UNIVERSITY OF CALIFORNIA
RIVERSIDE

Likelihood-Free Estimation for Some Flexible Families of Distributions

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Matthew Alexander Arvanitis

March 2018

Dissertation Committee:

Dr. Barry C. Arnold, Chairperson
Dr. Aman Ullah
Dr. James M. Flegal

The Dissertation of Matthew Alexander Arvanitis is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to begin by thanking my thesis advisor, Distinguished Professor Barry C. Arnold, whose mentorship is unmatched, and whose vast knowledge and experience has shaped my path to success. He was sure to allow me to pave my own way and carve my own niche in the scientific community. His contribution to this work is truly incalculable. But also, to his wife, Carole, a person whom I have never met, but feel as though I know well, I thank her from the bottom of my heart for her patience and understanding of the countless hours of time Professor Arnold spent away from her on my behalf; my sincerest thanks.

My other four committee members have been supportive in many ways, and have inspired me to work hard and believe in myself: Professors Aman Ullah, Xingping Cui, Matthew Mahutga, and particularly James Flegal.

I thank my professors. Professors James Flegal, Ki-shin Lii, Jun Li, Daniel Jeske, Subir Ghosh, and, of course, Barry Arnold, thank you for the quality teaching; you have advanced my understanding substantially. Thank you, Professor Vyjayanthi Chari, who supported my decision to move from mathematics to statistics and from whom I learned so much. Professor K. B. Reid, who was the first to ignite my sense of scientific and mathematical thought, you are the first character in the story of my higher education. And, Professor Saleem Watson, the most thought-provoking instructor I have ever had, thank you.

Linda Penas and Analisa Flores, you have not only been the finest teaching mentors of my career, you are two individuals with whom I have been able to have the most frank and open discussions. You both have made me the effective teacher that I have become.

To my fellow graduate students, Sakar Sigdel, Arnab Chowdhurry, Roberto Crackel, Ash-

ley Cacho, Rodrigo Gaitan, Ying Liu, and Luke Klein, whose constant challenges to go above and beyond and unconditional collaboration have been essential to my understanding of these difficult concepts and propelled me through the most difficult times over the early years of this work.

It cannot be put into words how thankful I am to my extended family. To my lifelong friend and mentor, Phyllis Bourque, who has supported me even in the most difficult times of her life, you epitomize the meaning of the word, friend. Chrissy and Bonnie, your support, your love, your sensitivity, your emotional strength, and your belief in me, have inspired me to become what I am today. My brother, Christopher, the most intelligent and most talented person I have ever known, has taught me the deepest meaning of integrity; he has supported me in every endeavor of my life; and he has shown me that the perceived impossible is actually often possible. And, to my mother, Nancy, thank you for your unconditional love and support; thank you for your patience, your tolerance of my stress-driven short-tempered moments, and for your willingness to do anything to make me happy. You do too much, and I love you, my family.

To my Mother, Nancy

ABSTRACT OF THE DISSERTATION

Likelihood-Free Estimation for Some Flexible Families of Distributions

by

Matthew Alexander Arvanitis

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, March 2018
Dr. Barry C. Arnold, Chairperson

Historically, the collection of available statistical models for fitting data has been, more or less, restricted to those which are analytically tractable. However, computing power today permits us to use models that, while more complex and often devoid of closed-form distribution or density functions, provide better fits to data. In this thesis, statistical theory, primarily parameter estimation, is developed for three such models: Arnold & Ng's bivariate beta family and its Arnold & Ghosh subfamily of copulas, an 8-parameter family of bivariate Asymmetric Laplace distributions, and a collection of compound random variables. All are distributions whose densities, in general, cannot be written down, but whose realizations can easily be generated via simulation. I apply, and adapt, several methods of likelihood-free statistical inference; including Modified Maximum Likelihood Estimation (MMLE), Approximate Bayesian Computation (ABC), and Markov-Chain Monte-Carlo (MCMC); to achieve various forms of parameter estimates.

For each of the models studied, sub-models were identified and cataloged. In doing so, care is taken to assure that a reasonable balance between the dimensionality of the parameter spaces and the flexibility of the resulting models is maintained. Moreover, in one case, a collection of

sub-models is formed, each of which permits simple parameter estimation, while one of the models from the collection is chosen to provide the best fit according to a particular metric. This is done to simplify parameter estimation through dimension reduction while maintaining a high level of diversity of available models. In other cases, a more direct approach is taken, where the model is selected based on prior knowledge.

Contents

List of Figures

List of Tables

Chapter 1

Likelihood-Free Estimation

1.1 Overview

The Gaussian distribution is arguably the most ubiquitous distribution applied in the statistics discipline. In academics, it is a benchmark for elementary statistics courses, where it is taught to be used in many applications, including those unrelated to the Central Limit Theorem. In practice, possibly because of its ubiquity in college courses, it is also a popular tool. Clearly the Central Limit Theorem provides justification for its application to countless practical problems. We raise the question, however, as to whether it is overused when the Central Limit Theorem does not provide justification to do so. In this thesis, I entertain the argument that Gaussian distributions, as well as other popular distributions, are applied to many problems for which other, namely more complex, models may be justifiably more appropriate. As part of the larger effort to expand the availability of applicable statistical methods, this thesis is aimed at developing statistical theory, particularly parameter estimation, for several models that, with few exceptions, do not have closed-form distribution or density functions, but lend themselves to a wide array of applications. In this chapter, a

comprehensive overview of current methods is studied, setting the stage for the specific applications in subsequent chapters.

Many methods for dealing with analytically intractable distributions have been proposed and implemented. The primary interest herein is on methods that provide a reasonable estimation framework for these distributions. In the following sections, we summarize four such methods: Modified Maximum Likelihood Estimation (MMLE), Approximate Bayes Computation (ABC), Markov Chain Monte Carlo methods, and Statistical Learning techniques. A brief summary of the methods and some examples will be provided.

Consider some family of distributions $\mathcal{F} = \{F(\boldsymbol{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is a set of possible m-dimensional vectors, $\boldsymbol{\theta}$. For the following sections, assume $\boldsymbol{X} := (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K)$ is a collection of iid d-dimensional random vectors, each with distribution $F(\boldsymbol{x}; \boldsymbol{\theta}_0) \in \mathcal{F}$, and $\boldsymbol{x} := (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_K)$ is a realization of $\boldsymbol{X}$. Notation: for any k, denote $\boldsymbol{X}_k = (X_{1k}, X_{2k}, ..., X_{dk})$, and $\boldsymbol{x}_k = (x_{1k}, x_{2k}, ..., x_{dk})$.

1.2 Modified Maximum Likelihood (MMLE)

Often, the marginal distributions of $\boldsymbol{X}_1$ are well-known and tractable. In order for MMLE to be applicable, the maximum likelihood estimates for the marginal parameters must be analytically accessible, and those marginal parameters must exhibit some analytical relationship with the parameter, $\boldsymbol{\theta}$. Plugging in the marginal MLEs results in k_1 equations with k_2 unknowns, $k_1 \leq k_2$. Under typical circumstances, this inequality is strict. Therefore, $k_2 - k_1$ additional equations must be introduced. A common way to do this is to obtain additional estimates using the Method of Moments. This method is most readily illustrated with a simple example.

Example: A Bivariate Poisson Distribution

Suppose $\boldsymbol{X}$ satisfies

$$\boldsymbol{X}_1 \stackrel{d}{=} \begin{pmatrix} Y_1 + Y_3 \\ \\ Y_2 + Y_3 \end{pmatrix}$$

where Y_1, Y_2, and Y_3 are independent Poisson random variables with parameters θ_1, θ_2, and θ_3, respectively. Then the marginal distributions of X_{11} and X_{21} are Poisson with parameters $\alpha_1 = \theta_1 + \theta_3$, and $\alpha_2 = \theta_2 + \theta_3$, respectively. Therefore, the marginal MLEs are given by the vector $\hat{\boldsymbol{\alpha}}_{MLE} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_k$. However, this gives only two equations, while we need to estimate three parameters. To do this we choose the mixed moment, $\mu_{12} = E(X_{11}X_{21})$, which can be estimated by $(\hat{\mu}_{12})_{MOM} = \frac{1}{K}\sum_{k=1}^{K} x_{1k}x_{2k}$, and apply MOM. The moment is given by

$$\begin{aligned} E(X_{11}X_{21}) &= E((Y_1 + Y_3)(Y_2 + Y_3)) \\ &= E(Y_1)E(Y_2) + E(Y_3)(E(Y_1) + E(Y_2)) + E(Y_3^2) \\ &= \theta_1\theta_2 + \theta_3(\theta_1 + \theta_2) + \theta_3^2 + \theta_3 \\ &= \theta_1\theta_2 + \theta_3(1 + \theta_1 + \theta_2) + \theta_3^2 \end{aligned}$$

Now, solving the three equations in three unknowns, we obtain MMLE estimates:

$$\begin{aligned} (\hat{\theta}_1)_{MMLE} &= (\hat{\alpha}_1)_{MLE}(1 + (\hat{\alpha}_2)_{MLE}) - (\hat{\mu}_{12})_{MOM} \\ (\hat{\theta}_2)_{MMLE} &= (\hat{\alpha}_2)_{MLE}(1 + (\hat{\alpha}_1)_{MLE}) - (\hat{\mu}_{12})_{MOM} \\ (\hat{\theta}_3)_{MMLE} &= (\hat{\mu}_{12})_{MOM} - (\hat{\alpha}_1)_{MLE}(\hat{\alpha}_2)_{MLE} \end{aligned}$$

At no point did we need the joint discrete density function to complete this process, and, thus, it qualifies as a likelihood-free method of parameter estimation. This method does, however, have some pitfalls. Most notably, if the space of *marginal* parameters corresponding to $\boldsymbol{\Theta}$ is not a product

space, this method can be inappropriate. For example, the 3-parameter bivariate Beta distribution [31] (discussed in more detail in the next section) is a family of bivariate beta distributions that does not include all possible combinations of Beta marginals (in fact, completely defining one marginal limits the possibilities for the other to only a one-dimensional set). So, with the exception of cases with large K, this method may not (probably wouldn't) yield marginal MLEs that map to possible parameters for the distribution.

1.3 Approximate Bayesian Computation (ABC)

ABC is a method of parameter estimation which applies a carefully-chosen collection of summary statistics to compare the original data and proposed simulated data sets through some specified distance function. Specifically, assume that $F_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ is an appropriate prior distribution for $\boldsymbol{\theta}$. We are interested in obtaining a sample of size N from the posterior distribution, $F_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$. Define $\boldsymbol{S_y} := (S_1, S_2, ..., S_M)$ to be a collection of summary statistics that are thought to be strongly informative about the unknown value of the parameter and can be calculated directly from the data, $\boldsymbol{y}$. Also, define $\rho(\boldsymbol{S_x}, \boldsymbol{S_y})$ to be a distance function, preferably capable of detecting differences between data sets whenever they are drawn from different elements of $\mathcal{F}$. In addition, declare a certain positive value, ϵ_0, as an acceptance threshold. Then the ABC algorithm proceeds as follows:

Step 1. Set $t = 0$.

Step 2. Generate a single proposal value, $\boldsymbol{\theta}^*$, from $F_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$.

Step 3. Simulate an iid sample, $\boldsymbol{y}^*$, of K realizations from $\boldsymbol{Y} \sim F(\boldsymbol{y}; \boldsymbol{\theta}^*)$.

Step 4. Compute $\rho(\boldsymbol{S_x}, \boldsymbol{S_{y^*}})$. If $\rho(\boldsymbol{S_x}, \boldsymbol{S_{y^*}}) < \epsilon_0$, then step t to $t + 1$ and accept $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^{(t)}$, a

draw from $F_{\tilde{\theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$. Otherwise, reject it.

Step 5. If $t < N$, repeat Step 2. Otherwise, stop.

This is the most basic form of this algorithm, and it has been adapted to many specific applications. It has become an increasingly popular tool for parameter estimation in the absence of tractable density functions. Obvious pitfalls include the need to judiciously select the set of summary statistics and ρ; in many cases, it may not be known or may be difficult to assess the quality of those ultimately selected, due, in part, to the absence of a density. See [42] for a recent detailed description of this method, some adaptations, and applications. Additional recent work and applications may be found in [13], [8], and [44]. To illustrate the effectiveness of the method, we consider an example where the joint density is analytically tractable so that a comparison with a more standard estimation strategy can be made.

Example: Olkin & Liu 3-Parameter Bivariate Beta Distribution

Olkin and Liu [31] proposed a bivariate beta distribution with the following density:

$$f(x_1, x_2; \alpha, \beta, \gamma) = \frac{x_1^{\alpha-1} x_2^{\beta-1} (1-x_1)^{\alpha+\gamma-1} (1-x_2)^{\beta+\gamma-1}}{B(\alpha, \beta, \gamma)(1-x_1 x_2)^{\alpha+\beta+\gamma}} I\{0 < x_1, x_2 < 1\} \tag{1.1}$$

where $B(\alpha, \beta, \gamma)$ is the generalized beta function,

$$B(\alpha, \beta, \gamma) = \frac{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\alpha+\beta+\gamma)} \tag{1.2}$$

In a Bayesian setting we consider the prior distribution

$$f_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}(\alpha, \beta, \gamma) = \frac{\alpha^{a_1-1} e^{-\frac{\alpha}{b_1}} \beta^{a_2-1} e^{-\frac{\beta}{b_2}} \gamma^{a_3-1} e^{-\frac{\gamma}{b_3}}}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3) b_1^{a_1} b_2^{a_2} b_3^{a_3}} I\{\alpha, \beta, \gamma > 0\} \tag{1.3}$$

Note that the marginals in Equation 1.3 are all gamma distributions and are independent. This is a natural choice for a joint prior for the three positive parameters, but Equation 1.3 is clearly not a

conjugate prior for the model in Equation 1.1. The posterior distribution is thus obtained.

$$f_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}|\boldsymbol{X}}(\alpha,\beta,\gamma|\boldsymbol{X}=\boldsymbol{x}) \propto f_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}(\alpha,\beta,\gamma)\prod_{j=1}^{K} f(\boldsymbol{x}_j;\alpha,\beta,\gamma). \tag{1.4}$$

In order to demonstrate the use of ABC in this example, we assume that the density function (1.1) is unknown. First, assume the actual parameter values are $\alpha = 2, \beta = 3$, and $\gamma = 4$, and set the hyper-parameters of the prior to $a_1 = a_2 = a_3 = 3$, and $b_1 = b_2 = b_3 = 1$, so that

$$f_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}(\alpha,\beta,\gamma) = \frac{(\alpha\beta\gamma)^2 e^{-(\alpha+\beta+\gamma)}}{8} I\{\alpha,\beta,\gamma > 0\}, \tag{1.5}$$

and $E(\tilde{\alpha}) = E(\tilde{\beta}) = E(\tilde{\gamma}) = 3$. We will then generate a single simulated data set of size 100; this data set will be assumed to be $\boldsymbol{x}$ in this example. Since the marginals are beta distributions, we consider the sufficient statistics for the marginals, as well as one additional mixed statistic, i.e.,

$$\boldsymbol{S_y} = \frac{1}{K}\left(\sum_{i=1}^{K}\log(y_{1i}), \sum_{i=1}^{K}\log(y_{2i}), \sum_{i=1}^{K}\log(1-y_{1i}), \sum_{i=1}^{K}\log(1-y_{2i}), \sum_{i=1}^{K}\log(1-y_{1i}y_{2i})\right) \tag{1.6}$$

We define ρ to simply be the sum of squared distances.

$$\rho(\boldsymbol{S_x},\boldsymbol{S_y}) = \left(\boldsymbol{S_x}-\boldsymbol{S_y}\right)\left(\boldsymbol{S_x}-\boldsymbol{S_y}\right)' \tag{1.7}$$

An important question at this point is whether $\boldsymbol{S_y}$ provides sufficient identifiability, that is whether a difference in the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ induces a ρ significantly greater than zero. In this example, we have the advantage of knowing the density, and thus, we know that $\boldsymbol{S_y}$ is, in fact, by the factorization criterion, a set of sufficient statistics, and we are therefore guaranteed identifiability with sufficiently large K. But, this is often not the case when applying ABC, so this step is generally considered the most important, and most difficult, in the process of applying ABC. In subsequent chapters, some examples of this will be exhibited.

Lastly, we must define the distance threshold for acceptance, ϵ_0. Generally, the smaller this value is, the more accurately the resulting sample will resemble a sample from the posterior

distribution. The trade-off is that small values of ϵ_0 will reduce the acceptance rate, and hence increase the processing time. Therefore, this parameter can be thought of as a desired accuracy setting. We set it to 0.001.

We set the desired sample size from the posterior to $N = 100$ (it need not be equal to K). Using the means of the sampled parameter values, we may obtain estimates of the parameters. For comparison, we also compute the maximum likelihood estimates. The results are shown in Table 1.1.

Estimator	Acceptance Rate	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
ABC	0.0001038122	2.239767	2.948822	3.968311
MLE	N/A	2.270081	3.063669	4.062024

Table 1.1: ABC Results for Olkin-Liu Bivariate Beta with $\alpha = 2, \beta = 3$, and $\gamma = 4$.

It should be noted that this distribution produces estimators that are heavily correlated (It will be revisited in Section 1.4.2 for this reason.). This correlation contributes to the low acceptance rate, but, the largest contributor, aside from the desired accuracy level, ϵ_0, is the prior distribution. Drawing from a distribution in which a greater level of (correct) prior information is contained leads to a better acceptance rate. Ultimately, while there a several components of the procedure that must be decided upon for ABC, the method is highly effective if good choices of the summary statistics are made together with a well-tuned selection of ϵ_0.

1.4 Markov Chain Monte Carlo (MCMC)

One of the most popular numerical techniques for inference, or more generally, modeling distributions, MCMC is a method which ostensibly builds a Markov Chain with the target distri-

bution as its long-run distribution, so that each step eventually should produce a draw from the target distribution. Upon sufficient burn-in and applying any necessary thinning, what results is, for all practical purposes, a random sample from the target distribution. In general, MCMC requires knowledge of the likelihood function, or at least some scalar multiple of it. However, with an increasingly rich selection of density estimation techniques now available, this is becoming less of a problem. In addition, methods of density estimation can be selected to maximize the utility of simulated data, further enhancing the efficiency of MCMC algorithms. In the following section, the most general form of MCMC is discussed. Following this, some adaptations of MCMC relevant to the distributions studied in this thesis are explained.

1.4.1 General MCMC Algorithm

The standard MCMC algorithm, also known as the Metropolis-Hastings Algorithm, is particularly simple. As with ABC, assume that $F_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$, with density $f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$, is an appropriate prior distribution for $\boldsymbol{\theta}$. We are, again, interested in obtaining a sample of size N from the posterior distribution, $F_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$. Assume, in addition, that we have a method of estimating the likelihood function

$$\hat{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{k=1}^{K} \hat{f}(\boldsymbol{x}_k|\boldsymbol{\theta})$$

where $\hat{f}$ is an estimate of the density, f, of $\boldsymbol{X}_1$, so that the estimated posterior density becomes

$$\hat{f}_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})\hat{L}(\boldsymbol{\theta}|\boldsymbol{x})$$

Then the Metropolis-Hastings algorithm proceeds as follows:

Step 1. Set $t = 1$. Draw an initial value $\boldsymbol{\theta}^{(1)}$, from $F_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$, or possibly from some other desired proposal distribution.

Step 2. Step t to $t+1$. Generate a single proposal value, $\boldsymbol{\theta}^*$, from the desired proposal distribution.

Step 3. Simulate an iid sample, $\boldsymbol{y}^*$, of M realizations from $\boldsymbol{Y} \sim F(\boldsymbol{y}; \boldsymbol{\theta}^*)$, where M is sufficiently large to accurately estimate f at all points in $\boldsymbol{x}$.

Step 4. Estimate $\hat{f}(\boldsymbol{x}_k|\boldsymbol{\theta}^*)$ for all $k \in \{1, 2, ..., K\}$, using the chosen density estimation technique based on the data, $\boldsymbol{y}^*$, and compute the estimated likelihood, $\hat{L}(\boldsymbol{\theta}^*|\boldsymbol{x})$, and posterior, $\hat{f}_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}^*|\boldsymbol{x})$.

Step 5. Simulate a single realization, u, from $U \sim U(0,1)$. If $u < \frac{\hat{f}_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}^*|\boldsymbol{x})}{\hat{f}_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{x})}$, then accept $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^{(t)}$ as a draw from $F_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$. Otherwise, set $\boldsymbol{\theta}^{(t-1)}$ as $\boldsymbol{\theta}^{(t)}$.

Step 6. If $t = N$, then stop. Otherwise, repeat Step 2.

Typically, $\hat{f}$ should be chosen to maximize the utility of available resources (usually simulated data) according to known characteristics of the family of parametric distributions being studied. One of the methods for choosing a density estimator is found in Appendix A.

Typically, burn-in and thinning will be applied to avoid bias in the posterior sample. Burn-in involves the removal of a number of initial draws of the chain, and thinning involves considering as part of the posterior sample only a sparse subset of the realizations of the chain subsequent to burn-in. As a result, the length of the chain must be increased in order to obtain a final sample of size N after the particular burn-in and thinning strategies have been applied.

1.4.2 Differential Evolution MCMC (DE-MCMC)

Often the posterior estimators of $\boldsymbol{\theta}$ can be heavily correlated. Failing to take this into account can slow the convergence rate of MCMC significantly. To deal with this problem, Storn and Price [41] proposed DE-MCMC.

While this method may be adapted to yield approximate maximum likelihood estimates, we elect to cover Bayesian estimation here. The primary reason for this is that information about well-understood marginal distributions may be applied to the prior distribution to increase the efficiency of the process. The method proceeds as follows:

1. Begin the Markov chain by simply drawing $\left\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right\}$ as an iid sample from the prior, $F_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$. Set $t = 2$.

2. Set t to $t + 1$. Sample $\left(\boldsymbol{\theta}^{(t_1)}, \boldsymbol{\theta}^{(t_2)}\right)$ from $\left\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(t-1)}\right\}$, without replacement.

3. Set

$$\boldsymbol{\theta}^{*} = \boldsymbol{\theta}^{(t-1)} + \tau\left(\boldsymbol{\theta}^{(t_1)} - \boldsymbol{\theta}^{(t_2)}\right) + \boldsymbol{\epsilon}^{*} \tag{1.8}$$

   where $\tau = \frac{2.38}{\sqrt{2d}}$ and $\boldsymbol{\epsilon}^{*}$ is a random error term.

4. Generate $U \sim U(0, 1)$. If $U < \frac{\hat{f}_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}^{*}|\boldsymbol{X})}{\hat{f}_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{X})}$, then set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{*}$. Otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

5. If the required chain length is achieved, then stop. Otherwise, and return to Step 2.

Upon stopping, the set $\left\{\boldsymbol{\theta}^{(b+q)}, \boldsymbol{\theta}^{(b+2q)}, ..., \boldsymbol{\theta}^{(b+Nq)}\right\}$, where q is a thinning parameter, and the values $\left\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(b)}\right\}$ are considered "burn-in," and are thus omitted, is then considered a sample from the actual posterior, $f_{\tilde{\boldsymbol{\theta}}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{X})$.

Example: Olkin & Liu 3-Parameter Bivariate Beta Distribution

We now return to the example used to demonstrate the use of ABC. Recall that it was previously stated that, with the Olkin & Liu Bivariate Beta distribution, the estimators tend to be highly correlated. So, the MCMC-DE method will be applied in order to deal with this issue. We will apply the same prior distribution given by

$$f_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}(\alpha,\beta,\gamma) = \frac{(\alpha\beta\gamma)^2 e^{-(\alpha+\beta+\gamma)}}{8} I\{\alpha,\beta,\gamma > 0\} \tag{1.9}$$

Also, we will assume the same actual parameter values, $\alpha = 2$, $\beta = 3$, and $\gamma = 4$, and $K = 100$ (in fact, we will assume the same data as was used in Example 1.3).

For this method, we also need to choose an error term, $\boldsymbol{\epsilon}^*$. It is necessary to explain how this will be done, and thus, how Step 3 of the algorithm will be implemented. This term is intended only to provide perturbations. Also, since the parameters must be positive, it must depend on the proposed parameters. For this reason, a pre-selected variance for the perturbations is applied; in this case, we set it to 1, and apply a tri-variate Gamma distribution. Set the constant value,

$$\boldsymbol{\theta}^{**} = \begin{pmatrix} \tilde{\alpha}^{(t-1)} + \tau\left(\tilde{\alpha}^{(t_1)} - \tilde{\alpha}^{(t_2)}\right) \\ \tilde{\beta}^{(t-1)} + \tau\left(\tilde{\beta}^{(t_1)} - \tilde{\beta}^{(t_2)}\right) \\ \tilde{\gamma}^{(t-1)} + \tau\left(\tilde{\gamma}^{(t_1)} - \tilde{\gamma}^{(t_2)}\right) \end{pmatrix}, \tag{1.10}$$

as indicated in Step 3. Then, we obtain a draw for $\boldsymbol{\theta}^*$ from the distribution,

$$\boldsymbol{\theta}^* = \boldsymbol{\epsilon}^* + \boldsymbol{\theta}^{**} \sim \boldsymbol{\Gamma}_3\left(\text{shape} = [\boldsymbol{\theta}^{**}]^2, \text{rate} = \boldsymbol{\theta}^{**}\right), \tag{1.11}$$

where $\boldsymbol{\Gamma}_3$ indicates a tri-variate Gamma distribution with independent marginals. This results in a distribution for each element of $\boldsymbol{\epsilon}^*$ with mean 0, and variance 1, as desired.

As this is intended to be an MCMC procedure, a density estimation technique is necessary. For this, the Turner & Sederberg [45] method will be applied. We will apply a burn-in of 10000

iterations, and thin by taking every 100th iteration. Hence, for a sample of size, $N = 100$, we need 20000 iterations in the Markov chain. For this comparison, using the same data as was used for Example 1.3; the results are displayed in Table 1.2.

Estimator	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
MCMC-DE	1.963814	2.726850	3.491819
ABC	2.239767	2.948822	3.968311
MLE	2.270081	3.063669	4.062024

Table 1.2: MCMC-DE Results for Olkin-Liu Bivariate Beta with $\alpha = 2, \beta = 3$, and $\gamma = 4$.

There are two potential reasons why the MCMC-DE results compare less favorably to the ABC results. First, the ABC procedure did, in fact, benefit from the presence of a sufficient statistic for the parameters, which is generally not the case. It therefore had maximal information to maximize efficiency. Unlike the ABC procedure, the MCMC-DE procedure required the use of an estimator for the density, which was based on simulated data, and may not provide the most accurate representation of the density and cause the chain to either take much longer than anticipated to burn in, or introduce some unwanted bias in the accept/reject decision. Second, the MCMC-DE procedure draws deeply into the collection of previous values in the chain to form new proposals. This may create the necessity to expand, maybe vastly, the length of the burn-in, as well as require more stringent thinning.

Overall, the MCMC-DE estimation procedure requires a careful selection of search parameters. With sufficient burn-in and thinning, the procedure may be applicable to at least some situations where the parameter estimators are highly correlated.

1.4.3 Gibbs Sampler

A popular method of drawing samples, the Gibbs Sampler is particularly useful when the distribution of the parameters is more tractably described in terms of the conditional distributions of each parameter given the values of the other parameters. Suppose, for example, that $p = 3$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, where the (prior and posterior) distributions of $\tilde{\theta}_1 | \left(\tilde{\theta}_2, \tilde{\theta}_3\right)$; $\tilde{\theta}_2 | \left(\tilde{\theta}_1, \tilde{\theta}_3\right)$; and $\tilde{\theta}_3 | \left(\tilde{\theta}_1, \tilde{\theta}_2\right)$ are all known, at least to the extent that draws from these distributions are easily obtained. The Gibbs Sampler proceeds by drawing an initial $\boldsymbol{\theta}^{(0)}$ from a prior distribution. Then, for each $t \in \{1, 2, ..., N + b\}$,

1. Draw $\theta_1^{(t)}$ from $\tilde{\theta}_1 | \left(\tilde{\theta}_2 = \theta_2^{(t-1)}, \tilde{\theta}_3 = \theta_3^{(t-1)}\right)$;

2. Draw $\theta_2^{(t)}$ from $\tilde{\theta}_2 | \left(\tilde{\theta}_1 = \theta_1^{(t)}, \tilde{\theta}_3 = \theta_3^{(t-1)}\right)$;

3. Draw $\theta_3^{(t)}$ from $\tilde{\theta}_3 | \left(\tilde{\theta}_1 = \theta_1^{(t)}, \tilde{\theta}_2 = \theta_2^{(t)}\right)$,

where b is the desired burn-in. See [10] for an overview of the Gibbs Sampler and its foundational theory.

1.5 Statistical Learning

Statistical learning, also known as machine learning, techniques involve, most generally, the use of presented data to 'learn' about the underlying distribution, restricted by a certain set of assumptions. In particular, and in the context of the general problem presented at the beginning of this chapter, we are interested in inferring the value of $\boldsymbol{\theta}_0$ from $\boldsymbol{x}$. We do this by attempting to 'learn' the relationship $\phi : \boldsymbol{x} \longrightarrow \boldsymbol{\Theta}$ through what is most intuitively viewed as a process of 'trial and error.'

Ultimately, we wish to construct ϕ so that it provides the best possible estimates of $\boldsymbol{\theta}$. There are two major extremes in the continuum of statistical learning methods: supervised, and unsupervised.

Supervised learning presents *training data*,

$$\boldsymbol{Y}_{test} = \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_K \\ \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \cdots & \boldsymbol{\theta}_K \end{pmatrix}, \tag{1.12}$$

where each $\boldsymbol{y}_k$ is known to be data drawn from $F(\boldsymbol{x}; \boldsymbol{\theta_k})$, for all $k \in \{1, 2, ..., K\}$. The algorithm attempts to fit the best ϕ according to a certain set of assumptions about ϕ. The simplest form of this is linear regression, but sometimes far less must be assumed about ϕ, rendering regression inadequate, and instead requiring a complex algorithm or process to obtain ϕ.

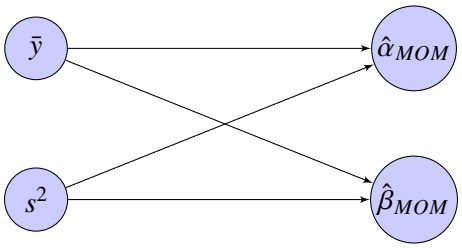
Unsupervised learning is more difficult, for the second row of the training data is absent. The algorithm is, instead, tasked with finding important anomalies across the datasets that *may* ultimately be important to understanding where they came from. This method may sometimes precede supervised learning as a way to obtain a relevant set of measurements (which are usually in the form of statistics) that appear to change from one dataset to the next. Subsequently varying these statistics in a supervised learning context could lead to a reasonable ϕ.

There are many approaches to statistical learning, but this thesis will focus on only one: Neural Networks. These are discussed in the next section.

1.5.1 Neural Networks

A neural network is a construct with an input (statistics), and an output (parameters). In the context of the problem, the neural network *is* ϕ, and we use statistical learning to construct it. To illustrate, we use a simple example: suppose $y_1, y_2, ..., y_K \sim B(\alpha, \beta)$, and we wish to estimate (α, β) via the method of moments, that is, we want ϕ, such that $\phi(\boldsymbol{y}) = (\hat{\alpha}_{MOM}, \hat{\beta}_{MOM})$. A one-step neural

network may be implemented to achieve this.



This neural network is about as simple as one would get, so much so that it is uninteresting. It is predetermined by knowledge of the estimators, and no learning is necessary. In general, the decisions, that is the calculations, in each step are not known *a priori*. Instead, the training data, $\boldsymbol{Y}_{test}$, can be used as a measure of quality for the network. In this way, until it becomes maximally successful for the test data, e.g., the sum of squared errors is minimized, we can change its calculations; that is, change the internal functions. For example, suppose that, rather than $\bar{y}$ and s^2, an arbitrary collection of summary statistics is placed in the first stage. The learning process would be to manipulate the calculations made from the summary statistics to the estimators of the parameters. We needn't be restricted to only one stage; a second set of summary measures, based on the summary statistics, may be computed, forming a second stage from which the parameter estimates may be obtained. In this sense, just about any statistical inference problem can be implemented by a neural network, but only a small subset of them are *best* implemented by this tool.

Now, to reduce the problem of 'learning' to one that is manageable, we add some restrictions to the network's structure, which, as it happens, form the most common form of a neural network. First, each function is restricted to a linear function of all values in the previous stage. Second, a sigmoid function is applied to these linear functions, resulting in all values in each (intermediate) stage having a value in $[0, 1]$. Third, a cost function, usually a sum of squared errors, is

computed from the last stage results and the training parameters. With these restrictions, the method of steepest decent may be applied.

Returning to the Beta example, suppose two intermediate stages, one with six nodes and one with four nodes, are added, forming a total of four stages. We shall restrict the first stage to the following collection of moments:

$$\begin{aligned} m_{11} &= \sum_{k=1}^{K} y_k & m_{15} &= \sum_{k=1}^{K} (1-y_k) \\ m_{12} &= \sum_{k=1}^{K} y_k^2 & m_{16} &= \sum_{k=1}^{K} (1-y_k)^2 \\ m_{13} &= \sum_{k=1}^{K} y_k^3 & m_{17} &= \sum_{k=1}^{K} (1-y_k)^3 \\ m_{14} &= \sum_{k=1}^{K} y_k^4 & m_{18} &= \sum_{k=1}^{K} (1-y_k)^4 \end{aligned}$$

The second and third stages will learn. This results in the network shown in Figure 1.1.

Note, here, that we deliberately choose less-than-optimal statistics, that is, non-minimally-sufficient statistics, in order to form a challenging scenario for the learning process. For the sigmoid function, we choose the commonly-used Logistic function (the distribution function of a standard Logistic random variable).

We can apply the learning algorithm (using the `R` package `neuralnet`) to a set of 500 Beta distributions with various parameters, of which 375 are randomly chosen to be training sets, and the other 125 are to be used for cross-validation. All data sets are of size $K = 100$. The average bias and MSE on the cross-validation sets for the resulting Neural Network are recorded in Table 1.3; included for comparison are the average bias and MSE for the corresponding MLEs.

Clearly not as good as the MLE, as is to be expected, the Neural Network is nevertheless a means by which reasonable estimates can be obtained, without the need for a density. In addition,

the Neural network need only "learn" once. After this, it is simple to apply.[1]

	Bias		MSE	
Estimator	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
Neural Network	−7.778	−2.889	13.480	18.301
MLE	−2.940	−3.281	6.917	7.307

Table 1.3: Bias and MSE for Neural Network applied to Beta Distributions.

