Spring 2021

Bias of Rank Correlation Under A Mixture Model

Russell Land

BIAS OF RANK CORRELATION UNDER A MIXTURE MODEL

by

RUSSELL CHARLES LAND

(Under the Direction of Stephen Carden)

ABSTRACT

This thesis project will analyze the bias in mixture models when contaminated data is present. Specifically, we will analyze the relationship between the bias and the mixing proportion, p, for the rank correlation methods Spearman's Rho and Kendall's Tau. We will first look at the history of the two non-parametric rank correlation methods and the sample and population definitions will be introduced. Copulas will be introduced to show a few ways we can define these correlation methods. After that, mixture models will be defined and the main theorem will be stated and proved. As an example, we will apply this theorem to the Marshall-Olkin distribution. This will allow us to show the bias graphically for each of the different correlation methods.

INDEX WORDS: Rank correlation, Copula, Mixture model, Spearman Rho, Kendall Tau, Cumulative distribution function, Bias

2009 Mathematics Subject Classification: 15A15, 41A10

BIAS OF RANK CORRELATION UNDER A MIXTURE MODEL

by

RUSSELL CHARLES LAND

B.S., Georgia Southern University, 2019

M.S., Georgia Southern University, 2021

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

BIAS OF RANK CORRELATION UNDER A MIXTURE MODEL

by

RUSSELL CHARLES LAND

Major Professor:	Stephen Carden
Committee:	Divine Wanduku
	Arpita Chatterjee

Electronic Version Approved:
May 2021

ACKNOWLEDGMENTS

I would like to acknowledge my advisor Dr. Stephen Carden. He spent a large amount of time with me in meetings teaching me and guiding me through this process. Not only was he a great advisor but also a great lecturer and friend, and I cannot speak highly enough regarding his knowledge and dedication to his job.

Along with Dr. Carden, I would like to acknowledge my other professors who sparked my interest in statistics. Both Dr. Arpita Chatterjee and Dr. Divine Wanduku are amazing professors and mentors. I could always count on both of them to put a smile on my face.

Lastly, I would also like to acknowledge my parents who helped me through my college education and beyond. They were by my side whenever I needed them and had concerns about my future. I love you and thank you for all that you do.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The goal of statistics is to use data from a sample to learn as much as possible about a population. However, it is often the case that a sample might be contaminated with individuals not belonging to the target population. Statisticians are always trying to improve sampling methods, sample analysis, and parameter estimation, because the end goal is the most representative sample possible. However, statisticians recognize that these methods cannot be perfected and other methods have to be developed to achieve that goal. This thesis paper hopes to achieve a method for characterizing bias, and behavior of bias, whenever contaminated data is present in a model. We will later define this as a mixture model. Specifically, we will analyze two non-parametric measures of association between two variables in a sample (and populations as we will see). These two measures are Spearman's rank correlation coefficient (Rho) and Kendall's rank correlation coefficient (Tau). There is a long history of these coefficients, and there are a few ways to define them as well. We will also see a general correlation coefficient where Rho and Tau are special cases. Copulas are also introduced in Chapter 2, where we define equivalent forms of both Rho and Tau in terms of copulas. There is a classification of integrals introduced, called Riemann-Stieltjes integrals, which have integral differentials that are not the identity function. We will use a clever technique to solve these with minimal effort.

The main result in this thesis will be defining and characterizing the bias for both Rho and Tau under mixture models. We will introduce a bivariate mixture model ($\vec{M}$) that consists of bivariate measurements from a valid population ($\vec{V}$) and a contaminating population ($\vec{C}$). A key question we will investigate is how this bias is affected as the contamination changes. The bias will be defined in terms of a mixing proportion, p and will have the form

$$\text{Bias}_\theta\left(p\right) = \theta_{\vec{M}} - \theta_{\vec{V}}$$

for some statistic θ. A natural question one may ask is what form these equations will take.

Will they have a simple closed form? If so, can we classify the direction the bias might take as contamination is introduced? These are both questions that we will explore.

We will apply our results to the bivariate Marshall-Olkin distribution. This distribution has a neat and easy analytical form that we can later define explicitly. Furthermore, the distribution has closed form cumulative distribution functions (CDF), probability density functions (PDF), and copulas. This adds to the simplicity of the analytical solutions. While some calculations are very drawn out and tedious, it is primarily elementary algebra and calculus. To wrap things up, we will demonstrate the technique by visualizing the bias for randomly simulated parameter values.

CHAPTER 2

BACKGROUND OF SPEARMAN'S RHO AND KENDALL'S TAU

2.1 THE HISTORY OF SPEARMAN'S RHO

In statistics, it is a common to study the correlation between variables. It is important to understand the relationship between variables to check for dependence or lack of dependence. A common statistic used in linear regression, Pearson's correlation coefficient (ρ) can quantify the linear dependence between two random variables. A variant of Pearson's correlation coefficient is Spearman's Rho (ρ_S), which can be defined as the same quantity using the rank of the variables. The rank of a number is the corresponding index of an observation in the ordered data set. For example, the numbers $\{5, 7, 8, 2, 4\}$ have corresponding ranks $\{3, 4, 5, 1, 2\}$. We will begin by defining the sample version of Spearman's Rho, and consider the population version shortly after.

Definition 2.1. *Let* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *be real-valued observations. Pearson's correlation coefficient can be defined as*

$$\widehat{\rho}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\widehat{Cov}(x, y)}{s_x s_y}.$$

Definition 2.2. *Let* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *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*

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

This idea of rank correlation was first introduced into the psychology community by Charles Spearman in 1904. The motivation behind this measurement was to erase the quantity in observations and analyze how data increased/decreased. In the original paper, Spearman commented "it can often be altogether escaped in the case of quantities not admitting absolute measurement, by substituting instead *comparison*"[1]. As a consequence, Rho not only measures linear correlation, but any type of positive or negative monotone

behavior (exponential, quadratic, linear, etc.). By using the ranks of the data, the quantity is erased from the calculation.

For random variables, the population version of Spearman's Rho will be needed. The population definition was introduced by William Kruskal in 1958 [2]. The following definition uses the concept of concordance and discordance, which are essential throughout this paper.

Definition 2.3. *A concordant pair occurs when there are two points,* (x_1, y_1) *and* (x_2, y_2)*, that have the same sign when subtracted. In other words,* $\operatorname{sign}(x_2 - x_1) = \operatorname{sign}(y_2 - y_1)$*. Discordant pairs occur when the opposite is true, or when they have opposite signs. In other words,* $\operatorname{sign}(x_2 - x_1) = -\operatorname{sign}(y_2 - y_1)$.

Graphically, a concordant pair will form an increasing, positively sloped line between the two points. Similarly, a discordant pair will form a decreasing, negatively sloped line between the two points.

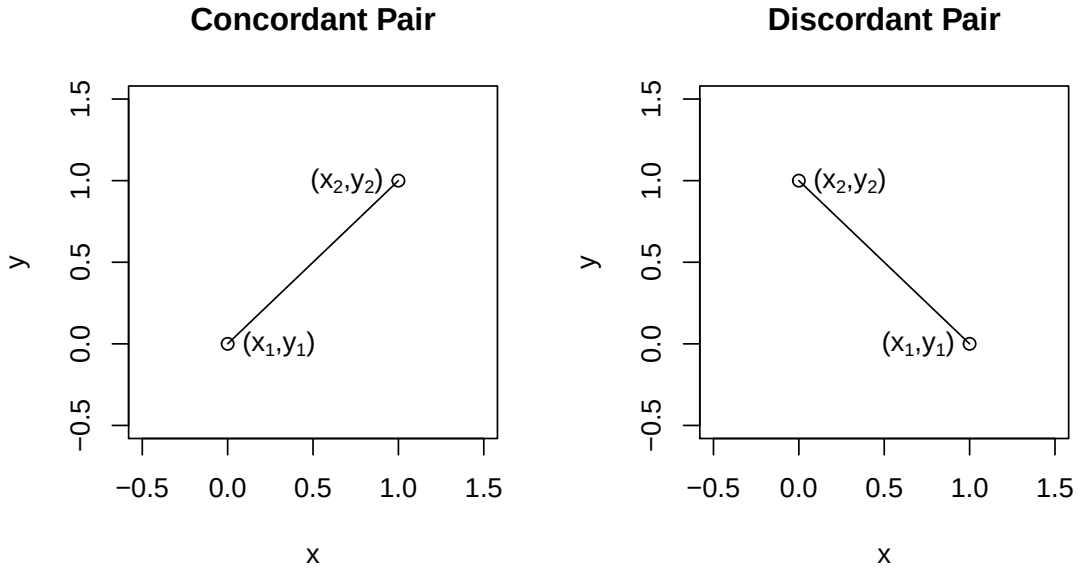

Figure 2.1: Two graphs showing examples of concordant and discordant pairs.

Definition 2.4. *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*

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

It is important to notice the dependence between X_1 and Y_1 and the independence between X_2 and Y_3. The reason for this will become more clear in Section 2.4. Also note that the quantity in the parentheses is multiplied by 3. This is because the quantity inside ranges from $-\frac{1}{3}$ to $\frac{1}{3}$ (for detail see [2] p. 824). The first term in the parentheses is the probability of concordance and the second term is probability of discordance. Also notice that the second term is the compliment of the first. There are several ways to write Spearman's Rho by taking advantage of this result. We will see in Section 2.4 that Spearman's Rho can be written in terms of probabilities, copulas, and CDFs. Although not obvious, we will later see how the sample definition and the population definition relate to each other.

2.2 The History of Kendall's Tau

In 1938, the statistics community was introduced to Kendall's Tau (τ), another measure of rank correlation [3]. There are several parallels between Rho and Tau. They both have the ability to measure monotone behavior between two variables and also both require ranks to calculate. If both Rho and Tau are calculated on a set of data, often the quantities are relatively close, and usually the same sign. The idea of concordant and discordant pairs are used in the calculation of Tau.

Definition 2.5. *Given a sample of raw data,* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$*, 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.*

There is an alternate definition for both Rho and Tau that accounts for tied pairs, which is when the sign equals zero. Throughout the paper, continuous data will be studied. Therefore, the probability of a tie happening is zero. The alternative definitions of Tau can be useful for discrete data.

For the population definition we have to find the difference between the probability of concordance and the probability of discordance. Just like Rho, there will be several results that take advantage of the symmetry of the following definition.

Definition 2.6. *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*

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

This result looks intuitively symmetric and, much like Rho, we will define this quantity in Section 2.4 in terms of probabilities, copulas, and CDFs. It is also important to notice the dependence between X_1 and Y_1, and the dependence between X_2 and Y_2.

Let's briefly relate this definition to the sample definition. Applying the Law of Large Numbers, think of the sample size approaching infinity.

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

This is a relaxed way to think about how the two definitions relate. To summarize, so far we have been introduced to the sample and population definitions of Spearman's Rho and Kendall's Tau.

