LIMIT THEOREMS FOR FUNCTIONS OF MARGINAL QUANTILES AND ITS APPLICATION

SU YUE

(B.Sc.(Hons.), Northeast Normal University)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2010

Acknowledgements

I would like to thank my advisor and friend, Professor Bai Zhidong and Associate Professor Choi Kwok Pui.

My thanks also goes out to the Department of Statistics and Applied Probability.

On the thesis edition technical aspects, I would like to thank Mr.Deng Niantao ,appreciate for his warmhearted assistance.

Su Yue

March 9 2010

Contents

Summary

A broken sample problem has been studied by statistician, which is a random sample observed for a tow-component random variable (X , Y), however, the link (or correspondences information) between the X-components and the Y-components are broken (or even missing). A method for re-pairing the broken sample is proposed as well as making statistical inference.

Meanwhile, multivariate data ordering schemes has a successful application in the color image processing. So in this paper, we extended the broken sample formulation to study the limit theorem for functions of marginal quantiles. We mainly studied how to explore multivariate distribution using the joint distribution of marginal quantiles. Limit theory for the mean of functions of order statistics is presented. The result include multivariate central theorem and strong law of large numbers. A result similar to Bahadurs representation of quantiles, is established for the mean of a function of the marginal quantiles. In particular, it shown that

$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(X_{n:i}^{(1)},...,X_{n:i}^{(d)}\right)-\bar{\gamma}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_{n:i}+O_p(1)$

as n tends to infinity, where is a constant, and for each n, are i.i.d. random variables. This leads to the central limit theorem. A weak convergence to a

Gaussian process using equicontinuity of functions is indicated. The conditions ,under which these results are established. Simulation results of the Marshall-Olkin bivariate exponential distribution and the Farlie-Gumbel-Morgenstern family of copulas are demonstrated to show our two main theoretical results satisfy in many examples that include several commonly occurring situations.

List of Figures

Chapter 1

Multivariate Data Ordering Schemes

1.1 The ordering of Multivariate data

A multivariate signal is a signal where each sample has multiple components.It is also called a vector valued,multichannel or multispectral signal.Color images are typical examples of multivariate signals.A color image represented by the three primaries in the RGB coordinate system is a two-dimentional three-variate(three-channel) signal. Let X denote a p-dimensional random variable,e.g. a p-dimensional vector of random variables $X = [X_1, X_2, ..., X_p]^T$. The probability density function(pdf)and the cumulative density function (cdf) of this p-dimensional random variable will be denoted by $f(X)$and $F(X)$ respectively. Now let $x_1, x_2, ..., x_n$be n random samples from the multivariate X. Each one of the x_i are p-dimensional vectors of observations $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]^T$.The goal is to arrange the n values $(x_1, x_2, ..., x_n)$ in some sort of order.The notion of data ordering,which is natural in the one dimensional case, does not extend in a straightforward way to multivariate data,since there is no unambiguous ,universally acceptable way to order n multivariate samples. Although no such unambiguous form of ordering exists, there are several ways to order the data,the so called sub-ordering principles.

Since ,in effect,ranking procedures isolate outliers by properly weighting each ranked multivariate sample,these outliers by properly weighting each ranked multivariate sample,these outlier can be discorded. The sub-ordering principles are useful in detecting outliers in a multivariate sample set.Univariate data analysis is sufficient to detect any outliers in the data in terms of their extreme value relative to an assumed basic model and then employ a robust accommodation method of inference. For multivariate data however,an additional step in the process is required,namely the adaption of the appropriate sub-ordering principle as the basis for expressing extremeness of observations. The sub-ordering principles are categorized in four types:

1.marginal ordering or M-ordering

2.conditional ordering or C-ordering

3.partial ordering or P-ordering

4.reduced(aggregated) ordering of R-ordering.

Marginal Ordering

In the marginal ordering (M-ordering) scheme,the multivariate samples are ordered along each of the p-dimensions independently yielding:

$x_{1(1)} \leq x_{1(2)} \leq \ldots \leq x_{1(n)}$

$x_{2(1)} \leq x_{2(2)} \leq \ldots \leq x_{2(n)}$

$\ldots\ldots\ldots\ldots\ldots$

$x_{p(1)} \leq x_{p(2)} \leq \ldots \leq x_{p(n)}$

According to the M-ordering principle,ordering is performed in each channel of the multichannel signal independently. The vector $x_1 = [x_1(1), x_2(1), \ldots, x_p(1)]^T$ consists of the minimal elements in each dimension and the vector,

$x_n = [x_1(n), x_2(n), \ldots, x_p(n)]^T$

consists of the maximal elements in each dimension. The marginal median is defined as $x_{v+1} = [x_1(v), x_2(v), \ldots, x_p(v)]^T$ for $n = 2v+1$,which may not correspond to any of the original multivariable samples. In contrast, in the scalar case there is a one-to-one correspondence between the original samples x_i and the order statistics x_i.

Conditional Ordering

In conditional ordering(C-ordering) the multivariate samples are ordered conditional on one of the marginal sets of observations. Thus,one of the marginal components is ranked and the other components of each vector are listed according to the position of their ranked component. Assuming that the first dimension is ranked,the ordered samples would be represented as follows:

$$x_1(1) \leq x_1(2) \leq \ldots \leq x_1(n)$$

$$x_{2[1]} \leq x_2(2) \leq \ldots \leq x_2(n)$$

$$\ldots\ldots\ldots$$

$$x_p(1) \leq x_p(2) \leq \ldots \leq x_p(n)$$

where $x_1(i), i = 1, 2, \ldots, n$ are the marginal order statistics of the first dimension ,and $x_j[i], j = 2, 3, \ldots, p, i = 1, 2, \ldots, n$are the quasi-ordered samples in dimensions $j = 2, 3, \ldots, p$, conditional on the marginal ordering of the first dimension. These components are not ordered,they are simply listed according to the ranked components.In the two dimensional case(p=2) the statistics $x_2(i), i = 1, 2, \ldots, n$ are called concomitants of the order statistics of x_1. The advantage of this ordering scheme is its simplicity since only one scalar ordering is required to define the order statistics of the vector sample. The disadvantage of the C-ordering principle is that since only information in one channel is used for ordering, it is assumed that all or at least most of the improtant ordering information is associated with that dimension. Needless to say that if this assumption were not to hold,considerable loss of useful information may occur. As an example,the problem of ranking color signals in the YIQ color system may be considered. A conditional ordering scheme based on the luminance channel (Y) means that chrominace information stored in the I and Q channels would be ignored in ordering. Any advantages that could be gained in identifying outliers or extreme values based on color information would therefore be lost.

Partial Ordering,

In partial (P-ordering),subsets of data are grouped together forming minimum convex hulls. The first convex hull is formed such that the perimeter contains a minimum number of points and the resulting hull contains all other points in the given set. The points along this perimeter are denoted c-order group1.These points form the most extreme group.The perimeter points are then discarded and the process repeats.The new perimeter points are denoted c-order group 2 and then removed in order for the process to be continued. Although convex hull or elliptical peeling can be used for outlier isolation,this method provides no ordering within the groups and thus it is not easily expressed in analytical terms. In addition,the determination of the convex hull is conceptually and computationally difficult,especially with higher-dimension data.Thus,although trimming in terms of ellipsoids of minimum content rather than convex hull has been proposed,P-ordering is rather infeasible for implementation in color image processing.

Reduced Ordering

In reduced (aggregating) or R-ordering,each multivariate observation x_i is reduced to signal,scalar value by means of some combination of the component sample values.The resulting scalar values are then amenable to univariate ordering.Thus,the set $x_1, x_2, \ldots, x_n$ can be ordered in terms of the values $R_i = R(x_i), i = 1, 2, \ldots, n$. The vector x_i which yields the maximum value $R_{(n)}$ can be considered as an outlier,provided that its extremeness is obvious comparing to the assumed basic model. In contrast to M-ordering ,the aim of R-ordering is to effect some sort of overall ordering on the original multivariate samples,and by ordering in this way,the multivariate ranking is reduced to a simple ranking operation of a set of transformed values.The type of ordering cannot be interpreted in the same manner as the conventional scalar ordering as there are no absolute minimum or maximum vector samples.Given that multivariate ordering is based on a reduction functon R(.),points which diverge from the'center'in opposite directions may be in the same order ranks.Furthermore,by utilizing a reduction function as the mean to accomplish multivariate ordering,useful information may be lost.Since distance measures have a natural mechanism for identification of outliers,the reduction function most frequently employed in R-ordering is the generalized (Mahalanobis) distance:

$R(x, \overline{x}, \Gamma) = (x - \overline{x})^T \Gamma^1 (x - \bar{x})$

where $\bar{x}$ is a lacation parameter for the data set,or underlying distribution,in consideration and Γ is a dispersion parameter with Γ^{-1} used to apply a differential weighting to the components of the multivariate observation inversely related to the population variability.The parameters of the reduction function can be given arbitrary values,such as $\bar{x} = 0$ and $\Gamma = I$,or they can be assigned the true meanμ and dispersion $\sum$ settings. Depending on the state of knowledge about these values,their standard estimates:

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

and

$S = \frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})(x - \bar{x})^T$

can be used instead. Within the framework of the generalized distance,different reduction functions can be utilized in order to identify the contribution of an individual multivariate sample. A list of such functions include,among others,the following:

$q_i^2 = (x - \bar{x})^T(x - \bar{x})$

$t_i^2 = (x - \bar{x})^T S(x - \bar{x})$

$u_i^2 = \frac{(x-\bar{x})^T S(x-\bar{x})}{(x-\bar{x})^T(x-\bar{x})}$

$v_i^2 = \frac{(x-\bar{x})^T S^{-1}(x-\bar{x})}{(x-\bar{x})(x-\bar{x})}$

$d_i^2 = (x - \bar{x})^T S^{-1}(x - \bar{x})$

$d_k^2 = (x - x_k)^T S^{-1}(x - x_k)$

with $i < k = 1, 2, \ldots, n$.Each one of the these functions identifies the contribution of the individual multivariate sample to specific effects as follows:

1.q_i^2 isolates data which excessively inflate the overall scale.

2.t_i^2 determines which data has the greatest influence on the orientation and scale of the first few principle components.

3.u_i^2 emphasizes more the orientation and less the scale of the principle components.

4.v_i^2 measures the relative contribution on the orientation of the last few principle components.

5.d_i^2 uncovers the data points which lie far away from the general scatter of points.

6.d_k^2 has the same objective as d_i^2 but provides far more detail of interobject separation.

The following comments should be made regarding the reduction functions discussed in this section:

1.If outliers are present in the data then $\bar{x}$ and $\sum$ are not the best estimates of the location and dispersion for the data,since they will be affected by the outliers. In the face of outliers,robust estimators of both the mean value and the covariance matrix should be utilized.A robust estimation of the matrix S is important because outliers inflate the sample covariance and thus may mask each other making outlier detection even in the presence of only a few outliers.Various design options can be considered.Among them the utilization of the marginal midian(median evaluated using M-ordering) as a robust estimate of the location.However,care must be taken since the marginal median of n multivariate samples is not necessarily one of the input samples.Depending on the estimator of the location used in the ordering procedure the following schemes can be distinguished.

a)R-ordering about the mean(Mean R-ordering)

Given a set of n multivariate samples $x_i, i = 1, 2, \ldots, n$in a processing window and $\bar{x}$ the mean of the multivariate ,the mean R-ordering is defined as:

$(x_{(1)}, x_{(2)}, \ldots, x_{(n)} : \bar{x})$

where$(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ is the ordering defined by:

$d_i^2 = (x - \bar{x})^T (x - \bar{x})$ and $(d_{(1)}^2 \leq d_{(2)}^2 \leq d_{(n)}^2)$.

b) R-ordering about the marginal median(Median R-ordering)

Given a set of n multivariate samples $x_i, i = 1, 2 \ldots, n$ in a processing window and x_m the marginal median of the multivariates,the median R-ordering is defined as:

$(x_{(1)}, x_{(2)}, \ldots, x_{(n)} : x_m)$

where $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ is the ordering defined by:

$d_i^2 = (x - x_m)^T (x - x_m)$ and $(d_{(1)}^2 \leq d_{(2)}^2 \leq d_{)(n)}^2)$.

c) R-ordering about the center sample (Center R-ordering) G

Given a set of n multivariate samples $x_i, i = 1, 2, \ldots, n$ in a processing window and $x_{\bar{n}}$ the sample at the window center $\bar{n}$, the center R-ordering is defined as:

$(x_{(1)}, x_{(2)}, \ldots, x_{(n)} : x_{\bar{n}})$

where $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ is the ordering defined by:

$d_i^2 = (x - x_{\bar{n}})^T (x - x_{\bar{n}})$ and $(d_{(1)}^2 \leq d_{(2)}^2 \leq \ldots \leq d_{(n)}^2)$.Thus ,$x_{(1)} = x_{\bar{n}}$.

