Bias of Correlation Under Mixture Models

Russell Land

under the direction of Stephen Carden

Georgia Southern University

March 2, 2021

Introduction

Consider measuring the correlation between two variables, but the sample is contaminated with observations from an unintended population. We will answer the following questions:

- As the proportion of contamination p changes, how will this bias the correlation?
- What form does the bias, as a function of p, take?
- Can we characterize the behavior of the bias in some way?
- Do different measures of correlation behave in different ways?

Introduction

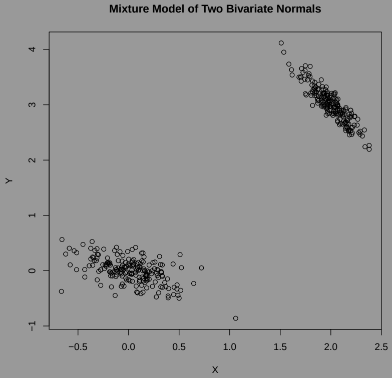
Consider the following example. A statistician collects data, where some is valid and some is unwanted, contaminated data. The two populations are bivariate normal distributions with parameters,

$$\mu_1 = (0,0)\,, \quad \Sigma_1 = \begin{pmatrix} 0.07 & -0.03 \\ -0.03 & 0.05 \end{pmatrix}, \quad \rho_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$$\mu_2 = (2,3)\,, \quad \Sigma_2 = \begin{pmatrix} 0.03 & -0.05 \\ -0.05 & 0.1 \end{pmatrix}, \quad \rho_2 = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$$

Introduction

A simulation will give us a feel for the data...



Let's finish this example in R...

Outline

1. Introduction

2. Mixture Models

3. Rank Correlation
   - Concordance & Discordance
   - Rank Correlation Methods
   - Copulas

4. Bias of Correlation Under Mixtures
   - Bias of Mixtures
   - The Marshall-Olkin Distribution

5. Conclusion

Mixture Models

Mixture models can arise from sampling unwanted populations. Consider a statistician collecting grade point averages (GPA) from undergraduate students on the Georgia Southern University campus. Without vetting, graduate students may enter the sample unintentionally. When unwanted observations are introduced into the sample, it becomes a mixture of valid data and contaminated data. For example, if graduate students have higher GPA's, this will affect any statistics that are calculated from the sample.

Mixture Models

Definition (Mixture Model)

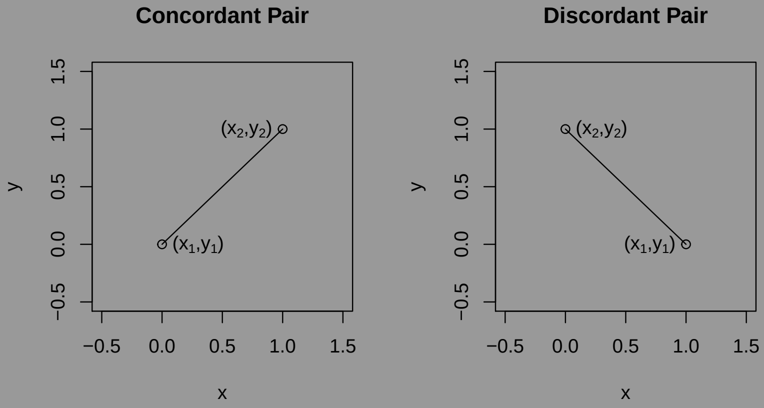
Let $\vec{V} = (V_1, V_2)$ and $\vec{C} = (C_1, C_2)$ be random vectors. Let $W \sim \text{Bernoulli}\,(p)$ for $p \in [0, 1]$, and let W be pairwise independent of the components of $\vec{V}$ and $\vec{C}$. Define the mixture of $\vec{V}$ and $\vec{C}$ as the random vector

$$\vec{M} = (1 - W)\,\vec{V} + W\vec{C}.$$

Concordant and Discordant Pairs

A concordant pair occurs when there are two points, (x_1, y_1) and (x_2, y_2), that have the same sign when the components are subtracted. In other words, $\text{sign}(x_2 - x_1) = \text{sign}(y_2 - y_1)$. A discordant pair occurs when the opposite is true, or when they have opposite signs. In other words, $\text{sign}(x_2 - x_1) = -\text{sign}(y_2 - y_1)$.

Concordant and Discordant Pairs



Kendall's Tau

Definition (τ Sample Version, Kendall 1938)

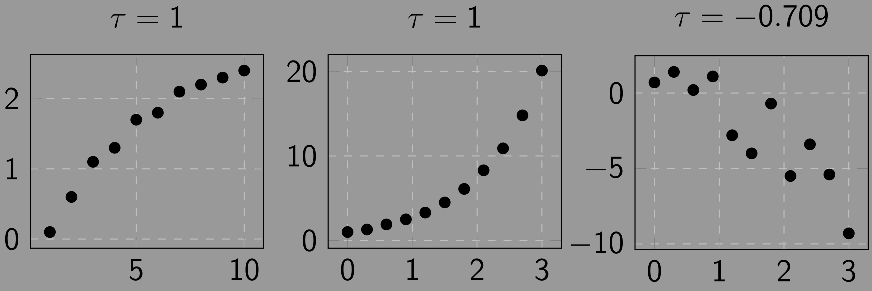
Given a sample of raw data, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, calculate the number of concordant pairs c and number of discordant pairs d. Kendall's Tau of a sample can be defined as

$$\widehat{\tau} = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{2}}$$

where n is the sample size.