2.3 A General Case for Rank Correlation

Apart from the correlation coefficients already introduced, in 1948 a general correlation coefficient was introduced by Maurice Kendall in his book *Rank Correlation Methods* [4]. He proposed that this correlation coefficient is a generalization of coefficients such as Kendall's Tau (τ), Spearman's Rho (ρ_S), and Pearson's Rho (ρ).

Definition 2.7. *To any pair of individuals, say the i^{th} and the j^{th}, we will allot an X-score, denoted by a_{ij}, subject to the condition that $a_{ij} = -a_{ji}$. Similarly, we will allot a Y-score, denoted by b_{ij}, where $b_{ij} = -b_{ji}$. We define a generalized correlation coefficient Γ by the equation*

$$\Gamma = \frac{\sum a_{ij} b_{ij}}{\sqrt{(\sum a_{ij}^2 \sum b_{ij}^2)}}$$

and regard a_{ij} as zero if $i = j$.

We can define Kendall's Tau using this definition. Let (X, Y) be a random bivariate vector with realized values (x, y). Then let r_i denote the rank of the i^{th} object and r_j denote the rank of the j^{th} object, both ranked according to the variable x. Similarly, let p_i denote the rank of the i^{th} object and p_j denote the rank of the j^{th} object, both ranked according to the variable y. Define

$$a_{ij} = \begin{cases} 1 & , r_i < r_j \\ -1 & , r_i > r_j \end{cases}$$

and

$$b_{ij} = \begin{cases} 1 & , p_i < p_j \\ -1 & , p_i > p_j \end{cases}.$$

Using these parameters, Kendall's Tau is now defined in terms of the general correlation coefficient. Definition 2.5 does not account for ties, but this general coefficient does. Because of this, the general coefficient is more powerful than the sample version. Again, our focus will be continuous data which has probability zero of a tie happening.

Spearman's Rho can be defined in a similar way. The parameters for Rho are much easier to implement. Using the same notation from the Tau case, define

$$a_{ij} = r_j - r_i$$

and

$$b_{ij} = p_j - p_i.$$

Spearman's Rho is now defined via parameters of the general correlation coefficient. It can easily be proved through algebraic manipulation that this way of defining it is equivalent to Definition 2.2.

This general coefficient that Maurice Kendall discovered is a pleasing result. In addition to the similarities between the two rank correlation methods that we have already seen, there will be more parallels throughout the entirety of this paper.

2.4 Modern Expressions for Rank Correlations

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. We can define copulas more formally below, and follow with some related results and important theorems.

Definition 2.8. *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 (for more detail, see [5] p. 8). 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 [6].

2. $C(u, 1) = u$ *and* $C(1, v) = v$*.*

3. $C(u, 0) = 0$ *and* $C(0, v) = 0$*.*

Definition 2.9. *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.

The next Theorem (Sklar's) will allow us to use copulas in a practical manner. We incorporate CDFs and copulas so we can apply them to the definitions of Tau and Rho. Using the next Theorem we will be able to input marginal distributions for some arbitrary joint distribution function and output a quantity that can capture the dependence structure between each marginal distribution.

Theorem 2.10 (Sklar's Theorem [7])**.** *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 $(x, y) \in \mathbb{R}^2$,*

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

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 $F_X(x), F_Y(y)$.

Corollary 2.11. *Using the same notation as in Theorem 2.10, 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 $(u, v) \in \operatorname{dom} C$,*

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

A quasi-inverse can be thought of as a traditional inverse function with weaker conditions. Now that we have defined copulas and how to apply copulas to probability distributions, we can harness their advantages and derive the rank correlation methods in terms of

copulas. To assist in future calculation and notation, we will introduce the "Q" construct. This will make the notation simpler.

The rest of the paper is scattered with a generalize integral of the form

$$\int_a^b f(x)\,\mathrm{d}g(x).$$

This is called a Riemann-Steiljes integral, and by having the differential as a non-identity function it will essentially "weight" the area of the curve with respect to the function $g(x)$. This idea can be extended into the bivariate case that we will utilize. This idea works when the function $g(x)$ satisfies general properties. However, when $g(x)$ does not satisfy these general properties, we will introduce a lemma that will fix that issue.

Theorem 2.12 (The "Q" Construct [5] p. 159)**.** *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*

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

Then

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

Proof. The random variables being used are continuous. Because of this, we can use the law of compliments to the following:

$$\begin{aligned} Q(C_1, C_2) &= P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0) \\ &= 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1 \\ &= 2\left[P(X_1 > X_2, Y_1 > Y_2) + P(X_2 > X_1, Y_2 > Y_1)\right] - 1. \end{aligned}$$

It remains to show that

$$P(X_1 > X_2, Y_1 > Y_2) + P(X_2 > X_1, Y_2 > Y_1) = 2P(X_1 > X_2, Y_1 > Y_2)$$
$$= 2\int\int_{\mathbb{R}^2} C_2(u,v)\,\mathrm{d}C_1(u,v).$$

We will spend the rest of the proof showing this. Start with the first term.

$$\begin{aligned} P(X_1 > X_2, Y_1 > Y_2) &= P(X_2 < X_1, Y_2 < Y_1) \\ &= \int\int_{\mathbb{R}^2} P(X_2 < x, Y_2 < y \mid X_1 = x, Y_1 = y)\, f_1(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int\int_{\mathbb{R}^2} F_2(x,y)\,\mathrm{d}F_1(x,y) \\ &= \int\int_{\mathbb{R}^2} C_2(F_X(x), F_Y(y))\,\mathrm{d}C_1(F_X(x), F_Y(y)). \end{aligned}$$

The previous line invokes Sklar's Theorem. Using what we call a probability-integral transformation [8] we can introduce the transformations $u = F_X(x)$ and $v = F_Y(y)$. After applying the transformation, it follows that

$$P(X_1 > X_2, Y_1 > Y_2) = \int\int_{[0,1]^2} C_2(u,v)\,\mathrm{d}C_1(u,v).$$

Moving on to the second half of the proof, we will prove the next result. Let $S_i(x,y)$ be the survival function for the i^{th} joint CDF.

$$\begin{aligned} P(X_1 < X_2, Y_1 < Y_2) &= P(X_2 > X_1, Y_2 > Y_1) \\ &= \int\int_{\mathbb{R}^2} P(X_2 > x, Y_2 > y \mid X_1 = x, Y_1 = y)\, f_1(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int\int_{\mathbb{R}^2} S_2(x,y)\,\mathrm{d}F_1(x,y) \\ &= \int\int_{\mathbb{R}^2} [1 - F_X(x) - F_Y(y) + F_2(x,y)]\,\mathrm{d}F_1(x,y) \\ &= \int\int_{\mathbb{R}^2} [1 - F_X(x) - F_Y(y) + C_2(F_X(x), F_Y(y))]\,\mathrm{d}C_1(F_X(x), F_Y(y)) \end{aligned}$$

The previous line invokes Sklar's Theorem. Using the same probability-integral transformation introduced earlier, it follows that

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

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

Thus,

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

■

Corollary 2.13 ("Q" corollary [6])**.** *Using the same notation from Theorem 2.12, 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,* $Q\left(C, \overline{C}\right) = Q\left(\overline{C}, C\right)$.

With the help of all the previous definitions, theorems, and corollaries, we can define both Spearman's Rho and Kendall's Tau in terms of copulas. Along with each definition, we will have a short discussion about each rank correlation method. To overcome potential confusion, we will finally connect the sample definition and the population definition of Spearman's Rho, as we have already seen with Tau.

Theorem 2.14. *Let* (X, Y) *be a continuous random vector and let* C *be a copula for* (X, Y)*. Spearman's Rho can be defined as*

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

where Π *is an independence copula.*

Proof. Recall from Definition 2.4, Spearman's Rho can be defined as

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

From Theorem 2.12, we know the difference between the probability of concordance and discordance is the Q construct. The following form may appear different than Theorem

2.12, but because X_2 and Y_3 are defined as independent, they will have an independence copula. Hence,

$$\rho_S = 3\left[4\int\int_{[0,1]^2} C\left(u,v\right)\mathrm{d}u\mathrm{d}v - 1\right] = 12\int\int_{[0,1]^2} C\left(u,v\right)\mathrm{d}u\mathrm{d}v - 3 = 3Q\left(C,\Pi\right).$$

■

An interesting result from the previous theorem can help us tie together the sample definition and the population version. It is important to restate that copulas of a probability distribution have uniform marginal distributions. Hence, $U \sim \text{Uni}\left(0,1\right)$ and $V \sim \text{Uni}\left(0,1\right)$. Note that the probability-integral transformation is used in the below derivation. Also recall the expected value and variance of a uniform distribution on $(0,1)$ are $\frac{1}{2}$ and $\frac{1}{12}$, respectively.

$$\begin{aligned}
\rho_S &= 3Q\left(C,\Pi\right) \\
&= 12\int\int_{[0,1]^2} uv\mathrm{d}C\left(u,v\right) - 3 \\
&= 12\int\int_{[0,1]^2} uv\mathrm{d}F_{X,Y}\left(F_X^{-1}\left(u\right),F_Y^{-1}\left(v\right)\right) - 3 && \text{(by Corollary 2.11)} \\
&= 12\int\int_{[0,1]^2} uv\mathrm{d}P\left(X < F_X^{-1}\left(u\right), Y < F_Y^{-1}\left(v\right)\right) - 3 \\
&= 12\int\int_{[0,1]^2} uv\mathrm{d}P\left(U < u, V < v\right) - 3 && \text{(transformation)} \\
&= 12\int\int_{[0,1]^2} uvf_{U,V}\left(u,v\right)\mathrm{d}u\mathrm{d}v - 3 \\
&= 12\cdot\mathrm{E}\left[UV\right] - 3 \\
&= \frac{\mathrm{E}\left[UV\right] - \frac{1}{4}}{\frac{1}{12}} \\
&= \frac{\text{Cov}\left(U,V\right)}{\sqrt{\text{Var}\left(U\right)}\sqrt{\text{Var}\left(V\right)}} \\
&= \rho\left(F_X\left(x\right),F_Y\left(y\right)\right).
\end{aligned}$$

Even through the population definition in terms of copulas, we are still able to define it in terms of Pearson's correlation coefficient. In summary, the population version of Spearman's Rho is Pearson's evaluated in terms of the marginal CDFs.

Theorem 2.15. *Let (X, Y) be a continuous random vector and let C be a copula for (X, Y). Kendall's Tau can be defined as*

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

Proof. Recall from Definition 2.6, the population definition of Kendall's Tau is

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

Apply Theorem 2.12 and we arrive at

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

■

Using the results from above, we can finally define both Spearman's Rho and Kendall's Tau in terms of CDFs. Although the bias that we will study in later sections can be achieved by using copulas, it is much easier to work with the definitions in terms of densities. The copulas will help during Chapter 4 when we introduce the Marshall-Olkin distribution. The intuition behind the following theorem comes easily from the results above, so we will not prove them.

Definition 2.16. *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 Spearman's Rho as*

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

and Kendall's Tau as

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

Proof. For both expression we will introduce the substitution $u = F_X(x)$ and $v = F_Y(y)$. Then, invoking Sklar's Theorem (2.10) we get the following expressions.

