3.1 Dependence Between Variables

So far we have discussed probability distributions and multivariate distributions in some detail. For collections of random variables, we are often interested in how they vary together. We already have a measure for this: the covariance. One issue with the covariance between two random variables, X and Y,

$$\Sigma(X,Y) = E[XY] - E[X]E[Y], \tag{3.1}$$

is that it has units that are the product of the units of X and Y. This can make it difficult to compare covariances. For instance, $\Sigma(X,Y) > \Sigma(X,Z)$ does not imply that there is a stronger relationship between X and Z than X and Y because of the units.

3.1.1 Pearson Correlation

A normalized measure of the relation between two random variables, is the Pearson correlation coefficient, ρ. Oftentimes, this is simply called the correlation coefficient or correlation. Considering two random variables, X, and Y, the correlation coefficient is

$$\rho(X,Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}. \tag{3.2}$$

That is, the Pearson correlation is the covariance normalized by the standard deviation of each variable. On this normalized scale, we can say things about how two variables change together. If the variables are independent, then $\rho(X,Y) = 0$. As with covariance, a correlation of zero between variables does not imply that the variables are independent.

One property of the correlation coefficient is that if X and Y are linearly related, i.e., there exist an $a > 0$ and b such that $Y = aX + b$, then $\rho(X,Y) = 1$. As a corollary, we have the relation

$$\rho(X,Y) = \mathrm{sign}(a)\rho(aX+b,Y).$$

When we have a collection of random variables, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$, we can define a correlation matrix $\mathbf{R}$ in terms of the covariance matrix as

$$R_{ij} = \frac{\Sigma_{ij}}{\sigma_{X_i}\sigma_{X_i}}, \tag{3.3}$$

where $\sigma^2_{X_i} = \Sigma(X_i, X_i)$ is the variance in X_i.

The benefit of the Pearson correlation coefficient is that it is easy to calculate, as simple as the covariance matrix. However, there are some downsides. One is that it is not defined if the expected value of XY is not defined. This is the case with Cauchy random variables, given by the PDF with parameters x_0, γ,

$$f(X) = \frac{1}{\pi\gamma}\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]^{-1}. \tag{3.4}$$

The mean and variance of the distribution are undefined because the distribution goes to zero too slowly, but the median and mode are x_0. The PDF for a Cauchy distribution, and it's comparison to the standard normal, are given in Figure 3.1.

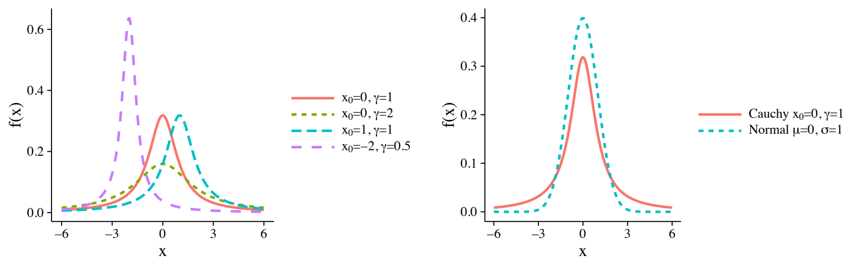

Fig. 3.1 The Cauchy distribution with various parameters and compared with the standard normal.

Another, potentially more important, downside of the Pearson correlation coefficient is that if X is transformed by a nonlinear, strictly increasing function, $g(X)$, the correlation $\rho(X,Y)$ will be different than $\rho(g(X),Y)$. This means that if there is a nonlinear relation between X and Y the Pearson correlation coefficient may under- or over-estimate the relation between the two variables.

3.1.2 Spearman Rank Correlation

An alternative to the Pearson correlation is the Spearman rank correlation, or Spearman correlation. In this measure we look for general, monotonic relationships between two variables. This is defined by looking at the correlation between the marginal CDF of each variable:

$$\rho_S(X,Y) = \rho(F_X(x), F_Y(y)). \tag{3.5}$$

If we do not know the marginal CDF, but we have samples of the random variables we can still estimate the Spearman correlation. Given N samples of X and Y we create a function that takes sample x_i or y_i and gives the rank of the that sample amongst the N samples:

$$\text{rank}(x_i) = \text{Rank of } x_I \text{ in sample population.}$$

Using this function we then define the Spearman correlation coefficient for the samples:

$$\rho_S(X,Y) = \frac{\sum_{i=1}^{N}(\text{rank}(x_i) - \bar{r}_X)(\text{rank}(y_i) - \bar{r}_Y)}{\sqrt{\sum_{i=1}^{N}(\text{rank}(x_i) - \bar{r}_X)^2}\sqrt{\sum_{i=1}^{N}(\text{rank}(y_i) - \bar{r}_Y)^2}}, \tag{3.6}$$

where

$$\bar{r}_X = \sum_{i=1}^{N} \text{rank}(x_i).$$

When computing ρ_S any ties in the data are assigned the average rank of the tied scores.

