Apply Gaussian Copula to the Decathlon Data

Author: Luchen Liu

Supervisor: Richard Stridbeck

D-level Essay in Statistics, June 2011

Dalarna University, Sweden

Contents

Apply Gaussian Copula to the Decathlon Data

Luchen Liu

Abstract

This paper intends to present a cutting-edge concept of Copulas, and demonstrate a way of testing copulas by using bootsrtap. Copulas are a currently introduced useful tool for the study of the relationship among several random variables, and bootstrap is a good method in testing whether a given distribution belongs to a distribution family. When doing analysis I used the decathlon data. I illustrated a bootstrap algorithm for testing the existence of Gaussian copulas by using the Cramer-von Mises function, and calculated some association statistics to show the relationship among decathlon data.

Key words: Copulas, Bootsrtap, Goodness-of-fit, Measures of Association

1. Introduction

While watching the summer Olympic games, we care a lot about the athletic games, which include these three kinds of events: running events, jumping events and throwing events. The event decathlon is consisted of ten athletic events of those three kinds. The aim of decathlon games is to test the speed, strength, skill, stamina and endurance of individual athletes, and give full play to the game player's physical limits. So the performance is judged on a points system in each event, not by the position achieved. Sometimes we may question that, if all the ten events are necessary for testing the athletes' capacity?

The question I raised above may lead to a statistical problem of whether there exist obvious dependence among the decathlon data. When we commit tests of independence, we always think about Pearson's (1904) correlation. But when analyzing the decathlon data, it faces some limits, such as it only measures linear dependence, and it also needs the marginal distributions which are hard to obtain. The results of the decathlon events are hard to be compared without the scoring table, because of the difference among the units of measurement of each events. So it's properly for us to transform the results into rank variables, and analysis them with some other measures of associations which can avoid such limits.

The concept of Copula was carried out by Sklar A. (1959) to study the linkage between the high-dimensional distribution functions and their lower dimensional marginal functions. Copulas are functions associating the joint distribution functions to the marginal distribution functions. The study of copulas is a quite recent event, and its applications are also cutting-edge projects.

The main applications are in the fields of Insurance, Finance, and Environment. Damarta S.(2002) explained the application of copulas in the field of Finance, suggesting that copula is usually used to commit risk management through modeling extreme risks by analyzing the tail dependence. Environmentalists also apply copulas to study extreme weathers, such as the recent earthquake and tsunami in Japan. In the field of Insurance, statisticians find it properly to apply copulas in actuarial studies. As insurance companies now prefer to make insurance for the entire family, or husband and wife, parents and children. Actuaries now would like to test the joint mortality of a family more than individuals. It is a good tool to apply copulas when studying the joint mortality, because if we know the copula of the family members' lifetimes, it is easier to know the relationship between the lifetimes. Clayton D. G.(1978) introduced a model for the association in bivariate life tables in his article, and David X. Li(2000) introduced Gaussian copula for solving the problem of default correlation when defining correlation coefficient between the survival times.

So why are we interest in copulas these years? I think it is because it can show the high-dimensional association in one real number without knowing the marginal distributions of the variables. And it can also evaluate the qualitative data beside the quantitative data, and we can prove that the transformation in increasing function of the variables won't change their copula. So it is properly to apply copula and related measures of association with the decathlon data.

In order to measure the association, I have to test if there exist some particular copulas that can associate the joint distribution functions to the marginal distribution functions. To commit goodness-of-fit test, I first think about Cramer-von Mises criteria, as Juan Carlos Pardo-Fernández, et. Al. (2007) and Jian-Jian Ren(2003) all approximated the critical values of tests based on Kolmogorov-Smirnov and Cramer-von Mises type statistics to commit the goodness-of-fit test. And for the incomplete of the data information, I have to do the test without knowing the distribution of Cramer-von Mises statistic. In order to get the distribution, I used the bootstrap method, to approximate the distribution of Cramer-von Mises statistic. B. Efron(1979) introduced the computer-based method of bootstrap, and it is widely used by statisticians. Christan Genest and Bruno Remillard(2007) illustrated a mechanisms implementing a two-level bootstrap for goodness-of-fit test, and proved its validity. This paper illustrated a algorithim for testing Gaussian Copula using Cramer-von Mises statistics based on parametric bootstrap.

This paper firstly introduced copula together with some measures of association, secondly illustrated an algorithm of testing Gaussian copula using bootsrap method, thirdly applied the test on the decathlon data and estimated some copula based measures of association, then made discussion on the result and method in the end.

