On the Use of Copulas for Multivariate Distribution Inference for Direct Cosimulation

Zhou Lan and Oy Leuangthong

Centre for Computational Geostatistics
Department of Civil and Environmental Engineering,
University of Alberta

Abstract

Research in the area of direct simulation is strongly motivated by the potential to improve the characterization of natural phenomena. This promises to avoid the need for transformation, while permitting the integration of multiscale information and the construction of property models directly onto realistic, complex grids. While most of the research energies in this area have been directed towards univariate modeling, the significant potential of multivariate modeling is key to the practicality of direct simulation.

Numerous methodologies have been proposed to infer the shape of the univariate conditional distributions, on which Monte Carlo simulation is performed; however, inference of the multivariate distribution remains a longstanding challenge not only in geostatistics but statistics in general. To address this issue, we can draw on the experience from the economic and financial industries, where the use of copulas theory has been actively explored in the last 50 years to address this problem specifically. A copulas function is a multivariate function constructed with the express purpose of accessing the shape of the multivariate distribution, given knowledge of the univariate marginal distributions. This paper explores the background and theory of copulas functions, and presents a methodology to construct a copulas function. A framework for application in direct sequential cosimulation for geostatistical modeling is finally presented.

1 Introduction

The simulation of multiple variables has wide application in various industries and research domains. Traditional methods of simulation assume that the variables either come from a multiGaussian (MG) distribution, or can be properly transformed to obtain a MG distribution. Useful transformations in this regard include the stepwise conditional transformation [6], alternating conditional expectation (ACE) [1], principal components analysis (PCA), etc. In many cases, this may be achieved using one or a combination of the previous transforms; however, in *most* cases, conforming to the MG assumption is unsuitable and difficult to achieve. Nevertheless, the MG assumption presents a greatly simplified simulation approach since the multivariate distribution is fully defined based on very limited information: the mean vector and variance-covariance matrix.

Greater flexibility in multivariate simulation warrants consideration of a non-Gaussian approach. This is not without it's own challenges, the biggest of which is the inference of the multivariate distribution. Direct simulation allows for modeling without the need for data transformation. Soares and Deutsch have presented methodologies to access the univariate conditional distributions. Soares further extended this result for collocated cosimulation of multiple variables; the use of collocated cokriging and hence adoption of the Markov screening assumption has the inherent result of assuming a bivariate (or multivariate) Gaussian distribution.

The motivation for this paper is to address the issue of inferring the *shape* of the multivariate distribution, that will permit a true departure from the MG framework. One of the important keys to this problem may lie in the theory of copulas. Sklar (1959) first developed the term copulas for a multivariate distribution with marginal distribution as uniform [0,1]. Since then, research in this area has rapidly developed and played an important role in the study of multivariate distributions, particularly for non-MG cases. This has been especially relevant and applicable to the financial and insurance industries over the last few decades. This paper first gives an overall brief introduction on copulas, followed by an application of the copulas to estimate the multivariate distribution of several analytical and real data distributions. A discussion on and framework for implementation into direct sequential cosimulation is also provided.

2 Background on Copulas Theory

Consider the case of two continuous random variables (RV), X and Y, with cumulative distribution functions (cdf) given by $F_X(x)$ and $F_Y(y)$, respectively. Let $U = F_X(x)$, then U is a uniformly distributed RV on the unit interval $\mathbf{I} = [0,1]$; similarly, let $V = F_Y(y)$, then V is also uniformly distributed on the interval $\mathbf{I} = [0,1]$. Working within the unit space of U and V(or equivalently the probability space of X and Y), we can apply the copulas function to describe and model the bivariate distribution (see Figure 1).

Sklar (1959) and Schweizer and Sklar (1974) give a definition for a two-dimensional (bivariate) copula as a function $C : \mathbf{I}^2 \rightarrow \mathbf{I}$, such that:

1. For every $u, v \in \mathbf{I}$, $C(u,0) = C(0,v) = 0$, $C(u,1) = u$ and $C(1,v) = v$
2. For every $u_1, u_2, v_1, v_2 \in \mathbf{I}$ such that $u_1 \leq u_2, v_1 \leq v_2$:

$$\begin{aligned} C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) &\geq 0 \\ C(u_2,v_2) + C(u_1,v_1) &\geq C(u_1,v_2) + C(u_2,v_1) \end{aligned} \quad (1)$$

Figure 2 shows a schematic illustration of the above conditions for the copulas function. One can consider the copulas function as a multivariate cdf expressed as a function of probability units, not the data units as would be expected of a conventional multivariate cdf.

The above bivariate conditions can be extended to the more general n-variate definition [15]. An n-dimensional copula C is a function $C : \mathbf{I}^n \rightarrow \mathbf{I}$ with the following properties:

1. For all $\mathbf{u} = \{u_1, \ldots, u_n\} \in \mathbf{I}^n, C(\mathbf{u}) = 0$ if at least one coordinate $u_i = 0, i = 1, \ldots, n$.

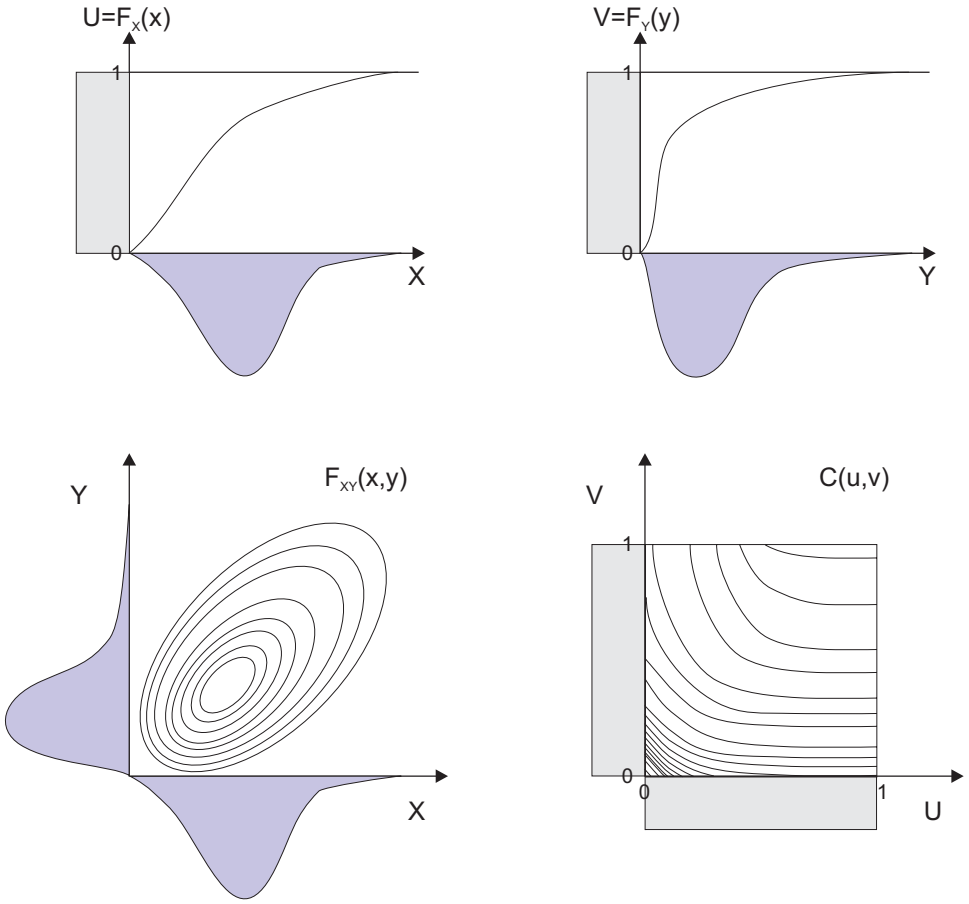

Figure 1: Schematic illustration of the relation between $U = F_X(x)$, $V = F_Y(y)$, $F_{XY}(x, y)$ and $C_{UV}(u, v)$.

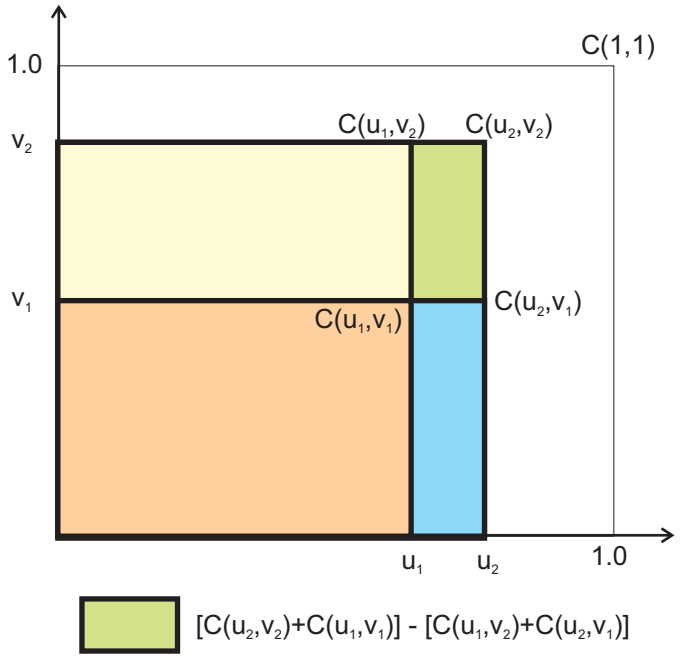

Figure 2: Schematic illustration of the definition of the copulas function $C_{UV}(u, v)$.

2. If all coordinates of $\mathbf{u}$ equals 1.0 except $u_i, i = 1, \ldots, n$, then $C(\mathbf{u}) = u_i, i = 1, \ldots, n$.

