Fit to a t – Estimation, Application and Limitations of the t-copula

Topic 1: Risk Evaluation

Gary G Venter

Guy Carpenter Instrat

One Madison Avenue, New York NY USA

+1-212-323-1739

Fax: +1-212-390-4739

gary.g.venter@guycarp.com

Abstract:

Copulas provide a convenient way to represent joint distributions. In fact the joint distribution function can be expressed as the copula function applied to the separate individual distributions. That is, $F(x_1, x_2, \ldots x_m) = C[F_1(x_1), F_2(x_2), \ldots F_m(x_m)]$ where C is the copula function. Background information on copulas is covered in a number of papers, and will be largely assumed here.

The starting point of this topic is bivariate copulas, but most of these do not extend well into higher dimensions. For a multivariate copula for insurance related variates you would like to be able to feed in a correlation matrix of the variates as well as have some control over the degree of correlation in the tails of the distributions. Often more than two related variates are needed, such as losses in several lines of insurance.

This paper focuses on the t-copula, which meets these minimum requirements, but just barely. You can input a correlation matrix and you do have control over the tail behavior, but you only have one parameter to control the tail, so all pairs of variates will have tail correlation that is determined by that parameter. The normal copula is a limiting case, in which the tails are ultimately uncorrelated if you go out far enough.

The structure of the paper is to jump right in to a discussion of the t-copula in the bivariate case, then extend this to higher dimensions. A trivariate example is given using cat model output for three lines of insurance. Methods for selecting parameters and testing goodness of fit are discussed in this context, using descriptive functions.

Keywords:
Copula, correlation, multi-variate

Acknowledgement

Much credit must go to Andrei Salomatov for solving the calculation issues for certain extreme cases discussed below.

Fit to a t – Estimation, Application and Limitations of the t-copula

The Bivariate t-Copula

The bivariate t-copula has two parameters that control the tail dependence and the degree of correlation separately. The t-distribution with n degrees of freedom is defined by:

$$f_n(x) = K_1(1+x^2/n)^{-(n+1)/2}, \text{ with } K_1=\Gamma(1/2+n/2)(n\pi)^{-1/2}/\Gamma(n/2).$$

Here n is often an integer, but doesn't have to be. The distribution is symmetric around zero and can be calculated by:

$$F_n(x) = ½ + ½ \text{ sign}(x)\text{betadist}[x^2/(n+x^2), ½, n/2],$$

where betadist defines the beta distribution, as in Excel.

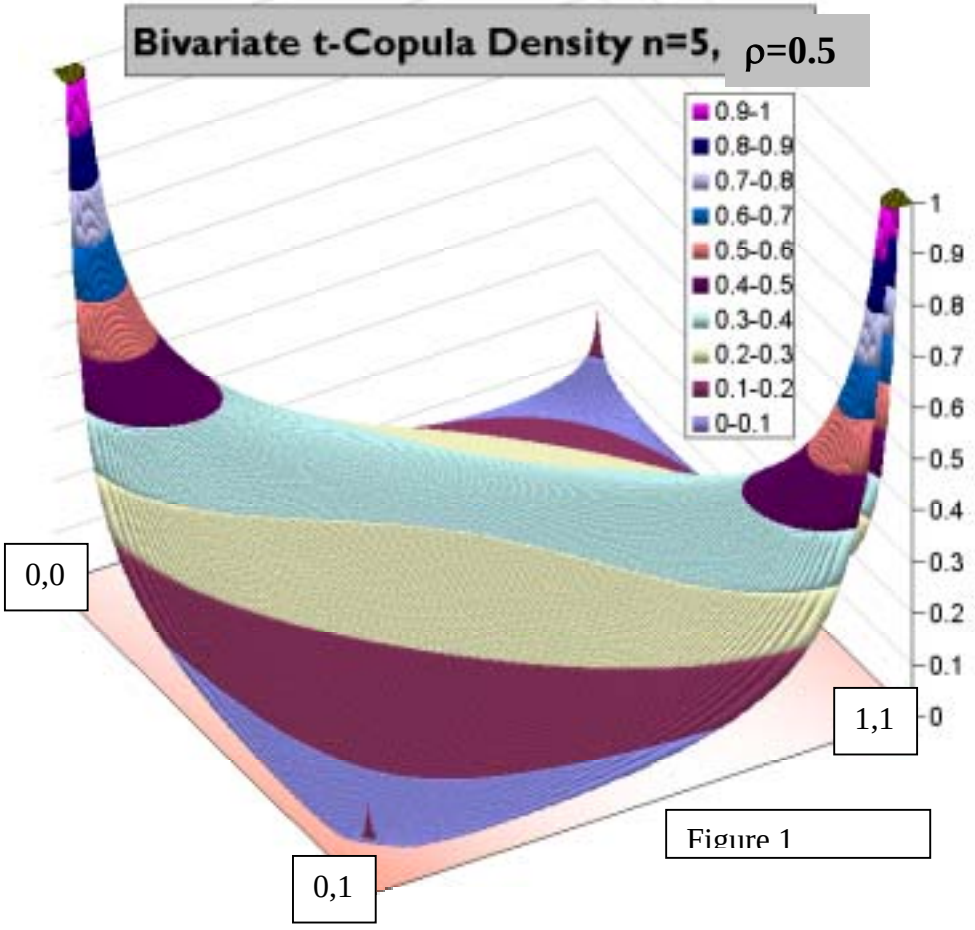


Figure 1

An example of the density of the t-copula is graphed in Figure 1. The density of the bivariate t-copula with n degrees of freedom and the correlation parameter ρ is defined as:

$$c(u,v; n,\rho) = K_2[(1+s^2/n)(1+t^2/n)]^{(n+1)/2}\{1+[s^2-2\rho st+t^2]/[(1-\rho^2)n]\}^{-1-n/2}$$

with $K_2 = ½[\Gamma(n/2)/\Gamma(0.5+n/2)]^2 n(1-\rho^2)^{-1/2}$ and $s=F_n^{-1}(u)$, $t=F_n^{-1}(v)$.

