Modeling Directional Dependence

using Order Statistics and Concomitants

Wahid Tawsif Ratul

Advisor : Engin Sungur

University of Minnesota, Morris

Table of contents

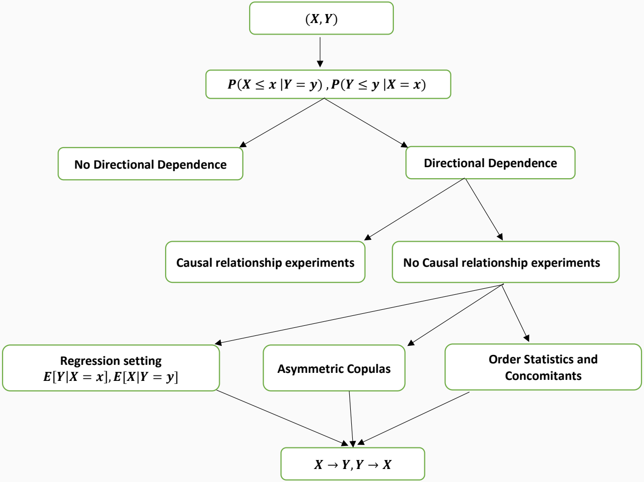
Introduction

- The main goal of this research is to carry out a simulation study that uses the fundaments of order statistics and concomitants to exhibit bivariate directional dependence property by sampling random numbers from probability distribution functions.

Directional Dependence

- **Directional dependence** is a method to determine the likely causal direction of effect between two variables. It also involves studying changes in distributional behavior of one variable initiated not necessarily caused by another random variable.
- In a bivariate setting, if the large and small values of the random variables tend to occur together then the **direction of dependence** is positive; and if the large values of one variable tend to occur with small values of the other then the **direction of dependence** is negative, Nelsen and Úbeda-Flores (2011).
- It is different than studying causality which can only be set up through carefully designed experiments.

The Flow Chart



Order Statistics and Concomitants

One possible way to understand the changes in distribution of one variable initiated by another is to study the concomitants of order statistics.

- Let (X, Y) be a continuous random vector with joint cumulative distribution function $F(x, y)$, and marginals $F_X(x)$ and $F_Y(y)$.
- Consider random sample of n observations from the distribution of X and Y, and let $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ be the order statistics of X.
- The value of Y associated with the $\mathbf{X_{i:n}}$ is called the concomitant of the i^{th} ordered statistics of X and will be represented by $\mathbf{Y_{[i:n]}}$. Similarly, $\mathbf{X_{[i:n]}}$ is the X-concomitant of $\mathbf{Y_{i:n}}$.

Order Statistics and Directional Dependence

Directional dependence is a phenomenon that can be demonstrated in different ways such as the following :

- Dodge and Rousson (2000) and Muddapur (2005) have suggested an approach to model directional dependence using the direction of the regression line. Their approach is based on the use of coefficients of skewness of the two random variables. The response variable will always have less skewness than the explanatory variable.
- Sungur and Celebioglu (2011) discussed generating a class of conditional copulas that exhibits directional dependence property by introducing direction covariate and set up the connection between directed graphs.
- In this research we will look at the components $\mathbf{Y}_{[\mathbf{i}:\mathbf{n}]}$ ($\mathbf{X}_{[\mathbf{i}:\mathbf{n}]}$), ordering on $Y(X)$ introduced by ordering in $X(Y)$, and compare it with $\mathbf{Y}_{(\mathbf{i}:\mathbf{n})}$ ($\mathbf{X}_{(\mathbf{i}:\mathbf{n})}$).

Distribution Cycle

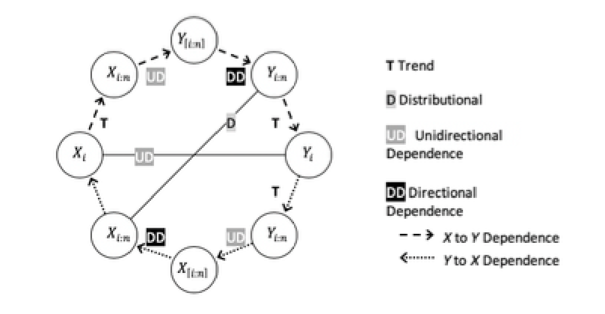


Figure: 2 Distribution Cycle.The key random variables and their roles on understanding marginal and joint behaviors of (X, Y).

