Statistica Sinica **11**(2001), 1031-1045

A FILE LINKAGE PROBLEM OF DEGROOT AND GOEL REVISITED

Hock-Peng Chan and Wei-Liem Loh

National University of Singapore

Abstract: This article is concerned with the file linkage problem first investigated by DeGroot and Goel (1980). Let $X_1,\ldots,X_n$ be a random sample from a bivariate normal distribution. Suppose that before the sample can be observed, it is broken into the components $X_{1,1},\ldots,X_{1,n}$ and $X_{2,\psi(1)},\ldots,X_{2,\psi(n)}$ where $X_j=(X_{1,j},X_{2,j})'$ and ψ is some unknown permutation of $\{1,\ldots,n\}$. The aim is to estimate the parameters (in particular the correlation coefficient) of the bivariate normal distribution using the above broken random sample. The main difficulty here is that direct computation of the likelihood is in general a NP-hard problem. Thus for n sufficiently large, standard likelihood or Bayesian techniques may not be feasible. This article proposes to reformulate the problem as a moment problem via Fisher's k-statistics. The resulting likelihood can be approximated as a product of bivariate normal likelihoods and consequently standard statistical methods can be applied. It is also shown that this approximation is very good in that very little Fisher information is lost.

Key words and phrases: Bivariate normal distribution, broken random sample, correlation coefficient, file linkage, Fisher information, k-statistics.

1. Introduction

Let $X_1,\ldots,X_n$, $X_j=(X_{1,j},X_{2,j})'$, $j=1,\ldots,n$, be a simple random sample from a bivariate normal distribution with mean $\mu=(\mu_1,\mu_2)'$ and covariance matrix

$$\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}.$$

In this article we suppose that before $X_1,\ldots,X_n$ can be observed, the sample is broken into the components $X_{1,1}$, $\ldots$, $X_{1,n}$ and $X_{2,\psi(1)},\ldots,X_{2,\psi(n)}$ where ψ is some unknown permutation of $\{1,\ldots,n\}$. The aim of this article is to estimate the parameters $\rho,\mu_1,\mu_2,\sigma_1,\sigma_2$ using only the broken random sample $X_{1,(1)},\ldots,X_{1,(n)},X_{2,1},\ldots,X_{2,n}$ where $X_{1,(1)}\le\cdots\le X_{1,(n)}$ denotes the order statistics of $X_{1,1},\ldots,X_{1,n}$.

The problem was first investigated by DeGroot and Goel (1980). One motivation comes from the increasing interest in the methodology of merging data files to create comprehensive files from multiple but incomplete sources of data. For a detailed survey of related file linkage literature, we refer the reader to Copas and Hilton (1990), Goel and Ramalingam (1989) and the references cited therein.

Let $\mathcal{S}_n$ denote the set of all permutations of $\{1, \ldots, n\}$. We note, for example from Vaughan and Venables (1972), that

$$\begin{aligned}
&f_{X_{1,(1)},\ldots,X_{1,(n)},X_{2,1},\ldots,X_{2,n}}(x_{1,(1)}, \ldots, x_{1,(n)}, x_{2,1}, \ldots, x_{2,n})\\
=&f_{X_{1,(1)},\ldots,X_{1,(n)}|X_{2,1},\ldots,X_{2,n}}(x_{1,(1)}, \ldots, x_{1,(n)}|x_{2,1}, \ldots, x_{2,n})f_{X_{2,1},\ldots,X_{2,n}}(x_{2,1}, \ldots, x_{2,n})\\
=&\sum_{\psi\in\mathcal{S}_n}\prod_{i=1}^n f_{X_{1,i},X_{2,i}}(x_{1,(i)}, x_{2,\psi(i)}), \quad x_{1,(1)} \leq \cdots \leq x_{1,(n)}, -\infty < x_{2,1}, \ldots, x_{2,n} < \infty,
\end{aligned}$$

where $f_{X_{1,(1)},\ldots,X_{1,(n)},X_{2,1},\ldots,X_{2,n}}$ denotes the joint density of $X_{1,(1)}, \ldots, X_{1,(n)}, X_{2,1}, \ldots, X_{2,n}$, $f_{X_{1,(1)},\ldots,X_{1,(n)}|X_{2,1},\ldots,X_{2,n}}$ denotes the conditional density of $X_{1,(1)}, \ldots, X_{1,(n)}$ given $X_{2,1}, \ldots, X_{2,n}$, etc. Hence the log-likelihood function based on the broken random sample $X_{1,(1)}, \ldots, X_{1,(n)}, X_{2,1}, \ldots, X_{2,n}$ is

$$\begin{aligned}
&l(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\\
&= \log\Big\{(2\pi\sigma_1\sigma_2)^{-n}(1-\rho^2)^{-n/2}\sum_{\psi\in\mathcal{S}_n}\prod_{i=1}^n \exp\Big[-\frac{1}{2(1-\rho^2)}\\
&\qquad \times[(\frac{X_{1,(i)}-\mu_1}{\sigma_1})^2 - 2\rho(\frac{X_{1,(i)}-\mu_1}{\sigma_1})(\frac{X_{2,\psi(i)}-\mu_2}{\sigma_2}) + (\frac{X_{2,\psi(i)}-\mu_2}{\sigma_2})^2]\Big]\Big\}\\
&= -n\log(2\pi\sigma_1\sigma_2) - \frac{n}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\sum_{i=1}^n[(\frac{X_{1,(i)}-\mu_1}{\sigma_1})^2 + (\frac{X_{2,i}-\mu_2}{\sigma_2})^2]\\
&\qquad + \log\Big\{\sum_{\psi\in\mathcal{S}_n}\exp[\frac{\rho}{1-\rho^2}\sum_{i=1}^n(\frac{X_{1,(i)}-\mu_1}{\sigma_1})(\frac{X_{2,\psi(i)}-\mu_2}{\sigma_2})]\Big\},\\
&\qquad\quad X_{1,(1)} \leq \cdots \leq X_{1,(n)}, -\infty < X_{2,1}, \ldots, X_{2,n} < \infty. \qquad (1)
\end{aligned}$$

Remark. Vaughan and Venables (1972) observed that the above likelihood can be expressed as a matrix permanent. It is well known (see Valiant (1979)) that, in general, computing the permanent is a NP-hard problem.

We note that the above log-likelihood function was obtained by DeGroot and Goel (1980) using a Bayesian argument with a uniform distribution on $\mathcal{S}_n$. In that paper, it was called the integrated (or summed) likelihood function. DeGroot and Goel further obtained an expression for the Fisher information matrix $\mathcal{I}^{(n)}(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2)$ for the broken bivariate sample. However, except for the case $\rho = 0$, they were unable to compute $\mathcal{I}^{(n)}(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2)$ when n is large.