The inverse t-distribution needed for this can be calculated with an inverse beta by $s = \text{sign}(u-½)n^{1/2}[^{-}1+1/\text{betainv}(|2u-1|,½, n/2)]^{-1/2}$.

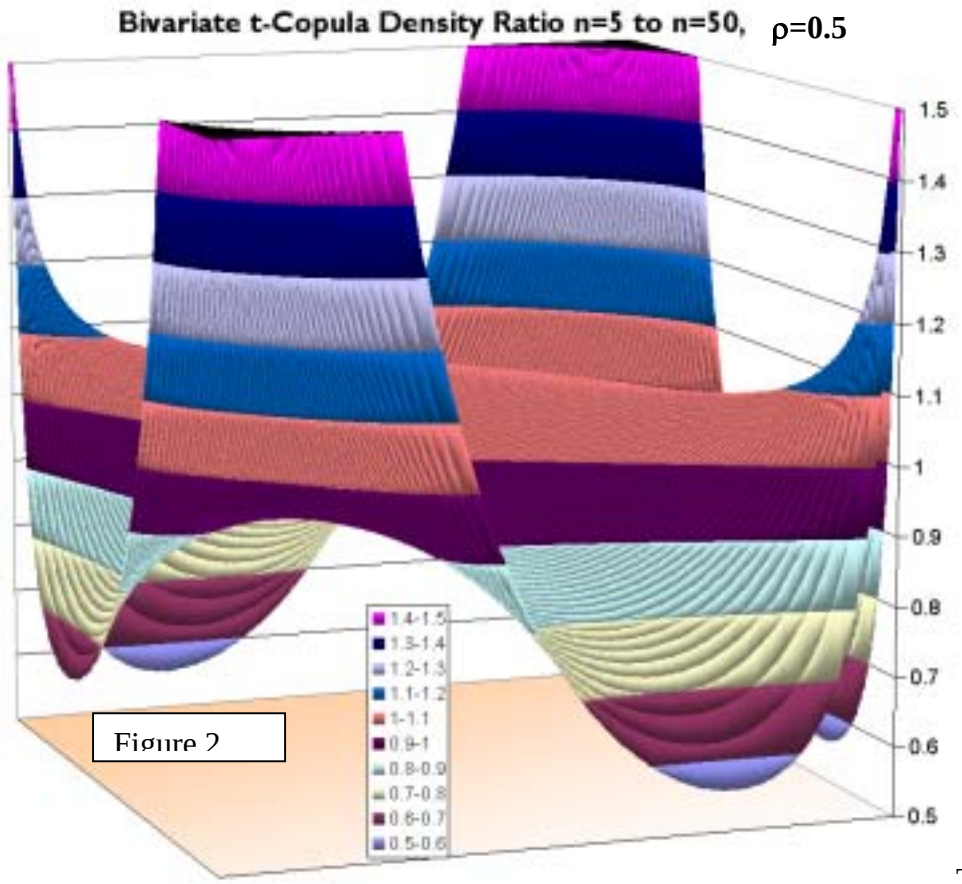


Figure 2

The concentrations of probability near [0,0] and [1,1] are seen in many copulas, but the smaller concentrations around [0,1] and [1,0] are more unusual. Figure 2 shows the ratio of densities for n=5 to n=50. The latter is similar to the normal copula density, which approaches zero at [0,1] and [1,0]. Thus the density ratio is highest in those regions, even though it is well above unity around [0,0] and [1,1]. With a given correlation parameter or matrix, the linear correlation and Kendall's τ for the t-copula are the same for any n as for the normal copula ($n \to \infty$). The lower values of n produce greater upper and lower tail dependence with the same overall correlation essentially because they put more weight in all the corners. The additional weight in the off-diagonal corners cancels out the additional tail dependence, keeping the overall correlation the same.

Kendall's τ is related to ρ by $\tau = (2/\pi)\arcsin(\rho)$. Also the right tail dependence measure R, defined as $\text{limit}_{z \to 1}\Pr(U>z|V>z)$, is given by:

$$R/2 = 1 - F_{n+1}\{[(n+1)(1-\rho)/(1+\rho)]^{0.5}\}.$$

Thus R can be expressed as a function of τ and n, as graphed in Figure 3. Even zero τ can give a positive tail dependence with this copula. The tail dependence can approach zero for any τ by taking n large, thus approximating the limit of the normal copula, which has tail dependence of zero.