2. Data Description

In this paper, I analyzed the decathlon data, the denotes and units of measuring is shown in the following table.

Table 1: Denotes and Units of the Decathlon Data

Name of Events	Units of Measurement	Denotes
100 m	Time in seconds	X_1
Long jump	Length in centimeters	X_2
Shot put	Length in meters	X_3
High jump	Hight in centimeters	X_4
400 m	Time in seconds	X_5
110 m h	Time in seconds	X_6
Discus	Length in meters	X_7
Pole vault	Hight in centimeters	X_8
Javelin	Length in meters	X_9
1500 m	Time in seconds	X_{10}

We can see that the data includes four running events measured by time in seconds: 100 meters, 400 meters, 110-meter hurdles and 1500 meters; three jumping events measured by length and height in centimeters: long jump, high jump and pole vault; three throwing events measured by length in meters: shot put, discus and javelin. The differences among the measures of units and the diversity of different sports make it pointless to compute Pearson's coefficients of the original data.

3. Methodology

In this part, I will introduce the main methods used in the study of decathlon data, and some supplement in detail will be in the discussion part.

3.1 Copulas

The study of copulas is quite modern, Sklar(1959) first described copulas in a mathematical and statistical way as functions that can link several one-dimensional distribution functions to form a multivariate distribution function.

3.1.1 Definition of Copulas

Definition 2.1.1. An *n*-dimensional copula is a function $C: [0,1]^n \to [0,1]$, with the following properties:

1. C is grounded, which means for every $u = (u_1, \cdots, u_n) \in I^n$, $C(u) = 0$ if at least one $u_i = 0, i = 1, \cdots, n$;
2. C is *n*-increasing, it means that for every $u \in I^n$ and $v \in I^n$ such that $u < v$, the *C*-volume $V_C([u,v])$ of the box $[u,v]$ is non-negative;
3. $C(1, \cdots, 1, u_i, 1, \cdots, 1) = u_i$ for all $u_i \in [0,1], i = 1, \cdots, n$.

3.1.2 Sklar's Theorem

Theorem 2.1.2. Let *H* be a joint distribution function with margins *F* and *G*. Then there exists a copula *C* such that for all *x*,*y* in *R*,

$$H(x,y) = C(F(x), G(y)). \qquad (2.1.1)$$

If *F* and *G* are continuous, then *C* is unique; otherwise, *C* is uniquely determined on Ran*F*×Ran*G*. Conversely, if *C* is a copula and *F* and *G* are distribution functions, then the function *H* defined by (2.1.1) is a joint distribution function with margins *F* and *G*.

3.1.3 Examples of Copulas

In this section I show two of the most known families of copulas.

3.1.3.1 Bivatiate Pareto Copula

Frees E. W. and Valdez E. A.(1998) introduced this copula when studying the application of copulas for actuary, which is defined by the formula

$$C_\alpha(u,v) = u + v - 1 + [(1-u)^{-\frac{1}{\alpha}} + (1-v)^{-\frac{1}{\alpha}}]^{-\alpha}, \qquad (2.1.2)$$

where α is a parameter $\alpha \in R / \{0\}$.

3.1.3.2 Gaussian Copula

Peter, et. Al. (2000) introduced Gaussian copula, which is derived from the multivariate Gaussian distribution function with mean zero and correlation matrx *C,* which is defined in the equation

$$C(u_1,\cdots,u_n) = \phi_C(\phi^{-1}(u_1),\cdots,\phi^{-1}(u_n))$$

$$= \frac{1}{2\pi^{\frac{n}{2}}|C|^{\frac{1}{2}}} \int_{-\infty}^{\phi^{-1}(u_1)} \cdots \int_{-\infty}^{\phi^{-1}(u_n)} \exp(-\frac{1}{2}(t_1\cdots t_n)C\begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix})dt_1\cdots dt_n . \qquad (2.1.3)$$

3.2 Copula Based Measures of Association

Pearson, K. defined that $D(u,v) = C(u,v) - uv$ is the function as the basis for measures of dependence, as the variables are independent from each other if and only if $D(u,v)$ is identically zero. He introduced the mean square contingency

$$\phi^2 = \int_0^1 \int_0^1 D^2(u,v)dudv . \qquad (2.2.1)$$

The most important difference between the copula based measures of association and the Pearson's correlation is that, the copula based association is only related to the rank variables. For example, the transformation in increasing function of the variables may change the value of the linear correlation coefficients, but the copula based measures of associations will remain the same. Therefore the copula based measures actually comes to its full right when the data is qualitative, although it works just fine for quantitative data as well.