One of the important properties of the Spearman correlation is that if there exists a strictly increasing function $g(X)$ that relates X to Y as $Y = g(X)$, then $\rho_S(X,Y) = 1$. Furthermore, a strictly monotonic transformation of X or Y will not affect the Spearman correlation

As with the Pearson correlation, we can compute a Spearman correlation matrix for a collection of random variables $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$. We will call this matrix $\mathbf{R}_S$ and it is given by

$$R_{S,ij} = \rho_S(X_i, X_j).$$

3.1.3 Kendall's Tau

The final measure of correlation that we will use is Kendall's tau or the Kendall rank correlation coefficient. Similar to the Spearman correlation it tries to measure the relation between two variables in terms of the ranks. It is best for looking at a sample population of random variables because it requires looking at pairs of samples of random variables. To define Kendall's tau be consider N samples of random variables X and Y. We examine all the pairs of samples (x_i, y_i) for $i \neq j$. There are $\frac{1}{2}n(n-1)$ such pairs. We look at each pair and say that a pair ij is concordant if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. A pair is discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i < y_j$. If either $x_i = x_j$ or $y_i = y_j$ then the pair is a tie.

Using this comparison of pairs we define Kendall's tau as

$$\tau = \frac{(\text{\# of concordant pairs}) - (\text{\# of discordant pairs})}{\frac{1}{2}n(n-1)}. \tag{3.7}$$

The range of τ is $[-1, 1]$. Kendall's tau has the property that it is not affected by performing a nonlinear, increasing transformation on either random variable: this is the same property Spearman correlation has. We can relate τ to the Pearson correlation coefficient if the variables X and Y are jointly normally distributed through the equation

$$\tau = \frac{2}{\pi} \arcsin \rho(X,Y).$$

We will use Kendall's tau when we want to relate two random variables through copulas.

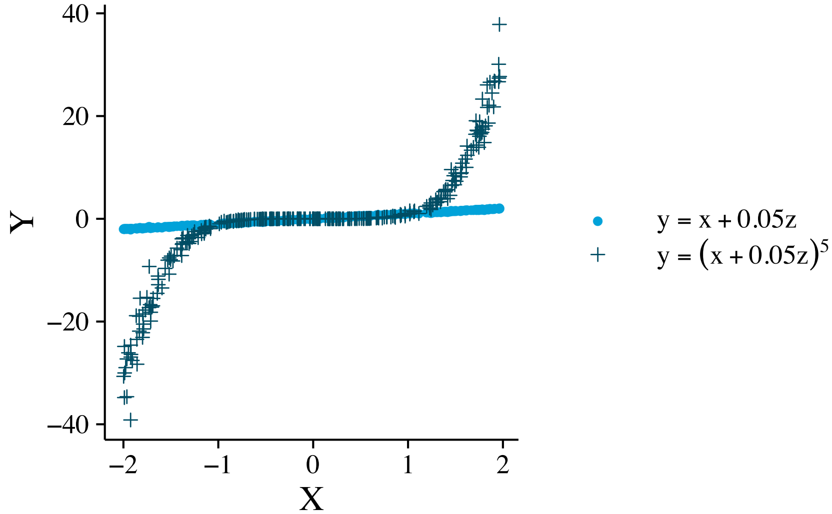

Fig. 3.2 The comparison of Pearson, Spearman, and Kendall's tau correlation measures on 300 samples of two pairs of random variables: $(X, X+0.05Z)$ and $(X, (X+0.05Z)^5)$, where Z is a standard normal random variable. The three measures give a correlation of $\rho = 0.999$, $\rho_S = 0.999$, and $\tau = 0.973$ for the correlation of $(X, X+0.05Z)$. The correlation of $(X, X+0.05Z)$, the Spearman correlation and Kendall's tau values do not change, but $\rho = 0.843$ for this data.

