Michal Kaut

A Copula-based Scenario Generation Heuristic

SINTEF Technology and Society

September 2011

SINTEF

Outline

Outline

Background – scenario generation

- In *stochastic programming*, we need to input the distributions of the stochastic parameters into the model.
- To use the standard mathematical-programming tools, these have to be discrete distributions with moderate number of support points.
- If we have the distributions in any other format (continuous distributions or discrete distributions with too many points), we have to approximate them.
- The process is called *scenario generation*.
- The goal is to find an approximation that
  - leads to a solution that is close to the true optimal solution,
  - has not too many points, so the model can be solved in a reasonable time.

Controlling the multi-variate 'structure'

With multi-variate random variables, we have to control the way the discrete margins are connected together, or what we call the 'multi-variate structure' of the distribution.

Usually, this is limited to:

- assumption of independence;
- using correlations/covariances.

Controlling the multi-variate 'structure'

With multi-variate random variables, we have to control the way the discrete margins are connected together, or what we call the 'multi-variate structure' of the distribution.

Usually, this is limited to:

- assumption of independence;
- using correlations/covariances.

Correlation (i.e. Pearson's correlation coefficient):

- Only one number for each pair of random variables.
- Captures only linear dependence.
- Does *not* capture any non-linear dependencies.
- Does *not* tell us anything about the "shape".

Controlling the multi-variate 'structure'

With multi-variate random variables, we have to control the way the discrete margins are connected together, or what we call the 'multi-variate structure' of the distribution.

Usually, this is limited to:

- assumption of independence;
- using correlations/covariances.

Correlation (i.e. Pearson's correlation coefficient):

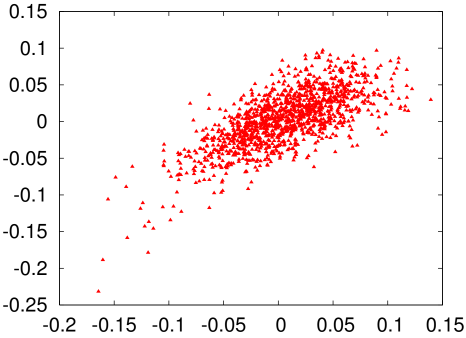
- Only one number for each pair of random variables.
- Captures only linear dependence.
- Does *not* capture any non-linear dependencies.
- Does *not* tell us anything about the "shape".

Using the Pearson correlation often implicitly means assuming the elliptical shape of the normal distribution.

Are data really elliptical?

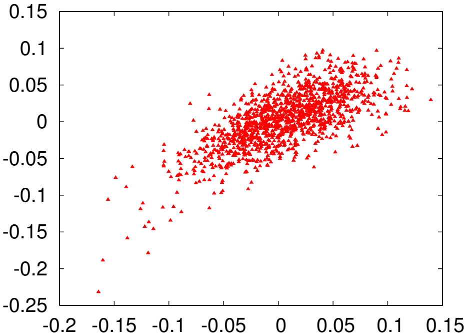
Are data really elliptical?

As an example, consider fortnightly returns of US and UK small cap stocks (data from MSCI):



Are data really elliptical?

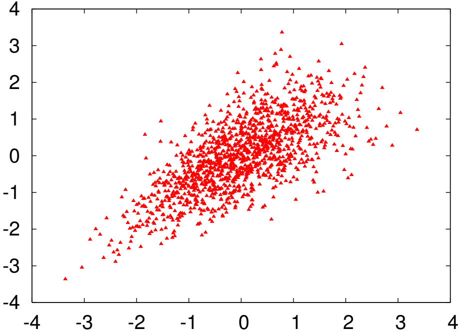
As an example, consider fortnightly returns of US and UK small cap stocks (data from MSCI):



- This could be due to skewed marginal distributions.
- So let's transforms the margins to N(0,1)

Are data really elliptical?