Dependence Cycle

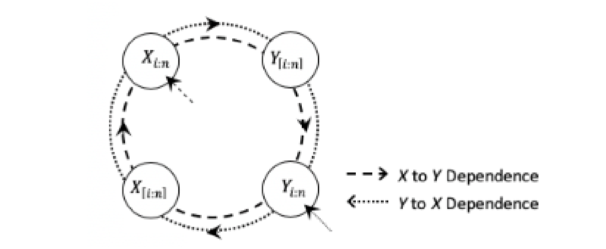


Figure: 3 Dependence Cycle. The key random variables and their roles on understanding dependence structure of (X, Y).

Quantitative Measurement of Directional Dependence

- Our objective is to focus on the key components from the Dependence Cycle: $\mathbf{X}_{(\mathbf{i:n})}$,$\mathbf{X}_{[\mathbf{i:n}]}$,$\mathbf{Y}_{(\mathbf{i:n})}$,$\mathbf{Y}_{[\mathbf{i:n}]}$ as r_{XY} is explained as the sample linear correlation coefficient. It measures the strength and direction of linear association between two quantitative variables and are the following:

$$r_{XY} = \frac{1}{(n-1)S_X S_Y} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \tag{1}$$

$$r_{X \to Y} = r_{Y_{i:n} Y_{[i:n]}} = \frac{1}{(n-1)S_Y^2} \sum_{i=1}^{n} (y_{i:n} - \bar{y_{[i:n]}})(x_i - \bar{x}) \tag{2}$$

$$r_{Y \to X} = r_{X_{i:n} X_{[i:n]}} = \frac{1}{(n-1)S_X^2} \sum_{i=1}^{n} (X_{i:n} - \bar{X_{[i:n]}})(x_i - \bar{x}) \tag{3}$$

- Normal Distribution is a function that represents the distribution of many random variables as a symmetrical bell-shaped graph.
- The Normal Distribution is often denoted by $N(\mu, \sigma)$. In this study, we considered $\mu = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and $\sigma = \begin{pmatrix} 9 & 6 \\ 6 & 16 \end{pmatrix}$ and randomly generated $n = 100$ observations for (X, Y).

X	Y	$X_{i:n}$	$Y_{[i:n]}$	$Y_{i:n}$	$X_{[i:n]}$
6.04	6.96	-5.74	3.22	-3.45	-7.98
-1.55	2.54	-4.70	-0.96	-0.65	-4.76
-1.35	2.67	-3.93	-1.43	2.87	-4.40
0.41	2.18	-3.12	-2.60	3.15	-3.77
7.87	1.63	-2.82	2.71	4.41	-3.23
-0.24	0.28	-2.51	1.81	8.31	-3.15

The Mystery of Data with Normal Distribution

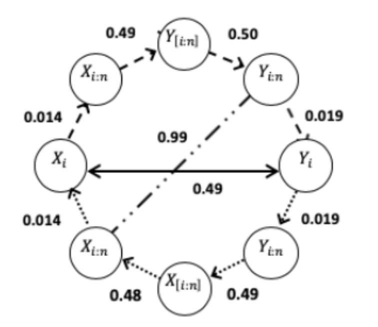


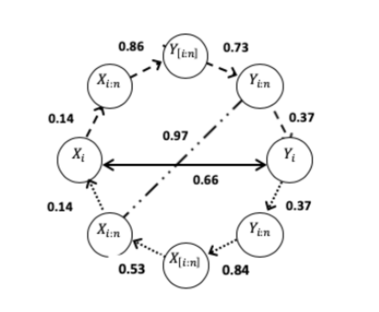
Figure: 4 The key random variables and their roles on understanding dependence structure for normally distributed (X, Y).

