sum_{n'=1}^{N} u(X_{ijn} - X_{ibn'})}{N} - \frac{\sum_{b\neq j}\sum_{n'=1}^{N} F_{ib}}{N}| \\
&= (J-1)E|\frac{1 + \sum_{b\neq j}\sum_{n'=1}^{N} u(X_{ijn} - X_{ibn'})}{(J-1)N+1} - \frac{\sum_{b\neq j}\sum_{n'=1}^{N} F_{ib} + F_{ij}}{(J-1)N+1}| + O(N^{-1}) \\
&= (J-1)E|\frac{R_A(X_i)}{N+1} - H_A(X_i)| + O(N^{-1}) = O(N^{-1})
\end{aligned} \tag{4.5.4}$$

Therefore, we can prove the convergence rate of $E|C_{ijn} - \Psi_{ij}| = O(N^{-1})$ as stated in the lemma. We can prove the convergence rate for the squared terms also, utilizing the fact that

$$|C_{ijn}^2 - \Psi_{ij}^2| = |C_{ijn} - \Psi_{ij}||C_{ijn} + \Psi_{ij}| \leq 2(J-1)|C_{ijn} - \Psi_{ij}|.$$

Thus

$$E(|C_{ijn}^2 - \Psi_{ij}^2|) = O(N^{-1})$$

□

Define $C_{ij.} = \sum_{n=1}^{N} C_{ijn}$ and $\overline{C_{ij.}} = \frac{1}{N} C_{ij.}$.

Lemma 4.5.2. *Let the random variables C_{ijn} and Ψ_{ij} be defined as above. Assuming the continuity of the cumulative distribution function F, it follows that $\overline{C_{ij.}} \to (E\Psi_{ij})$ in probability as $N \to \infty$..*

Proof.

$$C_{ijn} = \frac{1}{N} \sum_{b \neq j} \sum_{n'=1}^{N} u(X_{ijn} - X_{ibn'}).$$

$$\alpha_{n_{i.}}(r_{ijn}) = \frac{1}{n_{i.} + 1} \sum_{b=1}^{J} \sum_{n'=1}^{N} u(X_{ijn} - X_{ibn'}).$$

Since $S_N(i,j) = \sum_{n=1}^{N} \alpha_{n_{i.}}(r_{ijn})$, we have $S_N(i,j)(JN+1) - NC_{ij.} = \sum_{n=1}^{N} \sum_{n'=1}^{N} u(X_{ijn} - Xijn') = \frac{N(N+1)}{2}$. Thus $\overline{C_{ij.}} = \frac{JS_N(i,j)}{N} - \frac{N(N+1)}{2N^2}$. Since $\text{var}(\frac{S_N(i,j)}{N}) \to 0$ (Thm 3.2, Hájek [8]), we have $\text{var}(\overline{C_{ij.}}) \to 0$. Furthermore, lemma 4.5.1 implies $\lim_{n\to\infty} E(\overline{C_{ij.}}) = E(\Psi_{ij})$, and hence we have $\overline{C_{ij.}} \to (E\Psi_{ij})$ in probability. □

The result of lemma 4.5.1 and lemma 4.5.2 can be extended to the case of C_{ibn}, Ψ_{ib}, C_{ibn}^*, Ψ_{ib}^* for $b = 1, \ldots, J$.

Lemma 4.5.3. *Let the random variables C_{ijn} and Ψ_{ij} be defined as above. Assuming the continuity of the cumulative distribution function F, it follows that $E|C_{ibn}C_{ibn}^* - \Psi_{ib}\Psi_{ib}^*| = O(N^{-1})$*

Proof. We have shown $E|(C_{ibn} - \Psi_{ib})| = O(N^{-1})$ and $E|(C^*_{ibn} - \Psi^*_{ib})| = O(N^{-1})$. Ψ_{ib} and Ψ^*_{ib} are bounded. Therefore,

$$\begin{aligned} E|C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib}| &\leq E|(C_{ibn} - \Psi_{ib})\Psi^*_{ib}| + E|\Psi_{ib}(C^*_{ibn} - \Psi^*_{ib})| \\ &\quad + E|(C_{ibn} - \Psi_{ib})(C^*_{ibn} - \Psi^*_{ib})| \\ &\leq O(N^{-1}) + E|(C_{ibn} - \Psi_{ib})(C^*_{ibn} - \Psi^*_{ib})| \end{aligned} \tag{4.5.5}$$

We also have the convergence rate for the squared terms $E|(C_{ibn}-\Psi_{ib})^2| = O(N^{-1})$ and $E|(C^*_{ibn} - \Psi^*_{ib})^2| = O(N^{-1})$. By the Cauchy-Schwarz inequality,

$$\begin{aligned} &E|(C_{ibn} - \Psi_{ib})(C^*_{ibn} - \Psi^*_{ib})| \\ \leq &E|(C_{ibn} - \Psi_{ib})^2|^{\frac{1}{2}} E|(C^*_{ibn} - \Psi^*_{ib})^2|^{\frac{1}{2}} \\ = &O(N^{-1}) \end{aligned} \tag{4.5.6}$$

□

Lemma 4.5.4. *Let the random variables C_{ijn} and Ψ_{ij} be defined as above. Assuming the continuity of the cumulative distribution function F, it follows that*

$$\lim_{N\to\infty} E[\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn}] = E[\Psi_{ib}\Psi^*_{ib}].$$

Proof. According to 4.5.3,

$$E[\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn}] = \frac{1}{N}\sum_{n=1}^{N}\{E[\Psi_{ib}\Psi^*_{ib}] + O(N^{-1})\}$$

□

Lemma 4.5.5. $\lim_{n\to\infty} \mathrm{var}(\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn}) = 0$

Proof.

$$\begin{aligned}
&\mathrm{var}[\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn}] \\
&= E[\{\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn} - E(\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn})\}^2] \\
&= E[\{\frac{1}{N}\sum_{n=1}^{N}(C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib}) + O(\frac{1}{N})\}^2] \qquad (4.5.7) \\
&= E[\{\frac{1}{N}\sum_{n=1}^{N}(C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib})\}^2] + O(\frac{1}{N}) \\
&\le \frac{1}{N}\sum_{n=1}^{N} E[(C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib})^2] + O(\frac{1}{N})
\end{aligned}$$

Since $|C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib}|$ is bounded by $2(J-1)^2$, we have $E[(C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib})^2] \le 2(J-1)^2 E[C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib}]$. Lemma 4.5.3 implies that $E[(C_{ibn}C^*_{ibn} - \Psi_{ib}\Psi^*_{ib})^2] = O(\frac{1}{N})$, therefore $\lim_{n\to\infty} \mathrm{var}(\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn}) = 0$. □

Theorem 4.5.6. *Let the random variables C_{ijn} and Ψ_{ij} be defined as above. Assuming the continuity of the cumulative distribution function F, it follows that*

$$\frac{1}{N}\sum_{n=1}^{N}((C_{ibn} - \overline{C_{ib.}})(C^*_{ibn} - \overline{C^*_{ib.}}) \to \mathrm{cov}(\Psi_{ib}, \Psi_{ib'})$$

in probability.

Proof. Note

$$\frac{1}{N}\sum_{n=1}^{N}(C_{ibn} - \overline{C_{ib.}})(C^*_{ibn} - \overline{C^*_{ib.}}) = \frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn} - \overline{C_{ib.}C^*_{ib.}}$$

We have already shown $\overline{C_{ib.}} \to E\Psi_{ij}$ in probability and $\overline{C^*_{ib.}} \to E\Psi^*_{ij}$ in probability, which implies $\overline{C_{ib.}C^*_{ib.}} \to E\Psi_{ij}E\Psi^*_{ij}$ in probability. Lemma 4.5.4 and 4.5.5 imply $\frac{1}{N}\sum_{n=1}^{N} C_{ibn}C^*_{ibn} \to E(\Psi_{ib}\Psi^*_{ib})$ in probability. Combining these two results, we get $\frac{1}{N}\sum_{n=1}^{N}((C_{ibn} - \overline{C_{ib.}})(C^*_{ibn} - \overline{C^*_{ib.}}) \to \mathrm{cov}(\Psi_{ib}, \Psi_{ib'})$ in probability. The estimator for the variance is a special case of the covariance estimator.

□

Here we provide a numerical example to demonstrate the performance of this estimator. Suppose all α_i, β_j, γ_{ij} are all zero. We can explicitly obtain the value for the limiting variance and covariance.

$$\begin{aligned}
\sigma(i,j) &= \mathrm{var}(F(X_{ij}) - \frac{1}{J}\sum_{b=1}^{J} F(X_{ib})) \\
&= \mathrm{var}(\frac{J-1}{J}F(X_{ij}) - -\frac{1}{J}\sum_{b\neq j} F(X_{ib})) \\
&= \frac{(J-1)^2}{J^2}\mathrm{var}(U_{ij}) + \frac{1}{J^2}\sum_{b\neq j}\mathrm{var}(U_{ib}) \\
&= (\frac{(J-1)^2}{J^2} + \frac{J-1}{J^2})\frac{1}{12} \\
&= \frac{1}{12}\frac{J-1}{J}
\end{aligned} \tag{4.5.8}$$

Here $F(X_{ij})$ is a random variable following the uniform distribution on $[0,1]$, and its variance is equal to $\frac{1}{12}$.

$$\begin{aligned}
\sigma(i,j,i,j') &= \mathrm{cov}(F(X_{ij}) - \frac{1}{J}\sum_{b=1}^{J} F(X_{ib}), F(X_{ij'}) - \frac{1}{J}\sum_{b=1}^{J} F(X_{ib})) \\
&= \mathrm{cov}(\frac{J-1}{J}F(X_{ij}) - \frac{1}{J}\sum_{b\neq j} F(X_{ib}), \frac{J-1}{J}F(X_{ij'*/}) - \frac{1}{J}\sum_{b\neq j'} F(X_{ib})) \\
&= \frac{-(J-1)}{J^2}\mathrm{var}(U_{ij}) + \frac{-(J-1)}{J^2}\mathrm{var}(U_{ij'}) + \frac{1}{J^2}\sum_{b\neq j,j'}\mathrm{var}(U_{ib}) \\
&= (\frac{-2(J-1)}{J^2} + \frac{J-2}{J^2})\frac{1}{12} \\
&= \frac{1}{12}\frac{-1}{J}
\end{aligned} \tag{4.5.9}$$

Given $I = 3$, $J = 4$, $\sigma(i,j) = \frac{1}{16} = 0.0625$, $\sigma(i,j,i,j') = -\frac{1}{48} = -0.0208$. We conduct simulations with the I, J as defined and use standard normal as the common F. Cell sizes are simulated at 5, 10, 20, 50, 100 respectively. At each cell sizes, 10 replicates of data sets are simulated and estimators are calculated for each replicate. Let $\overline{\hat{\sigma}}(i,j)$ and $\overline{\hat{\sigma}}(i,j,i,j')$ designate the average of the estimators obtained from the 10 replicates, and let sd $\hat{\sigma}(i,j)$and sd $\hat{\sigma}(i,j,i,j')$ designate the standard deviation of the estimators among the 10 replicates. The convergence of the estimator is well demonstrated by table 1.

N	$\hat{\sigma}(i,j)$	sd $\hat{\sigma}(i,j)$	$\hat{\sigma}(i,j,i,j')$	sd $\hat{\sigma}(i,j,i,j')$
5	0.0614	0.0300	-0.0161	0.0148
10	0.0607	0.0107	-0.0192	0.0064
20	0.0635	0.0084	-0.0213	0.0031
50	0.0620	0.0041	-0.0214	0.0021
100	0.0621	0.0022	-0.0208	0.0015

Table 1: Convergence of the consistent estimators

4.6 Covariance structure between row rank sum and column rank sum

Now we derive the covariance structure for the row and column sum method. As before, we let $S_N(i,j)$ stand for the sum of row rank scores in cell (i,j) and $T_N(i,j)$ stand for the sum of column rank scores in cell (i,j).

According to theorem 3.2.2, we know that $S_N(i,j)$ can be projected onto a set of independent variable $W_{ibn}^{(i,j)}$, while $b = 1,\ldots,J$ and $n = 1,\ldots,N$.

Similarly, $T_N(i,j)$ can be projected onto a set of independent variables $W_{ajn}^{*(i,j)}$, while $a = 1,\ldots,I$ and $n = 1,\ldots,N$. Recall the definition of $C_{ibn}^{(i,j)}$ and $\Psi_{ib}^{(i,j)}$. Let $\Upsilon_{ij}^{(i,j)} = \sum_{a\neq i} F_{aj}(X_{ij1})$ and $\Upsilon_{aj}^{(i,j)} = -F_{ij}(X_{aj1})$ for $a \neq i$.

Define

$$D_{ijn}^{(i,j)} = \frac{1}{N}\sum_{a\neq i}\sum_{n'=1}^{N} \mu(X_{ijn} - X_{ajn'})$$

and for $a \neq i$,

$$D_{ajn}^{(i,j)} = -\frac{1}{N}\sum_{n'=1}^{N}\mu(X_{ajn} - X_{ijn'}).$$

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\operatorname{cov}(S_N(i,j), T_N(i',j')) &= \operatorname{cov}(\sum_{b=1}^{J}\sum_{n=1}^{N} W_{ibn}^{(i,j)}, \sum_{a=1}^{I}\sum_{n=1}^{N} W_{aj'n}^{*(i',j')}) \\ &= \operatorname{cov}(W_{ij'1}^{(i,j)}, W_{ij'1}^{*(i',j')}) \\ &= \frac{1}{IJ}\operatorname{cov}(\Psi_{ij'}^{(i,j)}, \Upsilon_{ij'}^{(i',j')})\end{aligned} \tag{4.6.1}$$

The corresponding estimator is $\frac{1}{IJN}\sum_{n=1}^{N}(C_{ij'n}^{(i,j)} - \overline{C_{ij'.}^{(i,j)}})(D_{ij'n}^{(i',j')} - \overline{D_{ij'.}^{(i',j')}})$. Up to now we have derived the consistent estimates for every entry of Σ_1,Σ_2, and Σ_{12}.

Chapter 5

Extension to Unbalanced Designs

5.1 Extension to unbalanced designs with proportional cell weights

In this section, our rank tests can be extended to accommodate the unbalanced design. Let n_{ij} represent the number of observations in cell (i, j). Let n_{max} and n_{min} be the maximum and minimum of n_{ij} respectively. Assume $0 < \lim_{n_{min} \to \infty} \frac{n_{min}}{nij} \leq 1$. Let the score generating function ϕ be the identity function. Let r_{ijk} represent the rank of observation x_{ijk} among the ith row, and c_{ijk} represent the rank of observation of x_{ijk} among the jth column.