Kendall's Tau has a range of $[-1, 1]$, where each extreme is perfect correlation of each sign. In fact, Kendall's Tau can capture any type of monotone behavior.

Kendall’s Tau Examples



$\tau = 1$

$\tau = 1$

$\tau = -0.709$

Kendall's Tau

Let's find Kendall's Tau for the following data:

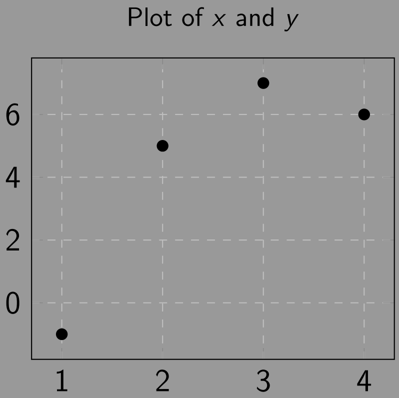
x	2	1	4	3
y	5	-1	6	7

$\downarrow$

x	1	2	3	4
y	-1	5	7	6

$$c = 5 \qquad d = 1$$

$$\widehat{\tau} = \frac{5-1}{5+1} = \frac{2}{3}$$



Kendall's Tau

Now we will introduce the population definition for Tau. As motivation, consider that as the sample size increases,

$$\lim_{n\to\infty} \frac{c-d}{\binom{n}{2}} = \lim_{n\to\infty} \frac{c}{\binom{n}{2}} - \lim_{n\to\infty} \frac{d}{\binom{n}{2}} \longrightarrow \Big(P\left(concordance\right) - P\left(discordance\right) \Big)$$

Definition (τ Population Version, Kruskal 1958)

Consider a bivariate distribution with random pair (X, Y). Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed pairs. Kendall's Tau of a population can be defined as

$$\tau = P\left(\left(X_1 - X_2\right)\left(Y_1 - Y_2\right) > 0\right) - P\left(\left(X_1 - X_2\right)\left(Y_1 - Y_2\right) < 0\right).$$

Spearman's Rho

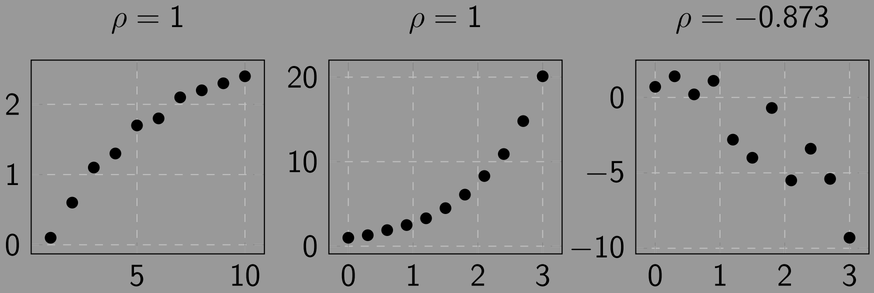
Definition (ρ Sample Version, Spearman 1904)

Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be real-valued observations and let r_x and r_y be the ranks of each respective variable. Spearman's Rho of a sample can be defined as

$$\widehat{\rho}_S(r_x, r_y) = \frac{\sum_{i=1}^{n}(r_{x_i} - \bar{r_x})(r_{y_i} - \bar{r_y})}{\sqrt{\sum_{i=1}^{n}(r_{x_i} - \bar{r_x})^2}\sqrt{\sum_{i=1}^{n}(r_{y_i} - \bar{r_y})^2}} = \frac{\widehat{\text{Cov}}(r_x, r_y)}{s_{r_x} s_{r_y}}.$$

Spearman's Rho is similar to Tau, where they can both capture general monotone behavior. Rho has a range $[-1, 1]$, where each extreme is perfect correlation of each sign. Also, observe that Spearman's Rho is Pearson's Correlation with the ranks.

Spearman's Rho



$\rho = 1$

$\rho = 1$

$\rho = -0.873$

Spearman's Rho

Let's find Spearman's Rho for the following data:

x	2	1	4	3
y	5	-1	6	7

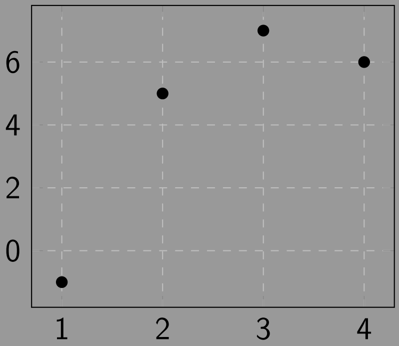
$$\downarrow$$

r_x	2	1	4	3
r_y	2	1	3	4

$$\rho = \frac{\operatorname{Cov}(r_x, r_y)}{s_{r_x} s_{r_y}} = 0.8$$

(via R)

Plot of x and y



Spearman's Rho

The population definition and sample definition are not connected intuitively or briefly, so we will skip that. In the mean time,

Definition (ρ Population Version, Kruskal 1958)

Consider a bivariate distribution with random pair (X, Y). Let (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) be independent and identically distributed pairs. Spearman's Rho of a population can be defined as

$$\rho_S = 3\left(P\left((X_1 - X_2)(Y_1 - Y_3) > 0\right) - P\left((X_1 - X_2)(Y_1 - Y_3) < 0\right)\right).$$

Introduction to Copulas

Both of the population definitions we have seen for Rho and Tau have several equivalent forms, including in terms of copulas. Copulas are very important tools, as they are able to isolate information about the dependence structure of jointly distributed random variables.