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

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

$$\begin{aligned}\tau &= 4 \int \int_{[0,1]^2} C(u,v)\, \mathrm{d}C(u,v) - 1 \\ &= 4 \int \int_{\mathbb{R}^2} F_{X,Y}(x,y)\, \mathrm{d}F_{X,Y}(x,y) - 1\end{aligned}$$

■

In summary, we can make the following observation. Ignoring the constants, Spearman's Rho can be defined as an integral over $\mathbb{R}^2$ of a joint CDF with respect to the two marginal CDFs. Similarly, ignoring the constants, Kendall's Tau can be defined as an integral over $\mathbb{R}^2$ of a joint CDF with respect to the joint CDF. Another way to think about these rank correlation methods is in terms of concordance and discordance. Ignoring constants, Spearman's Rho is the probability of concordance minus the probability of discordance, with the constraints that the marginals are independent. Similarly, ignoring the constants, Kendall's Tau is just the probability of concordance minus the probability of discordance.

In the next chapter we will see the expressions from Definition 2.16 in terms of survival functions. The following lemma will show this is valid, and will also help simplify future calculations.

Lemma 2.17. *Let $F_{X,Y}(x,y)$ be a joint CDF with marginals joint densities $f_X(x)$ 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

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

Proof. Starting with the first expression,

$$\begin{aligned}&\int \int_{\mathbb{R}^2} F_{X,Y}(x,y)\, \mathrm{d}F_X(x)\, \mathrm{d}F_Y(y) \\ =&\int \int_{\mathbb{R}^2} F_{X,Y}(x,y)\, f_X(x)\, f_Y(y)\, \mathrm{d}x\mathrm{d}y\end{aligned}$$

$$
\begin{aligned}
&= \int\int_{\mathbb{R}^2} \left[1 - S_X(x) - S_Y(y) + S_{X,Y}(x,y)\right] f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= \int\int_{\mathbb{R}^2} \left[1 - (1 - F_X(x)) - (1 - F_Y(y)) + S_{X,Y}(x,y)\right] f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= \int\int_{\mathbb{R}^2} f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y - \int\int_{\mathbb{R}^2} f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y - \int\int_{\mathbb{R}^2} f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&\quad + \int\int_{\mathbb{R}^2} F_X(x) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y + \int\int_{\mathbb{R}^2} F_Y(y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&\quad + \int\int_{\mathbb{R}^2} S_{X,Y}(x,y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

Now introduce the substitutions $u = F_X(x)$ and $v = F_Y(y)$, where $\mathrm{d}u = f_X(x)\,\mathrm{d}x$ and $\mathrm{d}v = f_Y(y)\,\mathrm{d}y$.

$$
\begin{aligned}
&= -\int\int_{\mathbb{R}^2} f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y + \int_{\mathbb{R}} f_Y(y) \int_0^1 u\mathrm{d}u\mathrm{d}y + \int_{\mathbb{R}} f_X(x) \int_0^1 v\mathrm{d}v\mathrm{d}x \\
&\quad + \int\int_{\mathbb{R}^2} S_{X,Y}(x,y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= -1 + \frac{1}{2}\int_{\mathbb{R}} f_Y(y)\,\mathrm{d}y + \frac{1}{2}\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x + \int\int_{\mathbb{R}^2} S_{X,Y}(x,y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= -1 + \frac{1}{2} + \frac{1}{2} + \int\int_{\mathbb{R}^2} S_{X,Y}(x,y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= \int\int_{\mathbb{R}^2} S_{X,Y}(x,y) f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

Now observe that

$$
\mathrm{d}S_X(x)\,\mathrm{d}S_Y(y) = \mathrm{d}\left(1 - F_X(x)\right)\mathrm{d}\left(1 - F_Y(y)\right) = f_X(x) f_Y(y)\,\mathrm{d}x\mathrm{d}y.
$$

Hence,

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

The second expression is proved in a similar way.

$$
\begin{aligned}
&\int\int_{\mathbb{R}^2} F_{X,Y}(x,y)\,\mathrm{d}F_{X,Y}(x,y) \\
&= \int\int_{\mathbb{R}^2} \left[1 - S_X(x) - S_Y(y) + S_{X,Y}(x,y)\right] f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= \int\int_{\mathbb{R}^2} \left[1 - (1 - F_X(x)) - (1 - F_Y(y)) + S_{X,Y}(x,y)\right] f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

$$
\begin{aligned}
&= \int\int_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y - \int\int_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y - \int\int_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\
&\quad + \int\int_{\mathbb{R}^2} F_X(x)\, f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y + \int\int_{\mathbb{R}^2} F_Y(x)\, f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\
&\quad + \int\int_{\mathbb{R}^2} S_{X,Y}(x,y)\, f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

Now introduce the substitutions $u = F_X(x)$ and $v = F_Y(y)$, where $\mathrm{d}u = f_X(x)\,\mathrm{d}x$ and $\mathrm{d}v = f_Y(y)\,\mathrm{d}y$.

$$
\begin{aligned}
&= -\int\int_{\mathbb{R}^2} f_{X,Y}(x,y) + \int_{\mathbb{R}} f_{X,Y}(x,y) \int_0^1 u\mathrm{d}u\mathrm{d}y + \int_{\mathbb{R}} f_{x,Y}(x,y) \int_0^1 v\mathrm{d}v\mathrm{d}x \\
&\quad + \int\int_{\mathbb{R}^2} S_{X,Y}(x,y)\, f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= -1 + \frac{1}{2} + \frac{1}{2} + \int\int_{\mathbb{R}^2} S_{x,Y}(x,y)\, f_{x,Y}(x,y)\,\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

Now observe that

$$
\mathrm{d}S_{x,Y}(x,y) = \mathrm{d}\left(1 - F_X(x) - F_Y(y) + F_{X,Y}(x,y)\right) = f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.
$$

Hence,

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

■

CHAPTER 3

BIAS UNDER A MIXTURE MODEL AND ITS BEHAVIOR

3.1 What Is a Mixture Model?

Mixture models can arise for multiple reasons. This could be due to bad sampling methods, diverse population, or it could be intended. In general, a mixture model represents a population with more than one distribution in it. For our purposes, we will be analyzing a bivariate mixture model that contains true, intended data, and contaminating, unwanted data. More formally, we can define a mixture model below.

Definition 3.1. *Let $\vec{V}$ and $\vec{C}$ be bivariate random vectors from the valid distribution and contaminated distribution, respectively. Let $W \sim$ Bernoulli(p), where p is the proportion of contamination. Then define a bivariate mixture model as*

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

Consider the following scenario. Statisticians are collecting two statistics for undergraduate students at a university. They administer surveys randomly throughout campus without screening. There is a possibility that graduate students accidentally enter the sample unintended. Although the probability may be low, it still must be considered. This sample will contain a proportion (p) of unwanted, contaminated data. Assume that the distribution of each population is known where undergraduates are the valid data and the graduate students are the contaminating data, and

$$\vec{V} \sim \text{Normal}\left(\vec{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ \Sigma = \begin{bmatrix} 0.07 & -0.03 \\ -0.03 & 0.05 \end{bmatrix}\right)$$

$$\vec{C} \sim \text{Normal}\left(\vec{\mu} = \begin{bmatrix} 2 \\ 3 \end{bmatrix},\ \Sigma = \begin{bmatrix} 0.03 & -0.05 \\ -0.05 & 0.1 \end{bmatrix}\right).$$

Taking a random sample from these populations and mixing accordingly, we can observe the results in Figure 3.1 below. In the first graph we are introducing no mixing, and is

all the valid population. In the second graph there is half-valid, half-contaminated. The third graph contains all contaminated data, which would occur if the statistician samples the wrong population, entirely. This example of mixing illustrates the famous Simpson's Paradox. Notice the correlation is negative with no mixing, and as mixing is introduced the correlation becomes positive and then back to negative as the contamination takes over. This relationship between mixing and positive/negative correlation will be similar to the main results introduced throughout this chapter.

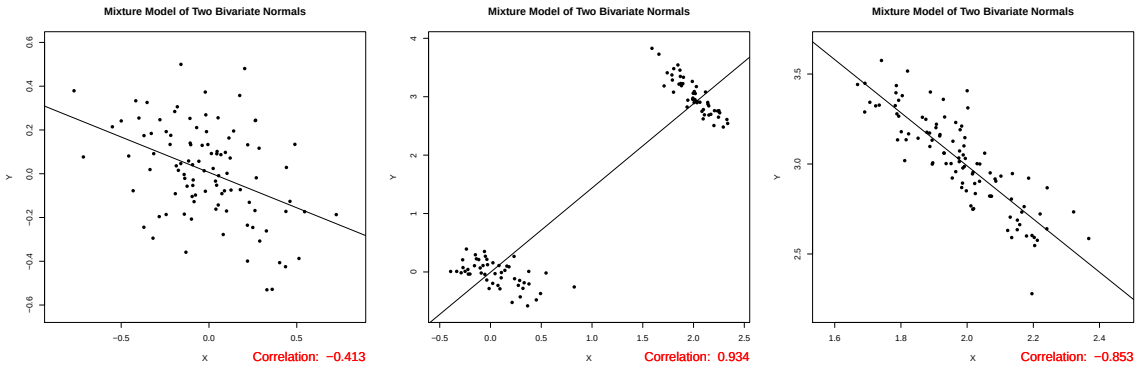

Figure 3.1: Three different results of changing the mixing proportion (p).

The first graph shows sampling the valid population. The second graph shows sampling half from the valid population and half from the contaminated population. The last graph shows sampling fully from the contaminated population. This diagram illustrates Simpson's Paradox. For visual convenience, the correlation is printed and the least-squares regression line is over-layed on each graph.

3.2 Bias Under Mixture Models

Traditionally, bias is defined as the difference between the expected value of an estimate and the true value of the parameter it is intended to estimate. In the context of Spearman's Rho, Kendall's Tau, and mixture models we define the bias as the difference between the parameter of the mixture and the parameter of the valid population. This can

be defined more formally below.

Definition 3.2. *Let $\vec{V}$, $\vec{C}$, $\vec{M}$, W, and p be described as in definition 3.1. The bias under the mixture is*

$$\textit{Bias}_\tau(p) = \tau_{\vec{M}} - \tau_{\vec{V}},$$
$$\textit{Bias}_\rho(p) = \rho_{\vec{M}} - \rho_{\vec{V}}.$$

Before the main result of the paper is introduced, there is one more lemma that will be used in the proof for Theorem 3.4. This lemma will use the property of the linearity of differentials and will help us prove later results.

Lemma 3.3. *Let $f(x), g(x)$, and $h(x)$ be real-valued functions and let $a, b, c \in \mathbb{R}$. Then*

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

Applying Definition 3.2 to the rank correlation methods, the main result of this paper can now be introduced.