- We will say that $\mathrm{X} \rightarrow \mathrm{Y}$ is the dominant direction of dependence in terms of ordering, if and only if $|\mathbf{r}_{\mathrm{Y}_{[i:n]}\mathrm{Y}_{i:n}}| < |\mathbf{r}_{\mathrm{X}_{[i:n]}\mathrm{X}_{i:n}}|$. In this case it is very challenging to conclude variables (X, Y) have any directional dependence.

The Mystery of Data

- The continuous uniform distribution or rectangular distribution is a family of symmetric probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable.
- The support is defined by the two parameters, a and b, which are its minimum and maximum values. The distribution is often abbreviated $U(a, b)$. In this study, we randomly generated $n = 100$ observations where $a = -300$ and $b = 300$ for (X, Y).

X	Y	$X_{i:n}$	$Y_{[i:n]}$	$Y_{i:n}$	$X_{[i:n]}$
140.5	-27.5	19.9	5.4	156.5	-191.2
203.0	-130.7	20.1	1.1	173.6	-186.9
95.2	-7.2	21.2	8.4	188.5	-186.3
62.3	10.5	33.4	-4.2	192.0	-183.6
206.2	-133.7	41.6	6.6	230.7	-178.3
199.1	-100.9	46.1	8.3	271.1	-160.8



- We will say that Y $\rightarrow$ X is the dominant direction of dependence in terms of ordering, if and only if $|\mathbf{r}_{\mathbf{Y}_{[i:n]}\mathbf{Y}_{i:n}}| > |\mathbf{r}_{\mathbf{X}_{[i:n]}\mathbf{X}_{i:n}}|$. In this case we (X, Y) have directional dependence property with a very strong evidence that $Y(X)$ causes a substantial change in marginal distribution as compared to $X(Y)$ in other words X $\rightarrow$ Y.

Introduction to the sampling algorithm

- Let us denote the transformed random pair with (X, Y) are the linear combinations of the independent uniform pair (U, V). Along with the positive semi-definite matrix of the form as demonstrated by Sungur and Celebioglu (2011): $\begin{pmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{pmatrix}$ such that $\boldsymbol{\alpha_1 > \alpha_2, \beta_1 \leq -\beta_2, \alpha_1, \alpha_2, \beta_1 > 0, \beta_2 < 0}$ have a common bivariate distribution and (X, Y) that can be expressed in the form:

$$X_i = \alpha_1 * U_i + \alpha_2 * V_i \tag{4}$$

$$Y_i = \beta_1 * U_i + \beta_2 * V_i \tag{5}$$

Extension to the sampling algorithm

- The model parameters $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}$ are randomly simulated using uniform distribution abiding by the conditions mentioned earlier in the making of the linear models (4) and (5) and are in the form:

$$\begin{pmatrix} U_1 & V_1 \\ U_2 & V_2 \\ . & . \\ . & . \\ . & . \\ U_n & V_n \end{pmatrix} * \begin{pmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{pmatrix} = \begin{pmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ . & . \\ . & . \\ . & . \\ X_n & Y_n \end{pmatrix} \tag{6}$$

- Using the model (6) we generated 100 combinations of parameters α_1, α_2, β_1, β_2 that satisfied the conditions to generate 100 list of data frames for (X_i, Y_i). Each set of (X_i, Y_i) were used to incorporate the key components of order statistics and concomitants $(\boldsymbol{X_{i:n}}, \boldsymbol{X_{[i:n]}}, \boldsymbol{Y_{i:n}}, \boldsymbol{Y_{[i:n]}})$.

Continuation of Extension to the sampling algorithm

- The key elements required to understand the dependence structure of (X, Y) are the $\boldsymbol{r_{x \to y}}$ and $\boldsymbol{r_{y \to x}}$ which are basically the correlation of $\boldsymbol{r_{y_{i:n} y_{[i:n]}}}$ and $\boldsymbol{r_{x_{i:n} x_{[i:n]}}}$. Strength of Directional Dependence can be computed as $\boldsymbol{\delta r = r_{x \to y} - r_{y \to x}}$.
- Our final objective was to find the different combinations of parameters α_1, α_2, β_1, β_2 that would exhibit the strongest directional dependence and how it behaves as we fluctuate the parameters satisfying the conditions.

