Shape-based Scenario Generation using Copulas

Kaut, Michal and Wallace, Stein W.
Norwegian University of Science and Technology, Trondheim, Norway
and
Lancaster University Management School, Lancaster, England

in: Computational Management Science. See also BibTeX entry below.

BibTeX:

```
@article{KautWallace2011,
  author = {Kaut, Michal and Wallace, Stein W.},
  title = {Shape-based Scenario Generation using Copulas},
  journal = {Computational Management Science},
  year = {2011},
  volume = {8},
  pages = {181--199},
  number = {1--2},
  doi = {10.1007/s10287-009-0110-y}
}
```


Document created on March 17, 2011.
Cover page generated with CoverPage.sty.

Lancaster University Management School

NTNU
Norwegian University of Science and Technology

Shape-based Scenario Generation using Copulas

Michal Kaut* Stein W. Wallace†

September 2009

Abstract

The purpose of this article is to show how the multivariate structure (the "shape" of the distribution) can be separated from the marginal distributions when generating scenarios. To do this we use the copula. As a result, we can define combined approaches that capture shape with one method and handle margins with another. In some cases the combined approach is exact, in other cases, the result is an approximation. This new approach is particularly useful if the shape is somewhat peculiar, and substantially different from the standard normal elliptic shape. But it can also be used to obtain the shape of the normal but with margins from different distribution families, or normal margins with for example tail dependence in the multivariate structure. We provide an example from portfolio management. Only one-period problems are discussed.

Keywords stochastic programming scenario generation copulas

Introduction

Stochastic programming has become a common tool to study and model decision problems with the presence of uncertainty. These models are usually based on the use of multivariate probability distributions describing the uncertainty in the input data. The exact or approximating methods that are important for applications mainly deal with discrete empirical probability distributions that are described by a list of realizations (called scenarios) and related probabilities. See Wallace and Ziemba (2005) for a discussion of modelling as well as applications.

In most applications, the multivariate distributions do not come in a form suitable for the optimization model, being either continuous, discrete with too many data points, or specified by a set of statistical properties. Hence, to use a stochastic programming model, one has to transform the given distribution to scenarios—a process known as *scenario generation*. There exist many different scenario-generation methods, each with its strengths and weaknesses, see for example Dupačová, Gröwe-Kuska, and Römisch (2003), Heitsch and Römisch (2003, 2009), Høyland and Wallace (2001), Høyland, Kaut, and Wallace (2003), Pflug (2001). For an overview, see Dupačová, Consigli, and Wallace (2000).

In recent years, we have been studying—and using—scenario-generation methods that use the first four moments to describe the marginal distributions and the correlation matrix to describe the multivariate structure; see Høyland et al. (2003), Kaut, Wallace, Vladimirou, and Zenios (2007). While our experience shows that in many applications four moments provide a sufficient control over

*Norwegian University of Science and Technology, Trondheim, Norway; michal.kaut@iot.ntnu.no.
†Lancaster University Management School, Lancaster, England.

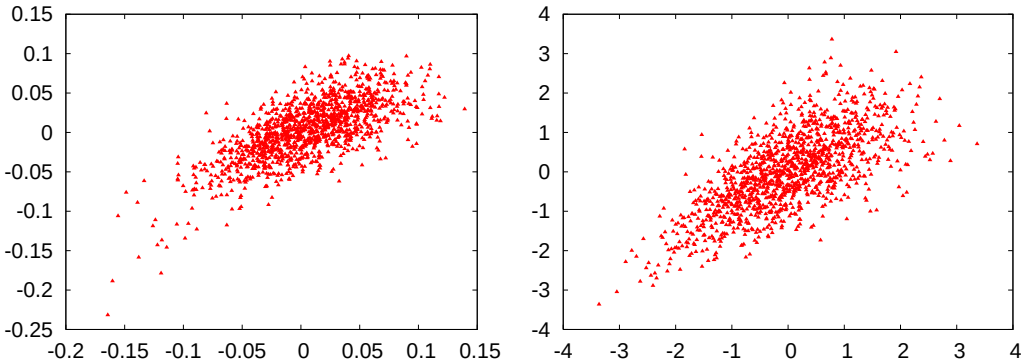

Figure 1: Scatter plot of fortnightly returns of US small caps vs. UK small caps. The left figure shows the actual data, the right figure the data with margins transformed to standard normal distribution, to demonstrate that the asymmetry is not caused by the marginal distributions.

the marginal distributions, the usefulness of correlations is much more limited. The reason is that a correlation—or more precisely the Pearson's correlation coefficient—describes only the degree of linear dependence between two random variables. It does not capture any non-linear dependencies, and it does not tell us anything about the "shape" of the multivariate structure. In a sense, using the Pearson correlation implicitly means assuming the elliptical shape of the normal distribution.

On the other hand, several recent studies—e.g. Hu (2006), Longin and Solnik (2001), Patton (2002, 2004)—point out that some financial data are not elliptical, showing for example higher correlations for downturns than for upturns (all markets tend to crash together). This is illustrated in Figure 1, which shows a scatter-plot of fortnightly returns of US and UK small cap stocks, using data from Morgan Stanley Capital International Inc. (MSCI). To demonstrate that the asymmetry does not come from the marginal distributions, we present also a plot of returns with margins transformed to the standard normal distribution.

Another example is the joint distribution of electricity prices and rainfall in countries with a significant proportion of hydro power (like Norway): in a dry year, the prices are almost guaranteed to be high, while in normal and wet years they can be both high and low, depending on other external factors. While we are not aware of studies from other areas, we find it likely that significantly non-elliptical structures can be found in many practical settings. Consider for example agricultural production in a given region: in "normal" years, the correlations between the productions of different crops can be expected to be small—or even negative, if the crops prefer different conditions. However, in a bad year (drought, flood) all the crops will fail, driving the correlations to one.

In this paper, we propose a general framework that can, at least in principle, generate scenarios with any multivariate structure. In addition, we propose several methods that fit into the framework and can be used in different cases. The framework is based on *copulas*, a concept that has been used in statistics and finance for some time—see for example Bouyé, Durrleman, Nikeghbali, Riboulet, and Roncalli (2000), Clemen and Reilly (1999), Rosenberg (2003)—yet remains virtually unknown in the rest of the OR community. To our knowledge, copulas have not yet been presented as a basis for scenario generation. We discuss only one-period problems in this paper, but some of the ideas can be used also in multi-period settings.

The rest of the paper is organized as follows: In the first section, we present the main results from the copula theory and discuss what it can offer for the scenario-generation problem. In the Section 2, we present the general framework, which is then tested in Section 3.

1 Copulas and their place in scenario generation

This section presents the notion of a copula and the main results from copula theory. In addition, it shows what this means for scenario generation.

1.1 Definitions and main results

The name *copula* was first used in Sklar (1959) to describe "a function that links a multidimensional distribution to its one-dimensional margins". The mathematical formulation comes from Sklar (1996) and Nelsen (1998).