Define $n_{i.} = \sum_{b=1}^{J} n_{ib}$ and $n_{.j} = \sum_{a=1}^{I} n_{aj}$. Let

$$S_N(i,j) = \sum_{k=1}^{n_{ij}} r_{ijk}/(n_{i.}+1)$$

$$T_N(i,j) = \sum_{k=1}^{n_{ij}} c_{ijk}/(n_{.j}+1)$$

In order to construct a quadratic form with zero expectation under the null hypothesis, we observe that:

$$\begin{aligned} E(S_N(i,j)) =& \frac{1}{n_{i.}+1}\sum_{k=1}^{n_{ij}} E(r_{ijk}) \\ =& \frac{1}{n_{i.}+1}\sum_{k=1}^{n_{ij}} E(\sum_{a=1}^{J}\sum_{b=1}^{n_{ia}} I(X_{iab} \le X_{ijk})) \\ =& \frac{n_{ij}}{n_{i.}+1}\sum_{a=1}^{J} n_{ia} P(X_{ia1} \le X_{ij1}) \\ =& \frac{n_{ij}}{n_{i.}+1}\sum_{a=1}^{J} n_{ia} \int F(x-\alpha_i-\beta_a)dF(x-\alpha_i-\beta_j) \end{aligned} \tag{5.1.1}$$

Define $V(a,j) = \int F(x-\alpha_i-\beta_a)dF(x-\alpha_i-\beta_j)$, and let $w_{ia} = \frac{n_{ia}}{n_{i.}+1}$, we have:

$$E(S_N(i,j)) = n_{ij}\sum_{a=1}^{J} w_{ia} V(a,j).$$

Under H_0,

$$\begin{aligned}
&|\frac{n_{min}}{n_{ij}}E(S_N(i,j)) - \frac{1}{I}\sum_{b=1}^{I}\frac{n_{min}}{n_{bj}}E(S_N(b,j))| \\
&= n_{min}|\sum_{a=1}^{J} w_{ia}V(a,j) - \frac{1}{I}\sum_{b=1}^{I}\sum_{a=1}^{J} w_{ba}V(a,j)| \\
&= n_{min}|\sum_{a=1}^{J}(w_{ia} - \frac{1}{I}\sum_{b=1}^{I} w_{ba})V(a,j)| \\
&\leq n_{min}\sum_{a=1}^{J}|w_{ia} - \frac{1}{I}\sum_{b=1}^{I} w_{ba}|V(a,j)
\end{aligned} \tag{5.1.2}$$

If $\lim_{n_{min}\to\infty} n_{min}|w_{ia} - \frac{1}{I}\sum_{b=1}^{I} w_{ba}| = 0$, then we have

$$\lim_{n_{min}\to\infty} E(\frac{n_{min}}{n_{ij}}S_N(i,j) - \frac{1}{I}\sum_{b=1}^{I}\frac{n_{min}}{n_{bj}}S_N(b,j)) = 0$$

Construct a diagonal matrix D with $D[(i-1)J+j,(i-1)J+j] = \frac{n_{min}}{n_{ij}}$. Let S_N be the vector of elements $S_N(i,j)$. Then the ijth element of DS_N's is equal to $\frac{n_{min}}{n_{ij}}S_N(i,j)$. Thus $(ADS_N)_{ij} = \frac{n_{min}}{n_{ij}}S_N(i,j) - \frac{1}{I}\sum_{b=1}^{I}\frac{n_{min}}{n_{bj}}S_N(b,j)$.

Let Σ_1 be the matrix with the $((i-1)J+j,(i'-1)J+j')$ element denoting $\sigma_1^2(i,j,i',j') = \lim_{n_{min}\to\infty}\frac{1}{n_{min}}\mathrm{cov}(S_N(i,j),S_N(i',j'))$.

Let $Z_{ian}^{(i,j)}$, $a = 1,\ldots,J$, $n = 1,\ldots,n_{ia}$ be the independent random variables constructed by the projection method on the linear rank statistic $S_N(i,j)$ (Theorem 3.2.2). Define

$$\Psi_{ia}^{(i,j)} = -\frac{n_{ij}}{n_{i.}}F_{ij}(X_{ia1})$$

for $a \neq j$, and

$$\Psi_{ij}^{(i,j)} = \sum_{a\neq j}\frac{n_{ia}}{n_{i.}}F_{ia}(X_{ij1}).$$

We have

$$\begin{aligned}
&\mathrm{var}(S_N(i,j)) \\
&= \mathrm{var}(\sum_{a=1}^{J}\sum_{n=1}^{n_{ia}} W_{ian}^{(i,j)}) + O(1) \\
&= \sum_{a=1}^{J} n_{ia}\,\mathrm{var}(W_{ia1}^{(i,j)}) + O(1) \\
&= \sum_{a\neq j} n_{ia}\,\mathrm{var}(-\frac{n_{ij}}{n_{i.}}F_{ij}(X_{ia1})) + n_{ij}\,\mathrm{var}(\sum_{a\neq j}\frac{n_{ia}}{n_{i.}}F_{ia}(X_{ij1})) + O(1) \\
&= \sum_{a=1}^{J} n_{ia}\,\mathrm{var}(\Psi_{ia}^{(i,j)}) + O(1)
\end{aligned} \tag{5.1.3}$$

Similarly we derive

$$\lim_{n_{min}\to\infty}\mathrm{cov}(S_N(i,j),S_N(i,j')) = \sum_{a=1}^{J} n_{ia}\,\mathrm{cov}(\Psi_{ia}^{(i,j)},\Psi_{ia}^{(i,j')}) + O(1)$$

Theorem 5.1.1. *let*

$$W_1 = \frac{1}{n_{min}}(ADS_N)'((AD)\hat{\Sigma}_1(AD)')^{-1}(ADS_N),$$

where $\hat{\Sigma}_1$ *is a consistent estimate of* Σ_1. *If* $\lim_{n_{min}\to\infty} n_{min}|w_{ia}-\frac{1}{I}\sum_{b=1}^{I} w_{ba}| = 0, \forall i,a,$ *then under the null hypothesis of no interaction,* W_1 *converges to a central chisquare distribution with* $(I-1)(J-1)$ *degrees of freedom as* $n_{min}\to\infty$.

Proof. According to theorem 3.2.2,

$$S_N \to N_{IJ}(E(S_N), \frac{1}{\sqrt{n_{min}}}\Sigma_1)$$

asymptotically. So if the condition of this theorem holds, we have

$$ADS_N \to N_{IJ}(0, \frac{1}{\sqrt{n_{min}}}AD\Sigma_1(AD)')$$

asymptotically. Since $\hat{\Sigma}_1$ converges to Σ_1 in probability element-wise, according to Slutsky's theorem, W_1 converges to a central chisquare distribution with $(I-1)(J-1)$ degrees of freedom. □

The condition under which the theorem holds imposes the requirement that asymptotically the deviations of $w_{ia} - \frac{1}{I}\sum_{i=1}^{I} w_{ia} = o(n_{min}^{-1})$. Similarly extensions exist for the column and the row-column sum statistics.

5.2 Extension to unbalanced designs with arbitrary cell weights

For experimental designs with arbitrary cell weight, there have not been to date any nonparametric tests available to test for interaction. The main difficulty appears to be that often the number of replications in a cell occurs in the formula for the expectation of the rank statistic under the null hypothesis and consequently the rank statistic will not have a central chisquare. In order to overcome this obstacle, we introduce a new weighted rank defined by $r^*_{ijk} = \frac{n_{i.}}{J}\sum_{a=1}^{J}\frac{1}{n_{ia}}\sum_{b=1}^{n_{ia}} I(X_{iab} \leq X_{ijk})$. The weighted rank reduces to the usual rank when the n_{ij} are equal.

Under this definition,

$$\begin{aligned} S_N^*(i,j) &= \sum_{k=1}^{n_{ij}} \frac{r_{ijk}^*}{n_{i.}} \\ &= \sum_{k=1}^{n_{ij}} \frac{1}{J} \sum_{a=1}^{J} \frac{1}{n_{ia}} \sum_{b=1}^{n_{ia}} I(X_{iab} \leq X_{ijk}) \end{aligned} \tag{5.2.1}$$

$$\begin{aligned} E(\frac{S_N^*(i,j)}{n_{ij}}) &= E(\frac{1}{J} \sum_{a=1}^{J} \frac{1}{n_{ia}} \sum_{b=1}^{n_{ia}} I(X_{iab} \leq X_{ijk})) \\ &= \frac{1}{J} \sum_{a=1}^{J} \frac{1}{n_{ia}} n_{ia} P(X_{ia1} \leq X_{ij1}) \\ &= \frac{1}{J} \sum_{a=1}^{J} \int F(x - \alpha_i - \beta_a) dF(x - \alpha_i - \beta_j) \end{aligned} \tag{5.2.2}$$

Under this construction,

$$E(\frac{S_N^*(i,j)}{n_{ij}}) - \frac{1}{I} \sum_{b=1}^{I} E(\frac{S_N^*(b,j)}{n_{bj}}) = 0$$

Therefore, $E(ADS_N^*) = 0$ under the null hypothesis. We will need to prove the asymptotic normality of ADS_N^* in order to justify the chisquare distribution of the test statistic.

Theorem 5.2.1. *Let $S_N^*(i,j)$ be the linear weighted rank statistic. Then, $S_N^*(i,j)$ has an asymptotically normal distribution, as $n_{min} \to \infty$.*

Proof. Consider the two cells (i,j) and (i,a) and let $r_a(X_{ijk})$ be the rank of X_{ijk} among the set of

$$(X_{ij1}, ...X_{ijn_{ij}}, X_{ia1}, ...X_{ian_{ia}}).$$

Let

$$\begin{aligned} S_a &= \sum_{k=1}^{n_{ij}} \frac{r_a(X_{ijk})}{n_{ia}+n_{ij}+1} \\ &= \sum_{k=1}^{n_{ij}} \frac{1}{n_{ia}+n_{ij}+1}[\sum_{k'=1}^{n_{ij}} I(X_{ijk'} \le X_{ijk}) + \sum_{b=1}^{n_{ia}} I(X_{iab} \le X_{ijk})] \end{aligned} \tag{5.2.3}$$

S_a is a linear rank statistic defined on set

$$(X_{ij1}, ...X_{ijn_{ij}}, X_{ia1}, ...X_{ian_{ia}}).$$

Then

$$\begin{aligned} S_N^*(i,j) &= \sum_{k=1}^{n_{ij}} \frac{r_{ijk}^*}{n_{i.}} \\ &= \sum_{k=1}^{n_{ij}} \frac{1}{J} \sum_{a=1}^{J} \frac{1}{n_{ia}} \sum_{b=1}^{n_{ia}} I(X_{iab} \le X_{ijk}) \\ &= \frac{1}{J} \sum_{a=1}^{J} \frac{1}{n_{ia}} (\sum_{k=1}^{n_{ij}} \sum_{b=1}^{n_{ia}} I(X_{iab} \le X_{ijk}) \\ &= const + \frac{1}{J} \sum_{a \ne j} \frac{n_{ia}+n_{ij}+1}{n_{ia}} S_a \end{aligned} \tag{5.2.4}$$

It follows that $S_N^*(i,j)$ is a sum of correlated linear rank statistics defined on subsets of the set

$$X_{i11}, \ldots, X_{iJn_{iJ}}.$$

Then by theorem 3.2.2, $S_N^*(i,j)$ is asymptotically normal.

□

Next, we derive the limiting covariance matrix of S_N^* by a projection argument.

$$\begin{aligned}\operatorname{var}(S_a) &= \sum_{b=1}^{n_{ia}} \operatorname{var}(Z_{iab}) + \sum_{k=1}^{n_{ij}} \operatorname{var}(Z_{ijk}) + O(1) \\ &= n_{ia} \operatorname{var}(\frac{-n_{ij}}{n_{ij}+n_{ia}} F_{ij}(X_{ib1}) + n_{ij} \operatorname{var}(\frac{n_{ia}}{n_{ij}+n_{ia}} F_{ia}(X_{ij1})) + O(1)\end{aligned} \tag{5.2.5}$$

$$\begin{aligned}\operatorname{var}(S_N^*(i,j)) &= \frac{1}{J^2} \sum_{a \neq j} \operatorname{var}(\frac{n_{ij}+n_{ia}+1}{n_{ia}} S_a) + O(1) \\ &= \frac{1}{J^2} [\sum_{a \neq j} n_{ia} \operatorname{var}(-\frac{n_{ij}(n_{ij}+n_{ia}+1)}{n_{ia}(n_{ij}+n_{ia})} F_{ij}(X_{ia})) \\ &+ n_{ij} \operatorname{var}((\sum_{a \neq j} \frac{n_{ij}+n_{ia}+1}{n_{ij}+n_{ia}} F_{ia})(X_{ij1})) + O(1) \\ &= \frac{1}{J^2} n_{ia} \operatorname{var}(\Psi_{ia}^{(j)}) + O(1),\end{aligned} \tag{5.2.6}$$

where

$$\begin{aligned}\Psi_{ia}^{(j)} &= -\frac{n_{ij}(n_{ij}+n_{ia}+1)}{n_{ia}(n_{ij}+n_{ia})} F_{ij}(X_{ia1}), \\ \Psi_{ij}^{(j)} &= \sum_{a \neq j} \frac{n_{ij}+n_{ia}+1}{n_{ij}+n_{ia}} F_{ia}(X_{ij1}).\end{aligned}$$

Similarly, we have

$$\operatorname{cov}(S_N^*(i,j), S_N^*(i,j')) = \frac{1}{J^2} n_{ia} \operatorname{cov}(\Psi_{ia}^{(j)}, \Psi_{ia}^{(j')}) + O(n_{min}^{-1})$$

For $a \neq j$, define

$$C_{iak}^{(j)} = -\frac{(n_{ij}+n_{ia}+1)}{n_{ia}(n_{ij}+n_{ia})} \sum_{n=1}^{n_{ij}} I(X_{ijn} \leq X_{iak}),$$

and

$$C_{ijk}^{(j)} = \sum_{a \neq j} \frac{(n_{ij}+n_{ia}+1)}{n_{ia}(n_{ij}+n_{ia})} \sum_{n=1}^{n_{ia}} I(X_{ian} \leq X_{ijk}),$$

Define

$$\overline{C_{ia}}^{(j)} = \frac{1}{n_{ia}} \sum_{n=1}^{n_{ia}} C_{ian}.$$

Let

$$\sigma_1(i,j,i,j) = \lim_{n_{min}\to\infty} \frac{1}{n_{min}} \mathrm{var}(S_N^*(i,j))$$

$$\sigma_1(i,j,i,j') = \lim_{n_{min}\to\infty} \frac{1}{n_{min}} \mathrm{cov}(S_N^*(i,j), S_N^*(i,j')).$$

Then

$$\hat{\sigma}_1(i,j,i,j) = \frac{1}{J^2} \sum_{a=1}^{J} \frac{1}{n_{min}} \sum_{k=1}^{n_{ia}} (C_{iak}^{(j)} - \overline{C_{ia}}^{(j)})^2,$$

and we have

$$\hat{\sigma}_1(i,j,i,j') = \frac{1}{J^2} \sum_{a=1}^{J} \frac{1}{n_{min}} \sum_{k=1}^{n_{ia}} (C_{iak}^{(j)} - \overline{C_{ia}}^{(j)})(C_{iak}^{(j')} - \overline{C_{ia}}^{(j')}).$$

The corresponding W_1 statistic takes the form of:

$$\frac{1}{n_{min}} (ADS_N^*)'((AD)\hat{\Sigma}_1(AD)')^{-1}(ADS_N^*)$$

Chapter 6

Monte Carlo Simulation Study

6.1 Monte Carlo simulation study for balanced designs

Monte Carlo simulations have been conducted to study the type I error and power of 6 tests: parametric, rank transform, aligned, row method (W_1), column method (W_2) and row-column method (W_3) respectively. Small and large sample properties are both examined for 5 different distributions: normal (NOR), lognormal (LOG), double exponential (DEX), uniform (UNI), cauchy (CAU). The parameters of the error distribution are listed in Table 2.

The simulation setting for the type I error is as follows: There are three rows and four columns, i.e. $I = 3$, $J = 4$. $\alpha_1 = \frac{4}{3}$, $\alpha_2 = \frac{2}{3}$, $\alpha_3 = -2$; $\beta_1 = \frac{3}{2}$, $\beta_2 = \frac{1}{2}$, $\beta_3 = \frac{-1}{2}$,

Normal	$N(0,1)$
Uniform	$U(-3,3)$
Cauchy	$\beta = 0.5$
Lognormal	$\log(\epsilon_{ijn}) \sim N(0, 0.009)$
Double exponential	$\beta = 1$

Table 2: Underlying distributions for the simulated errors

$\beta_4 = \frac{-3}{2}$. The $\gamma_{ij} = 0$ for all i, j's under the null hypothesis. Similar studies of other values give comparable results.

For each distribution, simulations are conducted for cell sizes equal to 5, 10, 20, 40. For each setting, 1600 simulation runs are performed. Figures 1 to 5 display the type I error (Y axis) against the cell sizes (X axis) for each method. The significance level is set to be 0.05.

