INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Université d'Ottawa • University of Ottawa

MULTIVARIATE NON-PARAMETRIC TESTS OF TREND IN THE PRESENCE OF MISSING DATA

By
Jincheol Park
January 2000

A Thesis
submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of
Master of Science in Mathematics[1]

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

Abstract

Statistics has become a crucial tool in various fields of life. The scope of statistical application has been expanding with high celerity. Among the many fields of applications there is an increasing interest in environmental and medical questions. One of the common questions arising in these fields is whether there is a time trend in one or more of the variables being measured.

In environmental and medical studies, data is often recorded over time in an effort to test for monotone trend. For example, we periodically measure the pH of a lake to test for trend in acidification. In monitoring recovering patients, we often look for trend in their vital signs. We may also simultaneously measure the pH of a few of lakes at close proximity or monitor the vital signs composed of several factors of a patient. In the multivariate case, correlations among variables have to be involved in the statistical procedure.

When testing for trend one may be interested in either a monotone trend or a step trend. The former assumes that the population shifts monotonically over time without specifing when the shift occurrs. The latter assumes that the observations recorded before some specific time belong to a different population from the one recorded after that time. Our interest will be focused on tests for monotone trend.

There exist parametric as well as nonparametric methods, univariate and multivariate, to test for monotone trend. Practically, it occurs more often than not that some portion of the collected data are missing. There is at present a way to analyze incomplete data in the univariate case. In this work, we introduce nonparametric multivariate test statistics to test for monotone trend in the presence of missing data and deduce some corresponding asymptotic properties.

In chapter one, we review both parametric and nonparametric univariate test statistics for monotone trend concluding with a discussion of the missing data case. Chapter two develops the theoretical background of general multivariate linear rank statistics and the necessary asymptotic tools. We review existing nonparametric multivariate test statistics for trend in chapter three. Two test statistics are considered. One is based on the Kendall similarity measure and the other is on the Spearman measure. The statistical procedures of chapter three assume that there are no missing observations. We extend, in chapter four, the test statistics to treat the incomplete data case. Specifically, Lemma 4.1.2, Theorem 4.1.1, Theorem 4.1.2 and 4.2.1 are new. The asymptotic results exploit results of Patel (1973) which extend to the multivariate case the Hájek and Šidák approach. In Chapter 5, we investigate the performance of the proposed test statistics when the sample size is small under some specific trend models.

Acknowledgement

I give thanks to the Lord, for by the grace of God I am what I am.

I wish to express my sincere appreciation to Professor Mayer Alvo, my supervisor, for his valuable guidance and help. I also wish to thank the University of Ottawa.

Contents

List of Figures

Chapter 1

Introduction

In this chapter we present some primary test statistics of time trend in the univariate case and describe their distributions under the null hypothesis of no time trend. We conclude with a discussion of the missing data case.

1.1 Univariate Trend Test

To test for monotone trend statistically, there are parametric as well as nonparametric methods. Let (t_i, X_{t_i}), $t_i < t_{i+1}$, $i = 1, ..., n$, denote samples with t_i measuring a time and X_{t_i} an observation at time t_i. Consider the null hypothesis that there is no time trend. Then the primary parametric method for trend is simple linear regression, where we assume

$$X_t = \beta_o + \beta_1 t + \epsilon_t$$

The terms (ϵ_t) represent uncorrelated error terms having mean zero and constant variance. The least squares estimator $\hat{\beta}_1$ of β_1 is $\sum_{i=1}^{n} X_{t_i}(t_i - \bar{t}) / \sum_{i=1}^{n}(t_i - \bar{t})^2$ and $\hat{\beta}_o$ of β_o is $\bar{X} - \hat{\beta}_1 \bar{t}$. Testing the null hypothesis that there is no time trend is equivalent to testing that $\beta_1 = 0$. To test it, we usually make the additional assumption that (ϵ_t) are independent and identically normally distributed with mean 0 and variance σ^2, $N(0, \sigma^2)$. If σ is known, we use the test statistic

$$\frac{\hat{\beta}_1}{\sqrt{\sigma^2 / \sum_{i=1}^{n}(t_i - \bar{t})^2}}$$

which is normally distributed under the null hypothesis. If σ is unknown, the test statistic is

$$t^* = \frac{\hat{\beta}_1}{\sqrt{\sum_{i=1}^{n}(X_{t_i} - \hat{X}_{t_i})^2/\{(n-2)\sum_{i=1}^{n}(t_i - \bar{t})^2\}}}$$

with $\hat{X}_{t_i} = \hat{\beta}_o + \hat{\beta}_1 t_i$. Under the null hypothesis, the distribution of t^* is that of a student's t-distribution with $(n-2)$ degrees of freedom.

Some nonparametric tests of trend are due to Mann-Kendall and to Spearman where it is supposed that observations are equally spaced in time. The observations are ranked from smallest to largest. The strength of a trend is then measured in terms of a similarity measure between the vector of ranks and the permutation $(1, 2, ..., n)$. More precisely, let $\mu = (\mu(1), ..., \mu(n))$ and $\nu = (\nu(1), ..., \nu(n))$ be any two permutations of the integers $(1, ..., n)$. The Spearman and Kendall distances between them are defined respectively by

$$d_S(\mu, \nu) = \frac{1}{2}\sum_{i=1}^{n}\{\mu(i) - \nu(i)\}^2$$

and

$$d_K(\mu, \nu) = \sum_{i<j}\{1 - s(\mu(i) - \mu(j))\}\{s(\nu(i) - \nu(j)\}$$

where $s(u)$ is the sign function defined by

$$s(u) = \begin{cases} -1 & u < 0 \\ 0 & u = 0 \\ 1 & u > 0 \end{cases}$$

These distances can be rewritten in terms of similarity measures $\mathcal{A}(\mu, \nu)$ as

$$d(\mu, \nu) = c - \mathcal{A}(\mu, \nu) \tag{1.1}$$

where for the Spearman and Kendall case respectively we have

$$c_S = \frac{n(n^2-1)}{12},\ \mathcal{A}_S(\mu, \nu) = \sum_{i=1}^{n}\left(\mu(i) - \frac{n+1}{2}\right)\left(\nu(i) - \frac{n+1}{2}\right)$$

$$c_K = \frac{n(n-1)}{2},\ \mathcal{A}_K(\mu, \nu) = \sum_{i<j} s(\mu(j) - \mu(i))s(\nu(j) - \nu(i))$$

The nonparametric tests for trend then reject the null hypothesis of no trend if the measure of similarity between the vector of ranks and the permutation $(1, ..., n)$ is large. For small samples, tables of the distributions of the test statistics can be used as given in Bradley (1968, p. 314, p. 364). For large samples, a normal approximation is available. According to Bradley (1968, p. 96, p. 288), the nonparametric tests due to Spearman and Kendall have an A.R.E of $(3/\pi)^{1/3}$ or .98 relative to the parametric test based on the test of β_1, when both tests are applied as tests of randomness against normal regression alternatives.

1.2 Missing Data

We may extend the notion of a measure of similarity in the presence of missing observations. In the following, we will denote an incomplete ranking of a subset of k-objects by $\mu^* = (\mu^*(1), ...\mu^*(k))$ or write this k-vector as a n-vector in which missing ranks are denoted by the symbol "-".

Definition 1.2.1 *(Alvo and Cabilio, 1991) The complete ranking μ of n objects is said to be compatible with an incomplete ranking μ^* of a subset of k of these objects, $2 \leq k \leq n$, if the relative ranking of every pair of objects ranked in μ^* coincides with their relative ranking in μ.*

To each fixed incomplete ranking μ^*, corresponds a compatibility class $C(\mu^*)$. For example, let $\mu^* = (1, 2, -)$. Then the compatibility class contains the rankings $\{(2, 3, 1), (1, 3, 2), (1, 2, 3)\}$. The distance between μ^* and ν^*, denoted by $d^*(\mu^*, \nu^*)$, is defined to be the average of all values of the distances $d(\mu_i, \nu_i)$ taken over all complete rankings μ_i, ν_i compatible with μ^* and ν^* repectively. From (1.1), we write $d^*(\mu^*, \nu^*) = c - A(\mu^*, \nu^*)$, where $A(\mu^*, \nu^*)$ is the average of the $\mathcal{A}(\mu_i, \nu_i)$ taken over all complete rankings μ_i, ν_i compatible with μ^* and ν^*.
Consider the case of two incomplete rankings μ^*, ν^* of k_1, k_2 objects respectively. For a fixed incomplete k_1-ranking μ^*, we define

$$\eta_\mu(i) = \frac{n+1}{k_1+1}\left(\mu^*(i) - \frac{k_1+1}{2}\right)\delta_\mu(i)$$

where $\delta_\mu(i)$ is 0 or 1 according to whether the object i is missing or not, and set

$$a_\mu(i,j) = \begin{cases} s(\mu^*(j) - \mu^*(i)) & \text{if both i and j are ranked} \\ 1 - 2\mu^*(i)/(k_1+1) & \text{if only i is ranked} \\ 2\mu^*(j)/(k_1+1) - 1 & \text{if only j is ranked} \\ 0 & \text{otherwise} \end{cases}$$

Define $\eta_\nu(i)$ and $a_\nu(i,j)$ in a similar way for a fixed incomplete k_2-ranking ν^*. Referring to Alvo and Cabilio (1995), in the Spearman case it can be shown that

$$A_S(\mu^*, \nu^*) = \sum_{i=1}^{n} \eta_\mu(i)\eta_\nu(i) \tag{1.2}$$

whereas in the Kendall case,

$$A_K(\mu^*, \nu^*) = \sum_{i<j} a_\mu(i,j)a_\nu(i,j) \tag{1.3}$$

For later use, it is useful to rewrite

$$d_S(\mu^*, \nu^*) = \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{(k_1+1)(k_2+1)} \sum_{i=1}^{n} u(i)v(i),$$

where $u^T = (u(1), ..., u(n))$ and $v^T = (v(1), ..., v(n))$ are respectively the augmented n-vector of μ^* and ν^* with $(k_i+1)/2$ replacing the missing observations. Recall from Alvo and Cabilio (1995) the null hypothesis H_1 whereby we assume that

1. k_1 and k_2, the number of ranked observations, are fixed with $k_1 \leq k_2$
2. the rankings for which we have incomplete data are uniformly distributed over the $n!$ permutations of $(1, ..., n)$
3. the pattern of the missing observations are randomly selected from the set of all possible patterns.

Consider test statistics A_S and A_K for H_1. Under H_1,

$$VarA_S = \frac{(n+1)^4}{144(n-1)} \frac{k_1(k_1-1)}{k_1+1} \frac{k_2(k_2-1)}{k_2+1} \tag{1.4}$$

The next theorem provides the asymptotics of the statistic $A_S(\mu^*, \nu^*)$ under the null hypothesis H_1.

Theorem 1.2.1 *(Alvo and Cabilio, 1995) Let $k_1 \to \infty$ (and hence $k_2 \to \infty$, $n \to \infty$) with $k_1/n \to \lambda > 0$. Then, under H_1, A_S is asymptotically normal with mean 0 and variance given by (1.4)*

Another interpretation of A_S and A_K in terms of conditional expectations is useful. Under H_1, for any pair of objects $i < j$,

$$\begin{aligned} E\left[\mu(i) - \frac{n+1}{2} \mid C(\mu^*)\right] &= \eta_\mu(i) \\ E[s(\mu(j) - \mu(i)) \mid C(\mu^*)] &= a_\mu(i,j) \end{aligned}$$

Here the conditional expectations are over the compatibility classes. Then it follows from the independence of the rankings that

$$A_S(\mu^*, \nu^*) = E[\mathcal{A}_S(\mu, \nu)|C(\mu^*), C(\nu^*)] \tag{1.5}$$

and

$$A_K(\mu^*, \nu^*) = E[\mathcal{A}_K(\mu, \nu)|C(\mu^*), C(\nu^*)] \tag{1.6}$$

Using this interpretation and results from the complete ranking situation, Alvo and Cabilio (1995) showed that

$$E\left(A_K - 4/nA_S\right)^2 = O(n^2)$$

Since $O(\sigma^2(A_S)) = O(n^5)$, we have

$$E\left(\frac{nA_K}{4\sigma(A_S)} - \frac{A_S}{\sigma(A_S)}\right)^2 = O(n^{-1}) \tag{1.7}$$

From Theorem 1.2.1 and (1.7), Corollary 1.2.1 follows.

Corollary 1.2.1 *(Alvo and Cabilio, 1995) Assume the conditions of the Theorem 1.2.1. Then, under H_1, A_K is asymptotically normal with mean 0 and variance* $16\ \sigma^2(A_S)/n^2$.

Chapter 2

Goals And Setting

2.1 Objectives

In chapter 1 we presented two nonparametric univariate statistics useful in testing for a monotone time trend. The statistics can be suitably modified in the presence of missing data. Efforts have been made to develop multivariate counterparts for complete data. Bhattacharyya and Klotz (1966) proposed a multivariate Spearman statistic whereas Dietz and Killeen (1981) proposed a multivariate Mann-Kendall statistic. These methods are suitable to analyze multivariate data recorded along approximately equal time intervals. Dietz and Killeen (1981) cited the following application. For one patient, four blood constituents were recorded at approximately monthly time interval over a period of two years. Another example can be taken from environmental studies. We measure pH simultaneously from four lakes at close proximity in order to test for trend in acidification of the region. Our main goal in this thesis is to extend the multivariate case in the presence of missing data and to investigate asymptotic properties of the corresponding test statistics .

2.2 Preliminaries

2.2.1 Multivariate Linear Rank Statistics

Here we present general limit theorems for multivariate rank statistics proved in Patel (1973).