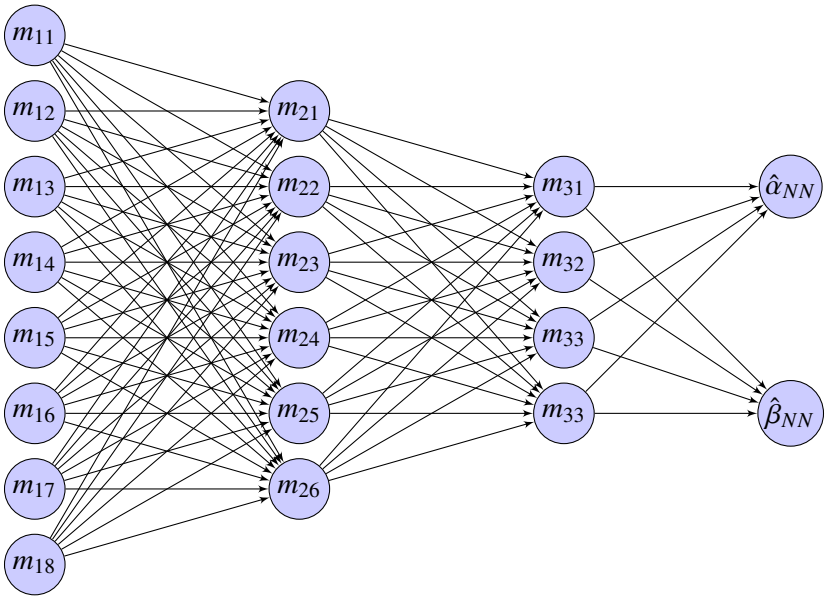


Figure 1.1: Neural Network for Obtaining estimates of a $B(\alpha,\beta)$ Model.

[1] However, the most advanced neural networks conceivable would undoubtedly continue to learn from all data to which they are applied.

Chapter 2

Bivariate Beta Distributions and Copulas

2.1 Introduction

The array of convenient families of bivariate beta distributions is vast. A comprehensive survey of the many variations can be found in [32]. Distributions of this type are commonly used as prior distributions for bivariate binomial parameter estimation, as well as for other modeling applications. One such application will be demonstrated in Chapter 3. In this chapter, we will begin with an 8-Parameter bivariate beta model and discuss a collection of its most interesting sub-families. This will be followed by an introduction to a flexible collection of 5-parameter sub-families. For this collection, we develop applicable parameter estimation techniques adapted from those methods discussed in Chapter 1. In addition, we will look at an interesting sub-family of the 8-parameter model whose marginals are uniform, that is, a family of copulas, which will be followed

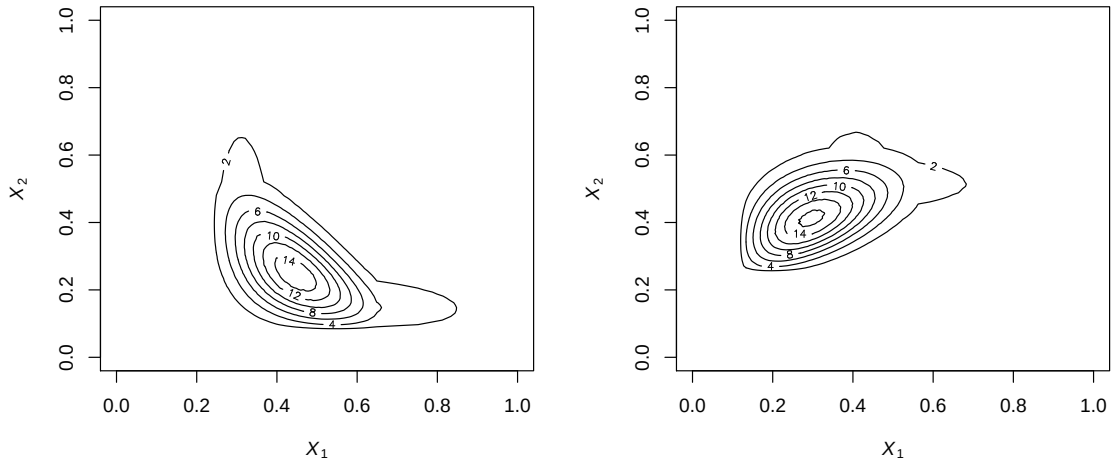


Figure 2.1: Densities for $\boldsymbol{X} \sim BB(3, 0, 8, 5, 0, 0, 6, 4)$ (left), and $\boldsymbol{X} \sim BB(0, 3, 0, 9, 5, 7, 0, 4)$ (right).

by a demonstration of an applicable parameter estimation technique. Lastly, some model-selection techniques will be explored.

2.2 General (8-Parameter) Bivariate Beta Distributions

Arnold & Ng [5] proposed a family of bivariate beta distributions which, most notably, can be easily simulated and include the full range of possible correlations. While these may be extended to higher dimensions, we focus on the bivariate case in this chapter. The most general family is the 8-Parameter family, constructed as follows.

Define $U_1, U_2, ..., U_8$ to be independent random variables with $U_i \sim \Gamma(\delta_i, 1)$ for all i. Also define the variables

$$V_1 := \frac{U_1 + U_5 + U_7}{U_3 + U_6 + U_8}, \text{ and } V_2 := \frac{U_2 + U_5 + U_8}{U_4 + U_6 + U_7}. \tag{2.1}$$

These V_i's have beta distributions of the second kind. Finally, define

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \\ X_2 \end{pmatrix} := \begin{pmatrix} \frac{V_1}{1+V_1} \\ \\ \frac{V_2}{1+V_2} \end{pmatrix},$$

This forms an 8-parameter family of bivariate beta random variables with distribution function $F_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\delta})$, where $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8)$. Though there are some special cases in which $F_{\boldsymbol{X}}$ may be written in closed form, this joint distribution, in general, has no known closed form, nor does the associated density, $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\delta})$. The marginal distributions are $X_j \sim B(\alpha_j, \beta_j)$, $j \in \{1, 2\}$, where

$$\begin{aligned} \delta_1 + \delta_5 + \delta_7 &= \alpha_1, \\ \delta_2 + \delta_5 + \delta_8 &= \beta_1, \\ \delta_3 + \delta_6 + \delta_8 &= \alpha_2, \text{ and} \\ \delta_4 + \delta_6 + \delta_7 &= \beta_2. \end{aligned} \tag{2.2}$$

Examples of the bivariate density are shown in Figure 2.1 (generated through simulation). By design, the last four parameters, $(\delta_5, \delta_6, \delta_7, \delta_8)$, impact the behavior of both random variables, so we shall call them the *dependence parameters*. By convention, we will reduce this model by stating $\delta_k = 0$ if and only if $U_k \equiv 0$, for any $k \in \{1, 2, .., 8\}$, while maintaining enough nonzero δ's to avoid either of the numerators or denominators in Equation 2.1 from being zero. Throughout this chapter, we will say that $\boldsymbol{X} \sim BB(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8)$, where some parameters are allowed to be zero, or the more compact form, $\boldsymbol{X} \sim BB(\boldsymbol{\delta})$.

The natural symmetry of the beta distribution provides some useful symmetries of the bivariate beta that will be exploited in subsequent sections. For example, if $\boldsymbol{X} \sim BB(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8)$,

and $\boldsymbol{Y} \sim BB(\delta_1, \delta_4, \delta_3, \delta_2, \delta_7, \delta_8, \delta_5, \delta_6)$, then $\boldsymbol{Y} \stackrel{d}{=} \begin{pmatrix} X_1 \\ 1 - X_2 \end{pmatrix}$. Applying these symmetries can be thought of as rotating the data about one or both of the lines $x_1 = \frac{1}{2}$ and $x_2 = \frac{1}{2}$.

2.2.1 Two Interesting Sub-Models

Two sub-models have been discussed in earlier literature. One is the 5-parameter family, also discussed in [5], where $\delta_3 = \delta_4 = \delta_5 = 0$, so that

$$V_1 := \frac{U_1 + U_7}{U_6 + U_8}, \text{ and } V_2 := \frac{U_2 + U_8}{U_6 + U_7},$$

which was actually the original motivation for the full 8-parameter model. Another subfamily is the Olkin & Liu 3-Parameter Bivariate Beta, discussed in Section 1.4.2, corresponding to the case in which $\delta_3 = \delta_4 = \delta_5 = \delta_7 = \delta_8 = 0$, which, as it happens, is a sub-model of the 5-parameter family.

Statistical Inference

Assume we obtain a sample $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_K)$ (using the same notation as in Chapter 1) from a $BB(\boldsymbol{\delta})$ distribution.

3-Parameter Olkin-Liu Sub-Model

First assume the Olkin-Liu 3-parameter sub-model, i.e. that $\delta_2 = \delta_4 = \delta_5 = \delta_7 = \delta_8 = 0$, and the other δ's are unknown. Recall that for this model, the density is known. We obtain maximum likelihood estimators for the three parameters. The likelihood is given by

$$L(\boldsymbol{\delta}|\boldsymbol{x}) = \prod_{k=1}^{K} \left[\frac{x_{1k}^{\delta_1 - 1} x_{2k}^{\delta_2 - 1} (1 - x_{1k})^{\delta_1 + \delta_6 - 1} (1 - x_{2k})^{\delta_2 + \delta_6 - 1}}{B(\delta_1, \delta_2, \delta_6)(1 - x_{1k} x_{2k})^{\delta_1 + \delta_2 + \delta_6}} \right], \tag{2.3}$$

so that the log-likelihood is

$$\ell(\boldsymbol{\delta}|\boldsymbol{x}) = (\delta_1 - 1)\sum_{k=1}^{K}\log(x_{1k}) + (\delta_2 - 1)\sum_{k=1}^{K}\log(x_{2k}) + (\delta_1 + \delta_6 - 1)\sum_{k=1}^{K}\log(1 - x_{1k})$$
$$+ (\delta_2 + \delta_6 - 1)\sum_{k=1}^{K}\log(1 - x_{2k}) - K\log\left[B(\delta_1, \delta_2, \delta_6)\right] - (\delta_1 + \delta_2 + \delta_6)\sum_{k=1}^{K}\log(1 - x_{1k}x_{2k})$$

Differentiating, we have

$$\frac{\partial \ell}{\partial \delta_1} = \sum_{k=1}^{K}\log(x_{1k}) + \sum_{k=1}^{K}\log(1 - x_{1k}) - \sum_{k=1}^{K}\log(1 - x_{1k}x_{2k})$$
$$+ \psi(\delta_1) - \psi(\delta_1 + \delta_2 + \delta_6)$$
$$\frac{\partial \ell}{\partial \delta_2} = \sum_{k=1}^{K}\log(x_{2k}) + \sum_{k=1}^{K}\log(1 - x_{2k}) - \sum_{k=1}^{K}\log(1 - x_{1k}x_{2k})$$
$$+ \psi(\delta_2) - \psi(\delta_1 + \delta_2 + \delta_6)$$
$$\frac{\partial \ell}{\partial \delta_6} = \sum_{k=1}^{K}\log(1 - x_{1k}) + \sum_{k=1}^{K}\log(1 - x_{2k}) - \sum_{k=1}^{K}\log(1 - x_{1k}x_{2k})$$
$$+ \psi(\delta_6) - \psi(\delta_1 + \delta_2 + \delta_6),$$

where $\psi(a) = \frac{d}{da}\log\Gamma(a), a > 0$ is the digamma function. Setting these derivatives to zero, a Newton-Raphson search is thus immediately feasible to obtain the solution, i.e. the wide-sense MLEs.

There are four Olkin-Liu models that may be obtained utilizing the symmetries discussed earlier:

Model 1: $\boldsymbol{\delta} = (0, 0, \delta_3, \delta_4, \delta_5, 0, 0, 0)$;

Model 2: $\boldsymbol{\delta} = (\delta_1, \delta_2, 0, 0, 0, \delta_6, 0, 0)$ (the original Olkin-Liu);

Model 3: $\boldsymbol{\delta} = (0, \delta_2, \delta_3, 0, 0, 0, \delta_7, 0)$;

Model 4: $\boldsymbol{\delta} = (\delta_1, 0, 0, \delta_4, 0, 0, 0, \delta_8)$.

The densities of these are all similar. These four models by themselves form a means of obtaining a 3-parameter estimate for various bivariate datasets on the unit square. Model selection is particularly easy in that the model producing the maximum value of the likelihood at its respective MLE is the one chosen. Also, combined, these four sub-families include multiple distributions with any correlation in $(-1, 1) \setminus \{0\}$. The benefit of this collection of sub-families is they involve a small number of parameters, and so are useful for smaller datasets. An unfortunate characteristic, however, is that only a small subset (in fact, one of measure zero) of possible marginal distributions are represented by this collection. To see this, consider Model 2 and any possible set of marginal parameters, $(\alpha_1, \alpha_2, \beta)$, where $X_1 \sim B(\alpha_1, \beta)$, and $X_2 \sim B(\alpha_2, \beta)$. Then exactly one distribution in this family has this particular set of marginal parameters: the one with $\delta_1 = \alpha_1$, $\delta_2 = \alpha_2$, and $\delta_6 = \beta$. Thus, in the 4-dimensional space of possible marginals, only a specific 3-dimensional subset encompasses the parameter space for this model, and, at most, only one representing a specific set of marginals.

5-Parameter Sub-Model

We now proceed to the more complex 5-Parameter sub-model. Arnold & Ng [5] demonstrated three methods of parameter estimation, the most successful of which was an MMLE procedure. Recently, Crackel [12] successfully applied ABC to obtain parameter estimates for this model, i.e., he found a collection of summary statistics, $\boldsymbol{S_x} := (S_1(\boldsymbol{x}), S_2(\boldsymbol{x}), ..., S_M(\boldsymbol{x}))$, with $M = 5$, and a distance function, $\rho(\boldsymbol{S_x}, \boldsymbol{S_y})$, that result in an efficient parameter estimator for $\boldsymbol{\delta} =$

$(\delta_1, \delta_2, 0, 0, 0, \delta_6, \delta_7, \delta_8)$. In particular, Crackel chose

$$
\begin{aligned}
S_1(\boldsymbol{x}) &= \frac{1}{K}\sum_{k=1}^{K}\log(x_{1k}) \\
S_2(\boldsymbol{x}) &= \frac{1}{K}\sum_{k=1}^{K}\log(x_{2k}) \\
S_3(\boldsymbol{x}) &= \frac{1}{K}\sum_{k=1}^{K}\log(1-x_{1k}) \\
S_4(\boldsymbol{x}) &= \frac{1}{K}\sum_{k=1}^{K}\log(1-x_{2k}) \\
S_5(\boldsymbol{x}) &= \frac{\sum_{k=1}^{K}(x_{1k}-\bar{x}_1)(x_{2k}-\bar{x}_2)}{\sqrt{\sum_{k=1}^{K}(x_{1k}-\bar{x}_1)^2\sum_{k=1}^{K}(x_{2k}-\bar{x}_2)^2}}
\end{aligned}
\tag{2.4}
$$

and

$$
\rho(\boldsymbol{S}_{\boldsymbol{x}}, \boldsymbol{S}_{\boldsymbol{y}}) = \sum_{m=1}^{5}|(S_1(\boldsymbol{x}) - (S_1(\boldsymbol{y})| \tag{2.5}
$$

For the proposal distribution, Crackel used a prior distribution, as is customary for ABC, though it is hoped that the choice of prior will have minimal effect on the final inferences. In this case, the prior was simply five independent gamma distributions[1].

For more general sub-models, and for the 8-parameter model, it is necessary to include additional statistics statistics in $\boldsymbol{S}$. For these, Crackel added three statistics to those found in Equation 2.4.

$$
\begin{aligned}
S_6(\boldsymbol{x}) &= 1 - \frac{6\sum_{k=1}^{K}(x_{1k}-x_{2k})^2}{K(K^2-1)} \text{ (Spearman's Rank Correlation Coefficient)} \\
S_7(\boldsymbol{x}) &= \frac{K_c - K_d}{\frac{1}{2}K(K-1)} \text{ (Kendall's Correlation Coefficient), and} \\
S_8(\boldsymbol{x}) &= \frac{1}{K}\sum_{k=1}^{K}\sqrt{x_{1k}x_{2k}}
\end{aligned}
\tag{2.6}
$$

These estimators are not expected to approach any well-known limiting distribution as K gets large, since the set of statistics is not known to be sufficient or provide identifiability. With this

[1] Others were tested, but this appeared to be the most successful.

said, simulation does strongly suggest that as $K \longrightarrow \infty$, $\hat{\boldsymbol{\delta}} \longrightarrow \boldsymbol{\delta}$, i.e. that these estimators appear to be consistent.

2.2.2 Other Families and their Limitations

There is a plethora of other sub-families of the 8-parameter model that may be obtained by setting various sets of parameters equal to zero. The two sub-families discussed to this point have a particular property that, as it turns out, permit a greater degree of accuracy for parameter estimation: their dependence parameters are, at least in part, defined by their marginal parameters, and in the case of the Olin-Liu distributions, completely defined by them. In particular, the Olkin-Liu model permits only a small subset of possible marginal distributions. The 5-parameter model also restricts its marginals, e.g. it cannot have $X_1 \sim B(1,2)$ and $X_2 \sim B(1,5)$. However, it is a bit more general than Olkin-Liu: for a given set of possible marginals, say $X_1 \sim B(3,5)$ and $X_2 \sim B(3,4)$, δ_6 must be in the interval $[2,4]$. Once defined, δ_6, along with the marginals, completely defines the other four parameters. In both of these cases, along with many similar cases, knowledge of the marginals greatly reduces the range of possible values for the remaining unknown parameters. Since MLEs for the marginal parameters are easily attainable, parameter estimation for these and similar models is also made much easier. An unfortunate characteristic of these sub-families is consequently made obvious: if there is no justification to limit the range of possible marginals, these models would be inadequate. For the remainder of this section, we will show that forming sub-families that eliminate this problem introduces additional problems.

The full, 8-parameter model is one such example. For this model, parameter estimation is impractical except in cases where enormous datasets are available, for more reasons than just the one highlighted in the previous paragraphs. While distinct parameter sets almost certainly map to

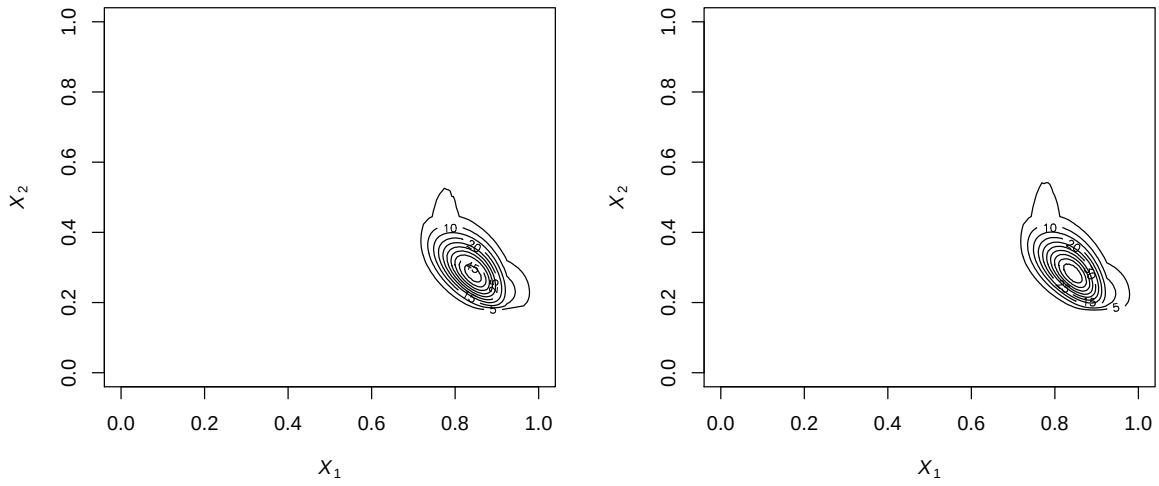


Figure 2.2: Densities for $\boldsymbol{X} \sim BB(1.43, 0.58, 0.87, 1.57, 7.02, 0.73, 30.41, 6.46)$ (left), and $\boldsymbol{X} \sim BB(17.44, 2.24, 0.034, 15.09, 4.44, 0.64, 16.98, 7.37)$ (right).

distinct distributions in even the full family, there is a clear identifiability problem from a practical point of view. A brief simulation study has shown this conclusively. For example, the (estimated) densities of two of these distributions with significantly different parameter sets are shown in Figure 2.2. While the marginal distributions of both are the same, the parameter vectors differ by a Euclidean distance of more than 25, yet the densities differ by less than 4%.[2] For comparison, this is a smaller difference than that between the univariate $\exp(\lambda = 0.50)$ and $\exp(\lambda = 0.53)$ densities. An additional example is shown in Figure 2.3. Here the Euclidean distance between the parameter vectors is more than 28, which exceeds the largest parameter value, while the difference in the densities is less than 1.2%, that is, a smaller difference than that between the univariate $\exp(\lambda = 0.50)$ and $\exp(\lambda = 0.51)$ densities. In addition, one of these has independent marginals, while the other's

[2]This was calculated as the mean absolute difference:

$$d = \int_0^1 \int_0^1 \left| f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\delta}^{(1)}) - f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\delta}^{(2)}) \right| dx_1 dx_2$$

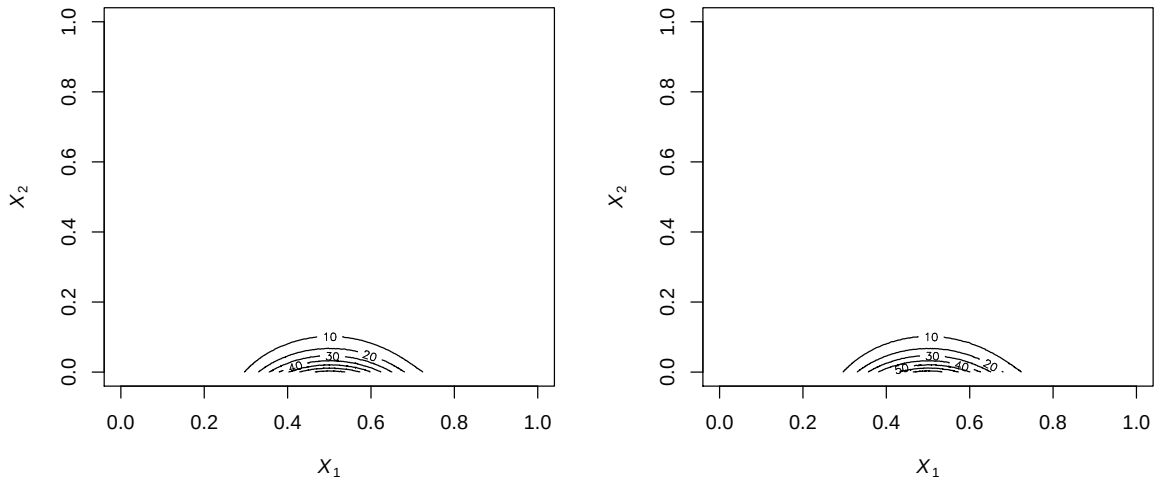


Figure 2.3: Densities for $\boldsymbol{X} \sim BB(1, 1, 1, 0, 0, 10, 10, 0)$ (left), and $\boldsymbol{X} \sim BB(11, 1, 11, 20, 0, 0, 0, 0)$ (right).

parameter make-up is dominated by the dependence parameters. Plenty of other examples exist.

We now show that even a collection of simple sub-families where all possible marginal distributions *can* be represented presents significantly more difficulty for estimation than the two models shown previously, and, in some cases, also presents identifiability issues. A simple sufficient condition for a sub-family to represent all possible marginal distributions is to require that all four of $\delta_1, \delta_2, \delta_3$, and δ_4 be active, that is, they are (permitted to be) non-zero. This leaves the dependence parameters as free parameters. Note that, by Equation 2.2, $\delta_1, \delta_2, \delta_3$, and δ_4 are completely defined by the marginal parameters and the dependence parameters:

$$\begin{aligned} \delta_1 &= \alpha_1 - \delta_5 - \delta_7 \\ \delta_2 &= \alpha_2 - \delta_5 - \delta_8 \\ \delta_3 &= \beta_1 - \delta_6 - \delta_8 \\ \delta_4 &= \beta_2 - \delta_6 - \delta_7, \end{aligned} \tag{2.7}$$

where $X_1 \sim B(\alpha_1, \beta_1)$, and $X_2 \sim B(\alpha_2, \beta_2)$. We establish a temporary alternative parameteriza-

tion with the marginal parameters and dependence parameters to form several five-parameter sub-families. These sub-families are given in Table 2.1.[3] While there are 15 of them, there are only 5 classes which behave differently from one another, outside of symmetries.

Some general observations about these families can be made before going into specifics. First, because the alternate parameterization includes the marginal parameters, the MLEs for these parameters can be used to obtain the final estimates for $\delta_1, \delta_2, \delta_3$, and δ_4. Second, the one additional parameter, δ, is in every case contained in a finite closed interval, making a search more tractable. This arrangement strongly suggests that parameter estimation should be trivial. This, as will be shown, is not the case. We will discuss methods for parameter estimation for these sub-families in the following sections.

Class A (Model 1)

For this class, only one of the dependence parameters is active, and it is associated with a positive correlation, though the relationship is not necessarily linear. This can be understood by recognizing that, for Model 1, if δ_5 is large (relative to the other parameters), then small values of U_5 will be associated with small values of both X_1 and X_2, and large values of U_5 will be associated with large values of both X_1 and X_2. A (not necessarily linear) correlation is often best measured through Spearman's Rank Correlation coefficient, ρ_S. In fact, for any fixed $(\alpha_1, \alpha_2, \beta_1, \beta_2)$, there is a one-to-one correspondence between δ and ρ_S. The estimate for δ can be obtained by implementing a bisection search over values for δ in the range $[0, \min\{\alpha_1, \alpha_2\}]$, each step simulating a large dataset and comparing the resulting estimated ρ_S to that obtained from the data. The number of steps is

[3]More general forms of these models can be constructed, where the δ's can be replaced by functions of some free parameter, e.g. $\delta_5 = g_5(\delta)$, $\delta_6 = g_6(\delta)$, $\delta_7 = g_7(\delta)$, and $\delta_8 = g_8(\delta)$, and any of these functions may be identically zero.

Class	Family	δ_5	δ_6	δ_7	δ_8	Restrictions
A	1	δ	0	0	0	$0 \le \delta \le \min\{\alpha_1, \alpha_2\}$
	2	0	δ	0	0	$0 \le \delta \le \min\{\beta_1, \beta_2\}$
	3	0	0	δ	0	$0 \le \delta \le \min\{\alpha_1, \beta_2\}$
	4	0	0	0	δ	$0 \le \delta \le \min\{\beta_1, \alpha_2\}$
B	5	δ	δ	0	0	$0 \le \delta \le \min\{\alpha_1, \beta_1, \alpha_2, \beta_2\}$
	6	0	0	δ	δ	$0 \le \delta \le \min\{\alpha_1, \beta_1, \alpha_2, \beta_2\}$
C	7	δ	0	δ	0	$0 \le \delta \le \min\{\frac{\alpha_1}{2}, \alpha_2, \beta_2\}$
	8	δ	0	0	δ	$0 \le \delta \le \min\{\alpha_1, \beta_1, \frac{\alpha_2}{2}\}$
	9	0	δ	δ	0	$0 \le \delta \le \min\{\alpha_1, \beta_1, \frac{\beta_2}{2}\}$
	10	0	δ	0	δ	$0 \le \delta \le \min\{\frac{\beta_1}{2}, \alpha_2, \beta_2\}$
D	11	δ	δ	δ	0	$0 \le \delta \le \min\left\{\frac{\alpha_1}{2}, \beta_1, \alpha_2, \frac{\beta_2}{2}\right\}$
	12	δ	δ	0	δ	$0 \le \delta \le \min\left\{\alpha_1, \frac{\beta_1}{2}, \frac{\alpha_2}{2}, \beta_2\right\}$
	13	δ	0	δ	δ	$0 \le \delta \le \min\left\{\frac{\alpha_1}{2}, \beta_1, \frac{\alpha_2}{2}, \beta_2\right\}$
	14	0	δ	δ	δ	$0 \le \delta \le \min\left\{\alpha_1, \frac{\beta_1}{2}, \alpha_2, \frac{\beta_2}{2}\right\}$
E	15	δ	δ	δ	δ	$0 \le \delta \le \min\left\{\frac{\alpha_1}{2}, \frac{\beta_1}{2}, \frac{\alpha_2}{2}, \frac{\beta_2}{2}\right\}$

Table 2.1: Five-Parameter Families of Bivariate Beta Distributions.

chosen based upon the desired accuracy level for $\hat{\delta}$. We therefore apply an MMLE procedure using ρ_S to predict δ. To form a $100(1-\tau)\%$ confidence region, we can apply a bootstrap procedure, simulating datasets of size K from the estimated distribution and applying the same procedure multiple independent times to obtain an approximate sample from the estimator.

We demonstrate the procedure for an example, $\boldsymbol{X} \sim BB(\boldsymbol{\delta})$, where $\boldsymbol{\delta} = (1, 2, 3, 4, 5, 0, 0, 0)$, and $K = 100$. The resulting parameter estimate is given by $\hat{\boldsymbol{\delta}} = (1.07, 0.63, 3.66, 3.75, 5.60, 0, 0, 0)$. The (bootstrap-generated) box plots are given in Figure 2.4. It should be observed that the increased variability of $\hat{\delta}_1$ and $\hat{\delta}_2$ is a result of the fact that these two estimators are calculated from the MLEs of α_1 and α_2 as well as the (significantly larger) $\hat{\delta}_5$, while $\hat{\delta}_3$ and $\hat{\delta}_4$ are calculated from the MLEs of β_1 and β_2 alone. In this way, it is important to keep in mind that different configurations of the marginal parameters could greatly impact the variability of the specific estimators. Now, these estimates (aside from δ_3 and δ_4 which are estimated directly from the marginal MLEs) are virtually indistinguishable from educated guesses. If, on the other hand, $K = 1000$, the estimates improve, but are still not ideal. This model, along with this method for parameter estimation, simple as they are, create a practically intractable estimation problem. Yet, it is the simplest conceivable sub-family of the full model whose marginal parameters are free.

Remark 1. *At this point, it may not be clear as to the difference between the impacts of δ_5 and δ_6 on the distribution, for they are both associated with positive correlation (and the same for δ_7 and δ_8 for negative correlation). For some distributions, this difference is quite subtle, and potentially insignificant, but not all. This will be made clear when copulas are discussed in Section 2.3.*

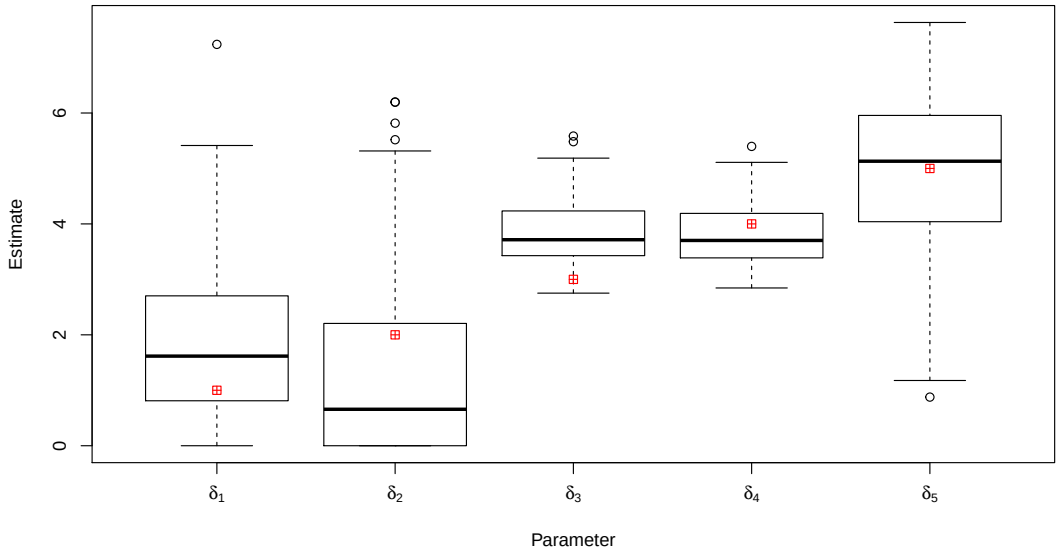


Figure 2.4: Parameter Estimation results for $\boldsymbol{X} \sim BB(1, 2, 3, 4, 5, 0, 0, 0)$ with $K = 100$.

Class B (Model 5)

We now consider the case where δ defines both δ_5 and δ_6. Stronger correlations can be observed between δ and ρ_S with Model 5 than Model 1. Therefore, the same method as that described for Model 1 can be applied. We use $\boldsymbol{\delta} = (1, 2, 3, 4, 3, 3, 0, 0)$ to generate data, $\boldsymbol{x}$, with $K = 100$. The resulting parameter estimate is given by $\hat{\boldsymbol{\delta}} = (0.66, 1.34, 1.89, 3.03, 2.87, 2.87, 0, 0)$. The (bootstrap-generated) box plots are given in Figure 2.5.

Of the five classes shown in Table 2.1, this class is arguably the easiest to deal with, for both δ_5 and δ_6 (in Model 5) affect correlation in concert with one another, so the estimates are reasonable. The next class, however, presents a very different problem.

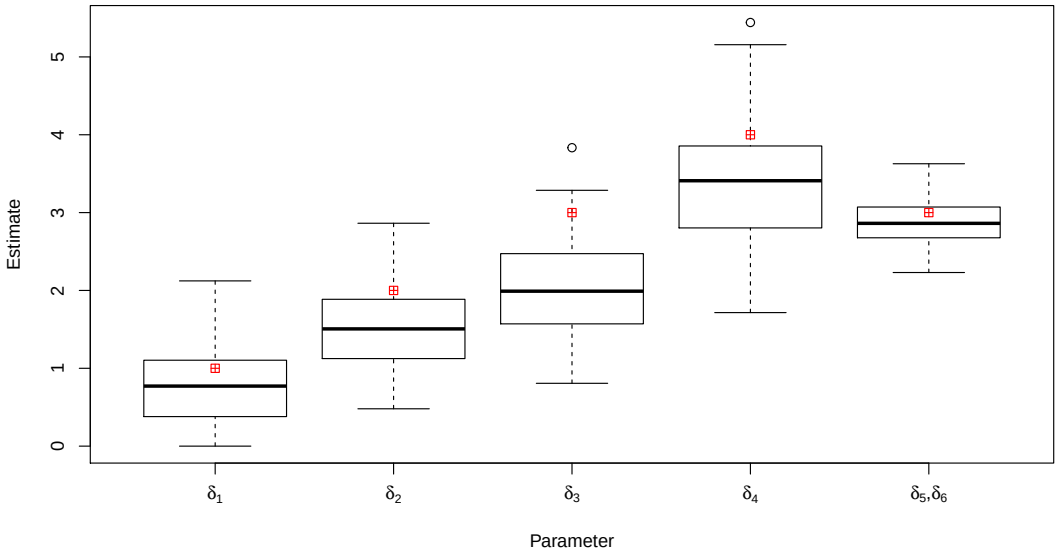


Figure 2.5: Parameter Estimation results for $\boldsymbol{X} \sim BB(1,2,3,4,3,3,0,0)$ with $K = 100$ with the actual values (red).

Class C (Model 7)

For this class of models, the two non-zero dependence parameters impact correlation in a conflicting way. Thus, a different collection of statistics for estimating δ is needed. The change to the distribution over the range of possible values for δ is, again, very subtle, as seen in the example in Figure 2.6. In the case shown, there is clearly the appearance of positive correlation when δ is at its maximum value. However, by symmetry, a similar distribution may be constructed which exhibits the exact opposite effect, and thus, by continuity, yet another where correlation is zero. So

correlation alone is insufficient. Rather, we use the four mixed moments

$$\begin{aligned} S_1(\boldsymbol{x}) &= \sum_{k=1}^{K} x_{1k}x_{2k}, \\ S_2(\boldsymbol{x}) &= \sum_{k=1}^{K} (1-x_{1k})x_{2k}, \\ S_3(\boldsymbol{x}) &= \sum_{k=1}^{K} x_{1k}(1-x_{2k}), \text{ and} \\ S_4(\boldsymbol{x}) &= \sum_{k=1}^{K} (1-x_{1k})(1-x_{2k}); \end{aligned} \tag{2.8}$$

which represent one method of capturing tail dependencies. Simulation suggests that these four moments provide a some level of identifiability for δ, given any fixed set of marginal parameters.