For the purposes of this project we only consider a bivariate case, but copulas can be extended to a *d*-dimensional case.

Introduction to Copulas

Definition

A two-dimensional copula is a function $C : [0,1]^2 \to [0,1]$ with bivariate inputs (u, v) such that the following conditions are satisfied:

1. C is a 2-increasing function, the bivariate analog of a univariate non-decreasing function. Equivalently, for every $u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

   $$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

   This has also been called quasi-monotone.
2. $C(u, 1) = u$ and $C(1, v) = v$.
3. $C(u, 0) = 0$ and $C(0, v) = 0$.

Sklar's Theorem

Theorem

Let $F_{X,Y}(x,y)$ *be a bivariate distribution function with marginals* $F_X(x)$ *and* $F_Y(y)$*. Then there exists a copula* C *such that for all* $(x,y) \in \mathbb{R}^2$,

$$F_{X,Y}(x,y) = C\left(F_X(x), F_Y(y)\right).$$

If $F_X(x), F_Y(y)$ *are continuous, then* C *is unique; otherwise* C *is uniquely determined on* $\operatorname{ran} F_X(x) \times \operatorname{ran} F_Y(y)$*. Conversely, if* C *is a copula and* $F_X(x), F_Y(y)$ *are distribution functions, then the function* $F_{X,Y}(x,y)$ *defined above is a bivariate distribution function with marginals* $F_X(x), F_Y(y)$*.*

Sklar's Corollary

Corollary

Using the same notation as in Sklar's Theorem, also let $F_X^{-1}(x)$ *and* $F_Y^{-1}(y)$ *be quasi-inverses of* $F_X(x)$ *and* $F_Y(y)$*, respectively. Then for any* $(u, v) \in \operatorname{dom} C$*,*

$$C(u, v) = F_{X,Y}\left(F_X^{-1}(u), F_Y^{-1}(v)\right).$$

Independence Copula

Definition

Let (U, V) be a bivariate random vector with uniform marginals. An independence copulas is defined as

$$\Pi(u, v) = uv.$$

In fact, random variables are independent if and only if their copula is the independence copula.

"Q" Construct

Theorem

Let (X_1, Y_1) and (X_2, Y_2) be independent vectors of continuous random variables with joint distribution functions F_1 and F_2, respectively, with common marginals $F_X(x)$ and $F_Y(y)$. Let C_1 and C_2 denote the copulas of (X_1, Y_1) and (X_2, Y_2), respectively, so that $F_1(x, y) = C_1(F_X(x), F_Y(y))$ and $F_2(x, y) = C_2(F_X(x), F_Y(y))$. Let Q denote the difference between the probability of concordance and discordance of (X_1, Y_1) and (X_2, Y_2), i.e. let

$$Q = P\left(\left(X_1 - X_2\right)\left(Y_1 - Y_2\right) > 0\right) - P\left(\left(X_1 - X_2\right)\left(Y_1 - Y_2\right) < 0\right).$$

Then

$$Q = Q(C_1, C_2) = 4 \int\int_{[0,1]^2} C_2(u, v)\, \mathrm{d}C_1(u, v) - 1.$$

"Q" Construct Corollary

Corollary ("Q" corollary)

Using the same notation from the previous theorem, also let $\overline{C}$ be a survival copula. A survival copula has the same properties as a typical survival function.

1. *Q is symmetric in its arguments. That is,* $Q(C_1, C_2) = Q(C_2, C_1)$.
2. *Copulas can be replaced by survival copulas in Q. That is,* $Q\left(C, \overline{C}\right) = Q\left(\overline{C}, C\right)$.

Equivalent Definition of Tau

Theorem

Let (X, Y) *be a continuous random vector and let* C *be a copula for* (X, Y)*. Kendall's Tau can be defined as*

$$Q(C, C) = 4 \int \int_{[0,1]^2} C(u, v) \, \mathrm{d}C(u, v) - 1.$$

Equivalent Definition of Rho

Theorem

Let (X, Y) be a continuous random vector and let C be a copula for (X, Y). Spearman's Rho can be defined as

$$3Q\left(C, \Pi\right) = 12\int\int_{[0,1]^2} C\left(u, v\right) \mathrm{d}u\mathrm{d}v - 3$$

where Π is an independence copula.

Now we can see this connection between the population definition and the sample definition of Rho.