3. For all $u_i, i = 1, \ldots, n$ in $n-$increasing: The $C-$volume of all hypercubes with vertices in $\mathbf{I}^n$ is positive, that is

$$\sum_{i_1=1}^{2} \ldots \sum_{i_n=1}^{2} (-1)^{i_1+i_2+\ldots+i_n} C(u_{1,i_1}, \ldots, u_{n,i_n}) \geq 0$$

for all $(u_{1,1}, \ldots, u_{n,1})$ and $(u_{1,2}, \ldots, u_{n,2})$ in $\mathbf{I}^n$ such that $u_{j,1} \leq u_{j,2}, \forall j \in \{1, 2, \ldots, n\}$. This is equivalent to the condition expressed in Equation 1 in the bivariate case above.

Essentially, the copulas function is a multivariate distribution function that represents the dependence structure of the joint cdf of $X_1, \ldots, X_n$, denoted as $F_{X_1,\ldots,X_n}$, with marginal cdfs $F_{X_1}, \ldots, F_{Xn}$:

$$\begin{aligned} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)) \\ &= C(u_1, \ldots, u_n) \end{aligned} \tag{2}$$

Of course, one could rewrite Equation 2 as

$$C(u_1, \ldots, u_n) = F_{X_1,\ldots,X_n}(F_{X_1}^{-1}(u_1), \ldots, F_{X_n}^{-1}(u_n))$$

In fact, if all marginal cdfs $F_{X_1}, \ldots, F_{X_n}$ are continuous, then $C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n))$ is unique [21]. Otherwise, $C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n))$ is uniquely determined on $Range(F_1) \times$

$Range(F_2) \times ... \times Range(F_n)$. Conversely, If C is an n-copula function and $F_{X_1}, \ldots, F_{X_n}$ are marginal cdfs, then $F_{X_1,...,X_n}(x_1, \ldots, x_n)$ defined by the above equation is an $n-$dimensional distribution function (joint cdf) with margins $F_1, F_2, ..., F_n$. This is known as Sklar's Theorem, which was established and proved by Sklar [18]; the corresponding theorem for the $n-$dimensional case was proved by Moore and Spruil [11], Deheuvels [2] and Sklar [19].

Based on Sklar's theorem, we can estimate a multivariate distribution $F_{X_1,...,X_n}(x_1, \ldots, x_n)$ by first estimating the marginal distribution for each variable, $F_{X_1}(x_1), \ldots, F_{X_n}(x_n)$, and then apply the copulas function to estimate the multivariate distribution. Directly from Sklar's theorem, the joint probability density function (pdf) for $F_{X_1,...,X_n}(x_1, \ldots, x_n)$ can be expressed as follows:

$$f_{X_1,...,X_n}(x_1, \ldots, x_n) = c(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)) \cdot \prod_{j=1}^{n} f_{X_j}(x_j)$$

where c is the density function for copulas C and $f_{X_i}(x_i)$ is the marginal density function for the i^{th} variable, $i = 1, \ldots, n$.

2.1 Classes and Families of Copulas

Not all copulas are created equal; in fact, there are various classes of copulas, each with specific definition and properties, and certain attributes related to and/or suitable for describing certain distributions. Following the taxonomic hierarchy for related groupings, classes denote a higher/larger grouping, within which one can find a family of lower/smaller groupings. In this fashion, the following important classes of copulas are identified, and certain typical families within these classes are discussed. Specifically, the class of elliptical, Archimedean, and Marshall and Olkin's copulas are discussed [22].

2.1.1 Elliptical Copulas

As the name suggests, this class of copulas apply for elliptical distribution, which include the normal or Gaussian distribution, Student's t distribution and Laplace distribution [14]. The first two distributions present the most important families within this class. The Gaussian copula applies for the MG distribution, wherein the joint and all lower-dimensional distributions are Gaussian. The copulas function is given as

$$C(u_1, u_2, ...u_n) = G_{X_1,...,X_n}[G_{X_1}^{-1}(u_1), \ldots, G_{X_n}^{-1}(u_n)]$$

where $G(\cdot)$ is the standard normal cdf.

Similarly, the Student's t Copula corresponds to Student's t distribution and is given by

$$C(u_1, u_2, ...u_n) = t_{\nu,X_1,...,X_n}[t_{\nu,X_1}^{-1}(u_1), \ldots, t_{\nu,X_n}^{-1}(u_n)]$$

where $t_{\nu,X_i}(\cdot)$ represents the standard multivariate Student's t distribution with ν degrees of freedom for the i^{th} variable.

2.1.2 Archimedean Copulas

In practice, we are often confronted with non-elliptical distributions; in these cases, we need a more extensive/flexible approach to estimate and model the multivariate distributions. Archemidean Copulas is one important class which meets this goal.

In a bivariate distribution, an Archemidean Copula, usually denoted as C^{ϕ} is a function given by

$$C^{\phi}(u, v) = \phi^{-1}[\phi(u) + \phi(v)]$$

where $u, v \in (0, 1]$, and $\phi(.)$ is called the generator, that is a continuous, convex and strictly decreasing function defined on interval [0,1] with range [0,+∞) such that $\phi(1) = 0$ [12]. This can be extended to the multivariate case [21]:

$$C(u_1, u_2, ..., u_n) = \phi^{-1}[\phi(u_1) + \phi(u_2) + ... + \phi(u_n)]$$

This class of copulas is attractive for a number reasons. Firstly, different generators can be adopted to suit the characteristics of different distributions. For this reason, there are a large number of families designed to fit wide range of bivariate and multivariate distributions. In fact, Nelsen (1999) lists 22 families under the Archimedean copulas [12] Secondly, this is a parametric copulas which can be used to infer the bivariate or multivariate distributions.

Some of the families that fall under the Archimedean class include (Frees and Valdes 1998):

- Clayton's family: In this case, $\phi(t) = t^{-\alpha} - 1$, with $\alpha > 1$ and $t \in [0, 1]$, which yields the copulas function $C_{\phi} = (u^{-\alpha} + v^{-\alpha} - 1)^{-\frac{1}{\alpha}}$. This family has a lower tail dependence.
- Gumbel's family: For this family, $\phi(t) = (-\ln t)^{\alpha}$, with $\alpha \geq 1$ and $t \in [0, 1]$, and the corresponding copulas is $C_{\phi} = \exp\{-[(-\ln u)^{\alpha} + (-\ln v)^{\alpha}]^{\frac{1}{\alpha}}\}$. This yields an extreme value copulas function with upper tail dependence. As a special case, when $\alpha = 1$, $\phi(t) = -\ln(t)$, and $C_{\phi} = uv = C^{\perp}$, and effectively represents independence between u and v.
- Frank's family: Consider $\phi(t) = \ln \frac{e^{\alpha t} - 1}{e^{\alpha} - 1}$, $\alpha \in (-\infty, +\infty), \alpha \neq 0$ and $t \in [0, 1]$. This yields the copulas function $C^{\phi} = -\frac{1}{\alpha} \ln[1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha u} - 1)}{e^{-\alpha} - 1}]$.

2.1.3 Marshall-Olkin's Copulas

Another class of copula which is closely related to Archemidean class is called Marshall and Olkin' copula [8, 9, 10]. The link between Marshall-Olkin's copulas and the Archimedean copulas lies in the Laplace transformation. If the inverse Laplace transform of the generator, $\phi(t)$, is taken, then the Archimedean copulas takes on the form of the Marshall-Olkin copulas. In fact the generators for Clayton's, Gumbel's and Frank's family (as discussed above) are all inverses of the Laplace transformation from certain distributions. Specifically, Clayton's family is from Gamma distribution, Gumbel's family is from Positive stable distribution, while Frank's family is from the Logarithmic series distribution on positive integers.

Marshall-Olkin copulas have been used as a survival copulas, wherein a system (comprised of multiple components) is subjected to shocks which may cause failure of one or more components. The idea is then to generate a copulas that characterizes the failure of this system [12].

2.2 Fitting and Simulation of Copulas

There are several approaches to fitting and modeling the copulas function. In general, two processes must be performed: (1) estimate the marginal distribution of each variable, and (2) estimate the copulas of the joint distribution. These two steps can be carried out simultaneously or they can be performed sequentially. Both parametric and non-parametric approaches are addressed below.

2.2.1 Parametric Approach

With the aid of parametric assumptions, this group of modeling approaches capitalizes on some elegant analytical forms for model fitting. The first approach is based on a Maximum Likelihood Estimator (MLE) while the second method takes the form of a Canonical Maximum Likelihood (CML).

Maximum Likelihood Estimator (MLE). Let us assume that each variable has a known parametric distribution, say $F_i(x_i;\vartheta_i), i = 1,\ldots,n$, where ϑ_i is the parameter vector for the i^{th} margin. Further, assume we know a parametric copulas function for the variable, for example a specific family in Archimedean class, with $\vec{\alpha}$ the parameter vector for the copulas. Suppose we have a data set $\mathbf{X}$=$\{x_1^t,\ldots,x_n^t\}, t = 1,\ldots,T$, where T is the number of observations for each variable.

We can then apply parametric copula function using its density to establish the Likelihood function and further obtain the log-likelihood function as

$$l[\vartheta] = \sum_{t=1}^{T} \ln c(F_1(x_1^t;\vartheta_1),\ldots,F_n(x_n^t;\vartheta_n);\vec{\alpha}) + \sum_{t=1}^{T}\sum_{i=1}^{n} f_i(x_i^t;\vartheta_i)$$

where $c(\cdot)$ is the copulas density function, and ϑ is the set of parameters for the marginal distribution functions. We then obtain the estimated $\hat{\vartheta}$ and $\hat{\vec{\alpha}}$ that maximizes $l[\vartheta]$, and thus obtain the estimated copula function.