The Multivariate t-Copula

To define the multivariate copula, suppose there are m variates, and **u** is a vector of m probability values (numbers in [0,1]). Let **s** be the vector of the univariate t-quantiles of **u** with n degrees of freedom, that is $s=F_n^{-1}(u)$ for each element of **s** and **u**. Also let Σ be an m x m

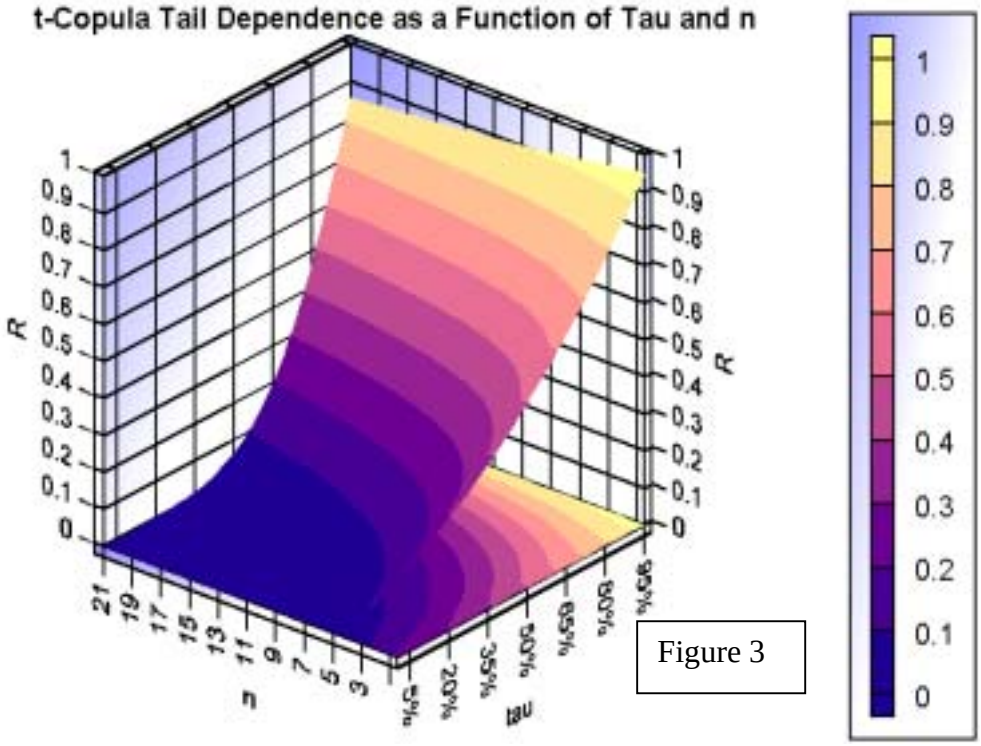


Figure 3

correlation matrix with determinant d. Then the m-dimensional t-copula has density:

$$c(\mathbf{u}; n, \Sigma) = K_m[\Pi_{i=1}^{m}(1+s_i^2/n)]^{(n+1)/2}(1+\mathbf{s}'\Sigma^{-1}\mathbf{s}/n)^{-(m+n)/2}$$

where $K_m = \Gamma[(m+n)/2][\Gamma(n/2)]^{m-1}[\Gamma(½ +n/2)]^{-m}d^{-1/2}$.

By starting with a Kendall's τ coefficient matrix T, the correlation matrix needed here can be specified by $\Sigma = \sin(T\pi/2)$. Thus this copula has complete flexibility in its correlation structure. However there is only one n used, so the tail dependence will be determined by that n for all pairs of variates. In the graph above, all the tail-dependence measures for all pairs of variates would fall on the same vertical cross section, determined by the value of n used for the copula. Thus the pairs with higher τ will have higher R as well.

The univariate or multivariate t distribution can be characterized (and simulated) by a (possibly multivariate) normal distribution divided by a multiple of the square root of an independent univariate chi-squared distribution. When generating it in this way, if a low draw comes up for the chi-squared variate, large values of the t variate can be produced, even if the normal values were not particularly large. In the multivariate case then all the t variates can be jointly large even if they were not originally correlated. That illustrates why the tail-dependence can be somewhat high even with zero τ. An example might be where the reciprocal of the t variate represents the inflation rate, which hits all the lines. This effect is sometimes called a common shock, i.e., the common shock of a large inflation rate can induce a correlation among otherwise independent lines.

More precisely, to generate a vector of probabilities from the multi-variate t-copula, first generate a multi-variate normal vector with the same correlation matrix, then divide it by $(y/n)^{0.5}$ where y is a number simulated from a chi-squared distribution with n degrees of freedom. This gives a t-distributed vector, and the t-distribution F_n can then be applied to each element to get the probability vector.

The ratio y/n is a scale transform of the chi-squared variate, so is a gamma variate. If the gamma density is parameterized to be proportional to $x^{\alpha-1}e^{-x/\beta}$, then y/n has parameters β = 2/n and α = n/2. This is a distribution with mean 1. It can be simulated easily if an inverse gamma function is available, as in some spreadsheets.

Because a power of the gamma deviate is a divisor, a factor is being applied that is actually inverse transformed gamma distributed. The inverse transformed gamma distribution in α, τ, θ has density proportional to $\exp(-(\theta/x)^{\tau})/x^{\alpha+1}$. The factor $(n/y)^{0.5}$ applied to the normal variates is distributed inverse transformed gamma in α = n, τ = 2, and $\theta = (n/2)^{1/2}$. This has a mean greater than unity and an inverse power tail with power n, and so is a heavy-tailed distribution. Especially when n is small, this gives the possibility of large values of the factor occasionally being applied to all the normal draws, giving simultaneous large values of all the variates.