An n-dimensional copula is the joint cumulative distribution function (cdf) of any n-dimensional random vector with standard uniform marginal distributions, i.e. a function $C : [0,1]^n \to [0,1]$. *Sklar's theorem* states that for any n-dimensional cdf F with marginal distribution functions $F_1, \ldots, F_n$, there exist a copula C such that

$$F(x_1, \ldots, x_n) = C\big(F_1(x_1), \ldots, F_n(x_n)\big)\,.$$

Moreover, if all the marginal cdfs F_i are continuous, then C is unique. For the proof, see Sklar (1996). An immediate consequence of the theorem is that, for every $\boldsymbol{u} = (u_1, \ldots, u_n) \in [0,1]^n$,

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

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

An important property of the copula is that it does not change under strictly increasing transformations of the margins. This allows us to transform margins from one continuous distribution to another, without changing the copula: if the margin $\tilde{X}_i$ has a cdf F_i, then $G_i^{-1}\big(F_i(\tilde{X}_i)\big)$ has cdf G_i, and the copula does not change since both F_i and G_i^{-1} are increasing.

This also means that any statistical property that depends only on the copula is invariant to strictly increasing transformations of the margins. An example of such a statistics is the Spearman's (rank) correlation—while the 'standard' Pearson's *linear* correlation is invariant only under positive linear transformations.

For the simplest example of a copula, consider two independent random variables $\tilde{X}_1$ and $\tilde{X}_2$ with $F(x_1, x_2) = F_1(x_1)F_2(x_2)$. The associated copula is then $C(u_1, u_2) = u_1 u_2$, i.e. the cdf of two independent standard uniform random variables.

Another example is the Gaussian copula, i.e. the copula of an n-variate standard normal distribution with correlation matrix Σ:

$$C_\Sigma(u_1, \ldots, u_n) = \Phi_\Sigma\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\right),$$

where Φ_Σ is the joint cdf of the multivariate normal distribution.

A special case, which will be needed later in the paper, is the so-called *empirical distribution*, i.e. a discrete distribution described by a matrix of equiprobable outcomes $\boldsymbol{X} = (x_{is})$. Its marginal cdfs are given by

$$F_i^{\mathrm{e}}(x) = \frac{\big|\{s : x_{is} \le x\}\big|}{n_S},$$

where $|A|$ denotes cardinality of a set and n_S is the number of scenarios (samples). Assuming that x_{is} are distinct in every margin $\boldsymbol{x}_i$, the cdfs evaluated at the sample points x_{is} are equal to

$$F_i^{\mathrm{e}}(x_{is}) = \frac{\mathrm{rank}(x_{is}, \boldsymbol{x}_i)}{n_S},$$

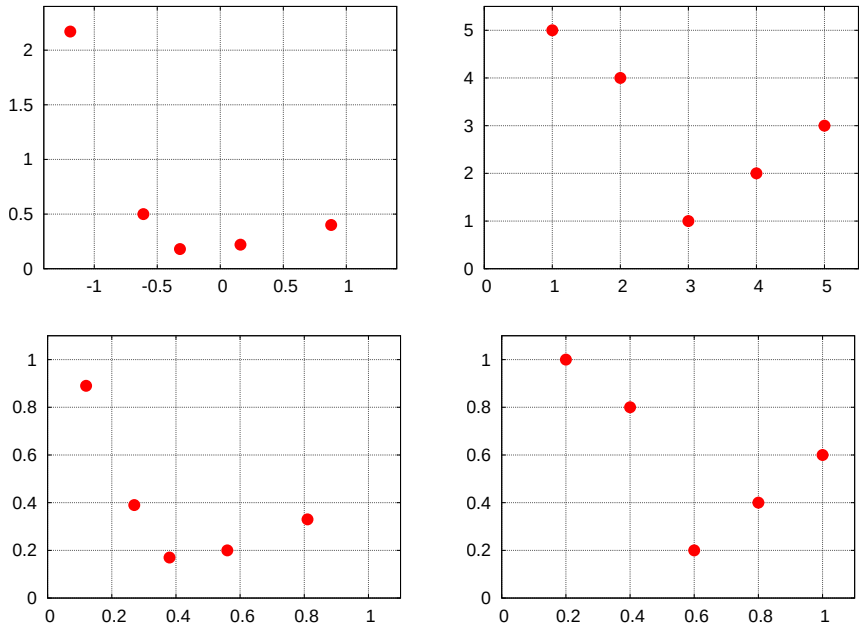

Figure 2: Empirical copula. The top-left figure presents five sample points from a distribution with margins $\tilde{X} \sim \mathrm{N}(0,1)$ and $\tilde{Y} \sim \exp(1)$ and the top-right figure the distribution of ranks of the sample. The bottom-left figure shows the data transformed to standard uniform margins by applying the marginal cdfs and finally the bottom-right figure shows the sample transformed using the empirical cdfs–the empirical copula. Note that the empirical copula is equivalent to the distribution of the ranks: the only difference between the two right-hand side figures is the scale of the axes.

where $\mathrm{rank}(x_s, \boldsymbol{x})$ is the rank (order) of value x_s in a vector $\boldsymbol{x}$ (also denoted $\mathrm{ord}(x_s, \boldsymbol{x})$), with values between 1 and n_S. Similar relation holds for the copula of the empirical distribution—which we refer to as an *empirical copula*—evaluated at the sample points:

$$\begin{aligned} C\left(\frac{k_1}{n_S}, \ldots, \frac{k_n}{n_S}\right) &= \frac{1}{n_S}\left|\left\{s : x_{is} \leq x_{(k_i)} \quad \forall i \in \{1, \ldots, n\}\right\}\right| \\ &= \frac{1}{n_S}\left|\left\{s : \mathrm{rank}(x_{is}) \leq k_i \quad \forall i \in \{1, \ldots, n\}\right\}\right| \end{aligned}$$

where $x_{(k)}$ is the k-th smallest element of vector $\boldsymbol{x}$. In other words, the empirical copula is uniquely described in terms of ranks of the original sample. This is further illustrated in Figure 2, where the empirical copula can be fully described as follows: create the bi-variate sample by pairing up the $x_{(1)}$ with $y_{(5)}$ (i.e. the smallest x_i with the biggest y_j), $x_{(2)}$ with $y_{(4)}$, $x_{(3)}$ with $y_{(1)}$, $x_{(4)}$ with $y_{(2)}$ and $x_{(5)}$ with $y_{(3)}$. We will use this way of describing an empirical copula later in Section 2.

For more information about copulas, see for example Clemen and Reilly (1999), Nelsen (1998), Sklar (1959, 1996). In addition, useful information can be found in User's Guide to the Statistics Toolbox for Matlab®.

1.2 Advantages of using copulas for scenario generation

Since the copula is obtained from the joint cdf by transforming the margins to the standard uniform distribution, it can be seen as the joint distribution stripped of all information about the margins. What

is left is information about the multivariate structure—none of this information is lost by transforming the margins. Copulas therefore allow us to de-couple the margins from the overall multivariate structure, and model these two independently.

This means, for example, that we can generate the margins using standard sampling and/or discretization methods for univariate distributions, giving us a degree of control that no general multivariate method (that we are aware of) can provide. Hence, even without any special copula-based tools, we can expect that sampling only the multivariate structure (shape) and using better tools for the marginal distributions will give better results than sampling directly from the multivariate distribution—an expectation that is confirmed by our numerical experiments later in the paper.