Canonical Maximum Likelihood (CML). Unlike the MLE, CML assumes that the copulas follows a parametric form, but the marginal distributions do not. This first requires the empirical distribution function for each variable, $\hat{F}_{X_i}(x_i), i = 1,\ldots,n$; however, this can usually be accomplished using a large data set that is completely different from that used to fit the copulas model. The empirical distributions, $\hat{F}_{X_i}(x_i)$, are then used to transform all observations T into a set of uniform variates, $\{\hat{u}_1^t,\ldots,\hat{u}_n^t\}$ where $\hat{u}_i^t = \hat{F}_i(x_i^t)$. This is finally used to construct the log-likelihood function

$$l[\vec{\alpha}] = \sum_{t=1}^{T} \ln c[\hat{u}_1^t,...,\hat{u}_n^t;\vec{\alpha}]$$

for which the parameter vector $\hat{\vec{\alpha}}$ can be determined to maximize the above function.

2.2.2 Nonparametric Estimation

Now we assume that neither the copula nor the marginal distributions are of parametric form. Deheuvels (1979) introduced an empirical copula for this purpose. Suppose we have a data set of T observations from a $n-$variate distribution, say $X_j^t, j = 1, \ldots, n; t = 1, \ldots, T$. First, we order the observations $\{x_1^{(t)}, \ldots, x_n^{(t)}\}$ and obtain the corresponding rank set $\{r_1^t, \ldots, r_n^t\}$, where r_j^t is the number of observations in X_j which is less than or equal to x_j^t. The empirical copulas is estimated as:

$$\hat{C}[\frac{t_1}{T}, \ldots, \frac{t_n}{T}] \quad = \quad \frac{1}{T} \sum_{t=1}^{T} \prod_{j=1}^{n} i(r_j^t \leq t_j) \tag{3}$$

where $t_p \in \{0, \ldots, T\}$, and $i(\cdot)$ represents the indicator function:

$$i(r_j^t \leq t_j) = \begin{cases} 1, & \text{if } r_j^t \leq t_j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

2.3 Model Selection

While the previous section discussed several ways to estimate the copulas function, this section focuses on the selection of the most suitable model to adequately fit the data.

Parametric Model. Given the large number of families that belong to the Archimedean class and its corresponding flexibility to address most problems, let's consider model selection under this type of copulas. For the purpose of simplification, we suppose a bivariate distribution with variables $\{X_1, X_2\}$ has an Archimedean copulas with generator $\phi(.)$, Genest and Rivest (1993) defined a function [4]:

$$K_\phi(z) = z - \frac{\phi(z)}{\phi'(z)} \tag{5}$$

where $\phi'(\cdot)$ is the first derivative of $\phi(\cdot)$ and z is a value within [0,1]. A non parametric estimate of K is given by

$$\hat{K}(z) = \frac{1}{T} \sum_{t=1}^{T} i[v_s \leq z] \tag{6}$$

where $i(\cdot)$ represents an indicator function, and $v_s = \frac{1}{T-1} \sum_{t=1}^{T} I[x_1^t < x_1^s, x_2^t < x_2^s], s = 1, \ldots, T$.

Frees and Valdez(1998) suggest a Q-Q plot between $K_\phi(z)$ and $\hat{K}(z)$ which will concentrates on the 45^o line if the model fits the data well. We can perform a parametric estimation for each possible family of the Archimedean class and select the most suitable family by employing this method of selection.

3 Application

In order to experiment and evaluate the usefulness of copulas functions, a number of multivariate cases are examined. Specifically, the analytical cases of Gaussianity and lognormality are tested. Further application to a real nickel laterite data set are evaluated. Both the non-parametric estimation approach and the parametric approach using an Archimedean copulas are applied.

Only three of the many possible families for an Archimedean Copulas is evaluated: Clayton, Gumbel and Frank. These tend to be fairly popular families, with widespread applicability. Choosing a suitable model requires selecting the family that yields results closest to the 45^o.

3.1 Analytical Case: Bivariate Gaussian Distribution

Consider the case of a bivariate Gaussian distribution with standard normal marginals and a correlation coefficient of $\rho = 0.7$. This is perhaps the most straightforward case. The analytical form of the multivariate distribution is known and the copulas estimation can be tested against this reference. The analytical form of the joint pdf is given by

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{1-\rho^2}\{\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(y-\mu_Y)^2}{2\sigma_Y^2} - \rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\}\right]$$

which is further simplified in the standard Gaussian case as

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp[-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}]$$

The joint cdf is then given by a numerical integration of the joint pdf:

$$\begin{aligned} F(x_i,y_j) &= \int_{-\infty}^{x_i}\int_{-\infty}^{y_j} f(x,y)dxdy \\ &= \sum_{s=1}^{i}\sum_{t=1}^{j} f(x_s,y_t) \end{aligned}$$

where $i,j = 1\ldots,ndisc$ and $ndisc$ is the number of discretizations of the bivariate space.

Empirical Approach. Inference of a copulas function is tested by first generating a sample of 100 pairs of data from a bivariate standard Gaussian number with correlation $\rho = 0.7$. This data set is used to estimate the joint cdf by constructing the copulas function using Equation 3:

$$\hat{F}(x_i,y_j) = \hat{C}[F_X(x_i), F_Y(y_j)]$$

where $F_X(x)$ and $F_Y(y)$ are the marginal cdf for X and Y, respectively.

Figure 3 shows the results of this comparison. The probability contours of reference and estimated joint distribution are similar; however, a close examination of the difference

in the two joint probabilities show a central region of overestimation by the estimated cdf. The crossplot between these two distributions clearly shows that this overestimation is systematic; the histogram of errors also confirms this observation. Further, the contour of differences shows an apparent banding that is likely due to the discretization of the 2D plane. Figure 4 shows the 3D visualization of the reference and estimated cdf.

Archimedean Copulas. This same analytical case is also evaluated using an Archimedean copulas, testing the Clayton, Gumbel and Frank families. Given the model selection criteria, the Clayton family was chosen with an α of 1.425. Table 1 gives a summary of the results of comparing all three families.

Figure 5 shows the results from an Archimedean copulas. A visual inspection of the 2D plot of the probability contours shows similar contours. Similar to Figure 3, the 2D plot of errors also shows that there is a large region near the middle of the distribution ranges that are consistently overestimated by the copulas function. This is further confirmed by the histogram of errors which clearly shows this bias in overestimation.

There are some differences between the empirical and Archimedean results. Unlike the empirical approach, no banding is apparent using the Archimedean copulas. Note further that the magnitude of the units in the histogram of errors is quite different; the units based on an empirical copulas reflect the accuracy of the estimate, which are directly related to the number of data used in estimation.

3.2 Analytical Case: Bivariate Log-Normal Distribution

Consider now another analytical case of the bivariate lognormal distribution. This case is particularly interesting because most real data distributions appear to be lognormal with long tails for extreme values, yet working with this analytical case permits us to evaluate the accuracy of the copulas funtion estimation.

For this example, consider a bivariate lognormal distribution both marginal distributions have a mean and standard deviation of 4 and 0.5, respectively, and the bivariate correlation coefficient is $\rho = 0.5$. This yields marginal distributions with values that range from 0 (exclusive) to approximately 270. This range of values is then partitioned into 100 classes for each marginal distribution.

Similar to the previous case, we can construct the reference joint distribution simply by using the following analytical relationship for the joint pdf:

$$f(x,y) = \frac{\exp[-\frac{1}{1-\rho^2}\{\frac{(\log x-\mu_{\log X})^2}{2(\sigma_{\log X})^2} + \frac{(\log y-\mu_{\log Y})^2}{2(\sigma_{\log Y})^2} - \rho\frac{(\log x-\mu_{\log X})(\log y-\mu_{\log Y})}{\sigma_{\log X}\sigma_{\log Y}}\}]}{2\pi xy\sigma_{\log X}\sigma_{\log Y}\sqrt{1-\rho^2}}$$

And as before, we can use numerical integration to determine the reference joint cdf over the 100x100 partitioned classes.

Empirical Approach. To begin inference of the joint distribution using a copulas approach, we first require some sample data from this bivariate lognormal distribution (say 100 sampled values using Monte Carlo simulation). Once again, Equation 3 is used to calculate an experimental copulas function to approximate the joint cdf.