A Monte Carlo study of the above likelihood was performed by DeGroot and Goel (1980) for $n = 5$, and versions of the Metropolis-type Markov chain Monte Carlo algorithm were implemented by Liu (1994) and Wu (1995) to compute posterior probabilities in a Bayesian setting of the broken bivariate normal sample problem when $n = 100$. In this article, we are concerned with much larger sample sizes, for example $n = 1000$ or more, for which direct computation of the above likelihood may be infeasible.

The rest of this article is organized as follows. In Section 2, a power series expansion of $\mathcal{I}^{(n)}(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2)$ about $\rho = 0$ is obtained. We note that $\mathcal{I}^{(n)}(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2)$ involves summation of approximately $O(n!)$ terms and so again direct computation of $\mathcal{I}^{(n)}(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2)$ appears infeasible for large n.

Section 3 proposes that we reformulate the broken sample problem as a moment problem. In particular, Corollary 1 shows that the broken sample and its Fisher's k-statistics contain the same amount of Fisher information about the unknown parameters. Furthermore (9) indicates that the set of low order Fisher's k-statistics appears to contain most of the information. Hence we suggest ignoring the high order k-statistics and take only into consideration the low order ones. Moreover, it is well known, from the practical standpoint of robustness, that k-statistics of order more than four are seldom required (see for example McCullagh (1987, p.110)). Another advantage gained in keeping only the low order k-statistics is that, for n large, the likelihood function of these k-statistics can be well approximated by a product of bivariate normal likelihoods and hence standard (either frequentist or Bayesian) statistical techniques can be brought to bear on this problem (thus bypassing the NP-hard problem of computing the original likelihood as given in (1)).

In Section 4 we consider the estimation of the correlation coefficient ρ with the other parameters known (for simplicity). Theorem 3 shows that it is unlikely that ρ can be consistently estimated given only a broken bivariate normal sample. By expressing the difference between the information from the original broken sample and that of its low order k-statistics (up to order j, $j = 1, \ldots, 4$) as a power series in ρ, Theorem 4 shows that the difference is of order $O(\rho^{2j})$ for large n. Our calculations suggest that this result should hold for larger values of j as well. The reason for stopping at $j = 4$ is that k-statistics of higher orders are generally seldom used in practice (see previous paragraph).

On the other hand, lest we give the impression that consistent estimation of the correlation coefficient is impossible for problems of this nature, we conclude with a simple, though somewhat artificial example where consistent estimation of ρ is indeed possible, and that the k-statistics approach of this article shows that the information in this example is unbounded with increasing n.

Finally, the Appendix contains the proof of Proposition 1.

2. Fisher Information

Let $X_{1,1},\ldots,X_{1,n},X_{2,\psi(1)},\ldots,X_{2,\psi(n)}$ be as in Section 1. Following DeGroot and Goel (1980) define

$$\begin{aligned}
\beta_\psi &\equiv \sum_{i=1}^{n}(\frac{X_{1,i}-\mu_1}{\sigma_1})(\frac{X_{2,\psi(i)}-\mu_2}{\sigma_2}), \qquad (2)\\
\alpha_\psi &\equiv \exp(\frac{\rho\beta_\psi}{1-\rho^2})[\sum_{\phi\in\mathcal{S}_n}\exp(\frac{\rho\beta_\phi}{1-\rho^2})]^{-1},\\
\gamma_1 &\equiv \frac{1+\rho^2}{n(1-\rho^2)^2}E[\sum_{\psi\in\mathcal{S}_n}\beta_\psi^2\alpha_\psi-(\sum_{\psi\in\mathcal{S}_n}\beta_\psi\alpha_\psi)^2],\\
\gamma_2 &\equiv \frac{\rho^2}{n(1-\rho^2)}E[\sum_{\psi\in\mathcal{S}_n}\beta_\psi^2\alpha_\psi-(\sum_{\psi\in\mathcal{S}_n}\beta_\psi\alpha_\psi)^2].
\end{aligned}$$

Writing $(\theta_1,\ldots,\theta_5)\equiv(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$, we further define $\mathcal{I}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ to be the 5×5 matrix whose (i,j)th element is $\mathcal{I}_{i,j}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)\equiv -E[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\mu_1,\,\mu_2,\,\sigma_1,\,\sigma_2,\,\rho)]$, $i,\,j=1,\,\ldots,\,5$, and $l(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ is as in (1). DeGroot and Goel (1980, p.274), showed that the Fisher information matrix based on the broken random sample $X_{1,1},\ldots,X_{1,n},X_{2,\psi(1)},\ldots,\,X_{2,\psi(n)}$ is given by

$$\mathcal{I}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)=\frac{n}{1-\rho^2} \qquad (3)$$

$$\times\begin{pmatrix}
\sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} & 0 & 0 & 0\\
 & \sigma_2^{-2} & 0 & 0 & 0\\
 & & \sigma_1^{-2}(2-\rho^2-\gamma_2) & -\sigma_1^{-1}\sigma_2^{-1}(\rho^2+\gamma_2) & -\sigma_1^{-1}\rho(1-\gamma_1)\\
 & & & \sigma_2^{-2}(2-\rho^2-\gamma_2) & -\sigma_2^{-1}\rho(1-\gamma_1)\\
 & & & & (1+\rho^2)(1-\rho^2)^{-1}(1-\gamma_1)
\end{pmatrix}.$$

Remark. We note that there is an error in the expression for one of the elements of the information matrix on page 274 of DeGroot and Goel (1980).

We proceed to obtain a power series expansion of $E(\sum_{\psi\in\mathcal{S}_n}\beta_\psi\alpha_\psi)^2$ in ρ. Once that is done, the elements of $\mathcal{I}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ can be similarly approximated.

Proposition 1. *Let α_ψ and β_ψ be as in* (2). *Then expanding as a power series about $\rho=0$, we have*

$$E[(\sum_{\psi\in\mathcal{S}_n}\beta_\psi\alpha_\psi)^2]=1+(1+4n+n^2)\rho^2-(\frac{32}{n-1}+11+4n)\rho^4+O(\rho^6),$$

where $O(\rho^6)$ denotes terms of the order ρ^6.

We defer the proof of Proposition 1 to the Appendix.

