Demystifying copulas and implications for diversification benefit

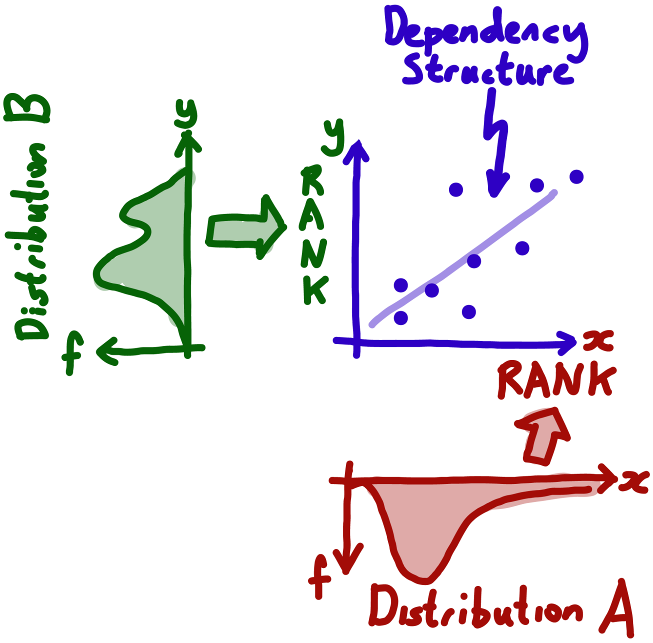


Fig 1: Illustration of what a copula is. It is a way of describing how ranks of two (or more) observed quantities (e.g. A and B) relate to each other, called their '*dependency structure*'. The shape of the distributions that are joined is irrelevant to how they are joined! A and B are shown as frequency distributions, and their data can be ranked.

Scenario & Relevance: A firm has assumed complete independence between 7 lines of business (A to F), and you're sceptical as they quantified tail dependency with copulas last year. Before talking to them, you would like to:

- remind yourself of correlation coefficients (i.e. Pearson, Spearman)
- really get your head around these things called 'copulas'
- estimate how much of an effect pairwise (i.e. between A and B) dependency *could* have upon their risk (e.g. at 1-in-200 level) to decide how important it is.

Summary: The aims of this practical are:

1. to demystify the essence of what a copula is, ignoring the underlying maths.
2. to better understand and quantify '*diversification benefit*'.

Specifically, you will follow instructions to:

- Simulate two correlated Gaussian 'random variables', and plot their ranks (**Task 1**) – this *is* a Gaussian copula.
- Investigate how Pearson's and Spearman's *r* relate to the Gaussian copula and start thinking about how tail risk, e.g. 1-in-200 year level, is affected (**Task 1**).
- Calculate the level of diversification benefit at the 1-in-200 year return period (**Task 2**).
- Investigate how this varies with (i) marginal distribution (ii) correlation *r* (**Task 2**).

Feedback & Support: An answer sheet is available or ask the trainer in your session.

Task 1: My first copula!

Aim: The aim of this task is to demystify the essence what a copula is, *via* a hands-on exercise, focussing upon concepts rather than the underlying maths.

Background Information

- Pearson's *r* reflects the whole of the data (i.e. dependency & marginal distribution shape)
- Normal and Gaussian are two names for the same distribution.
- In a distribution, the largest losses are said to form the '*upper tail*'. When these largest losses are stretched out or increased in value in comparison to a Normal distribution so that if there are more large losses, it is said to be '*heavy tailed*'.
- Spearman's is a rank correlation, reflecting only the *dependency* (i.e. the structure by which losses are linked)
- Copulas are functions that describe only the *dependency* structure, which are then applied to data (e.g. 'A' and 'B' Fig. 1).

Specifically you will

- Simulate two correlated Gaussian 'random variables' (*X* and *Z*) and plot their ranks – this *is* a Gaussian copula.
- Then, apply this copula to data (i.e., *A* and *B*)
- investigate how measures of correlation (i.e. *r*) relate to copulas, and start thinking about how tail risk (e.g. 1-in-200 year level) might be affected.

Practicalities

- Use the EXCEL sheet provided – tab '`Task 1`'.
- To answer the questions, follow the instructions on how to create your first copula.

To create your first copula you will follow the instructions below to complete the spreadsheet, filling in the cells highlighted in light blue as you work through. In the other cells some data, formulae and charts have been provided for you. Do the actions in the numbered steps. Text that is not in the numbered items describes more broadly the process.

Prepare your spreadsheet.