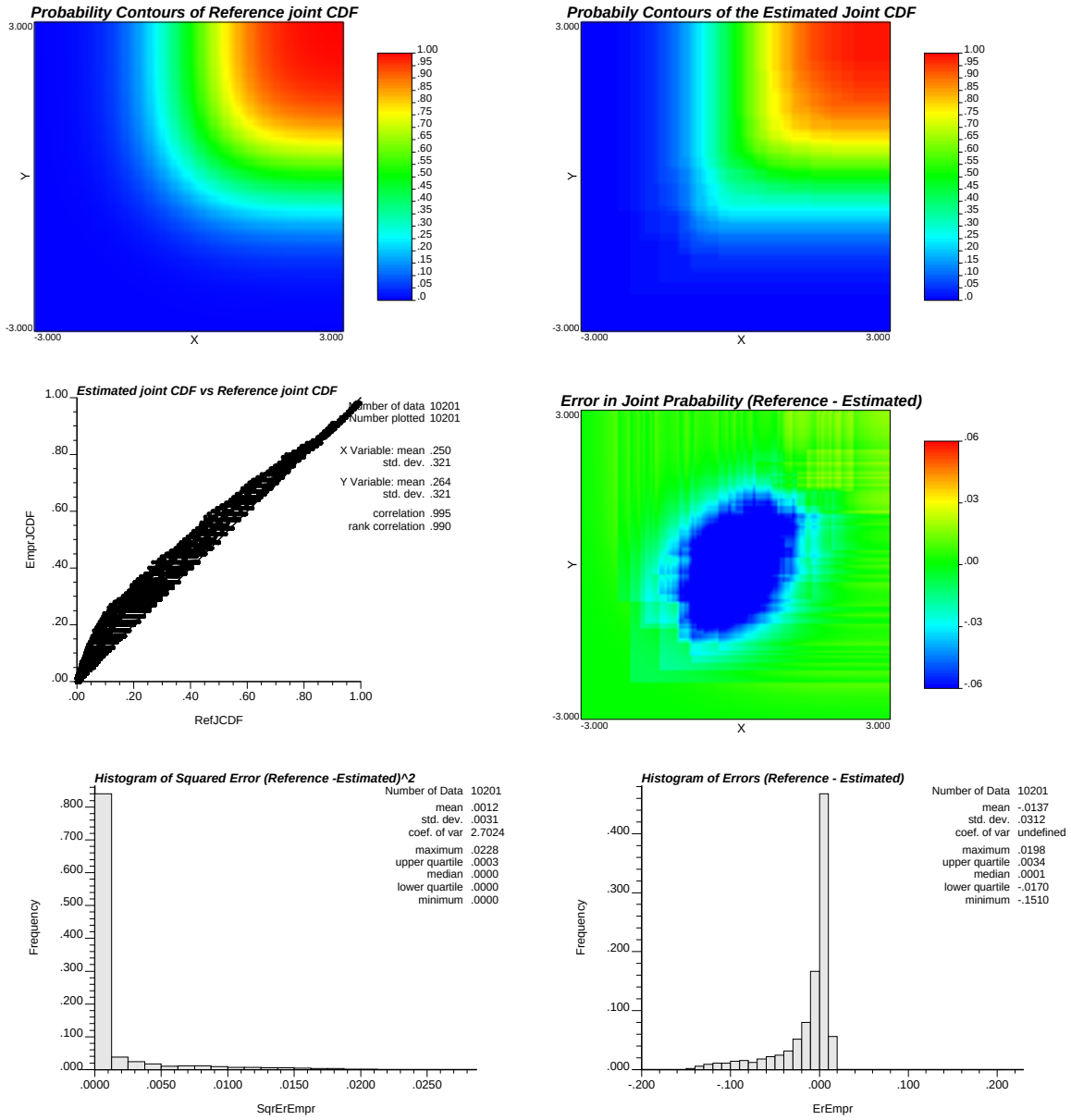

Figure 3: Comparison of reference and estimated joint distribution for a bivariate normal distribution using an empirical copulas: 2D probability contours(top row) of the reference (left) and estimated (right) joint distribution; crossplot of joint distributions and planar distribution of errors (middle row); histogram of squared error and errors (bottom row, left and right, respectively).

Figure 4: 3-dimensional surface plots of reference (left) and estimated (right) joint cdf for bivariate Gaussian case.

Inference of the copulas function yields the comparative plots shown in Figure 6. Similar to the Gaussian case, this 2D field of errors also shows an interesting artefact of the partitioning of the data field. Although in this case, we can see regions of over and under estimation. Overall, the empirical copulas yields an estimated joint cdf that underestimates the reference joint cdf; however, this underestimation does not appear to be systematic like that found in the Gaussian case.

Archimedean Copulas. For this lognormal case, the three Archimedean families were also evaluated, and once again, the Clayton family is chosen, but this time with an alpha of 1.00. Figure 7 shows the performance of the Clayton family in estimating the bivariate frequencies. The 2D distribution of errors shows only two areas of minor under or over-estimation, both in the lower left quadrant. In fact, the match reveals a correlation of 1.00 when the bivariate frequencies are cross plotted. Moreover, the histogram of errors is fairly well balanced about zero error with a significant spike at zero. Overall, the Archimedean copulas shows tremendously encouraging results for the lognormal scenario.

3.3 Real Data: Nickel Laterite Data

Now let's consider a real nickel laterite data set with Nickle (Ni) nd Iron(Fe) with over 9000 pairs of data. The main difference in this case is that we do not have any analytical formula for either the marginal or the joint density functions.

First, we need to determine the marginal distributions for both Ni and Fe, denoted as $\tilde{F}_{Ni}(x_i)$ and $\tilde{F}_{Fe}(y_i)$ respectively, for any given pair of values (x_i, y_i). Here the following formula can be applied:

$$\tilde{F}_{Ni}(x_i) = \frac{1}{T}\sum_{t=1}^{n} i(\tilde{x}_t \leq x_i) \text{ and } \tilde{F}_{Fe}(y_i) = \frac{1}{T}\sum_{t=1}^{n} i(\tilde{y}_t \leq y_i).$$

where T is the total number of observations in our data set, and $(\tilde{x}_i, \tilde{y}_i), i = 1, \ldots, T$ is the set of observed values for (x, y) in the data.

As for the reference joint distribution,$\tilde{F}_{XY}(x_i, y_i)$, we can adopt the following method:

$$\tilde{F}_{XY}(x_i, y_i) = \frac{1}{T}\sum_{t=1}^{T} i(\tilde{x}_t \leq x_i, \tilde{y}_t \leq y_i)$$

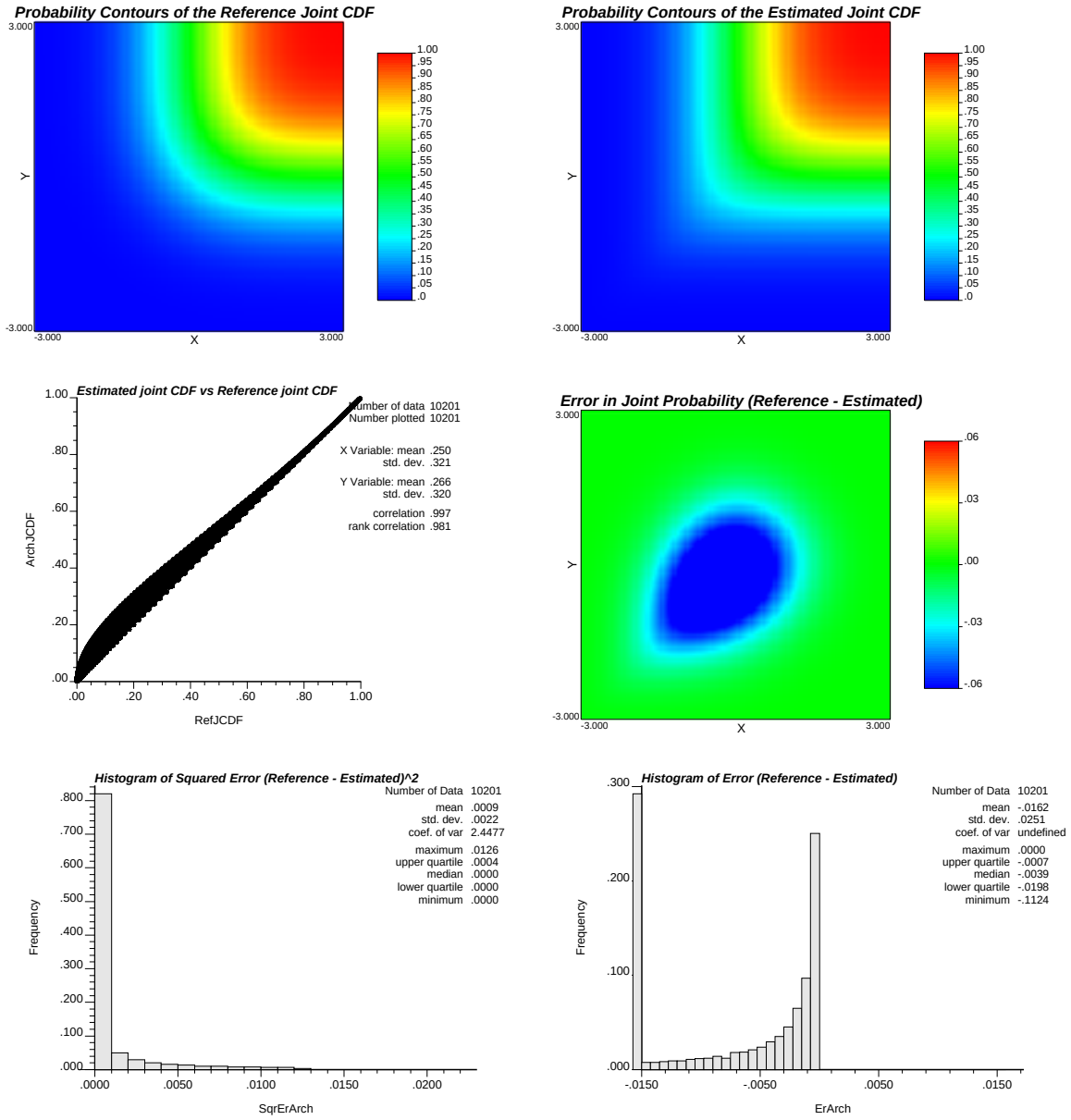

Figure 5: Comparison of reference and estimated joint distribution for a bivariate normal distribution using an Archimedean copulas: 2D probability contours(top row) of the reference (left) and estimated (right) joint distribution; crossplot of joint distributions and planar distribution of errors (middle row); histogram of squared error and errors (bottom row, left and right, respectively).

