

A Simplified, General Approach to Simulating from Multivariate Copula Functions *

Barry K. Goodwin

June 4, 2012

Abstract

Copulas have become an important analytic tool for characterizing multivariate distributions and dependence. One is often interested in simulating data from copula estimates. The process can be analytically and computationally complex and usually involves steps that are unique to a given parametric copula. We describe an alternative approach that uses “probability-proportional-to-size” (PPS) random sampling with weights formed from the copula likelihood function. The method is flexible and can be applied to parametric and nonparametric marginal density estimates. The precision of the simulation can be calibrated by adjusting the density of the multidimensional grid used in the simulation process. The approach is fully transparent to any copula function with continuous random variables. An example evaluates a number of goodness-of-fit criteria and provides strong support for the validity and practicality of the method.

*This work was supported by the U.S. Forest Service and the North Carolina Agricultural Research Service. The helpful comments of Andrew Patton are gratefully acknowledged. Goodwin is William Neal Reynolds Distinguished Professor in the Departments of Economics and Agricultural and Resource Economics at North Carolina State University. Direct correspondence to Box 8109, North Carolina State University, Raleigh, NC, 27695, E-mail: barry_goodwin@ncsu.edu.

A Simplified, General Approach to Simulating from Multivariate Copula Functions

Copulas are an important tool for modeling relationships among (possibly) dependent random variables. A copula relates marginal distributions for individual variables in a representation of the joint distribution function. Sklar (1959) showed that, for any group of continuous random variables, the joint distribution function for the variables can be uniquely represented in terms of the marginals and a copula function. Specifically, Sklar's Theorem implies that, for any continuous p -variate cumulative probability function F , a unique copula function $C(\cdot)$ exists for which

$$F(x_1, x_2, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p); \xi) \quad (1)$$

where $F_i(\cdot)$ are marginal distributions and ξ is a set of parameters that measures dependence. This implies that the joint density function can be written as

$$f(x_1, x_2, \dots, x_p) = c(f_1(x_1), \dots, f_p(x_p)) \cdot \prod_{i=1}^p f_i(x_i) \quad (2)$$

where $c(\cdot)$ is given by $\partial C(\cdot)/\partial x$ and the f_i terms are marginal density functions.

Copula models have recently realized widespread application in empirical models of joint probability distributions.¹ The models essentially use a “copula” function to tie together two marginal probability functions that may (or may not) be related to one another. Much of the work on copulas has been motivated by their applicability to the issues in risk management, insurance and financial economics (see, among others, Rodriguez (2003), Cherubini et al. (2004), Hu (2006), Patton (2006), and Jondeau and Rockinger (2006)). In the empirical literature, copula models have been used extensively in the design and rating of crop revenue insurance contracts, where the inverse correlation of prices and yields plays an important role in pricing revenue risk.

A p -dimensional copula, $C(u_1, u_2, \dots, u_p)$, is a multivariate distribution function in the unit hypercube $[0, 1]^p$ with uniform $U(0, 1)$ marginal distributions. As long as the marginal

¹For details on construction and properties of copulas, see among others, Joe (1997) and Nelsen (2006).

distributions are continuous, a unique copula is associated with the joint distribution, F , that can be obtained as:

$$C(u_1, u_2, \dots, u_p) = F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p)). \quad (3)$$

In a similar fashion, given a p -dimensional copula, $C(u_1, \dots, u_p)$, and p univariate distributions, $F_1(x_1), \dots, F_p(x_p)$, the equation 1 is a p -variate distribution function with marginals $F_1, \dots, F_p$ whose corresponding density function can be written as:

$$f(x_1, x_2, \dots, x_p) = c(F_1(x_1), \dots, F_p(x_p)) \prod_{i=1}^p f_i(x_i) \quad (4)$$

Provided that it exists, the density function of the copula, c , can be derived using equation 3 and marginal density functions, f_i :

$$c(u_1, u_2, \dots, u_p) = \frac{f(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p))}{\prod_{i=1}^p f_i(F_i^{-1}(u_i))}. \quad (5)$$

There is a large number of parametric families of copulas applied in the literature. Two of the most commonly used copula families are elliptical copulas and Archimedean copulas. Gaussian and t -copulas are examples of elliptical copulas while the Clayton and Gumbel are among Archimedean copulas.