1. Since you are using random numbers, it is probably best to set the sheets to manually recalculate (i.e. only recalculate when you specifically request rather than when you make any chance to a cell; ⌘ + = on a mac, f9 on a Windows machine).

Find the idealized 'data' for this exercise and add it to the EXCEL sheet that simulates a dependency by doing the following steps.

2. Type 'A' and 'B' into `cells C26` and `D26` of the '`Task 1 - Copula`' tab. This pulls through data from '`Data`' tab into `columns C and D`. Feel free to verify this.
3. Calculate Pearson's *r* in `cell B6`, e.g. using `=CORREL()`.
4. Calculate Spearman's *r* for A and B. Ranks of A and B are in `columns R & U` (they are needed later in this exercise). Answer **Q1**.
   - *Tip* – with no ties, Spearman's is =CORREL() of the ranks. See ***Session 2*** if you are unsure about this.

Now, completely ignore the data A and B, and create a Gaussian copula (i.e. a set of linked ranks in Columns O and P) that you will use to link the data later. In overview, you start by creating two uncorrelated variables (X & Y), then for each 'event' take some of X and some of Y in a weighted sum to create Z. Obviously, having some of X in it, Z is then correlated with X.

5. Set the correlation level you want to simulate by putting a value between 0 and 1 into cell F7 – 0.9 is used first in the questions.
6. Create X, a 'standard Normal' random variable, with mean of zero and standard deviation of 1.0.
   - Use =RAND() to generate random numbers equally distributed between 0 and 1 in column G.
   - Use =NORM.INV() to turn this into a Normal distribution (i.e. 'bell-shaped curve') in column H.
   - If in doubt, see how it is done for Y, i.e. in columns J to L.
   - See ***Session 2*** if you are still unsure how to do this.
7. Using detail in the bullet points below, add some of X to some of Y to create a Gaussian distribution Z that is partly correlated to X (i.e. Z = w_1X + w_2Y where w_1 and w_2 are appropriate weights in cells F9 and F8 respectively). Simply, Z is partly made up of X so must be related to, i.e. correlated with, X. Don't worry overly how the weights (w_1 and w_2) for this mixture are derived.
   - X is in column H. Put $X*w_1$ in column I.
   - w_2Y is calculated for you in column L. Put Z = w_1X + w_2Y into column M (i.e. column I + column L).
8. Now, calculate the ranks of Z in column P. See column O if in doubt how to do this.

Congratulations, you've created a copula. Examine the scatter plot directly above the table 'Dependent Ranks (i.e. X *vs* Z)' to see what it looks like. Now answer **Q2 – Q6**, and you're ready to apply it to your data below.

Your Gaussian copula is now in columns O and P.

Now, apply the copula to your data (i.e. to A and B). To do this, A must be ranked according to X and B according to Z. In other words, it is necessary to select a value from A that has a rank given by (i.e. matching) the rank of X, this has been done for you in column T.

9. To finish applying the copula to your data (i.e. to link A and B), select a value from B that has a rank given by (i.e. matching) the rank of Z in column P. Let's call the dataset where B is correlated B'.
   - In column W use =VLOOKUP() to do this. It has been done for you for dataset A to create dataset A' in column T, so copy what is done there.

10. To allow us to work out the Spearman's rank of the simulated correlation (cell F20), put the ranks for A' in column Y, and ranks for B' in column Z. Now, answer **Q7**

[Columns R,S,U,V are where they are simply to accommodate the way =VLOOKUP() works]

Finally, apply your copula to some different data. You might want to copy the spreadsheet or take screenshots before changing the data so that you can compare answers.

11. Copy input data C and D into the Task 1 – Copula' tab, and answer **Q8 & Q9**.

Questions

Q1: What is the rank correlation of the first two lines of business that you have been given 'A' and ‘B’?

Q2: Summarize briefly, in a sentence each, the three main steps you have undertaken to create a Gaussian copula.

Q3: Has creating the copula got *anything* to do with the data A and B?

Q4: In your own words, in a short paragraph, describe what a copula is.

Q5: Describe the shape of a Gaussian copula. What does it look like for a strong dependency with ρ = +0.9? And, for a weak dependency of ρ = +0.3? In particular comment on the upper right of the scatter plot (i.e. upper-tail relating to the largest risks) and the lower tail in the bottom left corner. Words or quick sketches are fine.

- *Tip* – remember you can change r in `cell F7`.

Q6: What do you think a copula with no lower tail dependence by high upper tail dependence might look like? Can you name any copulas like this?