Theorem 3.4. *The bias in Kendall's Tau and Spearman's Rho due to mixing can be expressed as*

$$\textit{Bias}_\tau(p) = 4\left(a_\tau p^2 + b_\tau p\right)$$
$$\textit{Bias}_\rho(p) = 12\left(a_\rho p^3 + b_\rho p^2 + c_\rho p\right),$$

where

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

and

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

Proof. Beginning with Kendall's Tau, we will start with the definition of bias.

$$\begin{aligned}
\mathrm{Bias}_\tau(p) =& \tau_{\vec{M}} - \tau_{\vec{V}}\\
=& 4\int\int_{\mathbb{R}^2} S_{\vec{M}}(x,y)\,\mathrm{d}S_{\vec{M}}(x,y) - 1 - \left(4\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) - 1\right)\\
=& 4\int\int_{\mathbb{R}^2} \left[(1-p)\,S_{\vec{V}}(x,y) + pS_{\vec{C}}(x,y)\right]\mathrm{d}\left[(1-p)\,S_{\vec{V}}(x,y) + pS_{\vec{C}}(x,y)\right]\\
&-4\int\int_{\mathbb{R}^2} S_{\vec{V}}\mathrm{d}S_{\vec{V}}(x,y)\,.
\end{aligned}$$

Now apply Lemma 3.3.

$$\begin{aligned}
=& 4\Bigg[\int\int_{\mathbb{R}^2} (1-p)^2\,S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) + \int\int_{\mathbb{R}^2} (1-p)\,pS_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\\
&+\int\int_{\mathbb{R}^2} p\,(1-p)\,S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) + \int\int_{\mathbb{R}^2} p^2 S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&-4\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y)\\
=&4\Big[\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) - 2p\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y)\\
&+p^2\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) + p\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\\
&-p^2\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y) + p\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y)\\
&-p^2\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) + p^2\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\Big]\\
&-4\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y)\\
=&4\Bigg[p^2\Big(\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) - \int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\\
&-\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y) + \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y)\Big)\\
&+p\Big(\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{C}}(x,y) + \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y)\\
&-2\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{\vec{V}}(x,y)\Big)\Bigg]\\
=&4\left(a_\tau p^2 + b_\tau p\right).
\end{aligned}
$$

Using the same strategy above, we can use it solve for the bias of Rho under a mixture.

$$
\begin{aligned}
\text{Bias}_\rho(p) =&\rho_{\vec{M}} - \rho_{\vec{V}}\\
=&12\int\int_{\mathbb{R}^2} S_{\vec{M}}(x,y)\,\mathrm{d}S_{M_1}(x)\,\mathrm{d}S_{M_2}(y) - 3\\
&-\left(12\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y) - 3\right)\\
=&12\int\int_{\mathbb{R}^2}\left[(1-p)\,S_{\vec{V}}(x,y) + pS_{\vec{C}}(x,y)\right]\\
&\qquad\qquad \mathrm{d}\left[(1-p)\,S_{V_1}(x) + pS_{C_1}(x)\right]\mathrm{d}\left[(1-p)\,S_{V_2}(y) + pS_{C_2}(y)\right]\\
&-\left(12\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\right)\\
=&12\Big[\int\int_{\mathbb{R}^2}(1-p)^3\,S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+\int\int_{\mathbb{R}^2}(1-p)^2\,pS_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)
\end{aligned}
$$

$$
\begin{aligned}
&+\int\int_{\mathbb{R}^2}(1-p)^2\,pS_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+\int\int_{\mathbb{R}^2}(1-p)\,p^2S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&+\int\int_{\mathbb{R}^2}(1-p)^2\,pS_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+\int\int_{\mathbb{R}^2}p^2(1-p)\,S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&+\int\int_{\mathbb{R}^2}p^2(1-p)\,S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+\int\int_{\mathbb{R}^2}p^3S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)\Big]\\
&-12\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y).
\end{aligned}
$$

$$
\begin{aligned}
=&12\Big[\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)-3p\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+3p^2\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)-p^3\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+p\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)-2p^2\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&+p^3\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)+p\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&-2p^2\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)+p^3\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+p^2\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)-p^3\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&+p\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)-2p^2\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+p^3\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)+p^2\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&-p^3\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)+p^2\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(Y)\\
&-p^3\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)+p^3\int\int_{\mathbb{R}^2}S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)\Big]\\
&-12\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y).
\end{aligned}
$$

$$
\begin{aligned}
=&12\Big[p^3\Big(-\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)+\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&+\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)-\int\int_{\mathbb{R}^2}S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)
\end{aligned}
$$

$$
\begin{aligned}
&+\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y) - \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&-\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y) + \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)\Bigg)\\
&+p^2\Bigg(3\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y) - 2\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&-2\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y) + \int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&-2\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y) + \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)\\
&+\int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y)\Bigg)\\
&+p\Bigg(\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y) - 3\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\\
&+\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y) + \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y)\Bigg)\Bigg]\\
=&12\left(a_\rho p^3 + b_\rho p^2 + c_\rho p\right).
\end{aligned}
$$

■

3.3 General Behavior of Quadratics and Cubics

In the case of Tau, the final form expression for the bias is a quadratic without a constant term. In the case of Rho, the final form expression for the bias is a cubic without a constant term. For the interest of statistical application, it is common to want the sign of the bias. That is, whether the contamination will result in a positive or negative bias. By analyzing and characterizing all the different root behaviors, we can easily find the sign of the bias. Because the polynomials have no constant terms, our analysis will be simplified by having a root at zero.

Proposition 3.5. *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. There are five cases below with subcases for some.*

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.*

   (a) $a \geq 0$ *and* $b \geq 0$ *(but not both equal to zero).*

   (b) $0 < -a \leq b$.

3. $f(p) < 0$ *for all* p *in* $(0, 1)$. *This has two subcases.*

   (a) $a \leq 0$ *and* $b \leq 0$ *(but not both equal to zero).*

   (b) $b \leq -a < 0$.

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* $-a < b < 0$.

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. There are 7 cases below with multiple subcases for some.*

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.*

   (a) $a > 0$, $|b| < 2\sqrt{ac}$, $c > 0$.

   (b) $a > 0$, $|b| > 2\sqrt{ac}$, $b \leq -2a$, $c \geq a$, $a + b + c \geq 0$.

   (c) $a > 0$, $b > 0$, $c > 0$, $|b| > 2\sqrt{ac}$.

   (d) $a < 0$, $c > 0$, $a + b + c \geq 0$.

3. $f(p) < 0$ *for all* p *in* $(0, 1)$. *This has four subcases.*

(a) $a < 0$, $|b| < 2\sqrt{ac}$, $c < 0$.

(b) $a < 0$, $|b| > 2\sqrt{ac}$, $b \geq -2a$, $c \leq a$, $a + b + c \leq 0$.

(c) $a < 0$, $b < 0$, $c < 0$, $|b| > 2\sqrt{ac}$.

(d) $a > 0$, $c < 0$, $a + b + c \leq 0$.

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.*

   (a) $a > 0$, $b \leq -a$, $c > 0$, $|b| > 2\sqrt{ac}$, $a + b + c \leq 0$.

   (b) $a < 0$, $c > 0$, $a + b + c \leq 0$.

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.*

   (a) $a < 0$, $b \geq -a$, $c < 0$, $|b| > 2\sqrt{ac}$, $a + b + c \geq 0$.

   (b) $a > 0$, $c < 0$, $a + b + c \geq 0$.

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* $a < 0$, $-2a \geq b \geq 0$, $c \geq a$, $|b| > 2\sqrt{ac}$, $a + b + c \leq 0$.

The proof of these cases are elementary, yet tedious, and will be skipped. The application of these cases will be shown graphically in Chapter 4 when we study the Marshall-Olkin distribution.

CHAPTER 4

THE MARSHALL-OLKIN DISTRIBUTION

4.1 INTRODUCING THE MARSHALL-OLKIN DISTRIBUTION

Albert W. Marshall and Ingram Olkin introduced the Marshall-Olkin (MO) distribution in 1967. They claimed that they were the first to have proposed a multivariate exponential distribution with an applicable use. This distribution arises from "shock models" and its ability to model the failing of a two-component system [9]. Define three independent random variables, where $Z_1 \sim \text{Exp}(\lambda_1)$, $Z_2 \sim \text{Exp}(\lambda_2)$, and $Z_3 \sim \text{Exp}(\lambda_3)$ represent the time until 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 $X = \min\{Z_1, Z_3\}$ and $Y = \min\{Z_2, Z_3\}$. These new random variables represent the lifetimes of component one and component two, respectively. We can now find the joint survival function. The survival function will allow for more convenient calculations than working with the joint CDF.

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

The visual of an exponential distribution curve is very well known. However, the MO distribution is difficult to visualize because of the "max" term. Let's show an example with three different sets of parameters.

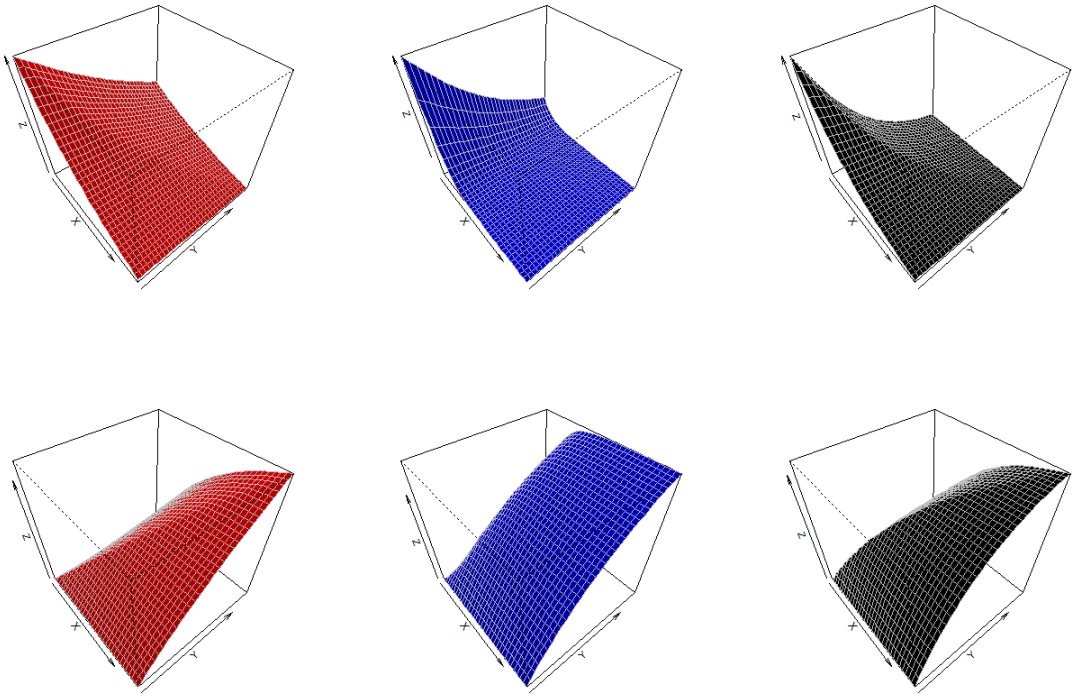

Figure 4.1: The red graph shows $\lambda_1 = 1$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.5$; the blue graph shows $\lambda_1 = 10$, $\lambda_2 = 1$, and $\lambda_3 = 0.5$; the black graph shows $\lambda_1, = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 2$; the top row shows the survival function and the bottom row shows the CDF.