Let $X = (X^{(1)}, ..., X^{(p)})^T$ be a p-dimensional random vector with density function $f(x)$ where $x \in R^p$, the p-dimensional real space and let $X_i = (X_i^{(1)}, ..., X_i^{(p)})^T, 1 \leq i \leq n$ be n independent observation vectors from density function $f(x)$. From now on, we assume that the density function is continuous so as to preclude ties. Ranking from smallest to largest, we denote the rank of $X_i^{(g)}$ among $X_1^{(g)}, ..., X_n^{(g)}$ by R_{ig} where $1 \leq g \leq p$. Consider the multivariate rank order statistic

$$T_{ng} = \sum_{i=1}^{n} (c_{ni} - \bar{c}_n) a_{ng}(R_{ig})$$

where $a_{ng}(\alpha)$, $1 \leq \alpha \leq n$, $g = 1, ..., p$, is a score function and the real vector $(c_{n1}, ..., c_{nn})$ satisfies

$$\sum_{i=1}^{n} (c_{ni} - \bar{c}_n)^2 > 0$$

where

$$\bar{c}_n = \sum_{i=1}^{n} c_{ni}/n$$

as well as the Noether condition

$$\lim_{n \to \infty} \left\{ \sum_{i=1}^{n} (c_{ni} - \bar{c}_n)^2 / max_{1 \leq i \leq n} (c_{ni} - \bar{c}_n)^2 \right\} = \infty \tag{2.1}$$

Consider the null hypothesis H_2 whereby the $X_i's$ are independent and identically distributed random vectors. Let $F(x^{(1)}, ..., x^{(p)})$ denote a common p-dimensional distribution function and $F^{(g,h)}$ denote the joint distribution function of random variables $X^{(g)}$ and $X^{(h)}$. The marginal distribution function of $X^{(g)}$ is denoted by $F^{(g)}$. Assume that there is some square integrable function $\varphi_g(u)$, $0 < u < 1$, $1 \leq g \leq p$ such that

$$\sigma_{\varphi_g}^2 = \int_0^1 \{\varphi_g(u) - \bar{\varphi}_g\}^2 \, du > 0$$

with $\bar{\varphi}_g = \int_0^1 \varphi_g(u)du$ and

$$\lim_{n\to\infty} \int_0^1 \{a_{ng}(1+[un]) - \varphi_g(u)\}^2 du = 0 \tag{2.2}$$

where [t] denotes the integer part of t.

For clarity of argument, let $R(.)$ be the collection of all n p-dimensional vectors of ranks such that $R(.) = \{(R_{i1}, ..., R_{ip}), i = 1, ..., n\}$. A permutation of a nonempty set B is a one-to-one mapping from B onto B. Let Π_n be the set of all permutations of the integers $\{1, ..., n\}$, and let $\pi(R(.))$, $\pi \in \Pi_n$, be an ordered set of n p-dimensional vectors such that

$$\pi R(.) = \{(R_{\pi(i)1}, ..., R_{\pi(i)p}), i = 1, ..., n\}$$

where the j-th element of the $\pi R(.)$ is $(R_{\pi(j)1}, ..., R_{\pi(j)p})$.

Then, under H_2, $P[\pi R(.)|R(.)] = 1/n!$. According to Patel (1973), under H_2, the conditional variance is

$$Var(T_{ng}|R(.)) = \sum_{i=1}^{n}(c_{ni} - \bar{c}_n)^2 \sum_{i=1}^{n}(a_{ng}(i) - \bar{a}_{ng})^2/(n-1), \tag{2.3}$$

and conditional covariance is

$$cov(T_{ng}, T_{nh}|R(.)) = \sum_{i=1}^{n}(c_{ni} - \bar{c}_n)^2 \sum_{i=1}^{n}(a_{ng}(R_{ig}) - \bar{a}_{ng})(a_{nh}(R_{ih}) - \bar{a}_{nh})/(n-1) \tag{2.4}$$

where $\bar{a}_{ng} = 1/n \sum_{i=1}^{n} a_{ng}(i)$.

Denote the conditional correlation of T_{ng} with T_{nh} by γ_{ngh} and let γ_n be the correlation matrix,

$$\gamma_n = (\gamma_{ngh})$$

Let

$$\sigma(\varphi_g, \varphi_h) = \int_0^1 \int_0^1 (\varphi_g(u) - \bar{\varphi}_g)(\varphi_h(u) - \bar{\varphi}_h) dP(U_g \le u, U_h \le v),$$

where $U_g = F^{(g)}(X^{(g)})$ and $U_h = F^{(h)}(X^{(h)})$ and we define the correlation matrix

$$\gamma = (\gamma_{gh}) \tag{2.5}$$

where $\gamma_{gh} = \sigma(\varphi_g, \varphi_h)/\sigma_{\varphi_g}\sigma_{\varphi_h}$

Lemma 2.2.1 *(Patel, 1973) Assume that (2.2) holds. Then, under H_2, γ_n converges in probability to γ as $n \to \infty$.*

Denote the joint conditional distribution of $T_{ng}/(var(T_{ng}|R(.)))^{1/2}, 1 \leq g \leq p$ by $G_n(x^{(1)}, ..., x^{(p)}|R(.))$

Theorem 2.2.1 *(Patel, 1973) Assume that γ given by (2.5) is positive definite and (2.2) holds. Then, under the null hypothesis H_2, for every $\epsilon > 0$ there exists β_ϵ such that*

$$\sum_{i=1}^{n}(c_{ni} - \bar{c}_n)^2 > \beta_\epsilon max_{1\leq i\leq n}(c_{ni} - \bar{c}_n)^2 \tag{2.6}$$

implies

$$P\left[sup_{x^{(1)},...,x^{(p)}}|G_n(x^{(1)}, ..., x^{(p)}|R(.)) - \Phi(x^{(1)}, ..., x^{(p)}|0, \gamma)| < \epsilon\right] > 1 - \epsilon$$

where $\Phi(.|0, \gamma)$ denotes the p-variate normal distribution function with mean vector 0 and dipersion matrix γ

Theorem 2.2.2 *(Patel, 1973) Under the assumptions of Theorem 2.2.1, for every $\epsilon > 0$ there exists β_ϵ such that (2.6) implies*

$$sup_{x^{(1)},...,x^{(p)}}|P\left[T_{n1} \leq x^{(1)}\sigma_{n1}, ..., T_{np} \leq x^{(p)}\sigma_{np}\right] - \Phi(x^{(1)}, ..., x^{(p)}|0, \gamma)| < \epsilon| > 1 - \epsilon$$

where

$$\sigma_{ng} = \sum_{i=1}^{n}(c_{ni} - \bar{c}_n)^2\sigma^2_{\varphi_g},\ 1 \leq g \leq p$$

Theorems 2.2.1 and 2.2.2 provide sufficient conditions in order for the multivariate linear rank statistics to have an asymptotic multivariate normal distribution.

2.2.2 Multivariate Tools

In this section, we collect some asymptotic results for multivariate random variables to be used later. First we state a multivariate central limit theorem for random variables having independent but possibly different distributions.

Theorem 2.2.3 *(Cramér, 1970, pp. 113-114) Let $V_1, V_2, ...$ be a sequence of independent random variables in R^p, p-dimensional real space, such that every V_n has the distribution function F_n with vanishing first order moments and finite second order moments m_{gh}. Suppose that, as $n \to \infty$, the following two conditions are satisfied:*

$$\frac{1}{n}\sum_{i=1}^{n} m_{i(gh)} \to m_{gh}, \ (g, h = 1, ..., p) \tag{2.7}$$

where the m_{gh} are not all equal to zero, and

$$\frac{1}{n}\sum_{i=1}^{n} \int_{|V|>\epsilon\sqrt{n}} |V|^2 dF_i \to 0 \tag{2.8}$$

for every $\epsilon > 0$, where $V = (v_1, ..., v_p)^T$ and $|V|$ denotes $\sqrt{v_1^2 + ... + v_p^2}$. Then the distribution function of $(V_1 + ... + V_n)/\sqrt{n}$ converges to a normal distribution with first order moments 0 and second order moments m_{gh}

Remark. If

$$lim_n \frac{\sum_{i=1}^{n} E|V_i|^3}{n^{3/2}} = 0, \tag{2.9}$$

it follows from

$$\frac{1}{n}\sum_{i=1}^{n} \int_{|V|>\epsilon\sqrt{n}} |V|^2 dF_i \leq \frac{1}{\epsilon n^{3/2}} \sum_{i=1}^{n} \int_{|V|>\epsilon n} |V|^3 dF_i \leq \frac{\sum_{i=1}^{n} E|V_i|^3}{\epsilon n^{3/2}}$$

that

$$lim_n \frac{1}{n}\sum_{i=1}^{n} \int_{|V|>\epsilon\sqrt{n}} |V|^2 dF_i \ = \ 0 \ \square$$

Here the notion of convergence in distribution can be expressed in several ways and equivalent definitions are of great use in obtaining asymptotic distributions.

Theorem 2.2.4 *(Hájek and Šidák, 1967, pp. 168-170) Let $V_i = (V_{i1}, ..., V_{ip})^T, 1 \leq i$ and $Z = (Z_1, ..., Z_p)^T$ be p-dimensional random vectors, each having a density function. The following definitions of convergence of $V_i, 1 \leq i$, to Z in distribution are*

equivalent.
D1: $Eh(V_i)$ converges to $Eh(Z)$ for every uniformly bounded function h which is continuous on a set C such that $P(Z \in C) = 1$.
D2: $h(V_i)$ converges to $h(Z)$ in distribution for every function h which is continuous on a set C such that $P(Z \in C) = 1$.
D3: $\sum_{j=1}^{p} \alpha_j V_{ij}$ converges in distribution to $\sum_{j=1}^{p} \alpha_j Z_j$ for every real vector $(\alpha_1, ..., \alpha_p)$.
D4: For each Borel subset A such that $P(Z \in$ the boundary of $A) = 0$

$$\lim_{i \to \infty} P(V_i \in A) = P(Z \in A) \tag{2.10}$$

holds.
D5: (2.10) holds for each p-dimensional interval A such that $P(Z \in$ the boundary of $A) =$ 0

We can also determine the limit distribution of a random variable by identifing it with that of a known random variable.

Theorem 2.2.5 *(Hoeffding, 1948) Let $V_1, V_2, ...$ be an infinite sequence of random vectors $V_n = (V_{n1}, ..., V_{np})^T$, and suppose that the distribution function $F_n(v)$ of V_n tends to a distribution function $F(v)$ as $n \to \infty$, where v is a p-dimensional vector. Let $V_{ng}^* = V_{ng} + d_{ng}$ where*

$$\lim_{n} E(d_{ng})^2 = 0, \ (g = 1, ..., p)$$

Then the distribution function of $V_n^* = (V_{n1}^*, ..., V_{np}^*)^T$ tends to $F(v)$

Chapter 3

Multivariate Trend Test: Complete Case

3.1 Multivariate Test of Trend Using Kendall Measure

Let

$$\mathfrak{X} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \dots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix}.$$

where p-dimensional vectors $X_i = (X_i^{(1)}, ..., X_i^{(p)})$ are observed over approximately equal time intervals. Dietz and Killeen (1981) suggested a multivariate nonparametric test for trend using the univariate Mann statistic. Their definition of the statistic was followed by the proof of the asymptotic distribution of the test statistic under the null hypothesis H_2. The limit distribution requires the notion of a U-statistic. Their test statistic is of the χ^2 type.

As in Fraser (1957, p. 259), we define the grade correlation coefficient κ_{gh} of $X^{(g)}$ and $X^{(h)}$ by

$$\kappa_{gh} = 3 \int \int [2F^{(g)}(x^{(g)}) - 1][2F^{(h)}(x^{(h)}) - 1] dF^{(g,h)}(x^{(g)}, x^{(h)})$$

Since $F(X^{(g)})$ represents the grade of the random variable $X^{(g)}$, κ_{gh} has been called the grade correlation coefficient of $X^{(g)}$ and $X^{(h)}$.
Replace $\mathfrak{X}$ by the matrix of ranks

$$R = \begin{bmatrix} R_{11} & \dots & R_{1p} \\ \vdots & \dots & \vdots \\ R_{n1} & \dots & R_{np} \end{bmatrix}.$$

where R_{ig} is the rank of $X_i^{(g)}$ among the observations $\{X_i^{(g)} | i = 1, ..., n\}$. The multivariate test statistic is obtained from univariate Mann statistics $K_{n(g)} = \sum_{i<j} s(X_j^{(g)} - X_i^{(g)}), g = 1, ..., p$. Under the null hypothesis H_2, $\sigma^2(K_{n(g)}) = (n-1)(2n+5)/18$, $g = 1, ..., p$ and each $K_{n(g)}/\sigma(K_{n(g)})$ is asymptotically $N(0,1)$. Define the multivariate Mann statistic $K_n = (K_{n(1)}/\sigma(K_{n(1)}), ..., K_{n(p)}/\sigma(K_{n(p)}))^T$. The next theorem provides the asymptotics for K_n.

Theorem 3.1.1 *Assume Σ is positive definite where*

$$\Sigma = \begin{bmatrix} 1 & \kappa_{12} & \dots & \kappa_{1p} \\ \kappa_{12} & 1 & \dots & \kappa_{2p} \\ \vdots & \vdots & \dots & \vdots \\ \kappa_{1p} & \kappa_{2p} & \dots & 1 \end{bmatrix}.$$

Under H_2, K_n is asymptotically p-variate normal with zero mean vector and dipersion matrix Σ and the asymptotic distribution of $K_n^T \Sigma^- K_n$ is $\chi^2(rank(\Sigma))$ where Σ^- is any generalized inverse of Σ.