As an example, consider fortnightly returns of US and UK small cap stocks (data from MSCI):



- The asymmetry gets even more pronounced.
- The asymmetry is in the *shape*, not in the margins.

Beyond correlations – what can we do?

Beyond correlations – what can we do?

- There is a tool for describing the multi-variate structure (shape) of a distribution, independent on the marginal distributions copulas.
- The name (proposed by Sklar in 1959) originates for the Latin term for a link or tie, as it describes how margins are linked to form a multi-variate distribution.

Beyond correlations – what can we do?

- There is a tool for describing the multi-variate structure (shape) of a distribution, independent on the marginal distributions copulas.
- The name (proposed by Sklar in 1959) originates for the Latin term for a link or tie, as it describes how margins are linked to form a multi-variate distribution.

- Copulas are not new, they have been used in statistics and finance for quite some time.

Outline

Copula – A formal definition

Formally, a copula is any function $C : [0,1]^n \to [0,1]$ such that

- $C(\boldsymbol{u}) = 0$ if there is a component i such that $u_i = 0$.
- $C(\boldsymbol{u}) = u_i$ if $u_j = 1$ for all $j \neq i$.
- $C(\boldsymbol{u})$ is n-increasing.
  - For definition, see literature or Wikipedia.
  - For n=2, this means that if $u_1 \leq u_2$ and $v_1 \leq v_2$, then

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 .$$

Note that this means that C is a cumulative distribution function (CDF) of an n-dimensional distribution with standard uniform margins.

More info: Bouyé et al. (2000); Clemen and Reilly (1999); Nelsen (1998)

Sklar's theorem

Sklar's theorem then states that for any n-dimensional CDF F with marginal CDFs F_i, there exist a copula C such that

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_n) = C\big(F_1(x_1), \ldots, F_n(x_n)\big) .$$

Moreover, C is unique for continuous marginal CDFs F_i.

An immediate consequence is that, for every $\boldsymbol{u} = (u_1, \ldots, u_n) \in [0, 1]^n$, we get

$$C(u_1, \ldots, u_n) = F\big(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\big) ,$$

where F_i^{-1} is the generalised inverse of F_i.

Copula - invariance to changing margins

Let us have

- random variable $\tilde{X}$ with CDF F
- another CDF G
- both CDFs continuous and strictly increasing

Then

- $\tilde{U} = F(\tilde{X})$ is a strictly increasing transformation.
- $\tilde{Y} = G^{-1}(\tilde{U})$ is a strictly increasing transformation.

This implies that changing the distribution of $\tilde{X}$ from F to G is a strictly increasing transformation and therefore does not change the copula.

Copula - invariance to changing margins

Let us have

- random variable $\tilde{X}$ with CDF F
- another CDF G
- both CDFs continuous and strictly increasing

Then

- $\tilde{U} = F(\tilde{X})$ is a strictly increasing transformation.
- $\tilde{Y} = G^{-1}(\tilde{U})$ is a strictly increasing transformation.

This implies that changing the distribution of $\tilde{X}$ from F to G is a strictly increasing transformation and therefore does not change the copula.

Note that we have already used this fact for the MSCI data, when we transformed the margins to the normal distribution.

Copulas – some more information

- Just like distributions, there are parametric families for copulas.
  - Copulas of the standard distributions, like normal or t-distr.
  - There are also some special copula families, though they are mostly limited to two random variables.
- This means that estimating a copula from data is analogous to estimating a distribution, i.e. we have to
  - choose a parametric family;
  - estimate its parameters.
- In addition, we can speak about an empirical copula of a data set.

Outline

Basic structure of a copula-based method

Assuming we know the marginal CDFs, or have some other way of transforming the margins to their correct distributions, we can do the following:

1. Generate scenarios for the copula, i.e. the multivariate structure with Unif[0,1] margins.
   - sampling from a "historical copula"
   - using some parametric copula family
2. Transform the margins to the desired marginal distributions.