Example – Hurricane Losses

Parameter estimation issues and applications can be illustrated by a sample of losses simulated from a hurricane model. The simulation generated losses under three lines of insurance: residential property (R), commercial property (C) and automobile (A). Naturally these are highly correlated losses, as hurricane losses from a stronger storm tend to be higher in all three lines. The strength of the storm could be considered to be the common shock that correlates all the lines. Having a large generated sample like this does not require a fitted copula to be useful in loss estimation, so in practice there would be little need to fit a copula to it. It is a useful dataset for illustrating fitting concepts, however.

The empirical trivariate copula can be calculated at any 3-vector of probabilities by counting the proportion of the sample triplets of empirical probabilities that are less in each index. Each of the three bivariate empirical copulas from the three pairs of variables can be calculated similarly. The averages of the bivariate copulas give estimates of the τ correlations by the relationship $\tau = 4E(C) - 1$, where E(C) is the expected value of the copula.. This scaling of the mean value of the copula can be extended to define higher dimensional analogues of τ by requiring that τ = 0 for the independence case and τ = 1 for perfect correlation. The scaling for m-dimensions that does this is $\tau = [2^m E(C)-1]/[2^{m-1}-1]$. (There are other possible multi-variate extensions of τ, but they will not be used here.) For the hurricane data, these τ's are:

	AC	**AR**	**RC**	**ARC**
τ	82.4%	84.4%	87.6%	84.8%

The bivariate τ's provide estimates of the correlation ρ for each bivariate copula, and thus for the correlation matrix for the trivariate copula, using $\rho = \sin(\pi\tau/2)$. For the sample, these are:

	AC	**AR**	**RC**
ρ	.96	.97	.98

To estimate n, the tail behavior is key. One possible approach might be to estimate R, the limit $_{z \to 1}\Pr(U>z|V>z)$. However this is difficult to estimate from data because the function

R(z), defined as Pr(U>z|V>z), can drop rapidly for z near 1, and there is less and less data to use the closer z gets to 1.

Note that R(z)= Pr(U>z & V>z)/Pr(V>z). Since Pr(V>z) = 1– z = Pr(U>z), U and V can be switched in the definition of R(z). A similar concept can be defined for the multivariate copula:

$$R(z)= Pr(U>z \ \& \ V>z \ \& \ W>z)/z = Pr(U>z \ \& \ V>z|W>z)$$

Because of the symmetry in the first equation, U, V, and W can be swapped around at will in the second equation. This function provides a measure of the overall tail dependency of the three variates, and it can be generalized to higher dimensions. A similar tail dependency function can be defined for the left tail:

$$L(z) = Pr(U<z \ \& \ V<z \ \& \ W<z)/z = C(z,z,z)/z,$$ and similarly in the bivariate case.

The empirical versions of these functions are graphed in Figure 4. From the graph, the right

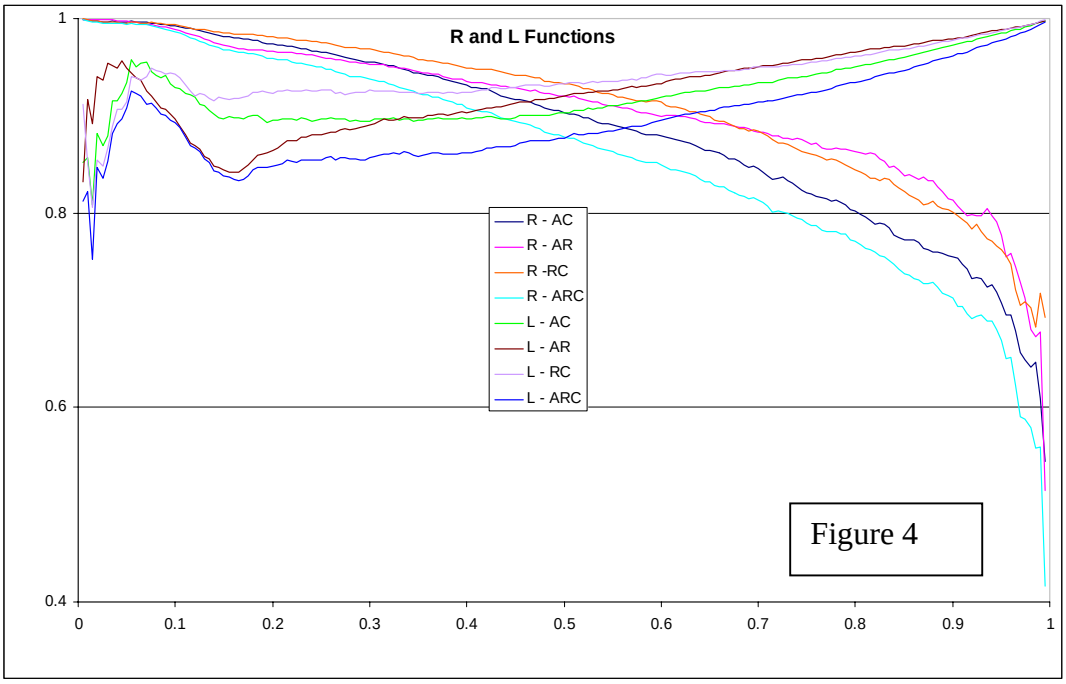
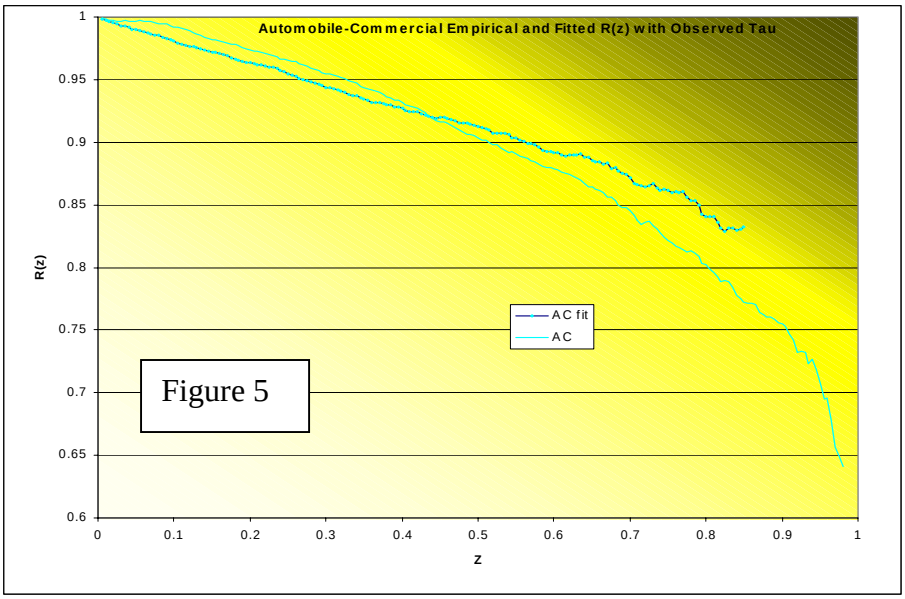