Copulas are generally used to represent joint probability distributions. This joint distribution may be based upon different parametric families for the marginals. The individual marginals may also be nonparametric and thus represented using empirical distribution functions. Upon estimation of the relevant copula function, one is typically interested in simulating or predicting dependent random variables from a joint probability distribution. To this end, the marginal distribution functions and the copula representation of the joint likelihood function must be inverted.

These methods make use of the fact that the inverse of the cumulative distribution function (cdf) for any random variable is uniformly distributed. Nelsen (2006) refers to methods that use this to simulate from a distribution as the “inverse distribution method.” Making use of the fact that $F^{-1}(x)$ is a uniform random variable u allows one to generate

uniform draws and plug these draws into the inverse cdf, such that $x = F^{-1}(u)$ yields a random draw x from $F(\cdot)$.

Extending this “inverse distribution method” to the case of two or more random variables becomes more complex in that the the inverse of the copula density must be evaluated conditional on values for all other uniform variates using the inverse copula density function. For example, consider a case of generating a single draw from a joint distribution for two random variables. This is accomplished by first drawing a pair of uncorrelated uniform distributed random variables (u_1, u_2) . The first uniform deviate is used to generate a random draw for $y_1 = F^{-1}(u_1)$. The second random variable will then be generated conditional on u_1 by plugging the values of u_1 and u_2 into the copula density function as follows

$$y_2 = c^{-1}(u_2|u_1). \quad (6)$$

Thus, an invertible form of the copula must be readily available in order to adequately simulate the dependent random variables. Further, the marginals must be fully specified in a parametric form in order to simulate the dependent random variables. Bouyé et al. (2000) describe this general iterative conditioning approach for simulating correlated random variables from copula functions.

In some special cases for certain copula functions, alternative algorithms that offer greater speed and efficiency can be applied in simulating random variables from copulas. Marshall and Okin (1988) and Frees and Valdez (1998) note that when the inverse of the generator function associated with an Archimedean copula is the Laplace transform of some other positive random variable from a known distribution, a simple and direct method of simulation exists by using simulations from the corresponding distribution. Yan and Kojadinovic (2012) note that the corresponding distributions for the Clayton, Frank, and Gumbel copulas are the gamma, log-series, and positively stable distributions, respectively. In spite of the simplicity and efficiency of this approach, it is limited by the fact that it only exists for a subset of Archimedean copulas for which the corresponding distributions are known.

This note suggests an alternative approach to simulation that may offer advantages in two important dimensions. First, one can work directly with the joint likelihood function implied

by the copula rather than the requiring conditional versions of the inverse of the copula density function. In addition, one can simulate directly from nonparametric marginals, thereby allowing simulation of dependent random variables without requiring one to specify parametric marginal distributions. The approach essentially uses values of the joint density function to form likelihood weights that can then be used in a probability-weighted sampling scheme to simulate dependent draws from the nonparametric densities.

The approach uses values of the joint likelihood function given by equation 5 to assign a probability weight associated with any relevant combination of random variables $(x_1, \dots, x_p)$. Values for the marginal densities $f_i(\cdot)$ for each observation can either be evaluated using the relevant parametric family (if known) or the value of a nonparametric density. The nonparametric marginals can be estimated using sample data and then evaluated over a grid that can be defined (between the maximum and minimum values of the observed range of data) to be as fine as needed to obtain a desired level of precision. In the case of nonparametric marginals, the simulation is limited to lie only within the range of observed data. However, in the case of parametric marginals, the simulation can encompass the entire domain of the parametric families. There are a number of methods available for probability-weighted random sampling (or as it is equivalently termed, “probability proportional to size” or PPS random sampling). We use the methods for PPS with replacement random sampling that are described in the classic reference text of Cochran (1977).