Equivalent Forms

Definition (Alternate definition in terms of CDF)

Let (X, Y) be a continuous, random vector with joint CDF $F_{X,Y}(x, y)$ and respective marginals $F_X(x)$, $F_Y(y)$. We can define Kendall's Tau as

$$\tau(X, Y) = 4 \int\int_{\mathbb{R}^2} F_{X,Y}(x, y)\, \mathrm{d}F_{X,Y}(x, y) - 1.$$

and Spearman's Rho as

$$\rho_S(X, Y) = 12 \int\int_{\mathbb{R}^2} F_{X,Y}(x, y)\, \mathrm{d}F_X(x)\, \mathrm{d}F_Y(y) - 3$$

Bias Introduction

Definition (Bias Under Mixtures)

Let $\vec{V}$, $\vec{C}$, $\vec{M}$, W, and p be described as in the Mixture Model definition. The **bias under the mixture** is

$$\begin{aligned} \text{Bias}_\tau(p) &= \tau_{\vec{M}} - \tau_{\vec{V}}, \\ \text{Bias}_\rho(p) &= \rho_{\vec{M}} - \rho_{\vec{V}}. \end{aligned}$$

Using this idea, we will now introduce the main result from this presentation.

Lemma

Let $F_{X,Y}(x,y)$ be a joint CDF with marginals $F_X(x)$ and $F_Y(y)$ with survival function $S_{X,Y}(x,y)$. Then

$$\int\int_{\mathbb{R}^2} F_{X,Y}(x,y)\,\mathrm{d}F_X(x)\,\mathrm{d}F_Y(y) = \int\int_{\mathbb{R}^2} S_{X,Y}(x,y)\,\mathrm{d}S_X(x)\,\mathrm{d}S_Y(y)$$

and

$$\int\int_{\mathbb{R}^2} F_{X,Y}(x,y)\,\mathrm{d}F_{X,Y}(x,y) = \int\int_{\mathbb{R}^2} S_{X,Y}(x,y)\,\mathrm{d}S_{X,Y}(x,y).$$

Bias in Rank Correlation Due to Mixing I

The following theorem will present this idea of bias in terms of the mixing proportion, p. Note that if there is no mixing, the bias is zero.

Theorem (Bias of τ and ρ)

The bias in Kendall's Tau and Spearman's Rho due to mixing can be expressed as

$$Bias_\tau(p) = 4\left(a_\tau p^2 + b_\tau p\right)$$
$$Bias_\rho(p) = 12\left(a_\rho p^3 + b_\rho p^2 + c_\rho p\right),$$

where...

Bias in Rank Correlation Due to Mixing II

Theorem (Bias of τ and ρ)

$$\begin{aligned} a_\tau = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{\vec{V}}(x,y) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{\vec{C}}(x,y) \\ &- \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{\vec{V}}(x,y) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{\vec{C}}(x,y), \\ b_\tau = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{\vec{C}}(x,y) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{\vec{V}}(x,y) \\ &-2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{\vec{V}}(x,y), \end{aligned}$$

Bias in Rank Correlation Due to Mixing III

Theorem (Bias of τ and ρ)

$$
\begin{aligned}
a_\rho = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{V_2}(y) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{C_2}(y)\\
&- \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{V_2}(y) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{C_2}(y)\\
&- \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{V_2}(y) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{C_2}(y)\\
&- \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{V_2}(y) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{C_2}(y),
\end{aligned}
$$

Bias in Rank Correlation Due to Mixing IV

Theorem (Bias of τ and ρ)

$$\begin{aligned}
b_\rho =&3\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{V_2}(y) - 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{C_2}(y)\\
&-2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{V_2}(y) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{C_2}(y)\\
&-2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{V_2}(y) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{C_2}(y)\\
&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{V_2}(y),\\
c_\rho =&-3\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{V_2}(y) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{C_2}(y)\\
&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{V_2}(y) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{C}}(x,y)\mathrm{d}S_{V_1}(x)\mathrm{d}S_{V_2}(y).
\end{aligned}$$

Bias in Rank Correlation Due to Mixing

The proof will be skipped, but the important step involves taking advantage of the linearity of integral differentials. The property can be shown below.

Lemma

Let $f(x), g(x)$, and $h(x)$ be real-valued functions and let $a, b, c \in \mathbb{R}$. Then

$$\int_s^t f(x)\mathrm{d}\left(ag(x) + bh(x)\right) = \int_s^t af(x)\mathrm{d}g(x) + \int_s^t bf(x)\mathrm{d}h(x)$$

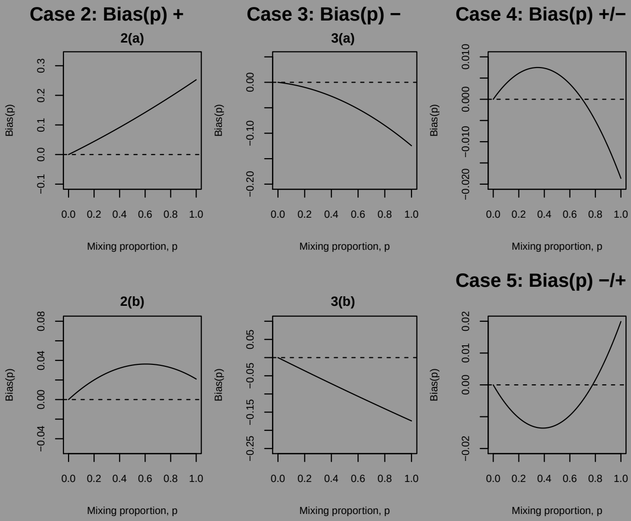
Bias Cases I

Consider a quadratic function without a constant term, $f(p) = ap^2 + bp$. The following inequalities between a and b serve as a partition of the coefficient space that characterizes root behavior and thus the regions in the unit interval where the function is positive and negative.