Q7: For Normally distributed data A & B, what are the Spearman’s and Pearson’s *r* values for the simulated ***data*** in `columns T & W` linked with a strength of ρ = 0.6? See `cells F19 & F20`.

Q8: For the heavy-tailed data (i.e. C & D), what are the Spearman’s and Pearson’s *r* values for the ***data*** with a dependency in the copula of ρ = 0.7. And, why do they differ?

Q9: Examine the scatter plot of dependent values (A’ *vs* B’, and C’ *vs* D’) along with the *r* values. From these simulations, do you think that having a heavier tail (e.g. log-Normal) will cause a larger or smaller effect on 1-in-200 year joint loss?

For more about why you might care about copulas, correlation as they impact diversification benefit at the 1-in-200 level please proceed to **Task 2**.

Task 2: Diversification Benefit: Differences are useful!

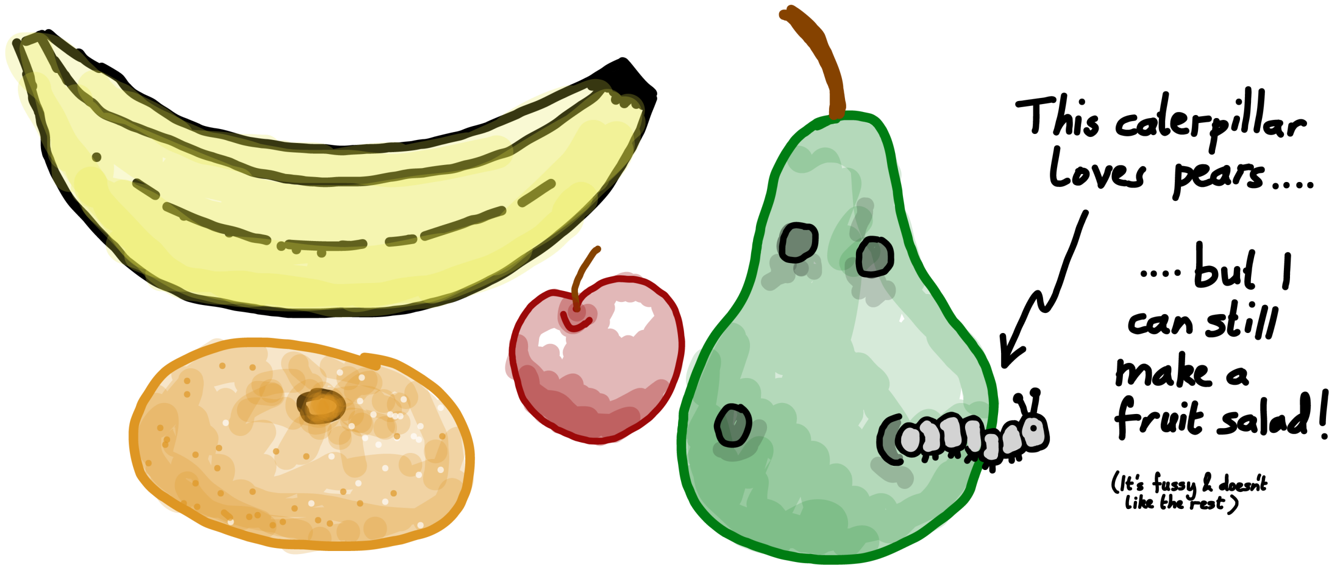
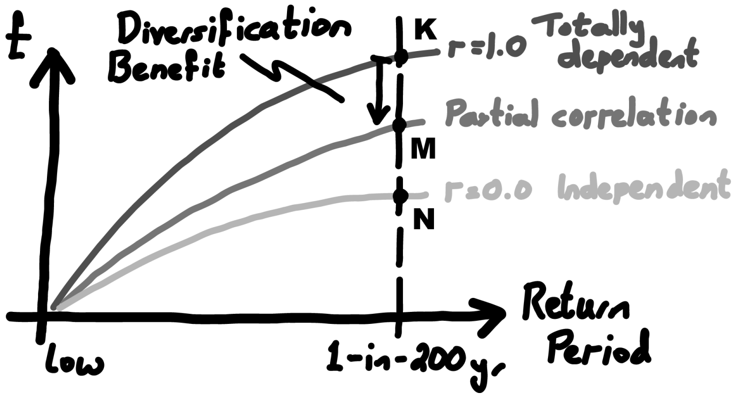

This caterpillar Loves pears....
.... but I can still make a fruit salad!
(It's fussy & doesn't like the rest)

Fig 1: Illustration of why diversification, the holding of assets that are impacted differently, at different times, can be useful. In short, it makes it less likely that everything will go bad at once, reducing the size of the maximum likely losses.