4.2 Rank Correlation of the Marshall-Olkin Distribution

The population definitions have been defined in multiple ways for both Spearman's Rho and Kendall's Tau. We can now use them to define these rank correlation methods for the MO distribution. We will begin with Spearman's Rho, because for Kendall's Tau we have to invoke a Lemma that will be defined soon.

Theorem 4.1. *Spearman's Rho for the Marshall-Olkin distribution can be defined as*

$$\rho_S = \frac{3\alpha_1\alpha_2}{2\alpha_1 + 2\alpha_2 - \alpha_1\alpha_2}$$

where $\alpha_1 = \frac{\lambda_3}{\lambda_1+\lambda_3}$ *and* $\alpha = \frac{\lambda_3}{\lambda_2+\lambda_3}$.

Proof. Recall Theorem 2.14, where Spearman's Rho is defined as

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

We can use this definition by finding the copula for the MO distribution. To find the copula, make the following observations. Notice that $\max\{x,y\} = x + y - \min\{x,y\}$. The marginal survival functions for X and Y are $S_X(x) = \exp\{-(\lambda_1+\lambda_3)x\}$ and $S_Y(y) = \exp\{-(\lambda_2+\lambda_3)y\}$, respectively. To make future calculations easier, let $\alpha_1 = \frac{\lambda_3}{\lambda_1+\lambda_3}$ and $\alpha_2 = \frac{\lambda_3}{\lambda_2+\lambda_3}$. Using these observations, we will start with the survival function, manipulate it, then apply the copula transformation.

$$\begin{aligned} S_{X,Y}(x,y) &= \exp\{-(\lambda_1 x + \lambda_2 y + \lambda_3 \max\{x,y\})\} \\ &= \exp\{-(\lambda_1+\lambda_3)x - (\lambda_2+\lambda_3)y + \lambda_3 \min\{x,y\}\} \\ &= \exp\{-(\lambda_1+\lambda_3)x\}\exp\{-(\lambda_2+\lambda_3)y\}\exp\{\lambda_3\min\{x,y\}\} \\ &= S_X(x)\,S_Y(y)\min\{\exp\{\lambda_3 x\},\ \exp\{\lambda_3 y\}\} \end{aligned}$$

Using the fact that $\exp\{\lambda_3 x\} = S_X(x)^{-\alpha_1}$ and $\exp\{\lambda_3 y\} = S_Y(y)^{-\alpha_2}$, now apply the copula transformation $U = F_X(x)$, $V = F_Y(y)$. Let $\overline{C}$ be the survival copula.

$$\begin{aligned} \overline{C}(u,v) &= u\,v\min\left\{u^{-\alpha_1}, v^{-\alpha_2}\right\} \\ &= \min\left\{v\,u^{1-\alpha_1}, u\,v^{1-\alpha_2}\right\} \\ &= \begin{cases} v\,u^{1-\alpha_1} & , u^{\alpha_1} \geq v^{\alpha_2} \\ u\,v^{1-\alpha_2} & , u^{\alpha_1} \leq v^{\alpha_2} \end{cases} \end{aligned}$$

The MO distribution contains a singular component. This is a concentrated cluster of density on the line $y = x$ (or the line $u^{\alpha_1} = v^{\alpha_2}$). Because of this, the integral in Theorem 2.14 will have to be split into two parts to account for this cluster. To visualize this, consider Figure 4.2 below, a simulation done using the software R. Now that the copula has been defined, apply to Theorem 2.14.

$$\begin{aligned} \rho_S &= 12\int\int_{[0,1]^2} C(u,v)\,\mathrm{d}u\mathrm{d}v - 3 \\ &= 12\int_0^1 \left[\int_0^{u^{\alpha_1/\alpha_2}} v\,u^{1-\alpha_1}\mathrm{d}v + \int_{u^{\alpha_1/\alpha_2}}^1 u\,v^{1-\alpha_2}\mathrm{d}v\right]\mathrm{d}u - 3 \end{aligned}$$

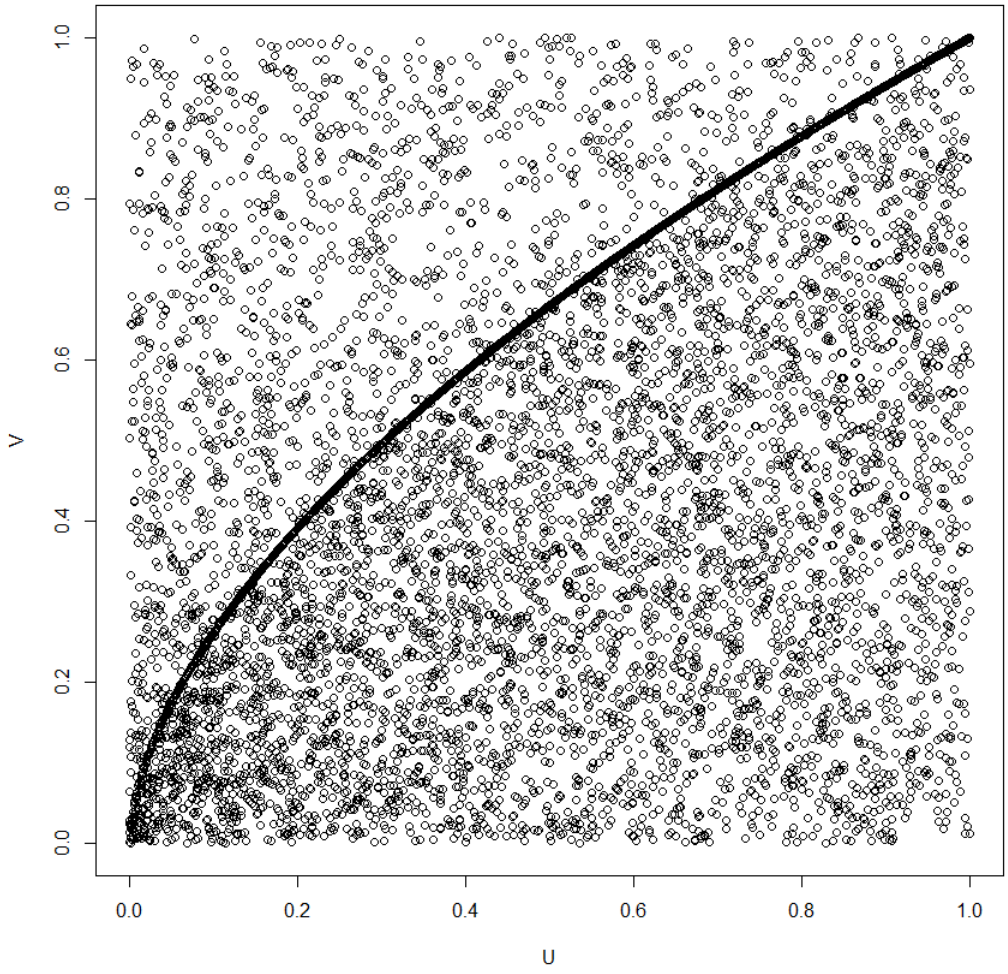

Figure 4.2: Simulation of the Marshall-Olkin survival copula using parameters: $\lambda_1 = 0.7$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.5$.

$$
\begin{aligned}
&= 12\int_0^1 \left[u^{1-\alpha_1}\frac{1}{2}v^2\Big|_0^{u^{\alpha_1/\alpha_2}} + u\,\frac{v^{2-\alpha_2}}{2-\alpha_2}\Big|_{u^{\alpha_1/\alpha_2}}^1 \right] \mathrm{d}u - 3 \\
&= 12\int_0^1 \left[\frac{u^{1-\alpha_1}u^{2\alpha_1/\alpha_2}}{2} + \left(\frac{u}{2-\alpha_2} - \frac{u\,u^{(\alpha_1/\alpha_2)(2-\alpha_2)}}{2-\alpha_2} \right) \right] \mathrm{d}u - 3 \\
&= 12\int_0^1 \left[u^{2\alpha_1/\alpha_2-\alpha_1+1}\left(\frac{1}{2} - \frac{1}{2-\alpha_2} \right) + \frac{u}{2-\alpha_2} \right] \mathrm{d}u - 3 \\
&= 12\left[\left(\frac{1}{2} - \frac{1}{2-\alpha_2} \right) \frac{u^{2\alpha_1/\alpha_2-\alpha_1+2}}{2\alpha_1/\alpha_2-\alpha_1+2}\Big|_0^1 + \frac{u^2}{2\,(2-\alpha_2)}\Big|_0^1 \right] - 3 \\
&= 12\left[\left(\frac{1}{2} - \frac{1}{2-\alpha_2} \right) \left(\frac{1}{2\alpha_1/\alpha_2-\alpha_1+2} \right) + \frac{1}{2\,(2-\alpha_2)} \right] - 3 \\
&= 12\left[\frac{2-\alpha_2-2+2\alpha_1/\alpha_2-\alpha_1+2}{2\,(2-\alpha_2)\,(2\alpha_1/\alpha_2-\alpha_1+2)} \right] - 3 \\
&= \frac{-12\alpha_2+24\alpha_1/\alpha_2-12\alpha_1+24}{2\,(2-\alpha_2)\,(2\alpha_1/\alpha_2-\alpha_1+2)} - \frac{6\,(2-\alpha_2)\,(2\alpha_1\alpha_2-\alpha_1+2)}{2\,(2-\alpha_2)\,(2\alpha_1/\alpha_2-\alpha_1+2)} \\
&= \frac{12\alpha_1-6\alpha_1\alpha_2}{2\,(2-\alpha_2)\,(2\alpha_1/\alpha_2-\alpha_1+2)} \\
&= \frac{3\alpha_1}{2\alpha_1/\alpha_2-\alpha_1+2} \\
&= \frac{3\alpha_1\alpha_2}{2\alpha_1+2\alpha_2-\alpha_1\alpha_2}
\end{aligned}
$$

In terms of lambda, we can define it as

$$
\rho_S = \frac{3\left(\frac{\lambda_3}{\lambda_1+\lambda_3}\right)\left(\frac{\lambda_3}{\lambda_2+\lambda_3}\right)}{2\left(\frac{\lambda_3}{\lambda_1+\lambda_3}\right)+2\left(\frac{\lambda_3}{\lambda_2+\lambda_3}\right)-\left(\frac{\lambda_3}{\lambda_1+\lambda_3}\right)\left(\frac{\lambda_3}{\lambda_2+\lambda_3}\right)}
$$

■

For the Kendall's Tau definition of the MO distribution, we will take the same route as Spearman's Rho by using the copula. We will first define a lemma to help us evaluate integrals with singular components. The original lemma was proved by Li, X. in 2002 [10] in terms of copulas. Here, we state it in terms of CDFs.

Lemma 4.2. *Let $F_{X,Y}(x,y)$ and $G_{X,Y}(x,y)$ be differentiable CDFs. Then*

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

Using Lemma 4.2, we will now derive Kendall's Tau for the MO distribution.