Figure 4

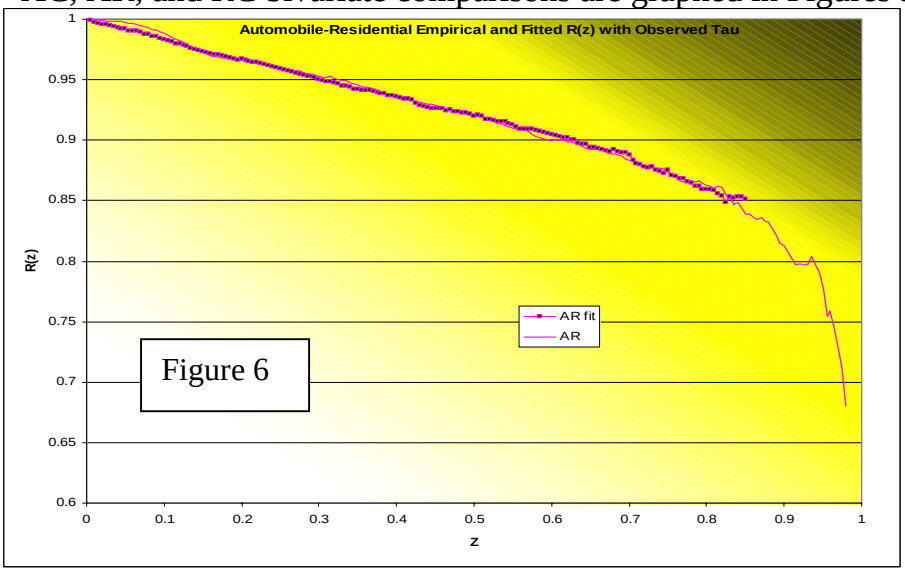
and left tail functions are clearly not symmetrical. This would rule out the t-copula, which is. However most issues concern the large loss cases, so a copula approximating the right side would be most appropriate, and the t-copula might work for this.

One possibility would be to estimate the t-copula correlations and degrees of freedom by maximum likelihood. The likelihood function for a parametric copula at a point in the sample is the density of the parametric copula computed at the empirical copula vector for that point, so can be readily calculated.

However in this case, MLE is not likely to give the intended fit, in that it would be affected by the smaller claims that do not appear to mirror the large claims. So the sample correlations come back as a starting reference point, even though they use the whole distribution.

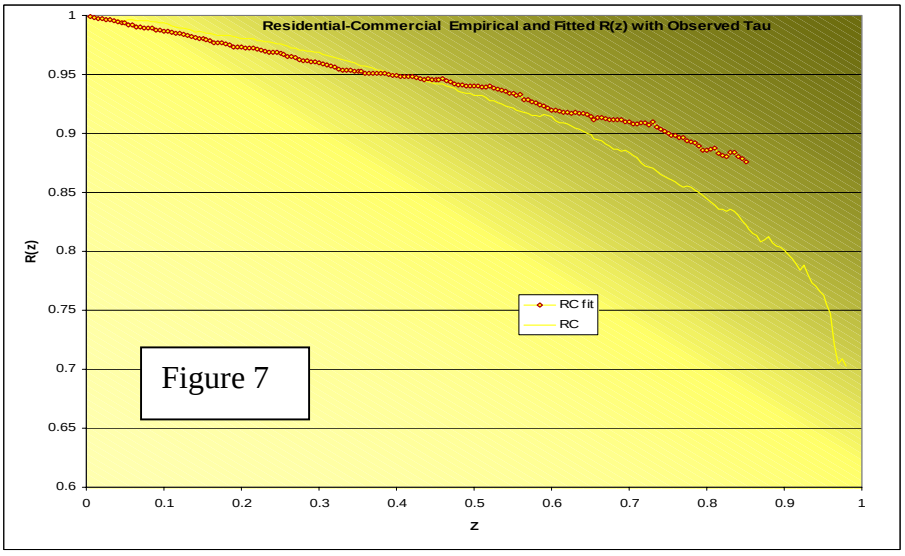


To test how well the sample correlations match the larger losses, a simulation can be performed with the sample correlations and a selected n, and the simulated R(z) compared to the sample's. The arbitrary choice of degrees of freedom most affects the extreme percentiles, so this comparison was initially cut off at the 85th percentile, with a selected n=20. The AC, AR, and RC bivariate comparisons are graphed in Figures 5, 6, and 7.