The de-coupling of the multivariate structure from the marginal distributions opens new possibilities for scenario generation, some of which are listed here:

Combining different (standard) copulas and margins

If we compare the normal distribution with t distributions with a small number of degrees of freedom ν, the most obvious difference is in the tails of the marginal distributions. There is, however, also one important difference between the two implied copulas, i.e. between the multivariate structures: the t distribution exhibits a *tail dependence*, defined as follows: A bivariate random vector $(\tilde{X}_1, \tilde{X}_2)$ with marginal cdfs F_1 and F_2 is lower-tail dependent if its lower-tail dependence coefficient

$$\lambda_L = \lim_{u \to 0^+} \mathrm{P}\left\{\tilde{X}_1 \leq F_1^{-1}(u) \,\middle|\, \tilde{X}_2 \leq F_2^{-1}(u)\right\}$$

is strictly positive (providing the limit exists). Upper-tail dependence λ_U is defined analogously and is equal to λ_L for all elliptical distributions. Figure 3 shows the tail-dependence for several t-distributions, computed using formulas from Embrechts, Lindskog, and McNeil (2003). As we can see, the t-distributions exhibit tail-dependence, while the normal distribution, i.e. the limit case $\nu \to \infty$, is tail-independent as long as the correlation is strictly smaller than one; for a formal proof, see Embrechts et al. (2003). In other words, if we draw a sample from a multivariate normal distribution and see that one of the variates is from the lower tail of its marginal distribution, the probability that another variate will also be in the tail is converging to zero as we move further "down" the tail. This implies that a simulation model using normal distributions is unlikely to produce scenarios with several margins with extreme values, even if we have enough samples to observe extreme values in each of the margins separately.

Using the definition of conditional probability, we can easily express the lower-tail dependence as (Joe, 1997, pp. 33)

$$\lambda_L = \lim_{u \to 0^+} \frac{\mathrm{P}\left\{\tilde{X}_1 \leq F_1^{-1}(u) \text{ and } \tilde{X}_2 \leq F_2^{-1}(u)\right\}}{\mathrm{P}\left\{\tilde{X}_2 \leq F_2^{-1}(u)\right\}} = \lim_{u \to 0^+} \frac{C(u,u)}{u}.$$

It follows that the tail dependence is a function of the copula and does not depend on the marginal distributions, so it is possible to create, for example, distributions with normal margins and t copula structure, i.e. normal margins with tail dependence. Note that the margins are not limited to normal distribution, each margin can even have a different type of distribution.

Introducing asymmetry

Instead of the standard t copula, we can use a copula from one of the skew-t distributions. These distributions allow for several types of asymmetric dependencies, the most important of which is the

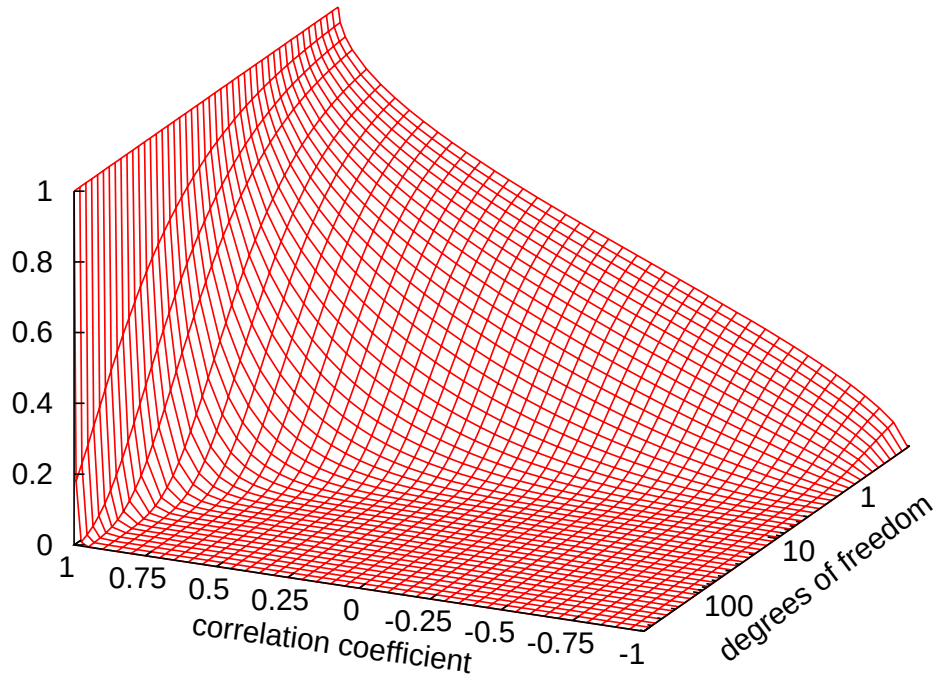

Figure 3: Tail-dependence coefficient λ of t-distribution, as a function of the correlation coefficient ρ and number of degrees of freedom ν. Note that $\lambda \to 0$ with increasing ν, as long as $\rho < 1$. Since the normal distribution is a limit case of t-distribution for $\nu \to \infty$, it implies that it has $\lambda = 0$ for all $\rho < 1$.

possibility to have higher correlation on the down-turn than on the up-turn—an effect we have already mentioned in connection with financial data.

Unfortunately, there are several different skewed versions of t distributions, each with different strengths and weaknesses. For information about the most important ones, see for example Adcock (2010), Azzalini and Capitanio (2003), Bauwens and Laurent (2005), Demarta and McNeil (2005), Jondeau and Rockinger (2003), Jones (2001). In addition, there is the non-central t distribution and Pearson Type IV distribution. For information on the latter, see Heinrich (2004).

Assuming that we are able to estimate the parameters for the chosen skew-t distribution, we can generate a sample from this distribution and then transform the margins, obtaining asymmetric dependency with arbitrary marginal distributions.

Using principal components

In many applications, it can be argued that there are too many random variables in the model, and the dimension could (and should) be reduced by techniques like principal components analysis (PCA). In addition to decreasing the dimension of the stochastic vector, the principal components are also uncorrelated—and therefore, *in the case of normal distributions*, independent. This means that scenarios for the individual principle components can be generated independently, converting the multivariate scenario generation to a much easier univariate generation problem. (The univariate margins can be combined into the multivariate vector in an all-against-all fashion, or by a random coupling of the margins. With the former, the number of scenarios grows exponentially with the dimension of the random vector, often resulting in a need to use a scenario-reduction procedure afterwards, while the latter yields scenarios that are only approximately independent.)