Bias Cases II

1. $f(p) = 0$ for all p in $(0, 1)$. This occurs when $a = b = 0$.
2. $f(p) > 0$ for all p in $(0, 1)$. This has two subcases.
    1. $a \geq 0$ and $b \geq 0$ (but not both equal to zero).
    2. $0 < -a \leq b$.
3. $f(p) < 0$ for all p in $(0, 1)$. This has two subcases.
    1. $a \leq 0$ and $b \leq 0$ (but not both equal to zero).
    2. $b \leq -a < 0$.
4. There exists a root p_1 in $(0, 1)$ such that $f(p) > 0$ for $p < p_1$, but $f(p) < 0$ for $p > p_1$. This occurs when $0 < b < -a$.
5. There exists a non-zero root p_1 in $(0, 1)$ such that $f(p) < 0$ for $p < p_1$, but $f(p) > 0$ for $p > p_1$. This occurs when $-a < b < 0$.

Bias Cases III



Bias Cases IV

Consider a cubic function without a constant term, $f(p) = ap^3 + bp^2 + cp$. The following inequalities between a, b, and c serve as a partition of the coefficient space that characterizes root behavior and thus the regions in the unit interval where the function is positive and negative.

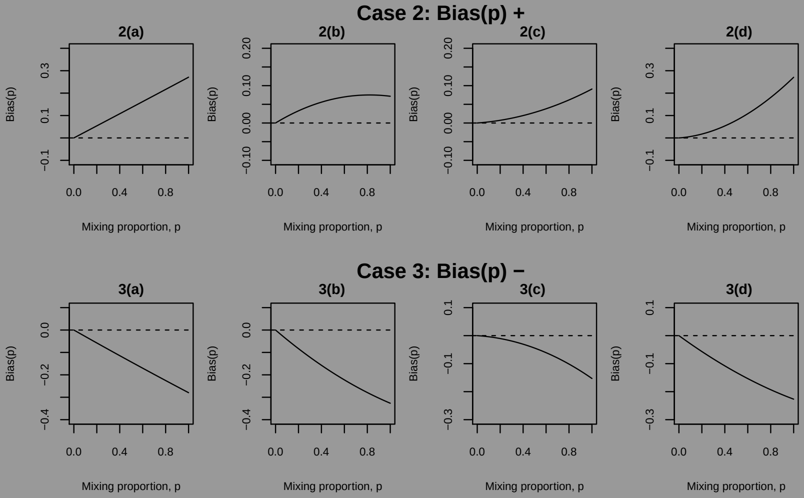
Bias Cases V

1. $f(p) = 0$ for all p in $(0,1)$. This occurs when $a = b = c = 0$.
2. $f(p) > 0$ for all p in $(0,1)$. This has four subcases.
    1. $a > 0$, $|b| < 2\sqrt{ac}$, $c > 0$.
    2. $a > 0$, $|b| > 2\sqrt{ac}$, $b \leq -2a$, $c \geq a$, $a + b + c \geq 0$
    3. $a > 0$, $b > 0$, $c > 0$, $|b| > 2\sqrt{ac}$
    4. $a < 0$, $c > 0$, $a + b + c \geq 0$
3. $f(p) < 0$ for all p in $(0,1)$. This has four subcases.
    1. $a < 0$, $|b| < 2\sqrt{ac}$, $c < 0$
    2. $a < 0$, $|b| > 2\sqrt{ac}$, $b \geq -2a$, $c \leq a$, $a + b + c \leq 0$
    3. $a < 0$, $b < 0$, $c < 0$, $|b| > 2\sqrt{ac}$
    4. $a > 0$, $c < 0$, $a + b + c \leq 0$
4. There exists a non-zero root p_1 in $(0,1)$ such that $f(p) > 0$ for $p < p_1$, but $f(p) < 0$ for $p > p_1$. This has two subcases.
    1. $a > 0$, $b \leq -a$, $c > 0$, $|b| > 2\sqrt{ac}$, $a + b + c \leq 0$
    2. $a < 0$, $c > 0$, $a + b + c \leq 0$

Bias Cases VI

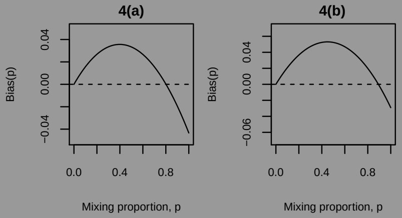
5. There exists a non-zero root p_1 in $(0,1)$ such that $f(p) < 0$ for $0 < p < p_1$, but $f(p) > 0$ for $p_1 < p < 1$. This has two subcases.
    1. $a < 0,\ b \geq -a,\ c < 0,\ |b| > 2\sqrt{ac},\ a + b + c \geq 0$
    2. $a > 0,\ c < 0,\ a + b + c \geq 0$
6. There exists two non-zero roots $p_1 < p_2$ in $(0,1)$ such that $f(p) > 0$ for $0 < p < p_1$, $f(p) < 0$ for $p_1 < p < p_2$, and $f(p) > 0$ for $p_2 < p < 1$. This occurs when $a > 0$, $-2a \leq b \leq 0,\ c \leq a,\ |b| > 2\sqrt{ac},\ a + b + c \geq 0.$
7. There exists two non-zero roots $p_1 < p_2$ in $(0,1)$ such that $f(p) < 0$ for $0 < p < p_1$, $f(p) > 0$ for $p_1 < p < p_2$, and $f(p) < 0$ for $p_2 < p < 1$. This occurs when $a < 0$, $-2a \geq b \geq 0,\ c \geq a,\ |b| > 2\sqrt{ac},\ a + b + c \leq 0.$