The simulation setting for power is as follows: All the parameters and noise models remain as before except for the interaction term: $\gamma_{11} = m$, $\gamma_{12} = -m$, $\gamma_{21} = -m$, $\gamma_{22} = m$, $\gamma_{ij} = 0$ for all other i, j's; m values are chosen to be 0.45, 0.52, 0.60. For each interaction level, the number of simulation runs are 3100, 2200, and 1500 respectively. We perform fewer simulation runs for higher interaction level because the rejection event is more frequent and to collect the same number of rejection events, we need fewer simulation runs. The justification is that in order to keep the same coefficient of variation of the Monte Carlo estimates for the power, which is equal to $\sqrt{\frac{1-p}{pn}}$, we need larger values of n for smaller value of p. Figures 6 to 10 plot the power curve

(Y axis) of the 5 methods (rank transformed method has been excluded for its poor performance in the Type-I error tests) against the interaction term m (X axis). The significance level is set to be 0.05.

When the underlying distribution is normal, the rank transform method has unacceptable huge type I error. When the cell size increases from 5 to 40, the type I error increases from 0.12 to over 0.90. This result reaffirms the result of Thompson [25] whereby the rank transform statistic has divergent mean as cell size increases. The other 5 methods have good type-I error behavior. When the cell size equals 5, the parametric method, the column method and the row method are very close to 0.05, the theoretical type I error rate. The aligned method and row-column method are in the range of $(0.06, 0.08)$. When the cell size increases to 20, all the methods except the rank transform method converge to the 0.05 level. The power of the row-column method is superior for cell sizes equal to 5 or 10. When the cell size is equal to 20 or more, the parametric is the most powerful. Under the same interaction level, the power increases as cell sizes increase. When the cell size equals 40, all the methods have power above 0.99 for all the interaction values.

When the underlying distribution is uniform, the rank transform method has unacceptable huge type-I error which ranges from 0.15 to 1 as the cell size increases from 5 to 40. When the cell size equals 5, the aligned and row-column method have type I error between 0.08 to 0.10. The other methods have type I error close to 0.05.

As the cell size increases to 20 or more, all the methods except for rank transform have type I error converging to 0.05. Regarding the power, for a sample size of 5, the row-column method has the highest power among all the methods. For cell sizes of 10 or more, parametric method is most powerful. This simulation demonstrates that the parametric method is valid for the uniform distribution. When the cell size is 40, all the tests have power greater than 0.99 at all three interaction levels.

When the underlying distribution is lognormal, the rank transform method has a type-I error in the range of 0.09 to 0.50 as the cell size increases from 5 to 40. For a cell size of 5, the column method's type I error is very close to 0.05. The parametric method has a type I error close to 0.02, while the row-column, aligned, row methods have type I errors in the range of $(0.06, 0.08)$. When the cell size increases, all the methods excluding the rank transform have good type I error rates below 0.10, which however, are higher compared to the normal and uniform distributions. The parametric method has very poor power ranging from less than 0.10 to 0.45 for $m = 0.45$, while the other methods range from 0.25 to above 0.90. For a cell size of 5, the row-column method has highest power. When the cell size equals 10 or more, the aligned tests take the lead. There are only small differences among the power of all the nonparametric methods excluding the rank transform method.

When the underlying distribution is double exponential, the rank transform method still has an unacceptable type I error. When the cell size equals 5, the type I error

for both the parametric method and aligned method is close to 0.05, while the row, column, and row-column methods have type-I errors in the range of $(0.07, 0.09)$. As the cell sizes increases to 20 or more, all the methods except for the rank transform have a type I error close to 0.05. The parametric method has very poor power for a cell size of 5. As the cell sizes increase, the power increases as expected. For a cell size of 5, the power of the row-column method is highest. When the cell size increases to 20 or more, the aligned test has higher power. This is due to the fact that when the cell size increases, the estimates of the row and column effect becomes more accurate and thus the power is improved.

When the underlying distribution is Cauchy, the rank transform method has a large type I error which ranges from 0.10 to 0.40 as the cell size increases from 5 to 40. For all the cell sizes, the aligned test nearly attains a type I error of 0.05. The parametric method has a type I error about 0.02. The row, column and row-column methods have type I errors in the range of 0.07 to 0.10. The power of the parametric method is below 0.10 for all settings, even when the cell size increases. This is because the Central Limit Theorem does not apply to the Cauchy. When the cell size is 5, the column and row-column method have highest power. When the cell size increases to 10, the aligned test takes has highest power, because the estimation becomes more accurate as the cell size increases. All the nonparametric methods behave well with good type-I error and high power in this extremely heavy tailed distribution.

In conclusion, the rank transform method leads to an invalid test for interaction due to the large type I error for large cell sizes. The parametric test is valid only for the normal distribution. It behaves reasonably well for light tailed distribution, such as the uniform distribution. For other distributions with bounded second moment, its type I error and power improve as cell sizes increase. The aligned test is a valid test for all distributions, however, it relies on the nonparametric estimation of the location parameter. Thus, its power is not best for small cell sizes. The type I errors of our proposed test statistics converge to the appropriate significance level for all distributions. Even for small cell sizes, they have satisfactory type I errors. They have high power in normal and all non-normal distributions.

6.2 Monte Carlo study for unbalanced designs

Simulation studies have been conducted to investigate the performance of the new test statistics in unbalanced designs. All the noise model and parameter settings are identical to balanced designs, except for the cell sizes. An unbalanced design is simulated with cell sizes ranging from 5 to 20. The replicates number in each cell are shown in Table 3. The significance level is set to be 0.05. The type I error and power of the three proposed methods for unbalanced designs are compared to the balanced designs. Figures 11 to 15 display the rejection probability of the methods (Y axis) against the different interaction levels (X axis). When the interaction level $m = 0$,

5	15	10	10
15	10	10	15
10	10	10	20

Table 3: Cell sizes for the simulated unbalanced design

the rejection probability is the type I error. Power is evaluated at interaction level $m = 0.45, 0.52, 0.60$ respectively. The simulation runs are 5000, 3100, 2200, 1500 for the 4 interaction levels. As m increases, the rejection probability increases and smaller numbers of simulation runs are required. The unbalanced design is compared with three balanced designs with fixed cell sizes 5, 10, 20 respectively. Different panels depict the performance of the row, column and row-column methods.

For all the 5 noise distributions, as m increases, the 4 power curves steadily increase. When $m = 0$, all the 3 methods have type I error close to 0.05 for unbalanced design. When $m > 0$, as the unbalanced design has cell size ranging from 5 to 20, its power curve is between the balanced designs with equal cell size 10 and 20. In conclusion, the simulation results verify that the 3 methods have conservative type I error for normal and non-normal distributions. The simulation also demonstrates that our method for the unbalanced design has the same level of power as a balanced design with its average cell size.

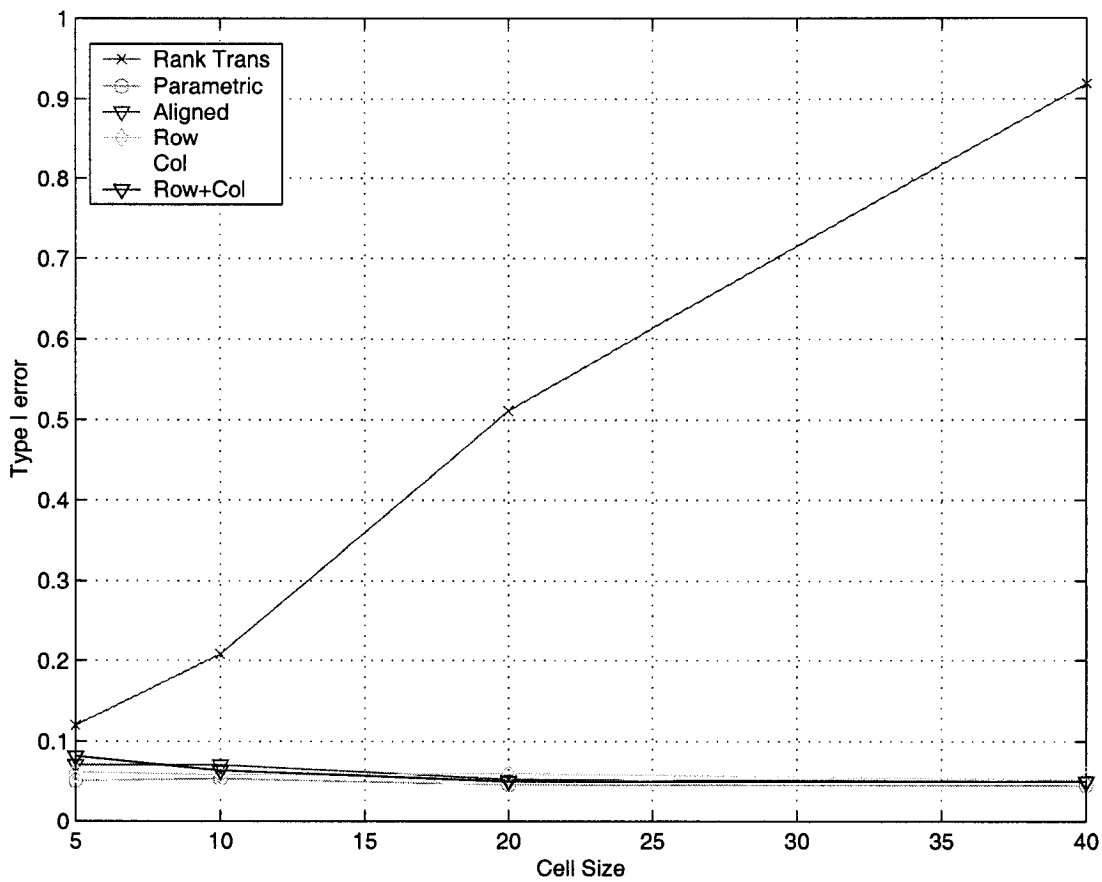

Figure 1: Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the normal distribution.

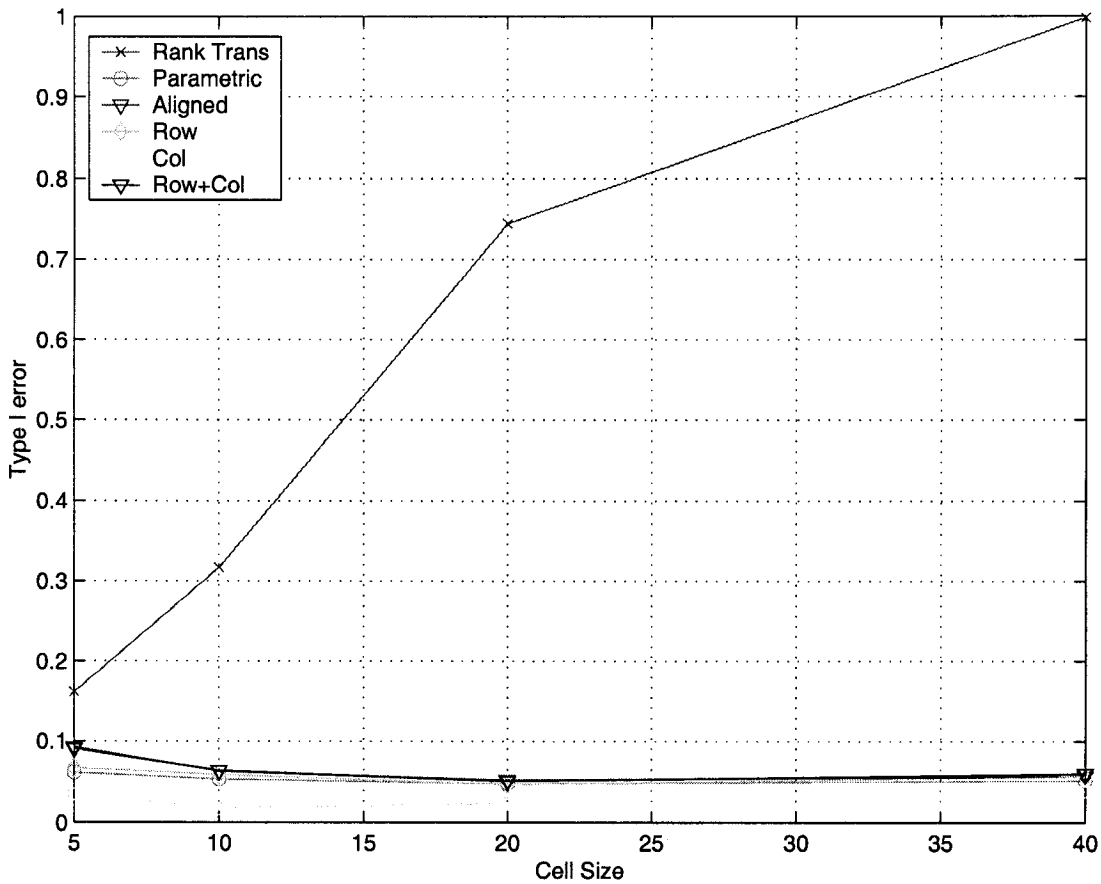

Figure 2: Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the uniform distribution.

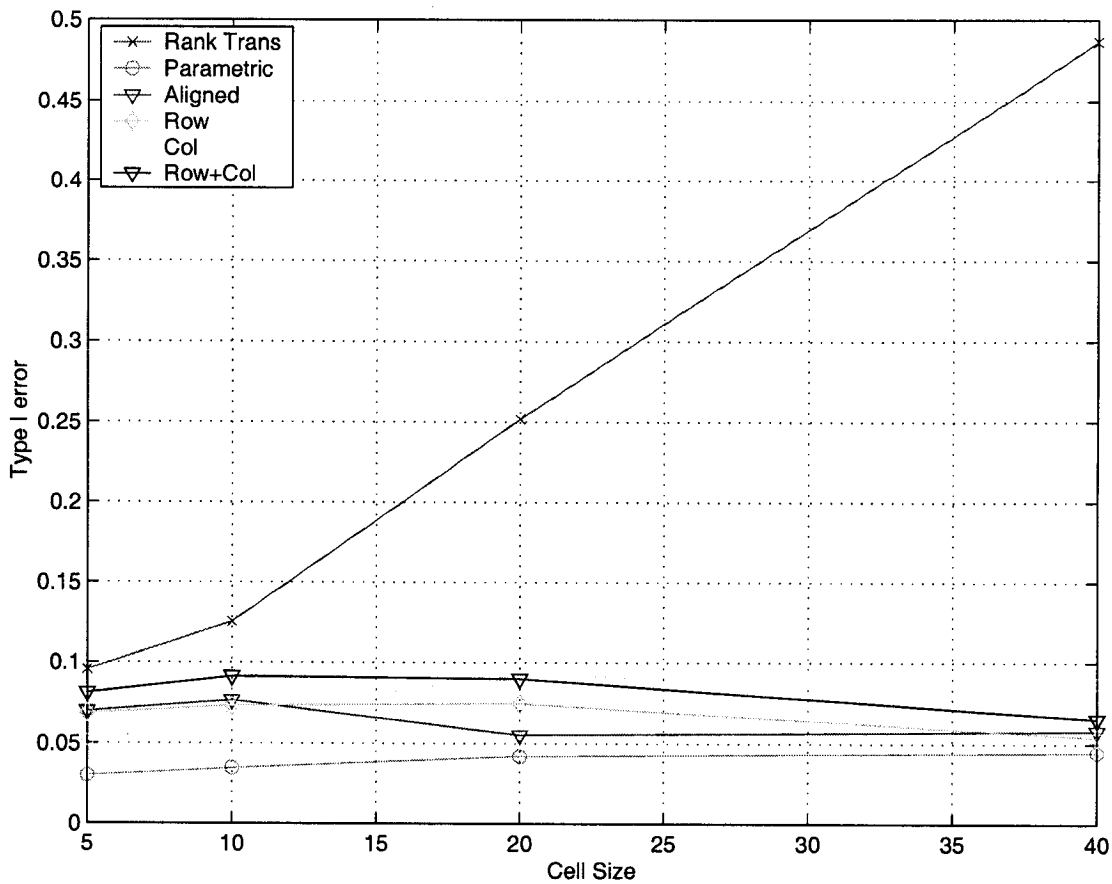

Figure 3: Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Lognormal distribution.