The approach can be summarized as follows. First, parameters characterizing the marginal distributions must be estimated. In the case of nonparametric marginals, kernel density estimation techniques (or equivalent nonparametric density estimation methods) can be used as an alternative. A multi-dimensional grid can then be generated and for each combination of simulated random variables the likelihood function given by equation 5 can be evaluated. The grid can be made as dense as desired in order to ensure adequate convergence of the simulation to the copula and marginals estimated for the sample data. Note that the exercise, though simplistic in design, can be computationally intensive as the dimension of the copula increases. In particular, if one generates a grid of size N for a p -dimensional copula,

the likelihood function must be evaluated N^p times. The resulting simulated data set from which the PPS random sampling will be applied to can be very large but is manageable given current computing capabilities. Once the p -dimensional grid of observations is generated, a sample of simulated data points is generated using the PPS with replacement random sampling methods. This is the most computationally intensive step of the procedure. The size of the simulated data set can be adjusted to the requirements of the problem, thereby reducing computational resources.

An Example

A simple example is considered in order to demonstrate the sampling algorithm. Simulating from highly skewed or multi-modal distributions is more complex in that the number of simulated data points needed to adequately capture thin tail behavior may be large. Likewise, the shape of such distributions may be sensitive to parametric estimates of the parameters characterizing the distributions. For example, small variations in the shape and scale parameter of a log-normal density may provoke substantial differences in the nature of the density. We chose a relatively platykurtic 2-dimensional copula model—a t-copula with a dependence parameter (ρ) of 0.50 and a degrees-of-freedom parameter (η) of 5.0. We chose a log-normal distribution with shape and scale parameters of 1.0 and 0.5 for one marginal, which we term the ‘X’ variable. For the other ‘Y’ marginal, we chose a skewed Weibull distribution with shape and scale parameters of 5.0 and 150. We generated a set of 1,000 observations, which we refer to as the “generated” sample by simulating from these marginals and copula. We then estimated the copula model using standard maximum likelihood methods. We used nonparametric kernel density estimates to model the marginal densities, with bandwidth parameters chosen using Silverman’s rule of thumb. We then applied the method described above to generate a simulated sample from the estimates, which we term the “simulated” sample.²

²Estimation and inferences were accomplished using the “pps” “copula” and “CDVine” packages of the *R* language. Details are available in Gambino (2012), Schepsmeier and Brechman (2012), and Yan and Kojadinovic (2012). Excellent overviews of the *R* copula packages and implementation issues are presented by Yan (2007) and Czado (2011).

Figure 1 presents kernel density estimates for the generated and simulated marginals. We then conducted a number of illustrative and inferential evaluations intended to validate the simulation procedure. Note that the goal of each step of this evaluation is to compare the results for the “simulated” data to those from the “generated” sample of data. To this end, estimates of evaluation criteria are presented side-by-side. In the case of the marginals, the estimates are very similar and indicate that the proposed method provides an adequate representation of the estimated densities based upon the generated data set. Figure 2 presents *qq*-plots of the marginal as compared to the true parameter values. Note that the log-normal data show some departures in the right-tail, reflecting the substantial skewness of the distribution. However, the important fact is that the plots are very similar for the generated and simulated data. Thus, the evidence again supports the validity of the proposed method for simulating data from a copula.

Figure 3 presents plots of Kendall’s K statistics, which is comparable to the *qq*-plots for a copula model. The figures represent the relatively strong degree of positive dependence inherent in the copula estimates.³ Again, the diagrams are very comparable and indicate that important patterns of dependence are replicated in the simulated data. Figure 4 presents contours of the estimated copula functions for the generated and simulated data. Once again, the implied multivariate distributions are similar.

Table 1 presents copula parameter estimates and a variety of goodness-of-fit statistics and other relevant criteria that compare results for the generated and simulated samples. At the top, estimates of the copula parameters are presented and confirm the similarities of the multivariate distributions. Summary statistics and Kolmogorov-Smirnov tests for the marginal distributions are very similar and confirm the validity of the method. Note that, by design, the minimum and maximum values observed in the data define the span of the simulated data. This is not guaranteed by the PPS sampling but is likely to be observed since the grid is defined by the minimum and maximum values. Standard measures of correlation are also very similar for the generated and simulated samples. Finally, the bootstrapped

³Values above the diagonal correspond to positive dependence. The outer envelope represents perfect correlation.