3.2.1 Kendall's Tau

Kendall's tau is defined as the difference between the probabilities of concordance and discordance for two independent pairs (X_1, Y_1) and (X_2, Y_2) each with distribution *H*, that is

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

$$= 4 \iint_{[0,1]^2} C(u,v)dC(u,v) - 1, \qquad (2.2.1.1)$$

where C is the copula associated to (X,Y). And it is generalized to the n-dimensional case as

$$\tau_n(C) = \frac{1}{2^{n-1}-1}(2^n \int_{[0,1]^n} C(u)dC(u)) - 1, \qquad (2.2.1.2)$$

3.2.2 Spearman's Rho

Let (X_1,Y_1), (X_2,Y_2) and (X_3,Y_3) be three independent random vectors, copies of a random vector (X,Y), with a common joint distribution function H, the Spearman's rho association is defined by

$$\rho = 3P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]$$

$$= 12\int_0^1\int_0^1 (C(u,v) - uv)dudv, \qquad (2.2.2.1)$$

where C is the copula associated to (X,Y). And it is generalized to the n-dimensional case as

$$\rho_n(C) = \frac{2^n(n+1)}{2^n - (n+1)} \int_{[0,1]^n} (C(u) - \prod u)du, \qquad (2.2.2.2)$$

3.3 Bootstrap

In statistics, bootstrap is a method that can simplify the methodology by resampling. Bootstrap method is to estimate the statistical values of an estimator by constructing random resamples from the original dataset. Bootstrap method can also be used to implement hypothesis tests. In this section I introduce some measures I used in the algorithm together with the bootstrap method.

3.3.1 Bootstrap Procedure

Efron, B. introduced bootstrap as a method of computing the sampling distribution of a statistics and its various functions without going through the large sample theory calculation. The procedure usually involves resampling the data values. Usually we have a statistic $T = T(X)$ based on the *i.i.d.* sample $X = (X_1, \cdots, X_n)$ from a distribution F, estimates a parameter $\theta = \theta(F)$, the sampling distribution is $R = \sqrt{n}(T - \theta)$. Bootstrap procedure is to commit a resampling Mont Carlo approximation as follows,

1) Pick B much larger than zero, and generate $X_1^*, \cdots, X_B^*$ (each size n) from

$\hat{F} = F_n$, where F_n is the empirical distribution of X,

2) Calculate $R_i^* = \sqrt{n}(T(X_i^*) - \hat{\theta})$ for each B, where $\hat{\theta} = T(X)$,

3) Histogram of bootstrap is given by $R : R_1^*, \cdots, R_B^*$.

3.3.2 Cramer-von Mises Criteria

Anderson, T. W.(1962) introduced the Cramer-von Mises ω^2 criterion for testing a sample $(x_1, \cdots, x_n)$, is from a continuous distribution $F(x)$ is

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dF(x) = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dF_n(x)$$

$$= \sum_{i=1}^{n} [F_n(x_{(i)}) - F(x_{(i)})]^2 , \qquad (2.3.1)$$

where $F_n(x)$ is the empirical distribution function of the sample, which means $F_n(x) = \frac{k}{n}$ if exactly k observations are less than or equal to $x (k = 0, 1, \cdots, n)$, and $x_{(1)}, \cdots, x_{(n)}$ are the ordered observations.

3.3.3 Gaussian Rank Correlation

Hajek, J. and Sidak, Z. (1967) introduced the Gaussian rank correlation to obtain correlation estimators, which is the conventional correlation computed from the Gaussian rank scores(also called Van der Waerden scores). For a bivariate sample $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ the Gaussian rank correlation is in the following form,

$$\hat{\sigma}_{XY} = \frac{\sum_{i=1}^{n} \phi^{-1}(\frac{R(x_i)}{n+1}) \phi^{-1}(\frac{R(y_i)}{n+1})}{\sum_{i=1}^{n} \phi^{-1}(\frac{i}{n+1})^2}, \qquad (2.3.3)$$

where $R(x_i)$ and $R(y_i)$ is the rank of observation x_i and y_i respectively, ϕ^{-1} denotes the quantile function of standard normal distribution. The Gaussian rank score $\phi^{-1}(\frac{R(x_i)}{n+1})$ is obtained by scale the rank of the observation to range between 0 and 1, and plug the scaled rank into the quantile function.