This opens some new possibilities for scenario generation, some of which are described on the following slides:

Combining different (standard) copulas and margins

Example: normal margins with a copula from a t-distribution:
What do we get?

Combining different (standard) copulas and margins

Example: normal margins with a copula from a t-distribution:

What do we get? Tail dependence:

- Probability of the very extremes of different marginal distributions hapenning together
  - Formally, a *lower-tail dependence coefficient* λ_L of two random variables $\tilde{X}_1$, $\tilde{X}_2$ is defined as

$$\lambda_L = \lim_{v \to 0^+} \mathrm{P}\left\{\tilde{X}_1 \leq F^{-1}_{\tilde{X}_1}(v) | \tilde{X}_2 \leq F^{-1}_{\tilde{X}_2}(v)\right\} .$$

- Normal distribution has zero tail dependence.
- t-distributions have positive tail dependence.
- Tail dependence is a function of copula, so we can transform margins from t-distribution to normal distribution, without changing the tail dependence.

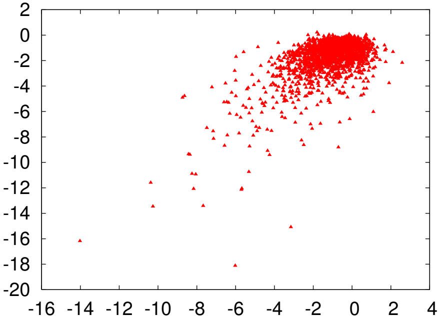
Introducing asymmetry

- t-copula is still elliptic and therefore symmetric.
- There are several skewed versions of t-distribution and some of them have asymmetric copulas.
- The asymmetry allows us, for example, to simulate the higher dependence (correlation) in the down-turns, as we have seen in the MSCI data.
- We then proceed the same way as with the standard t-distribution:
  1. Generate a sample from the chosen skew-t distribution.
  2. Transform its margins to (for example) normal distribution.

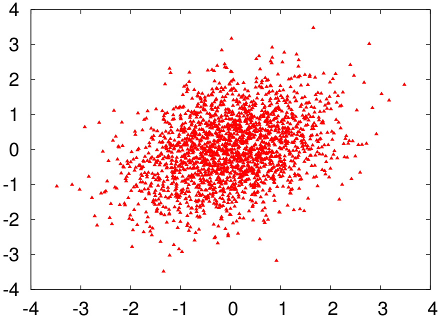
  (All this assuming we know how to choose the appropriate skew-t distribution and estimate its parameters)

Introducing asymmetry – example

Comparing skew-t and normal – the distributions



skew-t distribution

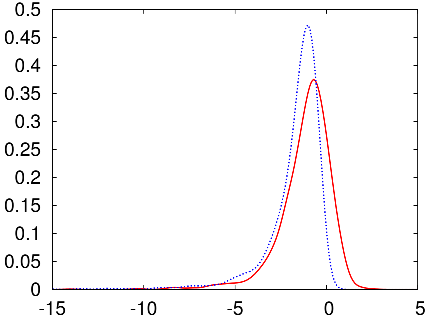


normal distribution

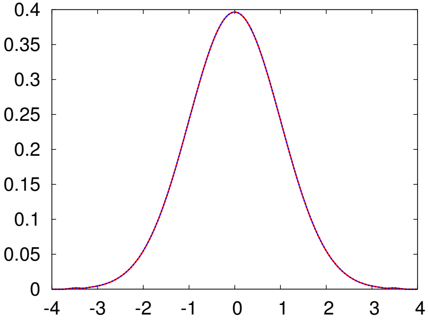
Note: both margins and copulas are different.