We reprove this result of Dietz and Killeen (1981) for completeness.
To obtain the asymptotic distribution of the multivariate statistic K_n, we first note that $2K_{n(g)}/n(n-1)$, $g = 1, ..., p$, are U-statistics. In fact,

$$\begin{aligned} 2K_{n(g)}/n(n-1) &= \frac{2}{n(n-1)} \sum_{i<j} s(X_j^{(g)} - X_i^{(g)}) \\ &= \frac{2}{n(n-1)} \sum_{i<j} \Phi_g(X_i', X_j') \end{aligned}$$

where $\Phi_g(X_i', X_j') = s(j-i)s(X_j^{(g)} - X_i^{(g)})$ and $X_i' = (i, X_i^{(1)}, ..., X_i^{(p)})^T$.
Since

$$\Phi_g(x_i', X_j') = s(j-i)s(X_j^{(g)} - X_i^{(g)}) \ ,$$

we have that the projection

$$E\Phi_g(x_i', X_j') = \begin{cases} 1 - 2F(x_i^{(g)}) & \text{if } i < j \\ -(1 - 2F(x_i^{(g)})) & \text{if } i > j \end{cases}$$

Let

$$\bar{\Psi}_{1(i)}^{(g)}(X_i') = \begin{pmatrix} n-1 \\ 1 \end{pmatrix}^{-1} \sum_{j=1, j\neq i}^{n} E\Phi_g(x_i', X_j')$$

Then

$$\bar{\Psi}_{1(i)}^{(g)}(X_i') = \frac{n+1-2i}{n-1}(1 - 2F(X_i^{(g)}))$$

Define

$$W_{n(g)} = \frac{2}{n}\sum_{i=1}^{n} \bar{\Psi}_{1(i)}^{(g)}(X_i')$$

Hoeffding (1948) showed that the statistics $W_{n(g)}$ and $K_{n(g)}$ standardized by their respective standard deviations $\sigma(W_{n(g)})$ and $\sigma(K_{n(g)})$ are asymptotically equivalent; that is

$$\lim_n E\left(\frac{W_{n(g)}}{\sigma(W_{n(g)})} - \frac{K_{n(g)}}{\sigma(K_{n(g)})}\right)^2 = 0 \tag{3.1}$$

Notice that $\sigma^2(\bar{\Psi}_{1(i)}^{(g)}(X_i')) = (n+1-2i)^2/3(n-1)^2$.
Since

$$\begin{aligned} \sum_{i=1}^{n}(n+1-2i)^2 &= \sum_{i=1}^{n}\left[(n+1)^2 - 4i(n+1) + 4i^2\right] \\ &= \frac{n(n+1)(n-1)}{3} \ , \end{aligned}$$

it follows that

$$\begin{aligned} \sigma^2\left(\sum_{i=1}^{n} \bar{\Psi}_{1(i)}^{(g)}(X_i')\right) &= \sum_{i=1}^{n} \sigma^2(\bar{\Psi}_{1(i)}^{(g)}(X_i')) \\ &= \sum_{i=1}^{n} \frac{(n+1-2i)^2}{3(n-1)^2} = \frac{n(n+1)}{9(n-1)}. \end{aligned}$$

Hence (3.1) is equivalent to

$$\lim_n E\left(\frac{K_{n(g)}}{\sigma(K_{n(g)})} - \frac{3\sum_{i=1}^n \bar{\Psi}^{(g)}_{1(i)}(X_i')}{\sqrt{n}}\right)^2 = 0$$

Let $V_i = (3\bar{\Psi}^{(1)}_{1(i)}(X_i'), ..., 3\bar{\Psi}^{(p)}_{1(i)}(X_i'))^T$, $i = 1, ...,$ be a sequence of random vectors. Then, in the light of Theorem 2.2.5, the limit distribution of K_n is that of $\sum_{i=1}^n V_i/\sqrt{n}$. Under H_2, $\sum_{i=1}^n V_i/\sqrt{n}$ is asymptotically multivariate normal with zero mean vector and covariance matrix Σ. In fact, since under H_2, the second order moments of V_i, $1 \le g, h \le p$, are

$$\begin{aligned} m_{i(gh)} &= 9\,\frac{(n+1-2i)^2}{(n-1)^2} E\left(1-2F(X_i^{(g)})\right)\left(1-2F(X_i^{(h)})\right) \\ &= 3\,\frac{(n+1-2i)^2}{(n-1)^2}\kappa_{gh}\,, \end{aligned}$$

it is true that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n m_{i(gh)} &= 3\,\frac{\sum_{i=1}^n (n+1-2i)^2}{n(n-1)^2}\kappa_{gh} \\ &= \frac{n+1}{n-1}\kappa_{gh} \to \kappa_{gh} \quad \text{as n} \to \infty \end{aligned}$$

It suffices to show that (2.9) is satisfied.
In fact,

$$\begin{aligned} \sum_{i=1}^n E|V_i|^3 &= 27\sum_{i=1}^n \frac{|n+1-2i|^3}{|n-1|^3} E\left[\sqrt{(1-2F(X_i^{(1)}))^2 + ... + (1-2F(X_i^{(p)}))^2}\right]^3 \\ &\le \frac{27}{|n-1|^3}\sum_{i=1}^n |n+1-2i|^3 p^3 \\ &\le \frac{27}{|n-1|^3} n|n-1|^3 p^3. \end{aligned}$$

Since $\lim_n np^3/n^{3/2} \to 0$, it follows that

$$\lim_n \sum_{i=1}^n \frac{E|V_i|^3}{n^{3/2}} = 0$$

and the result follows from Theorem 2.2.3.

Consequently the limit distribution of K_n is multivariate normal with zero mean vector and covariance matrix Σ and $K_n^T \Sigma^- K_n$ is asymptotically $\chi^2(rank\Sigma)$ □

When Σ is unknown, it can be replaced by a consistent estimator $\hat{\Sigma}_n$ so that the test statistic becomes $K_n^T \hat{\Sigma}_n^- K_n$. Dietz and Killeen (1981) defined

$$\hat{\Sigma}_n = Cov(K_n|R(.)),$$

and showed that this estimator is consistent.

Theorem 3.1.2 *(Dietz and Killeen, 1981) Under H_2, the conditional covariance of $K_{n(g)}$ and $K_{n(h)}$ is*

$$Cov\left(K_{n(g)}, K_{n(h)}|R(.)\right) = K_{n(g,h)}/3 + (n^3 - n)r_{n(g,h)}/9$$

where

$$K_{n(g,h)} = \sum_{i<j} s(X_j^{(g)} - X_i^{(g)})s(X_j^{(h)} - X_i^{(h)})$$

and

$$\begin{aligned} r_{n(g,h)} &= \frac{12}{n^3 - n} \sum_{i=1}^{n} \left(R_{ig} - \frac{n+1}{2}\right)\left(R_{ih} - \frac{n+1}{2}\right) \\ &= \frac{3}{n^3 - n} \sum_{i,j,k} s(X_j^{(g)} - X_i^{(g)})s(X_j^{(h)} - X_k^{(h)}) \end{aligned} \tag{3.2}$$

Theorem 3.1.3 *(Dietz and Killeen, 1981) Under H_2, $Cov(K_n|R(.)) \to \Sigma$ in probability*

In view of Theorems 3.1.1 and 3.1.3, the limit distribution of $K_n^T \hat{\Sigma}_n^- K_n$ is $\chi^2(rank(\Sigma))$. They proposed it as a test statistic for testing trend which rejects the null hypothesis for large values of the statistic.

3.2 Multivariate test of Trend Using Spearman Measure

Bhattacharyya and Klotz (1966) suggested a distribution-free test of trend for bivariate data based on the Spearman measure. From the construction of the statistic and the study of its asymptotic properties, it is evident that the extension to general multivariate data is straightforward. Bhattacharyya and Klotz (1966) applied the method to annual bivariate data collected for a lake composed of freezing days and thawing days and tested for a warming trend.

Let $Y_i = (X_i^{(1)}, X_i^{(2)}), 1 \leq i \leq n$, be n independent observations from a continuous density function $f(x)$ where $x \in R^2$, the 2-dimensional real space. By R_{ig}, we denote the rank of $X_i^{(g)}$ among $X_1^{(g)}, ..., X_n^{(g)}$ where $g = 1, 2$. Define $S_n = (S_{n(1)}/\sigma(S_{n(1)}), S_{n(2)}/\sigma(S_{n(2)}))^T$ where $S_{n(g)}$, $g = 1, 2$, are defined by

$$S_{n(g)} = \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)\left(R_{ig} - \frac{n+1}{2}\right)$$

Under H_2,

$$\begin{aligned} \sigma^2(S_{n(g)}) &= Var(S_{n(g)}) = \frac{n^2(n+1)^2(n-1)}{144}, \ g = 1, 2 \\ Cov(S_{n(1)}, S_{n(2)}|R(.)) &= \frac{n(n+1)}{12} S_{n(1,2)} \end{aligned}$$

where $S_{n(1,2)}$ is defined by

$$S_{n(1,2)} = \sum_{i=1}^{n} \left(R_{i1} - \frac{n+1}{2}\right)\left(R_{i2} - \frac{n+1}{2}\right)$$

Then, under H_2,

$$Cov(S_{n(1)}/\sigma(S_{n(1)}), S_{n(2)}/\sigma(S_{n(2)})|R(.)) = r_{n(1,2)}$$

where $r_{n(1,2)}$ is defined in (3.2).
Defining

$$\hat{\Gamma}_n = \begin{cases} \Gamma_n & \text{if } |r_{n(1,2)}| < 1 \\ I & \text{if } |r_{n(1,2)}| = 1 \end{cases}$$

where

$$\Gamma_n = \begin{bmatrix} 1 & r_{n(1,2)} \\ r_{n(1,2)} & 1 \end{bmatrix}$$

so that $\hat{\Gamma}_n$ is always non-singular, they considered the test statistic $S_n^T \hat{\Gamma}_n^{-1} S_n$. The following two theorems state the consistency of the statistic $\hat{\Gamma}_n$ and the limit distribution of the test statistic.

Theorem 3.2.1 *(Bhattacharyya and Klotz,1966) Under H_2, $\hat{\Gamma}_n$ converges in probability to Γ, where*

$$\Gamma = \begin{bmatrix} 1 & \kappa_{12} \\ \kappa_{12} & 1 \end{bmatrix}$$

Theorem 3.2.2 *(Bhattachayya and Klotz, 1966) Under H_2, if $|\kappa_{12}| < 1$, the test statistic $S_n^T \hat{\Gamma}_n^{-1} S_n$ is asymptotically $\chi^2(2)$.*

Chapter 4

Multivariate Trend Test: Incomplete Case

In the same spirit as Dietz and Killeen (1981) or Bhattacharyya and Klotz (1966), we will develop a method to analyze multivariate data in the presence of missing observations. Since in the complete data case, the multivariate test statistic is a function of univariate statistics, similarly in the multivariate case we define the multivariate statistic as a function of univariate Spearman statistics or of univariate Kendall statistics for missing data. We discuss the Spearman case first for the sake of simplicity. The proof of the asymptotic distribution of the Spearman measure will lead easily to that for the Kendall case. From the standpoint of multivariate rank statistics, we achieve an extension from complete to incomplete data.

4.1 Spearman Measure

Let

$$\mathfrak{X}^* = \begin{bmatrix} X_1^{(1)}\delta_{11}^* & \dots & X_1^{(p)}\delta_{1p}^* \\ \vdots & \dots & \vdots \\ X_n^{(1)}\delta_{n1}^* & \dots & X_n^{(p)}\delta_{np}^* \end{bmatrix}$$

where

$$X_i^{(g)}\delta_{ig}^* = \begin{cases} - & \text{if } X_i^{(g)} \text{ is missing} \\ X_i^{(g)} & \text{otherwise} \end{cases}$$

Allowing that some observations are missing, we let $k_g, g = 1, ..., p$, be the number of the non-missing values among $X_i^{(g)}, i = 1, ..., n$. Replace $\mathfrak{X}^*$ by the matrix of incomplete ranks,

$$R^* = \begin{bmatrix} R_{11}^* & \dots & R_{1p}^* \\ \vdots & \dots & \vdots \\ R_{n1}^* & \dots & R_{np}^* \end{bmatrix}.$$

where R_{ig}^* is the ranking of $X_i^{(g)}$ among the non-missing column values if $X_i^{(g)}$ is not missing, and is equal to $(k_g + 1)/2$ if $X_i^{(g)}$ is missing.

Consider the null hypothesis H_3 whereby we assume that

1. k_g, $g = 1, ..., p$, the number of ranked observations for each component, are fixed.
2. for each component, the rankings for which we have incomplete data are uniformly distributed over the n! permutations of $(1, ..., n)$.
3. for each component, the pattern of the missing observations are randomly selected from the set of all possible patterns.

Recalling (1.2), define the statistic $S_n^* = (S_{n(1)}^*/\sigma(S_{n(1)}^*), ..., S_{n(p)}/\sigma(S_{n(p)}^*))^T$ where

$$\begin{aligned} S_{n(g)}^* &= \frac{n+1}{k_g+1}\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)\left(R_{ig}^* - \frac{k_g+1}{2}\right) \\ &= \frac{n+1}{k_g+1}\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)R_{ig}^* \end{aligned} \tag{4.1}$$

with $(R_{1g}^*, ..., R_{ng}^*)$ being a permutation of $(1, 2, ..., k_g - 1, k_g, (k_g+1)/2, ..., (k_g+1)/2)$ Under H_3, $(R_{1g}^*, ..., R_{ng}^*)$ is the random vector uniformly distributed over the permutations of $(1, 2, ..., k_g - 1, k_g, (k_g + 1)/2, ..., (k_g + 1)/2)$ and it follows from (2.3) that

$$Var(S_{n(g)}^*) = \frac{k_g(k_g-1)}{k_g+1}\frac{n(n+1)^3}{144}$$

Let a_{ng} be a bijective monotone increasing function from $\{1, ..., n\}$ to $\{1/(k_g +$

$1), ..., 1/2, ..., 1/2, ..., k_g/(k_g+1)\}$ defined by

$$a_{ng}(i) = \begin{cases} \dfrac{i}{k_g+1} & i \leq \left[\dfrac{k_g+1}{2}\right] \\ \dfrac{1}{2} & \left[\dfrac{k_g+1}{2}\right] < i < \left[\dfrac{k_g+1}{2}\right] + n - k_g + 1 \\ \dfrac{i-n+k_g}{k_g+1} & \left[\dfrac{k_g+1}{2}\right] + n - k_g + 1 \leq i \leq n \end{cases}$$

Then, under H_3, we can identify (4.1) with (4.2)

$$S^*_{n(g)} = \sum_{i=1}^{n} (n+1)\left(i - \frac{n+1}{2}\right) a_{ng}(E_{ig}) \tag{4.2}$$

where $(E_{1g}, ..., E_{ng})$ is the random vector uniformly distributed over the permutations of $(1, ..., n)$. With this identification, it is only necessary to verify the conditions of Theorems 2.2.1 and 2.2.2 in order to obtain the asymptotic distribution of the statistic S^*_n.