The important feature of the Gaussian rank correlation is that, we can construct

an estimation matrix of multivariate normal distribution by estimating each element of this matrix by the Gaussian rank correlation. For qualitative decathlon data, I chose to estimate the correlation matrix by calculating Gaussian rank correlations.

4. Algorithm

To test that the *d*-dimensional variable *X* is of Gaussian copula, I used parametric bootstrap and illustrate the algorithm as follows. For a proof of the validity of this two-level parametric bootstrap for goodness-of-fit tests see Christan Genest and Bruno Remillard(2007).

I set the null hypothesis H_0 as the copula *C* belongs to Gaussian copula family. The algorithm is to calculate the Cramer-von Mises statistic and the *p*-value using parametric bootstrap, and reject H_0 if the *p*-value is less than 0.05.

While calculating the Cramer-von Mises statistic I used the following equation,

$$\Theta_n = n\int_{-\infty}^{\infty}[C_n(u)-C_\theta(u)]^2 dC_n(u) = \sum_{i=1}^{n}\left\{C_n(\hat{U}_i)-C_\theta(\hat{U}_i)\right\}^2, \qquad (3.1.1)$$

where $C_n(u)$ is the empirical copula distribution function and $C_\theta(u)$ is the theoretical copula distribution function based on the estimated Gaussian rank correlation matrix θ.

In this algorithm I calculate empirical copula frequency function by setting the marginal distributions uniform. Set *m* equals to the number of pairs (x, y) that $x \le x_{(i)}$ and $y \le y_{(i)}$, where $x_{(i)}$ and $y_{(i)}$ are the order statistics of *x* and *y* respectively, and so that the empirical copula frequency function is calculated by

$$C_n(\frac{i}{n},\frac{j}{n}) = \frac{m}{n}, 1 \le i \le n, 1 \le j \le n. \qquad (3.1.2)$$

Two-level parametric bootstrap for testing Gaussian copula

1) Transform the data into rank vectors $R_i = (R_{i1},\cdots,R_{id}), i = 1,\cdots,n$.

2) Calculate the estimated value $\hat{U}_i = \frac{R_i}{n+1}, i = 1,\cdots,n$.

3) Estimate the correlations matrix $\theta = (\hat{\sigma}_{jk})$ by Gaussian rank correlations as

$$\hat{\sigma}_{jk}=\frac{\sum_{i=1}^{n}\phi^{-1}(\frac{R_{ij}}{n+1})\phi^{-1}(\frac{R_{ik}}{n+1})}{\sum_{i=1}^{n}\phi^{-1}(\frac{i}{n+1})^{2}},j=1,\cdots,d,k=1,\cdots,d.$$

4) Pick $m>>n$

a) Generate a ramdom sample $V^{*}_{1},\cdots,V^{*}_{m}$ from joint Gaussian distrbution with $\mu=0$, and correlation matrix $\theta=(\hat{\sigma}_{jk})$.

b) Approximate the theoretical distribution $C_{\theta_n}(u)$ by

$C^{*}_{nv}(u)=\frac{1}{m}\sum_{i=1}^{m}1(V^{*}_{i}\le u)$, and the empirical distribution $C_n(u)$ by

$C_n(u)=\frac{1}{n}\sum_{i=1}^{n}1(\hat{U}_i\le u)$.

c) Compute the Cramer-von Mises statistic

$$\Theta_n=\sum_{i=1}^{n}\left\{C_n(\hat{U}_i)-C^{*}_{nv}(\hat{U}_i)\right\}^2.$$

5) Pick N large, and for $k=1,\cdots,N$:

a) Generate a ramdom sample $V^{*}_{1,k},\cdots,V^{*}_{n,k}$ from joint Gaussian distrbution with $\mu=0$, and correlation matrix $\theta=(\hat{\sigma}_{jk})$.

b) Transform the data into rank vectors $R^{*}_{i,k}=(R^{*}_{i1,k},\cdots,R^{*}_{id,k}),i=1,\cdots,n$, and Calculate the estimated value

$\hat{U}^{*}_{i,k}=\frac{R^{*}_{i,k}}{n+1},i=1,\cdots,n$.

c) Estimate the correlations matrix $\theta^{*}_{k}=(\hat{\sigma}^{*}_{jk})_k$ by Gaussian rank correlations as

$$\hat{\sigma}^{*}_{jk,k}=\frac{\sum_{i=1}^{n}\phi^{-1}(\frac{R^{*}_{ij,k}}{n+1})\phi^{-1}(\frac{R^{*}_{ik,k}}{n+1})}{\sum_{i=1}^{n}\phi^{-1}(\frac{i}{n+1})^{2}},j=1,\cdots,d,k=1,\cdots,d.$$

d) For the same m in step 4:

(i) Generate a ramdom sample $V^{**}_{1,k},\cdots,V^{**}_{n,k}$ from joint

Gaussian distrbution with $\mu = 0$, and correlation matrix $\theta^*_k = (\hat{\sigma}^*_{jk})_k$.