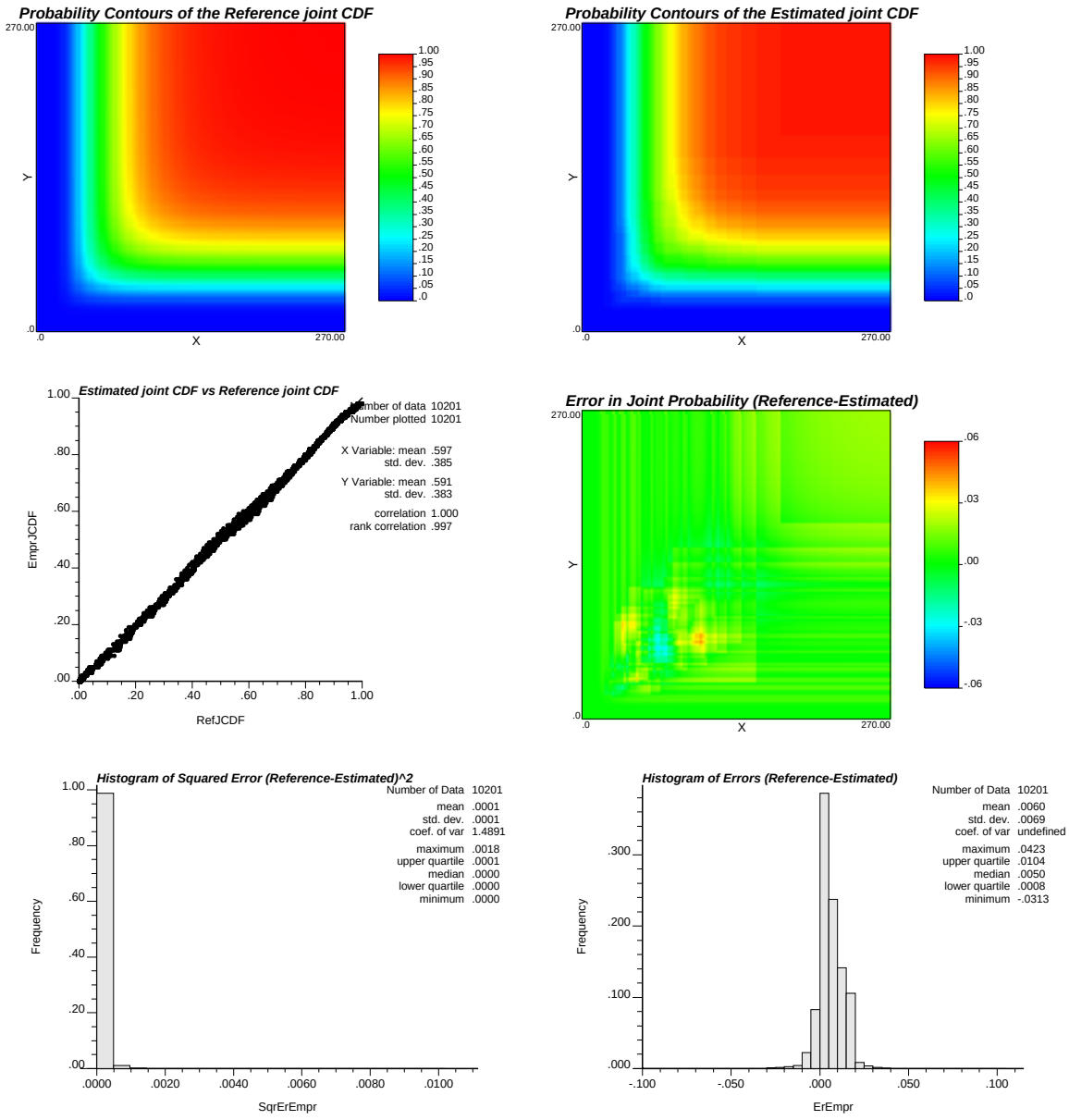

Figure 6: Comparison of reference and estimated joint distribution for a bivariate lognormal distribution using an empirical copulas: 2D probability contours(top row) of the reference (left) and estimated (right) joint distribution; crossplot of joint distributions and planar distribution of errors (middle row); histogram of squared error and errors (bottom row, left and right, respectively).

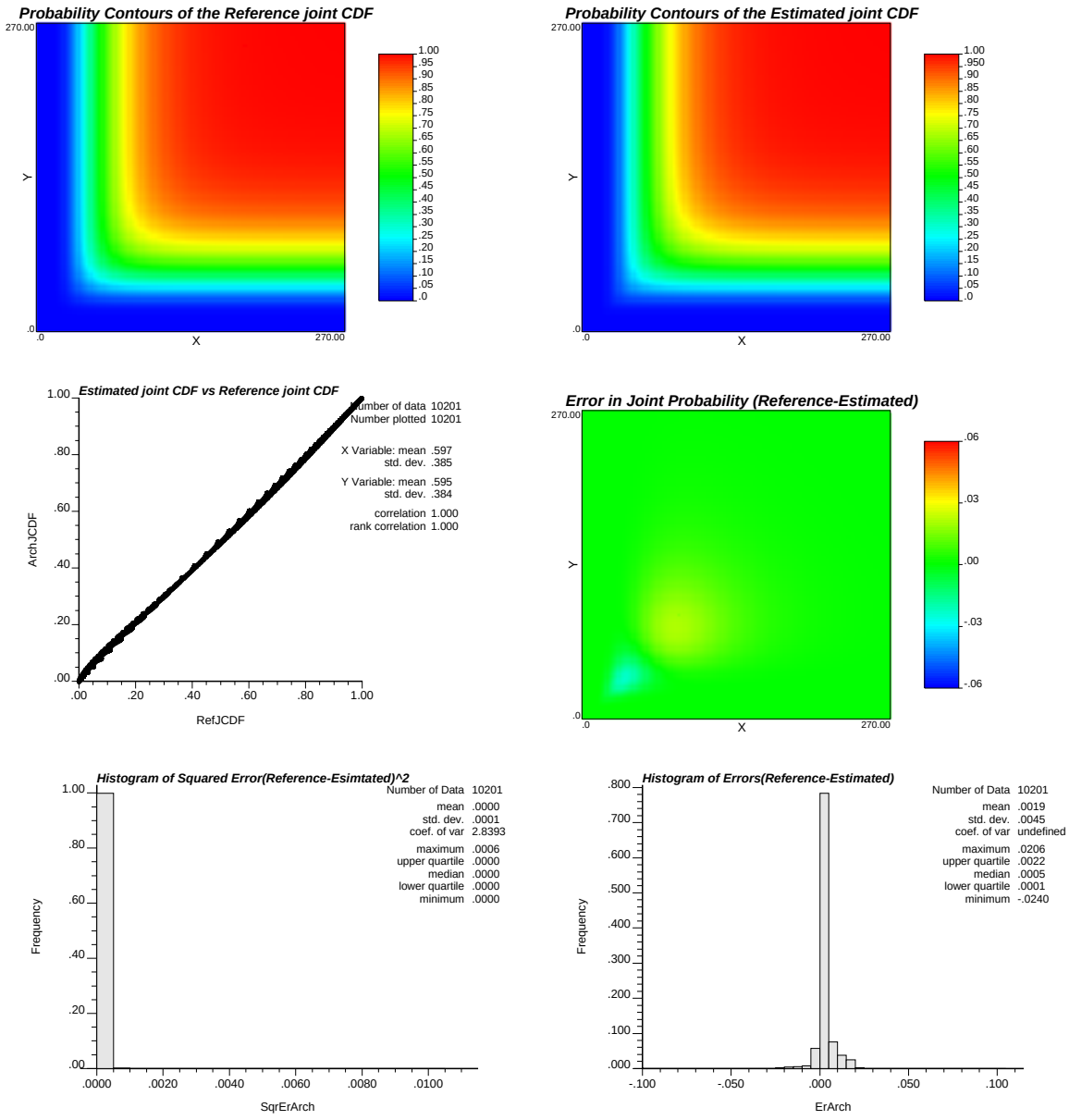

Figure 7: Comparison of reference and estimated joint distribution for a bivariate lognormal distribution using an Archimedean copulas: 2D probability contours(top row) of the reference (left) and estimated (right) joint distribution; crossplot of joint distributions and planar distribution of errors (middle row); histogram of squared error and errors (bottom row, left and right, respectively).

Empirical Approach. For this particular example, we use the full data set to infer the marginal distributions of Ni and Fe since the copulas approach assumes that these are already known. Estimation of the copulas function, however, will be based on a drawing of 100 data pairs of Ni and Fe, as we did for the analytical cases. The data lies within a range $(0, 5.6)$ for Ni and $(0, 55)$ for Fe. Similar to the previous cases, we partition each range of the marginal distributions into 100 classes and evaluate the approximated and reference joint cdf over the 100x100 partitioned classes where the approximated joint cdf, $\hat{F}_{XY}(x_i, y_i)$ is calculated by

$$\hat{F}_{XY}(x_i, y_j) = \hat{C}[\tilde{F}_{Ni}(x_i), \tilde{F}_{Fe}(y_j)]$$

with $\hat{C}(\cdot)$ representing the empirical copulas.

Figure 8 shows the output of our analysis. Again, we see that the cross plot of the reference and the estimated joint cdf form a 45^o line, suggesting almost perfect linear correlation. A comparison of the probability contours also confirm this good correlation; the 2D plot of the differences reveals significant banding due to discretization of the 2D space. A look at the histogram of errors shows that more than 75% of the region is underestimated by the empirical copulas. Despite this, the squared difference between estimated and reference joint cdf are small and that overall, these results suggests that the empirical copulas method works well in the Ni-Fe case.