For the statistic $S^*_{n(g)}$ of (4.2), it is easy to show that the constants $(1, ..., n)$ satisfy the Noether condition (2.1). For all n,

$$\sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 = \frac{n(n+1)(n-1)}{12}$$

$$\begin{aligned} max_{1\leq i\leq n}\left(i - \frac{n+1}{2}\right)^2 &= \left(1 - \frac{n+1}{2}\right)^2 \\ &= \frac{(n-1)^2}{4} \end{aligned}$$

Consequently,

$$\lim_n \left\{ \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 / max_{1\leq i\leq n}\left(i - \frac{n+1}{2}\right)^2 \right\} = \infty \tag{4.3}$$

Define

$$b_{nk_g}(i) = \frac{k_g+1}{n+1} a_{ng}(i)$$

and

$$\varphi_g(u) = \begin{cases} u & u < \lambda_g/2 \\ \lambda_g/2 & \lambda_g/2 \le u < 1 - \lambda_g/2 \\ u - (1-\lambda_g) & 1 - \lambda_g/2 \le u \end{cases}$$

Note that b_{nk_g} is a bijective monotone increasing function from $\{1, ..., n\}$ to $\{1/(n+1), ..., (k_g+1)/2(n+1), ..., (k_g+1)/2(n+1), ..., k_g/(n+1)\}$. Under the assumption that $k_g/n \to \lambda_g > 0$ as $n \to \infty$, we will show that

$$\lim_{n\to\infty} \int_0^1 \left[b_{nk_g}(1 + [un]) - \varphi_g(u)\right]^2 du = 0$$

Recall Lemma 4.1.1 from Hájek and Šidák (1967).

Lemma 4.1.1 *(Hájek and Šidák, 1967, p. 164) Define* $b_{ng}(i) = \varphi_g\left(\frac{i}{n+1}\right)$. *Then*

$$\lim_{n\to\infty} \int_0^1 \left[b_{ng}(1 + [un]) - \varphi_g(u)\right]^2 du = 0$$

We now show that b_{nk_g} and b_{ng} are asymptotically equivalent in the L_2 norm.

Lemma 4.1.2 *Let* $k_g/n \to \lambda_g > 0$ *as* $n \to \infty$, $g = 1, ..., p$. *Then*

$$\lim_{n\to\infty} \int_0^1 \left[b_{nk_g}(1 + [un]) - b_{ng}(1 + [un])\right]^2 du = 0$$

Proof.

Define

$$\varphi_{nk_g}(u) = \begin{cases} u & u \le \frac{1}{n+1}\left[\frac{k_g+1}{2}\right] \\ \frac{k_g+1}{2(n+1)} & \frac{1}{n+1}\left[\frac{k_g+1}{2}\right] < u < \frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right] + n - k_g + 1\right) \\ u - \frac{n-k_g}{n+1} & \frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right] + n - k_g + 1\right) \le u \le 1 \end{cases}$$

Note that from the definition $b_{nk_g}(i) = \varphi_{nk_g}\left(\frac{i}{n+1}\right)$, $i = 1, ..., n$.

Since $k_g/n \to \lambda_g$, there exists n_o such that if $n \geq n_o$

$$\left|\frac{1}{n+1}\left(\frac{k_g+1}{2}+1\right)-\lambda_g/2\right| = \left|\frac{k_g+3}{2(n+1)}-\lambda_g/2\right| < \epsilon \tag{4.4}$$

and

$$\left|\frac{1}{n+1}\left(\frac{k_g+1}{2}-1\right)-\lambda_g/2\right| = \left|\frac{k_g-1}{2(n+1)}-\lambda_g/2\right| < \epsilon \tag{4.5}$$

Let $n \geq n_o$. It follows from (4.4), (4.5) and

$$\frac{k_g-1}{n+1} < \frac{k_g}{n+1} < \frac{k_g+1}{n+1} < \frac{k_g+3}{n+1}$$

that

$$\left|\frac{k_g}{n+1}-\lambda_g\right| < 2\epsilon \text{ and } \left|\frac{k_g+1}{n+1}-\lambda_g\right| < 2\epsilon$$

It also follows from (4,4), (4.5) and

$$\frac{k_g-1}{2} < \left[\frac{k_g+1}{2}\right] < \left[\frac{k_g+1}{2}\right]+1 < \frac{k_g+3}{2}$$

that

$$\left|\frac{1}{n+1}\left[\frac{k_g+1}{2}\right]-\lambda_g/2\right| < \epsilon \text{ and } \left|\frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right]+1\right)-\lambda_g/2\right| < \epsilon$$

In the sequel, we will show that

$$|\varphi_{nk_g}(u)-\varphi_g(u)| \leq 2\epsilon \text{ for all } 0 \leq u \leq 1$$

To compare φ_g with φ_{nk_g}, we consider various possible cases.

Case 1. $u \leq min\left(\frac{1}{n+1}\left[\frac{k_g+1}{2}\right], \lambda_g/2\right)$.

Both φ_{nk_g} and φ_g are defined as u making the difference zero.

Case 2. $\frac{1}{n+1}\left[\frac{k_g+1}{2}\right] \leq u \leq \lambda_g/2$.

Since $\frac{1}{n+1}\left[\frac{k_g+1}{2}\right] \leq u$ and $\frac{1}{n+1}\left[\frac{k_g+1}{2}\right] \leq \frac{k_g+1}{2(n+1)}$,

$$\begin{aligned} |\varphi_{nk_g}(u) - \varphi_g(u)| &= \left|\frac{k_g+1}{2(n+1)} - u\right| \\ &\leq \left|\frac{1}{n+1}\left[\frac{k_g+1}{2}\right] - u\right| \\ &\leq \left|\frac{1}{n+1}\left[\frac{k_g+1}{2}\right] - \lambda_g/2\right| \\ &< \epsilon \end{aligned}$$

Case 3. $\lambda_g/2 \leq u \leq \frac{1}{n+1}\left[\frac{k_g+1}{2}\right]$.

$$\begin{aligned} |\varphi_{nk_g}(u) - \varphi_g(u)| &= |u - \lambda_g/2| \\ &\leq \left|\frac{1}{n+1}\left[\frac{k_g+1}{2}\right] - \lambda_g/2\right| \\ &< \epsilon \end{aligned}$$

Case 4.

$$max\left\{\frac{1}{n+1}\left[\frac{k_g+1}{2}\right], \lambda_g/2\right\} < u < min\left\{1 - \lambda_g/2, \frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right] + n - k_g + 1\right)\right\}.$$

It follows from $\varphi_{nk_g} = (k_g+1)/2(n+1)$ and $\varphi_g = \lambda_g/2$ that

$$|\varphi_{nk_g}(u) - \varphi_g(u)| = \left|\frac{k_g+1}{2(n+1)} - \lambda_g/2\right| < \epsilon$$

Case 5. $\frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right] + n - k_g + 1\right) \leq u < 1 - \lambda_g/2.$

Note that $\varphi_{nk_g}(u) = u - \frac{n-k_g}{n+1}$ and $\varphi_g(u) = \lambda_g/2$. Then

$$|\varphi_{nk_g}(u) - \varphi_g(u)| = \left|u - \frac{n-k_g}{n+1} - \lambda_g/2\right|$$

$$\begin{aligned} &\leq max\left\{\left|\frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right]+1\right)-\lambda_g/2\right| \ , \ \left|\frac{k_g+1}{n+1}-\lambda_g\right|\right\} \\ &\leq 2\epsilon \end{aligned}$$

Case 6. $1-\lambda_g/2 \leq u < \frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right]+n-k_g+1\right)$

Since $\varphi_{nk_g}(u) = \frac{k_g+1}{2(n+1)}$ and $\varphi_g(u) = u-(1-\lambda_g)$,

$$\begin{aligned} |\varphi_{nk_g}(u)-\varphi_g(u)| &= \left|u+\lambda_g-1-\frac{k_g+1}{2(n+1)}\right| \\ &\leq max\left\{\left|\frac{\lambda_g}{2}-\frac{k_g+1}{2(n+1)}\right|, \left|\lambda_g-\frac{1}{n+1}\left(k_g+\frac{k_g+1}{2}-\left[\frac{k_g+1}{2}\right]\right)\right|\right\} \end{aligned} \tag{4.6}$$

and it follows from

$$\frac{k_g}{n+1} \leq \frac{1}{n+1}\left(k_g+\frac{k_g+1}{2}-\left[\frac{k_g+1}{2}\right]\right) \leq \frac{k_g+1}{n+1}$$

that

$$\left|\lambda_g-\frac{1}{n+1}\left(k_g+\frac{k_g+1}{2}-\left[\frac{k_g+1}{2}\right]\right)\right| < 2\epsilon$$

Then (4.6) is less than 2ϵ.

Case 7. $max\left\{1-\lambda_g/2 \ , \ \frac{1}{n+1}\left(\left[\frac{k_g+1}{2}\right]+n-k_g+1\right)\right\} \leq u.$

Since $\varphi_{nk_g}(u) = u-\frac{n-k_g}{n+1}$ and $\varphi_g(u) = u-(1-\lambda_g)$,

$$\begin{aligned} |\varphi_{nk_g}(u)-\varphi_g(u)| &= \left|1-\lambda_g-\frac{n-k_g}{n+1}\right| \\ &= \left|\frac{k_g+1}{n+1}-\lambda_g\right| \\ &< 2\epsilon \end{aligned}$$

Since $b_{ng}(i) = \varphi_g\left(\frac{i+1}{n}\right)$ and $b_{nk_g}(i) = \varphi_{nk_g}\left(\frac{i+1}{n}\right)$, it follows that if $n \geq n_o$, for all $i = 1, ..., n$,

$$|b_{nk_g}(i) - b_{ng}(i)| \leq 2\epsilon$$

and consequently,

$$\int_0^1 \left[b_{nk_g}(1 + [un]) - b_{ng}(1 + [un])\right]^2 du \leq 4\epsilon^2 \ \square$$

Theorem 4.1.1 *Let $k_g/n \to \lambda_g > 0$ as $n \to \infty$, $g = 1, ..., p$. Then*

$$\lim_{n\to\infty} \int_0^1 \left[b_{nk_g}(1 + [un]) - \varphi_g(u)\right]^2 du \ = \ 0$$

Proof.

The proof requires some notions of functional analysis.

Let p be a real number which is greater than or equal to 1. Then a measurable function f defined on [0,1] is said to belong to the space $L^p[0,1]$ if $\int_0^1 |f|^p < \infty$. For $f \in L^p$, we define a norm (Royden, 1968, p. 111),

$$||f||_p \ = \ \{\int_0^1 |f|^p\}^{1/p}$$

One of the properties of the norm is

$$||f + g||_p \leq ||f||_p + ||g||_p \tag{4.7}$$

Using this notation, we can rewrite Lemmas 4.1.1 and 4.1.2 as

$$\lim_n ||b_{ng} - \varphi_g||_2^2 \ = \ 0 \tag{4.8}$$

and

$$\lim_n ||b_{nk_g} - b_{ng}||_2^2 \ = \ 0 \tag{4.9}$$

It follows from (4.7) that

$$||b_{ng} - \varphi_g||_2 \ \leq \ ||b_{nk_g} - b_{ng}||_2 + ||b_{ng} - \varphi_g||_2$$

From (4.8) and (4.9), it follows that $\lim_n ||b_{nk_g} - \varphi_g||_2 = 0$ and $\lim_n ||b_{nk_g} - \varphi_g||_2^2 = 0$ $\square$

Corollary 4.1.1 *Under the conditions of Theorem 4.1.1,*

$$\lim_n \int_0^1 \left[a_{ng}(1+[un]) - \lambda_g^{-1}\varphi_g(u)\right]^2 du = 0$$

Proof.

Since

$$\begin{aligned} a_{ng}(1+[un]) - \lambda_g^{-1}\varphi_g(u) &= \frac{n+1}{k_g+1} b_{nk_g}(1+[un]) - \lambda_g^{-1}\varphi_g(u) \\ &= \left\{\frac{n+1}{k_g+1} - \frac{1}{\lambda_g}\right\} b_{nk_g}(1+[un]) \\ &\quad +1/\lambda_g \left\{b_{nk_g}(1+[un]) - \varphi_g(u)\right\}, \end{aligned}$$

$$||a_{ng} - \lambda_g^{-1}\varphi_g||_2 \leq \left|\frac{n+1}{k_g+1} - 1/\lambda_g\right| ||b_{nk_g}||_2 + \lambda_g^{-1}||b_{nk_g} - \varphi_g||_2$$

Note that $||b_{nk_g}||_2$ is bounded by 1 and $k_g/n \to \lambda_g$. Hence

$$\left|\frac{n+1}{k_g+1} - 1/\lambda_g\right| ||b_{nk_g}||_2 \to 0 \text{ as } n \to \infty$$

and Theorem 4.1.1 completes the proof. □

Lemma 2.2.1 holds generally. In our case we can specify the matrix γ. From now on, we will be concerned with the computation of the correlation matrix γ.
Under H_3, from (2.4),

$$Cov(S^*_{n(g)}, S^*_{n(h)}|R^*(.)) = \frac{n(n+1)}{12} S^*_{n(g,h)}$$

where $S^*_{n(g,h)}$ is defined by

$$S^*_{n(g,h)} = \frac{(n+1)^2}{(k_g+1)(k_h+1)} \sum_{i=1}^{n} \left(R^*_{ig} - \frac{k_g+1}{2}\right)\left(R^*_{ih} - \frac{k_h+1}{2}\right)\delta_{ig}\delta_{ih}$$

and δ_{ig} is defined by

$$\delta_{ig} = \begin{cases} 0 & \text{if } X_i^{(g)} \text{ is missing} \\ 1 & \text{otherwise} \end{cases}$$