2.Statistic measures,such as d_i^2 and d_k^2 are invariant under non singular transformation of the data.

3.Statistics which measure the influence on the first few principle components,such as t_i^2,u_i^2,d_i^2 and d_k^2 are useful in detecting those outliers which inflate the variance,covariance or correlation in the data.Statistics measures ,such as v_i^2 will detect those outliers that add insignificant dimensions and/or singularities to the data.

Statistical descriptions of the descriptive measures listed above can be used to assist in the design and analysis of color image processing algorithms. As an example,the statistical description of the d_i^2 descriptor will be presented.Given the multivariate data set $(x_1, x_2, \ldots, x_n)$ and the population mean $\bar{x}$,interest lies in determining the distribution for the distances d_i^2 or equivalently for $D_i = (d_i^2)^{1/2}$.Let the probability density function of D for the input be denoted as f_D and the pdf for the i^{th} ranked distance be f_{D_i},If he multivariate data samples are independent and identically distributed then D will be also independent and identically distributed.Based on this assumption f_{D_i} can be evaluated in terms of f_D as follows

$f_{D_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F_D^{i-1}(x)[1 - F_D(x)]^{n-i} f_D(x)$

with $F_D(x)$ the cumulative distribution (cdf) for the distance D. As an example,assume that the multivariate samples x belong to a multivariate elliptical distribution with parameter μ_x,$\sum_x$ and of the form:

$f(x) = K_p|\sum_x|^{-1/2} h((x - \mu_x)^T \sum^{-1}(x - \mu_x))$

for some function h(.),where K_p is a normalizing constant and $\sum_x$ is positive definite.This class of distributions includes the multivariate Guassian distribution and all other densities whose contours of equal probability have an elliptical shape.if a distribution such as the multivariate Gaussian belonging to this class exists, then all its marginal distributions and its conditional distribution also belong to this class.

For the special case of the simple Euclidean distance $d_i = (x - \bar{x})^T (x - \bar{x})^{1/2} f_{D_{(x)}}$ has the general form of :

$f_{D_{(x)}} = \frac{2K_p \pi^{p/2}}{\Gamma(p/2)} x^{p-1} h(x^2)$

where $\Gamma(.)$ is the gamma function and $x \geq 0$.If the elliptical distribution assumed initially for the multivariate x_i samples is considered to be multivariate Gaussian with mean value μ_x and covariance $\sum_x = \sigma^2 I_p$,then the normalizing constant is $K_p = (2\pi\sigma^2)^{1/2}$and the $h(x^2) = exp(\frac{-x^2}{2\sigma^2})$,and thus $f_{D_{(x)}}$ takes the form of the Rayleigh distribution:

$$f_{D_{(x)}} = \frac{x^{p-1}}{\sigma^p 2^{\frac{p-2}{2}} \Gamma(\frac{p}{2})} exp(\frac{-x^2}{2\sigma^2})$$

Based on this distribution the k^{th} moment of D is given as:

$$E[D^k] = (2\sigma)^{\frac{k}{2}} \frac{\Gamma(\frac{p+k}{2})}{\Gamma(\frac{p}{2})}$$

with $k \geq 0$.It can easily be seen from the above equation that the expected value of the distance D will increase monotonically as a function of the parameter σ in the assumed multivariate Gaussian distribution.

To complete the analysis ,the cumulative distribution function F_D is needed. Although there is no closed form expression for the cdf of a Rayleigh random variable,for the special case where p is an even number, the requested cdf can be expressed as:

$$F_D(x) = 1 - exp(\frac{-x^2}{2\sigma^2}) \sum_{k=0}^{(\frac{p}{2}-1)} (\frac{1}{k!})(\frac{x^2}{2\sigma^2})^k$$

Using this expression the following pdf for the distance D_i can be obtained:

$$f_{D_{(i)}}(x) = Cx^{p-1} exp(\frac{-x^2}{2\sigma^2}) F_D(x)^{(i-1)}(1 - F_D(x))^{n-i}$$

where $C = \frac{(n!)\sigma^p \Gamma(\frac{p}{2})}{(i-1)!(n-i)!2^{\frac{p-2}{2}}}$ is a normalization constant.

In summary,R-ordering is particularly useful in the task of multivariate outlier detection,since the reduction function can reliably identify outliers in multivariate data samples.Also,unlike M-ordering,it treats the data as vectors rather than breaking them up into scalar components.Furthermore,it gives all the components equal weight of importance,unlike C-ordering.Finally,R-ordering is superior to P-ordering in its simplicity and its ease of implementation ,making it the sub ordering principle of choice for multivariate data analysis.

1.2 Color Image Processing and Applications

The probability distribution of p-variate marginal order statistics can be used to assist in the design and analysis of color image processing algorithms.Thus,the cumulative distribution function (cdf) and the probability distribution function (pdf) of marginal order statistics is described.In particular,the analysis is focused in the derivation of three-variate(three-dimensional) marginal order statistics,which is of interest since three-dimensional vectors are used to describe the color signals in the different color systems,such as the RGB.

The three-dimensional space is divided into eight subspaces by a point (x_1, x_2, x_3).The requested cdf is given as:

$F_{r1,r2,r3}(x_1, x_2, x_3) =$

$\sum_{i_1=r_1}^{n} \sum_{i_2=r_2}^{n} \sum_{i_3=r_3}^{n} P[i_1 of X_{1i} \leq x_1, i_2 of X_{2i} \leq x_2, i_3 of X_{3i} \leq x_3]$

of the marginal order statistic $X_1(r_1), X_2(r_2), X_3(r_3)$ when n three-variate samples are available.

Let $n_i, i = 0, 1, \ldots, 7$ denote the number of data points belonging to each of the eight subspace.In this case:

$P[i_1; X_{1i} \leq x_1, i_2; X_{2i} \leq x_2, i_3; X_{3i} \leq x_3] =$

$\sum_{n_0} \cdots \sum_{n_7} \frac{n!}{\prod_{i=0}^{7}} F_i^{n_i}(x_1, x_2, x_3)$

Given that the total number of points is $\sum_{i=0}^{7} = n$,the following conditions hold for the number of data points lying in the different subspaces:

$n_0 + n_2 + n_4 + n_6 = i_1$

$n_0 + n_1 + n_4 + n_5 = i_2$

$n_0 + n_1 + n_2 + n_3 = i_3$

Thus,the cdf for the three-variate case is given by:

$F_{r1,r2,r3}(x_1, x_2, x_3)$

$= \sum_{i_1=r_1}^{n} \sum_{i_2=r_2}^{n} \sum_{i_3=r_3}^{n} \sum_{n_0} \cdots \sum_{n_{2^3-1}} \frac{n!}{\prod_{i=0}^{2^3-1}} \prod_{i=0}^{2^3-1} F_i^{n_i}(x_1, x_2, x_3)$

which is subject to the constraints of the following conditions:

$n_0 + n_2 + n_4 + n_6 = i_i$

$n_0 + n_1 + n_4 + n_5 = i_2$

$n_0 + n_1 + n_2 + n_3 = i_3$

The probability density function is given by:

$f_{(r_1,r_2,r_3)}(x_1, x_2, x_3) = \frac{\partial^3 F_{r_1,r_2,r_3}(x_1,x_2,x_3)}{\partial x_1 \partial x_2 \partial x_3}$

The joint cdf for the three-variate case can be calculated as follows:

$F_{r_1,r_2,r_3,s_1,s_2,s_3}(x_1, x_2, x_3, t_1, t_2, t_3) = \sum_{j_1=s_1}^{n} \sum_{i_1=r_1}^{j_1} \cdots \sum_{j_3=s_3}^{n} \sum_{i_3=r_3}^{j_3} \phi(r)$

with

$\phi(r) = P[i_1 of X_{1i} \leq x_1, j_1 of X_{1i} \leq t_1, i_2 of X_{2i} \leq x_2, j_2 of X_{2i} \leq t_2, i_3 of X_{3i} \leq x_3, j_3 of X_{3i} \leq t_3]$

for $X - i < t_i$ and $r_i < s_i, i = 1, 2, 3$.The two points (x_1, x_2, x_3) and (t_1, t_2, t_3) divide the three-dimensional space into 3^3 subspace. If $n_i, F_i, i = 0, 1, \ldots, (3^3 - 1)$ denote the number of data points and the probability masses in each subspace then it can be proven that:

$\phi(r) = \sum_{n_0} \cdots \sum_{(n_3^3)-1} \frac{n!}{\prod_{i=0}^{(n_3^3)-1} n_i!} \prod_{i=0}^{(n_3^3)-1} F_i^{n_i}(x_1, x_2, x_3)$

under the constraints:

$\sum_{i=0}^{3^3-1} n_i = n$

$\sum_{I_0=0} n_i = i_1$

$\sum_{I_1=0} n_i = i_2$

$\sum_{I_2=0} n_i = i_3$

$\sum_{I_0=0,1} n_i = j_1$

$\sum_{I_1=0,1} n_i = j_2$

$\sum_{I_2=0,1} n_i = j_3$

where $i = (I_2, I_1, I_0)$ is an arithmetic representation of number i with base 3. Through above equation ,a numerically tractable way to calculate the joint cdf for the three-variate order statistics is possible.

Chapter 2

Two main theorem prove

2.1 Introduction

Let$\{(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}), i = 1, 2, \ldots\}$ be a sequence of random vectors such that for each j $(1 \leq j \leq d)$,$\{X_1^{(j)}, X_2^{(j)}, \ldots, \}$ forms a sequence of independent and identically distribution(i.i.d.) random variables with distribution function F_j. Let $X_{n:i}^{(j)}$denote the ith order statistic ($\frac{i}{n}$th quantile) of $\{X_1^{(j)}, X_2^{(j)}, \ldots X_n^{(j)}\}$. We study the asymptotic behavior of the mean of a function of marginal sample quantiles: $\frac{1}{n} \sum_{i=1}^{n} \phi \left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)} \right)$ as $n \to \infty$,where ϕ: $R^d \longrightarrow R$ satisfies some mild conditions.

We introduce condition on ϕ.

(C1) The function $\psi(u_1, \ldots, u_d)$ is continuous at $u_1 = u, \ldots, u_d = u, 0 < u < 1$. that is,$\psi$ is continuous at each point on the diagonal of $(0, 1)^d$.

(C2) There exist K and $c_0 > 0$ such that for $(x_1, \ldots, x_d) \in (0, c_0)^d \bigcup (1 - c_0, 1)^d$, $\mid \psi(x_1, \ldots, x_d) \mid \leq K \left(1 + \sum_{j=1}^{d} \mid \gamma(x_j) \mid \right)$.