Archimedean Approach. Evaluation of the three families reveals (yet again) that a choice of the Clayton family with an α of 1.00 yields the smallest sum of squared errors between $K_\phi(z)$ and $\hat{K}_\phi(z)$. Figure 9 shows the output for this Ni-Fe data set. A comparison of the 2D plot of probability contours does not reveal large differences; however, the 2D plot of errors shows that the Archimedean copulas overestimates the reference bivariate frequencies. This overestimation, however, occurs only in a small portion of the field (near the centre of the lower left quadrant). Overall, the crossplot shows very good correlation between the reference and the estimate, and the Archimedean copulas can be said to fit the data well.

4 Discussion

In all three cases, the Archimedean and empirical copulas performed quite well; however, the latter clearly shows a banding effect that is not apparent in the parametric Archimedean approach. Table 1 shows a comparison of the Archimedean and Empirical copulas approach. The quantitative measure used to select a suitable copulas family is the sum of squared errors in the K function (SS_K) as defined by Genest and Rivest (1993). The lower value indicates the best match between the estimated and reference $K_\phi(z)$.

The sum of squared errors (SSE) for the Archimedean and empirical approach shows the performance of each approach in inferring the reference bivariate distribution using the copulas-derived bivariate distribution. In all cases, the Archimedean performed better in matching the reference. This is not entirely unexpected, given that the Archimedean and the data in the first two cases correspond to parametric distributions. In the real data case of Ni-Fe, the banding artefact due to few data pairs to infer the empirical copulas function already makes the Archimedean a favourable alternative.

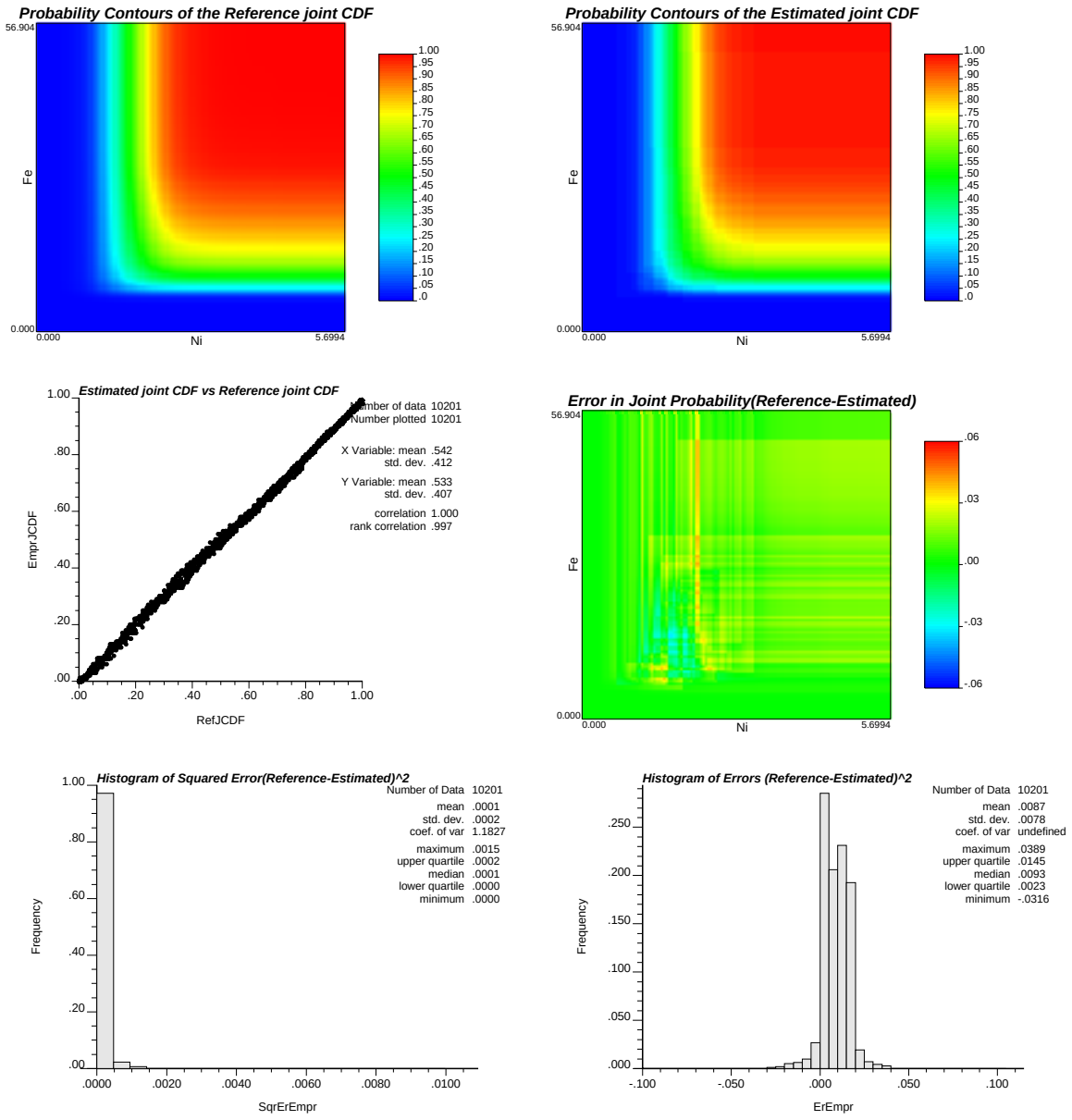

Figure 8: Comparison of reference and estimated joint distribution for the Ni-Fe data set using an empirical copulas: 2D probability contours(top row) of the reference (left) and estimated (right) joint distribution; crossplot of joint distributions and 2D distribution of errors (middle row); histogram of squared error and errors (bottom row, left and right, respectively).

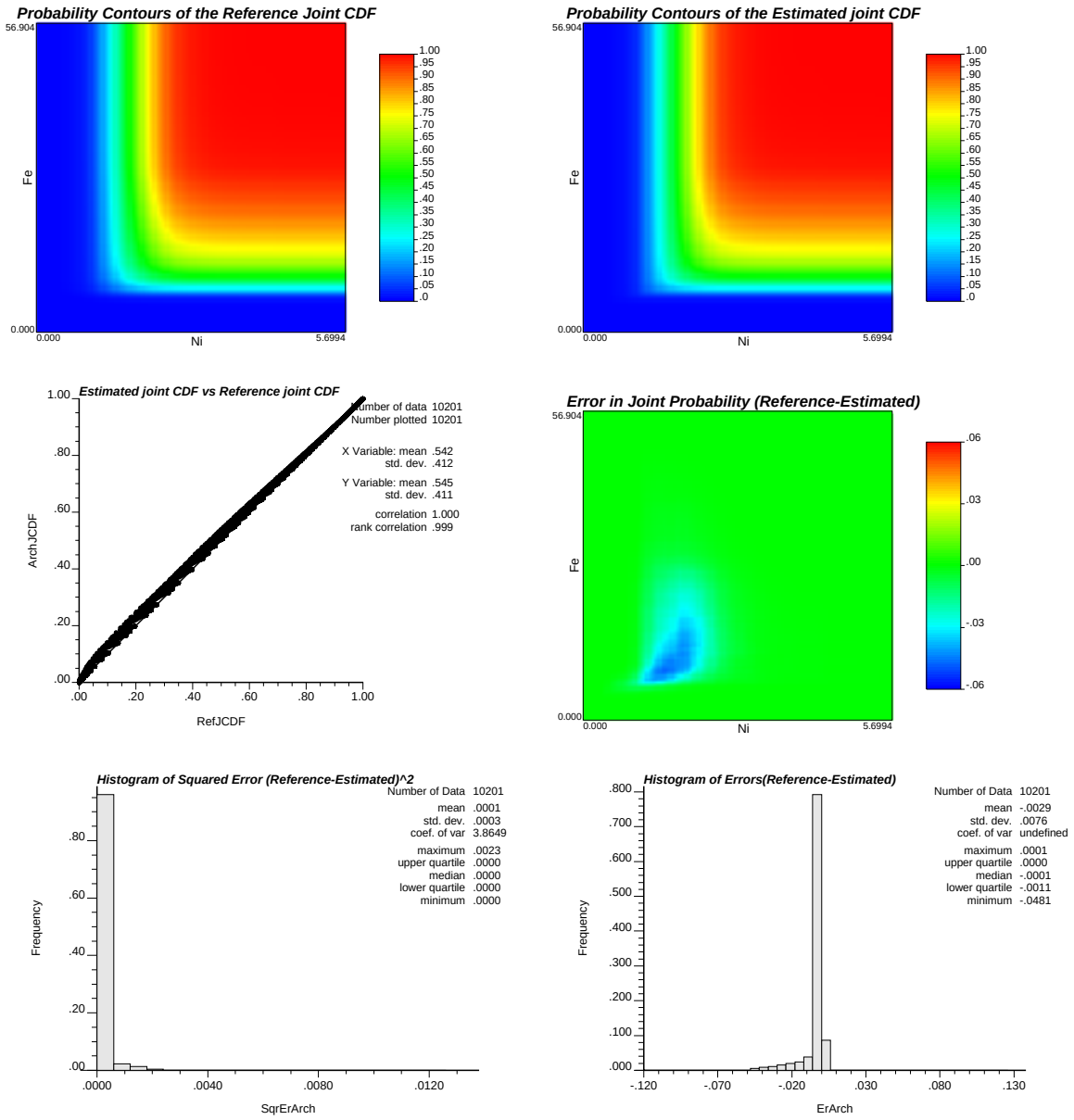

Figure 9: Comparison of reference and estimated joint distribution for Ni-Fe data set under Archimedean copulas: 2D probability contours(top row) of the reference (left) and estimated (right) joint distribution; crossplot of joint distributions and planar distribution of errors (middle row); histogram of squared error and errors (bottom row, left and right, respectively).