As comparison of the correlation measures, Figure 3.2 shows how a strictly increasing transformation of a variable changes the Pearson correlation, but not the Spearman correlation or Kendall's tau. In the figure the correlation between random variables $(X, X+0.05Z)$ and the correlation between $(X, (X+0.05Z)^5)$, where Z is a standard normal random variable is computed. The Spearman and Kendall measures do not change, whereas the Pearson correlation drops by 15%.

3.1.4 Tail Dependence

Another important characterization of how two variables vary together is tail dependence. This is a measure of the correlation between variables as their lower and upper bounds are approached. The lower tail dependence, λ_l is

$$\lambda_l(X,Y) = \lim_{q \to 0} \mathrm{P}(Y \le F_Y^{-1}(q) \mid X \le F_X^{-1}(q)). \tag{3.8}$$

This is the probability that Y goes to its lower bound as X goes to its lower bound. The upper tail dependence is

$$\lambda_u(X,Y) = \lim_{q \to 1} \mathrm{P}(Y > F_Y^{-1}(q) \mid X > F_X^{-1}(q)), \tag{3.9}$$

and measures the probability the X and Y go to their upper bound together.

Tail dependence is different than typical correlation measures in that it is only interested in extreme values. For example two variables could have a Pearson correlation of 0.5, but a tail dependence much larger, say 0.9. This has been observed, for example, in the returns of stocks. Many stocks that had low correlation in typical times had very high lower tail dependence during the financial crises (they all went down a lot).

3.2 Copulas

A common occurrence when evaluating collections of random variables, it is often much easier to determine the marginal CDF or PDF of each variable, rather than determine the joint distribution functions. Moreover, given a sample of data it is possible to compute correlations between the random variables (in either of the three flavors we mentioned in the previous section). The question is, given the scenario where one has

- An estimate of the marginal CDF of each random variable, and
- An estimate of the correlation between the random variables,

can one generate a joint distribution between the variables and generate samples from the joint distribution? Clearly, there is not a unique way of creating this joint distribution because many functions could replicate the marginal distributions and have a defined correlation.

To answer this question we turn to copulas (or copulæ if one is a fan of Latinisms). We will begin with discussing bivariate copulas before generalizing the idea to general collections of random variables. The word copula comes from a Latin for linking together; in our context it will link marginal distributions to a joint distribution.

A copula, $C(u,v)$, joins random variables X and Y if the joint CDF can be written as

$$F(x,y) = C(F_X(x), F_Y(y)). \tag{3.10}$$

This definition takes the marginal CDF for each variable and creates a joint CDF. A result known as Sklar's theorem tells us that such a copula will exist for any joint CDF and it is unique if the marginal CDFs are continuous. A copula has the domain $u, v \in [0,1]$, and a range of $[0,1]$. For a given copula we can define the joint PDF as

$$f(x,y) = c(F_X(x), F_Y(y)) f_X(x) f_Y(y), \tag{3.11}$$

where the copula density, $c(u,v)$, is given by

$$c(u,v) = \frac{\partial^2}{\partial u \partial v} C(u,v). \tag{3.12}$$

This definition is a special case of Eq. (2.25). Additionally, the conditional CDF $C(v|u)$ is

$$C(v|u) = \frac{\partial}{\partial u} C(u,v). \tag{3.13}$$

The simplest copula is the independent copula:

$$C_{\mathrm{I}}(u,v) = uv.$$

Copulas are widely used in the finance and insurance industries to model the joint distributions of risks. Because the fact that mapping marginal distributions to joint distributions is not unique, the way we use a copulas requires choices by the user. The considerations of ease of use, matching observed correlation, and tail dependence have to be weighed when choosing a copula.

3.2.1 Normal Copula

A simple, but useful copula, is the normal (or Gaussian) copula

$$C_N(u,v) = \Phi_{\mathbf{R}}(\Phi^{-1}(u), \Phi^{-1}(v)), \tag{3.14}$$

where $\mathbf{R}$ is a correlation matrix for the intended joint distribution. The normal copula is simple to sample. Given two random variables X and Y with marginal CDFs $F_X(x)$ and $F_Y(y)$, we can generate a sample from $C_N(F_X(x), F_Y(y))$ using the following procedure

1. Sample from the collection of two random variables $\mathbf{Z} \sim \mathscr{N}(\mathbf{0}, \mathbf{R})$ using the Cholesky factorization approach in the previous chapter.
2. Compute $u = \Phi(z_1)$ and $v = \Phi(z_2)$.
3. The samples are $x = F_X(u)$ and $y = F_Y(v)$.