(C3) Let $u_{n:i} = \frac{i}{n+1}$. For $1 \leq j, k \leq d$,

$\frac{1}{n}\sum_{i=1}^{n}\left[u_{n:i}(1-u_{n:i})\right]^{\frac{3}{2}}\left[\psi_j(u_{n:i})\right]^2 \longrightarrow \int_0^1 \left[x(1-x)\right]^{\frac{3}{2}}\left[\psi_j(x)\right]^2 dx$

and

$\frac{1}{n}\sum_{i=1}^{n}\left[u_{n:i}(1-u_{n:i})\right]^{\frac{3}{2}} \mid \tilde{\psi}_{j,k}(u_{n:i}) \mid\longrightarrow \int_0^1 \left[x(1-x)\right]^{\frac{3}{2}} \mid \tilde{\psi}_{j,k}(x) \mid dx$

Condition (3) holds if the function, $[x(1-x)]^{\frac{3}{2}}\left[\psi_j(x)\right]^2 (1 \le j \le d)$, and$[x(1-x)]^{\frac{3}{2}} \mid \tilde{\psi}_{j,k}(x) \mid (1 \le j,k \le d)$ are Riemann integrable over $(0,1)$,and satisfy $K-pseudo$ convexity. A function g is said to be K-pseudo convex if $g(\lambda x + (1-\lambda)y) \le K\left[\lambda g(x) + (1-\lambda)g(y)\right]$.

C4 For all large m , there exist $K = K(m) \ge 1$ and $\delta > 0$such that

$\mid \psi(y) - \psi(x) - \langle y - x, \nabla\psi(x)\rangle \mid$
$\le K \sum_{j,k=1}^{d} \mid (y_j - x)(y_k - x) \mid (1+ \mid \psi_{j,k}(x) \mid)$

if $x = (x,\dots,x)$,$y = (y_1,\dots,y_d) \in (0,1)^d$, $\parallel (y-x) \parallel_{l_1} < \delta$,and for $1 \le j \le d$, $y_j(1-y_j) > x(1-x)/m$. Here $\parallel y \parallel_{l_1} :=\mid y_1 \mid + \dots + \mid y_d \mid$ denotes the l_1-norm of y and $\bigtriangledown\psi(x)$ the gradient of ψ.

Following two theorem is our main results.

Theorem1. Let $\left\{(X_i^{(1)}, X_i^{(2)},\dots,X_i^{(d)}), i = 1,2,\dots\right\}$ be a sequence of random vectors such that for each $j(1 \le j \le d)$, $\left\{(X_1^{(j)}, X_2^{(j)},\dots,\right\}$ forms a sequence of $i.i.d.$ random variables with continuous distribution function F_j. Suppose that ϕ satisfies the conditions $C(1)$ and $C(2)$,function$\gamma(x) := \psi(x,x,\dots,x)$,$0 < x < 1$,is Riemann integrable,

then we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi\left(X_{n:i}^{(1)},\dots,X_{n:i}^{(d)}\right) \longrightarrow \bar{\gamma} \qquad a.s.$$

as $n \to \infty$. Here $\bar{\gamma} = \int_0^1 \gamma(y)dy = E\phi(F_1^{-1}(U), F_2^{-1}(U),\dots,F_d^{-1}(U))$ and U is uniformly distributed over $(0,1)$.

Note that we need only independence of marginal random variables. The result does not depend on the joint distribution of $(X_1^{(1)}, \ldots, X_1^{(d)})$.

Theorem2. Let $X_i = (X_i^{(1)}, \ldots, X_i^{(d)})$ be i.i.d. random vectors. Let $F_j 1 \leq j \leq d)$ denote the marginal distribution of X_i^j which is assumed to be continuous, and $F_{j,k}$,$(1 \leq j, k \leq d)$ the marginal distribution of (X_i^j, X_i^k). If ϕ satisfies condition $(C1) - C(4)$,and that $\gamma(x) := \psi(x, x, \ldots, x) := \phi(F_1^{-1}(x), \ldots, F_d^{-1}(x)), 0 < x < 1$ is Riemann integrable,then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi\left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\right) - \sqrt{n}\bar{\gamma} = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} Z_{n,l} + o_P(1)$$

where, for $1 \leq l \leq n$,$Z_{n,l} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} W_{j,l}\left(\frac{i}{n}\right) \psi_j \left(\frac{i}{n+1}\right)$.
Here$W_{j,l}(x) = I(U_l^{(j)} \leq x) - x$
and $\bar{\gamma} = \int_0^1 \gamma(y)dy = E\phi(F_1^{-1}(U), F_2^{-1}(U), \ldots, F_d^{-1}(U))$.

Hence,by Lindeberg-Levy central limit theorem,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi\left(X_{n:i}^{(1)}, \ldots, X_{n:i}^{(d)}\right) - \sqrt{n}\bar{\gamma} \xrightarrow{dist.} N(0, \sigma^2)$$

as $n \to \infty$,where
$\sigma^2 = \lim_{x \to \infty} Var(Z_{n,1}) = 2 \sum_{j=1}^{d} \int_0^1 \int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy$
$+2 \sum_{1 \leq j < k \leq d} \int_0^1 \int_0^1 \left[G_{j,k}(x, y) - xy\right] \psi_j(x)\psi_k(y)dxdy$
where$G_{j,k}(x, y) = F_{j,k}\left(F_j^{-1}(x), F_k^{-1}(y)\right)$.

This theorem can be extended to m function $\phi_1, ..., \phi_m$ simultaneously using Cramer-Wold device as in the corollary below.Let $\psi_j(x; r)$denote the partial derivative of $\phi_r\left(F_1^{-1}(x_1), \ldots, F_d^{-1}(x_d)\right)$ with respect to x_j evaluated at $x_1 = \ldots = x_d = x$.

Corollary.

Let $\phi_1, \ldots, \phi_m$ satisfy condition (C1)-(C4).For $1 \leq r \leq m$,define $T_n(\phi_r) = \sum_{i=1}^{n} \phi_r\left(X_{n:i}^1, \ldots, X_{n:i}^d\right)$ and $\overline{\gamma}_r = E\phi_r\left(F_1^{-1}(U), F_2^{-1}(U), \ldots, F_d^{-1}(U)\right)$,then

$\frac{1}{\sqrt{n}}\left(T_n(\phi_1), \ldots, T_n(\phi_m)\right) - \sqrt{n}(\overline{\gamma_1}, \ldots, \overline{\gamma_m}) \xrightarrow{dist.} N(0, \sum), asn \to \infty,$

where $\sigma_{r,s}$, the $(r,s)th$ element of $\sum$,is given by

$$\sum_{j=1}^{d} \int_0^1 \int_0^1 x(1-y)[\psi_j(x;r)\psi_j(y;s) + \psi_j(x;s)\psi_j(y;r)]dxdy$$

$$+ \sum_{1 \leq j < k \leq d} \int_0^1 \int_0^1 [G_{j,k}(x,y) - xy][\psi_j(x;r)\psi_k(y;s) + \psi_j(x;s)\psi_k(y;r)]dxdy.$$

Now,we prove above mentioned corollary.

Proof.

Use Cramer-Wold device.In computing $\sigma_{r,s}$,we used

$2\sigma_{r,s} = lim_{n\to\infty}[Var(Z_{n,1,r} + Z_{n,1,s}) - Var(Z_{n,1,r}) - Var(Z_{n,1,s})]$

where

$Z_{n,1,r} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} W_{j,1}(i/n)\psi_j(i/(n+1);r).$

2.2 Proof of the two main theorem

Proof of theorem 1.

If we define a new random variable $U_i^{(j)}$ and let $U_i^{(j)} = F(X_i^{(j)})$, it is easy for us to get the distribution formation of our new defined random variable U_i^j as uniform distribution,caused

$$P(U_i^{(j)} \leq x) = P(X_i^{(j)} \leq F_j^{-1}(x)) = F_j(F_j^{-1}(x)) = x$$

So $U_1^{(j)}, U_2^{(j)}, ...$ forms a sequence of i.i.d. uniformly distributed random variables and , with probability 1 , $F_j^{-1}(U_i^{(j)}) = X_i^{(j)}$.

We let $U_{n:i}^{(j)}$ denote the ith order statistic of $U_1^j, U_2^j, \ldots, U_n^j$.We use $\mu_{n:i}$to denote the expectation of the ith order statistics.We can also get the explicit expectation formulation of the ith order statistic.

Firstly,we can get the probability denstity function of the ith order statistic of $U_{n:i}^{(j)}$

$P(U_{n:i}^{(j)} = x)$

$= n\binom{n-1}{i-1}P(U_1^{(j)} = x, U_2^{(j)} \leq x, \ldots, U_i^j \leq x, U_{i+1}^{(j)} \geq x, \ldots, U_n^{(j)} \geq x)$

$= \frac{n!}{(i-1)!(n-i)!}x^{i-1}(1-x)^{n-i}$

So we can take advantage of above density function to get the explicit expatation formulation through the definition of expectation,

$\int_0^1 x\frac{n!}{(i-1)!(n-i)!}x^{i-1}(1-x)^{n-i}dx$

$= \int_0^1 \frac{n!}{(i-1)!(n-i)!}x^i(1-x)^{n-i}dx$

$= \frac{n!}{(i-1)!(n-i)!}\int_0^1 x^{(i+1)-1}(1-x)^{(n-i+1)-1}dx$

$= \frac{n!}{(i-1)!(n-i)!}Be(i+1, n-i+1)$

$= \frac{n!}{(i-1)!(n-i)!}\frac{\Gamma(i+1)\Gamma(n-i+1)}{\Gamma(n+2)}$

$= \frac{n!}{(i-1)!(n-i)!}\frac{i!(n-i)!}{(n+1)!}$

$= \frac{n!i(i-1)!(n-i)!}{(i-1)!(n-i)!(n+1)n!}$

$= \frac{i}{n+1}$

So we write

$\mu_{n:i} = EU_{n:i}^{(j)} = \frac{i}{n+1}$

Caused $\psi(x_1, ..., x_d) = \phi(F_1^{-1}(x_1), ..., F_d^{-1}(x_d))$,and $\gamma(x) = \psi(x, ..., x)$. Choose any $\epsilon \in (0, c_0)$. Then almost surely,

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_{n:i}^1, ..., X_{n:i}^d) = \frac{1}{n}\sum_{i=1}^{n}\psi(U_{n:i}^1, ...U_{n:i}^d)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\gamma(\mu_n^i) + R_{n,1} + R_{n,2} + R_{n,3}$$

where

$$R_{n,1} = \frac{1}{n}\sum_{1\le i<\epsilon n}[\psi(U_{n:i}^{(1)},...,U_{n:i}^{d}) - \gamma(\mu_{n:i})]$$

$$R_{n,2} = \frac{1}{n}\sum_{\epsilon} n \le i \le (1-\epsilon)n[\psi(U_{n:i}^{1},...,U_{n:i}^{d}) - \gamma(\mu_{n:i})]$$

$$R_{n,3} = \frac{1}{n}\sum_{(1-\epsilon)n<i\le n}[\psi(U_{n:i}^{(1)},...,U_{n:i}^{(d)}) - \gamma(\mu_{n:i})]$$

$\frac{1}{n}\sum_{i=1}^{n}\gamma(\mu_{n:i})$ is a Riemann sum of γ over $(0,1)$ and hence it goes to $\int_0^1 \gamma(y)dy = E\phi((F_1^{-1}(U),...,F_d^{-1}(U))$by Riemann integrability of γ. Thus it remains to show that $R_{n:i} \longrightarrow 0$a.s. as $n \to \infty$ for $i = 1, 2and3$.