Cramér-von-Mises and Kolmogorov-Smirnov goodness-of-fit statistics of Genest et al. (2009) for the copulas are similar and support the t-copula specification in both cases.

Concluding Remarks

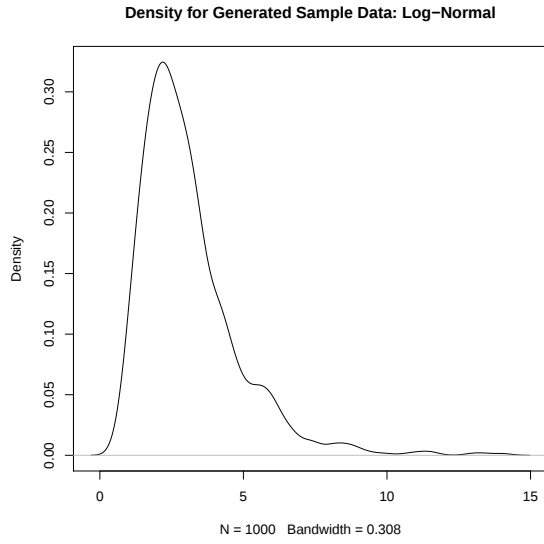
Copulas have become an important analytic tool for characterizing multivariate distributions and dependence. One is typically interested in simulating data from copula estimates. The process can be analytically and computationally complex and usually involves steps that are unique to a given parametric copula. We describe an alternative approach that uses “probability-proportional-to-size” (PPS) random sampling with weights formed from the copula likelihood function. The method is flexible and can be applied to parametric and nonparametric marginal density estimates. The precision of the simulation can be calibrated by adjusting the density of the multidimensional grid used in the simulation process. The approach is fully transparent to any copula function with continuous random variables. An example evaluates a number of goodness-of-fit criteria and provides strong support for the validity and practicality of the method.

References

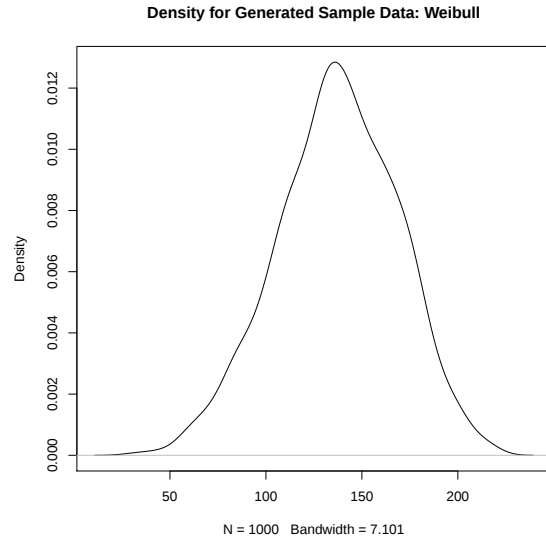
- Cherubini, U., Luciano, E. and Vecchiato, W. (2004), *Copula Methods in Finance*. John Wiley and Sons, Chichester.
- Gambino, J. (2012) “Users guide to R functions for PPS sampling,” unpublished manuscript provided with the R pps package documentation.
- Cochran, W.G. (1977). *Sampling Techniques*. John Wiley and Sons, New York.
- Czado, C. (2011) “The World of Vines,” Technische Universität München.
- Frees E.W. and E. A. Valdez (1998). “Understanding Relationships Using Copulas.” *North American Actuarial Journal*, 2(1), 125.
- Genest, C., Rémillard, R. and Beaudoin, D. (2009), “Goodness of Fit Tests for Copulas: A Review and Power Study.” *Insurance: Mathematics and Economics*, 44, 199–213.
- Joe, H. (1997), *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.
- Marshall A.W. and I. Olkin (1988). “Families of multivariate distributions.” *Journal of the American Statistical Association*, 83, 834–841.
- Nelsen, R.B. (2006), *An Introduction to Copulas*. Springer-Verlag, New York .
- Schepsmeier, U. and E. C. Brechmann. (2012) “Package CDVine: Statistical inference of C- and D-vine Copulas,” *R-Project CRAN Repository*, February, 2012.
- Yan, J. (2007) “Enjoy the Joy of Copulas: With a Package copula,” *Journal of Statistical Software*, Vol. 21, Issue 4.
- Yan, J. and I. Kojadinovic (2012). “Package ‘copula’: Multivariate Dependence With Copulas,” *R-Project CRAN Repository*, February, 2012.