(ii) Approximate the theoretical distribution $C_{\theta^*_n}(u)$ by $C^{**}_{nv,k}(u) = \frac{1}{m}\sum_{i=1}^{m} 1(V^{**}_{i,k} \le u)$, and the empirical distribution $C^*_n(u)$ by $C^*_{n,k}(u) = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{U}^*_{i,k} \le u)$.

(iii) Compute the Cramer-von Mises statistic

$$\Theta^*_{n,k} = \sum_{i=1}^{n} \left\{ C^*_{n,k}(\hat{U}^*_{i,k}) - C^{**}_{nv,k}(\hat{U}^*_{i,k}) \right\}^2 .$$

(iv) Approximate P-value for the test by

$$\frac{1}{N}\sum_{i=1}^{n} 1(\Theta^*_{n,k} > \Theta_n) .$$

5. Result and Discussion

In this part, I will give the results of testing Gaussian copula, and estimating measures of association. And then make discussion.

5.1 The *p*-values for the Tests of Gaussian Copula

The p-value for the test ten-dimentional Gaussian copula is 0.72, which indicates that we can't reject the null hypothesis that the copula is Gaussian at 5% significant level. And we can apply measures of association based on Gaussian copula on the ten-dimentional data.

And the *p*-values for testing two-dimentional Gaussian copulas are as follows,

Table 2: *p*-values for testing two-dimentional Gaussian copulas

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1	0.69	0.51	0.58	0.62	0.71	0.23	0.67	0.55	0.53
X_2	0.69	1	0.48	0.23	0.43	0.21	0.99	0.54	0.42	0.43
X_3	0.51	0.48	1	0.39	0.9	0.29	0.72	0.27	0.27	0.84
X_4	0.58	0.23	0.39	1	0.16	0.44	0.98	0.48	0.97	0.16
X_5	0.62	0.43	0.9	0.16	1	0.3	0.83	0.37	0.19	0.63

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_6	0.71	0.21	0.29	0.44	0.3	1	0.35	0.37	0.19	0.32
X_7	0.23	0.99	0.72	0.98	0.83	0.35	1	0.43	0.61	0.03
X_8	0.67	0.54	0.27	0.48	0.37	0.37	0.43	1	0.48	0.65
X_9	0.55	0.42	0.27	0.97	0.19	0.19	0.61	0.48	1	0.99
X_{10}	0.53	0.43	0.84	0.16	0.63	0.32	0.03	0.65	0.99	1

we can see from the table that, we can only reject the null hypothesis of Gaussian copula at 95% significant level for the group (X_7, X_{10}), for the *p*-value is 0.03. The copula of all the other groups are accept to belong to the Gaussian family.

As the algorithm really takes time to run in the programme R, I picked several groups of data that I interested in to check the higher dimensional Gaussian copulas. The results are as follows:

Table 3: *p*-values for testing Gaussian copulas

Variables tested	*p*-value	Variables tested	*p*-value
$X_1\ X_5\ X_6$	0.73	$X_1\ X_2\ X_4\ X_6$	0.24
$X_1\ X_5\ X_6\ X_{10}$	0.95	$X_4\ X_6\ X_8$	0.33
$X_1\ X_4\ X_6$	0.39	$X_2\ X_4\ X_8$	0.45
$X_1\ X_2\ X_4$	0.16	$X_3\ X_7\ X_9$	0.27
$X_2\ X_4\ X_6$	0.78	$X_3\ X_7\ X_8\ X_9$	0.84

as we can see from the table, none of the results in the picked group is to reject the null hypothesis of Gaussian copula at 95% significant level. So that, we can calculate the Gaussian copula based measures of association of these groups.