Therefore, via the normal copula we can create a joint distribution that has a prescribed Pearson correlation where the underlying marginal distributions do not have to be normal. This is different than saying that the two variables are a multivariate normal with a known correlation. Note that the matrix $\mathbf{R}$ has only one degree of freedom because the diagonal is 1 and it is symmetric, we can call this degree of freedom ρ. It can be shown that for a normal copula, the value of Kendall's tau is

$$\tau(X,Y) = \frac{2}{\pi} \arcsin \rho, \qquad F(X,Y) = C_N(F_X(x), F_Y(y)). \tag{3.15}$$

Therefore, given a desired value of Kendall's tau for the joint distribution, one can produce it using the normal copula.

The normal copula has zero tail dependence: as one variable approaches $\pm\infty$ the probability that the other variable does the same goes to zero. Therefore, if we are modeling a system where tail dependence could matter greatly, e.g., analyzing how the system behaves under input variables near their extremes, the normal copula may not be appropriate. In fact the normal has been blamed for the financial crisis of 2008 [Jones(2009)] because it does not account for the fact that mortgage defaults, while not being correlated under normal circumstances, have strong lower tail dependence because if everyone in a neighborhood is foreclosed, the housing prices fall, and more mortgages then default. A fact that risk assessors never understood. This needs to be carefully analyzed when quantifying uncertainty in a physical system. In many cases tail dependence could be present, and we need to understand how this may affect our predictions.

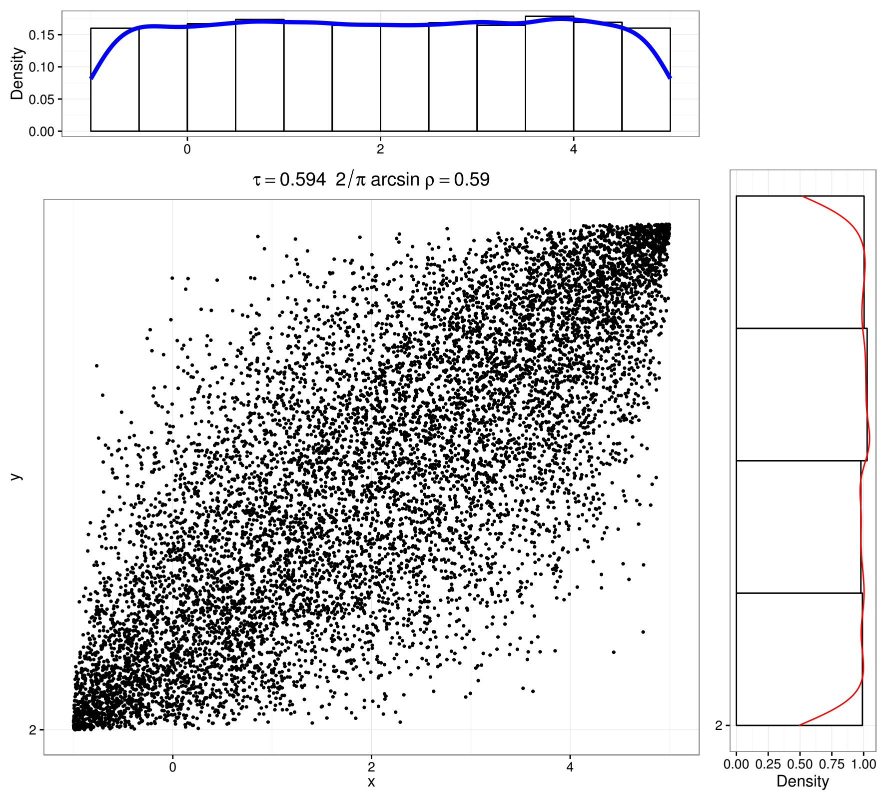

Fig. 3.3 Samples from uniform random variables $X \sim \mathscr{U}(-1,5)$ and $Y \sim \mathscr{U}(2,3)$ joined by a normal copula with $\rho = 0.8$. From these 10^4 samples the empirical value of τ and the predicted value from Eq. (3.15) are shown also.