Table 1: Summary Statistics for Generated Sample and Simulated Data

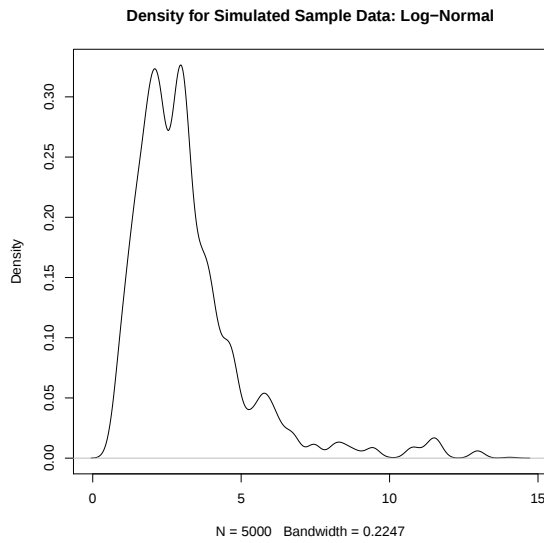
Parameter/Statistic	Generated	Simulated
ρ Estimate	0.4958	0.5086
ρ Standard Error	0.0253	0.0114
η Estimate	5.6275	4.6871
η Standard Error	1.1771	0.3864
Cramér von-Mises p-Value	0.90	0.33
Kolmogorov-Smirnov p-Value	0.81	0.13
X-Variable (Log-Normal)		
Kolmogorov-Smirnov test	0.0402	0.0406
Kolmogorov-Smirnov test (p-Value)	0.3930	0.3834
Minimum	0.62	0.62
1 st Quartile	1.94	1.95
Median	2.71	2.81
Mean	3.09	3.20
3 rd Quartile	3.76	3.79
Maximum	14.05	14.05
Y-Variable (Weibull)		
Kolmogorov-Smirnov test	0.0462	0.0446
Kolmogorov-Smirnov test (p-Value)	0.2359	0.2739
Minimum	31.95	31.95
1 st Quartile	115.90	117.00
Median	137.90	138.30
Mean	137.20	138.60
3 rd Quartile	159.90	163.40
Maximum	218.00	218.00
Pearson Correlation	0.3324	0.3416
Spearman Correlation	0.4707	0.4805
Kendall's Tau	0.4357	0.4486