Theorem 4.3. *Kendall's Tau for the Marshall-Olkin distribution can be derived as*

$$\tau = \frac{\alpha_1 \alpha_2}{\alpha_1 + \alpha_2 - \alpha_1 \alpha_2}$$

where $\alpha_1 = \frac{\lambda_3}{\lambda_1+\lambda_3}$ *and* $\alpha = \frac{\lambda_3}{\lambda_2+\lambda_3}$.

Proof.

$$\begin{aligned}
\tau &= 4 \int \int_{[0,1]^2} C(u,v)\, \mathrm{d}C(u,v) - 1 \\
&= 4 \left[\frac{1}{2} - \int \int_{[0,1]^2} \frac{\partial}{\partial u} C(u,v)\, \frac{\partial}{\partial v} C(u,v)\, \mathrm{d}u\mathrm{d}v \right] - 1 \\
&= 4 \left[\frac{1}{2} - \int_0^1 \left(\int_0^{u^{\alpha_1/\alpha_2}} \frac{\partial}{\partial u} u^{1-\alpha_1} v \frac{\partial}{\partial v} u^{1-\alpha_1} v \mathrm{d}v + \int_{u^{\alpha_1/\alpha_2}}^1 \frac{\partial}{\partial u} u v^{1-\alpha_2} \frac{\partial}{\partial v} u v^{1-\alpha_2} \mathrm{d}v \right) \mathrm{d}u \right] - 1 \\
&= 4 \left[\frac{1}{2} - \int_0^1 \left(\int_0^{u^{\alpha_1/\alpha_2}} (1-\alpha_1)\, u^{-\alpha_1} v u^{1-\alpha_1} \mathrm{d}v - \int_{u^{\alpha_1/\alpha_2}}^1 u v^{1-\alpha_2} (1-\alpha_2)\, v^{-\alpha_2} \mathrm{d}v \right) \mathrm{d}u \right] - 1 \\
&= 4 \left[\frac{1}{2} - \int_0^1 \left(\frac{1-\alpha_1}{2} u^{1-2\alpha_1} v^2 \Big|_0^{u^{\alpha_1/\alpha_2}} + \frac{1-\alpha_2}{2-2\alpha_2} u v^{2-2\alpha_2} \Big|_{u^{\alpha_1/\alpha_2}}^1 \right) \mathrm{d}u \right] - 1 \\
&= 4 \left[\frac{1}{2} - \int_0^1 \left(\frac{1-\alpha_1}{2} u^{2\alpha_1/\alpha_2 - 2\alpha_1 + 1} + \frac{1}{2} u - \frac{1}{2} u^{2\alpha_1/\alpha_2 - 2\alpha_1 + 1} \right) \mathrm{d}u \right] - 1 \\
&= 2 \left(1 + \alpha_1 \frac{u^{2\alpha_1/\alpha_2 - 2\alpha_1 + 2}}{2\alpha_1/\alpha_2 - 2\alpha_1 + 2} \Big|_0^1 - \frac{1}{2} u^2 \Big|_0^1 \right) - 1 \\
&= 2 \left(1 + \frac{\alpha_1}{2\alpha_1/\alpha_2 - 2\alpha_1 + 2} - \frac{1}{2} \right) - 1 \\
&= \frac{\alpha_1}{\alpha_1/\alpha_2 - \alpha_1 + 1} \\
&= \frac{\alpha_1 \alpha_2}{\alpha_1 + \alpha_2 - \alpha_1 \alpha_2}
\end{aligned}$$

In terms of lambda, we can define it as

$$\tau = \frac{\left(\frac{\lambda_3}{\lambda_1+\lambda_3}\right)\left(\frac{\lambda_3}{\lambda_2+\lambda_3}\right)}{\frac{\lambda_3}{\lambda_1+\lambda_3} + \frac{\lambda_3}{\lambda_2+\lambda_3} + \left(\frac{\lambda_3}{\lambda_1+\lambda_3}\right)\left(\frac{\lambda_3}{\lambda_2+\lambda_3}\right)}$$

■

For both of the previous proofs we could have arrived at the same solutions by using the CDF versions. We now have all the information we need to find the bias of the Marshall-Olkin distribution in terms of the mixing proportion, p.

4.3 Bias of Rank Correlation Under a Mixture Model

Consider two populations that follow a bivariate Marshall-Olkin distribution, one being a valid population and the other contaminating. Define the parameters for each to be

$$(X,Y)_V \sim \text{MO}\left(\lambda_{V1},\lambda_{V2},\lambda_{V3}\right) \qquad \text{and} \qquad (X,Y)_C \sim \text{MO}\left(\lambda_{C1},\lambda_{C2},\lambda_{C3}\right).$$

To find the bias in terms of the mixing proportion we can anticipate there will be a lot of integrals to solve in order to arrive at the desired analytical solution. However, by noticing a pattern for each integral in the coefficients from Theorem 3.4 we can formulate a general integral to solve. This is because each integral follows a similar form, the only difference being valid or contaminated survival function. Furthermore, some of the integrals are equivalent to the ones solved in both Theorem 4.1 and 4.3. Recall the bias from Theorem 3.4,

$$\text{Bias}_\tau\left(p\right) = \tau_{\vec{M}} - \tau_{\vec{V}} = 4\left(a_\tau p^2 + b_\tau p\right)$$

where

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

In the coefficients above, there are four unique integrals. Instead of solving all four integrals directly, we can solve a general integral. Beginning with the bias for Tau, we will generalize the integral using placeholder parameters, $(\alpha_1,\alpha_2,\alpha_3)$ and $(\beta_1,\beta_2,\beta_3)$. Think of these as being the valid and contaminated parameters, allowing us to rearrange them without loss of generality.

$$\int\int_{\mathbb{R}^2} S_\alpha\left(x,y\right)\mathrm{d}S_\beta\left(x,y\right)$$

$$
\begin{aligned}
=&\int_0^\infty \int_0^\infty e^{-\alpha_1 x-\alpha_2 y-\alpha_3 \max\{x,y\}}\mathrm{d}e^{-\beta_1 x-\beta_2 y-\beta_3 \max\{x,y\}}\\
=&\frac{1}{2}-\int_0^\infty \int_0^\infty \frac{\partial}{\partial x}e^{-\alpha_1 x-\alpha_2 y-\alpha_3 \max\{x,y\}}\frac{\partial}{\partial y}e^{-\beta_1 x-\beta_2 y-\beta_3 \max\{x,y\}}\mathrm{d}x\mathrm{d}y\\
=&\frac{1}{2}-\int_0^\infty \int_y^\infty (\alpha_1+\alpha_3)\,\beta_2 e^{-(\alpha_1+\alpha_3+\beta_1+\beta_3)x-(\alpha_2+\beta_2)y}\mathrm{d}x\mathrm{d}y\\
&-\int_0^\infty \int_x^\infty \alpha_1\,(\beta_2+\beta_3)\,e^{-(\alpha_1+\beta_1)x-(\alpha_2+\alpha_3+\beta_2+\beta_3)y}\mathrm{d}y\mathrm{d}x\\
=&\frac{1}{2}+\int_0^\infty \left[\frac{(\alpha_1+\alpha_3)\,\beta_2}{\alpha_1+\alpha_3+\beta_1+\beta_3}e^{-(\alpha_1+\alpha_3+\beta_1+\beta_3)x-(\alpha_2+\beta_2)y}\Big|_y^\infty\right]\mathrm{d}y\\
&+\int_0^\infty \left[\frac{\alpha_1\,(\beta_2+\beta_3)}{\alpha_2+\alpha_3+\beta_2+\beta_3}e^{-(\alpha_1+\beta_1)x-(\alpha_2+\alpha_3+\beta_2+\beta_3)y}\Big|_x^\infty\right]\mathrm{d}x\\
=&\frac{1}{2}-\frac{(\alpha_1+\alpha_3)\,\beta_2}{\alpha_1+\alpha_3+\beta_1+\beta_3}\int_0^\infty e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3)y}\mathrm{d}y\\
&-\frac{\alpha_1\,(\beta_2+\beta_3)}{\alpha_2+\alpha_3+\beta_2+\beta_3}\int_0^\infty e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3)x}\mathrm{d}x\\
=&\frac{1}{2}-\frac{(\alpha_1+\alpha_3)\,\beta_2}{\alpha_1+\alpha_3+\beta_1+\beta_3}\left(\frac{-1}{\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3}e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3)y}\Big|_0^\infty\right)\\
&-\frac{\alpha_1\,(\beta_2+\beta_3)}{\alpha_2+\alpha_3+\beta_2+\beta_3}\left(\frac{-1}{\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3}e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3)x}\Big|_0^\infty\right)\\
=&\frac{1}{2}-\frac{1}{\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_2+\beta_3}\left(\frac{(\alpha_1+\alpha_3)\,\beta_2}{\alpha_1+\alpha_3+\beta_1+\beta_3}+\frac{\alpha_1\,(\beta_2+\beta_3)}{\alpha_2+\alpha_3+\beta_2+\beta_3}\right)
\end{aligned}
$$

Observe that if the survival functions are the same for both the integrand and the differential, the general solution reduces to the form

$$\int\int_{\mathbb{R}^2} S_\alpha\,(x,y)\,\mathrm{d}S_\alpha\,(x,y)=\frac{1}{2}-\frac{\alpha_1+\alpha_2}{4\alpha_1+4\alpha_2+4\alpha_3}.$$

Now that the general integral is solved, the last step is to plug in the correct parameters for each corresponding integral. The coefficients for the bias of Tau can be defined as

$$
\begin{aligned}
a_\tau =&\frac{1}{2}-\frac{\lambda_{V1}+\lambda_{V2}}{4\lambda_{V1}+4\lambda_{V2}+4\lambda_{V3}}\\
&-\left(\frac{1}{2}-\frac{1}{\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+\lambda_{C3}}\left(\frac{(\lambda_{V1}+\lambda_{V3})\,\lambda_{C2}}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}}+\frac{\lambda_{V1}\,(\lambda_{V2}+\lambda_{V3})}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right)\right)\\
&-\left(\frac{1}{2}-\frac{1}{\lambda_{C1}+\lambda_{C2}+\lambda_{C3}+\lambda_{V1}+\lambda_{V2}+\lambda_{V3}}\left(\frac{(\lambda_{C1}+\lambda_{C3})\,\lambda_{V2}}{\lambda_{C1}+\lambda_{C3}+\lambda_{V1}+\lambda_{V3}}+\frac{\lambda_{C1}\,(\lambda_{C2}+\lambda_{C3})}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right)\right)\\
&+\frac{1}{2}-\frac{\lambda_{C1}+\lambda_{C2}}{4\lambda_{C1}+4\lambda_{C2}+4\lambda_{C3}}\\
b_\tau =&\left(\frac{1}{2}-\frac{1}{\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+\lambda_{C3}}\left(\frac{(\lambda_{V1}+\lambda_{V3})\,\lambda_{C2}}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}}+\frac{\lambda_{V1}\,(\lambda_{V2}+\lambda_{V3})}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right)\right)\\
&+\left(\frac{1}{2}-\frac{1}{\lambda_{C1}+\lambda_{C2}+\lambda_{C3}+\lambda_{V1}+\lambda_{V2}+\lambda_{V3}}\left(\frac{(\lambda_{C1}+\lambda_{C3})\,\lambda_{V2}}{\lambda_{C1}+\lambda_{C3}+\lambda_{V1}+\lambda_{V3}}+\frac{\lambda_{C1}\,(\lambda_{C2}+\lambda_{C3})}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right)\right)
\end{aligned}
$$

$$- 2\left(\frac{1}{2} - \frac{\lambda_{V1}+\lambda_{V2}}{4\lambda_{V1}+4\lambda_{V2}+4\lambda_{V3}}\right).$$