In Figure 3.3 two uniform distributions are joined by a normal copula with $\rho = 0.8$ are shown. Notice how there is a clear correlation between the two random variables and, as a result, a clustering in the corners of the distributions. An important property of these samples is that they are not normal, we have just used a normal copula to join them.

3.2.2 t-Copula

A distribution similar to the normal is the t-distribution: it is unimodal but has more kurtosis than a normal random variable. This distribution can be used to define a t-copula with a scale parameter $\nu > 0$, and a positive definite, symmetric scale matrix $\mathbf{S}$ with a diagonal of ones as

$$C_t(u,v) = F_{\mathbf{t}}(F_t^{-1}(u), F_t^{-1}(v)), \tag{3.16}$$

where $F_{\mathbf{t}}$ is the joint CDF for a t-distribution with parameters $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{S}$ and ν. The CDF $F_t(x)$ is the CDF of the t-distribution with parameter ν. The degree of freedom in the $\mathbf{S}$ matrix will be written as r.

To sample from random variables joined by the t-copula we use a similar procedure to that for the normal copula:

1. Sample from the collection of two random variables $\mathbf{Z} \sim \mathscr{N}(\mathbf{0}, \mathbf{S})$ using the Cholesky factorization approach in the previous chapter.
2. Compute $\hat{Z} = \sqrt{w}Z$, where w is a sample from the inverse gamma distribution, $W \sim \mathrm{IG}(\nu/2, \nu/2)$.
3. Compute $u = F_t(z_1)$ and $v = F_t(\hat{z}_2)$.
4. The samples are $x = F_X(u)$ and $y = F_Y(v)$.

The t-copula has the same form for Kendall's tau as the normal copula. In particular if we replace $r \to \rho$ in Eq. (3.15) we can relate Kendall's tau to the matrix $\mathbf{S}$.

In Figure 3.4 two uniform distributions are joined by a normal copula with $r = 0.8$ are shown. Notice how there is a clear correlation between the two random variables and, as a result, a clustering in the corners of the distributions. Notice there are more samples farther off the diagonal than in the normal case. This is due to the fact that the t-distribution with a small value of ν has more kurtosis than a normal distribution. Therefore, it is more likely to get anti-correlated values as samples. The fact that the t-copula has tail dependence can also be observed in this figure in the concentration of points near the lower-left and upper-right corners.

The tail dependence can be seen even more clearly if we use a t-copula to couple two

3.2.3 Fréchet Copulas

The Fréchet copula C_L and C_U are simple copulas that join random variables with Spearman correlation ± 1. Furthermore, any other copula is bounded by the relation $C_L \le C \le C_U$. The Fréchet copulas are

$$C_L(u,v) = \max(u + v - 1, 0), \qquad C_U(u,v) = \min(u,v). \tag{3.17}$$

C_L will give perfect negative dependence between variables and C_U will give perfect positive correlation between variables. We can then combine Fréchet copulas to