Theorem 1. *Let $\mathcal{I}^{(n)}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ be as in* (3). *Then expanding formally as a power series, we have*

$$\begin{aligned}
\mathcal{I}^{(n)}_{3,3}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= \sigma_1^{-2}[2n + \rho^2 + (2n+3)\rho^4] + O(\rho^6),\\
\mathcal{I}^{(n)}_{3,4}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= -\sigma_1^{-1}\sigma_2^{-1}[(2n-1)\rho^2 - 3\rho^4] + O(\rho^6),\\
\mathcal{I}^{(n)}_{3,5}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= -\sigma_1^{-1}[\rho + 5\rho^3 + (2 - \frac{32}{n-1})\rho^5] + O(\rho^7),\\
\mathcal{I}^{(n)}_{5,5}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= 1 + 7\rho^2 + (14 - \frac{32}{n-1})\rho^4 + O(\rho^6).
\end{aligned}$$

Proof. We observe from DeGroot and Goel (1980) that $E\sum_{\psi\in\mathcal{S}_n}\beta_\psi^2\alpha_\psi = n + n^2\rho^2 + n\rho^2$, which, together with (2) and Proposition 1, proves the theorem.

3. Fisher's k-Statistics

Let $X_{i,1}, \ldots, X_{i,n}$ and $X_{i,(1)} < \cdots < X_{i,(n)}$, $i = 1, 2$, be as in Section 1. Its rth sample moment is defined to be

$$m_{i,r} \equiv n^{-1}\sum_{j=1}^{n} X^r_{i,(j)}, \qquad r = 1, 2, \ldots, n. \tag{4}$$

Proposition 2. *Given $m_{i,1}, \ldots, m_{i,n}$, the solution $X_{i,(1)}, \ldots, X_{i,(n)}$ of* (4) *is unique a.s. if it exists.*

Proof. Given $m_{i,1}, \ldots, m_{i,n}$, suppose there are two solutions of (4), say, $X_{i,(1)} < \cdots < X_{i,(n)}$ and $\xi_{i,(1)} < \cdots < \xi_{i,(n)}$. With probability 1, we can assume without loss of generality that there are no ties. Writing $Q_{X_{i,(1)},\ldots,X_{i,(n)}}(z) \equiv (z - X_{i,(1)})\cdots(z - X_{i,(n)})$, we have

$$\begin{aligned}
&\frac{1}{Q_{X_{i,(1)},\ldots,X_{i,(n)}}(z)}\frac{d}{dz}Q_{X_{i,(1)},\ldots,X_{i,(n)}}(z)\\
&= \frac{1}{z - X_{i,(1)}} + \cdots + \frac{1}{z - X_{i,(n)}}\\
&= \frac{1}{z} + \frac{m_{i,1}}{z^2} + \cdots + \frac{m_{i,n}}{z^{n+1}} + \text{terms of order } z^{-n-2}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\frac{1}{Q_{X_{i,(1)},\ldots,X_{i,(n)}}(z)}\frac{d}{dz}Q_{X_{i,(1)},\ldots,X_{i,(n)}}(z) - \frac{1}{Q_{\xi_{i,(1)},\ldots,\xi_{i,(n)}}(z)}\frac{d}{dz}Q_{\xi_{i,(1)},\ldots,\xi_{i,(n)}}(z)\\
&= \text{terms of order } z^{-n-2}.
\end{aligned} \tag{5}$$

Integrating both sides of (5) with respect to z, we obtain

$$\log[\frac{Q_{X_{i,(1)},\dots,X_{i,(n)}}(z)}{Q_{\xi_{i,(1)},\dots,\xi_{i,(n)}}(z)}] = \text{terms of the order } z^{-n-1}.$$

Since $Q_{X_{i,(1)},\dots,X_{i,(n)}}(z)$ and $Q_{\xi_{i,(1)},\dots,\xi_{i,(n)}}(z)$ are both monic polynomials in z of degree n, we conclude that they must be identical and hence $X_{i,(j)} = \xi_{i,(j)}$ for $j = 1, \dots, n$.

In Fisher (1929), a new class of symmetric functions of the observations was proposed, the so-called k-statistics, and it was shown that in many ways they are analytically more tractible than the sample moments. For example, the sampling cumulants of the k-statistics can be obtained by combinatorial methods. For $i = 1, 2$, let $\{k_{i,r} : r = 1, 2, \dots, n\}$ denote the family of Fisher's k-statistics based on $X_{i,1}, \dots, X_{i,n}$, respectively. Then $Ek_{i,r}$ equals $\kappa_{i,r}$, the rth cumulant of $N(\mu_i, \sigma_i^2)$, for all $r = 1, 2, \dots$. An explicit expression for $k_{i,r}$ is given by

$$k_{i,r} = r! \sum \frac{(-1)^{\eta-1}(\eta-1)!}{n(n-1)\dots(n-\eta+1)} \sum \frac{X_{i,\gamma_1}^{r_1} \dots X_{i,\gamma_{\pi_1}}^{r_1} \dots X_{i,\gamma_\eta}^{r_s}}{(r_1!)^{\pi_1} \dots (r_s!)^{\pi_s} \pi_1! \dots \pi_s!},$$

where the second summation extends over all ways of assigning the $\pi_1 + \dots + \pi_s = \eta$ distinct subscripts $\gamma_1, \dots, \gamma_\eta$ (including permutations) from the n available, and the first summation extends over all partitions $(r_1^{\pi_1} \cdots r_s^{\pi_s})$ of the number $r = r_1\pi_1 + \dots + r_s\pi_s$. As illustrations, we have

$$k_{i,1} = \frac{1}{n} m_{i,1}, \qquad k_{i,2} = \frac{1}{n(n-1)}(nm_{i,2} - m_{i,1}^2),$$
$$k_{i,3} = \frac{1}{n(n-1)(n-2)}(n^2 m_{i,3} - 3nm_{i,2}m_{i,1} + 2m_{i,1}^3).$$

For a very readable account of k-statistics, we refer the reader to Chapter 12 of Stuart and Ord (1994). We further observe from Wishart (1929) that

$$\mathrm{Corr}(k_{i,r}, k_{j,s}) = \begin{cases} \rho^r, & \text{if } r = s \text{ and } i \neq j, \\ 0, & \text{if } r \neq s, \end{cases}$$

and

$$\mathrm{Var}(k_{i,r}) = \sigma_i^{2r} \Upsilon_r, \tag{6}$$

where

$$\Upsilon_r \equiv r! \sum_{s=1}^{r} \frac{(s-1)!\Delta^s 0^r}{sn(n-1)\cdots(n-s+1)}, \qquad r = 1, 2, \dots,$$

and $\Delta^s 0^r$ is defined as the sth difference of $|t|^r$ when $t = 0$.