To estimate the parameters, we may, once again, apply MMLE. We do so by borrowing from ABC to obtain an estimate of δ. To do this, we obtain δ^* from a $U\left(0, \min\left\{\frac{\hat{\alpha}_1}{2}, \hat{\alpha}_2, \hat{\beta}_2\right\}\right)$, where $\hat{\alpha}, \hat{\beta}_1, \hat{\alpha}_2$, and $\hat{\beta}_2$ are the marginal MLEs. Comparing the statistics in Equation 2.8 obtained from the original data to those from a simulated dataset with the proposed parameters, we can accept only the best values of δ by setting (ABC's) ϵ_0 appropriately. However, large sample sizes are still required to obtain reasonable assessments of the relationship. Not only this, but also some combinations of marginal parameters result in extremely poor estimation performance, even for reasonably large sample sizes. In particular, when the range of possible values for δ is small relative to the size of the largest marginal parameter, the distribution simply does not change significantly over the range of possible values for δ. As an example, consider the case where $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (1, 2, 3, 4)$. Then $\delta \in \left[0, \frac{1}{2}\right]$, making the range of possible distributions so similar, a very large dataset is required to provide any hope for identifiability within this family. The good news is that we can assess the severity of this when bootstrapping for the confidence regions.

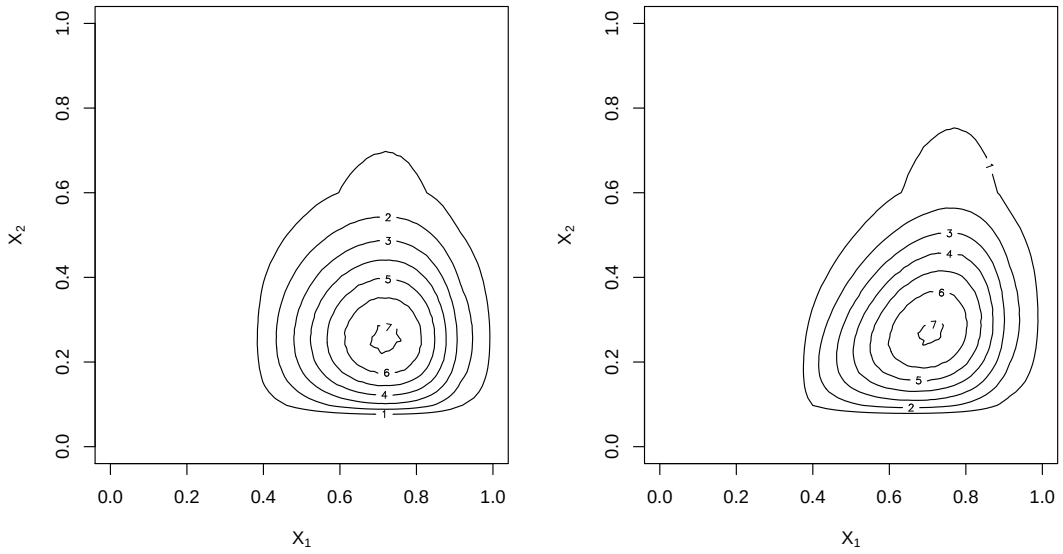


Figure 2.6: Densities (obtained by simulation) for $\boldsymbol{X} \sim BB(6, 3, 3, 7, 0, 0, 0, 0)$ (left), and $\boldsymbol{X} \sim BB(0, 0, 3, 4, 3, 0, 3, 0)$ (right).

We test this on the family discussed earlier, setting $\delta = 2$: $\boldsymbol{X} \sim BB(2, 1, 3, 5, 2, 0, 2, 0)$. The results, similar to those for Model 1, exhibit excessive variation from all parameters except δ_3, which is given by the marginal MLE only.

2.2.3 Summary of Bivariate Beta Sub-Models

The final two classes of models and the accompanying parameter estimation techniques are similar to Class C. While MMLE is the only method demonstrated here for these distributions, other methods were attempted with little improvement without adding significantly more complexity to the processes.

In light of this, some observations can be made about the 8-Parameter Bivariate Beta distribution and its sub-models. First, there is a clear trade-off between flexibility and a reasonable ability to perform parameter estimation. From the Olkin-Liu to the full, 8-Parameter, model, there

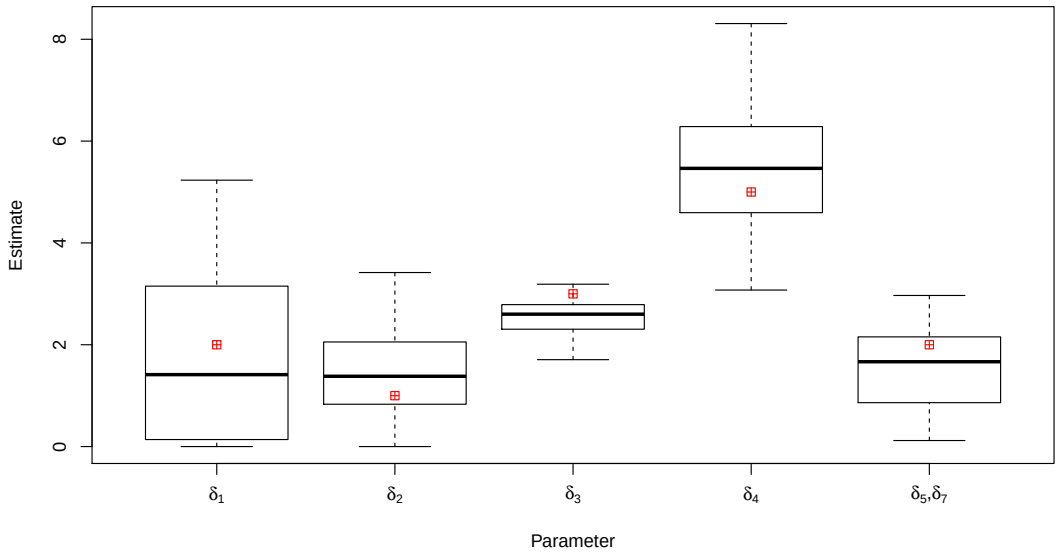


Figure 2.7: Parameter Estimation results for $\boldsymbol{X} \sim BB(2,1,3,5,2,0,2,0)$ with $K = 100$ with the actual values (red).

are many sub-models that may be useful for some applications. Arnold's 5-Parameter model only mildly restricts its marginals while also remaining flexible. On the other hand, any model allowing every collection of marginals, even when only one dependence parameter is active, is cumbersome when it comes to parameter estimation. The greatest problem manifests when more than one dependence parameter is introduced. How these dependence parameters affect the joint distributions of members of the same family with significantly different marginal distributions is vast, and thus it is an insurmountable problem without imposing some significant restrictions.

Ultimately, the most important lesson learned from this analysis is that if a dataset is small, that is to say, not "big data," then the marginally restricted models are preferred, but should be used only with caution, for their fitness may be questionable. For big data, however, the models which are not marginally restricted, such as those discussed in this section, may be useful, and parameter estimation is such cases is easy via MMLE methods for point estimations, and subsequent Bootstrap for confidence regions.

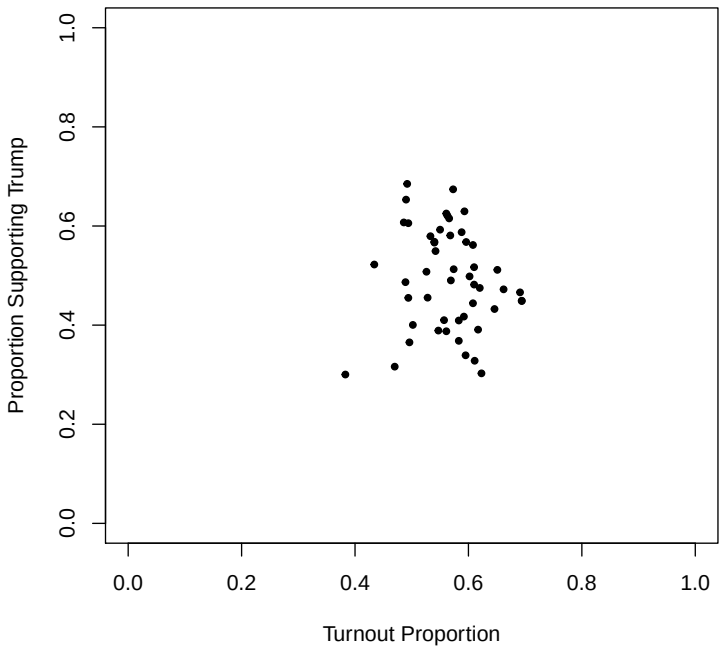


Figure 2.8: Voter turnout proportion vs. proportion supporting Trump for the 50 states.

In future work, much more can be done to explore this family, particularly for big data applications with tail dependence properties. In addition, more creatively formed sub-families may be capable of maintaining a high degree of flexibility while also eliminating the problem of practical identifiability. Specific sub-families should be constructed based on the known prior information about the phenomenon being studied.

2.2.4 An Example

We apply a few of the Bivariate Beta models discussed in this section to a dataset of size $K = 50$. In the 2016 election, voter turnout and the proportion of voters supporting Trump are recorded by state (D.C. is omitted as an obvious outlier). Set X_1 to be the voter turnout proportion and X_2 to be the proportion of voters supporting Trump. The data is given Figure 2.8. We apply

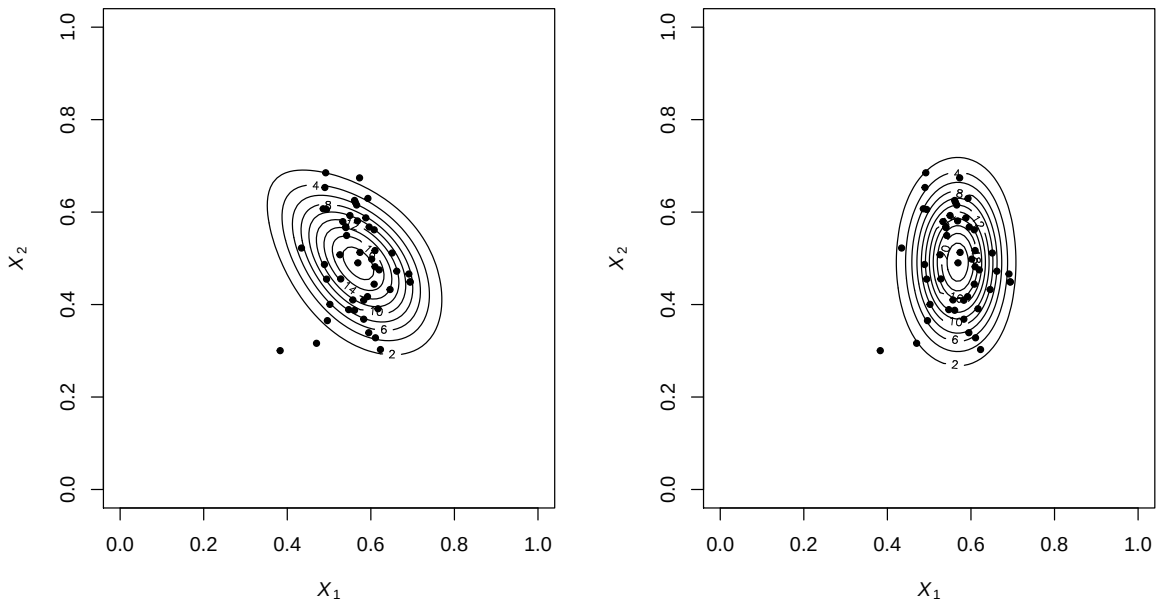


Figure 2.9: Olkin-Liu Model 3 (left) and Independent Model (right) Fitted to 2016 Election Data.

four models to this data: the 3-parameter Olkin-Liu Model 3, the 4-parameter independent model, a 5-parameter Arnold & Ng Model (rotated), and the 5-parameter Model 7 from Table 2.1; the results are given in Figures 2.9 and 2.10. In Table 2.2, the estimates for the appropriate parameters and the Akaike Information Criterion (AIC), which is computed from an estimated log-likelihood for the two 5-parameter models, are shown. The two plots in Figure 2.9 are constructed with the known densities and parameters given by the MLEs. To construct the densities in Figure 2.10, the ABC method applied by Crackel is used to estimate the parameters of the left density, and a similar ABC procedure is used to obtain estimates for the plot on the right. According to AIC, the 4-parameter independent model provides the best fit to the data.

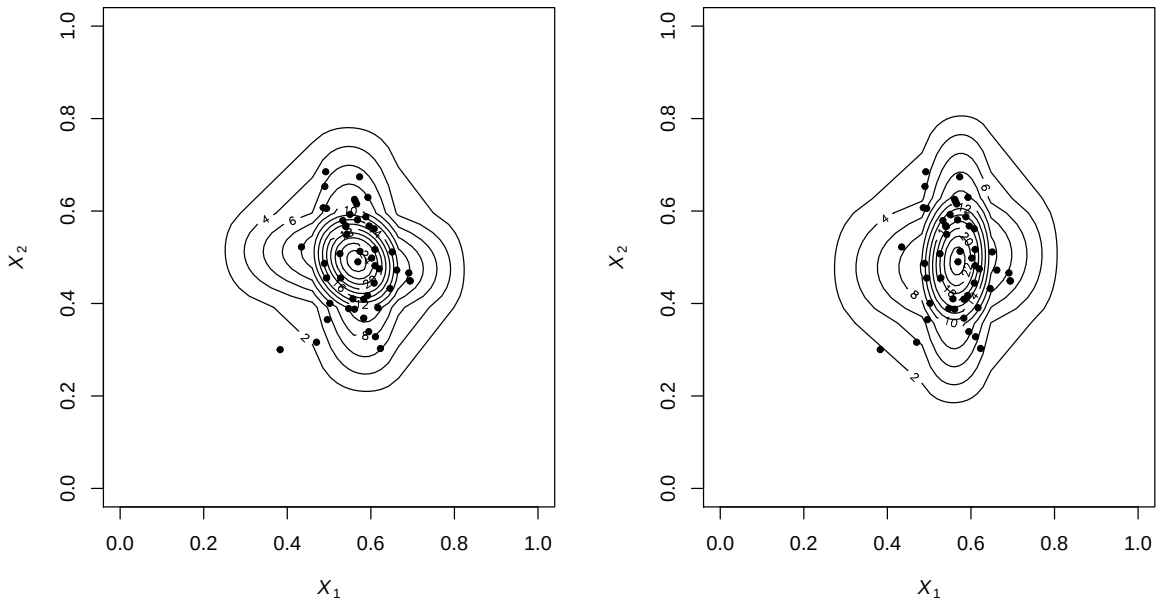


Figure 2.10: Arnold & Ng Model with $\delta_1 = \delta_2 = \delta_6 = 0$ (left) and Model 7 from Table 2.1 (right) Fitted to 2016 Election Data.

2.3 Bivariate Beta Copulas

When the marginal parameters of the 8-Parameter Bivariate Beta distribution are fixed, parameter estimation becomes a bit more tractable. As is shown in Equation 2.7, the number of free parameters is reduced to four. In addition, once the marginals are defined, the impacts of the four parameters are more predictable. In this section, we restrict the marginals to a special case: $X_1 \sim B(1,1)$, and $X_2 \sim B(1,1)$, that is, the marginals are $U(0,1)$. Formally, a multivariate distribution whose marginals are $U(0,1)$ is called a *copula*.[4] We, in this section, discuss exclusively this sub-family of copulas contained in the 8-Parameter Bivariate Beta model.

[4] A review of copulas can be found in [30].

Estimate	Olkin-Liu Model 3	Independent Model	Arnold & Ng (Rotated)	Model 7 from Table 2.1
δ_1	0	33.42	0	34.628
δ_2	0	11.72	0	10.088
δ_3	13.99	25.63	9.725	26.138
δ_4	11.13	12.09	11.257	10.461
δ_5	0	0	12.847	1.824
δ_6	0	0	0	0
δ_7	14.43	0	10.477	1.824
δ_8	0	0	8.327	0
AIC	−193.900	−213.2484	−185.0856	−209.9852

Table 2.2: Bivariate Beta models applied to Trump election data.

2.3.1 Definitions

The sub-family of copulas can be defined by expressing δ_1, δ_2, δ_3, and δ_4 in terms of δ_5, δ_6, δ_7, and δ_8, in order to generate uniform marginal distributions.

$$\begin{aligned} \delta_1 &= 1 - \delta_5 - \delta_7; \\ \delta_2 &= 1 - \delta_5 - \delta_8; \\ \delta_3 &= 1 - \delta_6 - \delta_8; \\ \delta_4 &= 1 - \delta_6 - \delta_7; \end{aligned} \tag{2.9}$$

so that the parameter space, Δ, may be defined by $(\delta_5, \delta_6, \delta_7, \delta_8) \in \Delta$ where

$$\begin{aligned} \delta_5 &\in [0, 1]; \\ \delta_6 &\in [0, 1]; \\ \delta_7 &\in [0, 1 - \max\{\delta_5, \delta_6\}]; \\ \delta_8 &\in [0, 1 - \max\{\delta_5, \delta_6\}]; \end{aligned} \tag{2.10}$$

Notation: Throughout this chapter, denote $(\delta_6, \delta_7, \delta_8)$ by $\delta_{(5)}$, $(\delta_5, \delta_7, \delta_8)$ by $\delta_{(6)}$, $(\delta_5, \delta_6, \delta_8)$

by $\delta_{(7)}$, and $(\delta_5, \delta_6, \delta_7)$ by $\delta_{(8)}$.

This subfamily will hereafter be called the Arnold & Ghosh 4-parameter family of copulas [4]. We will develop some useful methods for applying this model, and construct a method for parameter estimation for a small sub-collection of these models. The uniquely flexible characteristic of this family is a consequence of the fact that it is a multi-parameter copula, an unusual case in the world of copulas. The family includes the full range of possible correlations, i.e. it includes both the upper and lower Fréchet-Hoeffding bounds. In addition, it includes the product copula i.e. the bivariate $U(0, 1)$ distribution with independent marginals.

The cost associated with these benefits is complexity. Because the density is not available in general, maximum likelihood estimates of the parameters are inaccessible in closed form. In addition, mixed moments are also not analytically accessible and must be computed through simulation. And, as with any family of (continuous) copulas, accurate parameter estimation will require an unusually large amount of data. Thus, the problem of parameter estimation is far from trivial. Therefore, we propose a staged approach, in which the family is dismantled into smaller subclasses for which parameter estimation is more tractable, prior information and sample moments are used to eliminate obvious misfit subfamilies, and a final model is selected based on a reasonable metric. The following sections detail this process.

2.3.2 Sub-Families

A most useful way to break down the problem of parameter estimation into a more manageable form is to construct subfamilies with 1, 2, and 3 parameters. We will look more closely at some 1-parameter sub-families; larger families will be left for future work. These smaller families often allow for more accessible parameter estimation methods. As will be seen with a particular

collection of 1-parameter subfamilies, correlation can be strongly related to the single parameter, often monotonically. Figure 2.11 depicts this relationship for one such 1-parameter family, where the range of possible values for the correlation is complete. This will provide an easy method of elimination of some subfamilies whose correlation ranges do not include the data's sample correlation.

One-Parameter Subfamilies

We limit this discussion to a collection of 1-parameter sub-families whose Pearson correlation measures have one-to-one correspondence with the parameter. Table 2.3 gives a complete list of these copulas,[5] and Figures 2.11 and 2.12 show the corresponding correlation plots, for the first of each class, as functions of the single parameter (based on sample correlations of large simulated samples). Included is a quadratic model drawn from the data for each. Thus, given a specific model from this list, the parameter can be estimated with the sample correlation. In addition, confidence bounds can be constructed through a simple bootstrap procedure. It is easy to see that some of these families include one or both of the Fréchet-Hoeffding Bounds, while some include neither. Another important observation is that some are related by rotating about the horizontal and vertical lines through $\left(\frac{1}{2}, \frac{1}{2}\right)$, as with the general *BB* class, and thus are included within the same class.

Because the dependence structure is not always well-captured by a straight correlation measure, there are many other 1-parameter subfamilies with alternative defining characteristics. The number of such subfamilies is infinite; Table 2.4 provides a list of some of these cases.

Parameter estimation for these eleven models is less straightforward than the first nine-

[5]Roman class indexes in parentheses coincide with Arnold & Ghosh (2017).

Eq. Class	Family	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
A (IV)	1	0	0	0	0	δ	δ	$1-\delta$	$1-\delta$
B (V)	2	0	0	$1-\delta$	$1-\delta$	1	δ	0	0
	3	$1-\delta$	$1-\delta$	0	0	δ	1	0	0
	4	0	$1-\delta$	$1-\delta$	0	0	0	1	δ
	5	$1-\delta$	0	0	$1-\delta$	0	0	δ	1
C (VI)	6	$1-\delta$	$1-\delta$	1	1	δ	0	0	0
	7	1	1	$1-\delta$	$1-\delta$	0	δ	0	0
	8	$1-\delta$	1	1	$1-\delta$	0	0	δ	0
	9	1	$1-\delta$	$1-\delta$	1	0	0	0	δ
D (VII)	10	δ	δ	0	0	0	δ	$1-\delta$	$1-\delta$
	11	δ	0	0	δ	$1-\delta$	$1-\delta$	0	δ
	12	0	δ	δ	0	$1-\delta$	$1-\delta$	δ	0
	13	0	0	δ	δ	δ	0	$1-\delta$	$1-\delta$
E (VIII)	14	δ	1	$1-\delta$	0	0	δ	$1-\delta$	0
	15	$1-\delta$	0	δ	1	δ	0	0	$1-\delta$
F (IX)	16	1	δ	0	$1-\delta$	0	δ	0	$1-\delta$
	17	0	$1-\delta$	1	δ	δ	0	$1-\delta$	0
G	18	$1-\delta$	$1-\delta$	$1-\delta$	$1-\delta$	δ	δ	0	0
	19	$1-\delta$	$1-\delta$	$1-\delta$	$1-\delta$	0	0	δ	δ

Table 2.3: One-Parameter Families of Bivariate Beta Copulas with Parameter Monotonically Related to Correlation.

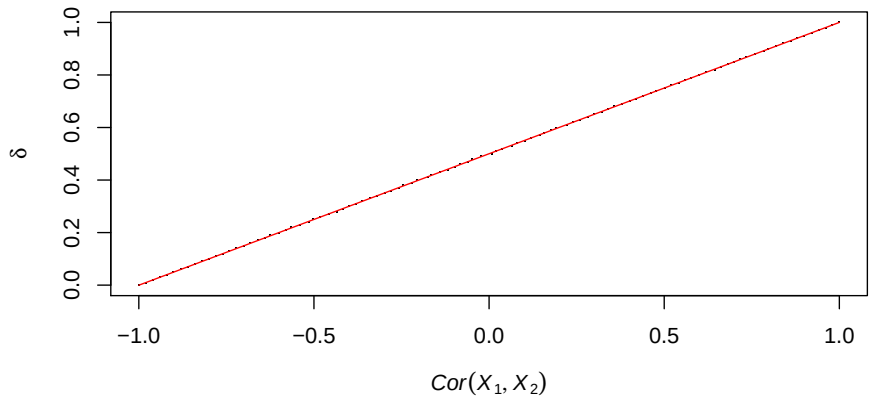


Figure 2.11: Correlation for Bivariate Beta Copula, Model 1. A linear model through $(0, 0)$ and $(1, 1)$ is included.

teen. It can be shown that Classes H and J have zero correlation for all possible values of δ, yet they have a clear dependence relationship for almost all values of δ. Classes H, I, and J include the product copula while Class K does not, and none of these contain the Fréchet-Hoeffding Bounds. While the members of Class I exhibit a monotonic relationship between δ and $Corr(X_1, X_2)$, it is a weak relationship, covering only about one eighth of the spectrum of possible correlation values, and correlation is not an important characteristic of this family. Lastly, the members of Class K exhibit a non-monotonic relationship between δ and $Corr(X_1, X_2)$, and only over a narrow window of possible correlations. In future work, many of these and other families can be further developed, and appropriate (sets of) statistics may be chosen specifically to form convenient one-to-one relationships with the parameter.

When it comes to model selection, prior information is very useful, providing a way of eliminating (or choosing) potential subfamilies. For example, some subfamilies include, say, the product copula and some do not. Also, prior knowledge may provide the expectation that the

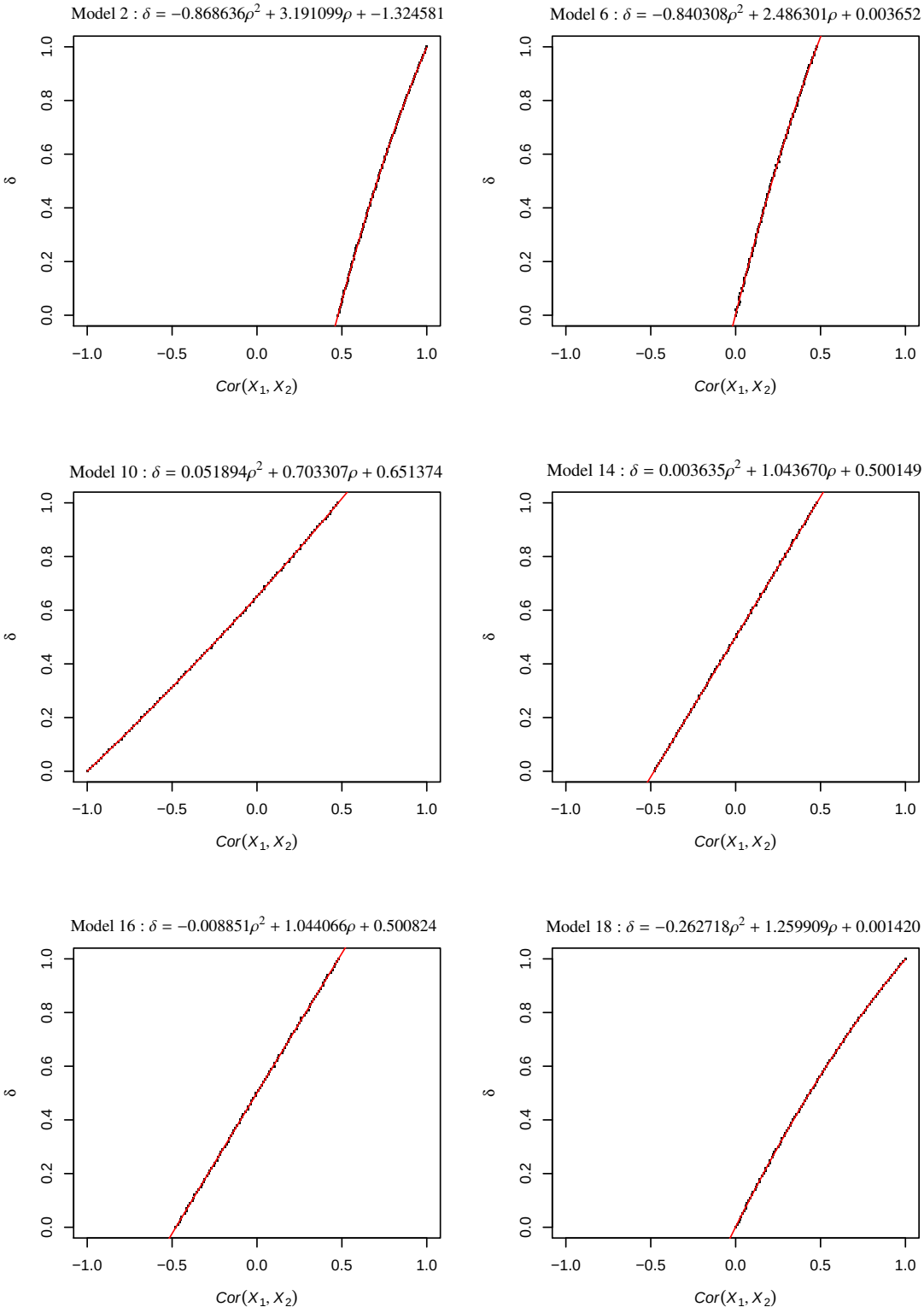


Figure 2.12: Correlation for Bivariate Beta Copula, Models 2, 6, 10, 14, 16, and 18.

Eq. Class	Family	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
H	20	$1-\delta$	$1-\delta$	$1-\delta$	$1-\delta$	$\frac{\delta}{2}$	$\frac{\delta}{2}$	$\frac{\delta}{2}$	$\frac{\delta}{2}$
I	21	$1-\frac{\delta}{2}$	$1-\frac{\delta}{2}$	$1-\delta$	$1-\delta$	0	$\frac{\delta}{2}$	$\frac{\delta}{2}$	$\frac{\delta}{2}$
	22	$1-\delta$	$1-\delta$	$1-\frac{\delta}{2}$	$1-\frac{\delta}{2}$	$\frac{\delta}{2}$	0	$\frac{\delta}{2}$	$\frac{\delta}{2}$
	23	$1-\frac{\delta}{2}$	$1-\delta$	$1-\delta$	$1-\frac{\delta}{2}$	$\frac{\delta}{2}$	$\frac{\delta}{2}$	0	$\frac{\delta}{2}$
	24	$1-\delta$	$1-\frac{\delta}{2}$	$1-\frac{\delta}{2}$	$1-\delta$	$\frac{\delta}{2}$	$\frac{\delta}{2}$	$\frac{\delta}{2}$	0
J	25	$1-\delta$	$1-\frac{\delta}{2}$	1	$1-\frac{\delta}{2}$	$\frac{\delta}{2}$	0	$\frac{\delta}{2}$	0
	26	$1-\frac{\delta}{2}$	$1-\delta$	$1-\frac{\delta}{2}$	1	$\frac{\delta}{2}$	0	0	$\frac{\delta}{2}$
	27	1	$1-\frac{\delta}{2}$	$1-\delta$	$1-\frac{\delta}{2}$	0	$\frac{\delta}{2}$	0	$\frac{\delta}{2}$
	28	$1-\frac{\delta}{2}$	1	$1-\frac{\delta}{2}$	$1-\delta$	0	$\frac{\delta}{2}$	$\frac{\delta}{2}$	0
K	29	$1-\delta$	$1-\delta$	δ	δ	δ	$1-\delta$	0	0
	30	$1-\delta$	δ	δ	$1-\delta$	0	0	δ	$1-\delta$

Table 2.4: Other One-Parameter Families of Bivariate Beta Copulas.

correlation is positive only, or provide knowledge of some type of symmetry for the data source. Many other prior assumptions can limit the applicable subfamilies as well. For example, Model 1 includes both Fréchet-Hoeffding Bounds, but does not include the product copula. Thus, knowledge that the phenomenon will never exhibit independence is useful for model selection. More generally, the known behavior of a system may allow for substantial narrowing of the potentially applicable models. Once a sub-collection of families is chosen (using prior information), parameter estimation is relatively easy via the method of moments.

Multi-Parameter Subfamilies

A vast collection of 2- and 3-parameter sub-families can be formed to possess various characteristics. A collection of them is given in Tables 2.5, and 2.6. We leave parameter estimation and model selection for these for future work.

Family	Parameters	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
31	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\delta-\gamma$	$1-2\gamma$	$1-\delta-\gamma$	$1-2\gamma$	δ	γ	γ	δ
32	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-2\delta$	$1-\delta-\gamma$	$1-2\gamma$	$1-\delta-\gamma$	δ	γ	δ	γ
33	$\delta, \gamma \in [0,1]\,; \delta+\gamma \leq 1$	$1-\delta-\gamma$	$1-\delta-\gamma$	$1-\delta-\gamma$	$1-\delta-\gamma$	δ	δ	γ	γ
34	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\gamma$	$1-\delta$	$1-2\delta$	$1-\delta-\gamma$	0	δ	γ	δ
35	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\delta$	$1-\gamma$	$1-\delta-\gamma$	$1-2\delta$	0	δ	δ	γ
36	$\delta, \gamma \in [0,1]\,; \delta+\gamma \leq 1$	$1-\gamma$	$1-\gamma$	$1-\delta-\gamma$	$1-\delta-\gamma$	0	δ	γ	γ
37	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\delta-\gamma$	$1-2\delta$	$1-\delta$	$1-\gamma$	δ	0	γ	δ
38	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-2\delta$	$1-\delta-\gamma$	$1-\gamma$	$1-\delta$	δ	0	δ	γ
39	$\delta, \gamma \in [0,1]\,; \delta+\gamma \leq 1$	$1-\delta-\gamma$	$1-\delta-\gamma$	$1-\gamma$	$1-\gamma$	δ	0	γ	γ
40	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\delta$	$1-2\delta$	$1-\delta-\gamma$	$1-\gamma$	δ	γ	0	δ
41	$\delta, \gamma \in [0,1]\,; \delta+\gamma \leq 1$	$1-\delta$	$1-\gamma$	$1-\delta-\gamma$	$1-\delta$	δ	δ	0	γ
42	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\delta$	$1-\delta-\gamma$	$1-2\gamma$	$1-\gamma$	δ	γ	0	γ
43	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-2\delta$	$1-\delta$	$1-\gamma$	$1-\delta-\gamma$	δ	γ	δ	0
44	$\delta, \gamma \in [0,1]\,; \delta+\gamma \leq 1$	$1-\delta-\gamma$	$1-\delta$	$1-\delta$	$1-\delta-\gamma$	δ	δ	γ	0
45	$\delta, \gamma \in \left[0, \frac{1}{2}\right]$	$1-\delta-\gamma$	$1-\delta$	$1-\gamma$	$1-2\gamma$	δ	γ	γ	0
46	$\delta, \gamma \in [0,1]$	$1-\gamma$	$1-\delta$	$1-\gamma$	$1-\gamma$	δ	γ	0	0
47	$\delta, \gamma \in [0,1]; \delta+\gamma \leq 1$	$1-\gamma$	1	$1-\delta$	$1-\delta$	0	δ	γ	0
48	$\delta, \gamma \in [0,1]; \delta+\gamma \leq 1$	$1-\delta-\gamma$	$1-\delta$	1	$1-\gamma$	δ	0	γ	0
49	$\delta, \gamma \in [0,1]$	$1-\delta$	$1-\gamma$	$1-\gamma$	$1-\delta$	0	0	δ	γ
50	$\delta, \gamma \in [0,1]; \delta+\gamma \leq 1$	1	$1-\gamma$	$1-\delta-\gamma$	$1-\delta$	0	δ	0	γ
51	$\delta, \gamma \in [0,1]; \delta+\gamma \leq 1$	$1-\delta$	$1-\delta-\gamma$	$1-\gamma$	1	δ	0	0	γ

Table 2.5: Two-Parameter Families of Bivariate Beta Copulas.