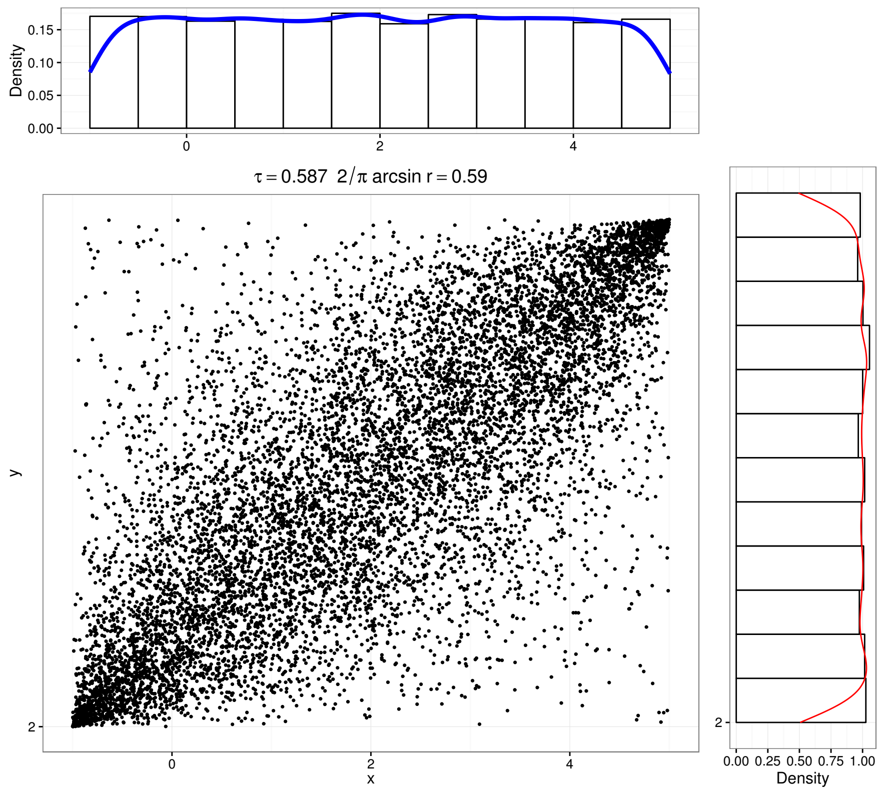

Fig. 3.4 Samples from uniform random variables $X \sim \mathscr{U}(-1,5)$ and $Y \sim \mathscr{U}(2,3)$ joined by a t-copula with $r = 0.8$ and $\nu = 4$. From these 10^4 samples the empirical value of τ and the predicted value from Eq. (3.15) are shown also.

describe something with a Spearman correlation between $[-1,1]$:

$$C_A(u,v) = (1-A)C_L(u,v) + AC_U(u,v), \qquad A \in [0,1]. \tag{3.18}$$

These are a simple combination and can give a Spearman correlation given by $2A-1$.

3.2.4 Archimedean Copulas

There is another class of copulas that easily generalize to an arbitrary number of dimensions and have an explicit formula. These copulas, called Archimedean copulas, and they are defined by a generator function, $\varphi(t)$ for $t \in [0,\infty)$. Given a generator, we define the quasi-inverse

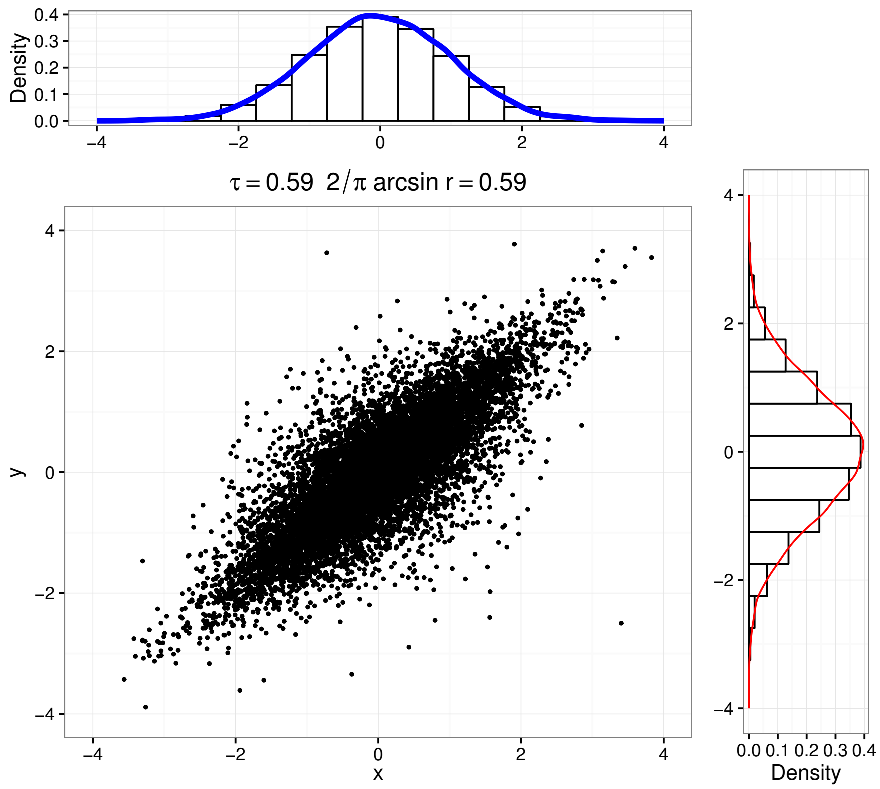

Fig. 3.5 Samples from standard normal random variables $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ joined by a t-copula with $r = 0.8$ and $\nu = 4$. From these 10^4 samples the empirical value of τ and the predicted value from Eq. (3.15) are shown also. Note the tail dependence: when one variable is close to ± 4 the other variable is also likely to be close to ± 4.