Proposition 3. *With probability* 1*, there is a one-to-one correspondence between the sample $X_{i,1}, \dots, X_{i,n}$ and its k-statistics $k_{i,1}, \dots, k_{i,n}$.*

Proof. This follows from Proposition 2 and the well known fact that $m_{i,1}$, …, $m_{i,n}$ determines $k_{i,1}$, …, $k_{i,n}$ and vice versa (see, for example, Stuart and Ord (1994, p.422)).

Corollary 1. *The broken random sample $X_{1,(1)}, \dots, X_{1,(n)}, X_{2,1}, \dots, X_{2,n}$, and its k-statistics $k_{1,1}, \dots, k_{1,n}, k_{2,1}, \dots, k_{2,n}$, have the same Fisher information matrix $\mathcal{I}^{(n)}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, where $\mathcal{I}^{(n)}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is defined as in* (3)*.*

It is convenient to define the normalized k-statistics

$$k^*_{i,r} = (k_{i,r} - \kappa_{i,r})/[\mathrm{Var}(k_{i,r})]^{1/2}, \qquad i = 1, 2, \qquad r = 1, 2, \dots, n. \tag{7}$$

Proposition 4. *Let m be a fixed positive integer. Then as $n \to \infty$, $(k^*_{1,1}, k^*_{2,1}, \dots, k^*_{1,m}, k^*_{2,m})'$ converges in distribution to the $2m$-variate normal distribution with mean* 0 *and covariance matrix*

$$\Sigma_m = \begin{pmatrix} 1 & \rho & 0 & 0 & \cdots & 0 & 0 \\ \rho & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \rho^2 & \cdots & 0 & 0 \\ 0 & 0 & \rho^2 & 1 & \cdots & 0 & 0 \\ \vdots & & & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & \rho^m \\ 0 & 0 & 0 & 0 & \cdots & \rho^m & 1 \end{pmatrix}.$$

Proof. See McCullagh (1987, p.135).

The following theorem gives an approximation for the Fisher information matrix $\mathcal{I}^{(n)}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

Theorem 2. *Let $m \geq 3$. Then an asymptotic "lower bound" approximation for the Fisher information matrix $\mathcal{I}^{(n)}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is given by the Fisher information matrix $\hat{\mathcal{I}}^{(n)}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ of a $2m$-variate normal random vector $Y = (Y_1, \dots, Y_{2m})'$ with mean ν and covariance matrix $D_m \Sigma_m D_m$ where Σ_m is as in Proposition* 4*, $\nu = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, 0, \dots, 0)'$, and $D_m = (d_{i,j})$ denotes a $2m \times 2m$ diagonal matrix such that for $i = 1, \dots, m$,*

$$d_{2i-1,2i-1} \equiv (\mathrm{Var}(k_{1,i}))^{1/2}, \qquad d_{2i,2i} \equiv (\mathrm{Var}(k_{2,i}))^{1/2}. \tag{8}$$

In particular,

$$\hat{\mathcal{I}}^{(n)}_{1,1}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{n}{(1 - \rho^2)\sigma_1^2},$$

$$\hat{\mathcal{I}}^{(n)}_{1,2}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = -\frac{n\rho}{(1-\rho^2)\sigma_1\sigma_2},$$

$$\hat{\mathcal{I}}^{(n)}_{3,3}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \frac{2n}{(1-\rho^4)\sigma_1^2} + \frac{\rho^2}{(1+\rho^2)\sigma_1^2} + \sum_{i=2}^m \frac{2i^2 - i^2\rho^{2i}}{(1-\rho^{2i})\sigma_1^2},$$

$$\hat{\mathcal{I}}^{(n)}_{3,4}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = -\frac{2n\rho^2 - \rho^2 + 5\rho^4}{(1-\rho^4)\sigma_1\sigma_2} - \sum_{i=3}^m \frac{i^2\rho^{2i}}{(1-\rho^{2i})\sigma_1\sigma_2},$$

$$\hat{\mathcal{I}}^{(n)}_{3,5}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = -\sum_{i=1}^m \frac{i^2\rho^{2i-1}}{(1-\rho^{2i})\sigma_1},$$

$$\hat{\mathcal{I}}^{(n)}_{5,5}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \sum_{i=1}^m \frac{i^2\rho^{2(i-1)}(1+\rho^{2i})}{(1-\rho^{2i})^2}.$$

Proof. The first statement of Theorem 2 follows from Corollary 1 and Proposition 4. We observe that the likelihood function of $Y = (Y_1,\ldots,Y_{2m})'$ is

$$f_Y(y_1,\ldots,y_{2m}) = \prod_{i=1}^m \frac{1}{2\pi d_{2i-1,2i-1}d_{2i,2i}(1-\rho^{2i})^{1/2}} \exp\{-\frac{1}{2(1-\rho^{2i})}[(\frac{y_{2i-1}-\kappa_{1,i}}{d_{2i-1,2i-1}})^2 + (\frac{y_{2i}-\kappa_{2,i}}{d_{2i,2i}})^2 - 2\rho^i(\frac{y_{2i-1}-\kappa_{1,i}}{d_{2i-1,2i-1}})(\frac{y_{2i}-\kappa_{2,i}}{d_{2i,2i}})]\},$$

and from (6) and (8), we have

$$\begin{aligned}
&\log f_Y(y_1,\ldots,y_{2m})\\
=& -\frac{1}{2(1-\rho^2)}[(\frac{y_1-\mu_1}{\sigma_1\Upsilon_1^{1/2}})^2 + (\frac{y_2-\mu_2}{\sigma_2\Upsilon_1^{1/2}})^2 - 2\rho(\frac{y_1-\mu_1}{\sigma_1\Upsilon_1^{1/2}})(\frac{y_2-\mu_2}{\sigma_2\Upsilon_1^{1/2}})]\\
&-\frac{1}{2(1-\rho^4)}[(\frac{y_3-\sigma_1^2}{\sigma_1^2\Upsilon_2^{1/2}})^2 + (\frac{y_4-\sigma_2^2}{\sigma_2^2\Upsilon_2^{1/2}})^2 - 2\rho^2(\frac{y_3-\sigma_1^2}{\sigma_1^2\Upsilon_2^{1/2}})(\frac{y_4-\sigma_2^2}{\sigma_2^2\Upsilon_2^{1/2}})]\\
&-\sum_{i=3}^m\{\frac{1}{2(1-\rho^{2i})}[(\frac{y_{2i-1}}{\sigma_1^i\Upsilon_i^{1/2}})^2 + (\frac{y_{2i}}{\sigma_2^i\Upsilon_i^{1/2}})^2 - 2\rho^i(\frac{y_{2i-1}}{\sigma_1^i\Upsilon_i^{1/2}})(\frac{y_{2i}}{\sigma_2^i\Upsilon_i^{1/2}})]\}\\
&-\sum_{i=1}^m\{\log(2\pi) + \log(\sigma_1^i\Upsilon_i^{1/2}) + \log(\sigma_2^i\Upsilon_i^{1/2}) + \frac{1}{2}\log(1-\rho^{2i})\}.
\end{aligned}$$

Now we observe from the definition of the information matrix that

$$\hat{\mathcal{I}}^{(n)}_{5,5}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = E[-\frac{\partial^2 \log f_Y(Y_1,\ldots,Y_{2m})}{\partial\rho^2}] = \sum_{i=1}^m \frac{i^2\rho^{2(i-1)}(1+\rho^{2i})}{(1-\rho^{2i})^2}.$$

The other elements of $\hat{\mathcal{I}}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ can be similarly evaluated and the proof of Theorem 2 is complete.