Consider

$$\begin{aligned} Cov(S^*_{n(g)}/\sigma(S^*_{n(g)}), S^*_{n(h)}/\sigma(S^*_{n(h)})|R^*(.)) &= \frac{n(n+1)}{12\ \sigma(S^*_{n(g)})\sigma(S_{n(h)})} S^*_{n(g,h)} \\ &= \frac{n(n+1)(n^3-n)}{144\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})}\ r^*_{n(g,h)} \end{aligned} \tag{4.10}$$

where

$$r^*_{n(g,h)} = \frac{12}{n^3-n} S^*_{n(g,h)} \tag{4.11}$$

We can rewrite

$$\begin{aligned} r^*_{n(g,h)} &= \frac{12}{n^3-n}\ \frac{(n+1)^2}{(k_g+1)(k_h+1)} \sum_{i=1}^{n} \left(R^*_{ig} - \frac{k_g+1}{2}\right)\left(R^*_{ih} - \frac{k_h+1}{2}\right)\delta_{ig}\delta_{ih} \\ &= \frac{(n+1)^2}{(k_g+1)(k_h+1)}\ \frac{3}{n^3-n} \sum_i \sum_j \sum_k s(X^{(g)}_i - X^{(g)}_j) s(X^{(h)}_i - X^{(h)}_k)\delta_{ig}\delta_{jg}\delta_{ih}\delta_{kh} \\ &= \frac{n+1}{k_g+1}\ \frac{n+1}{k_h+1}\ \frac{(n-2)U^*_{n(g,h)} + 3T^*_{n(g,h)}}{n+1} \end{aligned}$$

where

$$\begin{aligned} U^*_{n(g,h)} &= \frac{1}{n(n-1)(n-2)} \Sigma'\ 3 s(X^{(g)}_i - X^{(g)}_j) s(X^{(h)}_i - X^{(h)}_k)\delta_{ig}\delta_{jg}\delta_{ih}\delta_{kh} \\ T^*_{n(g,h)} &= \frac{2}{n(n-1)} \sum_{i<j} s(X^{(g)}_i - X^{(g)}_j) s(X^{(h)}_i - X^{(h)}_k)\delta_{ig}\delta_{jg}\delta_{ih}\delta_{jh} \end{aligned}$$

and Σ' is the summation over all three different integers i, j, k chosen from $\{1, ..., n\}$.

Lemma 4.1.3 *Assume that $k_g/n \to \lambda_g > 0$ as $n \to \infty$, g=1,..,p. Then, under H_3,*

$$\frac{n(n+1)^3(n^3-n)}{144\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})}\ r^*_{n(g,h)} \to \kappa_{gh}\,(\lambda_g\lambda_h)^{1/2} \text{ in probability}$$

Proof.

Under H_3,

$$E(U^*_{n(g,h)}) = \frac{1}{n(n-1)(n-2)} \Sigma'\ 3E\left(s(X^{(g)}_i - X^{(g)}_j) s(X^{(h)}_i - X^{(h)}_k)\right) E\left(\delta_{ig}\delta_{jg}\delta_{ih}\delta_{kh}\right)$$

Since

$$E\left(3s(X_i^{(g)} - X_j^{(g)})s(X_i^{(h)} - X_k^{(h)})\right) = 3\int \ldots \int s(x_i^{(g)} - x_j^{(g)})s(x_i^{(h)} - x_k^{(h)}) dF^{(g,h)}(x_i^{(g)}, x_i^{(h)})dF^{(g)}(x_j^{(g)})dF^{(h)}(x_k^{(h)}) = \kappa_{gh}$$

and

$$E\left(\delta_{ig}\delta_{jg}\delta_{ih}\delta_{kh}\right) = E\left(\delta_{ig}\delta_{jg}\right)E\left(\delta_{ih}\delta_{kh}\right) = \frac{k_g(k_g-1)}{n(n-1)}\frac{k_h(k_h-1)}{n(n-1)},$$

it follows that

$$E(U^*_{n(g,h)}) = \kappa_{gh}\,\frac{k_g(k_g-1)}{n(n-1)}\frac{k_h(k_h-1)}{n(n-1)}.$$

By the assumption that $k_g/n \to \lambda_g$ as $n \to \infty$, g=1,..,p, $E(U^*_{n(g,h)})$ converges to $\kappa_{gh}(\lambda_g\lambda_h)^2$, as $n \to \infty$. Since $(n+1)/(k_g+1)$ and $(n+1)/(k_h+1)$ converge respectively to λ_g^{-1} and λ_h^{-1} as n $\to \infty$ and $|E(T^*_{n(g,h)})| \leq 1$, it follows that $\lim_n E(r^*_{n(g,h)}) \to \kappa_{gh}\lambda_g\lambda_h$ as n $\to \infty$.

To show that $r^*_{n(g,h)} \to \kappa_{gh}\lambda_g\lambda_h$ in probability, it suffices to show $\lim_n Var(r^*_{n(g,h)}) = 0$. First we identify $U^*_{n(g,h)}$ as a U-statistic with symmetric kernel $\Phi_{(g,h)}$and vector arguments x_1, x_2, x_3

$$\Phi_{(g,h)}(x_1,x_2,x_3) = 1/6 \sum_{i\neq j\pm k, i\neq k}^{1,2,3} 3s(x_i^{(g)} - x_j^{(g)})s(x_i^{(h)} - x_k^{(h)})\delta_{ig}\delta_{jg}\delta_{ih}\delta_{kh}$$

It follows from Hoeffding (1948) that

$$Var(U^*_{n(g,h)}) = \frac{6}{n(n-1)(n-2)}\left(3\begin{pmatrix} n-3 \\ 2 \end{pmatrix}\zeta_1(\kappa_{gh}) + 3(n-3)\zeta_2(\kappa_{gh}) + \zeta_3(\kappa_{gh})\right)$$

$$\begin{aligned}
\zeta_1(\kappa_{gh}) &= E{\Phi'_{(g,h)}}^2(X_1) - \kappa_{gh}^2 \\
\zeta_2(\kappa_{gh}) &= E{\Phi''_{(g,h)}}^2(X_1,X_2) - \kappa_{gh}^2 \\
\zeta_3(\kappa_{gh}) &= E\Phi_{(g,h)}(X_1,X_2,X_3) - \kappa_{gh}^2 \\
\Phi'_{(g,h)}(x_1) &= E\Phi_{(g,h)}(x_1,X_2,X_3) \\
\Phi''_{(g,h)}(x_1,x_2) &= E\Phi_{(g,h)}(x_1,x_2,X_3)
\end{aligned}$$

Since $\Phi_{(g,h)}(x_1, x_2, x_3)$ is bounded in absolute value by 3, $\zeta_1(\kappa_{gh}), \zeta_2(\kappa_{gh}), \zeta_3(\kappa_{gh})$ are all bounded by some $M > 0$ and it folllows that

$$Var(U^*_{n(g,h)}) \leq \frac{6}{n(n-1)(n-2)} \left(3 \begin{pmatrix} n-3 \\ 2 \end{pmatrix} + 3(n-3) + 1 \right) M = O(n^{-1}) \tag{4.12}$$

It also follows from

$$\left| s(x_i^{(g)} - x_j^{(g)}) s(x_i^{(h)} - x_k^{(h)}) \delta_{ig}\delta_{jg}\delta_{ih}\delta_{jh} \right| \leq 1$$

that

$$\begin{aligned} Var\left(\frac{T^*_{n(g,h)}}{n+1}\right) &= \frac{1}{(n+1)^2} E\left[\frac{2}{n(n-1)} \sum_{i<j} s(X_i^{(g)} - X_j^{(g)}) s(X_i^{(h)} - X_k^{(h)}) \delta_{ig}\delta_{jg}\delta_{ih}\delta_{jh} \right]^2 \\ &\leq 1/(n+1)^2 \end{aligned} \tag{4.13}$$

From (4.12) and (4.13),

$$\begin{aligned} \left| Cov\left(\frac{n-2}{n+1} U^*_{n(g,h)} \, , \, \frac{3}{n+1} T^*_{n(g,h)} \right) \right|^2 &\leq Var\left(\frac{n-2}{n+1} U^*_{n(g,h)} \right) Var\left(\frac{3}{n+1} T^*_{n(g,h)} \right) \\ &\to 0 \text{ as } n \to \infty \end{aligned} \tag{4.14}$$

Combining (4.12), (4.13) and (4.14), we prove that $\lim_n Var(r_{n(g,h)}) = 0$.
As a consequence, under H_3,

$$\frac{n(n+1)^3(n^3-n)}{144\, \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})} r^*_{n(g,h)} \to \kappa_{gh} \, (\lambda_g \lambda_h)^{1/2} \quad \text{in probability}$$

since

$$\begin{aligned} \frac{n(n+1)(n^3-n)}{144\, \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})} &= \frac{n(n+1)(n^3-n)}{144} \frac{\sqrt{k_g+1}\sqrt{k_h+1}}{\sqrt{k_g(k_g-1)}\sqrt{k_h(k_h-1)}} \frac{144}{n(n+1)^3} \\ &\to (\lambda_g \lambda_h)^{-1/2} \text{ as n} \to \infty. \, \square \end{aligned}$$

The next Theorem provides the asymptotic distribution of S^*_n under H_3.

Theorem 4.1.2 *Assume $k_g/n \to \lambda_g > 0$, $g = 1, ..., p$, and the matrix Σ is positive definite, where*

$$\Sigma = \begin{bmatrix} 1 & \kappa_{12}(\lambda_1\lambda_2)^{1/2} & \dots & \kappa_{1p}(\lambda_1\lambda_p)^{1/2} \\ \vdots & \vdots & \dots & \vdots \\ \kappa_{1p}(\lambda_1\lambda_p)^{1/2} & \kappa_{2p}(\lambda_2\lambda_p)^{1/2} & \dots & 1 \end{bmatrix}.$$

Under H_3, S_n^* is asymptotically normal with zero mean vector and covariance matrix Σ

Proof.

It suffices to ensure that the statistic S_n^* satisfies the conditions of Theorem 2.2.2. In (4.3), we have shown the Noether condition is satisfied. Corollary 4.1.1 shows the existence of φ required in (2.2). Lemma 4.1.3 completes the proof. □

Let

$$\begin{aligned} \hat{\sigma}_{nS(g,h)} &= Cov(S^*_{n(g)}/\sigma(S^*_{n(g)}), S^*_{n(h)}/\sigma(S^*_{n(h)})|R^*(.)) \\ &= \frac{n(n+1)}{12\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})} S^*_{n(g,h)} \end{aligned}$$

As a Corollary to Theorem 4.1.2, we can provide the asymptotic distribution of the test statistic $S_n^{*T}\hat{\Sigma}_{nS}^{-}S_n^*$ where

$$\hat{\Sigma}_{nS} = \begin{bmatrix} 1 & \hat{\sigma}_{nS(1,2)} & \dots & \hat{\sigma}_{nS(1,p)} \\ \vdots & \vdots & \dots & \vdots \\ \hat{\sigma}_{nS(1,p)} & \hat{\sigma}_{nS(2,p)} & \dots & 1 \end{bmatrix}.$$

Corollary 4.1.2 *Under the same conditions as in Theorem 4.1.2, the limit distributions of $S_n^{*T}\Sigma^{-}S_n^*$ and $S_n^{*T}\hat{\Sigma}_{nS}^{-}S_n^*$ are $\chi^2\,(rank(\Sigma))$*

4.2 Kendall Measure

From (1.3), we can also define a similarity measure based on Kendall rank correlation in the presence of missing observations.

Let

$$R'^{*} = \begin{bmatrix} R'^{*}_{11} & \dots & R'^{*}_{1p} \\ \vdots & \dots & \vdots \\ R'^{*}_{n1} & \dots & R'^{*}_{np} \end{bmatrix}.$$

where R'^{*}_{ig} is the ranking of $X_i^{(g)}$ among the non-missing column values if $X_i^{(g)}$ is not missing, and is denoted by the symbol "-" if $X_i^{(g)}$ is missing. Let $R'^{*}(.) = \{(R'^{*}_{i1}, ..., R'^{*}_{ip}) | i = 1, ..., n\}$ be given. Recall that Π_n is the collection of all permutations of $\{1, ..., n\}$. For $\pi \in \Pi_n$, recalling (1.3), we define

$$K^{*}_{n\pi(g)} = \sum_{i<j} a_{n\pi(g)}(i,j) \tag{4.15}$$

where

$$a_{n\pi(g)}(i,j) = \begin{cases} s(R'^{*}_{\pi(j)g} - R'^{*}_{\pi(i)g}) & \text{if both } X_{ig} \text{ and } X_{jg} \text{ are ranked} \\ 1 - 2R'^{*}_{\pi(i)g}/(k_g + 1) & \text{if only } X_{ig} \text{ is ranked} \\ 2R'^{*}_{\pi(j)g}/(k_g + 1) - 1 & \text{if only } X_{jg} \text{ is ranked} \\ 0 & \text{otherwise} \end{cases}$$

Define the multivariate statistic

$$K^{*}_{n} = \left(\frac{K^{*}_{n(1)}}{4\sigma(S^{*}_{n(1)})/n}, ..., \frac{K^{*}_{n(p)}}{4\sigma(S^{*}_{n(p)})/n} \right)$$

where $K_{n(g)}$ is defined by (4.15) with π the identity permutation. Theorem 4.2.1 below provides the asymptotic distribution.

Theorem 4.2.1 *Assume $k_g/n \to \lambda_g > 0$ as $n \to \infty$, $g = 1, ..., p$ and Σ is positive definite, where Σ is defined in Theorem 4.1.2. Under H_3, the limit distribution of K^{*}_{n} is normal with mean vector zero and covariance matrix Σ*

Proof.