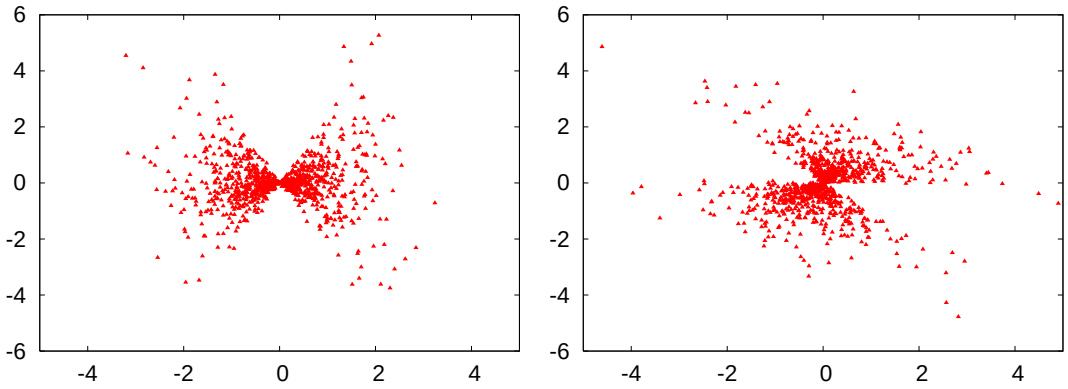

Figure 4: Bi-variate distribution with margins $\tilde{x}_1 = \tilde{\xi}_1,\ \tilde{x}_2 = \tilde{\xi}_1\tilde{\xi}_2$, with $\tilde{\xi}_1, \tilde{\xi}_2 \sim \mathrm{N}(0,1)$, independent. The left figure shows a sample from the random vector $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2)$, the right figure its principal components, scaled to variance equal to one. The principal components were computed from a sample of 25,000 points, but the plots show only the first 1000 points for better readability.

For other than normal distributions, however, the principal components are only uncorrelated, so there is still some dependency structure to be captured. This is illustrated in Figure 4, where the two principle components are clearly not independent, despite having zero correlations. Yet, as long as we use correlations as the only description of the multivariate structure, we are not able to make the distinction between uncorrelated and independent random variables and therefore cannot model the structure properly. It is therefore easy to forget the distinction between uncorrelated an independent.

Copulas, on the other hand, are capable of capturing the structure properly, allowing thus the use of principal components also for non-normal distributions. It is also possible that the distributions of principal components have qualitatively different structures than those of the underlying random variables, something that could be taken care of by the copula-based approach. This is, however, out of the scope of this paper and is left for future research.

1.3 Stability and Optimality gap

The ultimate test of the quality of a scenario tree will be how well it fits the corresponding stochastic program. This can be measured in terms of *optimality gap*, i.e. the difference in expected performances of the "true" optimal solution and the solution obtained from a stochastic program using the scenario tree in question. Unfortunately, this is usually impossible to measure exactly, because the true optimum is unknown. As a result, the quality of a scenario tree has to be estimated in some indirect way.

It is not the purpose of this paper to discuss that issue, but we would like to point out that there are several ways to estimate the quality of a scenario tree. Obvious possibilities are to compare the scenario tree directly to the underlying distribution, using metrics from probability theory—see for example Heitsch, Römisch, and Strugarek (2006)—or comparing (optimal) values of the relevant optimization problem, hence using the optimization problem as a metric. In the latter case, the performance of the scenario-based solutions can be evaluated using either a simulator—as in Kaut and Wallace (2007)—or a confidence interval on the optimality gap, obtained by solving several optimization problems—see Bayraksan and Morton (2006), Chiralaksanakul and Morton (2004), Linderoth, Shapiro, and Wright (2006).

2 The method

In this section, we present the scenario-generation method. The goal is to generate n_S samples from a given n-variate distribution, i.e. a matrix $\boldsymbol{X} \in \mathbb{R}^{n,n_S}$ of outcomes. We will call the empirical copula associated with the outcomes a *scenario copula*, to distinguish it from the empirical copula of the input data. We will use the fact that a discrete copula can be equivalently described as a coupling of ranks of the margins, as explained in Section 1.1.

2.1 Basic structure

The method consists of two parts, which can be done independently of each other:

1. Create the scenario copula, described in terms of the ranks of the margins. In other words, we want a set of n_S scenarios, each consisting of the ranks of values we want to use from each of the n margins. For example, one scenario may be "take the minimum of margin 1, third-smallest value of margin 2, etc". This can be done in several different ways:
   - *Sampling* from the true distribution (or its approximation) and computing the ranks of the values, one margin at a time. For each margin, we replace the values of the outcomes by their ranks inside the vector of all the outcomes for the given margin. This is the method used in tests in Section 3.
   - Using some *parametric family of copulas*, with parameters estimated from historical data. Since this is a more complex approach, we describe it in more detail in Section 2.3.
   - Using an optimization approach to *directly couple the ranks* in a way that minimizes some distance from the target distribution. One possible implementation could be a 'property-matching" NLP model along the lines of Høyland and Wallace (2001).
2. Generate the values of each margin. This is a standard and well-studied problem, so we only list the two approaches used in the tests in Section 3:
   - Using a prescribed discretization of the marginal distributions. For example, if we know the marginal cdfs F_i, we start with some discretization $\{u_1, \ldots, u_{n_S}\}$ of the standard uniform distribution and let $x_{is} = F_i^{-1}(u_s)$. In our case, we have used $u_s = \frac{2s-1}{2n_S}$, which is optimal in the Kolmogorov-Smirnov sense, as opposed to $u_s = \frac{s}{n_S+1}$ commonly used in the copula literature.
   - Compute marginal moments from the historical data and use a transformation-based moment-matching method to transform the scenarios to match the moments: for example, to match the first four moments we use the cubic transformation from Fleishman (1978), in the way described in Høyland et al. (2003).

Once we have both the structure (copula) and the values of the margins, the only thing that remains is to connect the margins in the way specified by the coupling of ranks. Note that this is easiest done if the margins are sorted.

2.2 Details and comments

Controlling the correlations/covariances

Controlling the structure using a copula means that we can only influence measures that depend on the copula, such as the Spearman's rank correlations. We cannot directly control the "standard" Pearson's

correlations, as they depend on both the copula and the margins.

If we need exact correlations, we can use the moment-matching algorithm from Høyland et al. (2003) as a post-process, setting the correlations to the desired values, while preserving (most of the) shape of the margins by controlling their first four moments. Since the process involves Cholesky transformation of the data, it will invariably distort the copula. The severity of the distortion will depend on the size of the errors to be corrected, small corrections should not change the structure noticeably.

Relation to the moment-matching algorithm by Høyland et al.

If we require control of moments and correlations, we can use the scenarios obtained by the copula-based method as a starting point for the moment-matching algorithm from Høyland et al. (2003). In the context of this paper, the algorithm can be seen as a method that takes a starting sample and transforms it to a sample with specified first four moments of the marginal distributions and a given correlation matrix. This is achieved by an iterative application of two transformations: a cubic transformation to correct the moments of the margins and a matrix transformation using a Cholesky component of the correlation matrix to correct the correlations.

In our implementation of the algorithm, the starting sample is a fixed discretization of $\mathrm{N}(0,1)$ distributions, but we also have the option to provide a starting sample externally. We can thus test whether the new copula-based approach provides a better starting point than the default sample. Note that the new starting sample can be expected to be better than the default sample both in terms of the copula and the marginal distributions. To test only the influence of a better copula, we can use the copula-based method to generate a starting sample with $\mathrm{N}(0,1)$, so the two starting samples differ only in the copula. Results of these tests are presented in Section 3.