Bias Cases VII

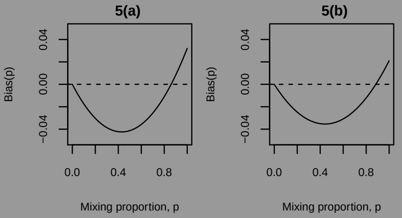


Bias Cases VIII

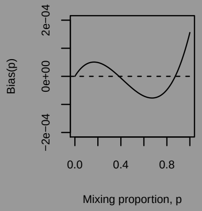
Case 4: Bias(p) +/−



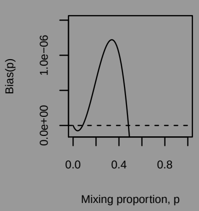
Case 5: Bias(p) −/+



Case 6: Bias(p) +/−/+



Case 7: Bias(p) −/+/−



Marshall-Olkin (MO) Distribution

This distribution (Marshall & Olkin 1967) arises from "shock models" and its ability to predict the failing of a two-component system. Define three independent random variables, where

$$Z_1 \sim \text{Exp}\left(\lambda_1\right), Z_2 \sim \text{Exp}\left(\lambda_2\right), \text{ and } Z_3 \sim \text{Exp}\left(\lambda_3\right)$$

which represent the occurrences of the shocks. The first two random variables are shocks to component one and component two, respectively, and the last random variable is a shock to both components. Now define two random variables

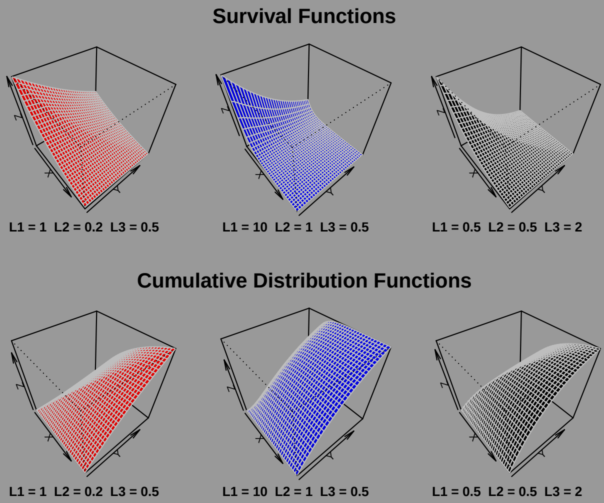
$$X = \min\{Z_1, Z_3\} \text{ and } Y = \min\{Z_2, Z_3\}.$$

MO Distribution

We can now find the joint survival function. The survival function will make future calculations more efficient while still arriving at the same result.

$$\begin{aligned} S_{X,Y}(x,y) &= P\left(X > x, Y > y\right) \\ &= P\left(\min\{Z_1, Z_3\} > x, \min\{Z_2, Z_3\} > y\right) \\ &= P\left(Z_1 > x, Z_3 > x, Z_2 > y, Z_3 > y\right) \\ &= P\left(Z_1 > x, Z_2 > y, Z_3 > \max\{x,y\}\right) \\ &= \exp\left\{-\left(\lambda_1 x + \lambda_2 y + \lambda_3 \max\{x,y\}\right)\right\}, \qquad x, y > 0 \end{aligned}$$

Visualization of the MO Distribution



Survival Functions

L1 = 1 L2 = 0.2 L3 = 0.5

L1 = 10 L2 = 1 L3 = 0.5

L1 = 0.5 L2 = 0.5 L3 = 2

Cumulative Distribution Functions

L1 = 1 L2 = 0.2 L3 = 0.5

L1 = 10 L2 = 1 L3 = 0.5

L1 = 0.5 L2 = 0.5 L3 = 2

Copula For MO Distribution

See demonstration in R...

Bias of Rank Correlation for MO Distribution

Using the definition for the bias defined earlier, we can find the bias for both Tau and Rho under the MO distribution in terms of the mixing proportion, p.

Let's begin with Tau. To save a lot of time and calculations, we can notice each integral follows a similar form. Therefore, by solving any integral with placeholder parameters, the actual integrals just need parameters plugged in to the correct spots. To do this, we will solve any integral with parameters α_1, α_2, α_3, β_1, β_2, and β_3.

Li's Lemma

Before we move forward, we need to introduce an important and required lemma needed to evaluate the non-differentiable cusp shown earlier.