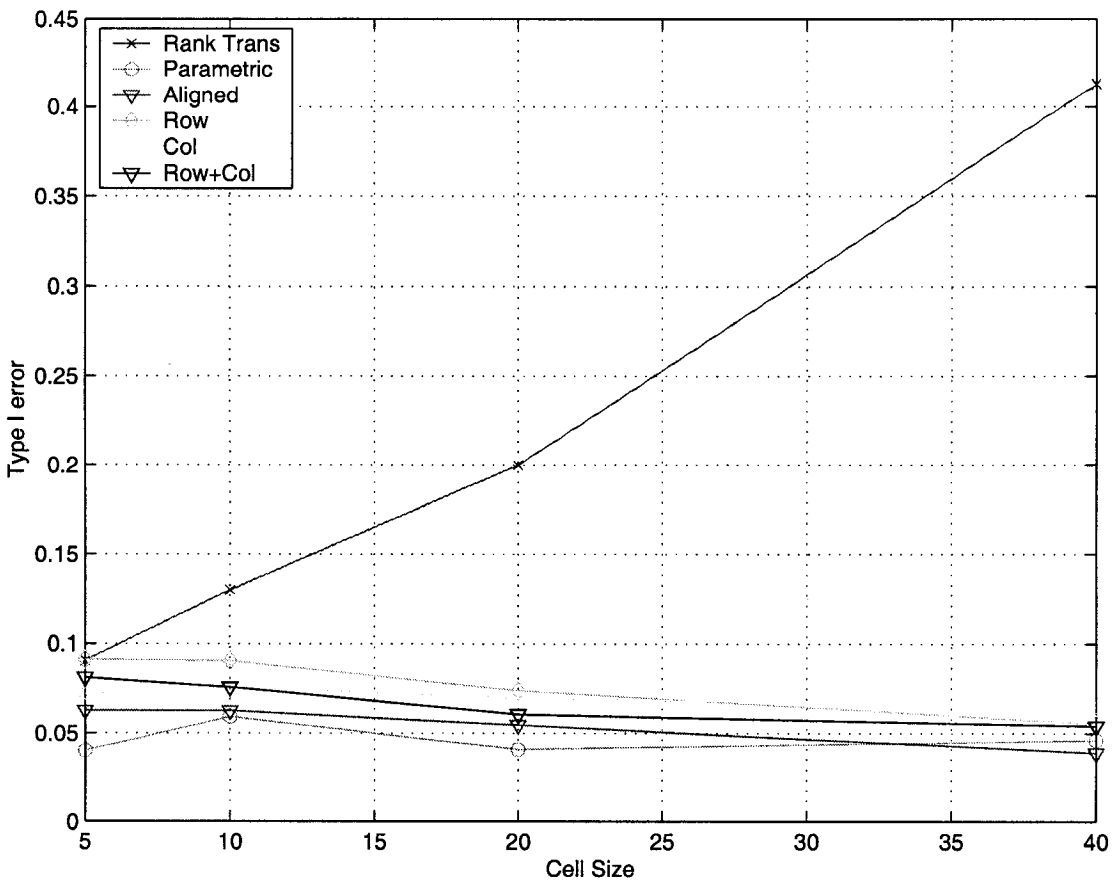

Figure 4: Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the double exponential distribution.

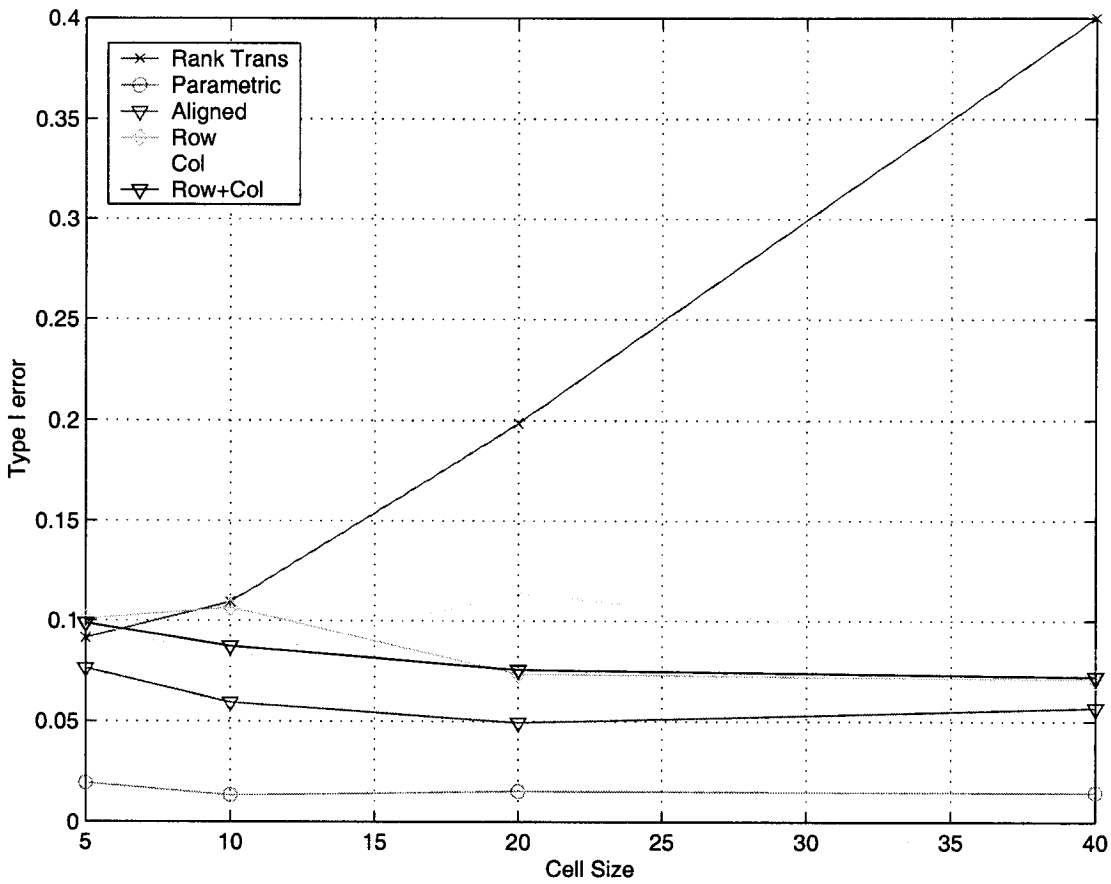

Figure 5: Type I error for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Cauchy distribution.

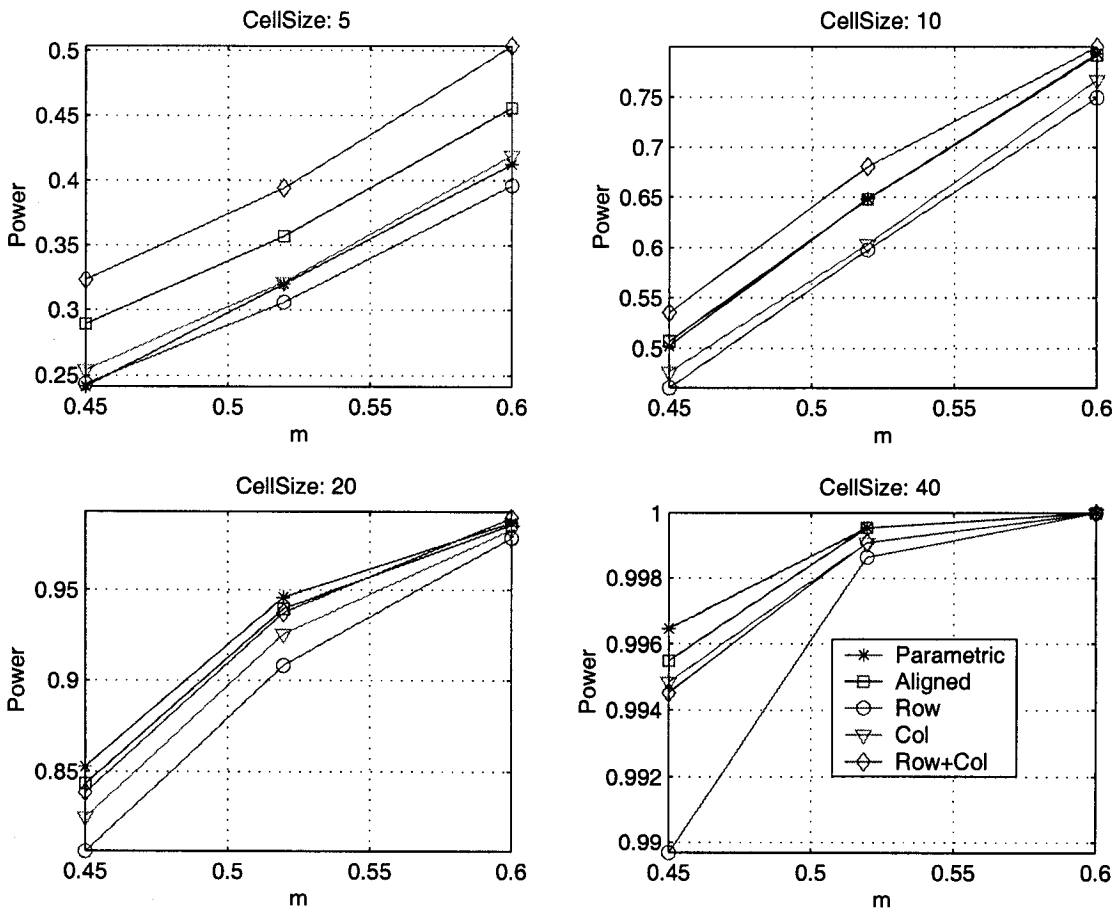

Figure 6: Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the normal distribution.

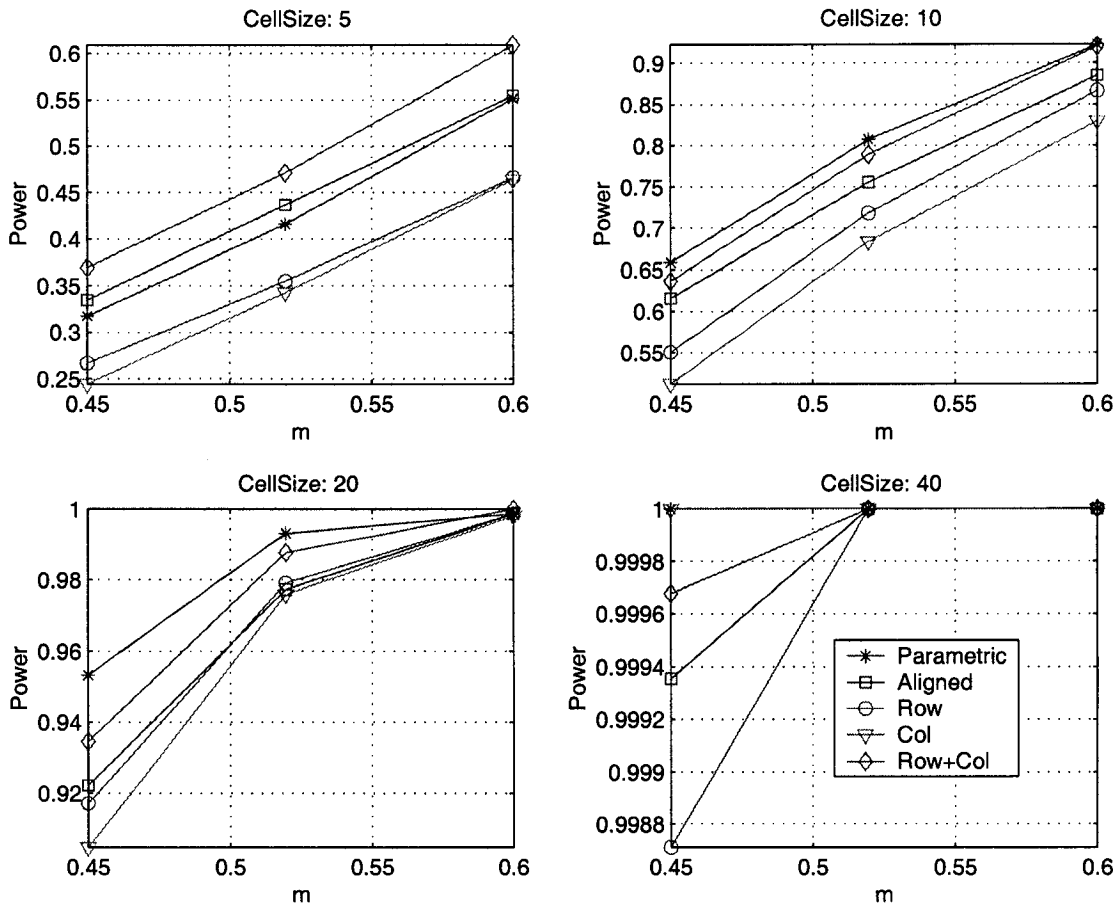

Figure 7: Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the uniform distribution.

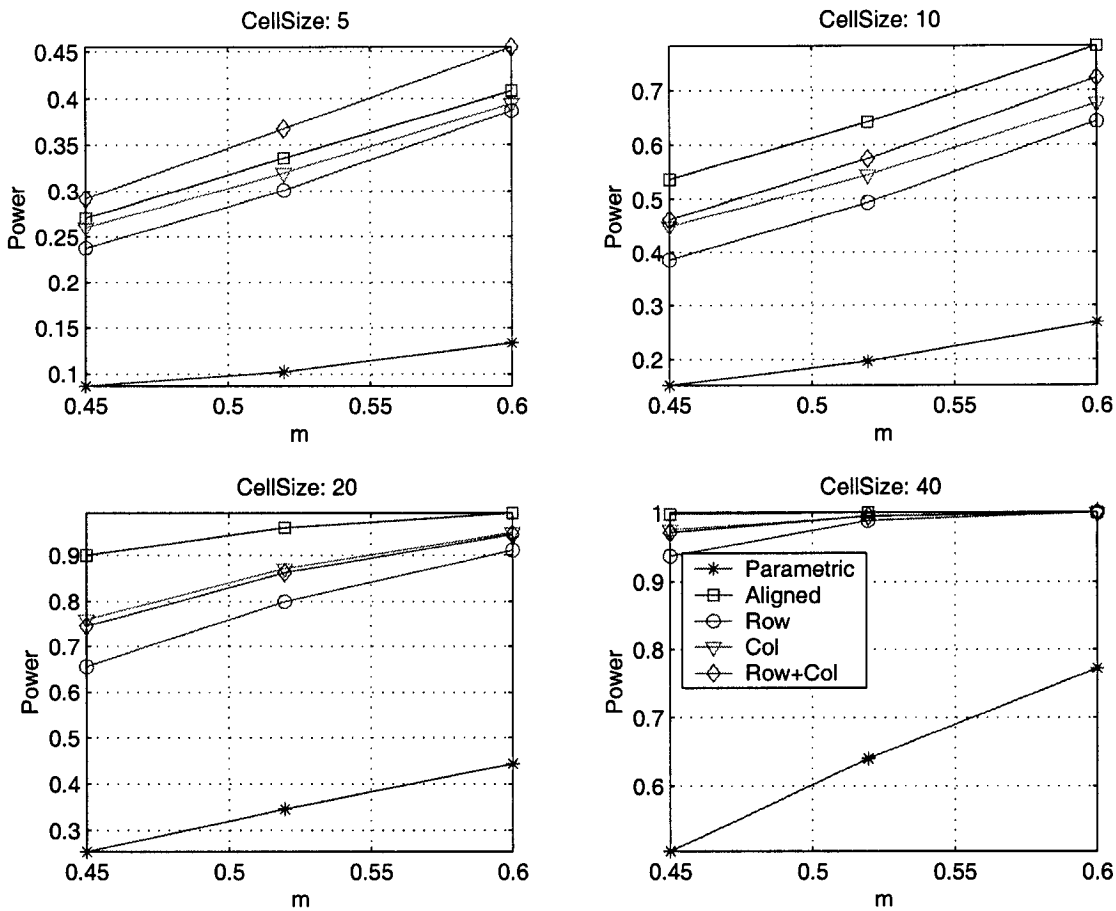

Figure 8: Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Lognormal distribution.