Aim: To better understand quantify 'diversification benefit' (i.e. reduction of 1-in-200 year loss due to losses being somewhat independent) and how correlation *r* and the marginal distribution (i.e. heavy tail or not?) can affect this.

Background Information

Diversification is the antithesis of accumulation or correlation, and it is the fundamental reason that insurance can exist[1]. So, it is of interest to quantify and understand the benefit that diversification in terms of reducing the size of the largest likely risks (e.g. at the 1-in-200 year return period).



Consider two loss distributions *A* and *B*. If, and *only* if, they are totally dependent (i.e. ranks always match, *r* = 1.0) the 1-in-200 losses can simply be added together. Arbitrarily call this value *K*. For any partial correlation the combined 1-in-200 loss, arbitrarily called *M* here, will be less than *K*. A lowest combined value is for independence.

The reduction from *K* is the *diversification benefit*, and can be an absolute value (i.e. $K - M$) or a fraction (i.e. $1 - M/K$). This depends on

- The relative magnitude of the two loss types.
- Level of correlation
- Other aspects of how the two are linked, the '*dependency structure*' (e.g. tail dependence) although this is not covered in this Session.

[1] For more detail on diversification, see pages 4-5 of Mitchell-Wallace et al (2017) '*Natural Catastrophe Risk Management and Modelling*'.

Practicalities

- Task 2 uses the output of Task 1 (i.e. the ability to simulate a Gaussian copula for the data provided).
- Use the EXCEL sheet provided – tab ‘`Task 2 – Diversification`’. Again, the cells highlighted light blue are those you need to fill.
- To answer the questions below, follow the instructions that are given below them. Again, actions are in the numbered points.

To examine the effects of correlation on *diversification benefit*: Complete the spreadsheet by filling in the cells highlighted in light blue. In the other cells some data, formulae and charts have been provided for you.

1. To record your existing beliefs, answer **Q1**.

To start the numerical work, we will use the output of Task 1, and need to reset this to use the simplest type of data i.e. A & B

2. Copy loss distributions A and B from the ‘`Data`’ tab into `columns C and D` of the ‘`Task 1 - Copula`’ tab.
3. Manually, or otherwise, ensure that EXCEL sheet has recalculated cell values.

Now return to the ‘`Task 2 – Diversification`’ tab.

First, consider the TOTALLY DEPENDENT case in `columns L to U`. Traditionally, this is the baseline against which benefit due to diversification is calculated (see figure in Background Information).

4. It is simple to induce a perfect correlation (i.e. *r* = 1.0). Sort both datasets in descending order. Then rank 1 in A is paired with rank 1 in B on the same row, in this case row 28, and so on.
   - We want this to update automatically rather than manually (i.e. ‘Data’ -> ‘sort’ -> ‘largest to smallest’ etc …), so use `=VLOOKUP()`
   - A is done for you (`Column N`)
   - Sort B similarly (`Column Q`)
5. In `cell L6` calculate the Pearson’s *r* value between and B (`columns N & Q`) using `=CORREL()`.
   - If it is not 1.0, please check your spreadsheet for errors.
6. However, we don’t care about correlation *per se*, we care about it as an indicator of how joint losses might behave. So, calculate the joint annual loss. Sum the sorted values of A and B into `Column T` (i.e. =N28+Q28).
7. Now, let’s select the 1-in-200 year value.
   - Answer **Q2 to Q4** and fill in `cells M9 to M11`.
     - And, yes, these questions are just a reminder about assumed background knowledge.
   - Use `=VLOOKUP()` to select the 1-in-200 year value of the joint losses (in `Column T`) into `cell P9`
     - `=VLOOKUP(M11,S$28:T$1027,2,FALSE)`
   - The joint losses (i.e. A+B) have been sorted for you in descending order in `Column U`. Check that you’re value in `P9` is correct, and answer **Q5**.

Now, consider the INDEPENDENT case in `columns B to J`. In terms of insurance, this is the best case scenario with maximum diversification.

8. Before starting, answer **Q6**.
9. In `cells B6 &B7` calculate the Pearson’s and Spearman’s *r* values using `=CORREL()`.

- If you are unsure, copy what was done in `Task 1`.
- *Tip* - The relevant columns are indicated in the sheet.