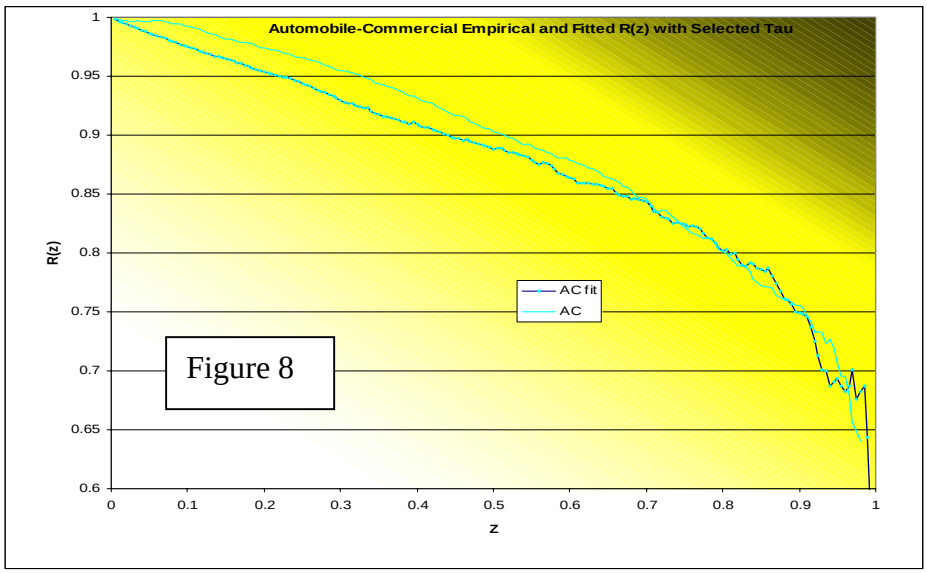
The AR R(z) comparison is very close by this measure, while the other two pairs do not fit very well.

To see how much this is influenced by the choice of correlations, a few correlations were tested by this same methodology to see how well they work. The better fitting selections for AC and RC were a bit lower than the overall correlations. The AC. AR, and RC fits are shown in Figures 8 – 10, again with n=20.