Let $\alpha = (\alpha_1, ..., \alpha_p)^T$ be any non zero real vector. Note that we can write

$$\left\{ E\left(\sum_{g=1}^{p} \alpha_g \frac{nK^*_{n(g)}}{4\sigma(S^*_{n(g)})} - \sum_{g=1}^{p} \alpha_g \frac{S^*_{n(g)}}{\sigma(S^*_{n(g)})} \right)^2 \right\}^{1/2} = \left\| \sum_{g=1}^{p} \alpha_g \frac{nK^*_{n(g)}}{4\sigma(S^*_{n(g)})} - \sum_{g=1}^{p} \alpha_g \frac{S^*_{n(g)}}{\sigma(S^*_{n(g)})} \right\|_2$$

$$\leq \sum_{g=1}^{p} |a_g| \left\| \frac{nK^*_{n(g)}}{4\sigma(S^*_{n(g)})} - \frac{S^*_{n(g)}}{\sigma(S^*_{n(g)})} \right\|_2 .$$

From (1.7),

$$\sum_{g=1}^{p} |a_g| \left\| \frac{nK^*_{n(g)}}{4\sigma(S^*_{n(g)})} - \frac{S^*_{n(g)}}{\sigma(S^*_{n(g)})} \right\|_2 = O(n^{-1/2})$$

and hence

$$E\left(\sum_{g=1}^{p} \alpha_g \frac{nK^*_{n(g)}}{4\sigma(S^*_{n(g)})} - \sum_{g=1}^{p} \alpha_g \frac{S^*_{n(g)}}{\sigma(S^*_{n(g)})} \right)^2 \to 0 \text{ as } n \to \infty$$

It follows from Theorem 2.2.5 that the limit distribution of $\alpha^T K^*_n$ is the same as that of $\alpha^T S^*_n$. Since the asymptotic distribution of S^*_n is multivariate normal with zero mean vector and covariance matrix Σ, $\alpha^T K^*_n$ has the same distribution as $\alpha^T N$, where N is multivariate normal with zero mean vector and dispersion matrix Σ. Reference to Theorem 2.2.4 completes the proof. □

It remains to determine a consistent estimator of Σ. We introduce some notation. For given $R'^*(.)$, let ${R'_g}^* \times {R'_h}^*$ be the collection of n vectors such that

$${R'_g}^* \times {R'_h}^* = \{({R'_{ig}}^*, {R'_{ih}}^*) | i = 1, ..., n\}$$

Let $\pi {R'_g}^*$ and $\pi({R'_g}^* \times {R'_h}^*)$ be ordered sets of n vectors such that

$$\pi {R'_g}^* = \{{R'_{\pi(i)g}}^* | i = 1, ..., n\}$$
$$\pi({R'_g}^* \times {R'_h}^*) = \{({R'_{\pi(i)g}}^*, {R'_{\pi(i)h}}^*) | i = 1, ..., n\}$$

where the j-th element is respectively ${R'_{\pi(j)g}}^*$ and $({R'_{\pi(j)g}}^*, {R'_{\pi(j)h}}^*)$. Then under H_3, $P[\pi(R'^*_g \times R'^*_h) | R'^*(.)] = 1/n!$.

Let $\mu = (\mu(1), ..., \mu(n))$ and $\nu = (\nu(1), ..., \nu(n))$ be permutations of $\{1, ..., n\}$. Also let $\pi\mu$ be the permutation $(\pi\mu(1), ..., \pi\mu(n))$ and define $K_n(\mu)$, $K_n(\mu, \nu)$ and $r_n(\mu, \nu)$ respectively by

$$\begin{aligned} K_n(\mu) &= \sum_{i<j} s(\mu(j) - \mu(i)) \\ K_n(\mu, \nu) &= \sum_{i<j} s(\mu(j) - \mu(i)) s(\nu(j) - \nu(i)) \\ r_n(\mu, \nu) &= \frac{3}{n^3 - n} \sum_{i,j,k} s(\mu(i) - \mu(j)) s(\nu(i) - \nu(k)) \end{aligned}$$

Lemma 4.2.1 *Under H_3,*

$$Cov\left[K^*_{n(g)}, K^*_{n(h)} | R'^*(.)\right] = K^*_{n(g,h)}/3 + (n^3 - n) r^*_{n(g,h)}/9$$

where $K^*_{n(g,h)} = \sum_{i<j} a_{n(h)}(i,j)\ a_{n(g)}(i,j)$ and $r^*_{n(g,h)}$ is defined in (4.11).

Proof.

Under H_3,

$$\begin{aligned} Cov\left[K^*_{n(g)}, K^*_{n(h)} | R'^*(.)\right] &= E\left[K^*_{n(g)} K^*_{n(h)} | R'^*_g \times R'^*_h\right] \\ &= \frac{1}{n!} \sum_{\pi \in \Pi_n} K^*_{n\pi(g)} K^*_{n\pi(h)} \end{aligned} \tag{4.16}$$

and

$$K^*_{n\pi(g)} = \frac{1}{\#(C(\pi R'^*_g))} \sum_{\mu \in C(\pi R'^*_g)} K_n(\mu)$$

where $\#(C(\pi R'^*_g))$ is the number of elements belonging to $C(\pi R'^*_g)$. Note that $\#(C(\pi R'^*_g)) = \#(C(R'^*_g))$ for any $\pi \in \Pi_n$. Similarly,

$$K^*_{n\pi(h)} = \frac{1}{\#(C(\pi R'^*_h))} \sum_{\nu \in C(\pi R'^*_h)} K_n(\nu)$$

Therefore (4.16) is equivalent to

$$\frac{1}{n!\ \#(C(R'^*_g))\ \#(C(R'^*_h))} \sum_{\pi \in \Pi_n} \sum_{\mu \in C(\pi R'^*_g)} \sum_{\nu \in C(\pi R'^*_h)} K_n(\mu) K_n(\nu) \tag{4.17}$$

It follows from

$$\sum_{\pi\in\Pi_n}\sum_{\mu\in C(\pi R'^*_g)} K_n(\mu) = \sum_{\mu\in C(R'^*_g)}\sum_{\pi\in\Pi_n} K_n(\pi\mu)$$

and

$$\sum_{\pi\in\Pi_n}\sum_{\nu\in C(\pi R'^*_h)} K_n(\nu) = \sum_{\nu\in C(R'^*_h)}\sum_{\pi\in\Pi_n} K_n(\pi\nu)$$

that (4.17) equals

$$\frac{1}{\#(C(R'^*_g))\ \#(C(R'^*_h))}\sum_{\mu\in C(R'^*_g)}\sum_{\nu\in C(R'^*_h)}\frac{1}{n!}\sum_{\pi\in\Pi_n} K_n(\pi\mu)K_n(\pi\nu)$$

From Theorem 3.1.2,

$$\frac{1}{n!}\sum_{\pi\in\Pi_n} K_n(\pi\mu)K_n(\pi\nu) = K_n(\mu,\nu)/3 + (n^3-n)r_n(\mu,\nu)/9$$

It is true from (1.6) that

$$\frac{1}{\#(C(R'^*_g))\ \#(C(R'^*_h))}\sum_{\mu\in C(R'^*_g)}\sum_{\nu\in C(R'^*_h)} K_n(\mu,\nu) = K^*_{n(g,h)}$$

and from (1.5)

$$\frac{1}{\#(C(R'^*_g))\ \#(C(R'^*_h))}\sum_{\mu\in C(R'^*_g)}\sum_{\nu\in C(R'^*_h)} r_n(\mu,\nu) = r^*_{n(g,h)}.$$

Summing up, we have shown that

$$Cov\left[K^*_{n(g)}, K^*_{n(h)}|R'^*(.)\right] \;=\; K^*_{n(g,h)}/3 + (n^3-n)r^*_{n(g,h)}/9\ \square$$

Lemma 4.2.2 *Assume the conditions of Theorem 4.2.1. Then, under H_3,*

$$Cov\left[nK^*_{n(g)}/4\sigma(S^*_{n(g)}), nK^*_{n(h)}/4\sigma(S^*_{n(h)})|R'^*(.)\right] \to \kappa_{gh}(\lambda_g\lambda_h)^{1/2} \text{ in probability}$$

Proof.

Note that

$$Cov\left[nK^*_{n(g)}/4\sigma(S^*_{n(g)}), nK^*_{n(h)}/4\sigma(S^*_{n(h)})|R'^*(.)\right] \;=\; \frac{n^2K^*_{n(g,h)}}{48\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})} + \frac{n^2(n^3-n)r^*_{n(g,h)}}{144\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})}$$

The fact that $\forall i < j,\ |a_{n(g)}(i,j)| \leq 1,\ g = 1, ..., p$ shows that

$$\begin{aligned}\left|\frac{n^2 K^*_{n(g,h)}}{48\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})}\right| &\leq \frac{n^2}{48\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})}\sum_{i<j}|a_{n(g)}(i,j)||a_{n(h)}(i,j)| \\ &\leq \frac{n^3(n-1)}{96\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})} \\ &= O(n^{-1}) \to 0\end{aligned}$$

because $\sigma(S^*_{n(g)}) = O(n^{5/2})$ and $\sigma(S^*_{n(h)}) = O(n^{5/2})$. On the other hand, comparison of $n^2(n^3 - n)r^*_{n(g,h)}/144\ \sigma(S^*_{n(g)})\sigma(S^*_{n(h)})$ with (4.10) shows that it converges to $\kappa_{gh}(\lambda_g\lambda_h)^{1/2}$ in probability because (4.10) does. The lemma follows. □

Let

$$\hat{\sigma}_{nK(g,h)} = Cov(nK^*_{n(g)}/4\sigma(S^*_{n(g)}), nK^*_{n(h)}/4\sigma(S^*_{n(h)})|R^*(.))$$

and

$$\hat{\Sigma}_{nK} = \begin{bmatrix} 1 & \hat{\sigma}_{nK(1,2)} & \cdots & \hat{\sigma}_{nK(1,p)} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{\sigma}_{nK(1,p)} & \hat{\sigma}_{nK(2,p)} & \cdots & 1 \end{bmatrix}.$$

In summary, Lemma 4.2.2 implies $\hat{\Sigma}_{nK} \to \Sigma$ in probability and we state the following counterpart to Corollary 4.1.2.

Corollary 4.2.1 *Under the same conditions as in Theorem 4.2.1, the limit distributions of* $K^{*T}_n \Sigma^- K^*_n$ *and* $K^{*T}_n \hat{\Sigma}^-_{nK} K^*_n$ *are* $\chi^2(rank(\Sigma))$

Chapter 5

Simulation

In this chapter, we report on the result of a simulation study where we compare the power of our test statistics with the naive statistic which ignores the missing observations.

We introduce some notation. Consider experiments of the following models denoted respectively by (a, ρ) and by (b, ρ):

$$(a, \rho) : y_i \sim \text{multinormal}\,((sp \times i, 0, 0), \Sigma) \;\; i = 1, ..., n$$

$$(b, \rho) : y_i \sim \text{multinormal}\,((sp \times i, -sp \times i, 0), \Sigma) \;\; i = 1, ..., n$$

where

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}.$$

and sp is the slope. We assume some missing observations occurr for the first and the second variables. The missing pattern obeys the conditions specified by H_3. By $(a, \rho, m = p)$, we denote the model (a, ρ) with the missing proportion p respectively for the first and the second variables.

We take two kinds of tests into account. The one is based upon statistic $S_n^* = (S_{n(1)}^*/\sigma(S_{n(1)}^*), S_{n(2)}^*/\sigma(S_{n(2)}^*), S_{n(3)}^*/\sigma(S_{n(3)}^*))$, and the other is based on the Spearman

measure after deleting the missing observations. Figure 1 exhibits the power of the test when the sample size is 20 and the models are $(a, \rho, m = p)$, where $\rho = .3, .5, .7$, $p = .1, .2$. Dashed lines denote our statistic and dotted ones denote the naive statistic. The figure makes a noticeable point. The weaker the correlation between variables and the higher the proportion of missing observations, the better the statistic S_n^* performs with respect to the naive statistic. Even though the naive appoach gains more power than S_n^* in the case of $(a, .7, m = .1)$, generally the proposed statistic is at least as good as the naive one. Figure 2 is the graph drawn under models $(b, \rho, m = p)$ with sample size 20, where the values ρ, p, n correspond to those in Figure 1. Figure 2 shows a similar pattern as Figure 1.

Figures 3 and 4 exhibit similar results but for a sample of size 30. Similarly our statistic does not perform well when the correlations between variables are high. Our statistic performs less efficiently than the naive statistic when the sample size is 20.

In order to determine when our statistic works well regardless of correlation, more simulations would need to be done. Since the general pattern of the graphs indicates that there is no significant difference between model (a, ρ) and (b, ρ), we will be concerned only with model (a, ρ). Suppose 20% of the observations are missing for each variable. Appling model (a, ρ), we obtain Figures 5 and 6. Figure 5 is done with a sample size of 30 and Figure 6 with a sample size of 40.

In summary, we conclude the following:

1. the choice of sample size and proportion of missing data has an effect upon the perfomance of our statistic.
2. when the sample size is large, our statistic is more powerful than the naive one whenever the proportion of missing observations is large.
3. when the correlations between variables is weak, our statistic performs better than the naive statistic irrespective of sample size and proportion of missing observations.