5.2 The Measures of Associations

I calculated the Kendall's tau and Spearman's rho of the groups that passed the hypothesis tests of Gaussian copula, and the results of Spearman's rho are as follows,

Table 4: Spearman's Rho of Two-dimensional Groups

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000	-0.406	-0.059	-0.011	0.566	0.472	-0.084	-0.125	-0.041	0.026

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_2	-0.406	1.000	0.189	0.247	-0.325	-0.421	0.207	0.217	0.21	-0.141
X_3	-0.059	0.189	1.000	0.124	-0.004	-0.209	0.684	0.243	0.388	0.106
X_4	-0.011	0.247	0.124	1.000	-0.028	-0.221	0.087	0.09	0.023	-0.052
X_5	0.566	-0.325	-0.004	-0.028	1.000	0.377	-0.022	-0.125	-0.005	0.442
X_6	0.472	-0.421	-0.209	-0.221	0.377	1.000	-0.225	-0.252	-0.125	0.077
X_7	-0.084	0.207	0.684	0.087	-0.022	-0.225	1.000	0.251	0.41	-
X_8	-0.125	0.217	0.243	0.09	-0.125	-0.252	0.251	1.000	0.151	-0.14
X_9	-0.041	0.21	0.388	0.023	-0.005	-0.125	0.41	0.151	1.000	-0.008
X_{10}	0.026	-0.141	0.106	-0.052	0.442	0.077	-	-0.14	-0.008	1.000

where we can see from the table that some pair of data shows association of dependence larger than 0.3, such as (X_1, X_2), (X_1, X_5), (X_1, X_6), (X_5, X_6), (X_2, X_6), (X_3, X_7), (X_3, X_9), (X_5, X_{10}), (X_7, X_9), which indicates some extent of dependence.

We can see from the table that some of the associations are negative, such as the one of group (X_1, X_2), that is because the two events are standing for 100 m and long jump separately. For the running events, the less seconds you take, the better you perform; but for the throwing events, the more centimeters you throw, the better you perform. I think that's why the association here is negative.

I calculated the Spearman's rho of groups of higher dimensions which have passed the hypothesis tests of Gaussian copula, the results are in the following form.

Table 5: Spearman's Rho of Some Higher Dimensional Groups

Variables	Spearman's rho	Variables	Spearman's rho
$X_1\ X_5\ X_6$	0.706	$X_1\ X_2\ X_4\ X_6$	-0.071
$X_1\ X_5\ X_6\ X_{10}$	0.610	$X_4\ X_6\ X_8$	-0.192
$X_1\ X_4\ X_6$	0.121	$X_2\ X_4\ X_8$	0.277
$X_1\ X_2\ X_4$	-0.088	$X_3\ X_7\ X_9$	0.743
$X_2\ X_4\ X_6$	-0.199	$X_3\ X_7\ X_8\ X_9$	0.552

As we can see from the table that the group (X_1, X_5, X_6), (X_1, X_5, X_6, X_{10}), (X_3, X_7, X_9), (X_3, X_7, X_8, X_9) are of association of dependence larger than 0.5, which is showing evidence dependence.

And we can see from the association of Spearman's Rho that the group (X_1, X_5, X_6, X_{10}), which involves four running events, and group (X_3, X_7, X_9), which involves three throwing events, and their two-dimensional sub-groups are all showing highly dependence.

5.3 Discussion

In this section, I made conclusions on the results, and illustrated some supplement of calculating the measures of associations.

5.3.1 Conclusions from the Results

From the results of testing Gaussian copulas and estimating of Spearman's rho, we can access the following conclusion:

The ten dimensional decathlon data and some of its sub-groups can be described by Gaussian copula;

All the running events are showing evidence of dependence, within the group, the sprint events like 100 m, 400 m, 110 m h are showing more dependence with each other, and the long-distance running 1500 m is only showing dependence with 400 m;

All the throwing events are showing evidence of dependence;

There are some groups such as 100 m and long jump showing dependence. I think it is because if the player can perform great in sprint, he is more likely to gain high speed when approaching before jump, and thus he is more likely to jump further than the other players;

Since the results showing the throwing events, the running events and their two-dimensional sub-groups are all showing highly dependence, I would suggest maybe not all the ten events are necessary for testing the athletes' capacity in the decathlon games, or we can make some redundancies in the ten events.

5.3.2 Algorithm for Calculating Measures of Associations

While calculating the measures of associations based on copulas, I used the Mont Carlo methods to compute the integral.