For $1 \le j \le d$, by Glivenko-Cantelli Lemma, $sup_{x\in(0,1)} \mid \hat{F_{n;j}}(x) - x \mid \xrightarrow{a.s.} 0$ as $n \to \infty$ where $\hat{F_{n:j}}$is the empirical distribution function of $\{U_i^{(j)} : 1 \le i \le n\}$ .For $1 \le i \le n$, $1 \le j \le d$,

$|U_{n:i}^{(j)} - \mu_{n:i}| \le |U_{n:i}^{(j)} - \frac{i}{n}| + \frac{1}{n} = |U_{n:i}^{(j)} - \hat{F}_{n;j}(U_{n:i}^{(j)})| + \frac{1}{n} \le \frac{1}{n} + sup_{x\in(0,1)}|x - \hat{F}_{n;j}(x)|$

So we have

$\delta_n := max\left(|U_{n:i}^{(j)} - \mu_{n:i}|, 1 \le i \le n, 1 \le j \le d\right) \xrightarrow{a.s.} 0.$

Caused for any $c \in (0, \frac{1}{2}), lim_{\delta\to 0}sup|\psi(x_1,\dots,x_d) - \gamma(y)| = 0$

and the supremum is taken over all $c \le x_1, x_2, \dots, x_d, y < 1-c$ such that $|x_i - y| < \delta$. And meanwhile $U_{n:i}^{j} \in (\mu_{n:i} - \delta_n, \mu_{n:i} + \delta_n)$for $1 \le j \le d$, and for each i such that $\epsilon \le \frac{i}{n} \le 1 - \epsilon$,we have

$|\psi(U_{n:i}^{(1)},\dots,U_{n:i}^{d}) - \gamma(\mu_{n:i})| \le \omega(\epsilon,\delta_n)$provided$\delta_n < \frac{\epsilon}{2}$.

So if $\delta_n < \frac{\varepsilon}{2}$,

$|R_{n,2}| \le \frac{1}{n}\sum_{\epsilon n\le i\le(1-\epsilon)n}|\psi(U_{n:i}^{(1)},\dots,U_{n:i}^{(d)}) - \gamma(\mu_{n:i})| \le \omega(\epsilon,\delta_n) \xrightarrow{a.s.} 0.$

By C(2),we have

$|R_{n,1}| \le K\sum_{j=1}^{d} R_{n,1,j} + \frac{1}{n}\sum_{1\le i<\epsilon n}|\gamma(\mu_{n:i})| + K\epsilon$

where

$R_{n,1,j} = \frac{1}{n}\sum_{1\leq i<\epsilon n}|\gamma(U^j_{n:i})|, 1 \leq j \leq d.$

If $U^{(j)}_{n:[\epsilon n]+1} \leq 2\epsilon$,then

$R_{n,1,j} \leq \frac{1}{n}\sum\limits_{1\leq i\leq n}|\gamma(U^{(j)}_i)|I(U^{(j)}_i \leq 2\epsilon).$

Caused with probability 1,$U^j_{n:[\epsilon n]+1} \leq 2\epsilon$ for all large n, and the right hand side of the above inequality goes to $\int_0^{2\epsilon}|\gamma(y)|dy a.s. as n \to \infty$. Hence

$limsup_{n\to\infty}|R_{n,1}| \leq (Kd+1)(\epsilon + \int_0^{2\epsilon}|\gamma(y)|dy)$ a.s.

As $|\gamma|$ is integrable,if we let ϵ tend to zero,we conclude that $R_{n,1} \xrightarrow{a.s.} 0$.Similar argument will show that $R_{n,3} \xrightarrow{a.s.} 0$ as $n \to \infty$.This completes the proof of theorem 1.

Proof of theorem 2.

As in the proof of Theorem1,we introduce $U^{(j)}_i = F_j(X^{(j)}_i)$ for $1 \leq i \leq n$,$1 \leq j \leq d$. It follows that $\left(U^{(1)}_i, \ldots, U^{(d)}_i\right)$, $1 \leq i \leq n$ are i.i.d. random vectors. For $1 \leq j,k \leq d$,define $G_{j,k}(x,y) = F_{j,k}\left(F^{-1}_j(x), F^{-1}_k(y)\right)$,and so $G_{j,k}$ is the joint distribution of $\left(U^{(j)}_1, U^{(k)}_1\right)$.In particular,$G_{j,j}(x,y) = min\{x,y\}, for 1 \leq j \leq d.$

Firstly,we start with some preliminary results in the following 4 lemmas.

Lemma1.

Let $U_{n:1} \leq \ldots \leq U_{n:n}$ be the order statistics of n independent random variables uniformly distributed over $(0,1)$. Then, for $1 \leq i \leq n$,

$var(U_{n:i}) = \frac{\mu_{n:i}(1-\mu_{n:i})}{(n+2)} \leq \frac{1}{n}$

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 \frac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i}dx \\ &= \frac{n!}{(i-1)!(n-i)!} \int_0^1 x^{i+1}(1-x)^{n-i}dx \\ &= \frac{n!}{(i-1)!(n-i)!} \int_0^1 x^{(i+2)-1}(1-x)^{(n-i+1)-1}dx \\ &= \frac{n!}{(i-1)!(n-i)!} Be(i+2, n-i+1) \\ &= \frac{n!}{(i-1)!(n-i)!} \frac{\Gamma(i+2)\Gamma(n-i+1)}{\Gamma(n+3)} \\ &= \frac{n!}{(i-1)!(n-i)!} \frac{(i+1)!(n-i)!}{(n+2)!} \\ &= \frac{n!(i+1)i(i-1)!}{(i-1)!(n+2)(n+1)n!} \\ &= \frac{i(i+1)}{(n+2)(n+1)} \end{aligned}$$

Caused we have already konwn that the expectation of the ith order statistics equals $\frac{i}{n+1}$ So we can calculate the variance of the ith order statistics X as following,

$$\begin{aligned} Var(X) &= E(X^2) - E^2(X) \\ &= \frac{i(i+1)}{(n+2)(n+1)} - \left(\frac{i}{n+1}\right)^2 \\ &= \frac{\frac{i}{n+1}(1-\frac{i}{n+1})}{n+2} \\ &= \frac{\mu_{n:i}(1-\mu_{n:i})}{n+2} \end{aligned}$$

So $Var(X) \leq \frac{1}{n}$.

Thus,we complete the proof of Lemma 1.

Lemma 2.

Under condition (C3), , the limiting variance σ^2 is well defined.

Proof.

It suffices to show that for $1 \le j, k \le d$

$\beta_1 := \int\int_{0<x<y<1} |G_{j,k}(x,y) - xy||\psi_j(x)\psi_k(y)|dxdy < \infty$

$\beta_2 := \int\int_{0<y<x<1} |G_{j,k}(x,y) - xy||\psi_j(x)\psi_k(y)|dxdy < \infty$

To prove above lemma,we introduce $W_j(x) := I\left(U_1^j \le x\right) - x$.Here $W_j(x) = I\left(U_1^j \le x\right) - x$.

$E\left(W_j(x)\right) = x(1-x) + (-x)(1-x) = 0$

$Var\left(W_j(x)\right) = x(1-x)^2 + (1-x)(-x)^2 = x(1-x)$

Furthermore,

$EW_j(x)W_k(y) = G_{j,k}(x,y) - xy$

In this way,$EW_j(x)W_j(y) = x(1-y)$ when $x < y$. By Cauchy-Schwarz inequality, β_1^2 is bounded above by

$\left(E\int_0^1\int_0^y \left(\frac{x}{1-y}\right)^{1/4}|\psi_j(x)||W_j(x)|\left(\frac{1-y}{x}\right)^{1/4}|\psi_k(y)||W_k(y)|dxdy\right)^2$

$\le E\int_0^1\int_0^y \left[\frac{x}{1-y}\right]^{1/2}[\psi_j(x)]^2[W_j(x)]^2dxdy \cdot E\int_0^1\int_0^y \left[\frac{1-y}{x}\right]^{1/2}[\psi_k(y)]^2[W_k(y)]^2dxdy$

$= \int_0^1\int_0^y x^{3/2}(1-x)(1-y)^{-1/2}[\psi_j(x)]^2dxdy \cdot \int_0^1\int_0^y x^{-1/2}y(1-y)^{3/2}[\psi_k(y)]^2dxdy$

$= 4\int_0^1 x^{3/2}(1-x)^{3/2}[\psi_j(x)]^2dx \cdot \int_0^1 y^{3/2}(1-y)^{3/2}[\psi_k(y)]^2dx < \infty$

By the similar way ,we can prove β_2 satisfy the above inequality.Thus,we completes the proof of above lemma2.

Lemma3

Let γ be the function associated with ϕ as defined as $\gamma(x) := \psi(x,x,\ldots,x) = \phi(F_1^{-1}(x), F_2^{-1}(x), \ldots, F_d^{-1})$.Suppose $\phi : (0,1)^d \to R$ satisfy (C3),and that γ is Riemann integrable, then we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma(\mu_{n:i}) - \sqrt{n}\int_0^1 \gamma(x)dx \to 0, asn \to \infty$$

.

Proof:

As $\gamma^{'}(x) = \psi_1(x) + \ldots + \psi_d(x)$,condition C(3) implies that $[x(1-x)]^{3/2}[\gamma'(x)]^2$ is Riemann integrable. We have

$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \gamma(\mu_{n:i}) - \sqrt{n} \int_0^1 \gamma(x)dx$

$= \sqrt{n} \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} [\gamma(\mu_{n:i}) - \gamma(x)]dx$

$= \sqrt{n} \sum_{i=1}^{n} [\int_{(i-1)/n}^{\mu_{n:i}} \int_{x}^{\mu_{n:i}} \gamma^{'}(y)dydx - \int_{\mu_{n:i}}^{i/n} \int_{x}^{\mu_{n:i}} \gamma^{'}(y)dydx]$

$= \sqrt{n} \int_0^1 g_n(y)\gamma^{'}(y)dy$

where

$$g_n(y) = \begin{cases} y-(i-1)/n, & \text{if } (i-1)/n \le y < i/(n+1), 1 \le i \le n, \\ y - i/n, & \text{if } i/(n+1) \le y < i/n, 1 \le i \le n. \end{cases}$$

Note that

$$|g_n(y)| \le \begin{cases} y, & \text{if } 0 < y \le 1/n, \\ 1/n, & \text{if } 1/n < y < 1-1/n, \\ 1-y, & \text{if } 1-1/n \le y < 1. \end{cases}$$

Therefore,

$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \gamma(\mu_{n:i}) - \sqrt{n} \int_0^1 \gamma(x)dx\right)^2$

$= n \left(\int_0^1 g_n(y)\gamma^{'}(y)dy\right)^2$

$\le n \int_0^1 [g_n(y)]^2[y(1-y)]^{-3/2}dy \cdot \int_0^1 [y(1-y)]^{3/2}[\gamma^{'}(y)]^2 dy.$

Since the second term above is finite by (C3), Lemma3 will follow if we can show that the first term goes to 0 as $n \to \infty$. Towards this end,

$n \int_0^1 [g_n(y)]^2[y(1-y)]^{-3/2}dy$

$\le 2^{3/2}n \left(\int_0^{1} \sqrt{y}dy + \int_{1-\frac{1}{n}}^{1} \sqrt{1-y}dy\right) + \frac{1}{n} \int_{\frac{1}{n}}^{1-\frac{1}{n}} y^{-3/2}(1-y)^{-3/2}dy$

$\le \frac{8\sqrt{2}}{3\sqrt{n}} + (1-n^{-1})^{-3/4}n^{-1/4} \int_0^1 y^{-3/4}(1-y)^{-3/4}dy \to 0.$

Lemma4

Let $U_{n:i}$ denote the *ith* order statistic of an i.i.d. sample of size n from the uniform distribution over $(0, 1)$.Define $\mathcal{A}_{m,n} = \bigcap_{1\leq i\leq n}\{U_{n:i}(1 - U_{n:i}) > \mu_{n:i}(1 - \mu_{n:i})/m\}$,then we have