Remark. From (3) and Theorem 2, $\hat{\mathcal{I}}_{i,j}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \mathcal{I}_{i,j}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ for $1 \leq i \leq 2, 1 \leq j \leq 5$, and $\lim_{|\rho|\to 1} \hat{\mathcal{I}}_{i,j}^{(n)}\ (\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \lim_{|\rho|\to 1} \mathcal{I}_{i,j}^{(n)}$ $(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$, for $1 \leq i,j \leq 5$. Furthermore Theorem 1 and Theorem 2 show that

$$\begin{aligned}
\mathcal{I}_{3,3}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) - \hat{\mathcal{I}}_{3,3}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) &= -\sigma_1^{-2}\sum_{i=2}^{m} 2i^2 + O(\rho^6),\\
\mathcal{I}_{3,4}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) - \hat{\mathcal{I}}_{3,4}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) &= 8\rho^4\sigma_1^{-1}\sigma_2^{-1} + O(\rho^6),\\
\mathcal{I}_{3,5}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) - \hat{\mathcal{I}}_{3,5}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) &= \frac{8(n+3)\rho^5}{\sigma_1(n-1)} + O(\rho^6),\\
\mathcal{I}_{5,5}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) - \hat{\mathcal{I}}_{5,5}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) &= -\frac{32\rho^4}{n-1} + O(\rho^6).
\end{aligned} \tag{9}$$

For $\rho \neq 0$, since $\mathcal{I}_{3,3}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ and $\mathcal{I}_{3,4}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ both tend to infinity as $n \to \infty$, the approximations given by $\hat{\mathcal{I}}^{(n)}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ appear to be remarkably good.

4. Estimating the Correlation Coefficient

As in the Introduction, let $X_{1,(1)},\ldots,X_{1,(n)},X_{2,1},\ldots,X_{2,n}$ denote a broken random bivariate normal sample with parameters $\rho,\mu_1,\mu_2,\sigma_1,\sigma_2$. For simplicity, we assume in this section that μ_1,μ_2,σ_1 and σ_2 are known and hence without lost of generality, set $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$.

Theorem 3. *Let* $\hat{\rho}(X_{1,(1)},\ldots,X_{1,(n)},X_{2,1},\ldots,X_{2,n}) \in [-1,1]$ *be an estimator for* $\rho \in (-1,1)$. *Suppose that there exists a constant* $\varepsilon > 0$ *such that*

$$\liminf_{n\to\infty} [\frac{\partial}{\partial\rho}E_\rho(\hat{\rho})]^2|_{\rho=0} \geq \varepsilon. \tag{10}$$

Then $\hat{\rho}$ *does not converge in probability to* 0 *as* $n \to \infty$ *when* $\rho = 0$. (E_ρ *denotes expectation when* ρ *is the value correlation coefficient*).

Proof. First we observe from the information inequality (see, for example, Theorem 6.4 of Lehmann (1983)) that

$$E_\rho(\hat{\rho} - E_\rho\hat{\rho})^2 \geq [\frac{\partial}{\partial\rho}E_\rho(\hat{\rho})]^2/i_n(\rho), \tag{11}$$

where $i_n(\rho)$ denotes the Fisher information about ρ in the broken random sample. DeGroot and Goel (1980, p.277), showed that $i_n(0) = 1$. Hence it follows from (10) and (11) that

$$\liminf_{n\to\infty} E_0(\hat{\rho} - E_0\hat{\rho})^2 \geq \varepsilon.$$

This implies that $\hat{\rho}$ does not converge in probability to 0 as $n \to \infty$ when $\rho = 0$.

Remark. Condition (10) is very weak and is satisfied for "sufficiently smooth" estimators of ρ that are asymptotically unbiased. Theorem 3 indicates that consistent estimation of ρ is unlikely.

Remark. With $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 1$, the Fisher information about ρ in the broken bivariate normal sample is $i_n(\rho) = \mathcal{I}_{5,5}^{(n)}(0,0,1,1,\rho)$, where $\mathcal{I}_{5,5}^{(n)}(0,0,1,1,\rho)$ is as in (3). We believe that the use of a set of low order k-statistics is the appropriate approach to summarize the amount of information about ρ in the broken random sample. The key observation is in (6) where it is seen that different k-statistics of unequal orders are uncorrelated and more importantly, the correlation between different k-statistics of the same order decreases geometrically with increasing order. This indicates that the information about ρ contained in k-statistics should decrease very rapidly with order and hence high order k-statistics could be omitted without significant information loss.

Theorem 4. *With the above notation, expanding as a power series about $\rho = 0$, we have*

$$\lim_{n\to\infty} i_n(\rho) = \sum_{i=1}^{j} \frac{i^2 \rho^{2(i-1)}(1+\rho^{2i})}{(1-\rho^{2i})^2} + O(\rho^{2j}), \qquad j = 1, \ldots, 4. \tag{12}$$

Proof. The proof is similar (though obviously more tedious) to that of (9), the reader is referred to Chan and Loh (2000b) for the proof.

Remark. We observe that the first term on the right hand side of (12) is asymptotically (as $n \to \infty$) equal to the information about ρ contained in the set of low order k-statistics up to and including order j. Our calculations seem to further indicate that (12) holds for larger values of j as well. Indeed we believe the following conjecture to be true.

Conjecture. $\lim_{n\to\infty} i_n(\rho) = \sum_{i=1}^{\infty} i^2 \rho^{2(i-1)}(1+\rho^{2i})(1-\rho^{2i})^{-2}$, for all $\rho \in (-1,1)$.

On the other hand, here is an example (though somewhat artificial) where consistent estimation of ρ is indeed possible.