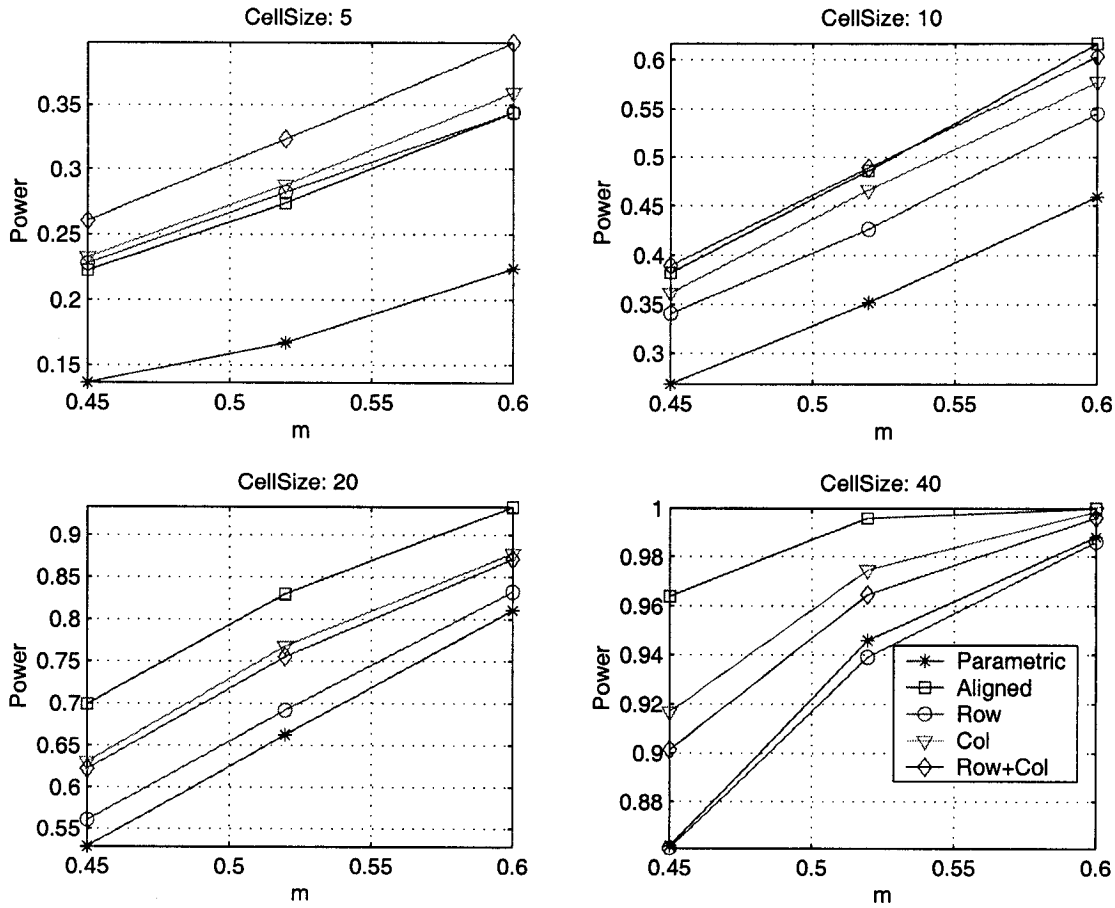

Figure 9: Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the double exponential distribution.

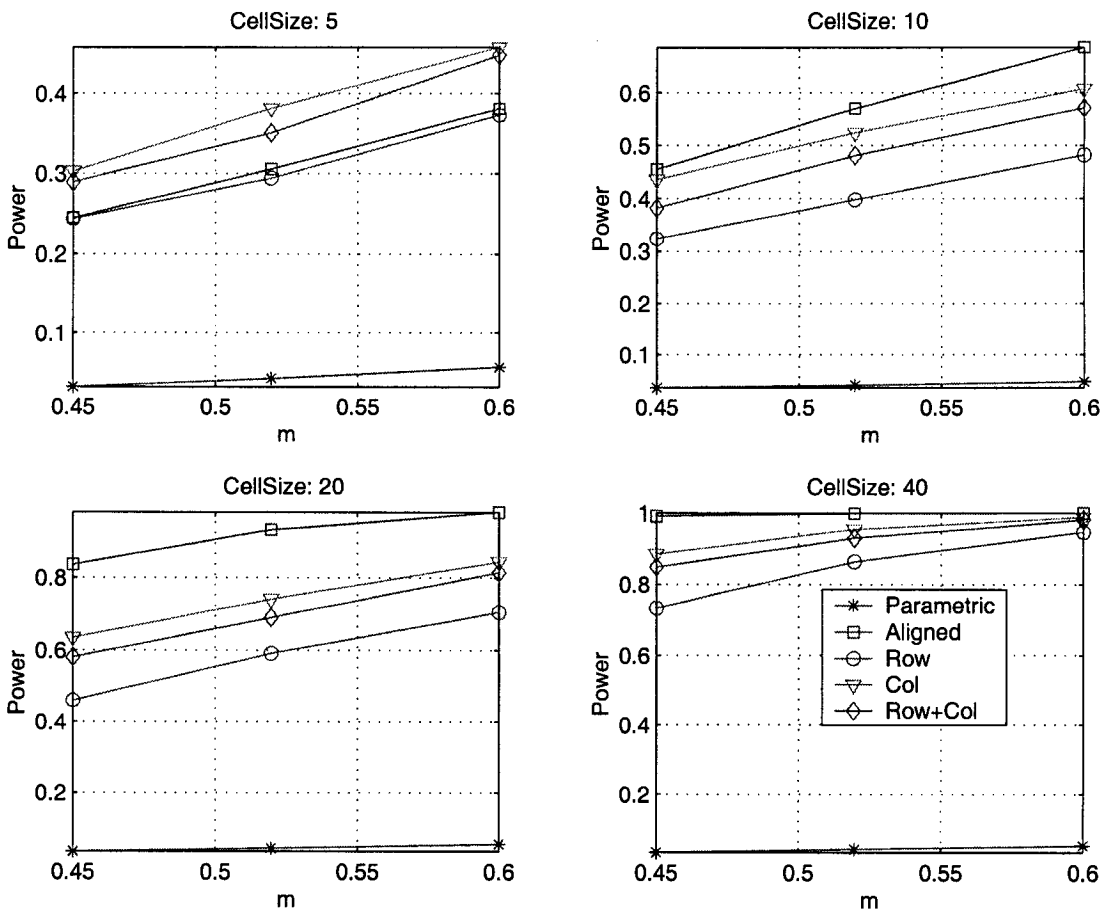

Figure 10: Power for balanced designs with cell sizes ranging from 5 to 40. Noises are simulated from the Cauchy distribution.

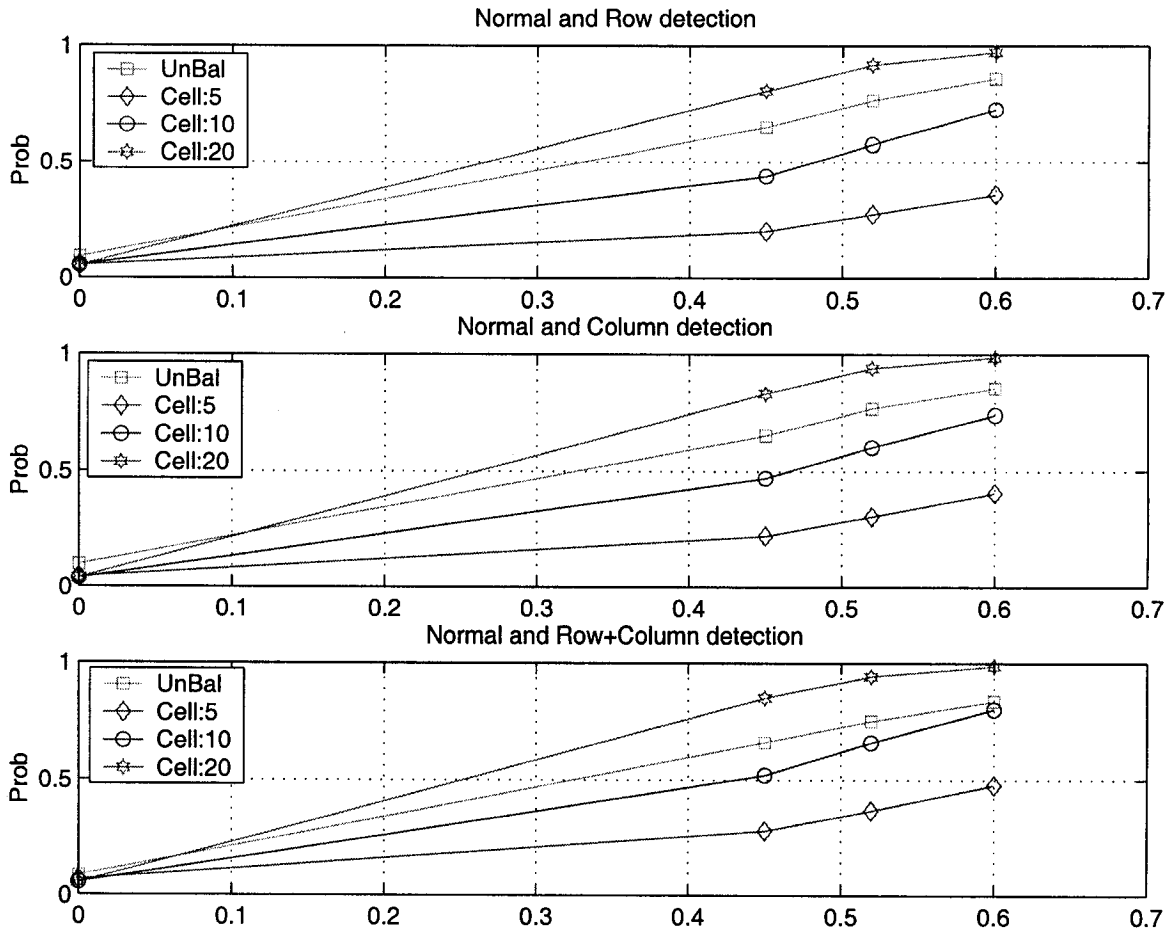

Figure 11: The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the normal distribution.

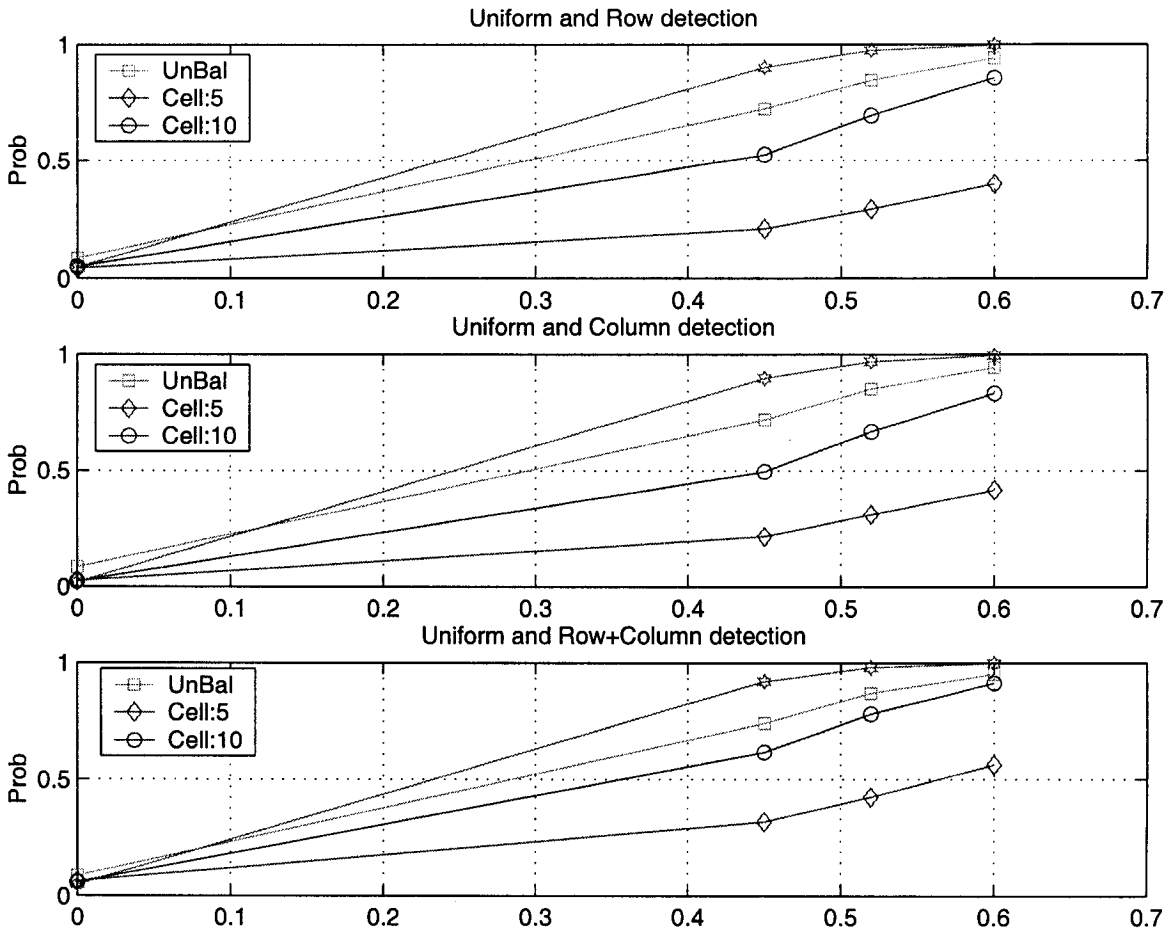

Figure 12: The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the uniform distribution.