Introducing asymmetry – example

Comparing skew-t and normal – the margins



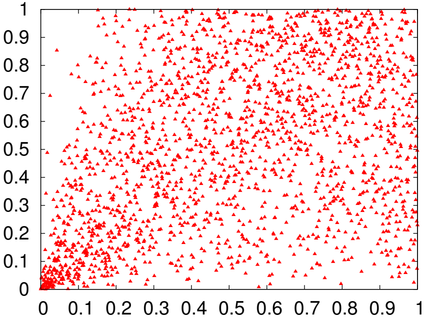
skew-t distribution



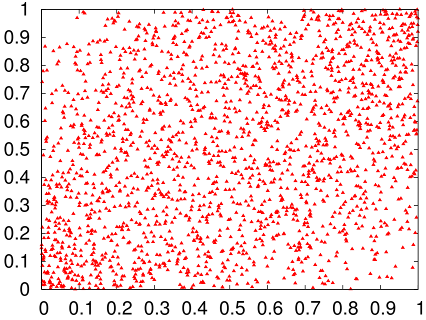
normal distribution

Introducing asymmetry – example

Comparing skew-t and normal – the copulas



skew-t distribution

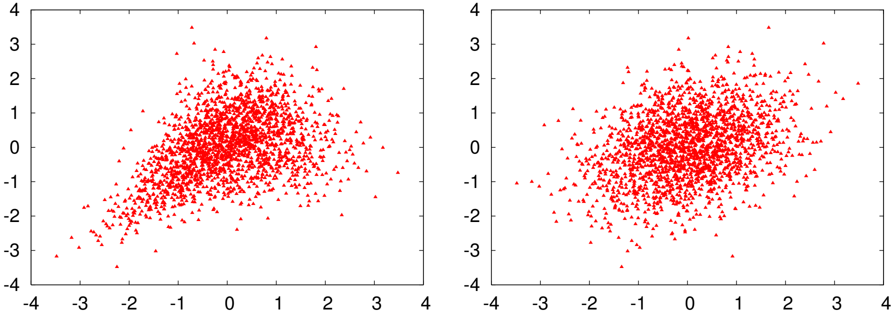


normal distribution

Note the difference in the lower-left corner.

Introducing asymmetry – example

Comparing skew-t and normal – both with normal margins



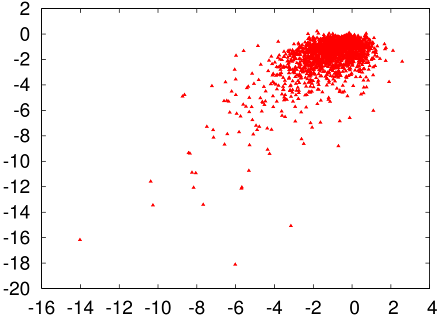
skew-t distribution

normal distribution

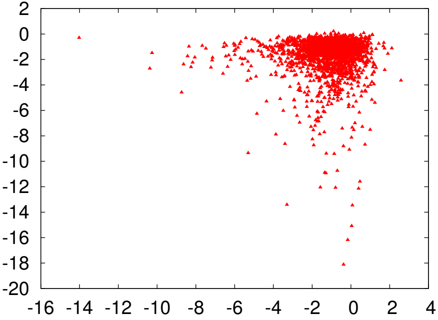
Note the clear difference in *tail dependence*.

Introducing asymmetry – example

Comparing skew-t and normal – both with skew-t margins



skew-t distribution



normal distribution

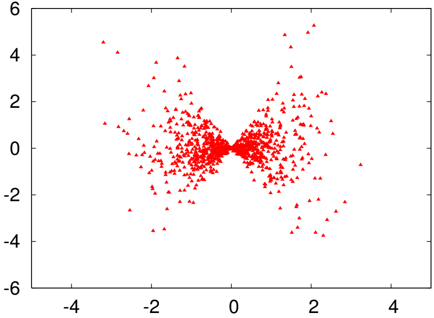
Note the clear difference in *tail dependence*.