Family	Parameters	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
52	$\delta, \gamma, \theta \in [0,1]$; $\gamma, \theta \leq 1-\delta$	$1-\delta-\gamma$	$1-\delta-\theta$	$1-\delta-\gamma$	$1-\delta-\theta$	δ	δ	γ	θ
53	$\delta, \gamma, \theta \in [0,1]$; $\gamma \leq \frac{1}{2}; \delta, \theta \leq 1-\gamma$	$1-\delta-\gamma$	$1-\delta-\theta$	$1-\gamma-\theta$	$1-2\gamma$	δ	γ	γ	θ
54	$\delta, \gamma, \theta \in [0,1]$; $\delta, \gamma \leq 1-\theta$	$1-\delta-\theta$	$1-\delta-\theta$	$1-\gamma-\theta$	$1-\gamma-\theta$	δ	γ	θ	θ
55	$\delta, \gamma, \theta \in [0,1]$; $\delta \leq \frac{1}{2}; \gamma, \theta \leq 1-\delta$	$1-2\delta$	$1-\delta-\theta$	$1-\gamma-\theta$	$1-\delta-\gamma$	δ	γ	δ	θ
56	$\delta, \gamma, \theta \in [0,1]$; $\delta \leq \frac{1}{2}; \gamma, \theta \leq 1-\delta$	$1-\delta-\theta$	$1-2\delta$	$1-\delta-\gamma$	$1-\gamma-\theta$	δ	γ	θ	δ
57	$\delta, \gamma, \theta \in [0,1]$; $\gamma \leq \frac{1}{2}; \delta, \theta \leq 1-\gamma$	$1-\gamma-\theta$	$1-\delta-\gamma$	$1-2\gamma$	$1-\gamma-\theta$	δ	γ	θ	γ
58	$\delta, \gamma, \theta \in [0,1]$; $\delta, \gamma \leq 1-\theta$	$1-\delta-\theta$	$1-\delta$	$1-\gamma$	$1-\gamma-\theta$	δ	γ	θ	0
59	$\delta, \gamma, \theta \in [0,1]$; $\delta, \gamma \leq 1-\theta$	$1-\delta$	$1-\delta-\theta$	$1-\gamma-\theta$	$1-\gamma$	δ	γ	0	θ
60	$\delta, \gamma, \theta \in [0,1]$; $\gamma, \theta \leq 1-\delta$	$1-\delta-\gamma$	$1-\delta-\theta$	$1-\theta$	$1-\gamma$	δ	0	γ	θ
61	$\delta, \gamma, \theta \in [0,1]$; $\gamma, \theta \leq 1-\delta$	$1-\gamma$	$1-\theta$	$1-\delta-\theta$	$1-\delta-\gamma$	0	δ	γ	θ

Table 2.6: Three-Parameter Families of Bivariate Beta Copulas.

2.3.3 Model Selection

We now return to the nineteen families shown in Table 2.3. If no prior information is known, model selection is still possible, though a bit more difficult.

To begin, we must gain an understanding of the interesting characteristics of the model classes A through G. In Figures 2.13, and 2.14 we show the densities[6] for three cases of each class: $\delta = 0$, $\delta = 0.5$, and $\delta = 1$.

An immediate observation from these densities is that nothing about them is particularly surprising. Each of the dependence parameters, in all cases, tends have a strong (positive) association with a specific tail dependence, illustrated by the densities exhibiting concentration at the corners. Specifically, when δ_5 is significant, the density includes concentration near the corner at $(0, 0)$; when δ_6 is significant, the density includes concentration near the corner at $(1, 1)$; when δ_7 is significant, the density includes concentration near the corner at $(0, 1)$; and when δ_8 is significant, the density includes concentration near the corner at $(1, 0)$. For model selection, this can be exploited, most notably by matching both known (or expected) cases of tail dependence, or any known lack of the same. For example, if a phenomenon exhibits a tail dependency, one where, say, X_1 tends to be either small or large whenever X_2 is large, then Model 14 may be a reasonable candidate. So, it would be very useful to have a strong measure of tail dependencies.

We propose the vector, $\boldsymbol{S}(\boldsymbol{x})$, of eight statistics that appear to not only be effective at identifying tail dependencies, but also at providing a general identifiability for the full, 4-parameter

[6] When the joint density is not analytically accessible, it was estimated through simulating large data sets.

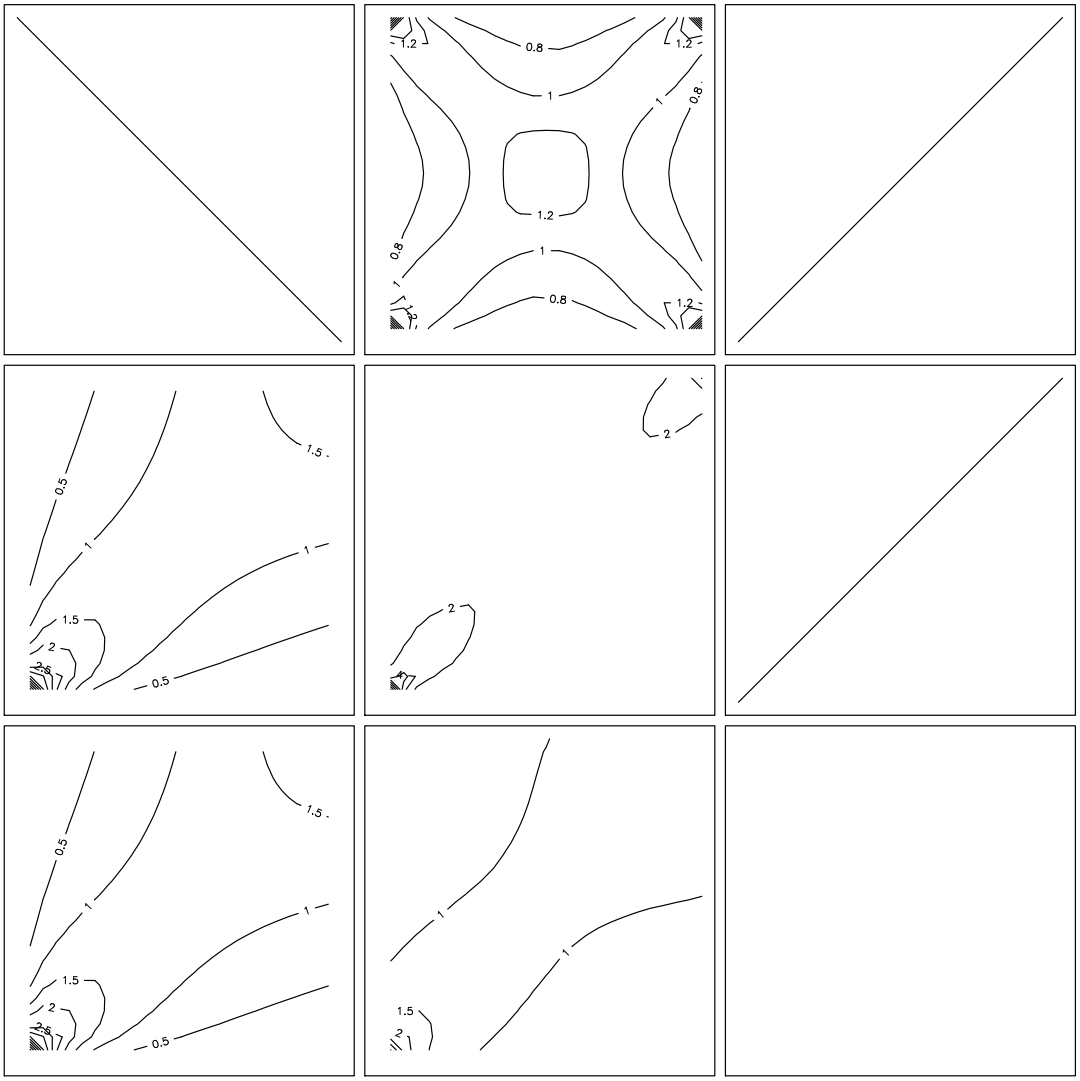


Figure 2.13: Densities for Models 1 (top), 2 (center), and 6 (bottom), for $\delta = 0$ (left), $\delta = 0.5$ (center), and $\delta = 1$ (right).

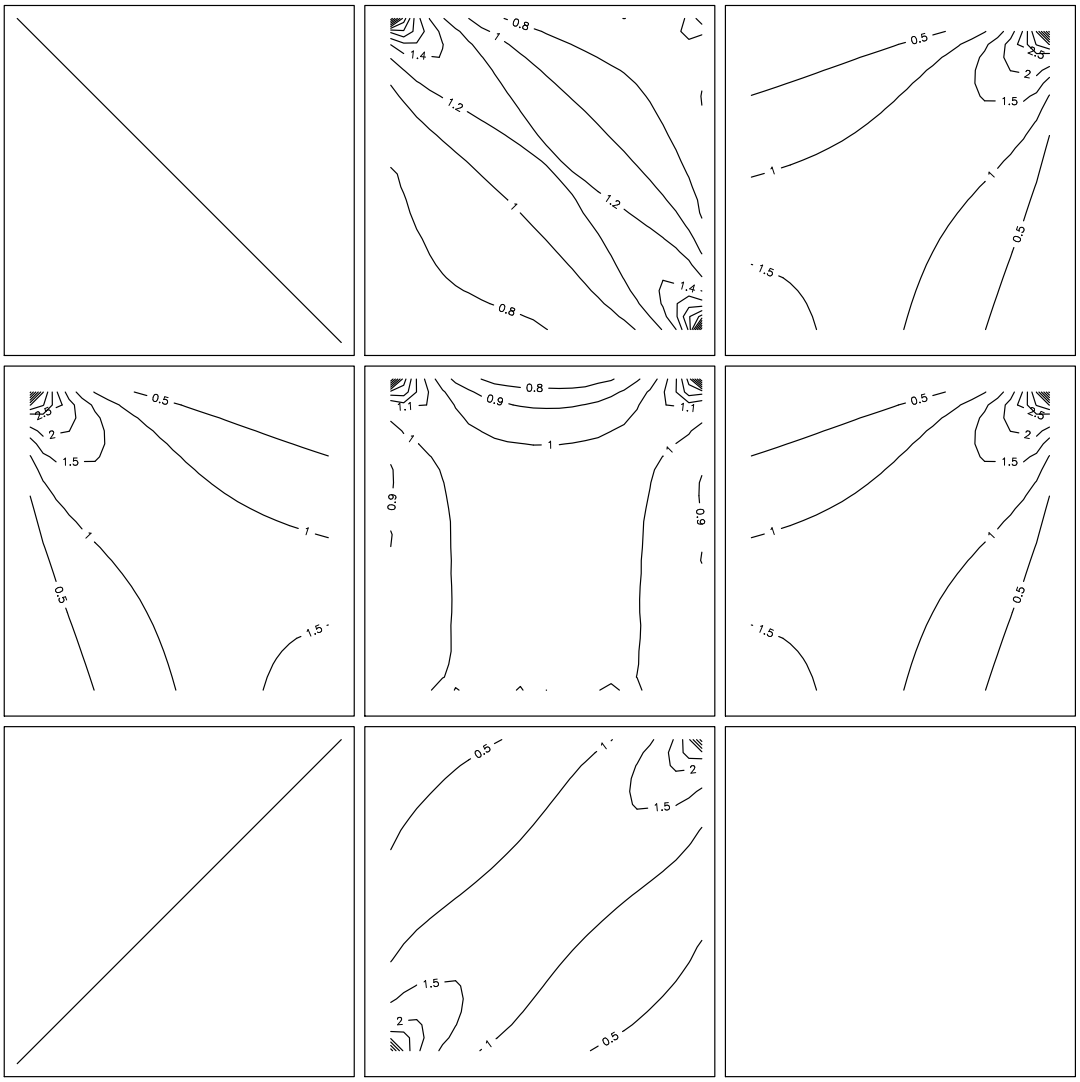


Figure 2.14: Densities for Models 10 (top), 14 (center), and 18 (bottom), for $\delta = 0$ (left), $\delta = 0.5$ (center), and $\delta = 1$ (right).

copula family. They are given in Equation 2.11.

$$
\begin{aligned}
S_1(\boldsymbol{x}) &= \frac{K_{11}}{K}, \text{ where } K_{11} = \sum_{k:x_{1k}\le\frac{1}{2}\&x_{2k}\le\frac{1}{2}} 1, \\
S_2(\boldsymbol{x}) &= \frac{K_{12}}{K}, \text{ where } K_{12} = \sum_{k:x_{1k}\le\frac{1}{2}\&x_{2k}>\frac{1}{2}} 1, \\
S_3(\boldsymbol{x}) &= \frac{K_{21}}{K}, \text{ where } K_{21} = \sum_{k:x_{1k}>\frac{1}{2}\&x_{2k}\le\frac{1}{2}} 1, \\
S_4(\boldsymbol{x}) &= \frac{K_{22}}{K}, \text{ where } K_{22} = \sum_{k:x_{1k}>\frac{1}{2}\&x_{2k}>\frac{1}{2}} 1, \\
S_5(\boldsymbol{x}) &= \frac{1}{K_{11}} \sum_{k:x_{1k}\le\frac{1}{2}\&x_{2k}\le\frac{1}{2}} \log\left[\frac{(1+x_{1k})(1+x_{2k})-2+(1-x_{1k})(1-x_{2k})}{(1-(1-x_{1k})(1-x_{2k}))^3}\right], \\
S_6(\boldsymbol{x}) &= \frac{1}{K_{12}} \sum_{k:x_{1k}\le\frac{1}{2}\&x_{2k}>\frac{1}{2}} \log\left[\frac{(1+x_{1k})(2-x_{2k})-2+(1-x_{1k})x_{2k}}{(1-(1-x_{1k})x_{2k})^3}\right], \\
S_7(\boldsymbol{x}) &= \frac{1}{K_{21}} \sum_{k:x_{1k}>\frac{1}{2}\&x_{2k}\le\frac{1}{2}} \log\left[\frac{(2-x_{1k})(1+x_{2k})-2+x_{1k}(1-x_{2k})}{(1-x_{1k}(1-x_{2k}))^3}\right], \\
S_8(\boldsymbol{x}) &= \frac{1}{K_{22}} \sum_{k:x_{1k}>\frac{1}{2}\&x_{2k}>\frac{1}{2}} \log\left[\frac{(2-x_{1k})(2-x_{2k})-2+x_{1k}x_{2k}}{(1-x_{1k}x_{2k})^3}\right],
\end{aligned}
\tag{2.11}
$$

The first four statistics are the proportions of data points in each of the four equally-sized quadrants in the unit square, and the last four are (partial) log-likelihood functions of a related distribution, which will be discussed further next.

The Ali-Michail-Haq copula, who's density is known, turns up in many applications. In the context of the 4-parameter Bivariate Beta Copula, it specifically appears when one of the dependence parameters is 1 and the other three are 0. The density in the case of $\delta_5 = 1$ is

$$
f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{(1+x_1)(1+x_2)-2+(1-x_1)(1-x_2)}{(1-(1-x_1)(1-x_2))^3} I\{0 \le x_1, x_2 \le 1\}
$$

This is the underlying density of the log-likelihood functions forming the last four statistics in Equation 2.11, and thus, the Ali-Michail-Haq distribution has an important tie to the larger 4-parameter

family of copulas. It turns out that these log-likelihood functions are useful measures of tail dependence for this particular family, better than Spearman's, Pearson's as well as other weighted correlation coefficients that were tested in this research.

The foundation is now set for the model selection process. Now, since the parameter space of this family of copulas (Equation 2.10) is compact; that is to say, it is a closed and bounded subset of $\mathbb{R}^4$ of volume $\frac{1}{6}$; a natural next step for any analyst would be to build a lattice of points within this space and measure the statistics in Equation 2.11 for a large simulated data set generated from each point in the lattice. Locating the point in the lattice producing those statistics with the closest (Euclidean) distance to the same statistics given by the original data should produce a reasonable starting point for model selection. In the context of Models 1 through 19, we can apply a similar routine, exploiting the monotonic relationship between correlation and δ for each to obtain an estimate, then record the statistics, $\boldsymbol{S}$, and choose the model yielding the closest value of $\boldsymbol{S}$ for the estimated value of δ. Denote the distribution function of Model $m \in \{1, 2, ..., 19\}$ with parameter δ by $F_m(\boldsymbol{x}; \delta)$. We detail the process below:

Step 1. Obtain estimates, $\hat{\delta}^{(1)}, \hat{\delta}^{(2)}, ..., \hat{\delta}^{(19)}$ using the method of moments.

Step 2. Compute $\boldsymbol{S}(\boldsymbol{x})$ from the original data, $\boldsymbol{x}$.

Step 3. Simulate samples, $\{\boldsymbol{y}_m\}_{m=1}^{19}$, of size $J \geq 10000$, from the distributions $F_m(\boldsymbol{x}; \hat{\delta}^{(m)})$ for all $m \in \{1, 2, ..., 19\}$.[7]

Step 4. Compute the sum of squared error, $[\boldsymbol{S}(\boldsymbol{x}) - \boldsymbol{S}(\boldsymbol{y}_m)]\,[\boldsymbol{S}(\boldsymbol{x}) - \boldsymbol{S}(\boldsymbol{y}_m)]'$, for all values of m.

[7]The value of J is chosen to be sufficiently large to provide overall identifiability for Models 1 though 19.

Step 5. Select Model m_0 with estimate $\hat{\delta}^{(m_0)}$, where m_0 is the subscript of the data set resulting in the smallest such SSE value.

We test this procedure by simulating 10000 data sets of size 100 from randomly-selected Models from Table 2.3 and random values of δ. Applying the above procedure, we find that it successfully chooses the correct model in 26% of the test cases. Of those correctly chosen, we computed the correlation from each data set, and used the models, examples of which are shown in Figures 2.11 and 2.12, to obtain estimates for δ. The average bias and mean squared error are given in Table 2.7. Given the effectiveness of the statistics $\boldsymbol{S}$, an ABC procedure may also be applied to those test cases the Model Search correctly identified, with the non-informative prior $(\tilde{\delta}_5, \tilde{\delta}_6, \tilde{\delta}_7, \tilde{\delta}_8) \sim Uniform(\Delta)$, SSE as ρ, and an $\epsilon_0 = 0.01$. The results of this are also provided in Table 2.7. In both cases, a bootstrap procedure can be applied to obtain interval estimates.

Estimator	Mean Bias	Mean Squared Error
Model Search with MOM	0.01677	0.01865
ABC	−0.04200	0.03179

Table 2.7: Model Search and ABC methods for 1-Parameter Bivariate Beta Copula, Models 1 through 19.

There may be some concern about the low percentage of correct model choices. There are two major reasons for this. First, the sample size is relatively small, so variations in the data can easily lead to different models becoming better fits. For large data sets, the proportion of cases in which the correct model is chosen can exceed 70%, but, even for extremely large data sets, it does not exceed 80%. This is likely attributable to the second reason: there is some practical overlap amongst the models. So it is expected that if a test case is within an overlap between two models,

either model could be selected without (practical) error. This reinforces the previous observation that use of prior knowledge is a valuable addition to the model selection process.

An optional addition to this procedure would be to engage in a search over some neighborhood of the estimate, in the unrestricted space, Δ. Definition of the neighborhood, however, is a subjective decision. One possibility is to define a prior distribution 'centered' at this estimate and proceed with ABC. In this way, parameter estimation can ultimately lead to any parameter set in Δ. The important question, left for future work, is whether or not the models in Table 2.3, provide sufficient coverage of Δ to justify this step. It can, however, be said with reasonable certainty that this step is applicable only to large data sets.

2.4 Conclusions

In this chapter we have investigated the 8-Parameter Bivariate Beta distribution, some of its sub-families, and particularly its sub-family of copulas.

We showed that difficulties arise for parameter estimation for not only the full model, but many of its sub-families, even some which are rather simple. This problem is made clear for sub-families whose marginal distributions are unrestricted because the behavior of the dependence structure changes with the marginal parameters. Sub-models for which parameter estimation is tractable will tend to have restricted marginals; that is, marginals which are functionally related to the dependence parameters. We conclude that the 8-parameter model should be viewed only as an omnibus of simpler models, each with both identifiable and useful characteristics.

For the sub-family of copulas, we exhibited the flexibility of the family, particularly in its ability to model tail dependencies and the full range of correlations. We reduced the problem of parameter estimation by reducing the number of parameters, resulting in nineteen models over seven

classes. This provided a means by which parameter estimation is accessible through the Method of Moments. Specifically, these sub-families were chosen, in part, for the parameter's monotonic relationship with a simple statistic: correlation. Model selection was accomplished by exploiting the compactness of the parameter space with a search based on a useful collection of statistics. ABC can also be applied using these same set of statistics. Multi-parameter models can be addressed in similar, though more tedious, methods.

Chapter 3

Bivariate Laplace Distributions

3.1 Introduction

The Laplace distribution, also known as the double exponential distribution, is often a useful alternative to the Gaussian distribution when heavier tails are desired. It is the error distribution associated with results obtained when absolute error, rather than squared error, is minimized in regression scenarios. Asymmetric Laplace distributions are, as the name would suggest, double exponential distributions where the scale parameters of the positive and negative parts of the density may be different. They are a more general form of the standard Laplace distribution, and are relevant for a larger variety of applications. In this chapter, we construct a family of bivariate Asymmetric Laplace ($BASL$) distributions, the most general form of which has 8 parameters, hereafter referred to as the "full model." We provide intuitive interpretations of the parameters, exhibit a similarity between the $BASL$ distribution and the Bivariate Beta Copula discussed in Chapter 2, and demonstrate a parameter estimation technique for a sub-family of this class of distributions. We finish this chapter with an example of a subfamily applied to a regression scenario with bivariate responses.

3.2 Definitions

Define the random variables $U_j \sim \Gamma(\delta_j, 1)$, $j = 1, 2, ..., 8$, where $\delta_j > 0$, $j \in \{1, 2, ..., 8\}$, $\beta_i > 0, i = 1, 2, 3, 4$, and

$$\begin{aligned} \delta_1 &= 1 - \delta_5 - \delta_7; \\ \delta_2 &= 1 - \delta_5 - \delta_8; \\ \delta_3 &= 1 - \delta_6 - \delta_8; \\ \delta_4 &= 1 - \delta_6 - \delta_7; \end{aligned} \tag{3.1}$$

Then, the random variable,

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \beta_3(U_3 + U_6 + U_8) - \beta_1(U_1 + U_5 + U_7) \\ \beta_4(U_4 + U_6 + U_7) - \beta_2(U_2 + U_5 + U_8) \end{bmatrix}, \tag{3.2}$$

has a bivariate asymmetric Laplace distribution, with a parameter space, $(\boldsymbol{\beta}, \boldsymbol{\delta})$ given by

$$\begin{aligned} \delta_5 &\in [0, 1]; \\ \delta_6 &\in [0, 1]; \\ \delta_7 &\in [0, 1 - \max\{\delta_5, \delta_6\}]; \\ \delta_8 &\in [0, 1 - \max\{\delta_5, \delta_6\}]; \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} \beta_1 &\in [0, \infty); \\ \beta_2 &\in [0, \infty); \\ \beta_3 &\in [0, \infty); \\ \beta_4 &\in [0, \infty). \end{aligned} \tag{3.4}$$

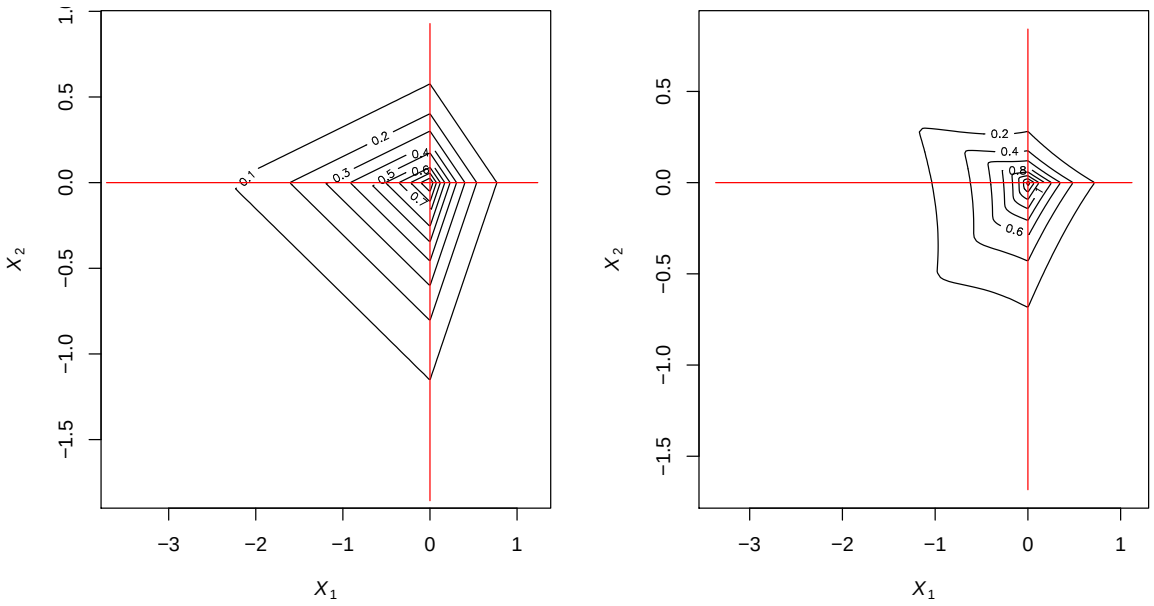


Figure 3.1: Densities of $BASL\left(\boldsymbol{\beta} = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\right), \boldsymbol{\delta} = (0,0,0,0)\right)$ (left), and $BASL\left(\boldsymbol{\beta} = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\right), \boldsymbol{\delta} = \left(\frac{2}{5}, 0, \frac{3}{5}, 0\right)\right)$ (right).

Notation: We will write $\boldsymbol{X} \sim BASL(\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4), \boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4))$, or more briefly $\boldsymbol{X} \sim BASL(\boldsymbol{\beta}, \boldsymbol{\delta})$.

Examples of the density are given in Figure 3.1.[1] It should be mentioned that even more general versions are possible, particularly those which include location parameters for each marginal. In this chapter, we restrict our study to the form shown in Equation 3.2, that is the version "centered" at $(0,0)$. We use the quotation marks here since the origin is not, in general, the mean of a $BASL$ distribution. The exact formula for the mean vector is given in the next section.

3.3 Moments

Moments are easy to compute for the $BASL$, though they become increasingly tedious to evaluate as the order of the moments increases. Some useful moments are given in Equations 3.5,

[1] The density on the left is for an independent case, so its density function is available in closed form. The one on the right has no closed-form density, and is generated though simulation and density estimation techniques.

3.6, 3.7, 3.8, and 3.9.

$$
\begin{aligned}
E[\boldsymbol{X}] &= \begin{pmatrix} E[X_1] \\ \\ E[X_2] \end{pmatrix} = \begin{pmatrix} E\left[\beta_3(U_3 + U_6 + U_8) - \beta_1(U_1 + U_5 + U_7)\right] \\ \\ E\left[\beta_4(U_4 + U_6 + U_7) - \beta_2(U_2 + U_5 + U_8)\right] \end{pmatrix} \\
&= \begin{pmatrix} \beta_3(\delta_3 + \delta_6 + \delta_8) - \beta_1(\delta_1 + \delta_5 + \delta_7) \\ \\ \beta_4(\delta_4 + \delta_6 + \delta_7) - \beta_2(\delta_2 + \delta_5 + \delta_8) \end{pmatrix} \\
&= \begin{pmatrix} \beta_3 - \beta_1 \\ \\ \beta_4 - \beta_2 \end{pmatrix}
\end{aligned} \tag{3.5}
$$

We next compute the covariance matrix:

$$
\begin{aligned}
Var[X_1] &= E[X_1^2] - E[X_1]^2 \\
&= E\left[\beta_3^2(U_3 + U_6 + U_8)^2\right] + E\left[\beta_1^2(U_1 + U_5 + U_7)^2\right] \\
&\quad - 2\beta_1\beta_3 E[(U_1 + U_5 + U_7)]E[(U_3 + U_6 + U_8)] \\
&\quad - \beta_3^2 E[(U_3 + U_6 + U_8)]^2 - \beta_1^2 E[(U_1 + U_5 + U_7)]^2 \\
&\quad + 2\beta_1\beta_3 E[(U_1 + U_5 + U_7)]E[(U_3 + U_6 + U_8)] \\
&= \beta_3^2 E\left[(U_3 + U_6 + U_8)^2\right] - \beta_3^2 E[(U_3 + U_6 + U_8)]^2 \\
&\quad + \beta_1^2 E\left[(U_1 + U_5 + U_7)^2\right] - \beta_1^2 E[(U_1 + U_5 + U_7)]^2 \\
&= \beta_3^2 Var[(U_3 + U_6 + U_8)] + \beta_1^2 Var[(U_1 + U_5 + U_7)] \\
&= 2\left[\beta_3^2 + \beta_1^2\right]
\end{aligned} \tag{3.6}
$$

Similarly,

$$
Var[X_2] = 2\left[\beta_4^2 + \beta_2^2\right] \tag{3.7}
$$

The covariance is

$$\begin{aligned} Cov[X_1, X_2] =& E\left[(\beta_3(U_3 + U_6 + U_8) - \beta_1(U_1 + U_5 + U_7)) \cdot \right. \\ & \left. (\beta_4(U_4 + U_6 + U_7) - \beta_2(U_2 + U_5 + U_8))\right] - E[X_1]E[X_2] \\ =& \beta_3\beta_4 E[U_6^2] - \beta_2\beta_3 E[U_8^2] - \beta_1\beta_4 E[U_7^2] + \beta_1\beta_2 E[U_5^2] \\ & - \beta_3\beta_4 E[U_6]^2 + \beta_2\beta_3 E[U_8]^2 + \beta_1\beta_4 E[U_7]^2 - \beta_1\beta_2 E[U_5]^2 \\ =& \beta_3\beta_4 Var[U_6] - \beta_2\beta_3 Var[U_8] - \beta_1\beta_4 Var[U_7] + \beta_1\beta_2 Var[U_5] \\ =& \beta_3\beta_4\delta_6 - \beta_2\beta_3\delta_8 - \beta_1\beta_4\delta_7 + \beta_1\beta_2\delta_5 \end{aligned} \tag{3.8}$$

Thus, the variance-covariance matrix is given by

$$Var[\boldsymbol{X}] = \begin{pmatrix} 2\left[\beta_3^2 + \beta_1^2\right] & \beta_3\beta_4\delta_6 - \beta_2\beta_3\delta_8 - \beta_1\beta_4\delta_7 + \beta_1\beta_2\delta_5 \\ \beta_3\beta_4\delta_6 - \beta_2\beta_3\delta_8 - \beta_1\beta_4\delta_7 + \beta_1\beta_2\delta_5 & 2\left[\beta_4^2 + \beta_2^2\right] \end{pmatrix} \tag{3.9}$$

3.4 Interpretation of Parameters

For this family, it is easy to see that, by construction, $\boldsymbol{\beta}$ completely determines the marginal distributions of $\boldsymbol{X}$, i.e.

$$X_1 \sim ASL(\beta_1, \beta_3), \text{ so that } f_{X_1}(x_1) = \frac{1}{\beta_1 + \beta_3} \begin{cases} e^{\frac{x_1}{\beta_1}}, \text{ if } x_1 < 0 \\ e^{-\frac{x_1}{\beta_3}}, \text{ if } x_1 \geq 0 \end{cases}, \tag{3.10}$$

and

$$X_2 \sim ASL(\beta_2, \beta_4), \text{ so that } f_{X_2}(x_2) = \frac{1}{\beta_2 + \beta_4} \begin{cases} e^{\frac{x_2}{\beta_2}}, \text{ if } x_2 < 0 \\ e^{-\frac{x_2}{\beta_4}}, \text{ if } x_2 \geq 0 \end{cases}. \tag{3.11}$$

where ASL represents the asymmetric Laplace distribution. The marginal MLEs are available in closed form:

$$\hat{\beta}_1 = \mu_1 + \sqrt{\mu_1\mu_3}; \quad \hat{\beta}_3 = \mu_3 + \sqrt{\mu_1\mu_3} \tag{3.12}$$

and

$$\hat{\beta}_2 = \mu_2 + \sqrt{\mu_2\mu_4}; \quad \hat{\beta}_4 = \mu_4 + \sqrt{\mu_2\mu_4} \tag{3.13}$$

where

$$\begin{aligned} \mu_1 &= \frac{1}{K}\sum_{x_1<0} x_1, \\ \mu_2 &= \frac{1}{K}\sum_{x_2<0} x_2, \\ \mu_3 &= \frac{1}{K}\sum_{x_1\geq 0} x_1, \text{ and} \\ \mu_4 &= \frac{1}{K}\sum_{x_2\geq 0} x_2. \end{aligned} \tag{3.14}$$

It should be noted that these are not simple means of the positive and negative values of X_1 and X_2, which, themselves, can be used to compute estimators of $\boldsymbol{\beta}$; in contrast, the denominators in all four expressions are K, the size of the entire dataset. These MLEs will be useful for parameter estimation in the next section.

The other four parameters, $\boldsymbol{\delta}$, bear a strong resemblance to the same parameters of the Bivariate Beta Copula, for they range over the same set of values, and each influences the two distributions in a similar way, particularly through tail dependencies. Referring, once again, to Figure 3.1, we see that when δ_5 and δ_7 are active, i.e., not equal to 0, we have two concurrent tail dependencies, as are seen by the stretching of the contour lines away from the origin in the third and fourth quadrants.

This is distinct from the situation with other bivariate asymmetric Laplace distributions. One such family was introduced by Kotz, et. al. [25], who constructed a multivariate asymmetric

Laplace family of distributions based on the Gaussian distribution's dependence properties. The two-dimensional version has density,

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{\exp\left[\frac{\left(\left(\frac{m_1\sigma_2}{\sigma_1}-m_2\rho\right)y_1+\left(\frac{m_2\sigma_1}{\sigma_2}-m_1\rho\right)y_2\right)}{\sigma_1\sigma_2(1-\rho^2)}\right]}{\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\cdot K_0\left(C(m_1,m_2,\sigma_1,\sigma_2)\sqrt{\frac{y_1^2\sigma_2}{\sigma_1}-2\rho y_1y_2+\frac{y_2^2\sigma_1}{\sigma_2}}\right),$$

where

$$C(m_1,m_2,\sigma_1,\sigma_2) = \frac{\sqrt{2\sigma_1\sigma_2(1-\rho^2)+\frac{m_1^2\sigma_2}{\sigma_1}-2m_1m_2\rho+\frac{m_2^2\sigma_1}{\sigma_2}}}{\sigma_1\sigma_2(1-\rho^2)},$$

and $K_0(\cdot)$ is a the modified Bessel function of the third kind (See Kotz).

This distribution has 5 parameters, $(m_1, m_2, \sigma_1, \sigma_2, \rho)$. m_1 and m_2 control skewness, and all five are responsible for dependence structure. Consequently this large family of distributions is capable of representing many unique combinations of dependence structure and marginals. In addition, observations from this distribution can easily be simulated through the relation,

$$\boldsymbol{Y} \stackrel{d}{=} \boldsymbol{m}W + \sqrt{W}\boldsymbol{Z}, \tag{3.15}$$

where $W \sim \exp(1)$, $\boldsymbol{Z} \sim BVN(\boldsymbol{0}, \Sigma)$, $W \perp \boldsymbol{Z}$, and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \tag{3.16}$$

Similar to the observations made in Chapter 2, the marginal distributions are related to the dependence structure. That is, given a particular pair of marginal distributions, the Kotz family is limited in its capacity to exhibit specific dependence structures. In contrast, though the $\boldsymbol{X} \sim BASL(\boldsymbol{\delta}, \boldsymbol{\beta})$ may provide additional flexibility, it has the defect that it fails to have closed form expressions for its densities. Hence, for it, likelihood-free techniques must be applied.