Lemma (Li's Copula Theorem for CDF's, Li et al. 2002)

Let $F_{X,Y}(x,y)$ and $G_{X,Y}(x,y)$ be joint distribution functions. Then

$$\int\int_{\mathbb{R}^2} F_{X,Y}(x,y)\,\mathrm{d}G_{X,Y}(x,y) = \frac{1}{2} - \int\int_{\mathbb{R}^2} \frac{\partial}{\partial x} F_{X,Y}(x,y)\,\frac{\partial}{\partial y} G_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

General Integral for Tau

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_\alpha(x,y)\mathrm{d}S_\beta(x,y)\\
&= \frac{1}{2} - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial}{\partial x}S_\alpha(x,y)\frac{\partial}{\partial y}S_\beta(x,y)\mathrm{d}x\mathrm{d}y\\
&= \frac{1}{2} - \int_{0}^{\infty}\int_{y}^{\infty} \beta_2\left(\alpha_1+\alpha_3\right)\exp\left(-\alpha_1 x-\alpha_3 x-\beta_1 x-\beta_3 x-\alpha_2 y-\beta_2 y\right)\mathrm{d}x\mathrm{d}y\\
&\quad - \int_{0}^{\infty}\int_{x}^{\infty} \alpha_1\left(\beta_2+\beta_3\right)\exp\left(-\alpha_1 x-\beta_1 x-\alpha_2 y-\alpha_3 y-\beta_2 y-\beta_3 y\right) dydx\\
&= \frac{1}{2} - \frac{\beta_2\left(\alpha_1+\alpha_3\right)}{\left(\alpha_1+\alpha_3+\beta_1+\beta_3\right)\left(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3\right)}\\
&\quad - \frac{\alpha_1\left(\beta_2+\beta_3\right)}{\left(\alpha_2+\alpha_3+\beta_2+\beta_3\right)\left(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3\right)}
\end{aligned}
$$

Example of Integral for Tau

Since we have both a validation distribution and a contaminated distribution, define parameters for each:

$$\mathsf{MO}_{\vec{V}}\left(\lambda_{V_1}, \lambda_{V_2}, \lambda_{V_3}\right) \qquad \mathsf{MO}_{\vec{C}}\left(\lambda_{C_1}, \lambda_{C_2}, \lambda_{C_3}\right).$$

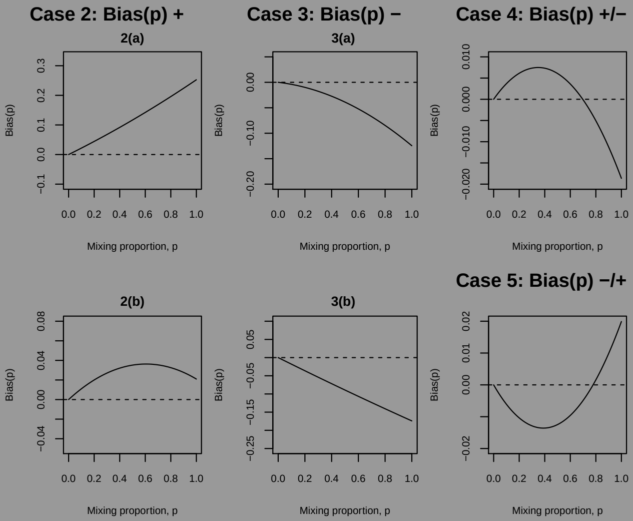
$$\begin{aligned}&\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\\ =\ &\frac{1}{2} - \frac{\lambda_{C2}\left(\lambda_{V1}+\lambda_{V3}\right)}{\left(\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}\right)\left(\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+\lambda_{C3}\right)}\\ &- \frac{\lambda_{V1}\left(\lambda_{C2}+\lambda_{C3}\right)}{\left(\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}\right)\left(\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+\lambda_{C3}\right)}\end{aligned}$$

Simulated Parameters for Tau for each Case

Case	λ_{V1}	λ_{V2}	λ_{V3}	λ_{C1}	λ_{C2}	λ_{C3}
2(a)	9.91	8.92	4.00	4.66	8.62	9.93
2(b)	6.20	7.52	7.45	1.23	5.83	4.20
3(a)	2.41	8.52	2.86	3.16	3.14	0.57
4(b)	3.05	0.69	7.24	3.64	4.48	7.64
5	6.47	7.69	6.34	5.29	3.99	3.80
6	2.80	7.44	6.69	7.67	3.24	7.74

Table: Parameter values for a mixture of Marshall-Olkin distributions that produce each scenario under Tau. Each value is rounded to two digits past the decimal. These are the parameter values used to produce the following figure.

Simulated Cases (Tau)



Bias of Rho for MO Distribution

Similar to Tau, we will solve a general integral to ease calculations. Rho's general integral will involve three general parameters, because the integral has two differentials. Let the general parameters be α_1, α_2, α_3, β_1, β_2, β_3, γ_1, γ_2, and γ_3.

General Integral for Rho

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\alpha}(x,y)\mathrm{d}S_X(x)\,\mathrm{d}S_Y(y) \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\alpha}(x,y)f_{\beta}(x)f_{\gamma}(y)\mathrm{d}x\mathrm{d}y \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\alpha_1 x-\alpha_2 y-\alpha_3\mathrm{max}\{x,y\}\right) \\
&(\beta_1+\beta_3)\exp(-(\beta_1+\beta_3)x)(\gamma_2+\gamma_3)\exp(-(\gamma_2+\gamma_3)y)\mathrm{d}x\mathrm{d}y \\
=&\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{(\alpha_2+\alpha_3+\beta_1+\beta_3)(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3)} \\
+&\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{(\alpha_1+\alpha_3+\gamma_2+\gamma_3)(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3)}
\end{aligned}
$$

Example Integral for Rho

Similar to Tau, we will consider the following integral to plug in the appropriate parameters.