Using principal components

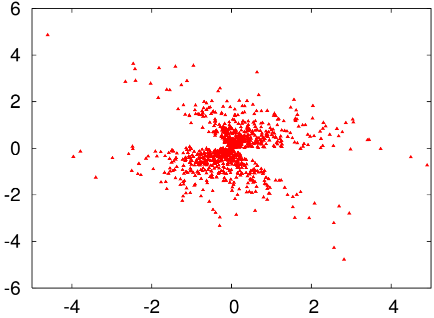
- Principle component analysis (PCA) is a technique used to decrease the dimension of the stochastic vector.
- The principal components are, by construction, orthogonal and therefore uncorrelated.
- Yet, except for the normal case, they are not independent.
  - It is tempting to forget/ignore the distinction and treat them as independent, as it makes the scenario generation significantly easier.
  - Copulas can be used to capture the dependence.
    (While using only correlations can not, obviously, capture the difference between uncorrelated and independent)

Using principal components – example

Let us have two independent random variables $\tilde{\xi}_1, \tilde{\xi}_2 \sim \mathrm{N}(0,1)$ and define $\tilde{x}_1 = \tilde{\xi}_1$, $\tilde{x}_2 = \tilde{\xi}_1 \tilde{\xi}_2$. Then we get:



random vector $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2)$

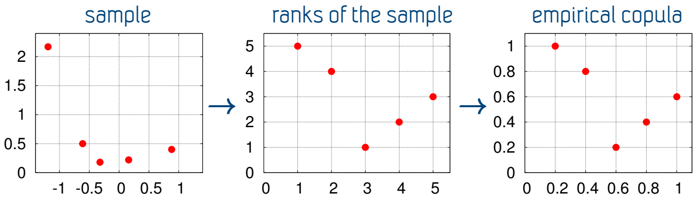


principal components, scaled to $\sigma = 1$

Outline

Copulas as a distribution of ranks

- Instead of correlations, we describe the co-variation using copulas.
- For our purpose, a copula of a given distribution F is a multivariate Uniform $(0, 1)$ distribution with with margins $C_i = F_i(X_i)$.
  - All information about the marginal distributions is removed.
  - All information about co-variation of the margins is intact.
- In a discrete case (scenarios), copula can be equivalently described as coupling of the ranks of the marginal distributions:



A copula-based scenario-generation method

If we have a way to generate a sample from the copula (*in terms of ranks* r_i^s) that is a good approximation of the target copula, the rest is easy:

- If we know the CDFs F_i and they are invertible, the scenario values are simply given as

$$X_i^s = F_i^{-1}\left(\frac{r_i^s - 0.5}{S}\right) .$$

- Otherwise, we generate samples for the margins by some other method, compute the ranks of the sample values and assign them to the scenarios as given by the copula sample.

Multivariate vs. bivariate approach

- For a full description of the co-variation (shape of the distribution), we should use the whole multivariate copula.
- Difficult to work with $\rightarrow$ use bivariate copulas for all the pairs instead.
  - We cannot describe higher-order dependencies
  - Still, it is better than using correlations, since each pair is described by a distribution, instead of just one number.

- The proposed method tries to match a specified subset of all the bivariate copulas.
- It assumes that we can compute target values $C_{i,j}(x_i, x_j)$ for all the specified pairs (i, j).

Matching a bivariate copula

What do we mean by 'matching a bivariate copula', for some (i, j)?

- We want to create pairs (r_i^s, r_j^s) for $s \in \{1, \ldots, S\}$, with $r_i^s, r_j^s \in \{1, \ldots, S\}$ and each value appearing exactly in once.
- The CDF of each (r_i^s, r_j^s) is given as

$$\left|\{s' : r_i^{s'} \leq r_i^s \;\&\; r_j^{s'} \leq r_j^s\}\right| / S\,.$$

- The target value is then

$$C_{i,j}\left(\frac{r_i^s - 0.5}{S}, \frac{r_i^s - 0.5}{S}\right).$$

- We then try ti minimize the total distance from the target.
  - In our case, we use the sum of squares of the differences.