3.5 Statistical Inference

Assume, first, that $\boldsymbol{\beta} = (1, 1, 1, 1)$. In this case, both marginals have standard Laplace distributions. Additionally, a method of moments solution for $\boldsymbol{\delta}$ is available in closed form. The 2nd- and 3rd-order mixed moments yield covariances with simple forms:

$$\begin{aligned}
\gamma_{11} &= Cov[X_1, X_2] = \delta_5 + \delta_6 - \delta_7 + \delta_8 \\
\gamma_{21} &= Cov[X_1^2, X_2] = 2(-\delta_5 + \delta_6 + \delta_7 - \delta_8) \\
\gamma_{12} &= Cov[X_1, X_2^2] = 2(-\delta_5 + \delta_6 - \delta_7 + \delta_8)
\end{aligned} \tag{3.17}$$

Higher-order moments become increasingly complex, but a 4th-order mixed moment ultimately yields

$$\gamma_{22} = Cov[X_1^2, X_2^2] = 2(2 + \delta_5^2 + 3\delta_6 + 3\delta_7 + \delta_5(3 + 2\delta_6 - 2\delta_7 - 2\delta_8) + 3\delta_8 + (-\delta_6 + \delta_7 + \delta_8)^2) \tag{3.18}$$

The corresponding sample moments may be computed to provide an estimate for $\boldsymbol{\delta}$:

$$\begin{aligned}
\hat{\delta}_5 &= \frac{1}{24}\left(6\hat{\gamma}_{11} - 2\hat{\gamma}_{11}^2 - 3\hat{\gamma}_{21} - 3\hat{\gamma}_{12} - \hat{\gamma}_{22} - 4\right) \\
\hat{\delta}_6 &= \frac{1}{24}\left(6\hat{\gamma}_{11} - 2\hat{\gamma}_{11}^2 + 3\hat{\gamma}_{21} + 3\hat{\gamma}_{12} - \hat{\gamma}_{22} - 4\right) \\
\hat{\delta}_7 &= \frac{1}{24}\left(-6\hat{\gamma}_{11} - 2\hat{\gamma}_{11}^2 + 3\hat{\gamma}_{21} - 3\hat{\gamma}_{12} - \hat{\gamma}_{22} - 4\right) \\
\hat{\delta}_8 &= \frac{1}{24}\left(-6\hat{\gamma}_{11} - 2\hat{\gamma}_{11}^2 - 3\hat{\gamma}_{21} + 3\hat{\gamma}_{12} - \hat{\gamma}_{22} - 4\right)
\end{aligned} \tag{3.19}$$

where $\hat{\gamma}_{ij}$ is the sample correlation corresponding to γ_{ij}, for $i, j \in \{1, 2\}$. These estimators provide reasonable estimates for $K \approx 10^5$ or larger. For smaller sample sizes, alternative methods are necessary.

Now, when we re-introduce $\boldsymbol{\beta}$ as unknown, we can obtain the marginal MLEs as described previously. However, the data cannot be transformed so that it has standard Laplace marginals.

Consider the most intuitive means by which to do this: normalizing the data in each of the four quadrants in $\mathbb{R}^2$ with the corresponding β's, i.e., set

$$\boldsymbol{x}^{++} = \{\boldsymbol{x}_k : x_{1k}, x_{2k} \geq 0\},$$

$$\boldsymbol{x}^{+-} = \{\boldsymbol{x}_k : x_{1k} \geq 0 \wedge x_{2k} < 0\},$$

$$\boldsymbol{x}^{-+} = \{\boldsymbol{x}_k : x_{1k} < 0 \wedge x_{2k} \geq 0\},$$

$$\boldsymbol{x}^{--} = \{\boldsymbol{x}_k : x_{1k}, x_{2k} < 0\}.$$

Then, defining

$$\boldsymbol{y}^{++} = (\boldsymbol{x}^{++})' B_{34},$$

$$\boldsymbol{y}^{+-} = (\boldsymbol{x}^{+-})' B_{32},$$

$$\boldsymbol{y}^{-+} = (\boldsymbol{x}^{-+})' B_{14},$$

$$\boldsymbol{y}^{--} = (\boldsymbol{x}^{--})' B_{12};$$

where B_{ij} is a two-row matrix with an appropriate number of columns to permit matrix multiplication, and β_i^{-1} in all entries of the first row, and β_j^{-1} in all entries of the second row, $i \in \{1, 3\}$, and $j \in \{2, 4\}$. Then, each of $|\boldsymbol{y}^{++}|$, $|\boldsymbol{y}^{++}|$, $|\boldsymbol{y}^{++}|$, and $|\boldsymbol{y}^{++}|$ can easily be shown to be from standard bivariate exponential distributions. However, the data $\boldsymbol{y} = \boldsymbol{y}^{++} \cup \boldsymbol{y}^{++} \cup \boldsymbol{y}^{++} \cup \boldsymbol{y}^{++}$ is from a distribution with standard Laplace marginals if and only if $\beta_1 = \beta_3$, and $\beta_2 = \beta_4$. Otherwise, the expected value will not be $(0, 0)$, that is there are (expected to be) different *amounts* of data in each quadrant. Attempting an additional transformation to resolve this is a risky endeavor, and will not be done here. Rather, we choose an alternative method for characterizing this family.

Suppose $\boldsymbol{x}$ is an observed random sample of size K from $\boldsymbol{X} \sim BASL(\boldsymbol{\beta}, \boldsymbol{\delta})$, where $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are unknown. With the marginal MLEs, we can obtain estimates for $\boldsymbol{\beta}$, setting the stage for a variety

of methods. We focus on one: an MCMC method with data-informed proposals.

Consider the following process. First, select a large sample, $\left(\boldsymbol{\delta}^{(1)}, \boldsymbol{\delta}^{(2)}, ..., \boldsymbol{\delta}^{(M)}\right)$. This sample can be randomly selected from $\tilde{\boldsymbol{\delta}} \sim Uniform(\Delta)$, where Δ is the parameter space defined in Equation 2.10, or be consciously decided upon, e.g. a lattice. Obtain simulated samples, $\left(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(M)}\right)$, from $F_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\beta}}, \boldsymbol{\delta}^{(m)}), m \in \{1, 2, ..., M\}$, respectively, where $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4\right)$, is a vector of the *marginal* MLEs given in Equations 3.12 and 3.13. Now assume $\hat{\boldsymbol{\beta}}$ is the true value of $\boldsymbol{\beta}$, and estimate the density, $f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\beta}}, \boldsymbol{\delta}^{(m)})$ based on each data set, $\boldsymbol{y}^{(m)}$, and using the method of density estimation outlined in Appendix A. Compute the likelihood of $\boldsymbol{x}$ for each $\boldsymbol{\delta}^{(m)}$ based on these estimated densities.

At this point there are a couple of options. If M is very large, set $\hat{\boldsymbol{\delta}} = \boldsymbol{\delta}^{(m_0)}$ where $f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\beta}}, \boldsymbol{\delta}^{(m_0)})$ is the density estimate resulting in the largest value of the likelihood, i.e. obtain a brute-force maximum likelihood estimate. The MMLE estimate would then be $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}\right)$. Alternatively, one can consider a sub-collection of the $\boldsymbol{y}^{(m)}$'s resulting in the 'best' likelihood values, and apply some carefully chosen model to these. This model can then be used as a proposal distribution for $\boldsymbol{\delta}$ in a MCMC process. A bootstrap process can then be applied to obtain simulated values from the marginal maximum likelihood estimators, $\hat{\boldsymbol{\beta}}$. In the next section, we demonstrate the latter on a 4-parameter sub-model of the 8-parameter family.

3.5.1 MCMC Parameter Estimation

Suppose it is known that, in reality, $\delta_7 = \delta_8 = 0, \beta_2 = \beta_4 = 1$, and that the observed sample, $\boldsymbol{x}$, is of size $K = 250$. In addition, suppose the true parameters are $\boldsymbol{\beta} = (3, 1, 5, 1)$ and $\boldsymbol{\delta} = (0.14, 0.71, 0, 0)$. We construct a Markov Chain whose limiting distribution will estimate the distribution of the true MLE of $(\boldsymbol{\beta}, \boldsymbol{\delta})$.

For the proposal distribution, we apply the procedure summarized at the end of the previous section. Experience suggests that the estimators, $\hat{\delta}_5$ and $\hat{\delta}_6$, exhibit a negative correlation. Thus, we choose the model for $(\hat{\delta}_5, \hat{\delta}_6)$ to be the Olkin-Liu Bivariate Beta, Model 4, denoted BB_{OL_4}, as defined in Section 2.2.1, particularly for its simplicity and its negative correlation. The process for obtaining a proposal distribution for $(\boldsymbol{\beta}, \boldsymbol{\delta})$ is detailed as follows:

Step 1. Compute the marginal MLEs, $\hat{\boldsymbol{\beta}}$, from the original data, $\boldsymbol{x}$.

Step 2. Set $m_0 \in \mathbb{N}$. This value will be the number of samples required to determine the MLE for the proposal distribution for $\boldsymbol{\delta}$. We set it at 40. Sample $n_0 = m_0^2$ values, $(d_{51}, d_{61}), (d_{52}, d_{62}), ..., (d_{5n_0}, d_{6n_0})$ from a $Uniform(S)$ distribution, where S is the unit square: $S = (0,1) \times (0,1)$; call the sample $\mathcal{L}$.

Step 3. Simulate large samples, $\left\{\boldsymbol{y}_{(d_5,d_6)}\right\}_{(d_5,d_6)\in\mathcal{L}}$, of size, say, $J \geq 25000$, each from the random variable $Y_{d_5,d_6} \sim BASL(\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \boldsymbol{\delta} = (d_5, d_6, 0, 0))$ corresponding to the ordered pair, $(d_5, d_6) \in \mathcal{L}$.

Step 4. Estimate the densities, $\hat{f}_{d_5,d_6}(\boldsymbol{y}; \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \boldsymbol{\delta} = (d_5, d_6, 0, 0))$ of Y_{d_5,d_6} using the method in Appendix A, from the datasets, $\left\{\boldsymbol{y}_{(d_5,d_6)}\right\}_{(d_5,d_6)\in\mathcal{L}}$.

Step 5. Compute $\hat{L}(\delta_5, \delta_6 | \boldsymbol{X} = \boldsymbol{x}) = \prod_{k=1}^{K} \hat{f}_{d_5,d_6}(\boldsymbol{x}_k; \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \boldsymbol{\delta} = (d_5, d_6, 0, 0))$, for all $(d_5, d_6) \in \mathcal{L}$.

Step 6. Select the m_0 ordered pairs, $(d_{51}, d_{62}), (d_{51}, d_{62}), .., (d_{5m_0}, d_{6m_0})$, from $\mathcal{L}$ that correspond to the largest values of $\hat{L}(\delta_5, \delta_6 | \boldsymbol{X} = \boldsymbol{x})$. From this dataset, obtain MLEs, $(\hat{a}, \hat{b}, \hat{c})$, for $(\tilde{\delta}_5, \tilde{\delta}_6) \sim BB_{OL_4}(a, b, c)$. The proposal distribution for (δ_5, δ_6) is $BB_{OL_4}(\hat{a}, \hat{b}, \hat{c})$.

The process of obtaining a single proposal value is as follows:

Step 1. Obtain a sample, $\boldsymbol{\delta}^* = (\delta_5^*, \delta_6^*, 0, 0)$, from $(\tilde{\delta}_5, \tilde{\delta}_6) \sim BB_{OL_4}(\hat{a}, \hat{b}, \hat{c})$.

Step 2. Obtain a sample, $\boldsymbol{y}^*$, of size K from $X \sim BAS\, L(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}, \boldsymbol{\delta} = \boldsymbol{\delta}^*)$, where $\hat{\boldsymbol{\beta}}$ is the set of marginal MLEs for $\boldsymbol{\beta}$. Compute the marginal MLEs, $\boldsymbol{\beta}^*$, from the new sample $\boldsymbol{y}^*$.

Step 3. The proposal is $(\boldsymbol{\beta} = \boldsymbol{\beta}^*, \boldsymbol{\delta} = \boldsymbol{\delta}^*)$.

For the MCMC process, we apply the standard Metropolis-Hastings algorithm, shown in Section 1.4.1, using the density estimation method detailed in Appendix A. We detail this process next:

Step 1. Set $t = 1$. Set the initial value $(\boldsymbol{\beta}, \boldsymbol{\delta})^{(0)} = (\hat{\boldsymbol{\beta}}, \boldsymbol{\delta} = (0, 0, 0, 0))$,. Since this parameter set represents independent marginals, we know the corresponding density and thus the likelihood: $\hat{L}((\boldsymbol{\beta}, \boldsymbol{\delta})^{(0)}|\boldsymbol{x}) = L(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\delta}}|\boldsymbol{x})$.

Step 2. Step t to $t + 1$. Generate a single proposal value, $(\boldsymbol{\beta}, \boldsymbol{\delta})^*$, from the proposal distribution detailed above.

Step 3. Simulate an iid sample, $\boldsymbol{y}^*$, of M realizations from $\boldsymbol{Y} \sim F(\boldsymbol{y}; (\boldsymbol{\beta}, \boldsymbol{\delta})^*)$, where M is sufficiently large to accurately estimate f at all points in $\boldsymbol{x}$. In this case, we choose $M = 25000$.

Step 4. Estimate $\hat{f}(\boldsymbol{x}_k|(\boldsymbol{\beta}, \boldsymbol{\delta})^*)$ for all $k \in \{1, 2, ..., K\}$ based on $\boldsymbol{y}^*$, and compute the estimated likelihood, $\hat{L}((\boldsymbol{\beta}, \boldsymbol{\delta})^*|\boldsymbol{x})$.

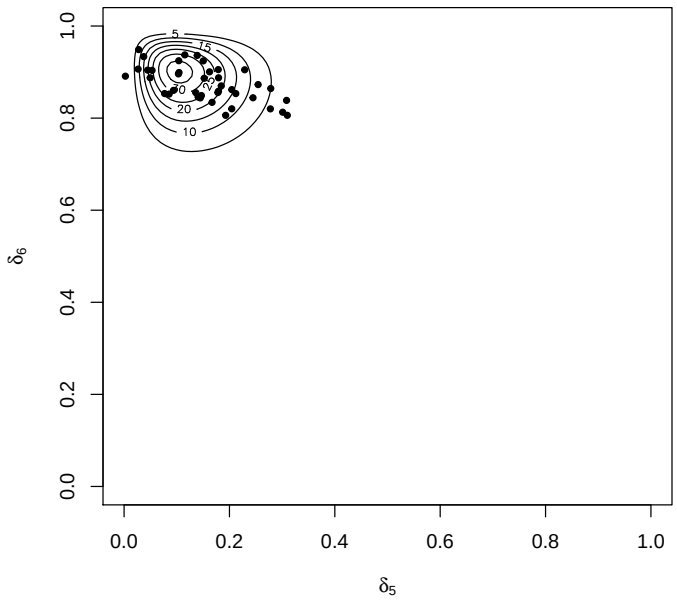


Figure 3.2: Initial $\boldsymbol{\delta}$ sample and corresponding $BB_{OL_4}(3.12, 3.26, 19.72)$ proposal model.

Step 5. Simulate a single realization, u, from $U \sim U(0,1)$. If $u < \frac{\hat{L}((\boldsymbol{\beta},\boldsymbol{\delta})^*|\boldsymbol{x})}{\hat{L}((\boldsymbol{\beta},\boldsymbol{\delta})^{(t-1)}|\boldsymbol{x})}$, then accept $(\boldsymbol{\beta},\boldsymbol{\delta})^*$ as $(\boldsymbol{\beta},\boldsymbol{\delta})^{(t)}$ as a draw from $F_{\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\delta}}|\boldsymbol{X}}(\boldsymbol{\beta},\boldsymbol{\delta}|\boldsymbol{x})$. Otherwise, set $(\boldsymbol{\beta},\boldsymbol{\delta})^{(t)}$ to $(\boldsymbol{\beta},\boldsymbol{\delta})^{(t-1)}$.

Step 6. If $t = N$, where N is the desired chain length, then stop. Otherwise, repeat Step 2.

We apply this process to the example presented at the beginning of this section. We wish to obtain a sample of size 200, apply a burn-in period of 100 steps, and thinning to every 100th step. This requires a chain length of $N = 20100$. For a typical MCMC scenario, a much larger chain length would be necessary, but due to the nature of this proposal distribution, we are able to use such a small N. Applying the above process to construct the proposal distribution, the corresponding BB_{OL_4} model is shown in Figure 3.2.

Remark 2. *A similar method would be useful for estimation involving different nonzero δ's, with*

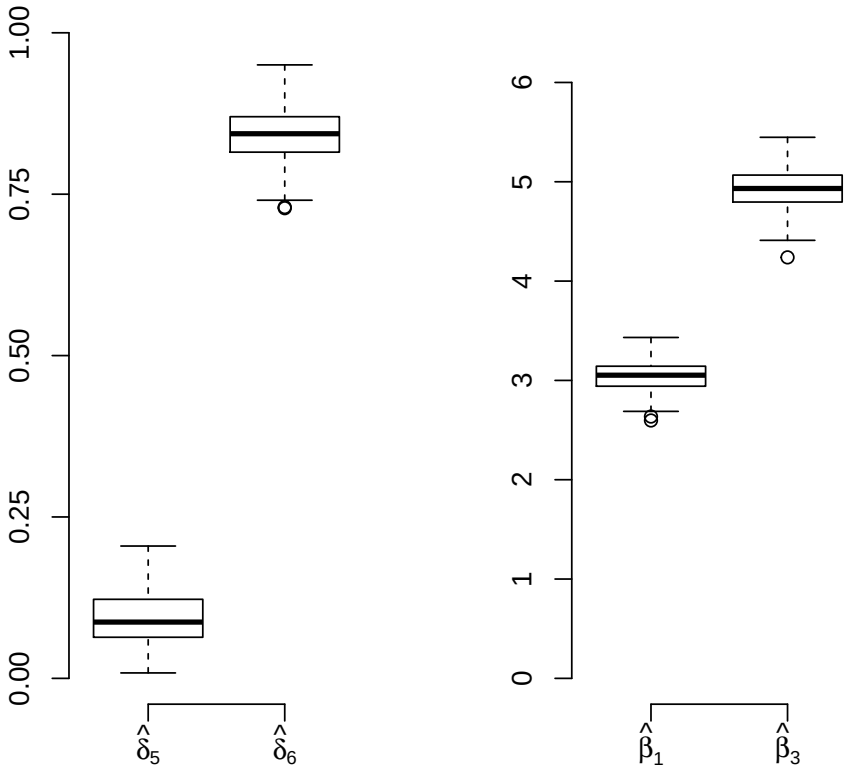


Figure 3.3: MCMC parameter estimation results for $\boldsymbol{X} \sim BAS\,L(\boldsymbol{\beta} = (3, 1, 5, 1), \boldsymbol{\delta} = (0.14, 0.71, 0, 0))$ and $K = 250$.

the possible exception of the proposal distribution for $\boldsymbol{\delta}$. For example, if δ_6 and δ_7 are nonzero, then the proposal distribution would need to have support in the lower triangle of the unit square. So, we may apply the family of distributions defined by considering $V_1 \sim B(\alpha_1, \gamma)$, and $V_2 \sim B(\alpha_2, \gamma)$, with $V_1 \perp V_2$. Then if we set $\tilde{\delta}_6 = V_{(1)}$, and $\tilde{\delta}_7 = 1 - V_{(2)}$, the support of $(\tilde{\delta}_6, \tilde{\delta}_7)$ is, in fact, the desired space, and the three-parameter family of distributions is sufficiently flexible for this application. Alternatively, to serve this purpose, we can truncate (V_1, V_2) to the lower triangle. Both of these distributions are difficult to characterize, but they are nevertheless tractable options for a proposal distribution, and they both provide easy simulation of proposal values.

The *BAS L* parameter estimates are computed as the means of the thinned MCMC sample, and variance is computed from the same. The results are shown in Figure 3.3.

3.5.2 Non-Linear Regression Example

Suppose a data set $(\boldsymbol{w}, \boldsymbol{y})$ is available; where $\boldsymbol{w}$ is a $K \times 2$ matrix, and $\boldsymbol{y}$ is a $K \times 2$ matrix, and the data set is one for which the following model is applicable:

$$\boldsymbol{y}_k = g(\boldsymbol{w}_{k\cdot}) + \boldsymbol{\epsilon}_{k\cdot} \tag{3.20}$$

where $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is some function, and $\boldsymbol{\epsilon}$ is a $K \times 2$ matrix of error parameters. Suppose, further that it is deemed that squared error is too stringent a restriction on the error term, maybe because positive errors in both response categories (the first and second columns of $\boldsymbol{y}$) will often be extreme, but negative errors cannot exhibit the same extreme behavior due to some known physical limitations. In addition, we suspect that the columns of $\boldsymbol{y}$ are correlated, and thus we should also suspect the same about $\boldsymbol{\epsilon}$. Then (a sub-model of) the 8-parameter bivariate asymmetric Laplace distribution may be an appropriate model for $\boldsymbol{\epsilon}$. In particular, consider volume data for Google, Inc. (GOOG) and Microsoft, Inc. (MSFT) stock from February, 23, 2017, to February, 23, 2018, a total of 253 trading days.[2] The data is shown in Figure 3.4. It can be seen that there is a slightly periodic underlying pattern for Microsoft, and, to a lesser degree, for Google. We apply non-linear regression to the data by numerically fitting a sinusoidal function to each time series separately,[3] that is, setting

$$g(w_1, w_2) = \begin{pmatrix} g_1(w_1) \\ \\ g_2(w_2) \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \sin(c_1 w_1 + d_1) \\ \\ a_2 + b_2 \sin(c_2 w_2 + d_2) \end{pmatrix}, \tag{3.21}$$

[2] Source: www.nasdaq.com

[3] It can be argued that the two sinusoidal functions are, themselves, correlated. However, for purposes of this analysis, we assume independence.

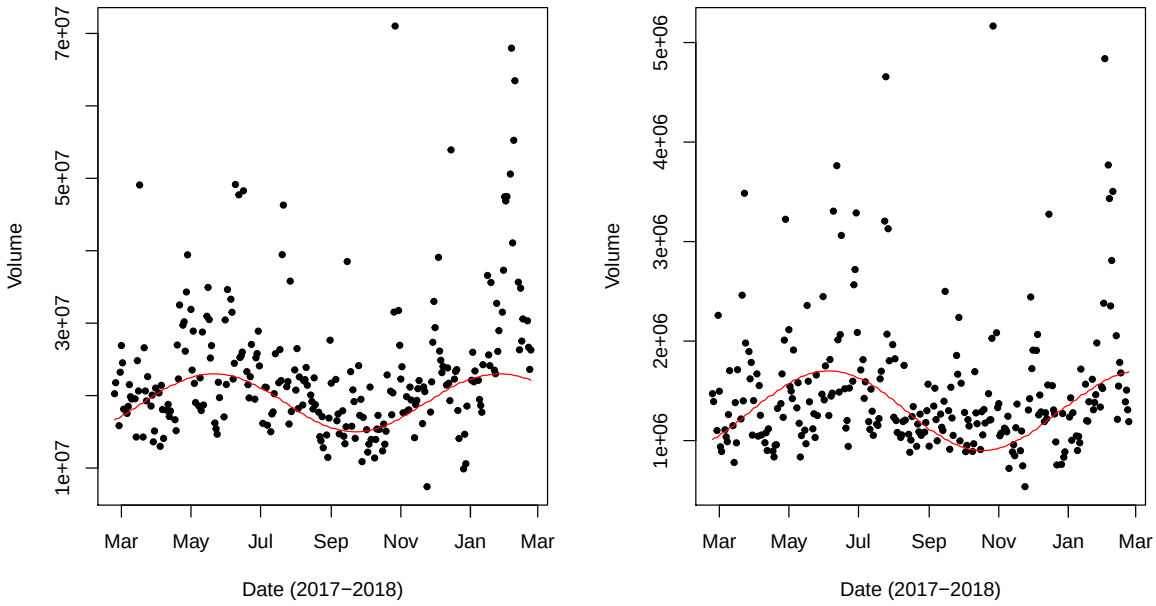


Figure 3.4: Microsoft Volume (left) and Google Volume (right) and corresponding periodic models (red).

where $(a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2)$ is a set of unknown parameters. We fit this regression model numerically where we minimize *absolute* error; for both stocks. The resulting fitted model is

$$\begin{pmatrix} \hat{y}_{k1} \\ \hat{y}_{k2} \end{pmatrix} = \hat{g} \begin{pmatrix} w_{k1} \\ w_{k2} \end{pmatrix} = \begin{pmatrix} (1.90 \times 10^7) + (4.00 \times 10^6)\sin(0.0363 w_{k1} + 0.855) \\ (1.30 \times 10^6) + (4.00 \times 10^5)\sin(0.0335 w_{k2} + 1.763) \end{pmatrix}. \tag{3.22}$$

Plots of these are overlaid on the data in Figure 3.4. From this, we obtain $\boldsymbol{\epsilon}$ for both stocks:

$$\boldsymbol{\epsilon}_{k\cdot} = \begin{pmatrix} y_{k1} - \hat{y}_{k1} \\ y_{k2} - \hat{y}_{k2} \end{pmatrix}. \tag{3.23}$$

Applying the above-outlined process, the proposal distribution for (δ_5, δ_6) is this time selected as the Olkin-Liu, Model 3, and shown in Figure 3.5. Using the MCMC process, the six unknown parameters are estimated; the estimates are $\beta_1 = 2931540, \beta_2 = 292311, \beta_3 = 7116668, \beta_4 = 507974.5, \delta_5 = 0.6167675$, and $\delta_6 = 0.5559745$. The results are shown in Figure 3.6, where box-plots are included. This *BAS L* model is overlaid on the $\boldsymbol{\epsilon}$'s in Figure 3.7. A bivariate normal model

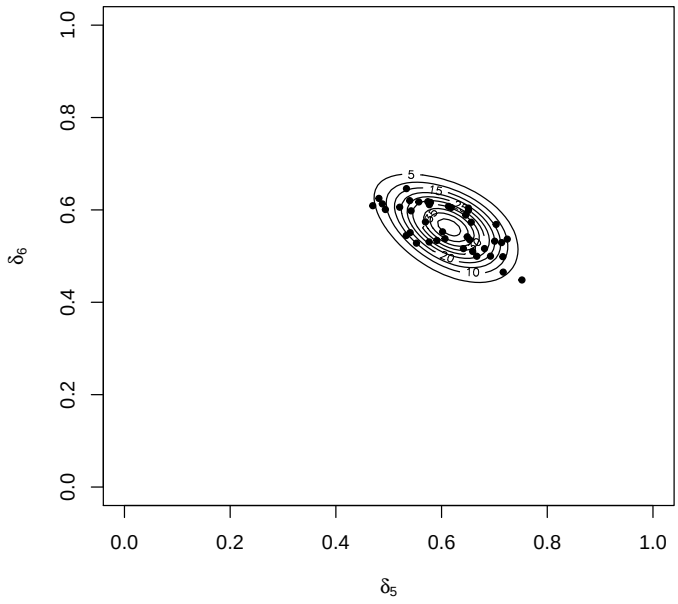


Figure 3.5: *BAS L* proposal distribution: $(\tilde{\delta}_5, \tilde{\delta}_6) \sim BB_{OL_3}(47.7, 23.6, 37.1)$.

with parameters equal to its MLEs is also shown for comparison.

For a means of comparing the resulting *BAS L* model to the stock data, that is, quantifying model fitness, we compute the mean and covariance matrix for the model, and compare these to the sample values of the same. These values are shown in Table 3.1. Also, according to the Akaike information criterion (AIC) criterion, the *BAS L* model (AIC=15925.09) provides a slightly better fit than the Gaussian model (AIC=16198.17). However, the Gaussian model is clearly a model misspecification for two reasons. First, there are two distinctly different tail dependencies in this data, but, as a symmetric model (about a set of rotated axes), the Gaussian model cannot capture this phenomenon. Second, the Gaussian's tails are not heavy enough for this data; Laplace distributions are far better equipped to deal with heavy-tailed data.

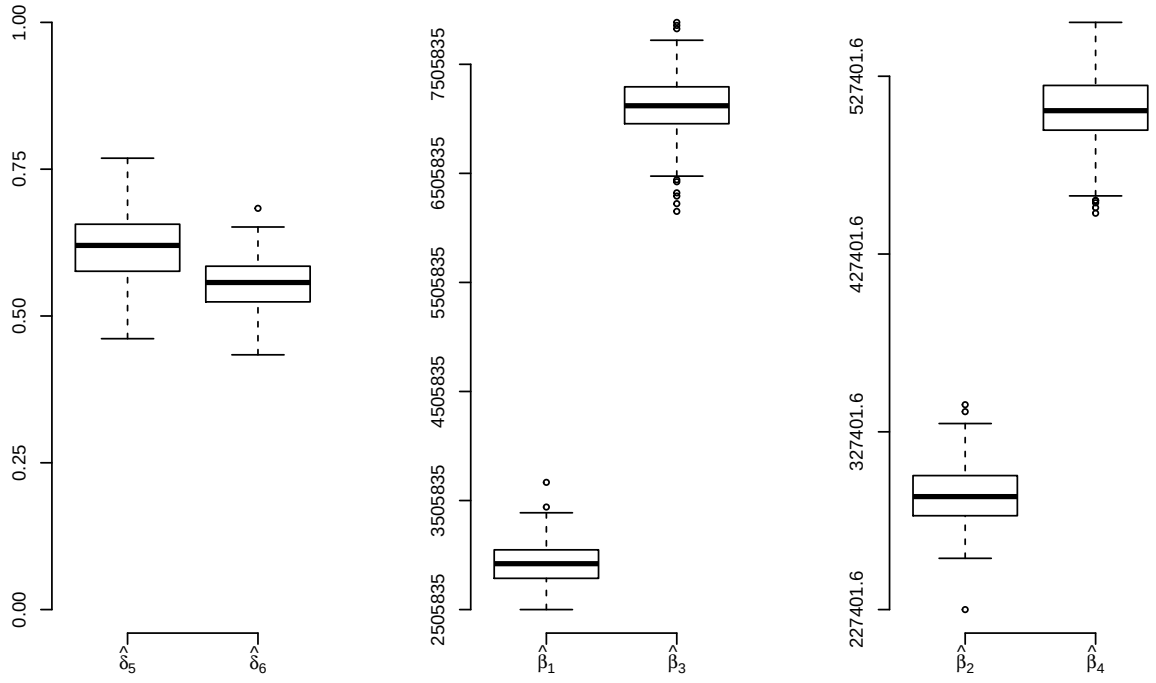


Figure 3.6: MCMC *BAS L* parameter estimates for Stock Volumes, and corresponding box plots.

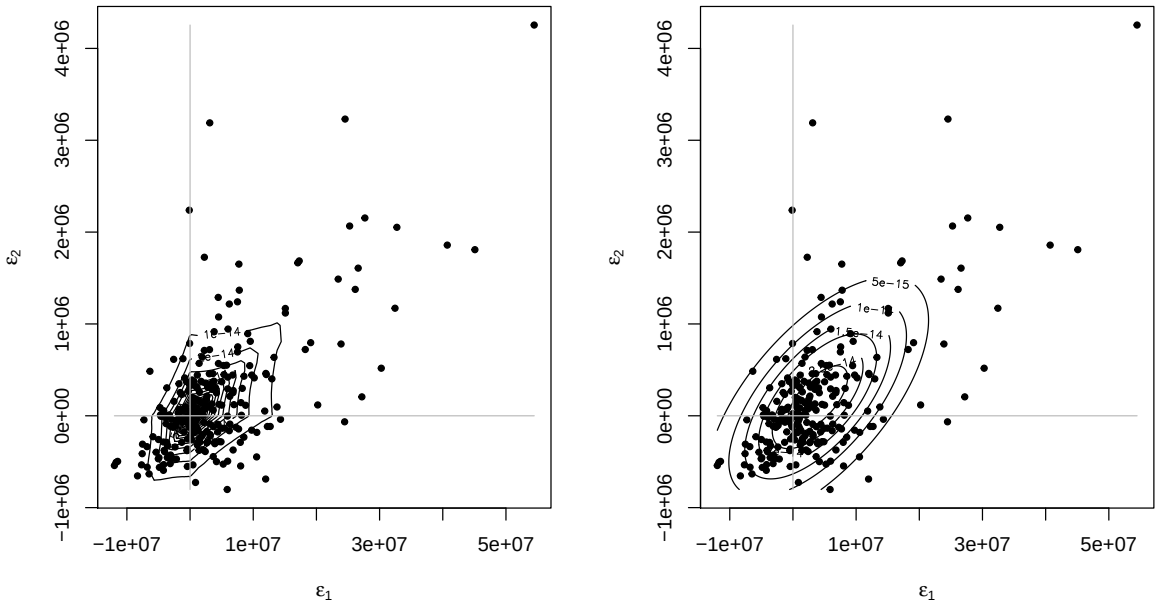


Figure 3.7: Residuals with fitted *BAS L* model (left) and Gaussian model (right).

Source	$\hat{E}(\boldsymbol{\epsilon})$	$\widehat{Var}(\boldsymbol{\epsilon})$
Model	$\begin{pmatrix} 4185128 \\ 215663.5 \end{pmatrix}$	$\begin{pmatrix} 1.184818 \times 10^{14} & 2.538417 \times 10^{12} \\ 2.538417 \times 10^{12} & 6.869677 \times 10^{11} \end{pmatrix}$
Data	$\begin{pmatrix} 3909728 \\ 184893.8 \end{pmatrix}$	$\begin{pmatrix} 7.970902 \times 10^{13} & 3.887233 \times 10^{12} \\ 3.887233 \times 10^{12} & 4.557925 \times 10^{11} \end{pmatrix}$

Table 3.1: Comparison of moments between *BAS L* model and data.

3.6 Conclusions

In this chapter we introduced a new 8-parameter bivariate Asymmetric Laplace distribution with a flexible dependence structure. We also showed that its parameters have intuitive interpretations, and we exhibited a closed-form method of moments solution for its 4-parameter sub-family with standard Laplace marginals. We finished by demonstrating parameter estimation of this model based on residuals from a regression model. We showed that, in this case where absolute error was minimized, the *BAS L* model provided a better fit to the residuals than a Gaussian model.

In general, the family of 8-parameter *BAS L* distributions constitutes a flexible collection of models appropriate for heavy-two-tailed bivariate data for which symmetric models are inadequate. Future work in this area would include a more thorough comparison the the *BAS L* and Kotz models, particularly identifying any overlap and properties unique to either. In addition, further exploration of specific sub-families may reveal additional useful characterizations of this family, and possibly reveal as yet unknown interrelationships between this family and other well-known families of distributions, whether asymmetric Laplace or other, entirely different, distributions.

Another avenue for further research is through Sklar's Theorem [40], that is, by separating each particular *BAS L* distribution into its marginals and its underlying copula. The form of

this copula may lead to interesting discoveries regarding the family of copulas associated with all members of the $BASL$ class. With this said, it can be asserted with a high level of certainty that, while the dependence parameters behave similarly to those of the Bivariate Beta Copula, the copulas that correspond to $BASL$ models are not the same as the Bivariate Beta Copulas, certainly not in general, though some conceivable non-empty intersection between the two classes of copulas may exist. This, however, does not guarantee that the $BASL$ copulas do not include some known copulas. In general, they can be expected to constitute new flexible family of copulas for general use.