After elementary algebra steps we can arrive at the final expression for the coefficients for the bias of Tau as

$$\begin{aligned}
a_\tau =& \frac{1}{\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+\lambda_{C3}} \\
&\cdot\left(\frac{(\lambda_{V1}+\lambda_{V3})\,\lambda_{C2}+(\lambda_{C1}+\lambda_{C3})\,\lambda_{V2}}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{\lambda_{V1}\,(\lambda_{V2}+\lambda_{V3})+\lambda_{C1}\,(\lambda_{C2}+\lambda_{C3})}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right) \\
&- \left(\frac{\lambda_{V1}+\lambda_{V2}}{4\lambda_{V1}+4\lambda_{V2}+4\lambda_{V3}} + \frac{\lambda_{C1}+\lambda_{C2}}{4\lambda_{C1}+4\lambda_{C2}+4\lambda_{C3}}\right) \\
b_\tau =& \frac{\lambda_{V1}+\lambda_{V2}}{2\lambda_{V1}+2\lambda_{V2}+2\lambda_{V3}} \\
&- \frac{1}{\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+\lambda_{C3}} \\
&\cdot\left(\frac{(\lambda_{V1}+\lambda_{V3})\,\lambda_{C2}+(\lambda_{C1}+\lambda_{C3})\,\lambda_{V2}}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{\lambda_{V1}\,(\lambda_{V2}+\lambda_{V3})+\lambda_{C1}\,(\lambda_{C2}+\lambda_{C3})}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right).
\end{aligned}$$

Solving for the coefficient for the bias of Rho is achieved following the same process. There are eight unique integrals in the case of Rho, so we will also solve a general integral. Recall the bias from Theorem 3.4,

$$\text{Bias}_\rho(p) = \rho_{\vec{M}} - \rho_{\vec{V}} = 12\left(a_\rho p^3 + b_\rho p^2 + c_\rho p\right)$$

where

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

$$c_\rho = -3\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y) + \int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{C_2}(y)$$
$$+\int\int_{\mathbb{R}^2} S_{\vec{V}}(x,y)\,\mathrm{d}S_{C_1}(x)\,\mathrm{d}S_{V_2}(y) + \int\int_{\mathbb{R}^2} S_{\vec{C}}(x,y)\,\mathrm{d}S_{V_1}(x)\,\mathrm{d}S_{V_2}(y).$$

For this general integral we will introduce three placeholder parameters, because each integral will involve three different arrangements of functions. These parameters will be $(\alpha_1,\alpha_2,\alpha_3)$, $(\beta_1,\beta_2,\beta_3)$, and $(\gamma_1,\gamma_2,\gamma_3)$.

$$\begin{aligned}
&\int\int_{\mathbb{R}^2} S_\alpha(x,y)\,\mathrm{d}S_\beta(x)\,\mathrm{d}S_\gamma(y)\\
=&\int\int_{\mathbb{R}^2} S_\alpha(x,y)\,f_\beta(x)\,f_\gamma(y)\,\mathrm{d}x\mathrm{d}y\\
=&\int_0^\infty\int_0^\infty e^{-\alpha_1 x-\alpha_2 y-\alpha_3\max\{x,y\}}(\beta_1+\beta_3)\,e^{-(\beta_1+\beta_3)x}(\gamma_2+\gamma_3)\,e^{-(\gamma_2+\gamma_3)y}\mathrm{d}x\mathrm{d}y\\
=&\int_0^\infty\int_0^\infty(\beta_1+\beta_3)(\gamma_2+\gamma_3)\,e^{-(\alpha_1+\alpha_3+\beta_1+\beta_3)x-(\alpha_2+\gamma_2+\gamma_3)y}\mathrm{d}x\mathrm{d}y\\
&+\int_0^\infty\int_0^\infty(\beta_1+\beta_3)(\gamma_2+\gamma_3)\,e^{-(\alpha_1+\beta_1+\beta_3)x-(\alpha_2+\alpha_3+\gamma_2+\gamma_3)y}\mathrm{d}y\mathrm{d}x\\
=&-\int_0^\infty\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_1+\alpha_3+\beta_1+\beta_3}\left(e^{-(\alpha_1+\alpha_3+\beta_1+\beta_3)x-(\alpha_2+\gamma_2+\gamma_3)y}\Big|_y^\infty\right)\mathrm{d}y\\
&-\int_0^\infty\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_2+\alpha_3+\gamma_2+\gamma_3}\left(e^{-(\alpha_1+\beta_1+\beta_3)x-(\alpha_2+\alpha_3+\gamma_2+\gamma_3)y}\Big|_x^\infty\right)\mathrm{d}x\\
=&\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_1+\alpha_3+\beta_1+\beta_3}\int_0^\infty e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3)y}\mathrm{d}y\\
&+\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_2+\alpha_3+\gamma_2+\gamma_3}\int_0^\infty e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3)x}\mathrm{d}x\\
=&\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_1+\alpha_3+\beta_1+\beta_3}\left(\frac{-1}{\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3}e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3)y}\Big|_0^\infty\right)\\
+&\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_2+\alpha_3+\gamma_2+\gamma_3}\left(\frac{-1}{\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3}e^{-(\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3)x}\Big|_0^\infty\right)\\
=&\frac{(\beta_1+\beta_3)(\gamma_2+\gamma_3)}{\alpha_1+\alpha_2+\alpha_3+\beta_1+\beta_3+\gamma_2+\gamma_3}\left(\frac{1}{\alpha_1+\alpha_3+\beta_1+\beta_3}+\frac{1}{\alpha_2+\alpha_3+\gamma_2+\gamma_3}\right)
\end{aligned}$$

Observe that if the survival function and marginal densities are just valid or just contaminating then the integral reduces to

$$\int\int_{\mathbb{R}^2} S_\alpha(x,y)\,f_\alpha(x)\,f_\alpha(y)\,\mathrm{d}x\mathrm{d}y = \frac{\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}}{4\lambda_{V1}+4\lambda_{V2}+6\lambda_{V3}}.$$

The general integral is solved, so the last step is to plug in the correct coefficients to each placeholder parameter. The resulting coefficients are as follows.

$$
\begin{aligned}
a_\rho = &- \frac{\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}}{4\lambda_{V1}+4\lambda_{V2}+6\lambda_{V3}} \\
&+ \frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{C2}+\lambda_{C3})}{2\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}+\lambda_{C2}+\lambda_{C3}} \left(\frac{1}{2\lambda_{V1}+2\lambda_{V3}} + \frac{1}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right) \\
&+ \frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{\lambda_{V1}+2\lambda_{V2}+2\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} \left(\frac{1}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{1}{2\lambda_{V2}+2\lambda_{V3}}\right) \\
&- \frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{C2}+\lambda_{C3})}{\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+2\lambda_{C3}} \left(\frac{1}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{1}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right) \\
&+ \frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{V2}+\lambda_{V3})}{\lambda_{C1}+\lambda_{C2}+\lambda_{C3}+\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}} \left(\frac{1}{\lambda_{C1}+2\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} + \frac{1}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right) \\
&- \frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{C2}+\lambda_{C3})}{\lambda_{C1}+2\lambda_{C2}+2\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} \left(\frac{1}{\lambda_{C1}+\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} + \frac{1}{2\lambda_{C2}+2\lambda_{C3}}\right) \\
&- \frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{2\lambda_{C1}+\lambda_{C2}+2\lambda_{C3}+\lambda_{V2}+\lambda_{V3}} \left(\frac{1}{2\lambda_{C1}+2\lambda_{C3}} + \frac{1}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right) \\
&+ \frac{\lambda_{C1}+\lambda_{C2}+2\lambda_{C3}}{4\lambda_{C1}+4\lambda_{C2}+6\lambda_{C3}} \\
b_\rho = &\, 3\frac{\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}}{4\lambda_{V1}+4\lambda_{V2}+6\lambda_{V3}} \\
&- 2\frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{C2}+\lambda_{C3})}{2\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}+\lambda_{C2}+\lambda_{C3}} \left(\frac{1}{2\lambda_{V1}+2\lambda_{V3}} + \frac{1}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right) \\
&- 2\frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{\lambda_{V1}+2\lambda_{V2}+2\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} \left(\frac{1}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{1}{2\lambda_{V2}+2\lambda_{V3}}\right) \\
&+ \frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{C2}+\lambda_{C3})}{\lambda_{V1}+\lambda_{V2}+\lambda_{V3}+\lambda_{C1}+\lambda_{C2}+2\lambda_{C3}} \left(\frac{1}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{1}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right) \\
&- 2\frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{V2}+\lambda_{V3})}{\lambda_{C1}+\lambda_{C2}+\lambda_{C3}+\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}} \left(\frac{1}{\lambda_{C1}+2\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} + \frac{1}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right) \\
&+ \frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{C2}+\lambda_{C3})}{\lambda_{C1}+2\lambda_{C2}+2\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} \left(\frac{1}{\lambda_{C1}+\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} + \frac{1}{2\lambda_{C2}+2\lambda_{C3}}\right) \\
&+ \frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{2\lambda_{C1}+\lambda_{C2}+2\lambda_{C3}+\lambda_{V2}+\lambda_{V3}} \left(\frac{1}{2\lambda_{C1}+2\lambda_{C3}} + \frac{1}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right) \\
c_\rho = &- 3\frac{\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}}{4\lambda_{V1}+4\lambda_{V2}+6\lambda_{V3}} \\
&+ \frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{C2}+\lambda_{C3})}{2\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}+\lambda_{C2}+\lambda_{C3}} \left(\frac{1}{2\lambda_{V1}+2\lambda_{V3}} + \frac{1}{\lambda_{V2}+\lambda_{V3}+\lambda_{C2}+\lambda_{C3}}\right) \\
&+ \frac{(\lambda_{C1}+\lambda_{C3})(\lambda_{V2}+\lambda_{V3})}{\lambda_{V1}+2\lambda_{V2}+2\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} \left(\frac{1}{\lambda_{V1}+\lambda_{V3}+\lambda_{C1}+\lambda_{C3}} + \frac{1}{2\lambda_{V2}+2\lambda_{V3}}\right) \\
&+ \frac{(\lambda_{V1}+\lambda_{V3})(\lambda_{V2}+\lambda_{V3})}{\lambda_{C1}+\lambda_{C2}+\lambda_{C3}+\lambda_{V1}+\lambda_{V2}+2\lambda_{V3}} \left(\frac{1}{\lambda_{C1}+2\lambda_{C3}+\lambda_{V1}+\lambda_{V3}} + \frac{1}{\lambda_{C2}+\lambda_{C3}+\lambda_{V2}+\lambda_{V3}}\right)
\end{aligned}
$$

In application, such as coding in the statistical language R, it would be most efficient to create a general formula for each of the eight integrals like we did above. Then the coefficient could be formed by writing the function as linear combination using the correct parameters for each. For the purposes of this thesis, a complete analytical solution is included. The same applies to Tau.