For estimating Spearman's rho in equation (2.2.2.2) for the tested *d*-dimensional variable *X* of Gaussian copula , the algorithm is elaborated as follows:

Mont Carlo Method for Calculating the Spearman's Rho Based on Gaussian Copula

1) Pick *B* a positive integer much larger than zero, and for each $i = 1, \cdots, B$,

a) Generate random numbers $u_i = (u_{1,i}, \cdots, u_{d,i})$ where $u_{k,i}, k = 1, \cdots, d$ is

uniformly distributed on the interval $[0,1]$,

b) Use the correlations matrix $\theta = (\hat{\sigma}_{jk})$ estimated former by Gaussian rank correlations as the correlation matrix, where the Gaussian ranks are transformed from the *d*-dimensional variable *X*,

c) Calculate $C_i = C_\theta(u_i) - \prod_{k=1}^{n} u_{k,i}$ for each *i*, where $C_\theta(u)$ is the Gaussian copula with correlation matrix θ,

2) The estimated Spearman's rho based on Gaussian copula is

$$\hat{\rho}_B(C) = \frac{2^d(n+1)}{2^d-(n+1)} \bullet \frac{C_i}{B}.$$

Although it seems the same way to calculate the Kendall's tau, there really exist a problem when considering the computing of $\int_{[0,1]^n} C(u)dC(u)$. We would like to put $dC(u)$ into the integral as $c(u)du$ where $c(u)$ denotes the Gaussian copula density in this paper. Peter, et. Al. (2000) gives the Gaussian copula density calculated in the following equation,

$$c_\phi(u|C) = |C|^{-\frac{1}{2}} \exp\{-\frac{1}{2} q^T \left(I_n - C^{-1}\right) q\}, \qquad (4.3.1)$$

where $q = (q_1, \cdots, q_n)^T$ with $q_i = \phi^{-1}(u_i), i = 1, \cdots, n$ denoting the normal quantiles.

The calculation of Kendall's tau is always resulting huge values of association in higher dimensional data, so I turned to Spearman's rho while analyzing association of three and four dimensional data. But I still listed the Kendall's tau in the appendix.

5.3.3 Two-dimensional Kendall's Tau and Spearman's Rho not Basing on Copulas

As there are $\binom{n}{2}$ distinct pairs of each two-dimensional data in the sample, and each pair is either concordant or discordant—let *c* denote the number of concordant pairs and *d* the number of discordant pairs. Then Kendall's tau for the sample is defined as

$$\tau = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{2}} \qquad (4.2.1)$$

without knowing the copula. And for Spearman's rho can be calculated as

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \qquad (4.2.2)$$

where d_i denotes the difference between the ranks of each pair of observations.

I calculated the measures of associations in two-dimensional twice by both equation (4.2.1) (4.2.2) and the copula form as their exist rejection in the testing of two dimensional copulas. If the two-dimensional copulas are Gaussian, then there shouldn't be large difference between the association calculated. And I listed the results in the appendix.

5.3.4 Some Inadequate in Testing the Gaussian Copulas

As Christian Genest andBruno Remillard suggest, to illustrate the validity of the parametric bootstrap for Cramer-von Mises statistic, if the size of random sample is n=500, the null hypothesis was test at 5% level, we should take N=1000 and m=100000 while committing the two-level bootstrap procedure. I failed to take m and N such large, as it will be too much time consuming. And I doubt if that is why so much Gaussian copulas are accepted. I this problem would be solved in further study.

References

[1] Pearson, K., 1904. *On the Theory of Contingency and its Relation to Association and Normal Correlation*. 1st ed. London: Drapers' Company Research Memoirs, Biometric Series.

[2] Nelsen, R. 2006. *An Introduction to Copulas*. 2nd ed. New York: Springer Science+Business Media, Inc.

[3] Sklar, A., 1973. Random Variables, Joint Distribution Functions and Copulas, *Kybernetika*, 9, pp.449-460.

[4] Sklar. A., 1959. *Fonctions de repartition ' n dimensions et leurs marges*. 8th ed. Paris: Publ. Inst. Statist. Univ.

[5] Efron B., 1979. Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7(1), pp.1-26.

[6] Clayton D.G., 1978. A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, 65, pp.141-151.

[7] Damarta S., 2002. Extreme Value Theory and Copulas. Master's thesis, Department of Mathematics, ETH Zurich, Switzerland.

[8] Juan Carlos Pardo-Fernández, Ingrid van Keilegom, Wenceslao Gonzá

lez-Manteiga, 2007. Goodness-of-Fit Tests for Parametric Models in Censored Regression. *The Canadian Journal of Statistics*, 35(2), pp.249-264.

[9] Jian-Jian Ren, 2003. Goodness of Fit Tests with Interval Censored Data. *Scandinavian Journal of Statistics*, 30(1), pp.211-226.