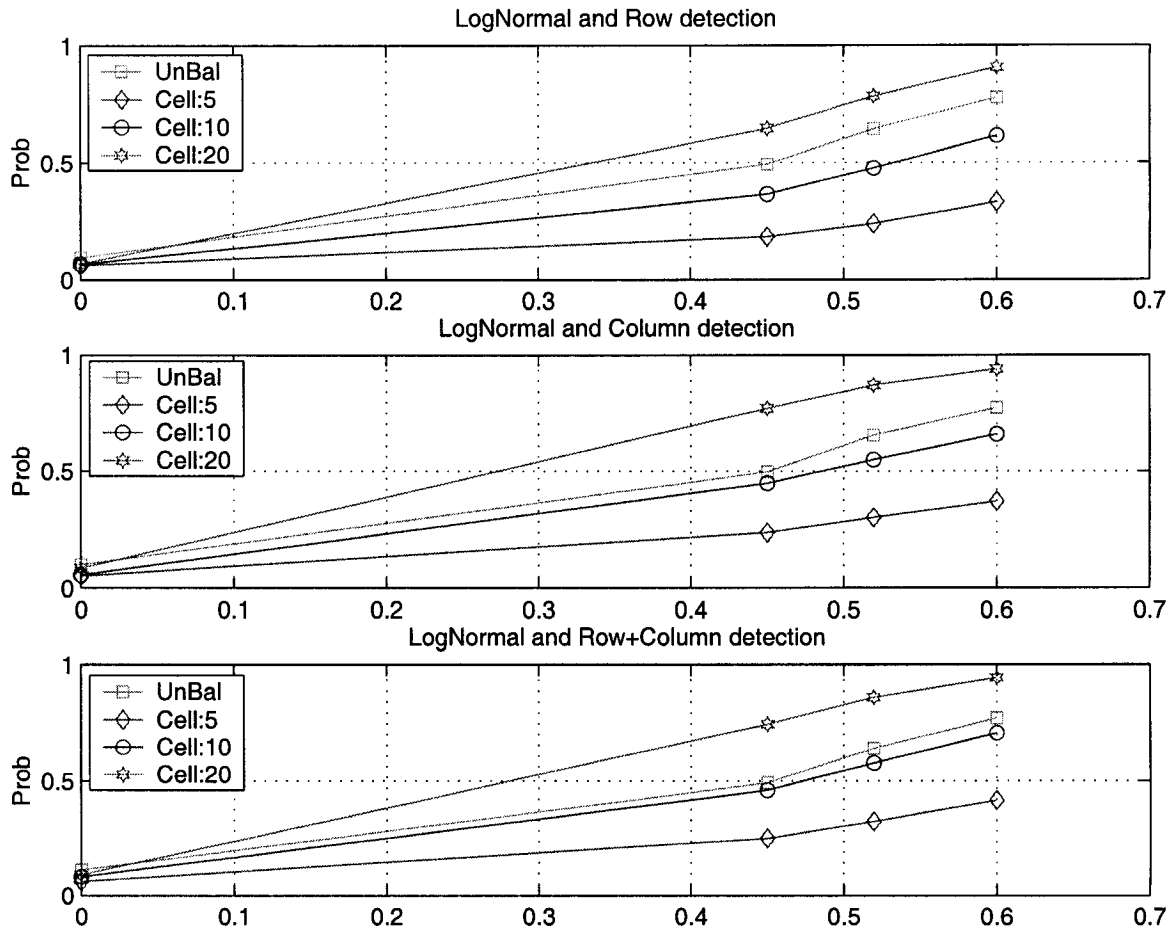

Figure 13: The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the Lognormal distribution.

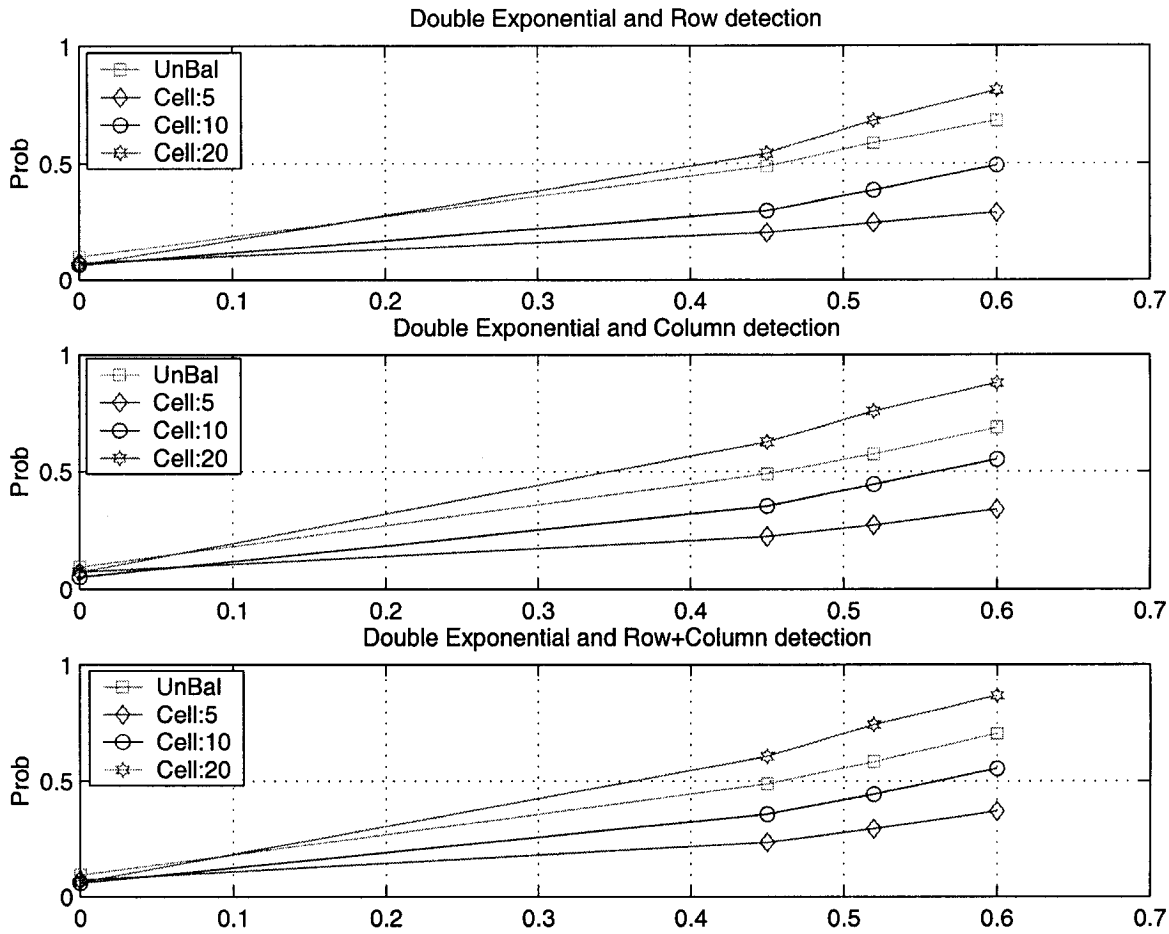

Figure 14: The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the double exponential distribution.

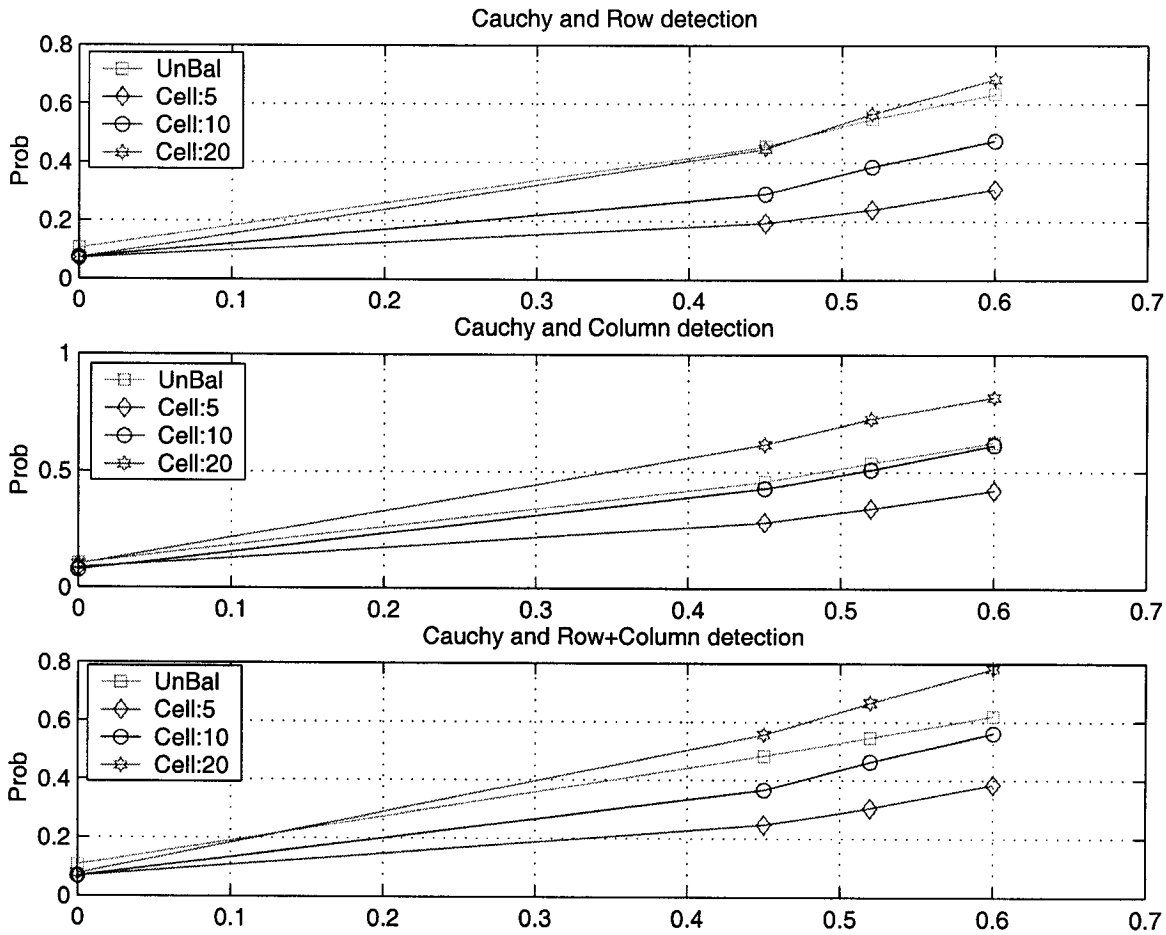

Figure 15: The power curves for unbalanced design with three other balanced designs of cell sizes equal to 5, 10 and 20. Three diagrams are provided for the row method, column method and row-column method respectively. Noises are simulated from the Cauchy distribution.

Chapter 7

Application and Discussion

7.1 Application of rank tests on gene interaction when genotypes are known

Human genome consists of 23 pairs of chromosomes carrying the genetic codes. A sequence of genetic material essential for a specific physiological or biological function is called a gene. By the Law of Inheritance, genes match in pairs. One inherited from the father, the other is inherited from the mother. When the gamete (egg or sperm) is formed, one of the two genes is randomly chosen and passed on to the offspring. This is the phenomenon of random segregation. For each gene, there exist different types of the gene called alleles. Genotype stands for the unordered pairs of alleles carried by individuals, such as AA, BA, AO. Homozygotes refer to individuals with 2 copies

of the same alleles, while heterozygotes refer to individuals with different copies. An important concept in statistical genetics is the Law of Hardy-Weinberg Equilibrium (HWE). Under the assumption of finite population, random mating, no selection, no migration, equal genotype frequencies for two sexes, the allele frequencies and the genotype frequencies in the population remain constant after one generation of random mating. Assume that we have S alleles, $A_1, A_2, \ldots, A_s$, according to the HWE, the genotype frequencies are $P_{A_iA_j} = P_{A_i}^2$, when $i = j$, and $P_{A_iA_j} = 2P_{A_i}P_{A_j}$,when $i \neq j$. Let us assume there are two genes A and B which influence the trait of interest. Gene A has only two alleles A and a with frequencies P_A and P_a for allele A,a, respectively. $P_A + P_a = 1$. Similarly we have diallelic gene B with allele frequencies P_B, P_b for allele B and b. Under an additive model, the genetic contribution to the quantitative trait can be written as:

$$X_{ijn} = \theta + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijn},$$

while i represents the genotype of gene A, j for gene B and ϵ_{ijn} stands for the noise. Let

$$\alpha_i = \begin{cases} a & \text{if genotype is AA,} \\ 0 & \text{if genotype is Aa,} \\ -a & \text{if genotype is aa.} \end{cases}$$

$$\beta_j = \begin{cases} b & \text{if genotype is BB,} \\ 0 & \text{if genotype is Bb,} \\ -b & \text{if genotype is bb.} \end{cases}$$

If the two genes are acting in an additive and independent manner, then, in the model, $\gamma_{ij} = 0$, for all i, j. If the two genes are interacting with each other, then, $\gamma_{ij} \neq 0$, for some i, j. For instance, if the genotype AABB causes the increase effect greater than the sum of their own contribution, then $\gamma_{11} > 0$. This can happen when AA and BB interact and cause a dramatic effect on the biological pathway greater than the linear effect expected if they behave independently. Detecting the gene interaction will shed light onto the understanding of the disease development mechanism at physiological and biological level. [We refer the reader to Lander and Green [13], Kruglyak and Lander [10], Wang et al [26] for more detailed introduction on linkage analysis and quantitative trait (QTL) mapping.]

If the genotypes of gene A and B are accessible, the problem becomes the hypothesis testing of interaction in repeated measurement two-way layout design. However, before we collect samples, we can not specify the number of replicates for each genotype combination, because the genotypes can only be determined after the samples are collected and analyzed in the laboratory by the genotyping machine. Unlike a classical experiment design, we do not have the option of choosing equal replicates. The detection of gene interaction remains largely unsolved theoretically. In this thesis

genotypes	BB	Bb	bb
AA	$p_a^2 p_b^2$	$2p_a^2 p_b(1-p_b)$	$p_a^2(1-p_b)^2$
Aa	$2p_a(1-p_a)p_b^2$	$4p_a(1-p_a)p_b(1-p_b)$	$2p_a(1-p_a)(1-p_b)^2$
aa	$(1-p_a)^2 p_b^2$	$2(1-p_a)^2 p_b(1-p_b)$	$(1-p_a)^2(1-p_b)^2$

Table 4: Genotype combinations frequencies at locus A and B

we are trying to propose a framework to analyze this type of data. The phenotypic contribution of gene A can be regarded as the row effect, and contribution of gene B can be regarded as the column effect. In Table 4, we represent the proportion of people with a certain genotype combination under the Hardy-Weinberg equilibrium. Let

$$w_{i1} = p_b^2, w_{i2} = 2p_b(1-p_b), w_{i3} = (1-p_b)^2, \forall i.$$

Therefore when we sample large number of individuals, the cell replicates for each genotype combination will satisfy the mild condition specified by Theorem 5.1.1. If the sample size is small, the cell frequency for each genotype combination will deviate from this frequency table. Nevertheless, we can use the test statistics developed for the unbalanced designs with arbitrary cell weights.

Using the extended method which can deal with this type of unbalanced design is more powerful than trimming down all the cell sizes to be equal to the n_{min}. Since by trimming, the number of total observations used in the ranking are decreased, the power will be affected significantly.

7.2 Discussion

Testing gene-gene interaction has been an important issue in the field of statistical genetics, but it remains largely unresolved due to the inherent difficulties residing in the methodology of testing interaction and the complex nature of human genetic data. In this thesis, we have explored a new framework to transform the biological problem into a statistical hypothesis testing and parameter estimation problem. When genotypes are accessible, we propose to test gene interaction as testing interaction in an unbalanced repeated measurement design. The traditional analysis of variance method has considerable appeal, in that it has good power under Gaussian noise. However, there is a need to provide a robust nonparametric method for testing interaction which makes no normality assumption. Aligned tests proceed by subtracting estimated row and column effect from the observations prior to ranking, thus they are not invariant under monotone transformations. The rank transform method is another popular but controversial test. All the observations are ranked together and the classical analysis of variance test is applied to these ranks. The rank transform has been proven to be an invalid test for interaction when the number of rows and columns each exceeds 2 and both main effects are present. Furthermore, the aforementioned tests are not applicable to unbalanced designs which constrains their use in testing gene-gene interaction. There are several other tests available which are based on innovative new definitions of interaction or limited to the situations when one of

the main effect is absent. In this paper, a new rank test which overcomes all the difficulties above is proposed. It has satisfactory type I error rate and superior power in all non-normal situations. It can be easily extended to accommodate unbalanced designs. As none of other competing methods are applicable to unbalanced designs with arbitrary cell weight, we believe that the new test statistics proposed here will be useful in testing for interaction in genetic setting. In regards to which of the three proposed methods should be employed in practice, we strongly recommend the use of W_3 statistic because it captures the rank discordance in both rows and columns caused by interaction and in most situations it is more powerful than W_1 and W_2. Nevertheless, we also admit that depending on the parameters' setting, there exist situations when W_3 could be outperformed by W_1 and W_2.

Bibliography

[1] M. G. Akritas, S. F. Arnold & E. Brunner (1997). Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs. *Journal of the American Statistical Association*, 92, 258-265.

[2] V. P. Bhapkar & A. P. Gore (1974). A Nonparametric Test for Interaction in Two-Way Layouts. *Sankhya: The Indian Journal of Statistics*, 36, 261-272.