$$\hat{\varphi}^{-1}(t) \equiv \begin{cases} \varphi^{-1}(t) & 0 \le t \le \phi(0) \\ 0 & \phi(0) < t < \infty \end{cases}. \tag{3.19}$$

With the generator and quasi-inverse the Archimedean copula for $\phi(t)$ is

$$C_\phi(u,v) = \hat{\phi}^{-1}\left(\phi(u) + \phi(v)\right). \tag{3.20}$$

The term Archimedean arises from the development of the triangle inequality for probability spaces, in that context Archimedes of Syracuse's name is attached a particular norm that has the form of Eq. (3.20).

Archimedean copulas are commutative:

$$C_\phi(u,v) = C_\phi(v,u),$$

associative

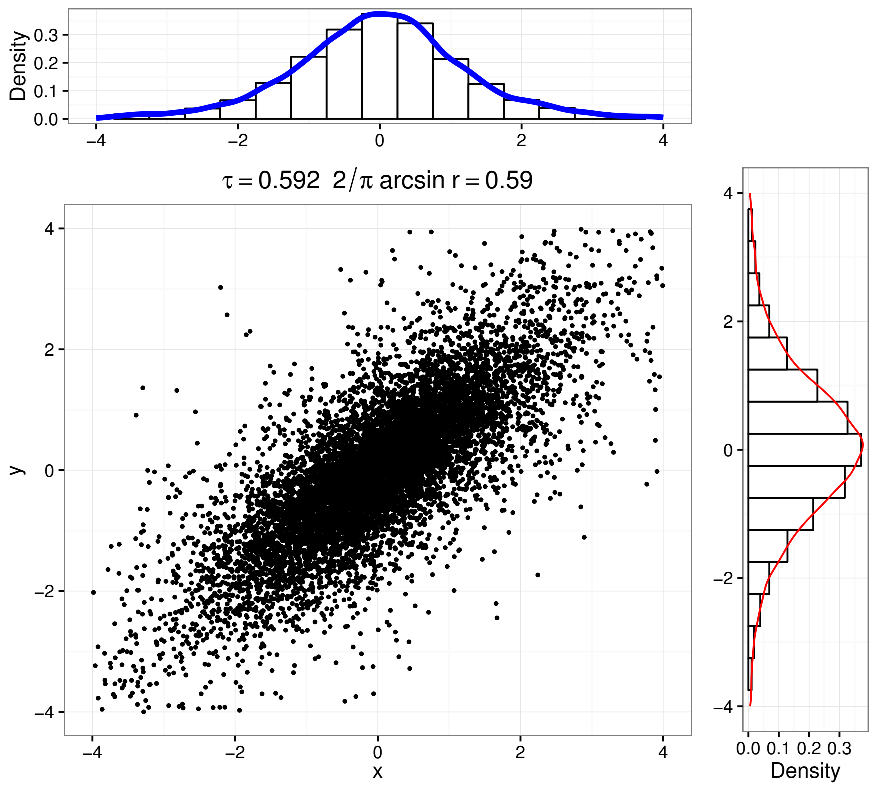

Fig. 3.6 Samples from standard normal random variables $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ joined by a normal copula with $\rho = 0.8$ and $\nu = 4$. From these 10^4 samples the empirical value of τ and the predicted value from Eq. (3.15) are shown also. Note the lack of tail dependence in the lack of concentration near the upper right and lower left corners.

$$C_\phi(C_\phi(u,v),w) = C_\phi(u,C_\phi(v,w)),$$

and are order preserving

$$C(u_1,v_1) > C(u_2,v_2), \qquad u_1 > u_2, \quad v_1 > v_2.$$

The associative property will be used later to easily create Archimedean copulas for arbitrary numbers of variables.

Furthermore, an Archimedean copula can be related to Kendall's tau via the formula

$$\tau(U,V) = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}\,dt. \tag{3.21}$$

There are many Archimedean copulas one could define, we will discuss two below that are commonly used.