Let $\rho \in [-1,1]$ and $X_{1,1}, \ldots, X_{1,n}, Z_1, \ldots, Z_n, J_1, \ldots, J_n$ be a sequence of independent random variables where, for $i = 1, 2, \ldots, n$, $X_{1,i} \sim N(0,1)$, $Z_i \sim N(0,1)$ and $P(J_i = 1) = |\rho|$, $P(J_i = 0) = 1 - |\rho|$. Now define for $i = 1, \ldots, n$,

$$X_{2,i} = \begin{cases} X_{1,i}J_i + Z_i(1-J_i) & \text{if } \rho \geq 0, \\ -X_{1,i}J_i + Z_i(1-J_i) & \text{if } \rho < 0. \end{cases}$$

We would like to estimate ρ using the broken sample $X_{1,(1)}$, …, $X_{1,(n)}$, $X_{2,1}$, …, $X_{2,n}$. It is easily seen that $E(X_{1,1}X_{2,1}) = \rho$. A key point to note is that both marginal distributions of the broken sample are standard normal, independent of ρ, but their joint distribution is not bivariate normal. As in Section 3, for $i = 1, 2$ and $r = 1, \ldots, n$, let $k_{i,r}$ denote the rth order k-statistic based on $X_{i,1}, \ldots, X_{i,n}$ and consequently $E[k_{i,r}] = \kappa_r$, the rth cumulant of the standard normal distribution. Define the normalized k-statistics $k^*_{i,r}$ as in (7).

Lemma 1. *Let r and s be fixed positive integers. Then*

$$\lim_{n\to\infty} \mathrm{Corr}(k_{1,r}, k_{2,s}) = \begin{cases} \rho & \text{if } r = s, \\ 0 & \text{if } r \neq s. \end{cases}$$

Proof. For any positive integer p, let $\mathcal{S}_{n,p}$ be the set of all injective functions $\psi : \{1, 2, \ldots, p\} \to \{1, 2, \ldots, n\}$. Thus $|\mathcal{S}_{n,p}| = n^{[p]} \equiv n(n-1)\cdots(n-p+1)$. Let $\mathcal{P}(r)$ denote the set of all partitions of an integer r (see, for example, Chapter 1 of Andrews (1976)). In any partition $\lambda \in \mathcal{P}(r)$, let $\lambda_i > 0$ denote the number of times r_i is repeated for $i = 1, \ldots, t_\lambda$. Then $\sum_{i=1}^{t_\lambda} r_i\lambda_i = r$ and we write

$$\#(\lambda) \equiv \sum_{i=1}^{t_\lambda} \lambda_i, \qquad C_\lambda \equiv \frac{(-1)^{\#(\lambda)-1}[\#(\lambda)-1]!r!}{(r_1!)^{\lambda_1}\cdots(r_{t_\lambda}!)^{\lambda_{t_\lambda}}\lambda_1!\cdots\lambda_{t_\lambda}!}.$$

It follows now from Stuart and Ord (1994, p.420), that for $i = 1, 2$,

$$k_{i,r} = \sum_{\lambda\in\mathcal{P}(r)} C_\lambda (n^{[\#(\lambda)]})^{-1} \sum_{\psi\in\mathcal{S}_{n,\#(\lambda)}} X^{r_1}_{i,\psi(1)} \cdots X^{r_{t_\lambda}}_{i,\psi(\#(\lambda))}.$$

Hence, via symmetry,

$$\begin{aligned} &Ek_{1,r}k_{2,s} - Ek_{1,r}Ek_{2,s} \\ &= \sum_{\lambda\in\mathcal{P}(r)} \sum_{\theta\in\mathcal{P}(s)} C_\lambda C_\theta (n^{[\#(\theta)]})^{-1} \sum_{\psi\in\mathcal{S}_{n,\#(\theta)}} \{E[X^{r_1}_{1,1}\cdots X^{r_{t_\lambda}}_{1,\#(\lambda)} X^{s_1}_{2,\psi(1)} \cdots X^{s_{t_\theta}}_{2,\psi(\#(\theta))}] \\ &\quad - E[X^{r_1}_{1,1}\cdots X^{r_{t_\lambda}}_{1,\#(\lambda)}]E[X^{s_1}_{2,\psi(1)}\cdots X^{s_{t_\theta}}_{2,\psi(\#(\theta))}]\}. \qquad (13) \end{aligned}$$

Note that $\mathcal{P}(r)$ and $\mathcal{P}(s)$ do not depend on n. Fix λ and θ. Each term in the inner sum is 0 if the subscripts of $X_{1,1}, \ldots, X_{1,\#(\lambda)}$ and $X_{2,\psi(1)}, \ldots, X_{2,\psi(\#(\theta))}$ do not match is 0. If exactly one subscript matches, say $\psi(1) = 1$, then it becomes $\rho(EX^{r_1+s_1}_{1,1} - EX^{r_1}_{1,1}EX^{s_1}_{1,1})Q(\lambda, \theta)$, where $Q(\lambda, \theta)$ is a quantity independent of n. The number of terms with one subscript matching is $\#(\lambda)\#(\theta)(n-\#(\lambda))^{[\#(\theta)-1]}$. Similarly for exactly $\nu > 1$ matches, the number of terms is $C^{\#(\lambda)}_\nu \#(\theta)^{[\nu]}(n - \#(\lambda))^{[\#(\theta)-\nu]}$. Thus by (13), we have

$$\mathrm{Cov}(k_{1,r}, k_{2,s}) = \frac{\beta_{r,s}\rho}{n} + O(n^{-2}) \qquad (14)$$

for some constant $\beta_{r,s}$. Letting $\rho = 1$ in (14), we get $\mathrm{Cov}(k_{1,r}, k_{1,s}) = \beta_{r,s} n^{-1} + O(n^{-2})$. Hence we conclude from (6) that $\beta_{r,r} \neq 0$, $\beta_{r,s} = 0$ for all $r \neq s$, and that $\mathrm{Var}(k_{1,r}) = \mathrm{Var}(k_{2,r})$ is exactly of order $O(n^{-1})$. This proves Lemma 1.

Hence as in Proposition 4, for an arbitrary but fixed m and suitably large n, the likelihood of $(k^*_{1,1}, k^*_{2,1}, \ldots, k^*_{1,m}, k^*_{2,m})'$ can be approximated by a product of bivariate normal likelihoods each with mean 0 and covariance matrix Σ where $\Sigma_{1,1} = \Sigma_{2,2} = 1$ and $\Sigma_{1,2} = \rho$. This implies that the amount of Fisher information about ρ is unbounded with increasing m. This is in direct contrast to the broken bivariate normal sample treated previously. In fact consistent estimation of ρ is indeed possible here and a simple consistent estimate is given by $\tilde{\rho} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n I\{X_{1,i} = X_{2,j}\} - \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n I\{X_{1,i} = -X_{2,j}\}$, where $I\{.\}$ denotes the indicator function.