An alternative interpretation of the method

There is an alternative (but equivalent) way of looking at the presented method in the case where we use the inverse cdfs to get the target distributions. Instead of representing the copula as the coupling of ranks (used to couple the pre-generated margins), we can transform the values from $\{1,2,\ldots,n_S\}$ to $\{\frac{1}{2n_S},\frac{3}{2n_S},\ldots,\frac{2n-1}{2n_S}\}$ to get a sample of the copula in the classical sense. Then we just apply the inversion method on the margins to get the target marginal distributions.

Instead of two independent parts plus a final assembly step, we now have a method with two steps: first generate scenarios for the copula and then transform the margins to their correct distributions. In this sense, the method is closer to the transformation-based algorithm from Høyland et al. (2003), with one important difference: the correction of margins does not change the copula, so there is no need for an iterative procedure.

In particular, if we use sampling to get the scenario copula, the method can be interpreted in the following way: sample one value x_{is} from the i-the marginal distribution and transform it to the standard uniform distribution using the discrete cdf F_i^{c} defined as

$$u_{is} = F_i^{\mathrm{c}}(x_{is}) = F_i^{\mathrm{e}}(x_{is}) - \frac{1}{2n_S},$$

where F_i^{e} is the standard empirical distribution function described in Section 1.1. The second term is due to the fact that F_i^{e} puts the probabilities at points $\frac{s}{n_S}$, while our discretization needs them at $\frac{2s-1}{2n_S} = \frac{s}{n_S} - \frac{1}{2n_S}$. The obtained u_{is} then becomes a sample from the i-th margin of the scenario copula,

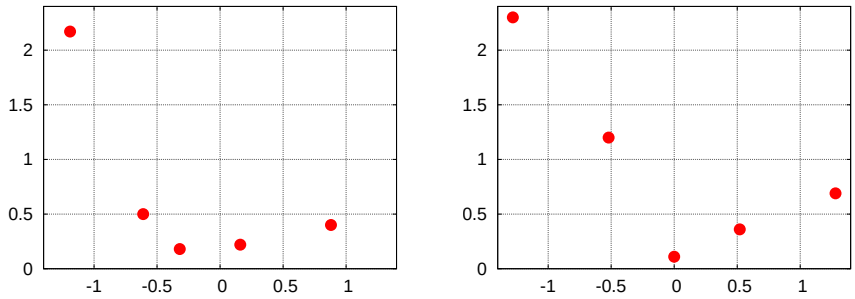

Figure 5: Standard sampling versus sampling the scenario copula and transforming the margins. The left figure is the same as the top-left plot in Figure 2 and shows the original sample. The right figure shows the scenario copula from the right-hand side of Figure 2, transformed back to the original distribution using the marginal cdfs. Note how the marginal values form a regular discretization of the respective marginal distributions, $\mathrm{N}(0,1)$ and $\exp(1)$—unlike the original sample, where the second component has too many small values.

so we can transform it back to the original distribution using the inverse cdf,

$$y_{is} = F^{-1}(u_{is}) = F^{-1}\left(F_i^{\mathrm{c}}(x_{is})\right) ,$$

to get a new sample from the i-th margin of the original distribution. We then repeat the procedure for all margins i and scenarios s.

Since F^{-1} is not an inverse of F_i^{c}, y's are different from x's. In other words, there is a difference between sampling the scenario copula and transforming its margins using the marginal cdfs, and doing standard sampling, as illustrated in Figure 5. The difference comes from the fact that F_i^{c} spreads the values evenly on the interval (0,1), so the inverse transformation results in a sample with values spread evenly in terms of percentiles.

2.3 Some information on copula families

One of the methods for generating the "scenario copula" mentioned in Section 2.1 involved using parametric families of copulas. Here, we present more details about both the copula families and how they fit in the presented scenario-generation framework. Readers who are not interested in this particular approach may proceed directly to the next section.

Copulas, just like distributions, have many parametric families with specialized methods for generation. Once we have decided for a particular copula, we have to estimate its parameters from the historical data and then use an appropriate method to create a sample from the copula. The best source of information on copula families is probably Nelsen (1998), other options include Bouyé et al. (2000), Hu (2006), Romano (2002).

In addition to the copula families, it is possible to use copulas from some standard distribution like normal or t, or the skewed versions of t distributions mentioned in Section 1.2. In this case, we generate a sample from the given distribution and then transform it to a copula in the same way as in the previous section.

Note that the transformation to copula removes all information of the marginal distributions, so only the copula (structure) of the chosen distribution remains. This means, for example, that we do not have to estimate the scale parameters, as they do not influence the copula. In other words, the normal copula depends only on the correlations, the t copula in addition on the degrees of freedom,

and the skewed version of t in addition on the skewness parameter(s). Furthermore, the skewness parameter(s) are used only to control the asymmetry of the skewed-t copulas, they have no relation to the skewness of the final distribution (again, because the marginal distributions are removed by the transformation to copula). This is illustrated in Figure 6 where the distribution remains skewed even when the margins are transformed to the standard normal distributions. For comparison, we present also a distribution obtained by combining the skewed-t margins with a standard normal copula. Note that unlike the skewed-t distributions, the extreme values do not happen together when we use the same margins with the normal copula. This is in concordance with the fact that the normal copula does not exhibit tail dependence, as mentioned in Section 1.2. For more information on using copulas of standard distributions, see for example Demarta and McNeil (2005), Romano (2002).

3 Case study – portfolio optimization with CVaR constraint

In this section, we test several variants of the scenario-generation method on a portfolio optimization model with a CVaR constraint. It is a one-period LP model, with positive variables (positions) that sum up to one. The LP formulation of the CVaR constraint comes from Rockafellar and Uryasev (2000) and Uryasev (2000).

Notation

$\boldsymbol{I}$	The set of financial instruments.
P^s	Probability of scenario $s \in \{1,\ldots,n_S\}$.
R_i^s	Return of asset i in scenario s.
β	Confidence level for CVaR; in our case, $\beta = 0.95$.
C	Minimal feasible value of CVaR.
x_i	Decision variables – proportion invested in instrument $i \in \boldsymbol{I}$.
α	Auxiliary variable; equal to VaR at the optimal solution.
z^s	Variables used for modelling CVaR; $z^s \geq 0$.

Using the above notation, the model can be formulated as

$$\max_{x_i} \sum_s P^s \sum_i R_i^s x_i \,,$$

subject to

$$\sum_i x_i = 1 \,, \tag{1}$$

$$z^s + \sum_i R_i^s x_i \geq \alpha \,, \tag{2}$$

$$\alpha - \frac{1}{1-\beta} \sum_s P^s z^s \geq C \,, \tag{3}$$

where (1) is a budget constraint and (2) and (3) are the CVaR-defining constraints from Rockafellar and Uryasev (2000) and Uryasev (2000).

The CVaR model has been chosen because it can be expected to react to differences in the shape of the distribution, particularly the shape of the lower tail of the return distribution. Two sets of data were used for the model: the main data set consists of daily prices of seven stock indices and three government bonds, from 1987-07-09 to 2005-04-05 (4476 points). This data set was kindly provided

by Kjetil Høyland from DNB Nor, Oslo, Norway. The second data set consists of 1302 daily prices of 10 stock indices, obtained from MSCI.