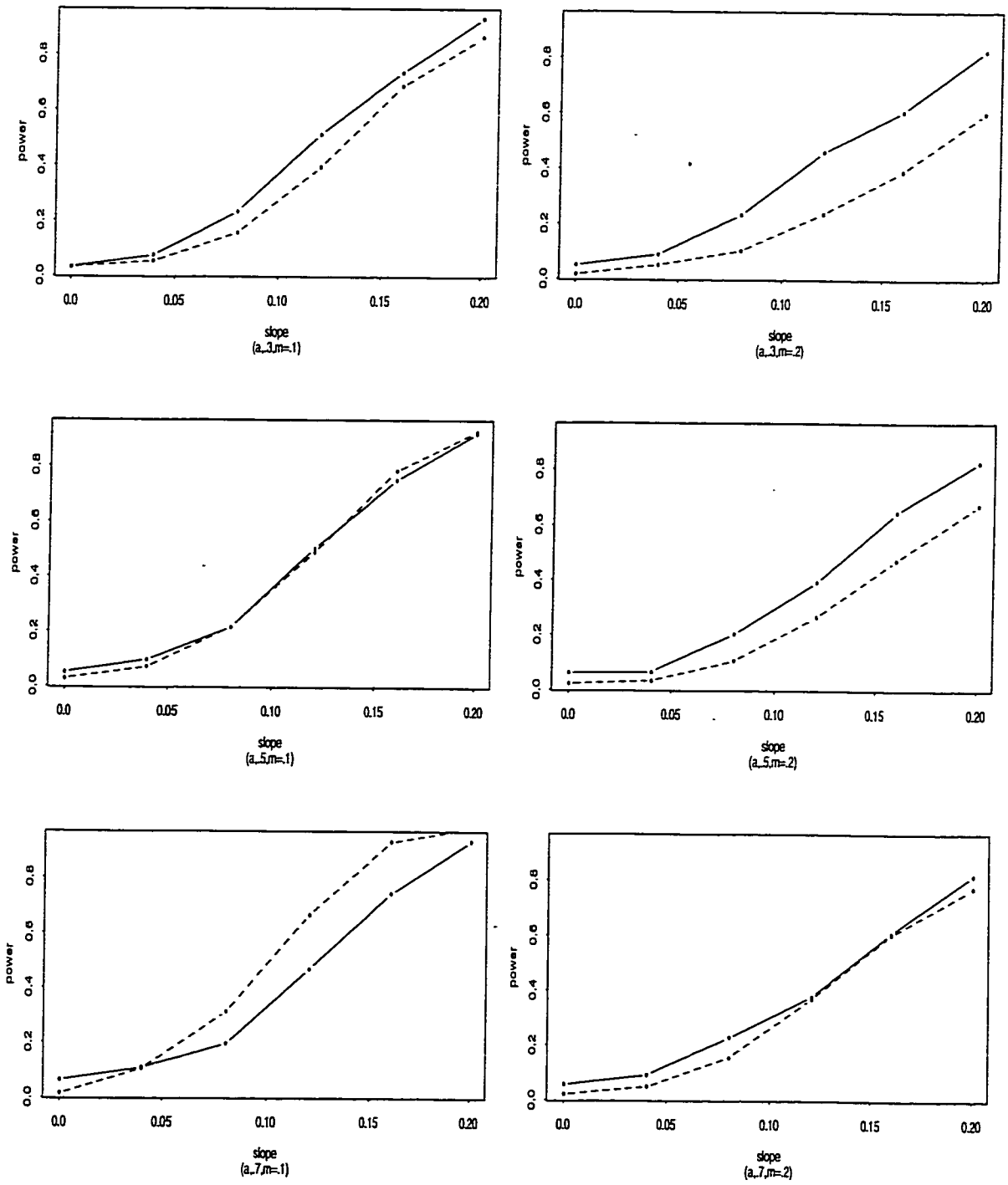

Figure 1: Power of $(a, \rho, m = p), \rho = .3, .5, .7, p = .1, .2$ when the sample size is 20

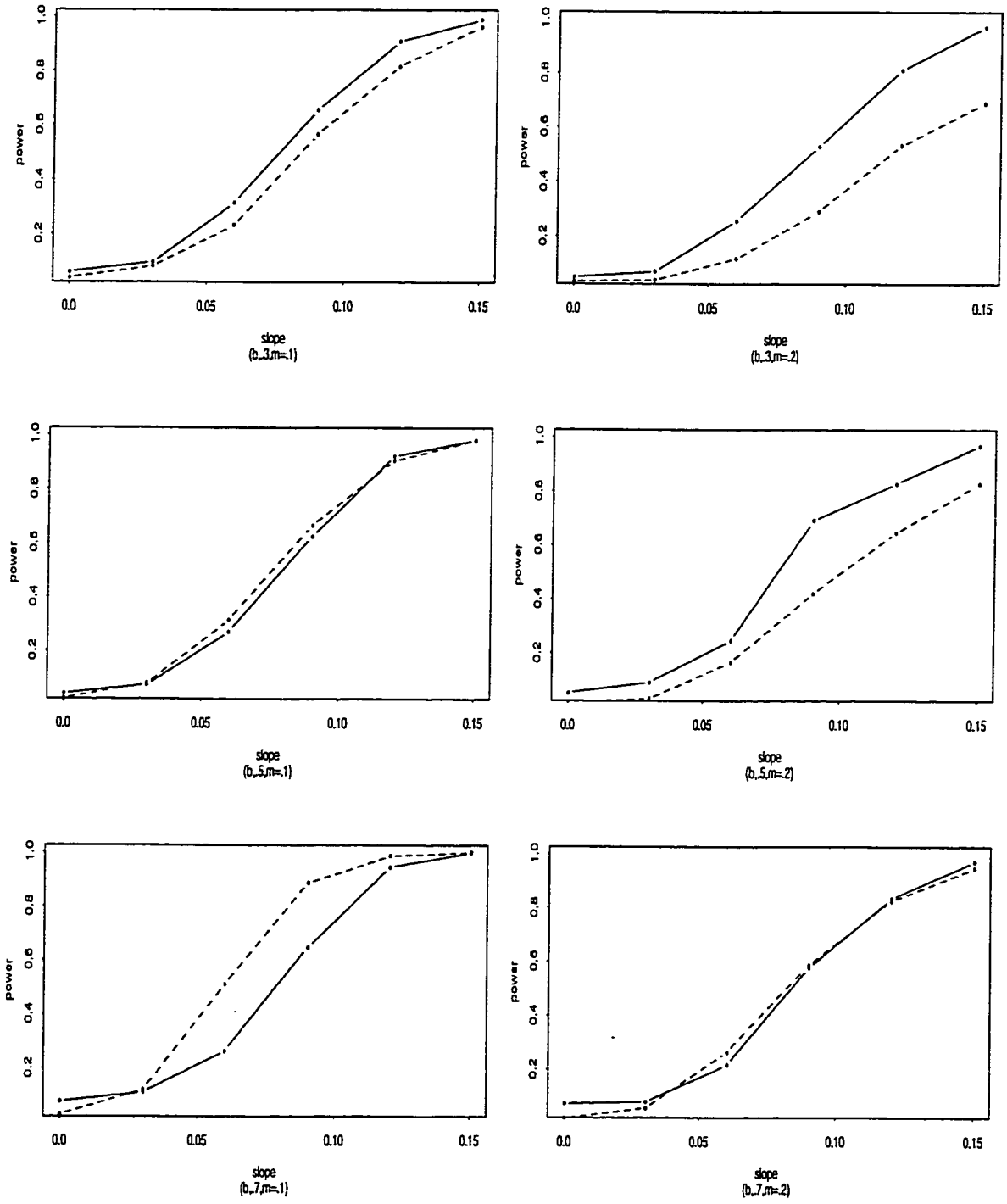

Figure 2: Power of $(b, \rho, m = p), \rho = .3, .5, .7, p = .1, .2$ when the sample size is 20

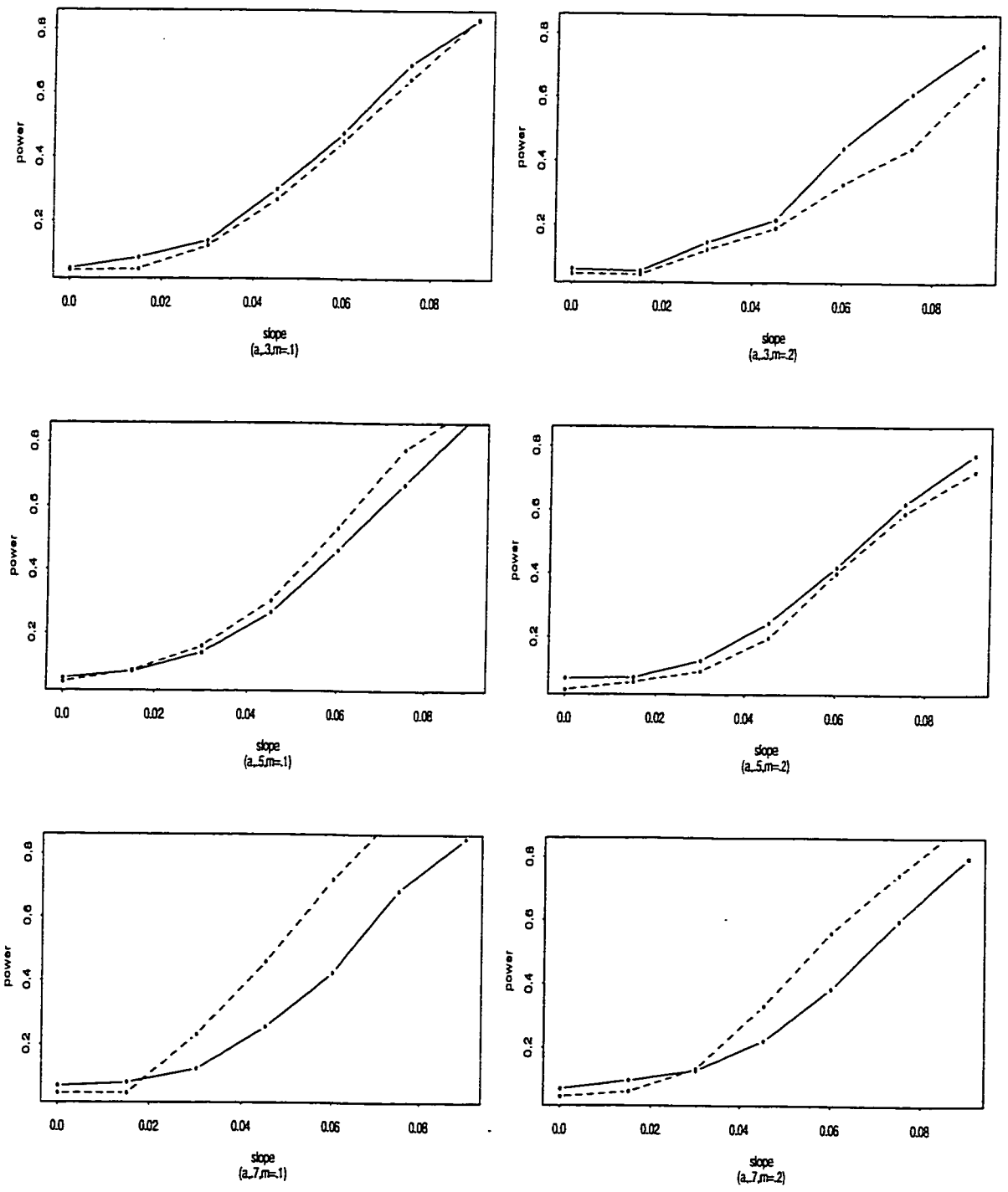

Figure 3: Power of $(a, \rho, m = p), \rho = .3, .5, .7, p = .1, .2$ when the sample size is 30

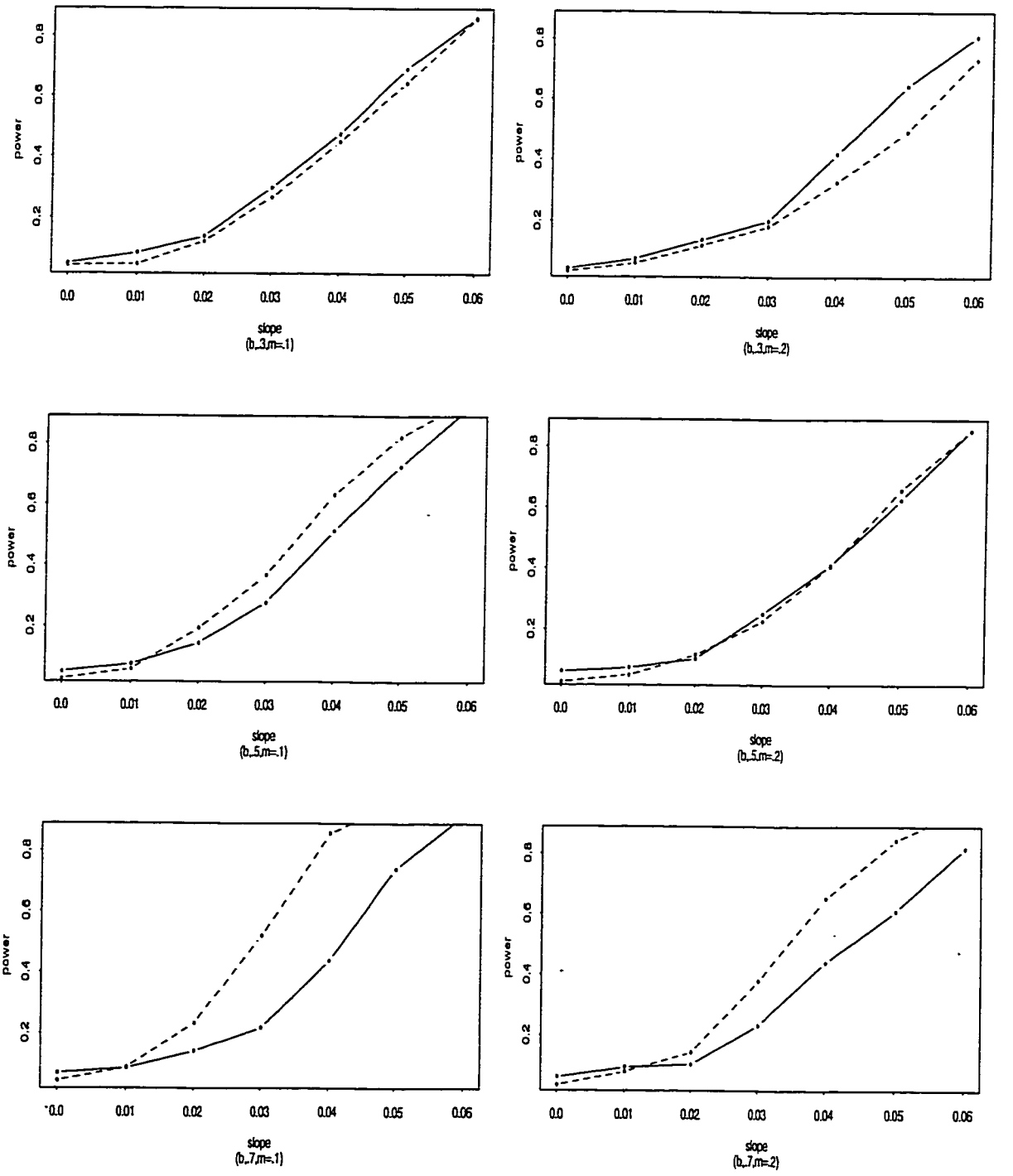

Figure 4: Power of $(b, \rho, m = p), \rho = .3, .5, .7, p = .1, .2$ when the sample size is 30