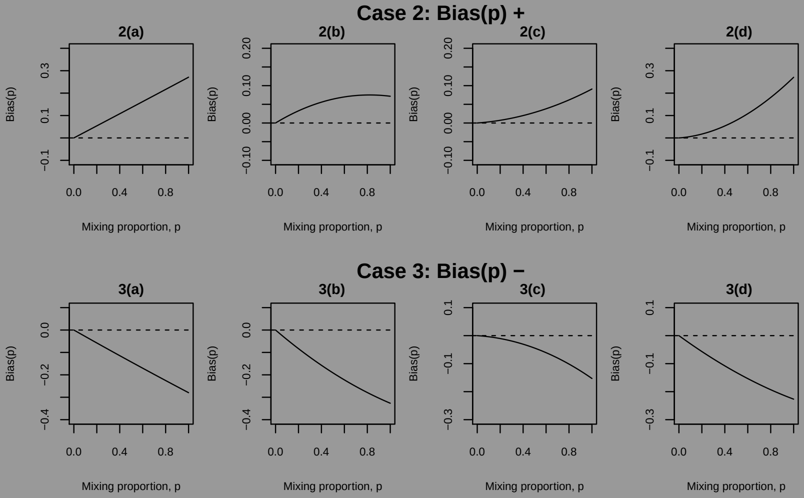
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_{\vec{V}}(x,y)\mathrm{d}S_{C_1}(x)\mathrm{d}S_{V_2}(y)$$
$$=\frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{(\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3})(\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3})}$$
$$+\frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{(\lambda_{V1}+\lambda_{V3}+\lambda_{V2}+\lambda_{V3})(\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3})}$$

Simulated Parameter for Rho for each Case

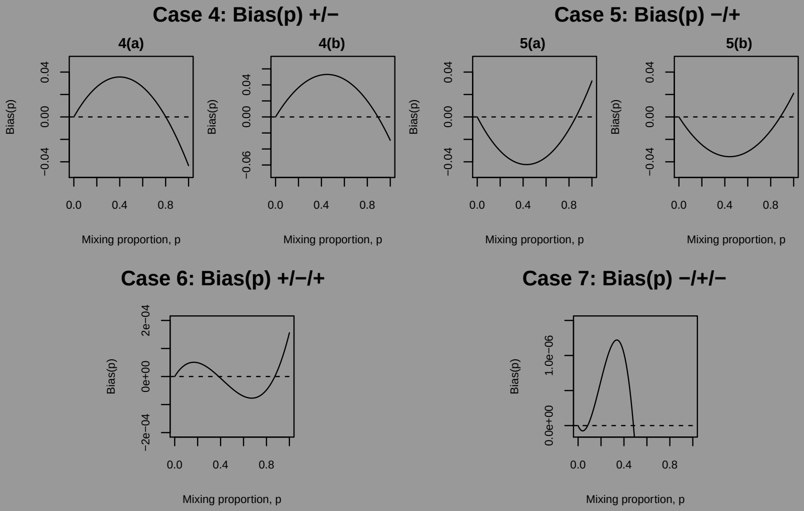
Case	λ_{V1}	λ_{V2}	λ_{V3}	λ_{C1}	λ_{C2}	λ_{C3}
2(a)	7.47	2.99	8.96	1.44	1.49	9.77
2(b)	4.01	5.47	4.26	4.92	9.36	8.58
2(c)	0.22	5.05	4.49	3.38	2.71	7.63
2(d)	4.51	9.60	2.48	8.32	2.39	6.58
3(a)	2.35	2.89	5.78	3.36	4.97	2.91
3(b)	1.35	5.89	4.67	4.74	2.72	0.98
3(c)	6.61	8.98	5.33	4.94	6.81	1.79
3(d)	1.75	5.93	7.00	9.80	5.73	5.59
4(a)	1.40	1.73	2.66	6.73	2.17	6.34
4(b)	6.72	4.48	5.20	3.90	1.47	2.21
5(a)	3.98	0.65	5.09	0.10	3.53	4.58
5(b)	0.20	9.31	6.26	6.01	4.68	7.66
6	3.45	2.07	6.09	8.84	0.18	9.95
7	3.06	8.03	8.73	3.77	5.35	7.17

Table: Parameter values for a mixture of Marshall-Olkin distributions that produce each scenario under rho. Each value is rounded to two digits past the decimal. These are the parameter values used to produce the following figure.

Simulated Cases I (Rho)



Simulated Cases II (Rho)



Conclusion

What question have we answered?

- Bias is introduced into models depending on a proportion of mixing.
- We can characterize the bias for Kendall's Tau under a mixture model as a quadratic with respect to the mixing proportion.
- We can characterize the bias for Spearman's Rho under a mixture model as a cubic with respect to the mixing proportion.
- Every case is feasibly possible, while some may be more common than others.

Conclusion

What question have not been answered?

- There is potential to represent these calculations in terms of copulas, a function that can capture the dependence structure between marginal distributions.
- There are other association measures that can be considered. Cronbach's Alpha has been analyzed in a previous paper by Stephen Carden, Trevor Camper, and Nick Holtzman.

Thank you

Questions?