[3] R.C. Blair, S.S. Sawilowsky & J.J. Higgins (1987). Limitations of the Rank Transform Statistic in Tests for Interactions. *Communications in Statistics. - Simulation*, 16, 1133-1145.

[4] D. F. Crouse (1967). A Class of Distribution-Free Analysis of Variance Testes. *South African Statistical Journal*, 1, 75-80.

[5] M. A. Fligner (1981). Comments on 'Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics'. *The American Statistician*, 35, 131-132.

[6] S. C. Hora & R. L. Iman (1988). Asymptotic Relative Efficiencies of the Rank-Transformation Procedure in Randomized Complete Block Designs. *Journal of the American Statistical Association*, 83, 462-470.

[7] J. Hájek & Z. Sidak (1967). *Theory of Rank Tests*, Academic Press, New York

[8] J. Hájek (1968). Asymptotic Normality of Simple Linear Rank Statistics Under Alternatives. *The Annals of Mathematical Statistics*, 39, 325-346.

[9] J. De. Kroon & P. Van. Der Lann (1981). Distribution-Free Test Procedures in Two-way Layouts: A Concept of Rank-interaction. *Statistica Neerlandica*, 35, 189-213.

[10] L. Kruglyak & E. S. Lander (1995). Complete Multipoint Sib-Pair Analysis of Qualitative and Quantitative Traits. *American Journal of Human Genetics*, 57, 439-454.

[11] L. Kruglyak & E. S. Lander (1995). A Nonparametric Approach for Mapping Quantitative Trait Loci. *Genetics*, 139, 1421-1428.

[12] E. L. Lehmann (1963). Robust Estimation in Analysis of Variance. *Annals of Mathematical Statistics*, 34, 957-966.

[13] E. S. Lander & P. Green (1987). Construction of Multilocus Genetics Linkage Maps in Humans. *Proceedings of National Academy of Science. USA*, 84, 2363-2367.

[14] H. H. Lemmer & D. J. Stoker (1967). A Distribution-Free Analysis of Variance for the Two-Way Classification. *South African Statistical Journal*, 1, 67-71.

[15] H. Mansouri & Z. Govindarajulu (1990). A Class of Rank Tests for Interaction in Two-Way Layouts. *Journal of Applied Statistics*, 17, 417-426.

[16] J. I. Marden & M. E. T. Muyot (1995). Rank Tests for Main and Interaction Effects in Analysis of Variance. *Journal of the American Statistical Association*, 90, 1388-1398.

[17] K.L. Mehra & P.K. Sen (1969). On a Class of Conditionally Distribution-Free Tests for Interaction in Factorial Experiments. *Annals of Mathematical Statistics*, 40, 658-664.

[18] K. M. Patel & D. G. Hoel (1973). A Nonparametric Test for Interaction in Factorial Experiments. *Journal of the American Statistical Association*, 68, 615-620.

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

[20] C. R. Rao & S. K. Mitra (1971). *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York

[21] P. K. Sen (1968). On a Class of Aligned Rank Order Tests in Two-Way Layouts. *The Annals of Mathematical Statistics*, 39, 1115-1124.

[22] G. L. Thompson & L. P. Ammann (1989). Efficacies of Rank-Transform Statistics in Two-Way Models with No Interaction. *Journal of the American Statistical Association*, 84, 325-330.

[23] G.L.Thompson and L.P. Ammann (1990). Efficiencies of Interblock Rank Statistics for Repeated Measures Designs. *Journal of the American Statistical Association*, 85, 519-528.

[24] G. L.Thompson (1991). A Unified Approach to Rank Tests for Multivariate and Repeated Measures Designs. *Journal of the American Statistical Association*, 86, 410-419.

[25] G. L. Thompson (1991). A Note on the Rank Transform for Interactions. *Biometrika*, 78, 697-701.

[26] D. Wang, S. Lin, R. Cheng, X. Gao & F. Wright (2001). Transformation of Sib-Pair Values for the Haseman-Elston Method. *American Journal of Human Genetics*, 68, 1238-1249.

[27] J. F. Xu, D. A. Meyers, C. Ober, M. N. Blumenthal, B. Mellen, K. C. Barnes, R. A. King, L. A. Lester, T. D. Howard, J. Solway, C. D. Langefeld, T. H. Beaty, S. S. Rich, E. R. Bleecker, N. J. Cox, & the Collaborative Study on the Genetics of Asthma (2001). Genomewide Screen and Identification of Gene-Gene Interactions for Asthma-Susceptibility Loci in Three U.S. Populations: Collaborative Study on the Genetics of Asthma. *American Journal of Human Genetics*, 68, 1437-1446.