$lim_{m\to}sup_{n\geq 1}P(\mathcal{A}_{m,n}) = 1.$

Proof:

By symmetry consideration ,we only need to prove

$lim_{m\to\infty}sup_{n\geq 1}P(\bigcap_{1\leq i\leq n/2}\{U_{n:i}(1 - U_{n:i}) > \mu_{n:i}(1 - \mu_{n:i})/m\}) = 1$

For any $\varepsilon > 0$,we can choose an integer n_0 such that for all $n > n_0$,

$P(U_{n:[(n+1)/2]} \geq 2/3) < \varepsilon/2$

Note that for all $n > n_0$,

$P(\bigcap_{1\leq i\leq n/2}\{U_{n:i}(1 - U_{n:i}) > \mu_{n:i}(1 - \mu_{n:i})/m\})$

$\geq P(\bigcap_{1\leq i\leq n/2}\{U_{n:i} > 3\mu_{n:i}/m\}) - P(\{U_{n:[(n+1)/2]} \geq 2/3\})$

$\geq P(\bigcap_{1\leq i\leq n/2}\{U_{n:i} > 3\mu_{n:i}/m\}) - \varepsilon/2.$

Obviously,we can find a constant m_0 such that for all $m > m_0$,

$sup_{1\leq n\leq n_0}P(\bigcap_{1\leq i\leq n/2}\{U_{n:i}(1 - U_{n:i}) > \mu_{n:i}(1 - \mu_{n:i})/m\}) > 1 - \epsilon.$

If we can choose a constant m_1 such that for all $m > m_1$,

$sup_{n>n_0}P(\bigcap_{1\leq i\leq n/2}\{U_{n:i} > 3\mu_{n:i}/m\}) \geq 1 - \epsilon/2,$

then for all $m > max(m_0, m_1)$,

$sup_{n\geq 1}P(\bigcap_{1\leq i\leq n/2}\{U_{n:i}(1 - U_{n:i}) > \mu_{n:i}(1 - \mu_{n:i})/m\}) > 1 - \epsilon.$

Therefore,the proof of Lemma4 reduces to show that

$lim_{m\to\infty}sup_{n\geq 1}P(\bigcap_{1\leq i\leq n/2}\{U_{n:i} > \mu_{n:i}/m\}) = 1.$

Recall the representation formula $U_{n:i} = \frac{e_1+\cdots e_i}{e_1+\cdots e_{n+1}}$, where $e_1, \cdots, e_{n+1}$ are independent exponentially distributed random variables of mean 1.Write $S_i = e_1 + \cdots + e_i, 1 \leq i \leq n + 1$. Let

$M = inf_{1\leq i\leq n<\infty}\frac{S_i/i}{S_{n+1}/(n+1)}.$

Since $S_n/n \to 1$,a.s. as $n \to \infty$,we have

$P(M > 0) = 1$.

Thus,when $M > 1/m$,for all $1 \leq i \leq n/2$,we have $\frac{S_i/i}{S_{n+1}/(n+1)} > 1/m$,which implies that as $m \to \infty$,

$lim_{m\to\infty} sup_{n\geq 1} P(\bigcap_{1\leq i\leq n/2}\{\frac{S_i/i}{S_{n+1}/(n+1)} > 1/m\}) \geq lim_{m\to\infty} P(M > 1/m) = 1$.

Then we complete the proof of lemma4.

Proof of Theorem2,we write,

$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\left(U_{n:i}^{(1)}, \cdots, U_{n:i}^{(d)}\right) - \sqrt{n}\bar{\gamma} = I_n + \epsilon_n = S_{n,1} + S_{n,2} + \epsilon_n$

where

$I_n = n^{-1/2} \sum_{i=1}^{n} \left[\psi\left(U_{n:i}^{(1)}, \cdots, U_{n:i}^{(d)}\right) - \gamma(\mu_{n:i})\right]$,

$S_{n,1} = n^{-1/2} \sum_{j=1}^{d} \sum_{i=1}^{n} \left(U_{n:i}^{(j)} - \mu_{n:i}\right) \psi_j(\mu_{n:i})$,

$S_{n,2} = I_n - S_{n,1}$

$\epsilon_n = n^{-1/2} \sum_{i=1}^{n} \gamma(\mu_{n:i}) - \sqrt{n} \int_0^1 \gamma(x)dx$.

By Lemma3 ,$\epsilon_n \to 0$ as $n \to \infty$. We shall first show that $S_{n,2} \to 0$ in probability as $n \to \infty$. Then we will prove that $S_{n,1} \to N(0, \sigma^2)$ in distribution as $n \to \infty$.

Since $\max\{|U_{n:i}^{(j)}| : 1 \leq i \leq n, 1 \leq j \leq d\} \to 0, a.s.$,and by C(4),we have

$|S_{n,2}| I_{\mathcal{A}_{m,n}} \leq \frac{K(m)}{\sqrt{n}} \sum_{j,k=1}^{d} \sum_{i=1}^{n} |(U_{n:i}^{(j)} - \mu_{n:i})(U_{n:i}^{(k)} - \mu_{n:i})|[1 + |\tilde{\psi}_{j,k}(\mu_{n:i})|]$

where $\tilde{\psi}_{j,k}(x) = \psi_{j,k}(x, \ldots, x)$. By condition (C3),Lemma1 and Cauchy-Schwarz inequality ,we obtain

$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E|(U_{n:i}^{(j)} - \mu_{n:i})(U^{(k)} - \mu_{n:i})|[1 + |\tilde{\psi}_{j,k}(\mu_{n:i})|]$

$\leq \frac{1}{n^{3/2}} \sum_{i=1}^{n} \mu_{n:i}(1 - \mu_{n:i})|\tilde{\psi}_{j,k}(\mu_{n:i})| + \frac{1}{\sqrt{n}}$

$:= J_1 + J_2 + J_3 + \frac{1}{\sqrt{n}}$

where $J_1 = n^{-3/2} \sum_{1\leq i\leq\sqrt{n}} \mu_{n:i}(1 - \mu_{n:i})|\tilde{\psi}_{j,k}(\mu_{n:i})|$,

and J_2 and J_3 are similarly defined over $\sqrt{n < i < n - \sqrt{n}}$ and $n - \sqrt{n} \leq i \leq n$ respectively .

We have

$J_1 \leq \frac{2}{n}\sum_{1\leq i\leq\sqrt{n}}[\mu_{n;i}(1-\mu_{n:i})]^{3/2}|\tilde{\psi}_{j,k}(\mu_{n:i})|$

$2\int_0^{1/\sqrt{n}}[x(1-x)]^{3/2}|\tilde{\psi}_{j,k}(x)|dx \to 0$ as $n\to\infty$.Similarly,$J_3 \to \infty$ as $n \to \infty$.Also ,as $n\to\infty$,

$J_2 \leq \frac{1}{n^{5/4}}\sum_{\sqrt{n}<i<n-\sqrt{n}}[\mu_{n:i}(1-\mu_{n:i})]^{3/2}|\tilde{\psi}_{j,k}(\mu_{n:i})| \to 0.$

That is, we have shown that as $n\to\infty$,for any given m,$S_{n,2}I_{\mathcal{A}_{m,n}} \to 0$ in probability.We can then choose a sequence of $m = m_n \to \infty$ such that $S_{n,2}I_{\mathcal{A}_{m,n}} \to 0$ in probability as $n\to\infty$.

Lemma 4 implies $I_{\mathcal{A}^c_{m,n}} \to 0$ in probability as (and hence $S_{n,2}I_{\mathcal{A}^c_{m,n}} \to 0$ in probability) as $m\to\infty$.Therefore $S_{n,2}\to 0$ in probability.

Modify the notation in the proof of Lemma2:$W_{j,l}(x) = I(U_l^{(j)} \leq x) - x$ for $1\leq j\leq d$ and $1\leq l\leq n$. We will abbreviate $W_{j,1}$ to W_j.Note also that $\hat{F}^{-1}_{n;j}(\frac{i}{n}) = U^{(j)}_{n:i}$.

By Bahadur's representation of quantile,we have

$sup_{0<t<1}|\hat{F}_{n;j}(t) - t + \hat{F}^{-1}_{n;j}(t) - t| = O(n^{-3/4}logn)$ a.s. for $1\leq j\leq d$,

we have

$S_{n,1} = \frac{1}{\sqrt{n}}\sum_{l=1}^{n} Z_{n,l} + o(1)$ a.s.,

where,for $1\leq l\leq n$,

$Z_{n,l} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} W_{j,l}(i/n)\psi_j(\mu_{n:i})$

are i.i.d. mean zero random variables satisfying

$var(Z_{n,1}) = \sum_{j,k=1}^{d}\frac{1}{n^2}\sum_{h,i=1}^{n} cov(W_j(h/n), W_k(i/n))\psi_j(\mu_{n:h})\psi_k(\mu_{n:i})$

$= \sum_{j,k=1}^{d}\frac{1}{n^2}\sum_{h,i=1}^{n}[G_{j,k}(h/n, i/n) - hi/n^2]\psi_j(\mu_{n:h})\psi_k(\mu_{n:i})$

$\to \sum_{j,k=1}^{d}\int_0^1\int_0^1[G_{j,k}(x,y) - xy]\psi_j(x)\psi_k(y)dxdy$

where $G_{j,k}$ is the joint distribution of $U_1^{(j)}$ and $U_1^{(i)}$.Note that $G_{j,j}(x,y) = min(x,y)$ for $1\leq j\leq d$. To establish the convergence above for $1\leq j,k\leq d$ fixed,one can split the sum into cases according to whetther h,or i,is less than ϵn,or between ϵn

and $(1-\epsilon)n$,or greater than $(1-\epsilon)n$. For example,when we sum over $\epsilon n \leq h, i \leq (1-\epsilon)n$, then it converges to $\int_{\epsilon}^{1-\epsilon}\int_{\epsilon}^{1-\epsilon} H(x,y)dxdy$ where $H(x,y) = [G_{j,k}(x,y) - xy]\psi_j(x)\psi_k(y)$. When $1 \leq h < \epsilon n$ and $\epsilon n \leq i \leq (1-\epsilon)n$, it can be shown to convergence to $\int_0^{\epsilon}\int_{\epsilon}^{1-\epsilon} H(x,y)dxdy$ which,from the method of proof of Lemma2 and condition (C3),can be shown to convergence to 0 as $\epsilon \to 0$.Similarly for the other ranges of h and i.

It is then easy to see that the limit above can be written in the form of σ^2 as stated in Theorem1.2.Note that $|Z_{n,1}| \leq \sum_{j=1}^{d} \frac{1}{n}\sum_{i=1}^{n} |\psi_j(\mu_{n:i})|$.If $(1/\sqrt{n})\frac{1}{n}\sum_{i=1}^{n}|\psi_j(\mu_{n:i})| \to$ 0

for $j = 1, 2, \ldots, d$, then Lindeberg-Levy condition holds.To see this,note that

$\left(\frac{1}{n}\sum_{i=1}^{n}|\psi_j(\mu_{n:i})|\right)^2$

$\leq \frac{1}{n}\sum_{i=1}^{n}[\mu_{n:i}(1-\mu_{n:i})]^{-3/2} \cdot \frac{1}{n}\sum_{i=1}^{n}[\mu_{n:i}(1-\mu_{n:i})]^{3/2}(\psi_j(\mu_{n:i}))^2$.