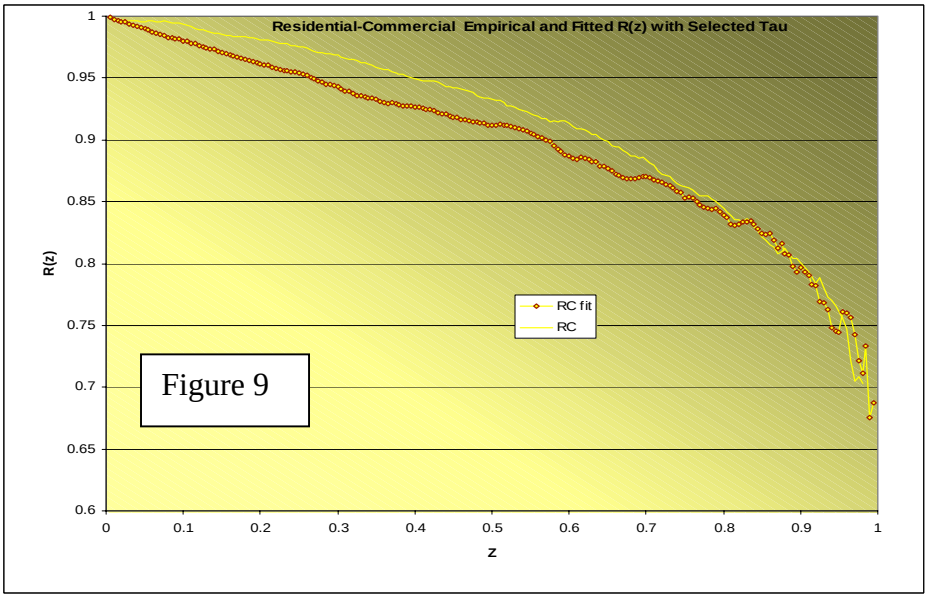


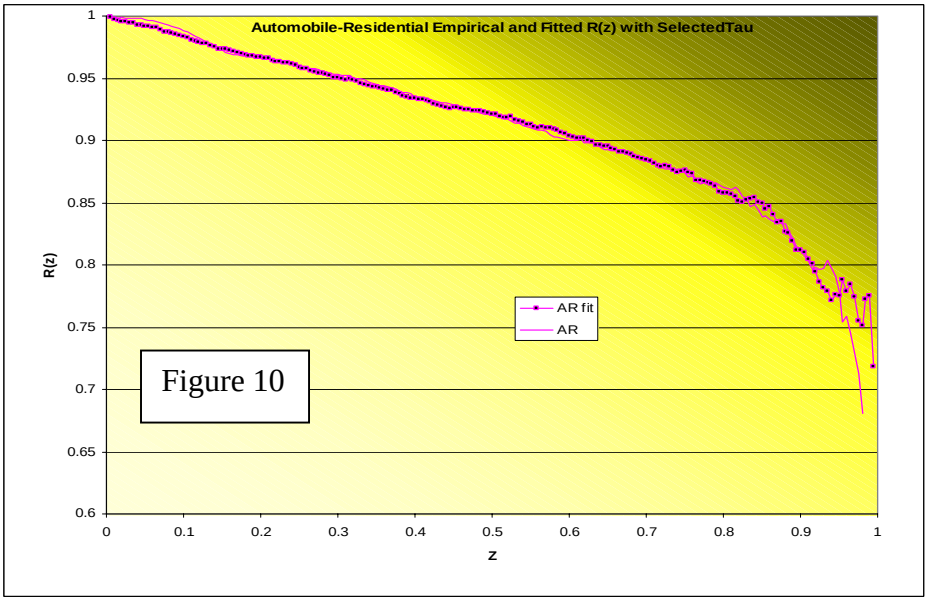
While there are still some issues with the fits for smaller values of z, they are much better for the large losses, as in-

intended. This case was not cut off at 85%, and some simulation instability shows for the larger values of z.



	AC	AR	RC
Sample ρ	.96	.97	.98
Selected ρ	.94	.97	.96





Due to the symmetry of the t-copula, R(z) = L(1 – z), and L can be calculated directly from the copula function. However, even though the t-copula is easy to simulate, the copula function is difficult to calculate, as the integration is difficult near 0. Numerical integration relying on simulation is often used for this. Figure 11 shows the L function for the ρ's fit above for a few values of n using this approach. It is only for small values of z that n makes a difference in the L and R functions, at least for large values of ρ. The ρ itself does affect the functions for all values of z.

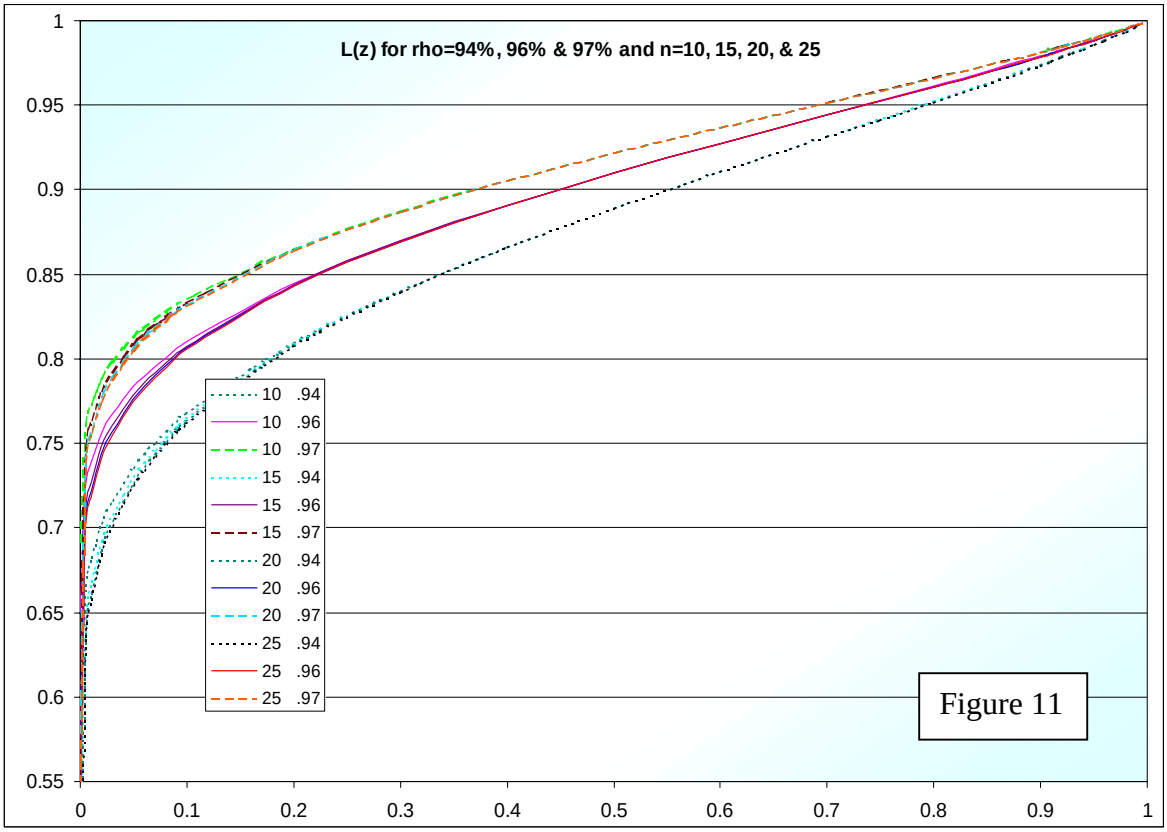
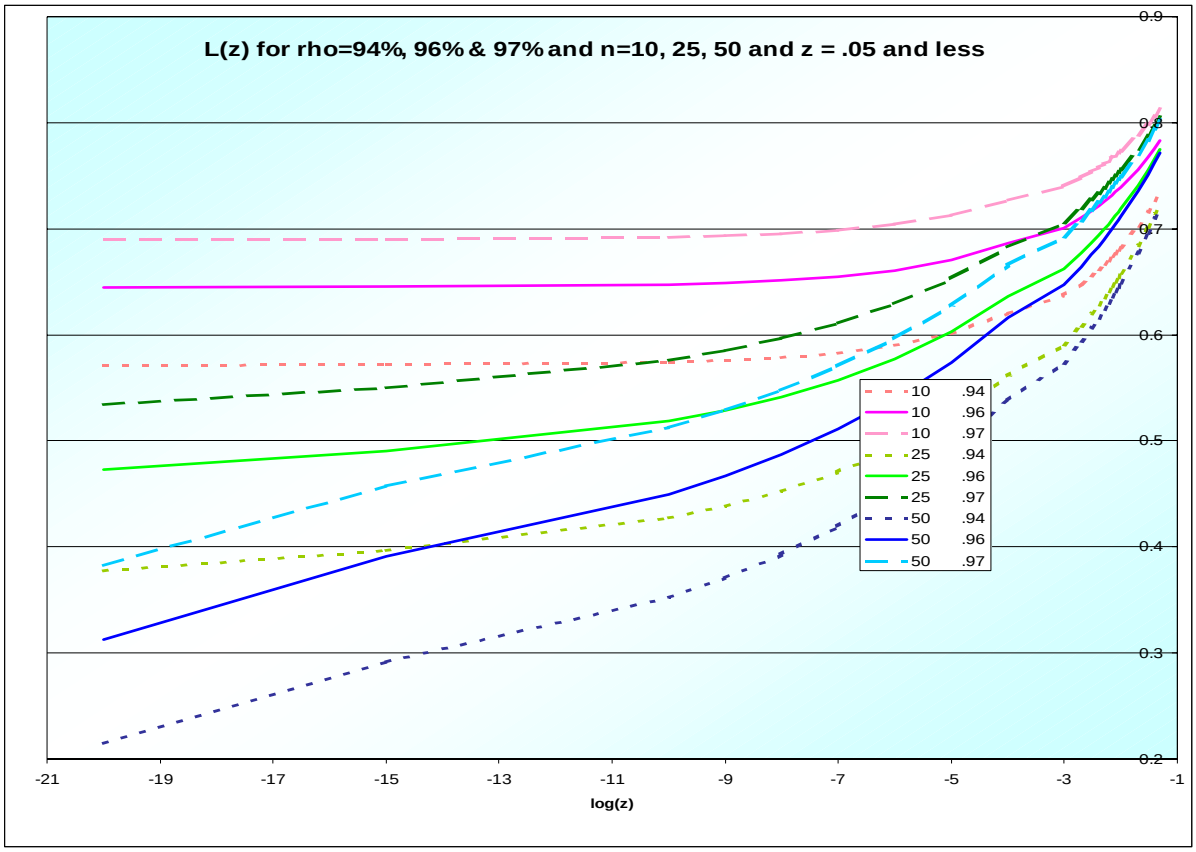