The price differentials (asset returns) from the historical data are used as the true distribution for the optimization problem, i.e. the distribution that the scenarios are supposed to be samples from. This implies that we can compute the true value of any statistical property of the return distribution. Hence, when we in the following text say that we "correct", for example, the mean of the scenarios, we mean that we transform the scenario distribution to match the mean of the historical returns. Naturally, if we did not know the correct values of these properties, the corrections would not be possible.

We have used sampling from the historical data to get the scenario copula in Step 1 of the algorithm presented in Section 2, and the fixed discretization of margins in Step 2. The method was compared to sampling, (both with and without correction of means and variances) and the moment-matching algorithm from Høyland et al. (2003). We have also tested the effect of correcting correlations—together with the first four moments—by running the moment-matching algorithm as a post-process. In addition, we have tested whether the moment-matching method benefits from starting from the scenario copula, in the way described in Section 2.2. Combining the different methods gave us thirteen different scenario-generation methods to test.

3.1 The tests

Since we model CVaR as a constraint, our objective function consists only of the expected return. The in- and out-of-sample objectives are therefore equal as long as scenarios have the correct means—which in our case they have, except in the case of direct sampling from historical data. In addition, the CVaR constraint is, almost by definition, active, so the in-sample CVaR values are on their bounds. As a result, we measure the performance of a solution by the difference between the in- and out-of-sample values of CVaR, instead of the objective values. In particular, we call a scenario-generation method *biased* if the tree it produces leads to a consistent difference between the in- and out-of-sample CVaR values. For example, the middle chart in the third row of Figure 7 shows a result of a biased scenario-generation method.

For the main data set, we have tested stability with two different values of the CVaR lower-bound C, one close to the minimum-risk value, and one more risky. For the MSCI data only one CVaR value was tested. Three different sizes of scenario trees were used in each case: 50, 250, and 1000 scenarios. In each case, one hundred scenario trees were generated, the model solved on them, and the solution evaluated on the reference tree consisting of the whole data set. We could thus perform both the in- and out-of-sample tests as described in Kaut and Wallace (2007), focusing on stability and bias. In addition, since it was possible to solve the model on the reference tree, we were able to obtain the "true" optimal solution and thus compute the optimality gap caused by the scenarios.

The CVaR model was written in the GNU MathProg language (a subset of AMPL) and solved by `glpsol`, both parts of GNU Linear Programming Kit (GLPK). The other tests were implemented in GNU Octave, a high-level language mostly compatible with Matlab. Finally, Gnuplot was used to produce the charts to visualize the results of the simulations.

3.2 The main result

Out of the thirteen tested combinations of the copula-based methods and post-processes, the one that performed consistently best in terms of both stability and bias was a combination of the copula-based method using inverse cdfs and a post-process correcting both the marginal moments and correlations.

Table 1: Out-of-sample stability for the results presented in Figure 7. The first part of the table shows the standard deviation of the objective function values, multiplied by 10 000, the second part shows the standard deviation of CVaR values, multiplied by 100.

	std. dev. of obj. ($\times$10000)			std. dev. of CVaR ($\times$100)		
n_S	sampl.	mom.	copula	sampl.	mom.	copula
50	8.72	5.66	4.89	1.81	1.33	1.41
250	3.49	2.74	2.01	0.73	0.52	0.42
1000	1.75	1.31	0.83	0.38	0.24	0.18

Table 2: Out-of-sample CVaR and its bias for the results presented in Figure 7. The first part of the table shows the average out-of-sample CVaR values, the second part shows the bias of CVaR, i.e. the difference between the average CVaR and its target value $C = -0.2$. Numbers in the second part are multiplied by 100 for better readability.

	average CVaR			bias of CVaR ($\times$100)		
n_S	sampl.	mom.	copula	sampl.	mom.	copula
50	-0.218	-0.224	-0.226	-0.83	-2.36	-2.63
250	-0.203	-0.206	-0.201	-0.33	0.40	-0.14
1000	-0.201	-0.190	-0.200	-0.09	0.97	-0.03

3.3 Other observations

It is not possible to present the results of all the tests, so we present results only for the most "important" methods: sampling with correction of means and variances, moment-matching, and the best copula-based method. The results of tests on trees with 50, 250, and 1000 scenarios, based on the main data set, are presented graphically in Figure 7 and numerically in Tables 1 and 2. Both the figure and tables can illustrate most of the following observations—even if the observations themselves are based on results of all the tests (all methods, all data sets, all sizes of scenario trees):

- For trees with 250 and 1000 scenarios, the copula-based method outperforms the corrected sampling and the moment-matching approach both in terms of stability and bias of CVaR. With 50 scenarios, however, it leads to a bias in CVaR that is comparable to the one of moment matching. This is probably due to the fact that with only 50 scenarios, the sample produced by the method itself has correlations that are significantly different from the target, so the correlation-correcting post-process distorts the copula too much.

- As expected, pure sampling of the historical returns performs poorly, though it can be improved significantly just by correcting the means and variances of the margins. Correcting moments and correlations further improves the performance of the sampled trees.

- The moment-matching method with default starting point leads to one of the most stable methods, but can introduce a bias. This is due to the fact that with no extra information, the method will use normal distributions as its starting point and thus generate scenarios with structure close to the normal copula. When the data has significantly different structure, this approach leads to a bias in the results. In Table 2, as well as in the last row of plots in Figure 7, we can see that the moment-matching for 1000 scenarios led to smaller risk than required. However, in the case of

CVaR constraint at $C = -0.25$ (instead of $C = -0.2$), the moment-matching resulted in a portfolio with a higher-than-required risk. This illustrates that the bias caused by moment-matching is unpredictable, including the sign of the bias.

- Using the sample copula and $\mathrm{N}(0,1)$ margins as a starting point for the moment-matching procedure decreased the bias compared to using the procedure on its own. This was to be expected, as it gives the moment-matching algorithm a better approximation of the true copula – see the previous point.

  Note, however, that this method was used only for the purpose of testing the effect of a better copula on the moment-matching procedure. Other than that, it makes little sense using normal margins, as long as we have a better approximation of the marginal distributions. (Even if we use the moment-matching as a post-process, it is beneficial to start as close to the true marginal distributions as possible, since the cubic transformation used to correct the margins may not be strictly increasing and can thus distort the copula.)

- As expected, using a post-process to control some statistical properties improves the stability of the optimization problem. Furthermore, the improvements increase with the number of controlled properties, i.e. controlling four moments is better than controlling just means and variances and can be further improved by controlling correlations as well (using the alg. from Høyland et al. (2003)).

Conclusions

In this article, we have shown how to separate marginal distributions from the multivariate structure—the copula—when generating scenarios. This way we can combine different approaches which, separately, may be good (or even applicable) for only one of these factors. By this separation of margins and copula it is, for example, possible to sample from the underlying distribution to obtain an approximation of the structure, while not having to rely on the same sample for margins. The margins can then be set up with methods that are better suited to handle them, but which are possibly even unable to handle multivariate structure. Our example from portfolio management indicates that such an approach is indeed a good idea: for the example at hand, the best approach was to combine sampling from the copula with a post-process correcting the marginal moments and correlations.