Lastly, higher-dimensional versions of the $BASL$ family rapidly become increasingly complex, as the dimension increases chiefly because of the rapid increase in the number of parameters in the model (in two dimensions there are 8 parameters, while in 3 and 4 dimensions there are 28 and 84, respectively). However, if a suitable method of model selection among nested submodels can be agreed upon, higher-dimensional models may be tractable, and thus useful for more general applications.

Chapter 4

Compound Random Variables

4.1 Introduction

Compound random variables, that is to say, random sums of random variables, have been studied in many forms. For example much work has been done to develop theory involving compound geometric and compound Poisson random variables, particularly for insurance and risk assessment applications. However, this is just a small subset of the vast assortment of such random variables. Many other forms constitute natural extensions that remain under the same umbrella. These include forms where the counting variable has some lesser-known distribution, the summand variables are not necessarily independent of the counting variable, multivariate forms, and others. Further, the summand random variable need not have only positive support, something generally assumed for compound random variables studied to date.

In this chapter, we begin with a thorough study of compound geometric random variables and review the vast literature on the subject. We proceed with some examples of parameter estimation for a specific case of compound geometric random variables. We then extend the study to two

distinct forms of bivariate compound random variables, and present some examples. Lastly, we will introduce a family of bivariate compound geometric distributions with exponential marginals and exhibit a method of likelihood-free parameter estimation.

4.2 Compound Geometric Random Variables

Consider the random variable, Y, defined by

$$Y = \sum_{j=1}^{M} X_j \tag{4.1}$$

where the X_j's are iid random variables and $M \sim \text{Geo}(1-p)$, $p \in (0,1)$, i.e. $P(M = k) = (1-p)p^{k-1}$, $k \in \mathbb{N}$, and where M is independent of the X_j's. Then Y is of the *compound geometric* type. In addition, throughout this chapter, it will be assumed that $X \stackrel{d}{=} X_j$ for all j, whenever X_j's are mentioned in the above context. Likewise, $Y \stackrel{d}{=} Y_j$ for all j, whenever Y_j's are mentioned in the above context. From Equation 4.1, by conditioning on M, we may obtain the following relationship between the characteristic functions, ϕ_X and ϕ_Y, of X and Y, respectively:

$$\phi_Y(t) = \frac{(1-p)\phi_X(t)}{1 - p\phi_X(t)}, \tag{4.2}$$

or equivalently,

$$\phi_X(t) = \frac{\phi_Y(t)}{1 - p + p\phi_Y(t)}. \tag{4.3}$$

These random variables are useful for several applications, including, but not limited to, insurance risk vs. ruin, p-thinning of point processes and aging. They also naturally arise in the study of certain types of branching processes.

4.2.1 Definitions

Define $\mathcal{G}(p)$ to be the collection of all random variables satisfying Equation (4.1) for some random variable, X and the given value of p. Also define

$$\overline{\mathcal{G}} := \bigcup_{0<p<1} \mathcal{G}(p) \qquad \text{and} \qquad \underline{\mathcal{G}} := \bigcap_{0<p<1} \mathcal{G}(p)$$

A random variable in $\underline{\mathcal{G}}$ is also sometimes termed *geometrically infinitely divisible (g.i.d.)*.

A random variable, V, is said to be *infinitely divisible* if for each $K \in \mathbb{N}$, there exists a set of K iid random variables $\left\{X_k{}^{(K)}\right\}_{k=1}^{K}$ such that

$$V \stackrel{d}{=} \sum_{k=1}^{K} X_k{}^{(K)}.$$

The characteristic function of any compound geometric random variable, Y, may be written in the form

$$\phi_Y(t) = \left(1 - \log \phi_V(t)\right)^{-1}$$

where V is some infinitely divisible random variable with characteristic function, $\phi_V(t)$ [22].

Consider the stationary process, $\{Y_j\}_{j=0}^{\infty}$, satisfying

$$Y_j = \begin{cases} Y_{j-1} + X_j \text{ w.p. } p \\ X_j \text{ w.p. } 1 - p \end{cases} \tag{4.4}$$

for $j \in \mathbb{N}$, where the sequence $\{X_j\}_{j=1}^{\infty}$ is iid. Then $\phi_Y(t)$ and $\phi_X(t)$, once again, satisfy Equations 4.2 and 4.3. In this way, the process defined by Equation 4.4 is related to a compound geometric random variable. Specifically, for some iid sequence $\{X_j\}$, each Y_j is compound geometric with parameter $(1 - p)$, i.e. $Y_j \in \mathcal{G}(p)$.

4.2.2 Exponential and Related Distributions

We wish to determine which characteristic functions, $\phi_Y(t)$, would, according to Equation 4.3, induce a valid characteristic function, $\phi_X(t)$.

If $X \sim \exp(\beta)$,

$$\begin{aligned}\phi_Y(t) &= \frac{(1-p)(1-i\beta t)^{-1}}{1-p(1-i\beta t)^{-1}} \\ &= \left[1 - i\left(\frac{\beta}{1-p}\right)t\right]^{-1}\end{aligned} \tag{4.5}$$

Since this is a valid characteristic function for all p, in fact, also exponential, the exponential distribution is in $\underline{\mathcal{G}}$.

Now, consider the special case where $\phi_Y(t) = \phi_X(ct)$. We have just seen that X is exponential if and only if Y is exponential, and thus setting $c = 1 - p$, the exponential is one such random variable. Is there any other random variable, Y, which satisfies $\phi_Y(ct) = \frac{(1-p)\phi_Y(t)}{1-p\phi_Y(t)}$? Well, a degenerate at 0 clearly satisfies it. Otherwise, define

$$\psi(t) = \frac{\phi(t) - 1}{\phi(t)}$$

so that

$$\psi(ct) = \frac{1}{1-p}\psi(t)$$

If ψ is representable as a power series, we may write

$$\psi(t) = \sum_{j=0}^{\infty} a_j t^j$$

and, by the uniqueness of power series representations, solve:

$$\sum_{j=0}^{\infty} a_j c^j t^j = \frac{1}{1-p}\sum_{j=0}^{\infty} a_j t^j$$

Clearly $a_0 = 0$. However, if more than one a_k is nonzero, there is no solution for c. Only one, in fact, exactly one, $a_k, k > 0$, must be nonzero, but any such k works. In general, we may pick $k \in \mathbb{N}$, and set $a_k = \theta \in \mathbb{C}$.[1] It would then follow that

$$\psi(t) = \theta t^k$$

We may then conclude that

$$\phi(t) = \frac{1}{1 - \theta t^k} \tag{4.6}$$

satisfies $\phi(ct) = \frac{(1-p)\phi(t)}{1-p\phi(t)}$, when $c = \sqrt[k]{\frac{1}{1-p}}$. However, note that the function $f(t) = 1 - \theta t^k$ forms either a semi-infinite line (when k is even) or an infinite line (when k is odd) on the complex plane which passes through 1. A necessary condition for ϕ to be a valid cf is that $f(t)$ must be outside the (open) unit disk for all t. Therefore, for odd k, it must be that the line is vertical (i.e. that θ is purely imaginary). For even values of k, it must be that the real part of θ is negative. Thus, the solution is given by Equation 4.6, where θ is in some subset of $\{z : \Re\{z\} \leq 0, k = 2m; \Re\{z\} = 0, k = 2m+1, m \in \mathbb{N}\}$. It should be noted that this is a necessary condition, but it holds under only the assumption that $\phi(t)$ is analytic on a neighborhood of its trace in $\mathbb{C}$. So, 1) further sufficiency conditions (due to the positive definiteness requirement) may be present, and 2) since characteristic functions generally do not need to be analytic, there may yet be more solutions. The case where $k = 1$ and $\theta = \alpha i$, $\alpha \in \mathbb{R}^+$ is the characteristic function for an exponential. Also, if $k = 2$ and $\theta < 0$ is real, then this represents a Laplace distribution centered at 0. Based on these results, it appears all solutions are related to the exponential distribution.

[1] We shall see in a moment that there are restrictions on what θ may be.

4.2.3 Characterizations

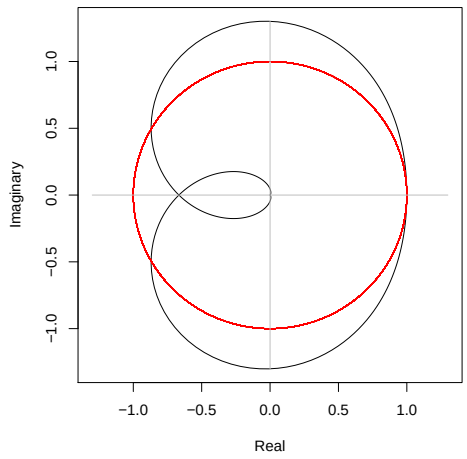
We now consider the more general question of finding random variables which are in $\mathcal{G}(p)$ for various values of p. As an example, consider $\phi_Y(t) = (1 - it\beta)^{-\alpha}$, i.e. $Y \sim \Gamma(\alpha, \beta)$. Plotting the trace of the corresponding $\phi_X(t)$ function for a few choices of p, α and β (Figure 4.1) indicates that sometimes it may be a valid characteristic function, and sometimes it is clearly not.

First, notice that for any $Y \in \mathcal{G}(p)$, $\phi_Y(t)$ is given by Equation 4.2. Therefore, we may immediately conclude that $Y \in \mathcal{G}(p)$ if and only if the left hand side of Equation 4.2 is a valid characteristic function. However, this is generally difficult to assess, for it involves analysis of complex-valued functions. If Y has non-negative support, we may avoid this by using Laplace transforms rather than characteristic functions. So, we begin with this special case.

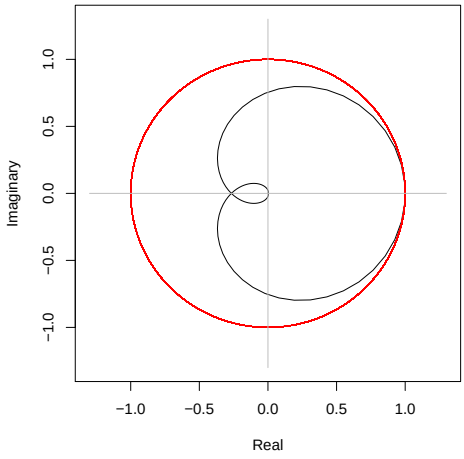
Suppose that Y is a non-negative random variable which can be written in the form given in Equation 4.1. Then the Laplace Transform of the X_j's can be written as

$$\mathcal{L}_X(t) = \frac{\mathcal{L}_Y(t)}{1 - p + p\mathcal{L}_Y(t)} \tag{4.7}$$

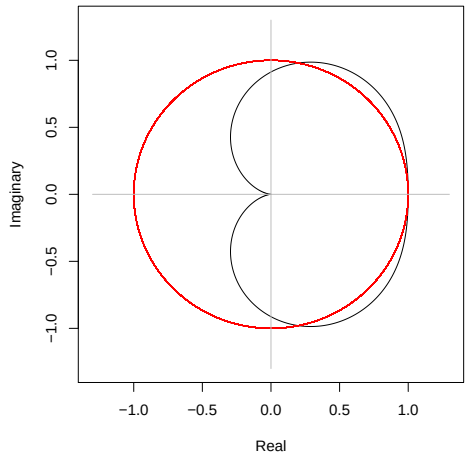
Now, note that the distribution of a random variable, Y, is completely monotone if and only if it satisfies $(-1)^n F_Y^{(n)}(y) \leq 0$ for all $n \in \mathbb{N}$. So, if Equation 4.7 does not yield a legitimate Laplace Transform, it must be that Y is not g.i.d. Hence, we must check that $\mathcal{L}_X(0) = 1$ and that $\mathcal{L}_Y(t)$ has complete monotonicity. Consider again the case where $Y \sim \Gamma(\alpha, \beta)$, where $\alpha \neq 1$. Then the Laplace



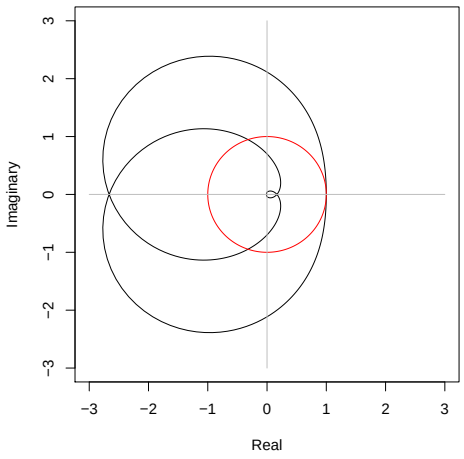
(a) $\alpha = 4; \beta = 2;$ and $p = 0.5$



(b) $\alpha = 4; \beta = 6;$ and $p = 0.05$



(c) $\alpha = 2; \beta = 0.75;$ and $p = 0.7$



(d) $\alpha = 9; \beta = 5;$ and $p = 0.5$

Figure 4.1: $\phi_X(t)$ for various choices of p, β, and α. Included in the plot is the unit circle (in red) to indicate that it is not a valid characteristic function based on the one requirement $|\phi_Y(t)| \leq 1$. In this case plots (a), (c), and (d) violate this requirement. Plot (b) *may* be a valid characteristic function.

Transform of X is given by

$$\begin{aligned}\mathcal{L}_X(t) &= \frac{\mathcal{L}_Y(t)}{1-p+p\mathcal{L}_Y(t)} \\ &= \frac{(1+\beta t)^{-\alpha}}{1-p+p(1+\beta t)^{-\alpha}} \\ &= ((1-p)(1+\beta t)^{\alpha}+p)^{-1}\end{aligned}$$

Applying the first three derivatives to this expression, and setting to zero, we have

$$\begin{aligned}\left.\frac{\partial \mathcal{L}_X(t)}{\partial t}\right|_{t=0} &= -(1-p)\alpha\beta \\ \left.\frac{\partial^2 \mathcal{L}_X(t)}{\partial t^2}\right|_{t=0} &= (1-p)\alpha(p-p\alpha+(1-p)(1+\alpha))\beta^2 \\ \left.\frac{\partial^3 \mathcal{L}_X(t)}{\partial t^3}\right|_{t=0} &= \Big[(p-1)\alpha\beta^3(1+t\beta)^{\alpha-3}(p^2(\alpha-2)(\alpha-1) \\ &\quad +4(p-1)p(\alpha^2-1)(1+t\beta)^{\alpha}+(p-1)^2(1+\alpha)(2+\alpha)(1+t\beta)^{2\alpha})\Big] \\ &\quad / \left[p-(p-1)(1+t\beta)^{\alpha}\right]^4\end{aligned}$$

From just these three derivatives, a necessary condition for X to be completely monotone is that $p < \frac{1+\alpha}{2\alpha}$. This is not a sufficient condition as can be seen from the traces of the characteristic functions in Figure 4.1 (see plots (a) and (d)).

We may also explore the moments of random variables in order to assess whether they are compound geometric. If we look at the first three moments of X, we have

$$\begin{aligned}\mu_X^{(1)} &= (1-p)\mu_Y \\ \mu_X^{(2)} &= (1-p)\left(\mu_Y{}^{(2)}-2p\mu_Y^2\right) \\ \mu_X^{(3)} &= 6\mu_Y^2(1-p)p\Big(2p\mu_Y{}^2+p\mu_Y{}^{(2)}-4\mu_Y \\ &\quad -\mu_Y{}^{(2)}-2p\mu_Y-2p^2\mu_Y^2\Big)\end{aligned}$$

Consider the Poisson distribution. Is this g.i.d.? Assume that $Y \sim Poi(\lambda)$ can be written in the form of Equation 4.1. Then the moments of X would be

$$\mu_X^{(1)} = (1-p)\lambda$$

$$\mu_X^{(2)} = (1-p)\lambda((1-2p)\lambda + 1)$$

$$\mu_X^{(3)} = 6p(1-p)\lambda^3(p(3-2p)\lambda - 5 - p - \lambda)$$

The second moment gives a necessary condition for $Y \in \overline{\mathcal{G}}$: $\lambda < \frac{1}{2p-1}$ if $p > \frac{1}{2}$. Also, the third moment gives a necessary condition for $Y \in \overline{\mathcal{G}}$ to be $\lambda < \frac{5+p}{3p-2p^2-1}$ if $p < \frac{1}{2}$, and $\lambda > \frac{5+p}{3p-2p^2-1}$ if $p > \frac{1}{2}$. No valid λ can satisfy these conditions simultaneously, so $Y \notin \overline{\mathcal{G}}$ for any λ.

4.2.4 Additional Properties

Consider the random variable, Y given by the following mass function:

$$P(Y = q_k) = \frac{6}{(\pi k)^2}, \text{ for } k \in \mathbb{N}$$

where $\{q_k\}_{k=1}^{\infty}$ is the sequence of positive prime numbers. Let $p \in (0,1)$. If $Y \in \mathcal{G}(p)$, notice that

$$0 < \frac{6}{4\pi^2} = P(Y = 3)$$
$$0 < \frac{6}{9\pi^2} = P(Y = 5)$$

This implies that there exist k_1 and k_2 such that

$$P(X_1 + X_2 + ... + X_{k_1} = 3, M = k_1) > 0, \text{ and}$$

$$P(X_1 + X_2 + ... + X_{k_2} = 5, M = k_2) > 0$$

so that

$$P(Y = 8) \geq P(X_1 + X_2 + ... + X_{k_1} + X_{k_1+1} + X_{k_1+2} + ... + X_{k_1+k_2} = 8, M = k_1 + k_2) > 0,$$

a contradiction. In fact, as is shown in the following proof, any random variable whose support is not closed under addition cannot be in $\mathcal{G}(p)$.

Proposition 3. *Let Y be any random variable with support not closed under addition. Then $Y \notin \overline{\mathcal{G}}$.*

Proof. Since supp(Y) is not closed under addition, there exists some set $I \subset \mathbb{R}$, disjoint from supp(X), such that $P(Y \in I) = 0$ and for some collection of sets $I_1, I_2, ..., I_m$ satisfying $P(Y \in I_j) > 0$ for all $j = 1, 2, ..., m$, $\sum_{i=1}^{m} y_i \in I$ for all $y_1 \in I_1, y_2 \in I_2, ..., y_m \in I_m$. Now, for any iid sequence of random variables $X_1, X_2, ...$ and $M \sim Geo(1-p)$, if there exist $k_1, k_2, ..., k_m$ such that

$$P(X_1 + X_2 + ...+X_{k_1} \in I_1, M = k_1) > 0,$$

$$P(X_1 + X_2 + ...+X_{k_2} \in I_2, M = k_2) > 0,$$

$$\vdots$$

$$P(X_1 + X_2 + ...+X_{k_m} \in I_m, M = k_m) > 0$$

then it immediately follows that

$$P(X_1 + X_2 + ... + X_{k_1} + X_{k_1+1} + X_{k_1+2} + ... + X_{k_1+k_2} + ... + X_{k_1+k_2+...k_m} \in I, M = k_1 + k_2 + ... + k_m) > 0$$

so that it cannot be that $Y \in \mathcal{G}(p)$. □

This is clearly not a sufficient condition since we have already seen many cases with support closed under addition; various gamma distributions (see Figure 4.1), for example; that have the property.

Proposition 4. *No bounded random variable other than $Y \equiv 0$ is in $\overline{\mathcal{G}}$.*

Proof. Suppose $Y \in \mathcal{G}(p)$ with $Y \not\equiv 0$. Then clearly $X_1 \not\equiv 0$. So there exists $y_0 > 0$ such that

$P(|X_1| > y_0) > 0$. Let $u_0 > 0$ be arbitrary. Then there is some $k \in \mathbb{N}$ such that $ky_0 > u_0$. We have

$$\begin{aligned}
P(|Y| > u_0) &= \sum_{j=1}^{\infty} P(X_1 + X_2 + ... + X_j > u_0, M = j) \\
&> \sum_{j=1}^{\infty} P(X_1 + X_2 + ... + X_j > ky_0, M = j) \\
&> P(X_1 + X_2 + ... + X_k > ky_0, M = k) \\
&> P(X_1 > y_0, X_2 > y_0, ..., X_k > y_0, M = k) \\
&= P(X_1 > y_0)P(X_2 > y_0) \cdots P(X_k > y_0)P(M = k) \\
&= \left[P(X_1 > y_0)\right]^k p(1-p)^{k-1} \\
&> 0
\end{aligned}$$

Since u_0 was arbitrary, Y is not bounded. □

Proposition 5. *If Y is of the compound geometric type satisfying Equation 4.1, and, in addition, X is non-degenerate and positive with a finite first moment, then for the random variable, $V = \frac{(1-p)Y}{\mu}$, where $\mu = E(X)$,*

$$\lim_{p \to 1^-} V \sim \exp(1), \tag{4.8}$$

that is, $\frac{Y}{E(Y)} \xrightarrow{d} W$, as $p \to 1^-$, where $W \sim \exp(1)$.

Proof. From Equation 4.2,

$$\phi_V(t) = \frac{(1-p)\phi_X\left(\frac{(1-p)t}{\mu}\right)}{1 - p\phi_X\left(\frac{(1-p)t}{\mu}\right)}$$

Thus, by L'Hopital's Rule, and the fact that $\phi'_X(0) = i\mu$,

$$\begin{aligned}\lim_{p\to 1^-} \phi_V(t) &= \lim_{p\to 1^-} \frac{(1-p)\phi_X\left(\frac{(1-p)t}{\mu}\right)}{1-p\phi_X\left(\frac{(1-p)t}{\mu}\right)} \\ &= \lim_{p\to 1^-} \frac{\phi_X\left(\frac{(1-p)t}{\mu}\right) + \frac{(1-p)t}{\mu}\phi'_X\left(\frac{(1-p)t}{\mu}\right)}{\phi_X\left(\frac{(1-p)t}{\mu}\right) - \frac{pt}{\mu}\phi'_X\left(\frac{(1-p)t}{\mu}\right)} \\ &= (1-it)^{-1}\end{aligned}$$

This concludes the proof. □

Compound Geometric Convolutions

Compound Geometric convolutions are of considerable interest. A positive continuous random variable of the form Equation 4.1 which also satisfies

$$\overline{F}_Y(x) = p\int_0^y \overline{F}_Y(y-t)d\left(F_X(t)\right) + p\overline{F}_X(y)$$

is known as a Compound Geometric Convolution. Close examination of the tail behavior of such random variables applies to insurance risk models, queuing theory, and reliability theory. See [11], [35], and [49] for the most recent developments in this area.

Relationship with Exponential Random Variables

Exponential random variables bear a resemblance to elements of $\mathcal{G}(p)$ when p is close to 1, as is suggested by Proposition 5. In fact, several authors have applied the exponential distribution as a bound for various forms of the compound geometric random variables; [9], [15], [16]. Several methods, including Cramer's and Stein's Methods, are applied to approximate the compound geometric distribution for various applications.

Identifiability

Lin and Stoyanov [29] showed that, under some regularity conditions, moments can be used to characterize compound geometric distributions.

Multivariate Geometric and Compound Geometric Distributions

A natural extension of the compound geometric random variables is a multivariate version, based on a multivariate geometric distribution. Arnold [2] constructed a version based on a process with multiple outcomes, and showed that the elements are independent if and only if the summand variables are (positive or negative) exponential.

4.2.5 General Compound Random Variables

Klebenov and Rachev [23], who construct a foundational theory of random sums of iid random variables, point out that the probability generating functions for the compounding variable, M, form a semigroup (in the algebraic sense). They use this fact to establish a method of approximating arbitrary geometric sums of random variables using infinitely-divisible random variables. In addition, they establish some means of measuring the goodness of approximations of M-infinite divisible random variables by applying a point-wise metric to the modulus of the characteristic functions. They also extend this theory to the multivariate case. Satheesh [39] further develops this theory by pointing out that the semigroup property does not allow for common applications, such as the case where $P(Y = 0) > 0$, and constructing a more flexible model.

Wang [46] describes multivariate compound random variables in their most general form, and also derives a compound Poisson distribution as the asymptotic limit of sums of independent

multivariate random variables.

In Section 4.3, we investigate bivariate compound random variables and, in Section 4.5, develop a method for statistical inference for data from one of these models.

4.2.6 Statistical Inference

The areas with surprisingly little development in the literature for compound random variables are parameter estimation and model adequacy checking. Patel and Patel [33] did discuss maximum likelihood estimation in such a setting. Their techniques pertain, however, strictly to the Geometric Competing Risks Failure Model, that is, one in which two causes of failure, which are independent and geometrically distributed, are applied to a series of independent test subjects. Estimation of the geometric parameters in light of censored data then proceeds. Beyond this, there is no apparent reference to statistical inference and assessing the quality thereof. A possible reason for this might be that the realized value of M is often known, in which case the problem of statistical inference for Y would become uninteresting. However, in many cases, it is quite conceivable that the only data available is, for example, that which is reported as public information, e.g. total claims expenditures, but without the number of claims or individual amounts of claims. So, the premise that only $Y_1, Y_2, ...$ are observed is reasonable in many cases.

In the interest of constructing tools which can reliably build accurate compound geometric models, several methods of parameter estimation are discussed. Suppose $Y_1, Y_2, ..., Y_n$ are iid compound geometric random variables, i.e.

$$Y_k = \sum_{i=1}^{M_k} X_i^{(k)}$$

where the $X_i^{(k)}$ are absolutely continuous and iid for all k and all i; and $M_1, M_2, ..., M_K \overset{iid}{\sim} \text{Geo}(1-p)$,

$p \in (0,1)$. Further, consider the special case in which the density of $X_{(j)}^{(k)} \coloneqq \sum_{i=1}^{j} X_i^{(k)}$ is $f_j(x;\boldsymbol{\theta})$ and is known up to an unknown parameter $\boldsymbol{\theta}$, for all k. For more general cases, that is to say the vast majority of cases, convolution of j iid X_is is not easily obtained for all j, and therefore $L(p,\boldsymbol{\theta})$ is not attainable analytically. However, numerical options may apply in many of these cases.

Maximum Likelihood

Recall that if $X \sim \exp(\beta)$, then $Y \sim \exp\left(\frac{\beta}{1-p}\right)$. So, from the above sample, no information can be discerned about p and β as separate parameters. However, as will be seen in Section 4.5, this is not necessarily true in multivariate cases.

So, suppose X is not exponential. Then the likelihood of $(p,\boldsymbol{\theta})$ is given by

$$\begin{aligned} L(p,\boldsymbol{\theta}) &= \prod_{k=1}^{n}\left(\sum_{j=1}^{\infty} f_j(y_k;\boldsymbol{\theta})(1-p)p^{j-1}\right) \\ &= (1-p)^n \prod_{k=1}^{n}\left(\sum_{j=1}^{\infty} f_j(y_k;\boldsymbol{\theta})p^{j-1}\right) \end{aligned}$$

So the log likelihood becomes

$$\ell(p,\boldsymbol{\theta}) = n\log(1-p) + \sum_{k=1}^{n}\log\left[\sum_{j=1}^{\infty} f_j(y_k;\boldsymbol{\theta})p^{j-1}\right] \tag{4.9}$$

Suppose $X \sim \Gamma(2,\beta)$. Then Equation 4.9 becomes

$$\begin{aligned} \ell(p,\boldsymbol{\theta}) &= n\log(1-p) + \sum_{k=1}^{n}\log\left[\sum_{j=1}^{\infty}\left(\frac{y_k^{2j-1}\exp\{\frac{-y_k}{\beta}\}}{\Gamma(2j)\beta^{2j}}\right)p^{j-1}\right] \\ &= n\log(1-p) + \sum_{k=1}^{n}\log\left[\left(\frac{1}{\exp\{\frac{y_k}{\beta}\}\sqrt{p}\beta}\right)\sum_{j=0}^{\infty}\frac{\left(\frac{y_k\sqrt{p}}{\beta}\right)^{2j+1}}{(2j+1)!}\right] \\ &= n\log(1-p) + \sum_{k=1}^{n}\log\left[\frac{\sinh\left(\frac{y_k\sqrt{p}}{\beta}\right)}{\sqrt{p}\exp\{\frac{y_k}{\beta}\}\beta}\right] \end{aligned}$$

To maximize this, we may attempt to use the Newton-Raphson method. However, this particular distribution for X causes some difficulty. The likelihood, if plotted as a surface in $\mathbb{R}^3$, forms a "ridge," along which the log likelihood is almost perfectly flat. In this case, Newton-Raphson becomes confused, and there is little likelihood that it will correctly identify the location of the true maximum without a very large n. The reason for this seems to be that α is too close to 1 (i.e. that the distribution is too close to being exponential), causing the negative correlation between p and β to be too strong to separate their estimates. For illustration, data was generated with $\beta = 6.1$, $p = 0.86$, and $n = 100$. A small portion of the log likelihood is plotted in Figure 4.2. It should be noted that

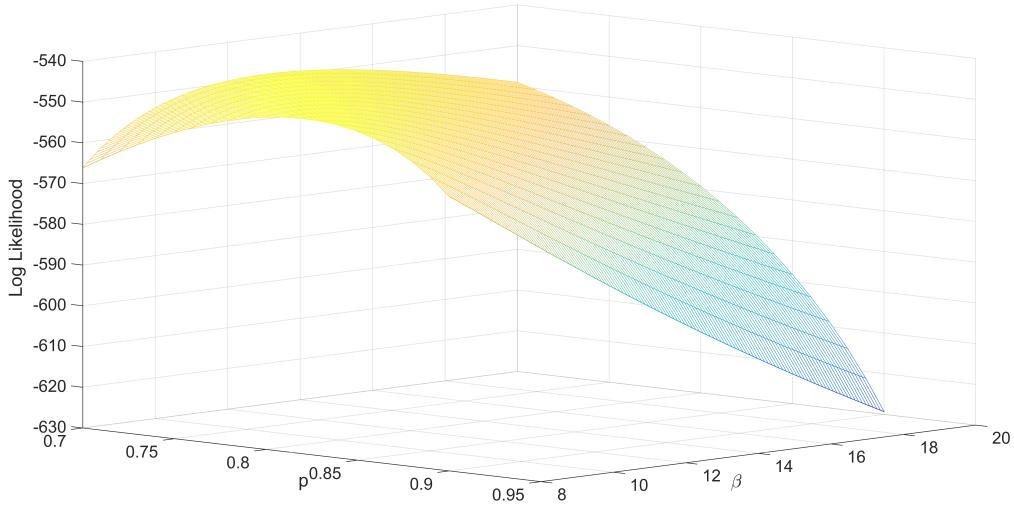


Figure 4.2: Log Likelihood Y such that $X \sim \Gamma(2, 6.1)$ and $M \sim \text{Geo}(0.14)$.

this issue may frequently arise with this type of random variable (the compound geometric). It will often occur when the distribution of X involves an unknown location parameter, θ. In this case, as either p or θ increases, Y will also tend to increase, and strong negative correlation between the estimators for these parameters manifests itself, potentially leading to an identifiability problem for

estimation. Though β, for the gamma example, is not a location parameter, the mean nevertheless is positively correlated with it. However, there are alternatives to maximum likelihood that may reduce the impacts of this problem, as will be exhibited next.

Bayesian Inference

We now apply Bayesian estimation to obtain estimates of p and β in the $X \sim \Gamma(2,\beta)$ example. Note that having an informative prior (or even one with a dependence structure between p and β) is important, for it will limit the natural tendency of the model to estimate the two parameters as one, as was (at least partially) the case with maximum likelihood. This is not an unreasonable assumption to make, for if it is not applied, little knowledge about p and β, as distinct parameters, is even contained in a typical sample. So, we strive to choose *informative* priors which form a reasonable model of reality. In this case, we consider the joint prior for p and β, where p and β are independent:

$$\pi(p,\beta) = \frac{\beta^{u_0-1}e^{-\frac{\beta}{v_0}}p^{a_0-1}(1-p)^{b_0-1}}{B(a_0,b_0)\Gamma(u_0)v_0^{u_0}}, \tag{4.10}$$

so that the marginal priors are $p \sim B(a_0,b_0)$, and $\beta \sim \Gamma(u_0,v_0)$. The posterior density becomes

$$\begin{aligned} h(p,\beta;\mathbf{y}) &= C\pi(p,\beta)\prod_{k=1}^{n}\left(\sum_{j=1}^{\infty} f_j(y_k;\beta)(1-p)p^{j-1}\right) \\ &= C\frac{\beta^{u_0-1}e^{-\frac{\beta}{v_0}}p^{a_0-1}(1-p)^{b_0-1}}{B(a_0,b_0)\Gamma(u_0)v_0^{u_0}}\prod_{k=1}^{n}\left(\sum_{j=1}^{\infty}\left(\frac{y_k^{2j-1}\exp\{\frac{-y_k}{\beta}\}}{\Gamma(2j)\beta^{2j}}\right)(1-p)p^{j-1}\right) \\ &= C\frac{\beta^{u_0-1}e^{-\frac{\beta}{v_0}}p^{a_0-1}(1-p)^{b_0-1}}{B(a_0,b_0)\Gamma(u_0)v_0^{u_0}}\prod_{k=1}^{n}\left[\frac{(1-p)\sinh\left(\frac{y_k\sqrt{p}}{\beta}\right)}{\sqrt{p}\exp\{\frac{y_k}{\beta}\}\beta}\right] \end{aligned}$$

where C is a normalization constant.

Now, we would like the (joint) loss function to be some two-parameter version of squared error loss. This could be as simple as the sum of two squared error loss functions, or one which

highlights a greater loss in poorly estimating one parameter over the other. Here, we consider a loss function which measures percentage error equally between the two parameters by simply adding the squared percentage errors:

$$L(\hat{p},\hat{\beta}) = \left[\log\left(\frac{\hat{p}}{p}\right)\right]^2 + \left[\log\left(\frac{\hat{\beta}}{\beta}\right)\right]^2$$

To obtain a Bayes estimate, the following must be minimized:

$$R(\pi,\hat{p},\hat{\beta}) = \int_0^\infty \int_0^1 L(\hat{p},\hat{\beta})h(p,\beta;\mathbf{y})\,dp\,d\beta$$

In order to do this; we combine two types of numerical methodology: 1) numerical integration, and 2) numerical optimization. For each step in the optimization, numerical integration is done for the current values of $\hat{p}$ and $\hat{\beta}$, and the first step uses the expected values of p and β from their marginal priors.

MOM Estimates

For the Method of Moments estimates, note that $E(M) = \frac{1}{1-p}$ and $E(M^2) = \frac{p}{(1-p)^2}$. We have

$$\mu_Y^{(1)} = \frac{\mu_X^{(1)}}{1-p}$$
$$\mu_Y^{(2)} = \frac{\mu_X^{(2)}(1-p) + 2p\left(\mu_X^{(1)}\right)^2}{(1-p)^2}$$

Setting the sample moments equal to these, we have

$$\overline{Y} \overset{\text{set}}{=} \frac{\mu_X^{(1)}}{1-p} \tag{4.11}$$
$$\overline{Y^2} \overset{\text{set}}{=} \frac{\mu_X^{(2)}(1-p) + 2p\left(\mu_X^{(1)}\right)^2}{(1-p)^2} \tag{4.12}$$

Noting that $E(X) = 2\beta$, $E(X^2) = 6\beta^2$, we have in this case

$$\overline{Y} = \frac{2\beta}{1-p} \tag{4.13}$$

$$\overline{Y^2} = \frac{6\beta^2(1-p) + 2p\,(2\beta)^2}{(1-p)^2} \tag{4.14}$$

So, the Method of Moments estimates are

$$\hat{p}_{\text{MOM}} = \frac{-3\left(\overline{Y}\right)^2 + 2\overline{Y}^2}{\left(\overline{Y}\right)^2}$$

$$\hat{\beta}_{\text{MOM}} = \frac{2\left(\overline{Y}\right)^2 - \overline{Y}^2}{\left(\overline{Y}\right)}$$

Checking this for various simulated realizations yields inaccurate results for both $\hat{p}$ and $\hat{\beta}$, even when n is in the thousands. Again, this issue may be as a result of the apparent negative correlation between them.