Teacher vs Student

```
cvect<-c(0,0,1,0,0,0,20)
elhat<-sum(cvect*coef(sec.mod))
stdelhat<-sqrt(t(cvect) %*% vcov(sec.mod) %*% cvect)
cbounds<-c(elhat - 1.96*stdelhat,elhat+1.96*stdelhat)
c(elhat,stdelhat,cbounds)
```

Figure: 6 Sensei Jon’s code

````
```{r}
# Creating Random U vector
randomU<-runif(100,min=1,max=10000)
randomV<-runif(100,min=1,max=10000)
Sim3v=cbind(c(randomU),c(randomV))
makeArray=function(n){
  swap=function(array){
    tmp<-array[1]
    array[1]<-array[2]
    array[2]<-tmp
    array
  }
  alpha=runif(2,2,50)
  if(alpha[1]<alpha[2]){alpha<-swap(alpha)}
  beta=runif(2,2,50)
  beta[2]=beta[2]*-1
  if(beta[1]>-beta[2]){beta<-swap(beta)}
  MAT=cbind(alpha,beta)
  Sim3=Sim3v%*%MAT
  colnames(Sim3)<-c("x","y")
  Sim3OX<-Sim3[order( Sim3[,1], Sim3[,2]),]
  colnames(Sim3OX)<-c("x(i:n)","y[i:n]")
  Sim3OY<-Sim3[order( Sim3[,2], Sim3[,1]),]
  colnames(Sim3OY)<-c("x[i:n]","y(i:n)")
  univariateframe=data.frame(Sim3,Sim3OX,Sim3OY)
  DeltaAlpha=(round(MAT[[1,1]] - MAT[[2,1]],digits=2))
  DeltaBeta=(round(MAT[[2,1]]-MAT[[2,2]],digits = 2))
  Pxy=round(cor(univariateframe$x.i.n.,univariateframe$x.i.n..1,method="pearson"),digits = 2)
  Pyx=round(cor(univariateframe$y.i.n.,univariateframe$y.i.n..1,method="pearson"),digits=2)
  DeltaCorrelation=round(abs(Pxy) -abs(Pyx),digits=2)
  return(data.frame(DeltaAlpha=DeltaAlpha, DeltaBeta=DeltaBeta,Pxy=Pxy,Pyx=Pyx, DeltaCorrelation=Pxy-Pyx))
}
results<-lapply(1:100,makeArray)
library(plyr)
Rtable=ldply(results,data.frame)
```
````

Figure: 7 Gakusei Wahid’s code

Model Parameters table

- Different range of combinations of model parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ were generated using the uniform distribution that helped us create a list of 100 data frames consisting of columns of $(X_{i:n}, X_{[i:n]}, Y_{i:n}, Y_{[i:n]})$. For every list of data frames, we measured the $\delta\alpha = \alpha_1 - \alpha_2$, $\delta\beta = \beta_1 - \beta_2$ and $\delta r = r_{x \to y} - r_{y \to x}$ as illustrated in the following table:

$\delta\alpha$	$\delta\beta$	$r_{X \to Y}$	$r_{Y \to X}$	δr
4.06	29.12	-0.31	-0.30	-0.01
22.4	51.28	0.37	0.31	0.06
12.51	16.75	0.05	0.06	-0.01
7.18	30.48	0.01	0.00	0.01
0.26	28.08	0.23	0.22	0.01
10.43	20.58	0.75	0.70	0.05

3D graphical displays

```
library(akima)
dens <- interp(Rtable$DeltaAlpha, Rtable$DeltaBeta, Rtable$DeltaCorrelation,duplicate="mean")
fancyplot=plot_ly(x=dens$x,y=dens$y,z=dens$z)%>%add_surface()
fancyplot
```

Figure: 8 3D surface plot code

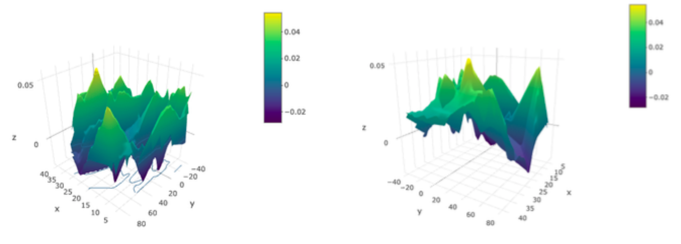


Figure: 9 3D surface plot showing Directional Dependence where X=$\delta\alpha$, Y=$\delta\beta$, Z=δr

- The plot exhibits the strongest directional dependence at different values of the model parameters. The strongest directional dependence is shown at $\delta\alpha$ =4.32, $\delta\beta$= -11.98, δr=0.06.