Acknowledgements We would like to thank Pavel Popela and Pavla Zemánková from Brno University of Technology, Brno, Czechia, for their help with the initial draft of the paper. The authors have been supported under grants no. 156315/530 and 171007/V30 from The Research Council of Norway. Furthermore, Michal Kaut has been supported by the project no. 103/05/0292 from the Grant Agency of the Czech Republic.

References

C. J. Adcock. Asset pricing and portfolio selection based on the multivariate extended skew-student-t distribution. *Annals of Operations Research*, 176(1):221–234, 2010. doi: 10.1007/s10479-009-0586-4.

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.

Luc Bauwens and Sébastien Laurent. A new class of multivariate skew densities, with application to GARCH models. *Journal of Business and Economic Statistics*, 23(3):346–354, 2005. Available at SSRN: `http://ssrn.com/abstract=691865`.

Güzin Bayraksan and David P. Morton. Assessing solution quality in stochastic programs. *Mathematical Programming*, 108(2–3):495–514, sep 2006. doi: 10.1007/s10107-006-0720-x.

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`.

Anukal Chiralaksanakul and David P. Morton. Assessing policy quality in multi-stage stochastic programs. Stochastic Programming E-Print Series, `http://www.speps.org`, 2004.

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

S. Demarta and A. J. McNeil. The *t* copula and related copulas. *International Statistical Review*, 73 (1):111–129, 2005.

Jitka Dupačová, Giorgio Consigli, and Stein W. Wallace. Scenarios for multistage stochastic programs. *Ann. Oper Res.*, 100(1–4):25–53, 2000. ISSN 0254-5330. doi: 10.1023/A:1019206915174. Research in stochastic programming (Vancouver, BC, 1998).

Jitka Dupačová, Nicole Gröwe-Kuska, and Werner Römisch. Scenario reduction in stochastic programming. *Mathematical Programming*, 95(3):493–511, 2003. doi: 10.1007/s10107-002-0331-0.

John W. Eaton. *GNU Octave Manual*. Free Software Foundation, Inc., 2006.

Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas and applications to risk management. In Svetlozar T. Rachev, editor, *Handbook of Heavy Tailed Distributions in Finance*, Handbooks in Finance, chapter 8, pages 329–384. Elsevier, 2003.

A. I. Fleishman. A method for simulating nonnormal distributions. *Psychometrika*, 43:521–532, 1978. doi: 10.1007/BF02293811.

Joel Heinrich. A guide to the pearson type IV distribution. Technical Report Memo 6820, The Collider Detector at Fermilab, Fermilab, Batavia, Illinois, 2004. Available at `http://www-cdf.fnal.gov/publications/cdf6820_pearson4.pdf`.

H. Heitsch and W. Römisch. Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications*, 24(2–3):187–206, 2003. doi: 10.1023/A:1021805924152.

H. Heitsch and W. Römisch. Scenario tree modelling for multistage stochastic programs. *Mathematical Programming*, 118(2):371–406, 2009. doi: 10.1007/s10107-007-0197-2.

H. Heitsch, W. Römisch, and C. Strugarek. Stability of multistage stochastic programs. *SIAM Journal on Optimization*, 17(2):511–525, 2006. doi: 10.1137/050632865.

K. Høyland and S. W. Wallace. Generating scenario trees for multistage decision problems. *Management Science*, 47(2):295–307, 2001. doi: 10.1287/mnsc.47.2.295.9834.

K. Høyland, M. Kaut, and S.W. Wallace. A heuristic for moment-matching scenario generation. *Computational Optimization and Applications*, 24(2–3):169–185, 2003.

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

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, London, 1997.

Eric Jondeau and Michael Rockinger. Conditional volatility, skewness, and kurtosis: Existence, persistence, and comovements. *J. Econ. Dynam. Control*, 27(10):1699–1737, 2003. ISSN 0165-1889. doi: 10.1016/S0165-1889(02)00079-9.

M. C. Jones. Multivariate t and beta distributions associated with the multivariate f distribution. *Metrika*, 54(3):215–31, 2001. doi: 10.1007/s184-002-8365-4.

Michal Kaut and Stein W. Wallace. Evaluation of scenario-generation methods for stochastic programming. *Pacific Journal of Optimization*, 3(2):257–271, 2007.

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.

Jeff T. Linderoth, Alexander Shapiro, and Stephen J. Wright. The empirical behavior of sampling methods for stochastic programming. *Ann. Oper. Res.*, 142(1):215–241, 2006.

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

Andrew Makhorin. *GNU Linear Programming Kit – Reference Manual, Version 4.9*. Free Software Foundation, Inc., 2006a.

Andrew Makhorin. *GNU Linear Programming Kit – Modeling Language GNU MathProg, Version 4.9*. Free Software Foundation, Inc., 2006b.

MathWorks. *Statistics Toolbox For Use with MATLAB® – User's Guide*. The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098, 2006.

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. Skewness, asymmetric dependence, and portfolios. In *Applications of Copula Theory in Financial Econometrics*, Ph.D. dissertation 3. Department of Economics, University of California, San Diego, 2002.

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.

G. C. Pflug. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming*, 89(2):251–271, 2001. doi: 10.1007/PL00011398.

R. Tyrrell Rockafellar and Stan Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

Claudio Romano. Calibrating and simulating copula functions: An application to the italian stock market. working paper 12, Centro Interdipartimale sul Diritto e l'Economia dei Mercati, 2002.

Joshua V. Rosenberg. Non-parametric pricing of multivariate contingent claims. *The Journal of Derivatives*, 10(3):9–26, 2003.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Abe Sklar. Random variables, distribution functions, and copulas – a personal look backward and forward. In L. Rüschendorff, B. Schweizer, and M. Taylor, editors, *Distributions with Fixed Marginals and Related Topics*, volume 28 of *Lecture Notes – Monograph*, pages 1–14. Institute of Mathematical Statistics, Hayward, CA, 1996. URL `http://www.jstor.org/stable/4355880`.

Stan Uryasev. Conditional value-at-risk: Optimization algorithms and applications. *Financial Engineering News*, 14:1–5, February 2000. doi: 10.1109/CIFER.2000.844598.

S.W. Wallace and W.T. Ziemba, editors. *Applications of Stochastic Programming*. MPS-SIAM Series on Optimization, Philadelphia, 2005.