Figure 12 looks at L(z) for z [0.05 for these same ρ's and a few n's. It is on a log scale to illustrate the behavior of L for very small z's. For z in this range, which is shown down to ln(z) = -20, n is at least as important as ρ in influencing the value of L(z). Also, the function declines very slowly even for n=50. With these values of z the tails become less determined by ρ, but only in the very extreme tail, beyond the area of practical concern.

To select a value of n, the empirical R functions were evaluated at z=0.005 and z=0.01, and n's sought to best match. For the RC pair with ρ = 96%, the best n was 41.5. For the other cases, n=500 worked as well as anything, suggesting a normal case. The target and fitted R(z)'s are shown for each pair in the table below. The RC fit is best. The other two fit ok at z = 0.01, but drop off for smaller z. That could be a sample size problem at this level.

Only one value of n is used in the t-copula, so a compromise value has to be selected. Perhaps a value near 42 would be appropriate. This works for RC but imposes too much tail association for the other pairs. This is only in the extreme tail, however, so might not be problematic.

Target/Fit	94% (AC)	97% (AR)	96% (RC)
0.005	.54 / .61	.52 / .68	.693 / . 695
0.01	.61 / .64	.68 / .70	.718 / .715



Summary

The functional form of the t-copula is somewhat complicated, but most of the key functions are readily available in spreadsheets and statistical packages. Simulating samples is quite easy, as this just uses a simple adjustment to normal copula samples.

Estimating parameters from data is more problematic. If the data is symmetric, maximum likelihood would be a good choice. In that case, a comparison of the empirical and fitted R and L functions could be used to evaluate goodness of fit.

When the right and left tails are quite different the t-copula would not usually be indicated, but if only the right tail behavior is important in practice, a fit to that could be sought. Finding parameters that match the empirical and fitted R function is a reasonable way to do that. In the sample data reviewed, finding a match for the correlation matrix was relatively straight-forward, but finding the best n was more difficult. For the high correlations found in this sample, different values of n affected R(z) only in the very extreme tail – even beyond where most reinsurance interest would be. Since that is where the data is most scarce, reliable fits are difficult. However the choice of n is not too critical for the same reason.

The main practical obstacle to the use of the t-copula is that there is only one parameter – n – to control tail association, and different pairs of variates might have different indicated n's. Computationally the biggest problem is calculating C for extreme values. This would be necessary only for trying to fit parameters to the extreme tail, however.