10. Add (`column I`) then rank, and sort A and B in `columns H and J`, exactly as in the dependent case above (i.e. see `columns S to U`).
11. In `cell C9` calculate the 1-in-200 year value as did previously (i.e. `cell P9`)
12. In `cell B11` calculate the diversification benefit of the independent case, using the 1-in-200 year values and looking at the Background Information if necessary. Answer **Q7**.
    - *Tip*: =100*(1-(independent/totally_correlated))

Now, consider the PARTIALLY DEPENDENT case (i.e. 0 > *r* > 1) in `columns W to AC`. This is something that is likely quite common in reality.

13. You have simulated this partially dependent case in '`Task 1 - Copula`' tab, specifically in `Columns T and W`. Insert a formula to put these data in `columns W and X` of the '`Task 2 – Diversification`' tab. (i.e. in `Column W =’Task 1 - Copula'!T31`)
14. In `cells W6 to W11` with light blue shading, add formulae to calculate metrics for this case. Do this as you have for the other cases.
15. Answer **Q8 to Q10**.
    - You might wish to examine visually the diversification benefit in `cell W11` for each 1000 run (i.e. manually recalculate to do another run), or copy out the values one by one to find and mean and its standard error.

You are now set up to examine the effects of marginal distribution on *diversification benefit*.

16. Copy loss distributions 'F' and 'G' from the '`Data`' tab into `columns B and C` of the '`Task 1 - Copula`' tab. These datasets are both log-normally distributed and of roughly equal magnitude. Use them to answer **Q11-Q14** by taking the case of ρ = 0.4.
    - Change *r* in `cell F7` of `Task 1 - Copula`
    - Recalculate to do another simulation run.
    - Examine simulated *r* values in `cells F19 & F20` of `Task 1 – Copula`
    - Examine diversification benefit in `B11` and `W11` of `Task 2 – Diversification`
    - For Q14 alter the size of F by adding *0.5 in the '`Data`' tab, then copy it across.
17. Copy loss distributions 'C' and 'D' from the '`Data`' tab into `columns B and C` of the '`Task 1 - Copula`' tab. C has a lighter tail than D, and they are of somewhat different sizes. Again by taking the case of ρ = 0.4, answer **Q15**. Answering the question is optional, but exploring this case is valuable even if not focussed on answering the question.
18. OPTIONAL – For interest, consider exploring distributions D & E. They are very heavy tailed and identical in size)

Questions

Q1: What do you think will produce the bigger 1-in-200 joint losses, case where types of loss A and B are independent or a case where they are correlated with *r* = 0.7? In 2-3 sentences explain why you think this?

Q2: How many years of data are there?

Q3: What is the probability of a 1-in-200 year loss occurring?

Q4: And so, what rank (i.e. rows form the top of the list) is the 1-in-200 value?

Q5: What is the 1-in-200 year joint loss for A and B?

Q6: From your experience how many percent do you think the maximum diversification benefit will be? And, briefly (i.e. 2-3 sentences) what are you basing this on? Consider the circumstances (e.g. type of data, presumed distribution and copula e.g. Gaussian).

Q7: For the Normally distributed loss data (i.e. A & B) joined by a Gaussian copula, what is the maximum possible diversification benefit as a percentage?

Q8: For this situation, what is the diversification benefit for a correlation of $\rho = 0.4$?

- *Tip* – remember that you can change ρ in the '`Task 1 - Copula`' tab.

[Here, if you are struggling for time, out of Q9-Q15 do Q11 and Q12 first]

Q9: How many simulation runs does it take to reduce the uncertainty in this (i.e. 2 standard errors) to <1%?

Q10: How many years of simulation would you like to give an answer that you are confident is stable?

Q11: Do longer tails (i.e. log-Normal for F & G) affect Pearson's and Spearman's correlation metrics? And if so, how and why? Give numbers for an induced dependency of $\rho = 0.4$ to illustrate your answer.

Q12: Do longer tails (i.e. log-Normal for F & G) affect diversification benefit? And if so, how and why? Give numbers for $r = 0.4$ to illustrate your answer. Compare this to your expectations in Q9 of Task 1.

Q13: How many years of simulation would you like to give an answer that you are confident is stable for F & G? E.g., uncertainty (i.e. 2 standard errors) of <1%, and how does this compare to your answer for Q9?

Q14: What effect does reducing the size of F by half have on correlation and diversification? Give numbers for an induced dependency of $\rho = 0.4$ to illustrate your answer.

Q15: [CHALLENGE] Describe and explain the correlation and diversification for datasets C and D. Given that the dependency structure is identical to Q13, what are the similarities and differences.