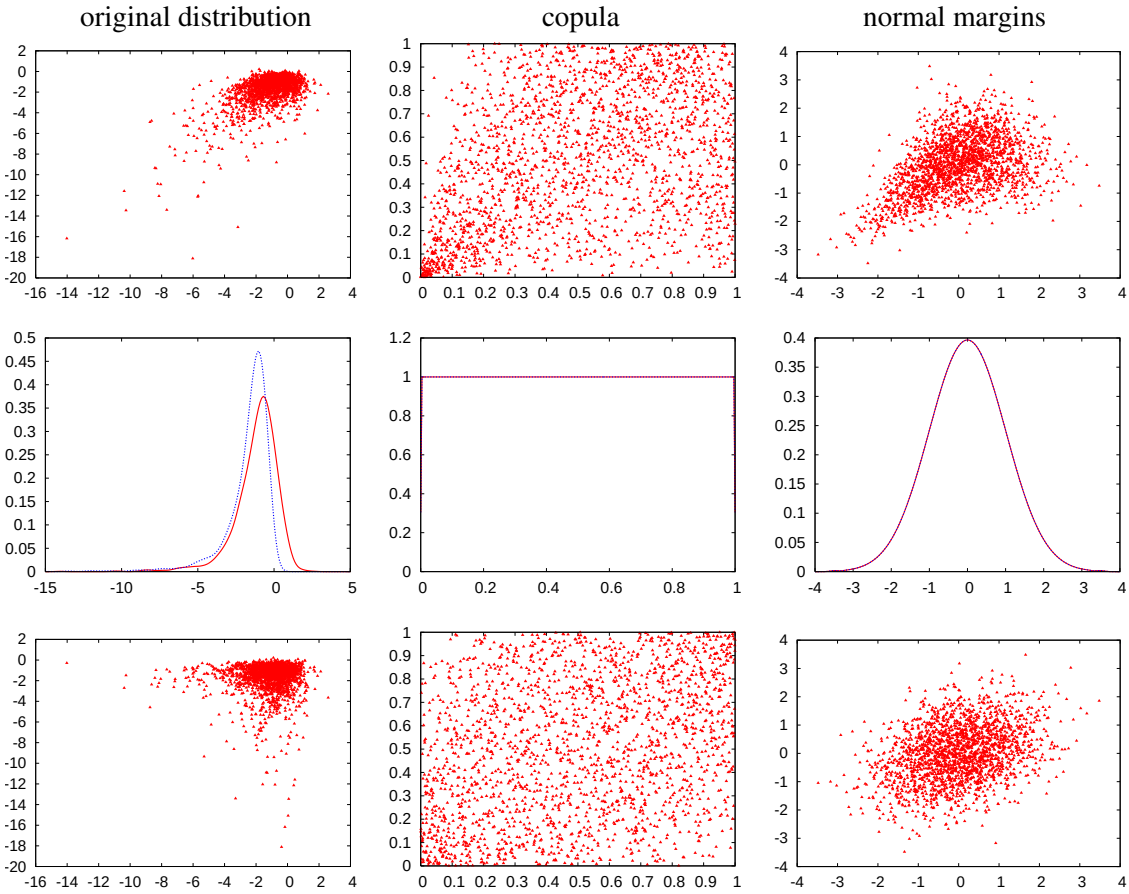

Figure 6: Skewed-t distribution and copula, using a skewed-t variant from Azzalini and Capitanio (2003) with 5 degrees of freedom and skewness parameters $(-0.5, -0.9)$. The first two rows show the two-dimensional scatter plots and marginal densities, respectively. The third row shows a distribution obtained by combining the skewed-t margins with a standard normal copula. In other words, the marginal distributions in the second row correspond both to the first and the third row. Note that the distribution shown in the bottom-right figure is a standard normal distribution. The reason there seems to be only one line in the second and third figure in the second row is that in those cases both margins have the same distributions, $\mathrm{U}(0,1)$ and $\mathrm{N}(0,1)$, respectively.

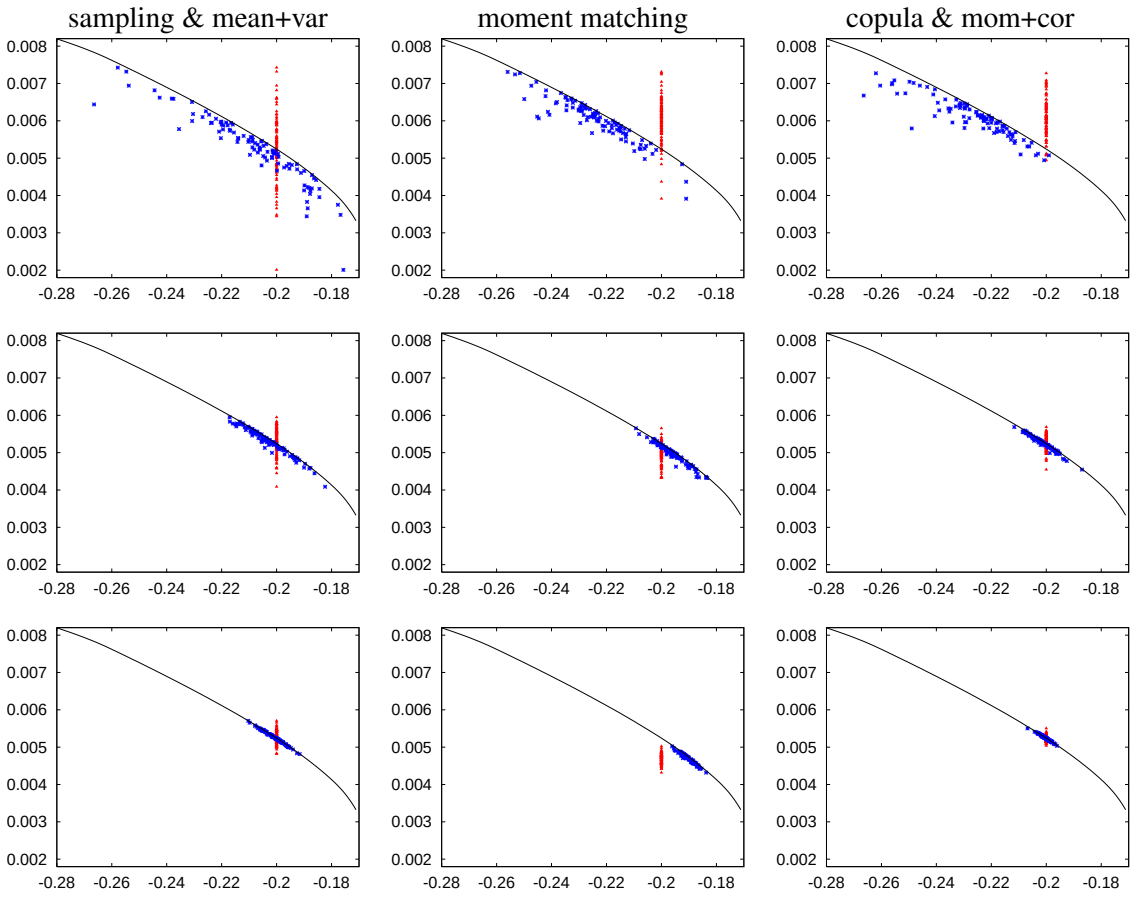

Figure 7: In-sample and out-of-sample properties of three selected scenario generation methods, on trees with 50, 250, and 1000 scenarios, based on the main data set. On the x-axis is CVaR, on the y-axis the objective function values. The in-sample values are scattered along a vertical line CVaR $= C = -0.2$, caused by the constraint on CVaR. The rest of the points represent the out-of-sample values and the line represents the "true" CVaR-efficient frontier. Note that the in-sample values can be above the efficient frontier, since they do not represent the true objective values of solutions.