Clearly, statistical inference of the compound geometric random variable is a complex problem. Correlation between estimates may also be a common problem, particularly when location parameters of X are to be estimated. However, the Bayesian method of estimation has built-in machinery to mitigate, though not completely eliminate, these issues.

4.3 Related Random Variables

In a more general form, M may not have a geometric distribution, or perhaps not even a well-known distribution. Moreover, the compounding machinery needn't be restricted to summation; other methods, such as considering order statistics and alternative functions of $\boldsymbol{X}$, produce many useful forms. Even more ambitious types could be compound geometric random variables where $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$ is not independent of M, or where the $\boldsymbol{X}_j$'s are either not independent of one another

or not identically distributed. In this more general setting, the range of possible random variables is much more broad. Some of these forms are briefly described in the following paragraphs.

Alternate Distributions for M

If M in Equation 4.1 has a Negative Binomial distribution, then Y is said to be *Negative Binomial Infinitely Divisible* and its characteristic function satisfies

$$\phi_Y(t) = \left(1 - \log \phi_V(t)\right)^{-r}$$

where V is some infinitely divisible random variable and $r \in \mathbb{R}^+$ [19]. Note that if $r = 1$, then Y is just g.i.d. This Y differs from the g.i.d. version in that there is a minimum number of X_js being summed to form Y. This type of random variable may be useful to model phenomena which have a minimum number of occurrences, each with a random magnitude. For example, a company can offer a service on an unlimited basis for a certain flat rate, and the number of times a customer uses the service is at least once (maybe it is natural to use it at the time of purchase). In this case, Y could model the overall cost of providing the services to its customers, where r is the number of customers who purchased the service at the flat rate, and the X_j's model the cost of each instance of a customer using the service.

Other distributions, such as the Poisson, the geometric with support including 0, or more exotic forms may include some viable options. These distributions can potentially form useful models for several applications.

Infinite support and independence between M and X are not necessities. Consider the case where $M \sim Bin(n, p)$, and $mX|M = m \sim U(0, 1)$ for all $m \in \{0, 1, 2, \cdots, n\}$, with the convention that $Y = 0$ whenever $M = 0$. It is easy to see, in this case, that the support of Y is $[0, 1)$, with

$P(Y = 0) > 0$.

Geometric Minimum and Maximum

To this point in the chapter, we have dealt with models involving convolution of the X's. However, this is not the only option. Instead, suppose that

$$Y = \max_{j \in \{1,2,\ldots,M\}} \{X_j\}.$$

If the density and distribution functions of X are available in closed form, so too will be the density and distribution functions for Y. Hence, for phenomena to which this model may apply, it is a potentially viable alternative to convolution.

Discrete X

Since the turn of the century, much interest has been brewing on the compound geometric random variable where X is assumed to be supported on $\mathbb{N}$. With applications in actuarial and risk assessment fields, this area is rich in literature. Willmot [48] is a starting point for further research.

Support on $(-\infty, \infty)$

To date, compound geometric random variables with both negative and positive support have received little attention. Insurance and actuarial applications generally consider nonnegative $X's$, but other applications may require a larger support set. For example, if $X \sim N(0, 1)$, then the geometric compounding of X forms a seamless continuum between normal and Laplace distributions. Explicitly, the family, $\{Y_p\}_{p \in (0,1)}$, where

$$Y_p = \sum_{j=1}^{M_p} X_j,$$

with $M_p \sim \text{Geo}(1-p)$ and $\{X_1, X_2, ...\}$ is a sequence of iid $N(0,1)$ random variables, independent of M_p, has two limiting distributions:

$$\lim_{p \longrightarrow 0^+} Y_p \sim N(0,1), \tag{4.15}$$

and

$$\lim_{p \longrightarrow 1^-} Y_p \sim Laplace(0,1), \tag{4.16}$$

that is, a standard Laplace distribution.[2]

4.4 Multivariate Compound Random Variables

Multivariate compound random variables come in many forms. Two types of such random variables are discussed here.

4.4.1 Compound Random Vectors of the First Kind

Consider the random variable, $\boldsymbol{M} := (M_1, M_2, ..., M_K)' \in S_{\boldsymbol{M}} \subset (\mathbb{N} \cup 0)^K$, with distribution $F_{\boldsymbol{M}}$, and the random vector, $\boldsymbol{X} := (X_1, X_2, ..., X_K)' \in S_{\boldsymbol{X}} \subset \mathbb{R}^K$, with distributions $F_1, F_2, ..., F_k$, respectively and $X_{k_1} \perp X_{k_2}$ for all $k_1 \neq k_2$. Then the random vector

$$\boldsymbol{Y} = \bigtimes_{k=1}^{K} \left[\sum_{m=1}^{M_k} X_{km} \right] \tag{4.17}$$

where $X_{km} \overset{d}{=} X_k$, and $X_{k_1 m_1} \perp X_{k_2 m_2}$ for all $(k_1, m_1) \neq (k_2, m_2)$, is a **compound random vector of the first kind**. This may be an appropriate model for the scenario described in Section 4.3, where M has a univariate negative binomial, but the costs are separated into K categories. We note here

[2] In fact, X needn't be normal for this to hold. Similar to what was shown in Proposition 5, if X 1) has a finite first moment; 2) has both positive and negative non-degenerate support; that is, at least two distinct negative and two distinct positive values are in the support set; 3) is appropriately scaled; and 4) is non-zero with probability 1; the latter limit will be some asymmetric Laplace distribution.

that it is possible for the cost in one area to be zero, and thus that a particular M_k could be zero. Generally, the distribution and density functions of Y are not expressible in closed form. However, some results are accessible in general.

Properties

The moments of Y can be computed directly:

$$\boldsymbol{\mu_Y} = E(\boldsymbol{Y}) = E(\boldsymbol{X}) \star E(\boldsymbol{M}), \tag{4.18}$$

where $\star$ denotes element-wise multiplication. Also,

$$\begin{aligned}
\Sigma_{\boldsymbol{Y}} &= Var(\boldsymbol{Y}) \\
&= Var(E(\boldsymbol{Y}|\boldsymbol{M})) + E(Var(\boldsymbol{Y}|\boldsymbol{M})) \\
&= \text{diag}\,(E(\boldsymbol{X}))^2\, Var\,(\boldsymbol{M}) + \text{diag}\,(E(\boldsymbol{M}))^2\, Var\,(\boldsymbol{X})
\end{aligned}$$

For $\boldsymbol{t} = (t_1, t_2, ..., t_K)'$, the characteristic function is

$$\begin{aligned}
\phi_{\boldsymbol{Y}}(\boldsymbol{t}) &= E\left[e^{i\boldsymbol{t}'\boldsymbol{Y}}\right] \\
&= E\left[e^{i\left(t_1 \sum_{m=1}^{M_1} X_{1m} + t_2 \sum_{m=1}^{M_2} X_{2m} + ... + t_K \sum_{m=K}^{M_K} X_{Km}\right)}\right] \\
&= E\left[E\left[e^{i\left(t_1 \sum_{m=1}^{M_1} X_{1m} + t_2 \sum_{m=1}^{M_2} X_{2m} + ... + t_K \sum_{m=K}^{M_K} X_{Km}\right)}\middle|\boldsymbol{M}\right]\right] \\
&= E\left[(\phi_{X_1}(t_1))^{M_1}\,(\phi_{X_2}(t_2))^{M_2} \cdots (\phi_{X_K}(t_K))^{M_K}\right]
\end{aligned}$$

Example

Suppose $K = 2$, and $M_1 \sim Bin(n = 6, p = 0.78)$ and $M_2 = 6 - M_1$, so that $M_2 \sim Bin(6, 0.22)$, and suppose $X_1, X_2 \overset{iid}{\sim} \Gamma(\alpha = 5, \beta = 3)$. Figure 4.3 shows 10,000 realizations of Y.

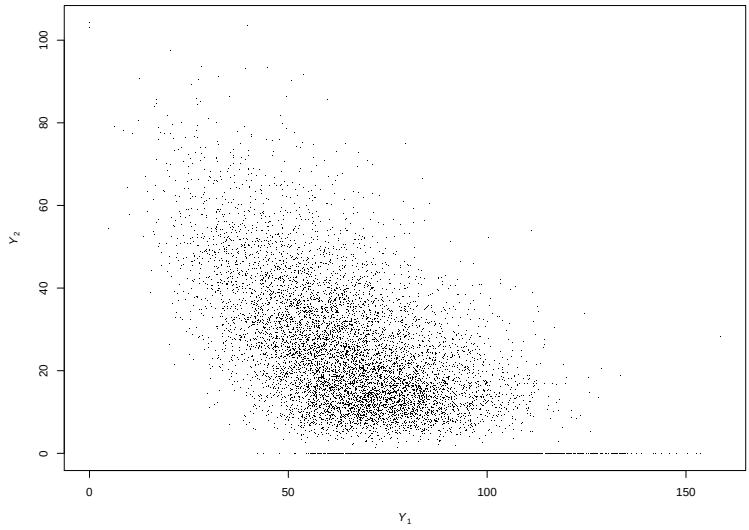


Figure 4.3: Ten Thousand Realizations of the Compound Random Vector of the First Kind.

Note that since $M_2 = 0$ with positive probability, the strip on the bottom edge appears on the plot. With all four parameters, n, p, α, and β, free, this family can produce a large array of forms. The mean and variance are

$$E(\boldsymbol{Y}) = \begin{bmatrix} E(X_1)E(M_1) \\ E(X_2)E(M_2) \end{bmatrix} = \begin{bmatrix} (15)(6)(0.78) \\ (15)(6)(0.22) \end{bmatrix} = \begin{bmatrix} 70.2 \\ 19.8 \end{bmatrix} \tag{4.19}$$

and

$$
\begin{aligned}
\Sigma_{\boldsymbol{Y}} &= \operatorname{diag}(E(\boldsymbol{X}))^2\, Var(\boldsymbol{M}) + \operatorname{diag}(E(\boldsymbol{M}))^2\, Var(\boldsymbol{X}) \\
&= \begin{bmatrix} (3)(5) & 0 \\ 0 & (3)(5) \end{bmatrix}^2 \begin{bmatrix} (6)(0.78)(0.22) & -(6)(0.78)(0.22) \\ -(6)(0.78)(0.22) & (6)(0.78)(0.22)) \end{bmatrix} \\
&+ \begin{bmatrix} (6)(0.78) & 0 \\ 0 & (6)(0.22) \end{bmatrix}^2 \begin{bmatrix} (3)(5)^2 & 0 \\ 0 & (3)(5)^2 \end{bmatrix} \\
&= 1.0296 \begin{bmatrix} 225 & 0 \\ 0 & 225 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + 75 \begin{bmatrix} 21.9024 & 0 \\ 0 & 1.7424 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1874.34 & -231.66 \\ -231.66 & 362.34 \end{bmatrix}
\end{aligned}
$$

4.4.2 Compound Random Vectors of the Second Kind

Consider the random variable, $M \in S_M \subset (\mathbb{N} \cup 0)$, with distribution F_M, and the random vector, $\boldsymbol{X} \in S_{\boldsymbol{X}} \subset \mathbb{R}^K$, with distribution $F_{\boldsymbol{X}}$. Then the random vector

$$\boldsymbol{Y} = \sum_{k=1}^{M} \boldsymbol{X}_k \tag{4.20}$$

where $\boldsymbol{X}_k \overset{d}{=} \boldsymbol{X}$, and $\boldsymbol{X}_{k_1} \perp \boldsymbol{X}_{k_2}$ for all $k_1 \neq k_2$, is a **compound random vector of the second kind**. It should be noted that $\boldsymbol{X}$ needn't be independent of M. This property opens the door to a vast array of forms, including those with compact support. The following example illustrates this.

Example

Suppose $M \sim Geo(1-p)$ (the version of the geometric with support $\mathbb{N}$), and $(m\boldsymbol{X}|M = m) \sim BB_{OL}(\alpha_U, \alpha_V, \alpha_W)$, where BB_{OL} denotes the Olkin-Liu 3-parameter bivariate beta distribution,

Model 2, as defined in Section 2.2.1. In this case, it is clear that the support of $\boldsymbol{Y}$ is the unit square. We consider the case where $p = 0.5$, $(\alpha_U, \alpha_V, \alpha_W) = (1, 2, 3)$. Ten thousand realizations of this is shown in Figure 4.4.

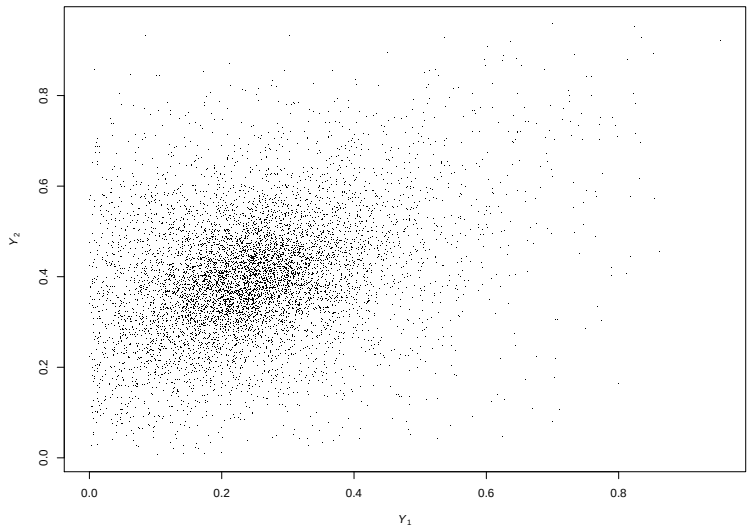


Figure 4.4: Ten Thousand Realizations of the Compound Random Vector of the Second Kind.

4.5 A Bivariate Compound Geometric Distribution

As was suggested in Section 4.2, exponential distributions turn up in many ways when compound geometric distributions are discussed. In the interest of extending these distributions to bivariate forms in a meaningful way, we introduce a bivariate Compound Random Vector of the Second Kind, where, in the context in Section 4.4.2, $\boldsymbol{X}$ has exponential marginals.

4.5.1 Pseudo-Exponential Distributions

To do this, we must first introduce the Filus [17] conditionally-specified exponential model. Suppose $V_1 \sim \exp(\alpha)$, and $V_2|V_1 = v_1 \sim \exp(\beta(v_1))$, where $\alpha > 0$ and $\beta : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is some differentiable-*a.e* function with domain $\mathbb{R}^+$. Then the joint density of $\boldsymbol{V}$ is

$$f_{\boldsymbol{V}}(\boldsymbol{v}) = \alpha\beta(v_1)e^{-\alpha v - \beta(v_1)v_2} I(v_1 > 0, v_2 > 0). \tag{4.21}$$

In general, while the marginal of V_1 is clearly exponential, the marginal of V_2 is not. However, Arnold [3] introduced a survival variation of this distribution, constructed in a similar manner to the Filus model. Set $X_1 \sim \exp(\alpha)$, and $X_2|X_1 > x_1 \sim \exp(\beta(x_1))$. Then, the survival function for $\boldsymbol{X}$ is

$$P(X_1 > x_1, X_2 > x_2) = e^{-\alpha x_1 - \beta(x_1)x_2}, \quad x_1, x_2 > 0. \tag{4.22}$$

In order for this to be a valid survival function, $\beta(x_1)$ must also be positive for all values of $x_1 > 0$, for, if $\beta\left(x_1^{(0)}\right) \leq 0$ for some $x_1^{(0)} > 0$, $\lim_{x_2 \to \infty} P\left(X_1 > x_1^{(0)}, X_2 > x_2\right) \neq 0$, a violation of the properties of a survival function. The mixed derivative gives the associated density function.

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = [\alpha\beta(x_1) + \beta'(x_1)(\beta(x_1)x_2 - 1)]e^{-\alpha x_1 - \beta(x_1)x_2}, \quad (x_1, x_2) \in S \times \mathbb{R}^+, \tag{4.23}$$

where $S = \mathbb{R}^+ \setminus S^-$, and S^- is the subset of $\mathbb{R}^+$ (of measure zero) over which $\beta'(x_1)$ is not defined.

Proposition 6. *A necessary and sufficient condition for (4.23) to be uniformly non-negative is*

$$0 \leq \beta'(x_1) \leq \alpha\beta(x_1) \quad \forall x_1 \in S \tag{4.24}$$

Proof. First, since $e^{-\alpha x_1 - \beta(x_1)x_2} > 0$ for all $x_1, x_2 > 0$, we need only check the coefficient. For

sufficiency, assume (4.24). Then, for any $(x_1, x_2) \in S \times \mathbb{R}^+$,

$$\begin{aligned}\alpha\beta(x_1) + \beta'(x_1)(\beta(x_1)x_2 - 1) &\geq \beta'(x_1) + \beta'(x_1)(\beta(x_1)x_2 - 1) \\ &= \beta'(x_1)(1 + \beta(x_1)x_2 - 1) \\ &= \beta'(x_1)(\beta(x_1)x_2) \\ &\geq 0\end{aligned}$$

For necessity, assume $\alpha\beta(x_1) + \beta'(x_1)(\beta(x_1)x_2 - 1) \geq 0 \quad \forall (x_1, x_2) \in S \times \mathbb{R}^+$. Notice that if for some $x_1^{(0)} \in S$, $\beta'\left(x_1^{(0)}\right) < 0$, then setting $x_2^{(0)} = \frac{2}{\beta\left(x_1^{(0)}\right)} - \frac{\alpha}{\beta'\left(x_1^{(0)}\right)}$ (which is positive since $\beta\left(x_1^{(0)}\right)$ is positive), then $\alpha\beta\left(x_1^{(0)}\right) + \beta'\left(x_1^{(0)}\right)\left(\beta\left(x_1^{(0)}\right)x_2^{(0)} - 1\right) = \beta'\left(x_1^{(0)}\right) < 0$, a contradiction. Hence, $\beta'(x_1) \geq 0 \quad \forall x_1 \in S$. Also, $\forall x_1 \in S$,

$$\lim_{x_2 \to 0^+} \alpha\beta(x_1) + \beta'(x_1)(\beta(x_1)x_2 - 1) = \alpha\beta(x_1) - \beta'(x_1) \tag{4.25}$$

So, by continuity in x_2, $\beta'(x_1) \leq \alpha\beta(x_1) \quad \forall x_1 \in S$. □

If, in addition, $\beta(x_1)$ is differentiable everywhere, then it must be that $\beta(0) > 0$ (that is, its continuous extension to 0 from the right exists and is greater than 0). To show this, recall from calculus the following lemma:

Lemma 7. *If $g(x_1)$ is a continuous measurable function with $g(a) = 0$ and $g(x_1) \geq 0$ for all $x_1 > a$, then for any $b > a$, there exists $x_1^{(0)} \in (a, b]$ such that*

$$\int_a^{x_1^{(0)}} g(x_1)dx_1 \leq g\left(x_1^{(0)}\right)\left(x_1^{(0)} - a\right) \tag{4.26}$$

Proof. By the Extreme Value Theorem and since $g(a) \leq g(x_1) \quad \forall x_1 \in (a, b]$, there exists $x_1^{(0)} \in (a, b]$ such that $g\left(x_1^{(0)}\right) \geq g(x_1)$ for all $x_1 \in [a, b]$. It follows that

$$\frac{1}{x_1^{(0)} - a} \int_a^{x_1^{(0)}} g(x_1)dx_1 \leq g\left(x_1^{(0)}\right) \tag{4.27}$$

The result immediately follows. □

Assume $\beta(0) = 0$ (and therefore, by necessity, $\beta'(0) = 0$), and set $a = 0$ and $b = \frac{1}{2\alpha}$. Since $\beta(x_1)$ is differentiable everywhere, $\beta'(x_1)$ is continuous and satisfies Lemma 7 for g. It follows from Lemma 7 and the Fundamental Theorem of Calculus that there exists $x_1^{(0)} \in (0, b]$ such that

$$0 \leq \beta'\left(x_1^{(0)}\right) \leq \alpha\beta\left(x_1^{(0)}\right) = \alpha \int_0^{x_1^{(0)}} \beta'(x_1)dx_1 \leq \alpha\beta'\left(x_1^{(0)}\right)x_1^{(0)} \leq \frac{\beta'(x_1^{(0)})}{2} \tag{4.28}$$

Since $\beta(x_1) > \beta(0)$ for $x_1 > 0$, it must be that $\beta'\left(x_1^{(0)}\right) > 0$. Thus, (4.28) is a contradiction, so it must be that $\beta(0) > 0$. Finally, a key result.

Proposition 8. *If $S = \mathbb{R}^+$, then $X_2 \sim \exp(\beta(0))$.*

Proof. By Equation 4.22, $P(X_2 > x_2) = \lim_{x_1 \longrightarrow 0+} P(X_1 > x_1, X_2 > x_2) = e^{-\beta(0)x_2}$. □

Remark 9. *Clearly the fact that $X_2 \sim \exp(\beta(0))$ requires that $\beta(0) > 0$. However, relaxing the requirement that $\beta(x_1)$ be differentiable everywhere to only being differentiable a.e. does form a larger family of distributions. For example, define $\beta(x_1)$ so that $\beta(x_1) = \frac{1}{k+1}$ whenever $x_1 \in \left[\frac{1}{k+1}, \frac{1}{k}\right)$ for $x_1 < 1$ and $k \in \mathbb{N}$, and $\beta(x_1) = \lfloor x_1 \rfloor$ otherwise. Clearly, this $\beta(x_1)$ satisfies (4.24) for all $x_1 \in S$ (in fact, all positive step functions do), where $S^- = \mathbb{N} \cup \{k^{-1} : k \in \mathbb{N}\}$, but $\beta(0)$ is not defined (though, by any reasonable definition, it would be zero). In this case, (X_1, X_2) is not an absolutely continuous random vector, nor does it have an exponential marginal distribution for X_2 (it is actually a mixture of an infinite number of exponentials). Undoubtedly, even more exotic forms of $\beta(x_1)$ are possible, but they nevertheless must satisfy (4.24).*

Example: Arnold Survival Pseudo-Exponential Nonlinear Model

Consider the family with joint distribution function given in Equation 4.22, in which $\beta(x_1) = \sqrt{2\alpha x_1 + \gamma^2}$, where $\gamma \geq 1$. We have that $0 \leq \beta'(x_1) = \frac{\alpha}{\sqrt{2\alpha x_1 + \gamma^2}} \leq \alpha\sqrt{2\alpha x_1 + \gamma^2} = \alpha\beta(x_1)$,

satisfying (4.24). In this case, the density of $\boldsymbol{X}$ is given by

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \alpha\beta(x_1) + \beta'(x_1)(\beta(x_1)x_2 - 1)e^{-\alpha x_1 - \beta(x_1)x_2} \\ &= \alpha\sqrt{2\alpha x_1 + \gamma^2} + \frac{\alpha}{\sqrt{2\alpha x_1 + \gamma^2}}\left(\sqrt{2\alpha x_1 + \gamma^2}x_2 - 1\right)e^{-\alpha x_1 - x_2\sqrt{2\alpha x_1 + \gamma^2}} \\ &= \alpha\left[\sqrt{2\alpha x_1 + \gamma^2} - \frac{1}{\sqrt{2\alpha x_1 + \gamma^2}} + x_2\right]e^{-\alpha x_1 - x_2\sqrt{2\alpha x_1 + \gamma^2}} \end{aligned} \tag{4.29}$$

Since $\beta(x_1)$ is differential everywhere, the marginals are, by Proposition 8,

$$f_X(x_1) = \alpha e^{-\alpha x_1}, \tag{4.30}$$

and

$$f_Y(x_2) = \gamma e^{-\gamma x_2} \tag{4.31}$$

This forms a 2-parameter family of bivariate exponential distributions. The corresponding survival copula is thus

$$\begin{aligned} &P\left(S_{X_1}(X_1) < u_1, S_{X_2}(X_2) < u_2\right) = P\left(e^{-\alpha X_1} < u_1, e^{-\gamma X_2} < u_2\right) \\ &= P\left(X_1 > -\alpha^{-1}\log(u_1), X_2 > -\gamma^{-1}\log(u_2)\right) \\ &= \alpha\left[\sqrt{-2\log(u_1) + \gamma^2} - \frac{1}{\sqrt{-2\log(u_2) + \gamma^2}} + -\gamma^{-1}\log(u_2)\right] \\ &\cdot e^{\log(u_1) - \gamma^{-1}\log(u_2)\sqrt{-2\log(u_1) + \gamma^2}} \end{aligned} \tag{4.32}$$

Example: Arnold Survival Pseudo-Exponential Linear Model

If, in Equation 4.22, $\beta(x_1) = \beta_0 + \beta_1 x_1$, where $\beta_0 > 0$ and $\beta_1 \leq \alpha\beta_0$, the joint density is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = [\alpha\beta_0 + \alpha\beta_1 x_1 + \beta_1(\beta_0 x_2 - 1 + \beta_1 x_1 x_2)]e^{\alpha x_1 - \beta_0 x_2 - \beta_1 x_1 x_2}, \quad x_1, x_2 > 0. \tag{4.33}$$

By Proposition 8, the marginals are

$$f_{X_1}(x_1) = \alpha e^{-\alpha x_1}, \tag{4.34}$$

and

$$f_{X_2}(x_2) = \beta_0 e^{-\beta_0 x_2}, \tag{4.35}$$

with joint survival function,

$$P(X_1 > x_1, X_2 > x_2) = e^{-\alpha x_1 - \beta_0 x_2 - \beta_1 x_1 x_2}, \quad x_1, x_2 > 0. \tag{4.36}$$

For the remainder of this chapter, we will write $\boldsymbol{X} \sim LPES(\alpha, \beta_0, \beta_1)$. It should be mentioned that if, as a special case, $\alpha = \beta_0 = 1$ and $\beta_1 = \theta \in [0, 1]$, then $\boldsymbol{X}$ has the Gumbel Bivariate Exponential Distribution [18].

Returning to the general case, given $S_{X_1}(x_1)$ and $S_{X_2}(x_2)$ are the survival functions for X_1 and X_2, respectively, the survival copula corresponding to this distribution is derived as follows:

$$\begin{aligned} P(S_{X_1}(X_1) < u, S_{X_2}(X_2) < v) &= P(e^{-\alpha X_1} < u, e^{-\beta_0 X_2} < v) \\ &= P(-\alpha X_1 < \log u, -\beta_0 X_2 < \log v) \\ &= P\left[X_1 > -\frac{\log u}{\alpha}, X_2 > -\frac{\log v}{\beta_0}\right] \\ &= e^{\alpha\left(\frac{\log u}{\alpha}\right) + \beta_0\left(\frac{\log v}{\beta_0}\right) - \beta_1\left(\frac{\log u}{\alpha}\right)\left(\frac{\log v}{\beta_0}\right)} \\ &= uve^{-\theta(\log u)(\log v)} \end{aligned} \tag{4.37}$$

where $\theta = \frac{\beta_1}{\alpha\beta_0}$. Since it must be that $0 \leq \beta_1 \leq \alpha\beta_0$, we have that $\theta \in [0, 1]$. This is an Archimedean copula with generating function $\psi(t) = \log(1 - \theta \log t)$, i.e., $P(S_{X_1}(X_1) < u, S_{X_2}(X_2) < v) = \psi^{-1}(\psi(u) + \psi(v))$ for this choice of ψ. Thus, when β_1 is close to $\beta_0\alpha$, the dependence between X_1 and X_2 is strong. In the next section, we introduce a compound geometric random vector of the second kind, with this bivariate exponential distribution as its summand variables.

4.5.2 A Compound Geometric Distribution with Bivariate Pseudo-Exponential Components

Define $\boldsymbol{Y}$ by

$$\boldsymbol{Y} = \sum_{j=1}^{M} \boldsymbol{X}_j \tag{4.38}$$

where $M \sim \text{Geo}(1-p)$, $p \in (0,1)$, and $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$ is a sequence of iid random variables, independent of M, with distribution $\boldsymbol{X}_j \sim LPES(\alpha, \beta_0, \beta_1)$ for all $j \in \mathbb{N}$. Since the marginals are exponential, by Equation 4.5, the marginals of $\boldsymbol{Y}$ are known:

$$Y_1 \sim \exp\left(\alpha(1-p)\right), \text{ and } Y_2 \sim \exp\left(\beta_0(1-p)\right) \tag{4.39}$$

This is a 4-parameter family of distributions, and we will write $\boldsymbol{Y} \sim CG_{LPES}(p, \alpha, \beta_0, \beta_1)$. While the joint density is not tractable in general, knowledge of the marginal distributions will set the stage for an MMLE method of parameter estimation. However, unlike the previous distributions to which likelihood-free methods were applied, simulating values from this distribution is not a trivial matter. The most accessible manner would be by way of Sklar's Theorem [40], that is by simulating realizations from the copula, Equation 4.37, and then applying the exponential quantile functions to obtain the marginal values. The procedure for simulating a single observation from $\boldsymbol{X} \sim LPES(\alpha, \beta_0, \beta_1)$ is outlined in the following process:

Step 1. Set $\theta = \frac{\beta_1}{\alpha\beta_0}$.

Step 2. Simulate u from $U \sim U(0,1)$, and, independent of U, w from $W \sim U(0,1)$.

Step 3. Find the $v \in (0,1)$ satisfying

$$F_{V|U}(v|u) = w, \tag{4.40}$$

where $F_{V|U}(v|u) = ve^{-\theta(\log u)(\log v)}(1 - \log v)$, via Newton-Raphson.

Step 4. Then, by the Probability Integral Transform, $\boldsymbol{x} = \left(S_{X_1}^{-1}(u), S_{X_2}^{-1}(v)\right)$ is from $\boldsymbol{X} \sim LPES(\alpha, \beta_0, \beta_1)$.

To obtain a simulated sample from $\boldsymbol{Y} \sim CG_{LPES}(p, \alpha, \beta_0, \beta_1)$, then, obtain m from $M \sim \text{Geo}(1-p)$, and simulate $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m)$, independently, from $\boldsymbol{X} \sim LPES(\alpha, \beta_0, \beta_1)$ using the process outlined above. Then

$$\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{x}_j \tag{4.41}$$

is a realization from $\boldsymbol{Y} \sim CG_{LPES}(p, \alpha, \beta_0, \beta_1)$.

Parameter Estimation for the CG_{LPES} Distribution

Now, as mentioned before, the marginal parameters for $\boldsymbol{Y}$ are estimable via MLEs. But, this provides estimates for only two of the four parameters. In fact, the copula of the $CG_{LPES}(p, \alpha, \beta_0, \beta_1)$ is dependent on α and β_0 only via $\theta = \frac{\beta_1}{\alpha\beta_0}$, i.e., p and θ completely define the dependence structure of this distribution, and the marginals have no impact on corresponding copula. We will therefore continue this discussion about only p and θ. Here, we must take care to avoid the problems illustrated in Section 4.2.6. Similar to that case, we have a trade-off, this time between p and θ. Is this trade-off direct enough to significantly restrict the space of these two parameters?

The answer to this question is evidently, yes. To see this, we recognize that the Archimedean copula in Equation 4.37 includes only non-positive correlations (it is the product copula when $\theta = 0$). However, the larger p is, the larger M will tend to be, and thus, the more likely Y_2 will tend to be large whenever Y_1 is. A plot of two cases with the same marginal distributions, Figure 4.5, makes this distinction clear. When p is small and θ is large, X_1 and X_2 will tend not to both

be small, that is, whenever one is small, the other will tend not to be. However, when p is large, this effect is overshadowed in $\boldsymbol{Y}$ by averaging, and thus, θ is difficult, if not practically impossible, to measure, even with very large sample sizes. We can deal with this problem in two ways. First, we

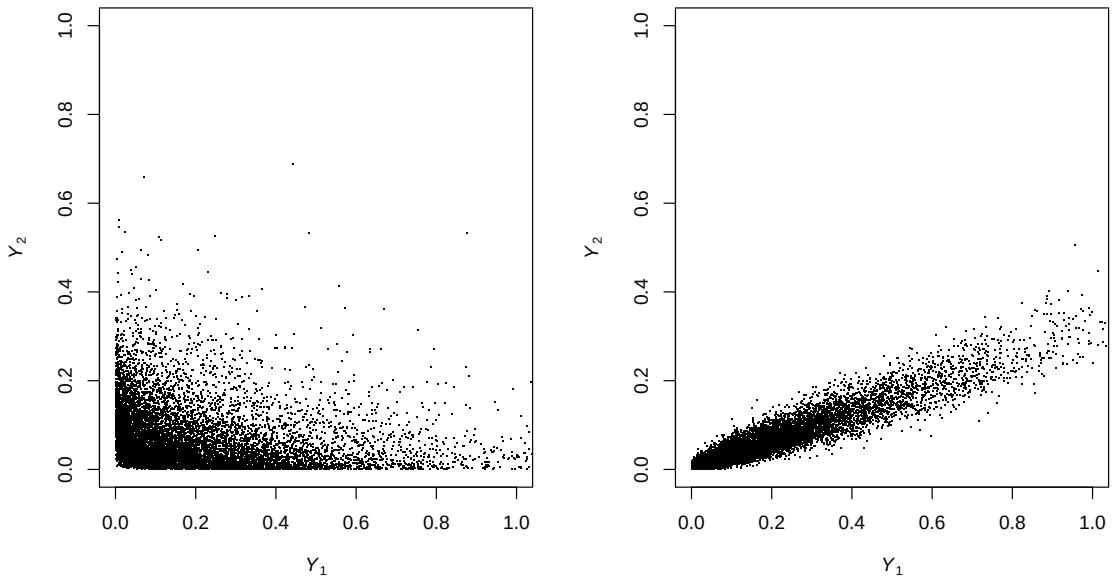


Figure 4.5: Ten Thousand Realizations of $\boldsymbol{Y} \sim CG_{LPES}(p = 0.05, \alpha = 5, \beta_0 = 15, \beta_1 = 50)$ (left) ,and $\boldsymbol{Y} \sim CG_{LPES}(p = 0.95, \alpha = 95, \beta_0 = 285, \beta_1 = 50)$ (right).

can apply Bayesian principles, restricting the expected ranges of values for both parameters through a carefully chosen prior. Second, we can reduce the model by setting both p and θ to be functions of a single parameter. Since the former was exhibited previously in the univariate case, we choose the latter in this example.

First, we choose to restrict the parameter space to a region (a curve), C, for (p, θ) to reflect the diminishing relevance of θ as p increases, while preserving its possible importance for small values of p. We choose C to be the curve given by $\theta = (1 - p)^2$. Simulation shows that the effects of

parameters θ and p interact strongly with Pearson correlation of the copula along C. Figure 4.6 is a plot of Pearson correlation, $\rho = Corr\left(e^{-\alpha(1-p)Y_1}, e^{-\beta_0(1-p)Y_2}\right)$, as a function of p, along C, modeled by a cubic polynomial, with ρ as the domain and p as the range. The ordered pairs $(p, \hat{\rho})$ used to generate this model (via regression) were obtained by generating very large samples for several values of p and computing the corresponding sample correlations. The correlations on this curve range from -0.54 to 0.65. The reason that ρ was chosen as the domain of the cubic model is that it is how the model will be applied to parameter estimation, making the estimation of p as simple as plugging in $\hat{\rho}$ to the polynomial model.