Case Study	Archimedean								Empirical
	Clayton		Gumbel		Frank		Choice	SSE	SSE
	α	SS_K	α	SS_K	α	SS_K			
Normal	1.425	0.263	1.0	2.625	5.850	10.007	Clayton	9.103	11.869
logNormal	1.0	0.0999	1.0	1.154	2.663	3.597	Clayton	0.240	0.850
Ni-Fe	1.0	0.185	1.0	0.444	0.9999	0.964	Clayton	0.6666	1.519

Table 1: Summary of results for Archimedean copulas and Empirical copulas. Note that SSE is the sum of squared errors between the reference and estimated distributions, for the Empirical copula and the chosen Archimedean Copulas; and $SS_K = \sum(K_\phi(z) - \hat{K}_\phi(z))^2$.

This required sample size for reliable inference of the empirical copulas function is an issue. The larger the sample size, such as 50 or 100 data pairs or more, the better the method works. However, large sample sizes may not be practical for real data and may cause accuracy issues in the estimation. The sample size is directly related to the number of quantiles that can be estimated; the fewer the samples, the more coarse the estimation of the distribution function. For example, a sample size of 10 data pairs will only permit estimation of the deciles of the joint cdf, while a sample size of 100 data pairs will permit estimation of the percentiles of the joint cdf. Figure 10 shows the effect of sample size on the accuracy of the empirical copulas in inferring the joint cdf. Despite the high correlations in all cases, it is clear that a larger sample size yields more accurate results.

5 Proposed Application in Direct Sequential Cosimulation

While the objective of this paper is to understand the construction and performance of copulas functions in multivariate distribution inference, the ultimate goal remains a multivariate direct sequential simulation methodology. In the previous sections, we have seen that a copulas function requires us to know the univariate marginal distributions corresponding to the multivariate distribution we are trying to infer. Several authors have already tackled the problem of univariate distribution inference [13, 20]. We can now apply copulas functions to take these results one step further and determine the conditional multivariate distribution. This presents an alternative to the multivariate scaling approach proposed in CCG Report 4 [7]. The implementation in a direct sequential cosimulation approach consists of:

1. Pick a random path visiting all locations.
2. At each location:
   (a) Search for all nearby data of different types and/or scale and previously simulated nodes.
   (b) Perform simultaneous cokriging (collocated or full) to determine the parameters corresponding to the conditional univariate distribution for each variable.
   (c) Using the cokriged parameters, determine the conditional univariate distribution for each variable using the approach proposed by Oz et. al. [13].

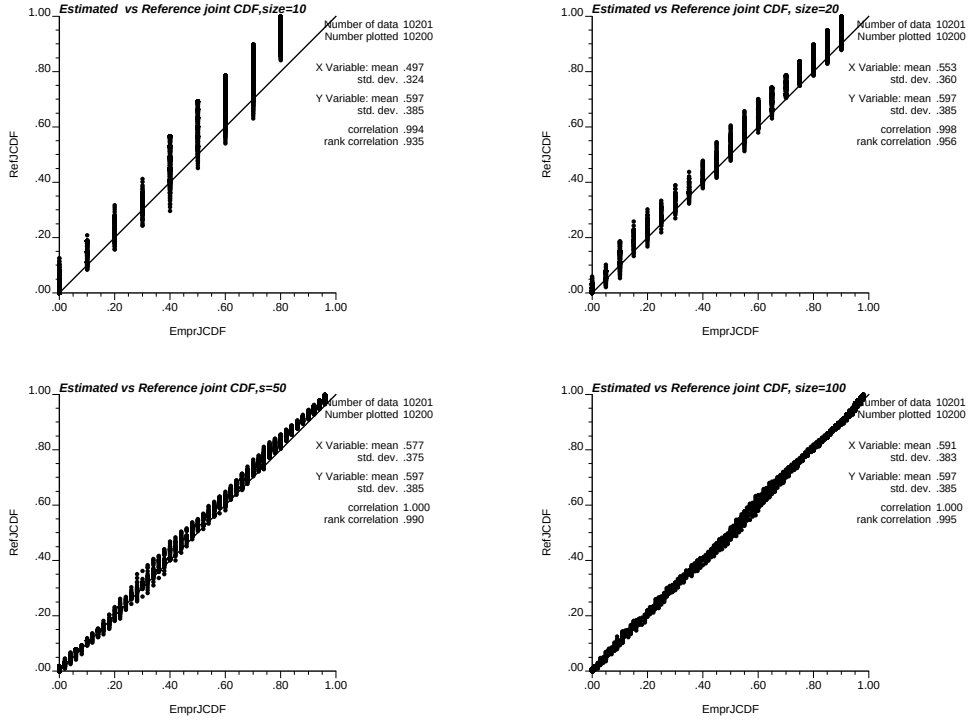

Figure 10: Accuracy of empirical copulas function in estimating joint cdf using different sample sizes: 10 (top left), 20 (top right), 50 (bottom left) and 100 (bottom right) data pairs.

(d) Determine a non-standard multivariate distribution via copulas functions, using both the empirical and Archimedean copulas function, and select the appropriate function based on SSE.

(e) Draw from the multivariate cdf in a stepwise manner:

i. Draw a simulated value y_1 from the conditional marginal distribution of $Y_1(y_1)$.

ii. From the conditional multivariate distribution determined in Step 2d, determine the conditional univariate distribution of $Y_2(y_2)$ given $Y_1 = y_1$, $f_{Y_2|Y_1} = y_1$. Draw y_2 from this conditional marginal distribution.

iii. If $p > 2$, then repeat repeat previous step with successively higher conditioning.

(f) Proceed to next node.

6 Future Work

Copulas theory presents a promising area of research for multivariate geostatistical simulation. The results presented in this report are interesting and show that copulas functions are quite flexible with both the parametric and non-parametric data. Further, implementation of the various models provides a subroutine for a future direct sequential cosimulation `dscosim` algorithm. This `dscosim` program will be developed to permit the choice of a

multivariate scaling or a copulas function approach for multivariate distribution inference. This will permit further testing of both approaches, and the robustness of each method can then be evaluated and compared.

Acknowledgements

The authors would like to acknowledge the support of the sponsor companies of the Centre for Computational Geostatistics.

References

[1] L. Brieman and J. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

[2] P. Deheuvels. Caractérisation complète des lois extrèmes multivariées et de la convergence des types extrèmes. *Pub. l'Institut de Statist. l'Université de Paris*, 3(23):1–36, 1978.

[3] P. Deheuvels. La function de dépendance empirique et ses propriétés-un test non paramétrique d'indépendance. *Bulletin de la Classes des Sciences*, 65(5):274–292, 1979.

[4] C. Genest and L.Rivest. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043, 1993.

[5] J. R. Lallena. A new class of bivariate copulas. *Statistics and Probability Letters*, 66(3):315–325, 2004.

[6] O. Leuangthong. *Stepwise Conditional Transformation for Multivariate Geostatistical Simulation*. PhD thesis, University of Alberta, Edmonton, AB, 2003.

[7] O. Leuangthong and C. Deutsch. Modeling multivariate multiscale data. Technical report, Centre for Computational Geostatistics, University of Alberta, Edmonton, AB, March 2002.

[8] A. Marshall and I. Olkin. A generalized bivariate exponential distribution. *Journal of applied probability*, 4(2):291–302, 1967.

[9] A. Marshall and I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 62(317):34–44, 1967.

[10] A. Marshall and I. Olkin. Families of multivariate distributions. *Journal of the American Statistical Association*, 83(403):834–841, 1989.

[11] D. Moore and M. Spruill. Unified large-sample theory of general chi-squared statistics for test of fit. *The annals of statistics*, 3(3):599–616, 1957.

[12] R. Nelsen. *An Introduction to Copulas*. Springer Verlag, New York, 1999.

[13] B. Oz, C. Deutsch, T. Tran, and Y. Xie. A fortran 90 program for direct sequential simulation with histogram reproduction. *Computers & Geosciences*, 29(1):39–51, 2003.

[14] M. Pelagatti and S. Rondena. Dynamic conditional correlation with elliptical distributions. In I. of Mathematical Statistics, editor, *Proceeding of the 2nd OxMetrics User Conference*, page 11. Cass Business School, London, 2004.

[15] A. M. R. Frey and M. Nyfeler. Copulas and credit models. *(Working Paper, Department of Mathematics, ETHZ, Zurich)*, page 8, 2001.

[16] C. Romano. Calibrating and simulating copula function: An application to the italian stock market. *working paper, Centro Interdipartmentale sul diritto e l'Economia dei Mercati*, 2002.

[17] B. Schweizer and A. Sklar. Operations on distribution fuctions not derivable from operations on random variables. *Studia Math.*, 52(1):43–52, 2002.

[18] A. Sklar. Functions de répartition a n dimensions et leurs marges. Technical Report 8, Inst. Statist.Univ.Parise.Publ., 1959.

[19] A. Sklar. Random variables, distributions, and copulas – a personal look backward and forward. In I. of Mathematical Statistics, editor, *Distribution with Fixed Marginals and Related Topics*, pages 1–14. Kluwer Academic Publishers, 1996.

[20] A. Soares. Sequential direct simulation and co-simulation. *Mathematical Geology*, 33(8):911–926, 2001.

[21] E. L. U. Cherubini and W. Vecchiato. *Copula Methods in Finance*. John Wiley & Sons Inc., New York, 2004.

[22] E. Valdez and E. Frees. Understanding relationships using copulas. *North American Acturial Journal*, 2(1):1–25, 1998.