Exact matching using integer programming

- The matching problem can be formulated as an integer-programming model, with $2\,S^2$ binary variables.
- Solvable only for small number of scenarios and this is just one bivariate copula, we need to match up to $N(N-1)/2$ of those.
- (The model can be formulated also for the multivariate case, but that would mean $N\,S^2$ binary variables)

Exact matching using integer programming

- The matching problem can be formulated as an integer-programming model, with $2\,S^2$ binary variables.
- Solvable only for small number of scenarios and this is just one bivariate copula, we need to match up to $N(N-1)/2$ of those.
- (The model can be formulated also for the multivariate case, but that would mean $N\,S^2$ binary variables)

→ We need a heuristic (Kaut, 2011).

Outline

Pseudo-code of the matching heuristic

```
initialize the first margin
for s ← 1...S do
    r_1^s ← s
end

generate the second margin
for r ← 1...S do
    for s ← 1...S do
        compute CDF of pair (r_1^s, r)
        compute the distance/error of s
    end
    r_2^s ← s that minimizes the error
end

generate all the other margins
```

More info about the heuristic

- Other margins generated just like the second one, with the distance/error computed w.r.t. all the already generated margins.
- Random selection of s in case of several minima.
- Can be made more random by choosing from the k best candidates at each iteration.

Outline

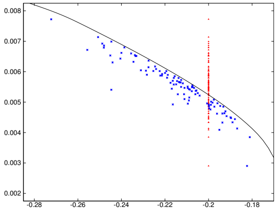
Test case – portfolio optimization

A simple portfolio optimization problem with CVaR constraints.

- CVaR was chosen because it can be expected to react to differences in the shape of the distribution, particularly in the lower tail.
- As data, we used daily prices of 10 assets (seven stock indices and three government bonds), for 4476 days.
- Based on the historical data, we generate scenarios using several different methods:
  - sampling with correction of means and variances
  - moment-matching alg. from Høyland et al. (2003)
  - the new copula-matching code, with the empirical bivariate copulas as targets; the margins' CDFs are then set using the empirical CDFs.
- As an *out-of-sample test*, the expected profit and CVaR of the portfolios are then evaluated using the whole data set.

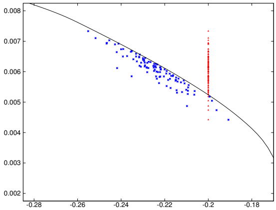
Test case – results for 50 scenarios

sampling



new code

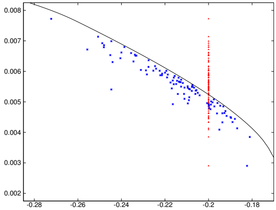
moment-matching



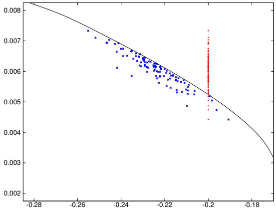
new code, 25–100 sc.

Test case – results for 50 scenarios

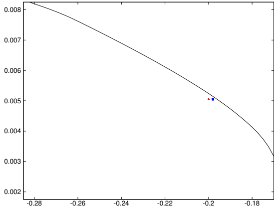
sampling



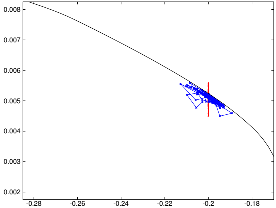
moment-matching



new code

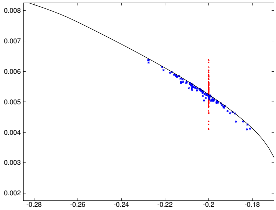


new code, 25–100 sc.