Model parameters and Directional Dependence

- The following is a scatterplot that will help us explain any relationship that exists (if any):

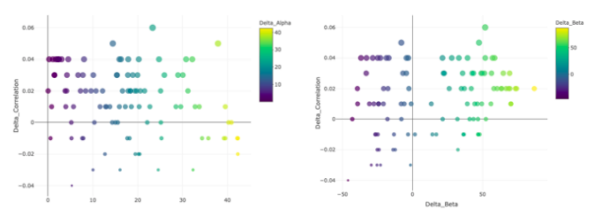


Figure: 10 Scatterplots explaining the relationship between $\delta\alpha$ and $\delta\beta$ on the horizontal axes and δr on the vertical axes.

- The scatterplot explains that $\boldsymbol{\delta\alpha}$ and $\boldsymbol{\delta\beta}$ have no concrete relationship with $\boldsymbol{\delta r}$. There are very few specific cases where the directional dependence is strongest when $\boldsymbol{\delta\alpha}$=22.4 and $\boldsymbol{\delta\beta}$=51.28.

Directional Dependence with strong correlation

- The last stage of the research is to look into cases when (U, V) exhibits characteristics of stronger correlation $0.7 < r_{xy} < 1$ and no correlation $r_{xy}=0$ between variables (X, Y). The following scatterplots will explain if any relationship exists when the randomly generated (X, Y) are strongly correlated $r_{xy}=0.9$.

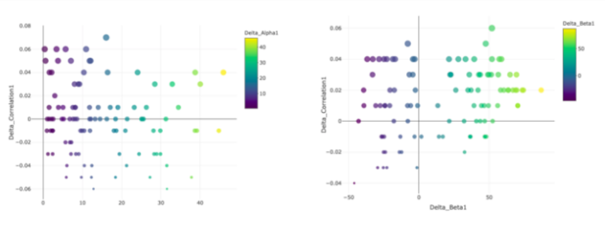


Figure: 11 Scatterplots explaining the relationship between $\delta\alpha$ and $\delta\beta$ on the horizontal axes and δr on the vertical axes when (X, Y) are strongly correlated.

Directional Dependence with no correlation

- The following scatterplots will explain if any relationship exists when the randomly generated (X, Y) are not correlated r_{xy}=0.

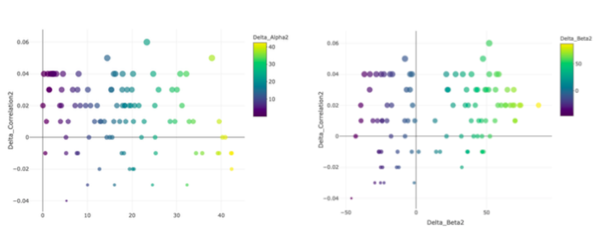


Figure: 12 Scatterplots explaining the relationship between $\delta\alpha$ and $\delta\beta$ on the horizontal axes and δr on the vertical axes when (X, Y) are not correlated.

- The scatterplots explains that correlation behavior between (X, Y) has no effect on the directional dependence. The random simulation suggests directional dependence are very similar at different combinations of model parameters when (X, Y) are both strongly and weakly correlated.

Future Work

- The sample algorithm introduced by Sungur and Celebioglu (2011) and in this study possess directional dependence property and the use of directional variable in discrete and continuous forms will give researches a chance to create more general dependence models that satisfies the conditions of the model parameters that are uniformly distributed.
- Increasing the dimensions of model parameters, $\alpha_1, \alpha_2, \beta_1, \beta_2$, and generating (U, V) with different probability distribution will be a research interest in future.
- Moreover, it will be very interesting to see how the dependence structure behaves if α_1, α_2 are kept constant with varying β_1, β_2 and vice versa.