By C(3),it is enough to establish $I_n = \frac{1}{n^2}\sum_{i=1}^{n}[\frac{i}{n+1}(1-\frac{i}{n+1})]^{-3/2} \to 0$. Now $I_n \leq \frac{4(n+1)^{3/2}}{n^2}[\sum_{1\leq i\leq (n+1)/2} i^{-3/2} + \sum_{(n+1)/2\leq i\leq n}(n+1-i)^{-3/2}] \to 0$.Hency by Linderberg-Levy central limit theorem,

$\frac{1}{\sqrt{n}}\sum_{l=1}^{n} Z_{n,l} \to N(0,\sigma^2)$.

Hence $S_{n,1} \to N(0,\sigma^2)$ which completes the proof of theorem2.

Chapter 3

Copula of marginal exponential and Morgenstain example

The univariate exponential distribution plays a central role in mathematical statistics because it is the distribution of waiting time in a standard Poisson process. The following bivariate exponential distribution,first described by Marshall and Olkin,plays a similar role in a two-dimentional Poisson process. Consider a two-component system-such as a two engine aircraft,or a desktop computer with both a CPU(central processing unit) and a co-processor. The computers are subject to ”shocks”,which are always ”fatal” to one or both of the components. For example,one of the two aircraft engines may fail,or a massive explosion could destroy both engines simultaneously;or the CPU or the co-processor could fail,or a power surge could eliminate both simultaneously. Let X and Y denote the lifetimes of the components 1 and 2,respectively. As is often the case in dealing with lifetimes,we will find the survival function $\bar{H}(x,y) = P[X > x, Y > y]$,the probability that component 1 survives beyond time x and that component 2 survives beyond time y. The ”shocks” to the two components are assumed to form three independent Poisson processes with(positive) parameters λ_1,λ_2,andλ_{12},respectively.So

$X = min(Z_1, Z_{12})$,$Y = min(Z_2, Z_{12})$,and hence for all $x, y \geq 0$,

$\bar{H}(x, y)$

$= P(Z_1 > x, Z_2 > y, Z_{12} > max(x, y))$

$= \left(\int_x^\infty \lambda_1 exp(-\lambda_1 \nu) d\nu\right) \left(\int_y^\infty exp(-\lambda_2 \nu) d\nu\right)$
$\left(\int_{max(x,y)}^\infty exp(-\lambda_{12} exp(-\lambda_{12}\nu) d\nu\right)$

$= exp(-\lambda_1 x) exp(-\lambda_2 y) exp(-\lambda_{12} max(x, y))$

$= exp[-\lambda_1 x - \lambda_2 y - \lambda_{12} max(x, y)]$

The marginal survival functions for the lifetimes of the components 1 and 2 are,

$\bar{F}(x) = \bar{H}(x, 0) = P(Z_1 > x, Z_2 > 0, Z_{12} > max(x, 0)) = exp[-(\lambda_1 x + \lambda_{12} x]$

$\bar{G}(y) = \bar{H}(0, y) = P(Z_1 > 0, Z_2 > y, Z_{12} > max(0, y)) = exp[-(\lambda_2 y + \lambda_{12} y)]$

respectively.

So it is easy for us to calculate the xth and yth survival quantile are $-\frac{lnx}{\lambda_1+\lambda_{12}}$,$-\frac{lny}{\lambda_2+\lambda_{12}}$ respectively.

And meanwhile,we can also get the xth and yth distribution quantile $-\frac{ln(1-x)}{\lambda_1+\lambda_{12}}and-\frac{ln(1-y)}{\lambda_2+\lambda_{12}}$seperately.

Let's find the survival copula,we write $\hat{C}(u, v)$ to denote the survival copula:

$\hat{C}(u, v) = \bar{H}(\bar{F}^{-1}(u), \bar{G}^{-1}(t))$

$= \bar{H}(-\frac{lnu}{\lambda_1+\lambda_{12}}, -\frac{lnv}{\lambda_2+\lambda_{12}})$

$= exp[-\lambda_1(-\frac{lnu}{\lambda_1+\lambda_{12}}) - \lambda_2(-\frac{lnv}{\lambda_2+\lambda_{12}}) - \lambda_{12} max(-\frac{lnu}{\lambda_1+\lambda_{12}}, -\frac{lnv}{\lambda_2+\lambda_{12}})]$
$= e^{lnu^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}} e^{lnv^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}} e^{-\lambda_{12} max(lnu^{-\frac{1}{\lambda_1+\lambda_{12}}}, lnv^{-\frac{1}{\lambda_2+\lambda_{12}}})}$

So we have

$$\hat{C}(u, v) = \begin{cases} uv^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}} & , \quad lnu^{-\frac{1}{\lambda_1+\lambda_{12}}} \geq lnv^{-\frac{1}{\lambda_2+\lambda_{12}}} \\ u^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}} v & , \quad lnv^{-\frac{1}{\lambda_2+\lambda_{12}}} \geq lnu^{-\frac{1}{\lambda_1+\lambda_{12}}} \end{cases}$$

above equation is equivalent to the following equation,

$$\hat{C}(u,v) = \begin{cases} uv^{1-\frac{\lambda_{12}}{\lambda_2+\lambda_{12}}} & , \quad u^{-\frac{1}{\lambda_1+\lambda_{12}}} \geq v^{-\frac{1}{\lambda_2+\lambda_{12}}} \\ u^{1-\frac{\lambda_{12}}{\lambda_1+\lambda_{12}}}v & , \quad v^{-\frac{1}{\lambda_2+\lambda_{12}}} \geq u^{-\frac{1}{\lambda_1+\lambda_{12}}} \end{cases}$$

so we can get a furthermore abbreviation formulation as following,

$$\hat{C}(u,v) = \begin{cases} u^{1-\alpha}v & , \quad u^{\alpha} \geq v^{\beta} \\ uv^{1-\beta} & , \quad u^{\alpha} \leq v^{\beta} \end{cases}$$

if we let $\alpha = \frac{\lambda_{12}}{\lambda_1+\lambda_{12}}$ and $\beta = \frac{\lambda_{12}}{\lambda_2+\lambda_{12}}$ seperately.

$$F_{12}(x,y) = 1 - e^{-(\lambda_1+\lambda_3)x} - exp^{-(\lambda_2+\lambda_3)y} + e^{-\lambda_1 x - \lambda_2 y - \lambda_3 max(x,y)}, x > 0, y > 0$$

$$\hat{F}(x,y) = exp\{-\lambda_1 x - \lambda_2 y - \lambda_3 max(x,y)\}, x > 0, y > 0$$

$\hat{F}(x) = exp\{-\lambda_1 x - \lambda_{12} x\}$ $\hat{F}(y) = exp\{-\lambda_2 y - \lambda_{12} y\}$

$F(x) = 1 - exp\{-(\lambda_1 + \lambda_{12})x\}$ $G(y) = 1 - exp\{-(\lambda_2 + \lambda_{12})y\}$

We can calculate the copula following the formulation,

$$\begin{aligned} C(u,v) &= F_{12}(F^{-1}(u), G^{-1}(v)) \\ &= F_{12}(-\frac{ln(1-u)}{\lambda_1+\lambda_{12}}, -\frac{ln(1-v)}{\lambda_2+\lambda_{12}} \\ &= 1 - exp\{-(\lambda_1+\lambda_{12})(-\frac{ln(1-u)}{\lambda_1+\lambda_{12}})\} - exp\{-(\lambda_2+\lambda_{12})(-\frac{ln(1-v)}{\lambda_2+\lambda_{12}})\} \\ &\quad + exp\{-\lambda_1(-\frac{ln(1-u)}{\lambda_1+\lambda_{12}}) - \lambda_2(-\frac{ln(1-v)}{\lambda_2+\lambda_{12}}) \\ &\quad -\lambda_{12} max(-\frac{ln(1-u)}{\lambda_1+\lambda_{12}}, -\frac{ln(1-v)}{\lambda_2+\lambda_{12}})\} \\ &= 1 - (1-u) - (1-v) + exp\{ln(1-u)^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}} + ln(1-v)^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}} \\ &\quad -\lambda_{12} max(-ln(1-u)^{\frac{1}{\lambda_1+\lambda_{12}}}, -ln(1-v)^{\frac{1}{\lambda_2+\lambda_{12}}}\} \\ &= u - 1 + v + exp\{ln(1-u)^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}\} exp\{ln(1-v)^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}\} \\ &\quad exp\{-\lambda_{12} max(ln(1-u)^{-\frac{1}{\lambda_1+\lambda_{12}}}, ln(1-v)^{-\frac{1}{\lambda_2+\lambda_{12}}})\} \end{aligned} \tag{3.1}$$

So we can get the corresponding copula as following,

$u-1+v+exp\{ln(1-u)^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}\}exp\{ln(1-v)^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}\}exp{-}\lambda_{12}ln(1-u)^{-\frac{1}{\lambda_1+\lambda_{12}}}$

if $ln(1-u)^{-\frac{1}{\lambda_1+\lambda_{12}}} \geq ln(1-v)^{-\frac{1}{\lambda_2+\lambda_{12}}}$

$u-1+v+exp\{ln(1-u)^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}\}exp\{ln(1-v)^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}\}exp{-}\lambda_{12}ln(1-v)^{-\frac{1}{\lambda_2+\lambda_{12}}}$

if $ln(1-v)^{-\frac{1}{\lambda_2+\lambda_{12}}} \geq ln(1-u)^{-\frac{1}{\lambda_1+\lambda_{12}}}$

3.1 Copula of marginal exponential

Consider the Marshall-Olkin Bivariate Exponential Distribution,the joint distribution function is,

$$F_{12}(x,y) = 1-exp[-(\lambda_1+\lambda_{12}x)]-exp[-(\lambda_2+\lambda_{12})y]+exp[-\lambda_1 x-\lambda_2 y-\lambda_{12}max(x,y)],$$

$x>0, y>0$

The marginal distribution of x is

$$F(x) = 1-exp[-(\lambda_1+\lambda_{12})x]$$

and the corresponding pth quantile x is $x = -\frac{ln(1-p)}{\lambda_1+\lambda_{12}}$.

The marginal distribution of y is

$$G(y) = 1-exp[-(\lambda_2+\lambda_{12})y]$$

and the corresponding qth quantile y is $y = -\frac{ln(1-q)}{\lambda_2+\lambda_{12}}$.