7.3 Appendix

The following codes are Splus programs to run the simulations.

```
varxin<-function(X,Y){
  if(length(which.na(X))==0){
       X1<-X
  }
  else {
    X1<-X[-which.na(X)]
  }
  if(length(which.na(Y))==0){
    Y1<-Y
  }
  else {
    Y1<-Y[-which.na(Y)]
  }
  if(length(X1)!=length(Y1)){
    a<-NA
  }
  else {
     a<-sum((X1-sum(X1)/length(X1))*(Y1-sum(Y1)/length(Y1)))/(length(Y1)-1)
```

```
  }
  return (a)
}

varVecNormal<-function(X){
  if(length(which.na(X))==0){
    return(sum(((X-sum(X)/length(X))^2)/(length(X)-1)))
  }
  else {
    X1<-X[-which.na(X)]
    return(sum(((X1-sum(X1)/length(X1))^2)/(length(X1)-1)))
  }
}
tomsonImp<-
function(R, C, N, dat, const, amat, bmat)
{
    rank1 <- array(rank(dat), dim = c(R, C, N))
    rank1 <- rank1/(R * C * N + 1)
    Q1 <- 0
    meanrank1 <- mean(rank1)
```

```
for(j in 1:C) {
    meanrank1j <- mean(rank1[, j, ])
    for(i in 1:R) {
        Q1 <- Q1 + (mean(rank1[i, j, ]) - mean(
            rank1[i, , ]) - meanrank1j +
            meanrank1)^2
    }
}
Q1 <- Q1 * N
D1 <- 0
for(j in 1:C) {
    for(i in 1:R) {
        tt1 <- rank1[i, j, ] - mean(rank1[i, j, ])
        D1 <- D1 + sum(tt1 * tt1)
    }
}
D1 <- D1/(R * C * N - R * C)
Q3 <- 0
meandat <- mean(dat)
for(j in 1:C) {
```

```
        meandatj <- mean(dat[, j,  ])
        for(i in 1:R) {
            Q3 <- Q3 + (mean(dat[i, j,  ]) - mean(dat[i,
                , ]) - meandatj + meandat)^2
        }
    }
    Q3 <- Q3 * N
    D3 <- 0
    for(j in 1:C) {
        for(i in 1:R) {
            tt3 <- dat[i, j,  ] - mean(dat[i, j,  ])
            D3 <- D3 + sum(tt3 * tt3)
        }
    }
    D3 <- D3/(R * C * N - R * C)
    dat5 <- array(dim = c(R, C, N))
    for(j in 1:C) {
        mediandatj <- LehmanEst(as.vector(dat[, j,  ]))
        for(i in 1:R) {
            dat5[i, j,  ] <- dat[i, j,  ] - LehmanEst(
```

```
            as.vector(dat[i,  ,  ])) -
            mediandatj
    }
}
rank5 <- array(rank(dat5), dim = c(R, C, N))
rank5 <- rank5/(R * C * N + 1)
Q5 <- 0
meanrank5 <- mean(rank5)
for(j in 1:C) {
    meanrank5j <- mean(rank5[, j,  ])
    for(i in 1:R) {
        Q5 <- Q5 + (mean(rank5[i, j,  ]) - mean(
            rank5[i,  ,  ]) - meanrank5j +
            meanrank5)^2
    }
}
Q5 <- Q5 * N
D5 <- 0
for(j in 1:C) {
    for(i in 1:R) {
```

```
            tt5 <- rank5[i, j,  ] - mean(rank5[i, j,  ])
            D5 <- D5 + sum(tt5 * tt5)
        }
    }
    D5 <- D5/(R * C * N - R * C)
    T1 <- Q1/D1
    T3 <- Q3/(D3 * (R - 1) * (C - 1))
    T5 <- Q5/D5
    return(c(pchisq(T1, (R * C - R - C + 1)), pf(T3, (R - 1) * (
        C - 1), R * C * (N - 1)), pchisq(T5, (R * C - R - C +

        1))))

}
LehmanEst<-
function(X)
{
    L <- length(X)
    Y <- numeric((L * (L + 1))/2)
    startL <- 1
```

```
    for(i in 1:L) {
        endL <- startL + L - i
        Y[startL:endL] <- X[i] + X[i:L]
        startL <- startL + L - i + 1
    }
    return(0.5 * median(Y))
}
generateAmat<-
function(R, C, N)
{
    amat <- array(dim = c(R * C, R * C))
    for(k in 1:(R * C)) {
        Rowk <- ceiling(k/C)
        Colk <- k - (Rowk - 1) * C
        for(l in 1:(R * C)) {
            Rowl <- ceiling(l/C)
            Coll <- l - (Rowl - 1) * C
            if((Rowk == Rowl) && (Colk == Coll)) {
                amat[k, l] <- (R - 1)/R
            }
```

```
            else if(Colk == Coll) {
                amat[k, l] <- -1/R
            }
            else {
                amat[k, l] <- 0
            }
        }
    }
    return(amat)
}
generateBmat<-
function(R, C, N)
{
    bmat <- array(dim = c(R * C, R * C))
    for(k in 1:(R * C)) {
        Rowk <- ceiling(k/C)
        Colk <- k - (Rowk - 1) * C
        for(l in 1:(R * C)) {
            Rowl <- ceiling(l/C)
            Coll <- l - (Rowl - 1) * C
```

```
            if((Rowk == Rowl) && (Colk == Coll)) {
                bmat[k, l] <- (C - 1)/C
            }
            else if(Rowk == Rowl) {
                bmat[k, l] <- -1/C
            }
            else {
                bmat[k, l] <- 0
            }
        }
    }
    return(bmat)
}
makedat<-
function(R, C, N, alpha, beta, std)
{
    dat <- array(dim = c(R, C, N))
    temp <- numeric(N)
    for(i in 1:R) {
        for(j in 1:C) {
```

```
            dat[i, j,  ] <- alpha[i] + beta[j] + rnorm(N,
                0, std)
        }
    }
    return(dat)
}
makeintedat.beta<-
function(R, C, N, alpha, beta, shape1, shape2, scale, interact)
{
    dat <- array(dim = c(R, C, N))
    for(i in 1:R) {
        for(j in 1:C) {
            dat[i, j,  ] <- (rbeta(N, shape1, shape2) -
                0.5) * scale + alpha[i] + beta[j] +
                interact[i, j]
        }
    }
    return(dat)
}
makeintedat.lognormal<-
```

```
function(R, C, N, alpha, beta, std, interact)
{
    dat <- array(dim = c(R, C, N))
    for(i in 1:R) {
        for(j in 1:C) {
            dat[i, j,  ] <- rlnorm(N, std) + alpha[i] +
                beta[j] + interact[i, j]
        }
    }
    return(dat)
}
makeintedat.cauchy<-
function(R, C, N, alpha, beta, std, interact)
{
    dat <- array(dim = c(R, C, N))
    for(i in 1:R) {
        for(j in 1:C) {
            dat[i, j,  ] <- rcauchy(N, 0, std) + alpha[i
                ] + beta[j] + interact[i, j]
        }
```

```
    }
    return(dat)
}
makeintedat<-
function(R, C, N, alpha, beta, std, interact)
{
    dat <- array(dim = c(R, C, N))
    for(i in 1:R) {
        for(j in 1:C) {
            dat[i, j,  ] <- rnorm(N, 0, std) + alpha[i] +

                beta[j] + interact[i, j]
        }
    }
    return(dat)
}
makeintedat.dexp<-
function(R, C, N, alpha, beta, rate, interact)
{
    dat <- array(dim = c(R, C, N))
```

```
    for(i in 1:R) {
        for(j in 1:C) {
            dat[i, j,  ] <- rexp(N, rate) - rexp(N, rate
                ) + alpha[i] + beta[j] + interact[i,
                j]
        }
    }
    return(dat)
}
tomsonaddrc<-
function(R, C, N, dat, const, amat, bmat)
{
    rank1 <- array(rank(dat), dim = c(R, C, N))
    rank2 <- array(dim = c(R, C, N))
    rank3 <- array(dim = c(R, C, N))
    for(i in 1:R) {
        rank2[i,  ,  ] <- array(rank(dat[i,  ,  ]), dim = c(
            C, N))
    }
    for(j in 1:C) {
```

```
    rank3[, j,  ] <- array(rank(dat[, j,  ]), dim = c(R,
        N))
}
rank1 <- rank1/(R * C * N + 1)
rank2 <- rank2/(C * N + 1)
rank3 <- rank3/(R * N + 1)
temp <- array(dim = c(R * C, N))
temp1 <- array(dim = c(R * C, N))
for(i in 1:R) {
    for(j in 1:C) {
        temp[(i - 1) * C + j,  ] <- rank2[i, j,  ] -
            mean(rank2[, j,  ])
        temp1[(i - 1) * C + j,  ] <- rank2[i, j,  ]
    }
}
temp2 <- array(dim = c(R * C, N))
temp3 <- array(dim = c(R * C, N))
for(i in 1:R) {
    meanrank3i<-mean(rank3[i,  ,  ])
    for(j in 1:C) {
```

```
            temp2[(i - 1) * C + j,  ] <- rank3[i, j,  ] -meanrank3i
            temp3[(i - 1) * C + j,  ] <- rank3[i, j,  ]
        }
    }
    cmatrix <- array(dim = c(R, C, C, N))
    sigma4 <- array(0,dim = c(R * C, R * C))
    for(i in 1:R) {
        for(j in 1:C) {
            semp1 <- array(rep(dat[i, j,  ], each = N), dim = c(N, N))
            if(j == 1) {
               emp <- dat[i, (j + 1):C,  ]
                  }
            else if(j == C) {
               emp <- dat[i, 1:(j - 1),  ]
                  }
            else {
               emp <- c(dat[i, 1:(j - 1),  ],dat[i, (j + 1):C,  ])
                  }
            cmatrix[i, j, j,  ] <- apply((dat[i, j,  ] >=
            (array(rep(emp, each = N),
```

```
dim = c(N, length(emp))))), 1, sum)/N
            for(b in 1:C) {
                if(b != j) {
                  cmatrix[i, j, b,  ] <- ( - apply((dat[i, b,  ]
                  >= semp1), 1, sum))/N
                }
            }
        }
    }
    for(k in 1:(R * C)) {
        Rowk <- ceiling(k/C)
        Colk <- k - (Rowk - 1) * C
        for(l in k:(R * C)) {
            Rowl <- ceiling(l/C)
            Coll <- l - (Rowl - 1) * C
            if((Rowk == Rowl) && (Colk == Coll)) {
                phivar <- cmatrix[Rowk, Colk,  ,  ]
                sigma4[k, l] <- sum(apply(phivar, 1, varVecNormal))
            }
            else if(Rowk == Rowl) {
```

```
            sigma4[k, l] <- sum(apply(cbind(
              cmatrix[Rowk, Colk,  ,  ], cmatrix[
              Rowk, Coll,  ,  ]), 1, varxinApply))
        }
        sigma4[l, k] <- sigma4[k, l]
    }
}
sigma4 <- sigma4/C^2
Impsigma8 <- amat %*% sigma4 %*% t(amat)
invsigma8 <- ginverse(Impsigma8)
vec <- apply(temp, 1, sum)
T8 <- vec %*% invsigma8 %*% vec/N
dmatrix <- array(dim = c(R, C, R, N))
sigma5 <- array(0,dim = c(R * C, R * C))
for(j in 1:C) {
    for(i in 1:R) {
        semp1 <- array(rep(dat[i, j,  ], each = N), dim = c(N, N))
        if(i == 1) {
            emp <- dat[(i + 1):R, j,  ]
        }
```

```
            else if(i == R) {
                emp <- dat[1:(i - 1), j,  ]
            }
            else {
                emp <- c(dat[1:(i - 1), j,  ],  dat[(i + 1):R, j,  ])
            }
            emp1 <- array(rep(emp, each = N),dim = c(N, length(emp)))
            dmatrix[i, j, i,  ] <- apply((dat[i, j,  ] >= emp1),1, sum)/N
            for(b in 1:R) {
                if(b != i) {
                  dmatrix[i, j, b,  ] <- ( - apply((dat[b, j,  ]
                  >= semp1), 1, sum))/N
                }
            }
        }
    }
    for(k in 1:(R * C)) {
        Rowk <- ceiling(k/C)
        Colk <- k - (Rowk - 1) * C
        for(l in k:(R * C)) {
```

```
            Rowl <- ceiling(l/C)
            Coll <- l - (Rowl - 1) * C
            if((Rowk == Rowl) && (Colk == Coll)) {
                phivar <- dmatrix[Rowk, Colk,  ,  ]
                sigma5[k, l] <- sum(apply(phivar, 1, varVecNormal))
            }
            else if(Colk == Coll) {
                sigma5[k, l] <- sum(apply(cbind(
                  dmatrix[Rowk, Colk,  ,  ], dmatrix[
                  Rowl, Colk,  ,  ]), 1, varxinApply))
            }
            sigma5[l, k] <- sigma5[k, l]
        }
    }
    sigma5 <- sigma5/R^2
    Impsigma9 <- bmat %*% sigma5 %*% t(bmat)
    invsigma9 <- ginverse(Impsigma9)
    vec <- apply(temp2, 1, sum)
    T9 <- vec %*% invsigma9 %*% vec/N
    cosig <- array(dim = c(R * C, R * C))
```

```
for(k in 1:(R * C)) {
    Rowk <- ceiling(k/C)
    Colk <- k - (Rowk - 1) * C
    for(l in 1:(R * C)) {
        Rowl <- ceiling(l/C)
        Coll <- l - (Rowl - 1) * C
        if((Rowk == Rowl) && (Colk == Coll)) {
            cosig[k, l] <- varxin(cmatrix[Rowk,
              Colk, Colk,  ], dmatrix[Rowk, Colk,
              Rowk,  ])
        }
        else if(Rowk == Rowl) {
            cosig[k, l] <- varxin(cmatrix[Rowk,
              Colk, Coll,  ], dmatrix[Rowk, Coll,
              Rowk,  ])
        }
        else if(Colk == Coll) {
            cosig[k, l] <- varxin(cmatrix[Rowk,
              Colk, Colk,  ], dmatrix[Rowl, Colk,
              Rowk,  ])
```

```
            }
            else {
                cosig[k, l] <- varxin(cmatrix[Rowk,
                  Colk, Coll,  ], dmatrix[Rowl, Coll,
                  Rowk,  ])
            }
        }
    }
    cosig <- cosig/(R * C)
    Impsigma6 <- amat %*% sigma4 %*% t(amat) +
    (bmat %*% sigma5 +2 * amat %*% cosig) %*% t(bmat)
    invsigma6 <- ginverse(Impsigma6)
    vec1 <- apply(temp + temp2, 1, sum)
    T10 <- vec1 %*% invsigma6 %*% vec1/N
    return(c(pchisq(T8, (R * C - R - C + 1)),
pchisq(T9, (R * C -R - C + 1)), pchisq(T10, (R * C - R - C + 1))))
}
varApply<-
function(X)
{
```

```
    L1 <- length(X)/2
    return(var(X[1:L1], X[(L1 + 1):length(X)]))
}
varxinApply<-
function(X)
{
    L1 <- length(X)/2
    return(varxin(X[1:L1], X[(L1 + 1):length(X)]))
}
coefunr<-function(R, C, Nmat){
    coef<-vector("list", R*C)
    for(i in 1:R) {
        for(j in 1:C) {
            coef[[(i-1)*C+j]]<-NULL
            for (a in 1:C){
              if (a!=j){
              coef[[(i-1)*C+j]]<-c(coef[[(i-1)*C+j]],
              -rep((Nmat[i,j]+Nmat[i,a]+1)/
(Nmat[i,a]*(Nmat[i,j]+Nmat[i,a])),Nmat[i,a]))
              }
```

```
            }
        }
    }
return(coef)
}
colrestack<-function(dat,dataIndexS,dataIndexE,R, C, Nmat){
datcol <- numeric(sum(Nmat))
dataIndexScol<-array(1:(R*C),dim=c(R,C))
dataIndexEcol<-array(1:(R*C),dim=c(R,C))
start<-1
    for(j in 1:C) {
        for(i in 1:R) {
            end<-start+Nmat[i,j]-1
            datcol[start:end]<-dat[dataIndexS[i,j]:dataIndexE[i,j]]
            dataIndexScol[i,j]<-start
            dataIndexEcol[i,j]<-end
            start<-start+Nmat[i,j]
        }
    }
return(list(datcol,dataIndexScol,dataIndexEcol))
```

```
}
makeintedatHong.gene<-
function(R, C, Nmat, alpha, beta, std, interact)
{

    data <- numeric(sum(Nmat))
    dataIndexS<-array(1:(R*C),dim=c(R,C))
        dataIndexE<-array(1:(R*C),dim=c(R,C))
        start<-1
    for(i in 1:R) {
        for(j in 1:C) {
            end<-start+Nmat[i,j]-1
            data[start:end]<-rnorm(Nmat[i,j], 0, std)+
             alpha[i] +beta[j] + interact[i, j]
            dataIndexS[i,j]<-start
            dataIndexE[i,j]<-end
            start<-start+Nmat[i,j]
        }
    }
    return(list(data,dataIndexS,dataIndexE))
```

```
}
makeintedatHong.genebeta<-
function(R, C, Nmat, alpha, beta, std, interact)
{

    data <- numeric(sum(Nmat))
    dataIndexS<-array(1:(R*C),dim=c(R,C))
        dataIndexE<-array(1:(R*C),dim=c(R,C))
        start<-1
    for(i in 1:R) {
        for(j in 1:C) {
            end<-start+Nmat[i,j]-1
            data[start:end]<-(rbeta(Nmat[i,j], 1, 1) -
                0.5) * std+ alpha[i] +beta[j] + interact[i, j]
            dataIndexS[i,j]<-start
            dataIndexE[i,j]<-end
            start<-start+Nmat[i,j]
        }
    }
    return(list(data,dataIndexS,dataIndexE))
```

```
}
makeintedatHong.genelog<-
function(R, C, Nmat, alpha, beta, std, interact)
{

    data <- numeric(sum(Nmat))
    dataIndexS<-array(1:(R*C),dim=c(R,C))
        dataIndexE<-array(1:(R*C),dim=c(R,C))
        start<-1
    for(i in 1:R) {
        for(j in 1:C) {
            end<-start+Nmat[i,j]-1
            data[start:end]<-rlnorm(Nmat[i,j], std)+
            alpha[i] +beta[j] + interact[i, j]
            dataIndexS[i,j]<-start
            dataIndexE[i,j]<-end
            start<-start+Nmat[i,j]
        }
    }
    return(list(data,dataIndexS,dataIndexE))
```

```
}
makeintedatHong.genecauchy<-
function(R, C, Nmat, alpha, beta, std, interact)
{

    data <- numeric(sum(Nmat))
    dataIndexS<-array(1:(R*C),dim=c(R,C))
        dataIndexE<-array(1:(R*C),dim=c(R,C))
        start<-1
    for(i in 1:R) {
        for(j in 1:C) {
            end<-start+Nmat[i,j]-1
            data[start:end]<-rcauchy(Nmat[i,j], 0, std)+
             alpha[i] +beta[j] + interact[i, j]
            dataIndexS[i,j]<-start
            dataIndexE[i,j]<-end
            start<-start+Nmat[i,j]
        }
    }
    return(list(data,dataIndexS,dataIndexE))
```

```
}
makeintedatHong.genedexp<-
function(R, C, Nmat, alpha, beta, std, interact)
{

    data <- numeric(sum(Nmat))
    dataIndexS<-array(1:(R*C),dim=c(R,C))
        dataIndexE<-array(1:(R*C),dim=c(R,C))
        start<-1
    for(i in 1:R) {
        for(j in 1:C) {
            end<-start+Nmat[i,j]-1
            data[start:end]<-rexp(Nmat[i,j], std) -
            rexp(Nmat[i,j], std
                )+ alpha[i] +beta[j] + interact[i, j]
            dataIndexS[i,j]<-start
            dataIndexE[i,j]<-end
            start<-start+Nmat[i,j]
        }
    }
```

```
    return(list(data,dataIndexS,dataIndexE))
}
cweight<-
function(R, C, Nmat)
{
    coef <- vector("list", R * C)
    for(i in 1:R) {
        for(j in 1:C) {
            coef[[(i - 1) * C + j]] <- NULL
            for(a in 1:C) {
                if(a != j) {
                  coef[[(i - 1) * C + j]] <-
                  c(coef[[(i - 1) * C + j]], rep(1/Nmat[
                    i, a], Nmat[i, a]))
                }
            }
        }
    }
    return(coef)
}
```

```
unbrow<-
function(R, C, Nmat, dat, datIndexS, datIndexE,coef,amat,cwe)
{   df<-(R-1)*(C-1)
    rankVec <- numeric(R*C)
    sigma <- array(dim = c(R*C, R * C))
    cmatrix <- array(dim = c(R, C, C, max(Nmat)))
    for(i in 1:R) {
        for(j in 1:C) {
            semp <- dat[datIndexS[i, j]:datIndexE[i,j]]
            self<-array(rep(semp, each = Nmat[i,j]),
                    dim = c(Nmat[i,j],Nmat[i,j]))
            if(j == 1) {
              emp <- dat[datIndexS[i, (j+1)]:datIndexE[i,C]]
                  }
            else if(j == C) {
              emp <- dat[datIndexS[i, 1]:datIndexE[i,(j-1)]]
                  }
            else {
              emp <- c(dat[datIndexS[i, 1]:datIndexE[i,(j-1)]]
```

```
                ,dat[datIndexS[i, (j+1)]:datIndexE[i,C]])
                }
            emp2 <- semp >= (array(rep(emp, each = Nmat[i,j]
),dim = c(Nmat[i,j], length(emp))))
            rankVec[(i-1)*C+j]<-mean(apply(cbind(
t(cwe[[(i-1)*C+j]]*t(emp2)),(semp >= self)/Nmat[i,j]),
                    1, sum))/C
            cmatrix[i,j,j,(1:Nmat[i,j])] <-
            apply(t(coef[[(i-1)*C+j]]*t(emp2)),
                    1, sum)
            for(b in 1:C) {
                if(b != j)  {
                  semp1 <- array(rep(semp, each = Nmat[i,b]),
                  dim = c(
                Nmat[i,b], Nmat[i,j]))
                  emp2 <- dat[datIndexS[i, b]:datIndexE[i,
            b]] >= semp1
                  cmatrix[i, j, b,(1:Nmat[i,b])]<- apply(
                    emp2, 1, sum)/(Nmat[i,b]*(Nmat[i,j]+
                    Nmat[i,b]))*(Nmat[i,j]+Nmat[i,b]+1)
```

```
                }
            }
        }
    }

    for(k in 1:(R * C)) {
        Rowk <- ceiling(k/C)
        Colk <- k - (Rowk - 1) * C
        for(l in k:(R * C)) {
            Rowl <- ceiling(l/C)
            Coll <- l - (Rowl - 1) * C
            sigma[k, l] <- 0
            if((Rowk == Rowl) && (Colk == Coll)) {
                sigma[k, l] <- sum(apply(cmatrix
                [Rowk, Colk,  ,  ], 1, varVecNormal)
*Nmat[Rowk,])/(Nmat[Rowk,Colk]^2)
            }
            else if(Rowk == Rowl) {
                 sigma[k, l] <- sum(apply(cbind(
                  cmatrix[Rowk, Colk,  ,  ], cmatrix[
```

```
                    Rowk, Coll,  ,  ]), 1, varxinApply)*
Nmat[Rowk,])/(Nmat[Rowk,Colk]*Nmat[Rowl,Coll])
            }
            sigma[l, k] <- sigma[k, l]
        }
    }
    sigma <- sigma/(C^2)
    Impsigma <- amat %*% sigma %*% t(amat)
    invsigma <- ginverse(Impsigma)

    T1<- t(amat%*%rankVec) %*% invsigma %*% (amat%*%rankVec)
    return(list(pchisq(T1, df),rankVec,cmatrix, sigma))
}
unball<-<-function(R, C, Nmat, dat, datIndexS, datIndexE,
cdat, cdatIndexS, cdatIndexE,
coef1,coef2,amat1,amat2,bmat,cw1,cw2){
reso<-unbrow(R, C, Nmat,dat, datIndexS, datIndexE,
coef1,amat1,cw1)
rankVec<-reso[[2]]
cmatrix<-reso[[3]]
```

```
sigma<-reso[[4]]

result<-unbrow(C, R, t(Nmat), cdat, t(cdatIndexS),

t(cdatIndexE),coef2,amat2, cw2)

cankVec<-result[[2]]

cankVec<-as.vector(t(array(cankVec,dim=c(R,C))))

dmatrix<-array(dim=c(R,C,R,max(Nmat)))

for (i in 1:R){

 for (j in 1:C){

 dmatrix[i,j,,]<-result[[3]][j,i,,]

 }

}


sigma2<-array(dim=c(R*C,R*C))

for(k in 1:(R * C)) {

        Rowk <- ceiling(k/C)

        Colk <- k - (Rowk - 1) * C

        kprime<-(Colk-1)*R+Rowk

        for(l in k:(R * C)) {

            Rowl <- ceiling(l/C)

            Coll <- l - (Rowl - 1) * C
```

```
            lprime<-(Coll-1)*R+Row1
            sigma2[k, l] <- result[[4]][kprime,lprime]
            sigma2[l, k]<-sigma2[k, l]

        }
}

cosig <- array(dim = c(R * C, R * C))
    for(k in 1:(R * C)) {
        Rowk <- ceiling(k/C)
        Colk <- k - (Rowk - 1) * C
        for(l in 1:(R * C)) {
            Row1 <- ceiling(l/C)
            Coll <- l - (Row1 - 1) * C
            uc<-Nmat[Rowk,Coll]/(Nmat[Rowk,Colk]*Nmat[Row1,Coll])
            if((Rowk == Row1) && (Colk == Coll)) {
                cosig[k, l] <- varxin(cmatrix[Rowk,
                  Colk, Colk,  ], dmatrix[Rowk, Colk,
                  Rowk,  ])*uc
            }
```

```
            else if(Rowk == Rowl) {
                cosig[k, l] <- varxin(cmatrix[Rowk,
                  Colk, Coll,  ], dmatrix[Rowk, Coll,
                  Rowk,  ])*uc
            }
            else if(Colk == Coll) {
                cosig[k, l] <- varxin(cmatrix[Rowk,
                  Colk, Colk,  ], dmatrix[Rowl, Colk,
                  Rowk,  ])*uc
            }
            else {
                cosig[k, l] <- varxin(cmatrix[Rowk,
                  Colk, Coll,  ], dmatrix[Rowl, Coll,
                  Rowk,  ])*uc
            }
        }
    }
    cosig <- cosig/(R * C)
    Impsigma3 <- amat1 %*% sigma %*% t(amat1) + (bmat %*% sigma2
+2 * amat1 %*% cosig) %*% t(bmat)
```

```
    invsigma3 <- ginverse(Impsigma3)
    T3 <- (t(amat1%*%rankVec+bmat%*%cankVec)) %*% invsigma3
%*%
(amat1%*%rankVec+bmat%*%cankVec)
    return(c(reso[[1]],result[[1]],pchisq(T3,(R-1)*(C-1))))
}
```