Acknowledgements

We thank Professors Prabir Burman, Tailen Hsing and Jun Liu for many fruitful discussions on this problem. We are also grateful to the Co-editor, Professor Ker-Chau Li, and a referee for their detailed comments and suggestions.

Appendix

Proof of Proposition 1. Let E_ψ denote the conditional expectation with respect to the uniform distribution on $\psi \in \mathcal{S}_n$ keeping the broken sample $X_{1,(1)}$, $\ldots$, $X_{1,(n)}$, $X_{2,1}$, $\ldots$, $X_{2,n}$ fixed. For simplicity, write $Y_{i,j} \equiv \frac{X_{i,j}-\mu_i}{\sigma_i\sqrt{1-\rho^2}}$, $i = 1, 2$, $j = 1, \ldots, n$, $\tilde{\beta}_\psi \equiv \sum_{i=1}^n Y_{1,i} Y_{2,\psi(i)}$, $\psi \in \mathcal{S}_n$, $\tilde{\alpha}_\psi \equiv \exp(\rho\tilde{\beta}_\psi)[E_\psi \exp(\rho\tilde{\beta}_\psi)]^{-1}$, $\psi \in \mathcal{S}_n$. Let E^* denote the expectation for which $Y_{1,1}, \ldots, Y_{1,n}, Y_{2,1}, \ldots, Y_{2,n}$ are i.i.d. $N(0,1)$ random variables. We observe from (2) that

$$
\begin{aligned}
&E[(\sum_{\psi\in\mathcal{S}_n} \beta_\psi \alpha_\psi)^2] \\
&= \frac{1}{n!}\int_{R^{2n}} [\sum_{\psi\in\mathcal{S}_n} \beta_\psi e^{\rho\beta_\psi(1-\rho^2)^{-1}}]^2 [\sum_{\psi\in\mathcal{S}_n} e^{\rho\beta_\psi(1-\rho^2)^{-1}}]^{-1} (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-n} \\
&\qquad \times \exp\{-\frac{1}{2(1-\rho^2)}\sum_{i=1}^n [(\frac{X_{1,i}-\mu_1}{\sigma_1})^2 + (\frac{X_{2,i}-\mu_2}{\sigma_2})^2]\} dX_{1,1}\cdots dX_{2,n} \\
&= \frac{(1-\rho^2)^{(n+4)/2}}{(2\pi)^n}\int_{R^{2n}} (E_\psi\tilde{\beta}_\psi\tilde{\alpha}_\psi)(\sum_{j=0}^\infty \frac{\rho^j E_\psi \tilde{\beta}_\psi^{j+1}}{j!}) e^{-\sum_{i=1}^n (Y_{1,i}^2+Y_{2,i}^2)/2} dY_{1,1}\cdots dY_{2,n} \\
&= \left\{\sum_{j=0}^\infty (-1)^j \rho^{2j} [\prod_{l=1}^j (\frac{n+6-2l}{2l})]\right\} E^*[(E_\psi\tilde{\beta}_\psi\tilde{\alpha}_\psi)(\sum_{j=0}^\infty \frac{\rho^j E_\psi\tilde{\beta}_\psi^{j+1}}{j!})]. \qquad (15)
\end{aligned}
$$

We further observe that

$$\frac{\partial}{\partial\rho} E_\psi \tilde{\beta}_\psi^l \tilde{\alpha}_\psi = E_\psi \tilde{\beta}_\psi^{l+1} \tilde{\alpha}_\psi - (E_\psi \tilde{\beta}_\psi^l \tilde{\alpha}_\psi)(E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi), \qquad l = 1, 2, \ldots. \tag{16}$$

Since $\tilde{\alpha}_\psi|_{\rho=0} = 1$ for all $\psi \in \mathcal{S}_n$, it follows from (16) that

$$\begin{aligned}
E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi|_{\rho=0} &= E_\psi \tilde{\beta}_\psi, \\
\frac{\partial}{\partial\rho} E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi|_{\rho=0} &= (E_\psi \tilde{\beta}_\psi^2) - (E_\psi \tilde{\beta}_\psi)^2, \\
\frac{\partial^2}{\partial\rho^2} E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi|_{\rho=0} &= (E_\psi \tilde{\beta}_\psi^3) - 3(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi) + 2(E_\psi \tilde{\beta}_\psi)^3, \\
\frac{\partial^3}{\partial\rho^3} E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi|_{\rho=0} &= (E_\psi \tilde{\beta}_\psi^4) - 4(E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi) - 3(E_\psi \tilde{\beta}_\psi^2)^2 \\
&\quad +12(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi)^2 - 6(E_\psi \tilde{\beta}_\psi)^4, \\
\frac{\partial^4}{\partial\rho^4} E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi|_{\rho=0} &= (E_\psi \tilde{\beta}_\psi^5) - 5(E_\psi \tilde{\beta}_\psi^4)(E_\psi \tilde{\beta}_\psi) - 10(E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi^2) \\
&\quad +20(E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi)^2 + 30(E_\psi \tilde{\beta}_\psi^2)^2(E_\psi \tilde{\beta}_\psi) \\
&\quad -60(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi)^3 + 24(E_\psi \tilde{\beta}_\psi)^5.
\end{aligned}$$

Now, by formally expanding $E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi$ as a power series about $\rho = 0$, we have

$$\begin{aligned}
E_\psi \tilde{\beta}_\psi \tilde{\alpha}_\psi &= (E_\psi \tilde{\beta}_\psi) + \rho[(E_\psi \tilde{\beta}_\psi^2) - (E_\psi \tilde{\beta}_\psi)^2] \\
&\quad + \frac{\rho^2}{2!}[(E_\psi \tilde{\beta}_\psi^3) - 3(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi) + 2(E_\psi \tilde{\beta}_\psi)^3] \\
&\quad + \frac{\rho^3}{3!}[(E_\psi \tilde{\beta}_\psi^4) - 4(E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi) - 3(E_\psi \tilde{\beta}_\psi^2)^2 \\
&\qquad +12(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi)^2 - 6(E_\psi \tilde{\beta}_\psi)^4] \\
&\quad + \frac{\rho^4}{4!}[(E_\psi \tilde{\beta}_\psi^5) - 5(E_\psi \tilde{\beta}_\psi^4)(E_\psi \tilde{\beta}_\psi) - 10(E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi^2) \\
&\qquad +20(E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi)^2 + 30(E_\psi \tilde{\beta}_\psi^2)^2(E_\psi \tilde{\beta}_\psi) \\
&\qquad -60(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi)^3 + 24(E_\psi \tilde{\beta}_\psi)^5] + O(\rho^6).
\end{aligned} \tag{17}$$