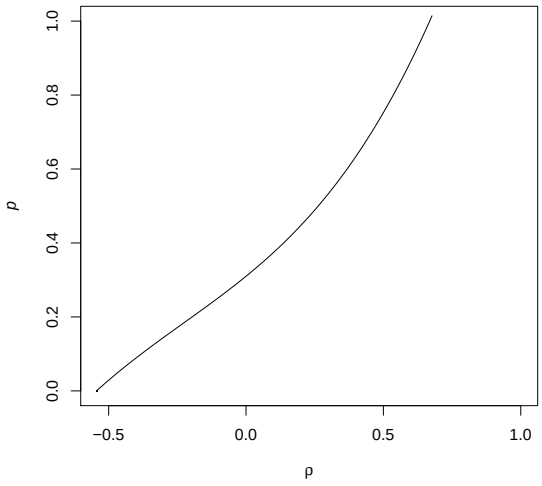


Figure 4.6: Cubic model of p as a function of ρ and where $\theta = (1-p)^2$.

This dimensional reduction of the parameter space results in a rather simple method of parameter estimation. Suppose $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_K)$ is an iid sample from $\boldsymbol{Y} \sim CG_{LPES}(p, \alpha, \beta_0)$. We apply an MMLE procedure, where the parameter, p (and, thus, θ), is determined by the Pearson correlation, in accord with Figure 4.6. We bootstrap this process to obtain a confidence region

for (p, α, β_0). This parameter-space reduction may appear to be somewhat unsatisfying; however, experience indicates that if a model corresponds to a parameter pair that falls below C, it can be closely represented by a model with parameters on C. One notable example is the product copula, which would be represented by the origin on this plot. The copula corresponding to $(p = 0.2938, \theta = 0.4987) \in C$ yields a copula almost indistinguishable from the product copula.

To test the MMLE approach for its capacity to identify the true distribution, we simulate 1000 samples from an arbitrary collection of parameter sets (p, α, β_0),[3] each sample of size 200. For each we perform the MMLE procedure, and record the mean bias and the MSE for each of the three parameter estimates. The results are shown in Table 4.1.

Parameter	Mean Bias	Mean Squared Error
p	0.0255	0.0041
α	0.0549	0.0209
β_0	0.0481	0.0204

Table 4.1: Parameter estimation results for 1000 arbitrary samples of size 100 from $\boldsymbol{Y} \sim CG_{LPES}(p, \alpha, \beta_0)$; bias and MSE for α and β_0 are computed from percentage errors.

4.6 Conclusions

In this chapter, we introduced compound random variables and discussed some of the literature related to them. We, in particular, discussed the important relationship between exponential random variables and compound geometric random variables. We also discussed some of the many possible variations of this type of random variable and how they may apply to practical problems. Two different forms of multivariate compound random variables followed, and we gave brief

[3]The p's were drawn from a $B(3, 13)$, that is, kept relatively small, so that it did not drown out the affects of β_1 for most samples.

examples of these. We finished with a bivariate compound geometric random variable with exponential marginals, and a dependence structure defined by a conditionally-specified survival model. We showed that one such version, the linear version, induces a convenient 3-parameter model with a wide range of correlations, and for which a simple and reliable MMLE method of parameter estimation is feasible

The possibilities to continue work in this area are vast. First, in light of the numerous applications in actuarial and risk assessment fields, further development of multivariate exponential compound geometric distributions, and survival models particularly, encompass promising areas for further research. The model discussed in Section 4.5 admits a new dependence structure for compound geometric exponential models. Extensions of this model to higher dimensions are straightforward, as Arnold [3] describes: for d dimensions, define $\boldsymbol{X}$ so that

$$\begin{aligned}
&X_1 \sim \exp(\alpha),\\
&X_2|X_1 < x_1 \sim \exp(\beta_1(x_1)),\\
&X_3|X_1 < x_1, X_2 < x_2 \sim \exp(\beta_2(x_1, x_2)),\\
&\qquad\qquad \vdots\\
&X_d|X_1 < x_1, X_2 < x_2, ..., X_{d-1} < x_{d-1} \sim \exp(\beta_{d-1}(x_1, x_2, ..., x_{d-1})).
\end{aligned} \tag{4.42}$$

Applying the Compound Random Vector of the Second Kind to $\boldsymbol{X}$ forms a multivariate version of the bivariate distribution discussed here. Though parameter estimation may be more difficult, the techniques highlighted and applied in this thesis may make it tractable for practical applications.

Chapter 5

Conclusions

This thesis has addressed a selection of questions regarding once intractable statistical models. With increasing computing capabilities, these models are becoming more accessible for common use. We have addressed three specific families of distributions that, in general, lack tractable density functions. In order to make full use of these models, unconventional methods must be constructed to estimate the parameters, based on data sets to which they are assumed to apply. Prediction and forecasting can then ensue unimpeded by the roadblocks of cumbersome analytical techniques. In essence, we are able to use the simplicity of simulating large amounts of data, and, through one or more of a variety of methods, infer a model's parameters, thus making the model useful.

This chapter will be organized into two parts. First will be a summary of the three models discussed in this thesis, some of the findings associated with them, and conclusions about their benefits and pitfalls. Second will be a discussion of likelihood-free techniques, including the current state of the discipline, and brief mentions of some underutilized methods. Throughout, we

will address possibilities for future work. In particular, we will address how some of the general findings about the models and methods studied in this thesis can be improved, adapted to specific applications, or augmented.

5.1 Specific Models

The models studied in this thesis comprise a minuscule sample of the vast array of statistical models that, only a few decades ago, would not be applicable to data because analysis involving the models was too computationally expensive. These three models were chosen for one or more of three primary reasons:

1. Their potential to be useful for an array of applications, that is, their flexibility;

2. Their lack of closed-form densities; and

3. Their ease of simulating realizations.

In Chapter 2, we discussed an 8-parameter family of bivariate beta distributions with gamma components. In chapter 3, we discussed a family of bivariate asymmetric Laplace distributions, also with gamma components, that are, in fact, a spin-off of the Bivariate Beta copula. In Chapter 4, we discussed a field with a much larger base in the literature, compound distributions. Primarily, we looked at such models related to the exponential distribution, and studied a bivariate compound geometric version based on a survival model. We address each of these in the following sections.

5.1.1 Bivariate Beta Distributions and Copulas

The 8-parameter Bivariate Beta model is a huge family, producing the full range of possible correlations and possible marginal beta distributions. There are, however, some issues that were identified in this work. First, the full model should not be applied to any practical estimation scenario, for the model exhibits a clear identifiability problem. For vastly different parameter configurations within this family, the corresponding densities can be, for all practical purposes, indistinguishable. This is a problem not only for parameter estimation, but also for understanding the effects of the parameters on the density shape. We cannot solve this problem entirely, for if we do, it will defeat the original purpose of the model: flexibility. What we can do is view the full model as an "omnibus" which includes various sub-models of potential interest.

In this thesis, this amounted to a starter kit. That is to say, we introduced what is arguably the most simple collection of 5-parameter sub-models whose set of possible marginal distributions is unrestricted. However, even this collection of models presents problems for parameter estimation. The reason for this appears to be a result of the nature of the parameters being estimated. To explain, recall that, given knowledge of the marginal parameters, the Olkin-Liu model is completely defined. No further investigation of the dependence structure is necessary, for it is completely defined by the marginal distributions. In contrast, for the 5-parameter Arnold model and the 5-parameter models constructed in Section 2.2.2, the situation is different. Here, upon defining the marginal parameters, a free parameter remains, one whose range depends on the marginals. It is also the most difficult to assess, for no help in its estimation is available from the well-established, not to mention density-based, estimation techniques available for the marginals. We can surely count parameters—three for Olkin-Liu, and five for the Arnold & Ng family—and even note the claim that the presence of

a density is significantly better than its absence. Nevertheless, the key take-away from this work is the following:

> **Models that contain dependence parameters which are independent of the marginal distributions present a significantly more difficult problem for estimation than models that do not.**

This is an illustration of the trade-off between flexibility and tractability. When multiple parameters are independent of the marginal distributions, the problem is compounded. In future work, further development of model-building strategies could prove useful, allowing for ways to identify, and incorporate into the final model, only the most significant parameters. Of course, there are hundreds of possible sub-models, all with parameter spaces of different shapes, sizes, and cross-sections of the full-model's space. Whenever possible, prior information should be incorporated not only into an informative prior distribution, but also into the design of the model selection process.

These issues with dependence parameters are nothing new. It is probably why the use of copulas is arguably one of the most rapidly growing fields of application of mathematical statistics today. And, that is why the second part of Chapter 2 was devoted to them. Since standard uniform distributions are $B(1, 1)$ distributions, there exists a sub-family of the full, 8-parameter, family of bivariate beta distributions which are also copulas. The particularly unique characteristic of this sub-family is the fact that it has four parameters, a rare characteristic of copulas, most of which have only a single parameter. Like other sub-families of the full model, parameter estimation presents difficulty. In fact, as was suggested in the previous paragraph, this difficulty is arguably greater than most, since all four parameters are dependence parameters. However, there are some interesting characteristics of the distribution that can be attributed to these parameters. Namely,

> **Each parameter of Arnold and Ghosh's 4-parameter family of copulas contributes to a specific and identifiable tail dependency.**

While there is certainly interaction amongst these contributions, a clear conclusion that can be made from the extensive simulation studies is that a particular tail dependency exists if and only if the corresponding parameter is nonzero. Specifically,

1. If $\delta_5 > 0$, then a tail dependency exists in the lower left, i.e. X_1 and X_2 tend to be small simultaneously;

2. If $\delta_6 > 0$, then a tail dependency exists in the upper right, i.e. X_1 and X_2 tend to be large simultaneously;

3. If $\delta_7 > 0$, then a tail dependency exists in the upper left, i.e. X_2 will tend to be large when X_1 is small; and

4. If $\delta_8 > 0$, then a tail dependency exists in the lower right, i.e. X_1 will tend to be large when X_2 is small.

Of course, some of these statements conflict with one another. Specifically, either of the first two statements conflicts with either of the last two. But, it is important to distinguish conflict from contradiction. For example, if $\delta_6 = \delta_7 = 0.5$, both corresponding statements are true, as is shown in the center density plot in Figure 2.14. In other words, if X_2 is large, X_1 will tend to be extreme, that is, either small or large, to a greater degree than in the independent case. These interactions, however, do complicate things for parameter estimation. In many cases, such as the one just highlighted, correlation measures or other measures of tail dependency do not work well, probably because the measures are sensitive to the interference caused by adjacent tail dependencies.

Another observation in this work is the persistence of the Ali-Michail-Haq copula throughout this research. This was something that piqued the interest of Roger Nelsen when hearing a talk

on this subject. This distribution (with its parameter equal to 1) is in the 4-parameter Arnold/Ghosh family, particularly in the case where $\delta_5 = 1$, and $\delta_6 = \delta_7 = \delta_8 = 0$. Hence, the Ali-Michail-Haq distribution represents a significant lower tail dependency. Also, its density function turned out to be useful for parameter estimation. The continuing ubiquity of this distribution is a mystery, but also an opportunity for further research.

In all, the work on the 8-parameter family, and its sub-families of copulas, is truly a work in progress. More must still be done than has been done. This research has as yet been unsuccessful in completely characterizing this family, but it has unearthed a host of new opportunities for research. Included in these are, most notably, characterizing the (classes of) families in Tables 2.4, 2.5, and 2.6, and forming model-selection processes for identifying useful models.

5.1.2 Bivariate Asymmetric Laplace Distributions

In this thesis, we studied a bivariate asymmetric Laplace distribution with eight parameters. With an intuitive interpretation of its parameters, this family is applicable to a range of practical problems, and its dependence parameters have a similar influence on the distribution to those of the Bivariate Beta copula. The difficulty with this distribution is in parameter estimation. While the marginal parameters are estimable via the MLEs, the dependence parameters, as has been the enduring theme of this thesis, present a problem for estimation.

In this work, while this problem was not solved, it was reduced. We developed a proposal distribution for MCMC which involves the Olkin-Liu Bivariate Beta Model. Specifically, we used the distribution to model two of the dependence parameters, based on large samples and a method of density estimation. The subsequent MCMC procedure provided reasonable estimates. However, it is by no means assumed that the method presented is ideal or maximally effective. The door is

open for improvements, particularly those that obviate the need for density estimation.

Higher-dimensional versions are cumbersome for this family of distributions. In order for the family to retain its flexibility, the number of parameters for three dimensions alone would more than triple, and for four, it would increase by an order of magnitude. However, if a suitable (forward) model selection technique can be constructed, it may permit utilization of higher-dimensional versions of this family.

An example of a regression scenario showed the applicability of this model to practical problems. The data included stock volumes over a year for two distinct stocks, and a regression model was applied to obtain residuals. The bivariate asymmetric Laplace model was applied to the residuals, and, according to the Akaine Information Criterion, it was preferred over a bivariate normal model.

Lastly, this model is an alternative to another bivariate asymmetric Laplace distribution, the Kotz [25] model. The Kotz model is easy to extend to higher dimensions, a reason why it may be preferred over the $BASL$. However, if multiple tail dependencies are present, the $BASL$ may be preferred. Much more work can be done to further this comparison, to include assessing similarities and differences, and identifying strengths of each.

In all, the $BASL$ distribution is a flexible family with, in general, no closed-form density. The parameters are easily interpretable, and parameter estimation is tractable for reduced models. Future work can focus on conducting a complete investigation of the model and its sub-models. One area is to apply advancements in data visualization methods by applying our intuitive understanding of the parameters' interpretations to augment the parameter estimation methods.

5.1.3 Compound Random Variables

Compound random variables include a large array of different random variables. In this work, the focus was on the overarching presence of exponential distributions, and bivariate versions of that distribution.

We constructed a model whose marginals are both exponential and whose copula is completely defined by only two of the parameters. This model is an example of the vast array of models that can be constructed through geometric compounding. In light of an evident identifiability problem, we reduced the space defining the copula to one of a single dimension, which led to a relatively simple method of parameter estimation on a new 3-parameter family of bivariate exponential distributions, constructed with conditionally-specified survival components.

An expansion of the parameter space is possible, particularly one that extends the space to at least some subset of the space *below* the curve, C, discussed in Section 4.5.2. This may prove useful in some cases, but should be done with caution to avoid reintroducing identifiability issues with the model.

Many possibilities exist for new models of the compound type, many of which will not result in tractable distribution functions, but will be uniquely applicable to specific types of datasets.

5.2 Likelihood-Free Methods

Distributions which are defined by the manipulation of random variables, rather than the manipulation of distribution functions, offer a particular advantage. When studying a phenomenon in this way, we can identify specific random variables to model specific components of the phenomenon, and manipulate them according to the perceived behavior of this overall system. Such

constructions of random variables are preferred because they provide intuitive justification for their application to common phenomena. However, as has been the case throughout this thesis, defining random variables in this way will often lead to intractable distribution and density functions, and have, in the past, been avoided for this reason. One important point to be observed is that we no longer need to avoid these types of constructions, and further advancement of the field of Likelihood-Free Statistical Inference will continue to make these constructions even more inferentially accessible.

The methods described and applied herein, by no means, form a comprehensive list. In this section, we briefly revisit each of the methods discussed, and express some areas where more development may be possible.

5.2.1 Modified Maximum Likelihood

If the marginal distributions of an otherwise intractable joint distribution are well-known, and their parameters can be easily estimated, then modified maximum likelihood estimation may be a viable method for estimating parameters of the joint distribution. This method was applied in several cases throughout this work, and proved useful. It is (often) simple and intuitive.

This method offers an opportunity to apply Sklar's Theorem, that is, to dismantle the joint distribution into its marginal distributions and a copula. Once done, the marginals can be estimated separately from the parameters defining the dependence structure. However, it should be noted that oftentimes a parameter can influence both marginal distributions and the copula. This can sometimes be resolved through an iterative process. For example, we can apply MMLE to estimate the marginal distributions, then apply these estimated distributions to the data to obtain an estimated copula, and ultimately estimate the copula parameters constrained by the marginal

parameters. This process can be repeated by bootstrapping the marginal estimates and repeating the constrained copula estimation. What results can be considered either an estimated sample from the distribution of the estimators, or a proposal distribution, as was the case in Chapter 3.

5.2.2 Approximate Bayesian Computation

We applied ABC in various ways throughout this thesis. It is a flexible and effective method for parameter estimation, and its use is justifiable even in the absence of prior information.

Most critical is the set of statistics, $\boldsymbol{S}(\boldsymbol{x})$. If no such set is known, and cannot be obtained though research, then this method may yield poor results. In addition, the distance function, ρ, must be large *in all cases* where the parameters are significantly different. A simple but costly error is to construct ρ so that it is not equally sensitive to important differences in all of the statistics within $\boldsymbol{S}$. For example, if $S_1(\boldsymbol{x})$ tends to be an order of magnitude larger than $S_2(\boldsymbol{x})$, and ρ is just the sum of squared errors, then ρ will end up ignoring errors in S_2 in favor of those in S_1. So scaling is important.

The choice of ϵ_0 is another concern. If too small, the only cost is excessive run times (due to a small acceptance rate). However, an ϵ_0 that is too large can cause excess bias and error in the posterior sample. Thus, simulation should be done to identify the smallest reasonable value of ϵ_0 to be used in the ABC procedure.

Some adaptations of ABC may prove useful. For example, in MMLE, if the marginal MLEs are easily obtainable but the remaining parameters are not by any means, a possible solution is to apply ABC in the following way. If a reasonable $\boldsymbol{S}(\boldsymbol{x})$ is still available, particularly for the remaining free parameter space after the marginal parameters are defined, do the following:

Step 1. Obtain the marginal MLEs, $\boldsymbol{\mu}_{MLE}$.

Step 2. Bootstrap the marginal MLEs to obtain a proposal set of marginal parameters, $\boldsymbol{\mu}^*_{MLE}$.

Step 3. Obtain the proposal from the remaining parameter space, which may be partially defined by the marginal parameters, $\boldsymbol{\mu}^*_{MLE}$, using any desired prior distributions.

Step 4. Perform the standard ABC step for accepting or rejecting the proposal.

Step 5. If $t = N$, then stop. Otherwise, repeat Step 2.

This method technically uses the data for its prior distribution. However, if we consider this method to be applied as the second stage of the MMLE parameter estimation process, one where the marginal MLEs are considered the true values, the method is justified, for the prior can legitimately be the distribution of the marginal MLEs, combined with any other prior information.

Another possible form for the prior distribution is a "learning" prior, one which, at some point in the process, begins to propose parameters which are similar to more successful previous proposals, and dissimilar to less successful ones. Many forms of this type of prior were attempted, but none worked. In all cases, they either led the process to converge illegitimately, or created excessive bias in the posterior sample. Whether it is possible to construct such a proposal distribution remains an open question.

5.2.3 Markov Chain Monte Carlo

A method applied in much more general settings where realizations are needed from an arbitrary distribution, MCMC can be a useful means to obtain a set of realizations from a posterior

distribution in a likelihood-free scenario. For the $BASL$ distribution, this method was applied, allowing for a specific proposal distribution to be constructed and applied. The result was a sample that exhibited minimal bias, and variation.

While this method was effective, it still did require density estimation. In future work either variations of this method that can avoid this, or other methods entirely, would be preferred, to form a purely likelihood-free method.

5.3 A Note About Pseudo-Randomness

The methods discussed herein are heavily reliant on the ability to simulate large amounts of data with computers. A well-known fact is that no random-number generator is truly random; it merely mimics randomness, and this pseudo-randomness becomes more of a factor as the size of a dataset grows. Thus, in engaging in these likelihood-free practices, we must remain cognizant of this, for the sizes of datasets generated in this endeavor will only continue to increase.

Bibliography

[1] Barry C Arnold. Some characterizations of the exponential distribution by geometric compounding. *SIAM Journal on Applied Mathematics*, 24(2):242–244, 1973.

[2] Barry C Arnold. A characterization of the exponential distribution by multivariate geometric compounding. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 164–173, 1975.

[3] Barry C. Arnold and Matthew A. Arvanitis. On bivariate pseudo-exponential distributions. *Unpublished Manuscript*, 2017.

[4] Barry C Arnold and Indranil Ghosh. Bivariate kumaraswamy models involving use of arnold-ng copulas. *Journal of Applied Statistical Science*, 22(3/4), 2014.

[5] Barry C Arnold and Hon Keung Tony Ng. Flexible bivariate beta distributions. *Journal of Multivariate Analysis*, 102(8):1194–1202, 2011.

[6] Matthew A Arvanitis. A 4-parameter copula with gamma components. In *Copulas and their Applications Conference; Almaria, Spain; July 5, 2017*, 2017.

[7] Matthew A Arvanitis. A bivariate asymmetric laplace distribtuion. In *Joint Statistical Meetings; Baltimore, MD; August 3, 2017*, 2017.

[8] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[9] J-L Bon. Error bounds for exponential approximation of large-system reliability. *Journal of Mathematical Sciences*, 138(1):5366–5376, 2006.

[10] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[11] Sung Nok Chiu and Chuancun Yin. On the complete monotonicity of the compound geometric convolution with applications in risk theory. *Scandinavian Actuarial Journal*, 2014(2):116–124, 2014.

[12] Roberto Carlos Crackel. *Likelihood Free Inference for a Flexible Class of Bivariate Beta Distributions*. PhD thesis, 2015.

[13] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[14] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[15] Fraser Daly. Compound geometric approximation under a failure rate constraint. *arXiv preprint arXiv:1504.06498*, 2015.

[16] Fraser Daly et al. Stein's method for compound geometric approximation. *Journal of Applied Probability*, 47(1):146–156, 2010.

[17] Jerzy K Filus and Lidia Z Filus. Pak. j. statist. 2006 vol. 22 (1) pp 21-42 on some new classes of multivariate probability distributions. *Pak. J. Statist*, 22(1):21–42, 2006.

[18] Emil J Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.

[19] Slobodanka Janković. Enlargement of the class of geometrically infinitely divisible random variables. *Publications de l'Institut Mathématique. Nouvelle Série*, 54(68):126–134, 1993.

[20] Zbigniew J. Jurek. 1d ising models, compound geometric distributions and selfdecomposability. *Reports on Mathematical Physics*, 47(1):21 – 30, 2001.

[21] George Kimeldorf, Allan R Sampson, et al. Monotone dependence. *The Annals of Statistics*, 6(4):895–903, 1978.

[22] LB Klebanov, GM Maniya, and IA Melamed. A problem of zolotarev and analogs of infinitely divisible and stable distributions in a scheme for summing a random number of random variables. *Theory of Probability & Its Applications*, 29(4):791–794, 1985.

[23] Lev Klebanov and Svetlozar Rachev. Sums of a random number of random variables and their approximations with ν-accompanying infinitely divisible laws. *Serdica Mathematical Journal*, 22(4):471–496, 1996.

[24] Lev Borisovich Klebanov, GM Maniya, and Joseph Aleksandrovich Melamed. A problem of zolotarev and analogs of infinitely divisible and stable distributions in a sheme for summing of a random number of random variables. *Teoriya Veroyatnostei i ee Primeneniya*, 29(4):757–760, 1984.

[25] Samuel Kotz, Tomaz J Kozubowski, and Krzysztof Podgórski. Asymmetric multivariate laplace distribution. In *The Laplace Distribution and Generalizations*, pages 239–272. Springer, 2001.

[26] Tomasz J Kozubowski and Anna K Panorska. Weak limits for multivariate random sums. *Journal of multivariate analysis*, 67(2):398–413, 1998.

[27] Andreas Nordvall Lagerås and Anders Martin-Löf. Genealogy for supercritical branching processes. *Journal of applied probability*, pages 1066–1076, 2006.

[28] David L Libby and Melvin R Novick. Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational Statistics*, 7(4):271–294, 1982.

[29] Gwo Dong Lin and Jordan Stoyanov. On the moment determinacy of the distributions of compound geometric sums. *Journal of Applied probability*, pages 545–554, 2002.

[30] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[31] Ingram Olkin and Ruixue Liu. A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412, 2003.

[32] Ingram Olkin and Thomas A Trikalinos. Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54–60, 2015.

[33] Nandita W Patel and MN Patel. Maximum likelihood estimation in compound geometric failure model with changing parameters from type-i two-stage progressively censored and group censored samples. *Communications in Statistics - Theory and Methods*, 36(13):2367–2375, 2007.

[34] RN Pillai and E Sandhya. Distributions with complete monotone derivative and geometric infinite divisibility. *Advances in Applied Probability*, pages 751–754, 1990.

[35] Georgios Psarrakos. A note on convolutions of compound geometric distributions. *Statistics & Probability Letters*, 79(9):1231–1237, 2009.

[36] Georgios Psarrakos. On the DFR property of the compound geometric distribution with applications in risk theory. *Insurance: Mathematics and Economics*, 47(3):428–433, 2010.

[37] A. Renyi. *Probability Theory*. Dover Books on Mathematics Series. Dover Publications, Incorporated, 2012.

[38] E Sandhya and RN Pillai. On geometric infinite divisibility. *arXiv preprint arXiv:1409.4022*, 2014.

[39] S Satheesh. Another look at random infinite divisibility. *Statistical Methods*, 6(2):123–144, 2004.

[40] Abe Sklar. Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449–460, 1973.

[41] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[42] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.

[43] Cajo JF ter Braak and Jasper A Vrugt. Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446, 2008.

[44] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[45] Brandon M Turner, Per B Sederberg, Scott D Brown, and Mark Steyvers. A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, 18(3):368, 2013.

[46] YH Wang. On the multivariate compound distributions. *Journal of multivariate analysis*, 59(1):13–21, 1996.

[47] Gordon E Willmot. On higher-order properties of compound geometric distributions. *Journal of Applied Probability*, pages 324–340, 2002.

[48] Gordon E Willmot and Jun Cai. Aging and other distributional properties of discrete compound geometric distributions. *Insurance: Mathematics and Economics*, 28(3):361–379, 2001.

[49] Gordon E Willmot, Jun Cai, et al. On applications of residual lifetimes of compound geometric convolutions. *Journal of Applied Probability*, 41(3):802–815, 2004.

Appendix A

Density Estimation

This appendix gives the details of the method of parameter estimation used in Chapter 3. Consider an absolutely continuous (*a.e.*) random variable, $\boldsymbol{X}$, supported on $\mathbb{R}^2$, with distribution function, $F_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})$ and density $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})$, both of which are unknown. However, the marginal distributions, $F_1(x_1;\boldsymbol{\theta})$ and $F_2(x_2;\boldsymbol{\theta})$, are known, and the corresponding quantile functions, $F_1^{-1}(u_1;\boldsymbol{\theta})$ and $F_2^{-1}(u_2;\boldsymbol{\theta})$, respectively, are known. We wish to estimate the density, $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})$, at the points, $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_K)$.

First, obtain quantiles $\left(q_{1,(n+1)^{-1}}, q_{1,2(n+1)^{-1}}, ..., q_{1,n(n+1)^{-1}}\right)$ and $\left(q_{2,(n+1)^{-1}}, q_{2,2(n+1)^{-1}}, ..., q_{2,n(n+1)^{-1}}\right)$, where $q_{i,p}$ satisfies

$$F_i(q_{i,p}) = p \tag{A.1}$$

for all $p \in (0,1)$ and $i \in \{1,2\}$, such that the convex hull of $\underline{\boldsymbol{x}}$ is contained in the rectangle with corners $\boldsymbol{c}_{11}$ and $\boldsymbol{c}_{nn}$, where $\boldsymbol{c}_{ij}$ is the center of the rectangle, $\left(q_{1,i(n+1)^{-1}}, q_{1,(i+1)(n+1)^{-1}}\right) \times \left(q_{2,j(n+1)^{-1}}, q_{2,(j+1)(n+1)^{-1}}\right)$, for all $i, j \in \{1, 2, ..., n-1\}$. Then generate a large sample, $\underline{\boldsymbol{y}} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_M)$, from $F_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})$, where M is sufficiently large to reduce error below some desired threshold. In addition, M must be suffi-

ciently large so that every sub-rectangle, $\left(q_{1,m_1(n+1)^{-1}}, q_{1,(m_1+1)(n+1)^{-1}}\right) \times \left(q_{2,m_2(n+1)^{-1}}, q_{2,(m_2+1)(n+1)^{-1}}\right)$, where $1 \leq m_1, m_2 \leq n$, contains at least one point in $\underline{\boldsymbol{y}}$. Define a_{ij} to be the area of the rectangle centered at $\boldsymbol{c}_{ij}$, for $i, j \in \{1, 2, ..., n\}$. Finally, define $\{d_{ij}\}_{i=1,j=1}^{n,n}$ by

$$d_{ij} = \frac{\text{(the number of points in } \underline{\boldsymbol{y}} \text{ in the rectange centered at } \boldsymbol{c}_{ij}}{Ma_{ij}} \tag{A.2}$$

Then d_{ij} is the approximate value of $f_{\boldsymbol{X}}(\boldsymbol{c}_{ij}; \boldsymbol{\theta})$.

Define r_{ij} to be the rectangle with corners $\boldsymbol{c}_{ij}$ and $\boldsymbol{c}_{i+1,j+1}$, for $i, j \in \{1, 2, ..., n-1\}$, and denote $\boldsymbol{c}_{ij} = \left((c_1)_{ij}, (c_2)_{ij}\right)$. Then, by design, every point in $\underline{\boldsymbol{x}}$ is in some r_{ij}. If $\boldsymbol{x}_k \in r_{i_0 j_0}$, then estimate the density by the weighted mean of the d_{ij}'s at the four corners of $r_{i_0 j_0}$.

$$\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}_k; \boldsymbol{\theta}) = \sum_{i=i_0}^{i_0+1} \sum_{j=j_0}^{j_0+1} \frac{(c_1)_{i_0+1,j_0+1} - (c_1)_{i_0 j_0} - |(c_1)_{ij} - x_{k1}|}{(c_1)_{i_0+1,j_0+1} - (c_1)_{i_0 j_0}} \cdot \frac{(c_2)_{i_0+1,j_0+1} - (c_2)_{i_0 j_0} - |(c_2)_{ij} - x_{k2}|}{(c_2)_{i_0+1,j_0+1} - (c_2)_{i_0 j_0}} d_{ij} \tag{A.3}$$

This approximation can easily be shown to be continuous everywhere, as well as differentiable *a.e.*, in the convex hull of $\underline{\boldsymbol{x}}$. It can also be shown to be converge to the actual density as both n and M increase, though the rate of convergence is dependent upon the distribution being measured. Below is an R function which computes the log-likelihood, where the density estimate for $\underline{\boldsymbol{x}}$ is obtained using this method, for a given *BASL* parameter set.

```
GetLogLikelihood <- function(X,params,minGridSize=M,N=simSz)
{
    maxSize <- 1000000
    K <- length(X[,1])
    xLim <- c(min(X[,1]),max(X[,1]))
    yLim <- c(min(X[,2]),max(X[,2]))
    gridSize <- max(c(3,minGridSize))
    tailSize <- ceiling(gridSize/10)
    centerSize <- gridSize-2*tailSize
    qSeq <- c(1/(centerSize+2)^(tailSize:1+1),
        (1:(centerSize+1))/(centerSize+2),
        1-1/(centerSize+2)^(1:tailSize+1))
    xBrks <- qASL(qSeq,c(params$beta1,params$beta3))
```

```
yBrks <- qASL(qSeq,c(params$beta2,params$beta4))
xP <- diff(xBrks)
yP <- diff(yBrks)
areas <- xP %*% t(yP)
estD <- matrix(0,gridSize,gridSize)
xVec <- xBrks[1:gridSize]+diff(xBrks)/2
yVec <- yBrks[1:gridSize]+diff(yBrks)/2
while ((xLim[1]<min(xVec))|(xLim[2]>=max(xVec))|
    (yLim[1]<min(yVec))|(yLim[2]>=max(yVec)))
{
    gridSize <- gridSize + 1
    tailSize <- ceiling(gridSize/10)
    centerSize <- gridSize-2*tailSize
    qSeq <- c(1/(centerSize+2)^(tailSize:1+1),
        (1:(centerSize+1))/(centerSize+2),
    1-1/(centerSize+2)^(1:tailSize+1))
    xBrks <- qASL(qSeq,c(params$beta1,params$beta3))
    yBrks <- qASL(qSeq,c(params$beta2,params$beta4))
    xP <- diff(xBrks)
    yP <- diff(yBrks)
    areas <- xP %*% t(yP)
    estD <- matrix(0,gridSize,gridSize)
    xVec <- xBrks[1:gridSize]+diff(xBrks)/2
    yVec <- yBrks[1:gridSize]+diff(yBrks)/2
}
n <- max(c(N,1000*(gridSize)^2))
numBlks <- n %/% maxSize
if (n>maxSize*(numBlks))
{
    n <- c(rep(maxSize,numBlks),n-maxSize*(numBlks))
    numBlks <- numBlks + 1
} else
{
    n <- rep(maxSize,numBlks)
}
delta <- rep(0,8)
delta[5:8] <-c(params$delta5,params$delta6,
    params$delta7,params$delta8)
delta[1] <- max(0,min(c(1,1-delta[5]-delta[7])))
delta[2] <- max(0,min(c(1,1-delta[5]-delta[8])))
delta[3] <- max(0,min(c(1,1-delta[6]-delta[8])))
delta[4] <- max(0,min(c(1,1-delta[6]-delta[7])))
totalCt <- 0
totalObs <- 0
```

```
    for (j in 1:numBlks)
    {
        locsLength <- gridSize^2
        locs <- data.frame(var1=integer(locsLength),
        var2=integer(locsLength),
        freq=integer(locsLength))
        U <- matrix(rgamma(n[j]*8,delta,1),nrow=8)
        Y <- cbind(colSums(U[c(3,6,8),])*params$beta3-
            colSums(U[c(1,5,7),])*params$beta1,
            colSums(U[c(4,6,7),])*params$beta4-
            colSums(U[c(2,5,8),])*params$beta2)
        Y <- Y[(Y[,1]>xVec[1])&(Y[,1]<xVec[gridSize])&
            (Y[,2]>yVec[1])&(Y[,2]<yVec[gridSize]),]
        eXp <- xP*n[j]
        eYp <- yP*n[j]
        totalObs <- totalObs + sum(locs[,3])
        den <- (totalCt+n[j])
        prop0 <- totalCt/den
        locs <- as.data.frame(table(findInterval(Y[,1],xBrks),
            findInterval(Y[,2],yBrks)))
        totalCt <- totalCt + n[j]
        estD[cbind(locs[,1],locs[,2])] <- estD[cbind(locs[,1],
            locs[,2])]*prop0 + locs[,3]/den
    }
    estD <- estD/areas
    ell <- rep(0,K)
    for (k in 1:K)
    {
        xPos <- sum(X[k,1]>xVec)
        xRw <- (X[k,1]-xVec[xPos])/(xVec[xPos+1]-xVec[xPos])
        xLw <- 1-xRw
        yPos <- sum(X[k,2]>yVec)
        yRw <- (X[k,2]-yVec[yPos])/(yVec[yPos+1]-yVec[yPos])
        yLw <- 1-yRw
        ell[k] <- log(xLw*yLw*estD[xPos,yPos] +
        xRw*yLw*estD[xPos+1,yPos] +
        xLw*yRw*estD[xPos,yPos+1] +
        xRw*yRw*estD[xPos+1,yPos+1])
    }
    return(sum(ell))
}
```