4.4 Root Behavior Cases for the Marshall-Olkin Distribution

In application of the example we introduced in this chapter, a natural question arises. Are all of the cases listed in Section 3.2 possible? By creating all the necessary functions and cases in the software program R, we let the computer generate random parameters until it matches each case. Because we have the analytical solutions, we can calculate the bias exact. We did this for both Tau and Rho. Some cases took much longer than others, but all were accounted for. The cases that took the longest were those that had the most inflection points. In Table 4.1 and 4.2 below there are listed the parameters that satisfied each case. Accompanying those tables we also have graphs for each case using those same parameters. These are important because they illustrate the behavior of the bias and have potential for each application. It is common for researchers wanting to know whether the bias is positive or negative and this can make it easier.

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 4.1: Parameter values for a mixture of Marshall-Olkin distributions that produce each scenario under Tau; these values are used to produce the figures below.

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 4.2: Parameter values for a mixture of Marshall-Olkin distributions that produce each scenario under Rho; these values are used to produce the figures below.

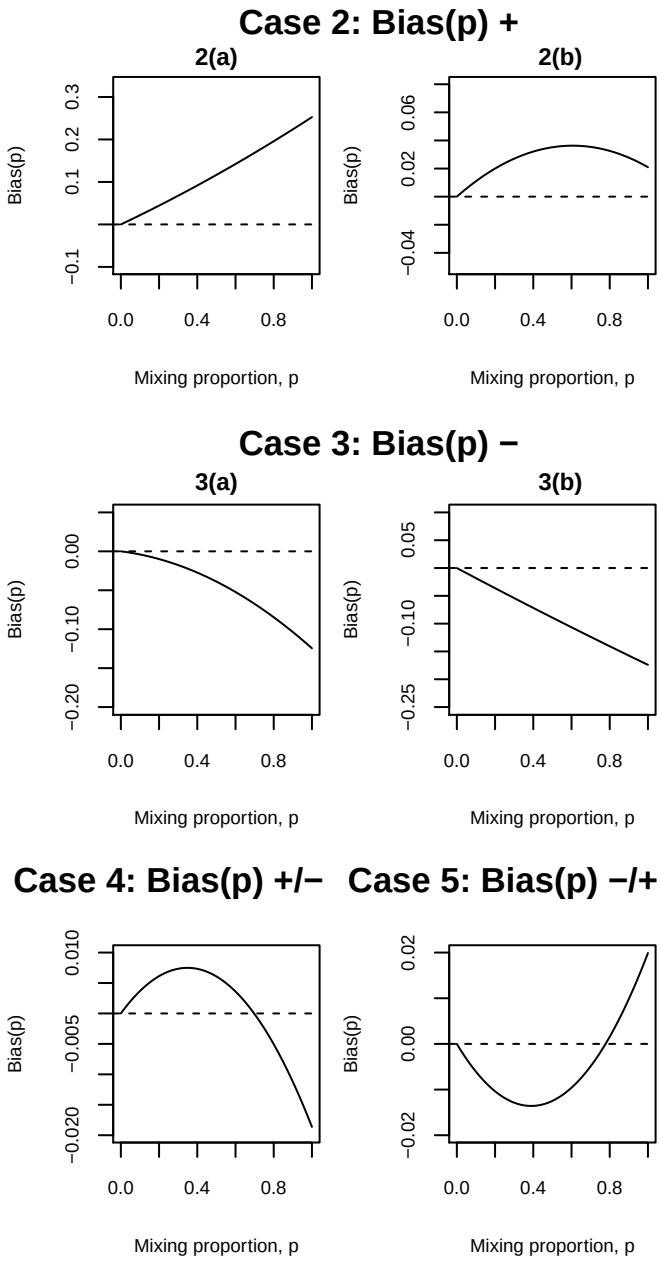

Figure 4.3: Using the values from Table 4.1, these are the possible scenarios for bias in Tau for a mixture of Marshall-Olkin distributions.

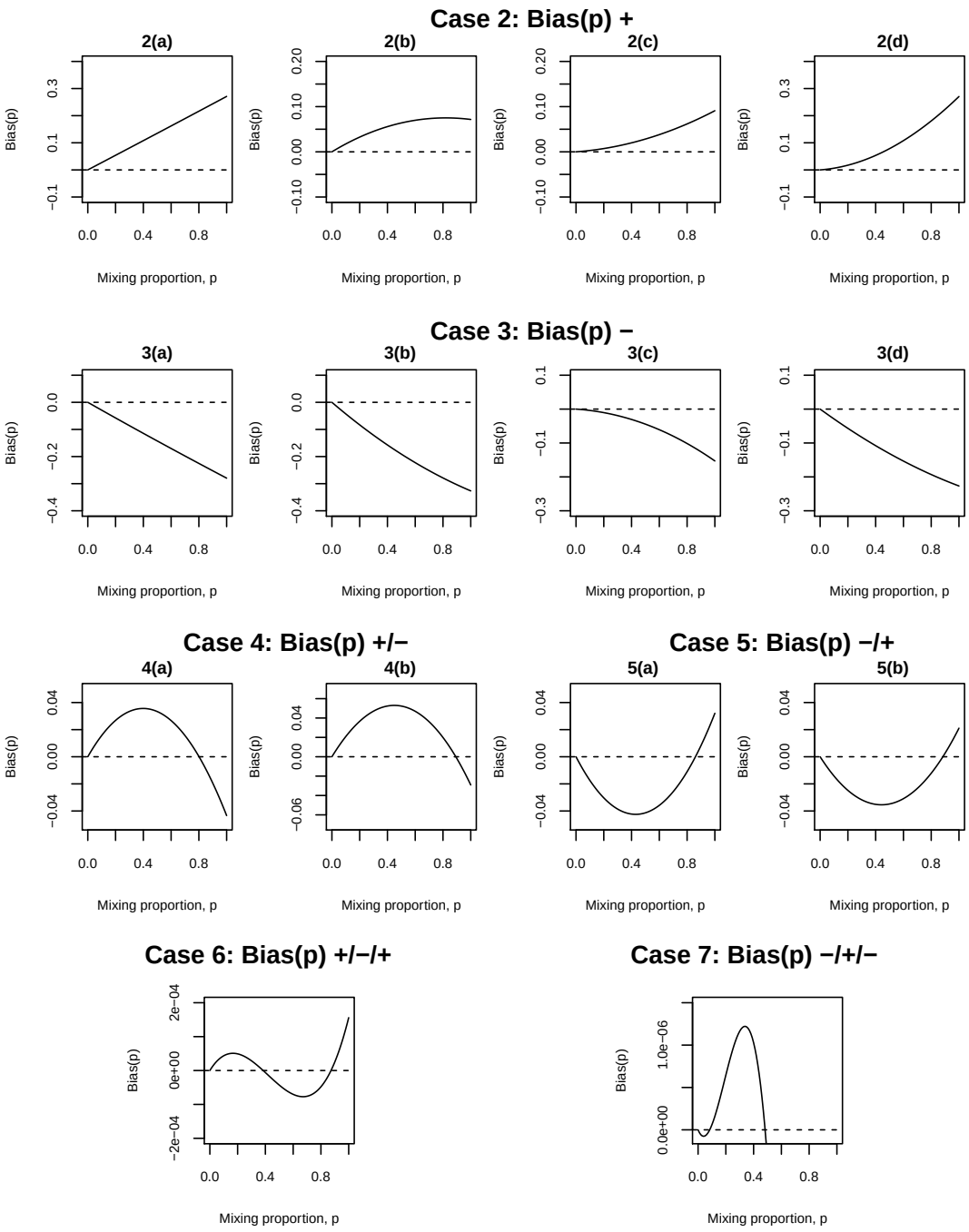

Figure 4.4: Using the values from Table 4.2, these are the possible scenarios for bias in Rho for a mixture of Marshall-Olkin distributions.

CHAPTER 5

CONCLUSION

In conclusion, we have answered the result we set out to show. The bias for both rank correlation methods under mixture models are described as polynomials of degrees two and three. Using that result we elaborated and extended the potential by showing the different cases for both quadratics and cubic. The importance of this comes with application. If a researcher has potentially contaminated data, based on parameters they estimate, they may be able to predict whether the bias is positive or negative relative to the mixing proportion. That is, if their data follows a Marshall-Olkin distribution. However, after laying this foundation for bias analysis of rank correlation under mixtures, it could now be extended using different distributions or even multivariate distributions. A multivariate extension of Rho can be found in Trevor Camper's thesis *Essays on Mixture Models*, although his motivation is quite different [11].

Along with our main results in Theorem 3.4, we also introduced a variety of different techniques throughout that readers may be unaware of. These include copulas, which are used extensively in the financial world, a generalized correlation coefficient, population extensions for Rho and Tau, Li's Theorem for CDFs, and some applications of Riemann-Stieljes integrals. The motivation of this project was to explore more techniques that statisticians can add to their toolbox when solving real-world problems. I believe this was accomplished in my thesis report. In reality, techniques are only as good as their ease to use and, with the R code that I provided and the analytical solution to the problems, it could make a successful statistical tool.

REFERENCES

[1] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, pp. 72–101, Jan. 1904.

[2] W. H. Kruskal, "Ordinal measures of association," *Journal of the American Statistical Association*, vol. 53, pp. 814–861, Dec. 1958.

[3] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, Jun. 1938.

[4] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*. Oxford University Press, fifth ed., 1990.

[5] R. B. Nelsen, *An Introduction to Copulas*. 233 Springer Street, New York, NY 10013, USA: Springer Science+Business Media, Inc., second ed., 2006.

[6] M. Scarsini, "On measures of concordance.," *Stochastica*, vol. 8, no. 3, pp. 201–218, 1984.

[7] A. Sklar, "Random variables, distribution functions, and copulas: A personal look backward and forward," *Lecture Notes-Monograph Series*, vol. 28, 1996.

[8] J. E. Angus, "The probability integral transform and related results," *SIAM Review*, vol. 36, no. 4, pp. 652–654, 1994.

[9] A. W. Marshall and I. Olkin, "A multivariate exponential distribution," *Journal of the American Statistical Association*, vol. 62, pp. 30–44, Mar. 1967.

[10] X. Li, P. Mikusiński, and M. D. Taylor, *Some Integration-by-Parts Formulas Involving 2-Copulas*, pp. 153–159. Dordrecht: Springer Netherlands, 2002.

[11] T. Camper, "Essays on mixture models," Master's thesis, Georgia Southern University, 2019.