References

- Sungur, Engin A., and Salih ÇelebİoĞlu. "Copulas with Directional Dependence Property." Gazi University Journal of Science, vol. 24, no. 3, July 2011, pp. 415–424.
- Nelsen, Roger, and Manuel Úbeda-Flores. "Directional Dependence in Multivariate Distributions." Annals of the Institute of Statistical Mathematics, vol. 64, no. 3, June 2012, pp. 677–685.
- von Eye, Alexander, and Richard P. DeShon. "Directional Dependence in Developmental Research." International Journal of Behavioral Development, vol. 36, no. 4, July 2012, pp. 303–312.
- Muddapur, M. V. "A Remark on the Paper 'A Note on Directional Dependence in Regression Setting' By Engin A. Sungur." Communications in Statistics - Theory and Methods, vol. 37, no. 3, 2008, pp. 386–387
- Schechtman, E., and S. Yitzhaki. "On the Proper Bounds of the Gini Correlation." Economics Letters 63.2 (2002): 133–138. Economics Letters. Web.

The Golden Circle

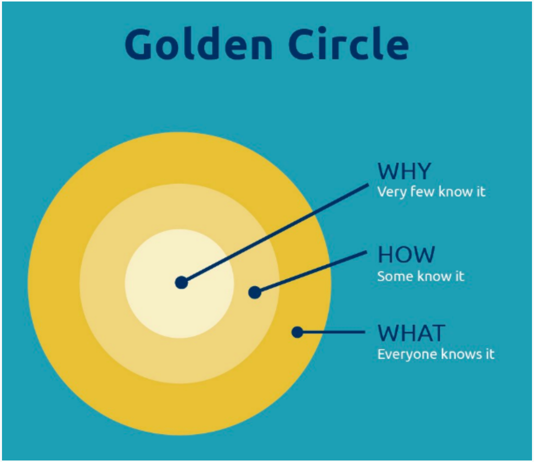


Figure: 13 The Golden Circle.

WHY I did what I did?

- To specialize in some sort of field of statistics for Masters / Ph.D.
- To continue Engin's tradition of Directional Dependence.
- To lay the foundation for fellow peers to think about their work.
- To show statistics is not **always** about p-values.
- To support the notion that **causation** $\neq$ **correlation**.

Memories

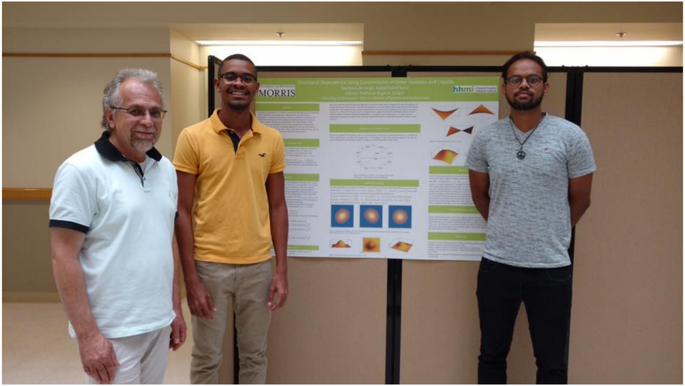


Figure: 14 HHMI 2017 Poster Presentation featuring Engin, Rocherno and Wahid.

Memories

Figure: 15 ASA Fall 2017 Conference featuring Jong-Min, Rocherno and Wahid.

Figure: 16 Michael Cagle, Dongmin Kim, Francisco Montanez.

Favorite classes

Favorite classes :

- Introduction to Econometrics
- Introduction to Data Science
- Interpersonal Skills

Least favorite classes :

- Probability and Stochastic Processes
- Mathematical Economics
- Mathematical Statistics

Future Plans

Future Plans :

- Start my own Non-Profit
- Become a Monk
- Find my **WHY**

Faculty

I would like to express my **eternal gratitude** to the following:

- Engin A. Sungur
- Stephen Burks
- Satis Devkota
- Peter Dolan
- The very BEST Jon Anderson

Questions

Any thoughts, comments, concerns?