[10] Christian Genest, Bruno Remillard, 2008. Validity of the Parametric Bootstrap for Goodness-of-fit Testing in Semiparametric Models. *Probabilites et Statistiques*, 44(6), pp.1096-1127.

[11] Anderson, T. W., 1962. *On the Distribution of the Two-sample Cramer-von Mises Criterion*. Stanford University, 12 April 1962. The Office of Naval Research: Stanford University.

[12] Hajek, J. and Sidak, Z., 1967. *Theory of Rank Tests*. 1st ed. New York: Academic Press.

Appendix

Table 6: Kendall's Tau of Two-dimensional Groups Based on Copulas

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000	-0.845	-0.591	-0.301	24.86	0.343	-0.402	-0.578	-0.531	-0.133
X_2	-0.845	1.000	-0.264	0.172	-0.8	-0.847	1.532	-0.074	-0.032	-0.673
X_3	-0.591	-0.264	1.000	-0.031	-0.371	-0.717	4.076	0.807	5.929	0.206
X_4	-0.301	0.172	-0.031	1.000	-0.053	-0.718	-0.246	-0.283	-0.189	-0.553
X_5	24.86	-0.8	-0.371	-0.053	1.000	-0.076	-0.601	-0.661	-0.488	0.146
X_6	0.343	-0.847	-0.717	-0.718	-0.076	1.000	-0.742	-0.767	-0.658	0.084
X_7	-0.402	1.532	4.076	-0.246	-0.601	-0.742	1.000	0.427	0.416	-
X_8	-0.578	-0.074	0.807	-0.283	-0.661	-0.767	0.427	1.000	0.372	-0.603
X_9	-0.531	-0.032	5.929	-0.189	-0.488	-0.658	0.416	0.372	1.000	-0.311
X_{10}	-0.133	-0.673	0.206	-0.553	0.146	0.084	-	-0.603	-0.311	1.000

Table 7: Spearman's Rho of Two-dimensional Groups not Based on Copulas

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000	-0.397	-0.053	-0.014	0.568	0.475	-0.078	-0.113	-0.036	-0.007
X_2	-0.397	1.000	0.188	0.269	-0.328	-0.425	0.199	0.196	0.202	-0.123
X_3	-0.053	0.188	1.000	0.124	0.004	-0.207	0.673	0.267	0.379	0.092
X_4	-0.014	0.269	0.124	1.000	-0.047	-0.204	0.099	0.118	0.012	-0.054
X_5	0.568	-0.328	0.004	-0.047	1.000	0.391	-0.006	-0.116	0.008	0.443
X_6	0.475	-0.425	-0.207	-0.204	0.391	1.000	-0.236	-0.247	-0.107	0.088
X_7	-0.078	0.199	0.673	0.099	-0.006	-0.236	1.000	0.273	0.419	0.061
X_8	-0.113	0.196	0.267	0.118	-0.116	-0.247	0.273	1.000	0.137	-0.150
X_9	-0.036	0.202	0.379	0.012	0.008	-0.107	0.419	0.137	1.000	-0.002
X_{10}	-0.007	-0.123	0.092	-0.054	0.443	0.088	0.061	-0.150	-0.002	1.000

Table 8: Kendall's Tau of Two-dimensional Groups not Based on Copulas

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000	-0.276	-0.035	-0.008	0.403	0.329	-0.05	-0.078	-0.023	-0.007
X_2	-0.276	1.000	0.127	0.185	-0.221	-0.293	0.132	0.138	0.137	-0.083
X_3	-0.035	0.127	1.000	0.087	0.003	-0.137	0.495	0.184	0.26	0.061
X_4	-0.008	0.185	0.087	1.000	-0.03	-0.14	0.069	0.083	0.008	-0.039
X_5	0.403	-0.221	0.003	-0.03	1.000	0.267	-0.004	-0.08	0.007	0.306
X_6	0.329	-0.293	-0.137	-0.14	0.267	1.000	-0.157	-0.174	-0.072	0.057
X_7	-0.05	0.132	0.495	0.069	-0.004	-0.157	1.000	0.193	0.288	0.042
X_8	-0.078	0.138	0.184	0.083	-0.08	-0.174	0.193	1.000	0.094	-0.102
X_9	-0.023	0.137	0.26	0.008	0.007	-0.072	0.288	0.094	1.000	-0.001
X_{10}	-0.007	-0.083	0.061	-0.039	0.306	0.057	0.042	-0.102	-0.001	1.000