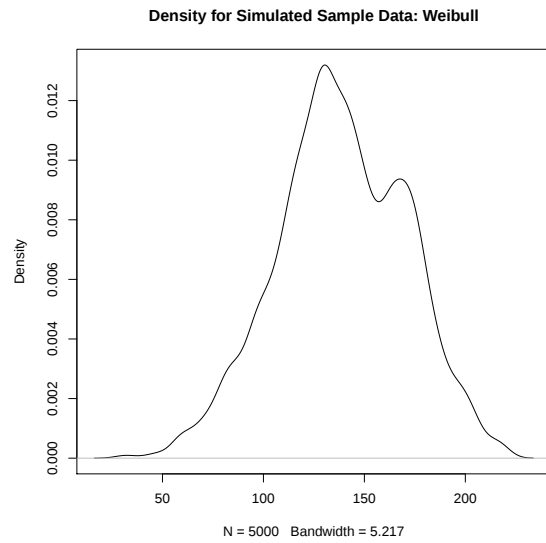
(a) Generated Log-Normal



(b) Generated Weibull

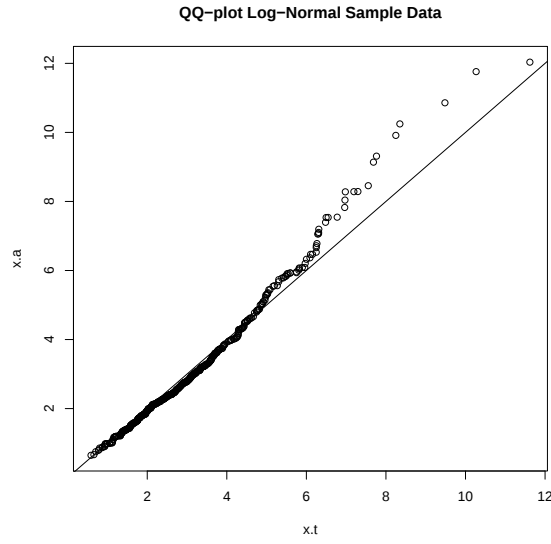


(c) Simulated Log-Normal

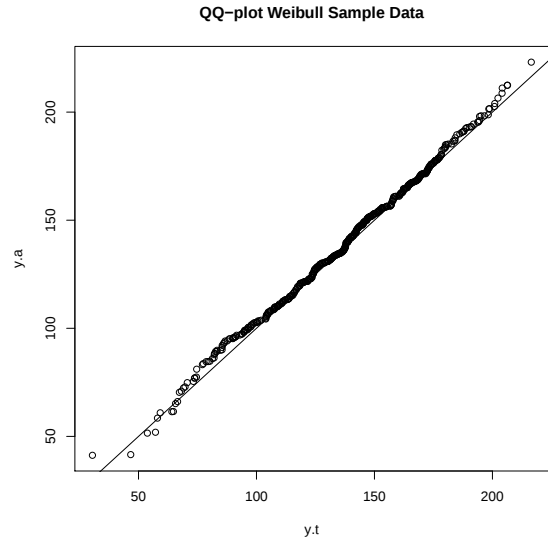


(d) Simulated Weibull

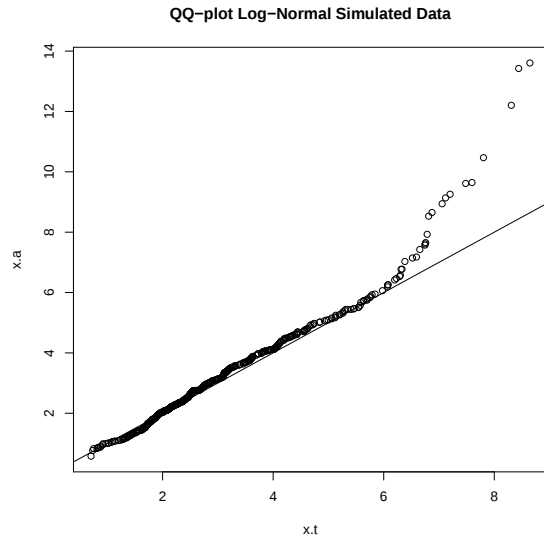
Figure 1: Generated and Simulated (Using Proposed Method) Data