Appendix A: Notation

The following notation is adopted for this paper:

- cdf, F: cumulative distribution function,
- jcdf, $F_{X_1}, \ldots, F_{Xn}$: joint comulative distribution function, in bivariate or multivariate distribution
- F_{X_i}: marginal CDF for the i^{th} variable in bivariate or multivariate distribution
- pdf, f: probability density function
- jpdf, f: joint probability density function, in bivariate or multivariate distribution
- f_{X_i}: marginal PDF for the i^{th} variable in bivariate or multivariate distribution
- $Range(G)$: the range of a function G
- RV: random variable
- X (upper case letter, except otherwise defined): notation for a random variable or a function
- x (lowercase letter, except otherwise defined): a real number, a particular outcome of RV X
- $\mathbf{I}$: interval $[0, 1]$
- $\mathbf{R}$: real number, interval $[-\infty, +\infty]$
- $\mathbf{u}$, $\vec{\alpha}$ (bold lowercase letter or arrow over a letter): notations for a vector
- $\mathbf{X}$ (bold uppercase letter, except otherwise defined): notation of a matrix
- C: notation for a copula function
- c: notation for the density function of a copula

Appendix B: Program Code Description

All the numeric calculations in this paper are carried out by a fortran program called `copulas`. In this program, we allow two options: 1) Approximating the joint cdf value for a given pair of (x, y); and 2) Carry out a bin test, in which the given intervals of variable X, Y are separated into a given number of sub-intervals and approximate the joint cdf value for each possible pair of separating point (x, y). All the above approximations are based on a given set of sample values $(\tilde{x}_i, \tilde{y}_i)$.

In the current program, we can compare the approximated joint cdf (using Empirical Copula or Archimedean copula) with three types of reference distributions:

- Bivariate Gaussian Distribution;
- Bivariate LogNormal Distribution;
- Non Parametric Sample Distribution with the joint CDF $\tilde{F}(x, y)$ defined as

$$\tilde{F}(x, y) = \frac{1}{T} \sum_{i=1}^{T} i[\tilde{x}_i \leq x, \tilde{y}_i \leq y]$$

  where T is the number of observation.

Similar to GSLIB, this program requires a parameter file (see Figure 13) which requires the following information:

- Name of the sample data file
- Name of the output data file
- Parameters for either the bivariate Gaussian or bivariate log-Normal distribution
- Interval and number of subintervals in the bin test
- Model selection choice for the Archimedean family: automatic determination or user specified model

The flowcharts in Figures 11 and 12 provides an overview of the detailed steps of the program.

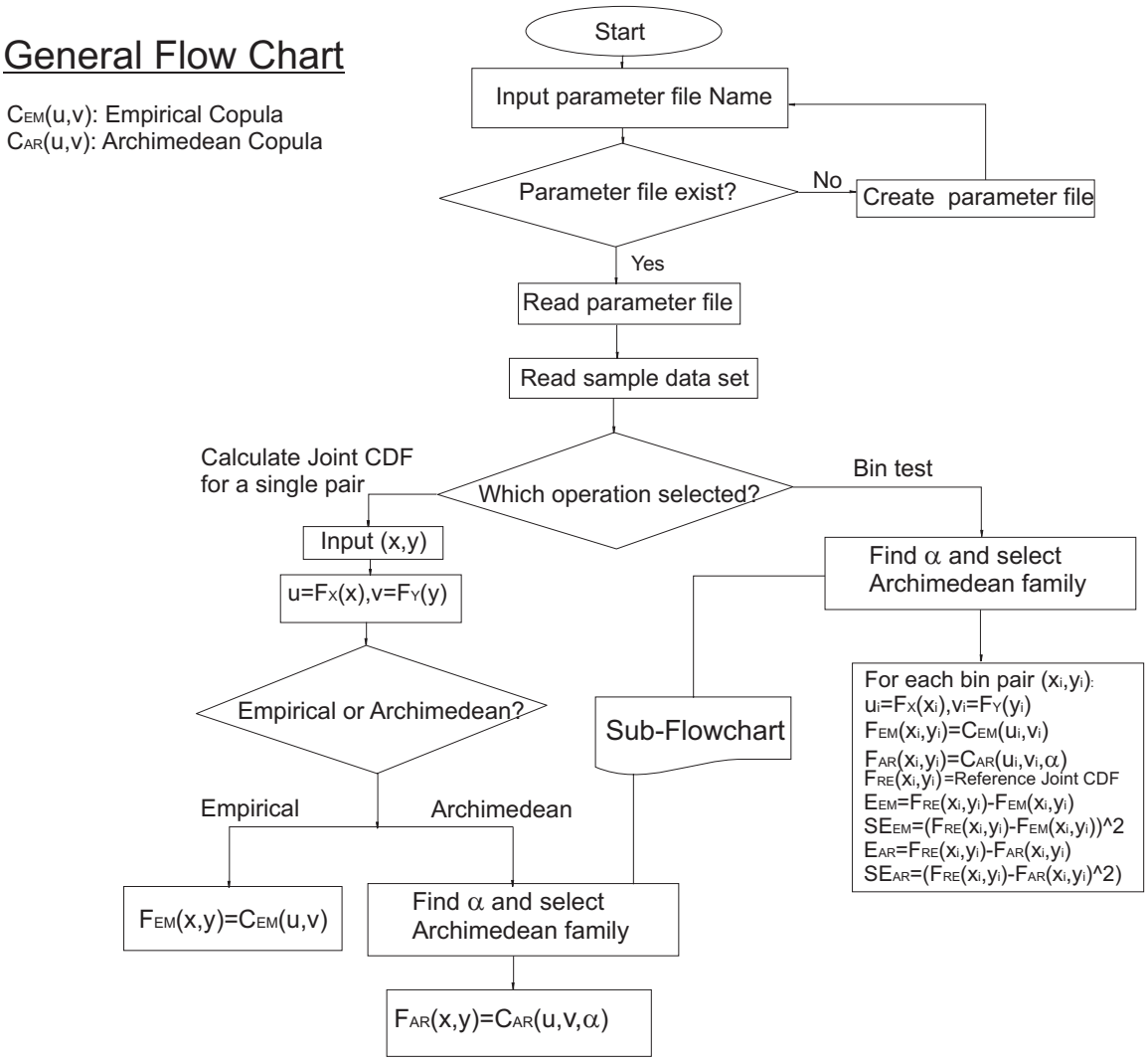

Figure 11: General flowchart for the copula program.

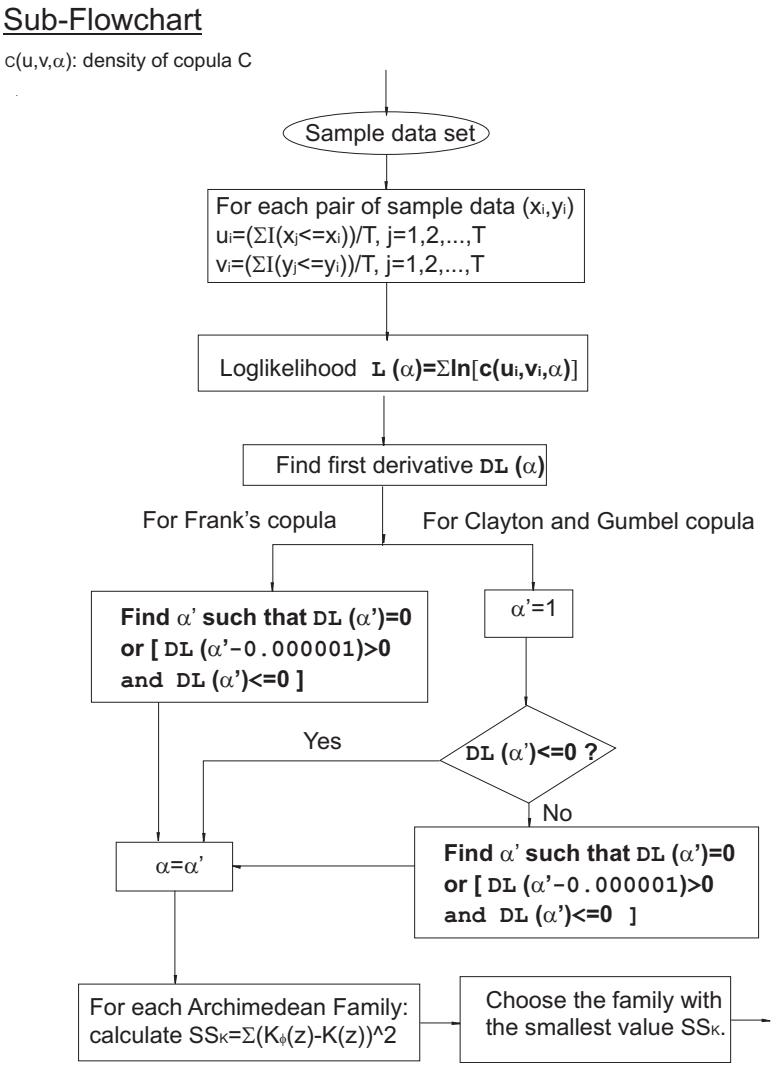

Figure 12: Sub-flowchart for Archimedean Model construction and selection.

```
                Parameters for COPULAS
                **********************

START OF PARAMETERS:
binorm.txt              - data file 1
2                       - columns for variable 1 and 2
1                       - type of ref dist (1=non-par,2=biGauss,3=bi logN)
copulas.out             - file for bin test output

PARAMETERS FOR REFERENCE DISTRIBUTION (biGauss and Bi logN):
0.7                     - correlation coefficient 0
1.0                     - mean, std dev of Var 1 0
1.0                     - mean, std dev of Var 2

BIN TEST:
-3.0  3.0               - min and max of Var 1
-3.0  3.0               - min and max of Var 2
100                     - number of subintervals

ARCHIMEDEAN FAMILIES:
1                       - Model (0=auto,1=Clayton,2=Gumbel,3=Frank)
```

Figure 13: Parameters for `copulas`.