$$\phi(x,y) = \left(-\frac{ln(1-x)}{\lambda_1+\lambda_{12}} + \frac{ln(1-y)}{\lambda_2+\lambda_{12}}\right)^2$$

$$\psi_1(x,y) = \frac{\partial\phi(x,y)}{\partial x} = \frac{2[(\lambda_1+\lambda_{12})ln(1-y)-(\lambda_2+\lambda_{12})ln(1-x)]}{(\lambda_1+\lambda_{12})^2(\lambda_2+\lambda_{12})(1-x)}$$

$$\psi_2(x,y) = \frac{\partial\phi(x,y)}{\partial y} = \frac{(-2)[(\lambda_1+\lambda_{12})ln(1-y)-(\lambda_2+\lambda_{12})ln(1-x)]}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-y)}$$

$$\sum_{j=1}^{2}\int_0^1\int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy$$

$$= \int_0^1\int_0^y x(1-y)\psi_1(x,x)\psi_1(y,y)dxdy + \int_0^1\int_0^y x(1-y)\psi_2(x,x)\psi_2(y,y)dxdy$$

$$= \int_0^1\int_0^y x(1-y)\frac{2ln(1-x)(\lambda_1-\lambda_2)}{(\lambda_1+\lambda_{12})^2(\lambda_2+\lambda_{12})(1-x)}\frac{2ln(1-y)(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(1-y)}dxdy+$$

$$\int_0^1\int_0^y x(1-y)\frac{(-2)(\lambda_1-\lambda_2)ln(1-x)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-x)}\frac{(-2)(\lambda_1-\lambda_2)ln(1-y)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-y)}dxdy$$

$$= \frac{4(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^4(\lambda_2+\lambda_{12})^2}\int_0^1\int_0^y \frac{xln(1-x)ln(1-y)}{1-x}dxdy+$$

$$\frac{4(\lambda_1-\lambda_2)^2}{(\lambda_2+\lambda_{12})^4(\lambda_1+\lambda_{12})^2}\int_0^1\int_0^y \frac{xln(1-x)ln(1-y)}{1-x}dxdy$$

$$= \left[\frac{4(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^4(\lambda_2+\lambda_{12})^2} + \frac{4(\lambda_1-\lambda_2)^2}{(\lambda_2+\lambda_{12})^4(\lambda_1+\lambda_{12})^2}\right]\int_0^1\int_0^y \frac{xln(1-x)ln(1-y)}{1-x}dxdy$$

$$= \frac{5}{2}\left[\frac{4(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^4(\lambda_2+\lambda_{12})^2} + \frac{4(\lambda_1-\lambda_2)^2}{(\lambda_2+\lambda_{12})^4(\lambda_1+\lambda_{12})^2}\right]$$

$$2\sum_{j=1}^{2}\int_0^1\int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy$$

$$= 5\left[\frac{4(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^4(\lambda_2+\lambda_{12})^2} + \frac{4(\lambda_1-\lambda_2)^2}{(\lambda_2+\lambda_{12})^4(\lambda_1+\lambda_{12})^2}\right]$$

$$\int_0^1\int_0^1 [G_{12}(x,y)-xy]\psi_1(x)\psi_2(y)dxdy$$

$$= \int_0^1\int_{1-(1-y)^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}}^1 \left([x-1+y+(1-x)(1-y)^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}]-xy]\right)\cdot$$

$$\frac{2ln(1-x)(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(1-x)}\frac{(-2)(\lambda 1-\lambda 2)ln(1-y)}{(\lambda_2)+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-y)}dxdy+$$

$$\int_0^1\int_0^{1-(1-y)^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}} \left((x-1+y+(1-x)^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}(1-y))-xy\right)\cdot\frac{2(\lambda_1-\lambda_2)ln(1-x)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(1-x)}\cdot$$

$$\frac{(-2)(\lambda_1-\lambda_2)ln(1-y)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-y)}dxdy$$

$$= \int_0^1 \int_{1-(1-y)^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}}^1 \left(-(1-x)(1-y) + (1-x)(1-y)^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}\right) \cdot \frac{2(\lambda_1-\lambda_2)ln(1-x)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(1-x)} \cdot \frac{(-2)(\lambda_1-\lambda_2)ln(1-y)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-y)} dxdy +$$

$$\int_0^1 \int_0^{1-(1-y)^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}} \left(-(1-x)(1-y) + (1-x)^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}(1-y)\right) \cdot \frac{2(\lambda_1-\lambda_2)ln(1-x)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(1-x)} \cdot \frac{(-2)(\lambda_1-\lambda_2)ln(1-y)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})(1-y)} dxdy$$

If we let $u = 1-x$,$v = 1-y$,above double integration is equivalent to the following double integration,

$$\int_0^1 \int_0^{v^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}} \left(-uv + uv^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}\right) \cdot \frac{2(\lambda_1-\lambda_2)lnu}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2 u} \cdot \frac{(-2)(\lambda_1-\lambda_2)lnv}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})v} dudv +$$

$$\int_0^1 \int_{v^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}}^1 \left(-uv + u^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}v\right) \cdot \frac{2(\lambda_1-\lambda_2)lnu}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2 u} \cdot \frac{(-2)(\lambda_1-\lambda_2)lnv}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})v} dudv$$

$$= \int_0^1 \int_0^{v^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}} \frac{2(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2} \frac{lnu}{u} \frac{(-2)(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})} \frac{lnv}{v} (-uv + uv^{\frac{\lambda_2}{\lambda_2+\lambda_{12}}}) dudv +$$

$$\int_0^1 \int_{v^{\frac{\lambda_1+\lambda_{12}}{\lambda_2+\lambda_{12}}}}^1 \frac{2(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2} \frac{lnu}{u} \frac{(-2)(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})} \frac{lnv}{v} (-uv + u^{\frac{\lambda_1}{\lambda_1+\lambda_{12}}}v) dudv$$

$$= \frac{2(\lambda_1-\lambda_2)(-2)(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})} \cdot$$

$$\left(-\frac{(\lambda_{12}+\lambda_2)^2(3\lambda_1+4\lambda_{12}+\lambda_2)}{(\lambda_1+2\lambda_{12}+\lambda_2)^3} + \frac{(\lambda_{12}+\lambda_2)^2(3(\lambda_1+\lambda_{12})+\lambda_2)}{(\lambda_1+\lambda_{12}+\lambda_2)^3}\right) +$$

$$\frac{2(\lambda_1-\lambda_2)(-2)(\lambda_1-\lambda_2)}{(\lambda_2+\lambda_{12})(\lambda_1+\lambda_{12})^2(\lambda_2+\lambda_{12})^2(\lambda_1+\lambda_{12})} \cdot$$

$$\left(-\frac{(\lambda_1+\lambda_{12})^2(\lambda_1+4\lambda_{12}+3\lambda_2)}{(\lambda_1+2\lambda_{12}+\lambda_2)^3} + \frac{(\lambda_1+\lambda_{12})^2(\lambda_1+3(\lambda_{12}+\lambda_2))}{(\lambda_1+\lambda_{12}+\lambda_2)^3}\right)$$

So according to the result of theorem2,

$$\sigma^2 = \lim_{n\to\infty} Var(Z_{n,1}) = 2\sum_{j=1}^{d} \int_0^1 \int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy$$

$$+2\sum_{1\leq j<k\leq d}\int_0^1\int_0^1[G_{j,k}(x,y)-xy]\psi_j(x)\psi_k(y)dxdy$$

$$whereG_{j,k}(x,y)=F_{j,k}(F_j^{-1}(x),F_k^{-1}(y))$$

.

we have,

$$2\int_0^1\int_0^1[G_{12}(x,y)-xy]\psi_1(x)\psi_2(y)dxdy$$

$$=\frac{(-8)(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^3(\lambda_2+\lambda_{12})^3}\cdot\left(-\frac{(\lambda_{12}+\lambda_2)^2(3\lambda_1+4\lambda_{12}+\lambda_2)+(\lambda_1+\lambda_{12})^2(\lambda_1+4\lambda_{12}+3\lambda_2)}{(\lambda_1+2\lambda_{12}+\lambda_2)^3}\right)$$

$$+\frac{(-8)(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^3(\lambda_2+\lambda_{12})^3}\cdot\left(\frac{(\lambda_{12}+\lambda_2)^2(3(\lambda_1+\lambda_2)+\lambda_2)+(\lambda_1+\lambda_{12})^2(\lambda_1+3(\lambda_{12}+\lambda_2))}{(\lambda_1+\lambda_{12}+\lambda_2)^3}\right)$$

$$\sigma^2=2\sum_{j=1}^2\int_0^1\int_0^y x(1-y)\psi_j(x)\psi_j(y)dxdy$$

$$+2\int_0^1\int_0^1[G_{12}(x,y)-xy]\psi_1(x)\psi_2(y)dxdy$$

$$=5\left(\frac{4(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^4(\lambda_2+\lambda_{12})^2}+\frac{4(\lambda_1-\lambda_2)^2}{(\lambda_2+\lambda_{12})^4(\lambda_1+\lambda_{12})^2}\right)+$$

$$\frac{(-8)(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^3(\lambda_2+\lambda_{12})^3}\cdot\left(-\frac{(\lambda_{12}+\lambda_2)^2(3\lambda_1+4\lambda_{12}+\lambda_2)+(\lambda_1+\lambda_{12})^2(\lambda_1+4\lambda_{12}+3\lambda_2)}{(\lambda_1+2\lambda_{12}+\lambda_2)^3}\right)$$

$$+\frac{(-8)(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^3(\lambda_2+\lambda_{12})^3}\cdot\left(\frac{(\lambda_{12}+\lambda_2)^2(3(\lambda_1+\lambda_2)+\lambda_2)+(\lambda_1+\lambda_{12})^2(\lambda_1+3(\lambda_{12}+\lambda_2))}{(\lambda_1+\lambda_{12}+\lambda_2)^3}\right)$$

$$\bar{\gamma}=\int_0^1\left(-\frac{ln(1-u)}{\lambda_1+\lambda_{12}}+\frac{ln(1-u)}{\lambda_2+\lambda_{12}}\right)^2du$$

$$=\frac{2(\lambda_1-\lambda_2)^2}{(\lambda_1+\lambda_{12})^2(\lambda_{12}+\lambda_2)^2}$$

Suppose random variable $u_0\sim exp(\lambda_{12}),u_1\sim exp(\lambda_1),u_2\sim exp(\lambda_2)$,so we have

$$P(min(u_0,u_1)>x,min(u_0,u_2)>y)$$

$$=P(u_0>x,u_1>x),u_0>y,u_2>y)$$

$$= P(u_1 > x, u_0 > max(x, y), u_2 > y)$$

$$= exp(-\lambda_1 x) \cdot exp(-\lambda_{12} max(x, y)) \cdot exp(-\lambda_2 y)$$

$$= exp(-\lambda_1 x - \lambda_2 y - \lambda_{12} max(x, y))$$

$$P(min(u_0, u_1) \leq x, min(u_0, u_2) \leq y)$$

$$= P(min(u_0, u_1) \leq x) - P(min(u_0, u_1) \leq x, min(u_0, u_2) > y)$$

$$= 1 - P(min(u_0, u_1 > x)) - P(min(u_0, u_2) > y) - (P(min(u_0, u_1) > x, min(u_0, u_2) > y))$$

$$= 1 - exp(-\lambda_1 x) \cdot exp(-\lambda_{12} x) - exp(-\lambda_2) exp(-\lambda_{12} y) + exp(-\lambda_1 x - \lambda_2 y - \lambda_{12} max(x, y))$$

So the random number generation mechanism of the aboved mentioned Marshall-Olkin Exponential Distribution is equivalent to the following random number generation mechanism.