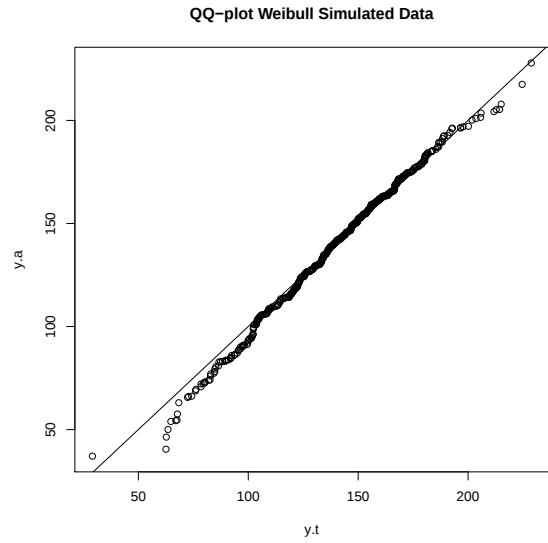
(a) Generated Log-Normal



(b) Generated Weibull

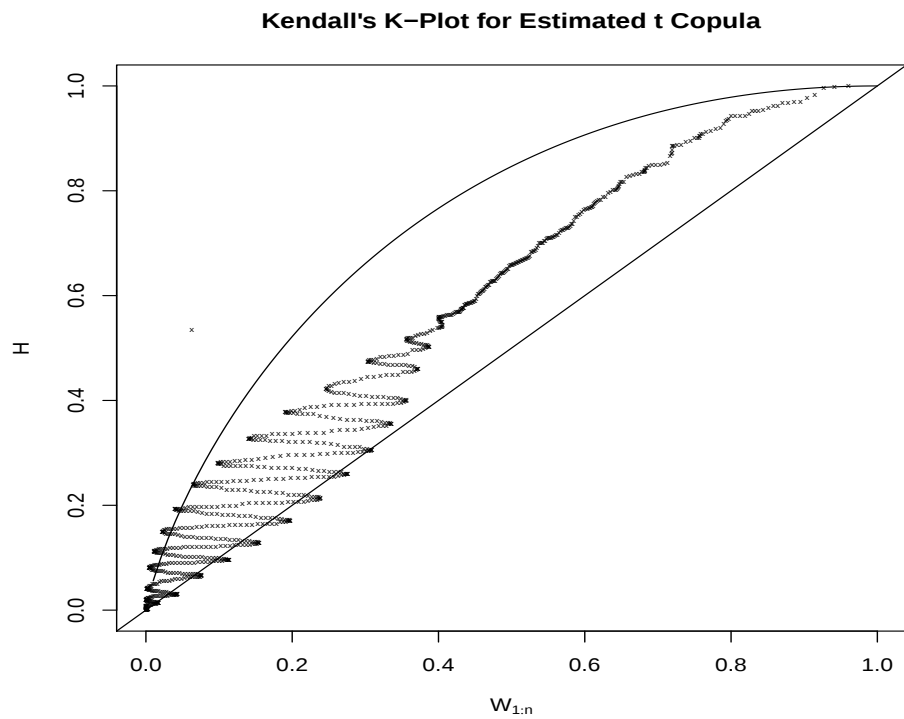


(c) Simulated Log-Normal

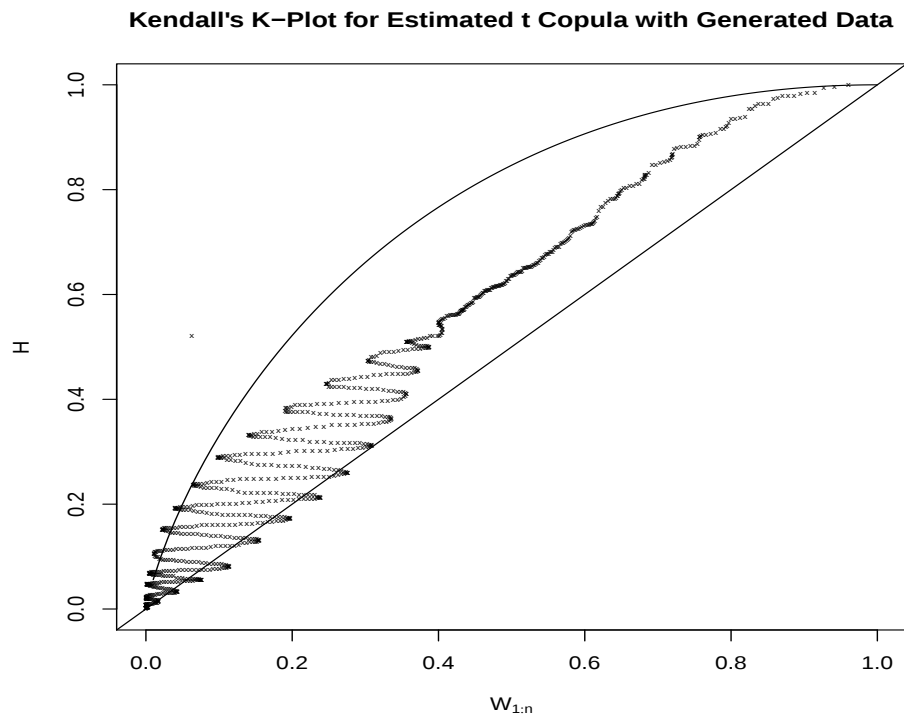


(d) Simulated Weibull

Figure 2: Generated and Simulated (Using Proposed Method) QQ-Plots