From (15) and (17), we obtain

$$\begin{aligned}
&E[(\sum_{\psi \in \mathcal{S}_n} \beta_\psi \alpha_\psi)^2] \\
&= E^*\{(E_\psi \tilde{\beta}_\psi)^2 + \rho[2(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi) - (E_\psi \tilde{\beta}_\psi)^3] \\
&\quad +\rho^2[(E_\psi \tilde{\beta}_\psi^2)^2 + (E_\psi \tilde{\beta}_\psi^3)(E_\psi \tilde{\beta}_\psi) - \frac{5}{2}(E_\psi \tilde{\beta}_\psi^2)(E_\psi \tilde{\beta}_\psi)^2 + (E_\psi \tilde{\beta}_\psi)^4 - \frac{(n+4)}{2}(E_\psi \tilde{\beta}_\psi)^2]
\end{aligned}$$

$$
\begin{aligned}
&+\rho^3[\frac{1}{3}(E_\psi\tilde{\beta}_\psi^4)(E_\psi\tilde{\beta}_\psi)+(E_\psi\tilde{\beta}_\psi^3)(E_\psi\tilde{\beta}_\psi^2)-\frac{7}{6}(E_\psi\tilde{\beta}_\psi^3)(E_\psi\tilde{\beta}_\psi)^2-2(E_\psi\tilde{\beta}_\psi^2)^2(E_\psi\tilde{\beta}_\psi)\\
&\quad+3(E_\psi\tilde{\beta}_\psi^2)(E_\psi\tilde{\beta}_\psi)^3-(E_\psi\tilde{\beta}_\psi)^5-(n+4)(E_\psi\tilde{\beta}_\psi^2)(E_\psi\tilde{\beta}_\psi)+\frac{(n+4)}{2}(E_\psi\tilde{\beta}_\psi)^3]\\
&+\rho^4[\frac{1}{3}(E_\psi\tilde{\beta}_\psi^4)(E_\psi\tilde{\beta}_\psi^2)-\frac{1}{2}(E_\psi\tilde{\beta}_\psi^2)^3-\frac{7}{2}(E_\psi\tilde{\beta}_\psi^2)(E_\psi\tilde{\beta}_\psi)^4+\frac{13}{4}(E_\psi\tilde{\beta}_\psi^2)^2(E_\psi\tilde{\beta}_\psi)^2\\
&\quad+\frac{4}{3}(E_\psi\tilde{\beta}_\psi^3)(E_\psi\tilde{\beta}_\psi)^3-\frac{11}{6}(E_\psi\tilde{\beta}_\psi^3)(E_\psi\tilde{\beta}_\psi^2)(E_\psi\tilde{\beta}_\psi)+\frac{1}{4}(E_\psi\tilde{\beta}_\psi^3)^2\\
&\quad-\frac{3}{8}(E_\psi\tilde{\beta}_\psi^4)(E_\psi\tilde{\beta}_\psi)^2+\frac{1}{12}(E_\psi\tilde{\beta}_\psi^5)(E_\psi\tilde{\beta}_\psi)+(E_\psi\tilde{\beta}_\psi)^6-\frac{(n+4)}{2}(E_\psi\tilde{\beta}_\psi^2)^2\\
&\quad+\frac{5(n+4)}{4}(E_\psi\tilde{\beta}_\psi^2)(E_\psi\tilde{\beta}_\psi)^2-\frac{(n+4)}{2}(E_\psi\tilde{\beta}_\psi^3)(E_\psi\tilde{\beta}_\psi)\\
&\quad-\frac{(n+4)}{2}(E_\psi\tilde{\beta}_\psi)^4+\frac{(n+4)(n+2)}{8}(E_\psi\tilde{\beta}_\psi)^2]\}+O(\rho^6) \qquad (18)\\
&=1+(1+4n+n^2)\rho^2-(\frac{32}{n-1}+11+4n)\rho^4+O(\rho^6).
\end{aligned}
$$

The last equality uses the mathematical computation software system *Mathematica* (Wolfram (1996)) to expand each term in the right hand side of (18) as a power series in ρ. These calculations are extremely tedious and we refer the reader to Chan and Loh (2000a) for the details. This completes the proof of Proposition 1.

References

Andrews, G. E. (1976). *The Theory of Partitions.* Cambridge University Press, New York.

Chan, H. P. and Loh, W. L. (2000a). A file linkage problem of DeGroot and Goel revisited. (http://www.stat.nus.edu.sg/~wloh/brokensample.ps)

Chan, H. P. and Loh, W. L. (2000b). A file linkage problem of DeGroot and Goel revisited, II. (http://www.stat.nus.edu.sg/~wloh/brokensampleII.ps)

Copas, J. B. and Hilton, F. J. (1990). Record linkage: statistical models for matching computer records (with comments). *J. Roy. Statist. Soc. Ser. A* **153**, 287-320.

DeGroot, M. H. and Goel, P. K. (1980). Estimation of the correlation coefficient from a broken random sample. *Ann. Statist.* **8**, 264-278.

Fisher, R. A. (1929). Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc. Ser. 2* **30**, 199-238.

Goel, P. K. and Ramalingam, T. (1989). *The Matching Methodology: Some Statistical Properties.* Springer, New York.

Lehmann, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

Liu, J. S. (1994). Fraction of missing information and convergence rate of data augmentation. In *Computationally Intensive Statistical Methods: Proc. 26th Symp. Interface* (Edited by J. Sall and A. Lehman), 490-497.

McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman and Hall, New York.

Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 1. 6th edition. Edward Arnold, London.

Vaughan, R. J. and Venables, W. N. (1972). Permanent expressions for order statistic densities. *J. Roy. Statist. Soc. Ser. B* **34**, 308-310.

Valiant, L. G. (1979). The complexity of computing the permanent. *Theor. Comp. Sci.* **8**, 189-201.

Wishart, J. (1929). The correlation between product moments of any order in samples from a normal distribution. *Proc. Roy. Soc. Edin.* **49**, 78-90.

Wolfram, S. (1996). *The Mathematica Book*. 3rd edition. Cambridge University Press, New York.

Wu, Y. (1995). Random shuffling: a new approach to matching problem. *ASA Proc. Statist. Comput. Sect.*, 69-74.

Department of Statistics and Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117543.

E-mail: stachp@nus.edu.sg

Department of Statistics and Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117543.

E-mail: wloh@stat.nus.edu.sg

(Received March 2000; accepted March 2001)