$x = min(u_0, u_1), y = min(u_0, u_2), u_0 \sim exp(\lambda_{12}), u_1 \sim exp(\lambda_1), u_2 \sim exp(\lambda_2)$ The simulation result is as following:

When $\lambda_1 = 4$,$\lambda_2 = 5$,$\lambda_{12} = 6$,$n = 1000$,the Q-Q plot is as following,

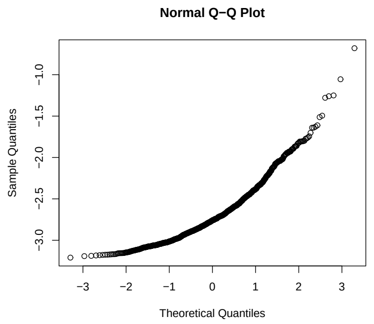

Figure 3.1: QQ plot when number of observation equals 1000

When $\lambda_1 = 4$,$\lambda_2 = 5$,$\lambda_{12} = 6$,$n = 5000$,the

Q-Q plot is as following,

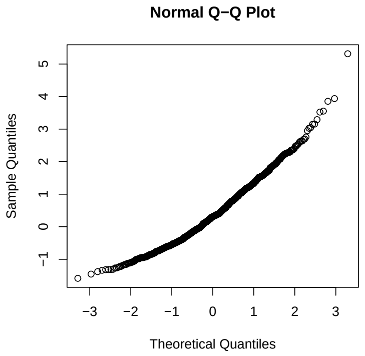

Figure 3.2: QQ plot when number of observation equals 5000

When $\lambda_1 = 4$,$\lambda_2 = 5$,$\lambda_{12} = 6$,$n = 10000$,the

Q-Q plot is as following,

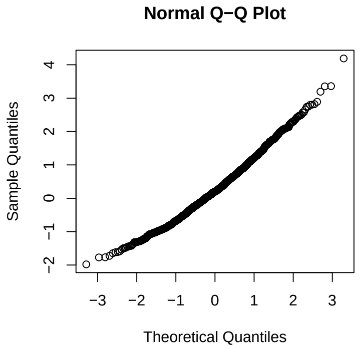

Figure 3.3: QQ plot when number of observation equals 10000

When $\lambda_1 = 4$,$\lambda_2 = 5$,$\lambda_{12} = 6$,$n = 50000$,the

Q-Q plot is as following,

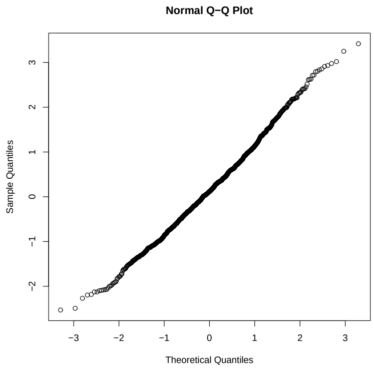

Figure 3.4: QQ plot when number of observation equals 50000

Histogram is as following when number of observation$n = 1000$,iteration times equals$= 1000$,

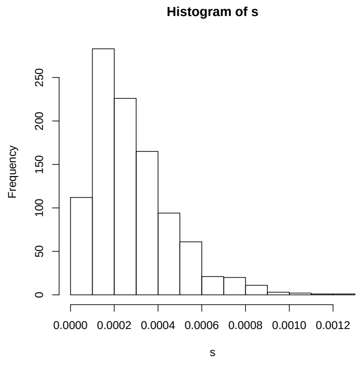

Figure 3.5: Histogram when number of observation equals 1000

Histogram is as following when number of observation$n = 5000$,iteration times equals$= 1000$,

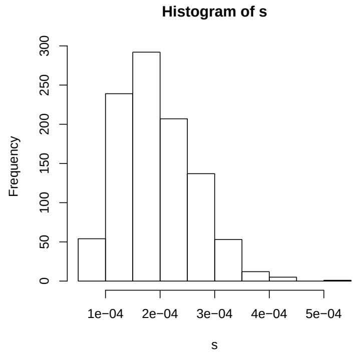

Figure 3.6: Histogram when number of observation equals 5000

Histogram is as following when number of observation$n = 10000$,iteration times equals$= 1000$,

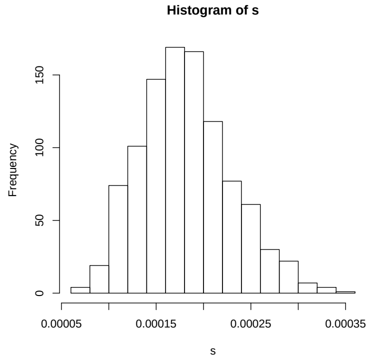

Figure 3.7: Histogram when number of observation equals 10000

Histogram is as following when number of observation$n = 50000$,iteration times equals$= 1000$,

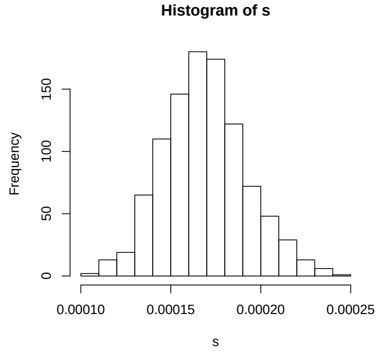

Figure 3.8: Histogram when number of observation equals 50000

When we take the number of observations takes value from 1000 to 50000,step takes 100,the corresponding simulation result of MSE is as following,

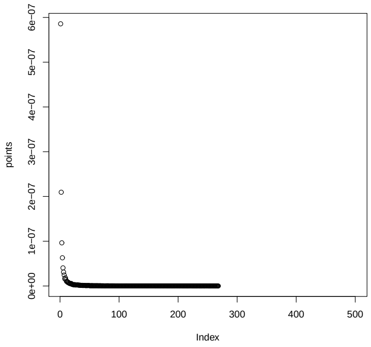

Figure 3.9: MSE when number of observations takes value from 1000 to 50000

3.2 Morgenstain

Copula Consider the Farlie-Gumbel-Morgenstern family of copula,the corresponding marginal distribution belong to uniform distribution.

$$C_\theta(u,v) = uv + \theta uv(1-u)(1-v)$$

Corresponding joint distribution is $F_\theta(x,y) = xy + \theta xy(1-x)(1-y)$ So the corresponding joint probability density function is as following,

$$p(x,y) = \frac{\partial F_\theta(x,y)}{\partial x \partial y} = 1 + \theta - 2\theta y - 2\theta x + 4\theta xy$$

$$0 < x < 1, 0 < y < 1, -1 \leq \theta \leq 1$$

Corresponding sample generation mechanism is as following,

$X \sim$ Uniform$(0,1)$,$y = \frac{(k+1)-\sqrt{(k+1)^2-4kz}}{2k}$,and $k = \theta(1-2x)$and $Z \sim$ Uniform $(0,1)$

Simulation result graph is as following,

When number of observation is 1000,QQ plot and histogramme is as following,

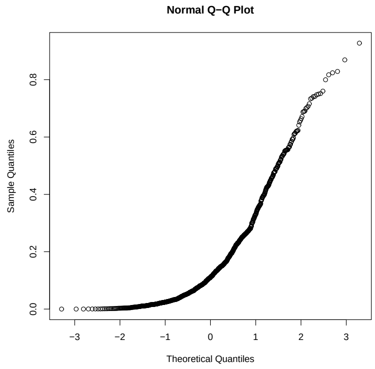

Figure 3.10: QQ plot when number of observation equals 1000

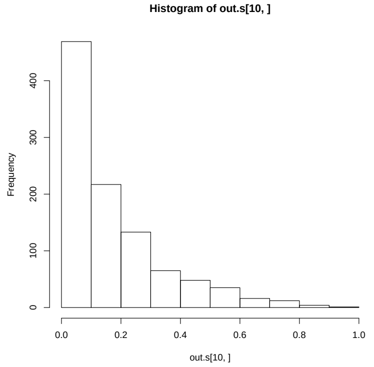

Figure 3.11: Histogramme when number of observation equals 1000

When number of observation is 5000,the corresponding QQ plot and histogramme is as following,

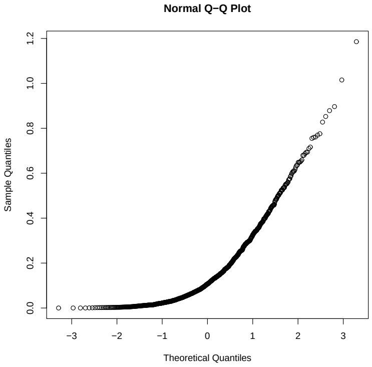

Figure 3.12: QQ plot when number of observation equals 5000

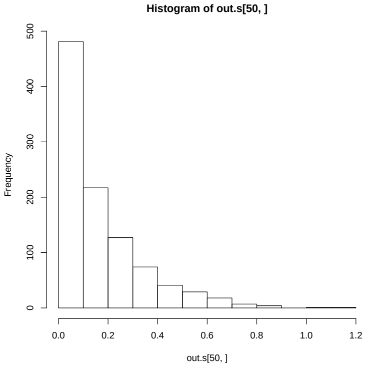

Figure 3.13: Histogramme when number of observation equals 5000

When number of observation is 10000,the corresponding QQ plot and histogramme is as following,

When number of observation is 50000,the corresponding QQ plot and histogramme is as following,

MSE When we take the number of observations takes value from 1000 to 50000,step takes 100,the corresponding simulation result of is as following,

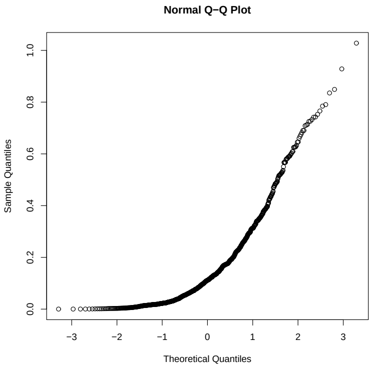

Figure 3.14: QQ plot when number of observation equals 10000

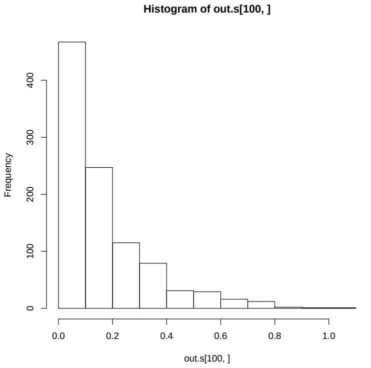

Figure 3.15: Histogram when number of observation equals 10000

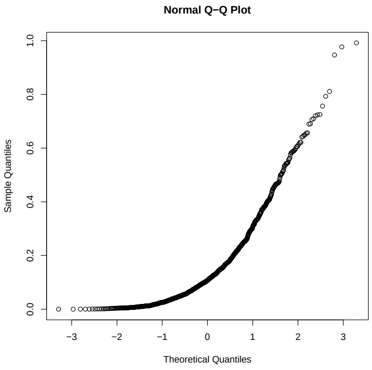

Figure 3.16: QQ plot when number of observation equals 50000

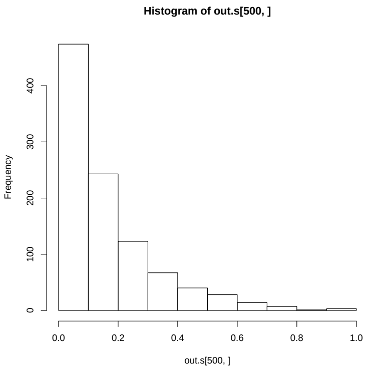

Figure 3.17: Histogram when number of observation equals 50000

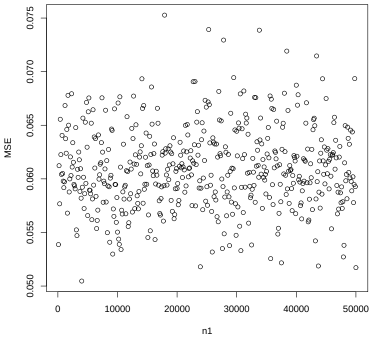

Figure 3.18: MSE when number of observation takes value form 1000 to 50000

Bibliography

[1] Babu, G.J.,Rao,C.R.(1988).Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population.J.Multivariate Anal.27,15-23.

[2] Bai,Z.D.,Hsing,T.(2005).The broken sample problem.Probab.Theory Related Fields. 131,528-552.

[3] Billingsley,P.(1999).Convergence of Probability Measures.John Wiley,New York.

[4] Copas,J.B.,Hilton,F.J.(1990). Record linkage:Statistical models for matching computer records. J.Roy.Statist.Soc.A 153,287-320.

[5] Chan,H.P.,Loh,W.L.(2001).A file linkage problem of DeGroot and Goel revisited.Statist.Sinica 11,1031-1045.

[6] David,H.A. (1991).Order Statistics .Wiley Series in Probability and Mathematical Statistics.

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

[8] Hardy,G.H.,Littlewood,J.E.,Polya,G.(1980). Estimation of the correlation coefficient from a broken sample.Ann.Statist.8,264-278.

[9] Hardy,G.H.,Littlewood,J.E.,Polya,G. (1959). Inequalities.Cambridge Univ.Press.

[10] Kiefer,J.(1970).Deviations between the sample quantile process and the sample d.f..In Nonparametric Techniques in Statistical Inference. Proc.Sympos.,Indiana Univ.,Bloomington,Ind.,1969, 299-319,Cambridge Univ.Press,London.

[11] Mangalam,V.(2008).Regression under Lost Association.In preparation.