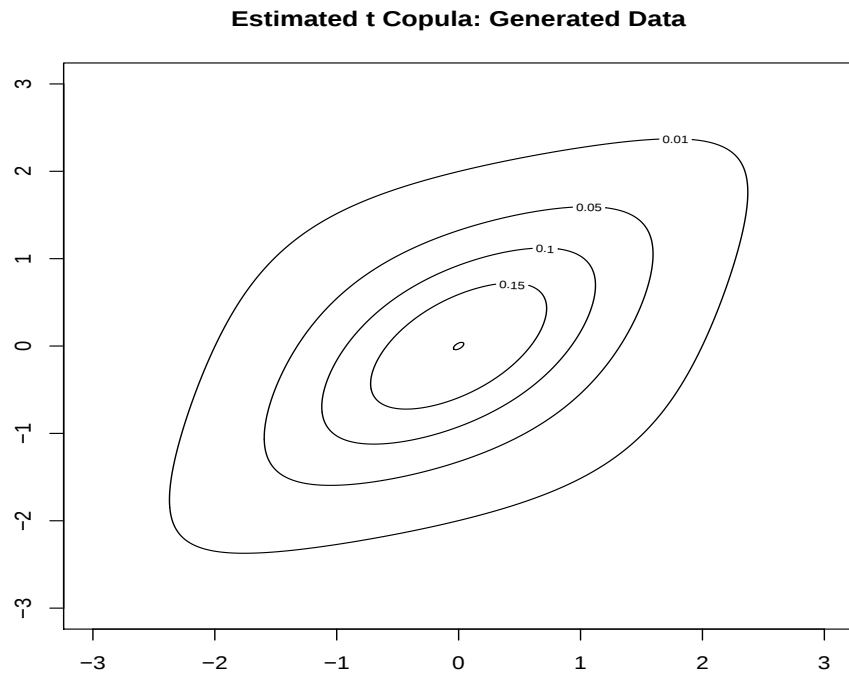


(a) Generated Data

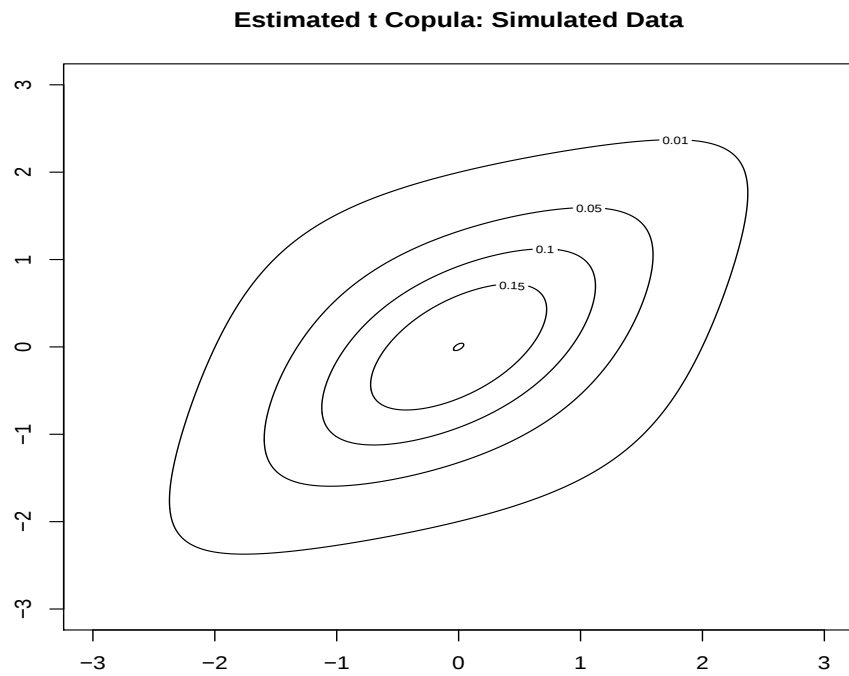


(b) Simulated Data

Figure 3: Generated and Simulated Copula Contours



(a) Generated Data



(b) Simulated Data

Figure 4: Contours for Copulas