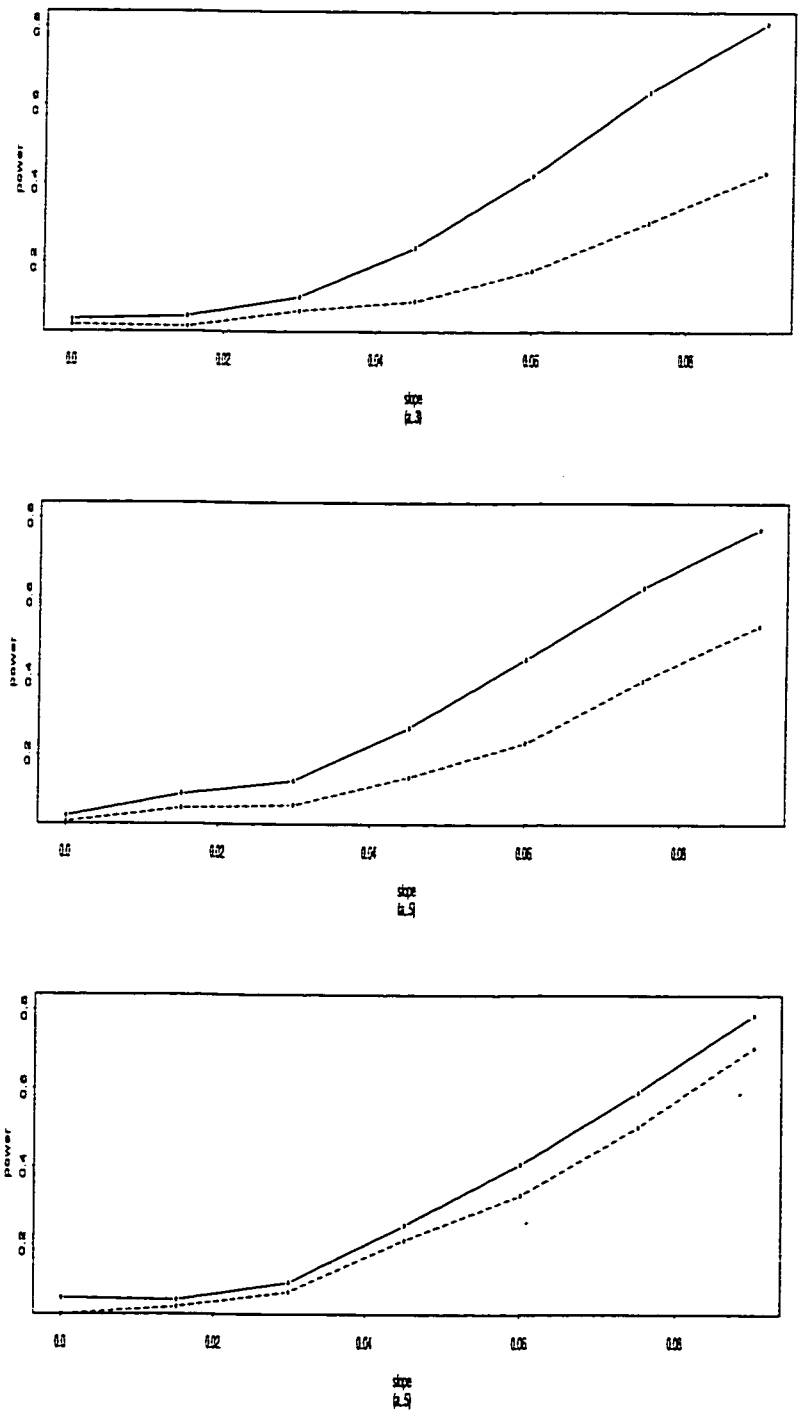

Figure 5: Power of $(a, \rho), \rho =, .3, .5, .7$ with 20 % missing for each variable when the sample size is 30

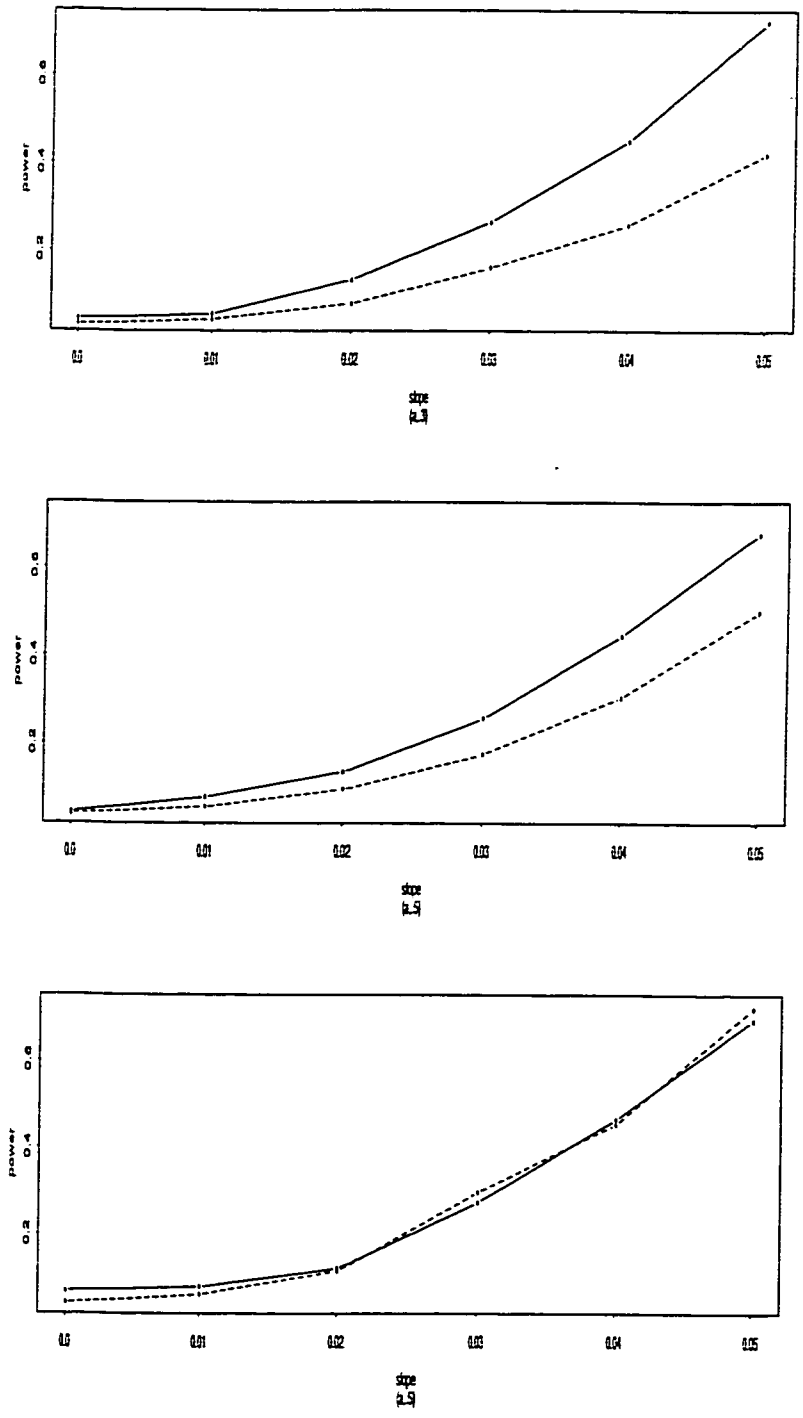

Figure 6: Power of $(a, \rho), \rho =, .3, .5, .7$ with 20 % missing for each variable when the sample size is 40

Appendix A

Splus Programme of The Simulation

We introduce the Splus programme of the simulation under the model $(a, .3, m = .1)$ with a sample of size 20.

Generate a sample from multivariate normal distribution with mean vector *mu* and covariane matrix *vmat*

```
rmultnorm_function(n,mu,vmat,tol=1e-07)
{
  p_ncol(vmat)
  if(length(mu)!=p)
         stop("mu vector is the wrong length")
  if (max(abs(vmat-t(vmat)))>tol)
         stop("vmat not symmetric")
  vs_svd(vmat)
  vsqrt_t(vs$v %*% (t(vs$u)*sqrt(vs$d)))
  ans_matrix(rnorm(n*p), nrow=n) %*% vsqrt
  ans_sweep(ans,2,mu,"+")
  dimnames(ans)_list(NULL, dimnames(vmat)[[2]])
```

```
  ans
}
```

Make an incomplete ranking with replacing missing observations with $(k+1)/2$

```
detect_function(x,r,k)
{
   for (i in 1:r){
   if (x[i,1]>k)  x[i,1]_0  else  x[i,1]_(x[i,1]-(k+1)/2)
   }
ans_x
ans
}
```

Calculate the Spearman rank correlation in the presence of misssing observations and calculate the variance.

```
spear_function(mu1,mu2)
{
 t_nrow(mu1)
 k1_length(mu1[!is.na(mu1)])
 k2_length(mu2[!is.na(mu2)])
 mu1_matrix(rank(mu1),ncol=1)
 mu2_matrix(rank(mu2),ncol=1)
 ans_((t+1)^2/((k1+1)*(k2+1)))*sum(detect(mu1,t,k1)*detect(mu2,t,k2))
 ans
}
vh1s_function(mu)
{
 t_nrow(mu)
 k1_length(mu[!is.na(mu)])
 ans_(k1*(k1-1)/(k1+1))*(t*(t+1)^3/144)
 ans
}
```

Main Programme

```
d_matrix(c(1:12),ncol=2)
y_matrix(1:60,ncol=3)
n_500
q1_0
q2_0
for (s in 1:6){
     for (i in 1:n){
            for (j in 1:20){
              mu_c(.04*(s-1)*j,0,0)
              vmat_c(1,.3,.3,.3,1,.3,.3,.3,1)
              vmat_matrix(vmat,ncol=3)
              y[j,]_rmultnorm(1,mu,vmat,tol=1e-07)
            }

          tau_matrix(c(1:20),ncol=1)
          t_nrow(tau)
          y1_matrix(y[,1],ncol=1)
          y2_matrix(y[,2],ncol=1)
          y3_matrix(y[,3],ncol=1)

          m11_sample(20,2)
          m12_sample(20,2)
          y1[c(m11),1]_NA
          y2[c(m12),1]_NA

          x_matrix(c(spear(tau,y1),spear(tau,y2),spear(tau,y3)),ncol=1)
          rho12_t(t+1)/12*spear(y1,y2)
          rho13_t(t+1)/12*spear(y1,y3)
          rho23_t(t+1)/12*spear(y2,y3)
```

```
x_matrix(c(spear(tau,y1),spear(tau,y2),spear(tau,y3)),ncol=1)
rho12_t(t+1)/12*spear(y1,y2)
rho13_t(t+1)/12*spear(y1,y3)
rho23_t(t+1)/12*spear(y2,y3)
rho_matrix(c(vh1s(y1),rho12,rho13,rho12,vh1s(y2),rho23,
rho13,rho23,vh1s(y3)),ncol=3)
result_t(x)%*%solve(rho)%*%x

if (pchisq(result,3)>.95) q1_q1+1


y1[c(m12),1]_NA
y1[c(m13),1]_NA
y2[c(m11),1]_NA
y2[c(m13),1]_NA
y3[c(m11),1]_NA
y3[c(m12),1]_NA
yy1_matrix(y1[!is.na(y1)],ncol=1)
yy2_matrix(y2[!is.na(y2)],ncol=1)
yy3_matrix(y3[!is.na(y3)],ncol=1)
kk_length(yy1)
tt_matrix(c(1:kk),ncol=1)

xx_matrix(c(spear(tt,yy1),spear(tt,yy2),spear(tt,yy3)),ncol=1)
rh12_kk*(kk+1)/12*spear(yy1,yy2)
rh13_kk*(kk+1)/12*spear(yy1,yy3)
rh23_kk*(kk+1)/12*spear(yy2,yy3)
rh_matrix(c(vh1s(yy1),rh12,rh13,rh12,vh1s(yy2),rh23,
rh13,rh23,vh1s(yy3)),ncol=3)
result_t(xx)%*%solve(rh)%*%xx
```

```
          if (pchisq(result,3)>.95) q2_q2+1
      }

       d[s,]_c(q1/500,q2/500)
}

print(d)
```

Bibliography

[1] M.Alvo and P.Cabilio (1991), On the Balanced Block Design for Rankings. Ann.Statist. 19, pp. 1597-1613

[2] M.Alvo and P.Cabilio (1995), Rank Correlation Methods for Missing Data. The Canadian Journal of Statistics 23, pp. 345-358

[3] G.K Bhattacharyya and J.H Klotz (1966), The Bivariate Trend of Lake Mendota, Technical Report No.98, University of Wisconsin, Madison, Dept. of Statistics.

[4] J.V. Bradley (1968), Distribution-Free Statistical Tests. Prentice-Hall.

[5] H.Cramér (1970), Random Variables and Probability Distributions. Third Edition. Cambridge Univ. Press.

[6] E.J. Dietz and T.J. Killeen (1981), A Nonparametric Multivariate Test for Monotone Trend With Pharmaceutical Applications. Journal of the American Statistical Association, 76, pp. 169-174

[7] D.A.S. Fraser (1957), Nonparametric Methods in Statistics. John Wiley & Sons.

[8] J.Hájek and Z.Šidák (1967), Theory of Rank Tests. Second Edition. Academic Press.

[9] W.Hoeffding (1948), A Class of Statistics With Asymptoticallly Normal Distribution. Ann.Statist. 19, pp. 293-325

[10] K.M Patel (1973), Hájek and Šidák Appoach to the Asymptotic Distribution of Multivariate Rank Order Statistics. Journal of Multivariate Analysis 3, pp. 57-70

[11] H.L Royden (1968) Real Analysis, Second Edition. The Macmilan Company.