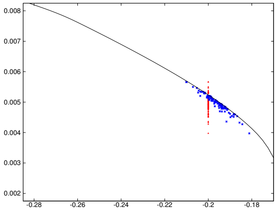


Test case – results for 250 scenarios

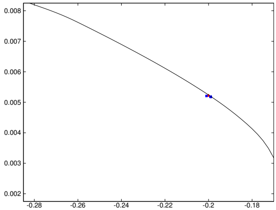
sampling



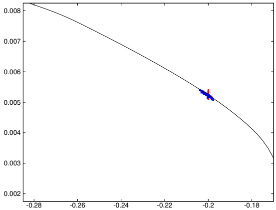
moment-matching



new code

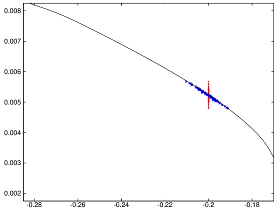


new code, 200–300 sc.

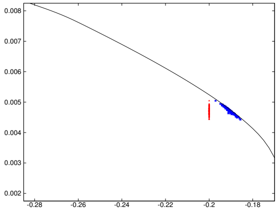


Test case – results for 1000 scenarios

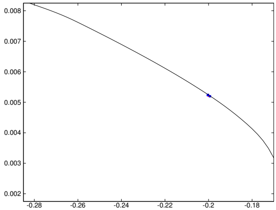
sampling



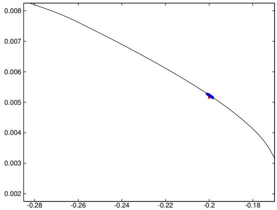
moment-matching



new code



new code, 900–1000 sc.



Test case – summary of the results

- The new copula-based method clearly outperforms both sampling and moment matching, both in terms of stability and out-of-sample value of CVaR.
- Even with just 25 scenarios, there is very little bias (systematic error) in the out-of-sample CVaR, once we fix the means and variances of the margins.
  - This is not bad, given that we have 95% CVaR and 10 assets (stochastic variables).
- Similar results with another data set.

Summary

- We have presented a method to generate scenarios for a multivariate copula, using its bivariate components.
- Together with some methods to generate samples from the marginal distributions, this makes a new way of generating scenarios.
- Our test case suggests that the method outperforms both sampling and moment matching both in terms of stability and (out-of-sample) error/bias.
- Generally, we expect this method to perform well for stochastic programs with with data that exhibit non-elliptical copulas.

The End

For Further Reading I

Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:367–389, 2003. doi: 10.1111/1467-9868.00391.

Eric Bouyé, Valdo Durrleman, Ashkan Nikeghbali, Gaël Riboulet, and Thierry Roncalli. Copulas for finance: A reading guide and some applications. working paper, Crédit Lyonnais, Paris, 2000. Available at SSRN: `http://ssrn.com/abstract=1032533`.

Robert T. Clemen and Terence Reilly. Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224, February 1999.

Kjetil Høyland, Michal Kaut, and Stein W. Wallace. A heuristic for moment-matching scenario generation. *Computational Optimization and Applications*, 24(2–3):169–185, 2003. doi: 10.1023/A:1021853807313.

Ling Hu. Dependence patterns across financial markets: A mixed copula approach. *Applied Financial Economics*, 16(10):717–729, 2006. doi: 10.1080/09603100500426515.

For Further Reading II

Michal Kaut. A copula-based heuristic for scenario generation. `http://www.optimization-online.org/DB_HTML/2011/09/3169.html`, 2011.

Michal Kaut and Stein W. Wallace. Shape-based scenario generation using copulas. *Computational Management Science*, 8(1–2):181–199, 2011. doi: 10.1007/s10287-009-0110-y.

Michal Kaut, Stein W. Wallace, Hercules Vladimirou, and Stavros Zenios. Stability analysis of portfolio management with conditional value-at-risk. *Quantitative Finance*, 7(4):397–409, 2007. doi: 10.1080/14697680701483222.

François Longin and Bruno Solnik. Extreme correlation of international equity markets. *The Journal of Finance*, 56(2):649–676, 2001.

MSCI. Morgan Stanley Capital International Inc. `http://www.msci.com/equity/`, 2006.

Roger B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, New York, 1998.

Andrew J. Patton. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, 2(1):130–168, 2004. doi: 10.1093/jjfinec/nbh006.