3.2.4.1 The Frank Copula

One common Archimedean copula is the Frank copula. This copula has a single parameter, $\theta \neq 0$, and a generator function given by

$$\varphi_{\mathrm{F}}(t) = -\log\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right). \tag{3.22}$$

The quasi-inverse is

$$\hat{\varphi}^{-1}(t) = -\frac{1}{\theta}\log\Big(1+e^{-t}(e^{-\theta}-1)\Big). \tag{3.23}$$

This makes the copula

$$C_{\mathrm{F}}(u,v) = \frac{1}{\theta}\log\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right). \tag{3.24}$$

One property of the Frank copula is that as $\theta \to \infty$, the copula becomes the upper Fréchet copula: $C_{\mathrm{F}} \to C_U$. As $\theta \to -\infty$, then the Frank copula approaches the lower Fréchet copula: $C_{\mathrm{F}} \to C_L$.

The value of Kendall's tau for a Frank copula is can be calculated from the Eq. (3.21) as

$$\tau_F(U,V) = 1 - \frac{2\left(3\theta^2 - 6i\pi\theta + 6\theta - 6\theta\log\left(e^{\theta}-1\right) - 6\mathrm{Li}_2\left(e^{\theta}\right) + \pi^2\right)}{3\theta^2}, \tag{3.25}$$

where $\mathrm{Li}(z)$ is the polylogarithm function. A table for matching a desired value of τ_{F} to θ is given in Table 3.2.4.1. Additionally, the value of τ_{F} as a function of θ is shown in Figure 3.7.

τ_{F}	θ
0.1	0.907368
0.2	1.860880
0.3	2.917430
0.4	4.161060
0.5	5.736280
0.6	7.929640
0.7	11.411500
0.8	18.191500
0.9	26.508600

Table 3.1 The corresponding value of θ for different values of Kendall's tau using the Frank copula. Note that negative values of τ_{F} will have a corresponding negative value of θ.

The Frank copula has a tail dependence of zero. In that sense it is similar to the normal copula.

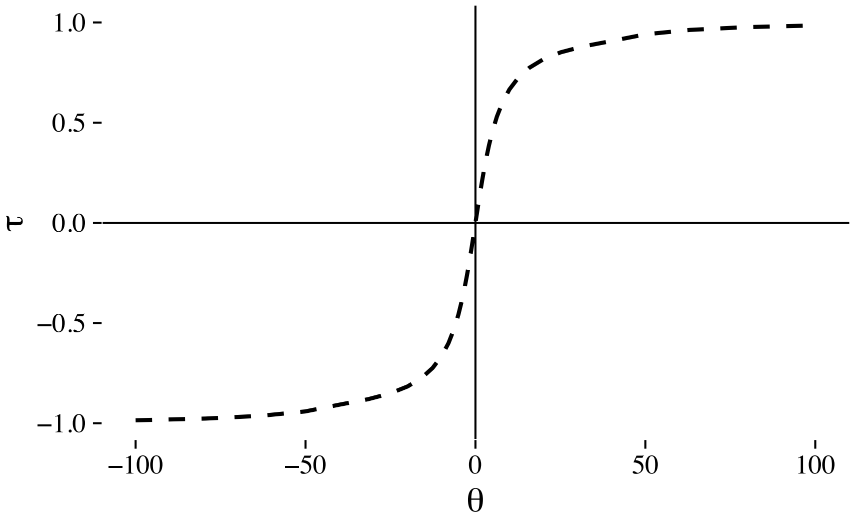

Fig. 3.7 Kendall's tau as a function of θ for the Frank copula.

3.2.4.2 The Clayton Copula

The Clayton copula has a single parameter, $\theta \geq -1$ and $\theta \neq 0$, with generator function

$$\varphi_C(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right), \tag{3.26}$$

and quasi-inverse

$$\hat{\varphi}_C^{-1}(t) = (1 + \theta t)^{-1/\theta}. \tag{3.27}$$

The resulting copula is

$$C_C(u,v) = \left[\max\left\{u^{-\theta} + v^{-\theta} - 1, 0\right\}\right]^{-1/\theta}. \tag{3.28}$$

The Clayton copula has Kendall's tau for the resulting joint distribution

$$\tau_C(U,V) = \frac{\theta}{\theta + 2}. \tag{3.29}$$

Additionally, the Clayton copula has zero upper tail dependence and non-zero lower tail dependence:

$$\lambda_l = 2^{-1/\theta}. \tag{3.30}$$

We can use the Clayton copula to produce joint distributions with upper tail dependence and no lower tail dependence by using the copula $C_C(1-u, 1-v)$.